

# **Mathematical Classification of the Modes of Tumour Evolution**

Veselin Manojlović

Doctor of Philosophy



School of Science and Technology  
Department of Mathematics

September 2023



# Contents

<b>Contents</b>	<b>iii</b>
<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>xi</b>
<b>Acknowledgements</b>	<b>xiii</b>
<b>Declaration</b>	<b>xv</b>
<b>Abstract</b>	<b>xvii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Trees and their applications . . . . .	2
1.1.1 Introduction . . . . .	2
1.1.2 Quantifying tree balance . . . . .	3
1.2 Agent-based modelling in oncology . . . . .	4
1.2.1 Introduction . . . . .	4
1.2.2 The <code>demon-warlock</code> framework . . . . .	6
1.3 Likelihood-free inference . . . . .	9
1.3.1 Introduction . . . . .	9
1.3.2 Approximate Bayesian Computation . . . . .	10
1.4 Fluctuating methylation clocks . . . . .	11
1.5 Aims and objectives . . . . .	12
1.5.1 Hypotheses . . . . .	12
1.5.2 Aims . . . . .	12
1.5.3 Objectives . . . . .	12

<b>2 Extreme and expected values of universal tree balance index <math>J^1</math></b>	<b>15</b>
2.1 Introduction . . . . .	15
2.2 Definitions . . . . .	16
2.3 Balancing binary trees according to $J^1$ . . . . .	18
2.4 Expected value of $J^1$ under simple evolutionary processes . . . . .	22
2.5 Analytic properties of the $J^1$ index . . . . .	25
2.5.1 Properties of $J^1$ on different tree families . . . . .	26
2.6 Discussion . . . . .	31
<b>3 Tracking cancer evolution <i>in silico</i> via evolutionary indices</b>	<b>33</b>
3.1 Introduction . . . . .	33
3.1.1 Why even bother with indices? . . . . .	34
3.1.2 A 3-dimensional index space — trees with uniform branch lengths . . . . .	34
3.1.3 A general set of indices — any rooted tree . . . . .	35
3.2 Tree resolution . . . . .	36
3.2.1 3-dimensional index space . . . . .	36
3.3 Computational methods . . . . .	37
3.3.1 Agent-based modelling framework - <i>warlock/demon</i> . . . . .	37
3.4 Results . . . . .	38
3.4.1 Sensitivity of evolutionary mode to parameter values . . . . .	38
3.5 Discussion . . . . .	38
<b>4 Agent-based workflow for inferring evolutionary parameters from molecular data using approximate Bayesian computation</b>	<b>49</b>
4.1 Introduction . . . . .	49
4.2 Colorectal adenocarcinoma . . . . .	50
4.3 Modifying existing simulation workflows . . . . .	50
4.4 Simulating fluctuating methylation arrays . . . . .	52
4.4.1 Overview . . . . .	52
4.4.2 Examples . . . . .	53
4.5 Fluctuating methylation arrays through the lens of ABC . . . . .	53
4.5.1 Overview . . . . .	53
4.5.2 Examples . . . . .	53

<b>5 Inferring evolutionary parameters of colorectal cancer from DNA methylation arrays</b>	<b>55</b>
5.1 Introduction . . . . .	55
5.2 Results . . . . .	55
5.2.1 A note on the fully neutral model . . . . .	56
5.2.2 Selective advantage . . . . .	56
5.2.3 Driver mutation rates . . . . .	60
5.2.4 fCpG flipping probabilities . . . . .	62
5.2.5 Gland fission rates . . . . .	65
5.3 Parameters . . . . .	67
5.4 Distance functions . . . . .	67
5.5 Next steps/work in progress . . . . .	71
5.5.1 Hypotheses . . . . .	71
<b>6 Discussion</b>	<b>73</b>
<b>A Title of the First Appendix</b>	<b>75</b>
<b>Appendix B</b>	<b>77</b>
<b>Bibliography</b>	<b>79</b>



# List of Figures

1.1	<b>a)</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.	4
	<b>b)</b> $J^1$ values for trees generated under the Yule process and the uniform model. Solid curves represent the approximate expected values, and the dashed lines the 5th and 95th percentiles.	4
1.2	Event hierarchy in the <b>demon-warlock</b> framework. Figure reproduced from (Bak et al. n.d.) with the authors' permission.	7
1.3	<b>a</b> ) Example output from <b>demon</b> , visualised using the <b>demonanalysis</b> package. <b>a</b> Muller plot of clonal dynamics over time. Each colour represents a clone with a distinct combination of driver mutations. <b>b</b> Final proportions and spatial plot of clones. <b>c</b> Fish plot of clone populations over time using the same colours as in <b>a</b> . <b>d</b> Muller plot showing evolution of tumour cell division rate. <b>e</b> Final spatial distribution of cell division rates. <b>f</b> Final spatial distribution of the mean numbers of passenger mutations per cell. <b>g</b> Final spatial distribution of the number of tumour cells per gland.	8
2.1	Comparison of probabilities for generation of trees on 4 leaves under the Yule and uniform models.	18
2.2	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.	22

2.3 <b>Top row:</b> 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 $n \log_2 n / \mathbb{E}(I_S)$ . <b>A</b> — uniform model, <b>B</b> — Yule model. <b>Bottom row:</b> The Jensen gap of $\mathbb{E}(J^1)$ calculated for trees up to 128 leaves under the uniform model ( <b>C</b> ), and the Yule model ( <b>D</b> ). 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 tables ?? and ?? for numerical values of the gap size for the first several values of $n$ .	26
2.4 If we limit our search to leafy trees, the least balanced tree on a given number of leaves is not necessarily the caterpillar. Pictured are the caterpillar trees on 4, 6, and 9 leaves, as well as minimally balanced brooms for 6 and 9 leaves, with corresponding $J^1$ values.	27
2.5 The labels used in the figures are as above - $n$ for number of leaves, $k$ for number of leaves in the broom head, $r = n/k$ . <b>A:</b> 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 $a = 0, 1, 2, \dots$ lie on the dashed grey lines. <b>B:</b> Behaviour of caterpillar and broom for different values of $n$ . <b>C:</b> $J^1$ for different broom trees on a given number of leaves using equation (2.24). The dashed lines indicate the value of $r$ for which $J^1$ is minimal.	30
3.1 Average trajectories of the three indices for different values of driver mutation rate and selective advantage for tumours progressing via boundary growth.	39
3.2 Average trajectories of the three indices for different values of driver mutation rate and selective advantage for well-mixed cancer cell populations.	40
3.3 Average trajectories of the three indices for different values of driver mutation rate and selective advantage for gland fission.	41
3.4 Average trajectories of the three indices for different values of driver mutation rate and selective advantage for invasive glandular tumours.	42

3.5	Average trajectories in index space for tumours progressing via boundary growth. . . . .	43
3.6	Average trajectories in index space for well-mixed cancer cell populations. . . . .	44
3.7	Average trajectories in index space for tumours progressing via gland fission. . . . .	45
3.8	Average trajectories in index space for invasive glandular tumours. .	46
4.1	The old workflow for simulating tumour growth and assigning methylation arrays. . . . .	51
4.2	The old workflow for simulating tumour growth and assigning methylation arrays. . . . .	52
5.1	correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.1$ . . . . .	57
5.2	correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.2$ . . . . .	58
5.3	correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.3$ . . . . .	59
5.4	correlation scatter plots, gland histograms and correlation heatmaps for driver mutation rate $10^{-6}$ . . . . .	60
5.5	correlation scatter plots, gland histograms and correlation heatmaps for driver mutation rate $10^{-4}$ . . . . .	61
5.6	correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities $10^{-4}$ . . . . .	62
5.7	correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities $10^{-3}$ . . . . .	63
5.8	correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities $10^{-2}$ . . . . .	64
5.9	correlation scatter plots, gland histograms and correlation heatmaps for fission rates 0.008. . . . .	65
5.10	correlation scatter plots, gland histograms and correlation heatmaps for fission rates 0.08. . . . .	66
5.11	Distances between glands from figure 5.10. . . . .	68
5.12	Distances between glands from data set J. . . . .	69

5.13 Data set J whose distances are shown in 5.12 . . . . . 70

# List of Tables

5.1 Default parameter values. . . . .	67
---------------------------------------	----



# Acknowledgements

I want to express my most sincere gratitude and recognition to all those anonymous persons around the world who generously share their knowledge, giving their time and effort to the community without expecting any reward. Without them, I would not have been able to develop this template and many other projects.



# Declaration

The present work is intended to be a base for a City University PhD dissertation latex template. Although it is ready to be used as it is, obtaining a suitable PhD dissertation which will fulfil the university requirements, many improvements can be done on it. All the users of this template are encouraged to share their modifications and improvements so a better template can be developed collaboratively.

In order to protect the freedom of this work and to guarantee the students' right to access, use, modification and distribution, this template is released under the terms and conditions of the GNU General Public License GPLv3. This license grants the user four key freedoms:

- 1 The freedom to use the software for any purpose.
- 2 The freedom to change the software to suit her/his needs.
- 3 The freedom to share the software with anyone else.
- 4 The freedom to share the changes he/she makes.

A copy of the terms and conditions of this license can be either found in the attached file License.pdf or accessed from the GNU Operating System web page:  
<http://www.gnu.org/licenses/gpl-3.0>.





# Abstract

In this work, the City University's Senate Regulation 25 (?) has been followed in order to obtain a L<sup>A</sup>T<sub>E</sub>X template providing the adequate format for a City University PhD dissertation.



# 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 (?). Despite advances in cancer research, the survival rates for many cancers remain low, with the disease being an increasing burden on healthcare systems (*Financial Burden of Cancer Care | Cancer Trends Progress Report* n.d.). 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, we consider different approaches to modelling cancer evolution via driver mutations, which includes the use of phylogenetic trees and agent-based models. Further, we employ methylation data to verify the accuracy of our models using Approximate Bayesian Computation (ABC).

Trees as a mathematical object have found use in a variety of fields, of which biology is our main focus. However, while writing this thesis, we have found trees interesting links to methods in computer science via information theory. In chapter 2, we expand upon three points. First, we further establish  $J^1$  as a universal index of tree balance by identifying fundamental connections to classical results in computer science, related to Huffman coding and self-balancing tree data structures. Second, we derive upper bounds on the error of the expected value approximations for the Yule process and the uniform models. For the Yule process, we prove that the approximation rapidly converges to the true expectation in the large tree limit.

Finally, we investigate the minimal values of  $J^1$  in important special cases, obtaining a counter intuitive result in the large tree limit.

In chapter 3, we 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 at tree shape indices. These results are compared to a new, more comprehensive system of tree shape indices (Noble & Verity n.d.) 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, we introduce a proprietary model for simulating a specific type of molecular data, methylation arrays, obtained from multi-site sequencing of colorectal cancer. We 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. We further explore how the model can be expanded for more general use due to its modular design. We 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, we use the new agent-based modelling framework to infer the evolutionary history of 10 colorectal cancer samples. Additionally, using tree shape indices, we compare trees of the inferred branching process in the data to the trees generated in the simulation. This provides us an additional tool for validating the model and the inference process. We also explore the potential of using the model for predicting the future evolution of tumours, and discuss the limitations of the workflow for our specific use case.

## 1.1 Trees and their applications

### 1.1.1 Introduction

In the most general sense, a tree is a connected graph with no cycles. In this thesis, when a tree is mentioned, we refer to a rooted tree, as formally defined in section 2.2. 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 n.d.) or the structure of a program's syntax (Knuth n.d.*b*), an approach that computer scientists share with linguists (Chomsky n.d.). 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 n.d., Knuth n.d.a). In biology, the first appearance 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, we can gain insights into the underlying processes driving the evolution of species (Mooers & Heard n.d.) or cancer (Scott et al. n.d., Noble, Burri, Le Sueur, Lemant, Viossat, Kather & Beerenwinkel n.d.). 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 n.d.). 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. n.d.), 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 knowe=ledge between the fields of computer science and biology, with certain results being rediscovered nigh on half a century later, as discussed in section 2.1.

### 1.1.2 Quantifying tree balance

In a recent paper (Lemant et al. n.d.), Lemant and Noble proposed a new robust, universal index,  $J^1$ , for quantifying the balance of rooted trees with arbitrary node degree and size distributions. This index is based on Shannon entropy and favours even distributions of node sizes. We 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 1.1A). We 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, Burri, Le Sueur, Lemant, Viossat, Kather & Beerenwinkel n.d.).

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. n.d.), we obtained analytical approximations to the expected values of  $J^1$  under the Yule process and the uniform model,

Chapter\_1/figures/ruindices\_paper\_figures.png

**Figure 1.1:** a)  $J^1$  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. b)  $J^1$  values for trees generated under the Yule process and the uniform model. Solid curves represent the approximate expected values, and the dashed lines the 5th and 95th percentiles.

and tested their accuracy numerically for trees with up to 128 leaves (figure 1.1B). In the same study, we proved that caterpillar trees minimise  $J^1$  among bifurcating trees but not when larger outdegrees are permitted.

## 1.2 Agent-based modelling in oncology

### 1.2.1 Introduction

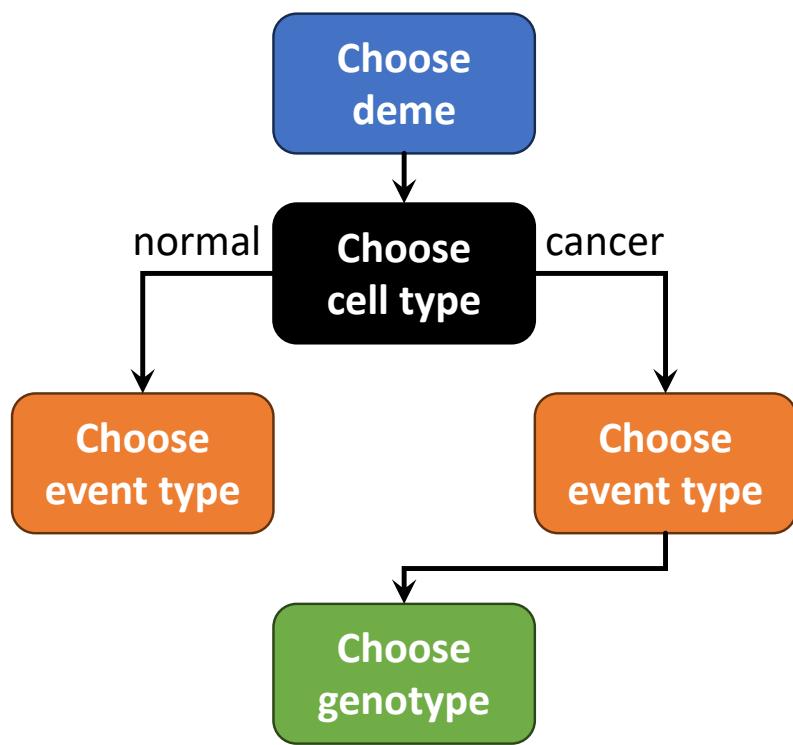
Agent-based models (ABMs) are a class of computational models that simulate the actions and interactions of individual agents within a system. These agents can

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. n.d.). However, therein lies their weakness, as oversimplification of a model can lead to rapid loss of its utility in modelling 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 our own, we have also been influenced by an alternative wording of this, that is *the best model is its own worst enemy*, by Philip Maini. Our 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 our demon-warlock framework (Bak et al. n.d.) to simulate the evolution of a tumour in space and draw conclusions on how spatial organisation will impact intratumour heterogeneity or patient outcomes (Noble, Burley, Le Sueur & Hochberg n.d., Noble, Burri, Le Sueur, Lemant, Viossat, Kather & Beerenwinkel n.d.). 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, we 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. n.d.) or PhysiCell (Ghaffarizadeh et al. n.d.), but they are not ideal either as simulating a realistically-sized tumour with these models is prohibitively expensive in terms of computational resources. Thus, our preferred approach is to develop a proprietary model which is informed by the literature and the data, and which has ample room for future expansion and improvement.

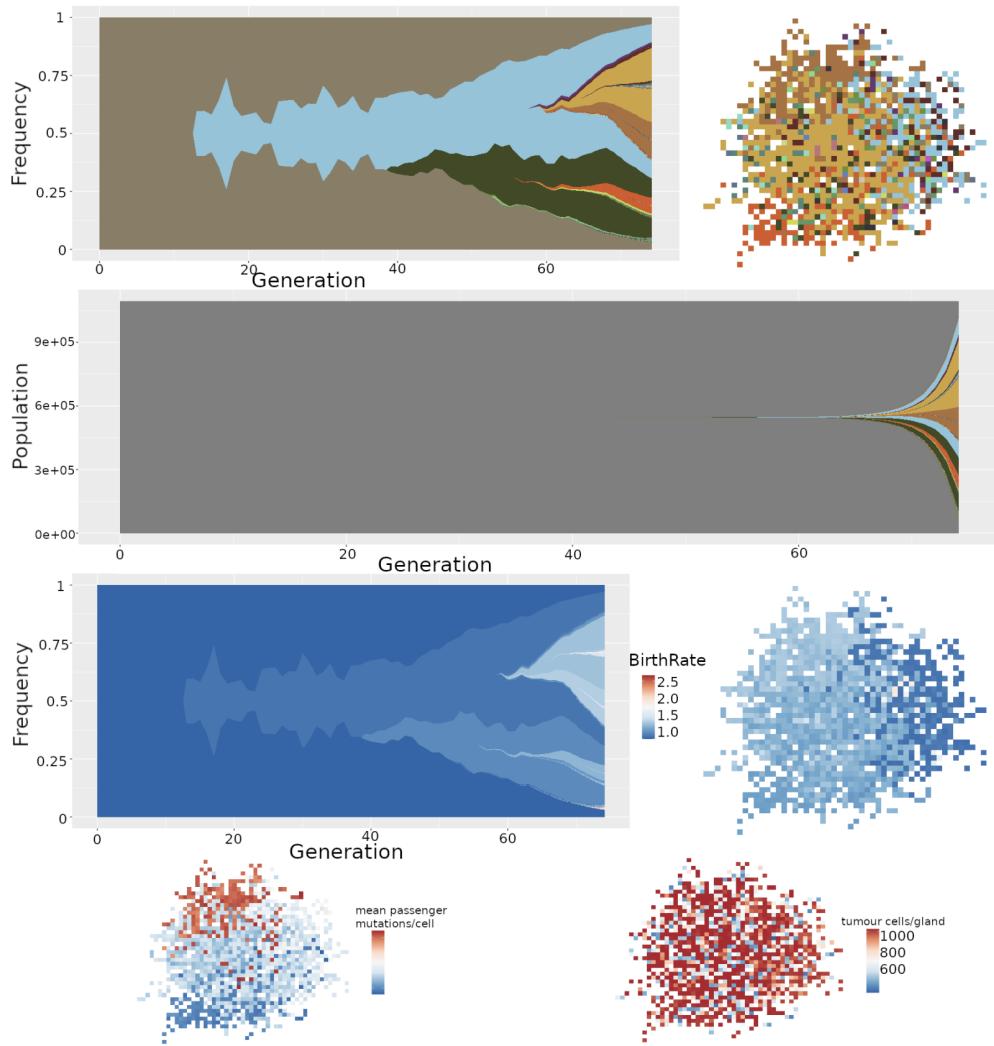
### 1.2.2 The demon-warlock framework

In a recent paper (Bak et al. n.d.), we formally introduced a new agent-based model for simulating the evolution of a tumour in space. 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. 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 invasion of tissue and other demes, and deme fission. 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 1.2. 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 1.3.

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 random numbers, it is not well-suited to simulating neutral stochastic markers, such as fluctuating methylation clocks (Gabbutt, Schenck, Weisenberger, Kimberley, Berner, Househam, Lakatos, Robertson-Tessi, Martin, Patel, Clark, Latchford, Barnes, Leedham, Anderson, Graham & Shibata n.d.). This is further discussed in section 4.3.



**Figure 1.2:** Event hierarchy in the `demon-warlock` framework. Figure reproduced from (Bak et al. n.d.) with the authors' permission.



**Figure 1.3:** 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 passenger mutations per cell. **g** Final spatial distribution of the number of tumour cells per gland.

## 1.3 Likelihood-free inference

### 1.3.1 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 ( $\theta$ ) 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 updated to the posterior distribution ( $P(\theta|D)$ ) using Bayes’ theorem (Bayes & Price n.d.):

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

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.

### 1.3.2 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. n.d., Sottoriva & Tavaré n.d., Jagiella et al. n.d.). 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  $\epsilon$ , 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. The choice of the distance function is crucial, as it determines how the simulated data is compared to the observed data. 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.  $d'(S(\hat{D}), S(D))$ . The choice of summary statistics is also crucial, as it determines how the data is reduced before the comparison.

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. n.d.). The choice

of the distance threshold is also essential, 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 n.d.). Finally, the computational aspects of ABC must be mentioned, 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, we 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, we introduce our simplified model of colorectal cancer gland fission and the accompanying ABC workflow. We 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.

## 1.4 Fluctuating methylation clocks

The concept of a molecular clock is a hallmark of 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 (?). 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 (?).

*TO DO:* Introduce and discuss FMCs.

## 1.5 Aims and objectives

### 1.5.1 Hypotheses

1. The  $J^1$  index is a universal measure of tree balance, and can be used, in conjunction with other tree shape indices, to differentiate between evolutionary modes in cancer.
2. Spatial agent-based models are a powerful tool for simulating the evolutionary dynamics of solid tumours.
3. These methods are useful for inferring the evolutionary history of colorectal cancer based on multi-site methylation array sequencing.

### 1.5.2 Aims

1. Calculate or approximate important properties of the  $J^1$  index, such as its expected value under standard tree-generating processes.
2. Determine the utility of sets of tree shape indices for differentiating between evolutionary modes in cancer.
3. Extend the fluctuating methylation clock model to multi-site sequencing of solid tumours on the example of colorectal cancer.

### 1.5.3 Objectives

1. Contextualise  $J^1$  within the broader field of tree shape indices in biology and computer science.
2. Investigate extreme and expected values of  $J^1$  under standard tree-generating processes.
3. Recapitulate the classification of evolutionary modes in cancer using a set of three tree shape indices.

4. Extend the discussion to a more interpretable and general system of tree shape indices.
5. Develop an agent-based model for simulating the evolutionary dynamics of colorectal adenocarcinoma.
6. 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.
7. Apply the model to colorectal cancer data and compare the inferred phylogenies to the trees generated by the model.



# Chapter 2

## Extreme and expected values of universal tree balance index $J^1$

**TO DO:** Finish introductions

### 2.1 Introduction

- there are different ways to define tree balance, depending on field and, consequently, type of tree
- however, it is an important concept **\*explain what it means for data structures and phylogenetics\***
- explain how evolutionary process differences can affect tree balance and mention the problem of phylogenetic trees in cancer

The balance of a tree can imply properties other than the symmetry in the tree's topology. Depending on the choice of index from at least 19 available in literature (Fischer et al. n.d.), the intuition and results may vary drastically or not at all between them. A common issue among them, however, is universality — some indices are only defined under certain topological restrictions (such as bifurcating trees), while others may not account for trees containing internal nodes with only one descendant or nodes with non-zero node sizes. As a consequence, it is difficult to use one index across different areas of research or even compare trees on different numbers of leaves from the same data source. Having a universal measure of balance applicable in this way would enable sensical comparison between any two trees.

In phylogenetics, balanced trees In a previous paper (Lemant et al. n.d.), we

introduced a new robust, universal balance index  $J^1$ . This index can be used on any tree topology and can account for trees whose nodes have non-zero size. The basic properties of the index are covered in the original paper, and in this follow-up we explore additional questions one might have when choosing a balance index.

## 2.2 Definitions

Before getting into the properties of  $J^1$  and related indices, we will briefly introduce these indices, tree naming suggestions and conventions, and notation used throughout the paper.

**Definition 2.2.1** (Rooted tree). A **rooted tree**  $T$  is a connected acyclic graph with node set  $V(T)$ , edge (branch) set  $E(T)$ , and node magnitudes assigned from  $\mathbb{R}_0^+$ , with an internal node designated as the root of  $T$ . We denote with  $\tilde{V}(T)$  the set of nodes with descendants of non-zero magnitude of tree  $T$  and with  $L(T)$  the set of its leaves, that is nodes with no descendants.

A **leafy tree** is a tree whose leaves are of equal non-zero size, and all internal nodes have size zero.

*Remark 2.2.1.* In general, a tree can have associated edge (or branch) lengths. We do not discuss such trees in this paper but focus on trees with equally sized branches.

**Definition 2.2.2** (The Sackin index). Let  $T$  be a rooted bifurcating tree on  $n$  leaves. We define the **Sackin index** of tree  $T$  as the sum of distances of the tree's leaves from its root i.e.,

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

where  $\nu(l)$  is the depth of leaf  $l$ .

*Remark 2.2.2.* The Sackin index can be generalised for trees with arbitrary node degree distributions. In this case, it is calculated as

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

where  $S_i^*$  is the size of the subtree rooted in internal node  $i$ , excluding node  $i$ .

**Definition 2.2.3** (Robust balance index). The robust balance index  $J^1$  of tree  $T$

is calculated as

$$J^1(T) = \frac{1}{\sum_{l \in \tilde{V}} S_l^*} \sum_{i \in \tilde{V}} S_i^* \sum_{j \in C(i)} W_{ij}^1, \quad (2.3)$$

where  $S_i^*$  is the magnitude of the subtree rooted in node  $i$ ,  $C(i)$  is the set of direct descendant nodes of node  $i$ , and  $W_{ij}^1$  is the node balance function defined as

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

where  $S_i$  is the magnitude of the subtree rooted in node  $i$ , including node  $i$ , and  $d^+(i)$  is the out-degree of node  $i$ .

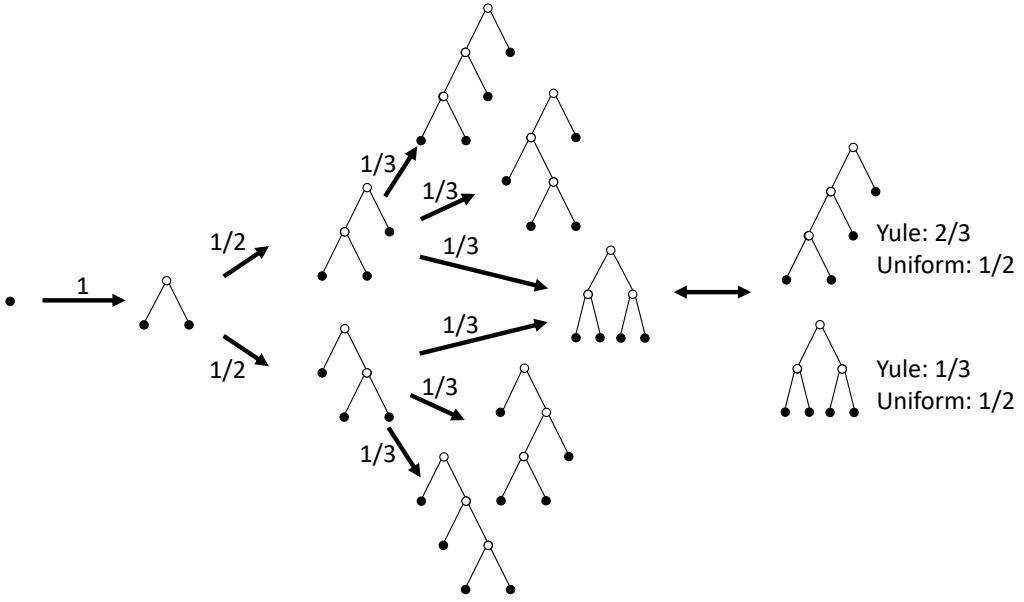
**Definition 2.2.4** (Cherry). A tree consisting only of a root and two leaves is called a **cherry**.

**Definition 2.2.5** (Yule model (Yule n.d.)). 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. The sum of probabilities of generating  $T$  in all possible ways is the probability of generating  $T$  under the Yule model.

**Definition 2.2.6** (Uniform model (Rosen n.d.)). Under the **uniform model** of tree generation, every bifurcating tree on  $n$  leaves has an equal probability of being generated, which is equal to  $n \binom{2n-2}{n-1}^{-1}$ .

*Remark 2.2.3.* We only consider leafy trees with equal leaf magnitudes in definitions 2.2.5 and 2.2.6.

The Yule and uniform models are statistical models which find their applications in evolutionary biology. Specifically, the Yule model, also known as the pure birth and coalescent model, is used when considering speciation rates and patterns (Alldous n.d., Steel & McKenzie n.d.). The uniform model is typically encountered in evolutionary process considerations (Mooers & Heard n.d., McKenzie & Steel n.d.). While simple, the two are valuable null models for studying different aspects of evolution.



**Figure 2.1:** Comparison of probabilities for generation of trees on 4 leaves under the Yule and uniform models.

## 2.3 Balancing binary trees according to $J^1$

The balance index  $J^1$  is defined on trees with arbitrary node out-degree distributions and node magnitudes (Lemant et al. n.d.). Depending on use case, this definition may be considered too broad. Specifically, in computer science the most commonly used tree is a binary tree.

**Definition 2.3.1** (Binary tree (Nievergelt n.d.)). A **binary tree** is a rooted tree such that each node can have 0, 1, and 2 children.

While  $J^1$  originally draws inspiration from the problem of quantifying properties of cancer phylogenetic trees (Noble, Burri, Le Sueur, Lemant, Viossat, Kather & Beerenwinkel n.d.), trees are encountered and used regularly in other fields, such as the broader realm of evolutionary biology, as well as computer science. However, one might notice that the notion of balance is used differently across those fields, being an effectively binary consideration in data structures, i.e. a tree can be balanced or unbalanced based on some measure (?). In evolutionary biology, a finer scale is used, and comparisons between trees are more relevant (Mir, Rosselló & Rotger n.d., Mir, Rotger & Rosselló n.d., Fischer et al. n.d.). Despite these differences, we have shown robustness of our method under common data gathering errors in biology (?) and can further show that it generalises tree balance in data structures. Let us begin by defining the different measures of balance encountered in computer science

literature.

**Definition 2.3.2** (Root balance score). The **root-balance score**  $\rho(T_n)$  of a binary tree  $T_n$  on  $n > 1$  nodes is

$$\rho(T_n) = \frac{l+1}{n+1}, \quad (2.5)$$

where  $l$  is the number of nodes in the left subtree of  $T_n$ .

Intuitively, one could imagine the root balance score as evaluating how well the tree could be balanced on a fulcrum placed under node  $n$ .

**Definition 2.3.3** (Weighted path length). Let  $T$  be a rooted tree on  $n$  nodes, and  $\alpha_i$ ,  $i = 1, 2, \dots$  the weights (or access frequencies) assigned to its nodes  $v_i$ . We define the **weighted path length** of tree  $T$  as

$$|T_n| = \sum_{v_i \in V(T)} \alpha_i \nu(v_i). \quad (2.6)$$

*Remark 2.3.1.* Let us rewrite the weighted path length in a more familiar way:

$$|T| = \sum_{v_i \in V(T)} \alpha_i \nu(v_i) = \sum_{i \in V(T)} p_i \nu(i) = \sum_{i \in V(T)} S_i^*,$$

which is just the generalised Sackin index for tree  $T$ .

Consider then the following proposition.

**Proposition 2.3.1** (Lemant et al. (?)). *Let  $T$  be a leafy tree on  $n$  leaves with  $d^+(i) = m > 1$ . Then*

$$J^1(T) = \frac{n \log_m n}{I_S(T)} \quad (2.7)$$

where  $I_S$  is the Sackin index.

*Corollary 2.3.1.* We can rewrite equation (2.7) for binary trees as

$$J^1(T) = \frac{n \log_2 n}{|\bar{T}|}. \quad (2.8)$$

This means that, for a fixed number of leaves  $n$ , minimising the weighted path length  $|\bar{T}|$  is equivalent to maximising  $J^1$ .

**Theorem 2.3.1** (Nievergelt and Wong (Wong & Nievergelt n.d.)). *Let  $T_n$  be a binary node-tree with  $n$  nodes and root balance score  $\beta$ . Then its total path length*

$|T_n|$  is bound from above by

$$|T_n|_{\text{upper}} = (H(\beta))^{-1}(n+1)\log_2(n+1) - 2n, \quad (2.9)$$

where  $H(\beta) = \beta \log_2 \beta^{-1} + (1-\beta) \log_2(1-\beta)^{-1}$ .

**Corollary 2.3.2.** There is a lower bound on  $J^1$  for binary trees with  $n$  nodes and root balance score  $\beta$ , and it equals

$$J_{\text{lower}}^1 = \frac{n \log_2 n}{|T_n|_{\text{upper}}}. \quad (2.10)$$

Corollary 2.3.2 follows trivially from proposition 2.10 but represents a result which has a deeper connection to the fundamentals of information theory.

**Proposition 2.3.2.** *The Huffman method (?) maximises  $J^1$  on binary trees for a given set of node magnitudes.*

*Proof.* Consider a set of non-negative real numbers  $F = \{\alpha_1, \dots, \alpha_m\}$ . The Huffman method places larger numbers, i.e. nodes of higher magnitude, closer to the root, minimising the weighted path length as a result. By corollary 2.3.2, the Huffman method maximises  $J^1$ .  $\square$

As Huffman coding is an optimisation algorithm, we can use  $J^1$  to measure how close a tree constructed using a given set of node magnitudes is to the maximally balanced binary tree on the same set. This means one can meaningfully measure how close an alternative algorithm which runs in a faster time complexity, such as arithmetic coding (?), gets to the optimal solution.

Let us now define the most common binary tree type one might encounter in computer science literature, binary search trees.

**Definition 2.3.4** (Binary search tree). A **binary search tree**  $T_n$  over  $n$  entries (w.l.g. numbers)  $x_1, \dots, x_n$  is a labelled binary tree, each of whose nodes have been labelled with a distinct number chosen from  $x_1, \dots, x_n$  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  $N$ .

The balance of binary search trees is usually measured by comparing the numbers of leaves in the left and right subtrees of the root. A more specific use case of

binary search trees is for implementing dynamic data structures such as dictionaries (Tsakalidis n.d.). For this purpose, weight-balanced trees are often used.

**Definition 2.3.5** (Weight-balanced tree). A weight-balanced tree is a binary search tree that stores the sizes of subtrees in the nodes. That is, a node contains:

- key, of any type;
- value;
- left and right pointers to the child nodes;
- size.

**Definition 2.3.6** (Node balance). A node  $i$ , with children  $i_l$  and  $i_r$  and corresponding weights  $w[i], w[i_l], w[i_r]$ , is  $\alpha$ -weight-balanced if

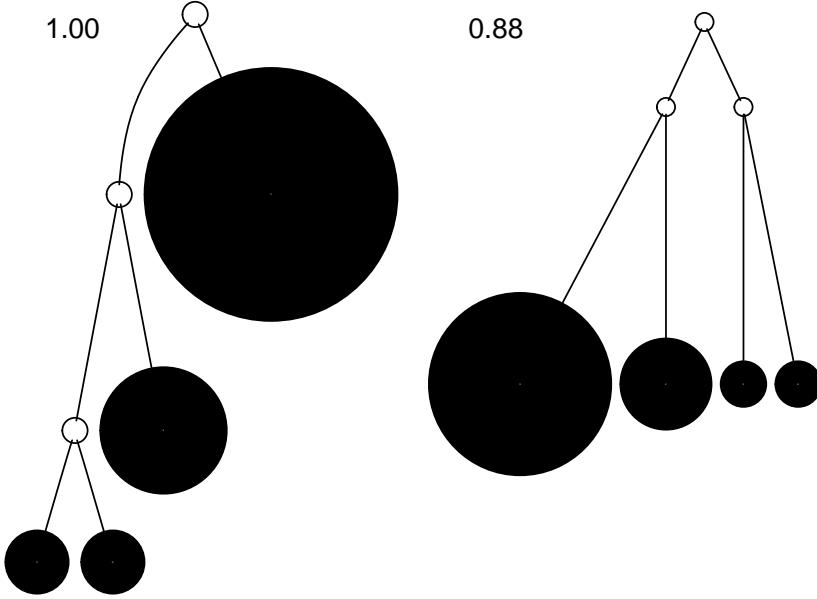
$$w[i_l] \geq \alpha w[i], \quad (2.11)$$

$$w[i_r] \geq \alpha w[i], \quad (2.12)$$

where  $\alpha$  is a numerical parameter to be determined when implementing weight-balanced trees.

Recall the general definition of  $J^1$ , where each internal node has an associated node balance. In the most general definition of  $J^1$ , equation (2.3), the node balance  $W_i$  of node  $i$  can be any function  $W_i : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$  according to some property of node  $i$  and its descendants. In other words, one may consider the node balance score  $W_i$  as a generalisation of the root balance score since it calculates how evenly a node's descendants split the subtree rooted in it, as opposed to the overall tree structure.

The balance of weight-balanced trees is optimised dynamically through operations called rotations, which keep the weights of left and right subtrees within  $\alpha$  of each other. Balance plays a major role in constructing good data structures in computer science, but computational efficiency is of much greater importance in the field, with balancing happening dynamically when new trees are initialised or nodes added to existing ones, other good examples being AVL trees and red/black trees (?). The role of balance indices for static data structures is thus quite niche. However, by having a mathematically robust definition underpinning  $J^1$ , and considering general features of rooted trees, we have shown that it can easily be used



**Figure 2.2:** 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.

far wider than just the original application of analysing cancer phylogenies. The connections between different tree use cases thus show how one can construct powerful methods for general use without sacrificing specificity.

## 2.4 Expected value of $J^1$ under simple evolutionary processes

For applications in evolutionary theory, an important property of balance indices is their expected value under an evolutionary process, as it tells us which process drove the tree into its current state. Equivalently to standard calculations in probability, one can obtain the expected value of a balance index by calculating the sum of the products of balance index values with the corresponding probability of the tree shape.

**Definition 2.4.1.** Let process  $P$  generate tree  $T_i$  with probability  $p(T_i)$  for  $T_i \in \mathcal{T}_n$ . The expected value of index  $I_B$  on  $n$  leaves under process  $P$  is defined as

$$\mathbb{E}_P^n(I_B) = \sum_{T_i \in \mathcal{T}_n} p(T_i) I_B(T_i). \quad (2.13)$$

While the above definition seems quite general, in practice the generating process can introduce constraints on the set of trees. Two of the simplest, and most widely studied, processes of tree generation are the uniform model (Rosen n.d.) and the Yule model (Yule n.d.), which generate bifurcating trees. The expected value of a few indices, and even some higher moments in certain cases, are known for each of these tree generation processes (Mir, Rosselló & Rotger n.d., M. Coronado et al. n.d., Goh et al. n.d.). Among these is the Sackin index, which is useful for our purposes.

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

$$\mathbb{E}_Y^n(I_S) = 2n \sum_{i=2}^n \frac{1}{i}, \quad (2.14)$$

as shown in (Kirkpatrick & Slatkin n.d.). Consider the relationship between  $J^1$  and the Sackin index for bifurcating trees, which is directly derived from equation (2.7),

$$J^1 = \frac{n \log_2 n}{I_S}. \quad (2.15)$$

This means that the expected value of  $J^1$  for a tree on  $n$  leaves, generated under the Yule model is

$$\mathbb{E}_Y^n(J^1) = \mathbb{E}_Y^n\left(\frac{n \log_2 n}{I_S}\right) = n \log_2 n \mathbb{E}_Y^n(1/I_S).$$

We can rewrite this equation as

$$\frac{1}{\mathbb{E}_Y^n(1/I_S)} = \frac{\mathbb{E}_Y^n(J^1)}{n \log_2 n}, \quad (2.16)$$

which is the harmonic mean of the Sackin index. The harmonic mean under the Yule process is not a standard result in literature, nor have we been able to obtain a closed-form solution for this problem so far. We can, however, compare the harmonic and arithmetic means of  $I_S$  by considering the Jensen gap

$$\mathcal{J}(f, X) = \mathbb{E}[f(X)] - f(\mathbb{E}[X]). \quad (2.17)$$

**Theorem 2.4.1** (Liao and Berg (Liao & Berg n.d.)). *Let  $X$  be a one-dimensional random variable with mean  $\mu$ , and  $P(X \in (a, b)) = 1$ , where  $\infty \leq a < b \leq \infty$ . If*

$\phi(x)$  is a twice differentiable function on  $(a, b)$ , and

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

then

$$\inf_{x \in (a, b)} \{h(x; \mu)\} \text{Var}(X) \leq \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]) \leq \sup_{x \in (a, b)} \{h(x; \mu)\} \text{Var}(X). \quad (2.18)$$

**Proposition 2.4.1.** 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:

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

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

as  $n \rightarrow \infty$ .

*Proof of proposition 2.4.1 (i).* Let  $\mu_Y$  be the expected value of the Sackin's index under the Yule process for trees on  $n$  leaves, and  $f(x) = \frac{1}{x}$ . By theorem 2.4.1

$$h(x; \mu_Y) = \frac{f(x) - f(\mu_Y)}{(x - \mu_Y)^2} - \frac{f'(\mu_Y)}{x - \mu_Y} = \frac{1}{x \mu_Y^2}. \quad (2.19)$$

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

$$\frac{n \log_2 n}{\frac{(n-1)(n+2)}{2} \mu_Y^2} \text{Var}_Y(I_S) \leq \mathbb{E}[J^1] - \frac{n \log_2 n}{\mathbb{E}[I_S]} \leq \frac{n \log_2 n}{\mu_Y^2 n \log_2 n} \text{Var}_Y(I_S), \quad (2.20)$$

where the supremum and infimum of  $h(x, \mu)$  are substituted with extremal values of the Sackin index on bifurcating trees (Fischer n.d.). The expectation of Sackin's index under the Yule model is a known quantity (Cardona et al. n.d.), and can be calculated as

$$\mathbb{E}_Y(I_S) = 2n \sum_{i=2}^n \frac{1}{i}, \quad (2.21)$$

as is its variance

$$\text{Var}_Y(I_S) = 7n^2 - 4n^2 \sum_{i=1}^n \frac{1}{i^2} - 2n \sum_{i=1}^n \frac{1}{i} - n. \quad (2.22)$$

Substituting these expressions into equation (2.20), we find limits

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

for the lower bound on the gap, and

$$\begin{aligned} \frac{n \log_2 n}{\mu_Y^2 n \log_2 n} \text{Var}_Y(I_S) &= \frac{7n^2 - 4n^2 \sum_{i=2}^n \frac{1}{i^2} - 2n \sum_{i=2}^n \frac{1}{i} - n}{4n^2 (\sum_{i=2}^n \frac{1}{i})^2} \\ &\xrightarrow{n \rightarrow \infty} \frac{1}{(\sum_{i=2}^n \frac{1}{i})^2} \rightarrow 0 \end{aligned}$$

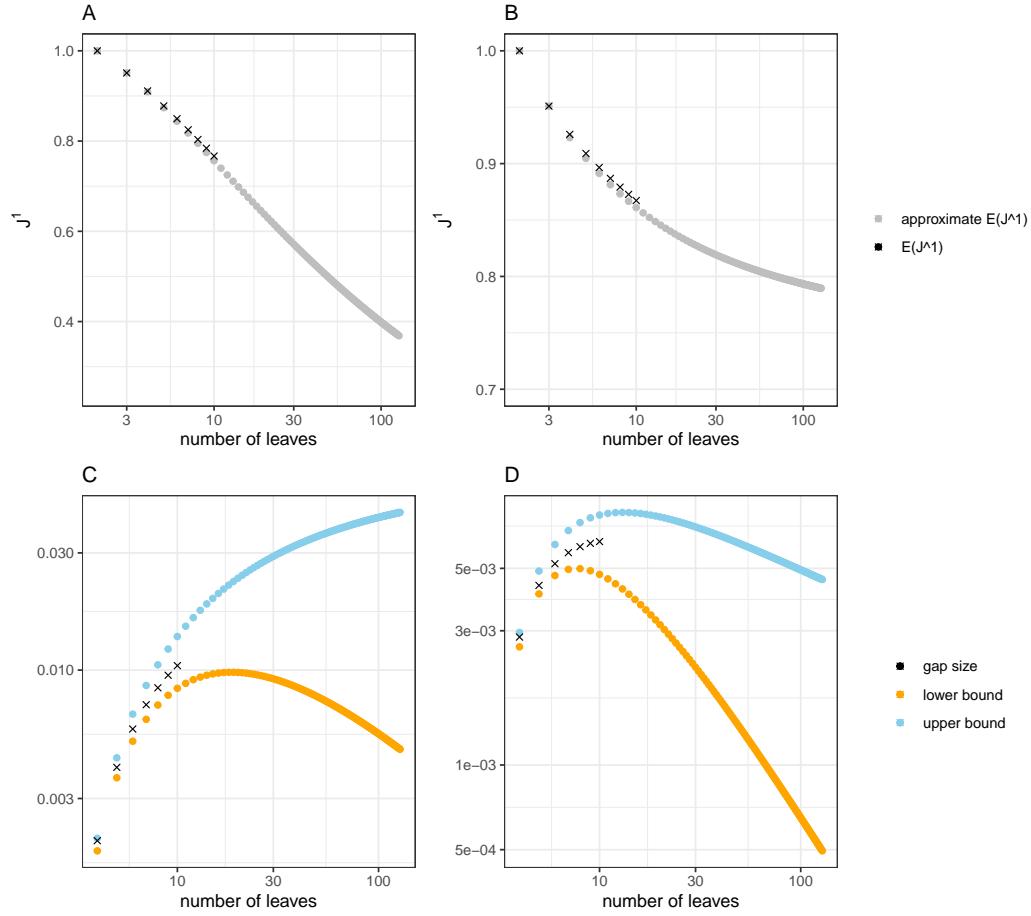
for the upper bound on the gap. The upper bound reaches a maximum at  $n = 13$  and is equal to approximately 0.0079, while the lower bound reaches a maximum at  $n = 8$  and equals approximately 0.005.  $\square$

In figure 2.3, we show behaviour of the Jensen gap and its bounds for  $J^1$  under the Yule and uniform models.

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.

## 2.5 Analytic properties of the $J^1$ index

A balance index is typically normalised by considering its extremal values for a given number of leaves (Fischer et al. n.d.). However, by its definition (?),  $J^1$  follows a different normalisation procedure which in turn makes comparison of its values on trees of different sizes valid. The way  $J^1$  was defined also makes it more complicated to define families of trees which maximise or minimise  $J^1$  if we do not impose restrictions on the tree topology or node size distribution. In this section we do a bit of both, and consider only leafy trees with the out-degree of each internal node greater than 1.



**Figure 2.3:** Top row: True values of  $E(J^1)$  for up to 10 leaves were calculated manually, and the approximations up to 128 leaves were calculated as  $n \log_2 n / E(I_S)$ . **A** — uniform model, **B** — Yule model.

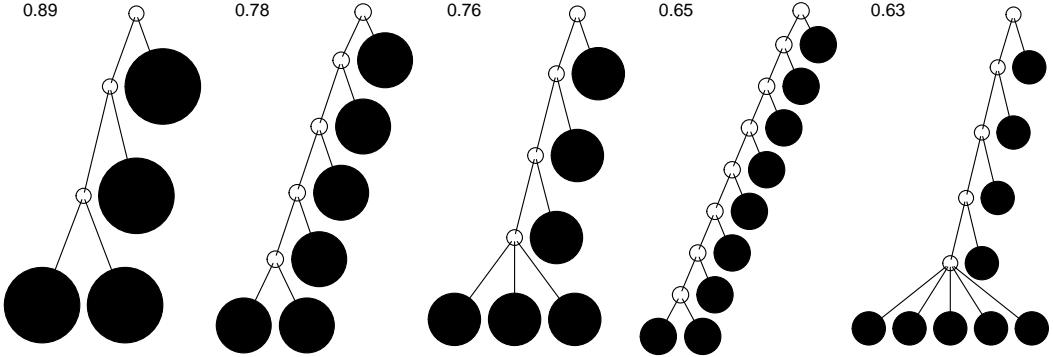
**Bottom row:** The Jensen gap of  $E(J^1)$  calculated for trees up to 128 leaves under the uniform model (**C**), and the Yule model (**D**). The size of the gap is calculated as the difference between the true and approximate expected value, with the gaps for 2 and 3 leaves equal to zero as there is only one possible bifurcating tree shape for each of those values. Refer to tables ?? and ?? for numerical values of the gap size for the first several values of  $n$ .

### 2.5.1 Properties of $J^1$ on different tree families

One may not encounter trees which yield extreme values of the balance index in practice often, if at all. However, it is still important to investigate the kinds of trees that maximise or minimise the index for the purposes of a complete mathematical discussion.

#### Some special tree topologies

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. We have previously derived



**Figure 2.4:** If we limit our search to leafy trees, the least balanced tree on a given number of leaves is not necessarily the caterpillar. Pictured are the caterpillar trees on 4, 6, and 9 leaves, as well as minimally balanced brooms for 6 and 9 leaves, with corresponding  $J^1$  values.

a general expression for leafy trees of this topology (?)

$$J^1(T_C) = \frac{2n \log_2 n}{(n-1)(n+2)}. \quad (2.23)$$

Most balance indices in literature define the caterpillar topology as the least balanced one (Fischer et al. n.d.). 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. We also have to consider 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 2.5.1.** Let  $T_B$  be a leafy tree on  $n$  leaves. Let every internal node of  $T_B$  except for the one with the highest depth have out-degree 2 such that one of its 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 can derive a general expression of  $J^1$  for this family of trees.

**Proposition 2.5.1.** *The value of  $J^1$  for a broom tree  $T_B$  on  $n$  leaves, of which  $k$  in the broom head is*

$$J^1(T_B) = \frac{2}{(n+k)(n-k+1)} (n \log_2 n - k \log_2 k + k). \quad (2.24)$$

We can also generalise the result of proposition 2.5.1 slightly in the following way.

**Proposition 2.5.2.** *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  $q$ ,  $q \in \mathbb{R}$ , and the leaves in the “handle” all of equal size 1, the value of  $J^1$  is*

$$\begin{aligned} J^1(T_{Bq}) &= \frac{1}{(n-k+1)(q+(n-k)/2)} \\ &\times \left( q \log_k q - \sum_{i=1}^k l_i \log_k l_i + (q+n-k) \log_2(q+n-k) - q \log_2 q \right), \end{aligned} \tag{2.25}$$

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

In figure 2.4 we 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 2.5.3.** *For leafy trees on  $n$  leaves and no linear parts, the caterpillar minimises  $J^1$  iff  $n < 5$ .*

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

$$J_B^1(n, 2) = \frac{2n \log_2 n}{(n+2)(n-1)}, \tag{2.26}$$

$$J_B^1(n, 3) = \frac{2}{(n+3)(n-2)}(n \log_2 n - 3 \log_2 3 + 3). \tag{2.27}$$

Consider the case when  $J_B^1(n, 2) < J_B^1(n, 3)$ . Plugging in equations (2.26) and (2.27), we can rearrange the inequality to find

$$8n \log_2 n - 6(n^2 + n - 2) \log_2 3 + 6(n^2 + n - 2) > 0, \tag{2.28}$$

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.  $\square$

This proposition gives us a threshold for the number of leaves at which the caterpillar is not longer the minimally balanced tree for the given number of leaves, which sets  $J^1$  apart from traditionally defined balance indices. This comes with the benefit of the new index's robustness to removal of small nodes, universality, and generality. However, we are yet to prove the following statement.

**Conjecture 2.5.1.** *For leafy trees on  $n$  leaves and no linear parts, the tree that minimises  $J^1$  belongs to the broom family.*

### Behaviour as $n \rightarrow \infty$

We have derived general behaviour of  $J^1$  on broom and caterpillar trees for a given number of leaves  $n$ . However, as some of the equations implied are not analytically solvable (e.g., conjecture 2.5.1) we also explore asymptotic behaviour of  $J^1$ . If we let  $n \rightarrow \infty$ , the value of  $J^1$  for the caterpillar from equation (2.23) will behave like

$$\lim_{n \rightarrow \infty} J^1(T_C) = \frac{2 \log_2 n}{n}. \quad (2.29)$$

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 2.5.4.** *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*

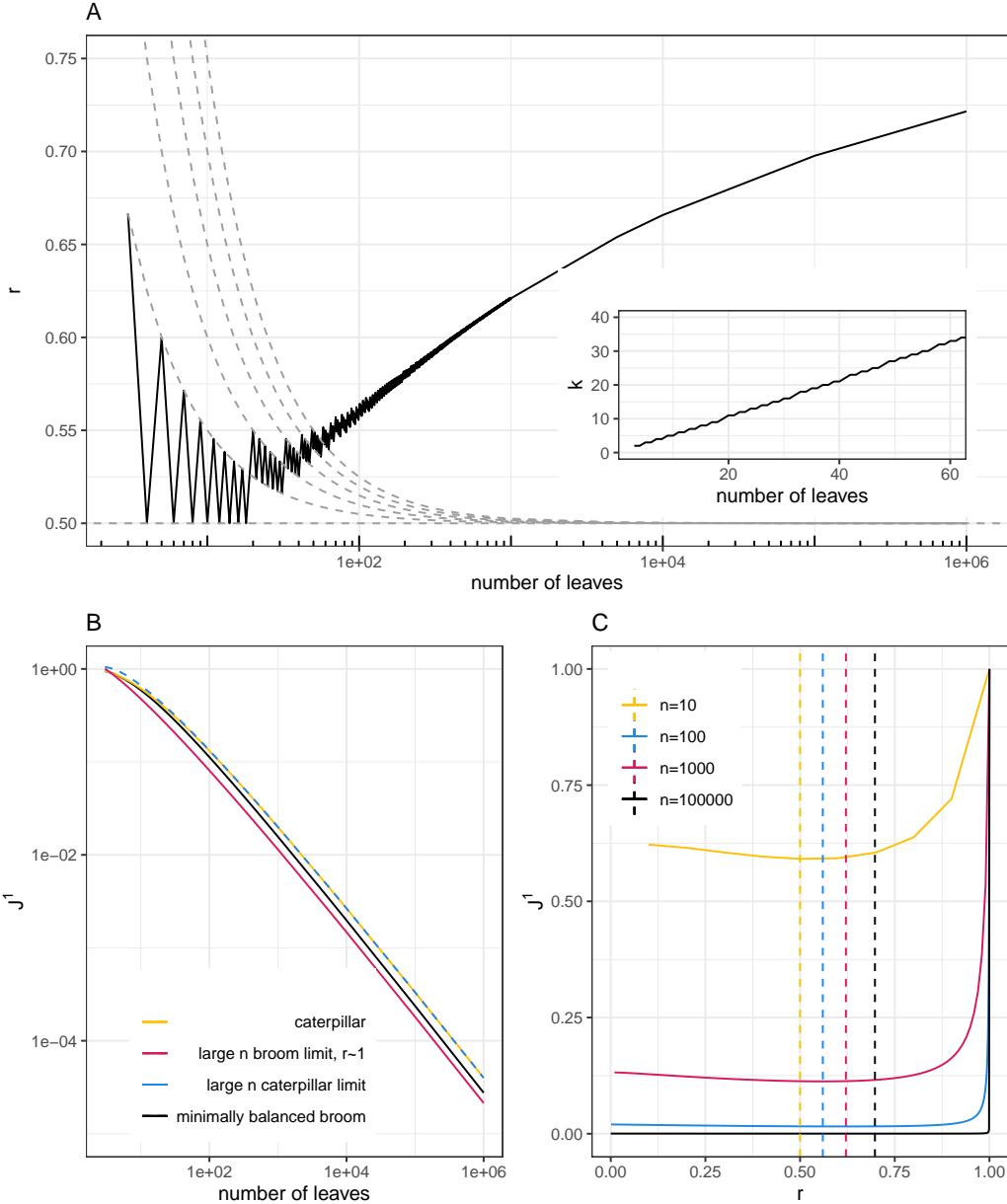
$$J^1(T_B) > J^1(T_C) \quad \text{iff} \quad p < \frac{1}{2}, \quad (2.30)$$

as  $n \rightarrow \infty$ .

For broom trees, the behaviour is a little more complicated and, perhaps, counterintuitive. Consider the following.

**Proposition 2.5.5.** *Let  $\mathcal{T}_B(n)$  be the set of all broom trees 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  $r_{opt} \rightarrow 1$  as  $n \rightarrow \infty$ .*

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



**Figure 2.5:** 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  $a = 0, 1, 2, \dots$  lie on the dashed grey lines. **B:** Behaviour of caterpillar and broom for different values of  $n$ . **C:**  $J^1$  for different broom trees on a given number of leaves using equation (2.24). The dashed lines indicate the value of  $r$  for which  $J^1$  is minimal.

start 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.

Finding the true value of  $k$  which minimises  $J^1(T_B)$  analytically is difficult. The derivative with respect to  $k$  of equation (2.24) yields a transcendental equation which

is not analytically solvable.

## 2.6 Discussion

The main aim of this paper was to explore deeper analytic properties of the robust, universal balance index  $J^1$  and start finding its place in the broader context of tree balance by extending past results and uncovering new connections.

There are still areas where the index  $J^1$  falls short in terms of generality. The balance of a tree whose branch lengths differ is not a case that the new index can handle. This opens up the possibility of further generalisation of the index and future research.

Finally, we only touched upon directly obtainable relationships without considering different real-world use cases of the index and the implications of equation (2.7). 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 cancer evolution *in silico* via evolutionary 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. We are 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. n.d.). 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.

In this chapter, we will exmine the utility of two different sets of evolutionary indices for tracking the evolution of tumours *in silico*.

### 3.1.1 Why even bother with indices?

Before introducing the sets of indices used to analyse properties of trees, let us consider a simpler question — can we map the set of all possible trees to the set of real numbers? For this purpose we should 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 n.d.). Each node then has a corresponding size, giving us an  $n$ -tuple of real numbers  $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ , and each edge (branch) has a corresponding length or  $(l_1, \dots, l_{(n-1)}) \in \mathbb{R}^{(n-1)}$ . This means we would need a family of maps

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

It would be easy to construct a mapping which would allow us to “enumerate” each possible tree with a real number. The only problem with this approach is that it is not at all useful, first and most importantly due to its lack of any interpretability. This chapter outlines an approach which uses real-valued summaries of trees’ properties in a way that is both intuitive and mathematically sound.

### 3.1.2 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 n.d.). 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,

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

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.

### 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, we 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,

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

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

### Balance index

As discussed in chapter ??, 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.

#### 3.1.3 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 n.d.). 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 (kimverity n.d.).

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.

## 3.2 Tree resolution

The first question we need to address is whether the indices we have chosen are sufficient to distinguish between different trees.

### 3.2.1 3-dimensional index space

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

$$N = \prod_{i=1}^k \alpha_i^{l_i}, \quad (3.4)$$

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

*Proof.* ... □

### 3.3 Computational methods

#### 3.3.1 Agent-based modelling framework - *warlock/demon*

There is no shortage of agent-based models of tumour evolution (Colyer et al. n.d.), 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, we want our 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, we also need a large number of simulations over which we can infer more general deterministic properties. 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 we decided to use is **warlock** (Bak et al. n.d.), a **snakemake** wrapper written for **demon** (Noble n.d.). 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 n.d.), 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 (cite github repos once finished), which allows reproducibility using identical parameters and random seeds. Parameter values for different batches can be found in the appendix (ref).

## 3.4 Results

### 3.4.1 Sensitivity of evolutionary mode to parameter values

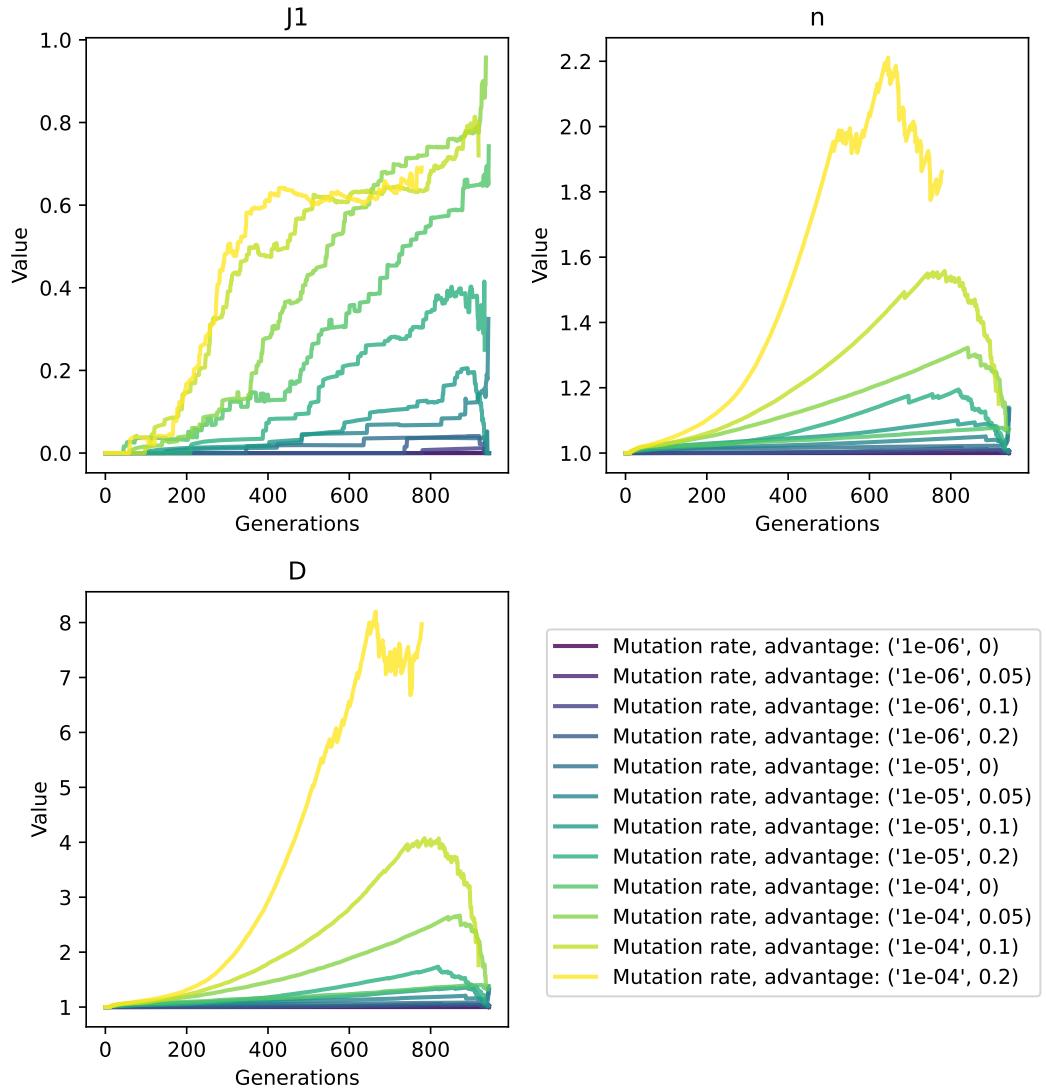
- there is clear variance in trajectories within a spatial config but less than one might expect for parameters within an order of magnitude of each other
- all things but spatial config being equal, the trajectories seem to be distinct in later stages of evolution
- should formalise somehow??

TO DO: add figures for index space, add figures to appendix, add expanded index set figures (both temporal and index space)

## 3.5 Discussion

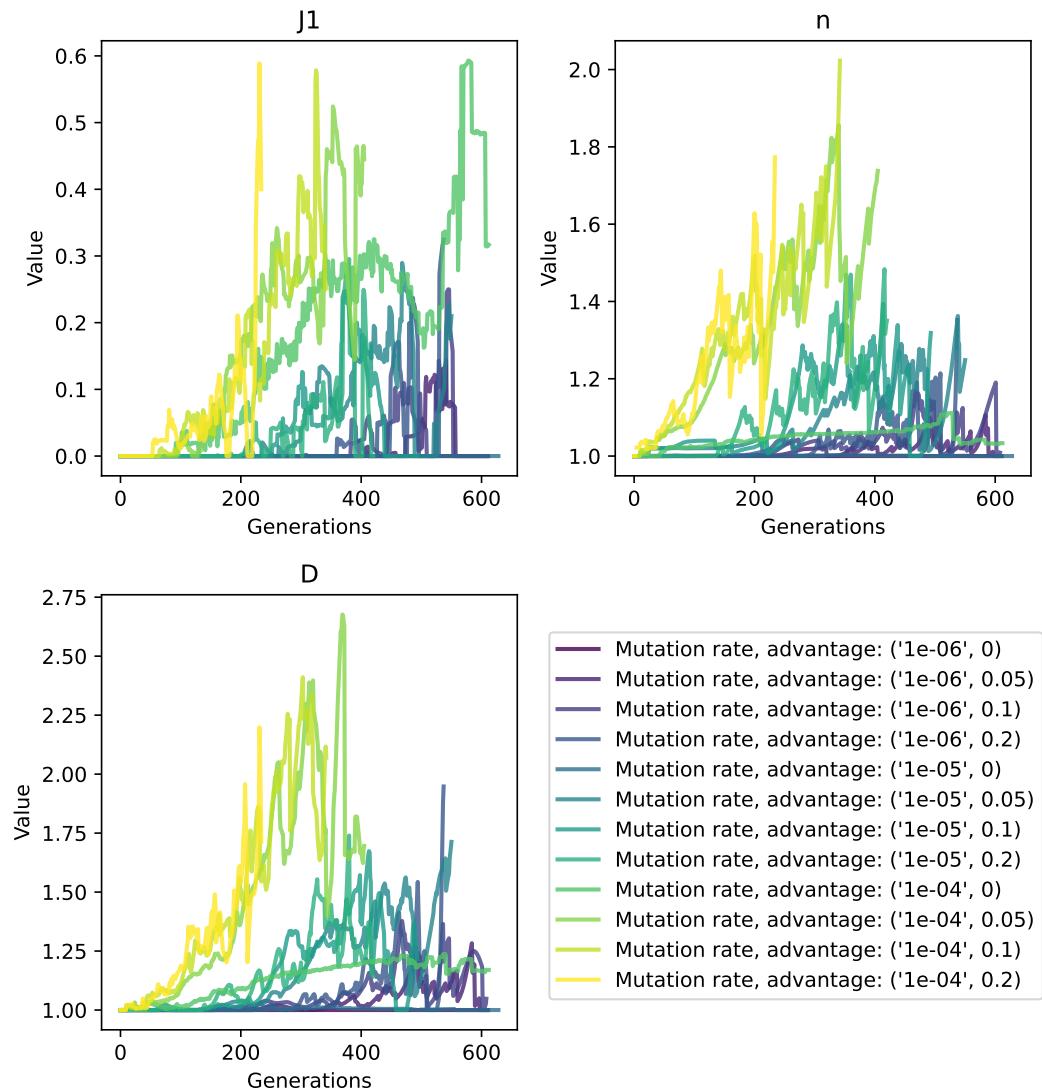
- clear differences between different tumour trajectories, but also decent amount of variance depending on parameters — which ones are realistic? (need to be inferred from real data)

Average temporal plots for boundary growth

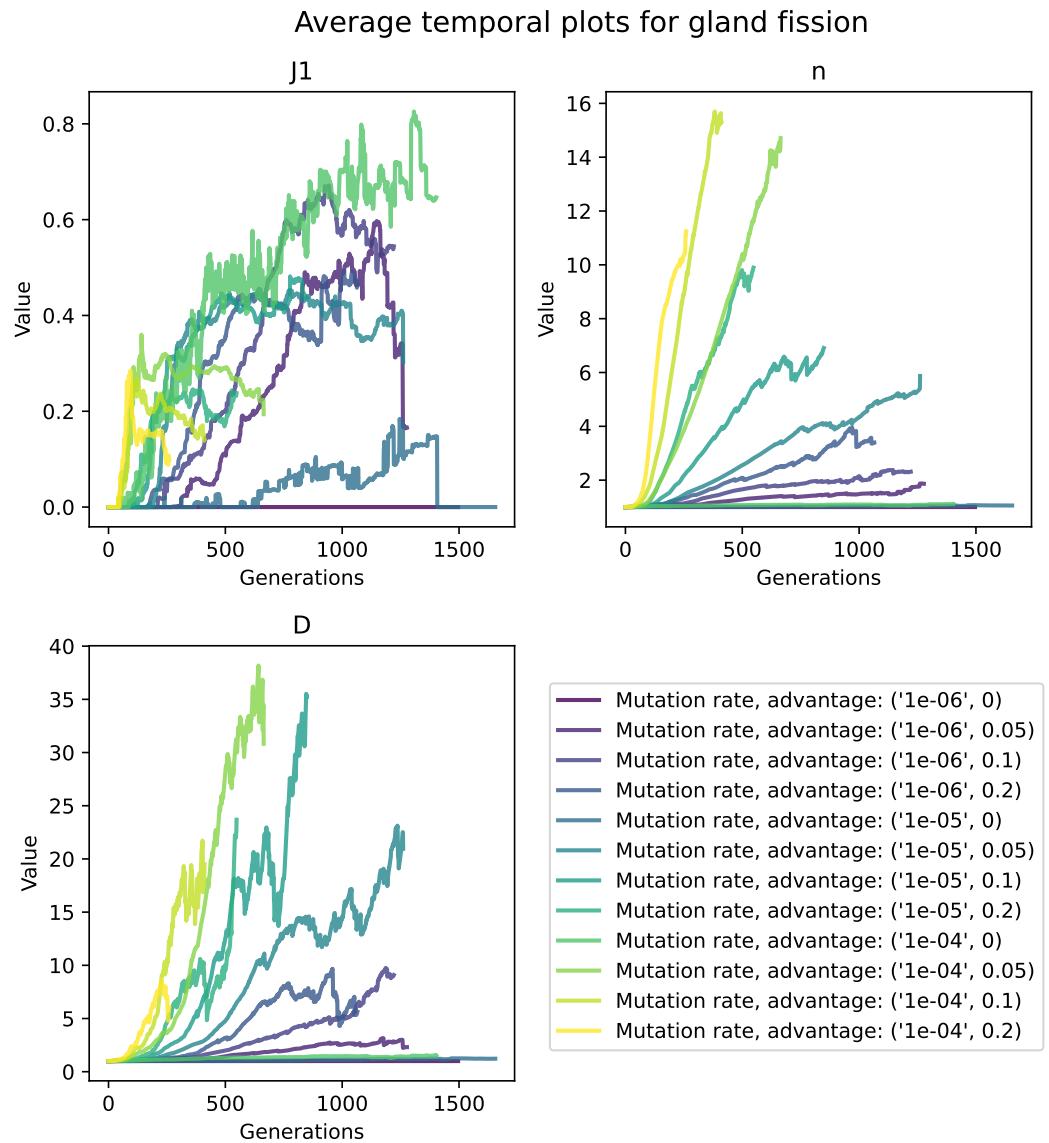


**Figure 3.1:** Average trajectories of the three indices for different values of driver mutation rate and selective advantage for tumours progressing via boundary growth.

Average temporal plots for non-spatial tumours

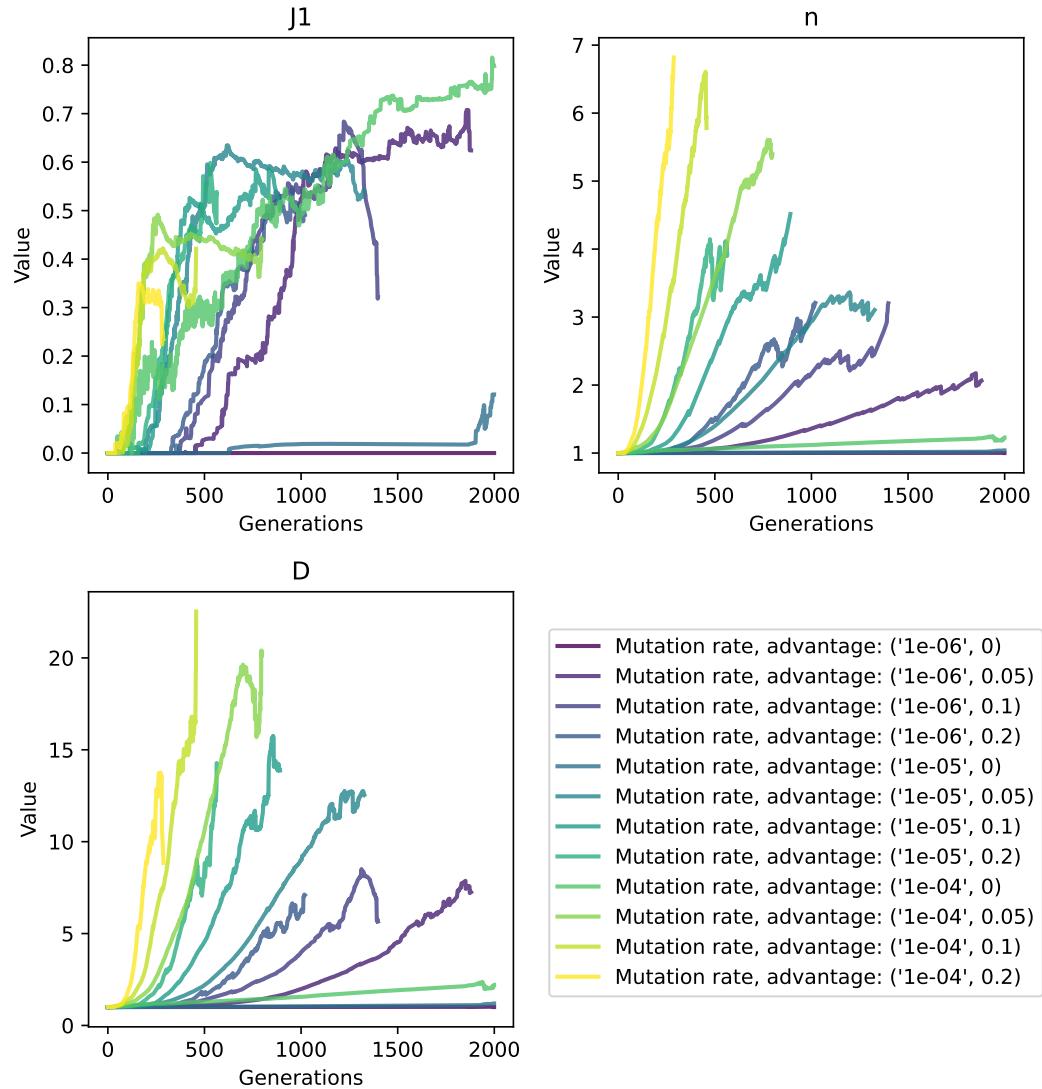


**Figure 3.2:** Average trajectories of the three indices for different values of driver mutation rate and selective advantage for well-mixed cancer cell populations.



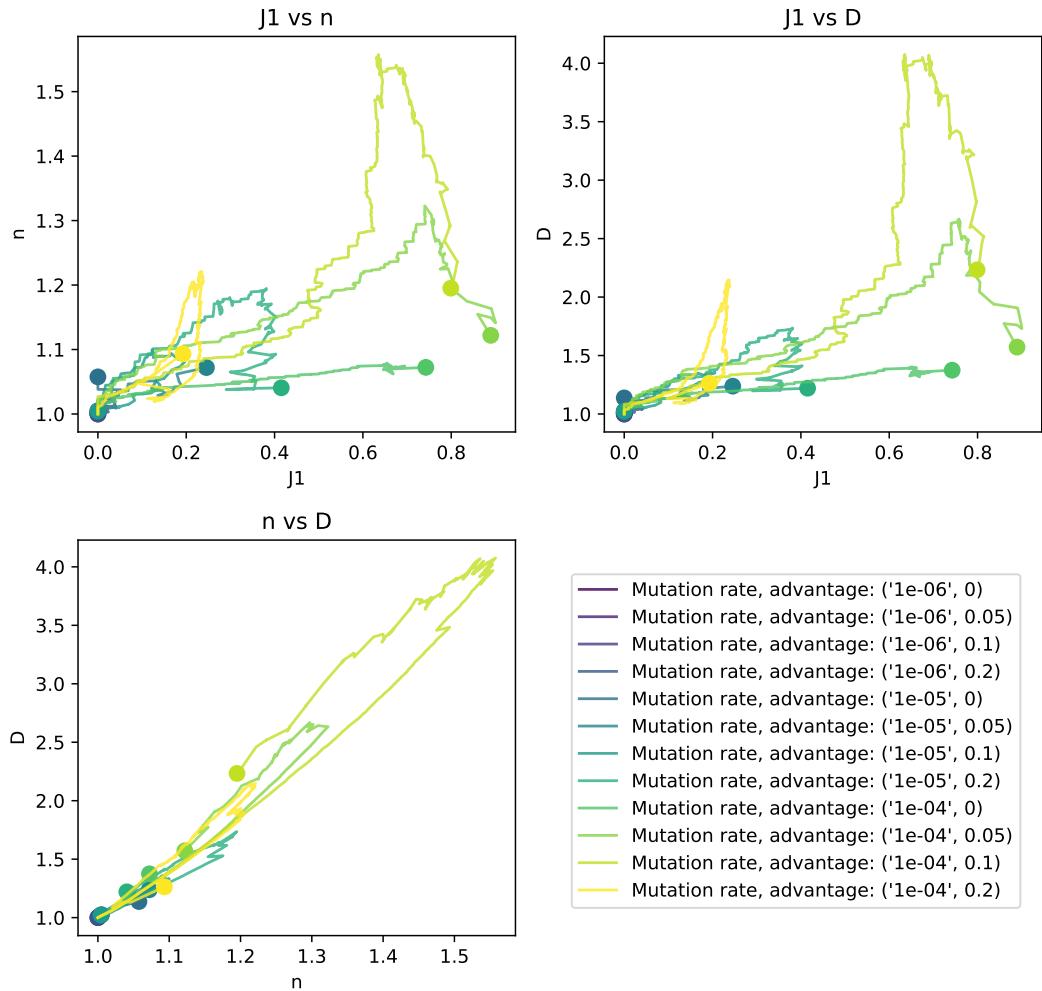
**Figure 3.3:** Average trajectories of the three indices for different values of driver mutation rate and selective advantage for gland fission.

Average temporal plots for invasive glandular tumours



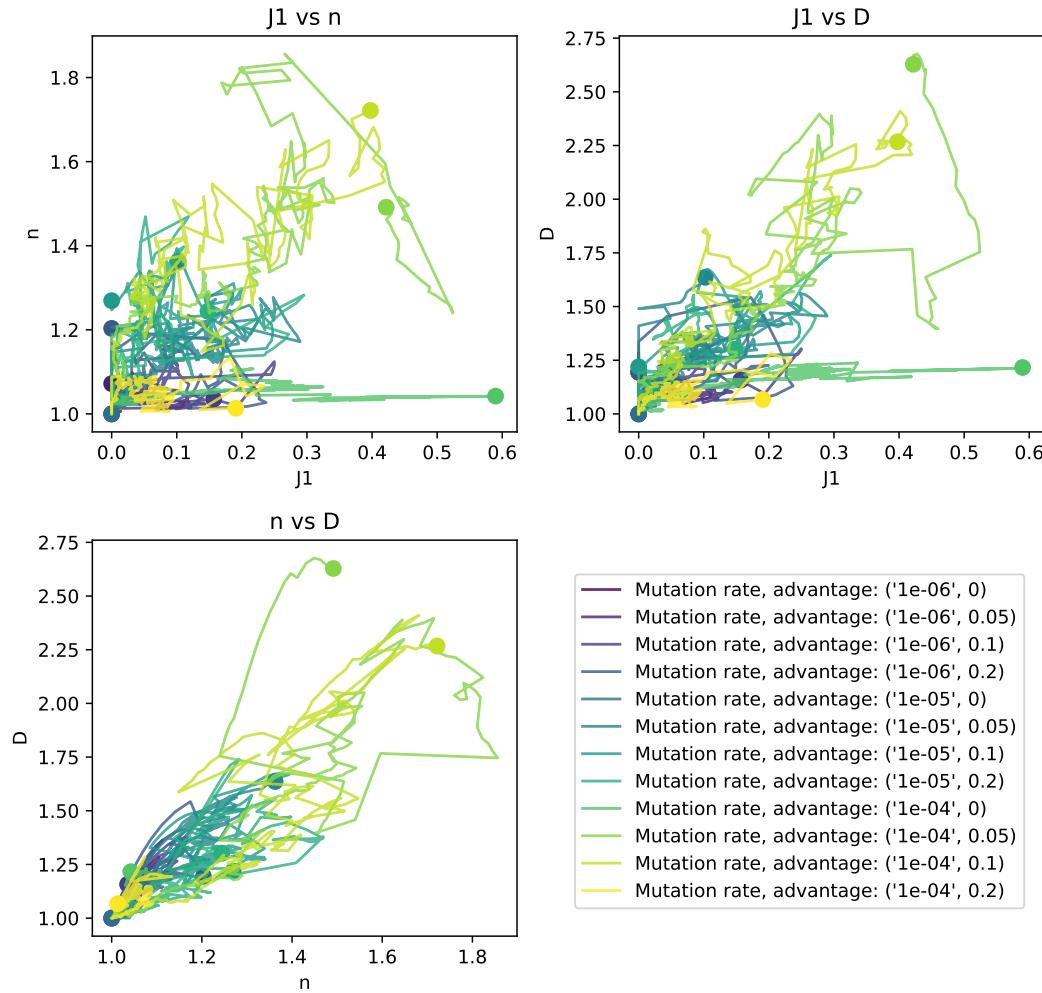
**Figure 3.4:** Average trajectories of the three indices for different values of driver mutation rate and selective advantage for invasive glandular tumours.

Average trajectories in index space for boundary growth

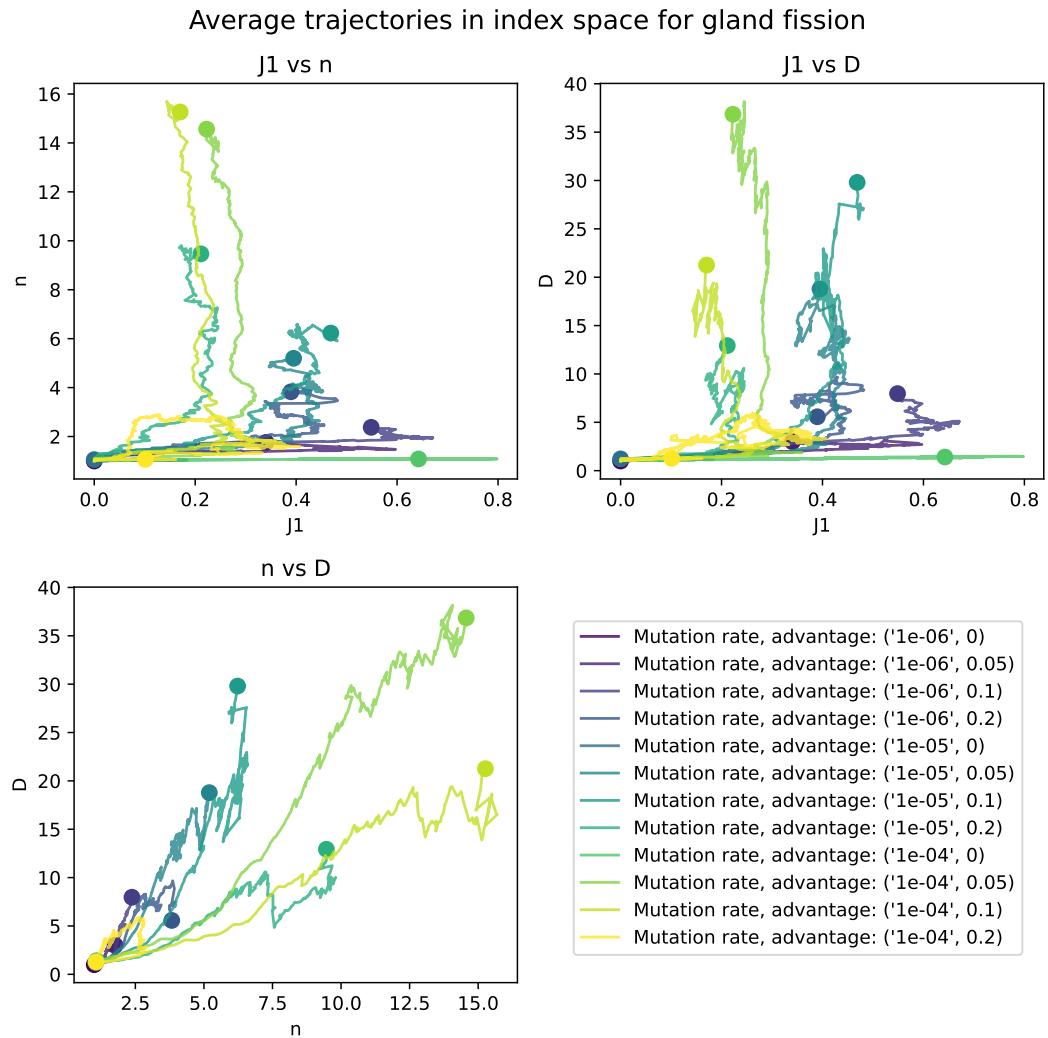


**Figure 3.5:** Average trajectories in index space for tumours progressing via boundary growth.

Average trajectories in index space for non-spatial tumours

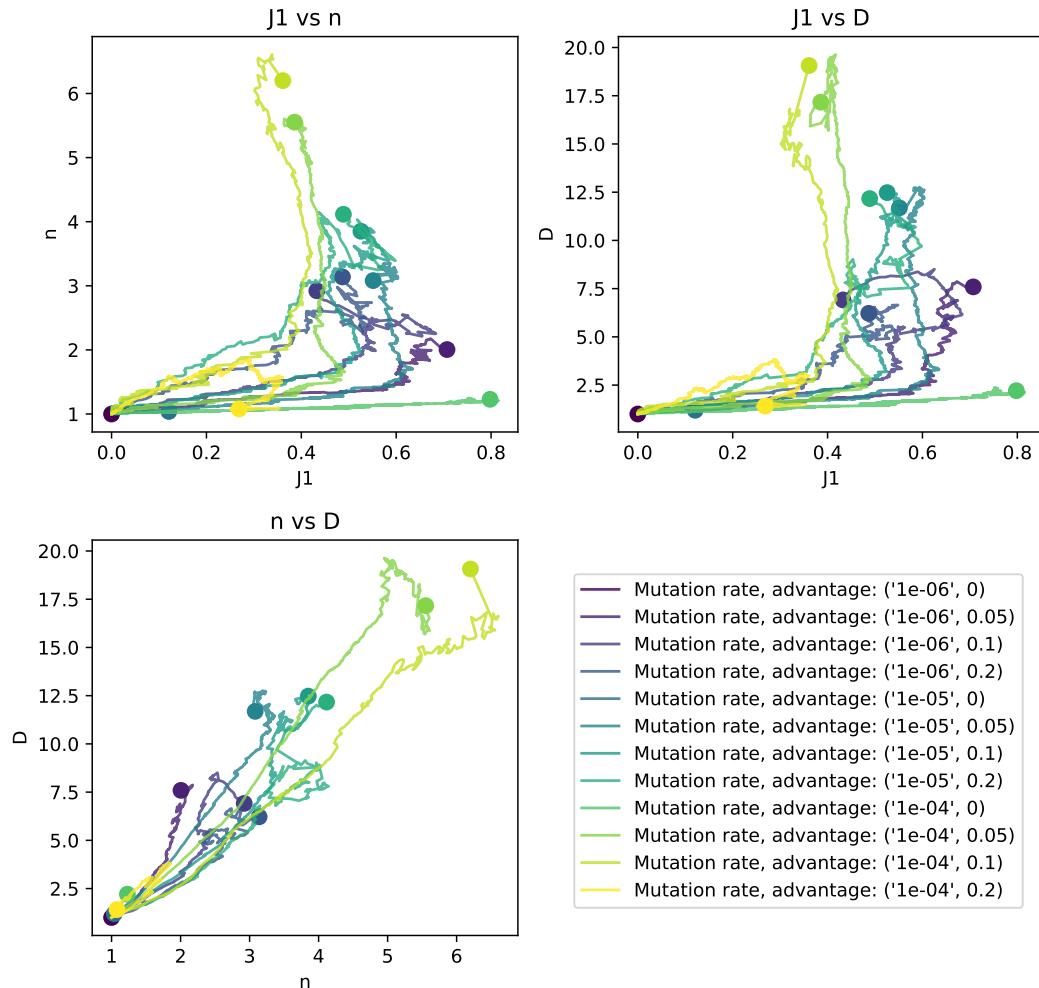


**Figure 3.6:** Average trajectories in index space for well-mixed cancer cell populations.



**Figure 3.7:** Average trajectories in index space for tumours progressing via gland fission.

Average trajectories in index space for invasive glandular tumours



**Figure 3.8:** Average trajectories in index space for invasive glandular tumours.

- what are the limitations of the approach? — clear starting point is data availability, but also general inter-patient variation of tumour progression
- next steps — further refining of the methods, sourcing and applying to more data (Kim's work in progress)



# **Chapter 4**

# **Agent-based workflow for inferring evolutionary parameters from molecular data using approximate Bayesian computation**

## **4.1 Introduction**

In chapter 3, we used a general 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 going beyond  $10^6$  cells. This means that using the model's outputs in comparison to or to draw inference from real data is not feasible.

There are a few ways to address this issue. For example, rather than simulating all clones in a tumour, we could take the approach of (Sottoriva et al. n.d.) and use demes (tumour glands) as the 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 have undergone a sweep the moment a mutation arises. If we wanted to study evolutionary dynamics on a finer scale we would need to at least simulate the dynamics of cell lineages, if not individual cells.

In (Gabbutt, Duran-Ferrer, Grant, Mallo, Nadeu, Househam, Villamor, Krali, Nordlund, Zenz, Campo, Lopez-Guillermo, Fitzgibbon, Barnes, Shibata, Martin-Subero & Graham n.d.), 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 does not suit our purposes either as it is not spatially explicit, and we need to be able to discriminate between tumour glands in colorectal cancer based on their spatial organisation.

## 4.2 Colorectal adenocarcinoma

Colorectal adenocarcinoma is a type of cancer that arises from the epithelial cells lining the colon. It is a solid tumour, which means we cannot ignore the spatial organisation of the cells. The tumour's origin, tumorigenesis, follows an accumulation of mutations in a cell's DNA, which leads to the cell's transformation from a healthy to malignant cell (Fearon & Vogelstein n.d.). Once the malignancy is established, the tumour forms hierarchical cell structures similar to those of normal tissue (Cernat et al. n.d.), organising into crypt-like glands. The tumour spread by the process of gland fission, which happens when a gland's population grows large enough to split into two glands (Preston et al. n.d.). This process is what informs our approach to simulating the tumour's growth.

## 4.3 Modifying existing simulation workflows

An analytical approach to modelling gland fission is difficult, due to the complexity of the process, as each gland has its own internal evolutionray dynamics in addition to spatial organisation in relation to other glands. Furthermore, fission splits a gland's population into two, with the daughter glands' spatial organisation further complicating the model. The issue arises when we consider the possibilities of boundary growth versus gland fissions being allowed throughout the tumour. If we limit the discussion to cells rather than glands, boundary growth seems to result in cell phylogenies with longer branches than unrestricted growth (Lewinsohn et al. n.d.), which would make the latter more consistent with neutral evolution (Sottoriva

et al. n.d., Williams et al. n.d.). However, as solid tumours often have spatial restrictions, one would not expect a purely unrestricted growth model to be realistic. To investigate how these spatial models affect a tumour's evolutionary dynamics and the corresponding methylation arrays, we used `demon` to simulate different scenarios. The output data was then assigned methylation arrays using `demonmeth`, a proprietary R package (figures 4.1 and 4.2).



Chapter\_4/figures/old\_workflow1.png

**Figure 4.1:** The old workflow for simulating tumour growth and assigning methylation arrays.

While informative on a high level, this approach had multiple issues. Firstly, the simulations were limited to a couple million cells which, after assigning across glands, resulted in a tumour too small to reasonably compare to real data. Secondly, previous work on fluctuating methylation arrays (Gabbott, Schenck, Weisenberger, Kimberley, Berner, Househam, Lakatos, Robertson-Tessi, Martin, Patel, Clark, Latchford, Barnes, Leedham, Anderson, Graham & Shibata n.d., Gabbott, Duran-Ferrer, Grant, Mallo, Nadeu, Househam, Villamor, Krali, Nordlund, Zenz, Campo, Lopez-Guillermo, Fitzgibbon, Barnes, Shibata, Martin-Subero & Graham n.d.) has shown that the methylation and demethylation rates are not necessarily equal, which is not a feature of the `demon` simulations, where we used passenger mutations as a proxy



**Figure 4.2:** The old workflow for simulating tumour growth and assigning methylation arrays.

for epigenetic changes. However, these issues may not have been insurmountable with clever tweaks to code. The final nail in the coffin came with the size of output files, as they were often unwieldy and difficult to handle by the R compiler. This meant that we would have to alter our approach in a way which allows for a leaner output, while keeping the spatial aspect of the evolutionary dynamics of the tumour intact. For that purpose, we developed a new agent-based model, `methdemon`.

## 4.4 Simulating fluctuating methylation arrays

### 4.4.1 Overview

- go over the simulation's inner workings
- provide estimates of running efficiency and memory requirements
- discuss possible upgrades and their potential computational costs

#### **4.4.2 Examples**

Provide example outputs (and their visualisations), parameter tables and a citation/link to the github repo.

### **4.5 Fluctuating methylation arrays through the lens of ABC**

#### **4.5.1 Overview**

- go over the `pyabc` package briefly (cite)
- explain the ABC workflow
- discuss computational costs and efficiency
- discuss whether this is the best approach (can we write down a likelihood for the problem?)

#### **4.5.2 Examples**

Provide example applications of the workflow to `methdemon` outputs - fit smaller simulations to a big one for example.



# Chapter 5

## Inferring evolutionary parameters of colorectal cancer from DNA methylation arrays

### 5.1 Introduction

- go over literature regarding colorectal adenocarcinoma evolution (pathology, big bang, etc.)
- discuss the relevant parts of the literature in the context of modelling
- explain how the data were collected (ask Darryl)

The model discussed in this report is the 1D version of the agent-based model `demon` developed by Rob. All rates are given relative to the birth rate which is assumed to be equal to 1 (as per the Gillespie algorithm).

### 5.2 Results

I performed the preliminary sensitivity analysis on a small set of simulations, more as a sanity check than a robust test. However, the results are interesting as the model is more sensitive to some parameters than expected. The parameters checked are ones controlling selective advantage, driver mutation rates, fCpG flipping probabilities, and gland fission rates.

### 5.2.1 A note on the fully neutral model

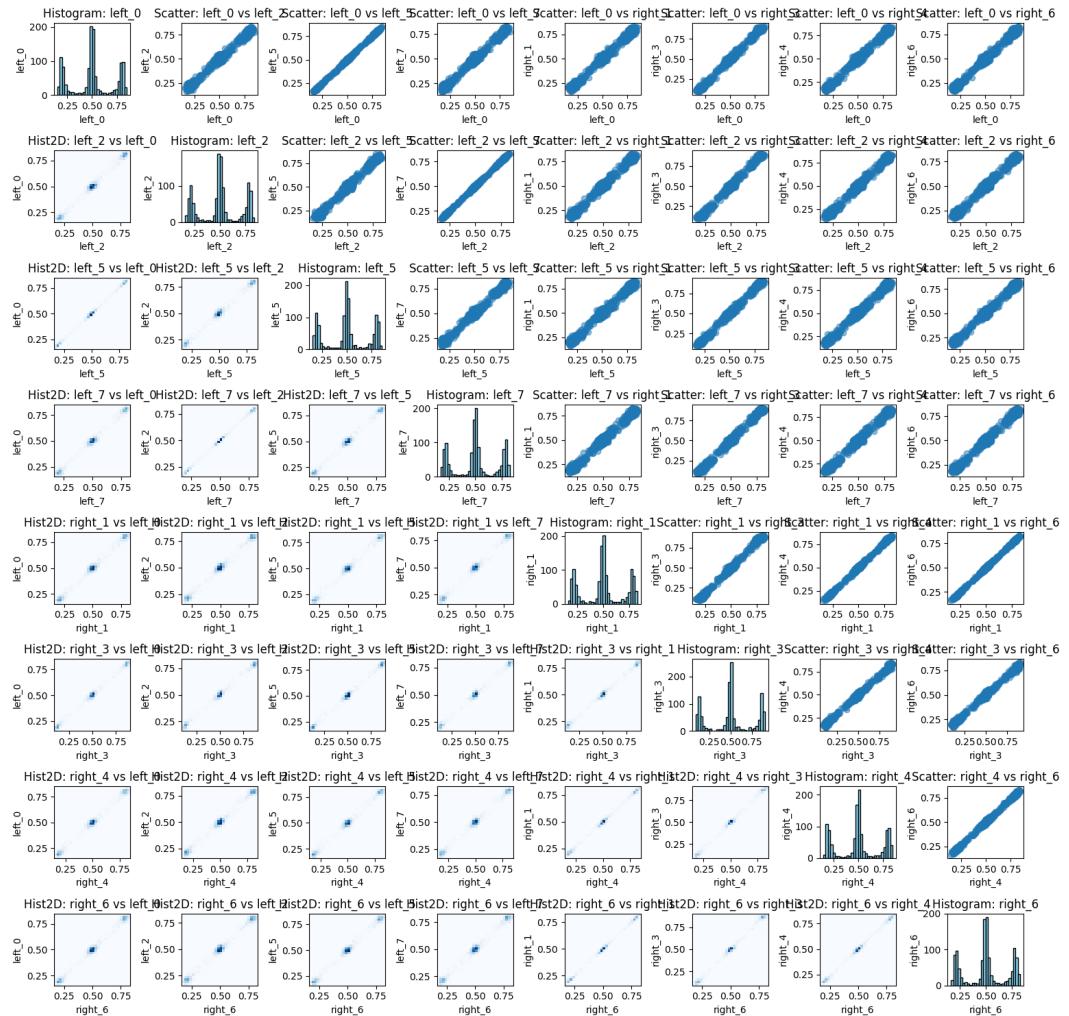
As discussed in previous meetings, the fully neutral model does not behave as expected. Instead, its outputs look like they are just oscillations around the initial fCpG array. This, I think, is due to the lack any preference for one lineage over any other within a gland, leading to a decoherence of the arrays after a long period of turnover, but without any resolution in space/time. However, even with selective advantage for driver mutations within glands, the glands themselves undergo fission in space neutrally as there is no competition for space in the model. The limitations of this assumption need to be discussed, but it seems reasonable for now.

### 5.2.2 Selective advantage

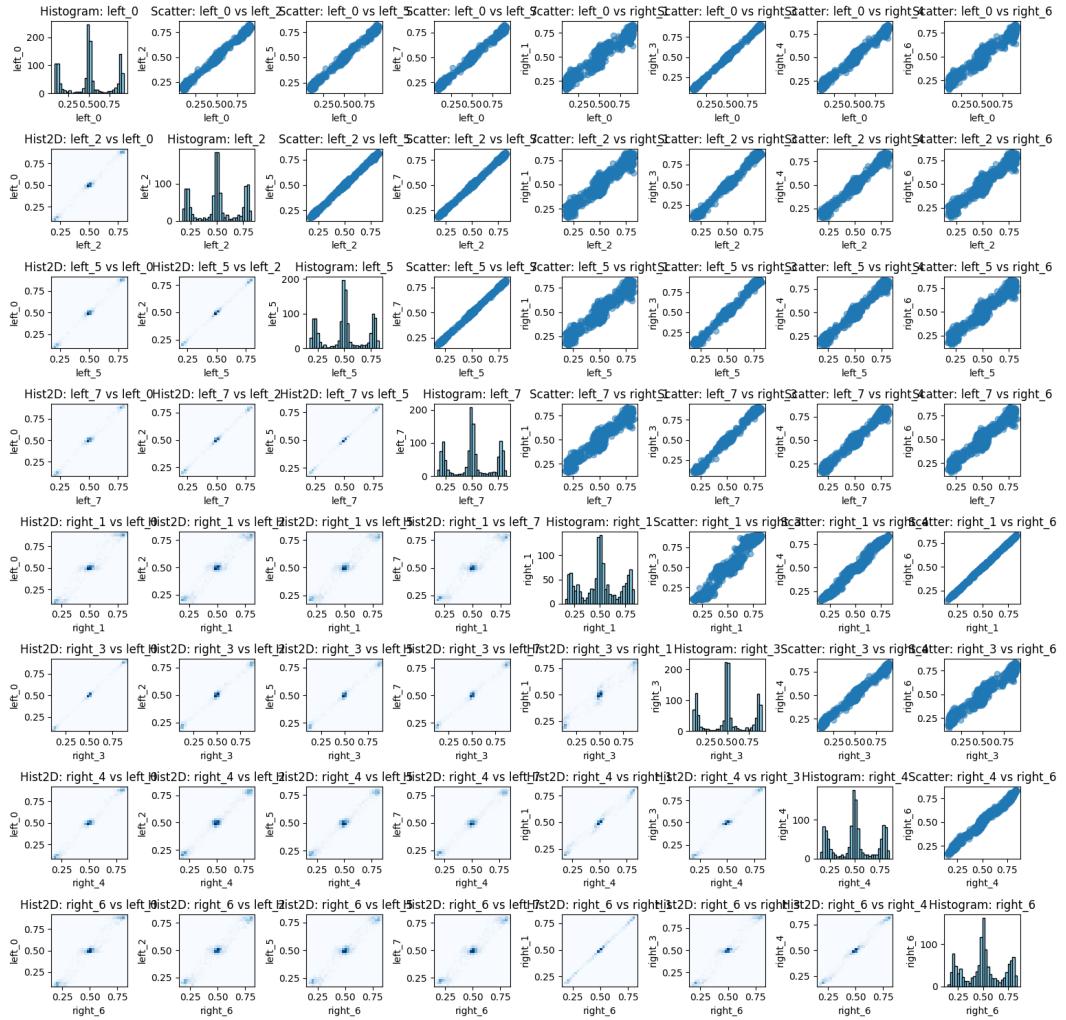
The selective advantage, accounted for in the model as

$$\lambda' = \lambda \left( 1 + s \left( \frac{\lambda}{\lambda_{max}} \right) \epsilon \right), \quad (5.1)$$

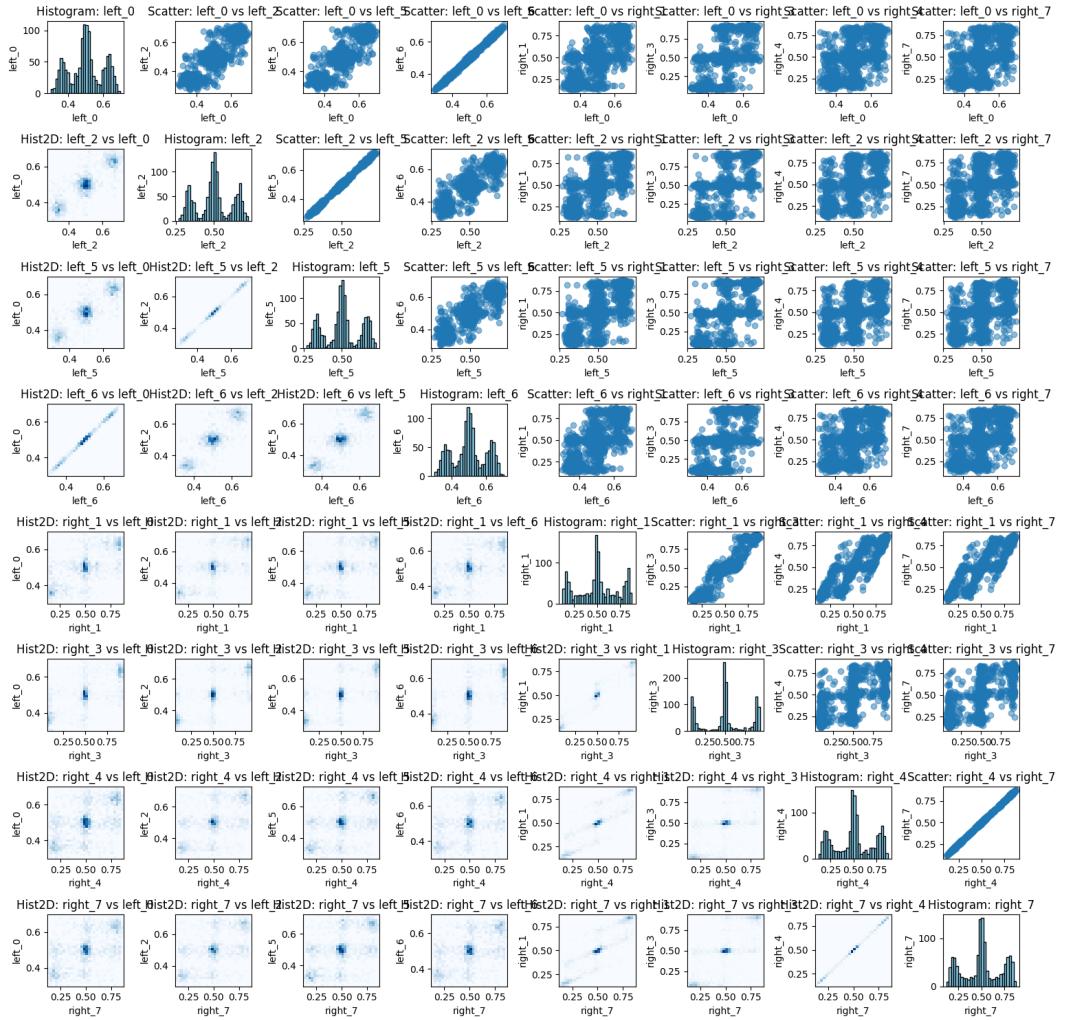
where  $\lambda$  is the birth rate before mutation,  $\lambda'$  is the birth rate after mutation,  $\lambda_{max}$  is the maximal allowed birth rate, and  $\epsilon$  is a unit exponentially distributed value. The values of  $s$  tested were 0.1, 0.2, and 0.3 with the other parameters kept at values given in table 5.1. In addition to the fully neutral model not recapitulating the patterns observed in data, it seems the weak selection regimes have a similarly hard time establishing lineages which would lead to decoupling between sides of the tumour.



**Figure 5.1:** correlation scatter plots, gland histograms and correlation heatmaps for  $s = 0.1$ .



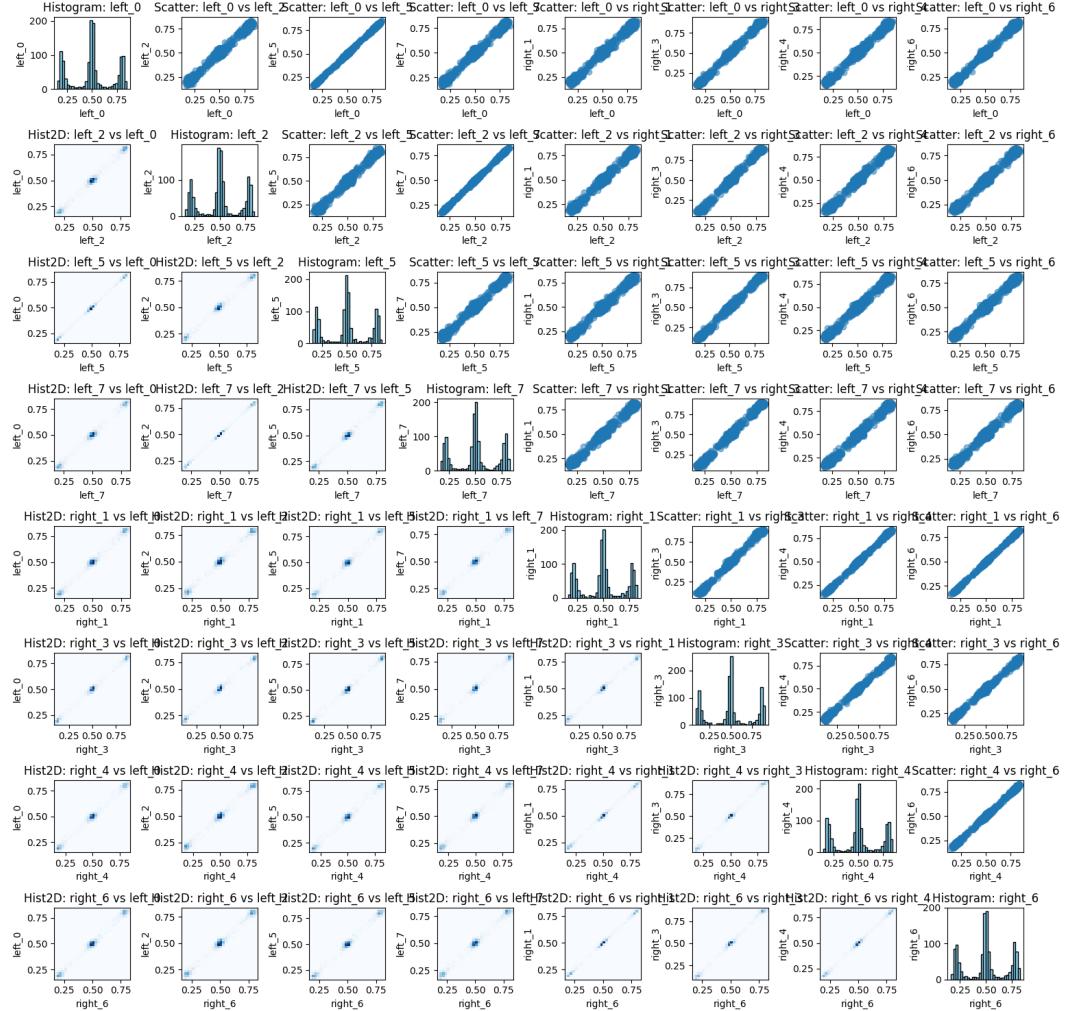
**Figure 5.2:** correlation scatter plots, gland histograms and correlation heatmaps for  $s = 0.2$ .



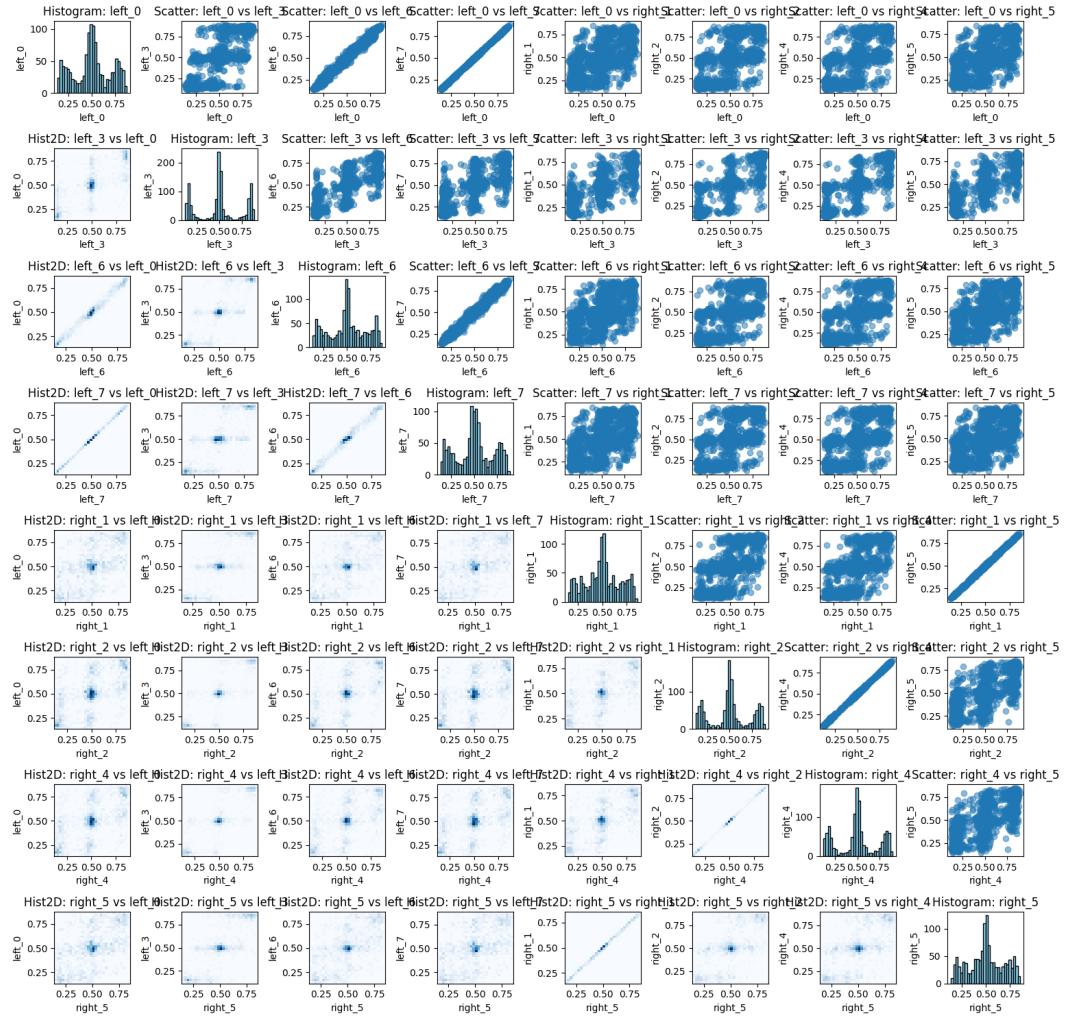
**Figure 5.3:** correlation scatter plots, gland histograms and correlation heatmaps for  $s = 0.3$ .

### 5.2.3 Driver mutation rates

The driver mutation rates were tested at  $10^{-6}$  and  $10^{-4}$  with the other parameters kept at values given in table 5.1, and the case for the mutation rate equal to  $10^{-5}$  covered in figure 5.3.



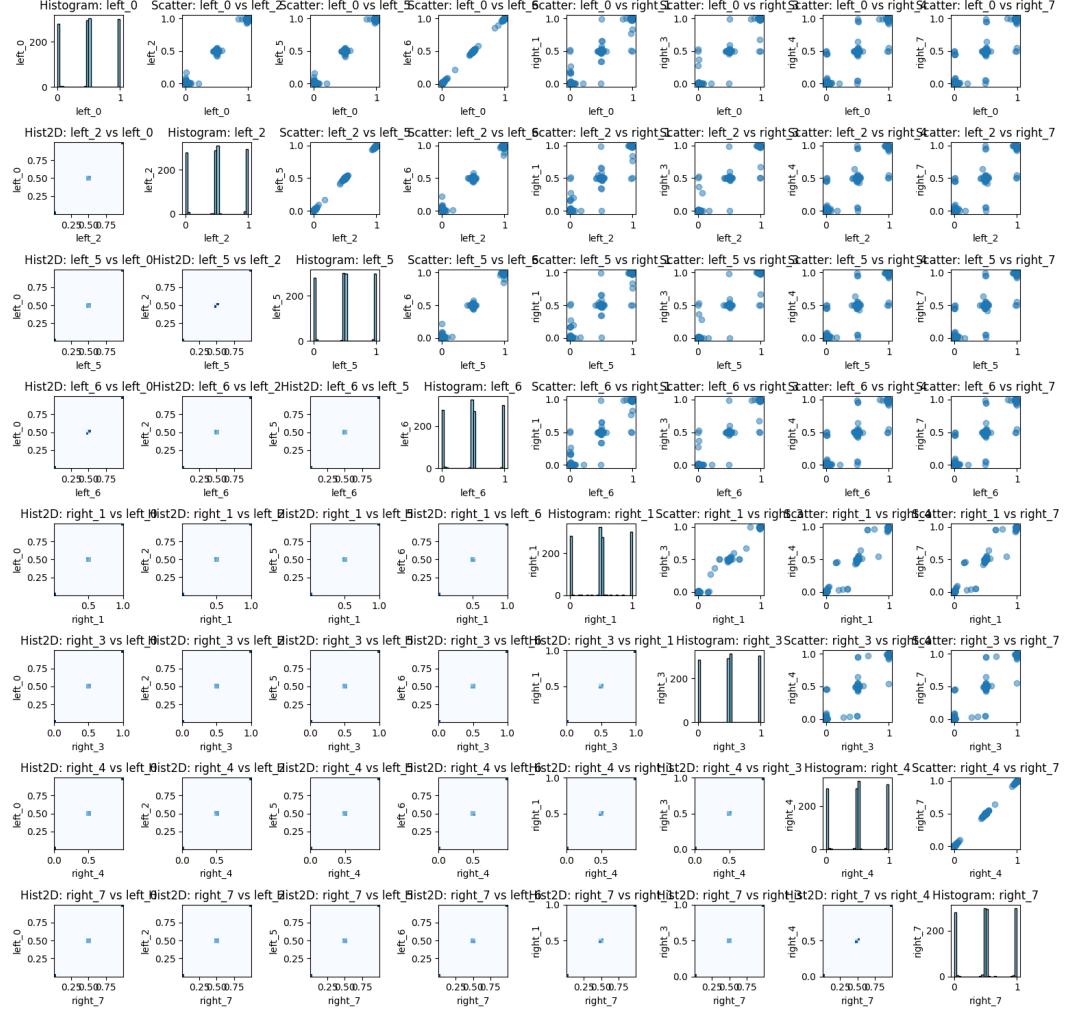
**Figure 5.4:** correlation scatter plots, gland histograms and correlation heatmaps for driver mutation rate  $10^{-6}$ .



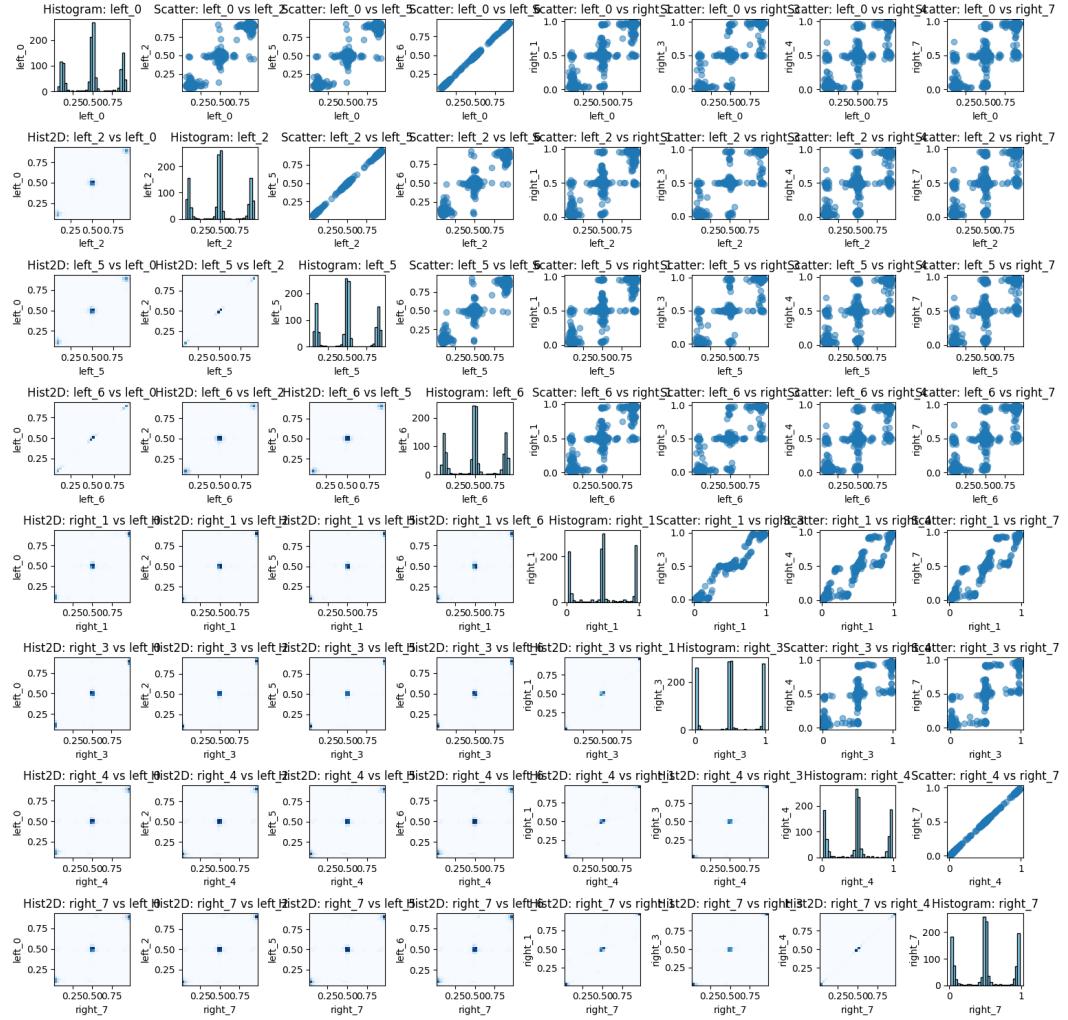
**Figure 5.5:** correlation scatter plots, gland histograms and correlation heatmaps for driver mutation rate  $10^{-4}$ .

### 5.2.4 fCpG flipping probabilities

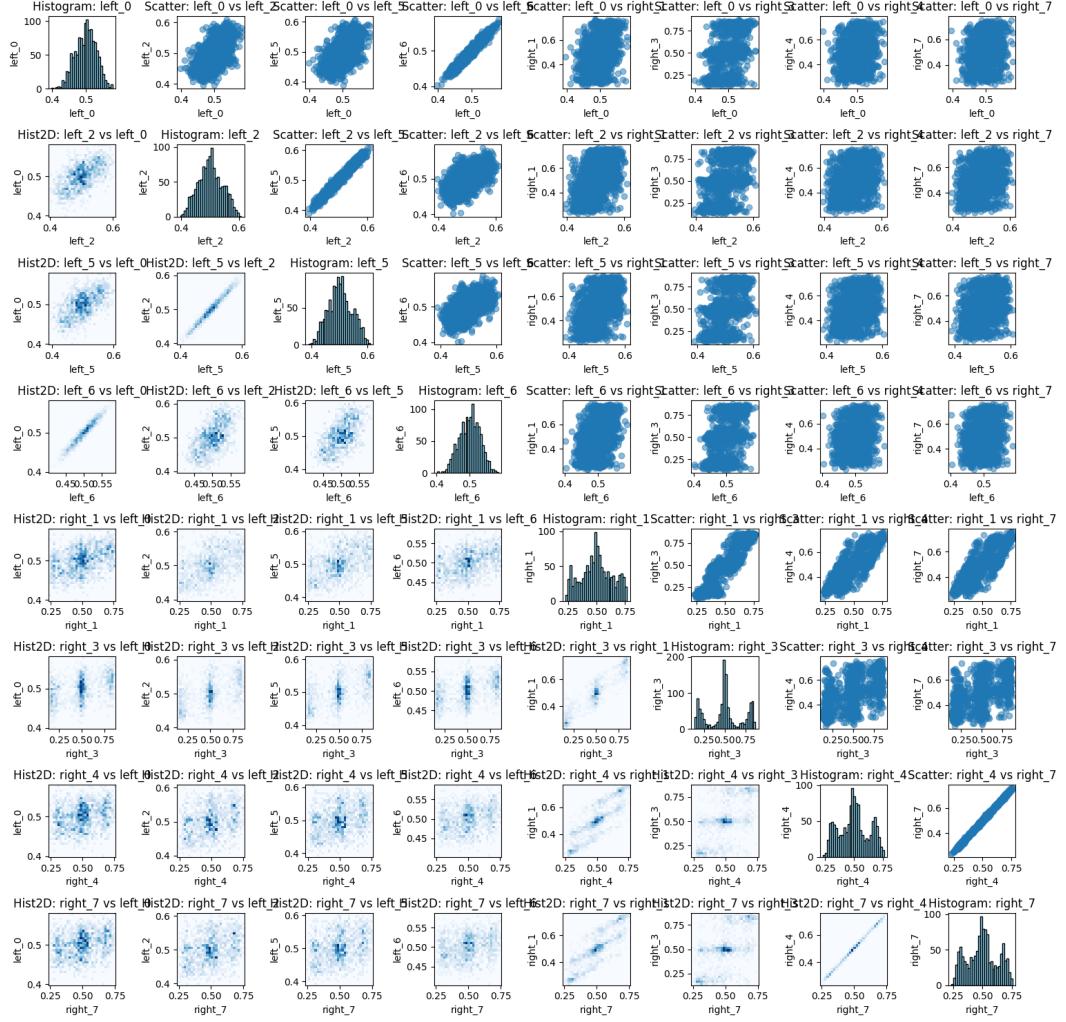
The fCpG flipping probabilities were tested at  $10^{-4}$ ,  $10^{-3}$ , and  $10^{-2}$  with the other parameters as in table 5.1. The case for  $5 \times 10^{-3}$  is covered by figure 5.3.



**Figure 5.6:** correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities  $10^{-4}$ .



**Figure 5.7:** correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities  $10^{-3}$ .

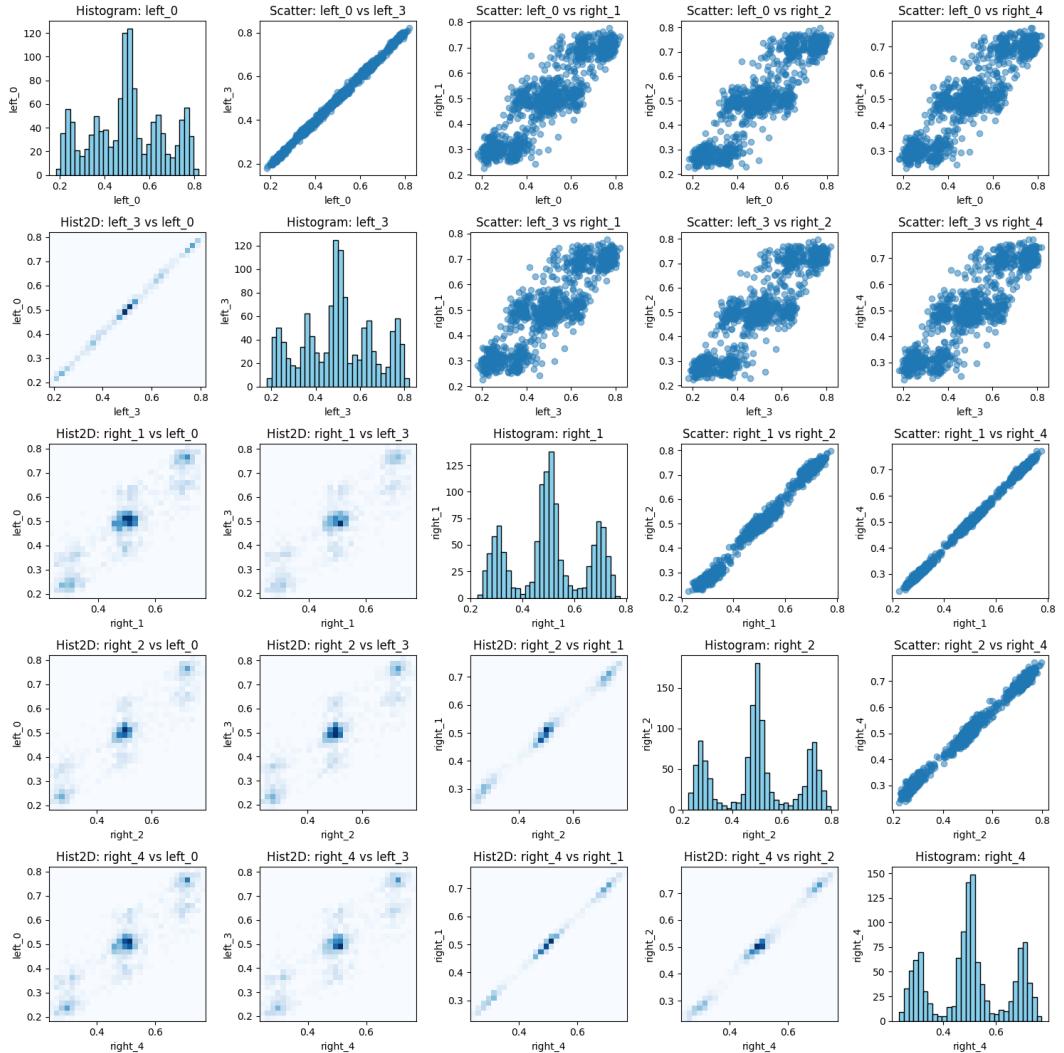


**Figure 5.8:** correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities  $10^{-2}$ .

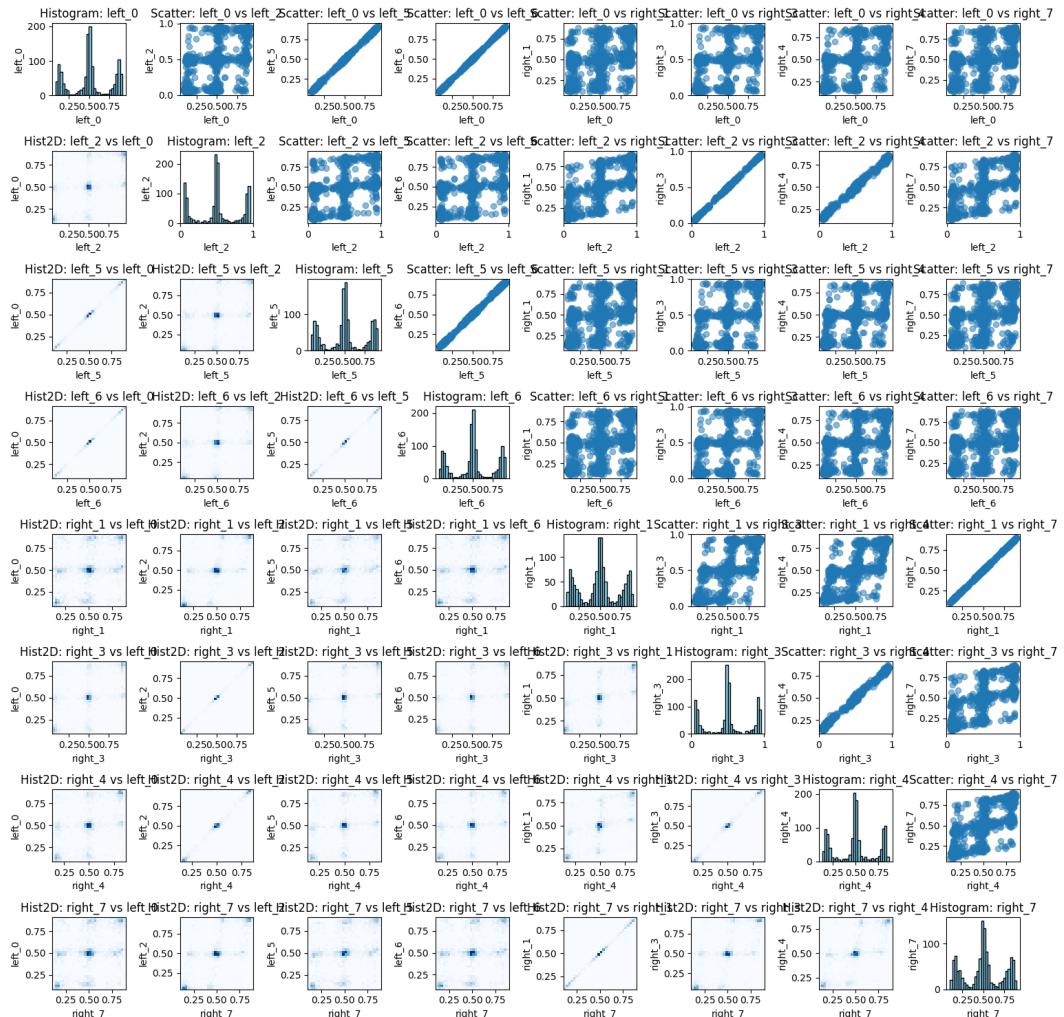
As expected, the higher probabilities lead to quicker decoupling, even between glands which are on the same side of the tumour. On the other hand, lower probabilities show very little change between the initial and final arrays.

### 5.2.5 Gland fission rates

The gland fission rates tested were 0.008, 0.08, and 0.8. The other parameters were kept as in table 5.1. The case of 0.008 did not produce any fissions and is not included below, and the case of 0.08 had a total of 4 fissions which resulted in 5 glands at the end of the simulation. The case for fission rates equal to 0.4 was covered in figure 5.3.



**Figure 5.9:** correlation scatter plots, gland histograms and correlation heatmaps for fission rates 0.008.



**Figure 5.10:** correlation scatter plots, gland histograms and correlation heatmaps for fission rates 0.08.

Nominally, faster fission rates should lead to less time spent in turnover before fission, and therefore less time for the fCpG sites to decouple. However, I set these simulations up with that in mind so that the glands spend around 40% of the simulated time in turnover.

### 5.3 Parameters

The loose default parameter settings used in the simulations are given in Table 5.1.

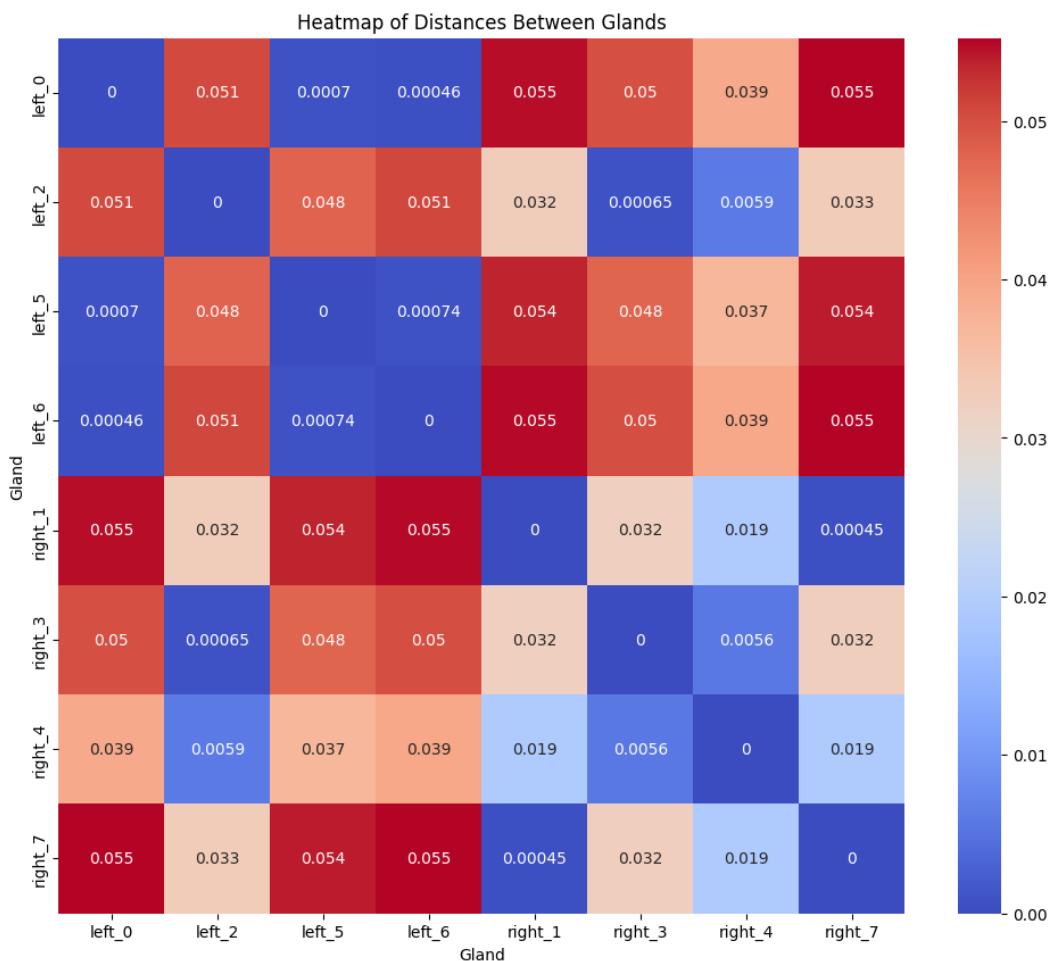
Parameter	Value
Driver mutation rate	$10^{-5}$
Methylation probability per fCpG site per cell division	$5 \times 10^{-3}$
Demethylation probability per fCpG site per cell division	$5 \times 10^{-3}$
Gland fission rate	0.4
Cells per gland	8192
fCpG loci per cell	1200
Selective advantage	0.3

**Table 5.1:** Default parameter values.

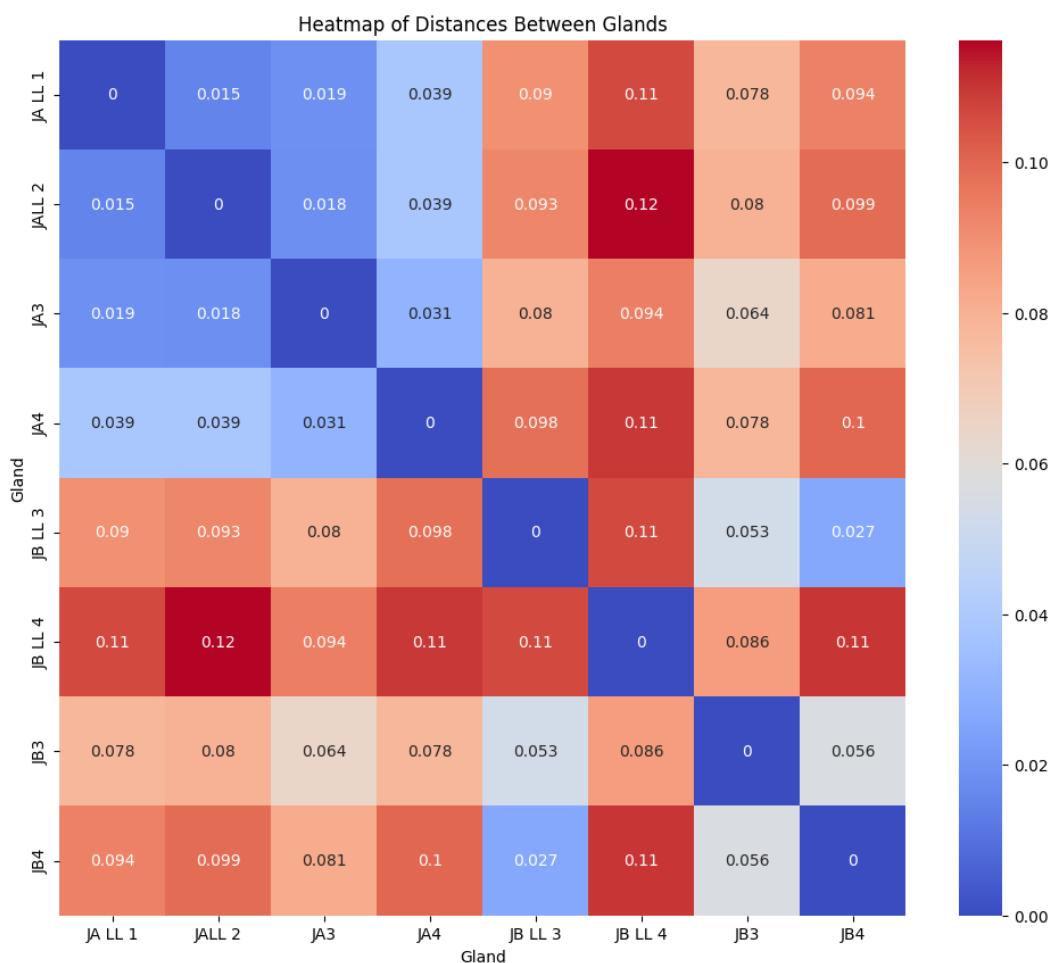
The values are educated guesses based on the two fCpG papers. The simulations were run for 50 Gillespie generations, which equated to tumours between 10 and 50 glands across. The tumours were allowed about 40% of the growth time in turnover. The simulations were run on my laptop, I am currently scaling the framework up for deployment on City’s computing cluster.

### 5.4 Distance functions

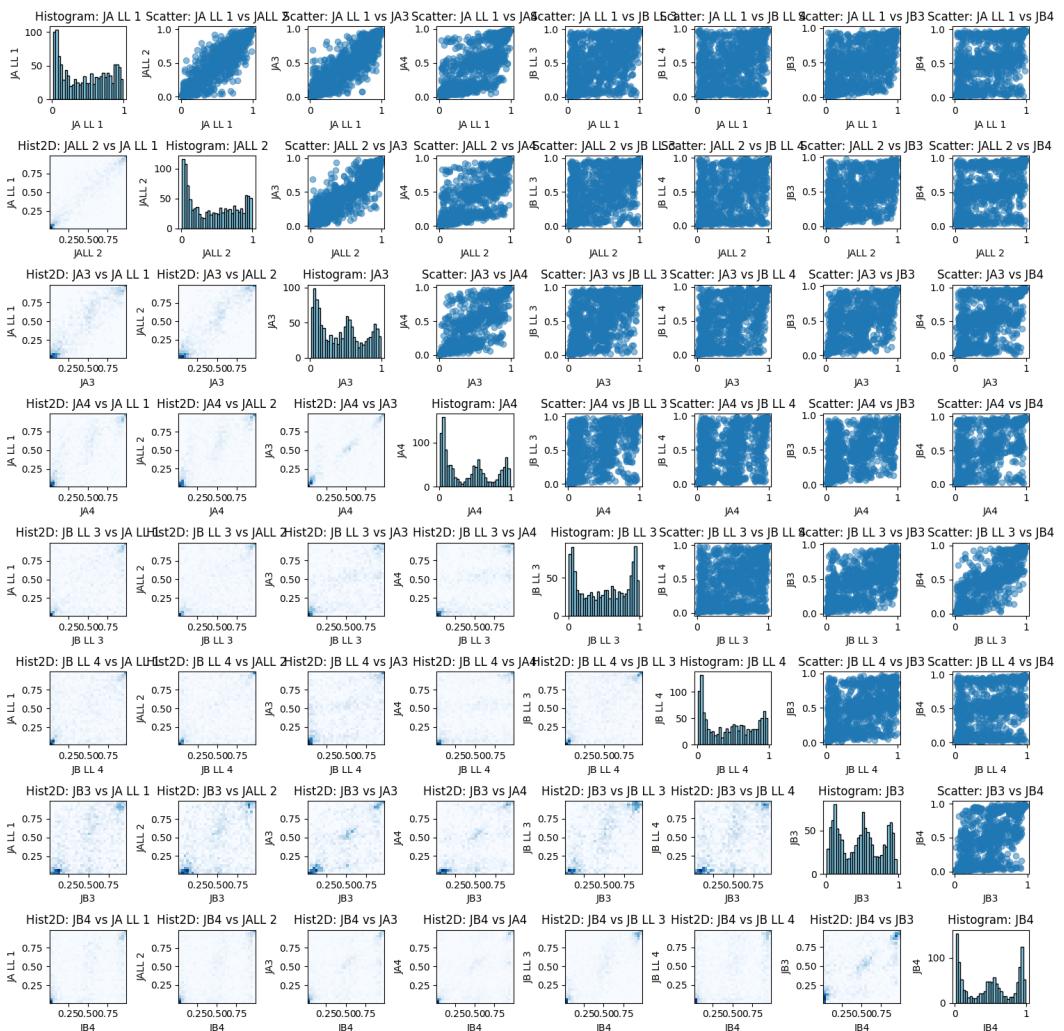
The basic distance functions I have started from are inspired by the Metropolitan distance. The distance between site  $A$  in gland 2 and site  $A$  in gland 2 is calculated as the classic Metropolitan distance with a modification that the values of  $A_1$  and  $A_2$  are put in bins based on their proximity to the values of 0, 0.5, and 1. The main difference between distance functions I’ve experimented with is adjusting the value added to the distance based on the difference between sites in different glands.



**Figure 5.11:** Distances between glands from figure 5.10.



**Figure 5.12:** Distances between glands from data set J.



**Figure 5.13:** Data set J whose distances are shown in 5.12.

## 5.5 Next steps/work in progress

Currently working on the following:

- Grid search over the above parameter space to check whether there are clusters which correspond to different tumour growth regimes.
- Refining the distance functions to better capture the differences between the arrays.

Next on the list:

- Inferring parameters from synthetic data based on the above two steps.
- Inference of parameters or qualitative properties of real data sets - depends on the results of the above.

### 5.5.1 Hypotheses

- `methdemon` recapitulates FMC patterns in colorectal cancer. **Test:** Extensive sensitivity analysis of `methdemon` and comparison to a different model (A/B model test with a simpler model, rule out trivial and, ideally, non-trivial models).
- `methdemon` reproduces evolutionary dynamics of colorectal cancer (effectively neutral). **Test:** Inferring parameters from simulations — under consideration are fission rate, mutation rate, time under turnover.
- stem cell hypothesis - assume expansion process for each lineage and draw all cells within glands from distribution (multinomial or whatever). polyclonal origin - easy to test (fully neutral).
- Spatial resolution is needed to recover evolutionary dynamics of colorectal cancer. **Test:** Compare `methdemon` to EVOFLUX (average over the data for each cancer to run the latter).
- FMC patterns imply evolutionary bottlenecks between distant glands in colorectal cancer. **Test:** Develop distance metric, run EVOFLUX on individual glands (or a variation of EVOFLUX). NOTE: ask Darryl if he can be more specific on which bottlenecks he means. Need specific things that can be implemented in the model.



## **Chapter 6**

## **Discussion**



## **Appendix A**

### **Title of the First Appendix**

Two possibilities for the appendices are presented in this template. The Appendix A is included in the main matter of the thesis after the `\appendix` command. This produces that the appendix input in the table of contents is labelled with the corresponding capital letter (in this case 'A'). The text 'Appendix A' will appear on top of the first page of the appendix, above the appendix's title, in case you have given a title to it.



# Appendix B

The Appendix B is included in the back matter of the thesis. No `\appendix` command is used. This produces that the appendix input in the table of contents is not labelled, neither with arabic numbers nor capital letters. The only text that will appear in the appendix title is that written in the `\chapter{}` command brackets. In this cases this is 'Appendix B'.



# Bibliography

Aldous, D. J. (n.d.), ‘Stochastic models and descriptive statistics for phylogenetic trees, from yule to today’, **16**(1), 23–34. Publisher: Institute of Mathematical Statistics.

**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>

Bak, M., Colyer, B., Manojlović, V. & Noble, R. (n.d.), ‘Warlock: an automated computational workflow for simulating spatially structured tumour evolution’.

**URL:** <http://arxiv.org/abs/2301.07808>

Bayes, M. & Price, M. (n.d.), *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. (n.d.), ‘A comparative review of dimension reduction methods in approximate bayesian computation’, **28**(2), 189–208. Publisher: Institute of Mathematical Statistics.

**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>

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. (n.d.), ‘Hybrid automata library: A flexible platform for hybrid modeling with real-time visualization’, **16**(3), e1007635. Publisher: Public Library of Science.

**URL:** <https://journals.plos.org/ploscombiol/article?id=10.1371/journal.pcbi.1007635>

Cardona, G., Mir, A. & Rossello, F. (n.d.), ‘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. (n.d.), ‘Colorectal cancers mimic structural organization of normal colonic crypts’, **9**(8), e104284.

**URL:** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4128715/>

Chomsky, N. (n.d.), *Syntactic structures*, Syntactic structures, Mouton. Pages: 116.

Colyer, B., Bak, M., Basanta, D. & Noble, R. (n.d.), ‘A seven-step guide to spatial, agent-based modelling of tumour evolution’.

**URL:** <http://arxiv.org/abs/2311.03569>

Fearon, E. R. & Vogelstein, B. (n.d.), ‘A genetic model for colorectal tumorigenesis’, **61**(5), 759–767.

**URL:** <https://linkinghub.elsevier.com/retrieve/pii/009286749090186I>

*Financial Burden of Cancer Care | Cancer Trends Progress Report* (n.d.).

**URL:** <https://progressreport.cancer.gov/after/economicburden>

Fischer, M. (n.d.), ‘Extremal values of the sackin tree balance index’, **25**(2), 515–541.

**URL:** <https://link.springer.com/10.1007/s00026-021-00539-2>

Fischer, M., Herbst, L., Kersting, S., Kühn, L. & Wicke, K. (n.d.), ‘Tree balance indices: a comprehensive survey’.

**URL:** <http://arxiv.org/abs/2109.12281>

Gabbatt, 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. (n.d.), ‘Evolutionary dynamics of 1,976 lymphoid malignancies predict clinical outcome’. Pages: 2023.11.10.23298336. Pages: 2023.11.10.23298336v1

**URL:** <https://www.medrxiv.org/content/10.1101/2023.11.10.23298336v1>

Gabbatt, 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. (n.d.), ‘Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues’, **40**(5), 720–730. Number: 5 Publisher: Nature Publishing Group.

**URL:** <https://www.nature.com/articles/s41587-021-01109-w>

Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M. & Macklin, P. (n.d.), ‘PhysiCell: An open source physics-based cell simulator for 3-d multicellular systems’, **14**(2), e1005991. Publisher: Public Library of Science.

**URL:** <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005991>

Gillespie, D. T. (n.d.), ‘Exact stochastic simulation of coupled chemical reactions’, **81**(25), 2340–2361. Publisher: American Chemical Society.

**URL:** <https://doi.org/10.1021/j100540a008>

Goh, G., Fuchs, M. & Zhang, L. (n.d.), ‘Two results about the sackin and colless indices for phylogenetic trees and their shapes’, **85**(6), 69.

**URL:** <https://link.springer.com/10.1007/s00285-022-01831-2>

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

kimverity (n.d.), ‘kimverity/RUIIndices’. original-date: 2023-09-27T09:57:19Z.

**URL:** <https://github.com/kimverity/RUIIndices>

Kirkpatrick, M. & Slatkin, M. (n.d.), ‘Searching for evolutionary patterns in the shape of a phylogenetic tree’, **47**(4), 1171–1181. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1558-5646.1993.tb02144.x>.

**URL:** <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1993.tb02144.x>

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

Knuth, D. E. (n.d.b), ‘Semantics of context-free languages’, **2**(2), 127–145.

**URL:** <https://doi.org/10.1007/BF01692511>

Lemant, J., Le Sueur, C., Manojlović, V. & Noble, R. (n.d.), ‘Robust, universal tree balance indices’, **71**(5), 1210–1224.

**URL:** <https://academic.oup.com/sysbio/article/71/5/1210/6567363>

Lewinsohn, M. A., Bedford, T., Müller, N. F. & Feder, A. F. (n.d.), ‘State-dependent evolutionary models reveal modes of solid tumour growth’, **7**(4), 581–596. Number: 4 Publisher: Nature Publishing Group.

**URL:** <https://www.nature.com/articles/s41559-023-02000-4>

Liao, J. G. & Berg, A. (n.d.), ‘Sharpening jensen’s inequality’.

**URL:** <http://arxiv.org/abs/1707.08644>

M. Coronado, T., Mir, A., Rosselló, F. & Rotger, L. (n.d.), ‘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/s12859-020-3405-1>

McKenzie, A. & Steel, M. (n.d.), ‘Distributions of cherries for two models of trees’, **164**(1), 81–92.

Mir, A., Rosselló, F. & Rotger, L. (n.d.), ‘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. (n.d.), ‘Sound colless-like balance indices for multifurcating trees’, **13**(9), 1–27. Publisher: Public Library of Science.

**URL:** <https://doi.org/10.1371/journal.pone.0203401>

Mooers, A. O. & Heard, S. B. (n.d.), ‘Inferring evolutionary process from phylogenetic tree shape’, **72**(1), 31–54.

**URL:** <https://www.journals.uchicago.edu/doi/10.1086/419657>

Nakano, S.-i. (n.d.), Tree enumeration, in M.-Y. Kao, ed., ‘Encyclopedia of Algorithms’, Springer, pp. 2252–2254.

**URL:** [https://doi.org/10.1007/978-1-4939-2864-4\\_26](https://doi.org/10.1007/978-1-4939-2864-4_26)

Nievergelt, J. (n.d.), ‘Binary search trees and file organization’, **6**(3), 195–207.

**URL:** <https://dl.acm.org/doi/10.1145/356631.356634>

Nievergelt, J. & Reingold, E. M. (n.d.), 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>

Noble, R. (n.d.), ‘demon’. original-date: 2019-03-22T10:29:24Z.

**URL:** [https://github.com/robjohnnoble/demon\\_model](https://github.com/robjohnnoble/demon_model)

Noble, R., Burley, J. T., Le Sueur, C. & Hochberg, M. E. (n.d.), ‘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. (n.d.), ‘Spatial structure governs the mode of tumour evolution’, **6**(2), 207–217.

**URL:** <https://www.nature.com/articles/s41559-021-01615-9>

Noble, R. & Verity, K. (n.d.), ‘A new universal system of tree shape indices’. Pages: 2023.07.17.549219 Section: New Results.

**URL:** <https://www.biorxiv.org/content/10.1101/2023.07.17.549219v2>

O’Meara, B. C. (n.d.), ‘Evolutionary inferences from phylogenies: A review of methods’, **43**(1), 267–285.

**URL:** <http://www.annualreviews.org/doi/10.1146/annurev-ecolsys-110411-160331>

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

Prangle, D. (n.d.), ‘Adapting the ABC distance function’, **12**(1), 289–309. Publisher: International Society for Bayesian Analysis.

**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. (n.d.), ‘Bottom-up histogenesis of colorectal adenomas: origin in the monocryptal adenoma and initial expansion by crypt fission’, **63**(13), 3819–3825.

Rosen, D. E. (n.d.), ‘Vicariant patterns and historical explanation in biogeography’, **27**(2), 159–188. Publisher: [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.].

**URL:** <https://www.jstor.org/stable/2412970>

Scott, J. G., Maini, P. K., Anderson, A. R. A. & Fletcher, A. G. (n.d.), ‘Inferring tumour proliferative organisation from phylogenetic tree measures in a computational model’.

**URL:** <http://biorxiv.org/lookup/doi/10.1101/334946>

Shannon, C. E. (n.d.), ‘A mathematical theory of communication’, **27**(3), 379–423. Conference Name: The Bell System Technical Journal.

**URL:** <https://ieeexplore.ieee.org/document/6773024>

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. (n.d.), ‘A big bang model of human colorectal tumor growth’, **47**(3), 209–216. Number: 3 Publisher: Nature Publishing Group.

**URL:** <https://www.nature.com/articles/ng.3214>

Sottoriva, A. & Tavaré, S. (n.d.), 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/10.1007/978-3-7908-2604-3\\_5](https://link.springer.com/10.1007/978-3-7908-2604-3_5)

Steel, M. & McKenzie, A. (n.d.), ‘Properties of phylogenetic trees generated by yule-type speciation models q’, p. 22.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (n.d.), ‘Inferring coalescence times from DNA sequence data’, **145**(2), 505–518.

**URL:** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1207814/>

Tsakalidis, A. K. (n.d.), ‘Maintaining order in a generalized linked list’, **21**(1), 101–112.

**URL:** <https://doi.org/10.1007/BF00289142>

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. (n.d.),

‘Identification of neutral tumor evolution across cancer types’, **48**(3), 238–244.

**URL:** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4934603/>

Wong, C. K. & Nievergelt, J. (n.d.), ‘Upper bounds for the total path length of binary trees’, **20**(1), 1–6.

**URL:** <https://dl.acm.org/doi/10.1145/321738.321739>

Yule, G. U. (n.d.), ‘II.—a mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f. r. s’, **213**(402), 21–87. Publisher: Royal Society.

**URL:** <https://royalsocietypublishing.org/doi/10.1098/rstb.1925.0002>