Mathematical theory of tumour evolutionary modes

Kimberley Verity

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

City St George's, University of London

Department of Mathematics

January 2026

Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my supervisor, Rob Noble, for his support throughout my PhD journey. Not only have his insights, skills and advice been invaluable in completing this thesis, but his kindness, patience and respect for a healthy work-life balance have also made this time incredibly enjoyable.

I am grateful to my family for their constant belief in me and for supporting me throughout this journey. To my brother, Mathew, who I informed at the age of 12 that I was going to get a PhD when I was older, and who replied that there was no way I was still going to want to do that when the time came, I told you so!

I am incredibly grateful to my close friends, Issy, Aimee, Jason and Emma, and my partner Charlie. Thank you for your encouragement, support and unwavering friendship.

I would also like to dedicate this work to the memory of Paul, who saw me begin this journey but will not see me finish it. He was a wonderful step-dad who always saw the best in me, believed I was capable of great things, and was ultimately my inspiration to pursue mathematical biology.

Declaration

I, Kimberley Verity, confirm that this thesis, "Mathematical theory of tumour evolutionary modes", and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- This thesis has not previously been submitted for a degree or any other qualification at this University or any other institution.
- Where the thesis is based on work done in collaboration, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Kimberley Verity

Date: 29/01/2025

Abstract

The analysis of rooted trees is important in biology and many other fields of study. Despite this, tree shape as a whole is not fully exploited, and the insights it can provide and its limitations are poorly understood. Additionally, the indices that are currently used to quantify and describe tree shape have properties that complicate interpretation and limit their applications.

This thesis uses and evaluates a newly proposed system of indices designed to address these limitations. The indices account for branch lengths and type abundances as well as tree topology, are robust to small changes to the tree, are universally applicable, and have straightforward interpretations. Using this system, the thesis investigates tree shape and the insights it can provide in macroevolution and cancer settings.

The results demonstrate that tree shape can capture meaningful signals across a range of contexts, including branching rate heterogeneity, differences between modes of evolution and associations with disease-free survival in cancer trees. Although tree balance often performs well, it is not the only index that is informative, and reliance on balance alone can neglect important aspects of tree shape. Across the settings I considered, the system of indices used here frequently outperforms existing alternatives.

Taken together, this thesis strengthens the case for tree shape as a broadly applicable tool for studying evolutionary processes, and highlights the importance of using a system of indices that are also robust, universal, and interpretable to fully exploit the information captured by tree shape.

Chapter 1

Introduction

The analysis of rooted trees is fundamental to many areas of research and has the potential to be an effective new method for others. Rooted tree data is abundant and readily available, making it a valuable data source for studying the processes by which it was generated. Despite its potential and the wealth of data, rooted trees are frequently underutilised.

Trees could provide valuable insights within cancer research. Well-studied and popular concepts from evolutionary biology are now being increasingly applied to cancer, including the modes of evolution and diversity, both of which have been shown to have important clinical implications. In macroevolution, imbalance is frequently observed in the Tree of Life, yet understanding the causes and, in particular, disentangling the factors that contribute to it, remains an open challenge.

Two popular methods for studying trees are, analysing tree shape using indices and analysing branch lengths, though these approaches are often not being distinct as tree indices frequently incorporate branch length information. Many of the currently used indices have properties that lead to confusion in their interpretation and limit their application, and this lack of suitable indices is one reason trees are not fully exploited as a method. This motivated the development of a new system of indices that does not suffer from any of these issues. This system accounts for branch lengths and type abundances as well as tree topology. They are robust to small changes to the tree, are universally applicable, and have straightforward interpretations. This thesis investigates this new system of indices and explores what evolutionary tree shape can provide insights into in both cancer and macroevolution.

1.1 Mode of evolution

The term ‘mode of evolution’ was originally introduced in palaeontology by George G. Simpson. In his seminal work ‘Tempo and Mode in Evolution’, Simpson defines two crucial topics. The first is ‘tempo’, which concerns the rate of evolution, and it is “practically defined as amount

of morphological change relative to a standard" [1]. The second is 'mode', which Simpson defines as "involving the study of the way, manner, or pattern of evolution" [1]. Additionally, he describes tempo as a factor that contributes to the mode, but mode constitutes much more than just tempo [1].

Simpson defines three modes of evolution to generalise the many evolutionary events and types that are observed. These are speciation, phyletic evolution and quantum evolution [2]. Speciation describes a "low-level process of iterating diversity" which does not have a "significant input to trends or other large scale patterns" [1]. Phyletic evolution is the "sustained, directional ... shift of the average characteristics of a population" [2]. Quantum evolution describes "rapid and rare ... 'all-or-nothing' transitions" [1]. In 1972, Eldredge and Gould proposed a fourth evolutionary mode called 'stasis'. This mode represents the "absence of net change in a species" [3]. They also hypothesised that a 'punctuated equilibrium' existed, which includes speciational evolution both preceded and followed by stasis [3]. Additionally, 'random walks' were later suggested as a fifth mode of evolution [3]. Random walks can either be biased or unbiased. A biased random walk represents directional evolution, where evolutionary divergence occurs continuously, and members of the same lineage are frequently discriminated from each other [4]. An unbiased random walk represents evolution which is not inherently directional; the divergence is expected to increase over time due to the accumulation of phenotypic differences [4].

The definition of mode provided by Simpson is general and unspecific. When defining his three modes, Simpson stated that they did not "represent the only or ultimate elements of evolutionary patterns" [2]. He himself knew that this concept would encapsulate many elements. As an increasing number of research areas consider an evolutionary perspective, and given the nature of its definition, the exact interpretation and what is deemed a 'mode' often changes between fields of study and the application. This presents a variety of issues, a major one being that research into the mode of evolution using different definitions of the term is fundamentally researching different concepts. A recent article argued the benefit and need of reuniting philosophical approaches with science, particularly in interdisciplinary fields [5]. Such an approach would utilise classical tools of philosophy including, conceptual clarification, critical assessment and connection between different disciplines [5], with the benefits of using these tools being self-explanatory. This framework would be useful for terms such as the mode of evolution, which are used across different fields and whose definitions have become increasingly diffuse.

Generally, how the mode of evolution is interpreted falls into one of four categories. The first is the overall pattern of evolution observed, such as linear, neutral, branched and punctuated evolution (see [6–19]). The second is the method of modelling such as directional change, unbiased random walks and stasis (see [20–24]). The third is the mechanism or process of evolution, such as natural selection and genetic drift (see [25–39]). Finally, the fourth interpretation is anything that falls under the umbrella of Simpsons' original definitions (see [40–43]). It should be emphasised that these four categories were defined in an attempt to describe what is observed in the literature. They are not necessarily distinct, as many concepts are interlinked, and there will be some crossover. For example, in the case of simulated data, the overall pattern observed is a direct result of the model used.

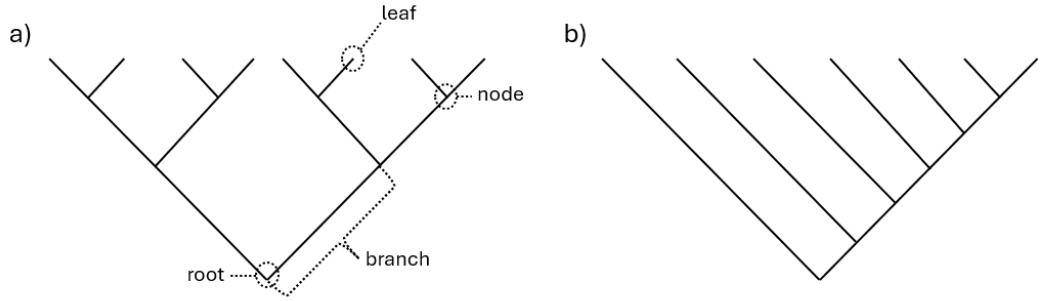

Figure 1.1: a) A perfectly balanced tree, demonstrating basic tree terminology. b) An unbalanced tree. Figure recreated from [44].

I completed a small review of 45 articles containing either the phrase 'mode of evolution' or 'modes of evolution', and this highlighted a clear disparity in how the term is defined, both generally and in cancer research (see Supplementary Table 8.1 for a full breakdown). In non-cancer cases, papers span all categories, with the most common being mechanisms of evolution. When studies are about cancer, they usually fall into the first or third category, with the first being more common. The overall observed pattern and the process of evolution are highly linked. It is the combination of all the underlying processes that leads to the overall observed pattern. Despite this, using both to mean the mode of evolution can be problematic, as they are inherently different. In cancer, the chosen definition is crucial, as it can significantly affect the interpretation of results and clinical implications.

The popularity of the pattern-based definition of evolutionary modes is not coincidental. In real biological systems, such as cancer evolution and macroevolution (explored further in sections 1.4 and 1.5), we typically observe only a single evolutionary history, or, when multiple histories are available, they are biologically distinct and inherently noisy. It is the resulting patterns that are directly observable, whereas the underlying evolutionary process is not and must therefore be inferred from these patterns. This inference is fundamentally uncertain, as completely different processes can, by chance, lead to the same pattern. Hence, when only a single tree, or multiple noisy trees, are available, the true process can not be known with certainty, and evolutionary modes are consequently defined in terms of the pattern rather than process. This is in contrast with simulation studies, where the full information is available about the true process, and the evolutionary mode can therefore be defined in terms of either the underlying process or the resulting pattern.

1.2 Trees

Concepts such as the tempo and mode of evolution were traditionally inferred by observing fossil records [44]. Simpson first defined tempo and mode to explain variations seen in the fossil record [1]. However, another popular and powerful way to study the tempo and mode of evolution is with phylogenetic trees [44]. The use of phylogenetic trees was not favoured historically due to a lack of data, but the increasing volume of data has made them a new and

effective tool for studying tempo and mode [44].

Trees typically have three sets of information: branch lengths, node sizes and topology. The first two can be chosen to be proportional to values of particular interest. Branch lengths could be proportional to time, and including branch lengths would then add a time axis and allow the study of diversification rates among lineages and their variations [44]. In this way, branch length is typically studied by observing patterns and comparing these with null models [44]. However, branch lengths may also represent taxonomic or phylogenetic distance, or even some other measure of change [45]. Nodes could correspond to biological taxa or clones, and hence node sizes would be equal to the corresponding sizes, such as taxonomic group size or cancer clone population size [46]. Or, if one only wants to consider whether each type is extant or extinct, nodes would be assigned size zero for extinct taxa and one for extant [46].

The study of trees can be broadly categorised into two methods. The first considers the topology or shape of the trees, and the second considers the branch lengths [47], though these approaches are often combined into a single one. To study and compare tree topology, concepts that can be quantified numerically using tree information are defined. One example is tree balance [44]. Tree balance is defined as the "extent to which nodes define subgroups of equal sizes", and to mathematically define tree balance, a summary index is used [44]. Multiple indices have been proposed, but they all have the same fundamental aim: to summarise the extent to which a tree is balanced or imbalanced in a single number [44]. Figure 1.1 illustrates tree balance by showing two trees, one which is perfectly balanced and one which is not balanced.

1.3 Indices

In this section, I will give a brief overview of the 'jungle' of indices, as it is commonly referred to as, and the issues with the continued use of many of the current indices. See [48–50] for detailed index reviews.

Before discussing indices in further detail, I will first outline some important definitions. I only consider rooted trees, which are trees where one node is the root and all branches are directed away from it. A node that has no descendants is a leaf and a non-leaf is an internal node. The out-degree of a node is the number of direct descendants of non-zero size. A bifurcating tree is a tree where every internal node has an out-degree of two, and tree size is the number of leaves in the tree. Finally, two common tree shapes are star and caterpillar trees. A star tree is a tree where all leaves are attached to the root. A caterpillar tree is a bifurcating tree where every internal node except one has one leaf. Figure 1.1 illustrates some of these concepts; it shows two rooted, bifurcating trees of sizes 8 and 7, respectively, and b is a caterpillar tree.

Indices aim to quantify a concept, which could be a tree property, such as tree balance, or a wider notion, such as diversity, all with just a single value. Indices quantifying tree balance (or imbalance) and diversity are the most commonly used and they will be the focus of this section. To quantify such concepts, indices use different components or 'units'. Common phylogenetic units used for diversity indices include branch lengths, pairwise phylogenetic distances between

taxa, nearest distances, and other measures of tree topology [48]. However, they do not have to consider only phylogenetic information; they can also be based on taxonomic, genetic, or even functional information. Tree balance indices rarely use such units and are typically based on components such as: comparison of subtree sizes, symmetry and node or subtree depths [50].

Tree balance indices are used in many fields of study, including systematic biology [44, 51], virology [52, 53], epidemiology [54, 55] and oncology [56–58]. They are typically used to compare models and infer parameter values [59]. Diversity indices are also used in multiple fields of study, including ecology [48] and oncology [60, 61]. Conservation biologists use diversity indices to help inform decisions regarding biodiversity [62], and in oncology, they are primarily used to study intratumour heterogeneity. A variety of indices have been created and are used for these purposes [48, 50], but many have properties that lead to confusion in their interpretation and limit their applications.

Two popular imbalance indices are Sackin's index and Colless' index. If we let T be a tree with a set of leaves $L(T)$. Then Sackin's index is defined as,

$$I_S(T) = \sum_{l \in L(T)} v_l,$$

where for a leaf $l \in L(T)$, v_l is the number of nodes between l and the root, including the root [46, 63]. This is an imbalance index, as the larger the value, the more imbalanced the tree. Sackin's index increases with the size of the tree, so it is often useful to normalise it. One way to normalise it for trees with $n > 2$ is by subtracting the minimum value and dividing through by the difference between the maximum and the minimum, [46]. The minimum value is achieved on the star tree where $\min_n(I_S) = n$, and the maximum is achieved on the caterpillar tree, where [46],

$$\max_n(I_S) = \frac{(n-1)(n+2)}{2}. \tag{1.1}$$

The normalised Sackin's index is then,

$$I_{S,norm}(T) = \frac{I_S(T) - n}{(n+2)(n-1)/2 - n}.$$

If we have a bifurcating tree T with internal node i, define n_{i_1} as the number of leaves of the left branch of the subtree rooted at i, and n_{i_2} as the number of leaves of the right branch. Then Colless' index is defined as,

$$I_C(T) = \sum_{i \in \tilde{V}(T)} |n_{i_1} - n_{i_2}|,$$

where $\tilde{V}(T)$ is the set of all internal nodes of T [46, 64]. This index can also be normalised for trees with $n > 2$ by dividing through by its maximum value, which is achieved on the caterpillar tree [46],

$$I_C(T) = \sum_{i \in \tilde{V}(T)} \frac{|n_{i_1} - n_{i_2}|}{\binom{n-1}{2}}.$$

The standard tree balance and imbalance indices, including Sackin's and Colless' indices, have properties that can lead to problems and limit their application [46, 59]. Common issues include not being defined for all trees, not allowing for meaningful comparison of all trees and being highly sensitive to the addition and removal of rare types [59], where a type is a biological lineage represented by a node in the tree (e.g. a species or clone). For example, Sackin's index can only be used to compare trees with the same number of leaves in its normal form, and Colless' can only be used on bifurcating trees [46]. Additionally, neither accounts for abundances and so are sensitive to the addition and removal of rare types [46]. A new balance index was recently introduced in [46], which does not suffer from any of these issues; however, it does not account for branch lengths.

Diversity indices have similar shortcomings. Popular diversity indices include richness, the Simpson index and the Shannon diversity index [49]. Let $p = (p_1, \ldots, p_N)$ be the vector of relative abundances where N is the total number of species. Richness is the simplest index and is given by,

$$H_R(p) = N,$$

which is just the number of species [49]. As this neglects to account for abundances, it is sensitive to the addition and removal of rare types, as the previous balance indices were. The Shannon diversity index is given by,

$$H_{Sh}(p) = -\sum_{i=1}^{N} p_i \ln(p_i),$$

and the Simpson index is given by [49],

$$H_{Si}(p) = 1 - \sum_{i=1}^{N} p_i^2.$$

Among other issues, none of these indices consider branch lengths. Many people have previously argued that different species make unequal contributions to diversity, such as one species being more genetically distinct from another [45]. Hence, diversity indices should incorporate information on similarity or dissimilarity, or equivalently, they should be defined on trees using branch lengths [65]. The most sophisticated diversity indices are the family introduced by Chao et al. [66]. They generalise many other indices, creating a unifying system rather than an entirely new one, and account for both branch lengths and abundances; however, they do not assign meaningful values to all trees [59]. Despite these issues, many of these indices are still used.

Due to the growing amount of phylogenetic information available, there has been a large increase in different phylogenetic indices [48]. One review article found 70 indices within ecological literature defining phylogenetic diversity alone [48], and another found at least 21 balance or imbalance indices introduced in the literature [50]. This profusion has led to confusion about which indices should be used. In addition to many of these indices not being ideal, the use of multiple different indices quantifying the same thing within fields can be an issue in itself. Take richness, Simpson's index and the Shannon diversity index as an example. Each one of

these indices has a different meaning when it comes to the diversity it is measuring. Richness is in units of species, the Simpson index is a probability and the Shannon index is a measure of entropy, so has units of 'bits of information' [49]. Any diversity index calculated using these different indices would not be immediately comparable as they, although attempting to quantify the same concept, are fundamentally different things. These types of indices hinder scientific study as they do not easily, if ever, allow for cross-comparison. This motivates the need for diversity indices in terms of 'effective numbers'. Converting classical diversity indices into effective numbers, as can be done for the three stated here, means diversity is measured in units of 'effective number of species', allowing for easy comparison and interpretation [49]. Additionally, it has been argued that diversity indices must be formulated in a way that is unambiguous and readily interpretable if they are to be meaningfully used by ecologists and conservationists, which further motivates frameworks based on effective numbers [67]. Despite most popular diversity indices being convertible into effective units, they are still regularly used in their classical form.

Given the profusion of indices and lack of a coherent framework surrounding them, it is easy to see how confusion over which indices to use arises. If there are many indices available to define a quantity, many of which are similar or are generalisations of each other and so behave similarly, how do you choose which one to use? As a result, many researchers use familiar indices, even if the index has properties that lead to issues or is not particularly suitable for the application. This motivates the need for indices which do not suffer from any of these shortcomings. These indices should be robust, universal and interpretable. Robust, meaning that small changes to the tree, such as the addition or removal of rare types, only have a small effect on the index value [59]. Universal meaning that they are defined for all rooted trees, and interpretable meaning they have both a simple and consistent interpretation [59]. Indices of this type would allow the comparison of any set of rooted trees.

1.4 Application to cancer

Cancer progression is an evolutionary process in which mutation, selection, genetic drift and cell dispersal result in a heterogeneous collection of cell subpopulations (clones) [57]. The subpopulations typically have different extents of aggressiveness and sensitivity to treatment [57]. Modern cancer research aims to characterise this process to improve clinical decision-making, enabling precise, patient-specific prognoses and the design of targeted therapy strategies [57]. Two ways to study this process are through the modes of evolution and diversity. However, it is crucial to note that in cancer research, trees, diversity measures, tree shape indices and other methods, such as Muller plots, all depend on how the term clone is defined.

For example, Merlo et al. observed that changing their definition of a clone, including from only driver mutations or only neutral (passenger) mutations, changed their diversity metric [69]. Where driver mutations are defined as mutations that provide a selective growth advantage to the cell, and passenger mutations are mutations that do not [70]. In this section, I focus on this particular change in definition, where clones are defined by driver mutations only or not (i.e.

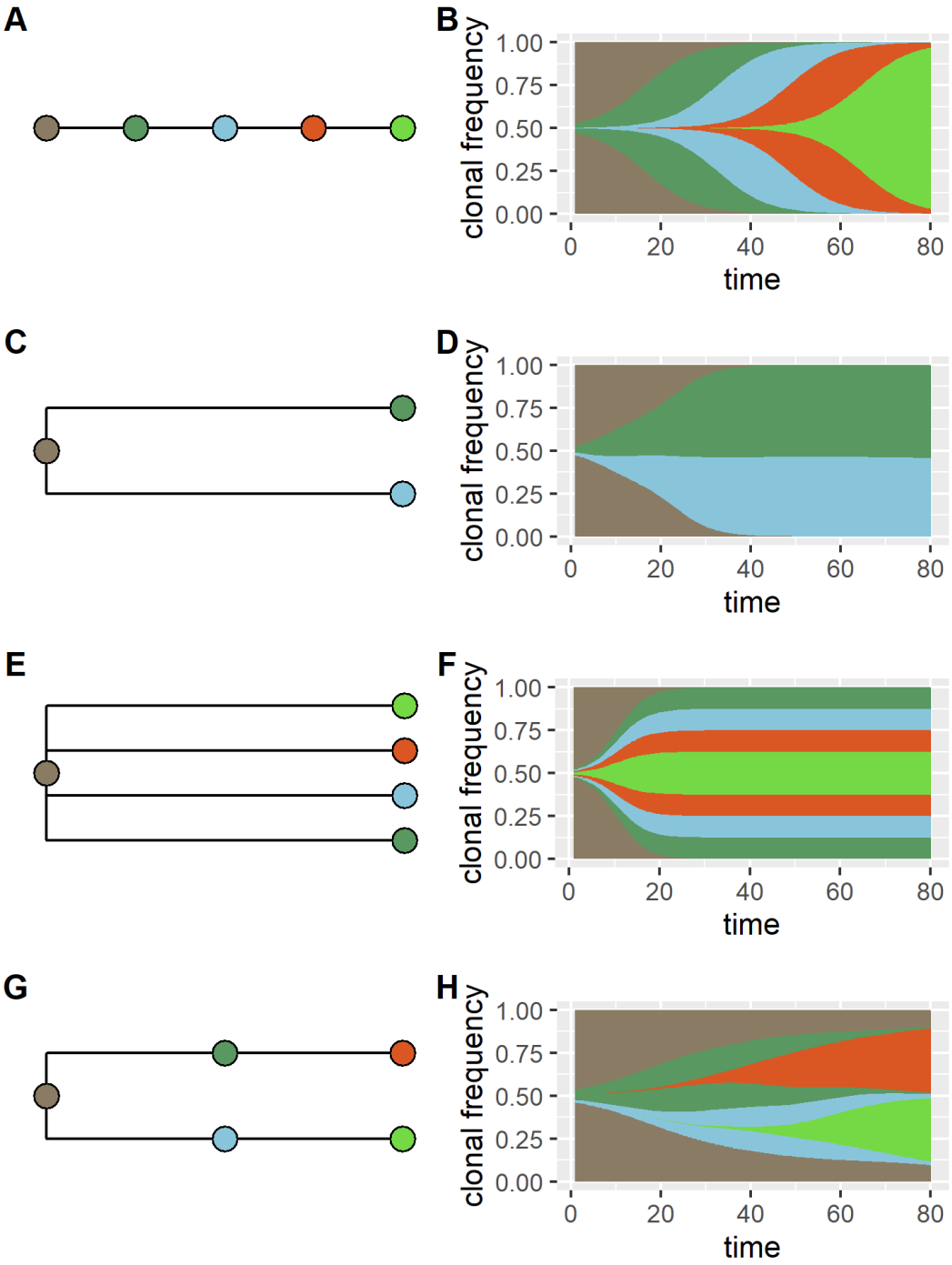

Figure 1.2: Schematic trees and the corresponding Muller plots illustrating the modes of tumour evolution. A-B) Linear evolution, C-D) neutral evolution. E-F) punctuated evolution, and G-H) branching evolution. Colours represent clones which are conceptually defined by driver mutations. (The data and code to create this figure were generated using ChatGPT-5.1. As it is schematic, the data was not model-based but was designed to create Muller plots that were visually as desired.)

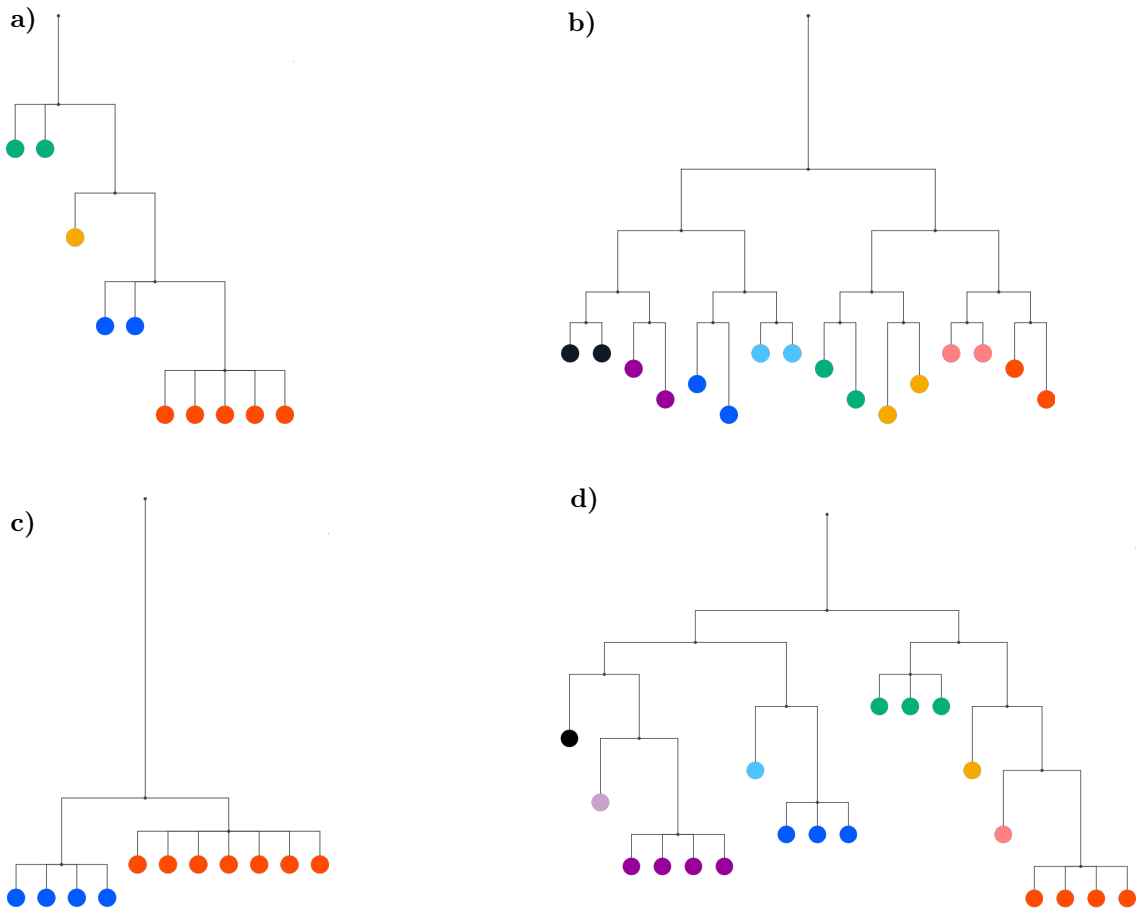

Figure 1.3: Schematic trees illustrating the difference in structure for each mode of evolution. a) Linear evolution, b) neutral evolution, c) punctuated evolution, and d) branching evolution. (Adapted from [68]; ordering modified).

clones could be defined by passenger mutations or all mutations). This choice affects both the methods and figures discussed. This issue is similar to my earlier discussion of the term mode of evolution. In fact, one of the authors of the paper that argued for reuniting philosophy and cancer has applied these ideas to the term clone, emphasising that the term should be carefully considered, as how we define and think about clones directly shapes how we treat cancer [71].

In cancer research, the mode of evolution is commonly defined in terms of the overall pattern observed (see section 1.1 for further detail); hence, this is the definition used in this section. Defined as such, evolutionary modes include patterns such as linear, branching, neutral and punctuated evolution. There are typical underlying processes that often lead to such patterns, which I will outline in terms of clones defined by driver mutations. The pattern of linear evolution is produced when new mutations provide a strong selective advantage and all previous clones are out-competed, leading to one dominant clone [68]. The pattern of branching evolution is formed by clones diverging from a common ancestor and then evolving in parallel, resulting in multiple clonal lineages [68]. Neutral evolution is a particular case of branching evolution and the pattern is formed when there are no selection or fitness changes throughout most of the evolutionary history. This results in very few clones present. Finally, the pattern

of punctuated evolution is formed when a large number of mutations occur early in tumour progression, subsequently, one or a few dominant clones then expand, forming the tumour [68]. Punctuated evolution is a term borrowed from Eldredge and Gould's 'punctuated equilibrium', and the latter is simply an analogy to the former as it is highly probable that the mechanisms behind each are very different [68].

Figure 1.3 and Figure 1.2 both show schematic phylogenetic trees for each of the modes of evolution, with the latter also showing the corresponding Muller plots. A Muller plot allows the visualisation of the genetic composition of a population and how this changes over time. In a Muller plot, each polygon represents a genotype or clone, and the thickness of this polygon represents the total number of individuals with that genotype or the frequency of the genotype in the population. Additionally, the nesting of genotypes represents evolutionary relationships [72]. In Figure 1.2, the trees and Muller plots are what would be expected when clones are defined by driver mutations. In contrast, the trees in Figure 1.3 show the typical shapes associated with each evolutionary mode, and there is not a single, consistent definition of clone corresponding with them. Both Figures 1.3 and 1.2 are schematic and intended to convey illustrative examples of phylogenetic trees and Muller plots for each mode of evolution rather than results of a model.

The development and evolution of tumours is associated with distinct tempo and modes [8], and the mode by which a tumour develops is important as different they have different diagnostic, prognostic and therapeutic implications [7]. For example, in terms of driver mutations, a tumour following linear evolution will have a dominant clone and low intratumour heterogeneity (ITH), and as a result would require a simpler biopsy as a few samples would be representative of the whole tumour [7]. A tumour following branching evolution will contain multiple clones and as a result, have higher ITH, meaning more biopsy samples would be required for the whole tumour to be represented accurately [7].

The study of diversity in cancer is typically done using a diversity index to quantify ITH. ITH refers to distinct tumour cell populations, which differ molecularly or phenotypically, within the same tumour [73]. ITH has been observed by pathologists for a long time, in the late 1800s Rudolf Virchow observed morphological differences between single tumour cells [68]. The subsequent introduction of next-generation sequencing (NGS) techniques revealed ITH is common in many human cancers [68]. Figure 1.2 demonstrates ITH visually, where clones consider driver mutations only. Linear evolution has extremely low or no ITH, shown by the almost entirely one colour at the end. Branching and punctuated evolution result in a few clones being present and will have some level of ITH, shown by the multiple colours at the end of the plots. Neutral evolution results in very few clones and consequently has a low ITH, as indicated by the two colours at the end.

A current area of research is the use of ITH as a prognostic biomarker, as ITH is associated with faster tumour progression, therapeutic resistance, treatment failure, and is a primary reason for poor survival in patients with metastatic disease [73, 74]. However, the relationship between ITH and clinical outcomes is complicated and inconsistent between different cancers [75]. Research into this relationship utilises a variety of indices to quantify ITH, including tree

indices [76], indices that use other data [69, 74, 77] (e.g. clone proportions from sequencing data, or copy number counts), and sometimes even a combination of the two [61, 78].

In macroevolution, tree indices are frequently used to study the mode of evolution (typically through studying patterns in shape and comparing to models) and diversity [44, 48, 50], providing valuable insights, and they have the potential to do so in cancer research too. Evolutionary mode and ITH are important topics in cancer, both with important clinical applications. However, tree indices are underutilised in these areas. Although there is increasing interest in tree indices, their use is limited, and as a result, it remains unclear if their potential will translate into real and useful results. When tree indices are used, the general issues with the current index landscape as described in the previous section persist. The indices have properties that lead to confusion in their interpretation and limit their application, potentially affecting the results of work using them.

1.5 Macroevolution and the tree of life

The Tree of Life is a phylogenetic tree showing the relationship between all extant species on Earth. Such a tree offers a clear view of the diversity of life on our planet. As the shape of every phylogenetic tree reflects, to some extent, the evolutionary processes that produced it, studying the shape of the Tree of Life can provide insight into the evolutionary process that led to this diversity [79, 80]. With this aim in mind, tree balance has been studied. Historically, tree balance was related to variation in speciation and extinction rates, and this rate variation is known to be associated with important macroevolutionary phenomena [79]. Therefore, quantifying past speciation and extinction rates with tree balance could provide crucial insights into these macroevolutionary processes [79]. Patterns in tree balance are not only the result of the evolutionary processes that produced the tree. They may be the result of things including stochasticity from models, the method used to estimate the tree, and the definition and inclusion of taxa [44]. Therefore, disentangling these effects to understand how they affect tree shape is vital when using tree balance to study the Tree of Life.

Stadler expanded on the fundamental idea of phylogenetic trees containing information about the underlying process in terms of speciation and extinction [47]. Stadler did this by examining the two classes of methods typically used: methods using true tree shape (i.e. tree topology without branch length information), and methods using branch length information [47]. She identified that through true tree shape, we can only detect variation in diversification that is lineage-specific, and we cannot quantify the magnitude of variation. This is where methods utilising branch lengths are required. For example, it has been repeatedly observed that empirical trees are less balanced than expected under simple speciation models [44], but the causes and consequences of this are not well understood. One reason for this is that topology-based measures cannot capture features that would help to explain the underlying dynamics, such as whether rates are increasing or decreasing globally, or the magnitude of those changes. Our indices bridge the gap between the two methods as our "tree shape" indices are not just topological, as they account for branch lengths.

Taking advantage of this fundamental idea, researchers have studied how diversification has varied across time periods, clades and geographic regions, and attributed the causes of this rate variation to organismal features, biotic interactions and abiotic conditions [81]. Surprisingly, many of these studies identified patterns in the shape of trees and the branch length distributions that occur repeatedly [81]. These patterns include: as I already outlined, empirical trees are often less balanced than expected under simple speciation models, branch lengths within time trees are often shorter early in a clade's history and longer towards the present, and the average branch lengths are usually shorter in young groups relative to older groups [81]. It remains a challenge to understand how the different patterns relate to each other. Additionally, many of the studies identifying such patterns were conducted on very different datasets. As there has been a great effort recently to construct large-scale phylogenetic trees for many taxonomic groups, this creates an opportunity to revisit many of the observed patterns in a now unified way [81]. This was the motivation for and purpose of the recent paper "The Major Features of Macroevolution" by Henao-Diaz and Pennell [81]. They find that their large-scale trees still exhibit structural patterns, including: greater imbalance than simple models predict, later sections of trees are more imbalanced and have more heterogeneous branch lengths than earlier sections, scale-invariance and weak dependence on taxonomy [81]. Although they were able to identify these patterns, the underlying causes of them are still not understood.

It is important to note that while phylogenetic trees are often considered direct representations of evolutionary history, different tree diagrams encode different relationships between time, ancestry and topology [82]. Podani provided a graph-theoretical classification of rooted trees used in evolutionary biology, distinguishing diachronous trees, the "true" evolutionary trees, from synchronous and achronous trees that represent abstractions of evolutionary history [82]. Most phylogenetic trees used in macroevolutionary analysis fall into the latter category and recover at best the backbone topology of the underlying evolutionary process. This is an important distinction that must be kept in mind when interpreting any summary of a tree, such as tree shape indices, as they are actually summarising the abstract representation, rather than the true evolutionary history.

Together, these results show that phylogenetic trees used in macroevolutionary analyses are neither direct representations of evolutionary history nor completely random objects. Podani highlighted that most trees recover only a structural abstraction of the underlying evolutionary process [82]. Stadler demonstrated that there are limits on what can be inferred from topology alone, with true tree shape statistics being sensitive only to lineage-specific diversification dynamics [47]. Despite being abstractions, there are persistent occurrences of patterns in empirical phylogenies which suggest that this abstraction still retains biologically meaningful information, and the causes of such repeating patterns are unknown. This motivates further study of these patterns, particularly with methods that incorporate topology and branch lengths, such as our system of indices.

1.6 Outline

In this section, I outline the work in the subsequent chapters. In chapter 2, I provide the definitions of our new system of indices and show how they are calculated for a simple example. They are then applied to empirical trees, showing two things: using tree shape indices that account for branch lengths is crucial for discriminating between trees, and that our indices can discriminate between trees generated by different tumour growth models. Finally, comparing the empirical trees to the Yule process, I show that our set of index values appears to be specific to the process that generated them. This chapter is based on work presented in a preprint [59].

In chapter 3, I investigate how our indices behave for non-spatial tree-generating models. In particular, I show that averaging over trajectories in index space massively increases correlations between index pairs and leads to higher index redundancy. I derive the expected values of our indices for 2 and 3-leaf trees, and show that under the Yule process, the distributions and expectations of our indices are independent of birth rate. Additionally, using the Yule process, I demonstrate that our indices are robust to unresolved polytomies.

In chapter 4, I explore whether our indices can detect branching rate heterogeneity using two different models. I show that our indices can distinguish between the models and the null model, and they can discriminate between the two models themselves. Also, I use both a Bayesian and a frequentist approach and find they lead to very similar results.

In chapter 5, I consider two definitions of the mode of evolution and investigate whether our indices can distinguish between modes in both cases. I show that when the mode of evolution is defined as the observed pattern, our indices cluster the modes within different index spaces and so can distinguish between them, but this is not the case when the mode of evolution is defined as the underlying process. Additionally, I explore whether we can detect model parameters using our indices and find that generally our indices perform poorly, but parameters can be detected in certain situations.

In chapter 6, I explore how tree balance varies with taxonomic level, reinvestigating the questions studied by Purvis and Agapow in [51]. Contrary to the previous study, I find that trees with leaves of a higher taxonomic level than species are more unbalanced when species richness is considered than when it is not. I also find that, for nodes of a given size, those with species as leaves are more unbalanced than those with higher taxa as leaves. Finally, I show how the imbalance index used by Purvis and Agapow is not ideal for their use, and that its supposed null model property is flawed with typical use. This chapter is based on a paper in progress.

In chapter 7, I apply our indices to evolutionary cancer trees from TRACERx's non-small cell lung cancer cohort. I show that there is a significant relationship between disease-free survival and our indices, and in particular, tree balance performs the best. I find that tree balance remains a significant predictor of disease-free survival even after stage is adjusted for. I show that our results are robust to the removal of rare nodes. Finally, I find that in high-quality data, tree topology accounts for the main diagnostic signal, but as data quality decreases, abundances and branch lengths are crucial for patient stratification. This chapter is based on

a preprint [83].

Chapter 2

A new system of indices

2.1 Introduction

This chapter is primarily based on work presented in [59], a preprint co-authored with my supervisor, Robert Noble. I developed the computational implementations, performed the analyses on empirical data, and contributed to the interpretation of results. The material has been adapted and integrated into this thesis to provide a foundation for subsequent chapters.

Here, I will outline our new system of indices defined in [59]. This system resolves the problems that can arise from properties of popular indices, some of which were previously outlined in the introduction (see [59] for a more detailed account). The indices do this by accounting for node sizes and branch lengths, being robust to small changes in either, assigning interpretable values to all trees, and enabling meaningful comparison of any set of trees. Our indices quantify all aspects of tree shape, including diversity (effective number of leaves) and tree balance, and belong to a coherent framework such that the mathematical relationships between the indices are well characterised.

To define the new system of indices, we first need to recall the family of diversity indices credited to Hill [84]. These indices are known as Hill numbers and are functions of a set of proportions $P = \{p_1, \ldots, p_n\}$ with $0 \leq p \leq 1$ for all $p \in P$ and $\sum_{i=1}^{n} p_i = 1$. The Hill numbers of order $q \geq 0$ are then defined as,

$$ {}^{q}D(P) := \left(\sum_{i=1}^{n} p_i^q \right)^{\frac{1}{1-q}} \text{ with } {}^{1}D(P) := \lim_{q \to 1} {}^{q}D(P) = \exp\left(-\sum_{i=1}^{n} p_i \log p_i \right). $$

An important special case is when $q = 0$,

$$ {}^{0}D(P) := |\{p \in P : p > 0\}|, $$

Type of average	Richness	Diversity (with $q > 0$)	Evenness (with $q > 0$)
Longitudinal mean	${}^0D_L$ = average branch count across the tree	${}^qD_L$ = effective number of maximally distant leaves	${}^qJ_L$ = evenness of branch sizes across the tree
Node-wise mean	${}^0D_N$ = average effective out-degree, ignoring branch sizes	${}^qD_N$ = average effective out-degree, accounting for branch sizes	${}^qJ_N$ = tree balance
Star mean	${}^0D_S$ = effective number of non-root nodes	${}^qD_S$ = effective number of branches, accounting for branch sizes	${}^qJ_S$ = evenness of all branch sizes

Table 2.1: Interpretations of the new indices.

which is the number of nodes or richness. We then define the evenness indices,

$${}^qJ(P) := \begin{cases} \frac{\log {}^qD(P)}{\log {}^0D(P)} \in [0,1] & \text{if } {}^0D(P) > 1 \\ 1 & \text{otherwise.} \end{cases}$$

These indices can then be applied to a rooted tree T by equating $P(T) = \{p_1, \ldots, p_n\}$ to the proportional sizes of the n nodes of T. The index ${}^0D(T) = {}^0D(P(T))$ is a richness index and quantifies the number of non-zero-sized nodes in the tree, we will refer to these as counted nodes. Counted nodes in an evolutionary tree correspond to extant types. Then for $q > 0$, the diversity index ${}^qD(T) = {}^qD(P(t))$ can be interpreted as the effective number of counted nodes, and ${}^qJ(T) = {}^qJ(P(T))$ measures the degree of the evenness of the counted node sizes. To keep notation simple, the tree as a function argument will typically be omitted.

Despite being robust to small changes in proportional node size, universal (as they can be applied to any set of nodes), and interpretable, they are not suitable for assessing tree shape as they only depend on node sizes and ignore both topology and branch lengths. To resolve this, these indices are extended to account for topology and branch lengths, creating our new system of indices. This system utilises three types of weighted mean, which we refer to as the longitudinal mean, node-wise mean and star mean.

2.2 Additional useful definitions

For a rooted tree, the depth of a node is the sum of the branch lengths along the unidirectional path from the root to the node, and the height of the tree is the maximum depth of its non-zero-sized nodes. The size of a branch will be defined as the sum of the proportional node sizes that descend, either directly or indirectly, from the branch. Take Figure 2.1a as an example. The branch 4-5 has size $\frac{2}{3}$ and all the other branches have size $\frac{1}{3}$. Note that the size of any

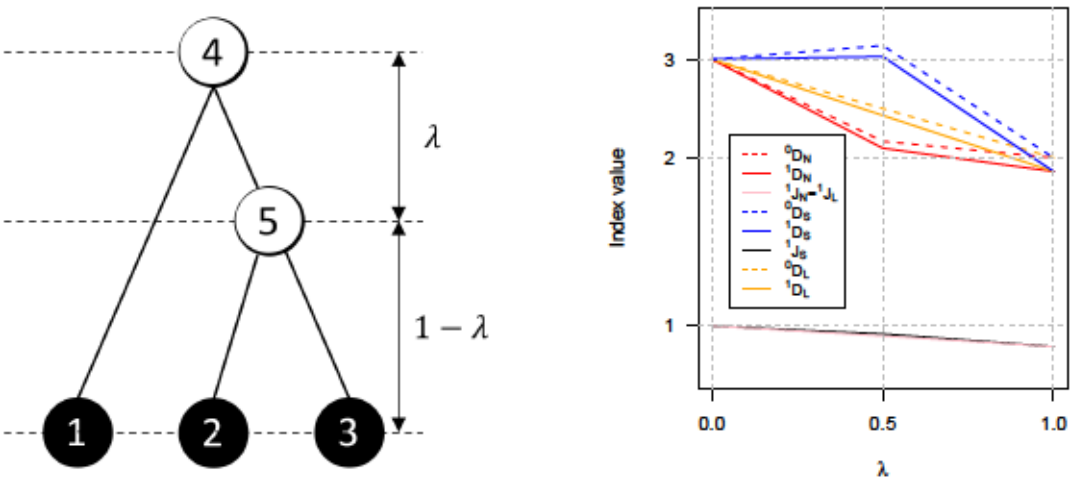

Figure 2.1: a) This is the simplest interesting case that can be considered. The leaves labelled 1, 2, and 3 each have size $\frac{1}{3}$ and the internal nodes 4 and 5 have size 0. The branch 4-5 has length λ, the branch 4-1 has length 1 and the branches 5-2 and 5-3 both have length $1-\lambda$. b) How the new indices vary for the tree in a) as the value of λ is changed.

segment of a branch is the same as the size of the branch.

A leafy tree has all internal nodes of size zero. A tree is ultrametric if, after the removal of all subtrees that contain only zero-sized branches, all its leaves have the same depth. A caterpillar tree is a bifurcating tree where every internal node except one has exactly one child leaf, and a star tree is a tree where all non-zero-sized branches are attached to a common node. Take Figure 2.1a as an example. If $\lambda = 0$ or $\lambda = 1$, then the tree is a star tree, and for all other cases, it is a caterpillar tree.

The new indices are weighted means, and our preferred weights require a normalising factor to be defined,

$$\bar{h} := \sum_{b \in B} s_b l_b \leq h,$$

where B is the set of all branches in the tree, $s_b \in [0,1]$ is the size of branch b, l_b is the length of branch b, and h is the height of the tree. The value $\bar{h}$ can be interpreted as the effective height of the tree, or as the average counted node depth, and $\bar{h} = h$ if and only if the tree is leafy and ultrametric.

In general, the new indices are defined for an arbitrary order q, where $q = 0$ does not account for branch lengths and considers all branches to have equal size. For $q > 0$, the indices account for both branch lengths and branch sizes, and as q increases, less weight is given to smaller branches. The general definitions of ${}^qD_L$, ${}^qD_N$ and ${}^qD_S$ and the derivation of all the indices from weighted means are provided in Chapter 2 supplementary material. In this thesis, I consider the cases $q = 0$ and 1.

2.3 Longitudinal mean

For the longitudinal mean, the idea is that the tree is split into transverse intervals (see the dashed lines in Figure 2.1a). The index value is then calculated within each interval, and a weighted average of the within-interval index values is taken.

Every interval i contains a set of branch segments B_i, all of the same length, which is referred to as the interval height h_i. Then define,

$$S_i := \sum_{b \in B_i} s_b \in (0, 1],$$

where s_b is the size of the branch segment b. $S_i = 1$ for all intervals if and only if the tree is leafy and ultrametric. It follows that,

$$\sum_{i \in I} S_i h_i = \bar{h}.$$

For every $b \in B_i$, the within-interval proportional branch size is defined as $p_b := s_b / S_i$ and let $P_i := \{p_b : b \in B_i, p_b > 0\}$. Then, $\sum_{p \in P_i} p = \sum_{b \in B_i} p_b = 1$ for all intervals $i \in I$. The longitudinal diversity indices are then,

$${}^0D_L = \begin{cases} \exp\left(\frac{1}{\bar{h}} \sum_{i \in I} S_i h_i \log |P_i|\right), & \text{if } h > 0 \\ 0 & \text{otherwise,} \end{cases}$$

$${}^1D_L = \begin{cases} \exp\left(-\frac{1}{\bar{h}} \sum_{i \in I} h_i \sum_{b \in B_i} s_b \log \frac{s_b}{S_i}\right), & \text{if } h > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The index ${}^0D_L$ can be interpreted as the average tree width. For $q > 0$, ${}^qD_L$ can be interpreted as the effective number of counted nodes maximally distinct from the root. As all maximally distinct counted nodes must be leaves, ${}^qD_L$ may also be interpreted as the effective number of maximally distinct leaves.

The balance index is given by,

$${}^qJ_L := \begin{cases} \frac{1}{\bar{h}} \sum_{i \in I} S_i h_i \, {}^qJ(P_i), & \text{if } h > 0 \\ 1 & \text{otherwise.} \end{cases}$$

${}^qJ_L$ measures the average evenness of branch sizes across the tree. If the tree is leafy and ultrametric then ${}^qJ_L = 1$ for $q > 0$ if and only if the tree is fully symmetric. Therefore, when applied to leafy ultrametric trees, ${}^qJ_L$ can be interpreted as a symmetry index.

To see how these indices are derived from a weighted mean, see Chapter 2 supplementary material. Figure 2.1b shows how ${}^0D_L$, ${}^1D_L$ and ${}^1J_L$ varies with branch length λ for the tree

shown in Figure 2.1a.

2.4 Node-wise mean

The basic idea for the node-wise mean is to calculate an index value for each node and then take a weighted average of the node index values. So that the index applies to all rooted trees, the node-wise mean needs to change continuously as we vary branch length. To do this, the index needs to depend not only on the branches descending from a node but the branches that also run alongside the node. Take the tree in Figure 2.1a. For node 5 we will take into account branches 5-3 and 5-2, but also the branch 4-5 as it runs alongside it. In general terms, when the distance between a node k and any ancestor j of k is less than the height of the subtree C_k, which contains k and its children only, then the index value assigned to k depends not only the branches of C_k but also on the branch segments that descend from j and that coexist in the transverse intervals with the branches of C_k. The weight assigned to k depends only on C_k but the index value assigned to k is a weighted average of index values across k and all ancestors of k.

A non-negative density f_b is assigned to every branch b at every depth x, with $f_b = 0$ for every x at which b is absent. The tree height is then defined as $h := \max\{x : f_b(x) > 0, b \in B\}$, where B is the set of all branches. Then the branch size s_b can be defined as the non-increasing function of depth x,

$$\bar{h} := \sum_{b\in B} \int_0^h f_b(x)\,dx, \quad s_b(x) := \begin{cases} \frac{1}{\bar{h}} \sum\limits_{b\in G_b} \int_x^h f_b(t)\,dt, & \text{if } h > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where G_b is the set containing b and all branches that descend from b. Let

$$S_{T_j}(x) := \sum_{b\in B_j} s_b(x) \in [0, 1],$$

where j is a node and T_j is the subtree containing j and all its descendants. For each $b \in B_j$, the proportional branch size is defined as

$$p_{b_j}(x) := \begin{cases} \frac{s_b(x)}{S_{T_j}(x)}, & \text{if } S_{T_j} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let $P_j(x) := \{p_{bj}(x) : b \in B_j, p_{bj}(x) > 0\}$. Then define,

$$S_{C_k}(x) := \sum_{b\in C_k} s_b(x),$$

and define the normalisation factor

$$\bar{h}_{C_j} := \int_0^h S_{C_j}(x)\,dx, \quad \Rightarrow \quad \bar{h} = \sum_{j\in V} \bar{h}_{C_j} = \int_0^h S(x)\,dx,$$

where V is the set of nodes. Let d_k denote the depth of node k and let $d_{jk} = d_k - d_j$ denote the distance from j to k. Let j' denote the parent of node j. Then the ancestor weight is defined as,

$$v_{jk} = \int_{\alpha_{jk}}^{\beta_{jk}} S_{C_k}(x)\,dx,$$

where $\alpha_{jk} := d_k + d_{jk}$ and

$$\beta_{jk} := \begin{cases} \alpha_{jk} + d_{j'j} & \text{if } j \text{ is not the root} \\ \infty & \text{otherwise.} \end{cases}$$

Finally, the node-wise mean diversity indices are given by,

$${}^0D_N = \begin{cases} \exp\left(\frac{1}{\bar{h}} \sum_{k\in V} \frac{1}{\bar{h}_{C_k}} \sum_{j\in A_k} v_{jk} \int_0^h S_{C_k}(x) \log|P_j(x)|\,dx \right) & \text{if } h > 0 \\ 0 & \text{otherwise,} \end{cases}$$

$${}^1D_N = \begin{cases} \exp\left(\frac{1}{\bar{h}} \sum_{k\in V} \frac{1}{\bar{h}_{C_k}} \sum_{j\in A_k} v_{jk} \int_0^h S_{C_k}(x)\, {}^1H(P_j(x))\,dx \right) & \text{if } h > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where

$${}^1H(P_j(x)) := \sum_{b\in B_j} \frac{s_b(x)}{S_{T_j}(x)} \log \frac{s_b(x)}{S_{T_j}(x)}.$$

The index ${}^qD_N$ can be interpreted as an average effective out-degree that accounts for branch lengths only when $q = 0$, or for both branch lengths and branch sizes when $q > 0$.

The universal tree balance ${}^qJ_N$ is given by,

$${}^qJ_N := \begin{cases} \frac{1}{\bar{h}} \sum_{k\in V} \frac{1}{\bar{h}_{C_k}} \sum_{j\in A_k} v_{jk} \int_0^h S_{C_k}(x)\, {}^qJ(P_j(x))\,dx & \text{if } h > 0 \\ 1 & \text{otherwise.} \end{cases}$$

This balance index assigns a balance score of 1 to any node that has out-degree 1, hence it considers linear trees to be maximally balanced. Other than this, for trees with uniform branch lengths ${}^qJ_N$ is identical to the previously defined J^1 balance index in [46]. Figure 2.1b shows how ${}^0D_N$, ${}^1D_N$ and ${}^1J_N$ as branch length varies for a three-leaf tree.

The longitudinal mean can also be written in integral form as the node-wise mean is here, see [59] for details. Also, see Chapter 2 supplementary material for the details of the derivation of the node-wise mean from a weighted mean.

2.5 Star mean

The star mean depends on branch sizes as the longitudinal and node-wise means do, but unlike them, the star mean ignores tree topology. The idea is that one effectively rearranges the tree by reattaching all branches to the root node to form a star tree while retaining branch sizes and lengths. Then the longitudinal, or equivalently node-wise, mean index value of the star tree is calculated.

Define $P^*(x) := \{p_b^*(x) : b \in B, p_b^*(x) > 0\}$, where

$$p_b^*(x) := \begin{cases} \frac{s_b(x+d_b)}{S^*(x)} & \text{if } S^*(x) > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad S^*(x) := \sum_{b \in B} s_b(x + d_b)$$

and d_b is the depth of the parent node of branch b. Also,

$$\int_0^h S^*(x)\, dx = \int_0^h S(x)\, dx = \bar{h}.$$

Then the star-mean diversity indices are given by,

$${}^0D_S = \begin{cases} \exp\left(\frac{1}{\bar{h}} \int_0^h S^*(x) \log |P^*(x)|\, dx\right) & \text{if } h > 0 \\ 0 & \text{otherwise,} \end{cases}$$

$${}^1D_S = \begin{cases} \exp\left(\frac{1}{\bar{h}} \int_0^h S^*(x)\, {}^1H(P^*(x))\, dx\right) & \text{if } h > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The index ${}^qD_S$ quantifies the effective number of branches in the tree, either accounting for only branch lengths when $q = 0$, or for both branch lengths and branch sizes for $q > 0$. Additionally, ${}^qJ_S$ which quantifies the evenness of all branch sizes is defined as,

$${}^qJ_S := \begin{cases} \frac{1}{\bar{h}} \int_0^h S^*(x)\, {}^qJ(P^*(x))\, dx & \text{if } h > 0 \\ 1 & \text{otherwise.} \end{cases}$$

2.6 A simple example

Figure 2.1a shows one of the simplest interesting trees that can be considered. For this tree, the leaves are equally sized, each has size $\frac{1}{3}$, and both internal nodes have size zero. Branch 4-5 has length λ, branch 4-1 has length 1, and branches 5-2 and 5-3 have length $1 - \lambda$. We will use this example to demonstrate how to calculate two of the new indices, ${}^1D_L$ and ${}^1D_N$. As the star mean can be thought of as a rearrangement of the tree and then calculating either the longitudinal or node-wise mean, its calculation will not be shown here.

2.6.1 Longitudinal mean

This tree has two regions, shown by the dashed lines in Figure 2.1a. If we let x be the distance from the root node, then we denote $0 < x < \lambda$ as region 1 and $\lambda < x < 1$ as region 2. Region 1 contains two branches of size $\frac{1}{3}$ and $\frac{2}{3}$ which gives,

$$S_1 = 1 \quad \text{and} \quad h_1 = \lambda.$$

Region 2 contains three branches all of size $\frac{1}{3}$ giving

$$S_2 = 1 \quad \text{and} \quad h_2 = 1 - \lambda.$$

We also have that

$$\overline{h} = 1.$$

Putting these together gives,

$${}^1D_L = \exp\left(\log 3\left(1 - \frac{2}{3}\lambda\right)\right).$$

2.6.2 Node-wise mean

For the internal node labelled 4, which is the root node, there is only one ancestor that can be considered which is itself. For this node, for $x \in (0, \lambda)$ there are two branches of size $\frac{1}{3}$ and $\frac{2}{3}$, and for $x \in (\lambda, 1)$ there are three branches all with size $\frac{1}{3}$. This gives,

$$S_{C_4}(x) = \begin{cases} 1, & 0 \leq x \leq \lambda \\ \frac{1}{3}, & \lambda < x \leq 1 \end{cases},$$

$$\overline{h}_{C_4} = \frac{1}{3}(2\lambda + 1),$$

$${}^1H(P_4(x)) = \begin{cases} -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3}), & 0 \leq x \leq \lambda \\ -\log(\frac{1}{3}), & \lambda < x \leq 1 \end{cases}.$$

Putting these together gives,

$${}^1D_{N,4} = \frac{1}{3}(\lambda(2\log(3) - 2\log(2) + \log(3)),$$

where the extra numerical subscript denotes that it is the index value for node 4.

Node 5 has two ancestors to consider, node 4 and itself. Node 4 is only considered as an ancestor when $\lambda < \frac{1}{2}$. This is because if $\lambda > \frac{1}{2}$, then the node is closer to its children than to the ancestor and the ancestor integral will be zero. In either case, we have,

$$\overline{h}_{C_5} = \frac{2}{3}(1 - \lambda).$$

When considering itself as an ancestor, $k = 5$, for $x \in (0, 1-\lambda)$ there are two branches, both with size $\frac{1}{3}$. This gives,

$$S_{C_5}(x) = \frac{2}{3}, \qquad \text{for} \quad x \in (0, 1-\lambda),$$
$${}^1H(P_5(x)) = \log(2), \qquad \text{for} \quad x \in (0, 1-\lambda).$$

Therefore, when $\lambda \geq \frac{1}{2}$,

$${}^1D_{N,5} = \frac{2}{3}(1-\lambda)\log(2).$$

When $\lambda < \frac{1}{2}$, we consider both itself and node 4 as an ancestor. There are three branches each with size $\frac{1}{3}$, hence,

$${}^1H(P_4(x)) = -\log(3) \quad \text{for} \quad x \in (0, 1-\lambda),$$

which gives,

$${}^1D_{N,5} = \frac{2}{3}((1-2\lambda)\log(3) + \lambda\log(2)).$$

Finally,

$$\overline{h} = 1.$$

Then adding all corresponding cases together we get, for $\lambda \geq \frac{1}{2}$,

$${}^1D_N = \exp\left(\frac{1}{3}((2\lambda+1)\log(3) - 2(2\lambda-1)\log(2))\right),$$

and for $\lambda < \frac{1}{2}$,

$${}^1D_N = \exp\left(\frac{1}{3}(3-2\lambda)\log(3)\right).$$

Figure 2.1b shows how every index varies as λ varies. The cases for $\lambda = 0$ and $\lambda = 1$ are both star trees with three and two branches respectively. For $\lambda = 0$, ${}^qD_N = {}^qD_L = {}^qD_S = 3$ for $q = 0$ and 1 and ${}^1J_N = {}^1J_L = {}^1J_S = 1$, and for $\lambda = 1$, ${}^0D_N = {}^0D_L = {}^0D_S = 2$, ${}^1D_N = {}^1D_L = {}^1D_S = \exp(\log(3) - \frac{2}{3}\log(2)) \approx 1.890$ and ${}^1J_N = {}^1J_L = {}^1J_S = \log_2(3) - \frac{2}{3} \approx 0.918$. For all λ, ${}^1J_N = {}^1 J_L$.

2.7 Calculating the indices

To compute the indices, I have written an R package called RUIindices, which is available publicly on GitHub [85]. Most of the code was written by me, but some was generated using ChatGPT 3.5. The package contains 16 functions, and two of them (`abundance_phylo` and `compute_T_i_S_i`), in their current form, were mostly generated by ChatGPT. For the first function, I gave a general prompt, asking it to generate a function in R that could do what I wanted it to. The second was generated by providing pseudo code and asking it to change this into functioning R code. The functions `calculate_S_i_a` and `calculate_S_i_a_star` perform very similar tasks to `compute_T_i_S_i` and so I used the general form of this function for those functions too. The rest of the functions were written by me, except the `descendant_edges`

function. It was modified from the `TreeTools` R package written by Martin Smith, which is licensed under the GNU General Public License version 3 (GPL-3).

The package takes in a single tree, which can be in one of three common formats, Newick, NEXUS or a `phylo` object. It can output all indices, indices of a specific mean type or just a single index. Additionally, both the normalised and non-normalised versions of the diversity indices can be calculated.

The node-wise indices are calculated as follows:

1. Calculate branch sizes for the whole tree by traversing the tree
2. Calculate normalisation factor $\overline{h}$
3. Then for every node:
    a. Calculate all ancestors
    b. Calculate $S_{C_k}(x)$ and $\overline{h}_{C_k}(x)$ at same time
    c. Then for every ancestor j,
        i. If $d_k - d_j < h_{C_k}$, break, else, proceed
        ii. Calculate ancestor sum A
        iii. Calculate $S_{T_j}(x)$ and store $s_b(x)$
        iv. Calculate index value of interest
        v. Calculate integral I as piecewise sum
4. Return $\frac{1}{\overline{h}} \sum \frac{1}{\overline{h}_{C_k}} A * I$, either exponentiated or not

The longitudinal and star mean indices are calculated differently from the node-wise indices, and they are simpler. They are calculated as follows:

1. Calculate branch sizes for the whole tree by traversing it
2. For root node, k, calculate $S_{T_k}(x)$ and store $s_b(x)$. If the star mean is being calculated, treat every branch as originating from the root.
3. Calculate the sum I by traversing the intervals of the tree
4. Calculate the normalisation term $\overline{h}$
5. Return $\frac{1}{\overline{h}} \sum I$, either exponentiated or not

2.8 Comparison of HIV and languages tree

We next investigate the new indices by applying them to some real data. We use two trees, one representing the within-host evolution of HIV and the other the evolutionary history of

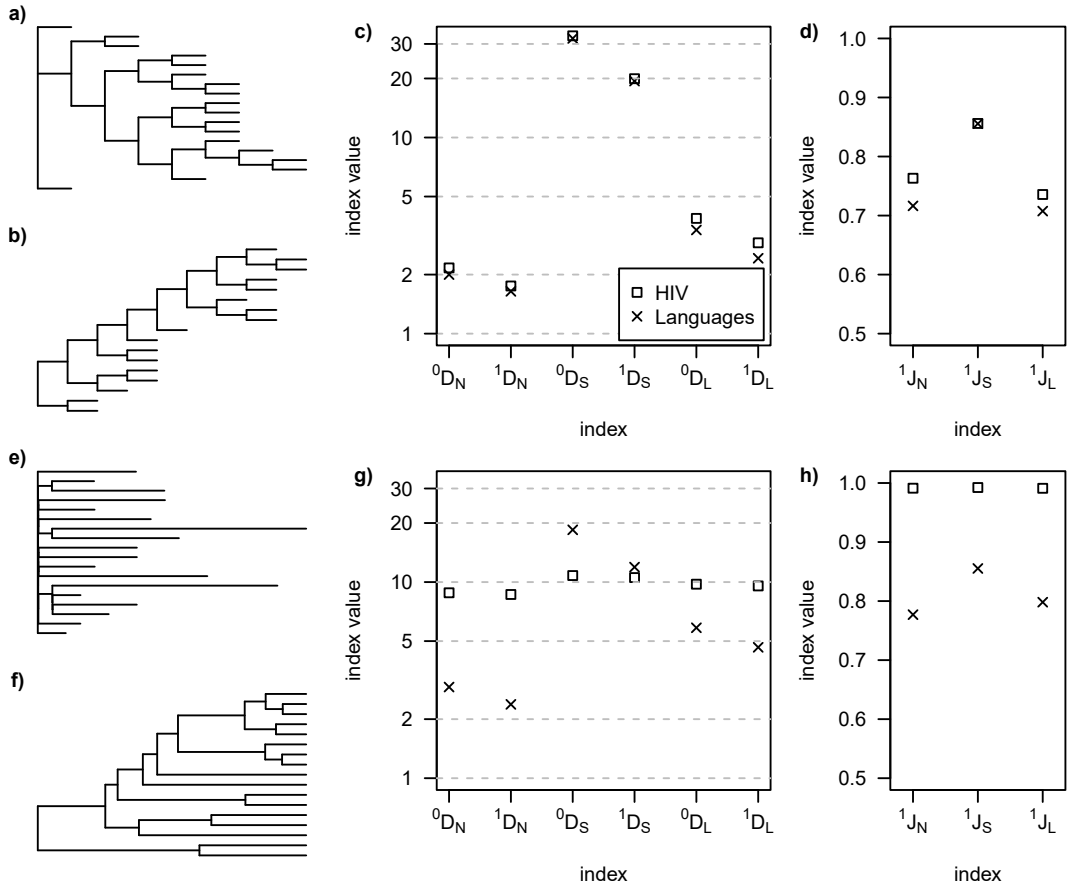

Figure 2.2: a-b) Trees with equal branch lengths representing a) the within-host evolution of HIV and b) the evolutionary history of the Uralic language. c) Diversity index values for the trees with equal branch lengths. d) Evenness index values for the trees with equal branch lengths. e-f) The same trees as in a-b but with their original inferred branch lengths. g) Diversity index values accounting for branch lengths. h) Evenness index values accounting for branch lengths. For all trees, leaves were assigned to be equally abundant and internal nodes were assigned to have size zero. The HIV tree is from the GitHub repository associated with [53] (file PIC38051.tre) and the languages tree from the D-PLACE database [86] (folder honkola_et_al2013).

the Uralic language [53, 87]. The trees are of similar size, both have 16 internal nodes, and the languages tree has 17 leaves, whereas the HIV tree has 18. The difference in the number of leaves is because the HIV tree is completely bifurcating, each internal node has two daughters, and the HIV is bifurcating except for the root node, which has three children, one of which is a leaf. For simplicity, the leaves are assumed to have equal size and the internal nodes are assumed to have size zero.

The HIV and languages trees are shown in Figures 2.2a and b with equal branch lengths, and in Figures 2.2e and f with their inferred branch lengths respectively. Visually, we can see that the trees appear to look similar when they have equal branch lengths and much more distinct when they do not. The drastic change in similarity is due to the HIV tree, which is approximately a non-ultrametric star tree when it has its inferred branch lengths. It has uneven branch lengths and setting the branch lengths to be equal changes its structure substantially. We might expect

then that the trees will have quite different indices with their inferred branch lengths and more similar values when they have equal branch lengths.

Consider the trees with equal branch lengths. As the trees look similar, it is difficult to visually gauge how the two trees will compare. The index values are shown in Figures 2.2c and d. Both trees have ${}^0D_N \approx 2$, it is not exact due to the HIV tree containing one non-bifurcating node. The trees have similar branch counts, where for total branch count ${}^0D_S = 32$ and 33, and at each depth across the tree $3 < {}^0D_L < 4$. The values for ${}^1D_N$, ${}^1D_S$ and ${}^1D_L$ are again similar for both trees, but slightly smaller than their $q = 0$ counterparts due to imbalances within the trees. The imbalances are quantified by the balance indices where for both trees, ${}^1J_S \approx 0.86$, $0.72 < {}^1J_N < 0.76$ and $0.71 < {}^1J_L < 0.74$.

Now consider the trees with their inferred branch lengths. For the HIV tree, the average effective out-degree (for $q = 0$) is now much greater than two, ${}^0D_N \approx 9$. This is due to multiple long branches close to the root node. The effective number of total branches is a third of what it was when inferred branch lengths were ignored, ${}^0D_S \approx 11$, and the branch count at each depth across the tree over doubles, ${}^0D_L \approx 10$. Additionally, as the HIV tree is approximately a star tree with equal-sized nodes, its diversity indices are approximately equal, and its balance indices are close to one, ${}^1J_N \approx {}^1J_S \approx {}^1J_L \approx 0.99$. For the languages tree, accounting for the inferred branch lengths only has a small effect on most of the index values. The indices ${}^0D_N$, ${}^1D_N$ ${}^0D_L$ and ${}^1D_L$ all increase slightly; this is due to the tree now having longer branches from nodes that overlap their children. The indices ${}^0D_S$ and ${}^1D_S$ decrease slightly as these indices now 'see' a non-ultrametric star tree. Despite these changes, the diversity indices remain largely different.

We have demonstrated using our indices that when branch length information is ignored, these two trees appear very similar, however when we account for branch lengths, that is not the case. The HIV tree has a larger effective out-degree, number of total branches and branch count across the tree, and is much more balanced than the languages tree. In conclusion, we have shown that the distinct differences between these two trees are only captured when indices that account for their branch lengths are used. If indices that do not account for branch lengths were used, it would lead to the assumption that these trees were very similar in terms of shape and hence were likely generated from similar processes.

2.9 Another example

Here we reanalyse results from a study of tumour evolution [88]. Using computational modelling, the paper aimed to deduce differences between the shapes of evolutionary trees which corresponded to different modes of tumour expansion, either boundary-driven growth (BDG) or unrestricted growth. It was found that on average, the BDG model generated ultrametric time trees with a higher variance in terminal branch lengths, and non-ultrametric gene trees with a higher variance in their leaf depths. To investigate how our indices vary with different modes of tumour growth, we consider the two simulated tumours from Figure 1 in [88].

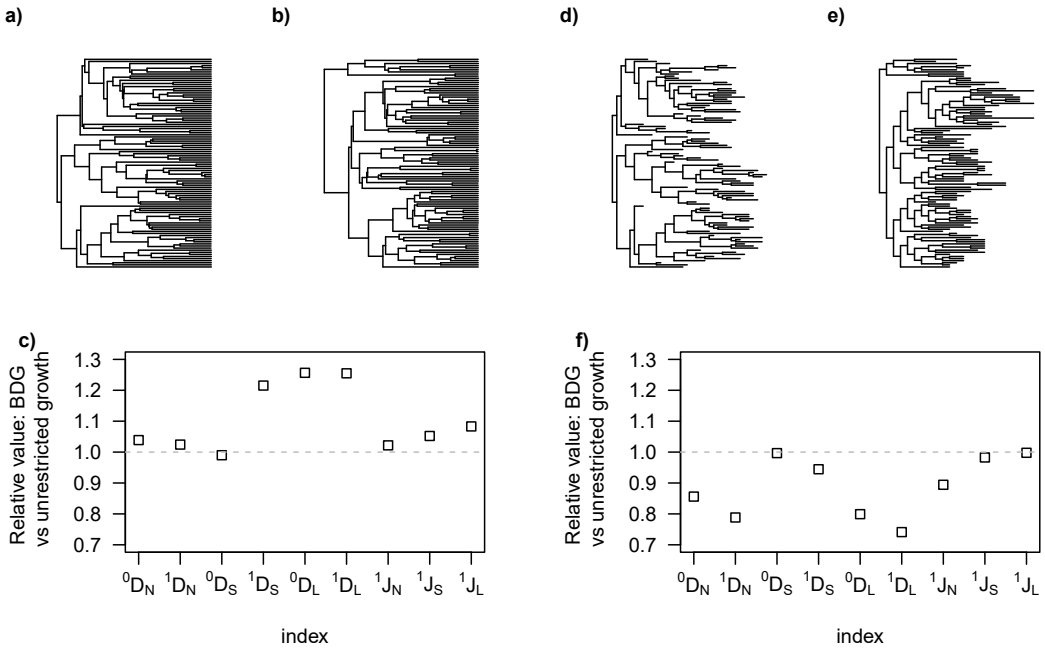

Figure 2.3: a-b) Time trees generated by computational models of tumour evolution with a) boundary-driven growth or b) unrestricted growth. The leaves represent extant cells and the branch lengths are proportional to the time elapsed between cell division events. c) The ratio of tree shape indices for the time trees. d-c) Gene trees generated by the same simulations as the time trees. The leaves represent extant cells and the branch lengths are proportional to genetic distances. All tree data was obtained from the GitHub repository associated with [88].

The time trees, shown in Figures 2.3a and b, both have the same number of leaves, similar effective out-degrees, very similar effective number of non-root nodes, and similar balance ($^0D_N \approx 4.5$ and 4.3, $^1D_N \approx 3.7$ and 3.6, $^0D_S \approx 117$ and 118, and $^1J_N \approx 0.84$ and 0.82 for BDG and unrestricted growth respectively). Despite these similarities, the BDG time tree has a higher effective branch count, branch count across the tree, and leaf diversity ($^1D_S \approx 65$ and 53, $^0D_L \approx 28$ and 22, and $^1D_L \approx 21$ and 17 for BDG and unrestricted growth respectively).

The gene trees, shown in Figures 2.3d and e, again have the same number of leaves and virtually the same effective number of non-root nodes ($^0D_S \approx 136$ for both trees). However, the BDG growth gene tree has a much lower effective out-degree, fewer branches across the tree, lower leaf diversity and is less balanced ($^0D_N \approx 2.6$ and 3.0, $^1D_N \approx 2.1$ and 2.6, $^0D_L \approx 17$ and 21, $^1D_L \approx 11$ and 15, and $^1J_N \approx 0.76$ and 0.84 for BDG and unrestricted growth respectively).

For BDG, the lineages trapped in the centre divide slowly and as a result exhibit longer times since diverging from another sampled cell [88]. These trapped lineages lead to long terminal branches in the time tree, resulting in a more star-like tree. This is why the BDG time tree has a higher, effective total branch count, branch count across the tree and leaf diversity, than the unrestricted growth time tree. Fewer divisions in the centre trapped lineages also lead to low mutation accumulation which creates short branches in a genetic tree, and as a result of the differences in birth rate, ladder-like patterns in the genetic trees [88]. This results in the

Index	HIV	Lanuages	BDG (m)	Unrestricted (m)	BDG (t)	Unrestricted (t)
${}^0D_N$	12.0	-0.2	-2.0	-0.9	2.5	2.1
${}^1D_N$	14.0	-0.6	-2.2	-0.7	2.5	2.2
${}^1J_N$	1.8	-1.0	-1.2	0.4	0.3	0.0
${}^0D_S$	-1.4	1.5	4.1	4.1	2.8	2.9
${}^1D_S$	0.1	0.8	4.2	4.6	3.8	2.4
${}^1J_S$	3.0	-0.2	1.6	2.0	2.1	1.0
${}^0D_L$	3.5	0.3	1.0	2.3	4.4	2.7
${}^1D_L$	4.4	0.0	0.4	1.9	4.2	2.5
${}^1J_L$	1.5	-0.9	-0.1	-0.1	0.7	-0.5

Figure 2.4: Comparison of tree indices to those of the Yule process. For each tree, to compare to the Yule process 1000 trees with the same number of leaves were generated and the average of their indices was taken. The values in the table shows how many standard deviations away the tree indices are from these values. Red shading indicates the index is smaller than the Yule case and blue indicates greater than. The shading goes from white for no deviation to the darkest for the most extreme value in the table in each case. For BDG and unrestricted growth, the 'm' denotes the molecular trees and the 't' the time trees.

BDG gene tree being much less star-like. Hence, the BDG gene tree becomes more unbalanced and has a lower effective out-degree, branch count across the tree and leaf diversity.

2.10 Comparison to the Yule process

I further analysed these trees to test the power of the new indices. I wanted to investigate whether the collection of these index values is specific to the tree generation process that created them, which is ultimately the question: can we use the index values to predict the process that generated them? I address this question in more detail in subsequent chapters; the work here was completed very early in this research and served as an exploration and proof of concept.

Here I look at the six trees that were considered in the previous part of this section: the HIV tree, the languages tree, the BDG molecular tree, the unrestricted growth molecular tree, the BDG time tree and the unrestricted growth time tree. To compare to the Yule process, for each example, I generated 1000 Yule trees with the same number of leaves as the example tree and averaged the index values. I then looked at how many standard deviations away the tree indices were from the Yule averages. Table 2.4 shows the results.

The Languages tree is the most Yule-like, with 8 out of 9 indices being within one standard deviation and 1 being within two standard deviations. This is not surprising as this tree was inferred using a Yule prior [87]. The HIV tree appears to be the least Yule-like, with the most indices (5) being more than three standard deviations away. Given the star-like nature of this tree, this is again not surprising. For the BDG and unrestricted growth trees, it is

much more unclear. Both molecular trees have 2 indices more than three standard deviations away, but the unrestricted tree has 4 indices within one deviation, whereas the BDG tree has only 2, suggesting that the unrestricted growth molecular tree is more Yule-like than the BDG molecular tree. The unrestricted time tree has 3 indices within one standard deviation, and the rest are within three. The BDG time tree has 2 indices within one standard deviation, 4 within three and 3 more than three standard deviations away. This again suggests that the unrestricted case is more Yule-like than the BDG case. Hence, the molecular and time trees are in agreement. All of the BDG and unrestricted growth trees were simulated from processes with selection; the simulated cells could gain mutations at the point of division [88]. As a result, we do not expect them to be entirely Yule-like, but we do expect the unrestricted growth to be more Yule-like than BDG, given the strong spatial constraints on the latter. Therefore, using rudimentary analysis of the new system of indices, we can distinguish between trees generated by different processes.

Chapter 3

Non-spatial tree generating models

3.1 Introduction

Non-spatial tree-generating models are the simplest tree-generating models that can be studied. Due to their simplicity, they are often well researched and understood, making them good candidates for null models when studying evolutionary processes [44]. Historically, the most popular null model is the Yule process (also called equal-rates-Markov or simply Markov); other null models include the Uniform model (also called proportional-to-distinguishable-arrangements) and the equiprobable-types model.

As the Yule process has a constant rate, it can be used to test for deviations in diversification rates. In fact, it has been well studied as a null model for evolutionary diversification. It has been repeatedly observed that empirical trees are more unbalanced than predicted by the Yule process [80, 89–91]. One explanation for this is due to the biological differences in organisms that lead to differences in the relative rates of speciation to extinction, making the Yule process a good null model to test for such differences [92]. Another explanation for the observed imbalance is errors in the tree inference process. When this is the case, it is thought that the trees will more likely represent the uniform model [24, 44]. The trees produced by the equiprobable-types model are not linked to any model of evolutionary processes or systematists' behaviour [44]. As such, the Yule process and the Uniform model are natural first models to investigate the new indices under, and the equiprobable types model is a simple extension that will further the understanding of how the indices behave.

The Yule process has various interpretations depending on the application. These include: a pure-birth process, in which individuals in a population may give birth but not die, a branching process where branches bifurcate and do not die, and cladogenesis, where species in a population may split into two species but never go extinct. In cancer research, 'species' would be replaced with 'clones' in the latter interpretation. To outline the Yule process mathematically, I will use the pure-birth process interpretation, but all are equivalent; individuals, branches and species, and 'giving birth', 'splitting into two branches', and 'splitting into two species' are analogous

respectively.

In the Yule process, each individual acts independently and gives birth at a constant rate λ, and hence times between births are exponentially distributed. The population as a whole has a birth rate $\lambda_n = n\lambda$ for $n \geq 0$, because if there are n individuals each giving birth at a rate λ, then the total rate at which births will occur is $n\lambda$ [93].

As the time for each individual to give birth is given by $\text{Exp}(\lambda)$, the inter-birth time, the time that it takes for the system to go from n individuals to $n+1$ individuals, will be given by the smallest birth time of the current individuals in the population. That is,

$$T_{n+1} - T_n = \min\{Z_i : Z_1, \ldots, Z_n \overset{i.i.d}{\sim} \exp(\lambda)\}. \tag{3.1}$$

In fact,

$$T_{n+1} - T_n \sim \text{Exp}(n\lambda). \tag{3.2}$$

These two statements are equivalent as,

$$\begin{aligned} P(\min\{Z_i : Z_1, \ldots, Z_n\} \leq z) &= 1 - P(\min\{Z_i : Z_1, \ldots, Z_n\} > z) \\ &= 1 - P(Z_1 > z, \ldots, Z_n > z) \\ &= 1 - (P(Z_i > z))^n \\ &= 1 - \mathrm{e}^{-n\lambda z}. \end{aligned}$$

The inter-birth times can then be used to calculate branch lengths for the tree corresponding to the birth process.

The uniform model can be thought of as either a tree-generating process that does not use rates or a distribution-based model [94]. It is usually presented as the latter, and in this case, the uniform model states that all labelled trees with the same number of leaves have the same probability [44]. For trees with four leaves, ignoring rotations about a node, there are two topologies: a caterpillar tree and a balanced tree. Considering the possible distinct arrangement of labels on the leaves, there are 12 different possibilities for the caterpillar tree and 3 for the balanced tree. Therefore, the caterpillar tree has a probability of 0.8, and the balanced tree has a probability of 0.2. As a process, it is initialised with an unrooted tree with 3 leaves, then at every step, an edge is chosen, and a branch terminating in a leaf is added to it. This is repeated until n leaves are reached, and lastly, an edge is chosen to add a root node to [94]. The equiprobable-types model is distribution-based [94]. It states that all tree topologies are equally likely [44]. For example, there are only two distinct topologies for four-leaf trees, and hence each has a probability of 0.5 under the equiprobable-types model. These models do not assign any branch lengths.

Previous work by Lemant et al. defined a balance index called J^1 [46]. This balance index is equivalent to the new index ${}^1J_N$ when branch lengths are equal, except for the value they assign to linear nodes (J^1 assigns them 0 and ${}^1J_N$ assigns them 1). In their work, they derived an approximation to the expected values of J^1 for random trees generated by the Yule process and uniform model. As they only generate trees with bifurcating nodes, this approximation

holds for ${}^1J_N$ when under equal branch lengths.

For a tree T with out-degree m and leaves of equal size,

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

and,

$$\mathbb{E}[J^1] = \mathbb{E}\left[\frac{n \log_m n}{I_S}\right] = \frac{n \log_m n}{\mathbb{H}[I_S]},$$

where $\mathbb{H}[I_S] = 1/\mathbb{E}[1/I_S]$ is the harmonic mean of the Sackin index [95]. As there is no closed-form expression of the harmonic mean of the Sackin index under the standard null models, the expectation can be approximated using $\mathbb{H}[I_S] \approx \mathbb{E}[\mathbb{I}_\mathbb{S}]$ [95]. The expectation of Sackin's index for the Yule process is [46],

$$\mathbb{E}_{Yule}(I_S) = 2n \sum_{i=2}^{n} \frac{1}{i} = 2n \ln n + (2\gamma - 2)n + o(n), \tag{3.3}$$

where γ is Euler's constant and n is the number of leaves. The expectation of Sackin's index for the uniform model is [46],

$$\mathbb{E}_{Unif}(I_S) = n\left(\frac{(2n-2)!!}{(2n-3)!!} - 1\right). \tag{3.4}$$

Jensen's inequality implies that,

$$\mathbb{E}\left(\frac{1}{I_S}\right) > \frac{1}{\mathbb{E}(I_S)},$$

therefore the approximation must be smaller than the true expected value. In fact, the difference between the true expectation and the approximation is less than 0.008 for the Yule process and less than 0.057 for the uniform model [95].

3.2 Methods

3.2.1 Tree generation

I simulated trees from the uniform model using the function `gemTrees()` from the package 'poweRbal' using the 'pda' mode. I simulated trees from the equiprobable-types model using the `rmtree()` function from the R package 'ape', setting 'equiprob' to be true. Neither of these models assigns branch lengths to trees, so I set all branch lengths to be equal. In terms of the indices, equal branch lengths mean that there are no ancestor contributions, as every node is equidistant from its children to its parent. From the R package 'TreeSim', I used the function `sim.bd.taxa()` to simulate trees from the Yule process. The function allows the user to specify the number of leaves desired for the simulated tree, the birth rate and the death rate, which is zero for the Yule process. For all three models, I assumed that the simulated trees had leaves of equal size and internal nodes of size zero.

3.2.2 Indices and correlations

I considered 9 indices, ${}^0D_L$, ${}^1D_L$, ${}^0D_N$, ${}^1D_N$, ${}^0D_S$, ${}^1D_S$, ${}^1J_L$, ${}^1J_N$ and ${}^1J_S$, or a subset of these throughout this section. For the expected values, I considered all 9. For the index correlations, I considered only $q = 1$, giving 6 indices.

I used three methods to analyse the trajectories, where a trajectory here is the path in index space formed as a tree grows. I calculated the Pearson correlation coefficient to assess the linear relationship between pairs of indices. I used hierarchical clustering. The hierarchical clustering was performed on a distance matrix given by,

$$d(i, j) = 1 - |r(i, j)|,$$

where $r(i, j)$ is the Pearson correlation coefficient. Therefore, perfectly correlated indices have a distance of 0 and uncorrelated indices have a distance of 1. Hierarchical clustering sequentially merges the closest index pair, with each step creating a branch in the dendrogram. The dendrogram is then cut at some chosen height to define clusters of redundant indices. A height of 0.1 was used here. Finally, I used principal component analysis (PCA). PCA enabled us to determine the number of principal components required to explain the variance across all our indices, indicating which indices were distinct and which were redundant.

3.3 Results

3.3.1 Uniform model

Index expected values

The expectation of ${}^1J_N$ can be calculated to a good approximation for the Uniform process with equal branch lengths by,

$$\mathbb{E}_U[{}^1J_N] \approx \frac{n \log_m n}{\mathbb{E}_U[I_S]},$$

where $\mathbb{E}_U[I_S]$ is the expectation of Sackin's index under the uniform model given in equation 3.4.

Calculating the expected values of the indices for trees generated by the uniform model is simple. Each labelled tree topology is equally likely, so to determine the expected value, one must calculate the index for each possible topology for a given number of leaves and take the mean. Here, I extend previous work for trees with 2 and 3 leaves. As the branch lengths were set to be equal in this case, it is possible to do this for any number of leaves; however, as the number of leaves increases, the number of topologies increases, making it increasingly taxing but not necessarily more complicated.

The case of $n = 2$ is trivial as it is a perfectly balanced star tree. All evenness indices are

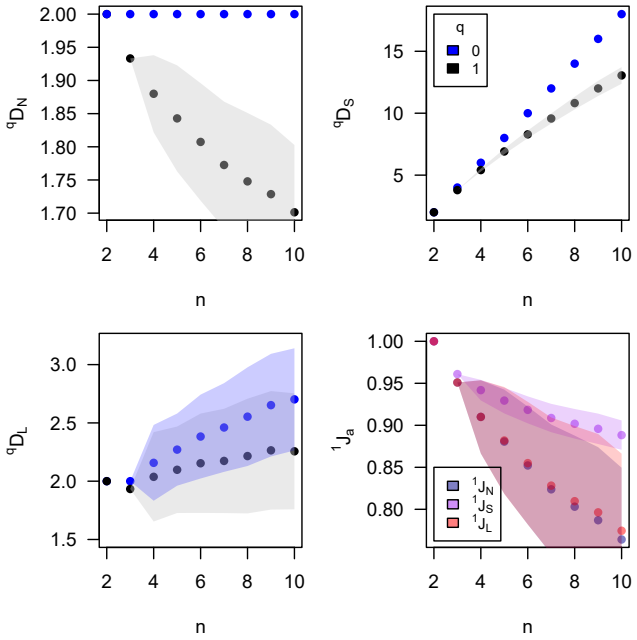

Figure 3.1: Index values for trees generated following the uniform model for different numbers of leaves, n. Each data point is the average of 1000 trees, and shaded regions show plus and minus one standard deviation. The internal nodes were assumed to have size zero and the leaves were assumed to be equally abundant.

exactly 1, and the diversity indices are exactly 2. The case of $n = 3$ is simple with only one distinct topology in terms of the indices, this is the tree shown in Figure 2.1a but with equal branch lengths ($\lambda = 1/2$ and branch 4-1 with length 1/2). As there is only one possible tree topology, the index values for this tree are the expectations of the index. Calculating these index values analytically gives, $\mathbb{E}({}^0D_N) = \mathbb{E}({}^0D_L) = 2$, $\mathbb{E}({}^1D_N) = \mathbb{E}({}^1D_L) = 1.933$, $\mathbb{E}({}^0D_S) = 4$, $\mathbb{E}({}^1D_S) = 3.789$, $\mathbb{E}({}^1J_N) = \mathbb{E}({}^1J_L) = 0.951$ and $\mathbb{E}({}^1J_S) = 0.961$.

Figure 3.1 shows how the indices vary as the number of leaves varies, where each data point is the average of 1000 trees. This figure demonstrates that the case of $q = 0$ is always greater than or equal to the case of $q = 1$. This is due to $q = 0$ only taking into account branch lengths, but $q = 1$ taking into account both branch lengths and branch sizes, where the size of a branch is the sum of the (proportional) node sizes for all nodes descendant from the branch (see proposition 1 in [59]). Every diversity index starts at 2, and the evenness indices at 1. This follows from property 0.8 in [59] as $n = 2$ is a star tree.

As ${}^0D_N$ is a measure of effective out-degree that does not account for branch sizes, it remains constant at 2 as the tree is bifurcating and has equal branch lengths. ${}^1D_L$ decreases when going from 2 leaves to 3, and ${}^0D_L$ remains constant at 2. This is because these cases both contain only one topology, and the 3-leaf topology has two regions, which are both bifurcating nodes, but one has equal branch sizes, and one has unequal branch sizes. Hence, when ignoring branch sizes, the weighted mean remains unchanged at 2, but when considering them, the region with unequal branch sizes reduces the weighted mean. After 3 leaves, the longitudinal indices monotonically increase. ${}^1D_N$ decreases monotonically from 2 as the trees are bifurcating, and

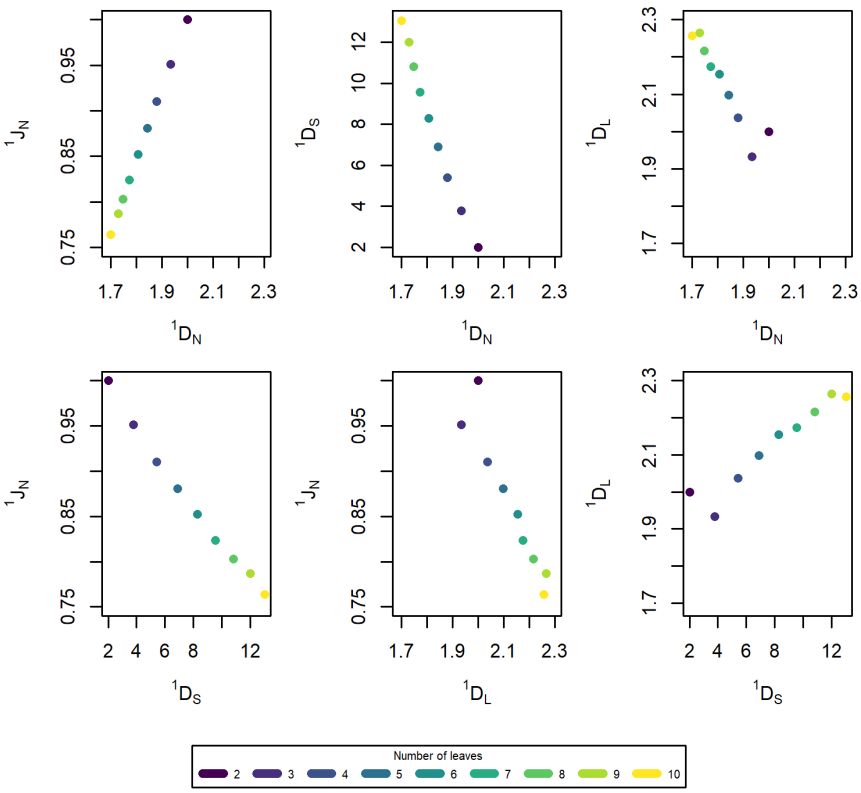

Figure 3.2: Index trajectories formed by four indices, ${}^1J_N$, ${}^1D_N$, ${}^1D_S$, ${}^1D_L$, for trees generated following the uniform model for varying numbers of leaves, n. Each index value is the average of 1000 trees.

as the trees grow, there is an increasing number of unbalanced nodes. It is for the same reason that all the evenness indices also decrease monotonically. The star mean indices increase monotonically, with ${}^0D_S$ being exactly equal to the number of branches in the tree and ${}^1D_S$ being less than this due to imbalances in the tree.

Index correlations

Every index pair is highly correlated with absolute correlation coefficients greater than or equal to 0.94 (see Figure 3.2). Pairs containing ${}^1D_L$ have a slightly lower correlation than other pairs (absolute range 0.94-0.96 versus 0.99-1.00). Hierarchical clustering assigns all the indices to the same cluster (see Figure 3.3a). From principal component analysis (PCA), I find that the first component describes 98% of the variance, with each index having an absolute loading of at least 0.40. Hence, all the indices are highly correlated, and when analysing average uniform model trajectories, only one index is needed to capture the changes in shape as the tree grows.

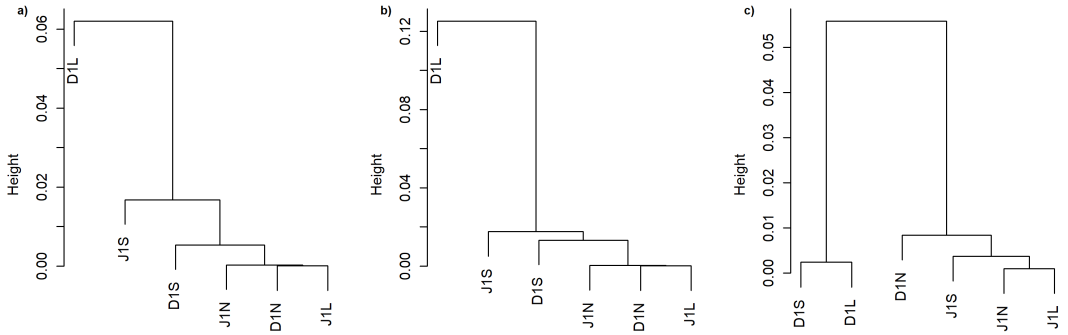

Figure 3.3: Hierarchical clustering dendrogram for our indices for average a) uniform, b) equiprobable-types and c) Yule trajectories.

3.3.2 Equiprobable-types model

Index expected values

For the equiprobable-types model, I do not have a general formula or approximation for the expectation of any of our indices. However, determining the expected values of the indices for trees generated by the equiprobable-type model is simple, and the method is the same as for the uniform model. Each unlabelled tree topology is equally likely; hence, to calculate the expected value, one computes the index for each possible topology for a given number of leaves and takes the mean. Here, I extend previous work for trees with 2 and 3 leaves. Again, the branch lengths are equal, so the expectation can be calculated for any number of leaves, but as the number of leaves increases, it becomes more demanding but not more complicated.

There is only one unlabelled tree topology for both 2 and 3 leaves, making the equiprobable-types and uniform models equal in these cases. Hence, as before, $n = 2$ is a perfectly balanced star tree, and all evenness indices are exactly 1, and the diversity indices are exactly 2. The case of $n = 3$ is the tree shown in Figure 2.1a with equal branch lengths and has index values: $\mathbb{E}({}^0D_N) = \mathbb{E}({}^0D_L) = 2.000$, $\mathbb{E}({}^1D_N) = \mathbb{E}({}^1D_L) = 1.933$, $\mathbb{E}({}^0D_S) = 4$, $\mathbb{E}({}^1D_S) = 3.789$, $\mathbb{E}({}^1J_N) = \mathbb{E}({}^1J_L) = 0.951$ and $\mathbb{E}({}^1J_S) = 0.961$.

Figure 3.4 shows how the indices vary as the number of leaves varies, where each data point is the average of 1000 trees. The indices ${}^1D_N$, ${}^0D_L$, ${}^1D_L$, ${}^1J_N$ and ${}^1J_L$ all exhibit the same pattern. When the number of leaves changes from odd to even, there is a greater change in the index value compared to when the number of leaves changes from even to odd. That is, if the index is generally decreasing, when the number of leaves goes from odd to even, the decrease is greater than when the number of leaves goes from even to odd, with the effect decreasing as the number of leaves increases. The pattern is due to the structure of the trees and the model that generates them. Every tree topology is equally likely, and as the branch lengths are equal, the only weighting for each index is the absolute abundance in the region being averaged over. To have a perfectly balanced tree in this case, there must be an even number of leaves, and going from even to odd creates more unbalanced topologies and going from odd to even adds

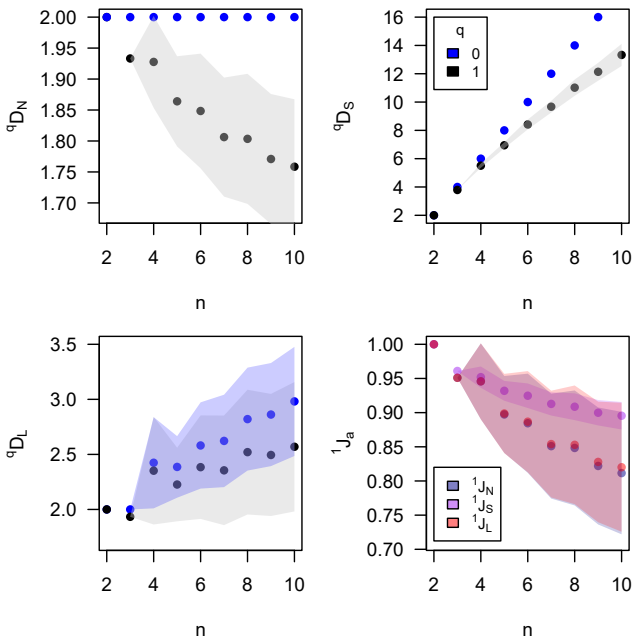

Figure 3.4: Index values for trees generated following the equiprobable-types model for different numbers of leaves, n. Each data point is the average of 1000 trees. The internal nodes were assumed to have size zero and the leaves were assumed to be equally abundant.

more balanced topologies. As the number of leaves increases, the number of tree topologies increases, and as the values here are averages over the topologies, the effect decreases. Overall, the indices exhibit the same pattern as they did with the uniform model. Specifically, ${}^0D_N$ remains constant at 2, ${}^1D_N$ decreases from 2, the longitudinal indices increase from 2, and the evenness indices decrease. The star mean indices increase monotonically, with ${}^0D_S$ the same as in the uniform case, and ${}^1D_S$ is also less than this due to imbalances in the tree, but not the same as the uniform case.

Index correlations

Every index pair has a high correlation, with absolute correlation coefficients greater than or equal to 0.860 (see Figure 3.5). Pairs containing ${}^1D_L$ have a slightly lower correlation than other pairs (absolute range 0.860-0.917 versus 0.988-0.999). Hierarchical clustering assigns ${}^1D_L$ to its own cluster and the rest of our indices to the same cluster (see Figure 3.3b). From principal component analysis (PCA), I find that the first component accounts for 96% of the variance with each index having an absolute loading of at least 0.38. The second principal component only accounts for 3% of the variance, but ${}^1D_L$ has an absolute loading of 0.89 with the next largest loading of 0.25. Hence, to capture changes in tree shape when analysing average equiprobable-type model trajectories, two indices are needed, ${}^1D_L$ and any one of the others.

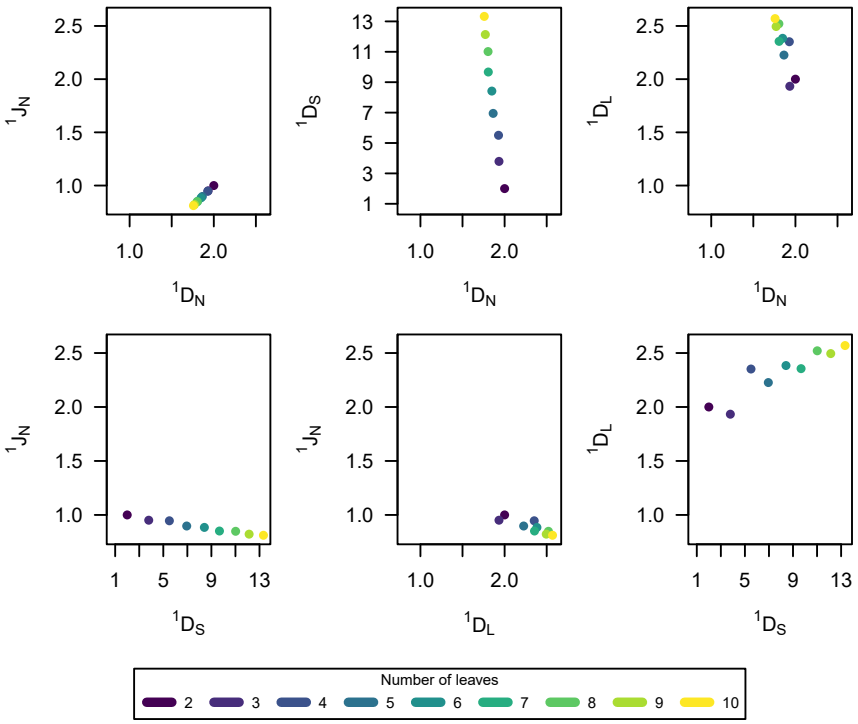

Figure 3.5: Index trajectories formed by four indices, ${}^1J_N$, ${}^1D_N$, ${}^1D_S$, ${}^1D_L$, for trees generated following the uniform model for varying numbers of leaves, n. Each index value is the average of 1000 trees.

3.3.3 Yule process

Index expected Values

When branch lengths are equal, the expectation of ${}^1J_N$ under the Yule process is known to good approximation (see Section 3.1), it is not known for exponentially distributed branch lengths. Here, I will extend on the previous work, but only in the case of $n = 3$ and the trivial case of $n = 2$, providing exact numerical values and no general formula for the expectation of the indices for the Yule process.

Consider the results outlined for ${}^1D_N$ for the simple case where $n = 3$ in Section 2.6.2. For this example, it was assumed the height of the tree was 1. Therefore, let,

$$\lambda = \frac{h_1}{h_1 + h_2}$$

where h_1 is the length of the branch connecting node 4 to node 5, and h_2 is the length of the branches connecting node 5 to nodes 2 and 3. Substituting this into the previous result, and

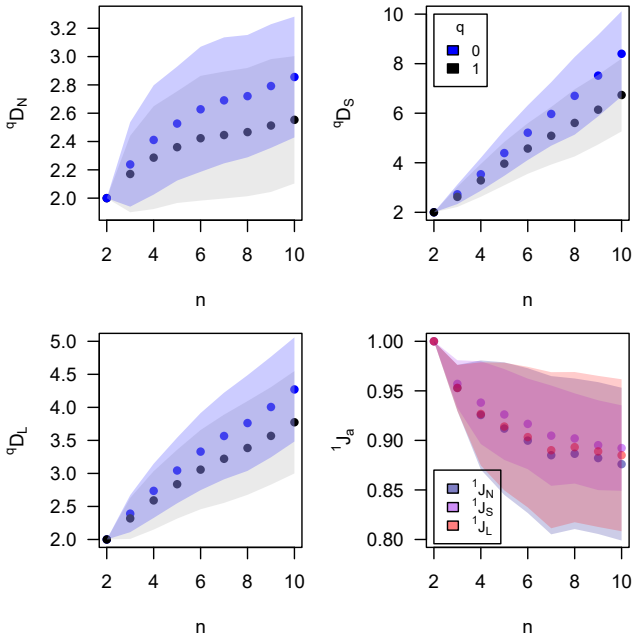

Figure 3.6: Index values for trees generated by the Yule process for different numbers of leaves, n. Every data point is the average of 1000 trees. The internal nodes were assumed to have size zero and the leaves were assumed to be equally abundant.

after simplification,

$$ {}^1D_N = \begin{cases} \exp\left(\frac{2}{3}\frac{h_1}{h_1+h_2}\left(\log(3)-2\log(2)\right)+\frac{1}{3}(\log(3)+2\log(2))\right), & \text{for} \quad h_1 \geq h_2, \\ \exp\left(\frac{1}{3}\left(3-2(\frac{h_1}{h_1+h_2})\right)\log(3)\right), & \text{for} \quad h_2 > h_1 \end{cases} $$

and the expectation of ${}^1D_N$ is then given by,

$$ \begin{aligned} \mathbb{E}({}^1D_N) = &\mathbb{E}\left(\exp\left(\frac{2}{3}\frac{h_1}{h_1+h_2}(\log(3)-2\log(2))+\frac{1}{3}(\log(3)+2\log(2))\right)\right) \\ &+\mathbb{E}\left(\exp\left(\frac{1}{3}\left(3-2\left(\frac{h_1}{h_1+h_2}\right)\right)\log(3)\right)\right) \end{aligned} $$

Using the law of the unconscious statistician, which states, if there is a random variable X which is continuously distributed with probability density function f_X, then the expected value of $g(X)$ is [96],

$$ \mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X \, dx. $$

The random variable X is $\lambda = h_1/(h_1+h_2)$ here. Along with the fact that when given two random variables X and Y, where $X \sim \text{Exp}(\alpha)$ and $Y \sim \text{Exp}(\beta)$, and letting $Z = \frac{X}{X+Y}$, the probability distribution function of Z is given by,

$$ f_Z(z;\alpha,\beta) = \frac{\alpha\beta}{(\alpha z+\beta(1-z))^2} $$

(see Chapter 3 supplementary material for proof of this). The expectation can then be calcu-

lated. As this is a tree generated by the Yule process, the distributions of the inter-birth times are known (given by equation 3.2). Hence, the branch length distributions are given by,

$$h_1 \sim \text{Exp}(2a) \quad \text{and} \quad h_2 \sim \text{Exp}(3a),$$

where a represents the birth rate and is used here so it is not confused with the random variable λ. Then, integrating over the correct regions of z, it is obtained that $\mathbb{E}({}^1D_N) = 2.180$ (see Chapter 3 supplementary material for a detailed outline). The same method can also be used to give the expectations of the other indices, giving $\mathbb{E}({}^1D_L) = 2.329$, $\mathbb{E}({}^1D_S) = 2.632$ and $\mathbb{E}({}^1J_N) = 0.954$.

To obtain the expectations computationally, I simulate 1000 Yule trees with three leaves and average each index. Doing this gives ${}^1D_N = 2.170 \pm 0.270$,${}^1D_L = 2.319 \pm 0.311$, ${}^1D_S = 2.622 \pm 0.407$ and ${}^1J_N = 0.953 \pm 0.023$, the error stated is the standard deviation. These are in agreement with the analytical results.

The case of $n = 2$ is trivial, as the value of each index is independent of the branch length and there is only a single tree topology in terms of our index. ${}^1D_N$, ${}^0D_N$ and ${}^1D_L$ have a value of 2 and ${}^1J_N$ has a value of 1. The case of $n = 4$ quickly becomes very complicated with non-linear dependence on branch lengths.

Figure 3.6 shows how the indices vary as the number of leaves increases, where each data point is the average of 1000 trees. The diversity indices all monotonically increase from 2. Yule trees are bifurcating and have uneven branch lengths with many ancestor branches overlapping descendant nodes; as a result, the effective out-degree, ${}^qD_N$, is greater than 2 and increases as trees grow, as the number of overlapping branches increases. ${}^qD_S$ is always less than the actual number of branches due to the uneven branch lengths. The evenness indices decrease from 1 as the star tree is maximally balanced, and the number of increasingly unbalanced tree topologies increases as the number of leaves increases.

Distribution and expectation are independent of birth rate

The distribution and expectation of our indices for trees generated by the Yule process are independent of the birth rate. In fact, this is true of any tree shape index that is invariant to a scaling that applies to the whole tree.

A tree with n leaves generated by the Yule process has intervals (time between having k nodes to $k+1$ nodes) given by,

$$T_k \sim \text{Exp}(k\lambda), \quad k = 2, 3, ..., n-1.$$

Using Theorem 14.2 from [97] and letting $X = S_k = cT_k$ for $c > 0$ gives the standard change of variable formula for continuous variables,

$$f_{S_k}(s) = \frac{1}{c} f_{T_k}(s/c), \quad s > 0.$$

Let $\lambda' > 0$ and define scaled intervals,

$$S_k = \frac{\lambda}{\lambda'} T_k.$$

Then by the standard change of variable formula,

$$S_k \sim \text{Exp}(k\lambda'),$$

hence, the sequence of S_k's has the same joint distribution of intervals as a tree with n leaves generated by the Yule process with birth rate λ'.

Now let I_λ denote the value of any scale-invariant index for a Yule tree with birth-rate λ, and let I_λ be a function of the weighting times and 'shape' of the tree,

$$I_\lambda = \Phi(T_2, ..., T_{n-1}, \text{shape}).$$

Then for a tree with rate λ', the corresponding intervals can be taken as $S_k = \frac{\lambda}{\lambda'} T_k$. This gives,

$$I_{\lambda'} = \Phi(S_2, ..., S_{n-1}, \text{shape}) = \Phi\left(\frac{\lambda}{\lambda'} T_2, ..., \frac{\lambda}{\lambda'} T_{n-1}, \text{shape}\right),$$

and by the scale invariance of I,

$$\Phi\left(\frac{\lambda}{\lambda'} T_2, ..., \frac{\lambda}{\lambda'} T_{n-1}, \text{shape}\right) = \Phi(T_2, ..., T_{n-1}, \text{shape}) = I_\lambda.$$

So,

$$I_{\lambda'} = I_\lambda,$$

that is the entire distribution of I_λ is independent of λ, and so all of its moments are also.

Index correlations

Every index pair exhibits a high correlation, with absolute correlation coefficients greater than or equal to 0.927 (see Figure 3.7). Hierarchical clustering assigns each index to the same cluster, showing not only that the pairs are highly correlated, but also that each index is strongly linearly related to every other index (see Figure 3.3c). Performing principal component analysis, I find that the first principal component explains 97% of the variance and each index has an absolute loading onto this component of at least 0.4.

When individual trajectories are used and not averages, the above no longer holds. I performed the same analysis on 1000 Yule trees with 10 leaves. Hierarchical clustering assigns each diversity index and ${}^1J_S$ to their own cluster and clusters ${}^1J_N$ and ${}^1J_L$ together. The first principal component explains 73% of the variance with absolute loadings of at least 0.38. The second principal component explains 16% of the variance, with the smallest absolute loading of 0.2, and four of the indices have an absolute loading of at least 0.45. Therefore, at this tree size, only ${}^1J_N$ and ${}^1J_L$ are strongly linearly related, and one of these, along with the remaining

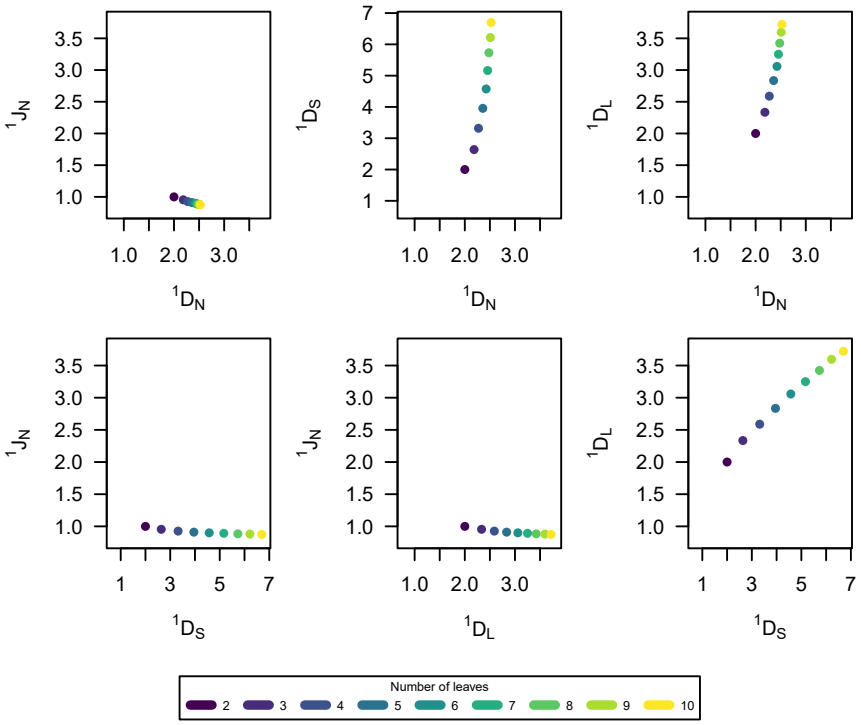

Figure 3.7: Index trajectories formed by four indices, ${}^1J_N$, ${}^1D_N$, ${}^1D_S$, ${}^1D_L$, for trees generated by the Yule process for varying numbers of leaves, n. Each index value is the average of 1000 trees.

indices, is needed to capture tree shape variation.

3.3.4 Comparing the three models

Unsurprisingly, the uniform and equiprobable types models have similar average index values and hence similar index trajectories (see Figures 3.8 and 3.9). In order of similarity: the ${}^1D_S$ values are almost identical, with the equiprobable types model being 1.1% larger on average; then the ${}^1D_N$ values are very close and the equiprobable types model is on average 1.9% larger; the ${}^1J_N$ values are similar with equiprobable types being 3.2% larger; finally, the ${}^1D_L$ values are the most different with equiprobable types being on average 8.7% larger. These two models' average indices are quite dissimilar from the Yule process, and as a result, their trajectories tend to occupy different regions in index space (see Figures 3.8 and 3.9). The indices ${}^1D_N$ and ${}^1D_L$ are particularly different here and this is due to the choice of branch lengths. Equal branch lengths and a bifurcating tree will always result in ${}^1D_N \leq 2$, and a bifurcating tree with branches from ancestor nodes overlapping descendant nodes (as is the case for the Yule process) will lead to ${}^1D_N \geq 2$ and a much bigger branch count across the tree ${}^1D_L$.

Given the correlations and clustering of the indices for these models, using ${}^1D_L$, one of ${}^1J_N$ or ${}^1J_L$ (given the former is a well-studied and known tree shape, it would be the logical choice

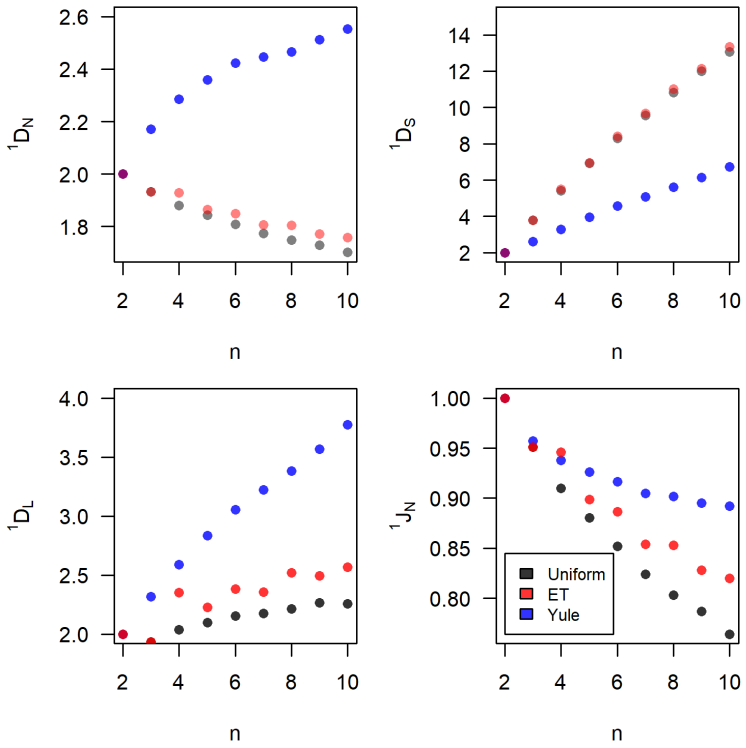

Figure 3.8: Comparison of index values produced by the three different tree-generating models: Yule process, uniform model and equiprobable-types model. Each data point is the average of 1000 trees.

of the pair), ${}^1D_N$ and ${}^1J_S$ should be a sufficient subset of indices that would work well across all three models.

3.3.5 Sensitivity to unresolved polytomies

The new indices have been created to be robust to small changes in tree structure. This section investigates the effect that unresolved polytomies may have on the indices. A polytomy, or a polytomous node, is a node which is multifurcating and not bifurcating [98]. A 'hard' polytomy reflects the true topology of the tree and is the result of 'multiple simultaneous splitting events' [98]. A 'soft' polytomy is the result of incomplete taxonomic resolution [98]. As soft polytomies are due to missing or ambiguous data, an ideal index would be minimally affected by this incomplete resolution.

For this analysis, I studied trees with 6, 8 and 12 leaves, generating 1000 Yule trees for each size. To create unresolved polytomies, I used the `di2multi.multiPhylo` function from the 'ape' package. This function takes a set of trees and a specified 'tolerance' value and collapses any branches with length less than or equal to the tolerance. Only branches whose removal would create polytomies are collapsed; branches terminating in leaves are therefore never considered. For example, a tree with n leaves that has all its (non-leaf-terminating) branches collapsed using this function will result in a star tree with n leaves. I will refer to any tree in which

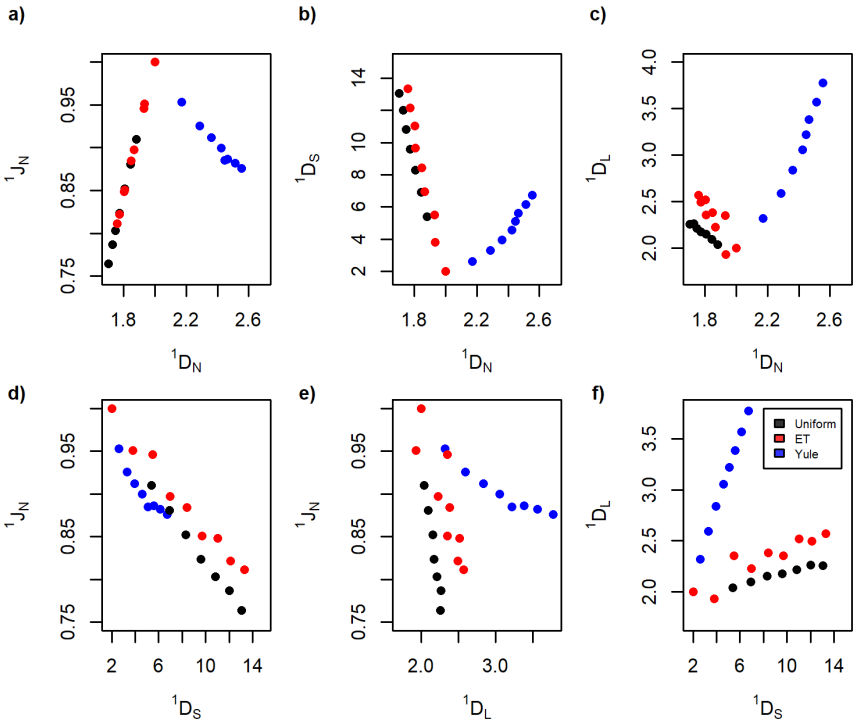

Figure 3.9: Comparison of index trajectories formed by the three different tree-generating models: Yule process, uniform model and equiprobable-types model. Each data point is the average of 1000 trees.

branches have been collapsed, creating polytomies, as an 'unresolved' tree. I investigated the effect of unresolved polytomies with four different tolerance values, 2, 4, 8 and 16. For the Yule process with a birth rate of $\lambda = 0.03$, using equation 3.2, $T_3 - T_2 \sim \text{Exp}(2\lambda)$ and $T_{13} - T_{12} \sim \text{Exp}(12\lambda)$, and hence $\mathbb{E}(T_3 - T_2) = 1/2\lambda \approx 16.67$ and $\mathbb{E}(T_{13} - T_{12}) = 1/12\lambda \approx 2.78$. Therefore, a tolerance value of 2 is reasonable, and 16 is a more extreme value.

I investigated the tree balance index, ${}^1J_N$, and the longitudinal mean diversity index, ${}^1D_L$, as these aim to quantify concepts that are well studied. For comparison with ${}^1J_N$, I used the normalised version of Sackin's index, $I_{S,norm}$, where the formula is as given in section 1.3. It is worth noting that Sackin's index is an imbalance index, so when ${}^1J_N$ increases, the normalised Sackin's index will decrease. For comparison with ${}^1D_L$, I ran into issues with finding a suitable index. To be fair, I ideally required a tree shape index that uses both abundances and branch lengths, is normalised and can be expressed in effective units as our index is. The best index for this would be ${}^1\overline{D}$ which belongs the family of indices defined by Chao et al. and are the only indices to satisfy my ideal requirements [66] (note that ${}^1\overline{D}$ is the normalised version of phylogenetic entropy [45] and so I do not consider this index separately). However, for leafy ultrametric trees, ${}^1\overline{D}$ is equivalent to our index ${}^1D_L$ [59]. The method I used to generate Yule trees creates ultrametric trees, and as there are no node sizes, I assigned leaves to have size one and internal nodes to have size zero, making the trees leafy and ultrametric, and hence ${}^1D_L = {}^1\overline{D}$. The only other index that was somewhat suitable was Shannon diversity (or Shannon entropy) but this index would remain unchanged when polytomies were introduced

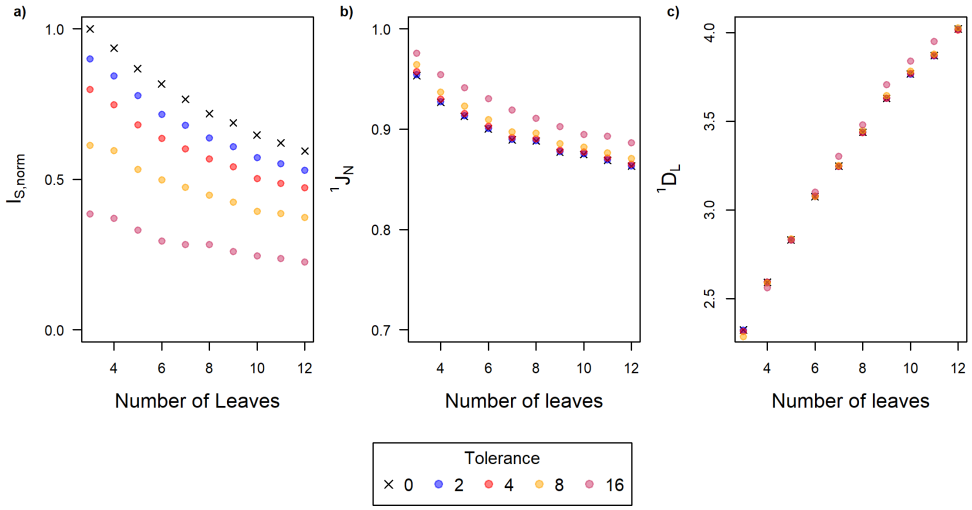

Figure 3.10: The effect of unresolved polytomies as trees grow, shown by the number of leaves, on a) $I_{S,norm}$, b) ${}^1J_N$ and c) ${}^1D_L$. Each data point is the average of 1000 trees generated by the Yule process with a birth rate of $\lambda = 0.03$. For each tolerance value, any branches with length less than or equal to the tolerance are collapsed to create an unresolved polytomy.

into the trees. Other diversity indices, such as Faith's phylogenetic diversity, are not normalised and so are not tree shape indices as they are not independent of tree size. As a result, I did not include a comparison for ${}^1D_L$.

Figure 3.10 shows the index values for the original set of trees, and for sets of unresolved trees with different tolerance values. For all tolerance values, the new indices, ${}^1J_N$ and ${}^1D_L$, are robust to unresolved polytomies, and ${}^1J_N$ appears to be more robust than $I_{S,norm}$. Figure 3.11 is a set of boxplots showing the relative difference between the index values of the original trees and the unresolved trees for different numbers of leaves for a tolerance value of 16. This demonstrates again that ${}^1J_N$ and ${}^1D_L$ are robust to polytomies, and that balance appears to be more robust than diversity. It can also be seen clearly that ${}^1J_N$ performs much better than $I_{S,norm}$.

3.4 Discussion

I have shown that for the three models investigated here, many of our indices are highly correlated and therefore redundant when analysing average trajectories. For the uniform model and Yule process, only one index is required to capture the change in shape as the trees grow, whereas for the equiprobable-types model, two are. In contrast, I found that for the Yule process at fixed tree size, 5 out of 6 indices are necessary to describe the differences in tree shape. Demonstrating that it is the averaging over trajectories that inflates redundancy. The latter analysis is not a perfect equivalent, as it was at a fixed tree size and not over trajectories. However, this approach was taken because the analysis of individual trajectories is non-trivial.

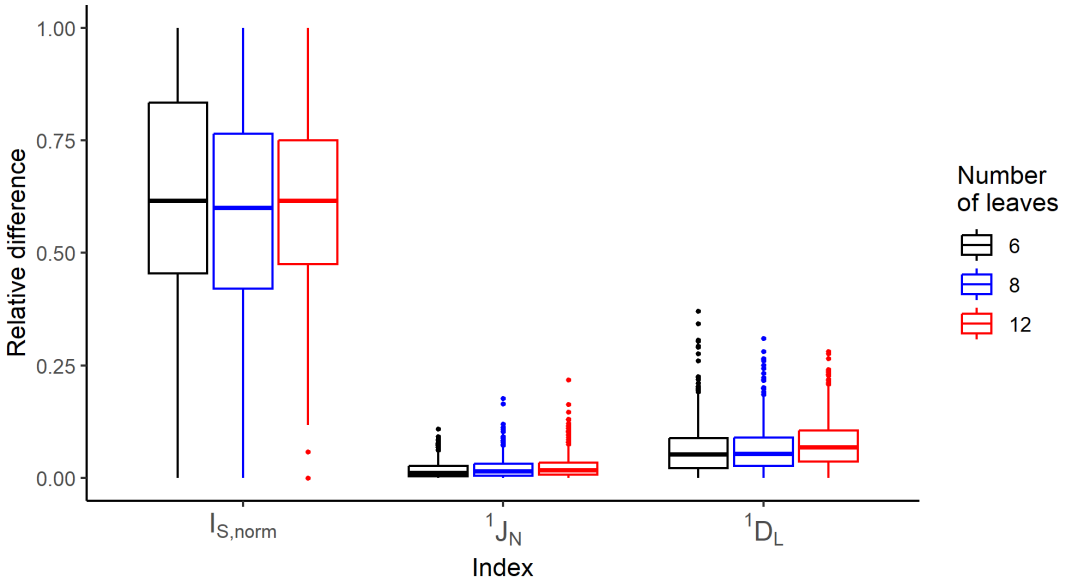

Figure 3.11: The absolute difference between each index value for the original tree and the unresolved tree with a tolerance value of 16.

I demonstrated empirically, using tree balance and effective out-degree, that our indices are not only robust to unresolved polytomies, but they are much more robust than other commonly used indices. This is a crucial property for indices used to study empirical data, as any soft polytomies will have a minimal effect on the indices and the subsequent results and conclusions drawn from them.

Future work could investigate more complex non-spatial models, not only to further our understanding of how the indices behave under different models, but also to move to more biologically informed models. Natural candidates include extending the Yule process to incorporate death, as well as the beta-splitting and the alpha-beta splitting models, which allow explicit control over tree balance and asymmetry. Together, these models would provide a broader framework for exploring how different diversification dynamics are reflected in tree shape and in the indices considered here.

In conclusion, this chapter shows that the apparent index redundancy is not an inherent property of the indices themselves but arises from averaging over trajectories. While the indices capture distinct aspects of tree structure at a fixed tree size, averaging across trajectories can substantially inflate redundancy, with the extent of this effect depending on the model. The chapter also demonstrates that our indices are robust to unresolved polytomies, and more so than other commonly used indices. Taken together, these results lay the groundwork for further analyses using the non-spatial models considered here, in particular the Yule process, which is frequently used as a null model, and provide a foundation for future work with more complex and biologically informed models.

Chapter 4

Detecting branching rate heterogeneity

4.1 Introduction

The shape of every phylogenetic tree is influenced by the evolutionary processes that produced it [79, 80]. In fact, tree balance has been related to variation in both speciation and extinction rates, and variation in such factors is connected with the underlying evolutionary processes [79]. In macroevolution, the signal left in tree shapes has been used to study the diversification process and how it has varied across time, clades and geological regions, and multiple patterns in tree shape and branch lengths have been observed [99]. For example, trees are frequently found to be more imbalanced than would be expected under simple null models [44, 80]. However, the reasons for such patterns are not well understood [99].

The study of signals in phylogenetic tree shape has applications beyond macroevolution. Cancer has long been considered an evolutionary process. It is known to exhibit widespread heterogeneity in both genotype and phenotype, which can influence key biological processes such as the rate of cell division [100]. Detecting such variations is crucial to providing further insight into the evolutionary processes underpinning cancer. A deeper understanding of these evolutionary processes should allow more accurate and patient-specific prognoses, and the optimisation of therapy regimens [57].

Although tree balance has been used to study cancer trees and can detect rate heterogeneity [56, 58, 101–105], it is rarely utilised in this way. In macroevolution, tree balance is frequently used to study rate variation [44, 79, 80]. However, these two areas have something in common. When tree balance is used, the indices chosen are not ideal. The most commonly used tree balance index is Sackin's index; however, this has various properties that can lead to issues (see Section 1.3 and [59]). Most relevant here is the fact that Sackin's index does not account for branch lengths or node sizes, leading to the index being sensitive to imperfections in data.

For example, in cancer, issues in data could arise from sampling errors, and in macroevolution, they could arise due to incomplete fossil records. Additionally, tree balance is not the only tree shape metric, but there is limited research using other tree shape indices to study rate variation. Hence, there is a need to study rate variation in cancer through tree balance, generally using a rigorously defined tree balance index, and using other tree shape indices.

Here, I study two models of branching rate heterogeneity and investigate whether our tree shape indices can detect it. I use both a Bayesian and a frequentist approach to compare the methods. I investigate whether our indices can distinguish between the two models of rate heterogeneity, and finally look at the trajectories they form in index space and correlations between the indices. I find that our indices can detect rate heterogeneity, and index performance depends on the model: tree balance performs the best for one model, and the effective number of branches in the tree performs the best for the other. I find that tree balance always outperforms Sackin's index when detecting rate heterogeneity, and that both Bayesian and frequentist approaches return very similar results. Lastly, I find that our indices can distinguish between the two models.

While completing this work, Feder and Gao released a preprint titled 'Detecting branching rate heterogeneity in multifurcating trees with applications in lineage tracing data' [58]. The paper is primarily focused on lineage tracing data. They demonstrate how differences in methods for lineage tracing result in distinct characteristics within the tree's shape. They use multiple indices, including the previous version of our balance index that does not use branch lengths (J^1) and Sackin's index, to try to detect branching rate heterogeneity for lineage tracing data. Most relevant to the work here is that they find that J^1 'effectively detects branching rate heterogeneity in simulated lineage tracing data' and 'tests based on other common statistics ... show inferior performance to J^1'. Aside from the clear differences in the context of lineage tracing data, there are several differences in the methods used in this paper and the work presented here. They use different models of branching rate heterogeneity, a frequentist approach exclusively, and they analyse larger and single trees (n = 50, 250, 1250 and 6250).

4.2 Methods

4.2.1 Models

Random time random branch

For the Yule process, the time for each branch to bifurcate is given by $\exp(\lambda)$, the inter-bifurcation time, the time that it takes for the system to go from n to $n+1$ terminal branches (or equivalently leaves), will be given by the smallest bifurcation time of the current terminal branches. That is,

$$T_{n+1} - T_n = \min\{Z_i : Z_1, \ldots, Z_n \overset{i.i.d}{\sim} \text{Exp}(\lambda)\}. \tag{4.1}$$

In fact,

$$T_{n+1} - T_n \sim \text{Exp}(n\lambda). \tag{4.2}$$

The R package 'Treesim' simulates Yule trees by two methods: either starting with the given number of leaves and performing coalescence for random pairs of branches, or by starting with the root node and then having branches bifurcate randomly. Whichever method is used, the branch lengths are obtained by using the inter-bifurcation times distribution, given by equation 4.2. To simulate a Yule-type process with branching rate heterogeneity, I modified the code to obtain the inter-bifurcation times using equation 4.1. This removed the random choice of which branch would bifurcate and, instead, the next bifurcating branch would be given by the branch with the current shortest bifurcation time. To add branching rate heterogeneity, I included a given chance of obtaining a 'mutation' at each bifurcation event. Each descendant branch of the bifurcating branch could independently mutate or not, allowing for parallel evolution. Acquiring the mutation meant an increase in the branching rate by a given factor. Once a mutation occurred, it could not be lost, so when a branch has acquired the mutation, all of its descendants would also carry it. This model allows branching rate heterogeneity to occur on a random branch at a random time; hence, I will refer to it as the random time random branch (RTRB) model.

I carried out simulations for 32 pairs of parameter values. For the branching rate factor increase, I considered 2.5, 5, 7.5 and 10, corresponding to medium to strong heterogeneity. For the probability of mutation, I considered values of 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, and 0.4.

Fixed time full tree

For this model of branching rate heterogeneity, the branching rate of the full tree changes at some fixed time point. Treesim has a function called `sim.bd.rateshift.taxa` that already does this; however, it is not ideal for use here. The function takes the time at which the change in branching rate occurs, with the time set to 0 at the present and it increases as it moves backwards in time. Hence, you specify some time point in the past where the branching rate changes. I aim to investigate the effect of varying both the time at which the branching rate changes and the factor by which the branching rate changes. As a result, going back in time is not as simple as it means the times at which the branching rate should be changed will differ depending on the factor by which the branching rate changes. For example, consider two trees with heights 12 and 20. Looking 2 and 10 units of time back from the present is the same as looking 10 units of time from the root of the tree. As all the trees start with the same branching rate, it is much simpler to go forward in time rather than backwards. Using the code I had already edited in the previous section, I further modified this such that at a given time, the branching rate across the whole tree could be changed by a given factor. This model reflects what is seen for trees inferred from the fossil record, where large speciation events will likely be due to large environmental changes affecting all species. I will refer to this as the fixed time full tree (FTFT) model.

I look at trees with 10 and 20 leaves. The inter-bifurcation times are given by equation 4.2,

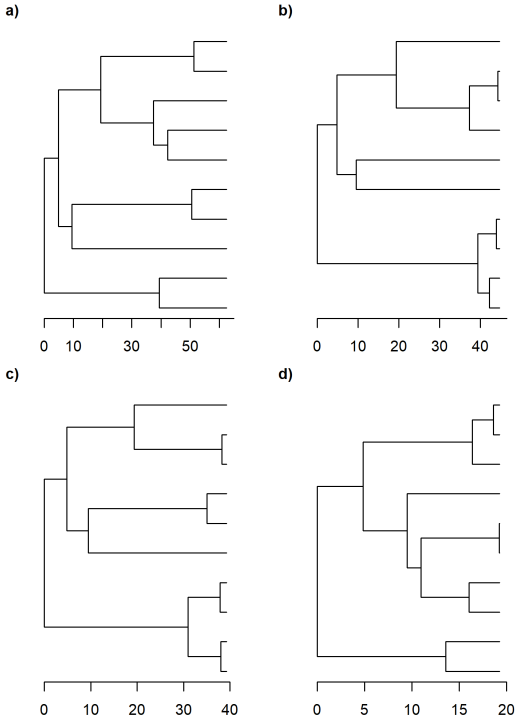

Figure 4.1: Trees showing the effect of changing the time at which the branching rate changes for the full tree fixed time model. a) Yule tree, and b-d) FTFT model trees where the time of the branching rate change is b) 10, c) 30 and d) 40. The branching rate increase is 5.

the expectation of the inter-bifurcation time (equivalently, the height of the tree between n and $n+1$ terminal branches) is given by,

$$\mathbb{E}[T_{n+1} - T_n] = \frac{1}{n\lambda}, \tag{4.3}$$

where λ is the bifurcation rate (or equivalently the birth rate). Trees with n leaves are generated by simulating the Yule process up until a tree has n terminal branches and terminating at the moment one of these would bifurcate. Therefore, the expectation of the height of a Yule tree with n leaves generated in this way is,

$$\mathbb{E}[\text{height of tree}|n \text{ leaves}] = \sum_{i=2}^{n} \frac{1}{i\lambda}.$$

Hence, $\mathbb{E}[\text{height of tree}|10 \text{ leaves}] = 63.30$ and $\mathbb{E}[\text{height of tree}|20 \text{ leaves}] = 86.59$.

For the branching rate factor increase, I again considered 2.5, 5, 7.5 and 10, corresponding to medium to strong heterogeneity. For the time of the rate change, I considered intervals of 10 starting from 10 for both tree sizes and ranging to 50 for 10-leaf trees and 70 for 20-leaf trees. This is equivalent to very early to very late in the average tree's lifetime.

4.2.2 Bayesian approach

For this approach, I require a null model for the Yule process and a null model for Yule with branching rate heterogeneity. I form empirical nulls by generating 1000 trees for the respective models and calculating the index values. I then use kernel density estimation to approximate the probability density function from the index values. The Bayesian approach requires us to generate distributions for the models of braching rate heterogeneity for each set of parameter values I are investigating. This is computationally intensive, and so I use 1000 trees for these distributions.

I will denote the null models for the Yule process and the Yule process with branching rate heterogeneity as θ_Y and θ_H, respectively. If I have a set of trees A and I assume independence of features, then the likelihoods are given by,

$$P(A|\theta_Y) = \prod_{i=1}^{n} P(A_i|\theta_Y),$$
$$P(A|\theta_H) = \prod_{i=1}^{n} P(A_i|\theta_H).$$

Bayes' theorem gives the posterior probabilities,

$$P(\theta_Y|A) = \frac{P(A|\theta_Y)P(A)}{P(\theta_Y)},$$
$$P(\theta_H|A) = \frac{P(A|\theta_H)P(A)}{P(\theta_H)}.$$

If I assume equal priors,

$$P(\theta_Y) = P(\theta_H),$$

then the ratio of the posterior distributions simplifies to,

$$\frac{P(\theta_Y|A)}{P(\theta_H|A)} = \frac{P(A|\theta_Y)}{P(A|\theta_H)} = \frac{\prod_{i=1}^{n} P(A_i|\theta_Y)}{\prod_{i=1}^{n} P(A_i|\theta_H)}.$$

This ratio can then be turned into 'confidence scores'. For example, the confidence score for the set of trees A being from distribution θ_Y is,

$$\frac{P(A|\theta_Y)}{P(A|\theta_Y) + P(A|\theta_H)}.$$

The confidence scores are simply the likelihood over the total likelihood.

I calculate the confidence scores for correctly assigning a tree to have or not have branching rate heterogeneity. For example, if I generated samples for Yule trees, I calculated the confidence score that those samples are from the Yule null, θ_Y.

4.2.3 Frequentist approach

For the frequentist approach, I only require null models for the Yule process for the two tree sizes I investigate. Therefore, for the null models here, I use 10000 Yule trees for each tree size. For each sample size, I compute a p-value for each sample by comparing it to the null distribution. Then the p-values are combined using Fisher's method, which I do empirically, and the null is either accepted or rejected. Using the empirical null, I generate the null Fisher distribution for the given sample size, employing a two-tailed test for the individual p-values. From this distribution, I calculate the 95% threshold. The details of this are: I sample from the empirical null the desired sample size, I then compare these to the null distribution using a two-tailed test. As our indices tend to be skewed, I perform the two-tailed test by taking the minimum of the right and left tail probabilities and doubling it. This gives a conservative p-value estimation suitable for skewed data. I then calculate the Fisher statistic for these p-values. I perform this 1,000 times, generating an empirical Fisher distribution for our null model, and from this, I calculate the 95% threshold. This threshold is then used to accept or reject the null model for a given set of samples.

To assess power, I generate samples of our models with branching rate heterogeneity, and the power is the proportion of rejections. For the type 1 error, I generate samples of the Yule process, and the proportion of rejections is the type 1 error. I use this method, rather than methods such as the Kolmogorov-Smirnov test, as I am testing small sample sizes. Additionally, empirically generating the distribution of Fisher statistics calibrates the test and controls the type 1 error.

4.2.4 Comparison to other indices

Out of our tree shape indices, I only compare our balance index to a commonly used alternative, Sackin's imbalance index. With our diversity index ${}^1D_L$, I run into the same issues I did in section 3.3.5. Although I would also like to have a diversity index to compare to, there exists only one tree-shape based (or normalised) diversity index that would be suitable here. This index belongs to the family of indices defined by Chao et al, however, for leafy ultrametric trees, this index is equivalent to our longitudinal diversity index ${}^1D_L$ [59, 66]. The method I use to generate trees here creates ultrametric trees, and as there are no node sizes, I assign leaves to have size one and internal nodes to have size zero, making the trees leafy and ultrametric. This makes the only tree-shape based diversity index equivalent to our ${}^1D_L$. Other diversity indices, such as Faith's phylogenetic diversity or phylogenetic entropy (defined by Allen et al.), are not normalised and so are not tree shape indices as they are not independent of tree size.

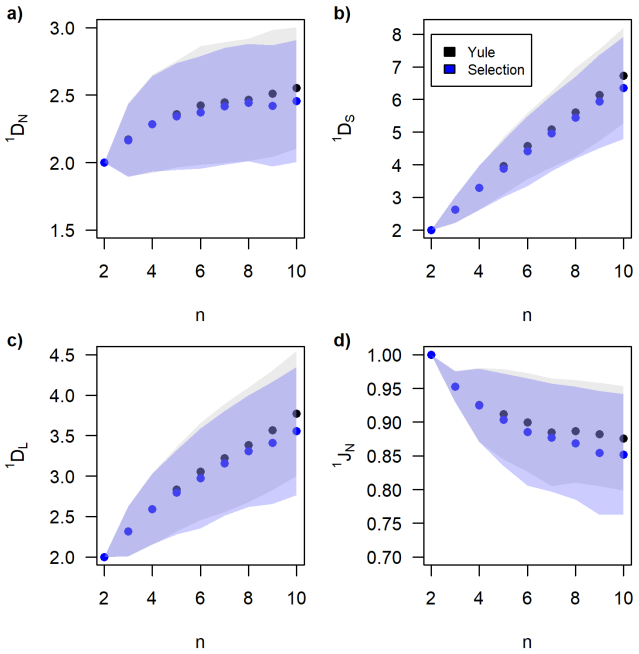

Figure 4.2: Average index values for 1000 Yule trees for varying numbers of leaves, n, with and without branching rate heterogeneity (RTRB). Shaded regions show plus and minus one standard deviation.

4.2.5 Comparison of random-time random-branch and full-tree fixed-time models

To compare the ability of indices to detect between the two types of branching rate heterogeneity, I need to categorise the data in such a way that the two parameters, the probability of mutating and the time at which the branching rate changes, become equivalent. To do this, I bin the data for RTRB based on when the first mutation occurred. I create 5 time groups, which are selected to align with existing times for FTFT and to have approximately 1000 trees for RTRB. For RTRB, I have groups 1, 2, 3, 4, 5, which correspond to trees where the time the mutation occurs, t, is, $t \leq 15$, $15 < t \leq 25$, $25 < t \leq 35$, $35 < t \leq 55$ and $55 < t \leq 75$, respectively. For FTFT, these groups correspond to time 10, 20, 30, 40 and 50, and 60 and 70.

4.2.6 Index correlations

The analysis of index correlations follows the methodology described in Section 3.2.2.

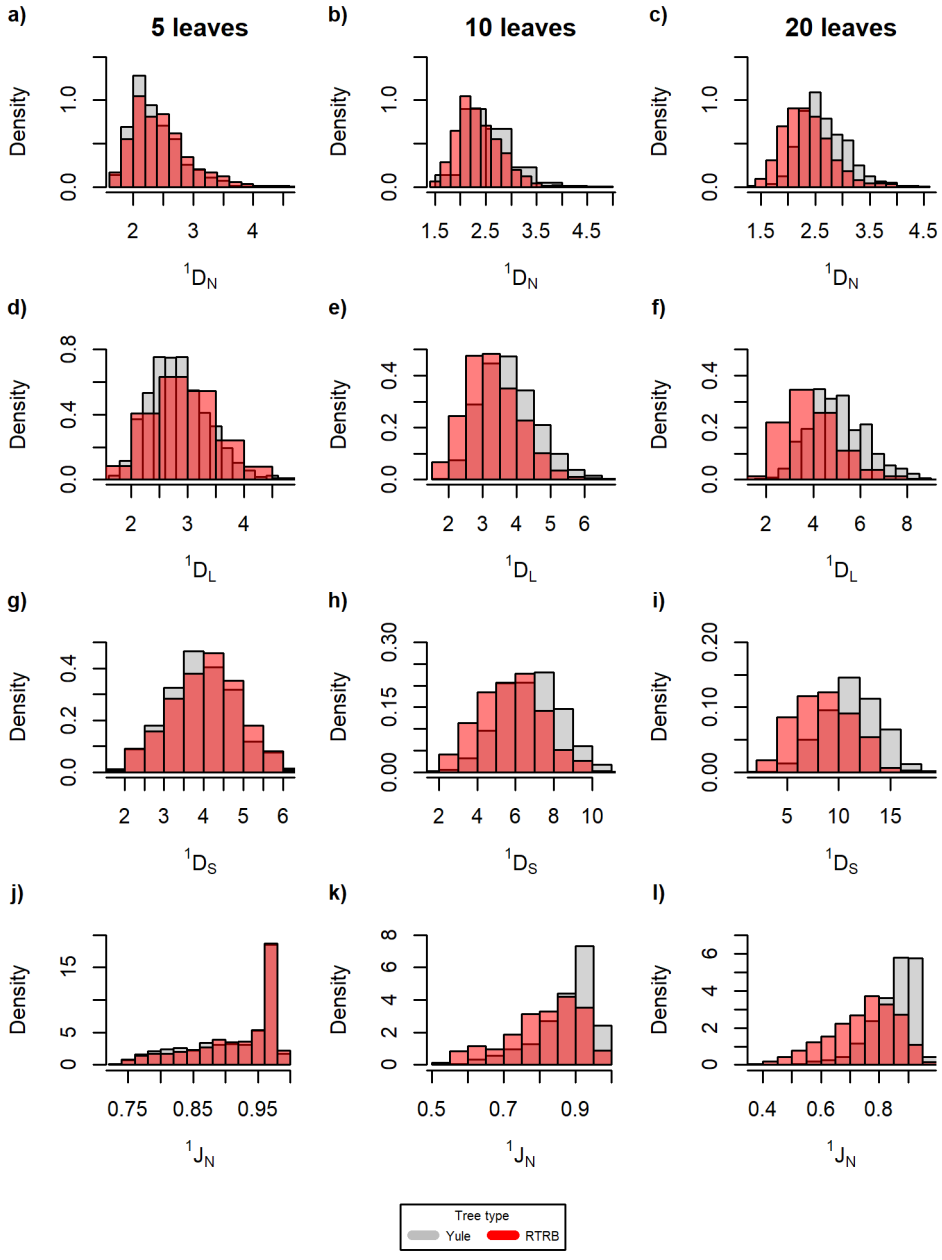

Figure 4.3: Distribution of index values for 1000 either Yule trees or RTRB trees with a,d,g,j) 5 leaves, b,e,h,k) 10 leaves and c, f, i,l) 20 leaves. The probability of mutation is 0.3 and the branching rate increase from a mutation is a factor of 5.

4.3 Results

4.3.1 Initial work

The work here was the first analysis I carried out for this project, serving as a proof-of-concept exploratory analysis before the larger body of work was carried out.

To investigate the effect of branching rate heterogeneity on our tree indices, I first examined the average index values for the RTRB model and compared them to those of the Yule process. The parameter values were chosen somewhat arbitrarily, but with the aim that the heterogeneity should theoretically leave a detectable signal. I set the probability of branching rate change to be moderate at 30% and the branching rate increase to 2.5. I observed that, on average, trees with branching rate heterogeneity had smaller index values than trees without (Figure 4.2). This did not include the case of two leaves, which are identical, as the root node branching rate is fixed. As the tree size increases, the difference in the index values increases. Notably, from $n = 5$ the difference begins to increase quite significantly.

Figure 4.3 shows the distribution of index values for the RTRB model for different numbers of leaves. Given Figure 4.2 the distributions are as I would expect. The index values for trees with branching rate heterogeneity have a left-skewed distribution compared with the Yule trees, and the shift increases as tree size increases. Considering the model, this is also what I would expect. When a branch's branching rate increases, it will bifurcate faster and hence have a shorter length. This means fewer long overlapping branches from ancestors, resulting in the effective out-degree (${}^1D_N$) being closer to two than in the Yule case, and a smaller effective number of branches across the tree (${}^1D_L$). There will be a large difference between branch sizes with an increased number of shorter branches, leading to the effective number of branches (${}^1D_S$) being smaller than for the Yule process. Another way of thinking about this is that the tree is becoming less star-like, and hence the effective number of maximally distinct leaves (${}^1D_S$) will be smaller than for Yule trees. All of this together, with the fact that once the branching rate change has occurred, it cannot be lost, resulting in sections of the tree beginning to effectively dominate, results in tree balance (${}^1J_N$) being smaller for trees with branching-rate heterogeneity compared with Yule tree.

4.3.2 Random-time random-branch model

Bayesian approach

For all parameter values, every index had over a 50% probability of correctly detecting RTRB or not. I observed that the probability increased as the number of samples increased and that the probability was generally higher for trees with 20 leaves versus 10 leaves. Additionally, as both the branching rate factor and probability of mutation increased, the probability increased (see Figure 4.4a and Supplementary Figure 8.3). I found that our tree balance index, ${}^1J_N$, performed the best, then ${}^1D_S$ and ${}^1D_L$ performed similarly, and finally ${}^1D_N$ performed the

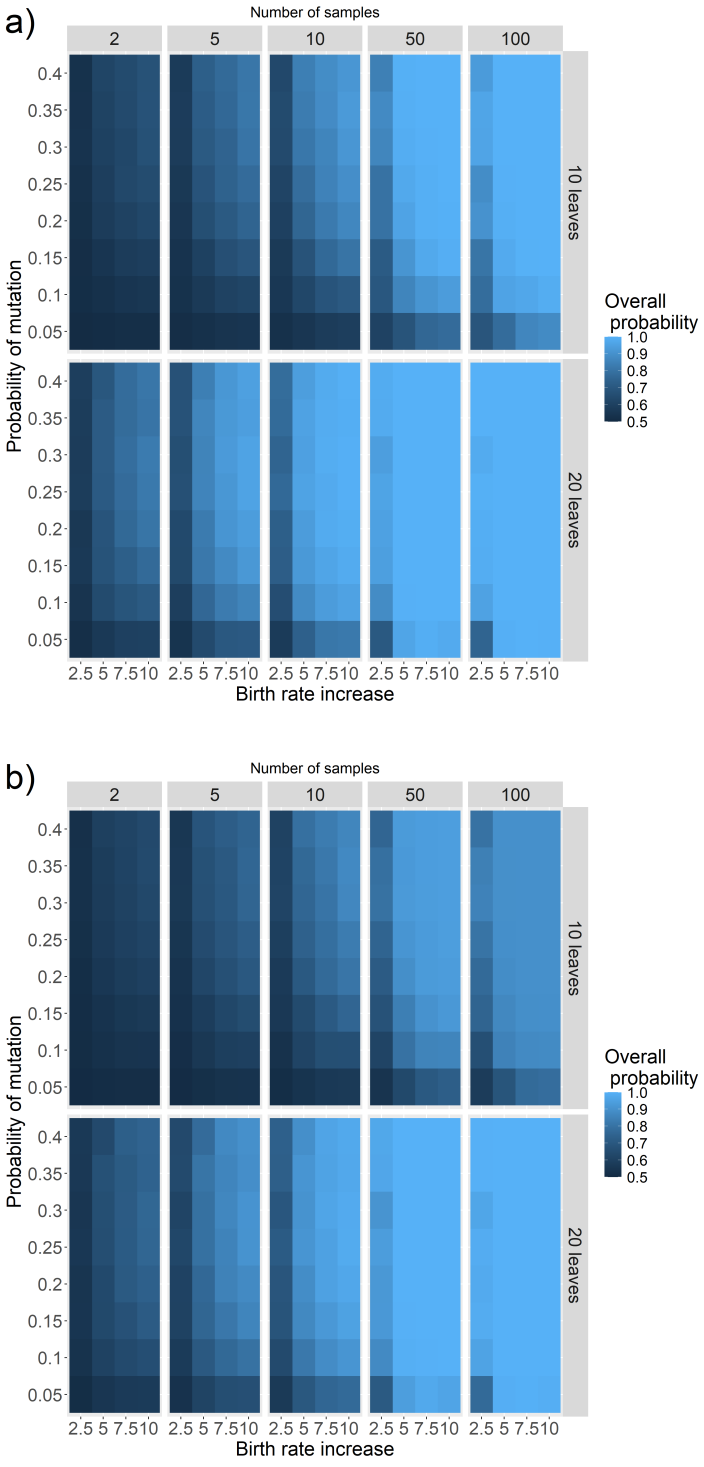

Figure 4.4: Heat maps showing the overall probability of a) our tree balance index ${}^1J_N$ or b) Sackin's index I_S, correctly detecting branching rate heterogeneity under the RTRB model. The heat maps illustrate the impact of varying the mutation probability, the increase in branching rate resulting from the mutation, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the overall probability.

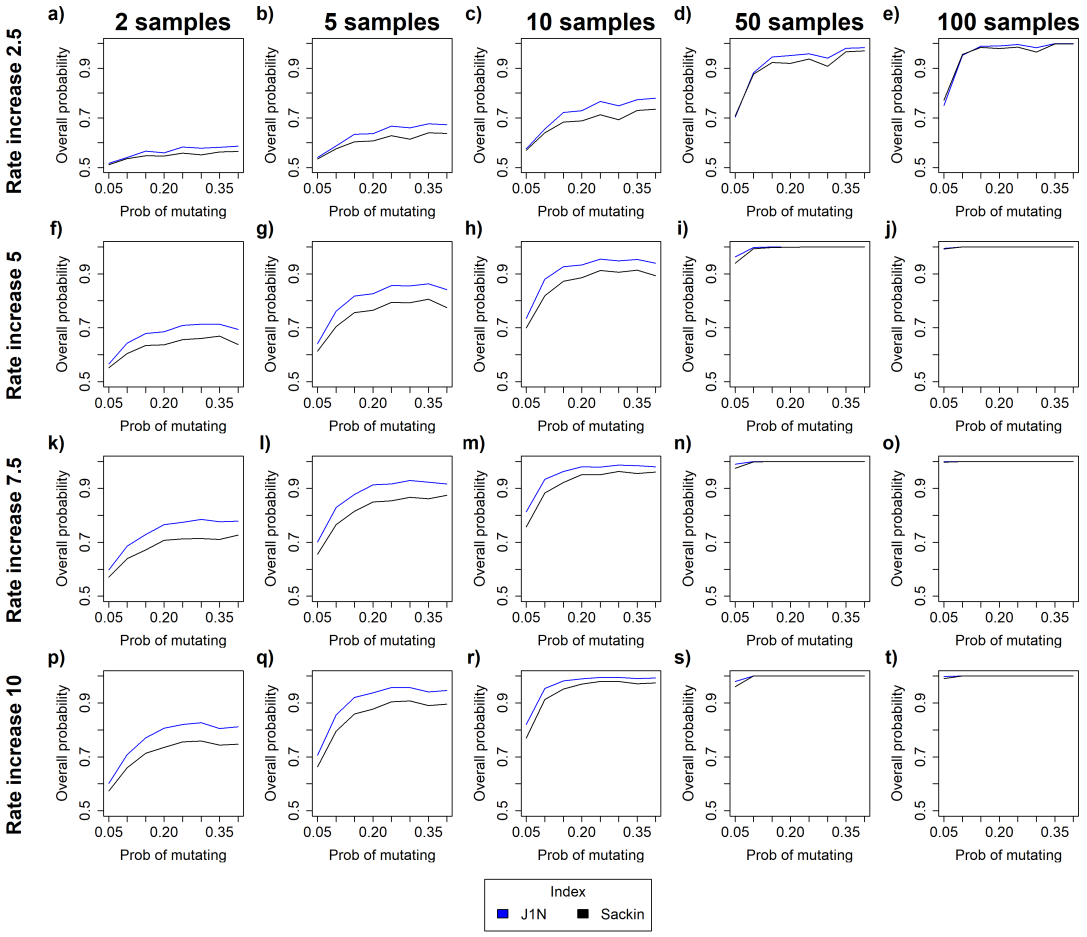

Figure 4.5: Overall probability of our balance index and Sackin's index correctly detecting branching rate heterogeneity under the RTRB model for trees with 20 leaves. The x-axis represents the probability of mutating. Vertical columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and horizontal columns correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

worst (see Supplementary Figure 8.4). Sackin's index I_S performed well and exhibited all the same patterns as our indices (see Figure 4.4b). ${}^1J_N$ performed better than Sackin's index in all cases other than when they both have probability 1 (see Figure 4.5).

Frequentist approach

I observed largely the same results using the frequentist approach as I did using the Bayesian approach. Power increased as the number of samples increased, was higher for the trees with 20 leaves compared with 10, and increased as the branching rate factor and probability of mutation increased (see Figure 4.6a). With this approach, the difference between the indices' abilities was not as distinct as it was with the Bayesian approach. Generally, ${}^1J_N$ performed the best with ${}^1D_L$ close behind and sometimes outperforming tree balance, and ${}^1D_N$ and ${}^1D_S$ performed similarly (see Supplementary Figure 8.6). There was a lot of crossover and violations of this order. I again used the normalised Sackin's index for comparison. The Sackin's index performed well, but ${}^1J_N$ outperformed it in virtually all cases (see Figures 4.6b and 4.7).

${}^1J_N$ had the best type 1 error rates out of our indices, with the other indices performing similarly to each other (see Figure 4.8). There was no clear difference in the type 1 error rate for the different-sized trees for our indices. Sackin's index also had reasonable type 1 errors, but performed much better for trees with 20 leaves compared with 10 leaves (see Figure 4.8e).

4.3.3 Full-tree fixed-time model

Bayesian approach

I again found that for all parameter values, every index had over a 50% probability of correctly detecting whether there had been a change in the branching rate or not. I observed that the probability increased as the number of samples increased, and there were minimal differences between tree sizes (see Figure 4.9a and Supplementary Figure 8.7). Generally, the overall probability increased as the branching rate change increased and as the time of the rate change decreased. However, for some parameter values, primarily for small sample sizes and or a lower branching rate increase, the power decreased again when the rate change occurred early. The index ${}^1D_S$ performed the best, which was closely followed by ${}^1D_L$, then ${}^1D_N$ and ${}^1J_N$ performed the worst (see Supplementary Figure 8.10).

I found that Sackin's index performed poorly. The overall probability ranged from 0.43 to 0.56 and no longer exhibited the typical patterns I had observed (see Figure 4.9b). The probability remained roughly constant and did not change systematically as the parameters were varied. It is unsurprising, then, that despite ${}^1J_N$ being the worst-performing of our indices, it still massively outperformed Sackin's index (see Figure 4.10).

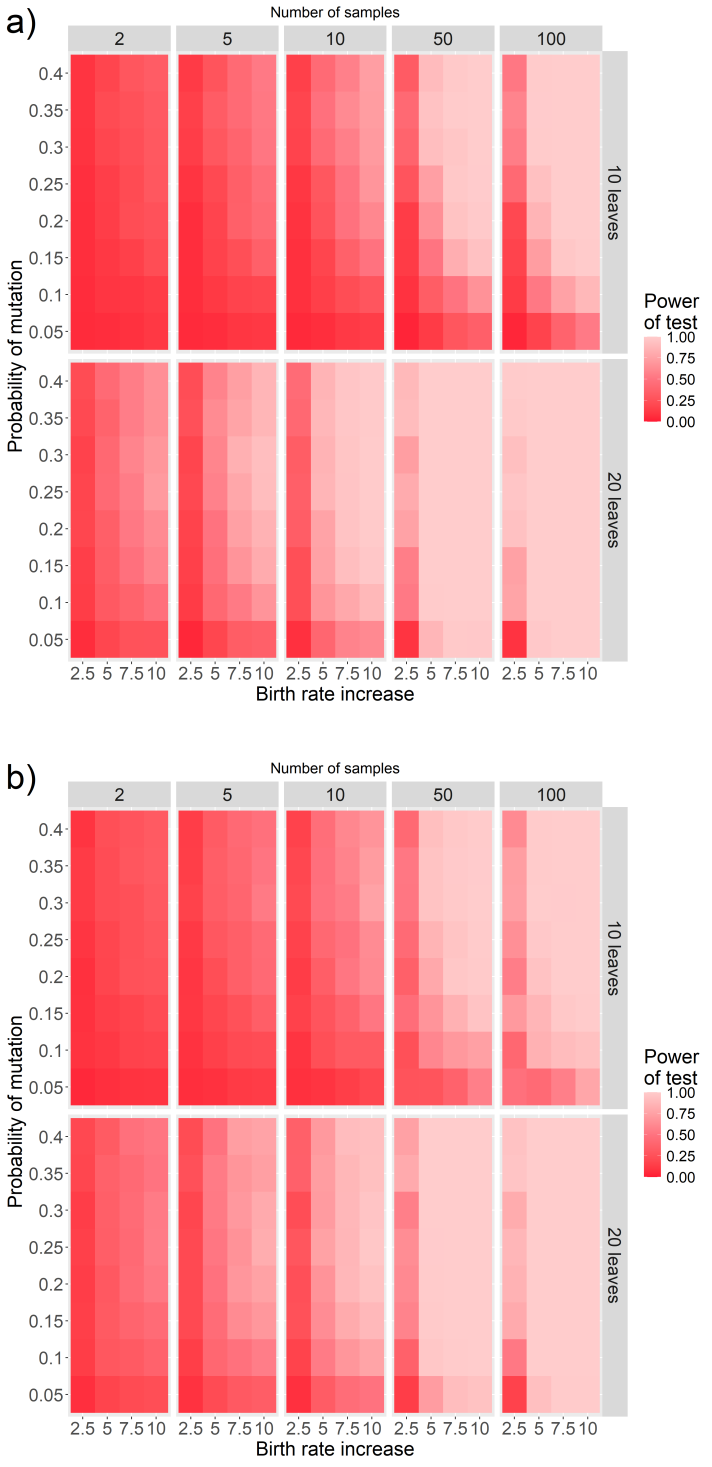

Figure 4.6: Heat maps showing the power of a) our tree balance index ${}^1J_N$ or b) Sackin's index I_S to correctly detect branching rate heterogeneity under the RTRB model. The heat maps illustrate the impact of varying the mutation probability, the increase in branching rate resulting from the mutation, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the power.

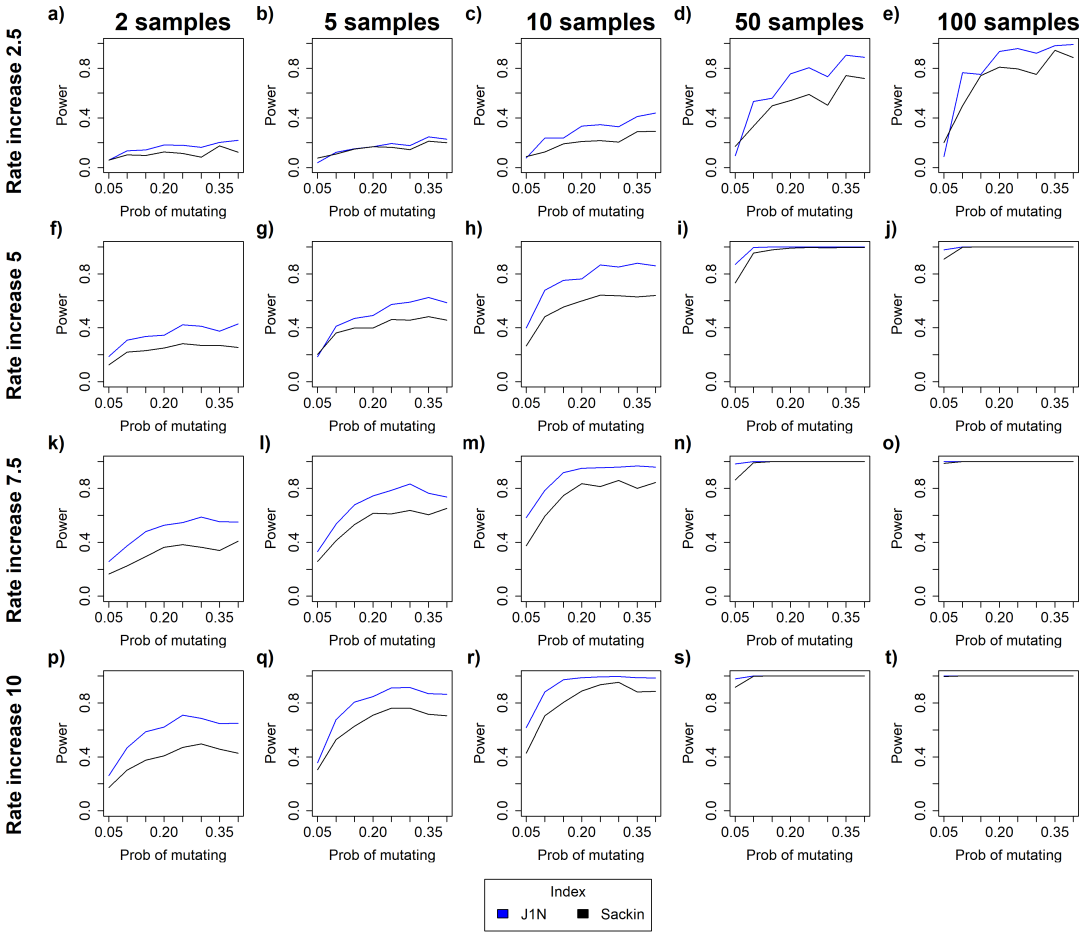

Figure 4.7: Power of our balance index and Sackin's index to correctly detect branching rate heterogeneity under the RTRB model for trees with 20 leaves. The x-axis represents the probability of mutating. Vertical columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and horizontal columns correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

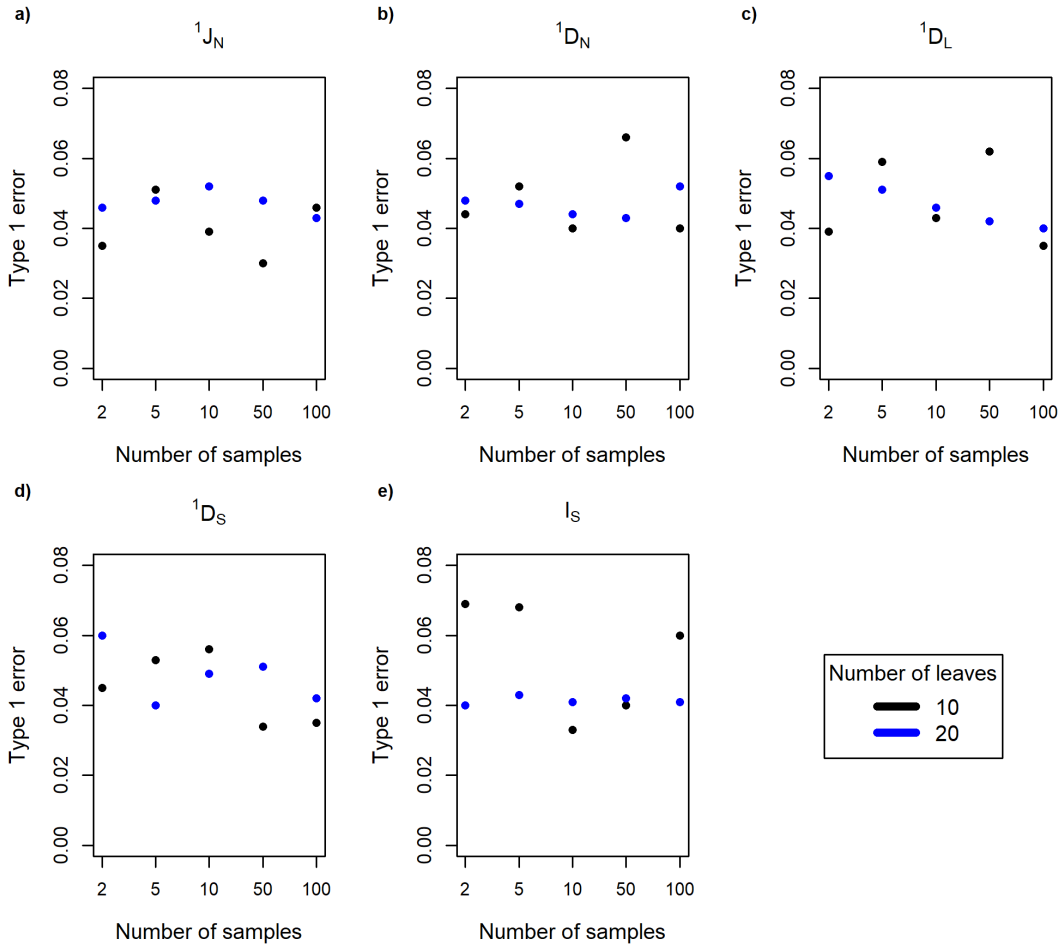

Figure 4.8: Type 1 errors for indices a) ${}^1J_N$, b) ${}^1D_N$, c) ${}^1D_L$, d) ${}^1D_S$ and e) Sackin (I_S).

Frequentist approach

I found virtually the same results using the frequentist approach as I did using the Bayesian approach. Power increased as the number of samples increased, and there was minimal difference between tree sizes (see Figure 4.11a and Supplementary Figure 8.9). Generally, the power increased as the branching rate increased and as the time of the branching rate change decreased. However, I observed here more strongly than with the Bayesian approach that, primarily for small sample sizes and a lower branching rate increase, the power decreased again when the rate change happened early. The relative performance of the indices remained unchanged with ${}^1D_S$ being the best, followed by ${}^1D_L$, then ${}^1D_N$ and finally ${}^1J_N$ (see Supplementary Figure 8.10). Sackin's index still performed poorly, reaching a maximum power of only 18.1%, and remained roughly constant, no longer exhibiting the typical pattern as the parameter values are varied (see Figure 4.11b). ${}^1J_N$ still massively outperformed Sackin's index (see Figure 4.12).

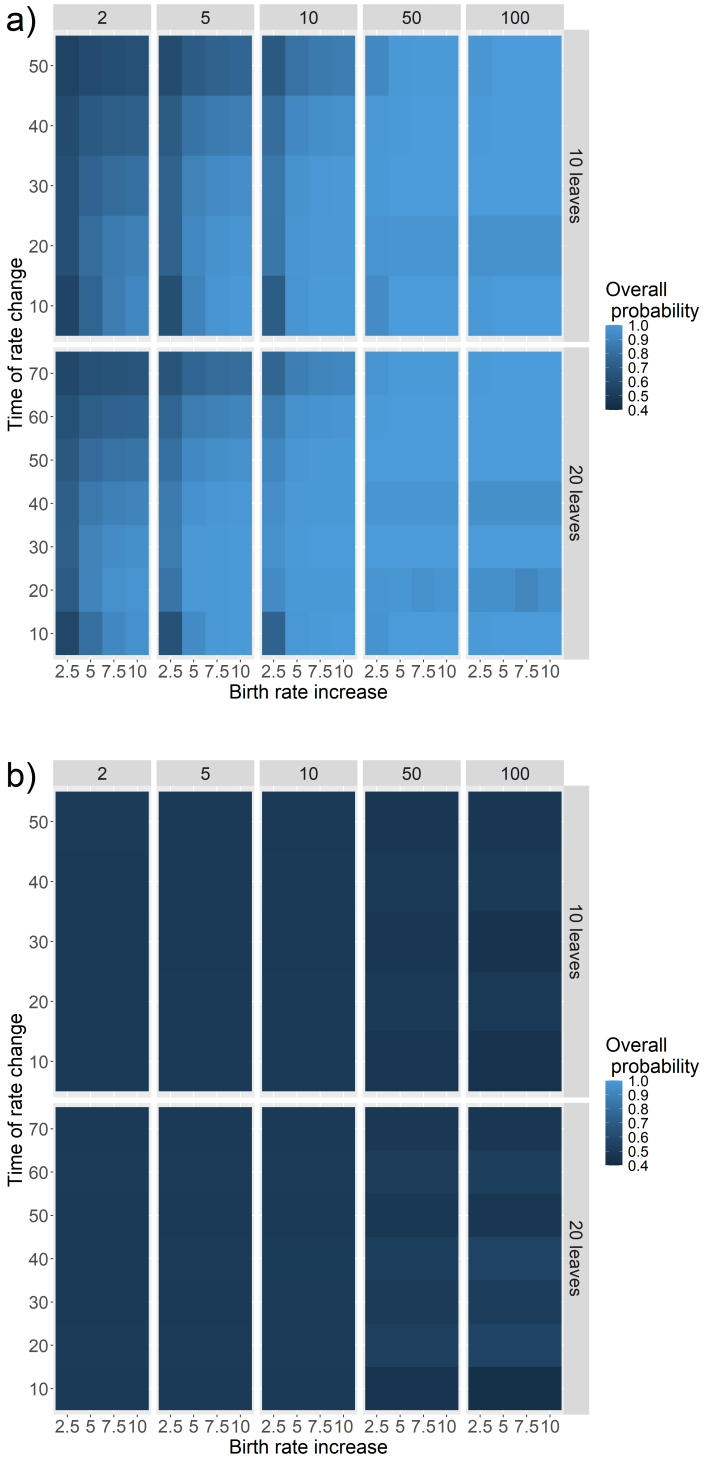

Figure 4.9: Heat maps showing the overall probability of a) our index ${}^{1}D_S$ and b) Sackin's index I_S, correctly detecting branching rate heterogeneity under the FTFT model. The heat maps illustrate the impact of varying the time at which the branching rate across the whole tree changes, the increase in branching rate, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the overall probability.

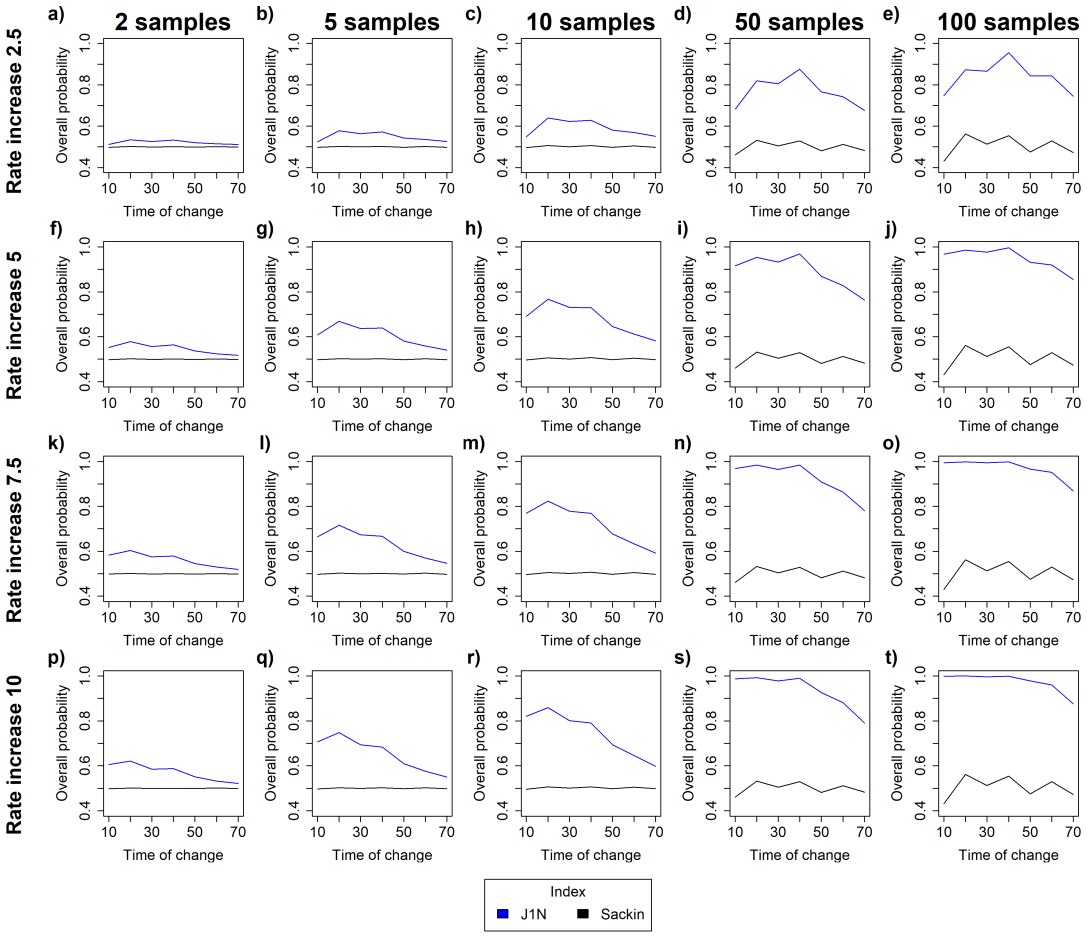

Figure 4.10: Overall probability of our balance index and Sackin's index correctly detecting branching rate heterogeneity under the FTFT model for trees with 20 leaves. The x-axis represents the time of the branching rate change. Vertical columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and horizontal columns correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

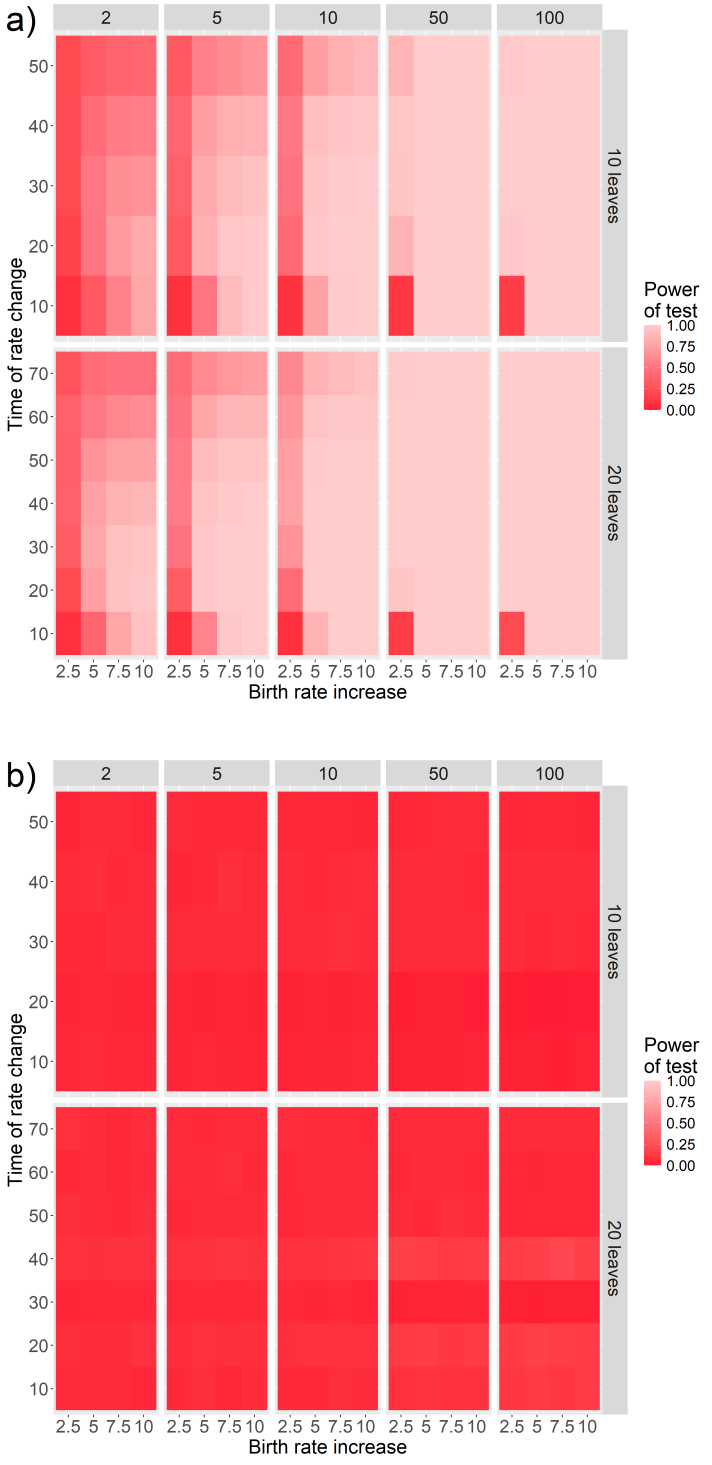

Figure 4.11: Heat maps showing the power of a) our index ${}^1D_S$ or b) Sackin's index I_S to correctly detect branching rate heterogeneity under the FTFT model. The heat maps illustrate the impact of varying the time at which the branching rate across the whole tree changes, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the power.

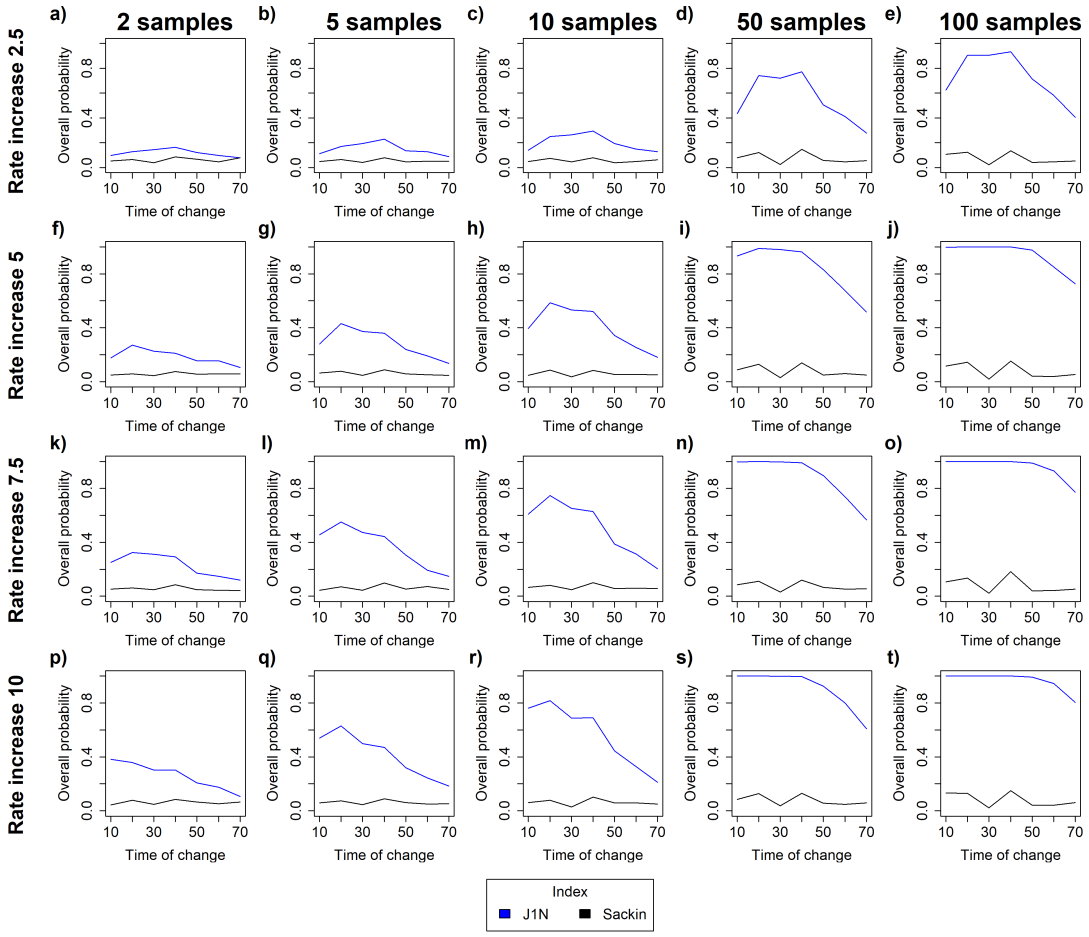

Figure 4.12: Power of our balance index and Sackin's index to correctly detect branching rate heterogeneity under the FTFT model for trees with 20 leaves. The x-axis represents the time of the branching rate change. Vertical columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and rows correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

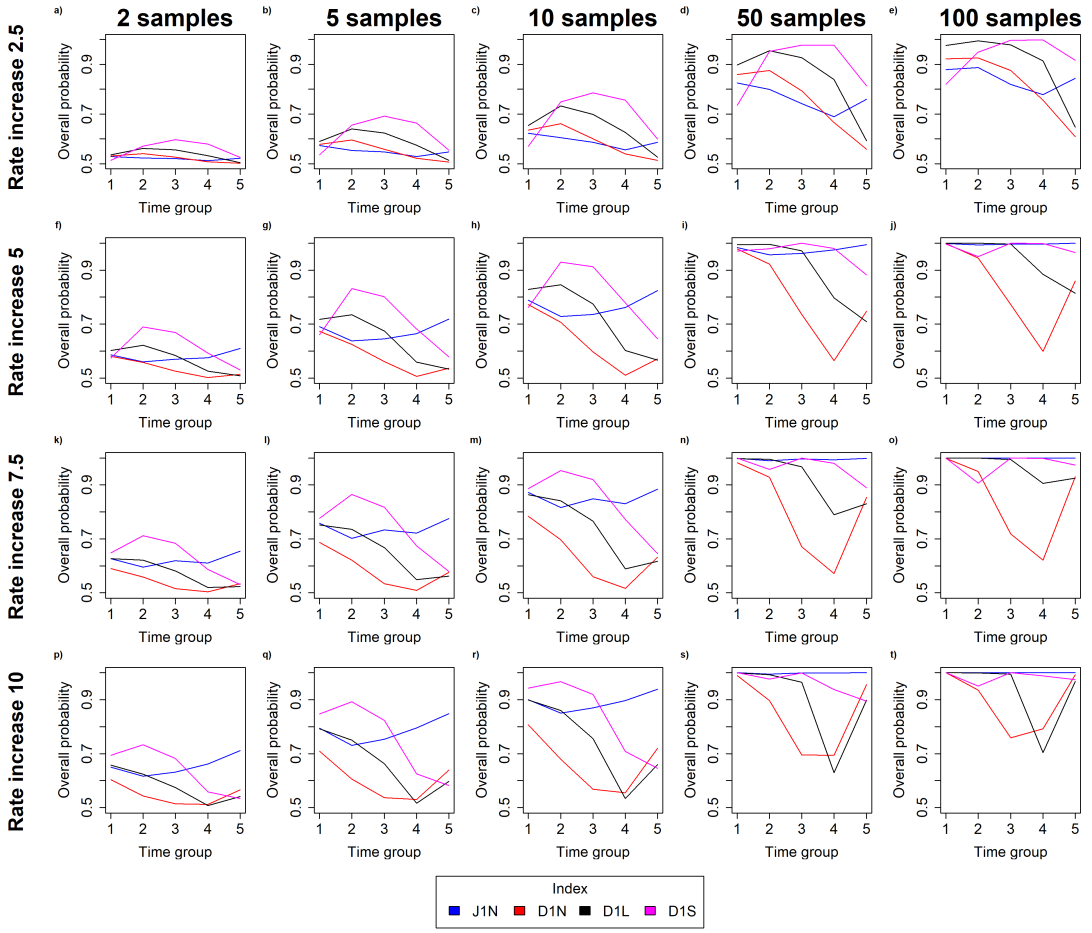

Figure 4.13: Overall probability of our indices correctly detecting between the two types of branching rate heterogeneity for trees with 20 leaves. The x-axis represents the time group. Columns of figures correspond to the number of samples (2, 5, 10, 50, 100, respectively), and rows correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

4.3.4 Our indices distinguish between different methods of branching rate heterogeneity.

Using the Bayesian approach, I found that our indices could distinguish between the two types of branching rate heterogeneity. All indices had a minimum probability of correctly assigning a tree of 50% (see Supplementary Figure 8.11). The pattern of the relative performance of the indices was affected by both the branching rate increase and the number of samples, and the relative performance varied as the time at which the branching rate increase occurs changed (see Figure 4.13). ${}^1D_S$ performed the best most of the time, but its performance followed a concave path, peaking in the middle, and performing worse at the earlier and later times. It is at the later times where ${}^1J_N$ performed better. Given the previous results that ${}^1D_S$ performed well for RTRB and the best for FTFT, where ${}^1J_N$ was the best for RTRB but the worst for FTFT, the results here are unsurprising. After these two indices, ${}^1D_L$ was generally the penultimate performing index and ${}^1D_N$ was the worst.

Sackin's index always outperformed ${}^1J_N$, but Sackin's and ${}^1D_S$ regularly switched between which was best (see Supplementary Figure 8.12). As the number of samples increased, Sackin's index tended to perform increasingly better compared with ${}^1D_S$. Additionally, I observed the same pattern as I did for ${}^1D_S$ and ${}^1J_N$, Sackin's index tended to perform better at early and later time groups, and ${}^1D_S$ performed better in the middle.

4.3.5 Trajectories and correlations

The average trajectories did not differ largely between the Yule model and the two types of branching rate heterogeneity, RTRB and FTFT, but there were a few small differences. RTRB produced the most unbalanced trees as tree size increased, followed by FTFT and then Yule (minimum ${}^1J_N$ reached: 0.81, 0.84, and 0.87, respectively). The diversity-based indices all showed the same pattern, with Yule reaching the largest values as tree size increased, then RTRB, followed by FTFT (maximum ${}^1D_N$: 2.55, 2.38, 2.17, maximum ${}^1D_L$: 3.71, 3.2 and 2.91, and maximum ${}^1D_S$: 6.67, 5.75 and 4.91, respectively).

Although the large linear correlations between index pairs were virtually always maintained, some pairs changed much more than others (see Figures 4.14). ${}^1J_N$ paired with either ${}^1D_S$ or ${}^1D_L$ always had a correlation coefficient ≥ 0.93, the strongest correlation was for RTRB. The pair ${}^1D_L$ and ${}^1D_S$ remained virtually perfectly correlated with coefficient ≥ 0.99. For all pairs with ${}^1D_N$, the linear relationship decreased from Yule to RTRB to FTFT.

For RTRB, hierarchical clustering assigned all indices to the same cluster as was the case for Yule. Although the overall clustering remained the same, the dendrogram showed that the linear relationships were actually quite different, with ${}^1D_N$ now fairly separated from the rest of the indices (see Figure 4.15a). The first principal component explained 97% of the variance and the absolute loadings were ≥ 0.39. Hence, this leads to the same conclusion as for the Yule process. For average trajectories, the indices all measure the same change between tree sizes, and only one needs to be used.

For FTFT, hierarchical clustering assigned ${}^1D_N$ to its own cluster and clustered the other indices together (see Figure 4.15b). The first principal component explained 93% of the variance with absolute loadings ≥ 0.37. The second principal component explained 5% of the variance and ${}^1D_N$ has the largest absolute loading of 0.88 (the next largest is 0.42). Hence, for FTFT, ${}^1D_N$ quantifies a different change in tree shape than the other indices do, and this index should be used along with one of the others.

4.4 Discussion

I have demonstrated that our indices can detect the two types of branching rate heterogeneity investigated here, and the index performance is determined by the type of heterogeneity. Tree balance was the best index in the random-time random-branch model, outperforming the

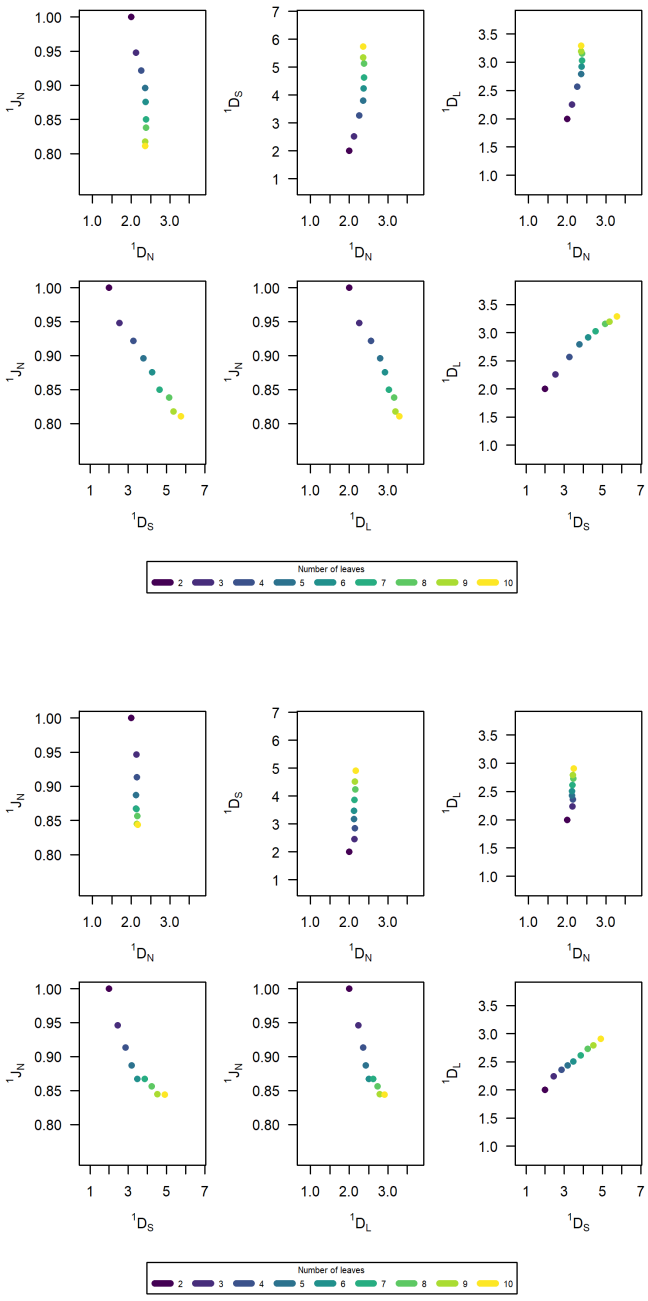

Figure 4.14: Index trajectories averaged over 1000 trees formed by our indices ${}^1J_N$, ${}^1D_N$, ${}^1D_S$ and ${}^1D_L$, for trees either generated by a) RTRB or b) FTFT model, for varying numbers of leaves, n. For both the branching rate increase is 5, and for RTRB the probability of mutating is 0.3, and for FTFT the time of the branching rate change is 0.1.

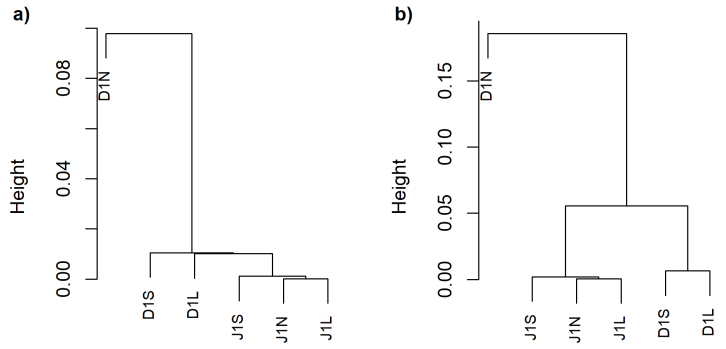

Figure 4.15: Hierarchical clustering dendrogram for our indices for average a) RTRB and b) FTFT trajectories.

widely used Sackin's index. In this model, branches randomly acquire mutations at branching events, increasing their branching rate and ultimately decreasing the balance of their parent and ancestor nodes. In contrast, for the fixed-time full-tree model, our diversity indices outperformed tree balance; in fact, ${}^1D_S$, which measures the effective number of branches, performed the best. The increase in branching rate here leads to the full tree branching faster. This does not affect balance as the branches remain unchanged relative to each other, but it leads to shorter branch lengths and a smaller effective number of branches.

I found that Sackin's index performed well for RTRB and badly for FTFT and in both cases was outperformed by ${}^1J_N$. Sackin's index's ability to detect FTFT remains effectively constant and does not change with parameter values, or even when more samples are used. These results highlight the fundamental differences in these balance and imbalance measures. Sackin's index measures imbalance by counting leaf depth, and when the branching rate across the whole tree changes, the counted leaf depth remains unchanged as the whole tree is now branching faster. This does, however, lead to shorter branch lengths for the tree after the branching rate changes, which is detected by our balance index as it uses branch lengths in its calculation.

Despite the previous result, Sackin's index outperforms ${}^1J_N$ in distinguishing between the types of branching-rate heterogeneity. For FTFT, Sackin's index never achieves an overall probability much higher than that obtained by guessing. This is because this method of branching rate heterogeneity does not change the tree structure compared to Yule in a way that is measurable using Sackin's index. Hence, when Sackin's index is used to distinguish between RTRB and FTFT, this is essentially the same as testing for RTRB against the Yule process, where Sackin's index performs well. ${}^1J_N$, however, becomes more unbalanced (compared with Yule) for both RTRB and FTFT, so it is not as good at telling them apart as it is at distinguishing them from the Yule process.

Bayesian and frequentist approaches lead to very similar results. The Bayesian approach is an appealing and informative method to consider, allowing a probability to be assigned to whether a tree came from a certain distribution or not, rather than having to choose a significance level and then reject or accept the null distribution. However, the Bayesian approach is a

comparison of two distributions. It requires you to not only know the null distribution (or another alternative distribution) but also the distribution from which the tree is drawn. In applications to real data, this is rarely known, and so it is more practical to use the frequentist approach. Hence, knowing that they lead to consistent conclusions is encouraging.

Our results are in agreement with those of Feder and Gao. They found that the two largest contributors to whether deviation from the Yule process could be detected were tree size and the size of the heterogeneity. I find that the size of the heterogeneity is the biggest factor and that there were slight differences between tree size, with detection being better for bigger trees. I likely did not observe as strong an effect due to tree size as they did, as I used trees of size 10 and 20, where they used trees of sizes 50, 250, 1250 and 6250. A benefit to the tree sizes I investigated is that many empirical trees that are of interest tend to be smaller and around this order of magnitude. For example, the phylogenetic trees I consider in chapter 6 range from 5 to 81 leaves and the cancer clone trees I consider in chapter 7 range from 1 to 10 leaves. Additionally, I would expect that it is harder to detect branching rate heterogeneity in smaller trees than in larger ones. Hence, if I can detect heterogeneity in smaller trees, this naturally carries over to larger trees.

The models I used here are deliberately abstract and are not intended to be special cases of any existing models; however, there are similarities with some models. The RTRB model introduced random and irreversible changes in the branching rate to lineages, and can be viewed as a simplified pure-birth analogue of state-dependent speciation models (e.g. [106, 107]). The FTFT model introduces a global time-dependent shift in the branching rate and is related to time-segmented or birth-death-shift models (e.g. [108]), but in a pure-birth setting. Although there do exist analogies between the models defined here and others in the literature, they were chosen as they are a simple setting for studying the ability of our tree shape indices to detect branching rate heterogeneity.

The rate-change factors I considered correspond to medium to strong branching rate heterogeneity. The models I used here are abstract and describe differences in branching rates rather than corresponding directly to biological parameters, particularly in the context of cancer. Despite this, it is still useful to consider analogous parameters to assess the biological relevance of the values I chose and the regimes they represent. In cancer evolution, studies that infer clone growth rates or fitnesses tend to find relative differences in the range of a few to tens of [109, 110], however, multi-fold differences are occasionally observed [111]. In macroevolution, inferred speciation and diversification rates vary largely across clades and lineages, and global shifts of several-fold are commonly reported [99, 112, 113]. Hence, the parameters I considered are much more relevant to the type of heterogeneity observed in macroevolution rather than cancer.

An important direction of future work is to investigate weaker regimes of heterogeneity that are more applicable to cancer. In addition, the models considered in this chapter could be extended in several natural ways. One such extension is to allow the branching rate to change more than once, better representing the biology in both cancer and macroevolution. Furthermore, while the underlying process in both models considered here is Yule process, this could be extended

to include death. More broadly, future work could move beyond non-spatial branching models to more generative models where trees and rate heterogeneity emerge from the underlying dynamics. In cancer, this includes agent-based models in which mutations affect cellular birth rates and naturally generate clone sizes. In macroevolution, analogous extensions include trait or ecology-dependent diversification models, in which quantities such as population sizes may be available. Studying the behaviour of the indices under such models would not only allow the full properties of our indices to be exploited, but would also provide a more biologically grounded assessment of signal in tree shape.

In conclusion, this chapter shows that, in the regimes considered, our indices can distinguish between the Yule process and models with branching rate heterogeneity, as well as between the two models. It also demonstrates that the different methods of introducing heterogeneity lead to different indices performing the best, highlighting that aspects of tree shape beyond balance are informative, and in some cases, more sensitive. Finally, this chapter shows that when distinguishing between the Yule process and models, our tree balance index consistently outperforms Sackin's index. Taken together, these results demonstrate that our indices are capable of detecting signals in tree shape, that this ability is not restricted to tree balance, and that they typically outperform widely used alternatives.

Chapter 5

Mode of evolution for non-spatial and spatial models

5.1 Introduction

Cancer has been considered an evolutionary process for a long time where cell-intrinsic changes and interactions with the tumour microenvironment lead to selection and facilitate adaptation [8, 68]. Cell-intrinsic changes include genetic and epigenetic changes. In cancer, only a small number of genetic changes are actively selected, known as 'driver' events; the majority have no effect on fitness, known as 'passenger' events [8]. Selective pressures in the microenvironment include the immune system, pH changes, treatment, nutrient depravation and spatial structure [68]. Current cancer research aims to characterise this evolutionary process as it has important applications clinically, from enabling patient-specific prognoses to informing how biopsies should be taken [7, 57].

Tumour evolution is associated with distinct evolutionary modes [8, 68], and studying these modes provides insight into the underlying evolutionary processes. However, how studies define the mode of evolution changes frequently, and this affects the applications and importance of the results (e.g. [13, 26, 59, 68], see section 1.1 for a more detailed review of the term). Two of the most common definitions are the resulting pattern that is observed or the underlying process. Since the historical explosion of the availability of phylogenetic data, phylogenetic trees have been a crucial way in which the mode of evolution (MOE) is studied in macroevolution [44]. Despite this, in cancer research, studying the MOE in cancer research using trees is an underutilised method.

Tumour architecture, or equivalently spatial structure, has been shown to determine the mode of tumour evolution in certain cases, where the mode is the underlying process [57]. Scott et al. observed that tree indices are affected by spatial structure and behave differently in spatial and non-spatial models [56]. In addition, Lewinsohn et al. found that boundary-driven growth

leaves patterns in trees not seen in unrestricted growth [88]. Hence, different spatial structures lead to noticeable differences in phylogenetic tree shape, and these differences can be seen using tree indices. As a result, where spatial structure determines the mode of evolution, these differences can also be used to distinguish between the modes.

In fact, when the MOE is defined as the underlying process, both tumour trajectories and their end-points have been observed to cluster according to MOE in certain index spaces. Noble et al. found that the tree topology at the end of the tumour simulation cluster according to MOE when summarised using clonal diversity and mean driver mutations per cell [57]. Manojlovic subsequently extended this work and found that tumour trajectories - their path in some index space as time increases - also cluster by MOE [114]. Additionally, both Scott et al. and Manojlovic observed that tree shape can detect relevant model parameters [56, 114]. The work by Manojlovic used a variety of indices, including four of our new tree shape indices ${}^1J_N$, ${}^1D_N$, ${}^1D_L$ and ${}^1D_S$, but only investigated index pairs. Moreover, due to the non-trivial challenge of comparing individual trajectory runs, no formal statistical analyses were performed.

Defining the MOE as the pattern, Edrisi et al. found that a recursive neural network could distinguish between modes when given a tree and a genotype matrix [7]. To the best of my knowledge, this is the only study that has explicitly investigated distinguishing between modes of evolution defined in this way using trees. However, a similar method that has been developed is an unsupervised learning method called `oncotree2vec` [115]. It is designed to embed tumour mutation trees into a low-dimensional space where trees with similar evolutionary patterns are close together, and dissimilar trees are far apart [115]. When applied to a small cancer data set, it clustered the trees into four, pattern-based, evolutionary modes. Both of these studies show that tree structure does capture evolutionary signals, in particular, the modes of evolution, that can then be detected using machine learning methods. Still, no studies have looked at whether these evolutionary modes can be detected using tree-shape indices.

The questions this chapter aims to answer are: do trees cluster in index space according to the MOE when defined as the pattern? For both definitions, what index space results in the best clustering? Do the results vary depending on how we define the MOE? And, where applicable, can we detect model parameters?

Here, I consider two ways of defining the mode of evolution: the resulting pattern and the underlying process. When defining the mode as the resulting pattern, I consider four modes of evolution: linear, branching, neutral and punctuated. When defining the mode as the underlying process, I also consider four modes, or more specifically, growth patterns: non-spatial, gland fission, invasive glandular and boundary. Noble et al. showed that tumour architecture determined the mode of evolution and that these growth patterns corresponded with evolutionary modes: selective sweeps, progressive diversification, branching and effectively almost neutral, respectively [57]. Since these spatial structures reliably map to specific modes of evolution, and really are the 'underlying process' here, I will use them when referring to the MOE as the process in this work.

In a real-world setting, tumour phylogenies are usually inferred from sequencing data taken

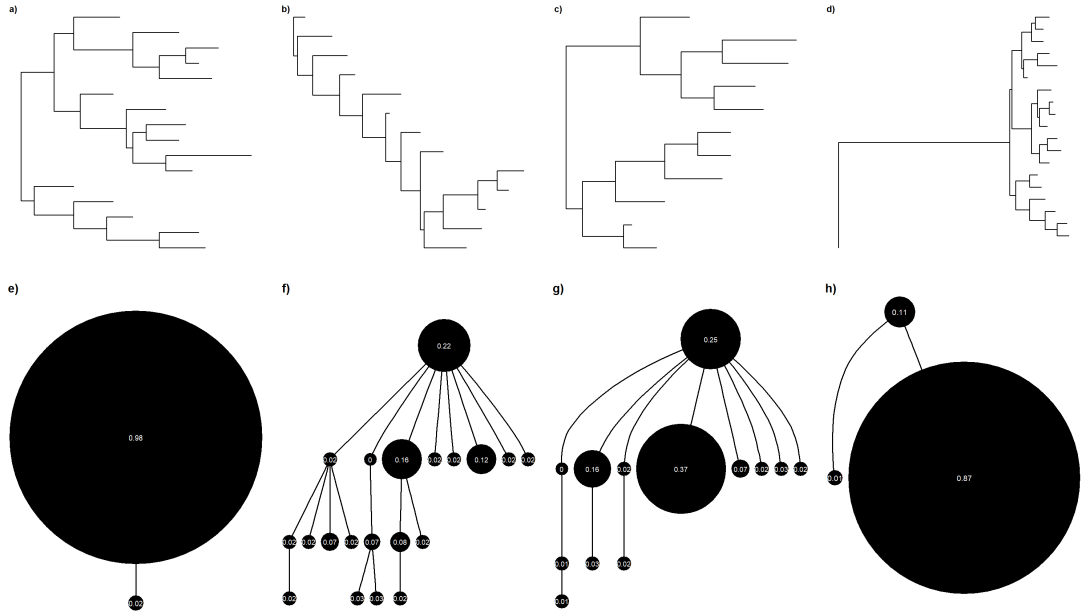

Figure 5.1: Trees showing each evolutionary mode for both definitions of the mode of evolution. For modes as the pattern, a) branching evolution, b) linear evolution, c) neutral evolution and d) punctuated evolution, simulated using the tree-generating method from MoTERNN [7]. The branch lengths are proportional to the number of mutations. For modes as the process, e) boundary growth, f) glandular growth, g) invasive glandular growth and h) non-spatial growth, where tree data was obtained from [114]. Node sizes are relative population sizes; branch lengths shown are arbitrary.

from a tumour at a single point in time. The current clinical methods, typically a biopsy, mean that obtaining the data necessary to construct a tumour's evolutionary trajectory through time would not be ethical. Therefore, I decided to consider a single time point in the tumour's trajectory as Noble et al. did [57], as this will have greater clinical application. Additionally, it has been observed that patterns in tree indices are maintained throughout tumour growth [56]. Hence, it is likely that what I observed at one time point will be applicable to others.

I investigate whether we can detect the mode of evolution in both cases using our tree-shape indices and whether we can detect model parameters for the underlying process. I extend on previous work by using rigorous tree-shape indices and studying more than two dimensions. I find that when the MOE is defined as the pattern, we can distinguish between the modes, but not when the MOE is defined as the underlying process. Additionally, I can only detect model parameters in specific cases.

5.2 Methods

5.2.1 The pattern data

For the MOE defined as the pattern we observe, we simulated trees using the method and the associated code available from MoTERNN [7]. This code allows trees from four modes of evolution to be simulated: linear evolution (LE), branching evolution (BE), neutral evolution (NE) and punctuated evolution (PE). We generated 4000 trees, 1000 for each mode, with 8 to 25 leaves (note that a tree with 8 leaves is the smallest this method can generate). These trees have branch lengths, which represent the number of mutations, but no node sizes, so we assigned leaves a size of 1 and internal nodes a size of 0. Examples of simulated trees are shown in Figures 5.1a-d.

This method uses a modified version of the beta-splitting model (BSM) to simulate the tree branching structure, and then assigns branch lengths by sampling from a Poisson distribution. The modification of the BSM results in a process that is not truly random. The method of manipulating a random process is essential for simulating trees when the MOE is defined as the pattern we observe. We want the scheme for each mode to generate trees that correctly fit the expected pattern. If the method of tree generation is truly random, then stochasticity will lead to instances in which a scheme for a given mode produces a tree that appears to belong to another mode.

For each MOE, the BSM is used in a specific scheme to create trees with the desired shape. The BSM generates trees with n leaves using the following steps [7]:

1 **Sampling generative sequences:** sample a sequence of $n-1$ independent and identically distributed (i.i.d) random variables, $B = (b_1, ..., b_{n-1})$, from the beta distribution $\mathcal{B}(\alpha+1, \beta+1)$, where $\alpha > 0$ and $\beta > 0$. Then, sample a sequence of i.i.d. random variables $U = (u_1, ..., u_{n-1})$ where $u_i \sim \mathcal{U}(0,1)$.

2 **Initialisation:** create the root node of the tree and assign it the interval $[0,1]$. Next split the root node into left and right child nodes and assign them the intervals $[0, b_1]$ and $[b_1, 1]$ respectively.

3 **Iteration:** in iteration i, where the initialisation step is the first and second iterations, from the current leaves, find which leaf interval $[x, y]$ contains u_i. Select this leaf, split it into left and right children nodes, and assign $[x, x+(y-x)b_i]$ and $[x+(y-x)b_i, y]$ to the children respectively. Stop at iteration $n-1$.

The parameters α and β control the shape of the tree, with equal values having a high probability of generating balanced topologies, and the difference between the parameters determining the imbalance [7]. To generate their trees, Edrisi et al. use BSM in balanced and unbalanced modes with $(\alpha, \beta) = (10^4, 10^4)$ and $(\alpha, \beta) = (10^4, 10^{-4})$ for the former and latter modes respectively.

		Driver rate								
		1×10^{-4}			1×10^{-5}			1×10^{-6}		
		Selection coefficient								
Mode	Total	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
B	450	50	50	50	50	50	50	50	50	50
G	450	50	50	50	50	50	50	50	50	50
IG	420	46	45	44	48	45	48	50	49	45
NS	445	50	50	46	50	50	49	50	50	50

Table 5.1: Number of trees for each mode and set of parameter values for the data from [114].

In linear evolution (LE), the tree is assumed to grow in two events, before and after the emergence of a dominant clone. The first event is two-thirds of the $n-1$ speciations, and the tree is simulated in the 'unbalanced' mode of BSM. The second event is the final third of the speciations, and the tree is simulated in the 'balanced' mode. The branch lengths are then generated by sampling from a Poisson distribution with a mean of 5 for each branch; this is also how branch lengths are generated for all other schemes unless stated otherwise.

In branching evolution (BE), the tree is generated using two steps. First, the total number of clones, C, is determined by sampling uniformly from the set $\{2, 3, 4\}$. The tree then grows using the balanced mode of BSM until it has C leaves. Next, the number of leaves descendant from each clone is sampled, where the sum of the number of leaves descendant from all clones has to equal n. The number of leaves here is sampled from a multinomial distribution with n trials and C categories, where each category has a success probability of $\frac{1}{C}$. Then, each clonal lineage is generated using the LE scheme.

In neutral evolution (NE), there is no selection or dominant clones so the entire tree is simulated using BSM's balanced mode.

For punctuated evolution (PE), first the number of clones, C, is determined by sampling uniformly from $\{2, 3\}$. The number of leaves belonging to each clone is determined using the same method used in BE. Then, the clonal lineages are simulated separately using BSM's balanced mode. The branch length of the long root branch is determined by sampling from a Poisson distribution with $\lambda = 100$, and the clonal branch lengths are generated using the method used in the LE scheme.

5.2.2 The process data

For the MOE defined as the process, I used data previously generated in [114]. The data were generated using `warlock`, an automated computational workflow that encapsulates a deme-based oncology model called `demon` [116]. `demon` is an agent-based model with parameters that control for spatial arrangement, mutation rates, selective advantage and migration, allowing for the modelling of different modes of evolution. The simulations are two-dimensional, and this is not an issue as the cell population is approximated to undergo stochastic isotropic growth (the probability of the tumour expanding in all directions is equal) [114]. The model tracks the population of genotypes and demes (a deme is a well-mixed patch of cells) rather than individual

cells, and cells that belong to the same deme-genotype intersection are considered identical [116]. Cell events are simulated using the Gillespie algorithm, first a random deme is selected, then a cell type, then an event type and finally, for cancer cells, a genotype [116]. The model allows for two types of mutations, driver and passenger mutations [114, 116]. Driver mutations are associated with a selective advantage, including increased proliferation or migration rates, and passenger mutations are typically neutral but can be deleterious [114].

As previously stated, we will refer to the process here by the spatial structure of the tumour growth. Four different spatial configurations of tumour evolution were simulated, these are [114]:

1. **Glandular:** the initial tumour cell forms the first deme, and when it reaches a given size, it splits into two daughter demes. This process is repeated until the final population size is achieved.
2. **Invasive glandular:** the initial tumour cell forms the first deme, and the daughter demes are formed by individual cells migrating away from the parent deme. This process us repeated until the final population size is achieved.
3. **Non-spatial:** there is no spatial structure imposed on the tumour and the cells are well-mixed.
4. **Boundary:** cells are only allowed to divide if they are not fully surrounded by other cells.

The data contains 450 trees for both the boundary and glandular growth modes, 420 trees for invasive glandular growth and 445 trees for non-spatial growth (see Figures 5.1e-h for example trees for each mode). For each mode, three different values were used for both model parameters, driver rate and selection coefficient (see Table 5.1 for a further breakdown of the data, and see Supplementary Figures 8.16, 8.17, 8.18 and 8.19 for examples of trees broken down by parameter values). Due to the differences in parameter values, the simulations were run with the aim to reach a target population of 10^6 cells between 500 and 1000 cell cycle events or generations [114].

The trees generated by the models for the pattern and the process are very different. The method of simulating the pattern trees does not assign or generate any node sizes; hence, I assign the leaves to have a size of 1 and the internal nodes to have a size of 0. The process trees' node sizes are the population sizes. The branch lengths in the pattern trees correspond to the number of mutations, and in the process trees, they correspond to time.

5.2.3 Silhouette analysis

I scaled the data using a robust scaling method as the indices are not typically normally distributed. Let X be a set of unscaled data, for $x \in X$, the robust scaling is given by,

$$\frac{x - \text{median}(X)}{\text{IQR}(X)},$$

where $\text{IQR}(X)$ is the interquartile range of X. We then calculated the correlation of every index pair, and for any pair with a correlation greater than 0.95, one of the pair was removed. This gave us a reduced dataset. We then carried out silhouette analysis on both the full set of index pairs and the reduced dataset.

5.2.4 Random Forest

I fit random forests to the reduced datasets and used 5-fold cross-validation. Cross-validation is a resampling method used to evaluate machine learning models on small data samples [117]. The data is split into k groups, and for each group, the group is taken as the test data set and the remaining groups are taken as the training data set. The model is then fit on the training set and evaluated on the test data set. Then the performance of the model is summarised using the individual model scores.

The performance scores that I use in this research are, accuracy, mean decrease accuracy, mean decrease Gini and Cohen's kappa. The accuracy I state is the mean proportion of correctly classified observations across 5 cross validations [117]. The mean decrease accuracy is the average reduction in predictive accuracy when the values of a given predictor are permuted [118]. It reflects the predictive performance of the variable, and larger decreases indicate more important predictors. The mean decrease Gini is the average total reduction in Gini impurity attributable to splits on a given predictor across all trees in a random forest [118]. It reflects how often and effectively a variable is used for splitting, and larger values indicate a greater contribution to class separation. Cohen's kappa quantifies classification performance by measuring the agreement between the predicted and true labels after accounting for agreement expected by chance [119].

5.2.5 Alternative indices

For the mode as the pattern, I use two alternative indices to evaluate how other indices perform in comparison to ours. As a measure of tree balance/imbalance, I use the normalised Sackin's index $I_{S,norm}$, defined in section 1.3. As a measure of diversity, I use Chao et al.'s phylogenetic diversity with $q = 1$ [66]. In terms of the notation defined in section 2, the index is defined as [59, 66],

$${}^1\overline{D} = \exp\left(-\frac{1}{\overline{h}} \sum_{i \in I} h_i \sum_{b \in B_i} s_b \log s_b\right),$$

where $\overline{h}$ is the effective height of the tree and shown below. I is the set of intervals, h_i is the height of the interval i, and s_b is the size of branch b.

For the mode as the process, I evaluate how two alternative indices, which have previously been found to cluster the modes we consider here well [57], perform for this data set. The first index is clonal diversity D, defined as [57],

$$D = \frac{1}{\sum_i p_i},$$

where p_i is the frequency of the ith combination of driver mutations. The second index is the mean number of driver mutations per cell, n, which is equivalent to the effective tree height $\overline{h}$, previously outlined in section 2.2 and first defined in [59], given by,

$$\overline{h} = \sum_{b \in B} s_b l_b,$$

where l_b is the length of branch b.

5.3 Results

5.3.1 Initial analysis

Individual indices and index pairs

In this work, the very first thing I did was for each definition of evolutionary mode, inspect the distribution of index values when split by mode, and investigate how index pairs clustered the different modes.

For both definitions, the distribution of the index values showed that the indices have varying power to distinguish between the modes of evolution, and the indices also distinguished between different modes of evolution (see Supplementary Figures 8.13 and 8.14). For example, for the pattern definition, ${}^1D_S$ separated PE well, but could not differentiate between the other modes, whereas ${}^1J_S$ did not separate PE nearly as well, but could differentiate between LE and BE or NE. From the distributions, it appeared the indices are better able to separate the modes when defined as the pattern rather than the process.

Plotting every index pair, I observed a similar pattern. Some index pairs exhibited more distinct clusters for some modes of evolution compared with others. For the pattern, PE consistently occupied a distinct cluster, while the other modes varied. Additionally, the indices seemed to cluster the modes defined as the pattern better than the process.

In plots with index pair ${}^1D_L$ and ${}^1D_N$, there is a diagonal boundary formed by the line ${}^1D_L = {}^1D_N$ that can be observed. This is due to the known inequality ${}^1D_L \geq {}^1D_N$ (Proposition 2 in [59]). From these plots, further diagonal boundaries can be seen for pairs ${}^1D_S$ and ${}^1D_N$ and

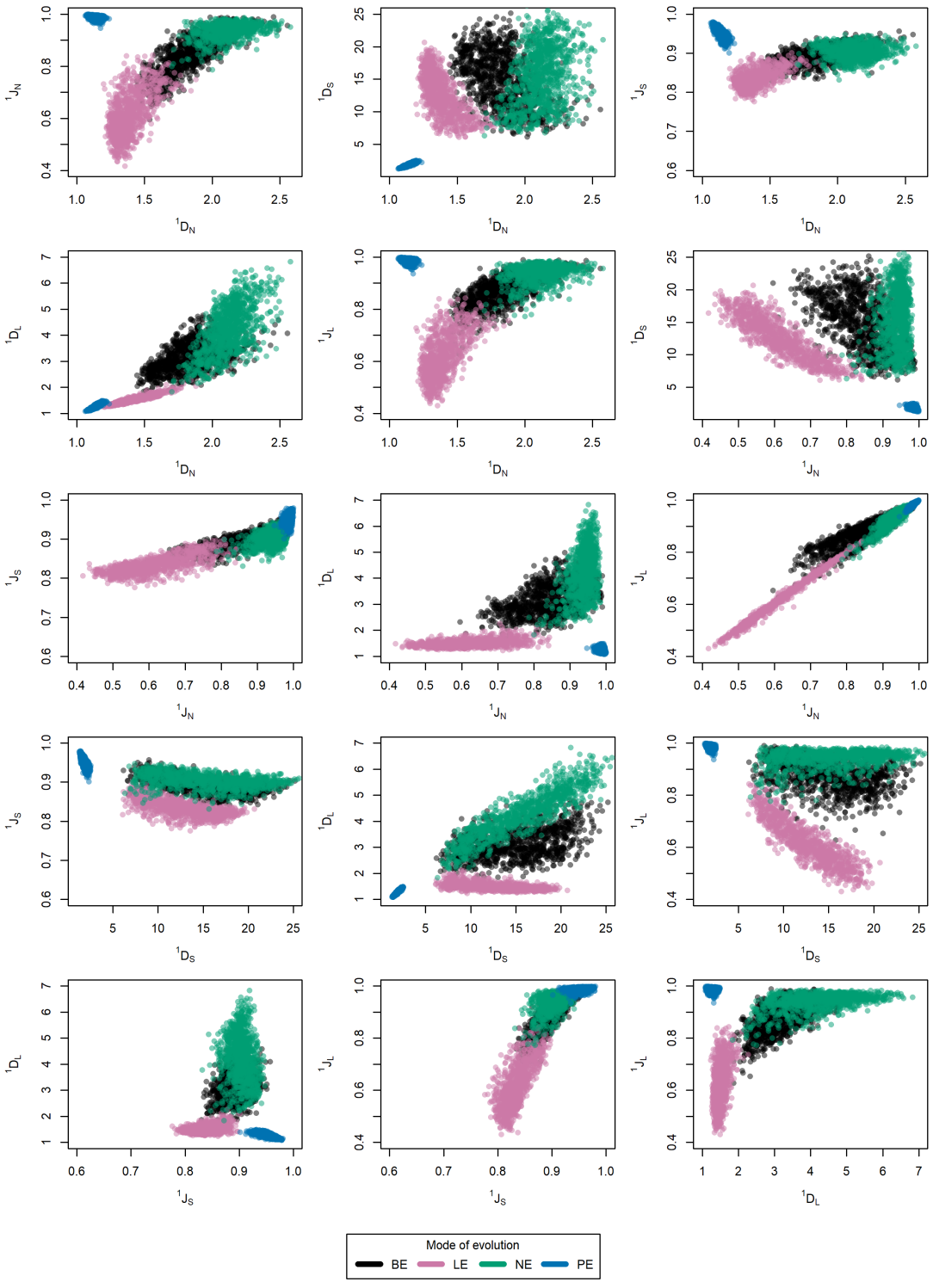

Figure 5.2: Pairs of index values where the colour shows the mode of evolution, which is defined as the pattern (BE: branching evolution, LE: linear evolution, NE: neutral evolution, PE: punctuated evolution).

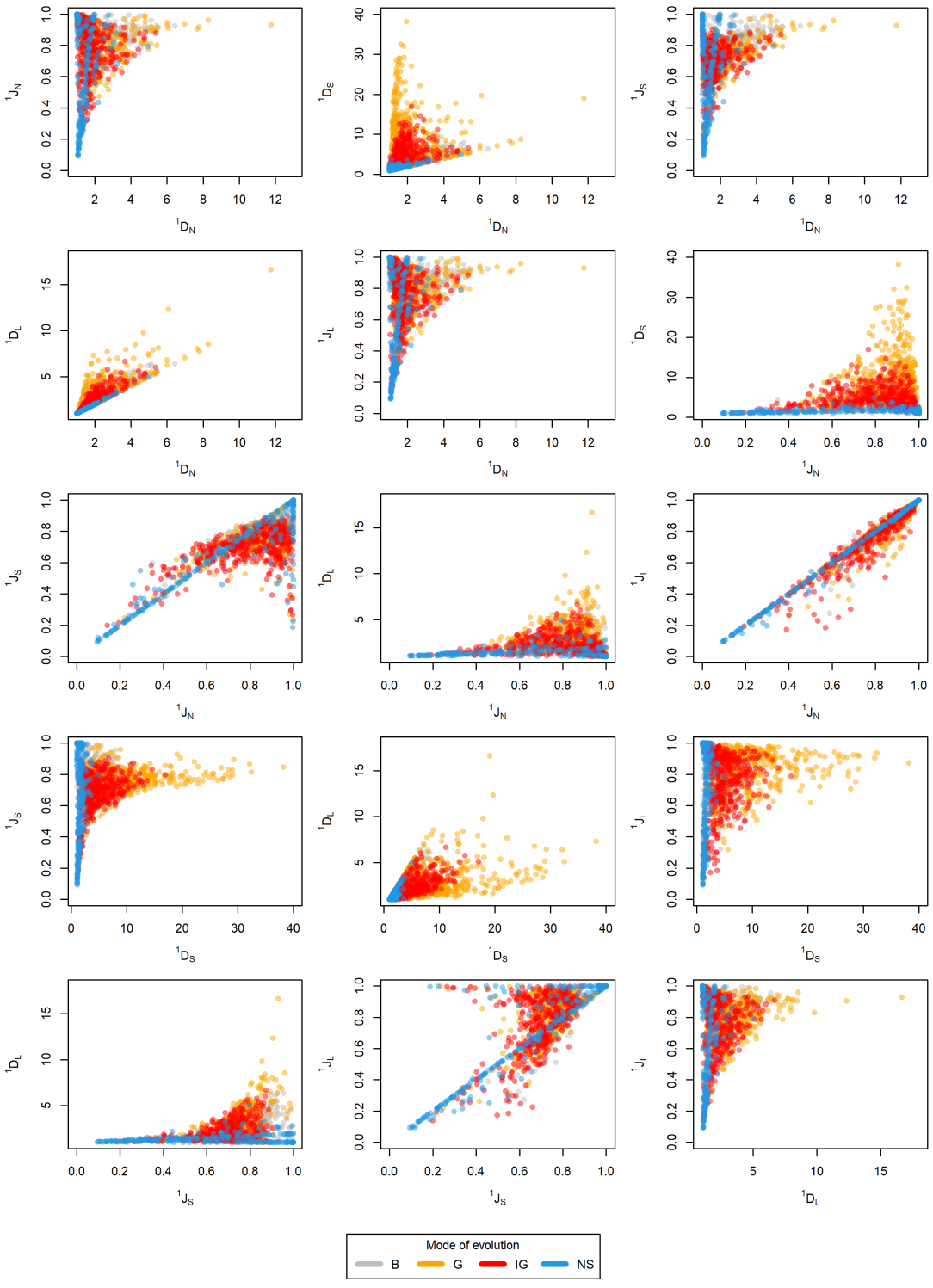

Figure 5.3: Pairs of index values where the colour shows the mode of evolution, where the mode of evolution is the process. (B: boundary growth, G: glandular growth, IG: invasive glandular growth, NS: non-spatial growth).

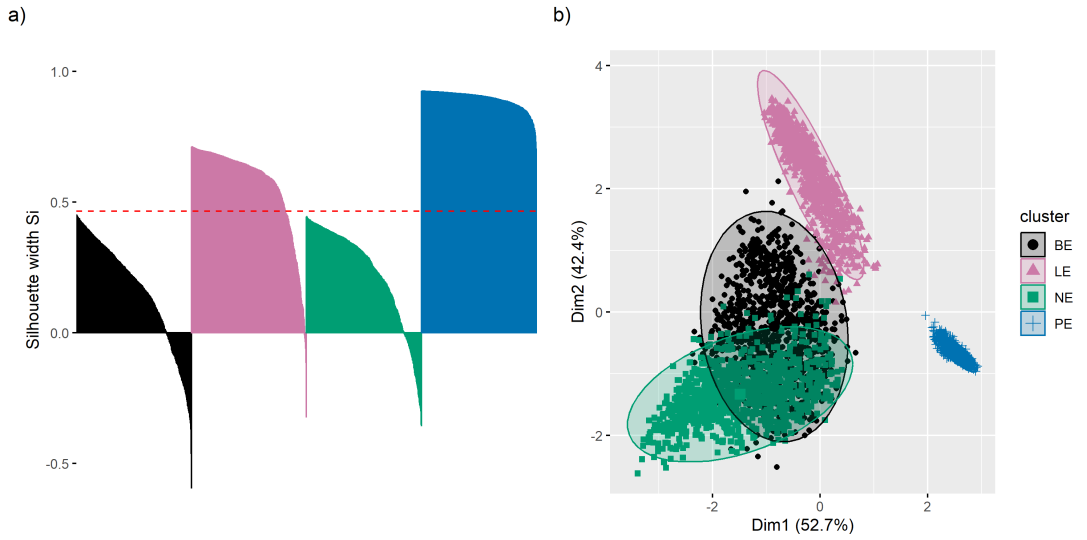

Figure 5.4: Modes of evolution, defined as the pattern, clustered by the reduced index set (${}^1D_N$, ${}^1D_L$, ${}^1D_S$, ${}^1J_N$, ${}^1J_S$). a) Silhouette width for each tree, dashed line is the average silhouette width of 0.47. b) Clustering of trees based on the reduced index set, visualised in 2D using PCA. True evolutionary modes are colour-coded, with ellipses showing the spread of each group. (BE: branching evolution, LE: linear evolution, NE: neutral evolution, PE: punctuated evolution).

${}^1D_S$ and ${}^1D_L$. These suggest that further inequalities for our indices exist that we have yet to identify and prove. By property 0.7 from [59], trees that are close to the line ${}^1D_L = {}^1D_N$ or ${}^1J_L = {}^1J_N$ are close to being piecewise star trees (a piecewise star tree is a 'tree that can be divided into transverse intervals such that, within each interval, all the non-zero-sized branches are attached to a common node' [59]). Additionally, by property 0.8 in [59], any trees that are close to the equality ${}^1D_S = {}^1D_N = {}^1D_L$ or ${}^1J_S = {}^1J_N = {}^1J_L$ are close to being star trees.

5.3.2 The pattern we observe

Here, I investigate whether the modes of evolution, defined as the pattern, cluster based on our tree shape indices, which is equivalent to saying, can we detect the mode of evolution of a tree using tree shape indices? First, I look at how index pairs cluster and at their correlations. Next, I look at how the modes cluster based on silhouette widths using 5 indices. Then, I investigate how a random forest trained on the same 5 indices to predict the evolutionary mode performs. Finally, using our indices, I also train a random forest on the index pair that clustered the data the best. Lastly, I look at how a pair of alternative tree shape indices perform.

Can we detect the mode of evolution?

The index pairs ${}^1D_N$ with ${}^1J_N$ or ${}^1J_L$ formed the most distinct clusters (both had an average silhouette width of 0.54), closely followed by ${}^1D_L$ and ${}^1J_N$ or ${}^1J_L$ (average silhouette widths of 0.53 and 0.52 respectively). The index pairs ${}^1D_S$ and ${}^1J_S$ formed the least distinct clusters

(average silhouette width of 0.33, see Figure 5.2). I find that ${}^1J_N$ and ${}^1J_L$ were highly correlated, with a correlation coefficient of 0.99. Therefore, the two best-performing pairs, ${}^1D_N$ with either of these, were virtually equivalent and separating data in the same way. This analysis also identified any redundancies, which were ${}^1J_L$ and ${}^1J_N$ here, as they were highly correlated. I formed the reduced dataset by removing ${}^1J_L$. It does not necessarily matter which index of the pair is used; however, as ${}^1J_N$ is our balance index, a critical and commonly used tree shape metric, I decided to choose this one out of the pair.

Silhouette analysis on the reduced data set (${}^1D_N$, ${}^1D_L$, ${}^1D_S$, ${}^1J_N$, ${}^1J_S$) showed that the modes formed fairly distinct clusters with a silhouette width of 0.47. PE formed the most distinct cluster with an average silhouette width of 0.89, and all trees had a positive silhouette width (see Figure 5.4). The next most distinct cluster was LE, which had an average silhouette width of 0.57, and a negligible number of trees had a negative silhouette width (0.87%). NE and BE were the two least distinct clusters, with average silhouette widths of 0.26 and 0.16, respectively, and BE had the largest number of trees with negative silhouette width (0.46% and 37%, respectively). In summary, overall, the reduced data set clustered the modes well, but the performance of clustering did change between modes, with some clustering better than others.

The random forest trained on the reduced dataset had an accuracy of 93%. The index ${}^1D_L$ had the largest mean decrease accuracy and mean decrease Gini (see Supplementary Figures 8.15a-b). Hence, it is an important index as it has a large impact on both model accuracy and class splitting. ${}^1J_N$ and ${}^1D_N$ are the next two most important, with the former having a slightly higher mean decrease accuracy and the latter having a slightly higher mean decrease Gini. ${}^1D_S$ and ${}^1J_S$ are the last two, with ${}^1D_S$ having a slightly higher mean decrease accuracy and a much higher mean decrease Gini. We tested the random forest on a set of unseen simulated data that contained 500 trees for each mode. The model again had an accuracy of 93% and Cohen's kappa of 0.90. Unsurprisingly, PE is perfectly classified, LE is classified with very few errors and BE and NE are sometimes confused with each other. The random forest trained on our 5 indices performs very well with good accuracy and a high kappa value, showing the model frequently correctly predicts the mode from the index values, and it is much better than what would be achieved by chance.

I also trained the random forest on the best-performing index pair, in terms of silhouette width, ${}^1D_N$ and ${}^1J_N$. This model had an accuracy of 88%, less than the model trained on the full reduced set. Given how important ${}^1D_L$ was for both model accuracy and class splitting, this is unsurprising. Training a random forest on ${}^1D_L$ and ${}^1J_N$ resulted in a model accuracy of 0.91. This is an improvement on the previous pair, but not quite as good as the full reduced set. Hence, very good performance can be achieved with only two of our indices, but extending this to the full reduced set leads to the best results.

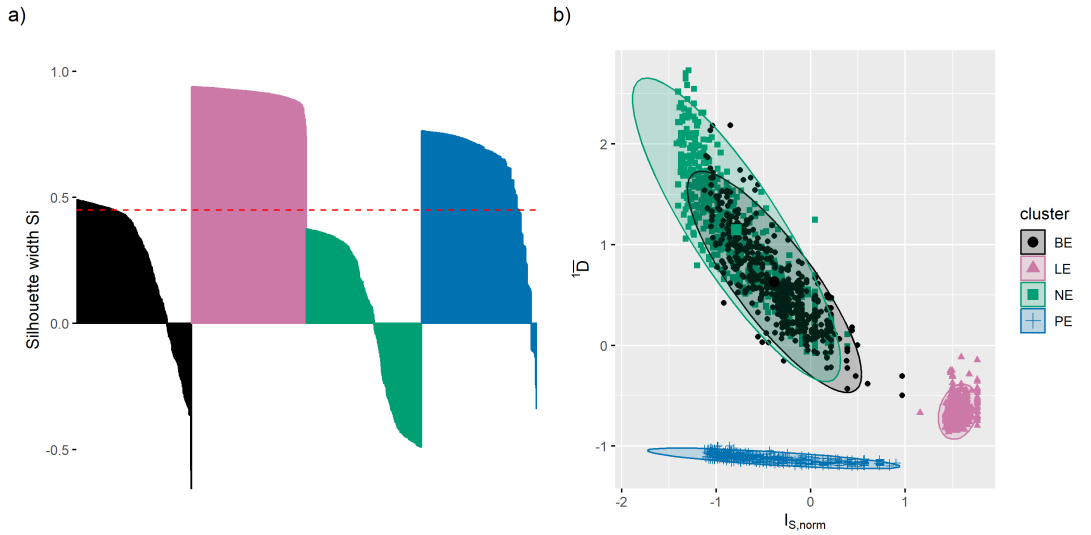

Figure 5.5: Modes of evolution, defined as the pattern, clustered based on alternative indices $I_{S,norm}$ and ${}^1\overline{D}$. a) Silhouette width for each tree, dashed line is the average silhouette width of 0.45. b) Clustering of trees based on $I_{S,norm}$ and ${}^1\overline{D}$. True evolutionary modes are colour-coded, with ellipses showing the spread of each group. (BE: branching evolution, LE: linear evolution, NE: neutral evolution, PE: punctuated evolution).

Alternative indices

The normalised Sackin's index, $I_{S,norm}$, and Chao et al.'s phylogenetic diversity with $q = 1$, ${}^1\overline{D}$, clustered the trees with an average silhouette width of 0.45 (see Figure 5.5). LE formed the most distinct cluster with an average silhouette width of 0.91, and all trees had positive silhouette widths. PE was the next most distinct cluster with an average silhouette width of 0.62, and a small number of trees had a negative silhouette width. BE was the second-worst cluster with an average silhouette width of 0.21 and some trees with a negative silhouette width. NE was the worst cluster with an average silhouette width of 0.02 and a large number of trees with a negative silhouette width. A random forest trained using these indices had an accuracy of 86%. Hence, these alternative indices perform worse than our equivalent index pair, ${}^1J_N$ and ${}^1D_L$, with worse clustering and a lower accuracy for the random forest model.

5.3.3 The process

In the first section here, I investigate whether the modes of evolution, defined as the process, cluster based on our tree shape indices. I first look at index pairs. Then, I explore the clustering using silhouette widths for 5 of our indices. Next, I investigate how the model parameters affect this clustering. Finally, I examine how a random forest trained on the 5 indices performs.

In the second section, I briefly evaluate the clustering based on two indices, D and n, which were previously found to cluster the modes we are considering here well. I also explore how our diversity index ${}^1D_L$ and n cluster the data.

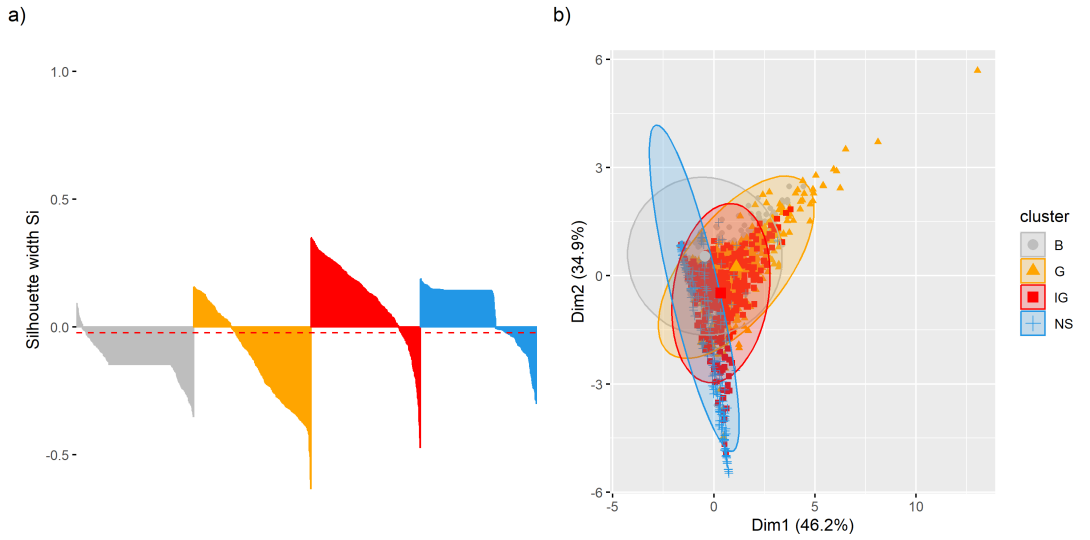

Figure 5.6: Modes of evolution, defined as the process, clustered based on the reduced index set (${}^1D_N$, ${}^1D_L$, ${}^1D_S$, ${}^1J_N$, ${}^1J_S$). a) Silhouette width for each tree, dashed line is the average silhouette width of -0.023. b) Clustering of trees based on the reduced index set, visualised in 2D using PCA. True evolutionary modes are colour-coded, with ellipses showing the spread of each group. (B: boundary growth, G: glandular growth, IG: invasive glandular growth, NS: non-spatial growth).

	Driver rate		
Selection coefficient	1×10^{-4}	1×10^{-5}	1×10^{-6}
0.05	0.289	0.122	0.009
0.1	0.218	0.080	0.049
0.2	0.102	0.054	0.052

Table 5.2: Average silhouette width of the data when it is split by the model parameter values and clustered by the mode of evolution.

In the third section, I investigate whether we can detect the model parameters using our indices, which is equivalent to whether the trees generated by certain parameter values cluster based on tree shape indices. I do this by first ignoring the mode and looking at the clustering across all trees simply based on each parameter value individually and as a pair, again using silhouette widths and training a random forest. Next, I perform the same analysis but within each mode of evolution individually.

Can we detect the mode of evolution?

No index pair formed distinct clusters. Every pair had a negative silhouette width, ranging from -0.10 to -0.019. ${}^1J_N$ and ${}^1J_L$ were again highly correlated (correlation coefficient 0.98, see Figure 5.6). To avoid redundancy, I followed what I previously did and chose to use ${}^1J_N$ and removed ${}^1J_L$ from our set of indices, forming my reduced dataset.

Using the reduced dataset, silhouette analysis demonstrated that these modes did not form distinct clusters in the reduced index space (silhouette width -0.023, see Figure 5.6). Invasive

glandular and non-spatial modes performed similarly with average silhouette widths of 0.12 and 0.07, respectively. The other two modes also performed similarly but much worse than the previous two, with average silhouette widths of -0.13 for glandular growth and -0.14 for boundary growth. Invasive glandular growth had a relatively small number of trees with negative silhouette width (18.7%). Then it was non-spatial growth with around a third having a negative silhouette width (34.9%). Glandular growth performed much worse, with the majority of trees having a negative silhouette width (67.8%). Boundary growth performed badly with almost all trees having a negative silhouette width (94.9%). Hence, overall, the clustering was poor and depended upon the mode, with some performing better than others.

I find that the performance of the clustering depended upon the model parameter values, with a bigger driver rate and smaller selection coefficient resulting in better clustering (see Table 5.2). Additionally, for the parameter values used here, the driver rate had a bigger effect on clustering than the selection coefficient.

The random forest trained on the reduced dataset had an accuracy of 65%. The index ${}^1D_S$ had the largest mean decrease accuracy and mean decrease Gini, showing it was a crucial index in the model as it has a large impact on both model accuracy and class splitting (see Supplementary Figures 8.15c-d). The other indices were clustered at much lower values of both scores, and the order of these indices was varied, but ${}^1J_N$ had the lowest value in both cases. Overall, the model accuracy was bordering on good, but it performed much better than simply clustering by the index values.

Alternative indices

I found that with this data set, clonal diversity, D, and the mean driver mutations per cell, n, clustered the modes in a similar index space as was found previously with these modes and indices (see Figure 5.7a). However, this data set did not show distinct clusters (see Supplementary Figure 8.21). The modes had an average silhouette width of -0.067, but with a large range across widths for individual modes. Boundary and non-spatial growth showed weak clustering with widths of 0.16 and 0.18, respectively. There was then a large decrease, with invasive-glandular growth having a width of -0.20, and again another large decrease as glandular growth had a width of -0.42. Hence, although the data does occupy the index space found previously, I found overall weak clustering and a big variation in how well the different modes cluster.

Using ${}^1D_L$ in place of D resulted in the indices clustering in a different index space (see Figure 5.7b). Overall, the average clustering remained almost unchanged, with an average silhouette width of -0.056; however, the silhouette widths of individual modes are quite different. Boundary growth had a negative average silhouette width of -0.27, and non-spatial growth had a much larger average silhouette width of 0.48. Glandular and invasive glandular growth both increased slightly, but both still had negative silhouette widths of -0.31 and -0.13, respectively. Therefore, although the overall average silhouette width changes minimally, the clustering of the modes using ${}^1D_L$ appears to be worse than with D, with three out of the four modes

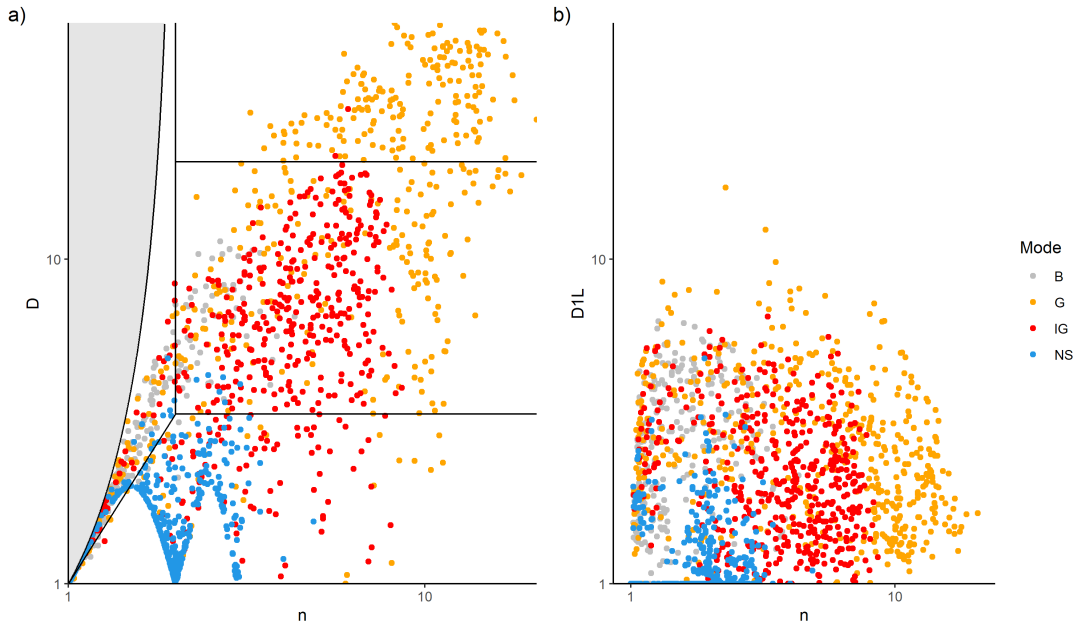

Figure 5.7: a) Clonal diversity, D, against mean number of drivers per cell, n. Lines correspond to clustering regions identified in [57]. It is impossible to construct trees for points within the shaded region. b) Diversity, ${}^1D_L$, against mean number of drivers per cell. (B: boundary growth, G: glandular growth, IG: invasive glandular growth, NS: non-spatial growth).

having a negative average silhouette width.

Can we detect the model parameters?

Not taking into account the different modes, I found that neither the driver mutation rate nor the selection coefficient clustered the data well, with average silhouette widths of -0.02 and -0.052, respectively (see Figures 5.8a-d). The clustering was even worse when I used both parameters together to cluster (silhouette width of -0.11, see Figures 5.8e and f). I observed similar results with the random forest. As outcome variables, neither the driver rate nor the selection coefficient led to a model with good accuracy; however, here there was a noticeable difference between the two, with driver rate being better (accuracy 58.2% and 46.1% respectively). Additionally, the model performed much worse when I used both parameters (accuracy 28.6%).

Splitting by the mode and using the driver mutation to cluster, I found that overall, the clustering was weak. Boundary growth exhibited the best clustering, followed by glandular growth, then invasive glandular growth, and finally non-spatial growth (average silhouette widths: 0.218, 0.066, -0.006, and -0.032, respectively, see Supplementary Figure 8.23). Splitting by the mode and using the selection coefficient to cluster led to even worse results. The average silhouette widths were very similar for all modes and were all negative (average silhouette widths: boundary growth -0.064, glandular growth -0.019, invasive glandular growth -0.001 and non-spatial growth -0.073, see Supplementary Figure 8.24). As before, when I used both

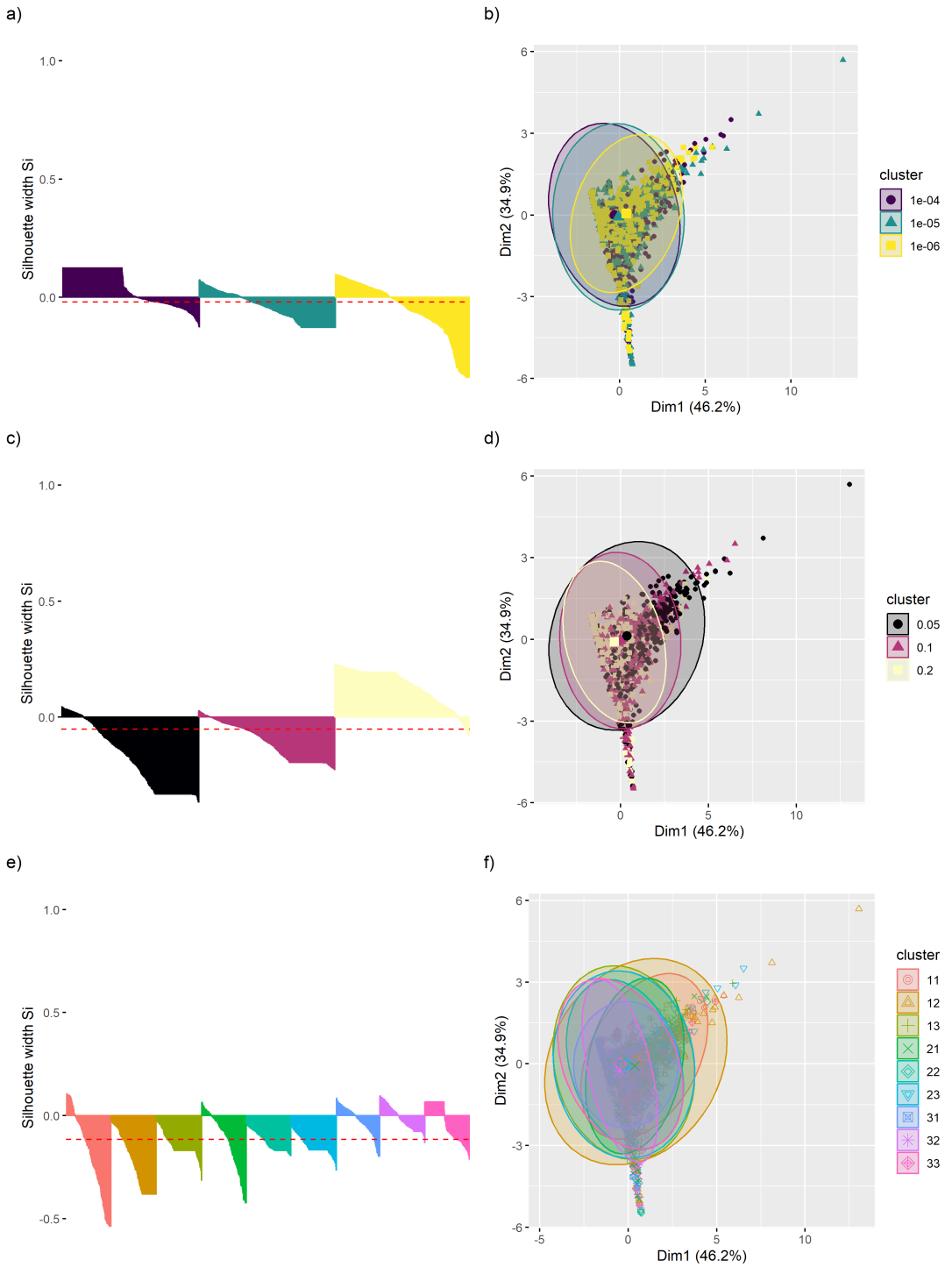

Figure 5.8: Clustering based on a-b) driver mutation rate, c-d) selection coefficient and e-f) both. For e-f, the first number corresponds to the position of the driver mutation in $(1 \times 10^{-4}, 1 \times 10^{-5}, 1 \times 10^{-6})$, and the second number corresponds to the position of the selection coefficient in (0.05, 0.1, 0.2). For example, 11 is equivalent to parameter pair $(1 \times 10^{-4}, 0.05)$. Average silhouette widths a-b) -0.02, c-d) -0.052 and e-f) -0.11.

parameters together, the clustering was poorer than either alone (average silhouette widths: boundary growth -0.345, glandular growth -0.035, invasive glandular growth -0.113 and non-spatial growth -0.313).

Random forest models on the data split by mode maintained the same pattern within each mode as I saw across all modes. The models had the best accuracy when the driver rate was the outcome, and using both parameters resulted in the worst accuracy. Glandular growth had the best accuracy across all parameters (driver rate: 86.9%, selection coefficient: 67.6% and both: 61.1%), then boundary and invasive glandular growth performed similarly (driver rate: 74.9% and 73.1%, selection coefficient: 41.6% and 56.0% and both: 34.4% and 38.6%, respectively), and non-spatial growth had the lowest accuracy (driver rate: 58.2%, selection coefficient: 33.0% and both: 24.5%).

5.4 Discussion

Our reduced index set achieves reasonable clustering overall when the MOE is defined as the pattern. BE and NE perform the worst among the modes; however, this is not an unexpected outcome and will likely always be the case when defining the MOE in this way. NE is a special case of BE in which there is no selection advantage, so BE includes trees that, by chance, appear to be NE. Despite our reduced index set performing well, the best average silhouette width is achieved with index pairs ${}^1D_N$ and either ${}^1J_N$ or ${}^1J_L$. The extra dimensions provided by the other indices dilute the signal relative to the noise. Besides ${}^1J_N$ and ${}^1J_L$, no other indices are highly correlated, but some are moderately correlated, and including these can distort distances. Additionally, extra indices add noise and unrelated variation. These can both lead to unfavourable changes to the geometry of the data, weakening the signal and increasing noise. We do not see the same for the random forest; it achieves the best accuracy on the full reduced index set.

For the MOE as the process, our reduced index set achieves poor clustering. Glandular and boundary growth are particularly bad, with the majority of their trees having a negative silhouette width. In contrast, the random forest performs well.

Crucially, our ability to detect the MOE based on our index values depends upon how we define the MOE. Additionally, the methodological approach results in the greatest disparity between the definitions. For the pattern, the reduced index set clusters the data reasonably, but for the process, the clustering is extremely poor. Conversely, the random forest performs well for the pattern and is reasonable for the process.

Clustering by model parameters, our indices perform poorly both across and within modes. On the other hand, random forests can detect model parameters in certain cases. Across and within modes, the random forests obtain much better accuracy for the driver rate compared with the selection coefficient. However, across modes, the accuracy when the driver rate is the outcome is not particularly good. For all other than non-spatial growth, good accuracy is achieved when considering the driver rate as the outcome. Hence, from the model parameters,

we can only detect the driver rate, and even then, this can only be done with reliable accuracy within certain modes.

These findings show that when the evolutionary mode is defined as the resulting pattern, different modes generate distinct tree shapes that our indices can capture, and they capture this better than commonly used alternative indices. However, when the mode is defined as the underlying process, the indices are much less able to distinguish between them. In cancer, detecting the mode of evolution in both cases is important because they have different clinical implications. Additionally, demonstrating that the evolutionary mode can be inferred from tree shape opens the door to the clinical use of such indices.

A key limitation here is the tree-generating model from MoTERNN, as it does not produce node sizes and assigns branch lengths after a tree's branching structure is simulated. Evaluating indices under more biologically driven models is essential to improve the clinical relevance of our results. Hence, an important direction for future work is studying the pattern through more biologically grounded models. Future work could also further investigate the ability of the indices to detect the process within the agent-based model used to generate the data analysed in this chapter, by exploring a wider range of simulations, parameter regimes and model settings than was done here.

In conclusion, this chapter has shown that the indices can distinguish between different modes of evolution when these are defined in terms of the resulting pattern, and our indices outperform commonly used alternatives. These results demonstrate that tree shape contains meaningful signals for characterising evolutionary dynamics that our indices can detect, providing a foundation for further methodological and clinical development.

Chapter 6

How tree balance varies with taxonomic level

6.1 Introduction

This chapter is based on work from a manuscript in preparation, co-authored with my supervisor. The analyses, simulations, and results presented here were carried out by me, with my supervisor providing editorial feedback and guidance.

Tree balance has been related to variation in speciation and extinction rates, and this rate variation is known to be associated with important macroevolutionary phenomena [79]. Hence, the analysis of phylogenetic tree shape promises to reveal patterns of variation in diversification rates associated with these important processes [79, 80]. A fundamental question that then arises is how phylogenetic tree shape varies with taxonomic level.

Katzourakis et al. tackled this question by analysing a genus-level phylogeny of hoverflies using a node imbalance index, I, and found that nodes were more balanced when the species richness of genera was used than when it was not [120]. Purvis et al. identified a flaw in I, which they corrected before reanalysing the phylogeny [121]. The results agreed with Katzourakis et al., motivating the study of other trees to determine the pervasiveness of this pattern. In a follow-up study [51], Purvis and Agapow aimed to answer two questions: Does using the species richness of leaves (number of species belonging to the taxonomic group of the leaves) make a significant difference to tree imbalance? And does the imbalance of nodes, with a given number of descendant leaves, depend on whether the leaves are species or higher taxa? In a meta-analysis, they found a significant effect of taxonomic level, such that trees were more balanced when the species richness of higher taxa was considered. They also found that for nodes of a given size, the mean node imbalance was significantly higher for nodes with higher taxa as leaves than for nodes with species as leaves.

The conclusions of these prior studies crucially depend on the nature of their chosen imbalance index. The index I was chosen primarily because it accounts for node sizes and it can be applied to trees with non-bifurcating nodes, yet the latter is only because such nodes are simply disregarded. Another reason for preferring this index was that its expectation under the Yule process is simple and independent of tree size. These are two examples of properties of this index that may lead to limitations that were not previously investigated and are examined further in this chapter.

This issue is not unique to the index used by Purvis and Agapow. Many commonly used tree balance indices suffer from limitations, as outlined in chapter 1. The previous study by Purvis and Agapow was conducted using the best methods available at the time. However, with the recent development of new methods that improve upon the previous ones, I revisit their questions.

Here, I repeat the analysis done by Purvis and Agapow on a new data set using their imbalance index and our balance index. I find that the results depend on which index you use. Using our index, I find that the meta-analysis shows that the use of species richness leads to more unbalanced trees. I also find that the balance of a node depends on whether the leaves are higher taxa or species, and species-level nodes are more unbalanced. Therefore, there is a significant difference in tree and node balance between the taxonomic level of analysis, and crucially, our results differ from Purvis and Agapow's in the direction of which taxonomic level is more balanced. I also investigate the use of the null model of an expectation of 0.5 for the imbalance indices under the Yule process, finding it to be a poor null model that leads to large type 1 errors. Finally, I find that the balance of the data compared with what is expected under the Yule process depends on which nodes are considered, and for this data, the difference is never significant.

6.2 Methods

Table 6.1: Comparison of the data used by Purvis and Agapow and the data used in this study from the Open Tree of Life.

Clade	Purvis & Agapow		Open Tree of Life	
	Number of leaves	Level of leaves	Number of leaves	Level of leaves
Vertebrates				
Anura	20	Families	41	Families
Salamanders	10	Families	10	Families
Ciconiidae	6	Genera	6	Genera
Trogonidae	6	Genera	5	Genera

Continued on next page

Table 6.1: Comparison of the data used by Purvis and Agapow and the data used in this study from the Open Tree of Life. (Continued)

Chaenopsidae	13	Genera	10	Genera
Bathyergidae	5	Genera	6	Genera
Odontoceti	33	Genera	30	Genera
Phyrnosomatinae	9	Genera	9	Genera
Pleurodira	16	Genera	15	Genera
Squamata	19	Families	36	Families
Plants				
Anthemidae	108	Genera	81	Genera
Barnadesioideae	9	Genera	9	Genera
Calenduleae	8	Genera	7	Genera
Chrozophoreae	11	Genera	8	Genera
Cunoniaceae	26	Genera	21	Genera
Cyclanthaceae	12	Genera	9	Genera
Hyoscyameae	7	Genera	7	Genera
Inuleae	35	Genera	48	Genera
Liabeae	14	Genera	19	Genera
Lythraceae	31	Genera	23	Genera
Nymphaeales	8	Genera	9	Genera
Podalyrieae	9	Genera	6	Genera
Restionaceae	55	Genera	50	Genera
Arthropods				
Augochlorini	33	Genera	27	Genera
Dolichoderinae	21	Genera	25	Genera
Helioconiiti	10	Genera	9	Genera

6.2.1 Data

I do not use the original data for the main analysis or comparison as I was unable to access a large number of the sources. The study by Purvis and Agapow is over 20 years old, with many of the sources for their trees dating back over 25 years. However, in regards to the main analysis, using new data is ultimately necessary. In the decades since this study, significant advancements have been made in the methods used to infer relationships and generate trees for species and other taxonomic groups. Historically done by looking at morphological features,

molecular systematics now leads the way [122]. This has led to changes in relationships between groups and further refinement for many trees. Therefore, using new data ensures I am using the most up-to-date versions of the trees for our desired taxonomic groups.

The new tree data and associated species richness were obtained from the Open Tree of Life [123, 124]. The Open Tree of Life is a summary tree that currently utilises 1362 phylogenies in the latest synthetic tree. I obtained subtrees with leaves of a given rank following the method outlined in [125]. The trees do not contain branch length information. I had two criteria for the trees: they must be the latest version available from the Open Tree of Life, and the trees must only contain tips of the same rank. As previous studies have, I do not create any restrictions based on tree completeness, where a complete tree contains all of the living members of the given group. The trees I used are not all complete and also vary in their degree of completeness (they are missing different amounts of members). When leaves are randomly omitted, incompleteness does not affect balance; it neither makes trees more or less balanced. This is not the case when the omission is not random [51, 90]. Hence, our trees not being complete may affect the results I obtain. Nevertheless, I aim to investigate the Tree of Life in its current form, so I do not exclude these trees, and I include an analysis of the effect of completeness on balance.

I used the same or similar analysis as Purvis and Agapow, using new data and our balance index. The imbalance indices used by Purvis and Agapow necessitated the use of non-parametric methods suitable for non-normal and non-symmetric data (Supplementary Figure 8.25). Similarly, the distributions of our balance index for this study, as is typical of our balance index, were also non-normal and non-symmetric (Supplementary Figure 8.26). Additionally, using the same (or similar) analysis allowed us to replicate their work with their imbalance indices and our new data to determine if their results changed, enabling a fair comparison. It is essential to note that the original study used an imbalance index, where a value of 1 is maximally unbalanced and 0 is maximally balanced, whereas I use a balance index, so 1 is maximally balanced and 0 is maximally unbalanced. All analysis was done in R 4.4.1.

6.2.2 Imbalance

The imbalance indices used by Purvis and Agapow use I as defined by Fusco and Cronk and modified by Purvis et al. [121, 126]. I measures the imbalance of every bifurcating node with at least four descendant leaves. It is defined as,

$$I = \frac{B - m}{S - m - 1},$$

where S is the number of descendant leaves, B is the size of the larger descendant node and m is the minimum value that B could take.

Two levels of analysis will be considered, 'higher-level' and 'species-level'. In the higher-level analysis, the leaves are treated as single leaves (Figure 6.1a). In the species-level analysis, the leaves are considered polytomies whose outdegree is equal to the number of species in the given

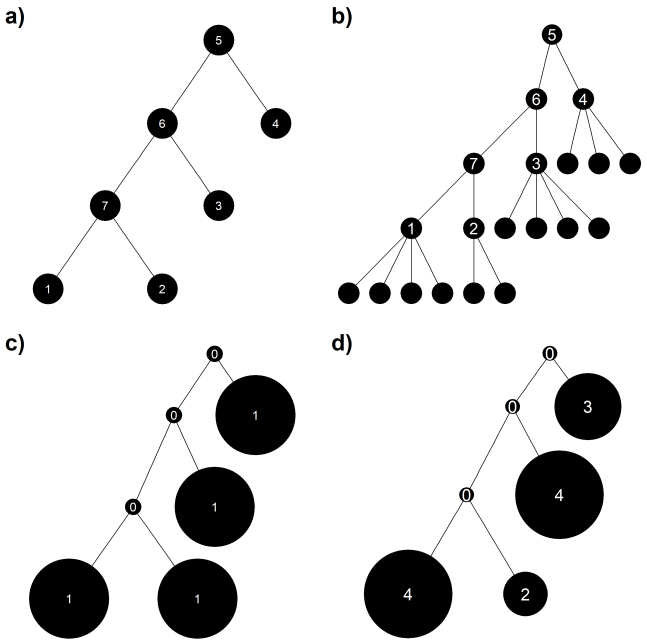

Figure 6.1: The meaning of the level of analysis for each index. For a) and b), the numbers name the nodes, but for c) and d), they represent node sizes. a) shows the tree I am considering, the tips are some given higher taxa where leaves 1, 2, 3 and 4 contain 4, 2, 4 and 3 species respectively. For the imbalance index, a) shows how the tree is considered for the higher-level analysis, the tree is considered as is and as the tips are higher taxa, the node sizes are the number of higher taxa. b) shows how it is considered for the species-level analysis, each leaf is considered to be a polytomy with outdegree equal to the number of species in that given taxonomic group, and sizes are then the number of species. For the balance index, c) shows the higher-level analysis, internal nodes are set to have size zero, and leaves are set to have size one. d) shows the species-level analysis where I set leaf sizes to be equal to the number of species in the given taxonomic group, and internal nodes have size zero. In this case, it is the total descendant abundance of a node that is equivalent to either the number of higher taxa or species.

taxonomic group (Figure 6.1b). Since the S for the imbalance index is based on the number of descendant leaves, this method allows species richness to be used without altering the tree shape in terms of the index. Moreover, S is the number of higher taxa of a given rank for the higher-level analysis and species in the species-level analysis.

Purvis et al. defined the following correction to I, denoted the node balance index I_w [126]. Each node imbalance score I has a weight w given by the following,

$$w = 1, \text{ if } S \text{ is odd,}$$
$$w = \frac{S-1}{S}, \text{ if } S \text{ is even and } I > 0,$$
$$w = \frac{2(S-1)}{S}, \text{ if } S \text{ is even, and } I = 0.$$

Given a set I of node balance scores and a set W of their associated weights, then for $I \in$ I and $w \in W$ each node has an I_w score given by,

$$I_w = \frac{wI}{\frac{1}{|W|}\sum_{w \in W} w}. \tag{6.1}$$

For Yule trees of a given size, the expectation of I_w across all nodes is 0.5. I_w can be calculated across any set of nodes, but in this study, the node balance scores are calculated across the set of nodes in each tree.

Finally, Purvis and Agapow defined the weighted mean of I, which I denote as $\overline{I}_w$. The weighted mean can also be taken over nodes from a set of trees or an individual tree. $\overline{I}_w$ has an expectation of 0.5 across all nodes under the Yule process. For this analysis, $\overline{I}_w$ is used as a tree imbalance index, and for a given tree it is calculated over all the eligible nodes in that tree. It should be noted that, when calculated on the same set of nodes, the mean of the set of I_w values is equivalent to the $\overline{I}_w$ of these nodes.

6.2.3 Balance

The general definition of our tree balance index was provided in Chapter 2, section 2.4. Here, I will outline our balance index in the simplified case where there is no branch length information, equivalent to all branch lengths being equal for our index. This is identical to the previous version of our balance index J^1, which does not use branch lengths, except for how they treat linear nodes [46]. ${}^1J_N$ assigns linear nodes to be maximally balanced and J^1 assigns them to be maximally unbalanced.

Let $V(T)$ be the set of internal nodes of tree T, $f(v)$ be the size of node v and T_i denote the subtree rooted at node i. Then S_i is the magnitude of T_i and S_i^* is the magnitude of T_i, excluding i, given by,

$$S_i = \sum_{v \in V(T_i)} f(v), \quad S_i^* = \sum_{v \in V(T_i),\, v \neq i} f(v).$$

Note that the total descendant abundance of a node i is equal to the magnitude of the subtree

T_i not including i, that is S_i^*.

Then for some node i, letting $C(i)$ be the set of direct descendants from i, and the number of descendants be given by $n = |C(i)|$, then,

$$ {}^1J = \begin{cases} -\sum_{k \in C(i)} \frac{S_k}{S_i^*} \log_n \frac{S_k}{S_i^*} & \text{if} \quad n > 1, \\ 1 & \text{otherwise,} \end{cases} $$

and our balance index is then,

$$ {}^1J_N = \frac{1}{\sum_{k \in V} S_k^*} \sum_{i \in V} S_i^{*\,1}J. \tag{6.2} $$

This is equivalent to the previous version of our balance index defined in [46], with the only difference being how they are defined for nodes of outdegree 1. For the purpose of this analysis, I then define a node balance index as is done for the I-based indices used by Purvis and Agapow. Given a set of internal nodes V, for $i \in V$, let,

$$ {}^1J_{N,i} = \frac{S_i^{*\,1}J}{\frac{1}{|V|}\sum_{k \in V} S_k^*}, \tag{6.3} $$

and, for a given set of nodes in a tree, the mean of the node balance scores is equivalent to the weighted mean index (tree balance), ${}^1J_N$.

I again consider two levels of analysis, but it is much simpler given the ability of our index to account for node sizes. In the higher-level analysis, I set the leaf sizes to be one and the node sizes to be zero (Figure 6.1c). This means the total descendant abundance of a node is the number of descendant higher taxa. For the species-level analysis, I set leaf sizes to be equal to the number of species in the given taxonomic group and nodes to be size zero (Figure 6.1d). Therefore, the total descendant abundance of a node is the number of descendant species.

For both indices, the tree size is the number of leaves in the tree. Node sizes, however, are counts of leaves for the imbalance index and some defined value for the balance index. For the balance index, there is also the magnitude of a tree and total descendant abundance from a node (the latter is equivalent to the former for the root node), where for a root node k, the tree magnitude is S_k, and the total descendant abundance from a node is defined above. The tree and node sizes of the imbalance index are equivalent to the magnitude of the tree and total descendant abundance from a node for the balance index (see Figure 6.1). In subsequent method sections, when referring to methods with both indices, I will, for brevity, use magnitude and total descendant abundance.

Purvis and Agapow and I did not use either of the widely used Sackin's or Colless' imbalance indices due to certain properties resulting in them being unsuitable for use here. Colless' index can only be applied to bifurcating trees; neither index allows for meaningful comparison between trees with different leaf counts, and both are sensitive to small changes in tree shape [46, 59]. Purvis and Agapow chose their imbalance index for its ability to be used with polytomous nodes, which allowed them to look at a larger sample of trees [51]. Our index

can also be used on trees containing polytomies, but accounts for them more robustly as the I-based indices are simply not calculated on polytomous nodes.

6.2.4 Completeness and tree balance

I performed two one-way analyses of covariances (ANCOVAs) for each index to examine whether the balance of a tree of a given magnitude depends on whether the tree is complete or not. To remove the potential effect of the taxonomic level of analysis on balance, I performed the ANCOVAs on the higher and species-level trees separately. I state the adjusted p-values, which were obtained using the Holm-Bonferroni correction. Following Purvis and Agapow, I define a tree to be complete if no more than 5% of the species of the desired group can not be assigned to a leaf. For species-level analysis, I perform the ANCOVA on trees with magnitude ≤ 500 to remove the significant relationship between magnitude and status; this removes 4 trees. The relationship between magnitude and status is simply an artefact of the data, as the incomplete trees contain more large trees.

6.2.5 Tree balance and taxonomic level

Consistent with the original study, I applied two methods to investigate the effect of taxonomic level on tree balance. First, I analysed the trees individually, performing a randomisation test on the node balance scores to test for differences in balance among taxonomic levels for each tree. This tests for significant differences in the mean of the node balance scores, which is equivalent to the tree balance.

Second, I combined the individual trees in a meta-analysis. To do this, I used a sign test to test whether the median of the differences in tree balance was significantly different from zero. I then used a weighted one-sample t-test to assess whether the mean of these differences was significantly different from zero, weighted by the inverse of their squared standard errors.

6.2.6 Node balance and taxonomic level

For the node analysis, Purvis and Agapow performed a standard ANCOVA. I used robust methods as the node balance and imbalance values for our data did not satisfy the conditions for a standard ANCOVA. They violate both the normality of residuals and the homogeneity of variances conditions, and transforming the data does not resolve this. I performed a robust linear regression to test whether the balance of a node depends on the level of analysis and the node's total descendant abundance. I used the function `lmrob` from the `robustbase` package [127]. I also conducted a robust ANCOVA to test whether the balance of a node depends upon the leaves' taxonomic level for nodes of total descendant abundance 4, 6, 8 and 12. To do this, I used the `ancova` function from the `WRS2` package [128].

The regression allows the overall trend to be observed and the ANCOVA allows the differences

in mean node balance to be observed at desired total descendant abundances. Due to the nature of the data here, the species-level analysis contains many more large nodes than the higher-level analysis. This creates an artificial interaction between taxonomic level and total descendant abundance, so I reduce the data to match the total descendant abundance distribution across taxonomic levels roughly. For both indices, I use only nodes with a total descendant abundance less than or equal to 20, and for ${}^1J_{N,i}$, I also restrict to nodes of total descendant abundance greater than or equal to 2.

6.2.7 Validation of I-based indices null model

To validate whether the null model of 0.5 for the I-based indices under the Yule process is suitable, I perform type 1 error analysis. I aim to test whether the mean node imbalance (I_w) equals 0.5, which is equivalent to testing whether the weighted mean $\overline{I}_w$ equals 0.5. I use a bootstrap method to carry out the analysis.

For each tree size and number of trees, I generate Yule trees and calculate the I_w values (taking into account all nodes present for this given number of trees). I then perform bootstrap sampling on this set of I_w values, sampling with replacement the given number of trees and calculating the mean 5000 times, generating a distribution of means. Each value is then shifted by the difference between the distribution mean and the null mean of 0.5, forcing the bootstrapped samples to have a mean of 0.5. Then, I perform a two-sided test, observing the proportion of bootstrapped means that are as extreme as or more extreme than the observed mean. I then carry out this test 1000 times, and the proportion of rejections is the type 1 error.

I also validate whether the tree balance index $\overline{I}_w$ is a suitable index to test for deviations from the Yule process using an empirical null model. The method in brief is, generate the empirical null model, compare samples to the null and combine the p-values using Fisher's method and then report the proportion of significant values. For the empirical null, I generate 10000 trees for each tree size and calculate the $\overline{I}_w$ values. I then generate the null Fisher distribution for the given sample size using a two-tailed test for the individual p-values, and from this distribution I calculate the 95% threshold. I then perform 1000 trials, generating samples of $\overline{I}_w$ for as many Yule trees as the sample size, comparing them to the empirical null and calculating the Fisher statistic. The Fisher statistic is then compared with the threshold I determined and rejected if it lies in this region. The proportion of rejections is then the type 1 error.

6.2.8 Comparison to the Yule process

As the trees here do not have branch lengths, for fair comparison, I consider Yule trees without branch lengths. For our balance index, ${}^1J_N$, this is equivalent to considering all branch lengths to be equal, and this index is then equivalent to the previous version of our tree balance index J^1 defined in [46]. The indices differ on how they treat linear nodes, but as the Yule process

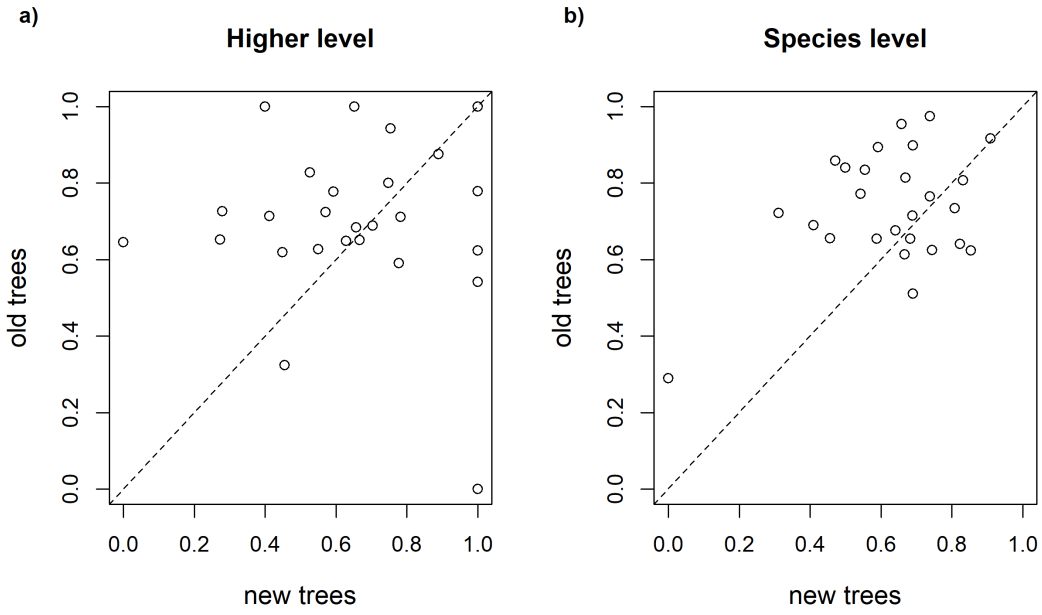

Figure 6.2: Comparison of $\overline{I}_w$ for the values obtained in [51] ('old trees') and the values obtained here ('new trees') for a) higher-level and b) species-level trees.

only produces bifurcating nodes, this does not matter here. For such trees, bifurcating with internal nodes of size zero and n equally sized leaves, the expectation of ${}^1J_N$ is known to a very close approximation. The expectation is [46, 95],

$$\mathbb{E}_Y[{}^1J_N] \approx \frac{n \log_2 n}{\mathbb{E}_Y[I_S]},$$

where $\mathbb{E}_Y[I_S]$ is the expectation of Sackin's index under the Yule process, given by,

$$\mathbb{E}_Y[I_S] = 2n \sum_{i=1}^{n} \frac{1}{i}.$$

The difference between the approximation and the true expectation is less than 0.008 for all n [95].

To compare the balance of our trees with the Yule process, for each observed value, I compare it to 1000 Yule trees of the same size. I calculate the proportion of these that are less than or equal to the observed data, giving a p-value. I then combine the p-values using Fisher's method to obtain an overall p-value. I perform this test three times, with three different node types considered when calculating the balance index. First, all nodes in the tree, second, all non-linear nodes (outdegree greater than 1) and third, only bifurcating nodes.

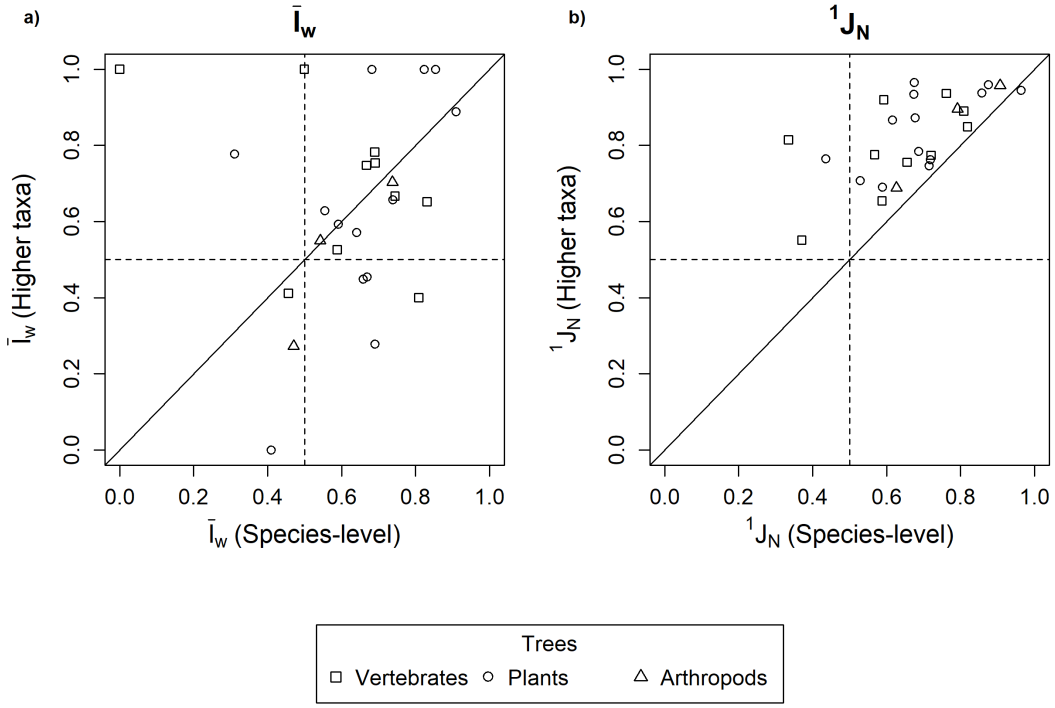

Figure 6.3: Higher-level versus species-level analysis for a) $\overline{I}_w$ and b) ^1J_N. The solid line is equality.

6.3 Results

6.3.1 Completeness of tree does not affect tree balance

I investigate whether the completeness of a tree affects tree balance as quantified by the two indices $\overline{I}_w$ and ^1J_N. For $\overline{I}_w$, both the higher-level and species-level trees showed that there is no significant relationship between tree balance and whether a tree was complete or not (adjusted p-values: P=0.77, P=0.77 respectively). There was also no significant relationship between balance and the magnitude of a tree (adjusted p-values P=0.86 and P=0.86). For ^1J_N, the higher-level trees showed there was a significant relationship between magnitude and balance (adjusted p-value P=0.03) but not whether the tree was complete (adjusted p-value P=0.66). For the species-level trees, there was no significant relationship between tree balance and whether a tree was complete (adjusted p-value P=0.36) and the magnitude of a tree (adjusted p-value P=0.06).

6.3.2 New index shows that tree balance varies with taxonomic level

I explore whether tree balance is affected by taxonomic level, using two different sets of indices, the I-based indices and our J-based indices. I first investigate trees individually, then I combine the trees in a meta-analysis.

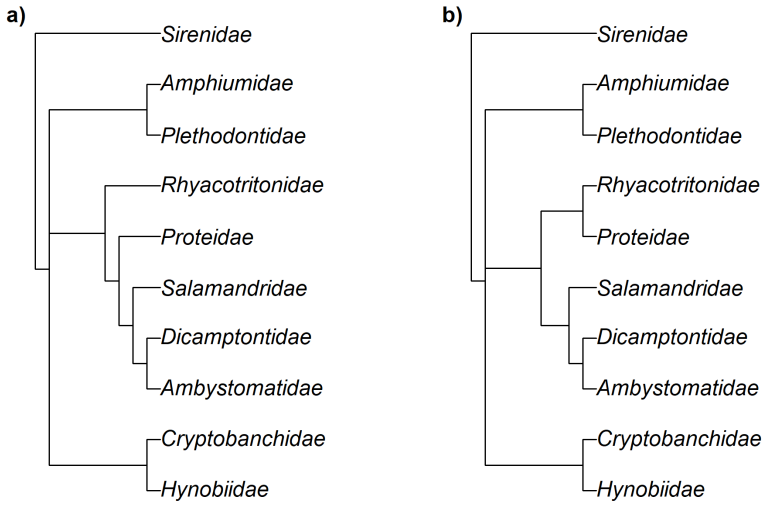

Figure 6.4: a) Urodela (Salamander) family-level tree obtained from the Open Tree of Life with $\overline{I}_w = 1$ and ${}^1J_N = 0.818$. b) The Urodela tree with the position of the family Proteidae altered, with $\overline{I}_w = 0.474$ and ${}^1J_N = 0.869$.

Using the I-based indices, for the individual tree analysis, I found that 1 of the 23 trees that qualify for a test has a $\overline{I}_w$ that differs significantly between taxonomic levels. The meta-analysis did not show a significant effect of taxonomic level on balance (sign test, P=0.85, Supplementary Figure 8.28a). The correlation between the higher-level and species-level $\overline{I}_w$ was 0.106 (Figure 6.3a).

For the J-based indices, I found that 5 out of 26 trees that qualify for a test had a ${}^1J_N$ that differs significantly between taxonomic levels. The meta-analysis showed a significant effect of taxonomic level, where for 25 of 26 trees, the higher-level analysis suggests a greater degree of balance (sign test, $P = 8 \times 10^{-7}$, Supplementary Figure 8.28a). I also found that the mean of these within-tree differences was significantly non-zero (weighted t-test, $P = 1 \times 10^{-6}$, Supplementary Figure 8.28b). The correlation between higher-level and species-level tree balance was 0.607 (Figure 6.3b).

Our balance index ${}^1J_N$ shows that the higher-level trees are, on average, significantly more balanced than the species-level trees. This is seen somewhat in individual trees, but it is clear when I consider all trees together. This is in direct contrast to what Purvis and Agapow found, that the species-level trees are more balanced. Three main reasons could be causing this difference: the different data used, how the indices measure balance and imbalance, and that $\overline{I}_w$ only takes into account bifurcating nodes with size ≥ 4, where ${}^1J_N$ takes into account all nodes.

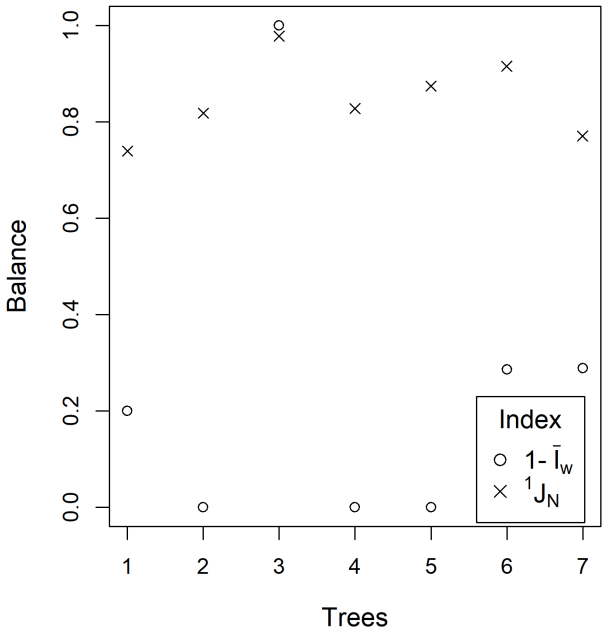

Figure 6.5: $1 - \overline{I}_w$ and ${}^1J_N$ for seven vertebrate trees used in [51]. The trees are Anura, Urodela (Salamander), Ciconiidae, Trogonidae, Bathyergidae, Phyrnosomatidae and Squamata respectively.

6.3.3 I-based indices are sensitive to small changes in tree shape

Here, I examine why I obtain different results from Purvis and Agapow using the same imbalance index. Using $\overline{I}_w$, I found no significant difference in balance between taxonomic levels. With this imbalance measure, Purvis and Agapow found that their trees were more unbalanced when analysed at higher levels rather than at the species level. Our trees generally have different imbalance values compared with the values obtained by Purvis and Agapow; in particular, their trees are typically more imbalanced than ours (see Figure 6.2). This is partly due to differences in tree topology, but also due to $\overline{I}_w$ itself. $\overline{I}_w$ is very sensitive to certain small changes in tree structure. For example, if I take the Salamander tree used by Purvis and Agapow and move the family Proteidae over one branch, $\overline{I}_w$ goes from 1 to 0.474 (see Figure 6.4). The imbalance index goes from assigning the tree to be maximally unbalanced to quite balanced from this small change. In contrast, ${}^1J_N$ goes from 0.818 to 0.869, seeing only a small increase in balance as a result of this change. In conclusion, we do not observe what Purvis and Agapow did with the I-based indices due to the difference in data and the index used.

6.3.4 Different data is not the cause of the contradictory results

Here, and in the next section, I examine why I obtain different results from Purvis and Agapow using our balance index. The fact that I used different tree data seems like the obvious and simplest answer to why I obtain my results; however, I will show here that the different trees are not a major contributor. Figure 6.5 compares the two index values (in particular $1 - \overline{I}_w$ to

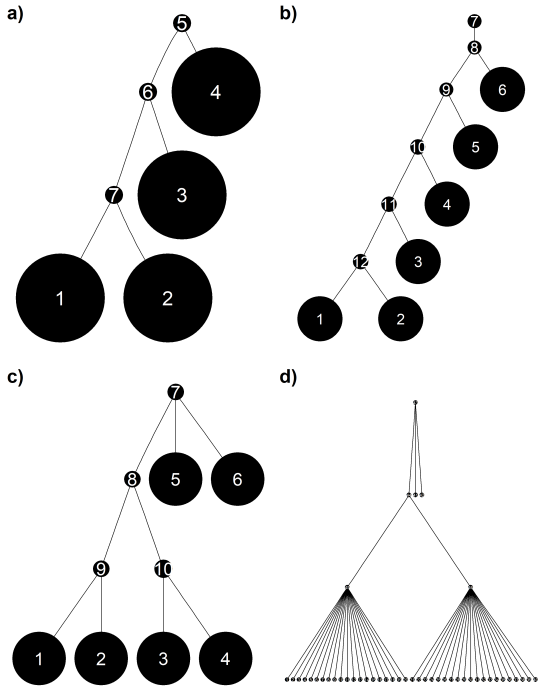

Figure 6.6: Four trees illustrating the differences in the imbalance index $\overline{I}_w$ and balance index ${}^1J_N$. The numbers are node labels. a) and b) are measured as perfectly imbalanced $\overline{I}_w = 1$. Our balance index assigns them both to be somewhat balanced, with a) being more balanced than b), a) ${}^1J_N = 0.909$ and b) ${}^1J_N = 0.827$. c) and d) are measured as perfectly balanced $\overline{I}_w = 0$, our balance index measures c) to be quite balanced, and much more balanced than d), c) ${}^1J_N = 0.910$ and d) ${}^1J_N = 0.726$.

make the comparison simpler) for seven of the vertebrate trees used in [51]. The index values are very different for all but one of the trees, demonstrating how the two indices measure balance differently. One example of this is the caterpillar tree. $\overline{I}_w$ always assigns caterpillar trees to be maximally unbalanced, where ${}^1J_N$ assigns them to be quite balanced for small trees and increasingly unbalanced as the tree grows (see Figure 6.6). The results with the Open Tree of Life data also demonstrate this difference; $\overline{I}_w$ does not find a difference between taxonomic levels, but ${}^1J_N$ does. In fact, for 11 of the 26 trees from the Open Tree of Life, $\overline{I}_w$ and ${}^1J_N$ assign opposite balance changes between the different levels of analysis. Therefore, it seems unlikely that I would reach the same conclusions as Purvis and Agapow even if I used the original set of trees.

6.3.5 Non-bifurcating nodes lead to counter-intuitive balance values for $\overline{I}_w$ but does not contribute to our results

One difference between the indices is that ${}^1J_N$ accounts for all nodes, but $\overline{I}_w$ only accounts for bifurcating nodes with size ≥ 4. Here, I explore whether this difference is the cause of our results. The ability of $\overline{I}_w$ to 'account' for polytomies can lead to some strange and counter-intuitive imbalance values for trees. Consider Figure 6.6. The tree in Figure 6.6c is assigned

as perfectly balanced ($\overline{I}_w = 0$) as the root node 7 is not bifurcating and so is ignored in the calculation. Increasing the number of descendant leaves from nodes 9 and 10 to some arbitrarily large value (Figure 6.6d) does not change the index value, as these nodes are not bifurcating and the extremely unbalanced tree is assigned to be perfectly balanced.

Our data contains 604 nodes, of which 233 are non-bifurcating. Of these 233 non-bifurcating nodes, 206 are linear, and 27 are polytomies. Repeating our analysis with ${}^1J_N$ using only the nodes that $\overline{I}_w$ considers changes our results slightly, but not the overall conclusion. I found that none of the trees individually has a balance that differs significantly between taxonomic levels. The meta-analysis showed a significant effect of taxonomic level on balance, with 22 out of 26 trees being more balanced when analysed at the higher level (sign test, $P = 0.0005$). The mean of these within-tree differences was significantly non-zero (weighted t-test, $P = 7 \times 10^{-4}$). Hence, it is not the non-bifurcating nodes that are responsible for the results.

6.3.6 The balance of a node depends on the taxonomic level of the leaves

I now investigate whether the balance of a node depends on the taxonomic level of the leaves, quantifying balance using the two indices I_w and ${}^1J_{N,i}$. For I_w, there was no significant effect of the taxonomic level of the leaves on node balance. The robust linear model showed that the interaction term, taxonomic level, and size were not significant predictors of node balance (P = 0.22, P = 0.13, and P = 0.65, respectively). The robust ANCOVA showed that for node sizes 4, 6, 8 and 12, there was no significant difference in balance between taxonomic levels (adjusted p-values: P = 0.97, P = 0.97, P = 0.97 and P = 0.83, respectively).

For ${}^1J_{N,i}$, there was a significant effect of the taxonomic level of the leaves on node balance. The robust linear model showed that the taxonomic level was a significant predictor (P=0.0001), but the interaction term and total descendant abundance were not (P=0.20 and P=0.28, respectively). The robust ANCOVA revealed a significant difference in node balance between taxonomic levels for nodes of total descendant abundance 4 (adjusted P = 0.03). For nodes of total descendant abundance 6, 8 and 12, there was no significant difference in balance between taxonomic levels (adjusted p-values: P= 0.09, and P= 0.07 and P= 0.28, respectively).

Using the node imbalance index I_w, I found that the taxonomic level and node size were not significant predictors of node imbalance. Purvis and Agapow also found this, however, they found a significant difference in the mean node imbalance between taxonomic levels, with species-level nodes being more balanced. Although no differences were significant, I found that for node sizes 4, 6, and 12, the mean node imbalance for species-level nodes was more balanced. Given the previous results that $\overline{I}_w$ showed there was no significant effect of taxonomic level on balance, and tree imbalance is the average node imbalance for the nodes in a tree, these results are as expected. The most likely reasons for the different results here are the same as outlined for tree balance, the different data used and the index I_w itself.

Using the node balance index ${}^1J_{N,i}$, I found that taxonomic level is a significant predictor of

node balance, and for the node with total descendant abundance 4, nodes with higher-level leaves were significantly more balanced than nodes with species as the leaves. Again, given our previous tree balance results, these results are mostly as expected. Although the linear model showed a significant effect of taxonomic level on node balance overall, the ANCOVA showed that the difference in balance between taxonomic levels was not significant across all node sizes. This could be the true effect, that the difference in balance between taxonomic levels is more prominent for smaller nodes than for bigger nodes. Or, it could be an artefact of the data. For all node total descendant abundances, higher-level nodes are more balanced than species-level nodes, and I have many more smaller nodes than larger nodes. For nodes of total descendant abundance 4, the comparison is between 169 and 191 nodes and for nodes of total descendant abundance 12, the comparison is between 29 and 22 nodes. Hence, at the larger total descendant abundances, I may not have enough data to detect significance.

It is worth noting that although I use slightly different methods here than Purvis and Agapow, robust versus classical, they are unlikely to be causing the different results. Robust methods can change the size and significance of effects, but to completely flip the direction of the effect, the data would need very extreme outliers such that when these are removed, the majority of the data shows the opposite pattern. As this is not the case for our data, and I also observe the opposite direction of which level is more balanced with different methods for tree balance, it can be assumed that it is not the robust methods causing these results.

6.3.7 Expected value of 0.5 is a poor null model

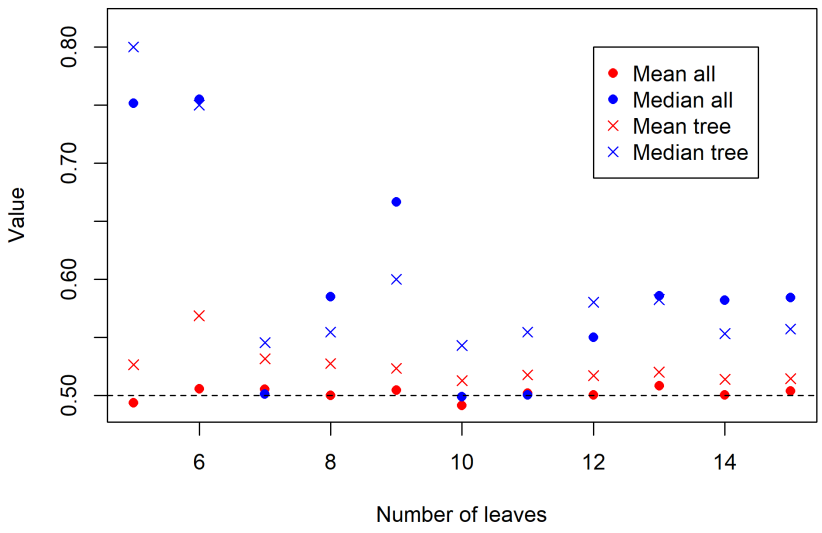

Figure 6.7: Mean and median values of I_w values for 1000 Yule trees. I_w is calculated two ways by changing the denominator. For 'all', I calculated the mean across the weights for every eligible node from all 1000 trees. For 'tree', for a given tree, the mean is calculated across the set of nodes in that tree.

Table 6.2: Type 1 errors, using either the null model of 0.5 or an empirical null, as tree size and sample size is varied.

Tree size	Number of trees	Type 1 error	
		Null 0.5	Empirical null
5	5	0.130	0.058
5	25	0.104	0.060
5	50	0.088	0.052
10	5	0.140	0.051
10	25	0.124	0.065
10	50	0.128	0.039
15	5	0.108	0.060
15	25	0.096	0.051
15	50	0.104	0.039
20	5	0.109	0.015
20	25	0.122	0.023
20	50	0.109	0.003

Here, I investigate one of the stated benefits of the I-based indices: their built-in property for comparison to the Yule process as a null model. Both the node and tree imbalance have an expected value of 0.5 when they are calculated across all nodes for Yule trees of a given size. However, this is misused in two ways. Either the indices are calculated on individual trees, or the median value for node imbalance is investigated, both of which mean that the null model assumption is no longer true. Additionally, I cannot find in the literature any validation of this as an appropriate null model, so I carry that out subsequently.

I find that when calculated on individual trees, as opposed to considering all nodes, the expected value of the indices was not 0.5 (see Figure 6.7). The deviation from 0.5 tended to be bigger for smaller trees and decreased as tree size increased. The distributions of the imbalance indices are non-normal and non-symmetric. As a result, the median and mean are often very different, and this is true regardless of whether the indices are correctly calculated across all nodes or not (see Figure 6.7). Again, the deviation tended to be larger for smaller trees and decreased as tree size increased.

Even when the indices were calculated as intended to satisfy the null model of 0.5, the test had large type 1 errors (see Table 2). I calculated type 1 errors for trees of size 5, 10, 15 and 20 using 5, 25 and 50 samples, and the type 1 error varied from 8.8% to 14%. However, using an empirical null model for the Yule process for these indices gave more reasonable type 1 errors of less than or equal to 6.5% for all the tree and sample sizes I used (see Table 2).

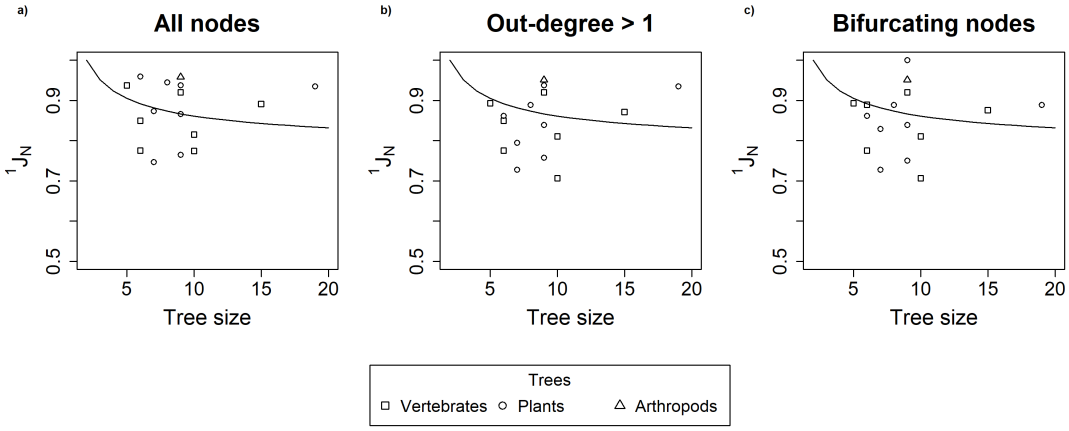

Figure 6.8: Balance of trees compared with the Yule process. The line is the expectation of ${}^1J_N$ for the Yule process. a) ${}^1J_N$ is calculated on all nodes, b) ${}^1J_N$ is calculated on nodes with out-degree greater than one, and c) ${}^1J_N$ is calculated on only bifurcating nodes.

6.3.8 Balance compared to Yule depends on nodes considered

Here, I explore how the balance of these trees compares with what I expect for our index under the Yule process. Due to the computational power required for larger trees and the fact that much of the previous work has only considered smaller trees, I restrict our data to trees with a size less than or equal to 20.

I find that how the balance of the trees compares to the Yule process depends on which nodes are considered. When all nodes were taken into account, I found the data split with 8 more balanced and 8 less balanced than the Yule expectation, with none of these being significant (see Figure 6.8a, overall p-value $P = 0.45$). I considered two subsets of nodes, nodes with out-degree greater than one and bifurcating nodes (see Figures 6.8b and c). Both had extremely similar results, with 6 trees being more balanced and 10 trees being less balanced than the Yule expectation (overall p-values $P = 0.15$ and $P = 0.14$, respectively). For the former set of nodes, one tree was significantly less balanced, and for the latter, two were.

6.4 Discussion

Using our J-based indices, I find that using species richness for trees with leaves of a higher taxonomic level leads to more unbalanced trees than if species richness were not used. I also find that the balance of a node depends on the taxonomic level of its leaves, and species-level leaves produce more unbalanced trees than higher taxonomic levels. Additionally, I found that the results depend on the index used and hence do not claim the pattern I found to be definitive, as our results demonstrate that how balance is measured affects the outcomes. The results I find with our tree balance are unsurprising. Given how our index calculates balance, and that when using species richness, it is much more likely that one will end up with more uneven splits than when assigning leaves to be equally abundant, I would expect the results

I find to almost always be the case. Hence, these properties are likely not the result of any biological process but a consequence of how our index measures balance and captures tree structure.

I find that whether a tree is complete or not does not significantly predict a tree's balance. Previous results on tree completeness are limited. Guyer and Slowinski found that the random omission of leaves does not bias imbalance, and Mooers found that for trees with 8 to 14 tips, complete trees were more balanced than incomplete trees [51, 90]. These results are not necessarily conflicting, as the leaves included (or excluded) in phylogeny reconstruction are unlikely to be truly random [51]. However, our results directly conflict with Mooers' results. This could be due to the different indices used and how they measure balance. I also use a different and smaller dataset. Potentially, the pattern I observed or Moores did is only true of the subset of trees used and not the true relationship. The leaves missing from our data could be more random, whereas for Mooers' data, they were more systematic and biased. The limited and conflicting nature of results on the relationship between tree completeness and balance motivates a need for further study to determine the true nature of the relationship.

I observed inconsistencies in the literature around the use of the I-based indices and their property of having an expected value of 0.5 for the Yule process. In the paper by Purvis et al., where they modify the node imbalance index I outlining their own I_w, they use mean and median interchangeably [121]. They modify I such that I_w has an expectation of 0.5 for the Yule process, but then proceed to test for the median imbalance, stating they are comparing the data with the 'median of 0.5 expected under the Markovian null model' [121]. However, I have shown here that under the Yule process, the node imbalance I_w does not have equal mean and median, and the median does not have an expectation of 0.5. In addition to this, the I-based indices are frequently calculated on the wrong set of nodes, invalidating the null model. In the paper I focus on here, Purvis and Agapow define I_w to be the product of I and w divided by the tree's mean w and state that this has an expectation of 0.5 for the Yule process. I have shown that when calculated on trees in this way, rather than across all nodes by dividing by the mean w for all nodes from trees being considered, the expectation is not 0.5. Calculating the indices in this way makes the null model invalid and no longer mathematically relevant.

I performed validation of the use of the null model of 0.5 for the Yule process, as this did not appear to have been done. I find that using reasonable sample sizes for trees, the null model of 0.5 has large type 1 errors. As tree size increases, the number of possible node arrangements increases rapidly, and as the null model requires all possible nodes to be used, using only a small sample of trees leads to large deviations from the true expected value. Hence, it is an unsuitable null model to use to test for deviations from the Yule process. In addition to this, I investigated the type 1 errors using an empirical null model for tree imbalance $\overline{I}_w$. This led to much more reasonable type 1 errors. Demonstrating, it is not that these indices cannot be used to test for deviations from the Yule process, but that using a null model of an expectation of 0.5 is unsuitable.

I found that how the balance compares to the Yule process depends on which nodes are con-

sidered. It has been previously and repeatedly observed that data is more unbalanced than the Yule process; however, the indices used for these analyses, as the I-based indices do, differ from ours [89–91]. One of the most important ways they differ is in which nodes they consider. Restricting our index to only account for nodes that these indices would, I find that more trees are less balanced than the Yule process, compared with considering all nodes. Additionally, the trees for this analysis have higher taxa as their leaves, which I have observed for this particular dataset, are more balanced than if the leaves were species. Studies comparing the balance of trees from data to the Yule process usually use trees of different ranks as the leaves [89–91]. Therefore, with this dataset, it is unsurprising that I do not find significant deviation from the Yule process.

I have demonstrated that there are issues with the I-based indices that can make them unideal for use. This is a perfect empirical example of why robust and universal indices should be used when studying tree shape. In the context of this work, robust indices are relatively insensitive to the effects of incomplete ancestral relationships [59]. Using robust indices will lead to results that are less likely to change if there are small changes in data, such as relationships being resolved or rearranged. A universal index can be applied to all trees (and nodes). An index needs to be truly applicable and give reasonable values in all the cases it is claimed it will, with the ideal index being applicable in all cases.

An important direction for future work is to explore such patterns more comprehensively using much larger trees. Using trees from many different studies is often problematic because trees can differ substantially in their methods of construction, data quality, and robustness. Exploring much larger, internally consistent trees avoids these issues and lays the groundwork for extending this analysis towards subtrees of the Tree of Life. More generally, future work should take advantage of the ability of our indices to account for branch lengths and should use trees that have reliable branch length data, not just node-size information. Additionally, future work should look at further papers using other balance and imbalance indices. This chapter demonstrates that changing the index can substantially affect the results. Since the I-based indices are not commonly used, it would be particularly informative to investigate whether other results from more widely used imbalance indices, such as Colless' and Sackin's, are similarly sensitive.

In conclusion, consistent with the findings of Purvis and Agapow, this chapter shows that the taxonomic level of the leaves affects tree balance, highlighting the need for caution when comparing phylogenetic trees at different taxonomic levels, and when treating such trees as equivalent when studying tree shape. More broadly, this chapter demonstrates that the results change depending on which index is used, emphasising the importance of being careful and intentional when choosing which tree balance index to use for the given application. Taken together, these results provide a foundation for studying larger subtrees of the Tree of Life. While the ultimate goal is to understand the biological processes behind the shape of the Tree of Life, disentangling and characterising the various contributing factors, whether biological or not, will enable more informed methodological choices and ultimately lead to more reliable and accurate results.

Chapter 7

Application to cancer trees

7.1 Introduction

This chapter is based on work presented in [83], a preprint co-authored with my supervisor. The analyses, simulations, and results presented here were carried out by me, with my supervisor providing editorial feedback and guidance.

Prognosis and treatment decisions in non-small cell lung cancer and many other cancer types are primarily based on how far the tumour has spread (stage) and morphological features (grade) [129]. While this system is broadly effective, it lacks precision and fails to account for the dynamic nature of tumour evolution. Proposals to enable more personalised prognostic forecasting include classifying tumours according to their evolutionary and ecological features [75, 130, 131].

Although numerous indices have been proposed for summarising the size and shape of evolutionary trees and similar structures [48, 66, 132], most studies seeking to develop evolutionary biomarkers in cancer have focused on describing intratumour heterogeneity (ITH) in relatively simple terms. ITH has been associated with tumour progression and therapeutic resistance in multiple cancer types [133]. The TRACERx renal consortium found in their own study and in the larger TCGA kidney cancer cohort that when tumours have low genomic instability, low tumour ITH correlates with longer progression-free and overall survival times [134]. After adjusting for known prognostic variables, including stage and grade, both ITH and genomic instability remained significant predictors in the TCGA kidney cancer cohort but not in their own cohort [134]. In the larger TRACERx non-small cell lung cancer (NSCLC) cohort that I reanalyse here, disease-free survival has been shown to be related to the ITH of somatic copy number alterations but not to mutational ITH [61]. In a multivariate Cox model controlling for stage and other clinical variables, the ITH of somatic copy number alterations was no longer predictive [61]. Measures of heterogeneity have also been found to be predictors of clinical outcome in prostate cancer and in a pan-cancer analysis across 28 cancer types [135, 136].

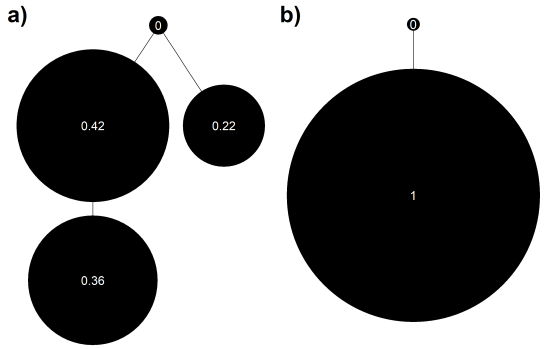

Figure 7.1: Tumour trees illustrating a case where mutational ITH is identical and cannot distinguish between the trees, but ${}^1J_N$ can. a) Tumour ID CRUK0254, mutational ITH = 0.03 and ${}^1J_N = 0.77$, b) tumour ID CRUK0092, mutational ITH = 0.03 and ${}^1J_N = 1$. Branch lengths are arbitrary, and node sizes are proportional abundances.

A complementary but comparatively under-investigated summary statistic is tree balance. As balance indices capture a fundamental aspect of tree shape that characteristically varies between evolutionary processes, tree balance can thus be used to infer parameter values or to compare empirical trees with those generated by mathematical models [44, 56, 58, 90, 105]. However, as I have repeatedly stated, many of the commonly used tree balance indices suffer from limitations, as outlined in chapter 1. Conventional tree balance indices, such as Sackin's and Colless' index, were designed to be applied to species trees and are less well suited to comparing the shapes of tumour phylogenies. The properties of our new system, outlined in chapter 2, mean they are more suited for such applications. Consider the two tumour trees shown in Figure 7.1. Mutational ITH – previously used in the TRACERx NSCLC study – assigns the same value 0.03 to both trees and so cannot distinguish between them. Our tree balance and diversity indices, ${}^1J_N$ and ${}^1D_L$, capture the obvious differences in shape (Figure 7.1a: ${}^1J_N = 0.77$ and ${}^1D_L = 1.70$; Figure 7.1b: ${}^1J_N = 1$ and ${}^1D_L = 1$).

Here, I show that our new tree balance index outperforms measures of intratumour heterogeneity in predicting disease-free survival in non-small cell lung cancer.

7.2 Methods

7.2.1 Indices

Tree shape indices

I have outlined our tree shape indices in chapter 2. Briefly, ${}^1D_N$ quantifies the average effective out-degree or, more informally, the "bushiness" of the tree; ${}^1D_L$ is a diversity index that accounts for phylogenetic relatedness; and ${}^1J_N$ is a tree balance index. All three indices account for tree topology, node sizes (here corresponding to subclone population sizes) and branch lengths (genetic distance). The two D indices can take any positive value, whereas ${}^1J_N$

varies between 0 (minimally balanced) and 1 (perfectly balanced).

Other evolutionary indices

I use three other evolution indices, mutational ITH, somatic copy number alteration (SCNA) ITH and Shannon diversity (or Shannon entropy). The first two indices were used in previous analyses of the TRACERx non-small cell lung cancer cohort [61, 78]. Mutational ITH is the percentage of subclonal mutations. It is calculated by dividing the number of mutations estimated to be subclonal by the total number of mutations classified as either truncal or subclonal in the phylogenetic tree [61]. Somatic copy number alteration (SCNA) ITH is the fraction of aberrant genome with SCNAs. It is calculated by dividing the percentage of the genome harbouring heterogeneous SCNA events, that is, those events that were not present in every region, by the percentage of the genome involved in any SCNA event in each tumour [61]. Shannon diversity is mathematically defined in section 1.3 and is calculated by taking the negative sum, over all non-zero relative abundances, of the relative abundance times its natural logarithm. In this analysis, I then take the exponential of this value, which converts the index into units of "effective amounts". This allows for easier interpretation and comparison with our diversity indices, which also have units of "effective amounts".

7.2.2 TRACERx data

The TRACERx 421 cohort contains 421 patients recruited across 19 hospital sites in the United Kingdom. The recruitment was broadly representative of an early-stage operable non-small cell lung cancer (NSCLC) population in the UK according to ethnicity, age, sex and smoking status [61]. The 421 patients had 432 genomically independent tumours: 248 lung adenocarcinomas (LUAD); 138 lung squamous cell carcinomas (LUSCs); and 46 'other' NSCLC subtypes. Pathological staging was available for all tumours but tumour grading was only available for LUADs [61, 137]. Tumour phylogenetic trees were reconstructed from multiregion whole-exome sequencing (WES) data using the CONIPHER computational framework to infer the evolutionary relationships between tumour clones [61, 138]. Nodes in the trees correspond to genetically defined clones, comprising tumour cells that are identical by descent in their somatic mutation history. The branching structure captures ancestral relationships between clones, where descendant clones inherit the mutational profile of their parent clones but may also acquire additional alterations or lose ancestral mutations through copy-number changes or other genomic events.

Phylogenetic trees could be reconstructed for 401 tumours; 9 were excluded from the analysis. One patient had two synchronous primary tumours, one of which was not sequenced. Additionally, for the other patients with synchronous primary tumours, I used the tumour with the highest stage; this removed a further 8 trees. For branch lengths, I used the absolute difference in the number of mutations between parent and child clones. Due to the tree construction method that allows for somatic copy number alterations (SCNAs) to remove mutations, child clones could have fewer mutations than their parent, hence the need to use the absolute num-

ber. Clone sizes were obtained from CONIPHER using the cancer cell fractions (CCF). 5 of the remaining 392 trees only contained the root node, a case in which our indices are not defined. To include these tumour trees in the analysis, I assigned a tree with only the root node to have index values of one. Figure 7.2 shows four of the tumour trees; a and b are completely balanced, and c and d are unbalanced.

7.2.3 Cutpoints for categorical analysis

Choosing cutpoints to group continuous variables is not a simple task, with no universally agreed way to do it, and crucially, the results of analyses can change drastically if different cutpoints are used [139]. Commonly used cutpoints are splitting around the median and the method of "optimising" the P-value, which is equivalent to minimising the P-value. However, just because they are commonly used does not mean they are without their pitfalls. Splitting around the median value gives even group sizes, but other than that, it is as arbitrary as splitting around any other value. The minimum P-value, although it may seem mathematically desirable, has issues from limiting the ability to compare studies, to an inflated type I error rate [139]. Altman et al. demonstrated the associated issues with the minimum P-value method using the example of S-phase fraction as a prognostic marker in breast cancer in [139]. Here I took their recommendation where 'the choice of cutpoints should be guided by biological reasoning, knowledge of measurement techniques, and simplicity' [139]. As our indices are mathematical and not biological, I used simplicity when selecting our cutpoints, where I chose the cutpoints such that the groupings made sense based on the index values and also kept reasonable group sizes. Given this, the groupings here may not be optimal, and I include the analysis for each index as a continuous variable to demonstrate the results are not just due to the chosen cutpoints.

7.3 Results

7.3.1 New tree shape indices predict disease-free survival

I applied tree shape indices to 392 phylogenetic trees reconstructed for tumours from patients from the TRACERx non-small cell lung cancer cohort. Trees ranged from having 1 to 10 leaves, and the average number of total nodes in a tree was 10.34. Node sizes correspond to the overall proportion of tumour cells belonging to each clone, and branch lengths correspond to genetic distance. Figure 7.2 shows four examples.

I initially investigated the relationship between disease-free survival (DFS) and three of our new indices: ${}^1D_N$,${}^1D_L$ and ${}^1J_N$. Treating the indices as continuous variables, I found a significant association between DFS and all three indices (${}^1D_N$ hazard ratio (HR) = 1.21, 95% confidence interval (CI) = 1.06-1.39, ${}^1D_L$ HR = 1.21, 95% CI = 1.05-1.38, and ${}^1J_N$ HR = 0.79, 95% CI = 0.69-0.90. When stating hazard ratios for continuous variables, the variables

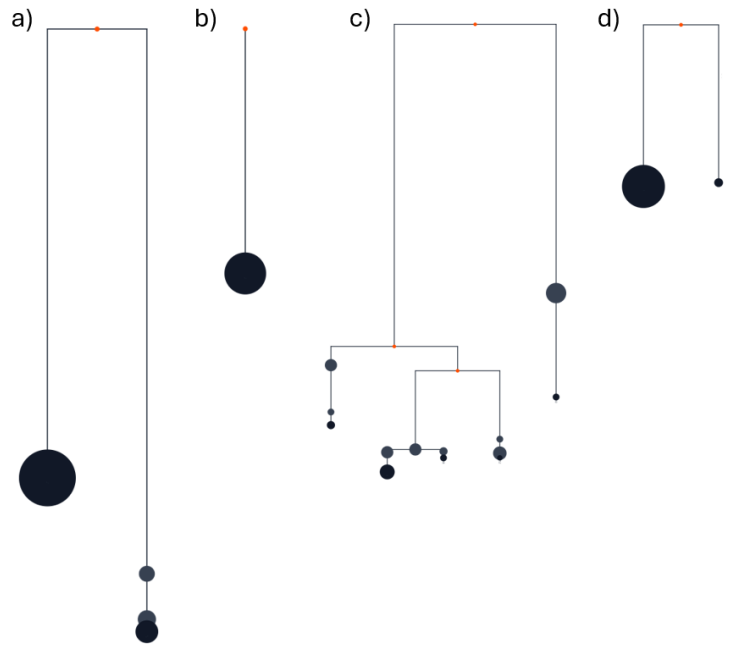

Figure 7.2: a-b) Completely balanced (${}^1J_N = 1$) and c-d) unbalanced (${}^1J_N < 0.73$) trees shown with proportional abundances (not consistent between plots) and branch lengths (consistent between plots). Red nodes either have zero abundance or are the root node. Tumour IDs a) CRUK0027, b) CRUK0061, c) CRUK0284 and d) CRUK0756.

have been scaled such that one unit change is equivalent to one standard deviation. This allows a fair comparison of the effect of variables across different scales.) In a multivariable Cox proportional hazards model accounting for all three indices, only ${}^1J_N$ predicted DFS (HR = 0.79, 95% CI = 0.69-0.91). I also explored the non-normalised version of ${}^1D_L$ but found no significant relationship with DFS (HR = 1.09 95% CI = 0.95-1.26).

Using the indices to assign patients to three categories (see Methods), I found a significant association between DFS and ${}^1J_N$ in all pairwise comparisons (Figure 7.3c). High ${}^1D_N$ and ${}^1D_L$ values also predicted shorter DFS when compared to low or intermediate values (Figures 7.3a and b). As the tree balance index ${}^1J_N$ performed best in these analyses, I will henceforth focus on this one index.

Of the 392 trees, 23 have ${}^1J_N = {}^1D_L = {}^1D_N = 1$, 18 are linear trees, and 5 contain only the root node, and so I defined the indices to be 1. Removing these trees had minimal effect on the results.

7.3.2 Tree balance and associations

I investigated the relationships between ${}^1J_N$ and clinical characteristics, such as age and smoking history, and between ${}^1J_N$ and disease and treatment variables, such as stage and surgery type. I found no significant association between ${}^1J_N$ and clinical characteristics. I found that

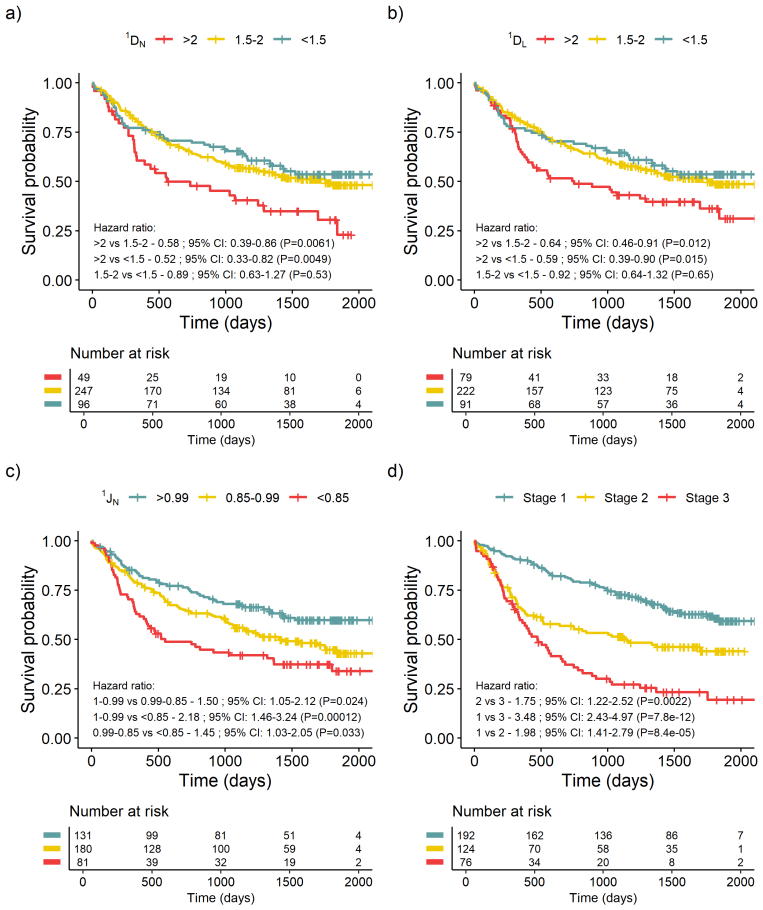

Figure 7.3: Survival curves showing the difference in disease-free survival for tumours based on tree shape indices a) ${}^1D_L$, b) ${}^1D_N$, c) ${}^1J_N$ and d) the stage.

${}^1J_N$ differs across stage on average, with balance in stages 2 and 3 being significantly lower than in stage 1, but not between stages 2 and 3 (1 vs 2: ${}^1J_N$ difference $= 0.041$, $P = 0.001$, 1 vs 3: ${}^1J_N$ difference $= 0.049$, $P = 0.001$ and 2 vs 3: ${}^1J_N$ difference $= 0.008$, $P = 0.86$, Supplementary Figure 8.29). I also found significant relationships between tree balance and several other disease and treatment factors, whether the patient had lymphovascular invasion, the surgery type and whether the patient had adjuvant treatment. However, these clinical factors are highly linked through clinical decision-making pathways, with treatment and surgical decisions being largely determined by tumour stage. Hence, these associations likely reflect the underlying relationship between tree balance and stage rather than anything else.

7.3.3 Tree balance remains prognostic after controlling for stage

As stage is also predictive of DFS (Figure 7.3d) and is only weakly associated with ${}^1J_N$, I next investigated whether stage and ${}^1J_N$ perform better in combination than either does alone. When I split the cohort according to stage, I observed a significant relationship between tree balance and DFS within stage 2 when comparing the most and the least balanced trees

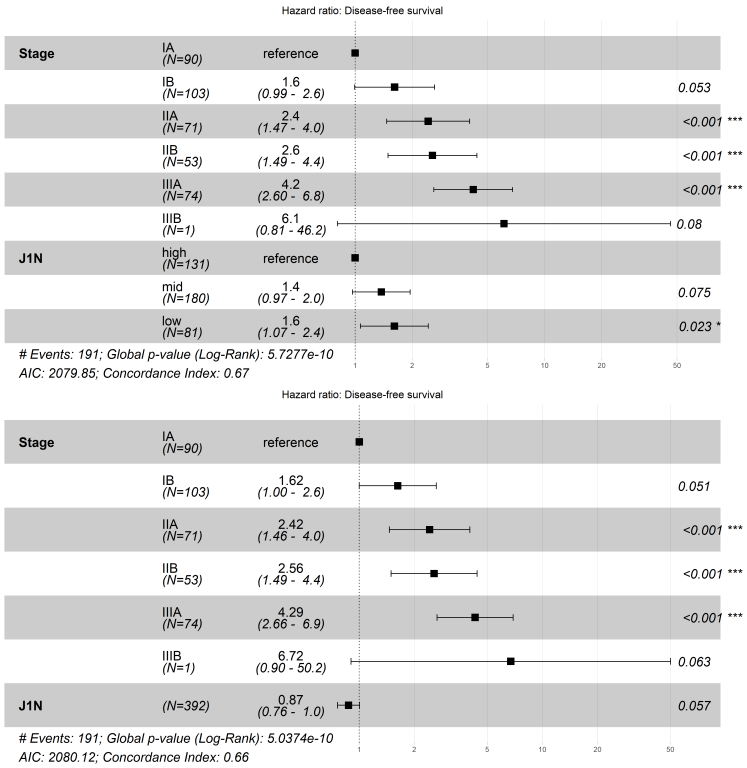

Figure 7.4: Multi-variable Cox proportional hazard models containing stage and tree balance, ${}^1J_N$, a) split into intervals, and b) as a continuous variable. The HR 95% CIs are shown in brackets and by the error bars. The asterisks indicate the P value ranges, where $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

(Figure 7.5b, HR = 2.12, 95% CI = 1.07-4.21). After controlling for stage in multivariate Cox models, tree balance remained a significant predictor of DFS when comparing the most and least balanced trees (HR = 1.6, 95% CI = 1.07-2.4; Figure 7.4). The hazard ratios for tree balance change little when I controlled for grade as well as stage but were no longer significantly different from unity (Supplementary Figure 8.31). The latter analysis suffered from a reduced sample size (191 patients and 86 events) because grade information is available only for lung adenocarcinomas.

7.3.4 Results are insensitive to the omission of rare clones

To test whether tree balance remained predictive when applied to poorer quality data, I examined the effect of omitting rare clones from the TRACERx trees. I identified nodes corresponding to tumour clones with low proportional abundances and then merged each such node with its parent (Figure 7.6; Figure 8.32). Merging nodes with abundances (including abundances of any descendant clones) less than 1%, 5% and 10% of the total abundance in the tree reduced the mean number of nodes per tree by 0.09 %, 8.8% and 25.7 %, respectively. For all three tolerance values, the results varied little. DFS remained significantly related to ${}^1J_N$ treated either as a continuous variable (HR = 0.79, 95% CI = 0.69-0.90; HR = 0.81, 95% CI

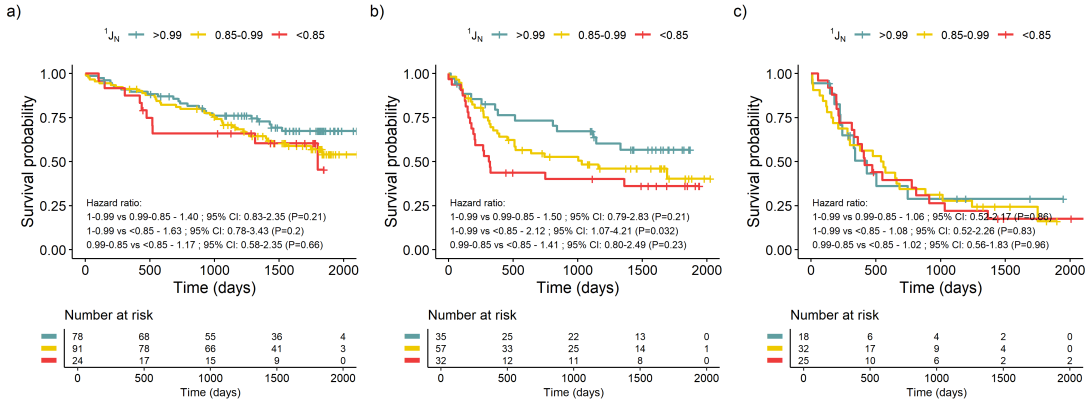

Figure 7.5: Survival curves showing the difference in DFS for tumours based on their phylogenetic tree balance, ${}^1J_N$, when split by stage, a) stage 1, b) stage 2 and c) stage 3.

= 0.71-0.92; HR = 0.84, 95% CI = 0.74-0.96; for thresholds 1% and 5% and 10% respectively) or (in all but one of the pairwise comparisons per threshold) as a categorical variable (Figure 7.7).

The removal of rare clones typically increases tree balance, resulting in more patients being initially assigned to the high-${}^1J_N$ category (Figure 7.7). As I increasingly removed rare types, the difference in DFS between the high- and medium-${}^1J_N$ categories diminished, while the difference between the medium- and low-${}^1J_N$ categories increased. Similar results pertain for multivariate models controlling for stage (HR = 0.87 95% CI = 0.76-1.00; HR = 0.89 95% CI = 0.78-1.02; HR = 0.92 95% CI = 0.80-1.05 for thresholds 5% and 10% respectively, with ${}^1J_N$ as a continuous variable).

7.3.5 Results are robust to the absence of clone size and branch length data

I next compared the predictive power of ${}^1J_N$ to that of three variants of our tree balance index that account for branch lengths but not node sizes (${}^1J_{N,a}$); account for node size but not branch lengths (${}^1J_{N,b}$); or account for neither branch lengths nor node sizes (${}^1J_{N,c}$). In the first and last cases (consistent with the convention for cladograms [46]), I assigned size one to the leaves of each tree and size zero to the internal nodes. For fairer comparison with ${}^1J_N$, I adjusted the lower cutpoint boundary for each alternative balance index to maintain approximately 80 trees in the low-balance category, while keeping the upper cutpoint unchanged (see Supplementary Figure 8.33 for results using the ${}^1J_N$ lower cutpoint for all indices). As categorical variables, all variants of our tree balance index gave similar results (Figures 7.8a-c). As continuous variables, both individually and in multivariate models with stage, the variants performed similarly but ${}^1J_{N,b}$ performed the best (${}^1J_{N,a}$: HR = 0.78, HR = 0.87, ${}^1J_{N,b}$ HR = 0.76, HR = 0.84, ${}^1J_{N,c}$ HR = 0.77, HR = 0.86 without and with stage respectively). For ${}^1J_{N,b}$, 75% of trees remained in the same category as for ${}^1J_N$; for ${}^1J_{N,a}$, 70% remained in the same category; and for ${}^1J_{N,c}$ 64% did. These results suggest that it is more important to account for node sizes (tumour

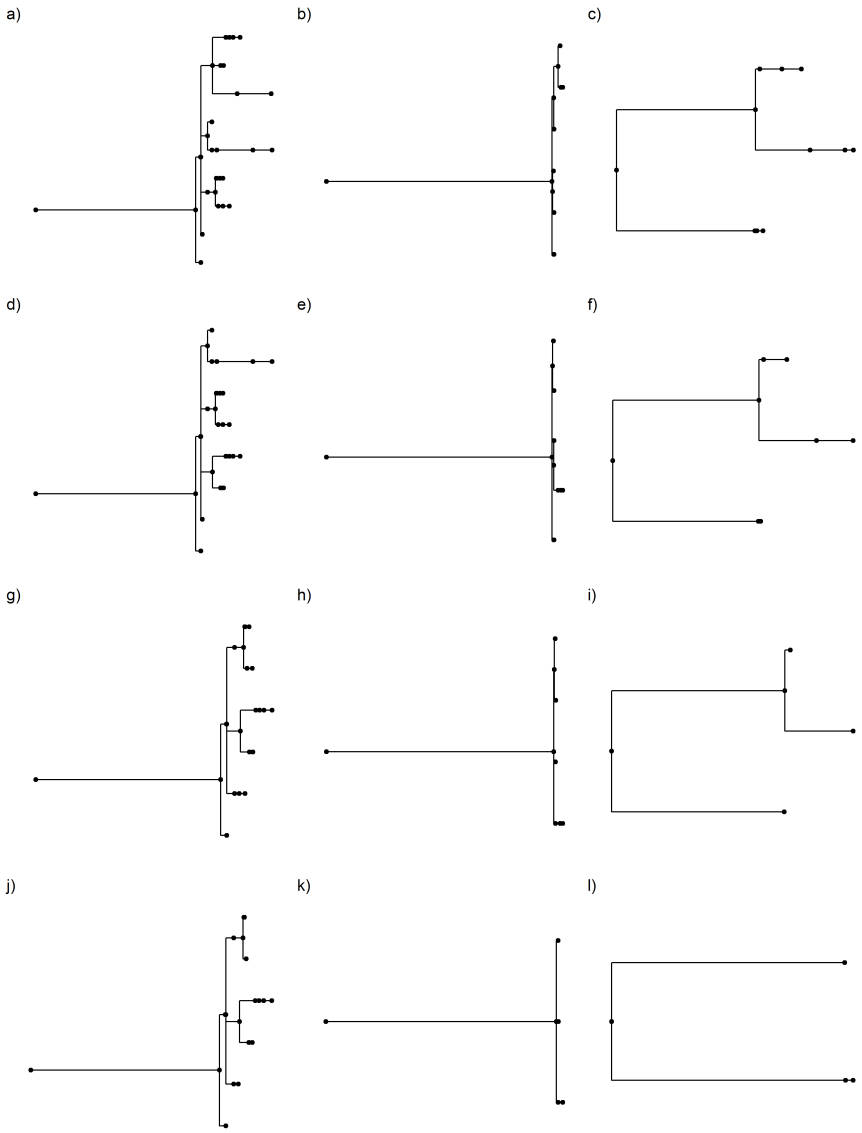

Figure 7.6: Three tumour trees at different levels of coarse-graining. a-c) the original trees, d-f) 1%, g-i) 5% and j-l) %10. (Tumour IDs CRUK0065, CRUK0462 and CRUK0496 respectively). Trees are shown with branch lengths only.

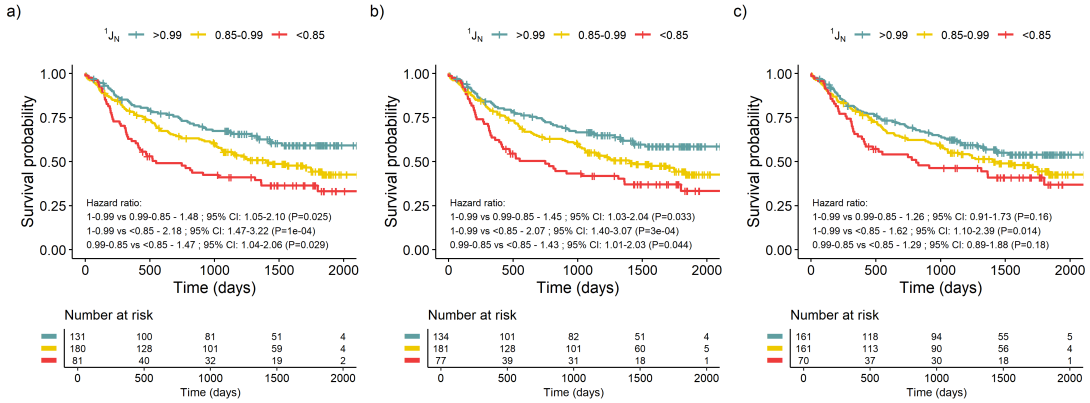

Figure 7.7: Survival curves showing the difference in DFS for tumours for varying levels of coarse-graining, a) 1%, b) 5% and c) 10%.

clone sizes) than for branch lengths (genetic distances between clones).

The variant indices as continuous variables appeared to perform slightly better than ${}^1J_N$, with smaller hazard ratios that are more significant. However, this was not true when the indices were transformed to categorical variables. Overall, no variants outperformed ${}^1J_N$. Without stage, only ${}^1J_{N,b}$ outperformed ${}^1J_N$ in one comparison having a larger HR that is more significant, the others did not outperform ${}^1J_N$ in both cases - larger HR that is also more significant - in any comparison (Figures 7.3c and 7.8). With stage, ${}^1J_{N,b}$ performed very similarly to ${}^1J_N$, and ${}^1J_N$ outperformed the other variants (Figures 7.4 and 8.34).

Coarse-graining with respect to either abundances or branch lengths (at 10% and 1.5% level respectively, which removes approximately the same number of branches) led to ${}^1J_{N,a}$ no longer being significant (HR = 0.90 95% CI = 0.78-1.02 and HR = 0.94 95% CI = 0.82-1.08 respectively). Coarse-graining with respect to branch lengths at the 1.5% level led to ${}^1J_{N,c}$ no longer being significant (HR = 0.93 95% CI = 0.69-1.06). For both types of coarse graining, ${}^1J_N$ and ${}^1J_{N,b}$ had HRs that changed little and remained significant. However, as categorical variables, ${}^1J_{N,b}$ performed worse and lost significance first.

7.3.6 New indices outperform prior evolutionary indices

Next, I compared the predictive power of our tree shape indices to three alternative measures of intratumour heterogeneity. I investigate mutational ITH and SCNA ITH, which were used in previous analyses of the TRACERx non-small cell lung cancer cohort [61, 78], and the Shannon diversity.

For mutational ITH, I found no significant relationship with DFS when the ITH index is treated as either a categorical variable (Figure 7.8d) or as a continuous variable. For ITH in terms of somatic copy number alterations (SCNA) as a continuous variable, I found a significant relationship with DFS (HR = 1.19 95% CI = 1.03-1.37). Treating SCNA ITH as a categorical variable, I found a significant difference in DFS when comparing the high- and low-value

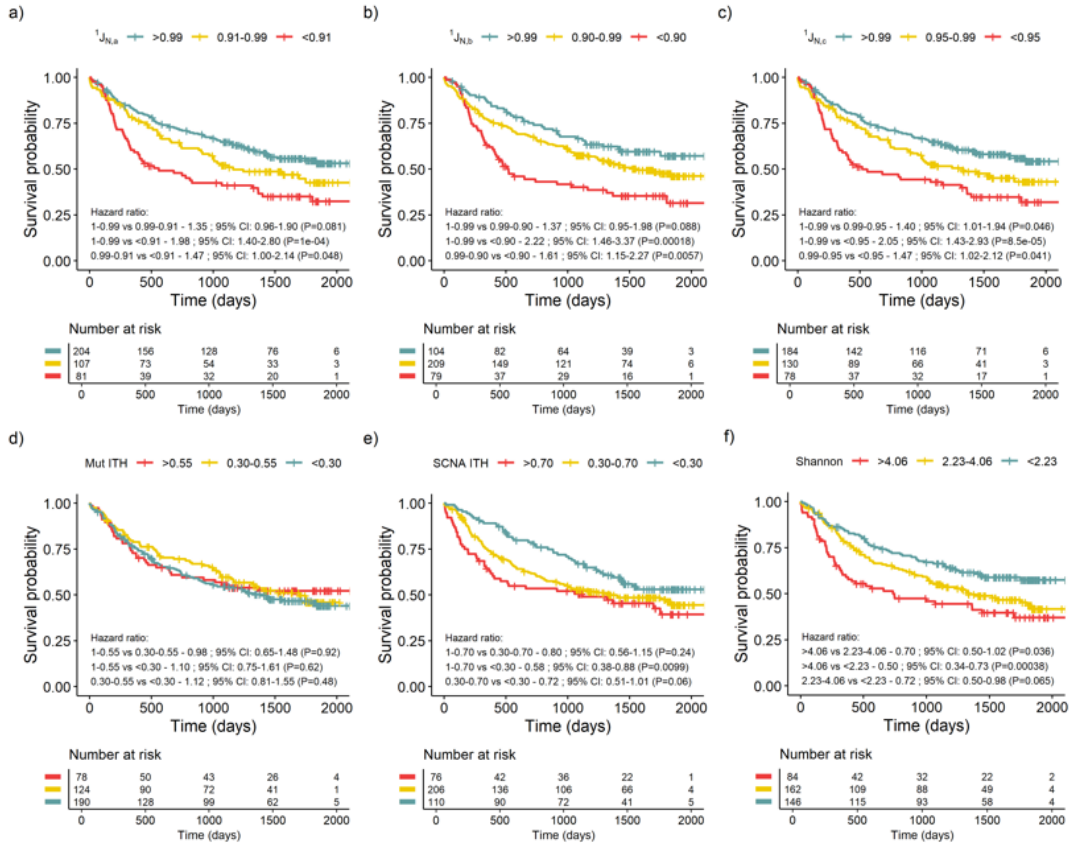

Figure 7.8: Survival curves showing the difference in disease-free survival for tumours based on alternative indices. a-c) are alternative versions of our tree balance index, where a) ${}^1J_{N,a}$ accounts for branch lengths but leaves are assumed to have equal abundance and internal nodes have size zero, b) ${}^1J_{N,b}$ accounts for node sizes but not branch lengths, and c) ${}^1J_{N,c}$ accounts for neither node sizes or branch lengths. d) Mutational ITH is the percentage of mutations that are subclonal. e) Somatic copy number alteration (SCNA) ITH is the fraction of aberrant genome with subclonal SCNAs, both d and e are taken from [61]. f) Shannon diversity in units of effective types calculated on leaves only.

categories (Figure 7.8e). But in a multivariable model controlling for stage, SCNA ITH was no longer significantly associated with DFS, consistent with previous results [61].

For the Shannon diversity calculated on the leaves of the trees, as a continuous variable, I found a significant relationship with DFS (HR = 1.28 95% CI = 1.12-1.46). As a categorical variable, I found a significant relationship with DFS in comparisons with the high-value categories (Figure 7.8f). However, in multivariable models controlling for stage, the Shannon diversity was no longer significantly associated with DFS. I also calculated the Shannon diversity on all nodes in the tree, finding that this was never significantly associated with DFS. This analysis confirms that our tree balance index ${}^1J_N$ outperforms prior evolutionary indices for predicting DFS in this cohort.

7.4 Discussion

I have shown that in the TRACERx non-small cell lung cancer cohort, there is a significant relationship between disease-free survival (DFS) and aspects of tumour clone tree shape, including the effective out-degree (${}^1D_N$), the effective number of maximally distinct leaves (${}^1D_L$), and tree balance (${}^1J_N$). Among these three indices, tree balance performs the best. Multivariable Cox models with both stage and tree balance showed that tree balance remains a significant predictor after accounting for cancer stage. The main advantage of tree balance is in stratifying patients in stage 2.

I demonstrated that the removal of rare nodes at small tolerance values has very little effect on our results. As the number of nodes removed increases, our balance index begins to lose power; however, for all three levels of coarse-graining, the relationship between tree balance and DFS remained significant. Therefore, I have shown the robustness of our method to the omission of rare types.

I find that tree topology accounts for the main diagnostic signal, with branch length and abundance data being unnecessary when fine-scale structure is intact, as is the case here with the high-quality data used to generate the trees. Branch lengths become important when I consider the categorical stratification of the coarse-grained data, suggesting that branch length information enhances discrimination between clinically distinct groups once short branches are removed. These results imply that, in well-resolved data, topology alone captures most of the prognostic information, but ${}^1J_N$, which accounts for abundances and branch lengths - clone size and genetic distance here - provides robustness and improved patient stratification as data quality decreases.

Evolutionary and ecological processes in cancer are known to be important, yet there is a need to develop methods that map the differences in tumour evolution into information that matters for patient outcomes [75]. Tree balance has long been used to study evolutionary processes [44], primarily in systematic biology but increasingly in other research areas such as cancer. It can detect branching rate heterogeneity in lineage-tracing data [58], and is associated with immunotherapy response in colorectal cancer [105]. These results, along with our finding that tree balance is significantly associated with DFS, suggest that the tree balance captures clinically relevant aspects of tumour evolution. Ultimately, highlighting tree balance as an emerging and informative lens for studying cancer evolution.

To understand why I find tree balance to have such prognostic power, I consider the evolutionary dynamics that may give rise to this pattern. Clonal diversity has been shown to predict tumour growth and outcome primarily as a proxy for intrinsic biological factors such as mutation rate and clonal turnover [140], while spatial models indicate that tumour architecture and cell dispersal dynamics influence these same processes [57]. Tree balance may therefore act as a proxy for the combined influence of biological and ecological constraints on tumour evolution. The better prognostic performance of tree balance compared with diversity-based indices may suggest that the evenness of evolutionary branching, rather than the effective out-degree or number of leaves, is more important for determining clinical outcomes.

An important direction for future work is to deepen the biological interpretation of the signals detected by these indices. One relatively simple way future work could attempt to do this is by looking at driver mutations. Future work should explore whether there is any relationship between tree balance and driver mutations, which could include simply the presence of mutations, the total number of mutations, the location of mutations, or some combination. A more complex direction for future work to tackle this question would involve devising and testing tumour growth models capable of reproducing the observed correlation between tree balance and disease-free survival. While this chapter demonstrates the presence of this association, understanding the underlying biology and mechanisms that generate it is crucial to assessing its robustness and potential clinical relevance.

In conclusion, this chapter has demonstrated that our indices show a significant association with disease-free survival, with tree balance exhibiting the strongest relationship. Moreover, the combination of tree balance and stage leads to better patient stratification than stage alone. Together, these results demonstrate that tree balance captures clinically relevant aspects of tumour evolution that are not fully reflected by traditional clinical variables. This suggests that phylogenetic tree shape has the potential to act as a complementary and biologically informed descriptor of tumour progression.

Chapter 8

Conclusion

This thesis aimed to explore our new system of tree shape indices and, through the use of these indices, investigate what insights tree shape can provide. Tree shape is an underexploited method with huge potential; however, a major issue was the lack of suitable indices. The introduction of our new system of indices solved this problem as they are robust, universal, and interpretable, providing a general method that can be applied to any rooted tree. In this thesis, I have considered tree shape from the perspective of macroevolution and cancer, though it has applications and potential in many other areas.

This thesis has demonstrated that tree shape indices can effectively detect meaningful signals in trees in varied contexts. Across multiple chapters, indices have been shown to detect branching rate heterogeneity, distinguish between different modes of evolution, and correlate with disease-free survival in a cancer data set. These results show that tree shape encodes information about the underlying processes in a way that is detectable using various tree shape indices.

Although tree balance was frequently the best-performing index, this was not always the case. In several scenarios, other measures of tree shape had comparable or superior power to detect the relevant signal, showing that using balance alone can neglect important aspects of tree shape. Previous work has largely focused on using tree balance to detect signals, but the analyses here have shown that this is not just a property of tree balance, but a more general one shared across multiple tree shape indices. This suggests that tree shape contains greater information about the underlying process than is typically exploited. Additionally, using only tree balance to study tree shape is overly restrictive, and a broader set of tree shape indices is needed.

Throughout this thesis, our indices generally outperform alternative measures across the majority of cases considered. This is not only reflected in their power, but in their fundamental properties of robustness, universality and interpretability. In several cases, commonly used alternative indices could not be applied to all trees, limiting their usefulness. By contrast, our indices are defined for all rooted trees and so can be applied across diverse contexts. These results demonstrate that the effectiveness of a tree shape index should not just be based on

performance in specific settings, but also on its robustness, universality and interpretability. They also show that our system of indices satisfies these stated properties and highlight the importance of such properties.

The primary contribution of this thesis is strengthening the case for tree shape indices as broadly applicable tools for detecting and comparing signals in tree shape, in doing so, providing insight into the underlying processes. In addition, this thesis demonstrates that although tree balance is informative, it is not always sufficient to fully capture tree shape and other indices are needed. Finally, the work utilises and evaluates a new system of indices that is robust, universal and interpretable, highlighting the importance of such properties in future work.

A major direction for future work is scaling the methods in this thesis and applying the indices to large-scale phylogenetic trees, in particular, the tree of life or substantial subsets of it. While the presented work has demonstrated their effectiveness on simulated and small-sized empirical trees, extending their use to such large trees will require improvements in computational efficiency. This naturally motivates the development of faster software to evaluate the indices. To increase ease of use and to encourage the use of our indices, this software should be integrable with tree inference pipelines.

Beyond enabling broader application, an equally important direction for future work is to deepen the biological interpretation of the signals detected by these indices. In particular, future work would involve devising and testing models to explain the correlation between tree balance and survival in the TRACERx data. Although this thesis demonstrates the presence of this association, establishing its clinical relevance requires moving beyond correlation to an understanding of the underlying biology and evolutionary mechanisms.

Another direction for future work is to investigate the relationship between tree shape indices and traits, with the aim of extending the framework to incorporate trait information. The indices considered in this thesis incorporate phylogeny, abundances and branch lengths; combining these with trait data could provide a better characterisation of evolutionary dynamics by linking patterns in tree shape to observable biological features. Developing such extensions would allow exploration of how variation in traits interacts with tree topology, and may help to clarify the biological interpretation of signals detected by tree shape indices.

Finally, an important methodological direction for future work is to compare the predictive and discriminatory performance of the indices used here with that of machine learning approaches. In this thesis, simple machine learning methods were shown to perform well in distinguishing modes of evolution on relatively small datasets, motivating further investigation into how these two methods compare.

This thesis set out to explore tree shape and the insights it could generate using a new system of tree shape indices. I have shown that tree shape contains detectable signals across a range of contexts. In doing so, the work demonstrated that tree shape has a much greater potential beyond just balance. Additionally, it highlighted the importance of tree shape indices with the properties ours have. Overall, these findings show the potential of tree shape as an analytical

method, with relevance in many fields.

A Chapter 1 supplementary material

Meaning of moe	Total	Cancer	Non-cancer
Overall pattern	15	11	4
Method of modelling	4	0	4
Mechanism/s	15	5	10
Simpsons examples	5	0	5
More than one	6	4	2

Table 8.1: Table of the results of a review of 45 articles containing either of the phrases 'mode of evolution' or 'modes of evolution'. More than one means the paper used concepts that fell into more than one of the defined categories.

B Chapter 2 supplementary material

Weighted means

The following sections outline how the new indices are derived from weighted means. Any terms that are not defined here are defined within the relevant sections in the main text.

To outline how the new index definitions are based on weighted means, it is useful to recall their definition. For a sequence of positive real numbers, $X = x_1, \ldots, x_n$, real number $r \neq 0$, and a set of positive weights, $W = w_1, \ldots, w_n$, the weighted power mean of exponent r is,

$$M_r(X; W) := \left(\frac{\sum_{i=1}^{n} w_i x_i^r}{\sum_{i=1}^{n} w_i} \right)^{\frac{1}{r}}.$$

For $r = 0$, M_0 is defined in the limit as,

$$M_0(X; W) := \exp\left(\frac{\sum_{i=1}^{n} w_i \log x_i}{\sum_{i=1}^{n} w_i} \right).$$

M_{-1}, M_0 and M_1 are the weighted harmonic, geometric and arithmetic means respectively. $M_{-\infty}$ and M_{∞} return the minimum and maximum respectively.

Longitudinal mean

For index F and tree T, we define the longitudinal mean of order r of F as the function $F \mapsto M_{long,r}(F)$ such that,

$$M_{long,r}(F)(T; w) := \begin{cases} \left(\frac{\sum_{i \in I(T)} w_i [F(P_i)]^r}{\sum_{i \in I(T)}} \right)^{\frac{1}{r}}, & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases}$$

where the weight $w > 0$ is a function of i that remains to be specified. Hence, $M_{long,r}(F)$ is a weighted power mean of the F values assigned to the intervals. The tree indices are then

longitudinal means of ${}^qD$ and ${}^qJ$ with $w_i = S_i h_i$, and the index value assigned to each interval i is weighted by the product of the length h_i and the summed sizes S_i of the branch segments that i contains. Where we define,

$$ {}^qD_L := M_{long,0}({}^qD), $$

and,

$$ {}^qJ_L := M_{long,1}({}^qJ). $$

Node-wise mean

Define the node-wise average as the triple power mean,

$$ M_{node,r,s,t}(F)(T;u,v,w) = \left(\frac{1}{\sum\limits_{k\in V(T)} u_k} \sum_{k\in V(T)} \left[\frac{1}{\sum_{j\in A_k} v_{jk}} \sum_{j\in A_k} v_{jk} \left(\frac{\sum_{i\in I(T) w_{ik}[F(P_{ij})]^r}}{\sum_{i\in (I(T)} w_{ik}} \right)^{\frac{s}{r}} \right]^{\frac{t}{s}} \right)^{\frac{1}{t}}, $$

where A_k is the set containing k and all ancestors of k, s is the exponent of the across-ancestors power mean, and v_{jk} are the ancestor weights. To satisfy the condition that for a piecewise star tree with $h > 0$, the index value of each internal node k is equivalent to the longitudinal mean index value of the subtree C_k, we must have (see [59] for further details),

$$ t = s = r, \quad \sum_{j\in A_k} v_{jk} = u_k = \sum_{i\in I(T)} w_{ik}, \quad \sum_{k\in V(T)} u_k = \sum_{i\in I(T)} w_i. $$

This then leads to the simpler general definition,

$$ M_{node,r}(F)(T;v,w) := \begin{cases} \left(\frac{1}{\sum\limits_{k\in V(T)} u_k} \sum\limits_{k\in V(T)} \sum\limits_{j\in A_k} \frac{v_{jk}}{u_k} \sum\limits_{i\in I(T)} w_k [F(P_j(x))]^r \right)^{\frac{1}{r}}, & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases} $$

The preferred ancestor weights are best expressed as integrals; therefore, we will express the node-wise mean by integrating over depths instead of summing over intervals. The node-wise mean of order r of an index F is then defined as,

$$ M_{node,r}(F)(T;v,w) := \begin{cases} \left(\frac{1}{\sum\limits_{k\in V(T)} u_k} \sum\limits_{k\in V(T)} \sum\limits_{j\in A_k} \frac{v_{jk}}{u_k} \int_0^h w_k(x) [F(P_j(x))]^r \, dx \right)^{\frac{1}{r}}, & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise,} \end{cases} $$

where $w_k(x)$ is the weight assigned to node k at depth x, and

$$ u_k = \int_0^h w_k(x) \, dx. $$

Then we set $w_k = S_{C_k}$, and

$$\sum_{j \in A_k} v_{jk} = u_k = \int_0^h w_k(x)\, dx.$$

Finally, we define,

$${}^qD_N := M_{node,0}({}^qD),$$

and

$${}^qJ_N := M_{node,0}({}^qJ).$$

Star mean

The star mean of order r of F is defined as,

$$M_{star,r}(F)(T; w^*) := \begin{cases} \left(\frac{\int_0^h w^*(x)[F(P^*(x))]^r\, dx}{\int_0^h w^*(x)\, dx} \right)^{\frac{1}{r}} & \text{if } h > 0 \\ F(\emptyset) & \text{otherwise.} \end{cases}$$

Then with $w^* = S^*$, we define the star mean diversity of order q as,

$${}^qD_S := M_{star,0}({}^qD).$$

We define the star-mean evenness index of order q as,

$${}^qJ_S := M_{star,0}({}^qJ).$$

C Chapter 3 supplementary material

Expectation of $\frac{X}{X+Y}$

Given two random variables X and Y, the probability that $X > Y$ is given by,

$$\mathbb{P}(X > Y) = \int_{-\infty}^{\infty} \int_{y}^{\infty} f_X(x) f_Y(y)\, dxdy.$$

Let $X \sim \exp(\alpha)$, $Y \sim \exp(\beta)$ and,

$$Z = \frac{X}{X+Y}.$$

As $0 \leq Z \leq 1$, have,

$$\begin{aligned} F_Z(z;\alpha,\beta) = \mathbb{P}(Z \leq z) &= \mathbb{P}\left(Y \geq \frac{1-z}{z} X\right) \\ &= \int_0^{\infty} \int_{\frac{1-z}{z}x}^{\infty} \alpha e^{-\alpha x} \beta e^{-\beta y}\, dydx \\ &= \int_0^{\infty} \alpha e^{-x(\alpha+\beta\frac{1-z}{z})}\, dx \\ &= \frac{\alpha z}{\alpha z + \beta(1-z)}. \end{aligned}$$

Finally,

$$f_Z(z;\alpha,\beta) = \frac{d}{dz} F_Z(z;\alpha,\beta) = \frac{\alpha\beta}{(\alpha z + \beta(1-z))^2}$$

Expectation for 3-leaf tree

We have,

$${}^1D_N = \begin{cases} \exp\left(\frac{2}{3}z\left(\log(3) - 2\log(2)\right) + \frac{1}{3}(\log(3) + 2\log(2))\right), & \text{for} \quad z \geq \frac{1}{2}, \\ \exp\left(\frac{1}{3}(3-2z)\log(3)\right), & \text{for} \quad z < \frac{1}{2} \end{cases}$$

then letting,

$$\begin{aligned} A &= \frac{2}{3}\log(3) - \frac{4}{3}\log(2), \\ B &= \frac{1}{3}\log(3) + \frac{2}{3}\log(2), \\ C &= \frac{1}{3}\log(3) \end{aligned}$$

we can write,

$$\begin{aligned} \mathbb{E}[{}^1D_N] &= \mathbb{E}[\exp(Az+B)] + \mathbb{E}[\exp(3C-2Cz)] \\ &= \exp(B)\mathbb{E}[\exp(Az)] + \exp(3C)\mathbb{E}[\exp(-2Cz)]. \end{aligned}$$

Then by the law of the unconscious statistics,

$$\begin{aligned} \mathbb{E}[{}^1D_N] &= \exp(B)\int_{\frac{1}{2}}^{1} \exp(Az) f_z \, dz + \exp(3c)\int_0^{\frac{1}{2}} \exp(-2Cz) f_z \, dz \\ &= \exp(B)\int_{\frac{1}{2}}^{1} \frac{\alpha\beta\exp(Az)}{(\alpha z+\beta(1-z))^2}\, dz + \exp(3C)\int_0^{\frac{1}{2}} \frac{\alpha\beta\exp(-2Cz)}{(\alpha z+\beta(1-z))^2}\, dz \\ &\approx 2.18 \end{aligned}$$

Yule process variation

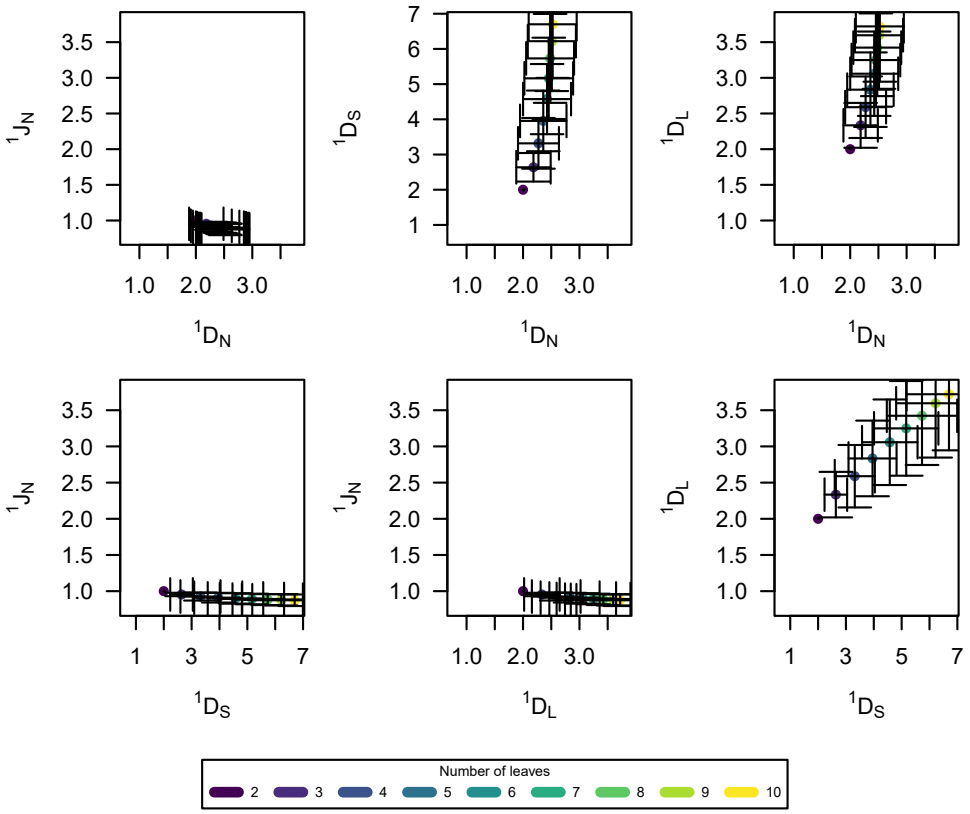

Figure 8.1: Index trajectories formed by four indices, ${}^1J_N$, ${}^1D_N$, ${}^1D_S$, ${}^1D_L$, for trees generated by the Yule process for varying numbers of leaves, n. The birth rate is $\lambda = 0.03$, and each index value is the average of 1000 trees. The error bars show the standard deviation of the respective value.

D Chapter 4 supplementary material

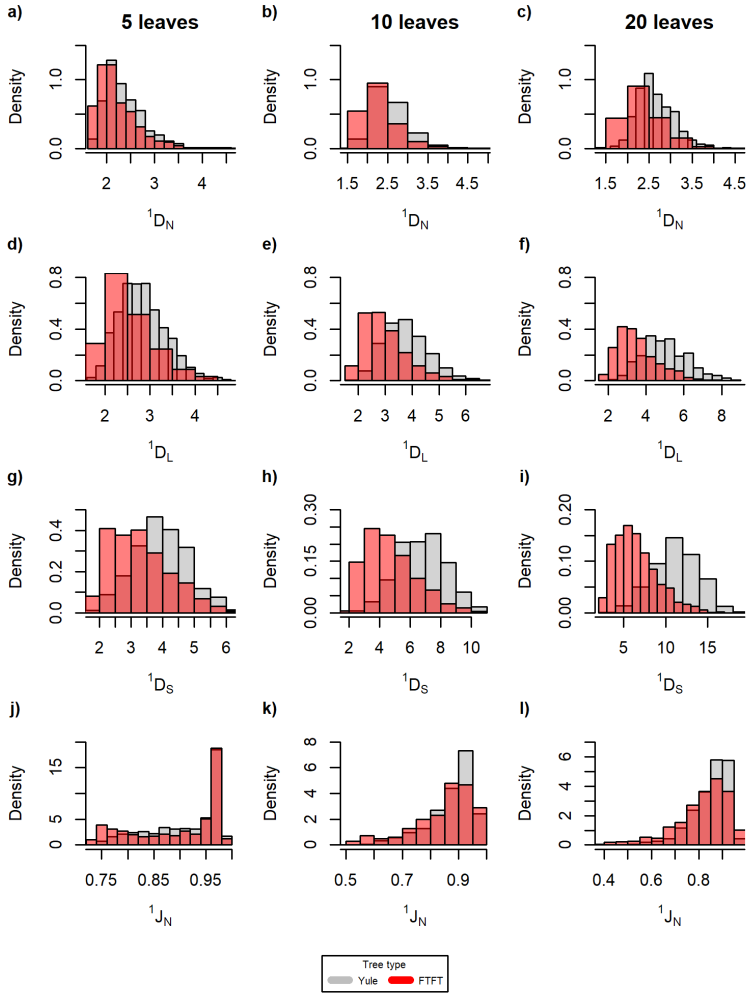

Figure 8.2: Distribution of index values for 1000 either Yule trees or FTFT trees with a,d,g,j) 5 leaves, b,e,h,k) 10 leaves and c, f, i,l) 20 leaves. The branching rate increase is 5 for all, and the time at which the branching rate change occurs is 20, 30 and 40 for trees with 5, 10, and 20 leaves respectively.

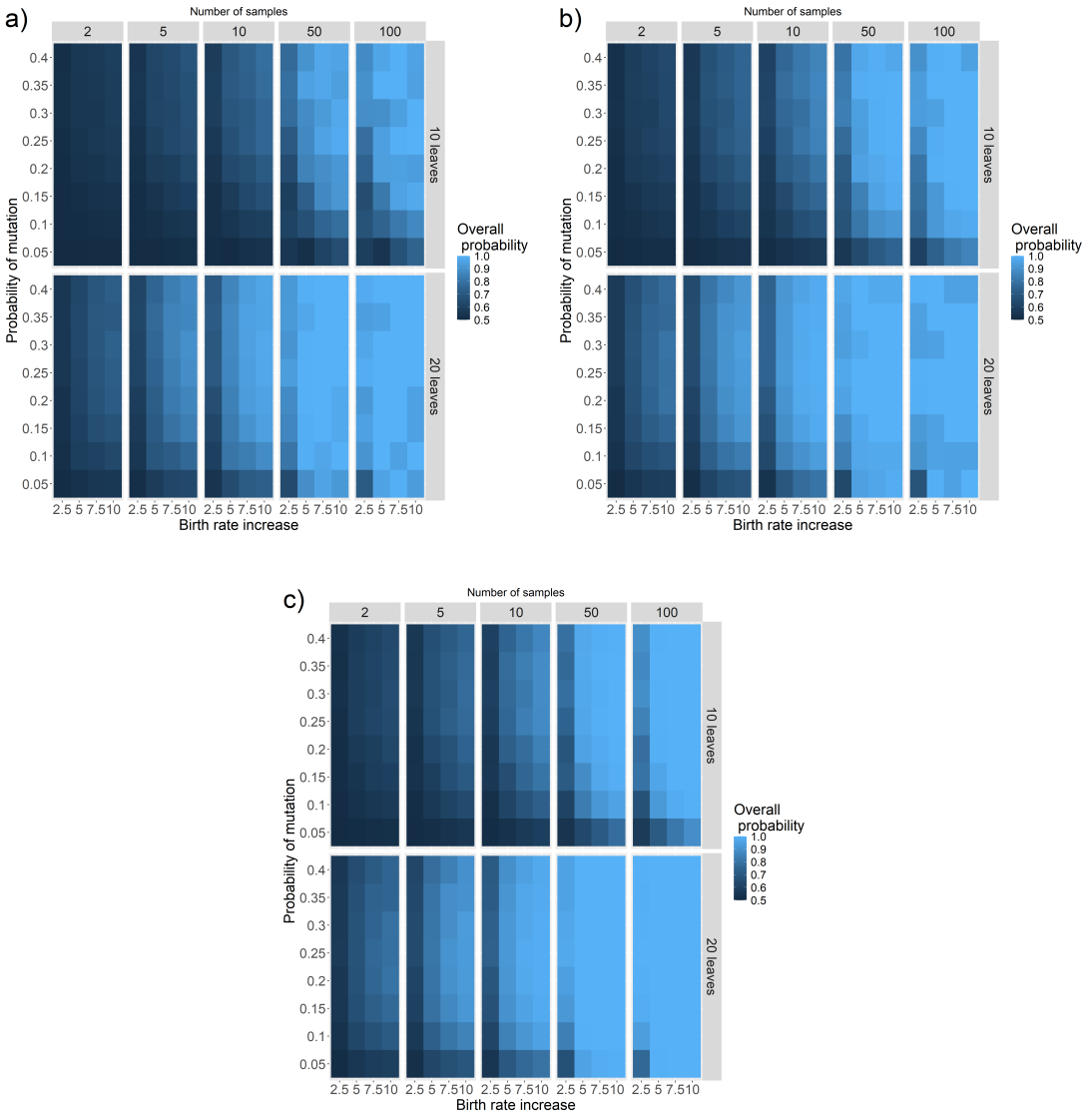

Figure 8.3: Heat maps showing the overall probability of a) ${}^1D_N$, b) ${}^1D_L$ and c) ${}^1D_S$, correctly detecting branching rate heterogeneity under the RTRB model. The heat map illustrates the impact of varying the mutation probability, the increase in branching rate resulting from the mutation, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the overall probability.

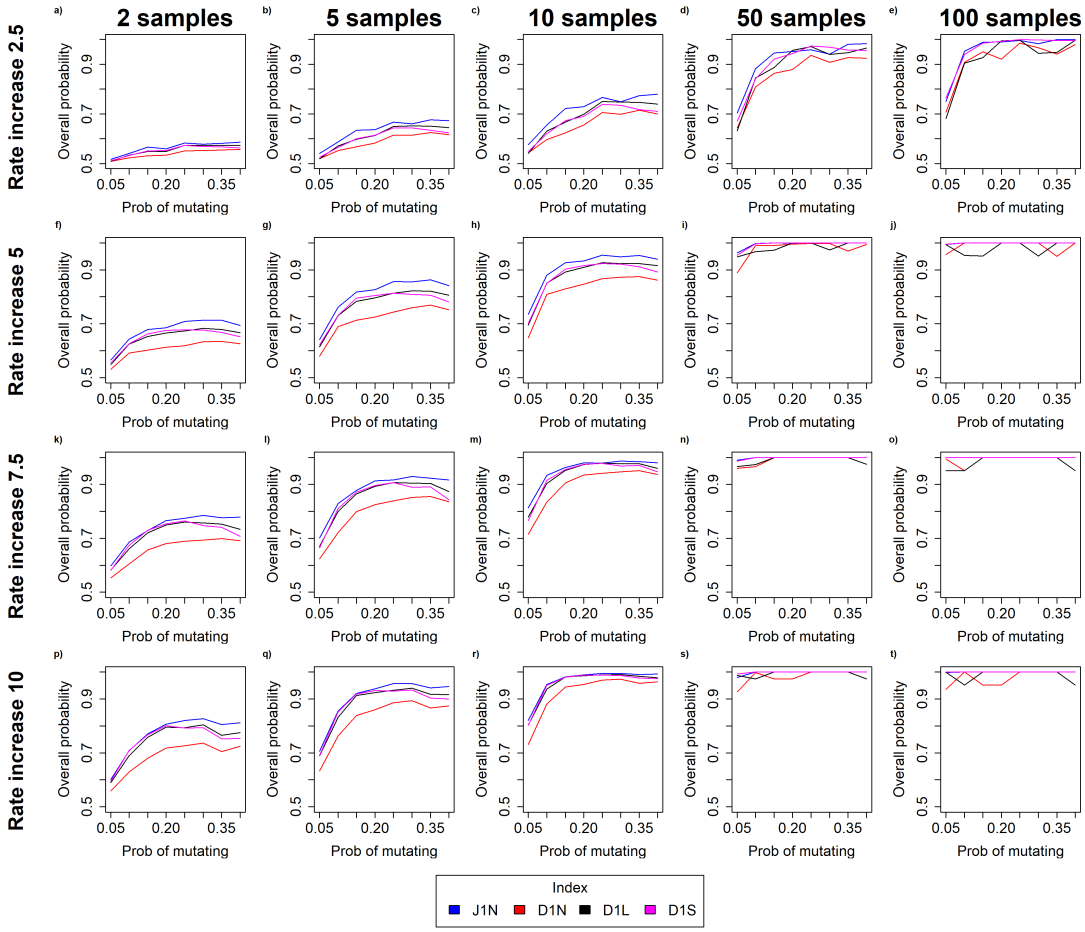

Figure 8.4: Overall probability of our indices correctly detecting branching rate heterogeneity under the RTRB model. The x-axis represents the probability of mutating. Columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and rows correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

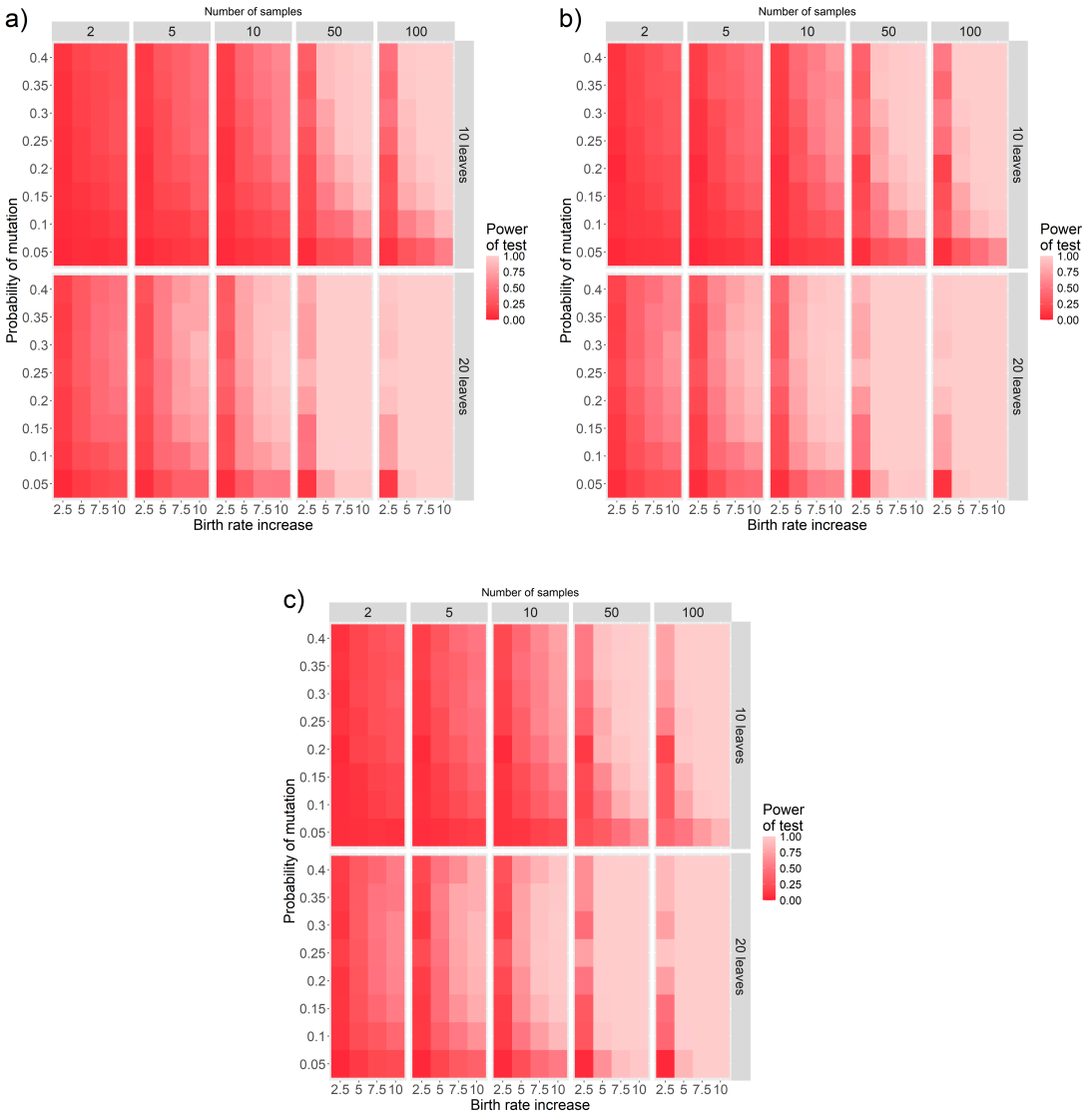

Figure 8.5: Heat maps showing the power for our indices, a) ${}^1D_N$, b) ${}^1D_L$ and c) ${}^1D_S$, to correctly detect branching rate heterogeneity under the RTRB model. The heat map illustrates the impact of varying the mutation probability, the increase in branching rate resulting from the mutation, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the power.

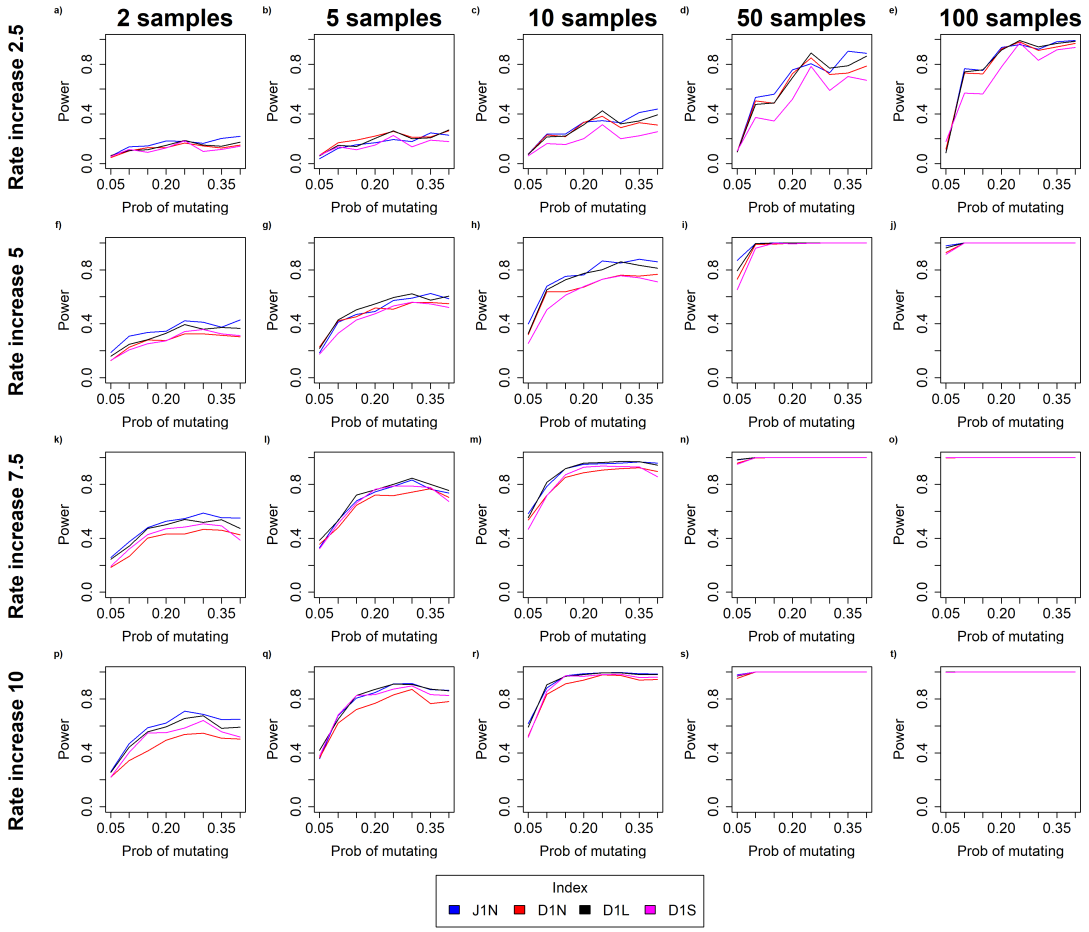

Figure 8.6: Power of our indices to correctly detect branching rate heterogeneity under the RTRB model for trees with 20 leaves. The x-axis represents the probability of mutating. Columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and rows correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

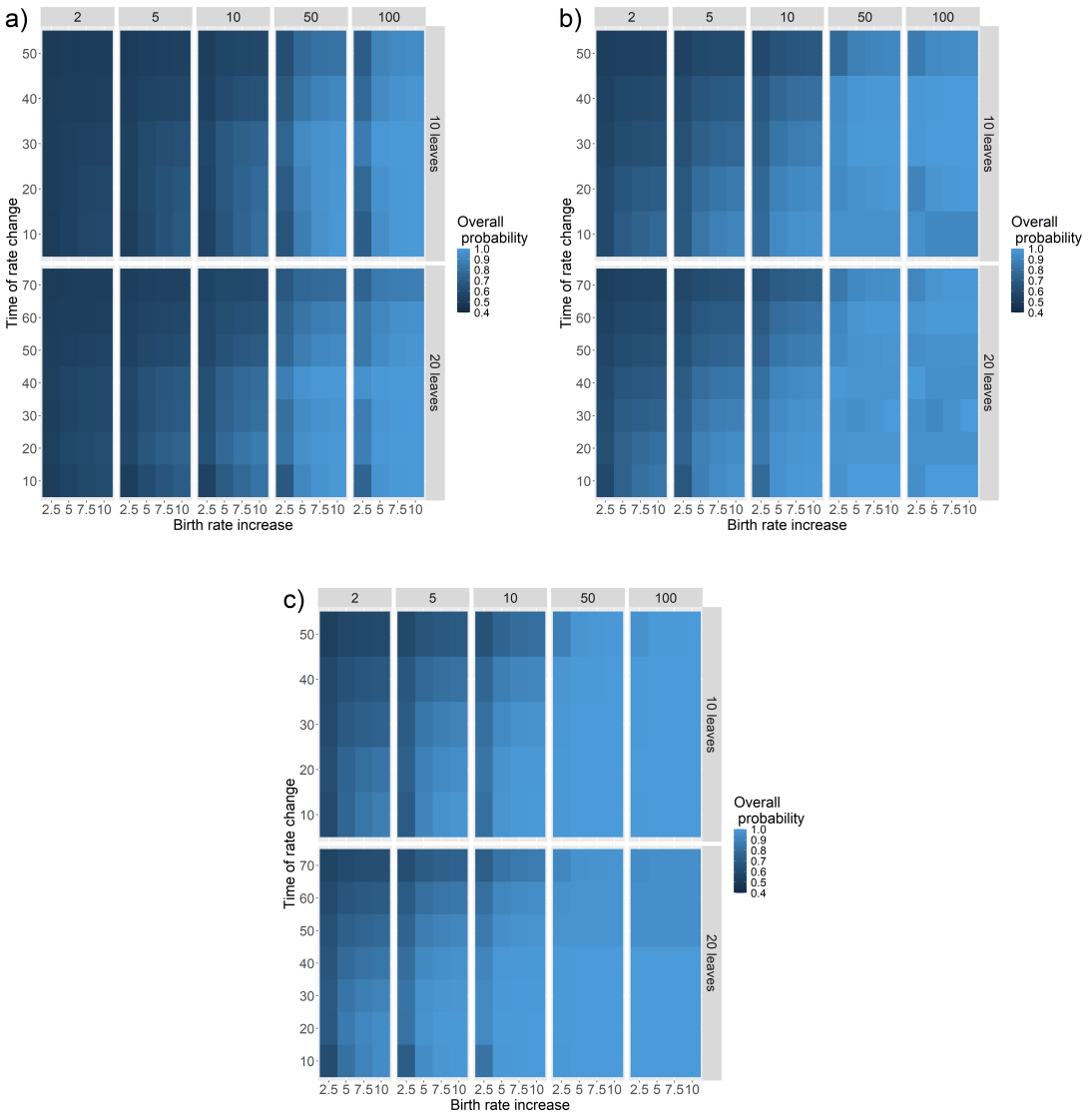

Figure 8.7: Heat maps showing the overall probability of a) ${}^1J_N$, b) ${}^1D_N$, and c) ${}^1D_L$, correctly detecting branching rate heterogeneity under the FTFT model. The heat map illustrates the impact of varying the time at which the branching rate across the whole tree changes, the increase in branching rate resulting from the mutation, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the overall probability.

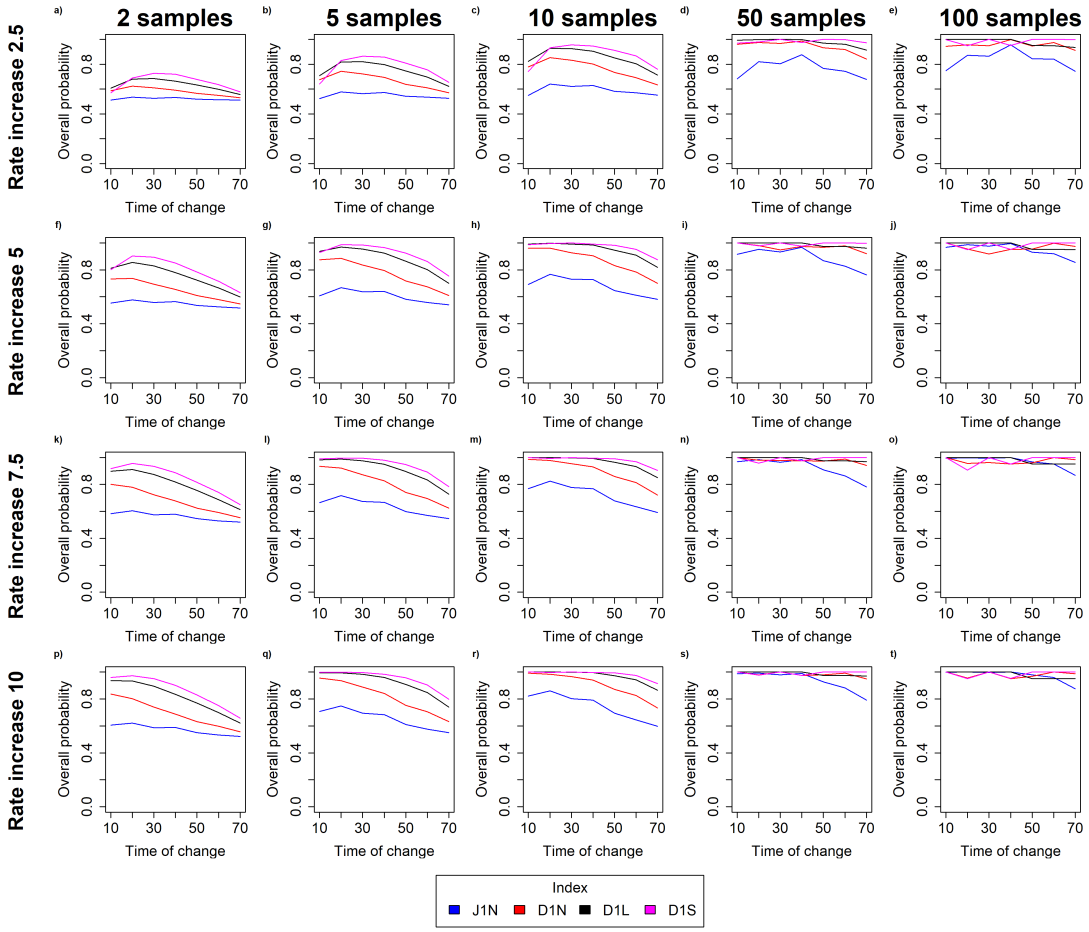

Figure 8.8: Overall probability of our indices correctly detecting branching rate heterogeneity under the FTFT model for trees with 20 leaves. The x-axis represents the time of the branching rate change. Columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and rows correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

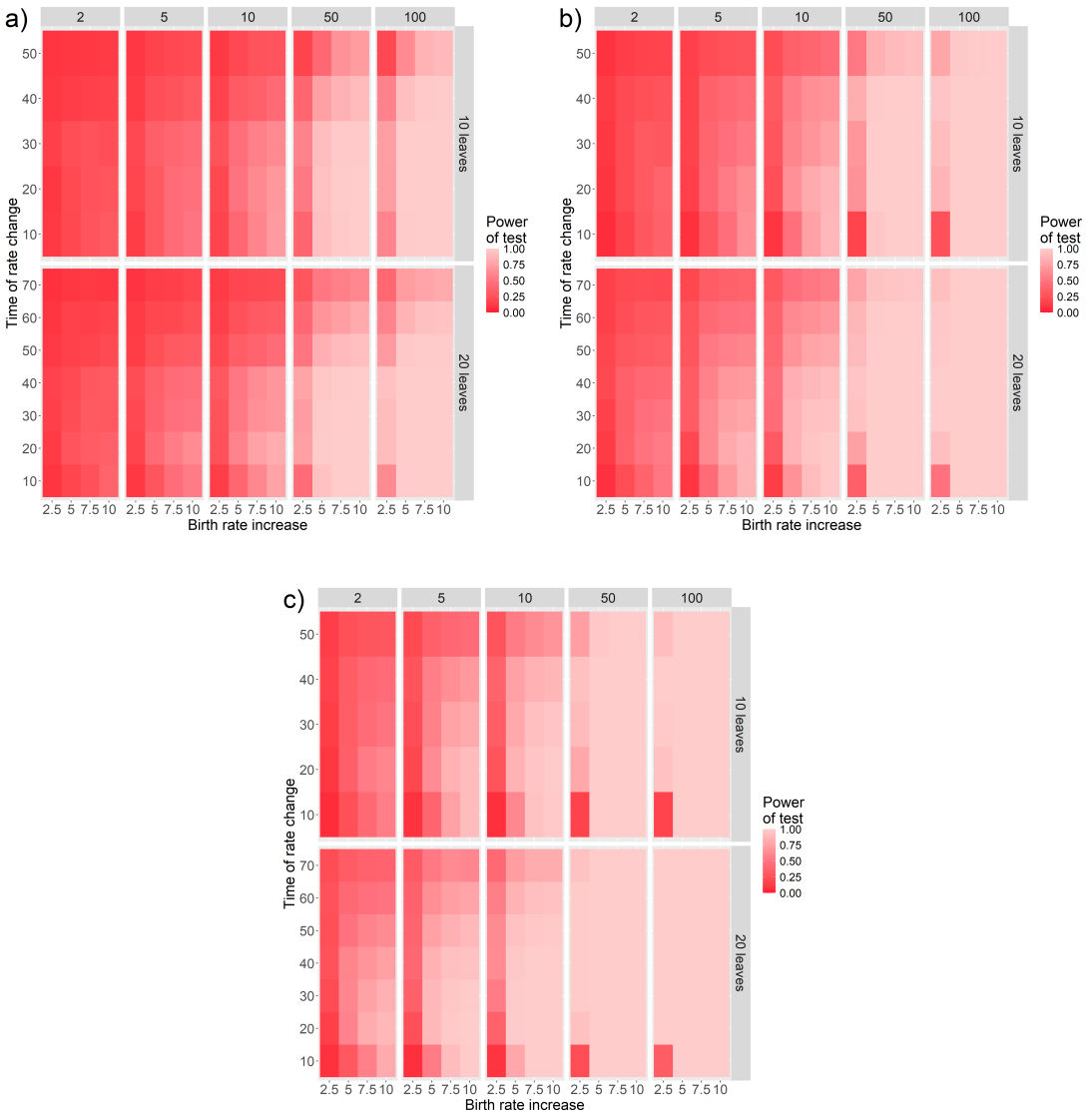

Figure 8.9: Heat maps showing the power of a) ${}^1J_N$, b) ${}^1D_N$, and c) ${}^1D_L$ to correctly detect branching rate heterogeneity under the FTFT model. The heat map illustrates the impact of varying the time at which the branching rate across the whole tree changes, the increase in branching rate resulting from the mutation, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the power.

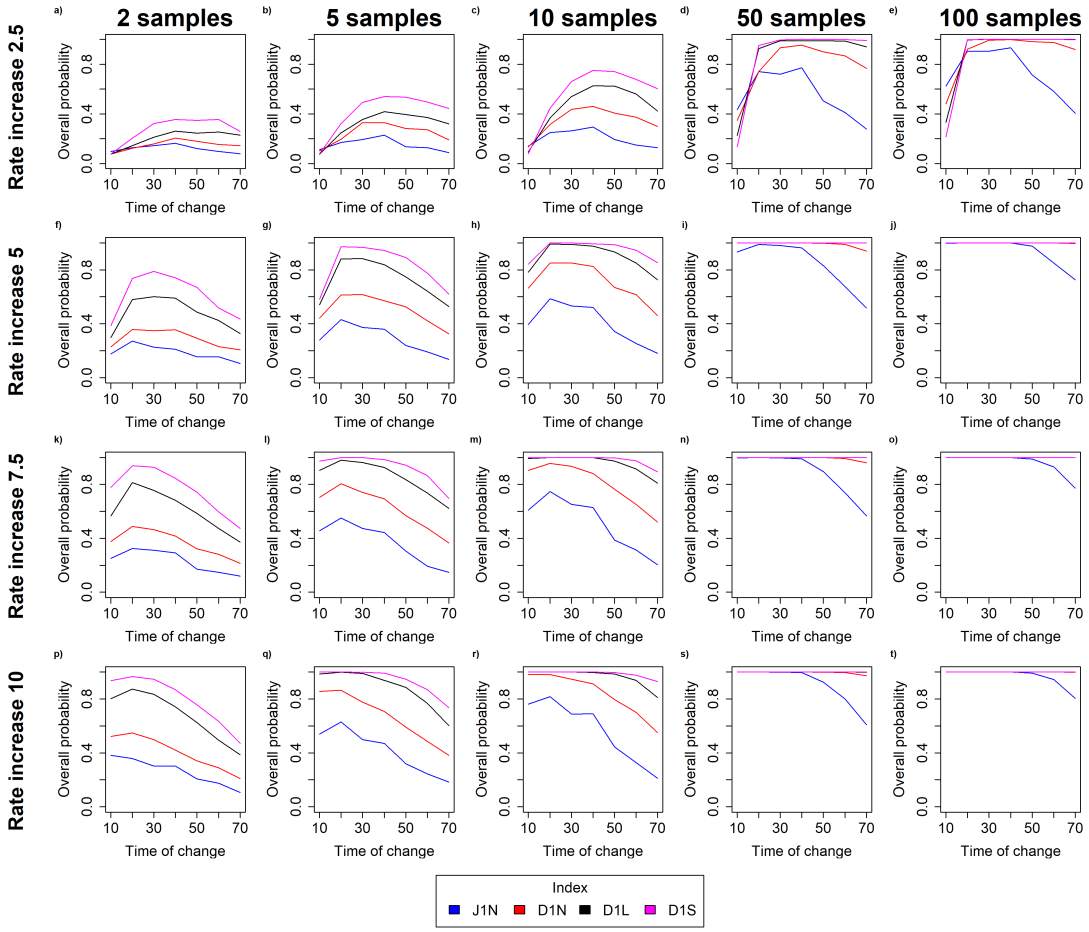

Figure 8.10: Power of our indices to correct detecting branching rate heterogeneity under the FTFT model for trees with 20 leaves. The x-axis represents the time of the branching rate change. Columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and rows correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

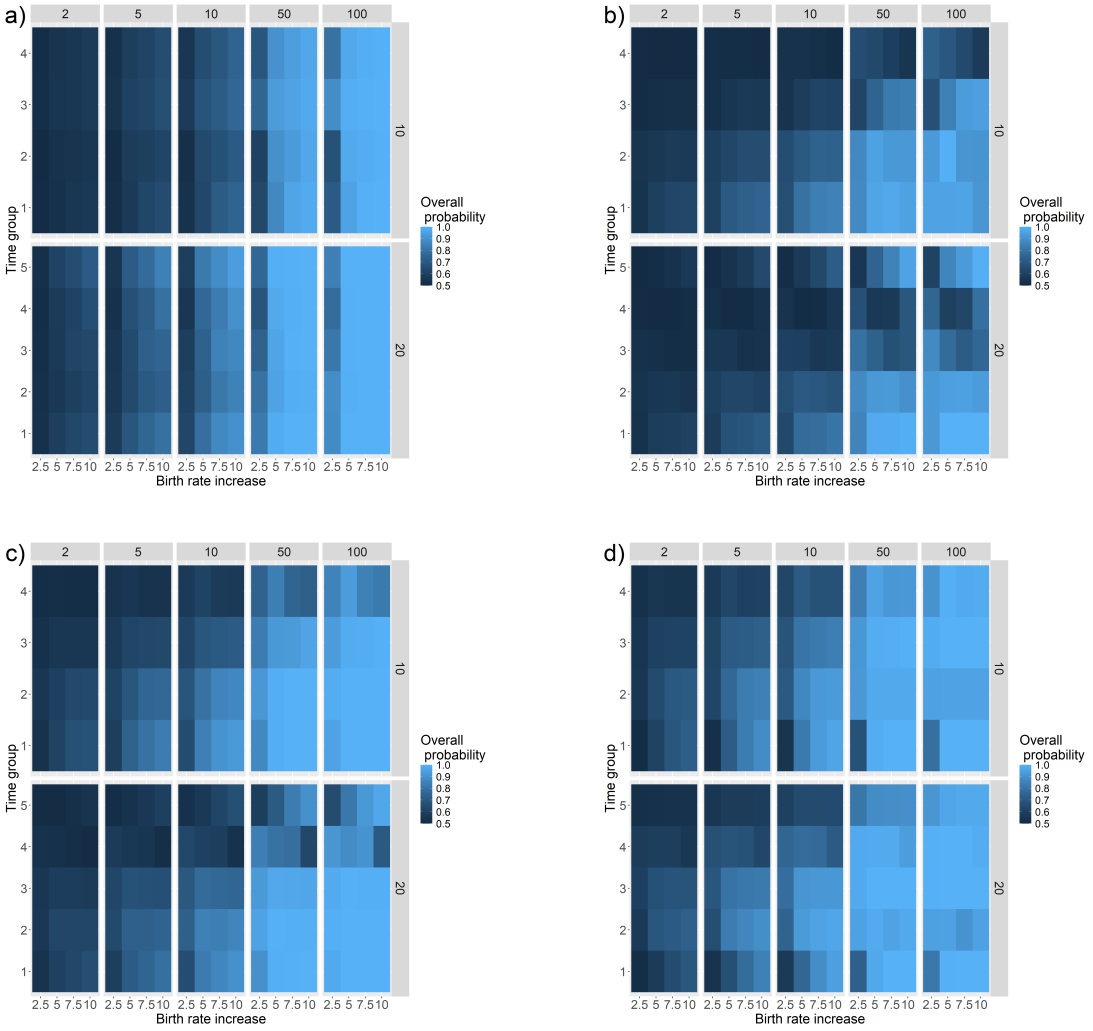

Figure 8.11: Heat maps showing the overall probability of correctly detecting between the two types of branching rate heterogeneity using our indices a) ${}^1J_N$, b) ${}^1D_N$, c) ${}^1D_L$, and d) ${}^1D_S$. The heat map illustrates the impact of varying the time at which the birth rate across the whole tree changes, the increase in birth rate resulting from the mutation, the number of samples (2, 5, 10, 50, or 100), and the number of leaves in the trees (10 or 20) on the overall probability.

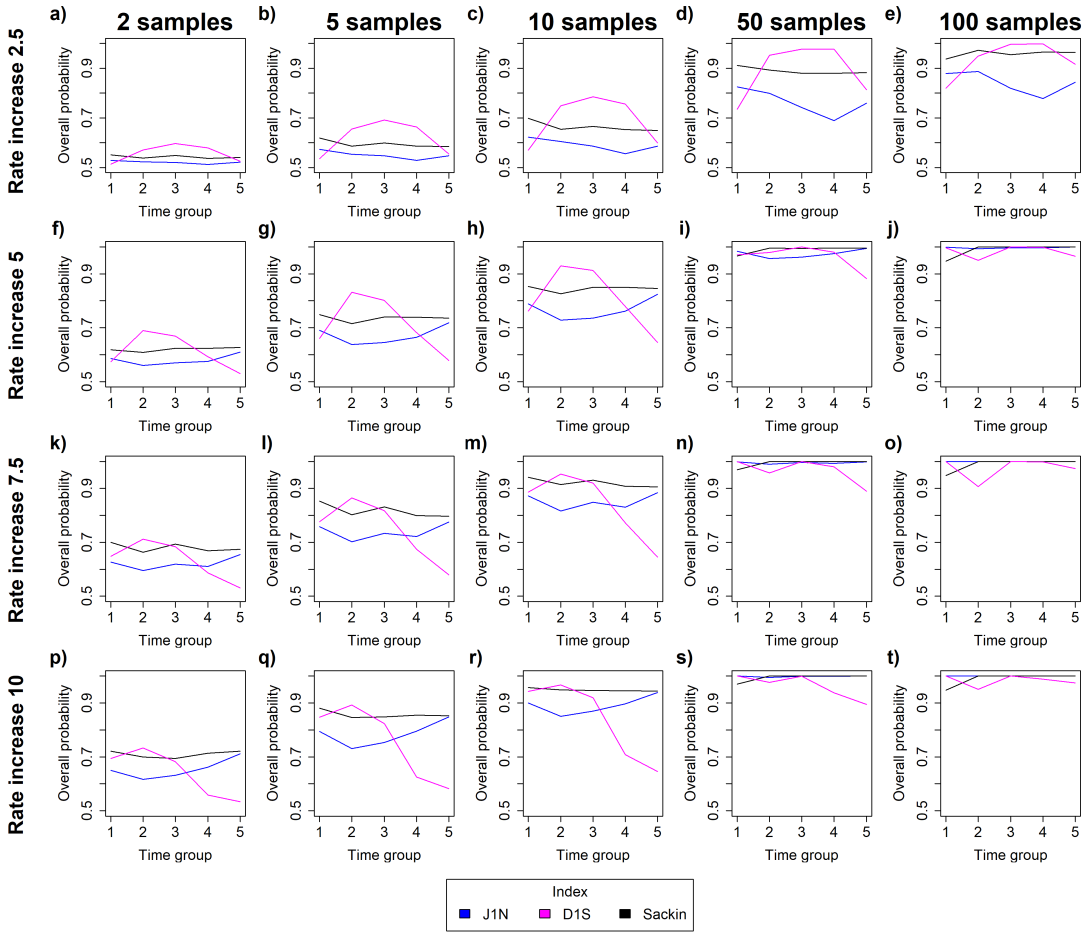

Figure 8.12: Overall probability of ${}^1J_N$, ${}^1D_S$ and Sackin's index correctly detecting between RTRB and FTFT for trees with 20 leaves. The x-axis represents the time of the branching rate change. Columns correspond to the number of samples (2, 5, 10, 50, 100, respectively), and rows correspond to the branching rate increase (2.5, 5, 7.5, 10, respectively).

E Chapter 5 supplementary material

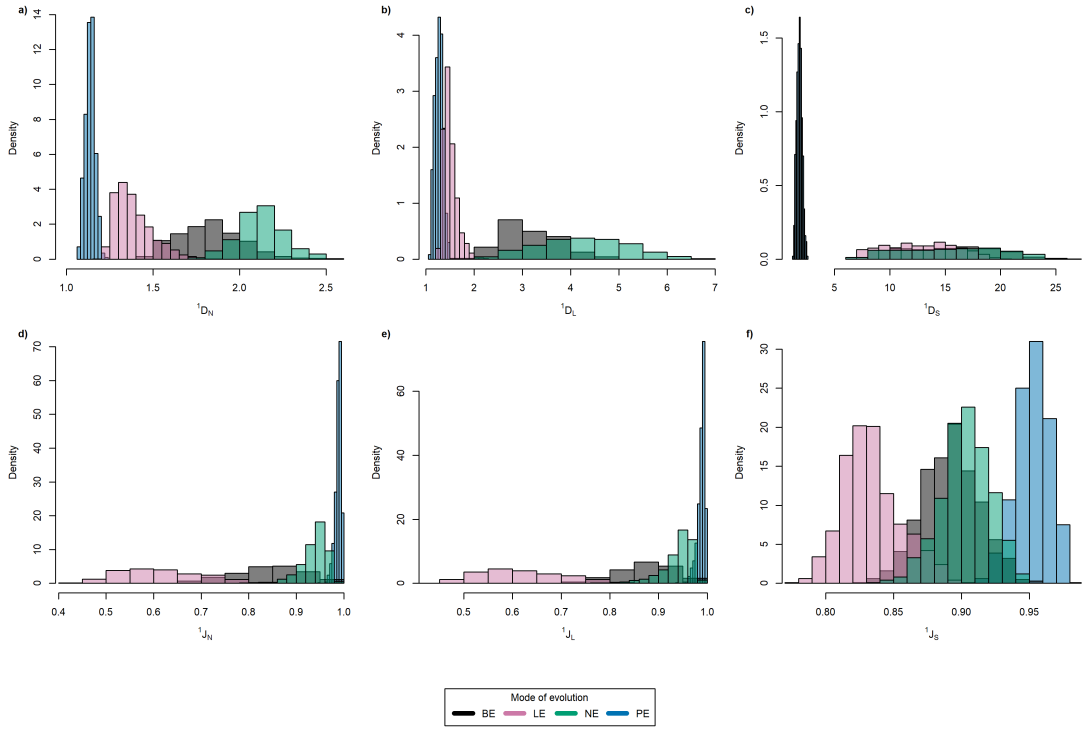

Figure 8.13: Histograms showing the distributions of each index value for the different modes of evolution, which is defined as the pattern. (BE: branching evolution, LE: linear evolution, NE: neutral evolution, PE: punctuated evolution).

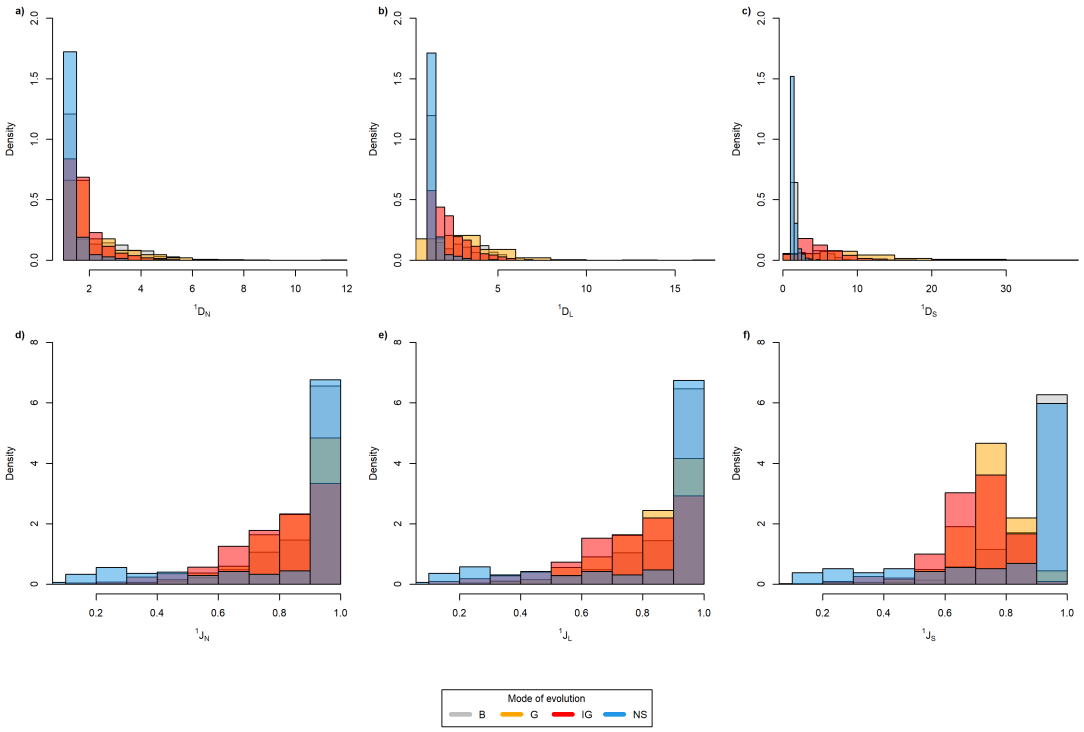

Figure 8.14: Histograms showing the distributions of each index value for the different modes of evolution, where the mode of evolution is the process. (B: boundary growth, G: glandular growth, IG: invasive glandular growth, NS: non-spatial growth).

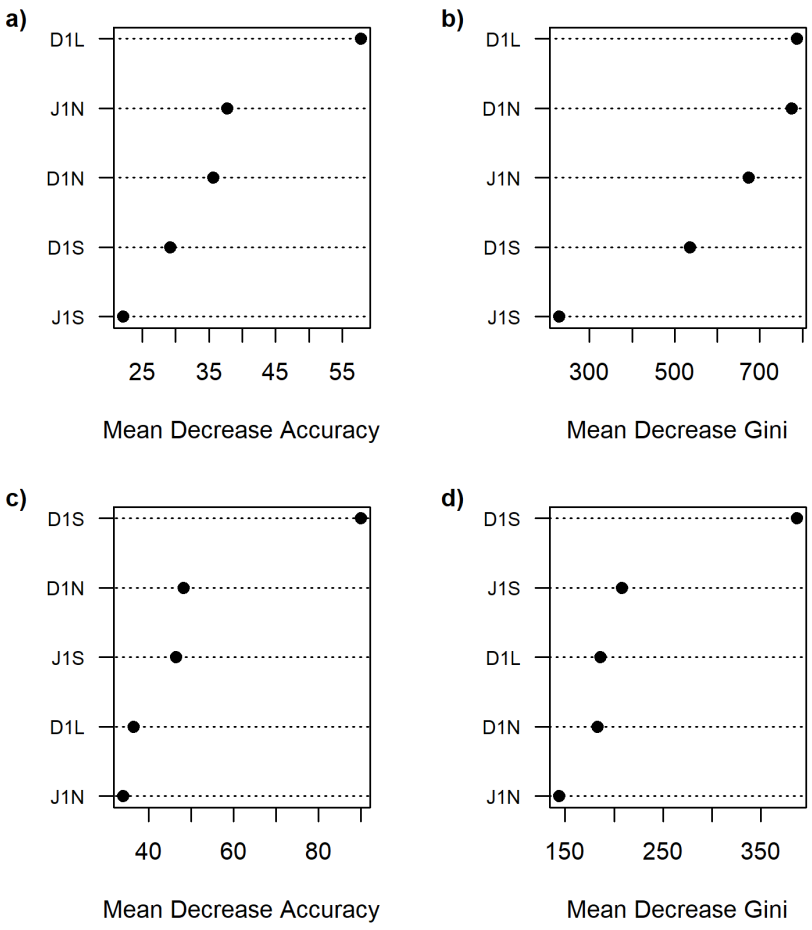

Figure 8.15: Variable importance for random forests on the reduced feature set - only using ${}^1D_N$, ${}^1J_N$, ${}^1D_L$, ${}^1D_S$ and ${}^1J_S$ - where the outcome variable is a-b) the pattern mode of evolution or c-d) the process mode of evolution.

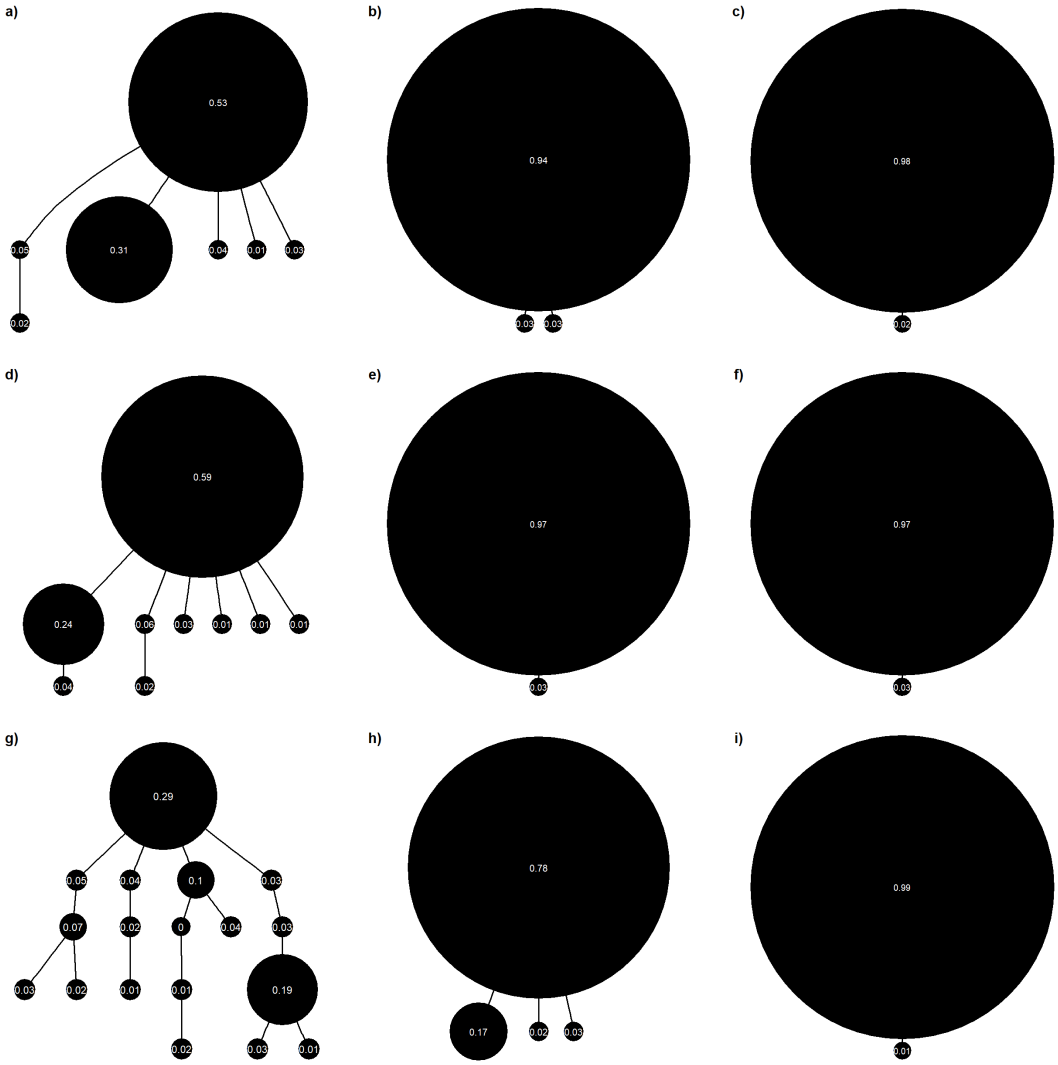

Figure 8.16: Boundary growth trees. Rows correspond to different selection coefficients, 0.05, 0.10 and 0.20 top to bottom respectively. Columns correspond to the driver mutation rate, 1×10^{-4}, 1×10^{-5} and 1×10^{-6} left to right respectively.

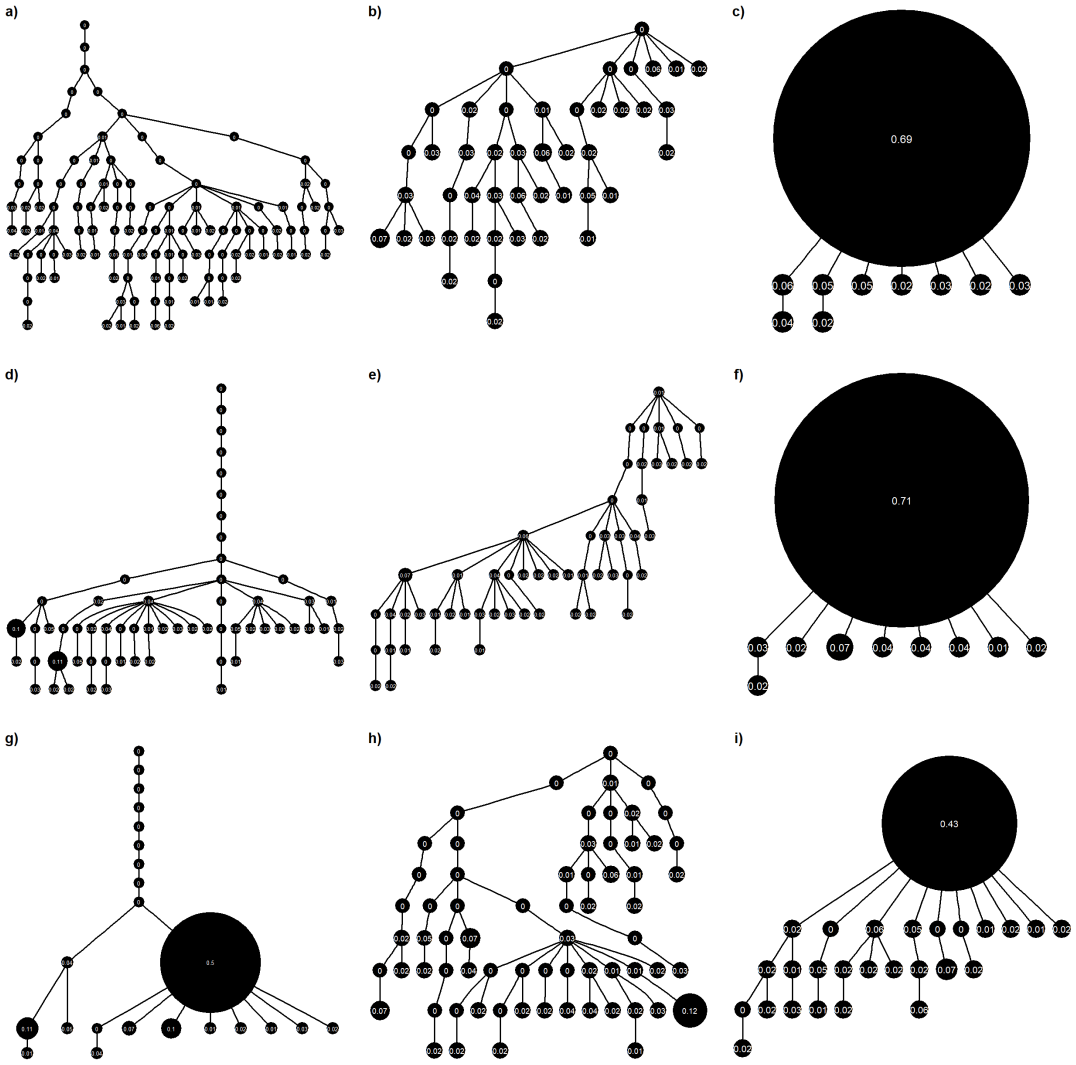

Figure 8.17: Glandular growth trees. Rows correspond to different selection coefficients, 0.05, 0.10 and 0.20 top to bottom respectively. Columns correspond to the driver mutation rate, 1×10^{-4}, 1×10^{-5} and 1×10^{-6} left to right respectively.

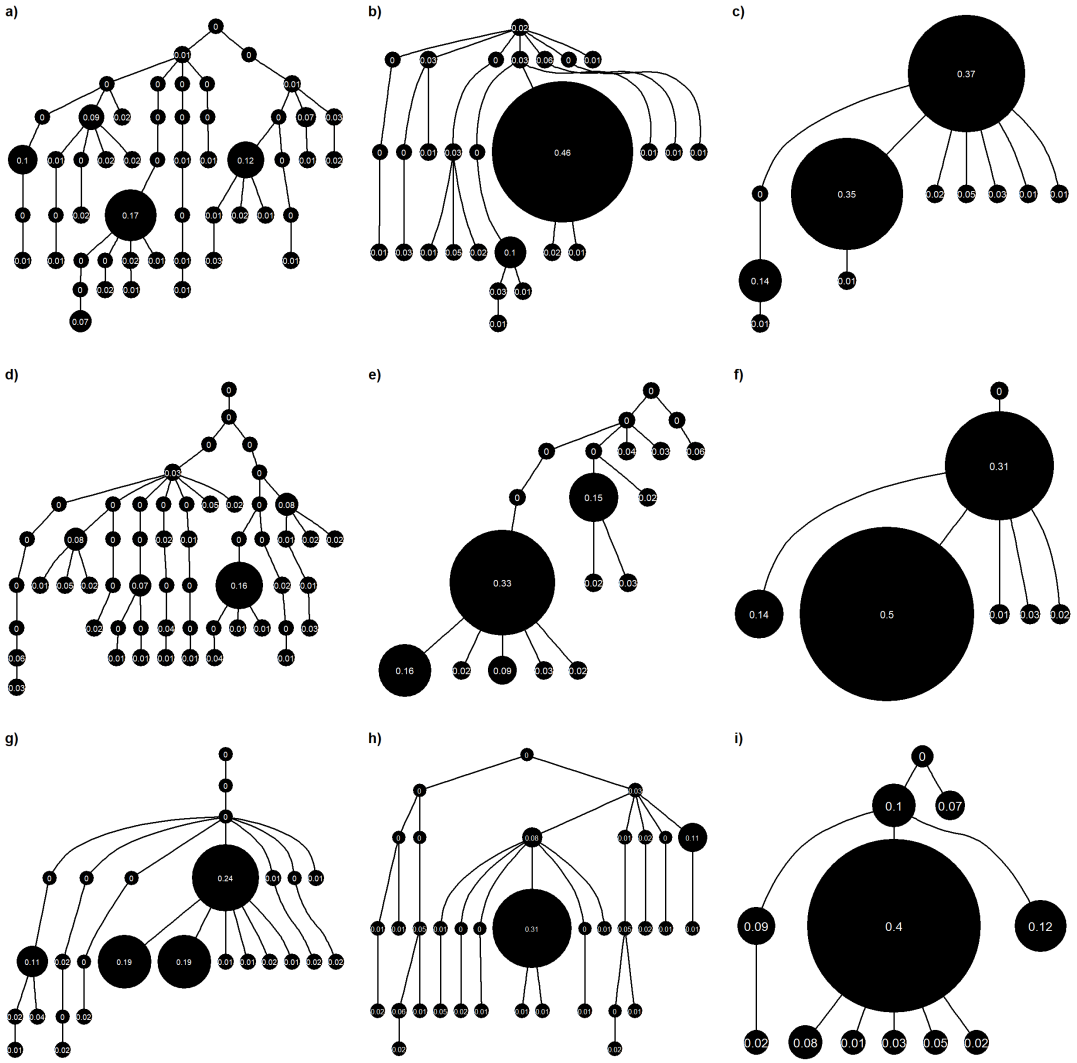

Figure 8.18: Invasive glandular growth trees. Rows correspond to different selection coefficients, 0.05, 0.10 and 0.20 top to bottom, respectively. Columns correspond to the driver mutation rate, 1×10^{-4}, 1×10^{-5} and 1×10^{-6} left to right respectively.

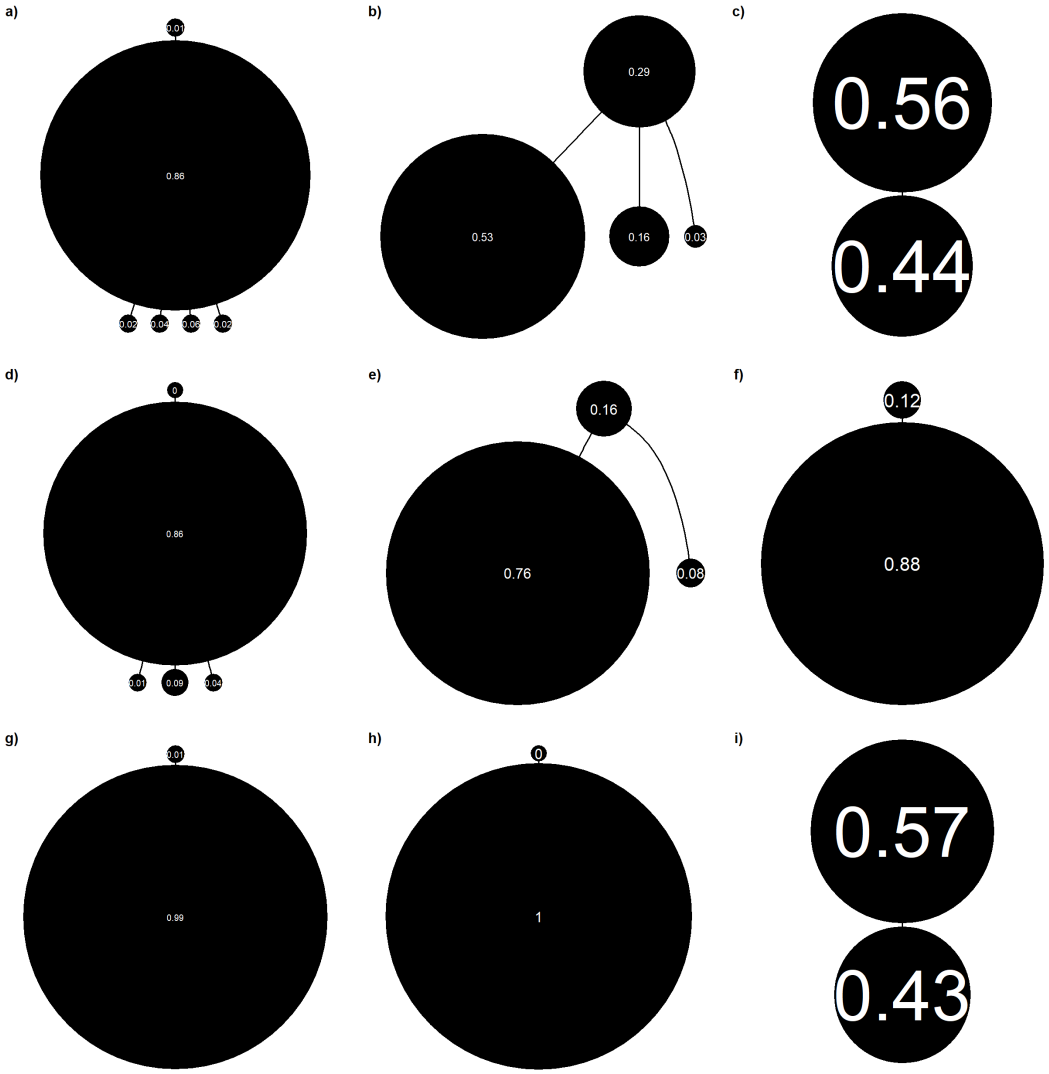

Figure 8.19: Nob-spatial growth trees. Rows correspond to different selection coefficients, 0.05, 0.10 and 0.20 top to bottom respectively. Columns correspond to the driver mutation rate, 1×10^{-4}, 1×10^{-5} and 1×10^{-6} left to right respectively.

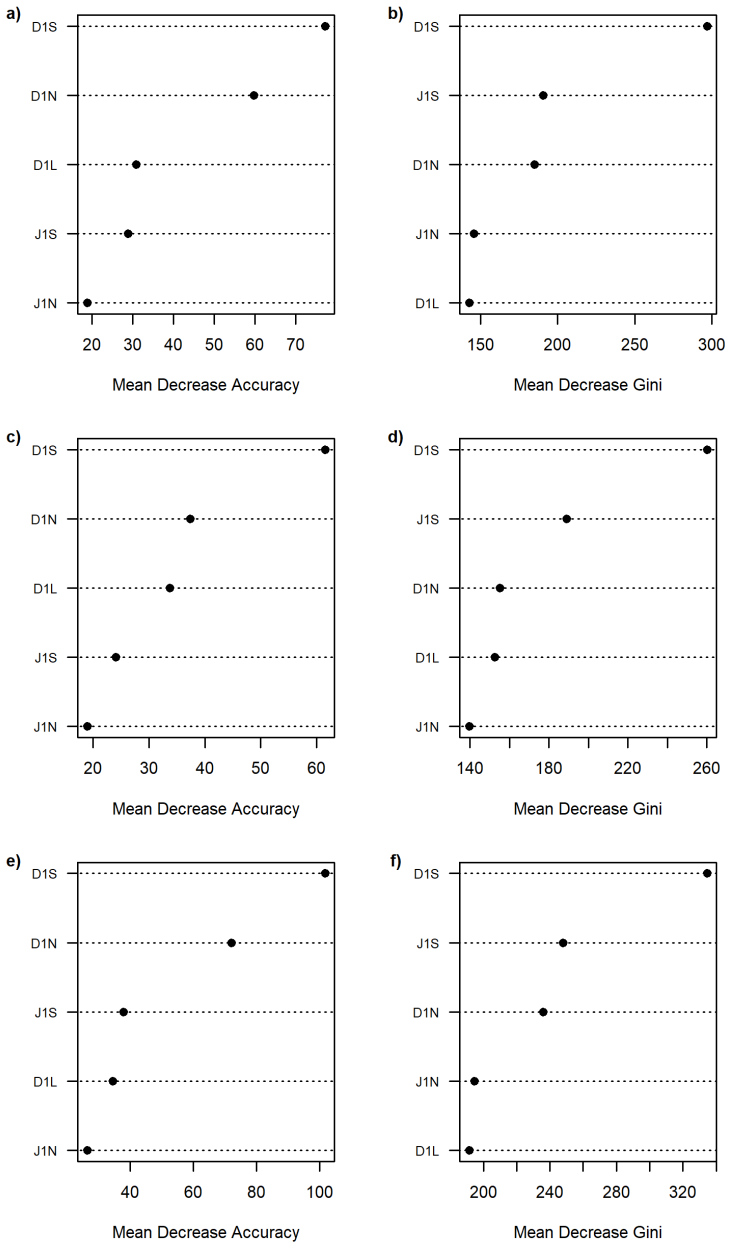

Figure 8.20: Variable importance for the random forest on the reduced feature set - only using ${}^1D_N$, ${}^1J_N$, ${}^1D_L$, ${}^1D_S$ and ${}^1J_S$ - for the process mode of evolution data, where the outcome variable is either a-b) the driver rate, c-d) the selection coefficient or e-f) both.

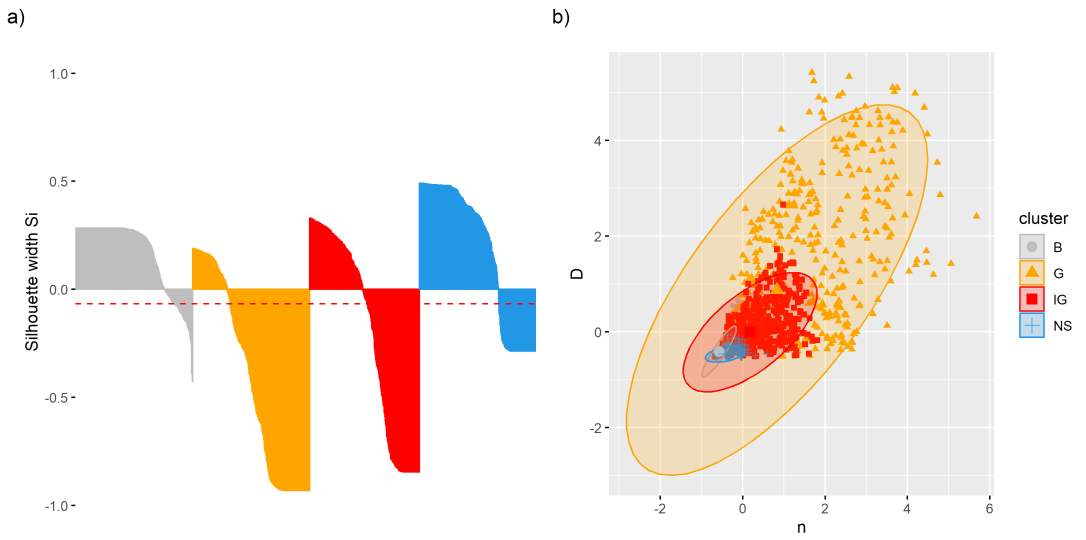

Figure 8.21: Modes of evolution, defined as the process, clustered based on clonal diversity, D, and the mean driver mutations per cell, n. a) Silhouette width for each tree, dashed line is the average silhouette score of -0.067. b) Clustering of trees based on D and n. True evolutionary modes are colour-coded, with ellipses showing the spread of each group. (B: boundary growth, G: glandular growth, IG: invasive glandular growth, NS: non-spatial growth).

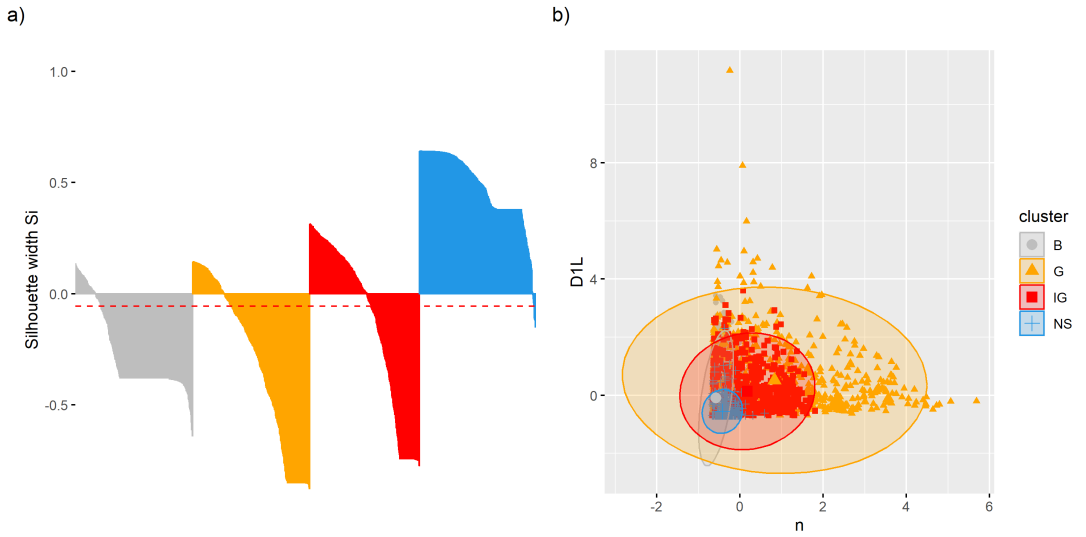

Figure 8.22: Modes of evolution, defined as the process, clustered based on diversity index, ${}^1D_L$, and the mean driver mutations per cell, n. a) Silhouette width for each tree, dashed line is the average silhouette score of -0.067. b) Clustering of trees based on ${}^1D_L$ and n. True evolutionary modes are colour-coded, with ellipses showing the spread of each group. (B: boundary growth, G: glandular growth, IG: invasive glandular growth, NS: non-spatial growth).

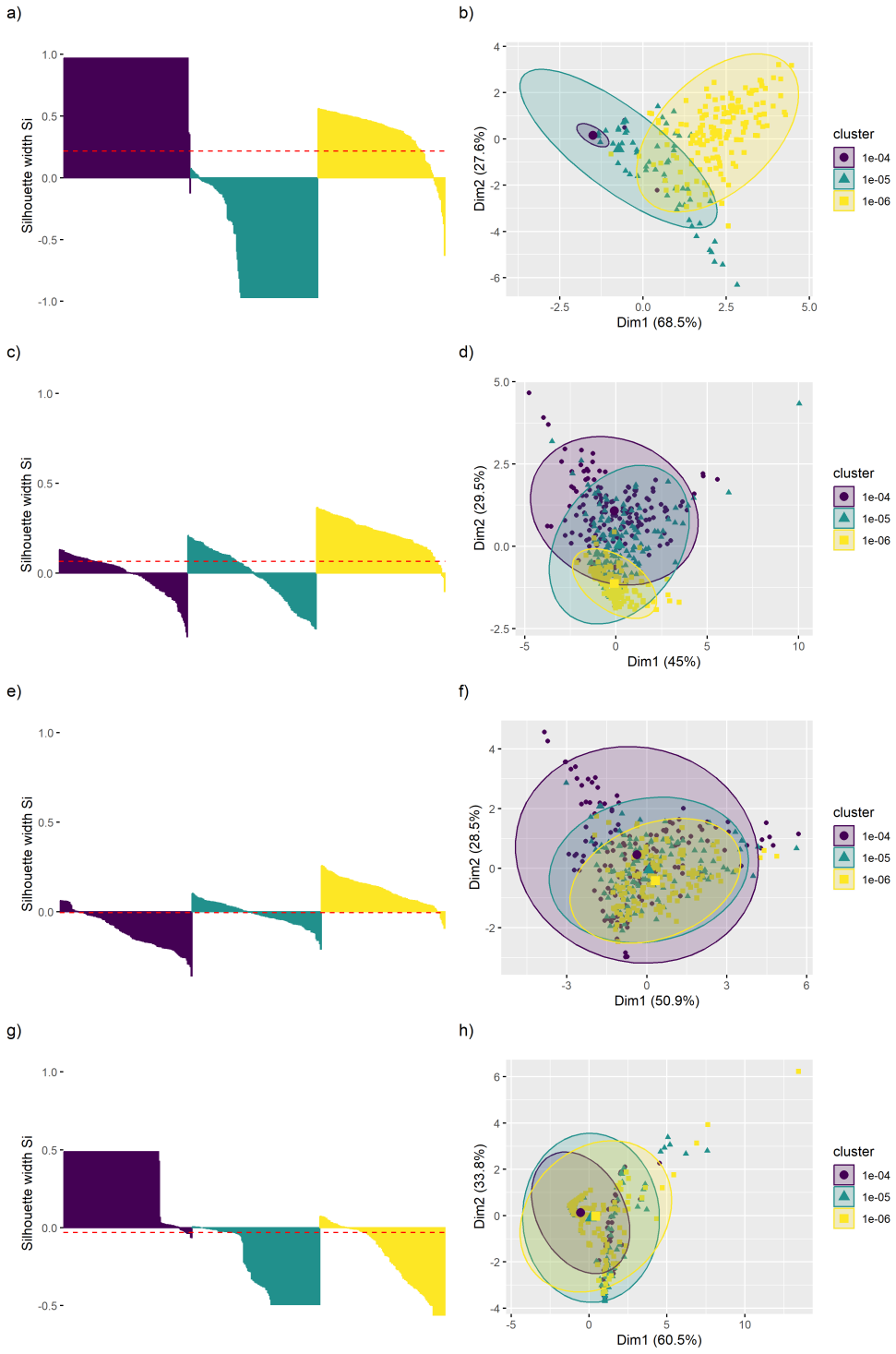

Figure 8.23: Clustering using driver mutation rate within each mode, where mode is defined as the process. a-b) Boundary growth, average silhouette width 0.22, c-d) glandular growth, average silhouette width 0.07, e-f) invasive glandular growth, average silhouette width -0.006, g-h) non-spatial growth, average silhouette width -0.03. Driver mutation rates are colour-coded, with ellipses showing the spread of each group.

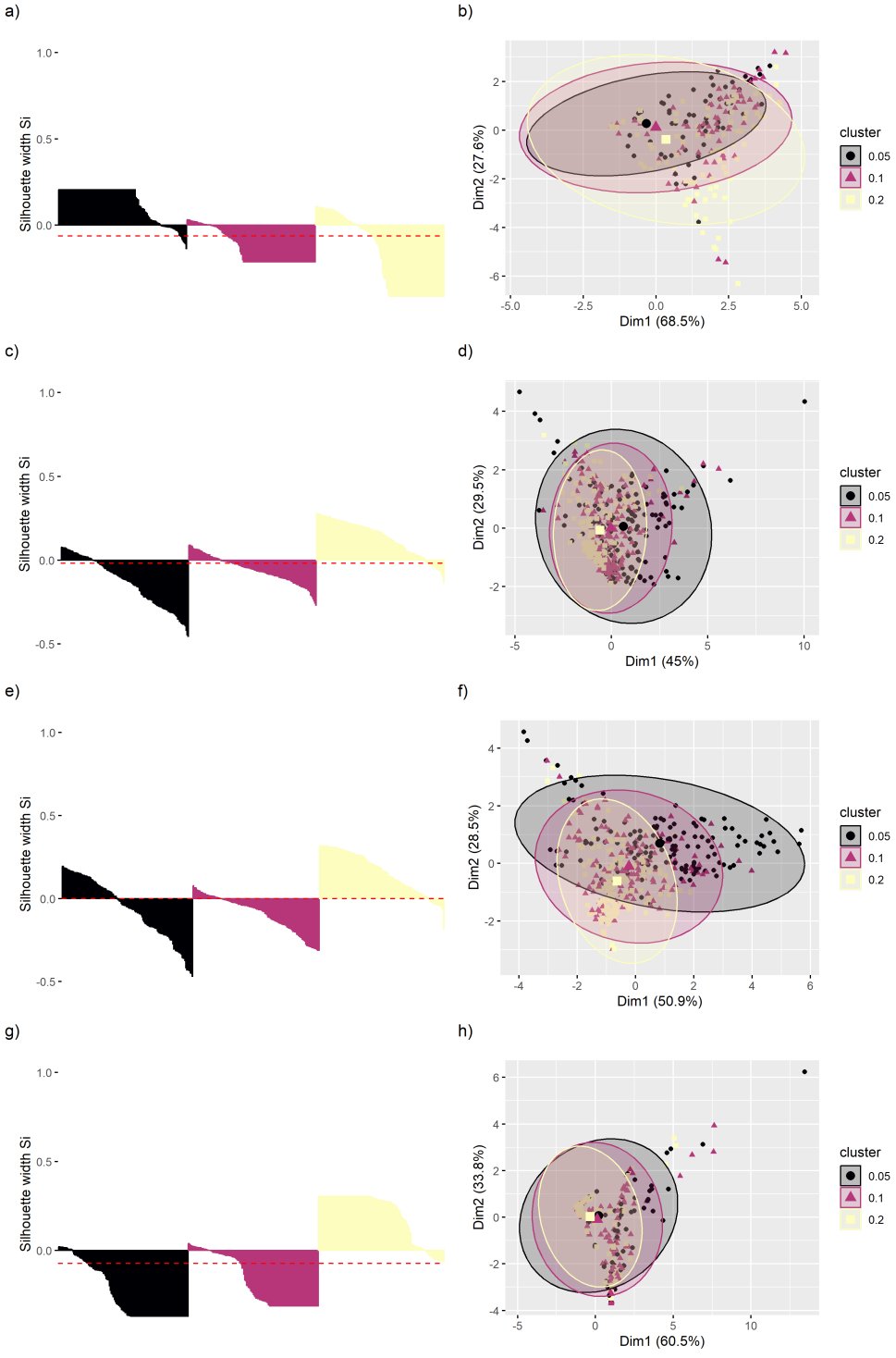

Figure 8.24: Clustering using selection coefficient within each mode, where mode is defined as the process. a-b) Boundary growth, average silhouette width -0.06, c-d) glandular growth, average silhouette width -0.02, e-f) invasive glandular growth, average silhouette width -0.001, g-h) non-spatial growth, average silhouette width -0.07. Selection coefficients are colour-coded, with ellipses showing the spread of each group.

F Chapter 6 supplementary material

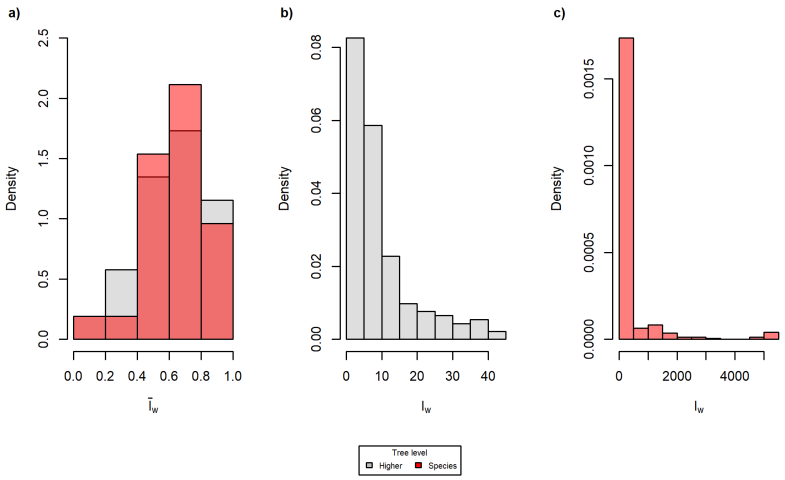

Figure 8.25: Histograms showing the distribution of the I-based indices, a) shows the distribution of tree imbalance $\overline{I}_w$, and b-c) show the distributions of node imbalance I_w.

Clade	Purvis & Agapow				Opentree			
	Higher		Species		Higher		Species	
	$\overline{I}_w$	n	$\overline{I}_w$	n	$\overline{I}_w$	n	$\overline{I}_w$	n
Vertebrates								
Anura	0.8	13	0.613	19	0.748	23	0.667	40
Salamanders	1	3	0.807	6	0.652	4	0.831	9
Ciconiidae	0	1	0.29	4	1	1	0	2
Trogonidae	1	3	0.734	5	0.4	2	0.808	4
Chaenopsidae	0.828	5	0.654	6	0.526	5	0.588	7
Bathyergidae	1	2	0.84	3	1	3	0.499	4
Odontoceti	0.943	8	0.898	10	0.754	17	0.690	24
Phyrnosomatinae	0.714	4	0.656	8	0.412	4	0.456	8
Pleurodira	0.651	8	0.625	13	0.667	5	0.744	11
Squamata	0.711	7	0.215	16	0.783	19	0.689	35
Plants								
Anthemidae	0.724	14	0.676	29	0.571	14	0.641	30
Barnadesioideae	0.619	2	0.954	5	0.449	3	0.658	8
Calenduleae	0.778	3	0.641	7	1	2	0.823	4
Chrozophoreae	0.59	7	0.722	5	0.449	3	0.658	8
Cunoniaceae	0.684	10	0.975	4	0.657	10	0.738	16
Cyclanthaceae	0.645	10	0.69	6	0	1	0.409	2
Hyoscyameae	0.542	6	0.654	4	1	4	0.682	6
Inuleae	0.777	21	0.894	9	0.593	9	0.591	17
Liabeae	0.624	9	0.624	5	1	1	0.854	3
Lythraceae	0.726	10	0.511	5	0.279	9	0.691	18
Nymphaeales	0.875	6	0.916	4	0.889	3	0.909	18
Podalyrieae	0.324	8	0.814	4	0.455	2	0.669	5
Restionaceae	0.649	32	0.834	22	0.628	21	0.554	40
Arthropods								
Augochlorini	0.627	13	0.772	25	0.55	3	0.542	4
Dolichoderinae	0.688	12	0.765	9	0.704	13	0.738	23
Helioconiiti	652	4	0.858	9	0.273	3	0.471	7

Table 8.2: Comparison of the imbalance for trees used by Purvis and Agapow and the trees used in this study. The number of nodes used in the imbalance calculation is given by n.

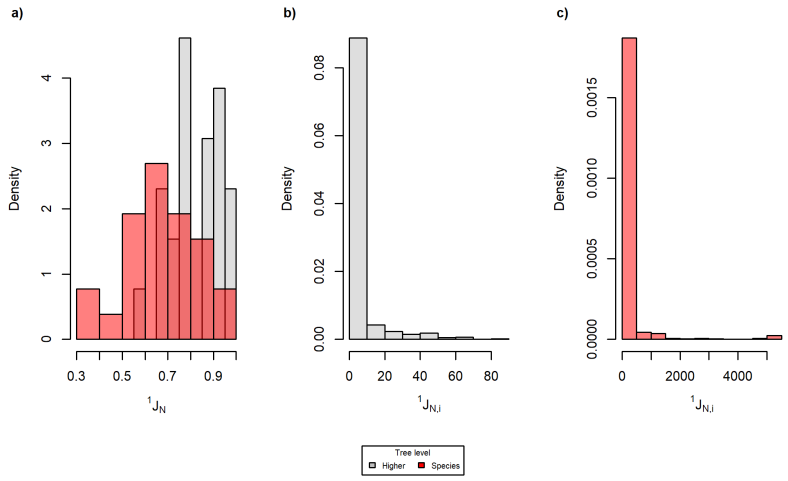

Figure 8.26: Histograms showing the distribution of the J-based indices, a) shows the distribution of tree imbalance ^1J_N, and b-c) show the distributions of node imbalance $^1J_{N,i}$.

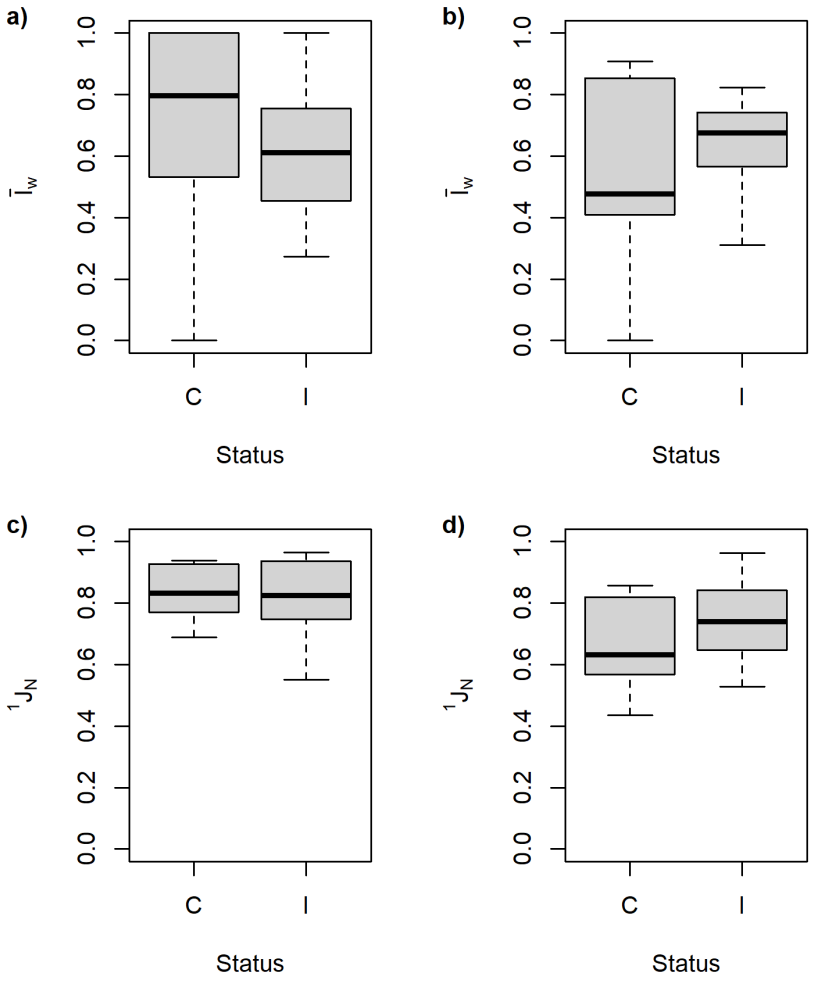

Figure 8.27: Distribution of tree balance values for complete (C) and incomplete (I) trees for a,c) higher-level trees and b,d) species-level trees.

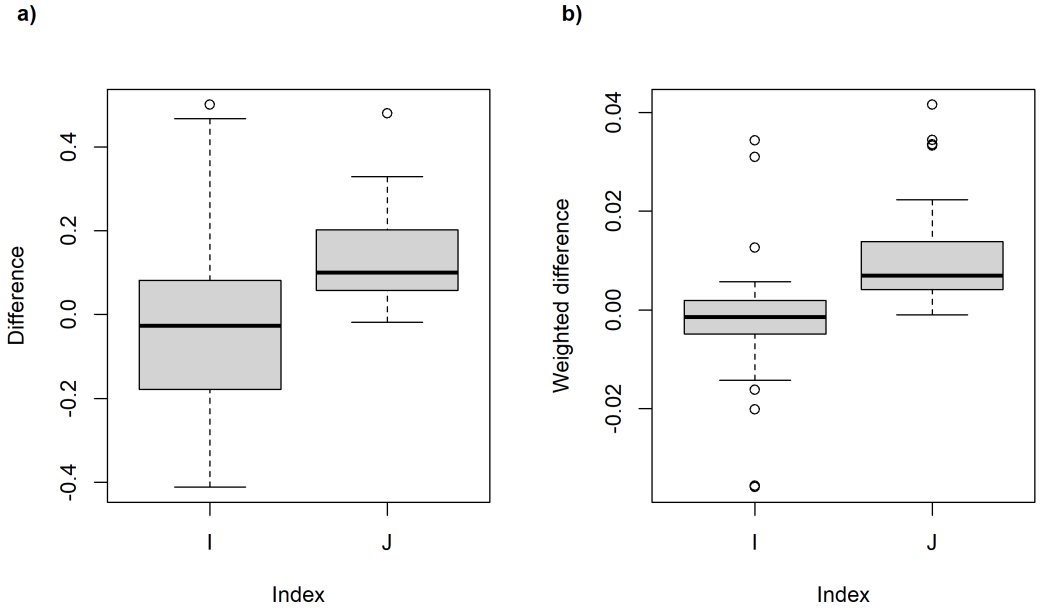

Figure 8.28: Boxplots showing the a) difference and b) weighted difference in $\overline{I}_w$ and ${}^1J_N$ between trees at different levels.

G Chapter 7 supplementary material

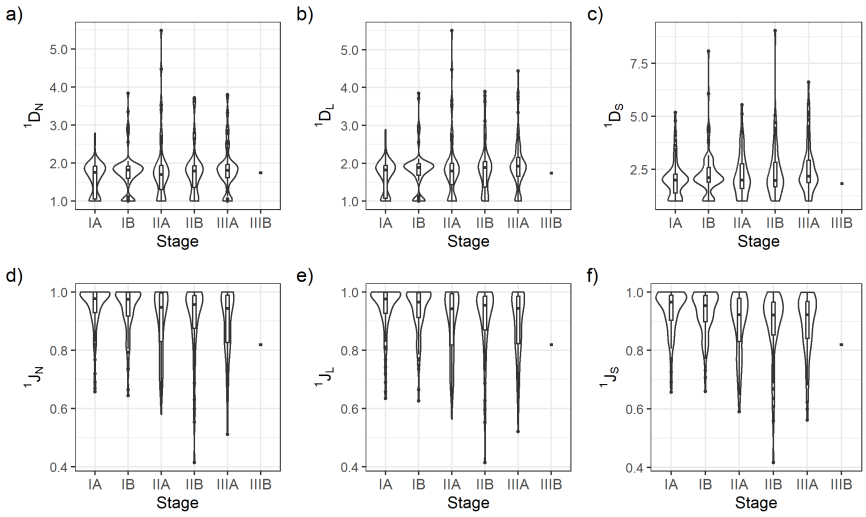

Figure 8.29: Violin and box plots for indices ${}^1D_N$, ${}^1J_N$, ${}^1D_L$, ${}^1J_L$, ${}^1D_S$, ${}^1J_S$ split based on stage. Stage IIIB contains only two patients and hence is not plotted.

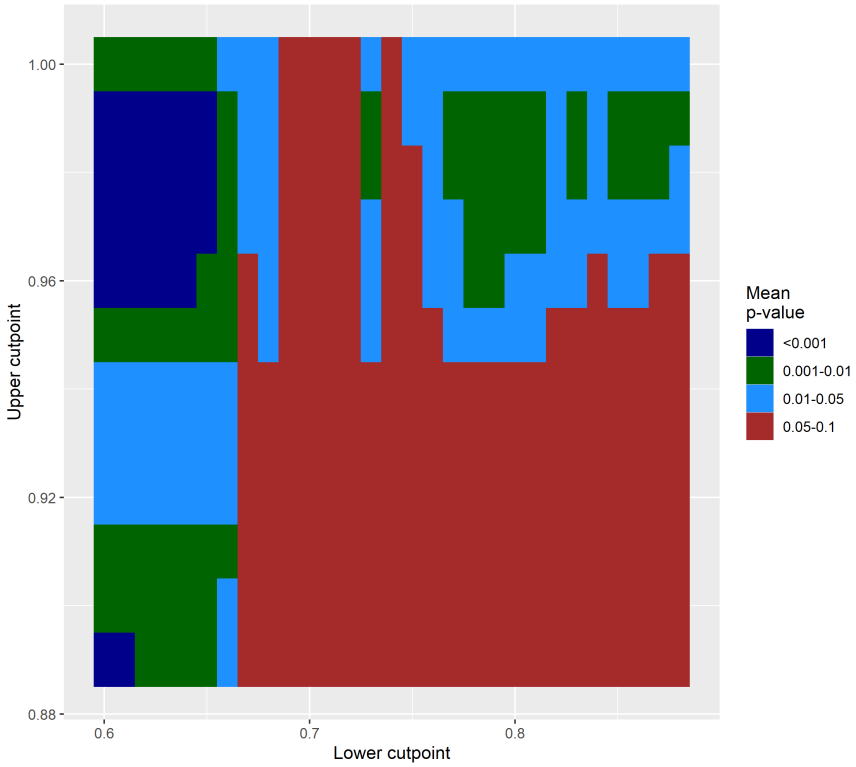

Figure 8.30: Heatmap showing the mean p-value when the lower and upper cutpoints are varied for tree balance, ${}^{1}J_{N}$.

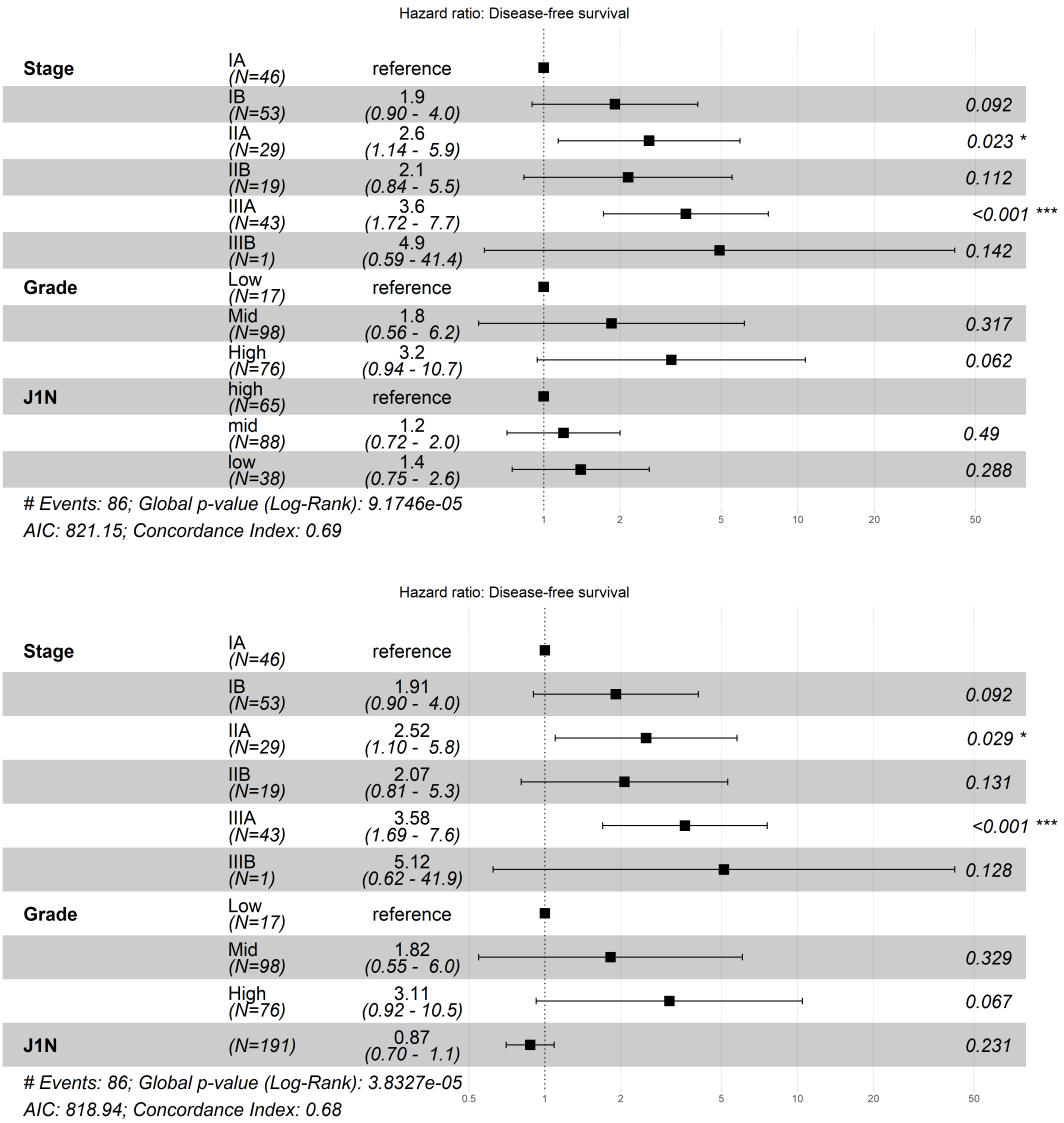

Figure 8.31: Multi-variable Cox proportional hazard models containing stage, grade and tree balance, ${}^1J_N$, a) split into intervals, and b) as a continuous variable. The HR 95% CIs are shown in brackets and by the error bars. The asterisks indicate the P value ranges, where $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

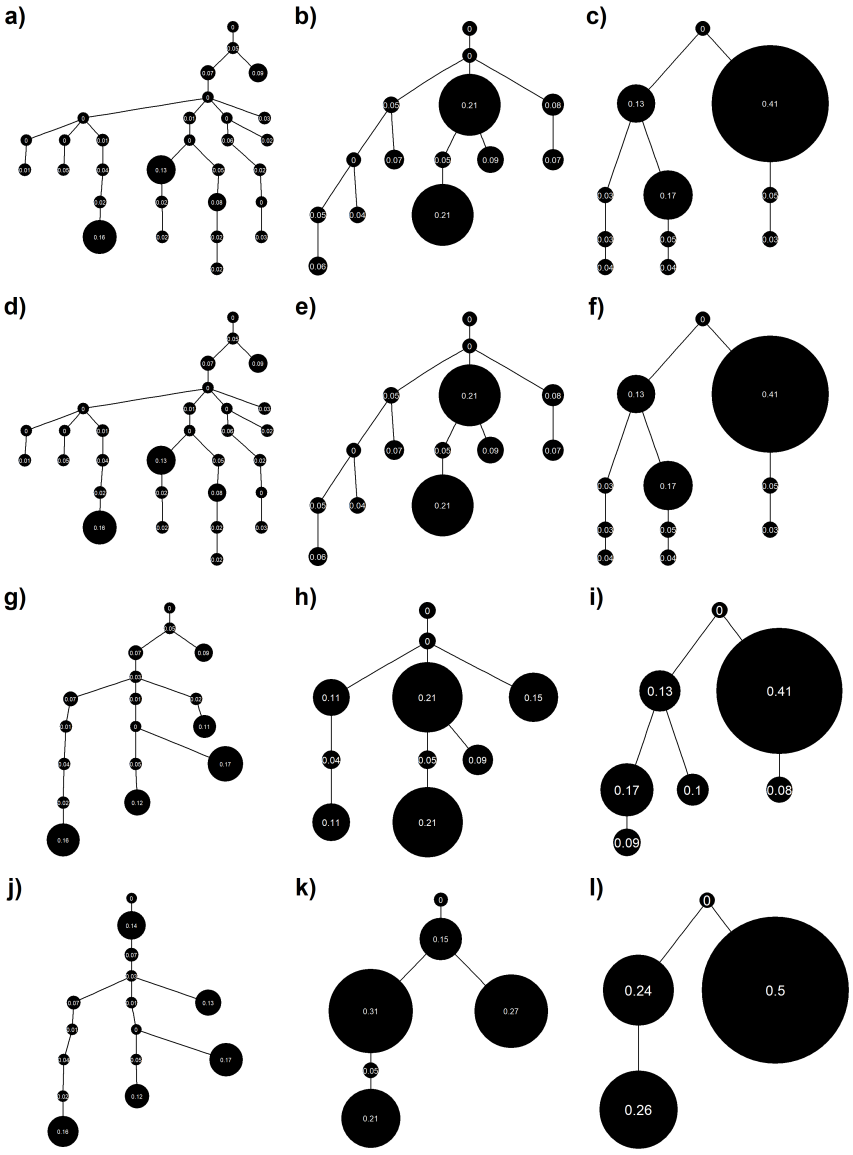

Figure 8.32: Three tumour trees at different levels of coarse-graining. a-c) the original trees, d-f) 1%, g-i) 5% and j-l) 10%. (Tumour IDs CRUK0065, CRUK0462 and CRUK0496 respectively). Trees are shown with proportional node sizes only (branch lengths are arbitrary).

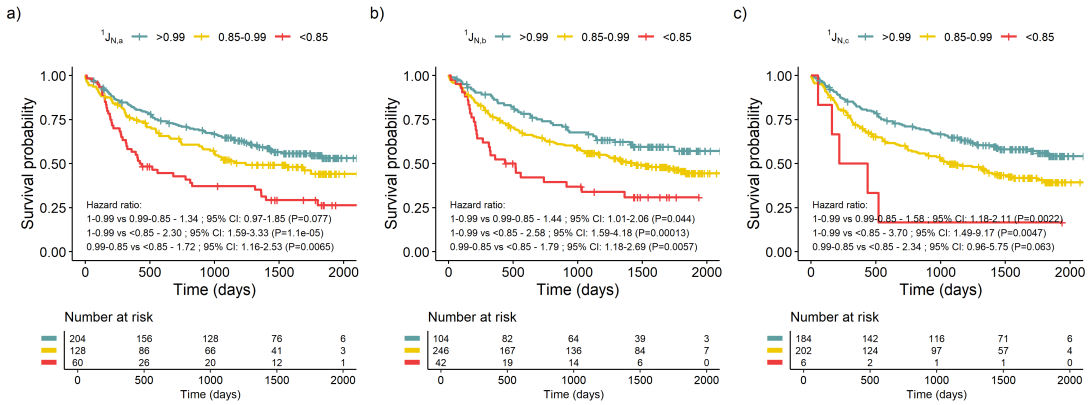

Figure 8.33: Survival curves showing the difference in DFS for tumours based on alternative tree shape indices with the original cut-points.

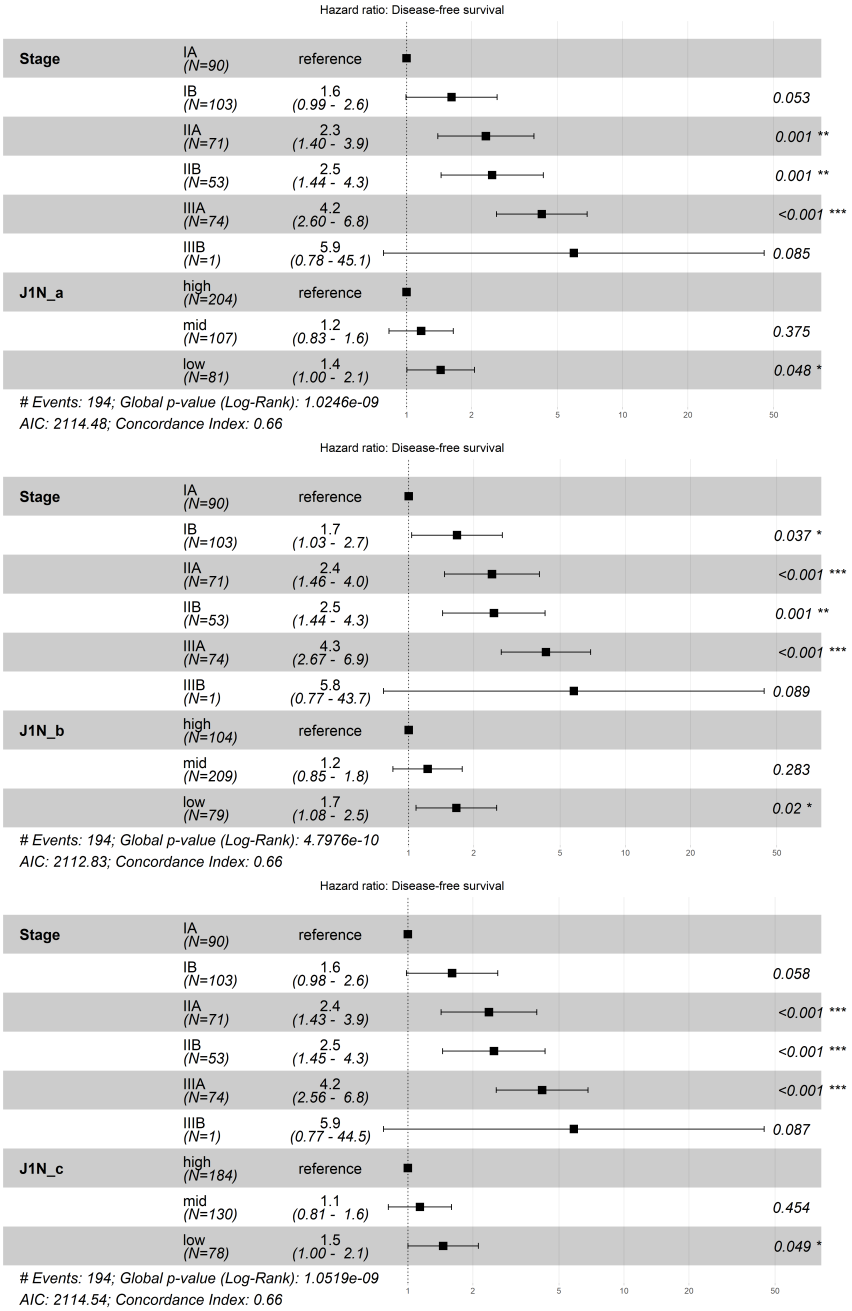

Figure 8.34: Multi-variable Cox proportional hazard models containing stage and alternative tree balance indices. The adjusted cutpoints used in Figure 7.8 are the cutpoints used here. The HR 95% CIs are shown in brackets and by the error bars. The asterisks indicate the P value ranges, where $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

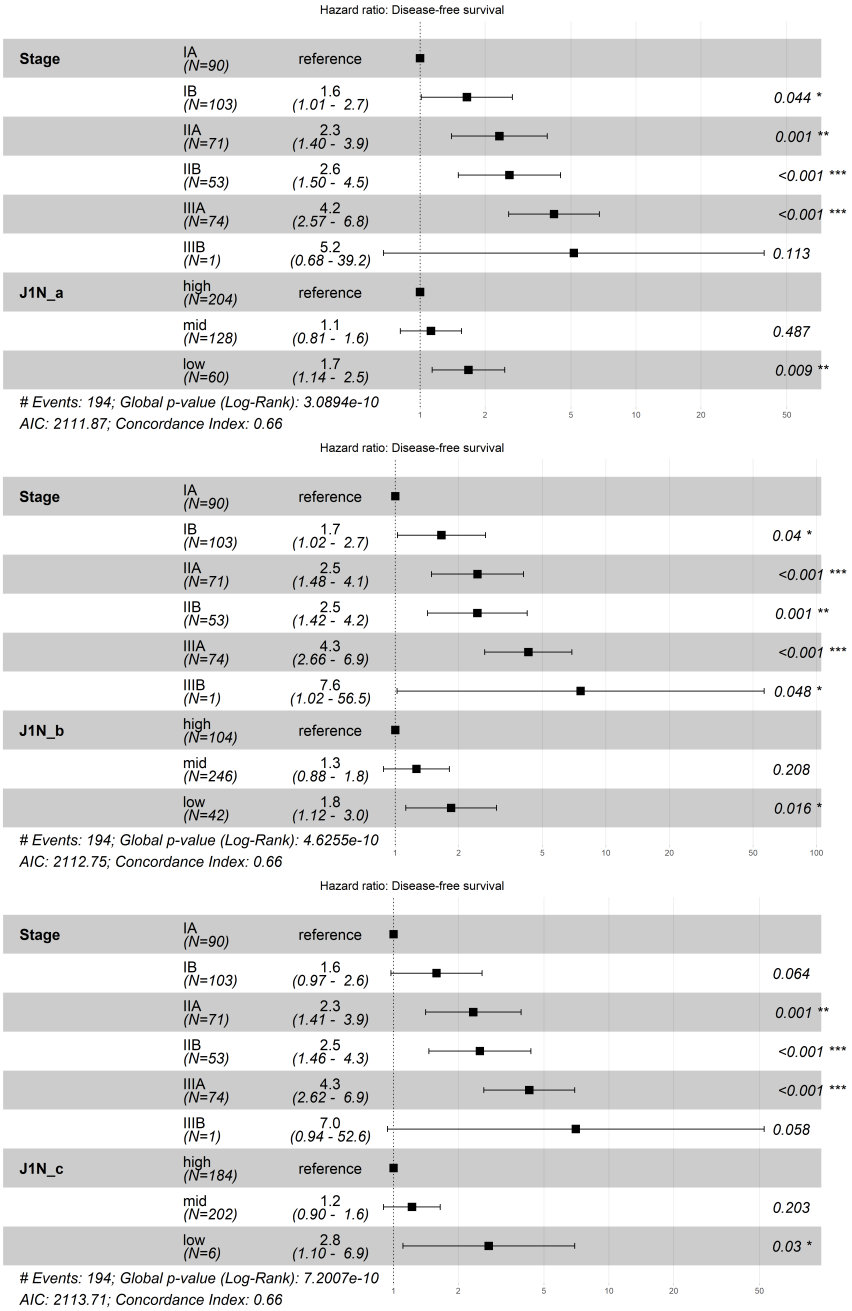

Figure 8.35: Multi-variable Cox proportional hazard models containing stage and alternative tree balance indices using the original cutpoints of 0.85 and 0.99. The HR 95% CIs are shown in brackets and by the error bars. The asterisks indicate the P value ranges, where $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

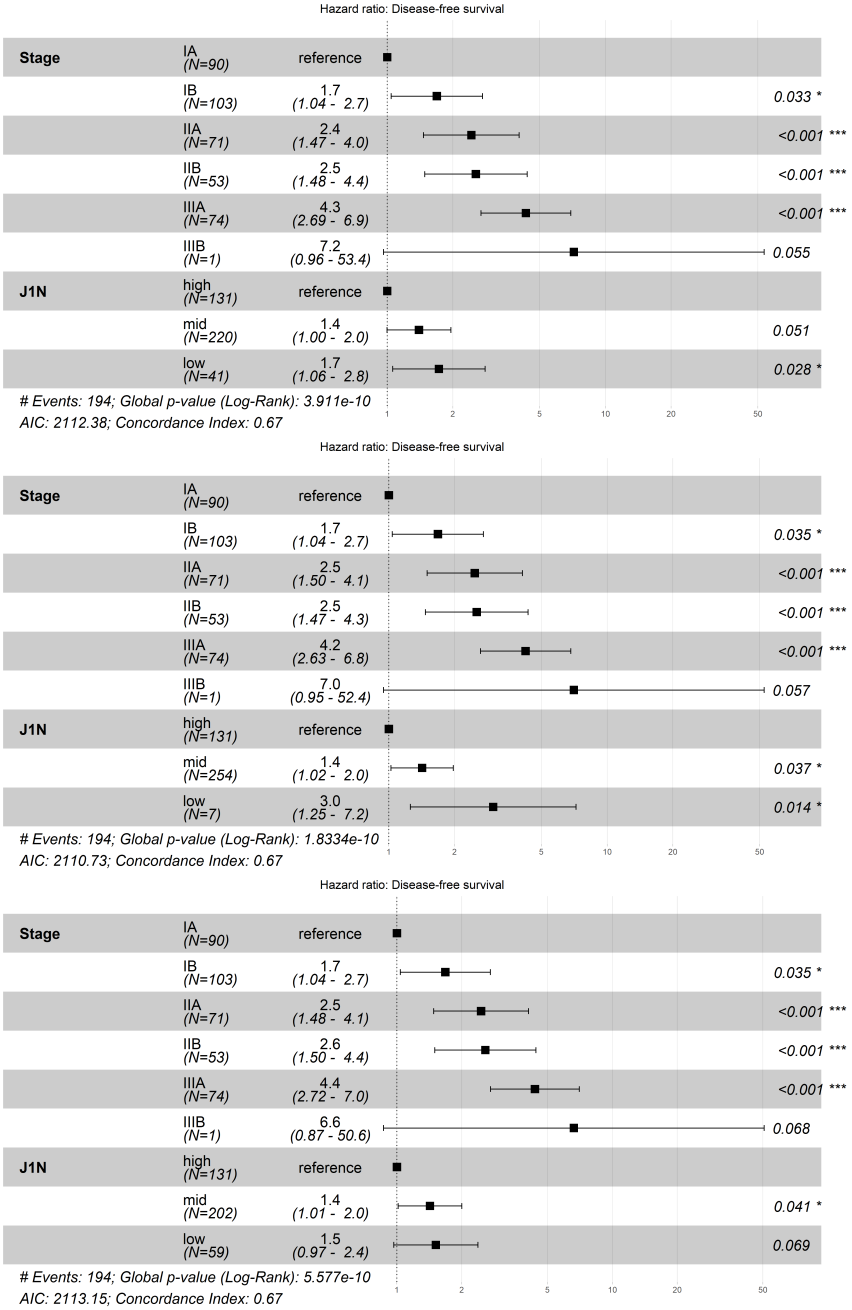

Figure 8.36: Multi-variable Cox proportional hazard models containing stage and the tree balance index, ${}^1J_N$. The lower cutpoints here are chosen such that they give a "low" size group as close to the groupings for the alternative indices with the original cut points (Figure 8.35. The HR 95% CIs are shown in brackets and by the error bars. The asterisks indicate the P value ranges, where $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Bibliography

[1] National Academy of Sciences (US). *Tempo And Mode In Evolution: Genetics And Paleontology 50 Years After Simpson*. Ed. by Walter M. Fitch and Francisco J. Ayala. Washington (DC): National Academies Press (US), 1995.

[2] George Gaylord Simpson. "Tempo and Mode in Evolution". In: *Tempo and Mode in Evolution*. Columbia University Press, 2019. DOI: 10.7312/simp93040.

[3] Joseph F. Pachut and Robert L. Anstey. "Inferring Evolutionary Modes in a Fossil Lineage (Bryozoa: *Peronopora*) from the Middle and Late Ordovician". In: *Paleobiology* 35.2 (2009), pp. 209–230. DOI: 10.1666/07055.1.

[4] Gene Hunt. "The Relative Importance of Directional Change, Random Walks, and Stasis in the Evolution of Fossil Lineages". In: *Proceedings of the National Academy of Sciences* 104.47 (2007), pp. 18404–18408. DOI: 10.1073/pnas.0704088104.

[5] Thomas Pradeu et al. "Reuniting Philosophy and Science to Advance Cancer Research". In: *Biological Reviews* (2023), brv.12971. DOI: 10.1111/brv.12971.

[6] Zayd Tippu, Lewis Au, and Samra Turajlic. "Evolution of Renal Cell Carcinoma". In: *European Urology Focus* 7.1 (2021), pp. 148–151. DOI: 10.1016/j.euf.2019.12.005.

[7] Mohammadamin Edrisi et al. "MoTERNN: Classifying the Mode of Cancer Evolution Using Recursive Neural Networks". In: *Comparative Genomics*. Ed. by Katharina Jahn and Tomáš Vinař. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 232–247. DOI: 10.1007/978-3-031-36911-7_15.

[8] Husayn Ahmed Pallikonda and Samra Turajlic. "Predicting Cancer Evolution for Patient Benefit: Renal Cell Carcinoma Paradigm". In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1877.5 (2022), p. 188759. DOI: 10.1016/j.bbcan.2022.188759.

[9] Christoph Röcken et al. "Multiscale Heterogeneity in Gastric Adenocarcinoma Evolution Is an Obstacle to Precision Medicine". In: *Genome Medicine* 13.1 (2021), p. 177. DOI: 10.1186/s13073-021-00975-y.

[10] Yuri Bakhtin et al. *Punctuated Equilibrium as the Default Mode of Evolution of Large Populations on Fitness Landscapes Dominated by Saddle Points in the Weak-Mutation Limit*. 2020. arXiv: 2007.10228. Pre-published.

[11] Minsoo Kim et al. “Inferring Modes of Evolution from Colorectal Cancer with Residual Polyp of Origin”. In: *Oncotarget* 9.6 (2017), pp. 6780–6792. DOI: 10.18632/oncotarget.23687.

[12] Andrea M. Muscat et al. “The Evolutionary Pattern of Mutations in Glioblastoma Reveals Therapy-Mediated Selection”. In: *Oncotarget* 9.8 (2017), pp. 7844–7858. DOI: 10.18632/oncotarget.23541.

[13] Ke Yuan et al. “BitPhylogeny: A Probabilistic Framework for Reconstructing Intra-Tumor Phylogenies”. In: *Genome Biology* 16.1 (2015), p. 36. DOI: 10.1186/s13059-015-0592-6.

[14] Kjetil Lysne Voje, Emanuela Di Martino, and Arthur Porto. “Revisiting a Landmark Study System: No Evidence for a Punctuated Mode of Evolution in Metrarabdotos”. In: *The American Naturalist* 195.5 (2020), pp. 899–917. DOI: 10.1086/707664.

[15] Eugene Rosenberg. “Rapid Acquisition of Microorganisms and Microbial Genes Can Help Explain Punctuated Evolution”. In: *Frontiers in Ecology and Evolution* 10 (2022).

[16] Yuri Bakhtin et al. “Evolution in the Weak-Mutation Limit: Stasis Periods Punctuated by Fast Transitions between Saddle Points on the Fitness Landscape”. In: *Proceedings of the National Academy of Sciences* 118.4 (2021), e2015665118. DOI: 10.1073/pnas.2015665118.

[17] Geoff Macintyre et al. “How Subclonal Modeling Is Changing the Metastatic Paradigm”. In: *Clinical Cancer Research* 23.3 (2017), pp. 630–635. DOI: 10.1158/1078-0432.CCR-16-0234.

[18] Haruhiko Takeda et al. “Genetic Landscape of Multistep Hepatocarcinogenesis”. In: *Cancers* 14.3 (3 2022), p. 568. DOI: 10.3390/cancers14030568.

[19] Yong Zhang et al. “Identifying Bifurcated Paths with Differential Function Impact in Glioblastomas Evolution”. In: *International Journal of Cancer* 147.11 (2020), pp. 3139–3151. DOI: 10.1002/ijc.33276.

[20] Karla S. C. Yotoko et al. “Does Variation in Genome Sizes Reflect Adaptive or Neutral Processes? New Clues from Passiflora”. In: *PLOS ONE* 6.3 (2011), e18212. DOI: 10.1371/journal.pone.0018212.

[21] Melanie J. Hopkins and Scott Lidgard. “Evolutionary Mode Routinely Varies among Morphological Traits within Fossil Species Lineages”. In: *Proceedings of the National Academy of Sciences* 109.50 (2012), pp. 20520–20525. DOI: 10.1073/pnas.1209901109.

[22] Callum Whitten. “Nature vs Nurture: Quantifying Evolution and Ecophenotypic Variation in Pelicaria Vermis”. thesis. Open Access Te Herenga Waka-Victoria University of Wellington, 2023. DOI: 10.26686/wgtn.22179053.

[23] Pablo Duchen et al. “On the Effect of Asymmetrical Trait Inheritance on Models of Trait Evolution”. In: *Systematic Biology* 70.2 (2021). Ed. by Luke Harmon, pp. 376–388. DOI: 10.1093/sysbio/syaa055.

[24] Niklas Hohmann et al. *Identification of the Mode of Evolution in Incomplete Carbonate Successions*. 2023. URL: https://www.biorxiv.org/content/10.1101/2023.12.18.572098v1 (visited on 01/09/2024). Pre-published.

[25] M.L. Glassman, N. De Groot, and A. Hochberg. "Cancer, Evolution and Birth: Reliving Our Ancestral Past". In: *Medical Hypotheses* 46.1 (1996), pp. 13–16. DOI: 10.1016/S0306-9877(96)90227-3.

[26] Shaoping Ling et al. "Extremely High Genetic Diversity in a Single Tumor Points to Prevalence of Non-Darwinian Cell Evolution". In: *Proceedings of the National Academy of Sciences* 112.47 (2015). DOI: 10.1073/pnas.1519556112.

[27] Bachisio Ziccheddu et al. "The Genomic and Transcriptomic Landscape of Double-Refractory Multiple Myeloma". In: *Blood* 134 (2019), p. 3056. DOI: 10.1182/blood-2019-122197.

[28] Yuri I. Wolf and Eugene V. Koonin. "Genome Reduction as the Dominant Mode of Evolution". In: *BioEssays* 35.9 (2013), pp. 829–837. DOI: 10.1002/bies.201300037.

[29] Deepak Singh et al. "Bioremediation of Nitroaromatic Compounds". In: *Wastewater Treatment Engineering*. IntechOpen, 2015. DOI: 10.5772/61253.

[30] Natalja Strelkowa and Michael Lässig. "Clonal Interference in the Evolution of Influenza". In: *Genetics* 192.2 (2012), pp. 671–682. DOI: 10.1534/genetics.112.143396.

[31] Vaibhav Bhandari, Hafiz Naushad, and Radhey Gupta. "Protein Based Molecular Markers Provide Reliable Means to Understand Prokaryotic Phylogeny and Support Darwinian Mode of Evolution". In: *Frontiers in Cellular and Infection Microbiology* 2 (2012).

[32] Shuang Jiang et al. "Primitive Genepools of Asian Pears and Their Complex Hybrid Origins Inferred from Fluorescent Sequence-Specific Amplification Polymorphism (SSAP) Markers Based on LTR Retrotransposons". In: *PLOS ONE* 11.2 (2016), e0149192. DOI: 10.1371/journal.pone.0149192.

[33] Meiying Fang et al. "Contrasting Mode of Evolution at a Coat Color Locus in Wild and Domestic Pigs". In: *PLOS Genetics* 5.1 (2009), e1000341. DOI: 10.1371/journal.pgen.1000341.

[34] Matthew R. E. Symonds, Adnan Moussalli, and Mark A. Elgar. "The Evolution of Sex Pheromones in an Ecologically Diverse Genus of Flies: PHEROMONE EVOLUTION IN BACTROCERA". In: *Biological Journal of the Linnean Society* 97.3 (2009), pp. 594–603. DOI: 10.1111/j.1095-8312.2009.01245.x.

[35] Qinghua Shi et al. "Autoploid Origin and Rapid Diploidization of the Tetraploid Thinopyrum Elongatum Revealed by Genome Differentiation and Chromosome Pairing in Meiosis". In: *The Plant Journal* 113.3 (2023), pp. 536–545. DOI: 10.1111/tpj.16066.

[36] Joachim M. Surm et al. "A Process of Convergent Amplification and Tissue-Specific Expression Dominates the Evolution of Toxin and Toxin-like Genes in Sea Anemones". In: *Molecular Ecology* 28.9 (2019), pp. 2272–2289. DOI: 10.1111/mec.15084.

[37] Ville Mustonen and Michael Lässig. "Adaptations to Fluctuating Selection in Drosophila". In: *Proceedings of the National Academy of Sciences* 104.7 (2007), pp. 2277–2282. DOI: 10.1073/pnas.0607105104.

[38] Monique Oliveira Freitas et al. "Genomic Instability in Circulating Tumor Cells". In: *Cancers* 12.10 (10 2020), p. 3001. DOI: 10.3390/cancers12103001.

[39] Nathaniel D. Anderson et al. “Rearrangement Bursts Generate Canonical Gene Fusions in Bone and Soft Tissue Tumors”. In: *Science* 361.6405 (2018), eaam8419. DOI: 10.1126/science.aam8419.

[40] Jason D. Pardo, Adam K. Huttenlocker, and Jonathan D. Marcot. “Stratocladistics and Evaluation of Evolutionary Modes in the Fossil Record: An Example from the Ammonite Genus Semiformiceras”. In: *Palaeontology* 51.4 (2008), pp. 767–773. DOI: 10.1111/j.1475-4983.2008.00794.x.

[41] Denver Warwick Fowler. “Revised Geochronology, Correlation, and Dinosaur Stratigraphic Ranges of the Santonian-Maastrichtian (Late Cretaceous) Formations of the Western Interior of North America”. In: *PLOS ONE* 12.11 (2017), e0188426. DOI: 10.1371/journal.pone.0188426.

[42] Alexandra Magro et al. “Oviposition Deterring Infochemicals in Ladybirds: The Role of Phylogeny”. In: *Evolutionary Ecology* 24.1 (2010), pp. 251–271. DOI: 10.1007/s10682-009-9304-6.

[43] Günter Theißen. “Saltational Evolution: Hopeful Monsters Are Here to Stay”. In: *Theory in Biosciences* 128.1 (2009), pp. 43–51. DOI: 10.1007/s12064-009-0058-z.

[44] Arne O. Mooers and Stephen B. Heard. “Inferring Evolutionary Process from Phylogenetic Tree Shape”. In: *The Quarterly Review of Biology* 72.1 (1997), pp. 31–54. DOI: 10.1086/419657.

[45] Benjamin Allen, Mark Kon, and Yaneer Bar-Yam. “A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats”. In: *The American Naturalist* 174.2 (2009), pp. 236–243. DOI: 10.1086/600101.

[46] Jeanne Lemant et al. “Robust, Universal Tree Balance Indices”. In: *Systematic Biology* 71.5 (2022), pp. 1210–1224. DOI: 10.1093/sysbio/syac027.

[47] T. Stadler. “Recovering Speciation and Extinction Dynamics Based on Phylogenies”. In: *Journal of Evolutionary Biology* 26.6 (2013), pp. 1203–1219. DOI: 10.1111/jeb.12139.

[48] Caroline M. Tucker et al. “A Guide to Phylogenetic Metrics for Conservation, Community Ecology and Macroecology: A Guide to Phylogenetic Metrics for Ecology”. In: *Biological Reviews* 92.2 (2017), pp. 698–715. DOI: 10.1111/brv.12252.

[49] Aisling Daly, Jan Baetens, and Bernard De Baets. “Ecological Diversity: Measuring the Unmeasurable”. In: *Mathematics* 6.7 (2018), p. 119. DOI: 10.3390/math6070119.

[50] Mareike Fischer et al. *Tree Balance Indices: A Comprehensive Survey*. Springer Nature, 2023. 398 pp.

[51] Andy Purvis and Paul-Michael Agapow. “Phylogeny Imbalance: Taxonomic Level Matters”. In: *SYSTEMATIC BIOLOGY* 51 (2002).

[52] Leonid Chindelevitch et al. “Network Science Inspires Novel Tree Shape Statistics”. In: *PLOS ONE* 16.12 (2021), e0259877. DOI: 10.1371/journal.pone.0259877.

[53] Lucia P. Barzilai and Carlos G. Schrago. “Signatures of Natural Selection in Tree Topology Shape of Serially Sampled Viral Phylogenies”. In: *Molecular Phylogenetics and Evolution* 183 (2023), p. 107776. DOI: 10.1016/j.ympev.2023.107776.

[54] Gabriel E. Leventhal et al. “Inferring Epidemic Contact Structure from Phylogenetic Trees”. In: *PLOS Computational Biology* 8.3 (2012), e1002413. DOI: 10.1371/journal.pcbi.1002413.

[55] Caroline Colijn and Jennifer Gardy. “Phylogenetic Tree Shapes Resolve Disease Transmission Patterns”. In: *Evolution, Medicine, and Public Health* 2014.1 (2014), pp. 96–108. DOI: 10.1093/emph/eou018.

[56] Jacob G Scott et al. “Inferring Tumor Proliferative Organization from Phylogenetic Tree Measures in a Computational Model”. In: *Systematic Biology* 69.4 (2020), pp. 623–637. DOI: 10.1093/sysbio/syz070.

[57] Robert Noble et al. “Spatial Structure Governs the Mode of Tumour Evolution”. In: *Nature Ecology & Evolution* 6.2 (2021), pp. 207–217. DOI: 10.1038/s41559-021-01615-9.

[58] Alison F Feder and Yingnan Gao. *Detecting Branching Rate Heterogeneity in Multifurcating Trees with Applications in Lineage Tracing Data*. 2024. URL: http://biorxiv.org/lookup/doi/10.1101/2024.06.27.601073 (visited on 07/03/2024). Pre-published.

[59] Robert Noble and Kimberley Verity. *A New Universal System of Tree Shape Indices*. 2023. URL: https://www.biorxiv.org/content/10.1101/2023.07.17.549219v3 (visited on 12/14/2023). Pre-published.

[60] Kiyomi Morita et al. “Clonal Evolution of Acute Myeloid Leukemia Revealed by High-Throughput Single-Cell Genomics”. In: *Nature Communications* 11.1 (2020), p. 5327. DOI: 10.1038/s41467-020-19119-8.

[61] Alexander M. Frankell et al. “The Evolution of Lung Cancer and Impact of Subclonal Selection in TRACERx”. In: *Nature* 616.7957 (2023), pp. 525–533. DOI: 10.1038/s41586-023-05783-5.

[62] Tom Leinster and Christina A. Cobbold. “Measuring Diversity: The Importance of Species Similarity”. In: *Ecology* 93.3 (2012), pp. 477–489. DOI: 10.1890/10-2402.1.

[63] M. J. Sackin. ““Good” and “Bad” Phenograms”. In: *Systematic Biology* 21.2 (1972), pp. 225–226. DOI: 10.1093/sysbio/21.2.225.

[64] Donald H. Colless. “Review of Phylogenetics: The Theory and Practice of Phylogenetic Systematics”. In: *Systematic Zoology* 31.1 (1982), pp. 100–104. DOI: 10.2307/2413420.

[65] Sandrine Pavoine and Carlo Ricotta. “A Simple Translation from Indices of Species Diversity to Indices of Phylogenetic Diversity”. In: *Ecological Indicators* 101 (2019), pp. 552–561. DOI: 10.1016/j.ecolind.2019.01.052.

[66] Anne Chao, Chun-Huo Chiu, and Lou Jost. “Phylogenetic Diversity Measures Based on Hill Numbers”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1558 (2010), pp. 3599–3609. DOI: 10.1098/rstb.2010.0272.

[67] Sandrine Pavoine, Eric Marcon, and Carlo Ricotta. “‘Equivalent Numbers’ for Species, Phylogenetic or Functional Diversity in a Nested Hierarchy of Multiple Scales”. In: *Methods in Ecology and Evolution* 7.10 (2016), pp. 1152–1163. DOI: 10.1111/2041-210X.12591.

[68] Alexander Davis, Ruli Gao, and Nicholas Navin. "Tumor Evolution: Linear, Branching, Neutral or Punctuated?" In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1867.2 (2017), pp. 151–161. DOI: 10.1016/j.bbcan.2017.01.003.

[69] Lauren M.F. Merlo et al. "A Comprehensive Survey of Clonal Diversity Measures in Barrett's Esophagus as Biomarkers of Progression to Esophageal Adenocarcinoma". In: *Cancer prevention research (Philadelphia, Pa.)* 3.11 (2010), pp. 1388–1397. DOI: 10.1158/1940-6207.CAPR-10-0108.

[70] Ivana Bozic et al. "Accumulation of Driver and Passenger Mutations during Tumor Progression". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.43 (2010), pp. 18545–18550. DOI: 10.1073/pnas.1010978107.

[71] Lucie Laplane and Eric Solary. "Towards a Classification of Stem Cells". In: *eLife* 8 (2019), e46563. DOI: 10.7554/eLife.46563.

[72] Chandler D. Gatenbee et al. "EvoFreq: Visualization of the Evolutionary Frequencies of Sequence and Model Data". In: *BMC Bioinformatics* 20.1 (2019), p. 710. DOI: 10.1186/s12859-019-3173-y.

[73] Santiago Ramón y Cajal et al. "Clinical Implications of Intratumor Heterogeneity: Challenges and Opportunities". In: *Journal of Molecular Medicine* 98.2 (2020), pp. 161–177. DOI: 10.1007/s00109-020-01874-2.

[74] Yul Ri Chung et al. "Diversity Index as a Novel Prognostic Factor in Breast Cancer". In: *Oncotarget* 8.57 (2017), pp. 97114–97126. DOI: 10.18632/oncotarget.21371.

[75] Carlo C. Maley et al. "Classifying the Evolutionary and Ecological Features of Neoplasms". In: *Nature Reviews Cancer* 17.10 (2017), pp. 605–619. DOI: 10.1038/nrc.2017.69.

[76] Jenny Karlsson et al. "Early Evolutionary Branching across Spatial Domains Predisposes to Clonal Replacement under Chemotherapy in Neuroblastoma". In: *Nature Communications* 15.1 (2024), p. 8992. DOI: 10.1038/s41467-024-53334-x.

[77] Vanessa Almendro et al. "Genetic and Phenotypic Diversity in Breast Tumor Metastases". In: *Cancer Research* 74.5 (2014), pp. 1338–1348. DOI: 10.1158/0008-5472.CAN-13-2357-T.

[78] Mariam Jamal-Hanjani et al. "Tracking the Evolution of Non–Small-Cell Lung Cancer". In: *New England Journal of Medicine* 376.22 (2017), pp. 2109–2121. DOI: 10.1056/NEJMoa1616288.

[79] Stephen B Heard. "Patterns in Phylogenetic Tree Balance with Variable and Evolving Speciation Rates". In: *Evolution* 50.6 (1996), pp. 2141–2148. DOI: 10.2307/2410685.

[80] Michael G. B. Blum and Olivier François. "Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance". In: *Systematic Biology* 55.4 (2006), pp. 685–691. DOI: 10.1080/10635150600889625.

[81] L Francisco Henao-Diaz and Matt Pennell. "The Major Features of Macroevolution". In: *Systematic Biology* 72.5 (2023), pp. 1188–1198. DOI: 10.1093/sysbio/syad032.

[82] János Podani. "Tree Thinking, Time and Topology: Comments on the Interpretation of Tree Diagrams in Evolutionary/Phylogenetic Systematics". In: *Cladistics* 29.3 (2013), pp. 315–327. DOI: 10.1111/j.1096-0031.2012.00423.x.

[83] Kimberley Verity and Robert John Noble. *Evolutionary Tree Balance Predicts Disease-Free Survival in the TRACERx Non-Small Cell Lung Cancer Cohort*. 2025. URL: https://www.medrxiv.org/content/10.1101/2025.11.22.25340797v1 (visited on 11/26/2025). Pre-published.

[84] M. O. Hill. "Diversity and Evenness: A Unifying Notation and Its Consequences". In: *Ecology* 54.2 (1973), pp. 427–432. DOI: 10.2307/1934352.

[85] Kimberley Verity. *GitHub repository*. https://github.com/kimverity/RUIindices. Version 0.1.

[86] Kathryn R. Kirby et al. "D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity". In: *PLOS ONE* 11.7 (2016), e0158391. DOI: 10.1371/journal.pone.0158391.

[87] T Honkola et al. "Cultural and Climatic Changes Shape the Evolutionary History of the Uralic Languages". In: *Journal of Evolutionary Biology* 26.6 (2013), pp. 1244–1253. DOI: 10.1111/jeb.12107.

[88] Maya A. Lewinsohn et al. "State-Dependent Evolutionary Models Reveal Modes of Solid Tumour Growth". In: *Nature Ecology & Evolution* 7.4 (4 2023), pp. 581–596. DOI: 10.1038/s41559-023-02000-4.

[89] Stephen B. Heard. "Patterns in Tree Balance among Cladistic, Phenetic, and Randomly Generated Phylogenetic Trees". In: *Evolution* 46.6 (1992), pp. 1818–1826. DOI: 10.2307/2410033.

[90] Arne Øyvind Mooers. "Tree Balance and Tree Completeness". In: *Evolution* 49.2 (1995), pp. 379–384. DOI: 10.2307/2410349.

[91] Ed Stam. "Does Imbalance in Phylogenies Reflect Only Bias?" In: *Evolution* 56.6 (2002), pp. 1292–1295. DOI: 10.1111/j.0014-3820.2002.tb01440.x.

[92] Eric W Holman. "Nodes in Phylogenetic Trees: The Relation Between Imbalance and Number of Descendent Species". In: *Systematic Biology* 54.6 (2005), pp. 895–899. DOI: 10.1080/10635150500354696.

[93] Sheldon Ross. *Continuous-Time Markov Chains*. 11th ed. Academic Press, 2014, pp. 357–407. DOI: 10.1016/B978-0-12-407948-9.00006-2.

[94] Sophie J. Kersting, Kristina Wicke, and Mareike Fischer. "Tree Balance in Phylogenetic Models". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 380.1919 (2025), p. 20230303. DOI: 10.1098/rstb.2023.0303.

[95] Veselin Manojlović et al. *Expected and Minimal Values of a Universal Tree Balance Index*. 2025. arXiv: 2507.08615 [q-bio]. URL: http://arxiv.org/abs/2507.08615 (visited on 12/02/2025). Pre-published.

[96] Scott A. Pardo. "Simulation and Random Variable Generation". In: *Empirical Modeling and Data Analysis for Engineers and Applied Scientists*. Ed. by Scott A. Pardo. Cham: Springer International Publishing, 2016, pp. 203–221. DOI: 10.1007/978-3-319-32768-6_14.

[97] Sujit K. Sahu. "Transformation and Transformed Distributions". In: *Introduction to Probability, Statistics & R: Foundations for Data-Based Sciences*. Ed. by Sujit K. Sahu. Cham: Springer International Publishing, 2024, pp. 287–308. DOI: 10.1007/978-3-031-37865-2_14.

[98] Tyler S. Kuhn, Arne Ø. Mooers, and Gavin H. Thomas. "A Simple Polytomy Resolver for Dated Phylogenies". In: *Methods in Ecology and Evolution* 2.5 (2011), pp. 427–436. DOI: 10.1111/j.2041-210X.2011.00103.x.

[99] L. Francisco Henao Diaz et al. "Macroevolutionary Diversification Rates Show Time Dependency". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.15 (2019), pp. 7403–7408. DOI: 10.1073/pnas.1818058116.

[100] Xiao Zhu et al. "Cancer Evolution: A Means by Which Tumors Evade Treatment". In: *Biomedicine & Pharmacotherapy* 133 (2021), p. 111016. DOI: 10.1016/j.biopha.2020.111016.

[101] Benjamin Werner et al. "Detecting Truly Clonal Alterations from Multi-Region Profiling of Tumours". In: *Scientific Reports* 7.1 (2017), p. 44991. DOI: 10.1038/srep44991.

[102] Ketevan Chkhaidze et al. "Spatially Constrained Tumour Growth Affects the Patterns of Clonal Selection and Neutral Drift in Cancer Genomic Data". In: *PLOS Computational Biology* 15.7 (2019), e1007243. DOI: 10.1371/journal.pcbi.1007243.

[103] Xiaowei Jiang and Ian P. M. Tomlinson. "Why Is Cancer Not More Common? A Changing Microenvironment May Help to Explain Why, and Suggests Strategies for Anti-Cancer Therapy". In: *Open Biology* 10.4 (2020), p. 190297. DOI: 10.1098/rsob.190297.

[104] Sohrab Salehi et al. "Cancer phylogenetic tree inference at scale from 1000s of single cell genomes". In: *Peer Community Journal* 3 (2023). DOI: 10.24072/pcjournal.292.

[105] Ivana Bozic, Alanna Sholokova, and Kamran Kaveh. *Neoantigen Evolution and Response to Checkpoint Inhibitor Immunotherapy in Colorectal Cancer*. 2024. URL: https://www.researchsquare.com/article/rs-4922340/v1 (visited on 08/28/2024). Pre-published.

[106] Wayne P. Maddison, Peter E. Midford, and Sarah P. Otto. "Estimating a Binary Character's Effect on Speciation and Extinction". In: *Systematic Biology* 56.5 (2007), pp. 701–710. DOI: 10.1080/10635150701607033.

[107] Jeremy M. Beaulieu and Brian C. O'Meara. "Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction". In: *Systematic Biology* 65.4 (2016), pp. 583–601. DOI: 10.1093/sysbio/syw022.

[108] Tanja Stadler. "Mammalian Phylogeny Reveals Recent Diversification Rate Shifts". In: *Proceedings of the National Academy of Sciences* 108.15 (2011), pp. 6187–6192. DOI: 10.1073/pnas.1016876108.

[109] Marc J. Williams et al. “Quantification of Subclonal Selection in Cancer from Bulk Sequencing Data”. In: *Nature genetics* 50.6 (2018), pp. 895–903. DOI: 10.1038/s41588-018-0128-6.

[110] Sohrab Salehi et al. “Clonal Fitness Inferred from Time-Series Modelling of Single-Cell Cancer Genomes”. In: *Nature* 595.7868 (2021), pp. 585–590. DOI: 10.1038/s41586-021-03648-3.

[111] Brian Johnson et al. “cloneRate: Fast Estimation of Single-Cell Clonal Dynamics Using Coalescent Theory”. In: *Bioinformatics (Oxford, England)* 39.9 (2023), btad561. DOI: 10.1093/bioinformatics/btad561.

[112] Sonal Singhal et al. “No Link between Population Isolation and Speciation Rate in Squamate Reptiles”. In: *Proceedings of the National Academy of Sciences of the United States of America* 119.4 (2022), e2113388119. DOI: 10.1073/pnas.2113388119.

[113] Andrew F. Magee et al. “Locally Adaptive Bayesian Birth-Death Model Successfully Detects Slow and Rapid Rate Shifts”. In: *PLOS Computational Biology* 16.10 (2020). Ed. by David A. Rasmussen, e1007999. DOI: 10.1371/journal.pcbi.1007999.

[114] Veselin Manojlovic. “Mathematical Classification of the Modes of Tumour Evolution”. City St Geroge’s, University of London, 2023.

[115] Monica-Andreea Baciu-Drăgan and Niko Beerenwinkel. “Oncotree2vec — a Method for Embedding and Clustering of Tumor Mutation Trees”. In: *Bioinformatics* 40 (Supplement_1 2024), pp. i180–i188. DOI: 10.1093/bioinformatics/btae214.

[116] Maciej Bak et al. *Warlock: An Automated Computational Workflow for Simulating Spatially Structured Tumour Evolution*. 2023. arXiv: 2301.07808 [q-bio]. URL: http://arxiv.org/abs/2301.07808 (visited on 01/30/2024). Pre-published.

[117] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. New York, NY, UNITED STATES: Springer New York, 2009.

[118] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[119] *A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960*. URL: https://journals.sagepub.com/doi/10.1177/001316446002000104 (visited on 12/29/2025).

[120] A. Katzourakis et al. “Macroevolution of Hoverflies (Diptera: Syrphidae): The Effect of Using Higher-Level Taxa in Studies of Biodiversity, and Correlates of Species Richness”. In: *Journal of Evolutionary Biology* 14.2 (2001), pp. 219–227. DOI: 10.1046/j.1420-9101.2001.00278.x.

[121] Andy Purvis, Aris Katzourakis, and Paul-Michael Agapow. “Evaluating Phylogenetic Tree Shape: Two Modifications to Fusco & Cronk’s Method”. In: *Journal of Theoretical Biology* 214.1 (2002), pp. 99–103. DOI: 10.1006/jtbi.2001.2443.

[122] Elizabeth Pennisi. “Large-Scale Gene Comparisons Boost Tree of Life Studies”. In: *Science* 342.6154 (2013), pp. 26–27. DOI: 10.1126/science.342.6154.26.

[123] Open Tree et al. *Open Tree of Life Synthetic Tree*. Version 15.1. DOI: https://doi.org/10.5281/zenodo.3937741.

[124] Open Tree et al. *Open Tree of Life Taxonomy*. Version 3.7. DOI: https://doi.org/10.5281/zenodo.3937750.

[125] Mark Holder Luna L. Sanchez Reyes Emily Jane McTavish. *McTavishLab/R_ OpenTree_ tutorials v0.9.1: Using the Open Tree of Life for your Research, with R*. https://github.com/McTavishLab/R_OpenTree_tutorials. Version 0.9.1.

[126] Giuseppe Fusco and Quentin C. B. Cronk. "A New Method for Evaluating the Shape of Large Phylogenies". In: *Journal of Theoretical Biology* 175.2 (1995), pp. 235–243. DOI: 10.1006/jtbi.1995.0136.

[127] Valentin Todorov and Peter Filzmoser. "An Object-Oriented Framework for Robust Multivariate Analysis". In: *Journal of Statistical Software* 32 (2010), pp. 1–47. DOI: 10.18637/jss.v032.i03.

[128] Patrick Mair and Rand Wilcox. "Robust Statistical Methods in R Using the WRS2 Package". In: *Behavior Research Methods* 52.2 (2020), pp. 464–488. DOI: 10.3758/s13428-019-01246-w.

[129] Jeffrey C. Liu and John A. Ridge. "Chapter 67 - What Is Cancer?" In: *Abernathy's Surgical Secrets (Seventh Edition)*. Ed. by Alden H. Harken and Ernest E. Moore. Elsevier, 2018, pp. 307–310. DOI: 10.1016/B978-0-323-47873-1.00067-X.

[130] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. "Intra-Tumour Heterogeneity: A Looking Glass for Cancer?" In: *Nature Reviews Cancer* 12.5 (2012), pp. 323–334. DOI: 10.1038/nrc3261.

[131] Mariam Jamal-Hanjani et al. "Translational Implications of Tumor Heterogeneity". In: *Clinical cancer research : an official journal of the American Association for Cancer Research* 21.6 (2015), pp. 1258–1266. DOI: 10.1158/1078-0432.CCR-14-1429.

[132] Marten Winter, Vincent Devictor, and Oliver Schweiger. "Phylogenetic Diversity and Nature Conservation: Where Are We?" In: *Trends in Ecology & Evolution* 28.4 (2013), pp. 199–204. DOI: 10.1016/j.tree.2012.10.015.

[133] Noemi Andor et al. "Pan-Cancer Analysis of the Extent and Consequences of Intratumor Heterogeneity". In: *Nature Medicine* 22.1 (2016), pp. 105–113. DOI: 10.1038/nm.3984.

[134] Samra Turajlic et al. "Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal". In: *Cell* 173.3 (2018), 595–610.e11. DOI: 10.1016/j.cell.2018.03.043.

[135] Javier Fernandez-Mateos et al. "Tumor Evolution Metrics Predict Recurrence beyond 10 Years in Locally Advanced Prostate Cancer". In: *Nature Cancer* 5.9 (2024), pp. 1334–1351. DOI: 10.1038/s43018-024-00787-0.

[136] Yujie Jiang et al. *Pan-Cancer Subclonal Mutation Analysis of 7,827 Tumors Predicts Clinical Outcome*. 2024. URL: https://www.biorxiv.org/content/10.1101/2024.07.03.601939v1 (visited on 01/14/2026). Pre-published.

[137] Takahiro Karasaki et al. "Evolutionary Characterisation of Lung Adenocarcinoma Morphology in TRACERx". In: *Nature medicine* 29.4 (2023), pp. 833–845. DOI: 10.1038/s41591-023-02230-w.

[138] Kristiana Grigoriadis et al. “CONIPHER: A Computational Framework for Scalable Phylogenetic Reconstruction with Error Correction”. In: *Nature Protocols* 19.1 (2024), pp. 159–183. DOI: 10.1038/s41596-023-00913-9.

[139] D. G. Altman et al. “Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors”. In: *JNCI Journal of the National Cancer Institute* 86.11 (1994), pp. 829–835. DOI: 10.1093/jnci/86.11.829.

[140] Robert Noble et al. “When, Why and How Tumour Clonal Diversity Predicts Survival”. In: *Evolutionary Applications* 13.7 (2020), pp. 1558–1568. DOI: 10.1111/eva.13057.