

Mathematical Classification of the Modes of Tumour Evolution

Veselin Manojlović

Doctor of Philosophy



School of Science and Technology

Department of Mathematics

September 2023

Contents

Contents	iii
List of Figures	vii
List of Tables	xi
Acknowledgements	xiii
Declaration	xv
Abstract	xvii
1 Introduction	1
1.1 Mathematical oncology	2
1.1.1 Introduction	2
1.1.2 Mathematical models of tumour evolution	3
1.2 Trees and their applications	4
1.2.1 Introduction	4
1.2.2 Quantifying tree balance	5
1.3 Agent-based modelling in oncology	7
1.3.1 Introduction	7
1.3.2 The <code>demon-warlock</code> framework	8
1.4 Likelihood-free inference	11
1.4.1 Introduction	11
1.4.2 Approximate Bayesian Computation	12
1.5 Fluctuating methylation clocks	13
1.6 Aims and objectives	14
1.6.1 Hypotheses	14

1.6.2	Aims	15
1.6.3	Objectives	15
2	Expected and extreme values of universal tree balance index J^1	17
2.1	Introduction	17
2.2	Prerequisites	18
2.2.1	Preliminary definitions from systematic biology	18
2.2.2	Preliminary definitions from computer science	19
2.3	Results	20
2.3.1	J^1 unites and generalises prior notions of tree balance	20
2.3.2	J^1 is maximised by Huffman coding	22
2.3.3	Expected value of J^1 under simple evolutionary processes	24
2.3.4	Analytic properties of the J^1 index	32
2.3.5	Properties of J^1 on different tree families	33
2.3.6	Behaviour as $n \rightarrow \infty$	36
2.4	Discussion	40
3	Tracking cancer evolution <i>in silico</i> via evolutionary indices	41
3.1	Introduction	41
3.1.1	Why even bother with indices?	42
3.1.2	A 3-dimensional index space — trees with uniform branch lengths	42
3.1.3	A general set of indices — any rooted tree	43
3.2	Tree resolution	44
3.2.1	3-dimensional index space	44
3.3	Computational methods	45
3.3.1	Agent-based modelling framework - <i>warlock/demon</i>	45
3.4	Results	46
3.4.1	Sensitivity of evolutionary mode to parameter values	46
3.5	Discussion	46
4	Agent-based workflow for inferring evolutionary parameters from molecular data using approximate Bayesian computation	57
4.1	Introduction	57
4.2	Background on colorectal cancer — model assumptions	58

4.2.1	Colorectal cancer evolution	58
4.2.2	Fluctuating methylation arrays	59
4.3	Existing simulation workflows	61
4.4	Simulating fluctuating methylation arrays	62
4.4.1	Overview	62
4.4.2	Examples	63
4.5	Fluctuating methylation arrays through the lens of ABC	63
4.5.1	Overview	63
4.5.2	Examples	63
5	Inferring evolutionary parameters of colorectal cancer from DNA methylation arrays	65
5.1	Introduction	65
5.2	Results	65
5.2.1	A note on the fully neutral model	66
5.2.2	Selective advantage	66
5.2.3	Driver mutation rates	70
5.2.4	fCpG flipping probabilities	72
5.2.5	Gland fission rates	75
5.3	Parameters	77
5.4	Distance functions	77
5.5	Next steps/work in progress	81
5.5.1	Hypotheses	81
6	Discussion	83
A	Title of the First Appendix	85
Appendix B		87
Bibliography		89

List of Figures

1.1	A J^1 values for trees generated under the Yule process and the uniform model. Solid grey curves represent the approximate expected values, and the dashed lines the 5th and 95th percentiles. B J^1 values for 100 random trees on 16 leaves using the alpha-gamma model, with $\alpha \sim \text{Unif}(0, 1)$ and $\gamma \sim \text{Unif}(0, \alpha)$. The values were calculated before and after applying a 1% population threshold, i.e. removing all leaves with sizes smaller than 0.01 times the total population. C Normalised Sackin index values for the same trees as in B	6
1.2	Event hierarchy in the <code>demon-warlock</code> framework. Figure reproduced from (Bak et al. n.d.) with the authors' permission.	9
1.3	Example output from <code>demon</code> , visualised using the <code>demonanalysis</code> 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.	10
2.1	A simple example of a binary search tree over the set of labels $S = \{2, 3, 4, 5, 6, 9, 11, 13\}$	23
2.2	Comparison of probabilities for generation of trees on 4 leaves under the Yule and uniform models. A: Arrows show generation under the Yule model. Each of the trees shown on 4 leaves has the same probability under the uniform model. B: Comparison of probabilities of tree topologies on 4 leaves under the Yule and uniform models. . .	25

2.3 Top row: True values of $\mathbb{E}(J^1)$ for up to 10 leaves were calculated manually, and the approximations up to 128 leaves were calculated as $n \log_2 n / \mathbb{E}(I_S)$. A — uniform model, B — Yule model. Bottom row: The Jensen gap of $\mathbb{E}(J^1)$ calculated for trees up to 128 leaves under the uniform model (C), and the Yule model (D). The size of the gap is calculated as the difference between the true and approximate expected value, with the gaps for 2 and 3 leaves equal to zero as there is only one possible bifurcating tree shape for each of those values. Refer to table 2.1 for numerical values of the gap size for the first several values of n . The red crosses in A and B represent sample mean J^1 values for 100000 trees generated under the uniform model and Yule process, and the difference between the approximate gap size and the sample mean, with standard error represented by error bars, in C and D .	29
2.4 The convergence of the upper bound to $4/3\pi$ is much slower than the convergence of the lower bound to 0, and the maximum it reaches over the plotted range is 0.0604 for $n = 128000$. The red crosses, as in figure 2.3, suggest convergence of the gap size.	30
2.5 Higher variance in the uniform model leads to a non-zero upper bound on the Jensen gap. Shown are frequencies of J^1 values on 10-leaf trees generated under the Yule and uniform models. The dashed lines represent the true expected value of J^1 , and the dotted lines the approximate value calculated as $\frac{n \log_2 n}{\mathbb{E}(I_S)}$	31
2.6 By including the node-balance function W^1 in J^1 , we allow for the possibility of perfectly balanced caterpillars (left) and less balanced fully symmetric trees (right) based on the node size distribution in the tree. The leaf sizes in these two trees are identical, with a ratio 4 : 2 : 1 : 1 from largest to smallest.	33
2.7 If we limit our search to leafy trees with equal leaf sizes, the least balanced tree on a given number of leaves is not necessarily the caterpillar. Pictured are the caterpillar trees on 6 and 9 leaves, as well as minimally balanced brooms for 6 and 9 leaves, with corresponding J^1 values.	34

2.8	The labels used in the figures are as above - n for number of leaves, k for number of leaves in the broom head, $r = n/k$. A: Value of r for which the minimum value of J^1 is obtained on leafy trees. Trees on n leaves which satisfy $r = \frac{n+a}{2n}$, for $a = 0, 1, 2, \dots$ lie on the dashed grey lines. The inset plot shows $k = rn$, the number of leaves attached at the broom head. B: Comparison of true and approximate values of J^1 for the caterpillar and minimally balanced broom trees as a function of n . C: The difference between values of J^1 of the minimally balanced broom and the caterpillar trees.	37
2.9	Values of J^1 on trees of different sizes calculated using equation (2.34) for different values of $r = k/n$. The dashed lines are at values of r which minimise J^1	39
3.1	Average trajectories of the three indices for different values of driver mutation rate and selective advantage for tumours progressing via boundary growth.	47
3.2	Average trajectories of the three indices for different values of driver mutation rate and selective advantage for well-mixed cancer cell populations.	48
3.3	Average trajectories of the three indices for different values of driver mutation rate and selective advantage for gland fission.	49
3.4	Average trajectories of the three indices for different values of driver mutation rate and selective advantage for invasive glandular tumours.	50
3.5	Average trajectories in index space for tumours progressing via boundary growth.	51
3.6	Average trajectories in index space for well-mixed cancer cell populations.	52
3.7	Average trajectories in index space for tumours progressing via gland fission.	53
3.8	Average trajectories in index space for invasive glandular tumours. .	54
4.1	The old workflow for simulating tumour growth and assigning methylation arrays.	61
4.2	The old workflow for simulating tumour growth and assigning methylation arrays.	62

5.1	correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.1$.	67
5.2	correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.2$.	68
5.3	correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.3$.	69
5.4	correlation scatter plots, gland histograms and correlation heatmaps for driver mutation rate 10^{-6} .	70
5.5	correlation scatter plots, gland histograms and correlation heatmaps for driver mutation rate 10^{-4} .	71
5.6	correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities 10^{-4} .	72
5.7	correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities 10^{-3} .	73
5.8	correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities 10^{-2} .	74
5.9	correlation scatter plots, gland histograms and correlation heatmaps for fission rates 0.008.	75
5.10	correlation scatter plots, gland histograms and correlation heatmaps for fission rates 0.08.	76
5.11	Distances between glands from figure 5.10.	78
5.12	Distances between glands from data set J.	79
5.13	Data set J whose distances are shown in 5.12.	80

List of Tables

2.1 Comparison of exact and approximate expected values of J^1 and I_S under the Yule and uniform models.	32
5.1 Default parameter values.	77

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^AT_EX 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, 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, we have found 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 through connections with data structures in computer science. Second, we derive upper bounds on the error of the expected value approximations for the Yule process and the uniform model. Finally, we investigate the minimal values of J^1 in important special cases, with special emphasis on 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 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 tailor-made 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 Mathematical oncology

TO DO: Better references in the introduction subsection. Write the modelling subsection

1.1.1 Introduction

Cancer emergence and progression is an evolutionary process (Nowell n.d., ?). This statement is now widely accepted, and the applications of quantitative methods found in evolutionary biology in cancer research are numerous (Rockne et al. n.d., Yin et al. n.d., Kourou et al. n.d.). The, now well established, area of mathematical oncology is informed by clinicians, computer scientists, mathematicians of all flavours, and biologists alike (Bull & Byrne n.d.), which has led to a rapid development of more specific avenues of research spanning from the initiation of the disease (Paterson et al. n.d.) to the optimisation of therapy protocols (?). This is a perfect

reflection of the complexity of the disease itself, as its rapid evolution, heterogeneity and constraints on how much information we can obtain from a patient take the combined efforts of thousands of scientists. Mathematics plays its own role in this effort, providing a common language through rigour and methods development (), and frameworks for the interpretation of data () .

1.1.2 Mathematical models of tumour evolution

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

As mentioned earlier, the applications of mathematics in oncology are diverse. Thus, our focus over the course of this PhD has been on modelling tumour evolution and progression from its early stages up to and excluding treatment. This makes the problem more of an exercise in population dynamics than strict oncology, as underlying assumptions of such models tend to focus less on the microenvironment impact and more on how mutations accumulate and spread in the tumour. A good example of one such model is the Big Bang model of tumour growth (Sottoriva et al. n.d.). Informed by multi-site sequencing, the authors' hypothesis was that colorectal cancer evolves neutrally after an initial period of rapid expansion and selection. Much like cosmic microwave background radiation is unevenly distributed across the observable universe, they observed an asymmetrical distribution of mutations across the tumour spheroid. This inspired a spatial branching process model based on gland fission, with each tumour gland approximated to rapid fixation in the event of a driver mutation, which showed good agreement with the data. A follow-up paper (Williams et al. n.d.) ignited a debate on neutral evolution in exponentially growing tumours within the community (Tarabichi et al. n.d., ?, Heide et al. n.d., Bozic et al. n.d.). However, theoretical considerations of the two-level model compared to the neutral model did, in fact, show that it is possible to distinguish the two based on

mutation frequency spectra (Tung & Durrett n.d.).

1.2 Trees and their applications

1.2.1 Introduction

In the most general sense, a tree is a connected graph with no cycles. In this thesis, when a tree is mentioned, we refer to a rooted tree, as formally defined in section ??.

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 (actually - pedigrees and family trees, make this more specific), 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 & Beerewinkel 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 knowledge between the fields of computer science and biology, with certain results being rediscovered nigh on half a century later, as discussed in section ??.

1.2.2 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 (ambiguous pronous, need to be very specific on what who's done, attribute appropriately) 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 1.1B, C). 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, and tested their accuracy numerically for trees with up to 128 leaves (figure 1.1A). In the same study, we proved that caterpillar trees minimise J^1 among bifurcating trees but not when larger outdegrees are permitted.

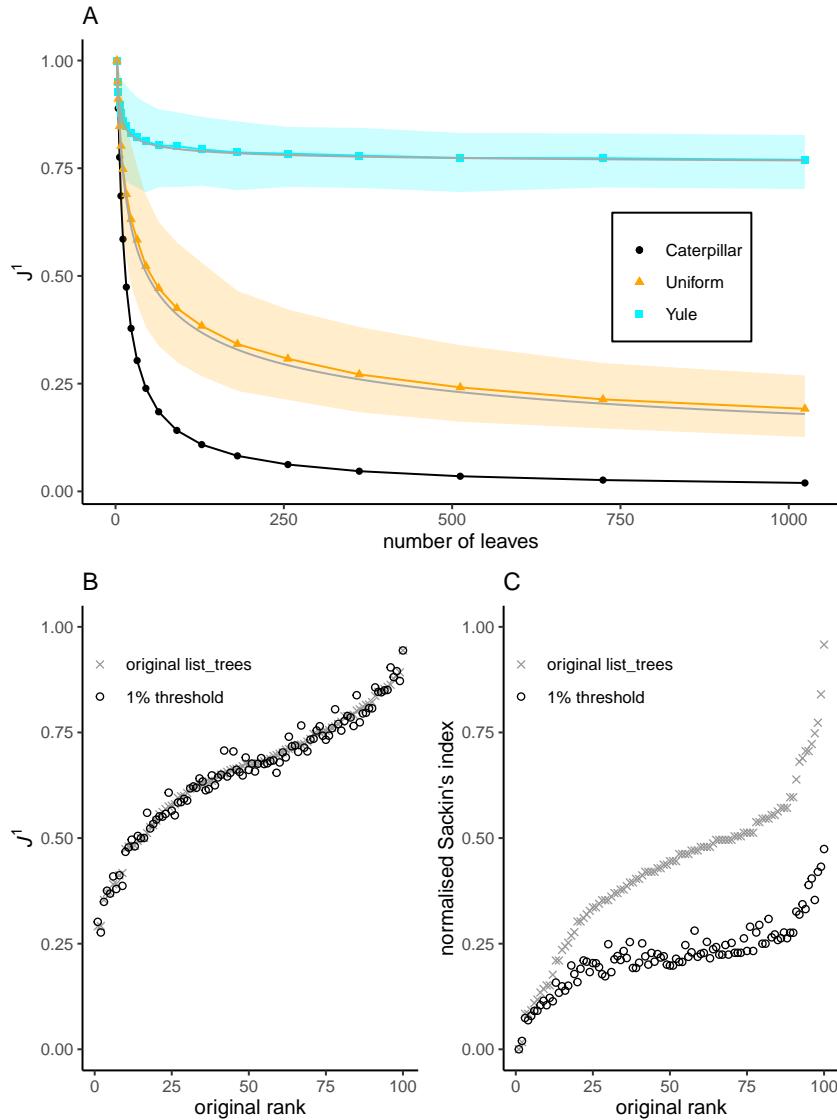


Figure 1.1: **A** J^1 values for trees generated under the Yule process and the uniform model. Solid grey curves represent the approximate expected values, and the dashed lines the 5th and 95th percentiles.

B J^1 values for 100 random trees on 16 leaves using the alpha-gamma model, with $\alpha \sim \text{Unif}(0, 1)$ and $\gamma \sim \text{Unif}(0, \alpha)$. The values were calculated before and after applying a 1% population threshold, i.e. removing all leaves with sizes smaller than 0.01 times the total population.

C Normalised Sackin index values for the same trees as in **B**.

1.3 Agent-based modelling in oncology

1.3.1 Introduction

Agent-based models (ABMs) are a class of computational models that simulate the actions and interactions of individual agents within a system. These agents can represent anything from cells in a tissue to animals in an ecosystem. ABMs are particularly useful in cancer research, as they can capture the complex interactions happening on the microscale in cancer. Spatial agent-based models (SABMs) are a subclass of ABMs that incorporate spatial information into the simulations. This is particularly useful for modelling solid tumours as it allows for the simulation of things like the spatial heterogeneity of the tumour microenvironment and the effects of spatial constraints on tumour growth. A strength of ABMs is that they can be as simple or as complex as the researcher needs them to be (Colyer et al. n.d.). However, therein lies their weakness, as oversimplification of a model can lead to rapid loss of its utility in capturing the behaviour of a complex system such as cancer. On the other hand, a model that is too complex, and attempts to include everything from epigenetic mutations to the effects of the immune system on the tumour, is likely too computationally expensive to be useful for modelling a tumour of reasonable size. This is an organic demonstration of many a researcher's favourite saying *all models are wrong, but some are useful, and some are more useful than others*. In parsing through the literature and developing a new model of 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 (cite talk and event, specify who Philip Maini is). 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 & Beerewinkel 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.

1.3.2 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 (adenocarcinoma?). 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.

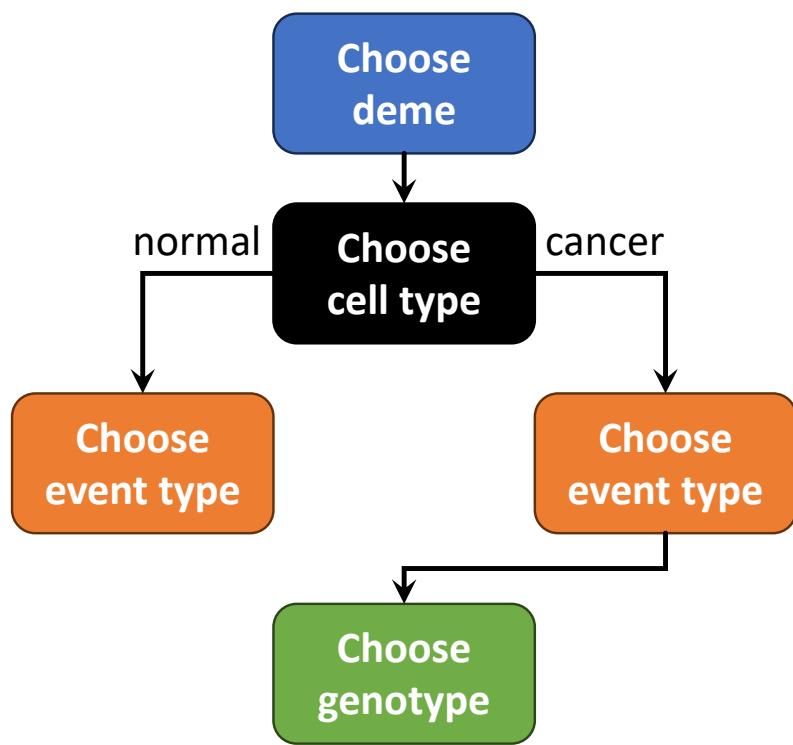


Figure 1.2: Event hierarchy in the `demon-warlock` framework. Figure reproduced from (Bak et al. n.d.) with the authors' permission.

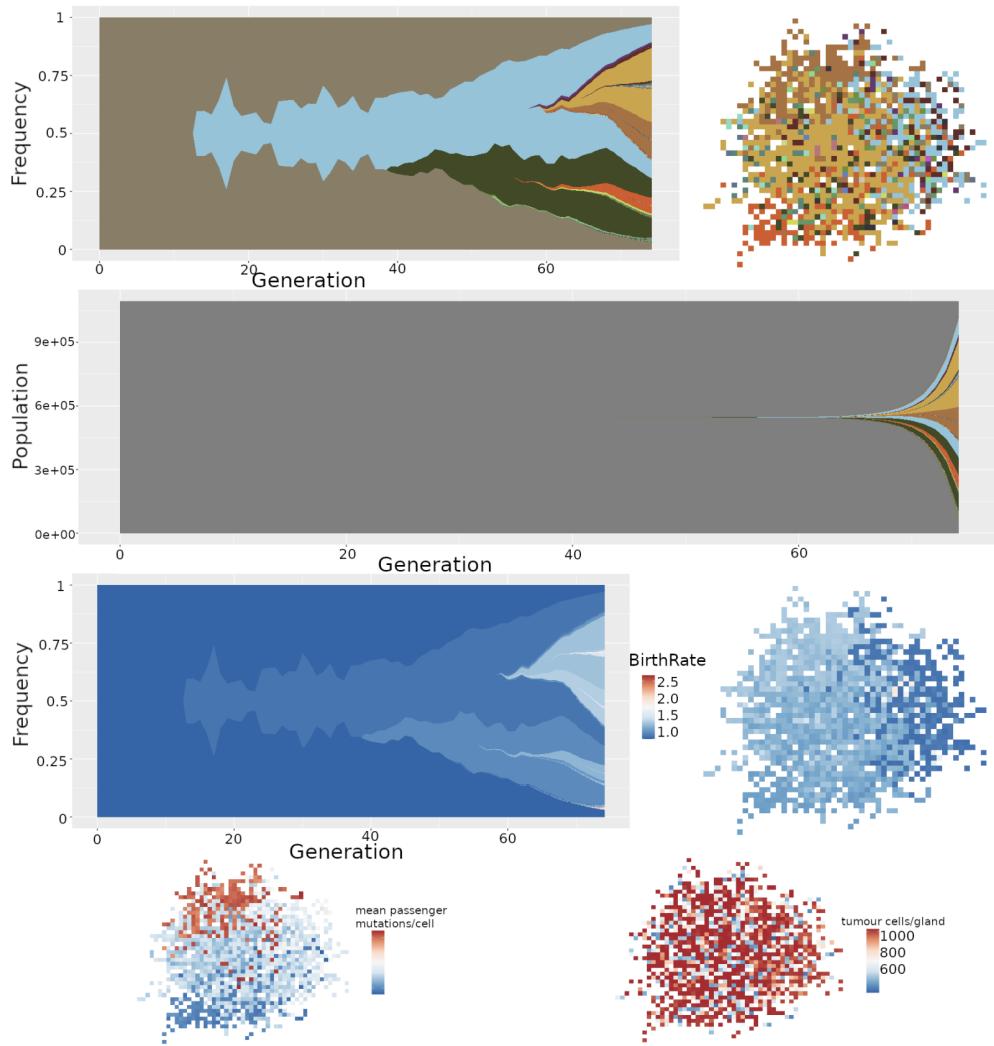


Figure 1.3: Example output from `demon`, visualised using the `demonanalysis` package. **a** Muller plot of clonal dynamics over time. Each colour represents a clone with a distinct combination of driver mutations. **b** Final proportions and spatial plot of clones. **c** Fish plot of clone populations 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.

1.4 Likelihood-free inference

1.4.1 Introduction

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

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

1.4.2 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a likelihood-free inference method that has gained popularity in the last three decades (Tavare et al. 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 ϵ , accept the parameters. Otherwise, reject them.
5. Repeat steps 1-4 until a sufficient number of accepted parameters have been obtained.

The distance threshold must be strictly positive, and is often chosen to be a small value. (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.

1.5 Fluctuating methylation clocks

The concept of a molecular clock is a (hallmark) of molecular evolutionary biology. It is based on the idea that the rate of evolution of a particular gene or set of genes is constant over time, and can be used to estimate the time of divergence between species or the time of a particular event in the evolutionary history of a species. The most famous example of a molecular clock is the mitochondrial DNA clock, which is used to estimate the time of divergence between species (?). 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 (?).

Recently, a new type of molecular clock has been proposed, the fluctuating methylation clock, based on the observation that the methylation status of certain CpG sites in the genome is heritable but fluctuates stochastically over time (Gabbutt, Schenck, Weisenberger, Kimberley, Berner, Househam, Lakatos, Robertson-Tessi, Martin, Patel, Clark, Latchford, Barnes, Leedham, Anderson, Graham & Shibata n.d., Gabbutt, Duran-Ferrer, Grant, Mallo, Nadeu, Househam, Villamor, Krali, Nordlund, Zenz, Campo, Lopez-Guillermo, Fitzgibbon, Barnes, Shibata, Martin-Subero & Graham n.d.). A CpG site is a cytosine nucleotide followed by a guanine nucleotide in the linear sequence of bases along its 5' → 3' direction, and is a common site of methylation in the genome. As citosine and guanine are complementary, each CpG site in the 5' → 3' direction has a corresponding pair in the 3' → 5' direction. This means that each CpG pair can be in one of three states: both methylated (homozygously methylated), both unmethylated (homozygously unmethylated), or one methylated and the other unmethylated (heterozygously methylated). Depending on tissue type, the fluctuating CpG (fCpG) sites can number somewhere between 1000 and 2000, which means each cell has a potentially unique barcode in its fCpG array. As the array is not constant, with methylations and demethylations of fCpG sites happening over the course of cell divisions, the authors of the two papers covering fCpGs so far have been able to reconstruct the evolutionary history of healthy colonic crypts and lymphoid malignancies with high accuracy. This is a promising development, as sequencing methylation arrays is a cheaper method than genome sequencing, but may offer finer temporal resolution. In chapter 5, we investigate whether multi-site methylation array sequencing can be used to reconstruct clonal dynamics of colorectal cancer.

1.6 Aims and objectives

1.6.1 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.) SABMs recapitulate molecular data observed in solid tumours (or sth like that)
3. These methods are useful for inferring the evolutionary history of colorectal cancer based on multi-site methylation array sequencing.

1.6.2 Aims

1. Calculate or approximate important properties of the J^1 index, such as its expected value under standard tree-generating processes.
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.

1.6.3 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

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

2.1 Introduction

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

Following the J^1 index paper (Lemant et al. n.d.), where a universal index was proposed, shown to be robust to the removal of small nodes and to outperform conventional tree balance indices as a summary statistic for comparing model output to empirical data, I examined several important properties of J^1 . Given any new tree shape index, the expected value under standard tree-generating processes and the extreme values need to be known for the index to be useful in practice. In figure 1.1A, I showed the sample mean of J^1 up to 128 leaves under the Yule and uniform models, which appears to be close to the inverse Sackin index expression derived by Noble in (Lemant et al. n.d.). Additionally, as a consequence of this relationship, the caterpillar trees minimises J^1 for bifurcating trees. However, I showed in (Lemant et al. n.d.) that the caterpillar topology does not minimise J^1 for multifurcating trees by providing a counterexample on 6 leaves.

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

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

2.2 Prerequisites

2.2.1 Preliminary definitions from systematic biology

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Definition 2.2.11 (Fully symmetric tree). If, for every internal node i , the subtrees rooted at the children of i all contain the same number of leaves then the tree is **fully symmetric**.

2.2.2 Preliminary definitions from computer science

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

at least three nodes is

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

where S_i is the magnitude of T_i , and S_{i_1} and S_{i_2} are the magnitudes of the subtrees rooted at the children of i . The balance score of a bifurcating leafy tree T is then defined as

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

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

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

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

2.3 Results

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

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

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

Consider then the following proposition.

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

for all internal nodes i . Then

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

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

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

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

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

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

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

If the node sizes sum to unity then this simplifies to

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

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

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

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

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

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

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

tree T can be defined as

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

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

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

the balance score of m -furcating leafy tree T .

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

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

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

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

2.3.2 J^1 is maximised by Huffman coding

Definition 2.3.1 (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 (e.g. figure 2.1).

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

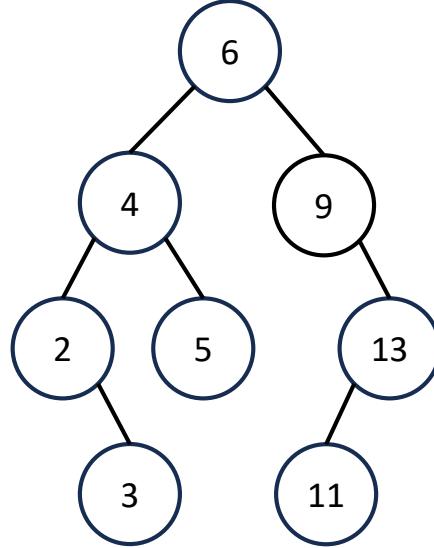


Figure 2.1: A simple example of a binary search tree over the set of labels $S = \{2, 3, 4, 5, 6, 9, 11, 13\}$.

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

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

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

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

As Huffman coding is an optimisation algorithm, J^1 can be used to measure how close a tree constructed using a given set of node sizes is to the maximally balanced binary tree on the same set. This means we can quantify how close an alternative algorithm which runs in a faster time complexity, such as arithmetic coding (Pasco n.d.), gets to the optimal solution.

2.3.3 Expected value of J^1 under simple evolutionary processes

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

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

Definition 2.3.3 (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 (figure 2.2A). The sum of probabilities of generating T in all possible ways is the probability of generating T under the Yule model.

Definition 2.3.4 (Uniform model, Rosen (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 $n \binom{2n-2}{n-1}^{-1}$ (figure 2.2B).

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

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

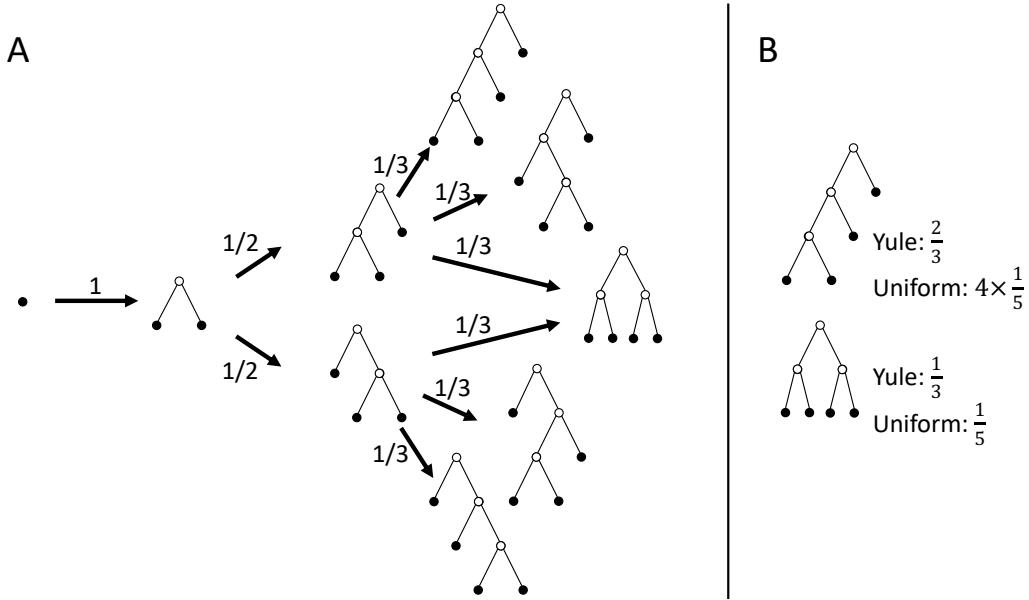


Figure 2.2: Comparison of probabilities for generation of trees on 4 leaves under the Yule and uniform models. **A:** Arrows show generation under the Yule model. Each of the trees shown on 4 leaves has the same probability under the uniform model. **B:** Comparison of probabilities of tree topologies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

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

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

as shown in Kirkpatrick & Slatkin (n.d.). Equation (2.9) implies then that the expected value of J^1 for a tree on n leaves is

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

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

compare the harmonic and arithmetic means of I_S by considering the Jensen gap

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

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

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

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

then

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

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

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

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

as $n \rightarrow \infty$.

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

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

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

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

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 the Sackin index under the Yule process is given in equation (2.28), and its variance is calculated

as (Cardona et al. n.d.)

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

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

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

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 approximately 0.008, while the lower bound reaches a maximum at $n = 8$ and is approximately 0.005.

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

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

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

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

analogously to (2.24). The expectation of Sackin's index under the uniform model is given by (Cardona et al. n.d.)

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

and its variance is

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

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

$$n! \xrightarrow{n \rightarrow \infty} \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \quad (2.30)$$

$$n! \xrightarrow{n \rightarrow \infty} \begin{cases} \sqrt{\pi n} \left(\frac{n}{e}\right)^{n/2}, & n \text{ even,} \\ \sqrt{2n} \left(\frac{n}{e}\right)^{n/2}, & n \text{ odd,} \end{cases} \quad (2.31)$$

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

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

and the variance

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

□

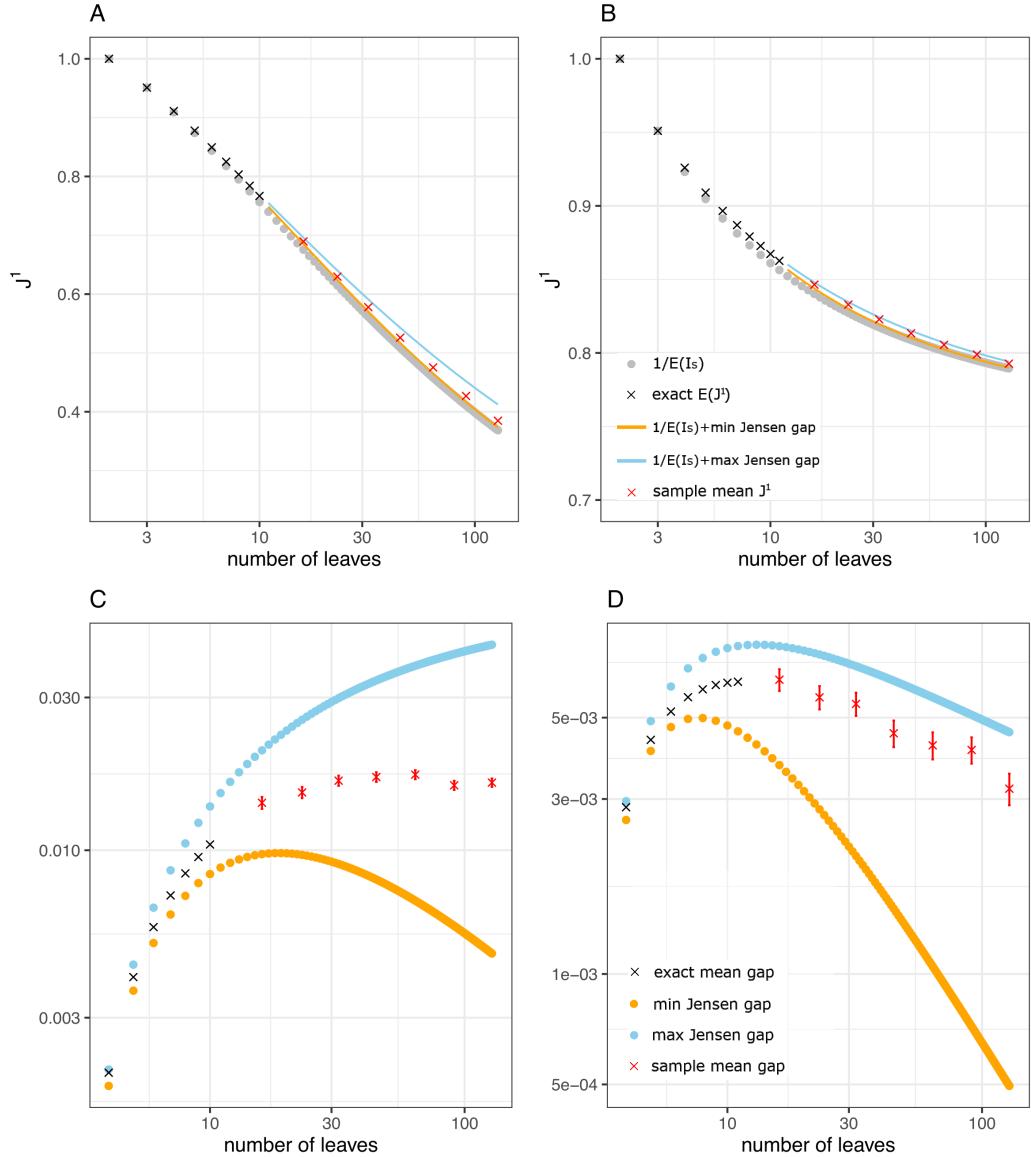


Figure 2.3: Top row: True values of $E(J^1)$ for up to 10 leaves were calculated manually, and the approximations up to 128 leaves were calculated as $n \log_2 n / E(I_S)$. **A** — uniform model, **B** — Yule model.

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 table 2.1 for numerical values of the gap size for the first several values of n . The red crosses in **A** and **B** represent sample mean J^1 values for 100000 trees generated under the uniform model and Yule process, and the difference between the approximate gap size and the sample mean, with standard error represented by error bars, in **C** and **D**.

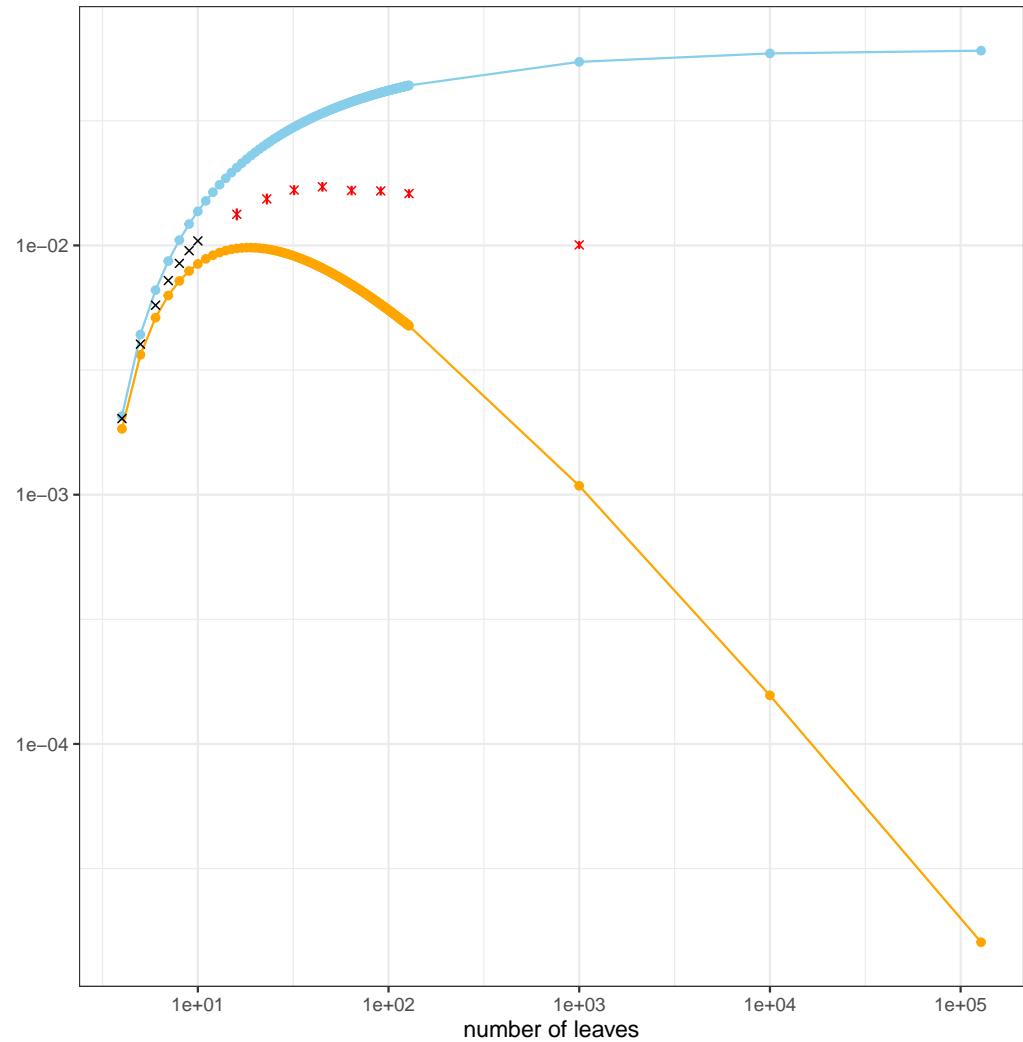


Figure 2.4: The convergence of the upper bound to $4/3\pi$ is much slower than the convergence of the lower bound to 0, and the maximum it reaches over the plotted range is 0.0604 for $n = 128000$. The red crosses, as in figure 2.3, suggest convergence of the gap size.

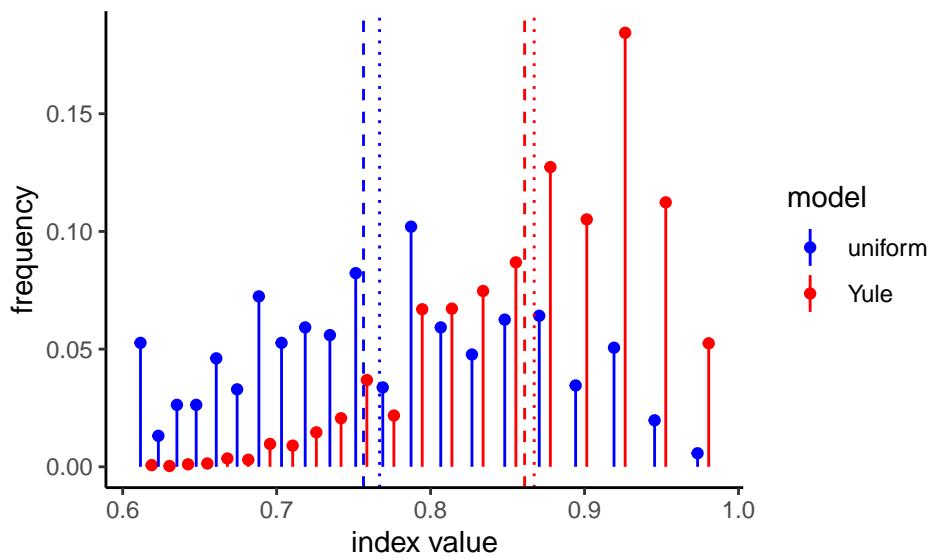


Figure 2.5: Higher variance in the uniform model leads to a non-zero upper bound on the Jensen gap. Shown are frequencies of J^1 values on 10-leaf trees generated under the Yule and uniform models. The dashed lines represent the true expected value of J^1 , and the dotted lines the approximate value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

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

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

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

n	$n \log_2 n / \mathbb{E}_Y(J^1)$	$\mathbb{E}_Y(I_S)$	$n \log_2 n / \mathbb{E}_U(J^1)$	$\mathbb{E}_U(I_S)$
2	2	2	2	2
3	5	5	5	5
4	216/25	26/3	360/41	44/5
5	728/57	77/6	3822/289	99/7
6	1162800/67217	87/5	18.25643	386/21
7	199806750/9017743	223/10	23.81979	793/33
8	27.29901	962/35	29.87282	12952/492
9	32.68993	4609/140	36.38201	26333/715
10	38.30246	4861/126	43.31989	106762/2431
11	44.11464	55991/1260	n/a	n/a

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

2.3.4 Analytic properties of the J^1 index

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

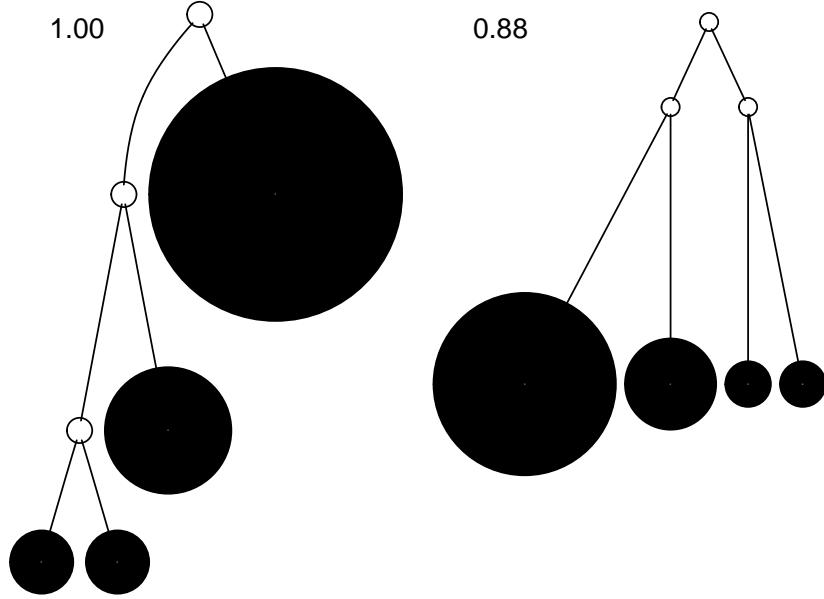


Figure 2.6: By including the node-balance function W^1 in J^1 , we allow for the possibility of perfectly balanced caterpillars (left) and less balanced fully symmetric trees (right) based on the node size distribution in the tree. The leaf sizes in these two trees are identical, with a ratio 4 : 2 :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

2.3.5 Properties of J^1 on different tree families

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

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

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. Also important to consider are node sizes and, more specifically, how the population is split across different subtrees in the tree of interest. Let us consider a slightly altered caterpillar topology.

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

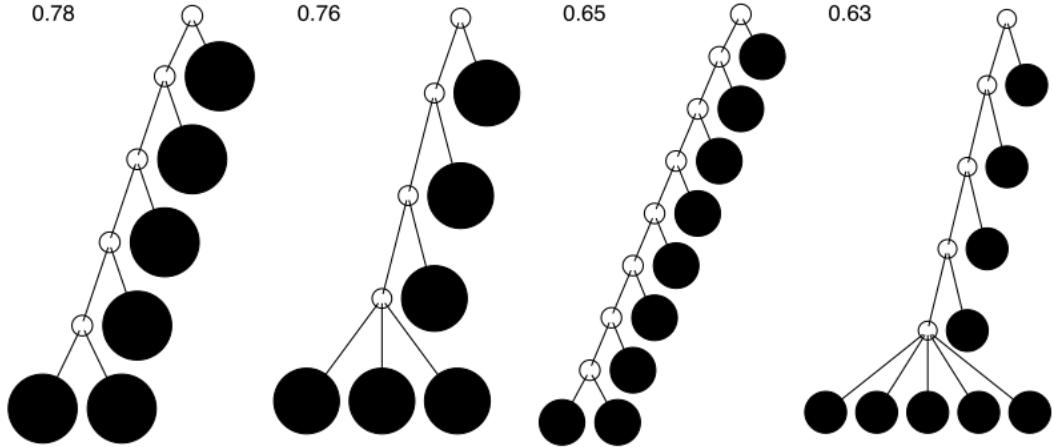


Figure 2.7: If we limit our search to leafy trees with equal leaf sizes, the least balanced tree on a given number of leaves is not necessarily the caterpillar. Pictured are the caterpillar trees on 6 and 9 leaves, as well as minimally balanced brooms for 6 and 9 leaves, with corresponding J^1 values.

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

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

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

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

Proof.

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

□

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

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

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

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

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

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

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

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

$$J_B^1(n, 2) = \frac{2n \log_2 n}{(n+2)(n-1)}, \quad (2.36)$$

$$J_B^1(n, 3) = \frac{2}{(n+3)(n-2)}(n \log_2 n - 3 \log_2 3 + 3). \quad (2.37)$$

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

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

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

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

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

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

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

2.3.6 Behaviour as $n \rightarrow \infty$

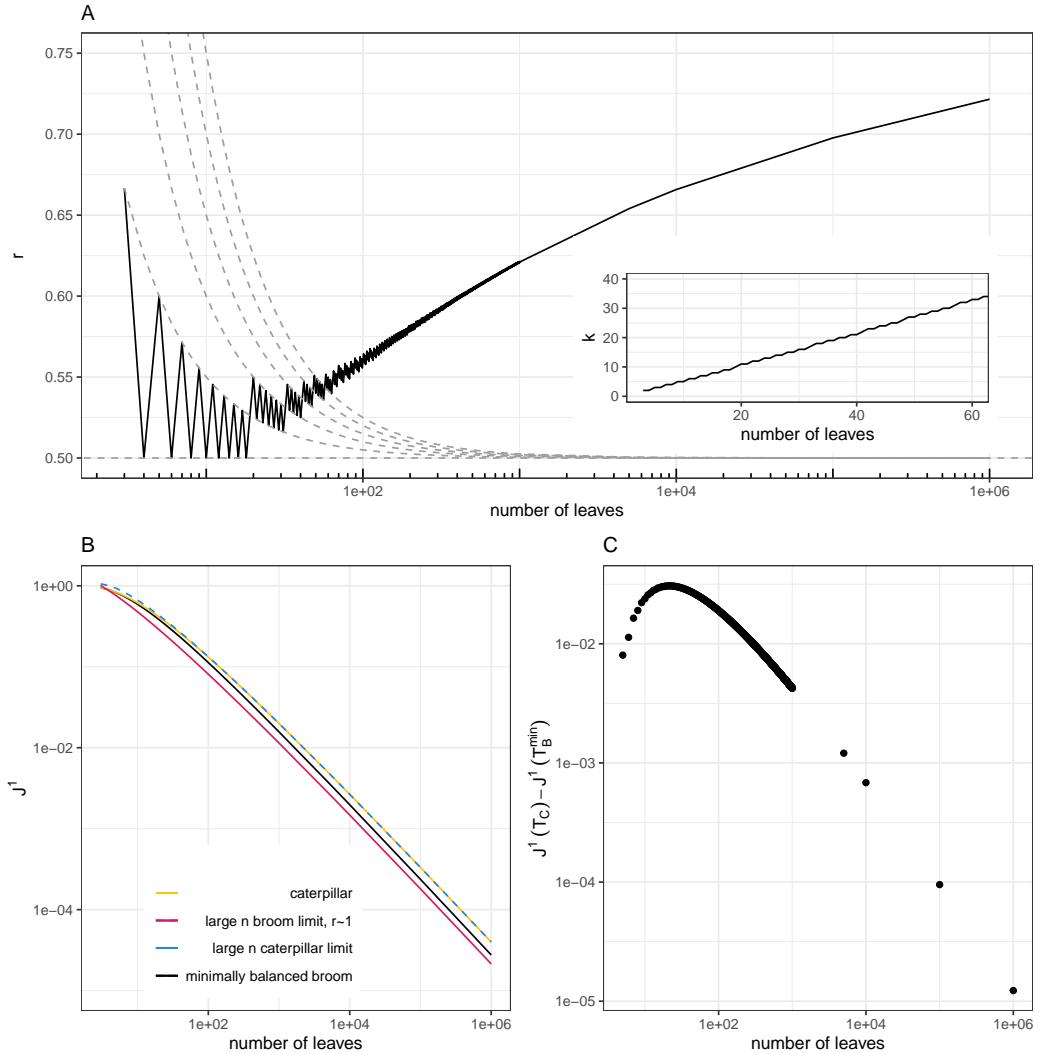


Figure 2.8: The labels used in the figures are as above - n for number of leaves, k for number of leaves in the broom head, $r = n/k$. **A:** Value of r for which the minimum value of J^1 is obtained on leafy trees. Trees on n leaves which satisfy $r = \frac{n+a}{2n}$, for $a = 0, 1, 2, \dots$ lie on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

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

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

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

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

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

as $n \rightarrow \infty$.

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

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

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

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

We can evaluate the difference between these expressions:

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

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

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

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

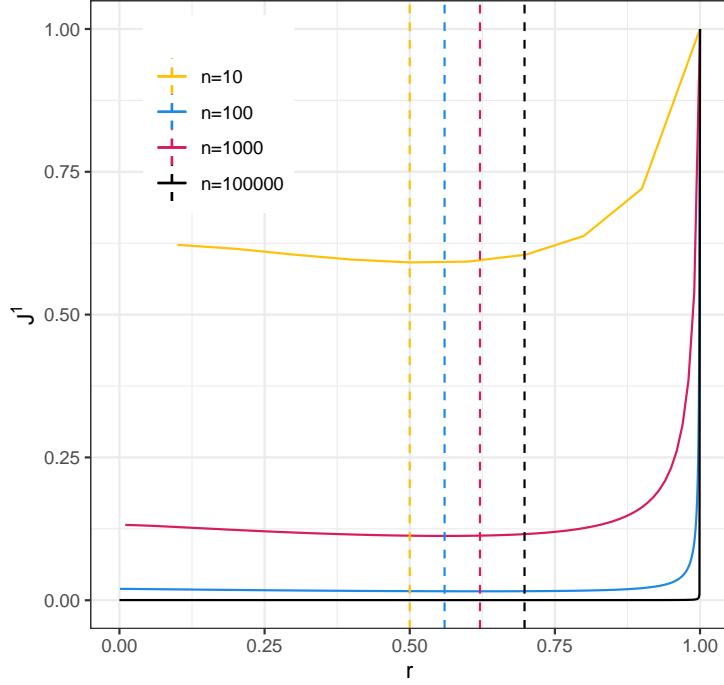


Figure 2.9: Values of J^1 on trees of different sizes calculated using equation (2.34) for different values of $r = k/n$. The dashed lines are at values of r which minimise J^1 .

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

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

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

$$J^1 \xrightarrow{n \rightarrow \infty} \frac{2}{n(1-r)(r+1)}((r+1)\log_2 n(r+1) - r\log_2 nr) \quad (2.41)$$

which is minimised for $r \rightarrow 1$. \square

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

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

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

2.4 Discussion

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

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

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

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

Chapter 3

Tracking 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)

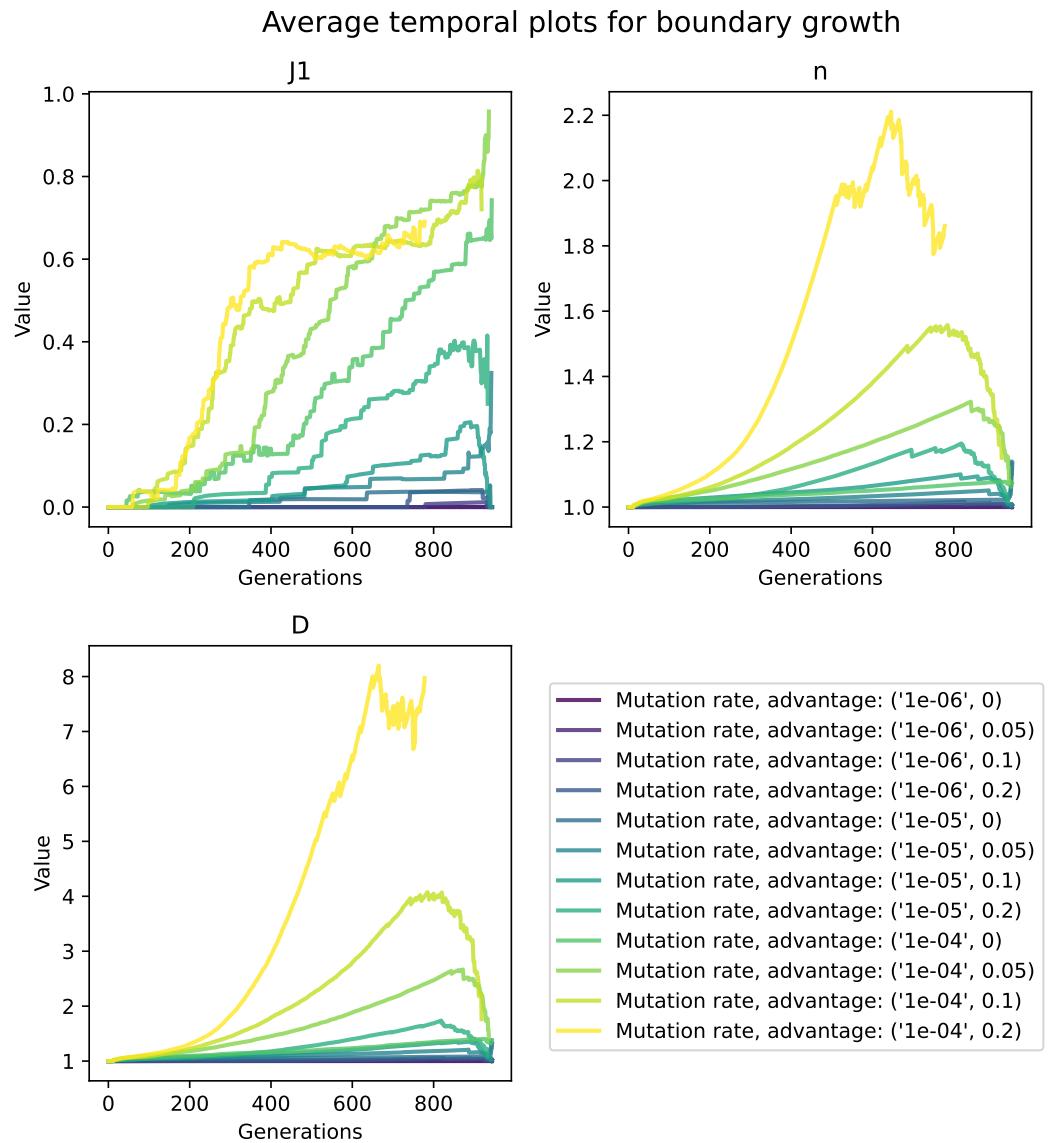


Figure 3.1: Average trajectories of the three indices for different values of driver mutation rate and selective advantage for tumours progressing via boundary growth.

Average temporal plots for non-spatial tumours

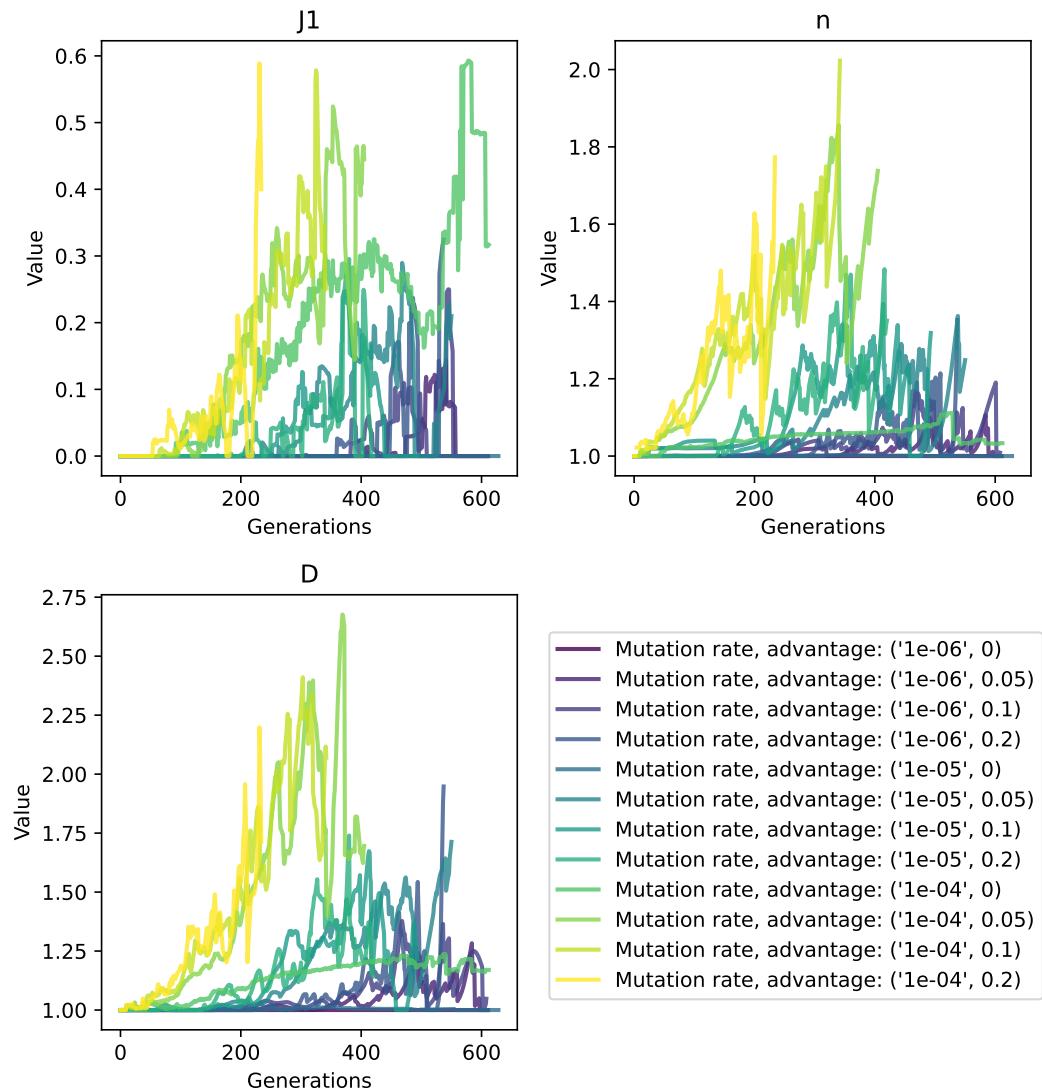


Figure 3.2: Average trajectories of the three indices for different values of driver mutation rate and selective advantage for well-mixed cancer cell populations.

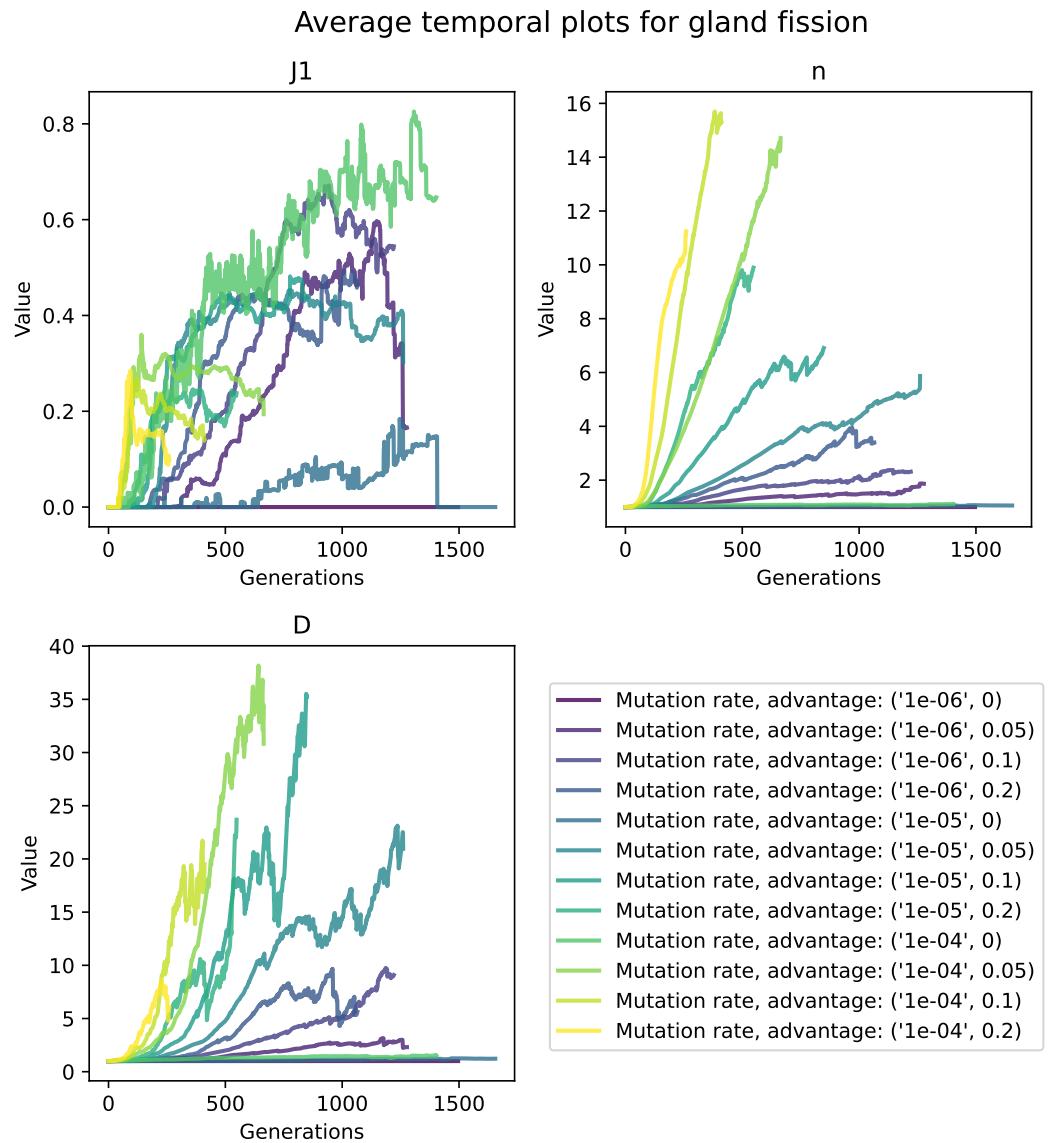


Figure 3.3: Average trajectories of the three indices for different values of driver mutation rate and selective advantage for gland fission.

Average temporal plots for invasive glandular tumours

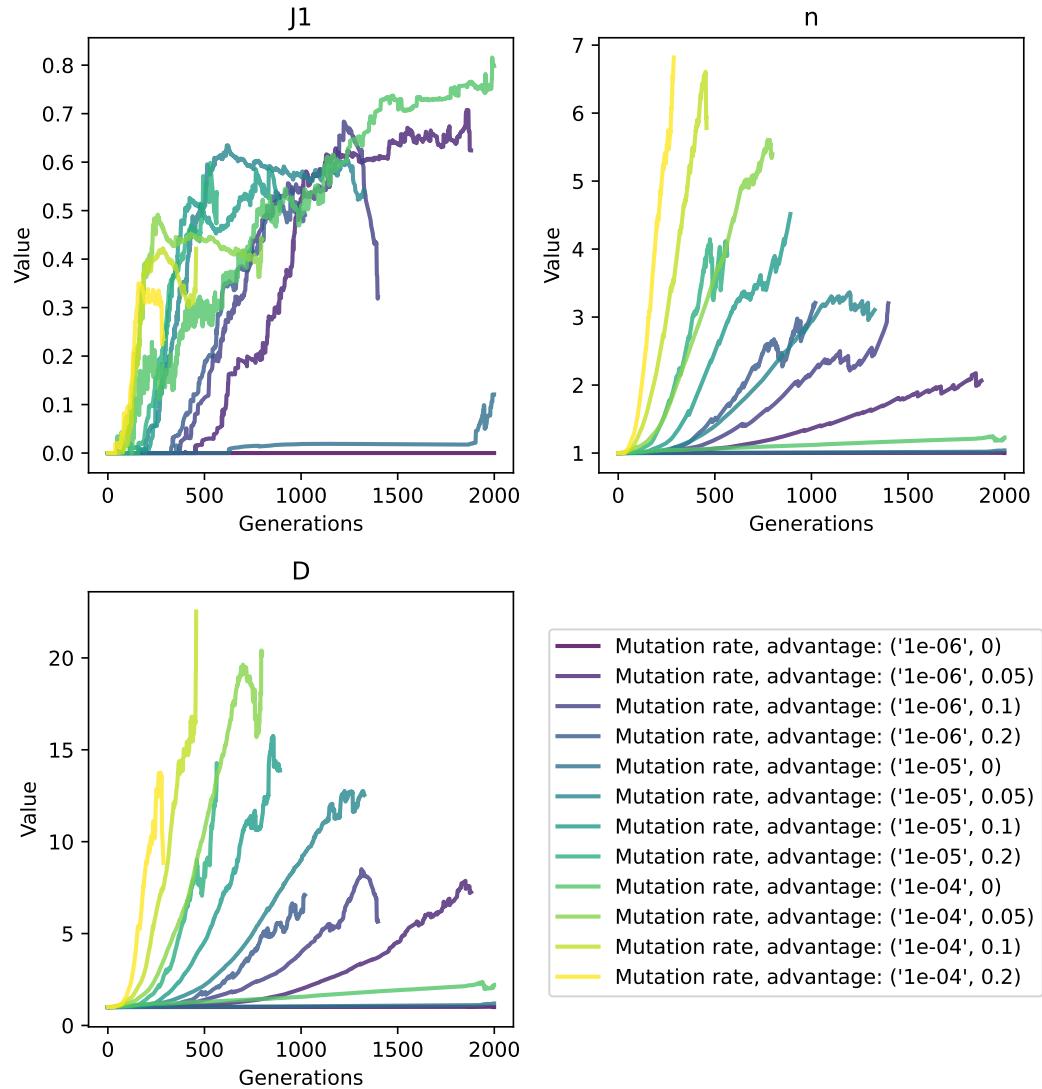


Figure 3.4: Average trajectories of the three indices for different values of driver mutation rate and selective advantage for invasive glandular tumours.

Average trajectories in index space for boundary growth

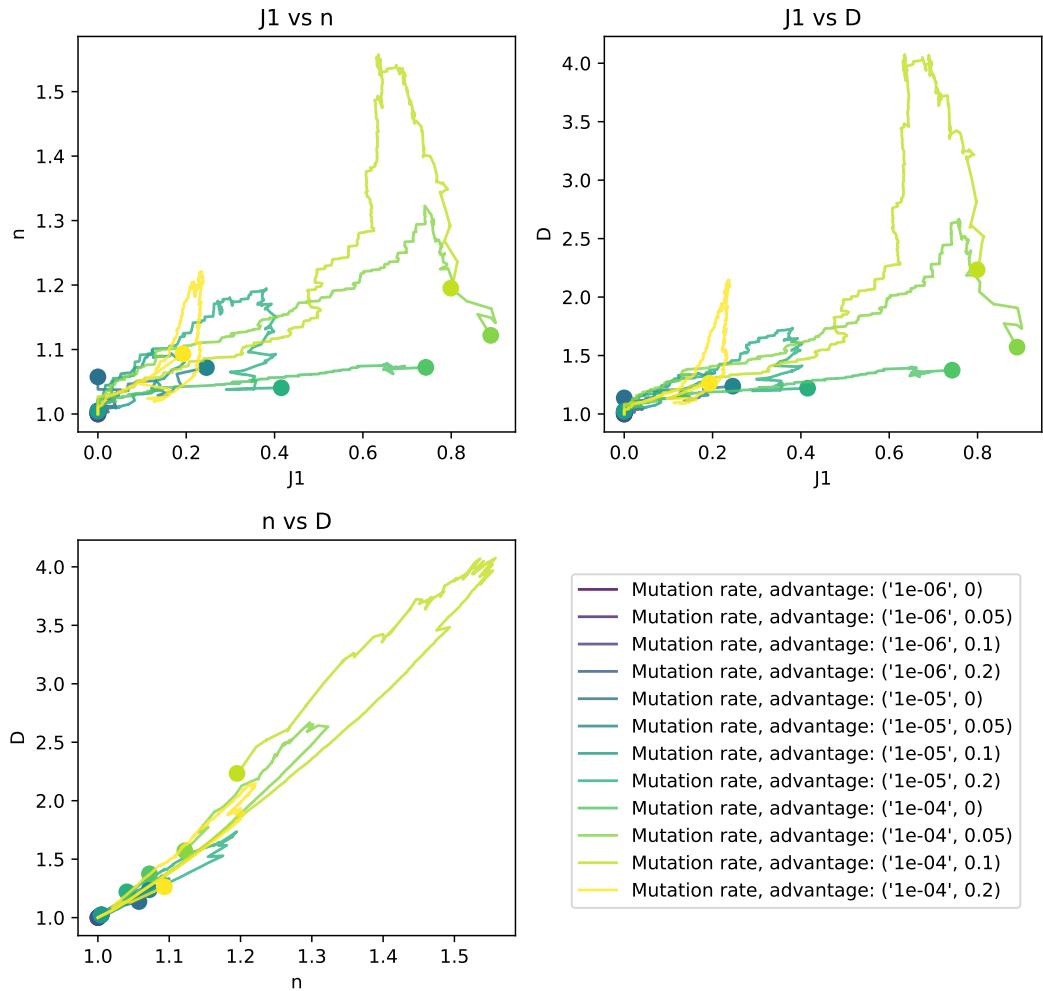


Figure 3.5: Average trajectories in index space for tumours progressing via boundary growth.

Average trajectories in index space for non-spatial tumours

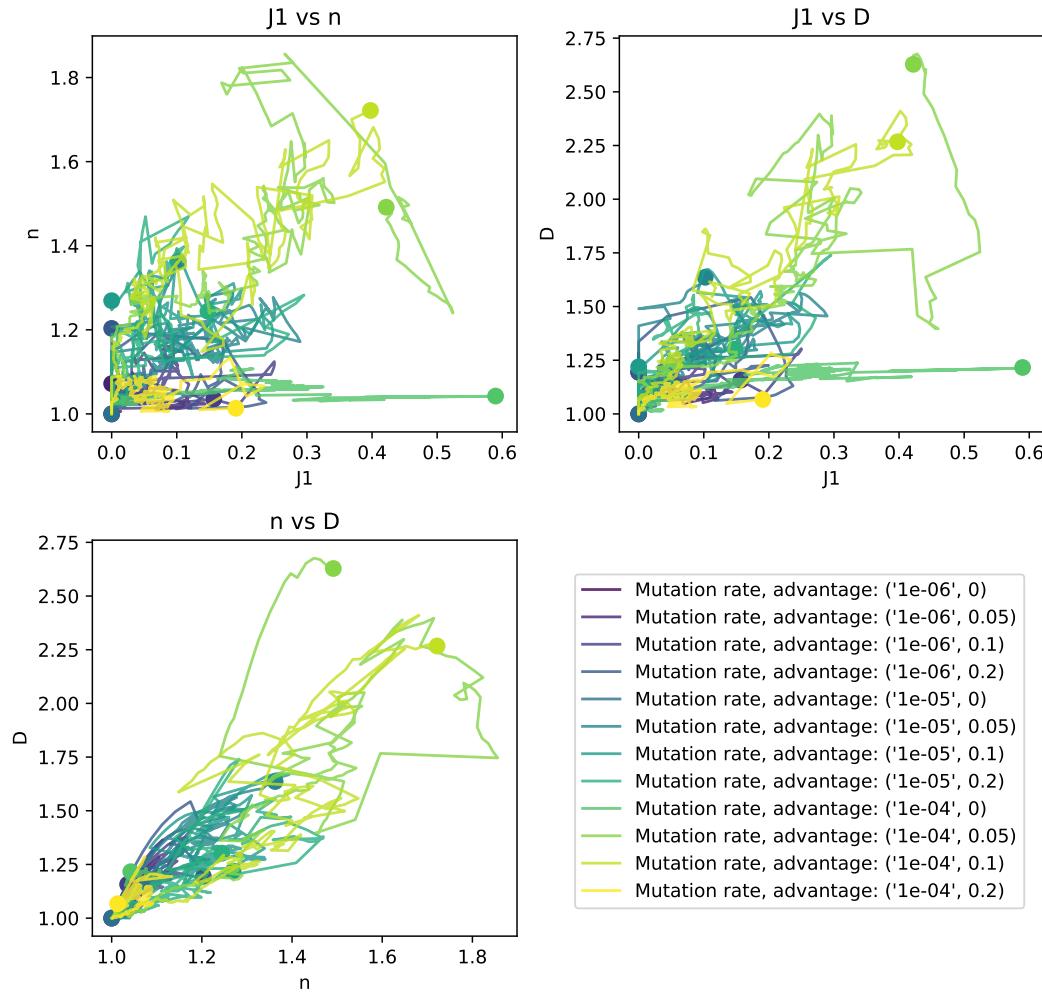


Figure 3.6: Average trajectories in index space for well-mixed cancer cell populations.

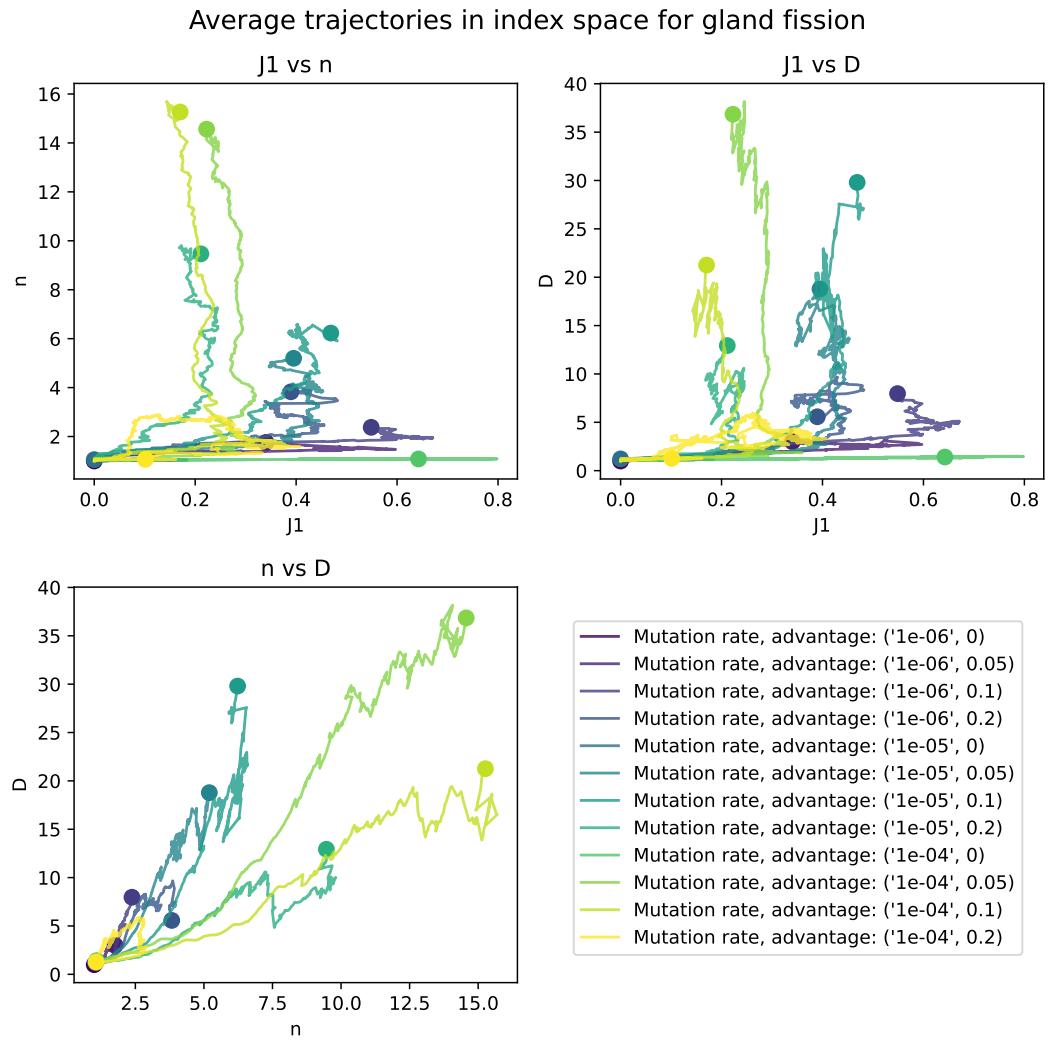


Figure 3.7: Average trajectories in index space for tumours progressing via gland fission.

Average trajectories in index space for invasive glandular tumours

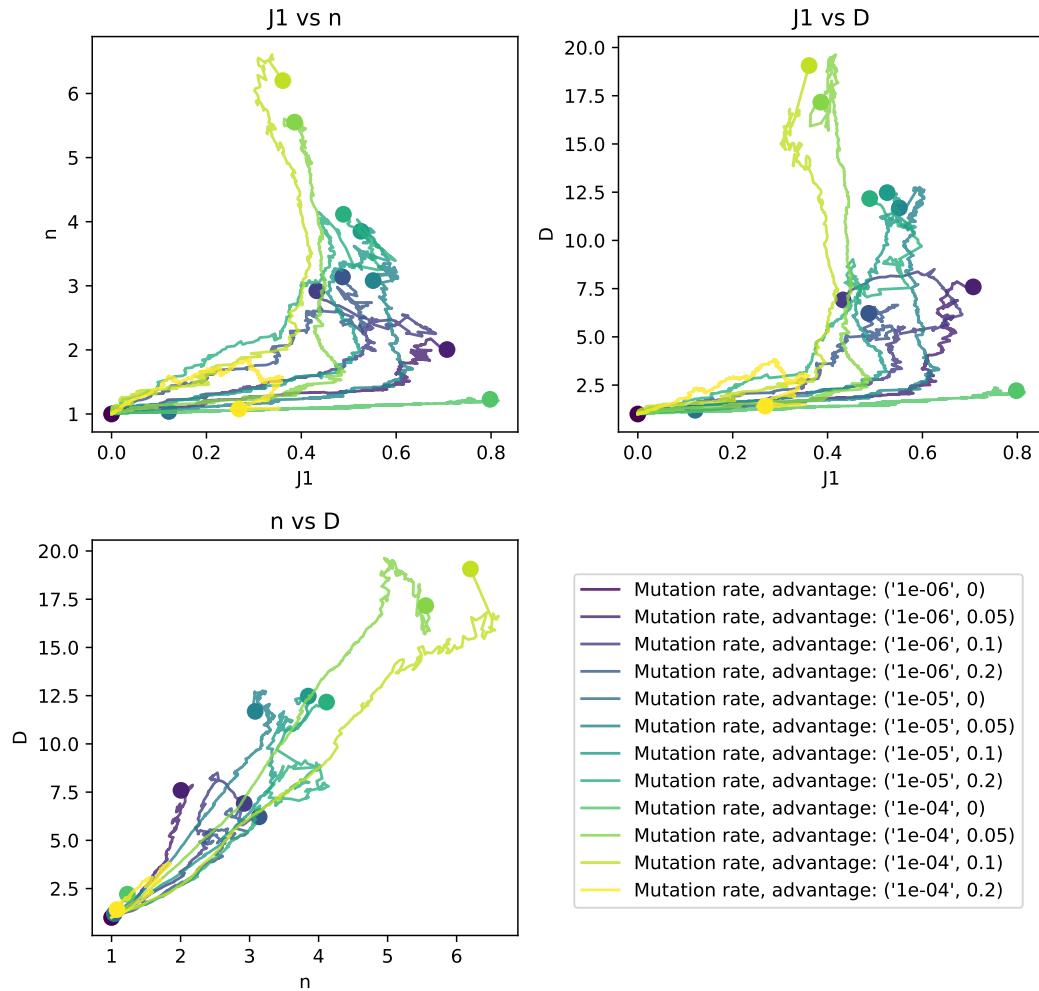


Figure 3.8: Average trajectories in index space for invasive glandular tumours.

- 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, I 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 simulating a large number of cells. This means that using the model’s outputs in comparison to or to draw inference from real data is not feasible.

There are a few ways to address this issue. For example, rather than simulating all clones in a tumour, one could take the approach of (Sottoriva et al. n.d.) and use demes (tumour glands) as the principal agent of our simulation. This would allow for a realistically-sized tumour to be generated as the number of glands would be around the right order of magnitude. A problem with this approach is that it loses resolution since a gland’s population is assumed to be clonal, undergoing rapid fixation in the case of an emerging mutant. If 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, as performed earlier.

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 is not spatially explicit, which is a feature we need to be able to discriminate between different glands in a solid tumour. In this chapter, I present a purpose-written agent-based model, `methdemon`, which reduces the computational cost of simulating a tumour's growth and models the fluctuating methylation arrays in colorectal cancer.

4.2 Background on colorectal cancer — model assumptions

4.2.1 Colorectal cancer evolution

The most common type of colorectal cancer is adenocarcinoma, which arises from the epithelial cells lining the colon, covering more than 90% of cases. Most diagnosed colorectal cancers are moderately to well differentiated, meaning that at least 50% of the tumour's cells form glands (Fleming et al. n.d.). As it is a solid tumour, one cannot ignore the spatial organisation of the cells. Its 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 spreads 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.).

Translated into the language of an agent-based model, we can write down our initial assumptions as follows:

- (i) **A single cell forms the first gland and initiates tumour growth.** This assumption skips over the process of tumorigenesis, during which a cell accumulates mutations and becomes malignant (?). This is a simplification to be sure, but a reasonable one, given that the focus of this work is on the

evolutionary dynamics of the tumour rather than its initiation.

- (ii) **The rate of driver mutations is Poisson distributed and identical for all cells.** This assumption is consistent with most models of tumour evolution (Niida et al. n.d.).
- (iii) **The cell population within a gland grows exponentially and is well-mixed.** While not necessarily consistent with the biology of a solid tumour, this assumption allows for more efficiency in the simulation as opposed to a multi-level spatial model. Further, as the data discussed in chapter 5 is obtained from bulk samples of tumour glands, this assumption is not unreasonable.
- (iv) **Once a gland reaches a certain size, which we call the carrying capacity, the population undergoes steady-state turnover according to the Moran process.**
- (v) **At carrying capacity, a gland has a certain probability of undergoing fission, which splits the gland's population randomly into two.** As a consequence of assumption (iii), fissions do not take into account a gland's spatial organisation.
- (vi) **Gland fission occurs as a neutral spatial branching process.** The previous two assumptions and this one together form the basis of the model's spatial dynamics. While there are other mechanisms of colorectal adenocarcinoma progression, gland fission is the principal way in which the tumour grows (Preston et al. n.d.). The assumption of neutrality in the spatial branching process is consistent with the findings of (Sottoriva et al. n.d.).

4.2.2 Fluctuating methylation arrays

To my shock, much like many mathematicians before me, the availability of perfectly clean data containing detailed information about the population structure of a tumour is non-existent. This is mainly because it is impossible to obtain it with current technology. What one learns very quickly when working with biological data is to compromise. Specifically, when it comes to cancer, a compromise has to be made between resolution and scale. Where single-cell data can provide a detailed view of the mutations accumulated in the genome, it is not feasible to obtain it for a

whole tumour. On the other hand, bulk data gives a high-level view of the tumour's population structure, but a lot of the details get lost in the process.

However, DNA sequencing is not the only way to obtain information about a tumour's population structure. Early work with methylation arrays in colorectal cancer has shown potential for inferring the ancestry and age of a tumour (??). In a way the genome shows more mutations in older populations, methylation arrays will also be more diverse as time goes on. Current techniques allow for the sequencing of some 850,000 CpG sites which, while a small fraction of the genome, is still enough to provide valuable insight into the underlying dynamics of the cell population. Initial studies on methylation as a tracker of evolution made use of the whole array (??). However, more recent work has shown that just a small subset of CpG sites is enough to infer the evolutionary dynamics of a cell population (Gabbutt, Schenck, Weisenberger, Kimberley, Berner, Househam, Lakatos, Robertson-Tessi, Martin, Patel, Clark, Latchford, Barnes, Leedham, Anderson, Graham & Shibata n.d., Gabbutt, Duran-Ferrer, Grant, Mallo, Nadeu, Househam, Villamor, Krali, Nordlund, Zenz, Campo, Lopez-Guillermo, Fitzgibbon, Barnes, Shibata, Martin-Subero & Graham n.d.). This is the set of fluctuating CpG (fCpG) loci, which is also the topic of chapter 5.

On top of the assumptions outlined in section 4.2.1, we include the following set for the modelling of fluctuating methylation arrays:

- (i') **Each cell has a corresponding fCpG array inherited from its parent cell.**
- (ii') **Upon cell division, each methylated fCpG site has an equal and independent probability of being demethylated, and vice-versa.** This assumption is based on the findings of (Gabbutt, Schenck, Weisenberger, Kimberley, Berner, Househam, Lakatos, Robertson-Tessi, Martin, Patel, Clark, Latchford, Barnes, Leedham, Anderson, Graham & Shibata n.d., Gabbutt, Duran-Ferrer, Grant, Mallo, Nadeu, Househam, Villamor, Krali, Nordlund, Zenz, Campo, Lopez-Guillermo, Fitzgibbon, Barnes, Shibata, Martin-Subero & Graham n.d.).
- (iii') **The rates of methylation and demethylation do not change over time.**

4.3 Existing simulation workflows

With the assumptions outlined in the previous section, my initial approach was to employ a general agent-based model with small modifications, due to the specific data type but broad evolutionary questions. **TO DO: discuss the shortcomings of the old approach.**



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 (Gabbutt, Schenck, Weisenberger, Kimberley, Berner, Househam, Lakatos, Robertson-Tessi, Martin, Patel, Clark, Latchford, Barnes, Leedham, Anderson, Graham & Shibata n.d., Gabbutt, 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 for epigenetic changes. However, these issues may not have been insurmountable



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

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 λ is the birth rate before mutation, λ' is the birth rate after mutation, λ_{max} is the maximal allowed birth rate, and ϵ 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.

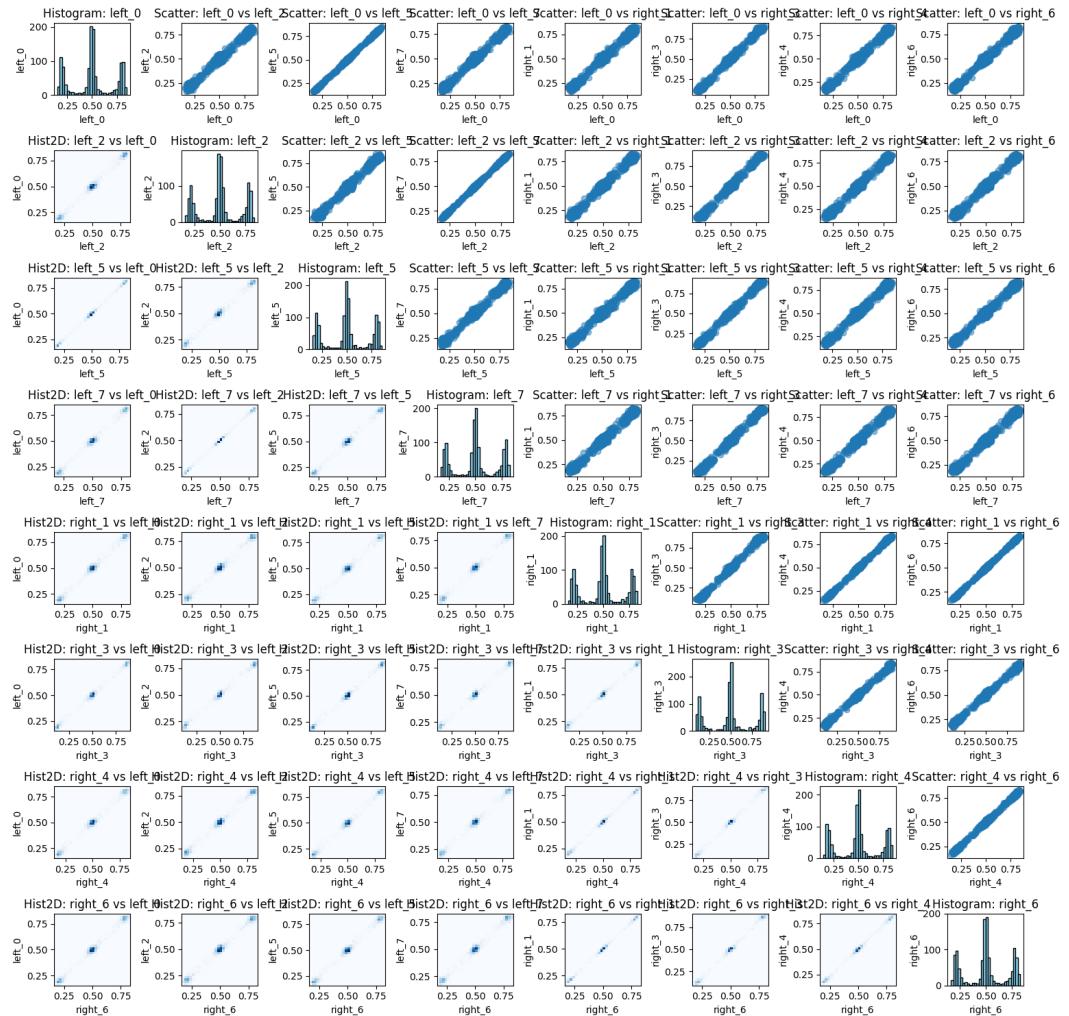


Figure 5.1: correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.1$.

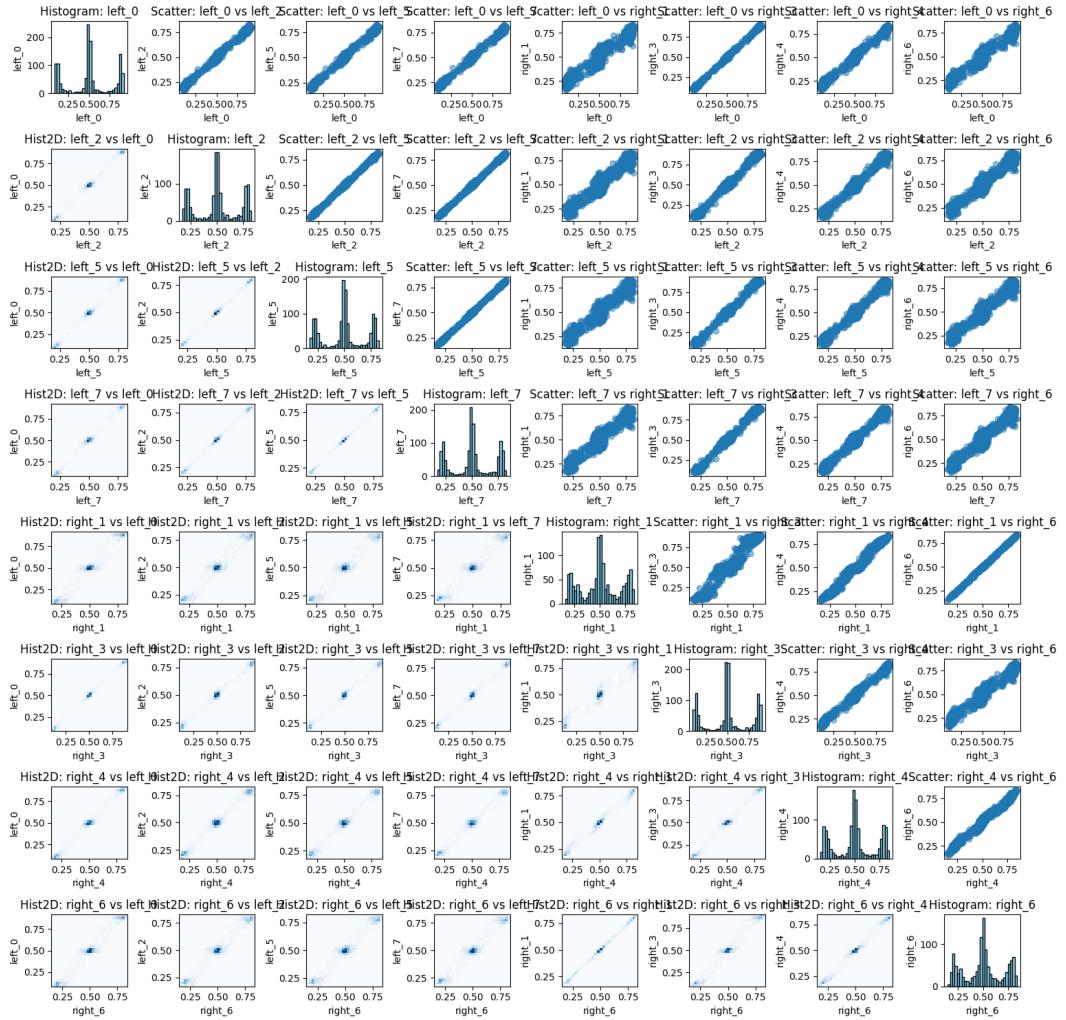


Figure 5.2: correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.2$.

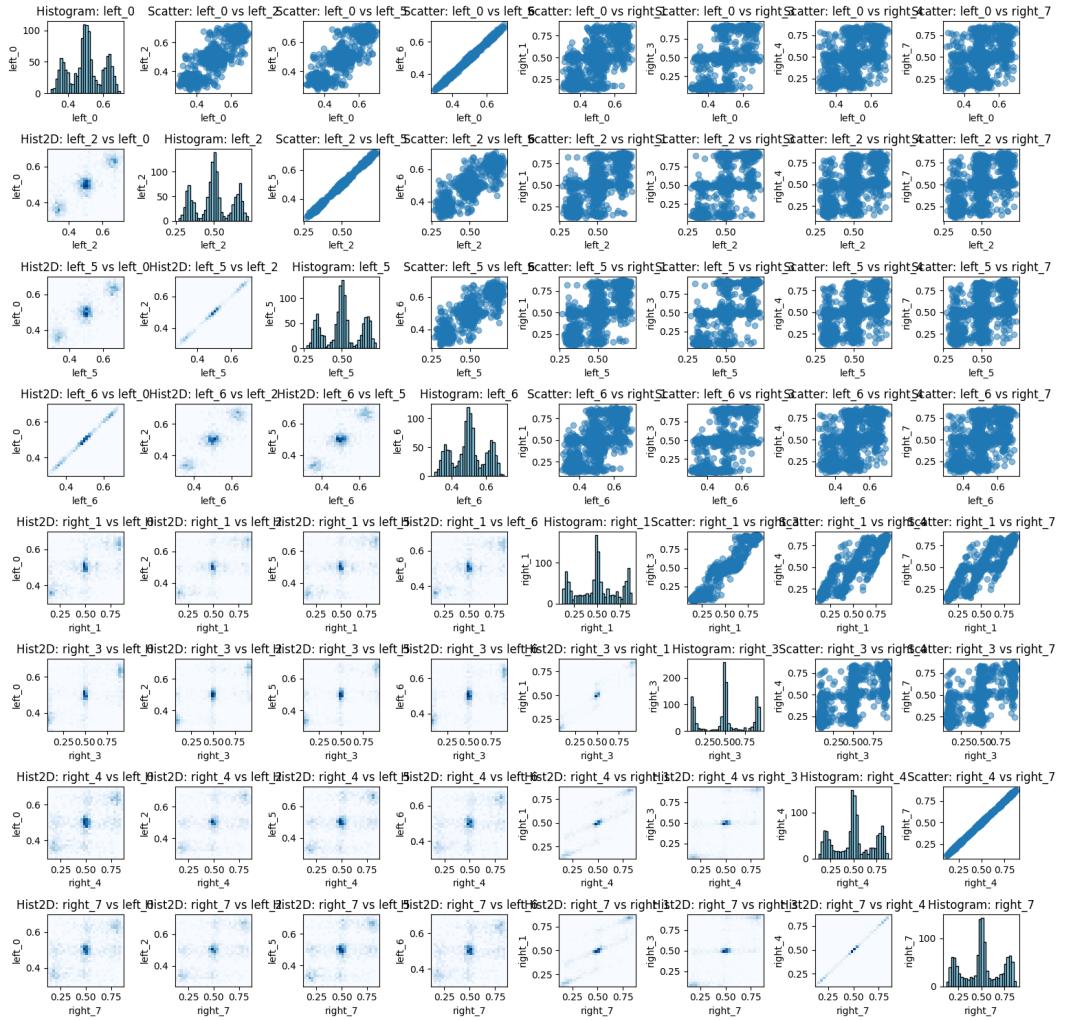


Figure 5.3: correlation scatter plots, gland histograms and correlation heatmaps for $s = 0.3$.

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.

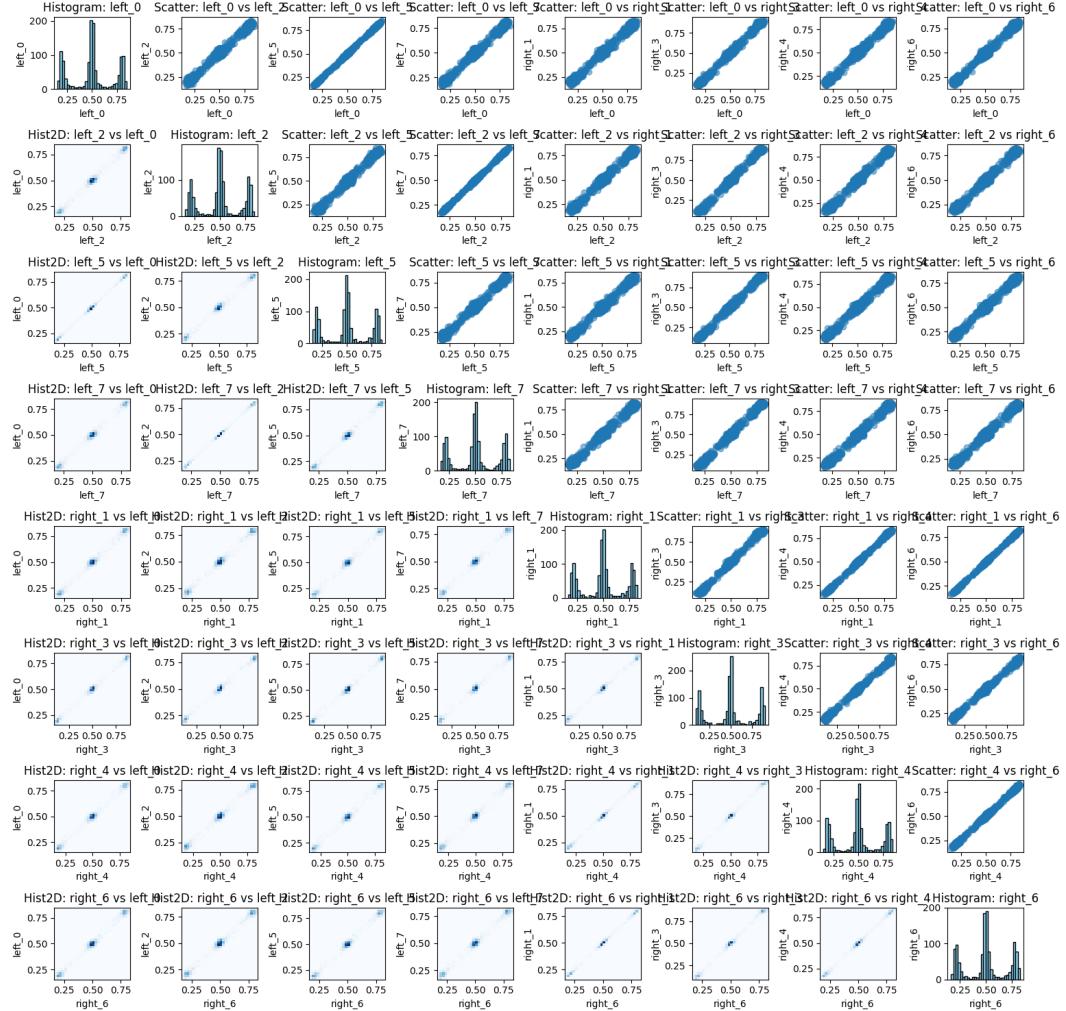


Figure 5.4: correlation scatter plots, gland histograms and correlation heatmaps for driver mutation rate 10^{-6} .

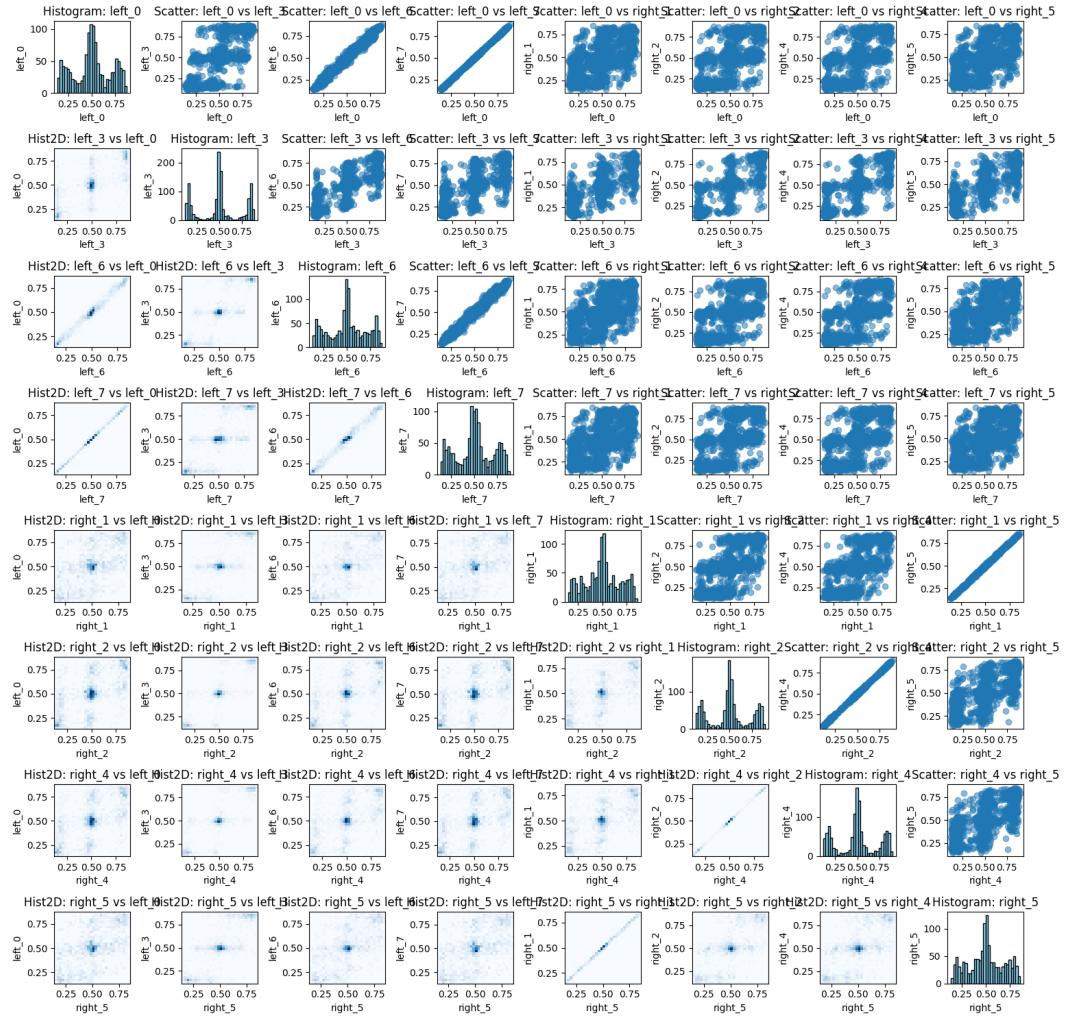


Figure 5.5: correlation scatter plots, gland histograms and correlation heatmaps for driver mutation rate 10^{-4} .

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 5×10^{-3} is covered by figure 5.3.

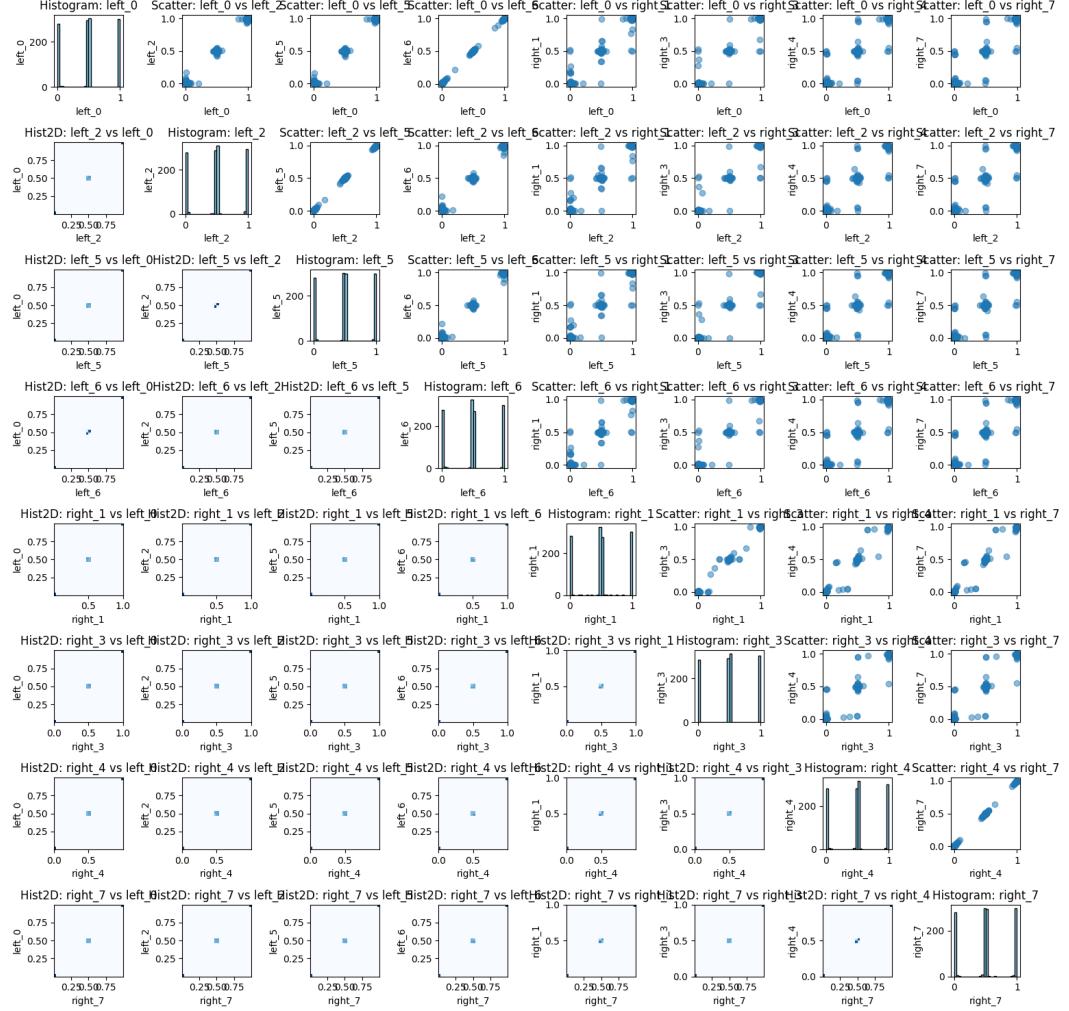


Figure 5.6: correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities 10^{-4} .

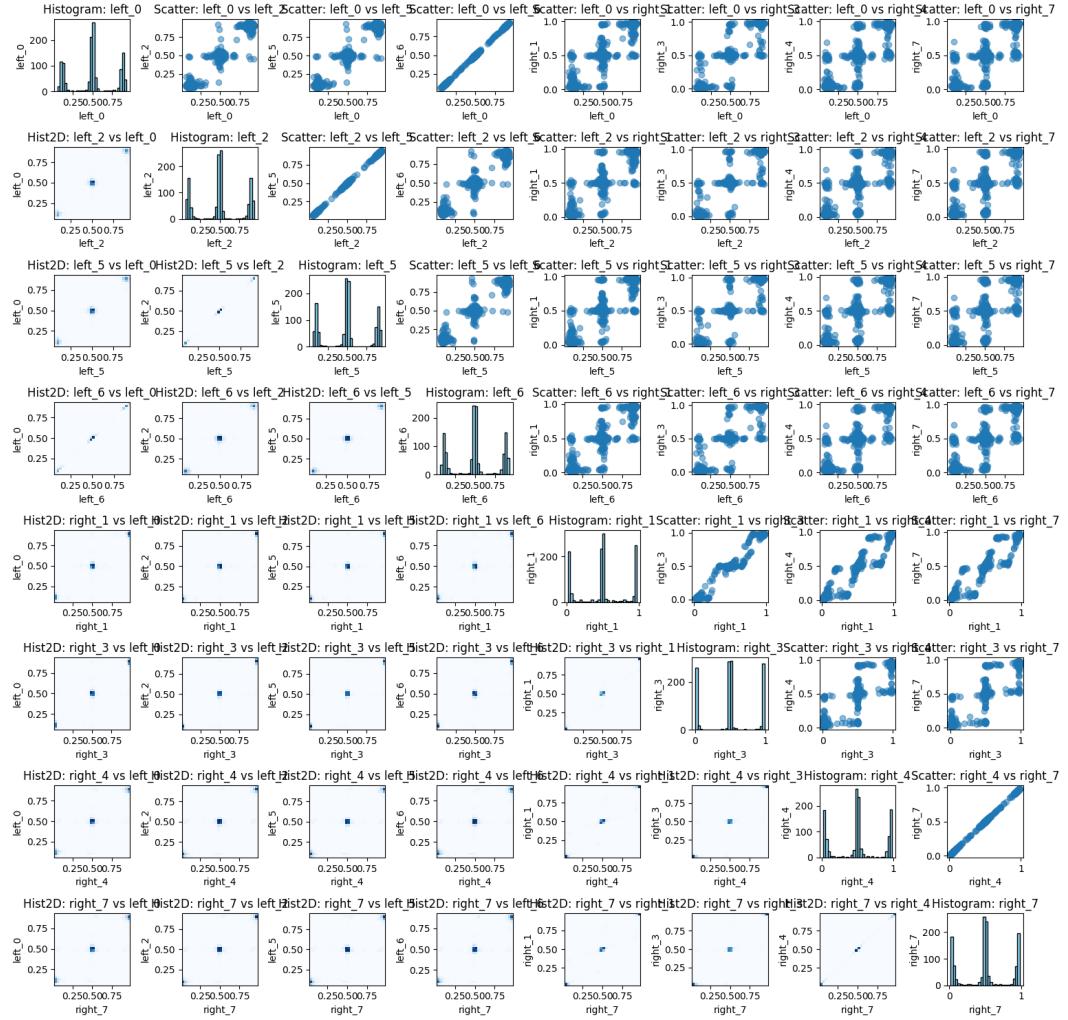


Figure 5.7: correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities 10^{-3} .

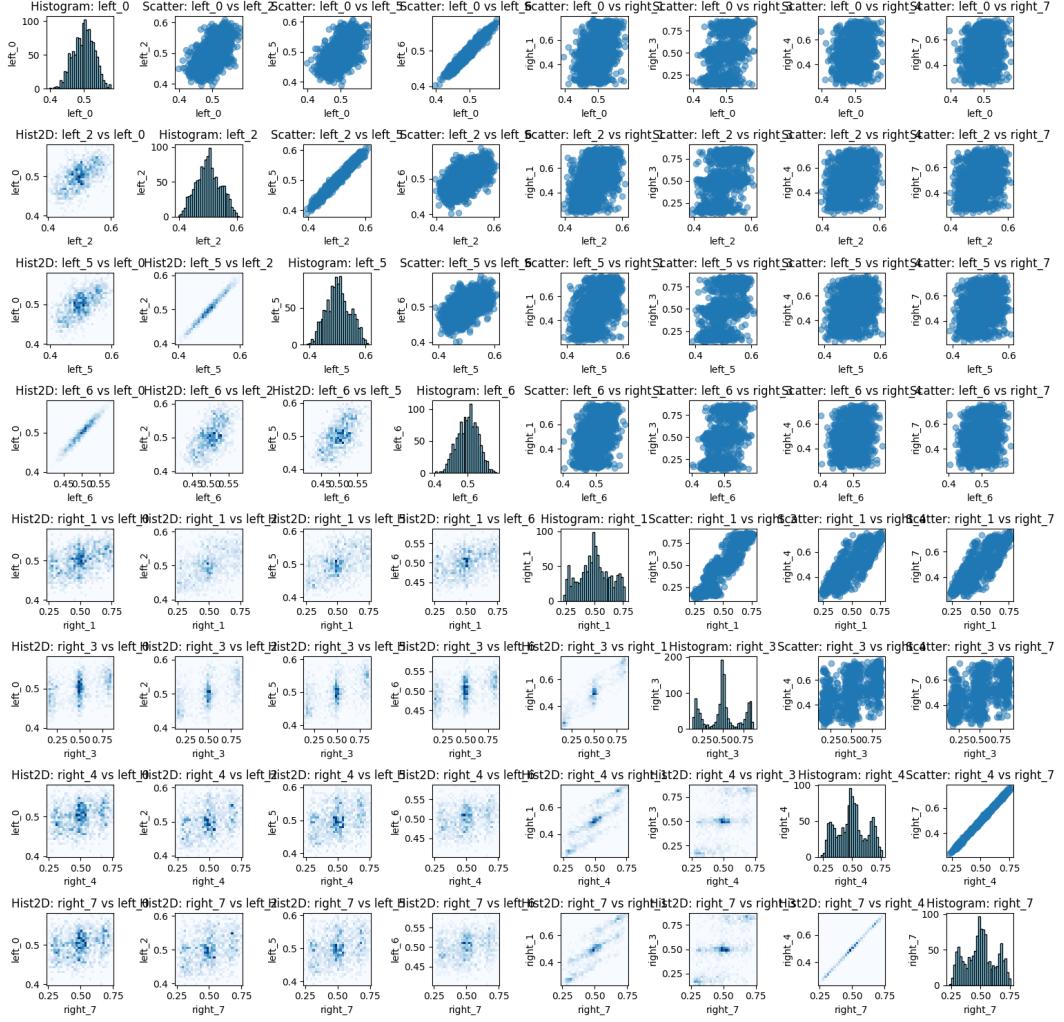


Figure 5.8: correlation scatter plots, gland histograms and correlation heatmaps for flip probabilities 10^{-2} .

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.

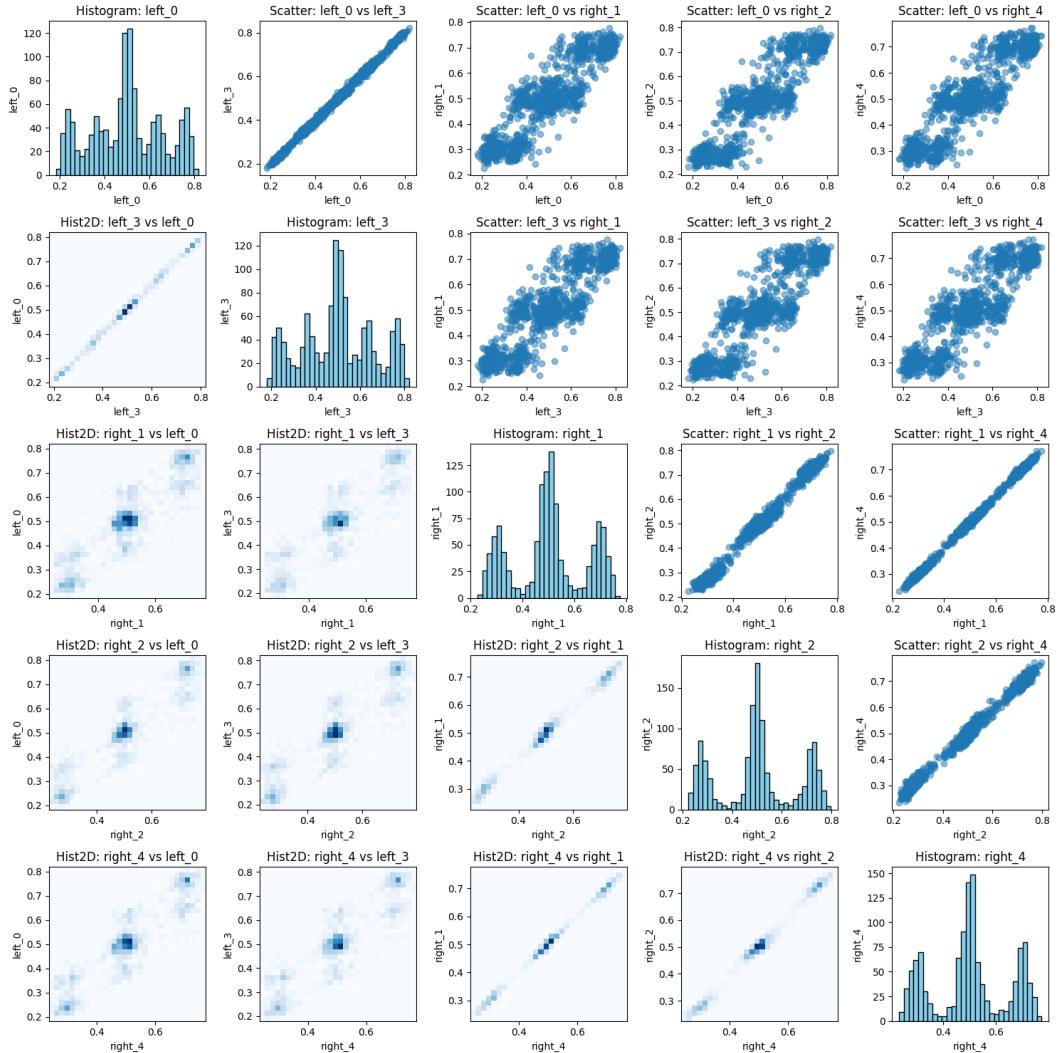


Figure 5.9: correlation scatter plots, gland histograms and correlation heatmaps for fission rates 0.008.

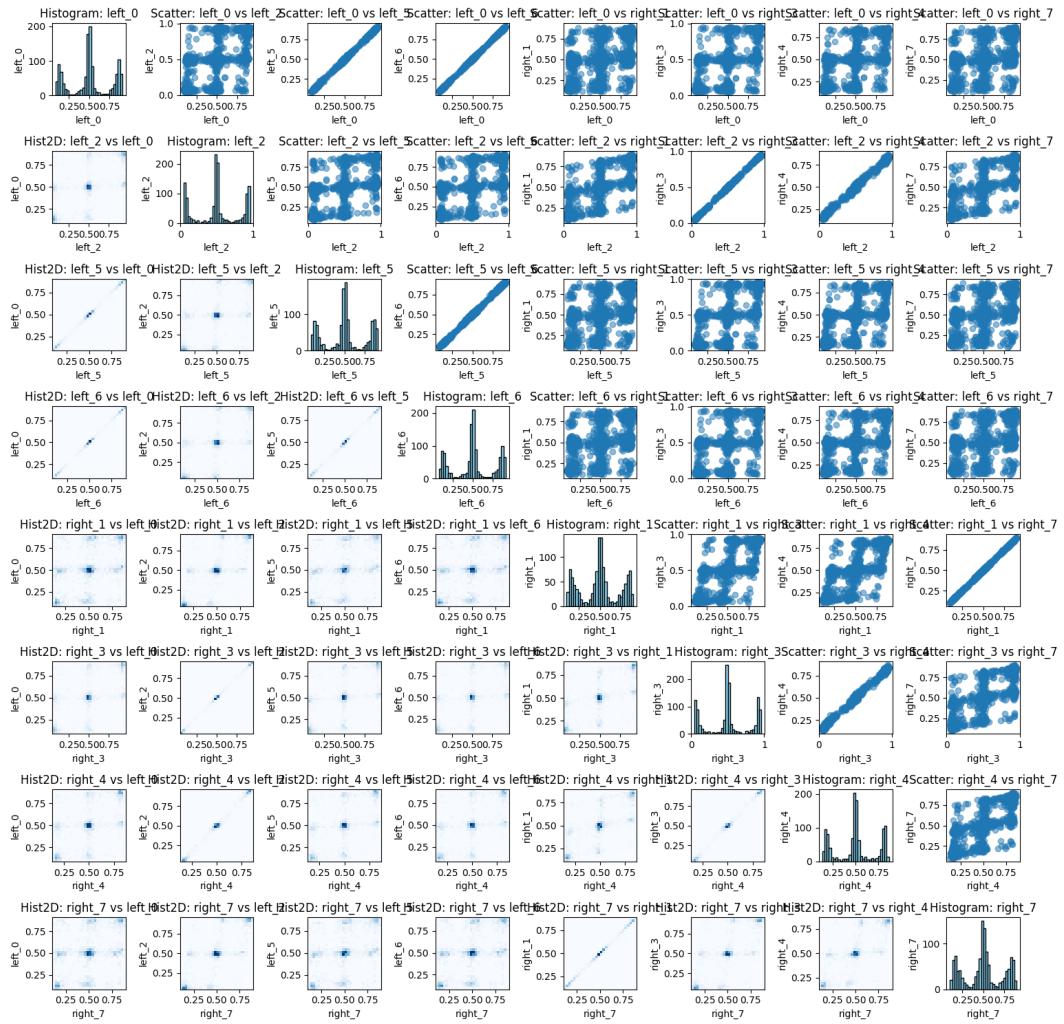


Figure 5.10: correlation scatter plots, gland histograms and correlation heatmaps for fission rates 0.08.

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	5×10^{-3}
Demethylation probability per fCpG site per cell division	5×10^{-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.

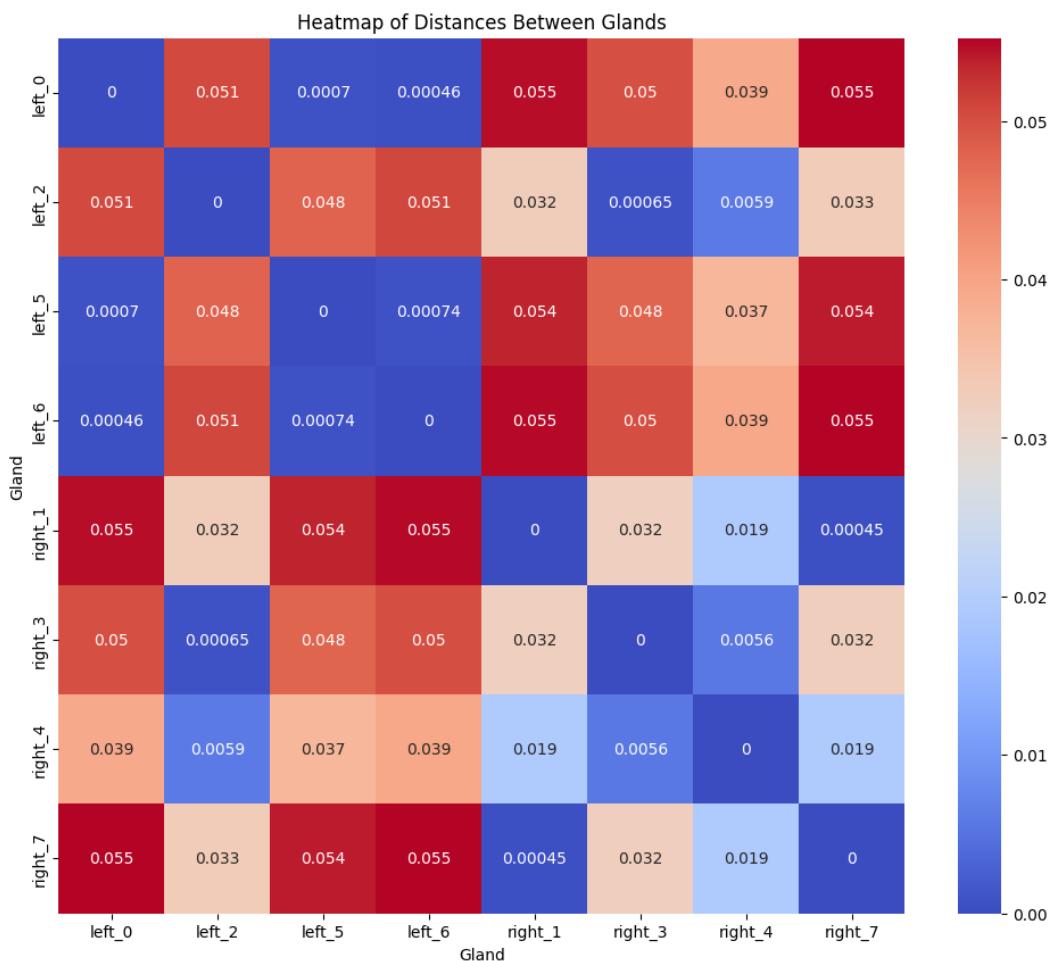


Figure 5.11: Distances between glands from figure 5.10.

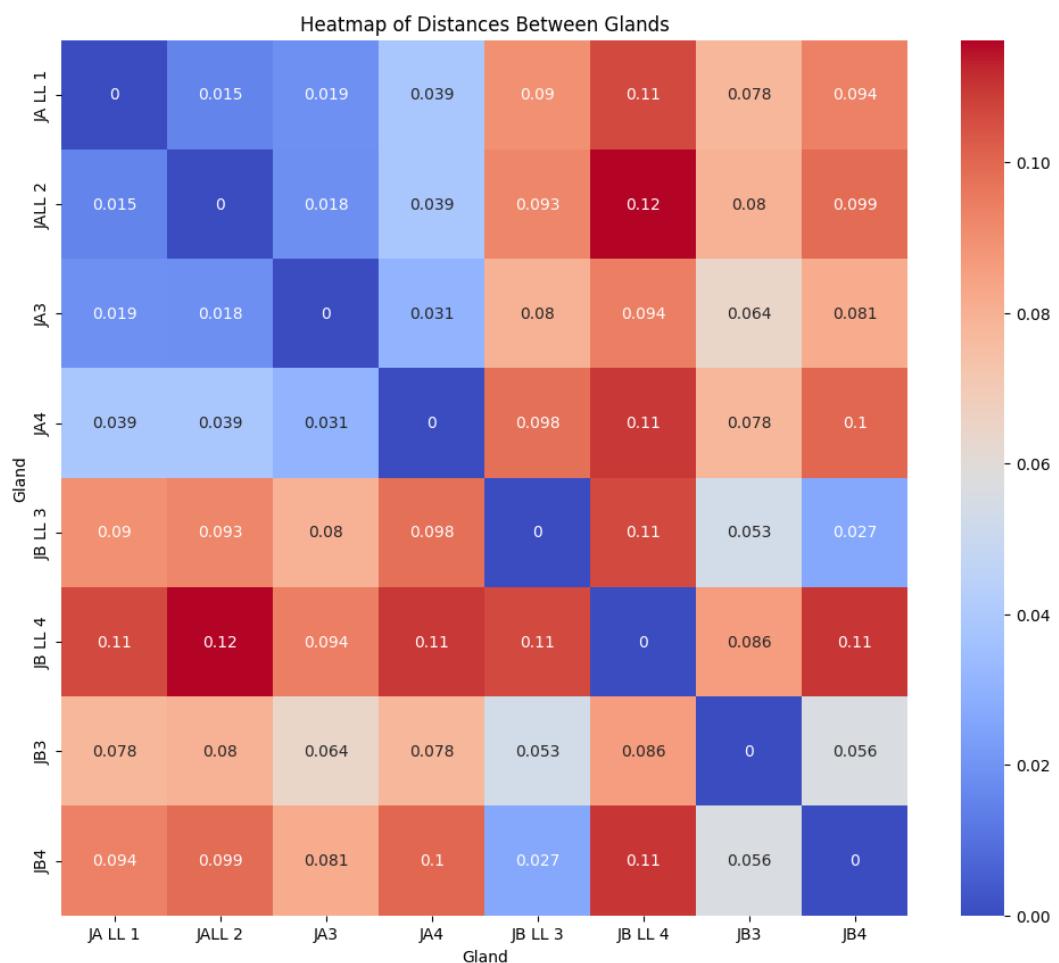


Figure 5.12: Distances between glands from data set J.

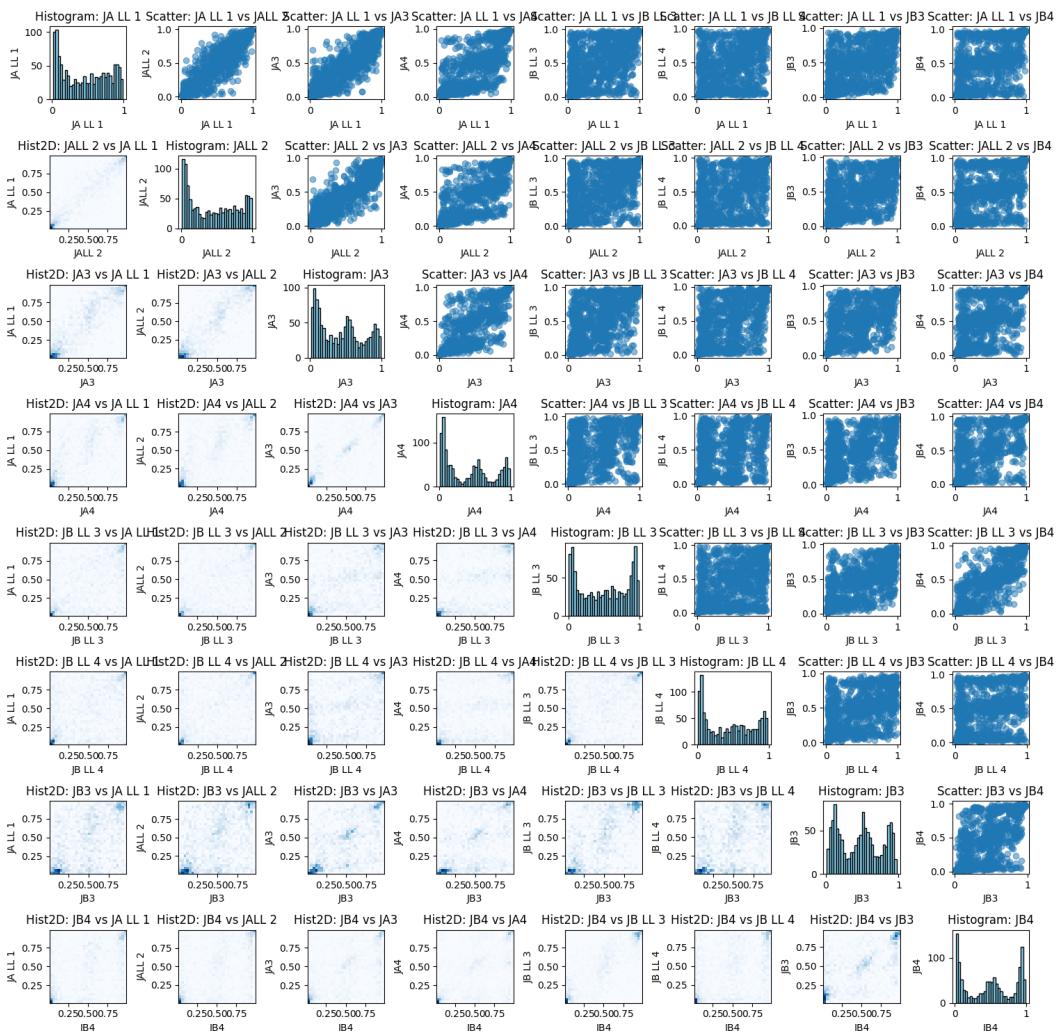


Figure 5.13: Data set J whose distances are shown in 5.12.

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>

Bozic, I., Paterson, C. & Waclaw, B. (n.d.), ‘On measuring selection in cancer from subclonal mutation frequencies’, **15**(9), e1007368. Publisher: Public Library of Science.

URL: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007368>

Bravo, R. R., Baratchart, E., West, J., Schenck, R. O., Miller, A. K., Gallaher, J.,

Gatenbee, C. D., Basanta, D., Robertson-Tessi, M. & Anderson, A. R. A. (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/ploscompbiol/article?id=10.1371/journal.pcbi.1007635>

Bull, J. A. & Byrne, H. M. (n.d.), ‘The hallmarks of mathematical oncology’, pp. 1–18. Conference Name: Proceedings of the IEEE.

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/>

Chauvin, B. & Rouault, A. (n.d.), ‘Connecting yule process, bisection and binary search tree via martingales’.

URL: <https://arxiv.org/abs/math/0410318v1>

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>

Fleming, M., Ravula, S., Tatishchev, S. F. & Wang, H. L. (n.d.), ‘Colorectal carcinoma: Pathologic aspects’, **3**(3), 153–173.

URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3418538/>

Gabbutt, C., Duran-Ferrer, M., Grant, H., Mallo, D., Nadeu, F., Househam, J., Villamor, N., Krali, O., Nordlund, J., Zenz, T., Campo, E., Lopez-Guillermo, A., Fitzgibbon, J., Barnes, C. P., Shibata, D., Martin-Subero, J. I. & Graham, T. A. (n.d.), ‘Evolutionary dynamics of 1,976 lymphoid malignancies predict clinical outcome’. Pages: 2023.11.10.23298336. Number: 5

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

Gabbutt, C., Schenck, R. O., Weisenberger, D. J., Kimberley, C., Berner, A., Househam, J., Lakatos, E., Robertson-Tessi, M., Martin, I., Patel, R., Clark, S. K., Latchford, A., Barnes, C. P., Leedham, S. J., Anderson, A. R. A., Graham, T. A. & Shibata, D. (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>

Heide, T., Zapata, L., Williams, M. J., Werner, B., Caravagna, G., Barnes, C. P., Graham, T. A. & Sottoriva, A. (n.d.), ‘Reply to ‘neutral tumor evolution?’’,

50(12), 1633–1637. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Cancer;Genome informatics Subject_term_id: cancer;genome-informatics.

URL: <https://www.nature.com/articles/s41588-018-0256-z>

Huffman, D. A. (n.d.), ‘A method for the construction of minimum-redundancy codes’, **40**(9), 1098–1101. Conference Name: Proceedings of the IRE.

URL: <https://ieeexplore.ieee.org/document/4051119>

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>

Kourou, K., Exarchos, K. P., Papaloukas, C., Sakaloglou, P., Exarchos, T. & Fotiadis, D. I. (n.d.), ‘Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis’, **19**, 5546–5555.

URL: <https://linkinghub.elsevier.com/retrieve/pii/S2001037021004281>

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>

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>

Nagaraj, S. V. (n.d.), ‘Optimal binary search trees’, **188**(1), 1–44.

URL: <https://www.sciencedirect.com/science/article/pii/S0304397596003209>

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

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., Pradels, J., Wong, C. K. & Yue, P. C. (n.d.), ‘Bounds on the weighted path length of binary trees’, **1**(6), 220–225.

URL: <https://www.sciencedirect.com/science/article/pii/0020019072900154>

Nievergelt, J. & Reingold, E. M. (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>

Niida, A., Mimori, K., Shibata, T. & Miyano, S. (n.d.), ‘Modeling colorectal cancer evolution’.

URL: <http://www.nature.com/articles/s10038-021-00930-0>

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

URL: <https://github.com/robjohnnoble/demonmodel>

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>

Nowell, P. C. (n.d.), ‘The clonal evolution of tumor cell populations’, **194**(4260), 23–28.

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>

Pasco, R. (n.d.), ‘Source coding algorithms for fast data compression (ph.d. thesis abstr.)’, **23**(4), 548–548. Conference Name: IEEE Transactions on Information Theory.

URL: <https://ieeexplore.ieee.org/document/1055739>

Paterson, C., Clevers, H. & Bozic, I. (n.d.), ‘Mathematical model of colorectal cancer initiation’, **117**(34), 20681–20688.

URL: <http://www.pnas.org/lookup/doi/10.1073/pnas.2003771117>

Patrone, M. V., Hubbs, J. L., Bailey, J. E. & Marks, L. B. (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.

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

URL: <https://iopscience.iop.org/article/10.1088/1478-3975/ab1a09>

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>

Sackin, M. J. (n.d.), ‘“good” and “bad” phenograms’, **21**(2), 225–226.

URL: <https://doi.org/10.1093/sysbio/21.2.225>

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

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

Tarabichi, M., Martincorena, I., Gerstung, M., Leroi, A. M., Markowetz, F., Spellman, P. T., Morris, Q. D., Lingjærde, O. C., Wedge, D. C. & Van Loo, P. (n.d.), ‘Neutral tumor evolution?’, **50**(12), 1630–1633. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Correspondence Publisher: Nature Publishing Group Subject_term: Cancer;Computational biology and bioinformatics;Genomics;Mathematics and computing Subject_term_id: cancer;computational-biology-and-bioinformatics;genomics;mathematics-and-computing.

URL: <https://www.nature.com/articles/s41588-018-0258-x>

Tavare, 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/>

Tung, H.-R. & Durrett, R. (n.d.), ‘Signatures of neutral evolution in exponentially growing tumors: A theoretical perspective’, **17**(2), e1008701. Publisher: Public Library of Science.

URL: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008701>

Werner, B., Dingli, D. & Traulsen, A. (n.d.), ‘A deterministic model for the occurrence and dynamics of multiple mutations in hierarchically organized tissues’, **10**(85), 20130349.

URL: <https://royalsocietypublishing.org/doi/10.1098/rsif.2013.0349>

West, J., Schenck, R. O., Gatenbee, C., Robertson-Tessi, M. & Anderson, A. R. A. (n.d.), ‘Normal tissue architecture determines the evolutionary course of cancer’, **12**(1), 2060. Number: 1 Publisher: Nature Publishing Group.

URL: <https://www.nature.com/articles/s41467-021-22123-1>

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>

Yin, A., Moes, D. J. A., van Hasselt, J. G., Swen, J. J. & Guchelaar, H. (n.d.), ‘A review of mathematical models for tumor dynamics and treatment resistance evolution of solid tumors’, **8**(10), 720–737.

URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6813171/>

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>