Estimating the Somatic Mutation Rate in Long-Lived Trees: Phylogenomic Approaches and ABC Alternatives

Andrea Maria Grecu

Student ID: 293287544, UPI: agre945
School of Biological Sciences

A thesis submitted in partial fulfilment of the requirements for the degree of Masters of Science in Biological Sciences, The University of Auckland, 2025

Abstract

Long-lived trees accumulate mutations throughout their lifetimes, creating intra-individual genetic diversity that persists for centuries. These mutations shape evolutionary potential and species adaptation, yet accurately estimating their rates remains challenging due to the complexity of tree growth and mutation inheritance. In this thesis, I evaluate the phylogenomic method developed by Orr et al. (2020), which links genetic variation to tree topology to estimate mutation rates. However, this method assumes mutations follow the tree's branching structure, an assumption that may not always hold. Using simulations based on Tomimoto and Satake's (2023) models of somatic mutation accumulation, the method is assessed across various tree topologies, mutation rates, and meristem dynamics.

Unbalanced long-terminal topologies performed best under low mutation rates ($<10^{-9}$ per site per year), as their extended terminal branches reduce mutation overlap and enhance genetic signal distinctiveness. In contrast, balanced long-terminal topologies systematically underpredicted mutation rates due to diluted mutation signals, while unbalanced short-terminal topologies overpredicted rates due to uneven mutation distributions. Additionally, I observed a substantial increase in shared mutations at higher mutation rates, with the coefficient of variation remaining nearly constant. This suggests that as mutation rates rise, branches share mutations more evenly, reducing phylogenetic distinctiveness. This effect was particularly pronounced in topologies with long shared internal nodes relative to terminal branches, further diminishing the phylogenomic method's effectiveness in resolving tree topologies.

Given these findings, I recommend using the phylogenomic method for low mutation rates and topologies with minimal shared mutations and distinct genetic signals.

To address its limitations, I present a prototype approach based on Approximate Bayesian Computation (ABC), which does not rely on topological assumptions. The ABC simulation framework approximates the highest posterior density (HPD) for somatic mutation rates and meristem parameters. Validation showed 99.4% of true parameter values were captured within the 95% HPD interval. This prototype offers significant potential as a flexible, scalable alternative for estimating somatic mutation rates, particularly in complex tree topologies. My findings suggest that while the phylogenomic method can be effective under certain conditions, the ABC approach provides a promising direction for future research in tree mutation dynamics.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Allen Rodrigo, and my co-supervisor, Dr. Teng Li. Their patience, support, and invaluable guidance have been instrumental throughout this journey—I truly could not have done it without them.

I am also immensely grateful to Sou Tomimoto and Professor Akiko Satake of the Mathematical Biology Laboratory, Kyushu University, for generously sharing their simulation code and taking the time to help me understand and implement it in my research. Their patience and expertise made a significant impact on this thesis.

To the many professors at the University of Auckland who have inspired me to continue in postgraduate study—thank you. I am also deeply appreciative of the colleagues and friends I have met during my time here, whose support and camaraderie have made this experience so meaningful.

A special thank you to my closest friends, the V.I.T.S., for their unwavering encouragement.

Lastly, I am forever grateful to my family, and especially my mother, Gabriela, who has given her all—without hesitation or judgment—to support me throughout my academic journey and beyond.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background

Far from being static elements of ecosystems, trees are dynamic, living records of genetic change, adapting and evolving across centuries. Each cell division in a tree carries the potential -or risk- for somatic mutation. In plants, these mutations can accumulate over their lifetimes due to the late or complete absence of germline segregation. This feature allows mutations to propagate through tissues and, in some cases, to offspring (Lanfear, 2018). This heritability makes somatic mutations a key driver of individual adaptation, as well as population-level genetic diversity (Klekowski, 1988; Schoen & Schultz, 2019).

Somatic mutations can have profound long-term evolutionary consequences. They contribute to the phenomenon of genetic mosaicism, in which different parts of a plant develop distinct genotypes and phenotypes. For example, in a long-lived *Eucalyptus melliodor*a individual in Canberra, Australia, somatic mutations enabled two of its eight major branches to develop chemical defences against herbivory by the Australian Christmas Beetle (Padovan et al., 2013; Orr et al., 2020). This localised adaptation highlights how genetic diversity within a plant can facilitate environmental adaptation within a single generation. Similarly, somatic mutations have been shown to enhance fitness in other long-lived species by influencing traits like flower morphology and reproductive timing, shaping ecological interactions with pollinators and herbivores over multiple generations (Gill et al., 1995).

In clonal plants and long-lived species, somatic mutations are critical for maintaining genetic diversity within populations. Clonal plants reproduce asexually, producing genetically identical offspring that risk genetic stagnation over time (Prach & Pyšek, 1994). Somatic mutations mitigate this by introducing new genetic variation, enhancing adaptability and resilience (Klekowski & Godfrey, 1989; Vincent et al., 2019). This diversity is particularly important in long-lived plants, where mutations accrued over centuries can affect key traits like pest resistance or reproductive timing, contributing to their evolutionary success. By acting as a reservoir of genetic novelty, somatic mutations not only sustain clonal lineages but also drive broader evolutionary patterns, such as shifts in molecular evolution rates and even speciation (Lanfear et al., 2013).

These evolutionary consequences are not confined to natural ecosystems; they also extend to human-managed environments. In agriculture, many commercially valuable plant varieties owe their traits- such as flower color or fruit morphology- to somatic mutations (Tilney-Bassett, 1986). However, these mutations can also pose challenges. In clonal forestry, for instance, they may reduce genetic stability or introduce harmful traits (Khoury et al., 2010). Conversely, in invasive or weedy species, somatic mutations have conferred adaptive advantages such as herbicide resistance, enabling their proliferation in adverse environments (Michel et al., 2004).

Despite their importance, somatic mutation rates in plants- particularly long-lived trees- remain poorly understood. These species face a unique evolutionary challenge; balancing the benefits of somatic mutations, such as increased genetic variability and localised adaptation, with the risks of accumulating deleterious mutations over centuries of growth. This balance is critical for their long-term survival, as excessive mutations could undermine genetic integrity, while too few mutations

might limit the potential for adaptation to environmental changes. Long-lives trees, like *Eucalyptus melliodora*, seem to maintain remarkably low mutation rates per unit of growth, suggesting that evolutionary mechanisms have evolved to optimise this trade-off (Orr et al., 2020).

Understanding the somatic mutation rates of long-lived plants has implications far beyond theoretical interpretation. For conservation, it offers insights into how species utilise somatic mutations to adapt and persist in changing environments, informing strategies to support resilience in the face of climate change and other environmental pressures (Miryeganeh & Armitage, 2024). In forestry, studies like those on Sitka spruce (*Picea sitchensis*) reveal a high per-generation somatic mutation rate, which contributes to local adaptation but also increases genetic load (Hanlon et al., 2019). This duality has direct implications for forest management, suggesting that selective propagation of beneficial mutations, particularly in seed orchards, could optimise productivity and adaptability. By integrating these findings into forestry practices, we can improve species resilience while maintaining genetic stability. Developing reliable methods to estimate somatic mutation rates across different species enables us to refine such strategies, offering practical tools for conservation, breeding and ecosystem management.

1.2 Methods of Somatic Mutation Estimation

1.2.1 Historical Methods of Somatic Mutation Estimation

Early efforts to estimate somatic mutation rates in plants heavily relied on phenotypic observations, providing foundational insights into genetic variation. Maize (*Zea may*s) served as a model organism, with variations in kernel pigmentation offering visible markers for mutational events (Darrah et al., 2019). Emerson (1913) documented variations in kernel color and proposed that somatic changes in genetic factors could influence the traits inherited through gametes. While his work was among the earliest to explore the potential role of somatic mutations in plants, other researchers built upon these findings, examining broader inheritance and mutational patterns in maize and other agricultural plants.

Recognising the importance of somatic mutations, researchers also explored artificial methods to induce them. Radiation and chemical treatments became widely used in the mid-20th century to accelerate genetic variation (Richter & Singleton, 1955; Sparrow & Cuany, 1959; Sekiguchi et al., 1971; Brunner, 1995). Cabrera-Ponce et al. (2019) highlighted how irradiation in corn breeding programs led to the discovery of novel traits, advancing agricultural and genetic research. The induction of artificial mutations and the underlying interest in producing new variants for crop improvement spurred efforts to better understand the mechanisms and rates of somatic mutations in plants.

Modern phenotypic methods have expanded upon early approaches by linking mutation rates to quantifiable outcomes. For example, Bobiwash et al. (2013) developed methods to estimate deleterious somatic mutation rates in the clonal woody plant *Vaccinium angustifolium* (lowbush berry) by analysing fruit set (the successful formation of a fruit from a flower) after autogamous (self-pollination) and geitonogamous (pollination between flowers of the same plant) pollination. Their maximum-likelihood model inferred a higher deleterious mutation rate in long-lived plants compared to annuals, without the use of any sequencing technologies. However, while such phenotypic studies may provide insights into somatic mutation rates, they are inherently limited by their reliance on observable traits- such as the fruit set of *Vaccinium angustifolium*. Phenotypic studies

can not capture the full spectrum of genetic changes or any additional molecular information of the mutations themselves.

Mutation reversion assays provided a hybrid approach to understanding somatic mutations, combining phenotypic simplicity with molecular precision (Quiroz et al., 2023). These assays were particularly well-studied in *Arabidopsis thaliana*, where defective reporter genes were used to detect specific de-novo mutations that restored gene function. For example, reversion of a stop codon in transgenic *Arabidopsis* lines containing the β-glucuronidase (β-GUS) gene resulted in blue fluorescence, offering a neutral and precise system to measure mutation rates (Kovalchuk, 2000; Yao & Kovalchuk, 2011). These experiments revealed key insights, such as higher mutation rates in low-expression genes and the variability of single-base pair substitution frequencies (Boyko et al., 2006; Golubov et al., 2010). Additionally, they demonstrated how somatic mutation rates increased in mature tissue and under stress, highlighting the dynamic nature of mutation processes in *Arabidopsis*. Despite their utility, these assays were limited to localised sites, leaving broader genomic patterns unexplored.

Cytogenetic techniques offered another avenue for estimating somatic mutation rates by analysing chromosomal abnormalities. Methods such as flow cytometry and DNA fingerprinting revealed large-scale changes, including aneuploidy, deletions, and translocations (Nybom, 1991; Ochatt, 2008). These analyses provided valuable insights into structural mutations but were constrained by their inability to resolve single-nucleotide changes. Additionally, cytogenetic methods were labor-intensive and not easily scalable, limiting their utility in large-scale studies (Ban & Jung, 2023).

Recent advancements in genomic technologies have provided transformative tools for somatic mutation estimation, particularly in long-lived trees. Whole-genome sequencing (WGS) revolutionised the field, allowing researchers to detect mutations with unparalleled accuracy. Although widely applied in cancer-research, where tumour-normal pair comparisons revealed mutation dynamics (Alioto et al., 2015), the principles of WGS have been applied to plant studies with significant success (Quiroz et al., 2023). For instance, Hofmeister et al. (2020) used a high-quality genome assembly of a 330-year-old Poplar tree (*Populus trichocarpa*) to estimate the individual's somatic mutation and epimutation rates. By sequencing and comparing branches within the tree, somatic mutations were identified and their accumulation linked to the estimated age of the tree. Their findings demonstrated a lower per-year mutation rate in perennials compared to annuals, further supporting the hypothesis that long-lived trees mitigate mutational overload by limiting meristematic cell divisions (Schmid-Siegert et al., 2017).

Additionally, other genomic approaches beyond WGS of individual plants have proven valuable for estimating somatic mutation rates. Feng et al. (2023) utilised population genomic techniques to analyse two ‘living fossil’ species; *Dipteronia sinensis* and *D. dyerinia*. By leveraging genomic assemblies and population-level data, they investigate mutational load, demographic history, and inbreeding effects, of the closely related long-lived species. These population genomic methods complement WGS by providing broader insights into mutation dynamics over time.

Somatic mutation estimation has evolved rapidly from simple phenotypic observations to sophisticated genomic tools in the last century. These advancements have enabled precise studies of mutation dynamics, as well as more accurate estimates of somatic mutation rates. The growing evidence of low mutation rates in long-lived trees has furthered research into more efficient somatic mutation techniques, such as the phylogenetic method examined in the following sections.

1.2.2 Mutation Rate Estimation in Phylogenetics

Phylogenetics, the study of evolutionary relationships among organisms, lies at the heart of understanding how genetic variation accumulates over time. Central to this field is the construction of phylogenetic trees - graphical representations of ancestral relationships that trace the divergence of species, genes or populations (Kapli et al., 2020). Phylogenetic trees are the basis for a variety of evolutionary analyses, from tracking the origins of novel traits to mutation rate estimation.

The methodology of phylogenetic tree construction has evolved drastically since its inception. Early efforts relied on morphological data, but the advent of molecular biology introduced genetic information as a more reliable unit of evolutionary inferences. Advances in sequencing technology have further expanded the scope and accuracy of phylogenetics, enabling the analysis of entire genomes to reconstruct evolutionary pathways with unprecedented precision (Bandelt et al., 2006).

The first step of phylogenetic tree construction involves sequence alignment, where homologous genetic sequences are arranged to identify conserved and variable regions. Algorithms such as MAFFT and MUSCLE align sequences by introducing gaps that account for insertions and deletions over evolutionary time (Edgar, 2004; Katoh et al., 2009). Following alignment, phylogenetic inference methods are employed, ranging from distance-based methods such as neighbour-joining to more sophisticated approaches such as maximum parsimony, maximum likelihood, and Bayesian inference (Brocchieri, 2001). Each method has unique strengths, tailored to the complexity and resolution of the data.

Central to many phylogenetic analyses is the molecular clock hypothesis, introduced by Zuckerkandl and Pauling (1965). This concept assumes a relatively constant rate of molecular evolution over time, enabling researchers to estimate the timing of divergence events. By calibrating phylogenetic trees with external data such as the fossil record, these trees can be transformed into chronograms which provide temporal data for evolutionary processes.

Phylogenetic methods have become essential tools for estimating mutation rates with genetic changes mapped on phylogenetic trees allowing researchers to infer rates of somatic mutations across lineages. Substitution models, such as the general time-reversible (GTR) model, are used to estimate the probabilities of nucleotide changes along branches (Weiss, 2003). Tools like BEAST and MrBayes have become integral to this work, offering robust frameworks for Bayesian inference of molecular evolution and tree calibration (Drummond & Rambaut, 2007; Ronquist et al., 2012).

One of the most transformative innovations in phylogenetics has been the development of relaxed molecular clocks, which accommodate variations in mutation rates among different lineages (Lepage et al., 2007). Implemented in software such as BEAST, these methods provide a more realistic depiction of evolutionary processes, particularly in taxa with complex histories. Relaxed clocks have proven invaluable in studying organisms with variable mutation rates, including plants with long generation times.

Despite their utility, phylogenetic methods face several challenges. Errors in sequence alignment, model selection, and phylogenetic tree topology can introduce biases that affect mutation rate estimates (Bruno & Halpern, 1999). Additionally, phenomena like incomplete lineage sorting and horizontal gene transfer (HGT) can obscure true evolutionary relationships (Kurland et al., 2003;

Maddison & Knowles, 2006). Overcoming these challenges requires continuous refinement of computational models and the integration of complementary data sources, such as ecological context.

The evolution of phylogenetic methodologies has expanded the understanding of mutation dynamics across taxa, providing a framework for estimating genetic variation over time. The principles of phylogenetics can be further adapted to estimate the somatic mutation rate in biological long-lived trees and other plant taxa, whose branching structures parallel those of phylogenetic trees.

1.2.3 The Phylogenomic Method

Orr et al. (2020) presented an innovative solution for measuring the somatic mutation rate of individual plants, which aimed to address the shortcomings of other methods pre-existing in scientific literature. This method leverages the phylogeny-like structure of individual plants, specifically in the application to a long-lived *Eucalyptus melliodora* tree. This individual exhibits pronounced phenotypic mosaicism, with two of its eight major branches having developed chemical defenses against herbivory by the Australian Christmas Beetle (Padovan et al., 2013). Based on this observation, the authors assumed that accumulated somatic mutations closely mirror the tree's physical structure.

With this assumption, Orr et al. (2020) developed their 'phylogenomic method'. They sequenced eight major branches of *E. melliodora,* each with three biological replicates (3 separate leaf samples of the same branch), generating high-coverage whole genomic data. The whole genomes of each branch were then aligned to a pseudo-reference genome and maximum-likelihood phylogenetic trees were subsequently reconstructed from this alignment. One of the highest-likelihood phylogenies closely matched *E. melliodora*'s topology, validating their reconstruction and prior assumption. This validated topology was then used to filter out any false positive variants, ensuring only mutations consistent with the tree's branching structure remained. Orr et al. (2020) reported an exceptionally low somatic mutation rate (1.16E-10 to 1.12E-09) per base per year for a single apical meristem, suggesting that *E. melliodora* may have evolved mechanisms to suppress mutational accumulation.

To date, there has only been one study published which has attempted to partially replicate Orr et al.'s phylogenomic method. Herrera et al. (2021) adapted the method to study *Lavandula latifolia* (commonly known as spike lavender), a Mediterranean perennial shrub characterised by its modular growth. Specifically, they applied the concept of reconstructing genealogical trees to map genetic and epigenetic divergence across branches, focusing on localised changes within modular units. By incorporating branching topology and linear distances, they analysed how these patterns align with the plant's developmental structure. While Orr et al. emphasized genome-wide mutation rates, Herrera et al. concentrated on epigenetic mosaicism, using DNA methylation data to capture cumulative changes within specific modules. This adaptation demonstrated the flexibility and potential of Orr et al.'s phylogenomic framework, showing its utility in exploring modular growth and epigenetic complexity in perennial plants.

However, Orr et al.'s phylogenomic method has not been without criticism. Iwasa et al. (2023) argued that the methodology might underestimate mutation rates by focusing only on phylogenetically consistent mutations while ignoring low-frequency somatic mutations. They proposed distinguishing between mutations accumulated along the "path after forking" and "path before forking", emphasising the need to consider the dynamics of stem cell replacement within the shoot apical meristem (SAM). Similarly, Plomion et al. (2024) highlighted limitations in Orr et al.'s reliance on physical

reconstructions of the tree and suggested that low-frequency somatic mutations in tropical trees like *Dicorynia guianensis* might not align neatly with phylogenetic frameworks.

As alternative perspectives emerge, Tomimoto and Satake (2023) have proposed a hierarchical architecture model focusing on meristem dynamics to predict mutation accumulation over time in long-lived trees. This mathematical framework incorporates branching and elongation processes, offering a theoretical approach that complements empirical somatic mutation studies. Their model predicts patterns of somatic genetic drift and mosaicism across branches, providing new insights into the structural and functional implications of mutation accumulation in long-lived plants.

The phylogenomic method presented by Orr et al. is a promising approach for understanding and estimating somatic mutation dynamics across plant taxa. However, the critiques discussed elucidate the need for a systemic evaluation of the method under diverse biological conditions. With these results, recommendations on the utility of the phylogenomic method can be made alongside refinements or alternative solutions.

1.3 Meristem Dynamics and Somatic Mutation Accumulation in Trees

The adaptability and longevity of plants are deeply rooted in the unique dynamics of their meristematic cells. Unlike animals, where the Weismann barrier rigidly separates somatic and germ cells, plants exhibit a more fluid interplay between these cell types. First described in 1885, the Weismann barrier delineates germline cells from somatic cells, functioning to prevent the inheritance of somatic mutations (Weismann, 1885). While this separation is strict in animals, it is less distinct in plants, where germline cells often arise from somatic cells late in development (Popov et al., 2022). This creates ‘leaks’ in the barrier, allowing somatic mutations to influence reproductive cells and propagate to offspring (Lanfear, 2018).

The partial or complete dissolution of the Weismann barrier has profound implications for somatic mutation accumulation, genetic mosaicism and plant evolution. Thus, a thorough understanding of meristem cells and their dynamics is critical for the estimation of somatic mutation rates in plant taxa.

1.3.1 Meristems; Stochastic or Structured

Meristems are clusters of undifferentiated, pluripotent cells responsible for producing all organs and tissues in plants (Tooke & Battery, 2003). During embryogenesis, the shoot apical meristem (SAM) and root apical meristem (RAM) are established as primary sites of cell division. The SAM governs above-ground organogenesis, while the RAM supports root architecture and function. Axillary meristems (AMs), derived from the SAM, are secondary meristems that initiate lateral branching, determining plant architecture and adaptability (Domagalska & Leyser, 2011).

The process of elongation, the vertical growth of a plant’s stem, is driven by successive divisions of stem cells within the SAM (Tomimoto & Satake, 2023). These stem cells are continuously dividing, producing all cell types for above-ground plant growth. The structure and stability of stem cells in the SAM vary significantly across plant taxa, influencing the accumulation and distribution of somatic mutations during elongation. More primitive plant taxa, such as the bryophytes and pteridophytes, contain only a single stem cell in the SAM, responsible for all above-ground plant growth (Klekowski, 1984). In contrast, seed plants possess a layered SAM, characterised as by a

tunica-corpus structure (Poething, 1989). In this structure, the tunica forms the outer layer, responsible for the surface growth, while the corpus lies beneath, contributing to volume growth.

Within this study, only seed plants capable of forming branching structures or trees are considered. There are two major classifications of seeded plants- the Angiosperms, or flowering plants and the Gymnosperms, or coniferous plants. Angiosperms typically maintain stable tunica-corpus layers, while gymnosperms often exhibit more flexible division patterns that disrupt this distinction, leading to intermixing between layers. This intermixing of tunica-corpus layers can lead to the loss of stem cell lineages throughout elongation, where stem cells in the SAM may be sampled randomly throughout iterative divisions. Meristems with a weak boundary between tunica-corpus layers can be defined as 'stochastic' meristems, introducing a level of stochasticity or randomness to the elongation process. In contrast, meristems with a strict boundary between tunica-corpus layers, more typical of Angiosperm taxa, are defined as 'structured' meristems that preserve stem cell lineage throughout elongation (Dermen, 1969).

The preservation of stem cell lineages in structured meristems induces somatic genetic fixation, where mutations accumulate within retained lineages over successive divisions (Chen et al., 2024). Structured meristems also employ somatic selection mechanisms, such as diplontic selection, to purge deleterious mutations by conferring a competitive disadvantage to affected cells (Otto & Orive, 1995). Stochastic meristems, which randomly mix stem cell lineages throughout division, induce somatic drift (aka epigenetic drift) as mutations are often lost before fixation. This lack of lineage retention also weakens the effects of somatic selection, resulting in lower mutation rates overall but less stringent mutation filtering.

Branching, the formation of axillary meristems from the SAM, further illustrates how structure and stochastic meristems influence lineage dynamics. Axillary meristems give rise to lateral branches through the radial expansion and division of a subset of stem cells in the SAM (Domagalska & Leyser, 2011). Not all stem cell lineages contribute equally, leading to spatially biased sampling, In structured meristems, typical of angiosperms, lineage preservation minimises this bias, maintaining a consistent contribution of stem cells. In stochastic meristems, common in gymnosperms, the lack of strict lineage maintenance allows for greater flexibility in contribution, amplifying biases and promoting variable branches.
The organisation of meristem cells greatly impacts the balance between genetic stability, adaptability, and thus, somatic mutation rates in plants. However, it is important to recognise that meristems across plant taxa do not strictly conform to these extremes of structure or stochasticity. Instead, they exist on a spectrum, where intermediate forms display varying degrees of lineage retention and flexibility. This spectrum enables greater balance between somatic drift and selection, optimising adaptability or stability depending on ecological and evolutionary pressures.

1.3.2 Genetic Mosaicism

Genetic mosaicism, the occurrence of distinct genetic variations within an individual organism, is especially pronounced in long-lived plants. Mutations arising in meristems propagate to entire branches or organs, causing intra-organism phenotypic variation. This phenomenon was evident in the previously mentioned *E. melliodora* tree, where somatic mutation(s) occurring in an axillary meristem rendered two branches resistant to Australian Christmas beetle defoliation (Padovan et al., 2013). Zahradníková et al. (2020) conducted a comprehensive review of genetic mosaicism across gymnosperms and angiosperms, synthesizing previously recorded examples of this phenomenon. This

study identified key patterns, such as age-related increases in intra-organismal mosaicism and how patterns of variation are influenced by branching dynamics and meristem structure.

Mosaicism reflects the structural design of axillary meristems, which distribute somatic mutation across branches and, in doing so, decrease fixation probabilities while promoting genotypic heterogeneity (Burian et al., 2016). Intra-organismal variation often corresponds to branching patterns, as mutations fixed in the SAM or AM propagate through daughter cells, leading to distinct genotypic sectors in the crown. However, Zahradníková et al. highlighted cases where mosaicism did not align with a plants branching topology, attributing them to sectorial or mericlinal chimeras- two distinct forms of genetic mosaicism. Sectorial chimeras involve variants that span entire longitudinal sectors of a plant, often affecting extensive 'sectors' of an organism due to mutation(s) throughout meristem layers. Mericlinal chimeras are characterised by mutations occurring in a localised patch of a single meristem layer, leading to more fragmented genetic patterns (Frank & Chitwood, 2016). These chimeric patterns highlight the intricacy of somatic mutation propagation within stratified meristems, further complicated by the spectrum of their structure across taxa. Structured meristems retain genetic lineages and tunica-corpus layers strictly, fixing mutations in place and promoting both sectorial and mericlinal chimeras as well as other types of genetic mosaicism. Stochastic meristems, with their intermixing of tunica-corpus layers and reduced lineage retention, are less likely to fix mutations thus hindering the formation of observable mosaicism. The random propagation of lineages (with unique mutations) in stochastic meristems promotes a broader distribution of genetic variability than structured meristems, limiting the impact of accumulated mutations for selection, fixation or the formation of mosaicism.

The phenomenon of genetic mosaicism is not simply a by-product of stratified meristems, but an adaptive evolutionary advantage. The Genetic Mosaic Hypothesis (GMH) posits that intra-organismal diversity enables plants to cope with environmental pressure, namely herbivory, by disrupting herbivore specialisation and increasing overall fitness (Whitham & Slobodchikoff, 1981; Folse & Roughgarden, 2012). Plants can respond to changing pressures at a modular level, with intra-organismal selection acting as a sieve purging deleterious mutations while retaining beneficial ones (Otto & Orive, 1995).

Understanding genetic mosaicism is essential for estimating somatic mutation rates in plants. Mosaicism occurs as a result of the structure and behaviour of meristems throughout branching and elongation, factors which influence the accumulation (or elimination) of somatic mutations. Studies such as Burian et al. (2016) and Zahradníková et al. (2020) highlight the need to incorporate spatial and temporal aspects of meristem development to better estimate somatic mutation rates in plants and further evaluate the ecological and evolutionary implications.

1.3.3 Modeling Meristem Dynamics

Tomimoto and Satake (2023) developed a mathematical framework to simulate the accumulation and propagation of somatic mutations in trees with a hierarchical modular architecture. Motivated by advancements in sequencing technologies that revealed the extent of plant intra-organismal variability, they aimed to integrate the dynamic processes of elongation and branching into models applicable to biological trees. Their work addresses critical gaps in understanding the proliferation and fixation dynamics of somatic mutations, providing insights into how these processes generate genetic mosaicism across a tree's architecture.

The mathematical model was applied to a 330-year old poplar (*Populus trichocarpa*) tree previously studied by Hofmeister et al. (2020). In this earlier study, genome sequencing of leaves from eight branches revealed single nucleotide polymorphism (SNPs) among different branches, enabling estimates of mutation rates per year per site based on branch age and topology. Leveraging this empirical data, Tomimoto and Satake simulated mutation accumulation and expansion within the poplar's modular architecture.

Their simulations demonstrated strong agreement with observed SNP counts across branches, validating the accuracy of their simulation framework. Specifically, they conducted 1,000 independent simulations to calculate the average number of mutated sites in each branch, comparing these predictions with empirical observations. These results highlighted that elongation, rather than branching, played the dominant role in mutation accumulation. Additionally, their model revealed that structured meristems, characterized by lineage stability, exhibited higher intra-meristem mosaicism compared to stochastic meristems. This distinction was further emphasised by the observation that mutations unrelated to the tree's topology were predicted exclusively by structured models.

Tomimoto and Satake's study illustrates the utility of computational approaches for interpreting mutation dynamics in long-lived trees. By accounting for both spatial and temporal aspects of mutation accumulation, their framework robustly explores the interplay between tree topology and genetic variability. In this study, I build upon their framework in order to evaluate the phylogenomic method for estimating somatic mutation rates in long-lived plants as defined by Orr et al. (2020). Tomimoto and Satake's critique of this method highlights its limitations in capturing mutation propagation across complex tree architecture and differing meristem structures.

1.4 Approximate Bayesian Computation

Studying somatic mutation rates in long-lived trees presents unique challenges due to the complexity of these biological systems. Traditional parameter estimation methods, namely maximum likelihood estimation (MLE), often face challenges with analytically intractable likelihood functions, such as in systems where mutations follow irregular patterns or topologies.
To overcome these challenges, alternative approaches like Bayesian inference and simulation-based frameworks have proven invaluable. This study employs Approximate Bayesian Computation (ABC), a simulation-based Bayesian inference method, to develop a prototype framework for estimating somatic mutation rates and developmental parameters in long-lived trees (see Chapter 3). The following sections outline the foundational principles of Bayesian statistics, their applications in biology, and the emergence of ABC.

1.4.1 Bayesian Statistics

Bayesian statistics, named after the 18th-century minister Thomas Bayes, represents a paradigm in statistical inference that combines prior information with observed data. Although Bayes' posthumously published essay (1763) laid the theoretical groundwork, the practical adoption of Bayesian methods remained limited until the mid-20th century when computational advancements enabled their broader application (Leonard, 2014). Today, Bayesian inference is a cornerstone of modern statistics, offering a coherent and flexible framework for addressing complex problems.

At its core, Bayesian inference uses Bayes' theorem to update beliefs about a parameter based on observed data. This process integrates prior knowledge, reflecting initial assumptions, with the

likelihood, which measures how well the data supports those assumptions. The result is the posterior distribution, an updated perspective that accounts for both sources of information (Efron, 2013).* Unlike frequentist approaches, which focus on long-term frequencies, Bayesian inference quantifies uncertainty about parameters directly (O'Hagan, 2003). This approach enables the incorporation of prior knowledge- whether derived from expert opinion, historical data, or plausible knowledge- into statistical analyses.

Scholars continue to debate the philosophical foundations of Bayesian statistics, particularly the divide between subjective and objective interpretations of probability (Machina & Schmeidler, 1992). John Maynard Keynes, an early 20th-century thinker, emphasised logical probability as an objective measure of support (Lukan, 2019). In contrast, Bruno de Finetti, a contemporary of Keyes, championed subjective probability as a reflection of personal belief (De Finetti, 1961; Lukan, 2019). Dennis Lindley, who followed later in the mid-20th century, further advanced Bayesian decision theory by emphasising the coherence of Bayesian methods, bridging philosophical debates into applied statistics (Lindley, 2000). Critics argue that poorly chosen priors, such as uniform distributions, can oversimplify complex systems and fail to account for real-world variability, leading to biased results (Gelman, 2008). Proponents highlight the strength of Bayesian methods in systematically updating beliefs with new data, mirroring the scientific method of inquiry where hypotheses are posted, tested, refined and validated with empirical evidence (Lukan, 2019). Despite the controversy, this iterative framework has become essential in disciplines like genetics, ecology, and epidemiology, where uncertainty and evolving data are rife.

Key milestones in Bayesian history begin with Harold Jeffreys who established a foundation for Bayesian methods with his development of Jeffrey priors. These priors are non-informative priors that ensure objectivity by being invariant under reparameterization, making them widely applicable in scientific inference (Jeffreys, 1998). The next major advance came with the development of Markov Chain Monte Carlo (MCMC) methods. The Metropolis algorithm, introduced in 1953 by Metropolis et al., laid the groundwork for MCMC by constructing Markov chains to sample from posterior distributions. A Markov chain is a stochastic process where the next state depends only on the current state, making it suitable for iterative sampling (Brooks, 1998). The Metropolis algorithm evaluates proposed samples by comparing their likelihood to the current sample, accepting moves based on a probability rule. This was generalised in 1970 with the Metropolis-Hastings algorithm, which extends the Metropolis method to accommodate asymmetrical proposal distributions, increasing its flexibility (Hastings, 1970). Gibbs sampling, introduced by Geman and Geman (1984), further advanced MCMC by breaking high dimensional sampling into smaller conditional updates, iterating over one parameter at a time. This approach reduced the computational burden of Bayesian methods in high-dimensional spaces (Qian et al., 2003). The development of tools like WinBUGS (Windows Version of Bayesian Inference Using Gibbs Sampling) by the UK Medical Research Council made Bayesian computation accessible for hierarchical models, while PyMC (Python Monte Carlo), created by the Python open-source community, introduced flexible, modern Bayesian modeling to a broader audience (Lunn et al., 2000; Abril-Pla et al., 2023).
These tools and advancements solidified Bayesian statistics as a cornerstone of modern science, enabling researchers to address some of the most complex and computationally demanding problems. In the next section, I explore the applications of Bayesian methods in biology and the development of Approximate Bayesian Computation, which extends these principles to scenarios involving analytically intractable likelihood functions.

**For the mathematical formulation of Bayes' theorem, refer to Section 3.2.1.*

1.4.2 Bayesian Applications in Biology

Bayesian methods have transformed how complex biological systems are analysed, offering robust tools to address uncertainty and noise inherent in biological data.

In phylogenetics, Bayesian inference has become a foundational tool for reconstructing evolutionary relationships. Unlike traditional methods, such as maximum likelihood, Bayesian inference allows researchers to impose prior distributions on tree topologies and evolutionary models, integrating assumptions about evolutionary relationships and patterns of genetic change. For instance, Huelsenbeck et al. (2002) highlight how MCMC methods enable the estimation of posterior probabilities of phylogenetic trees, making it possible to accommodate and analyse complex models of DNA evolution within a Bayesian framework. Bayesian phylogenetics has been used to address diverse questions, from divergence time estimation to detecting positive selection in genes (Mueller, 2006). Tools like MrBayes and BEAST have further popularized these by making Bayesian inference computationally accessible to researchers (Drummond & Rambaut, 2007; Ronquist et al., 2012).

The growing complexity of biological data, particularly in genomics and systems biology, has amplified the demand for Bayesian approaches. Wilkinson (2007) emphasised that Bayesian inference is ideal for modelling high-dimensional, noisy datasets, such as those encountered in microarray and sequencing studies. For example,Werhli et al. (2006) demonstrated the utility of Bayesian networks in reconstructing gene regulatory networks by integrating gene expression data within a probabilistic framework. Their approach outperformed alternatives like graphical Gaussian models in identifying directed interactions and mapping the structure of regulatory networks. By leveraging Bayesian networks, researchers gained deeper insights into cellular signaling pathways and the regulatory mechanisms governing gene expression.

Another application of Bayesian inference, of particular relevance to this thesis, is the estimation of somatic mutation rates in cancer genomes. Shiraishi et al. (2013) introduced Empirical Bayesian mutation Calling (EBCall), a method designed to enhance the accuracy of somatic mutation detection by explicitly modelling sequencing errors. EBCall utilises prior information from non-paired normal samples to estimate sequencing error distributions, enabling the identification of mutations even at low allele frequencies. This framework not only outperformed existing methods but also revealed minor tumor subpopulations and intratumoral heterogeneity, providing deeper insights into the clonal architecture of cancer. This example elucidates the power of Bayesian methods to integrate prior knowledge and tackles challenges in complex, noisy biological datasets.

Despite their versatility, the application of Bayesian methods to biological systems presents unique challenges. Gomez-Ramirez and Sanz (2013) reviewed key issues that Bayesian approaches face when applied to complex biological systems, particularly their difficulty in modeling the dynamism of such systems. For instance, gene expression datasets, where the number of genes vastly exceeds the number of samples, exemplify high-dimensional spaces with sparse probability densities (Secrier et al., 2009). In such cases, MCMC or other Bayesian sampling methods struggle to effectively explore parameter space, leading to computational inefficiencies. Additionally, Gomez-Ramirez and Sanz emphasise the ‘patchy’ nature of biological systems, such as spatial data in neurobiology, where probabilistic models must integrate geometric information from imaging techniques with physiological data to derive meaningful conclusions (Fiorillo, 2012).

Moreover, Bayesian approaches in biology are often constrained by the availability and quality of prior information. For example, integrating priors derived from genomic studies with environmental or experimental data can create inconsistencies due to variations in data scale, quality or relevance (Nikooienejad et al., 2016). Priors based on poorly curated datasets may reflect biases that propagate errors into posterior estimates, hindering analysis. Furthermore, the reliance on computationally intensive techniques like MCMC presents additional challenges for their application to large-scale biological datasets, such as whole genome sequences with millions or billions of nucleotide bases.

A particularly significant challenge arises when biological systems have intractable likelihood functions. These occur when the complexity of data or underlying biological processes makes direct evaluation of the likelihood difficult or computationally infeasible (Drovandi & Pettitt, 2013). Such cases are common in fields like evolutionary biology, where reconstructing complex phylogenies requires evaluating large amounts of genomic data, or cancer genomics, where identifying rare somatic mutations among vast amounts of sequencing information complicates likelihood calculations. Methods like Approximate Bayesian Computation (ABC), which bypass the need for explicit likelihood calculation, offer a promising solution to these challenges. In the next section, I explore ABC and its potential applications to complex biological systems, such as the accumulation of somatic mutations in long-lived trees.

1.4.3 The ABC's of Approximate Bayesian Computation

1.4.3.1. The History of ABC

Tavare et al. (1997) first introduced Approximate Bayesian Computation as we know it today, using simulations to infer coalescence times from observed DNA sequence data. Tavare and colleagues simulated genealogies under different values of parameters of interest including mutation rates, population sizes and divergence times. These simulated genealogies were then compared to a 'real' observed genealogy through summary statistics, such as the number of segregating sites and pairwise differences, to infer plausible coalescence times. Building on this foundation, Pritchard et al. (1999) formalised the ABC framework, applying it to human Y chromosome microsatellites. Their study showcased how ABC could be used to estimate demographic parameters such as population growth rates, migration patterns, and population bottleneck events without relying on maximum likelihood or traditional Bayesian methods. Pritchard et al. simulated datasets under varying demographic models and used relevant summary statistics- including the number of alleles, variance in repeat numbers, and heterozygosity- to compare with observed genetic data. Their analysis revealed that human Y chromosome data supported a history of recent population growth after a bottleneck event, providing a critical insight into historical demographic changes.

The work of Tavare et al. (1997) and Pritchard et al. (1999) not only advanced the field of demographic biology, but laid the foundation of ABC as an algorithm for solving previously intractable problems. It was Beaumont et al. (2002) who later coined the term 'Approximate Bayesian Computation', introducing regression adjustments which significantly improved the accuracy and efficiency of posterior approximations. Since then, ABC has spread to applications beyond demographic biology being widely applied throughout scientific fields.

1.4.3.2. Fundamental Considerations of ABC

At its core, ABC approximates posterior distributions through simulation rather than explicit likelihood calculation (Turner & Van Zandt, 2012). This method involves generating parameter values from prior distributions, simulating data under these parameters, and comparing the resulting summary statistics to observed data. If the distance between simulated and observed statistics, measured by a predefined distance metric, falls below a specific threshold, the parameter is accepted. This iterative process produces an approximation of the posterior distribution.

The accuracy and utility of ABC relies heavily on the selection of appropriate values of summary statistics, thresholds, distance metrics and priors for each dataset.

Summary statistics simplify high-dimensional data by retaining information relevant to the parameters of interest (Turner & Van Zandt, 2012). Ideally, these statistics should be sufficient, capturing all necessary information about the likelihood function. However, achieving sufficiency is rarely possible in practice. Nunes and Balding (2010) emphasised that optimal summary statistics should minimise the average squared error of the posterior distribution, ensuring that the chosen statistics retain as much information about the parameter as possible. They proposed methods like entropy minimisation to identify optimal summary statistics and demonstrated that the choice of statistics often depends on the specific dataset and parameter of interest. For instance, statistics such as the number of segregating sites, the number of haplotypes, and mean pairwise differences have been effective in population genetics. Poorly chosen statistics, by contrast, can bias posterior approximations or inflate uncertainty, diminishing interpretations. Principled, data-specific selection strategies should be employed to avoid these consequences (Fearnhead & Prangle, 2012).

The threshold, or epsilon (ϵ), defines the maximum allowable difference between simulated and observed summary statistics for a parameter to be accepted. This difference is measured using a relevant distance metric, such as Euclidean distance or the sum of squared errors, reflecting the structure of the data and the study's goals. The selection of appropriate thresholds and distance metrics is crucial in the ABC methodology (Turner & Van Zandt, 2012). Stricter thresholds may improve accuracy by focusing posterior distributions closer to the true data, but require significantly higher computation resources as many more selected parameters are rejected (Bertorelle et al., 2010). In contrast, lenient thresholds approximate meaningless posterior distributions which mirror prior distributions with little to no transformation. Sensitivity analyses are essential for validating threshold and metric choices. Liu and Niranjan (2021) emphasised the importance of systematically varying the threshold to assess its influence on posterior accuracy and computational feasibility. Adaptive distance metrics, which assign weights to summary statistics based on their predictive power, can also be used in conjunction with sensitivity analyses. By prioritising statistics most relevant to the parameters of interest, adaptive metrics reduce the impact of less informative data features.

Priors guide ABC simulations by representing pre-existing knowledge about the parameters. Selecting an appropriate prior is another crucial step that influences posterior distributions significantly. Non-informative priors, such as Jeffreys' prior, aim to minimise subjectivity, offering a starting point for inference without strongly influencing results (Jeffreys, 1998; Kass & Wasserman, 1994). However, the construction of such priors introduces challenges, particularly for models with high-dimensional parameter spaces, where improper or overly broad priors can lead to computational

inefficiency and bias posterior inferences (Roth et al., 2001). The flexibility of ABC allows for sensitivity analyses to evaluate the impact of different priors. Researchers may consider a class of candidate priors, assessing their effect on posterior approximations and inference robustness. Kass and Wasserman highlight the importance of 'reference priors', which can be tailored to specific data structures or experiment designs. While these priors serve as a default when subjective priors are unavailable, they may require iterative adjustments to align with empirical data and study-specific constraints.

1.4.3.3. Variants of ABC

Different classes of the ABC methodology have evolved to address limitations and improve efficiency.

The simplest form, ABC-Rejection as used by Tavare at al. (1997) and Pritchard et al. (1999), involves generating parameters from a prior distribution, simulating data under those parameters, and accepting parameters if the simulated data fell within a threshold of the observed data. No weight is put on accepted or rejected values, and all sampling occurs randomly from the prior. While conceptually straightforward, ABC-Rejection is computationally inefficient, especially for high-dimensional datasets, due to its high rejection rates. Marin et al. (2012) noted that rejection-based ABC is most effective for small datasets or models with low-dimensional parameter spaces.

Marjoram et al. (2003) introduced Markov Chain Monte Carlo (MCMC) into ABC , allowing for more efficient exploration of parameter space by using proposal distributions tailored to previously sampled regions. Unlike rejection sampling, ABC-MCMC reduced the rejection rate by iteratively refining parameter proposals. However, Marin et al. (2012) cautioned that ABC-MCMC requires careful tuning of the proposal distribution and is prone to convergence issues, particularly in models with complex or multimodal posteriors.

Sequential Monte Carlo (SMC) approaches, developed by Beaumont et al. (2009) and extended by Toni et al. (2009), iteratively refine posterior distributions by gradually reducing the threshold across generations. This method uses 'importance sampling' to focus on parameter prior regions with higher posterior probability, improving efficiency and scalability for high-dimensional problems. Marin et al. (2012) highlight that ABC-SMC's adaptive thresholding reduces computational demands while maintaining accuracy, making it a robust choice for large datasets.

Wilkinson (2008), introduced a modification to ABC that incorporates a kernel-based approximation to model error, addressing issues with inexact matching between simulated observed data. This method, termed 'Noisy ABC', explicitly models error distributions thus offering a more controlled approximation of the posterior. Marin et al. (2012) emphasises that Noisy ABC provides exact inference under certain conditions, particularly when the kernel matches the models' error distribution.

As previously mentioned, Beaumont et al. (2002) developed regression adjustments to improve the precision of ABC posteriors by correcting for the discrepancy between accepted parameters and the observed data. This ABC-Regression method assumes a linear relationship between summary statistics and parameters, which may not hold in all cases. Nonlinear regression approaches, as

discussed by Marin et al. (2012), offer more flexibility for complex models but can also increase computational demands.

Each method discussed offers unique advantages and limitations, with their suitability depending on their applications, computational demand and complexity of the model. A careful navigation of these trade offs is necessary for the selection of an appropriate ABC variant selection.

1.4.3.4. Posterior Interpretation in ABC

Posterior distributions are central to Bayesian inference, offering probabilistic insights into parameter values after incorporating prior knowledge and observed data. Each parameter sampled has its own approximated posterior distribution of accepted samples. Interpreting these distributions requires understanding of their shape, diagnostics, and credible intervals to ensure meaningful conclusions.

Posterior distributions can exhibit various characteristics, such as modality (unimodal, bimodal, or multimodal), skewness, and kurtosis (sharpness of the peak), depending on the data, model and parameters. A unimodal posterior suggests a clear single estimate for the parameter, whereas bimodal or multimodal distributions indicate ambiguity or competing hypotheses about the parameter value (Pick et al., 2023). Skewness often arises when parameters are constrained by boundaries of when the data provides limited information. For example, variance parameters, which cannot be negative, naturally produce asymmetric posteriors. In such cases, the posterior median is generally a more robust summary statistics than the mean because it is less influenced by long tails or extreme values. Meanwhile, modes (the most occurring value) may be affected by kernel density smoothing or sampling irregularities, especially in highly skewed distributions (Pick et al., 2023).

Trace plots visualise the sequence of accepted samples in parameter space, helping researchers evaluate two critical aspects of posterior sampling: convergence and mixing (Davis-Stober et al., 2016). Convergence refers to the point at which the sampling process stabilizes, indicating that the posterior distribution has been sufficiently explored and that additional samples are unlikely to alter the inferred distribution significantly.Convergence is important in ABC variants like ABC-SMC, where thresholds are iteratively refined, or particularly in ABC-MCMC, where chains may hit local maxima and not effectively explore parameter space (Nylander et al., 2007). Mixing assesses how well the sampling process explores the entire parameter space, with good mixing characterised by consistent fluctuations across the range of sampled values. In the context of ABC, trace plots track posterior samples that satisfy a chosen threshold. Good mixing is reflected by stable, evenly distributed traces without trends or clustering, whereas poor mixing manifests as plateaus or clusters, suggesting issues with thresholding, priors, or parameter sampling (Gabry et al., 2019). Overall, trace plots are useful in diagnosing whether the posterior distribution is representative and reliable.

Effective Samples Size (ESS) can be used in conjunction with trace plots. ESS quantifies the independent information in posterior samples, distinguishing between central (bulk ESS) and extreme (tail) regions of the posterior distribution (Thiébaux & Zwiers, 1984; Vehtari et al., 2020). High ESS and tail ESS values indicate reliable sampling across the distribution. Low bulk ESS suggests inefficient exploration of the central posterior region, potentially biasing estimates of mean or median values. Low tail ESS, by contrast, reflects poor sampling of the extremes, undermining the accuracy of credible intervals. If both values are low, fundamental issues like inefficient priors or overly strict thresholds may need to be addressed (Vehtari et al., 2020).

Highest Posterior Density (HPD) intervals summarise the most probable parameter ranges, offering Bayesian counterparts to frequentist confidence intervals (Joseph et al., 1995). Unlike frequentist intervals, which describe the likelihood of future data containing the true parameter, Bayesian intervals quantify the parameter's plausibility given the observed data (Turner & Van Zandt, 2012). For example, a 95% HPD interval for a posterior distribution of a mutation rate parameter reflects that 95% of the posterior density lies within this range. These intervals provide actionable insights into uncertainty while contextualising parameter estimates with the mean or median. Means, influenced by extreme values, often serve well for symmetric distributions, while medians are preferable under asymmetry for their robustness to skew (Pick et al., 2023).

Integrating these diagnostics enables robust posterior interpretation, guiding decisions on parameter plausibility, model fit, and areas requiring further refinement. By emphasising these metrics, researchers ensure that posterior distributions are not only statistically valid but scientifically meaningful.

1.4.3.5. Biological Applications of ABC

ABC has been successfully applied across diverse biological disciplines. For instance, Visani et al. (2021) utilised ABC to model COVID-19 hospital trajectories, estimating posterior distributions for transition and duration parameters to forecast hospital resources. Their approach highlighted the utility of ABC in addressing data-limited scenarios, generating actionable insights for public health decision-making. In conservation biology, Dittberner et al. (2022) employed ABC to study hybridization between two endangered plant species, *Arabis nemorensis* and *A. sagittata*. Their analysis revealed a history of low but persistent introgression punctuated by periods of isolation, providing insights into the evolutionary histories of these species which will aid in management. In this application, ABC helped to disentangle complex demographic and genetic patterns from genomic data.

These two recent studies provide a glimpse into the wide applicability of the ABC methodology throughout biology, and its potential to handle complex biological systems where intractable likelihoods limit traditional methods. In this thesis, ABC-Rejection sampling is applied in conjunction with Tomimoto and Satakes (2023) simulation framework to approximate the posterior distributions of somatic mutation rates and developmental parameters of long-lived trees. By incorporating ABC-Rejection sampling, I develop a 'prototype' method for estimating mutation rates, which aims to capture the complexity of meristem dynamics throughout a tree topology - without calculating a challenging likelihood.

1.5 Research Scope and Objectives

The phylogenomic method introduced by Orr et al. (2020) provides a promising framework for estimating somatic mutation rates in long-lived trees by aligning mutation accumulation with tree topology. However, as outlined in Section 1.2, critiques of this method suggest it may underestimate mutation rates by filtering out low-frequency mutations and making assumptions about stem cell replacement dynamics. Additionally, it has not been systematically tested across diverse tree architectures, meristem behaviours and mutation rate conditions throughout plant taxa.

This thesis addresses these gaps by employing Tomimoto and Satake's (2023) hierarchical modular growth model to simulate somatic mutation accumulation in a tree under controlled conditions, providing a framework to assess and refine somatic mutation rate estimation. The research is structured around two core objectives:

1. **Testing the Phylogenomic Method**: I aim to evaluate the accuracy and limitations of the methodology as described by Orr et al. (2020) by investigating two key factors through simulation.
    - **Mutation Rate Recovery**: The accuracy of the phylogenomic method in recovering inputted mutation rates across varying tree topologies and other biological parameters is assessed using simulated mutations.
    - **Topology Recovery**: A phylogenetic tree is reconstructed from simulated somatic mutations across a tree and compared to the true topology of the tree.
2. **Developing a Novel Approximate Bayesian Computation (ABC) Method:** I introduce a proof-of-concept, non-phylogenetic approach for estimating the somatic mutation rates of long-lived trees utilising Tomimoto and Satake's (2023) simulation framework and an ABC-Reject methodology. This novel method does not assume that mutations strictly follow topology, thus avoiding over-filtering. Instead, the ABC-Reject framework approximates a plausible interval of somatic mutation rates which produce observed mutation frequencies across branches, as well as approximating parameters of interest of meristem behaviour.

All computational experiments were performed under NeSI (New Zealand eScience Infrastructure). I wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support and/or training services as part of this thesis. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment Research Infrastructure programme (https://www.nesi.org.nz.). The execution of experiments primarily involves the use of Jupyter notebook and the programming languages Python and R. The scripts utilised throughout this thesis are available in an open GitHub Repository: https://github.com/andreag186/sim_tree_mut

I also acknowledge the use of OpenAI's ChatGPT-4 for assistance in structuring sections of the thesis, refining written expression, and facilitating idea development, particularly in relation to coding and methodological organization. All research design, intellectual content, methodological decisions, experimental implementations, analyses, and interpretations presented in this thesis are solely my own.

1.6 Thesis Structure

Chapter 2: Evaluation of the Phylogenomic Method

This chapter evaluates the phylogenomic method for estimating somatic mutation rates in long-lived trees. I employ Tomimoto and Satake's (2023) model of hierarchical modular growth to simulate mutation accumulation across 20 tree topologies that vary in branch number, balance, and internal-to-terminal branch ratios. These simulations incorporate four meristematic growth models, capturing variation in elongation (StD; structured or stochastic) and branching bias (biasVar; biased or unbiased).

To refine the phylogenomic approach, I implement a modified phylogenomic method optimized for high-throughput simulation analysis. Mutation rates are estimated using linear regression models, which relate genetic and physical distances between branch pairs, with an extended version incorporating branching events. The accuracy of phylogenetic reconstruction is assessed using Robinson-Foulds (RF) distances, evaluating whether mutations strictly follow topology and whether tree topology can be recovered from mutation patterns. Additionally, I examine the deviation between input and estimated mutation rates to assess the method's accuracy. Generalized linear models (GLMs) are then used to evaluate how tree topology, mutation rate, and meristem parameters influence mutation rate recovery and phylogenetic inference.

Finally, I apply the modified phylogenomic method to empirical *Eucalyptus melliodora* datasets (Orr et al., 2020) before and after topological filtering, comparing simulated mutations generated via the four meristematic growth models defined by Tomimoto and Satake (2023).

Chapter 3: A Novel ABC Approach For Somatic Mutation Inference in Long-Lived Trees

Chapter 3 introduces a prototype Approximate Bayesian Computation (ABC) framework for estimating somatic mutation rates as an alternative to Orr et al.'s (2020) phylogenomic method. The ABC-Rejection algorithm samples mutation rate (μ), elongation (StD), and branching bias (biasVar) from biologically informed priors, simulating mutation accumulation and comparing outputs to observed data using a Euclidean distance metric. Parameters are accepted if their discrepancy falls below a tolerance threshold (ϵ). An iterative sensitivity analysis refines priors and threshold values, revealing that $StD = 0$ behaves as a lower-dimensional model where only branching bias influences mutation accumulation. This model-jumping effect distorts posterior estimates, leading to its exclusion.

To validate the novel ABC method, I apply it to 169 simulated datasets with known true parameters. The method accurately recovers 99.4% of elongation values and 100% of mutation rate and branching bias values within 95% highest posterior density (HPD) intervals.

Finally, I apply the ABC method to empirical *E. melliodora* datasets before and after topological filtering, testing its robustness when applied to real-world data.

Chapter 5: General Discussions

This chapter summarizes my findings, discusses methodological limitations, and outlines potential directions for future research. I highlight the strengths and constraints of the phylogenomic method and the ABC-based prototype framework, considering their utility for estimating somatic mutation rates in long-lived trees.

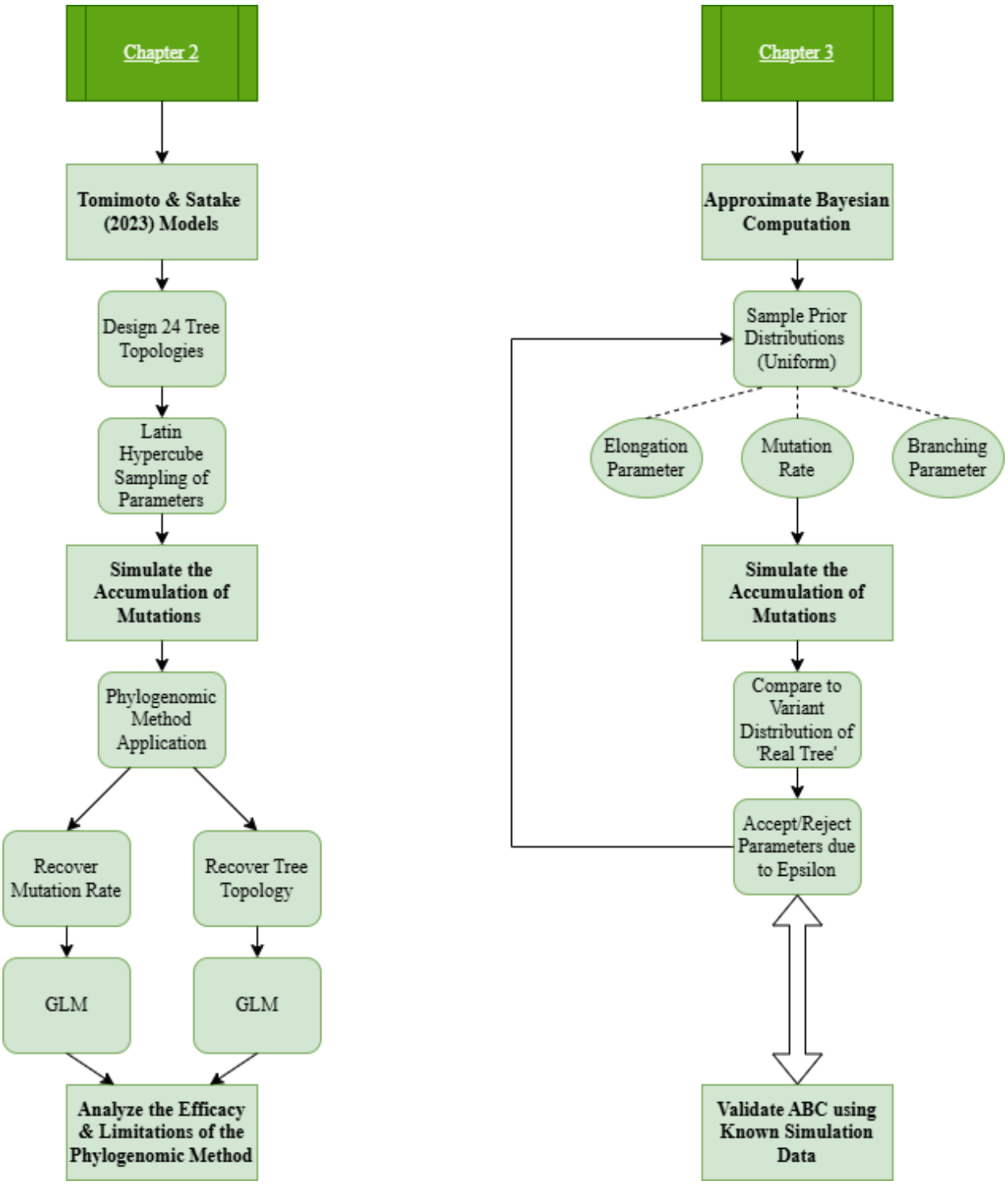


Figure 1.1 Overview of Thesis Structure, Methodology and Objectives
This flowchart illustrates the two main components of the thesis: evaluating the phylogenomic method using simulated mutations (Chapter 2) and optimizing somatic mutation rate estimation using Approximate Bayesian Computation (Chapter 3).

Chapter 2: Evaluation of the Phylogenomic Method

2.1 Overview

This chapter critically evaluates the validity of the phylogenomic somatic mutation rate estimation methodology, as introduced by Orr et al. (2020), under diverse biological conditions. To test the limitations of the phylogenomic method, I employ the hierarchical modular models defined by Tomimoto and Satake (2023) to simulate mutation accumulation across a range of input trees and topologies. Twenty-four unique tree topologies are defined, varying by the number of terminal branches, the topology balance and the ratio of internal to terminal branch lengths. These unique tree topologies are representative of the vast diversity of biological tree topologies, combined with varying somatic mutation rates and meristem behaviour parameters across simulations.

The mutations generated from simulations will then be applied to the 'modified' phylogenomic method (a modified approach designed for simulations) to recover the somatic mutation rate. I further assess the recovery of the input tree topology, constructing phylogenetic trees from the somatic mutations generated from the simulations. The accuracy of the estimated somatic mutation rate and topology will inform under which conditions, if any, the phylogenomic method is appropriate to estimate the somatic mutation rate of long-lived trees.

2.2 Methodology

2.2.1 Model

Tomimoto and Satake (2023) developed a modelling framework to describe the dynamic meristem processes which underpin the accumulation and expansions of somatic mutations in long-lived trees throughout their growth. Modelling meristem behaviour and resulting somatic mutations throughout growth processes can help explain the phenomenon of genetic mosaicism through a single tree and further aid in assessing the applicability of the phylogenomic method.

The model considers the modular nature of tree growth, in which the final topology of a tree is defined by the reiteration of two independent yet interconnected growth processes: elongation and branching (Figure 2.1). In the following sections, the components of Tomimoto and Satake's model are described in detail.

2.2.1.1 Elongation

Elongation can be defined as the vertical growth of the stem of a plant due to successive divisions of a collection of stem cells positioned in the shoot apical meristem (SAM).

Tomimoto and Satake (2023) define two types of elongation, structured and stochastic, representing the extremes of SAM structures across seed plants. For simplicity, the model assumes that the number of stem cells (α), remains constant throughout both elongation and branching processes (α = 5). The elongation rate is defined as r_e, relating to the number of cell divisions per generation of elongation. When r_e= 1, there is one division per year of growth. Tomimoto and Satake define a parameter StD, representing the number of stem cells (α) that do not maintain their lineage.

- When StD = 0, elongation is structured.

In structured elongation, throughout r_e cell divisions, one of the daughter cells is differentiated into a defined tissue type while the other remains, keeping the stem cell lineage constant (Figure 2.1). Each α cell is thus a direct descendant of its parent cell.

- When StD = α, elongation is stochastic.

In stochastic elongation, no α daughter cells are eliminated through r_e cell divisions. When the number of cell divisions during elongation per unit length is r_e, the size of the daughter cell pool becomes $\alpha 2^{Re}$. For each generation of elongation, α cell initials are sampled randomly from the accumulated $\alpha 2^{Re}$ daughter cells (Figure 2.1).

2.2.1.2 Branching

Branching can be defined as the formation of an axillary meristem from an apical meristem, which is responsible for the growth of a lateral branch (Figure 2.1).

Tomimoto and Satake define two types of branching, biassed and unbiased, representing the extremes of axillary meristem formation throughout plant taxa. In order to model the branching process, they assumed that α stem cell initials in the SAM proliferate radially over r_b successive cell divisions. The stem cell initials for the newly formed axillary meristem are sampled from $\alpha 2^{r_b}$cells.

The stem cells selected for the axillary meristem are then arranged along the circumference of a unit circle. Cells generated from the same mother cell are positioned closer to each other, reflecting their shared lineage. Let $x \in [0, 2\pi)$ represent the location of a stem cell on the circumference. The probability of a cell occupying position $x \in [0, 2\pi)$ being sampled to form the axillary meristem is described as $f(x; u, \sigma)R$ where;

- $f(x; u, \sigma)$ is the probability of a wrapped normal distribution
- $R = 2\pi/\alpha 2^k$ is the standardised constant representing the area occupied by a single cell

The probability density function of the wrapped normal distribution is further defined as:

$$f(x; u, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{s=-\infty}^{\infty} exp\left(-\frac{(x+2\pi s-u)^2}{2\sigma^2}\right)$$

where $x \in [0, 2\pi)$ is the mean position of a given stem cell, and $\sigma > 0$ is the standard deviation of the distribution. The value of σ (or biasVar) determines the nature of sampling.

- When $\sigma \geq 2\alpha$, the wrapped normal distribution approaches uniformity. This means the stem cell initials are sampled randomly from the entire circumference to populate the axillary meristem. Proliferating cells migrate extensively across lineage boundaries, leading to 'well-mixed' lineages in the axillary meristem. **Thus, branching is unbiased**.

- When $\sigma \leq \frac{1}{2}\alpha$, stem cells near the position $x = u$ are more likely to be selected to populate the axillary meristem. Stem cells closest to the 'flank' or edge of the SAM where an axillary meristem is formed are more likely to be selected. The lineages of the axillary meristem are spatially specific. **Thus, branching is biassed**.

For each branching event, a position u is randomly selected along the SAM circumference $[0, 2\pi)$.

2.2.1.3 Somatic Mutation Accumulation

The accumulation and distribution of somatic mutations throughout a single tree is greatly influenced by both the elongation and branching processes, as well as the structural configuration of meristems. In this model, mutations are introduced at a rate of v per cell division, with each mutation randomly affecting a unique genomic site (no mutations may hit the same genome site).

The mutation rate per site is defined as $\mu = \frac{v}{G}$ (where G represents the genome size) and must be low enough in order to disregard back mutations. For each cell in the meristem, the mutation state of a particular site k is represented as:

$$m_{ik}^{(n)} = \begin{cases} 0, & \text{if the site is non-mutated} \\ 1, & \text{if the site is mutated} \end{cases}$$

Where $m_{ik}^{(n)}$ denotes the mutation state of site k in the i^{th} stem cell of branch n.

In order to capture how mutations propagate through successive cell division, the probability distribution of mutated cells is represented using a state vector. Let $\pi(t) = \left(\pi_0(t), \pi_1(t), \dots, \pi_\alpha(t)\right)$ denote the state vector, where $\pi_i(t)$ represents the probability of having i mutated cells out of α stem cells at time t. The mutation model differs for structured and stochastic meristem configurations:

- **Structured Elongation Model**: In structured elongation, where cell lineages are preserved, the number of mutated cells changes deterministically through each division. Transition probabilities from i to j mutated cells follow a binomial distribution:

  $$q_{ij} = \binom{a-i}{pj-i} \mu^{j-i} (1-\mu)^{(a-i)-(j-i)}$$

  Here, q_{ij} describes the probability of gaining additional mutated cells without losing existing mutations. Using this transition matrix $Q = [q_{ij}]$, the probability distribution of mutated cells over time is updated as:

  $$\pi_{str}(t) = \pi(0) Q^t$$

- **Stochastic Elongation Model:** In stochastic elongation, stem cells are sampled randomly after each division, introducing the potential for loss of lineages. The probability of mutation accumulation and loss is calculated with a transition matrix P that accounts for the random sampling of daughter cells:

  $$\pi_{sto}(t) = \pi(0) P^t$$

  where P represents the likelihood of each mutated cell state transitioning under stochastic lineage selection.

The probability that a mutation is present at a given site within the meristem is given by $1 - \pi_0(t)$, indicating the likelihood that at least one cell in the meristem carries the mutation. This probability formula applies to both stochastic and structured models.

The expected number of mutated cells at any given time can be calculated by summing the probabilities of each mutation state in the state vector. This framework allows for the simulation of somatic mutation accumulation patterns throughout a tree under both structured and stochastic meristem behaviour.

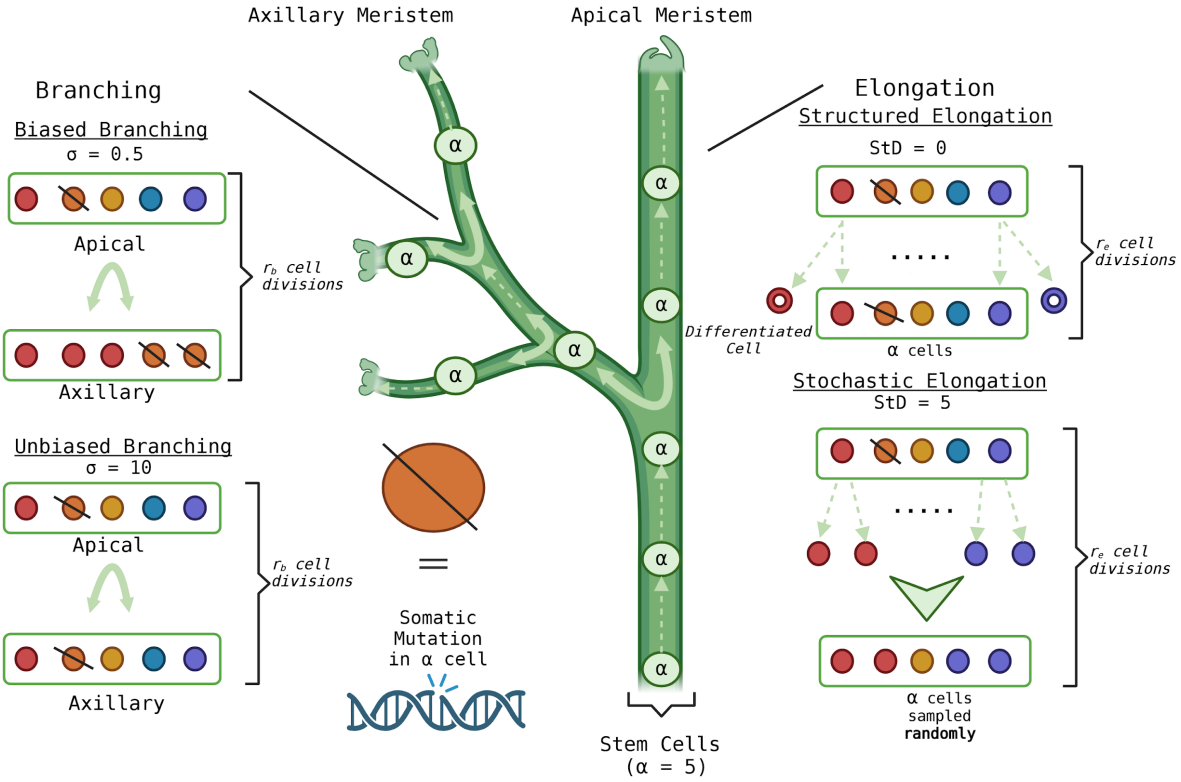


Figure 2.1 Schematic Representation of Elongation and Branching Models in the Apical Meristem and Axillary Meristem Formation. Meristems are represented by green circles denoted with α**;** each meristem contains five stem cells (α = 5). Stem cells are represented by circles, with their colour denoting their lineage. A stem cell which is 'dashed' with a diagonal line represents the presence of a somatic mutation.

A dashed arrow denotes elongation in the stem (and along branches). During elongation (right panel), each stem cell in the meristem divides to produce new cells. In structured elongation (StD = 0), lineages are preserved, with one daughter cell differentiating (shown as a circle with a white inner core) and the other retained. In stochastic elongation (StD = 5), cell lineages are randomly sampled after each division. Elongation is repeated for r_e cell divisions (in my application, $r_e = 1$).

Branching, the process of forming an axillary meristem from a shoot apical meristem, is represented by a solid curved arrow at the branch node. In biassed branching (σ = 0.5), cells closer in lineage are more likely to contribute to the new meristem (noted by colour). Unbiased branching (σ = 10) assumes equal probability for all cells to contribute, resulting in well-mixed lineages. r_b denotes the number of cell divisions involved in axillary meristem formation (in my application, $r_b = 7$).

Created with BioRender.com

2.2.1.4 Models Summary

Overall, the model examines how the structure and behaviour of meristems throughout elongation and branching affect the accumulation and distribution of somatic mutations within a tree. Two main factors are considered in this modelling framework: elongation (structured or stochastic) and branching (unbiased or biassed). Combining these elongation and branching types results in four extreme model configurations, each representing a different pattern of lineage maintenance and mutation accumulation across branches (Table 2.1).

	UNBIASED BRANCHING	**BIASSED BRANCHING**
STRUCTURED ELONGATION	Model I. Structured-Unbiased Model Stem Cell Lineage Maintained in Elongation Stem Cell Lineage Randomly Replaced in Branching	Model II. Structured-Biassed Model Stem Cell Lineage Maintained in Elongation Stem Cell Lineage Biasedly Replaced in Branching
STOCHASTIC ELONGATION	Model III. Stochastic-Unbiased Model Stem Cell Lineage NOT Maintained Stem Cell Lineage Randomly Replaced in Branching	Model IV. Stochastic-Biassed Model Stem Cell Lineage NOT Maintained Stem Cell Lineage Biasedly Replaced in Branching

Table 2.1 Classification of Four Models of Somatic Mutation Accumulation across Trees as defined by Tomimoto and Satake (2023).

These four models capture the extremes of possible elongation and branching behaviours, allowing us to simulate how each combination might affect the number and distribution of somatic mutations across branches. Furthermore, these models allow us to simulate how different meristem behaviours across taxa may influence the application of the phylogenomic method.

2.2.2 Parameter Testing and Latin Hypercube Sampling

2.2.2.1 - Parameters

The four extreme models Tomimoto and Satake (2023) developed were tested across a range of input parameters, each subjected to multiple treatment levels (Table 2).

i) **Tree Topology:** To identify potential limitations of the phylogenomic method, I examine tree topologies that vary across both taxa and individuals. This thesis considers two extreme categories of tree topologies: balanced and unbalanced. In this thesis, a balanced topology is characterised as having terminal branches equally distributed after the initial node split, leading to a symmetrical tree structure. In contrast, an unbalanced topology has more

branches on one side (e.g., more on the right side after the first split), resulting in an asymmetrical structure (Shao & Sokal, 1990). This asymmetry can cause mutations to concentrate in specific lineages and complicate the recovery of the branching structure from the phylogenomic topology constructed using mutations. *

ii) **Internal to External Branch Ratio:** The effect of differing ratios of internal branches to external (aka. terminal) branches will also be examined. Internal branches connect two nodes, while external branches terminate in a single node (Sereno, 2005). The ratio of internal nodes to external nodes impacts the accuracy with which mutations can be mapped across a phylogenetic tree, impacting estimated mutation rates (Nielsen, 2002). Thus, it is essential to account for introduced uncertainty due to the external to internal ratio of branches within the physical topology of a tree itself. *

iii) **Number of Terminal Branches**: To evaluate the limitations of the simulation models proposed by Tomimoto and Satake (2023), the number of physical branches was expanded beyond previous studies (4,6,8,10, 12 branches). This allows us to examine whether the model can accommodate more complex branching structures and identify any possible constraints.*

iv) **Mutation Rate Relative to Genome Length:** A composite parameter comprising the product of the somatic mutation rate of a given tree relative to its genome length will be utilised. Both the magnitude of the somatic mutation rate and genome length could introduce noise, impacting the accuracy of somatic mutation rate estimation. The magnitude of each component is proportional to one another.

*In order to vary parameters i-iii, 20 input trees were designed. For each value of terminal branches, four trees were constructed combining balanced/unbalanced topologies and short /long (terminal branch) internal:external ratios (Table 2.2). Each tree had an approximate age of 310 years and a total physical length (not height) of 31 metres. These values were arbitrarily selected.

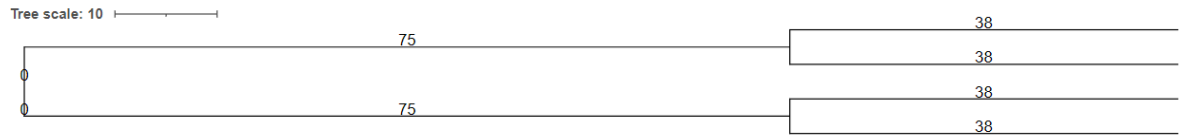
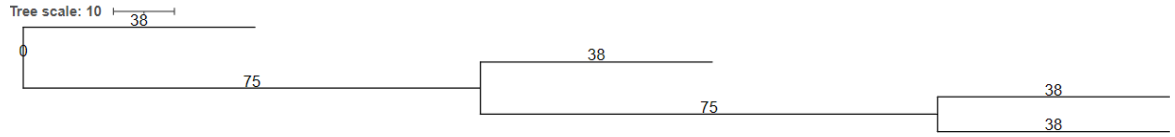
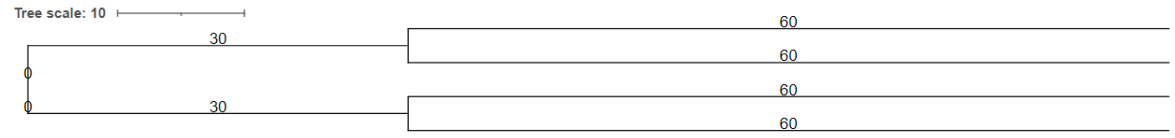
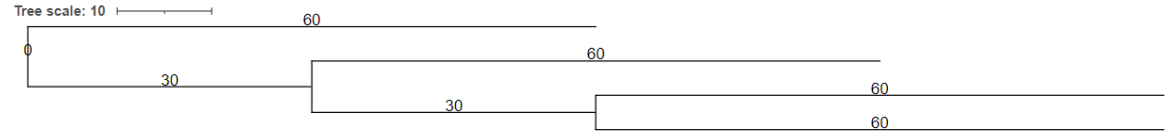
1) Balanced Tree with Short Terminal:Long Internal Branches

2) Unbalanced Tree with Short Terminal:Long Internal Branches

3) Balanced Tree with Long Terminal: Short Internal Branches

4) Unbalanced Tree with Long Terminal: Short Internal Branches


Table 2.2 Example Topologies of Input Trees (Number of Terminal Branches = 4). Topologies of 4 out of the 20 total input trees sampled, varying by topology balance and internal:terminal branch ratios. Branch lengths (shown in years) represent both physical length in centimetres and approximate age (10cm/year growth rate). Tree scale (100cm) shown above for reference; the total length of each tree = 3100 cm, and the age = 310 years (approximately).

2.2.2.2 - Latin Hypercube Sampling

In order to appropriately sample the high-dimensional data space, Latin Hypercube Sampling is employed, with a sample size of 12,000. Latin Hypercube Sampling (LHS) was selected as it is more computationally effective at sampling a large number of parameters than grid sampling (Shields & Zhang, 2016). Within LHS sampling, each parameter represents an axis in multi-dimensional space, collectively forming the 'sample space' (Loh, 1996). This sample space includes a combination of biological and technical parameters, each with distinct treatment levels. These treatment levels allow us to explore various combinations of parameter values systematically.

Each parameter's range is partitioned into equal intervals from which a value is selected, ensuring all regions of the sample space are well-represented. Traditional random sampling methods can lead to a biassed representation, with some regions of the parameter space being over or under-sampled. In order to achieve a comprehensive coverage of sample space, a large number of samples are typically

required to mitigate the inherent error of the random sampling method. In contrast, LHS guarantees that the whole sample space is explored within a smaller, pre-specified sample number, reducing the computational resources required to evaluate the sample space fully. By ensuring that all parameters are sampled representatively across their range of values, LHS ensures that critical parameter interactions are adequately represented, helping us better understand model performance under different scenarios.

Simulation Parameter	**Treatment Levels**	**Associated Methods**
T- Tree Topology Balance + i: e- Internal:External Branch Ratio	**4:** Combinations of Balanced and Unbalanced Topologies and Internal to External ratio distributions	Design 4 trees (per each terminal branch value) with combinations of balanced/unbalanced topologies and short/long i:e ratios. 20 Trees constructed in total.
B- The Number of Physical Branches within a Tree	**5:** 4, 6, 8, 10, and 12 terminal branches	Increase branch number of biological trees to extend upon prior Literature (Orr et al., 2020; Hoffmeister et al., 2020).
$G\mu$- Mutation Rate (per cell division) x Genome Length	**6:** 3 small mutation rates (4,6,8 × 10E-10) and 3 large mutation rates (4,6,8 × 10E-10). Genome size (G) = 500Mb	Represent range of estimated somatic mutation rates across taxa of long-lived trees.
StD - Elongation Parameter	**2:** 0 for structured elongation and 5 for stochastic elongation (discrete values)	Elongation parameter values for models assigned by Tomimoto and Satake (2023)
$biasVar$ (σ) - Branching Parameter	**2:** 0.5 for biassed branching and 10 for unbiased branching (discrete values)	Branching parameter values for models assigned by Tomimoto and Satake (2023)
α- Number of Stem Cells in Meristems	**1:** Value = 5	Constant Value Assumed (Tomimoto & Satake, 2023)
r_e - Number of cell divisions during elongation	**1:** Value = 1	Constant Value Assumed (Tomimoto & Satake, 2023)
r_b- Number of cell divisions during branching	**1:** Value = 7	Constant Value Assumed (Tomimoto & Satake, 2023)
u- Center position of newly formed branch	**1:** Value = $u \sim U(0, 2\pi)$	Constant Value Assumed (Tomimoto & Satake, 2023)

Table 2.3 Parameters and Associated Treatment Levels Sampled via LHS. Parameters highlighted in grey are constant values identical to those used in Tomimoto and Satake's (2023) Simulation

Method. Parameters highlighted in green are varied to assess the limitations/applicability of the phylogenomic somatic mutation rate estimation methodology.

2.2.3 Simulation and Model Implementation

This section details the simulation methodology used within Chapter 2 to assess the applicability and limitations of the phylogenomic method across varying trees. The simulation pipeline, implemented in Python (v3.11.4), is available in the open Github repository **sim_tree_mut** (https://github.com/andreag186/sim_tree_mut). The code consists of several functions, summarised in Table 2.4, which handle various stages of the pipeline, from parameter sampling to mutation simulation and modified phylogenomic method application.

Function Name	**Purpose**	**Key Outputs**
`latin_hypercube_sampling`	Generate diverse parameter sets for simulations- there are 6 parameters to define (refer to Appendix B)	Parameter sets for simulations
`write_samples_to_csv`	Save simulations results to a CSV file	CSV of input parameters and outputs
`create_tree_list_and_dict`	Decode input tree configurations and generate branch metadata	`tree_list` `tree_dict` `numBranch` `age`
`simulate_somatic_mutations`	Simulate somatic mutations using Tomimoto & Satake (2023) model	Mutation matrices for each branch
`mutInStemCells`	Simulate mutations during stem elongation prior to branch formation	Mutation history of stem cells (`tCells`)
`makeMutMatrix`	Create a binary mutation matrix indicating mutation presence/absence across branches	Binary mutation matrix
`calc_variants`	Estimate somatic mutation rate using regression of genetic vs. physical distance matrices	Mutation rate, regression equation

Table 2.4 Description of the Main Functions in the Simulation Code (sim_code.py)

The simulation begins with Latin Hypercube sampling (`latin_hypercube_sampling`), generating diverse parameter sets based on input variables (e.g., mutation rate, tree topologies, elongation, and branching parameters).

Input trees are decoded using `create_tree_list_and_dict`, and the accumulation of somatic mutations are simulated with `simulate_somatic_mutations`. Mutation matrices are constructed using `makeMutMatrix`, and genetic and physical distance matrices are generated. Finally, `calc_variants` applies the modified phylogenomic method to estimate somatic mutation rates.

Overall, the sim_code.py script enables the simulation of somatic mutations across a range of input trees and estimates a somatic mutation rate using the modified phylogenomic method. Running the script (with sufficient samples) enables a complete replication of my results. Other input trees of varying branch number, length and input values can also be assessed by modifying the tree_dict dictionary, following the nomenclature pattern defined by Tomimoto and Satake (2023) (Appendix B, Fig 1). A detailed explanation of the simulation code's functions, parameters and caveats can also be found in Appendix B.

2.2.4 Regression Analysis and Statistical Significance Testing

To evaluate the accuracy of the modified phylogenomic method in recovering somatic mutation rates, I performed linear regressions between input and e mutation rates across 12,000 simulations spanning 20 tree topologies (Figure 2.2). Each regression followed the model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where y is the estimated mutation rate, x is the input mutation rate, β_1 represents the regression slope, and ϵ is the residual error.

To assess whether regression slopes significantly deviated from 1, I performed two-tailed t-tests for the null hypothesis H_0: $\beta_0 = 1$, calculating the p-value as:

$$p = 2 \times (1 - T(|t|, df))$$

Where $t = \frac{(\beta_1 - 1)}{SE,SE}$ is the standard error of the slope, and df is the degrees of freedom. Bonferroni correction was then applied by multiplying raw p-values by the total number of tests ($n = 40$) to control for multiple comparisons (Sedgwick, 2012).

2.2.5 Robinson-Foulds Distance Calculation

Reconstructed phylogenetic trees are compared with the original input topologies using the Robison-Foulds distance metric to evaluate topology recovery from simulated mutations.
The RF distance metric measures dissimilarity between phylogenetic trees, counting unique bipartitions present in one tree but absent in the other (Böcker et al., 2013). A minimum score of zero indicates identical topologies, while the maximum RF distance score, representing 'opposite' topologies, is defined as follows (*where n = the number of terminal branches*):

$$Max\ RF\ Dist. = 2 \times (n - 3)$$

Using the simulated unique, shared and total mutations per branch, a distance matrix is constructed which quantifies genetic dissimilarity between branches. Unlike methods that rely on discrete site-based mutation data, this approach uses averaged mutation counts derived from state vectors in simulations. Since state vectors store probabilities of different mutation states rather than fixed counts, the resulting mutation estimates are continuous rather than discrete. This reflects an aggregate

measure of mutation events rather than precise mutation site patterns*. This matrix is the foundation for phylogenetic tree reconstruction through clustering algorithms like Neighbour-Joining (NJ). The NJ algorithm iteratively joins pairs of branches with the smallest genetic distance, minimising total branch lengths to fit the input distance matrix best and producing an output phylogenetic tree (Gascuel, 2006). The reconstructed tree topology of the input tree (represented in Newick format) is then compared to the input, using the RF distance metric to calculate the degree of similarity between the topologies. The RF distance was standardised by dividing the raw RF distance by its maximum possible value for each tree. This standardisation scales the RF distance to 0 and 1, where 0 indicates perfect recovery, and 1 represents maximum dissimilarity—using standardised RF distance allowed for direct comparisons of topology recovery across all of the 20 input tree topologies (Results 2.3.1.3, Figure 2.6).

In addition to NJ, I trialled alternative tree reconstruction methods, including UPGMA (Unweighted Pair Group Method) and parsimony-based approaches. UPGMA assumes a molecular clock or constant evolutionary rate, requiring a defined evolutionary model (Moulton et al., 2018). This assumption is incompatible with simulation data, with the simulated mutations produced from Tomimoto and Satake (2023)'s model not conforming to clock-like behaviour. Parsimony methods, which aim to find topology requiring the fewest evolutionary changes, rely on sequence-level data or unique mutation sites (Jin & Nei, 1990). While Orr et al. (2020) successfully employed parsimony-based methods for phylogenetic reconstruction, their approach was tailored to whole genome sequence data. My simulation framework, in its current implementation, as defined by Tomimoto and Satake (2023), does not provide discrete mutations for unique sites but instead yields estimates for unique, shared, and total mutations per branch by summing mutation state probabilities from state vectors. This feature complicates the direct application of parsimony and requires additional assumptions or weighting schemes, limiting its utility in this thesis.

Orr et al. (2020) further utilised the distance metric path distance (PD) to compare topologies. The PD distance metric calculates the total branch lengths between data pairs and can be applied across various tree reconstruction methods. However, since the data did not include accurate branch lengths but relied on mutation averages, path distance was also unsuitable for this analysis. Ultimately, the NJ algorithm, combined with the RF distance metric, was most appropriate for reconstructing and evaluating trees from the simulation data.

This analysis was implemented using the custom R script `full_newick.R`, included in my published repository (https://github.com/andreag186/sim_tree_mut). This script requires two inputs: a csv file (the output from the simulations directly generated by `sim_code.py`) and the Newick strings of each input tree defined in tree_dict. The script automates the creation of distance matrices, performs NJ clustering, and calculates and outputs RF distances (in a specified csv file) between reconstructed and input topologies.

**While the Neighbour-Joining (NJ) method does not require distance matrices to be derived from discrete site-based mutation data, it relies on pairwise measures of dissimilarity between taxa (Gascuel, 2006). Using continuous mutation counts smooths stochastic variability in simulations, aligning with NJ's flexibility in accepting diverse input distances (Kuhner & Felsenstein, 1994). Although I found no explicit objections to averaging mutation counts in NJ, this approximation may impact branch length estimation, requiring careful interpretation of topology recovery results.*

2.2.6 Generalised Linear Model Analysis

The generated simulation data involves a wide range of input parameters, including variations of input tree topologies and mutation models. These parameters influence two key outputs that assess the phylogenomic method: the estimated somatic mutation rate and the accuracy of phylogenetic tree recovery, measured by RF distance. Understanding how these input parameters- and their interactions- impact the key outputs is essential for understanding the applicability or feasibility of the phylogenomic method.

Generalised Linear Models (GLMs) were employed to address the multifactorial relationships across simulations. GLMs are a flexible extension of linear regression that can handle continuous and categorical predictors, model response variables with various distributions, and evaluate interaction effects between predictors (Nicholls, 1989). This flexibility makes GLMs well-suited for disentangling the complex, multifactorial relationships in the simulation data.

The GLM analyses were implemented in a Jupyter Notebook (utilising Python version 3.11.4), `glm.ipynb`, available in the repository (https://github.com/andreag186/sim_tree_mut). Here, I detail each step outlined in the Jupyter Notebook.

Data Preparation and Preprocessing

- Mutation Rate Conversion: Estimated somatic mutation rates (`output_mut_new`) were normalised to a per site, per year basis using the formula (and the total age of each input tree, which was set to 310 years):

$$Mutation\ rate\ (per\ site, per\ year) = \frac{Variants}{Genome\ Size\ x\ 310\ (years)}$$

  Observations with zero or negative total variant values were excluded (assumed outliers).

- RF Distance Standardisation: RF distances were standardised by dividing by the maximum RF distance for each topology -determined by terminal branch number. This ensured that RF distance comparisons across different topologies were on the same scale, facilitating fair assessments of tree recovery accuracy.

- Z-Score Transformation: To ensure comparability across predictors, all continuous variables were transformed into Z-scores using the formula:

$$Z = \frac{x-\mu}{\sigma}$$

  where x is the observed value, μ is the mean, and σ is the standard deviation. Under this transformation, each unit change in a parameter represents one standard deviation unit (Colan, 2013). Consequently, the slopes (coefficients) in the GLM represent the change in the response variable (in standard deviation units) for a one-standard-deviation change in the independent variable. This transformation facilitates interpretation and ensures model stability, particularly for interaction terms.

Statistical Modeling with GLMs

Two GLMs were implemented to assess different output variables:

1. Mutation Rate Analysis: The output mutation rate was modelled as the dependent variable, while input mutation rate, topology balance, branch ratio, terminal branch number, elongation parameter, branching parameter, and their two-way interactions were the independent

variables. This choice reflects the aim of assessing how The model was specified using the formula API in `statsmodels`:

Output Mutation Rate ~ Input Mutation Rate + Topology Balance + Branch Ratio + Terminal Branch Number + Elongation Parameter + Branching Parameter + 2-way interactions

2. RF Distance Analysis: The RF distance was modeled as the dependent variable, using the same variables as in the mutation rate analysis to ensure compatibility. The formula was:

   RF Distance ~ Input Mutation Rate + Topology Balance + Branch Ratio + Terminal Branch Number + Elongation Parameter + Branching Parameter + 2-way interactions

A Gaussian family was selected for both GLMs as it is the default choice for continuous response variables in the `statsmodels.formula.api.glm` package (Maia & Bonat, 2024). Maximum likelihood estimation (MLE) was used to fit the models, ensuring optimal parameter estimation. Model summaries were generated to interpret the results, including coefficients, significance levels, and goodness-of-fit metrics.

Partial Eta-Squared Effect Sizes

To measure the contribution of each predictor, partial eta-squared (η_p^2), was calculated. This statistic quantifies the proportion of variance in the response variable explained by a predictor, while adjusting for other terms in the model. Partial eta-squared is computed as:

$$\eta_p^2 = \frac{SS_{Effect}}{SS_{Effect} + SS_{Residual}}$$

Where SS_{Effect} represents the sum of squares for the predictor or interaction of interest, and $SS_{Residual}$ the residual sum of squares. This metric highlights the relative importance of predictors, enabling us to compare their effects directly.

Partial eta-squared (η^2) was calculated for main effects to quantify their contribution to mutation rate recovery. However, η^2 is not well-suited for two-way interactions due to shared variance between predictors, making independent partitioning problematic (Richardson, 2011). In factorial designs, partial η^2 values can sum to more than 1, complicating interpretation, and their magnitude depends on the number of predictors included (Kennedy, 1970). As a result, interaction terms were evaluated using p-values for significance and beta coefficients to describe their effect size and direction.

Diagnostic and Model Fit Evaluation

Diagnostic plots were generated to validate model assumptions and evaluate fit:

- Residuals vs. Fitted Values: To assess the linearity and homoscedasticity of residuals
- Predicted vs. Actual Values: To evaluate the model's predictive accuracy

These diagnostics ensured the reliability of the GLM analyses and provided an insight into model performance.

2.2.7 Back-mutation Analysis via Poisson Distribution

A back mutation, also known as a reverse mutation, occurs when a nucleotide that has mutated from its original state reverts to its ancestral state at the same time (Loewe & Hill, 2010). Potential back mutations were modelled as rare stochastic events using a Poisson distribution to estimate their presence throughout simulations. The Poisson probability of observing k back mutations at a given site is defined as:

$$P(k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

Where λ represents the expected number of back mutations per site. This was calculated as follows:

$\lambda = \mu^{2} \times Genome\ Size \times Total\ Divisions$

Here, μ is the mutation rate per size per division, genome size is fixed at 500Mb, and the total number of divisions included 300 elongation divisions (total age of each tree = 310 years, r_e = 1, meaning there is 1 division per year of growth) and 7 division per branching event (r_b = 7, the number of cell divisions per axillary meristem formation across 4-12 branches).

Back mutation probabilities, $P(k)$, were analysed across six defined input mutation rates spanning 4×10^{-10} to 8×10^{-9} across the 20 input tree topologies (Age: 310 years, 4-12 branches).

Although Tomimoto and Satake's (2023) model operates under an infinite sites assumption (where no site is hit twice), this analysis was performed to validate that back mutations remain negligible across the range of input mutation rates used in simulations. At larger mutation rates, deviations from the infinite site assumption could theoretically arise due to high mutation densities, influencing results. By quantifying back mutation probabilities using a Poisson model, I aimed to confirm that back mutations are not a significant confounding factor of the results.

2.2.8 Shared Mutation Dispersion Analysis

The distribution of shared mutations between branches can provide insight into how varying mutation rates can influence phylogenetic distinctiveness. In phylogenetic studies, excessive substitutions at the same sites can obscure evolutionary relationships, a phenomenon sometimes referred to as mutation saturation (Xie et al., 2023). While this study does not directly assess site-specific substitution accumulation, I apply a related approach to examine how shared mutations distribute across branches under different mutation rate conditions.

For each mutation rate category (small: $4, 6, 8 \times 10^{-10}$; large: $4, 6, 8 \times 10^{-9}$), shared mutation dispersion was quantified using the following metrics:

- Mean Shared Mutations: The average number of shared mutations across all branch pairs
- Variance of Shared Mutations: The extent of variability in shared mutations across branch pairs.
- Coefficient of Variation (CV) (Brown, 1998): A normalised measure of dispersion, calculated as:

$$CV = \frac{Standard\ Deviation}{Mean}$$

A decline in variability relative to the mean suggests an increasing uniformity of shared mutations across branches, potentially reducing the distinctiveness of individual branches. This analysis allows

us to determine whether higher mutation rates lead to a convergence in shared mutation patterns, impacting the applicability of the phylogenomic method.

2.2.9 Application of Models to *Eucalyptus melliodora*

2.2.9.1 - Application of Simulation Code

To apply the models defined by Tomitmoto and Satake (2023), I directly requested access to the original simulation source code. This code was kindly provided by Sou Tomimoto and Professor Akiko Satake of the Mathematical Biology Laboratory, Kyushu University, and was subsequently used in all further analyses described in this thesis (with modifications, where appropriate).

The cell division parameter values for elongation and branching (r_e and r_b) were kept identical to the constant values which Tomimoto and Satake (2023) defined due to the lack of additional empirical data specific to meristem behaviour of the *E. melliodora* individual. The rate of elongation cell division (r_e) was set to 1, meaning that the number of cell divisions during elongation was assumed to be one per year to produce the unit length of growth. The rate of cell division during the formation of an axillary meristem (r_b, associated with branching events) was set to 7, consistent with literature values for model species such as *Arabidopsis thaliana* and *Solanum lycopersicum* (aka. tomato) (Burian et al., 2016).

The simulation code requires several input parameters, including the number of branches, the relative age of each internal node and terminal branch, and an input somatic mutation rate per cell division per cell for each input tree. To adapt the model for the *E. melliodora* individual, the age of each internal and terminal branch was estimated by utilising the physical distances (in metres) measured by Orr et al. (2020). Although there is no direct literature on the average annual growth rate of *E. melliodora*, this estimate was based on studies of a closely related Eucalyptus species, specifically *E. grandis*. Intensively managed plantations of *E. grandis* have reported substantial annual increments in wood volume and height growth due to silvicultural practices like fertilising, thinning and pruning (Forrester et al., 2010). Using these findings, I adopted an approximate annual growth rate of 10 cm per year to convert measured branch lengths (in centimetres) into estimates of branch age (in years). I recognize that this approximation introduces potential errors due to environmental, life history and interspecies differences between *E. melliodora* and *E. grandis* (Baker, 2003). However, in the absence of better data, this estimate provides a reasonable approximation.

To determine the average somatic mutation rate per cell division per cell, an average of the estimated mutation rate per year per site reported by Orr et al. (2020) was calculated. This value was then divided by r_e (set to 1) and multiplied by the genome size of *E. melliodora*, estimated to be approximately 500Mb, as obtained from whole genome sequencing reads by Orr et al. (2020). This estimation provided the somatic mutation rate for model simulations (6.18×10^{-10} per cell division per cell).

I subsequently applied the four extreme models representing different meristematic structures and behaviour to the *E. melliodora* individual and compared the distribution of somatic mutations against empirical data derived from genomic sequencing reads provided by Orr et al. 2020. Specifically, I compared the distributions of variants per branch before and after filtering the data using

DeNovoGear (DNG) to account for the *E. melliodora* phylogeny - in other words, before and after applying the phylogenomic method (Ramu et al., 2013). The pre-dng distribution was determined using the distribution of 99 high-confidence variant sites containing 'putative' somatic mutations that remained after GATK variant calling, and applying the false positive and recovery rate as defined by Orr et al., resulting in a distribution of 330 'true' mutations. The post-dng distribution was taken directly from the 90 high-confidence variants Orr et al. defined after filtering with *E. melliodora* topology. By comparing the mutation distributions across branches for both pre-dng and post-dng filtered data, I aimed to assess the impact of the phylogenomic correction on the mutation distribution pattern.

2.2.9.2 - NMRSE Calculation

The comparisons were then extended to the simulated distributions from each extreme model. In order to quantify the discrepancy between real and simulated data, I utilise the Normalised Mean Root Square Error (NMRSE) as the metric for model evaluation. This choice was motivated by the significant variation observed between the stochastic and structured models and the empirical data, necessitating a standardised comparison metric (Otto, 2019). NMRSE is advantageous in this context as it provides a measure of deviation relative to the mean, allowing us to evaluate both the fit of simulated mutation distributions to the real data and the degree of variation across branches.

To calculate the NRMSE, the Root Mean Square Error (RMSE) is first calculated, which is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y})^2}{n}}$$

Where y is the ith observation of y and $\widehat{y}$ is the predicted y value given by the simulation model. A small RMSE value indicates that the simulated values (variants per branch) are close to the real values.

In order to normalise the RMSE, the difference between the maximum and minimum variants is calculated for each branch as follows:

$$NMRSE = \frac{RMSE}{y_{max} - y_{min}}$$

2.2.10 Phylogenomic Method Application

2.2.10.1 - Orr et al. (2020) Pipeline

Orr et al. (2020) developed a bioinformatic pipeline which aimed to measure somatic mutations across the branches of a phenotypically mosaic *E. melliodora* individual, estimating a final somatic mutation rate. The pipeline was published in an open-access repository (github.com/adamjorr/somatic-variation) alongside the input raw genome reads of each branch (https://www.ncbi.nlm.nih.gov/bioproject/553104) and a reference genome of a *Eucalyptus grandis* individual provided by Myburg et al. (2014) (https://www.ncbi.nlm.nih.gov/nuccore/AUSX00000000).

I first attempted to reproduce the results obtained by Orr et al. 2020 by running the pipeline as specified by the instructions in the provided repository. The pipeline is structured around sequential makefiles referencing numerous third-party scripts. In my attempt to replicate the pipeline, I encountered complications due to workflow decay, dependencies and substantial resource demands, limiting reproducibility. These challenges are described in detail in Appendix A.

These constraints emphasise why the phylogenomic method pipeline, defined by Orr et al. (2020), is ill-suited for large-scale simulation replication. Modifications are required to improve the applicability and scalability of the phylogenomic method.

2.2.10.2 - Phylogenomic Method Modification

In response to the computational challenges of the original pipeline (Orr et al., 2020), I developed a streamlined 'modified phylogenomic method' optimised for high-throughput simulations. This method leverages the unique, total and shared mutation counts generated by Tomimoto and Satake's (2023) simulation framework. By using these 'true' simulated mutations- where the complete mutation history is already known, one avoids the need for the error correction steps that constitute a large portion of the original Orr et al. pipeline. This elimination of error correction is possible because the simulated sequences are free from sequencing artefacts or unknown evolutionary events, simplifying the estimation process. I refer to this as an 'modified' phylogenomic method because it is a simplified version of the original Orr et al. (2020) method, adapted to leverage the known mutation distributions of simulated datasets.

Pairwise genetic and physical distance matrices are reconstructed using these 'true' simulated mutations, enabling linear regression models that estimate somatic mutation rates with significantly reduced computational overhead. Once these matrices are generated, the upper triangle portion of each matrix (excluding the main diagonal) is extracted, representing unique branch pairs without redundancy. A linear regression is then applied to the pairwise values, where the physical distance matrix is the predictor and the genetic distance matrix is the response variable. This linear regression equation allows us to estimate the total number of variants within the tree, similar to the phylogenomic concept 'isolation by distance' common in population genetics, where physical separation correlates with genetic divergence. To enhance the regression model's accuracy, I experimented with multiple regression, incorporating the number of branching events between branches as an additional predictor variable. This multi-linear approach allowed us to explore whether the number of nodes separating branches contributes independently to observed genetic distance. The outcomes of both the linear and multiple regressions and their implications are detailed in section 2.3.4 (See Figure 2.16 - 2.18).

It is important to note that this approach represents an idealised, best-case scenario for testing the phylogenomic method. In real-world applications, sequencing errors and other sources of noise would necessitate additional error correction steps, as in the original Orr et al. pipeline. However, by working with ideal data, I will determine whether the phylogenomic method works under optimal conditions. If the method does not perform well with perfect data, it is unlikely to succeed with real, noisy datasets.

2.3 Results

2.3.1 Simulation Results

2.3.1.1 Mutation Rate Regression

I conducted 12,000 simulations across 20 tree topologies, varying in terminal branch number (4–12 branches), balance (balanced vs. unbalanced), and internal-to-terminal branch ratio (short vs. long terminal branches). Six distinct input mutation rates (4,6,8 × 10^{-10} and 4,6,8 × 10^{-9}) were tested under four models of somatic mutation accumulation (Tomimoto & Satake, 2023). Output mutation rates were estimated using the modified phylogenomic method and regressed against input rates across all tree topologies.

Regression results across topologies, shown in Figure 2.2 and summarized in Table 2.5, demonstrate strong correlations between input and estimated mutation rates at low mutation rates (<10^{-9}). Among the tested topologies, unbalanced long-terminal trees with eight or more branches consistently achieve the most accurate recovery of mutation rates, with regression coefficients closest to one (ranging from 0.73 to 2.0), high R^2 values (0.75–0.80), and low RMSE values (~1.76 × 10^{-10}). These configurations provide strong genetic signals due to the extended terminal branches, which accumulate sufficient mutations to maintain a clear relationship between physical and genetic distances. Additionally, the asymmetry of unbalanced topologies reduces shared mutations at internal nodes, further improving estimation accuracy.

In contrast, balanced long-terminal trees consistently underpredict mutation rates, with regression coefficients ranging from 0.29 to 0.51. For instance, the six-branch balanced long-terminal topology produces a regression coefficient of 0.29 with an R^2 of only 0.38, reflecting a weak correlation between physical and genetic distances. Unbalanced short-terminal trees exhibit a tendency to overpredict mutation rates, with regression coefficients frequently exceeding 1.5.

To assess the statistical significance of these regression coefficients, I applied a Bonferroni correction to account for multiple comparisons. Figure 2.3 presents the regression coefficients along with their associated raw p-values, with colors representing Bonferroni-corrected significance levels. Unbalanced long-terminal topologies show regression slopes closest to one with non-significant p-values, reinforcing their accuracy in recovering input mutation rates. Conversely, balanced long-terminal trees show highly significant underestimation, with regression coefficients dropping to 0.29–0.51 ($p < 0.001$), indicating systematic structural biases. Unbalanced short-terminal topologies, on the other hand, consistently overestimate mutation rates, with regression slopes significantly greater than one ($p < 0.05$), further supporting the conclusion that their phylogenomic signals are distorted.

At large mutation rates (≥10^{-9}), regression coefficients decline sharply across all topologies, in some cases approaching or even falling below zero. Some unbalanced long-terminal configurations reach regression coefficients as low as -0.032, suggesting that mutation rate recovery is no longer meaningful under these conditions. As regression slopes approach zero or negative values, it becomes evident that phylogenomic estimates of somatic mutation rates become unreliable at high mutation loads.

These results indicate that unbalanced long-terminal topologies provide the most accurate recovery of mutation rates at small mutation rates, outperforming all other tested configurations. Balanced long-terminal trees consistently underestimate mutation rates, while unbalanced short-terminal trees tend to overestimate them. At high mutation rates, the phylogenomic method fails across all topologies, demonstrating a fundamental limitation.

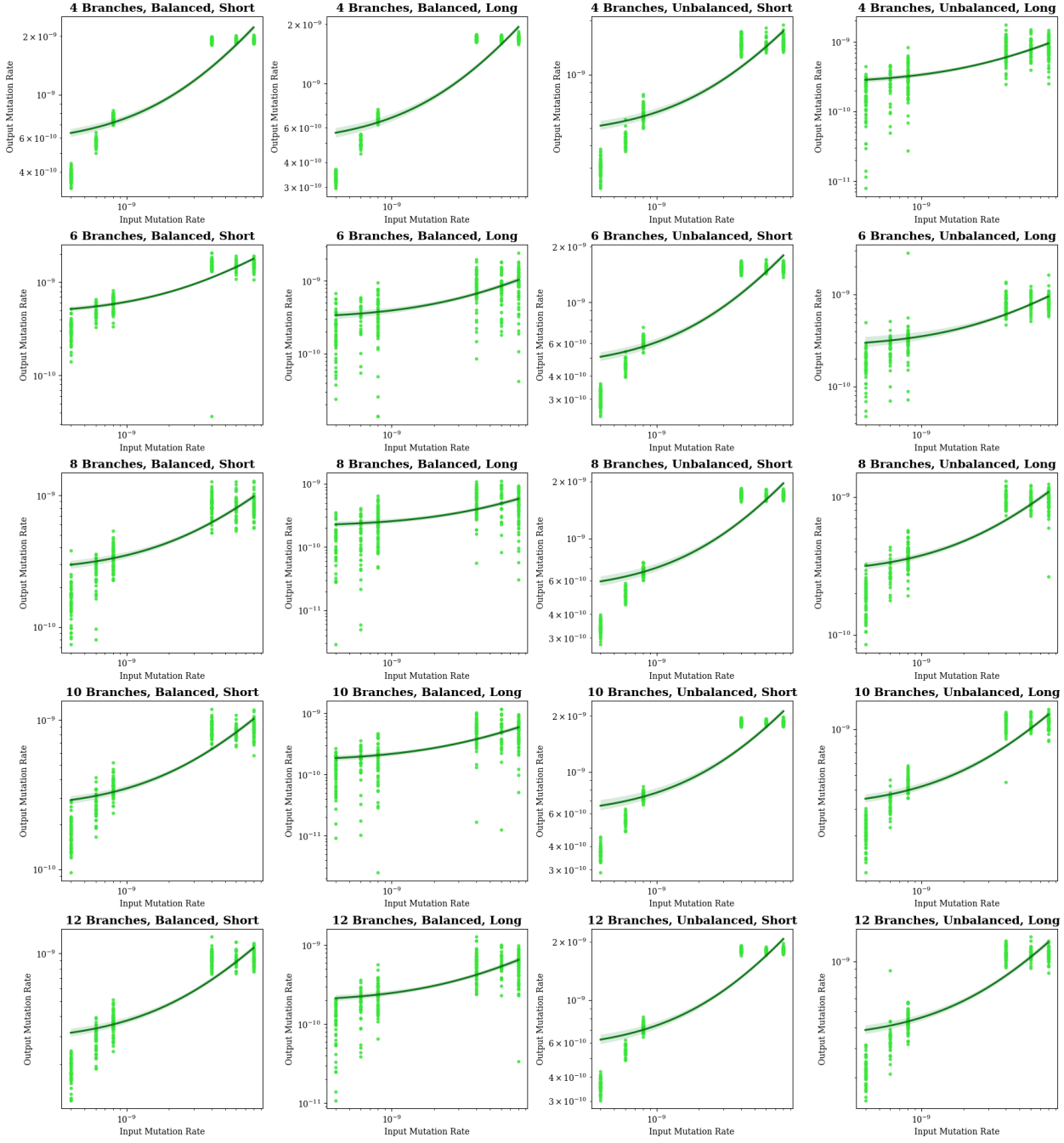


Figure 2.2 Regression of Estimated vs. Input Mutation Rates across 20 Tree Topologies. Six total input mutation rates ($4, 6, 8 \times 10^{-10}$, $4, 6, 8 \times 10^{-9}$) were inputted, as well as 20 tree topologies differing by the number of terminal branches (4-12), balance and internal:terminal branch ratio. Regression slopes evaluate how the phylogenomic method recovers true mutation rates under varying conditions. Each point represents an individual simulation trial. Overall, unbalanced topologies with long-terminal branches perform best (see Table 2.5).

Branch Number	Topology Balance	Internal: Terminal Branch Ratio	R^2 Value	RMSE Value
4	Balanced	Short Terminal	0.818218	2.867566e-10
4	Balanced	Long Terminal	0.823657	2.557958e-10
4	Unbalanced	Short Terminal	0.776401	2.574104e-10
4	Unbalanced	Long Terminal	0.539742	2.442527e-10
6	Balanced	Short Terminal	0.756492	2.772946e-10
6	Balanced	Long Terminal	0.378746	23.54286e-10
6	Unbalanced	Short Terminal	0.800129	2.481513e-10
6	Unbalanced	Long Terminal	0.552733	2.304865e-10
8	Balanced	Short Terminal	0.679049	1.818478e-10
8	Balanced	Long Terminal	0.290628	2.138007e-10
8	Unbalanced	Short Terminal	0.788812	2.772546e-10
8	Unbalanced	Long Terminal	0.747690	1.760113e-10
10	Balanced	Short Terminal	0.762316	1.542294e-10
10	Balanced	Long Terminal	0.414590	1.903570e-10
10	Unbalanced	Short Terminal	0.784816	3.007406e-10
10	Unbalanced	Long Terminal	0.796789	1.854420e-10
12	Balanced	Short Terminal	0.776570	1.581509e-10
12	Balanced	Long Terminal	0.456154	1.856046e-10
12	Unbalanced	Short Terminal	0.811098	2.785588e-10
12	Unbalanced	Long Terminal	0.765254	1.930626e-10

Table 2.5 Regression Metrics (R^2, RMSE) Corresponding to Figure 2.2
Metrics calculated for each of the 20 subplots in Figure 2.2, stratified by topology balance, internal:external branch ratio (short terminal vs. long terminal) and terminal branch numbers across all six input mutation rates ($4, 6, 8 \times 10^{-10}$, $4, 6, 8 \times 10^{-9}$).

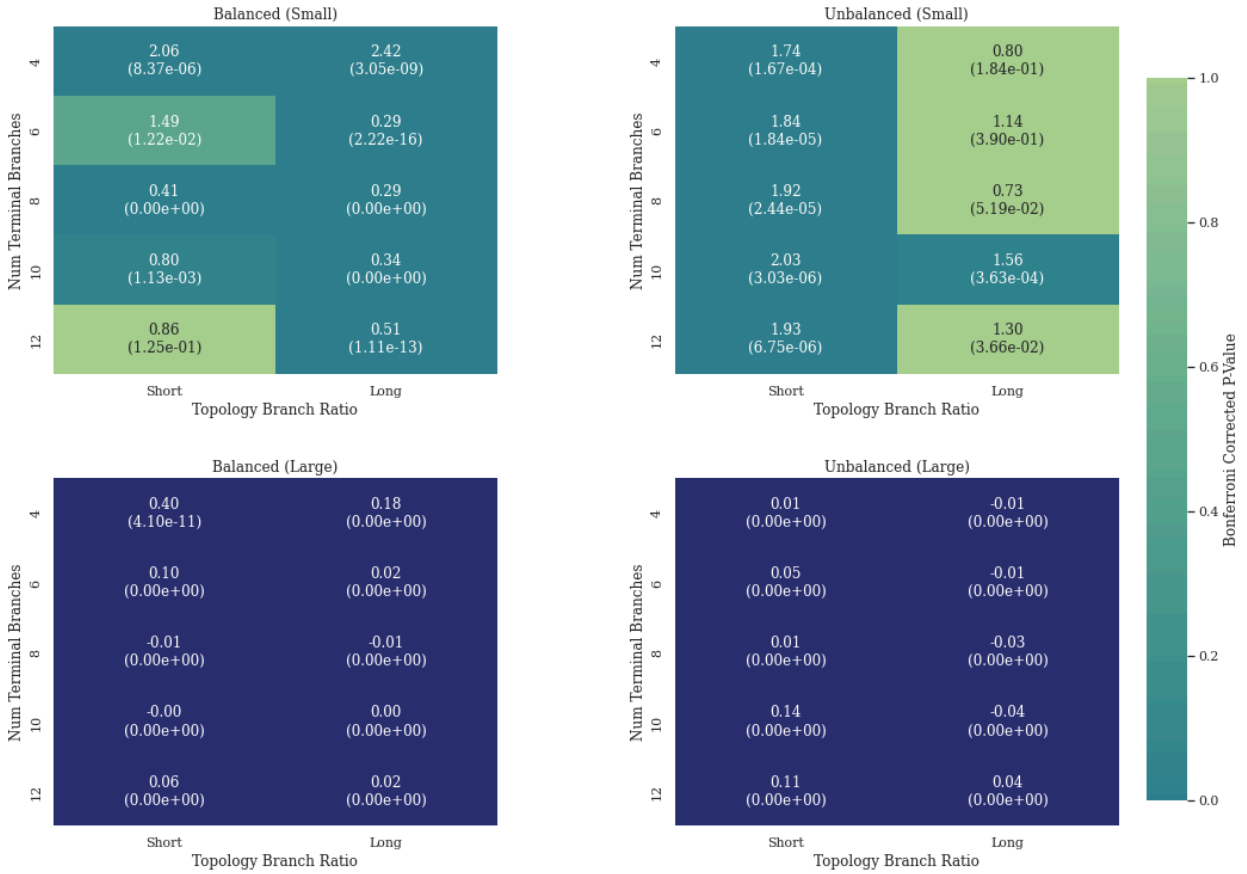


Figure 2.3 Heatmap of Regression β Coefficients and Bonferenni Corrected Significance for Small and Large Mutation Rates.
Regression analyses were performed between input and estimated mutation rates for small (4,6,8 × 10^{-10}) and large (4,6,8 × 10^{-9}) mutation rates across different tree topologies. Each cell shows the regression coefficient (top) and raw p-value (bottom, in parentheses), with colors indicating Bonferroni-corrected significance. Darker shades denote significant deviations from 1, while lighter shades indicate accurate mutation rate recovery. Results suggest that recovery is most reliable at small mutation rates- particularly for unbalanced topologies with long-terminal branches, while large mutation rates lead to systematic underestimation.

2.3.1.2 Mutation Rate GLM

To evaluate how both topology and mutation model parameters influence bias in mutation rate recovery, I employed a Generalised Linear Model (GLM) to the simulation data. Predictor variables include the number of terminal branches, topology balance (balanced vs. unbalanced), branch length ratio (short vs. long terminal branches), branching parameter (biased vs unbiased), elongation parameter (structured vs. stochastic) and input mutation rates. The two-way interactions of each combination of the stated variables were also included as predictors within the GLM. All predictors and interaction terms were Z-Score standardised, and partial eta-squared values were calculated alongside beta coefficients representing each predictor's significance, direction and relative contribution to the output mutation rate.

The results (Figure 2.4) reveal that topology variables and model parameters influence the mutation rate bias.

The elongation parameter (StD) is the most influential predictor, with the largest partial eta-squared value ($\eta^2 = 0.61$) and a negative beta coefficient ($\beta = -0.51$). This result suggests that stochastic elongation leads to lower estimated mutation rates, likely due to the loss of mutations through successive divisions in elongation where the stem cell lineage is not maintained. Alternately, structured elongation ($\beta = 0.51$) increases estimated mutation rates, potentially exacerbating overestimation. The branching parameter (biasVar) has a negligible effect ($\eta^2 = 0.013$), with a slight positive beta coefficient ($\beta = 0.0169$). This result suggests that the degree of bias in axillary meristem formation during branching minimally impacts the overall estimated mutation rate. The mode of elongation throughout growth remains the most influential model parameter, showing the greatest significance across Tomimoto and Satake's (2023) models of somatic mutation accumulation.

Across input tree topological parameters, the topology balance emerges as the most significant ($\eta^2 = 0.33$) with a positive beta coefficient ($\beta = 0.19$), suggesting a slight increase in estimated mutation rates for balanced topologies. However, given the small effect size and variability across topological configurations, this result should be interpreted with caution. Balanced short-terminal topologies slightly underestimate mutation rates, while balanced long-terminal topologies tend to underestimate even further, suggesting that factors like the proportion of unique to shared mutations may play a larger role in mutation rate recovery than balance alone. Terminal branch number shows similar significance ($\eta^2 = 0.28$), with a negative beta coefficient ($\beta = -0.15$), indicating that the addition of terminal branches results in lower estimated mutation rates. The proportionality of physical and genetic distance is disrupted with increasing branch number, as the number of shared mutations dominates unique mutations across branches in a topology. Lastly, the internal:terminal branch ratio shows moderate significance (η^2=0.19) with a small negative beta coefficient (β = -0.18), suggesting that topologies with longer terminal branches tend to recover slightly lower mutation rates. However, given the small effect size and variability across other topological effects, this result should be interpreted with caution.

Of the two-way interactions, interactions involving the elongation parameters (StD) are consistently highly significant, demonstrating its substantial impact on mutation rate recovery in combination with other predictors:

- Input Mutation Rate * StD ($P < 0.001, \beta = 0.26$): Amplifies estimated mutation rates as input mutation rates increase, suggesting structured elongation magnifies overestimation at higher mutation rates.
- Topology Balance * StD ($P < 0.001, \beta = -0.13$): Structured elongation reduces mutation bias in balanced topologies, slightly counteracting or correcting for overestimation.
- Topology Branch Ratio * StD ($P < 0.001, \beta = 0.084$): Structured elongation may (weakly) mitigate underestimation in long-terminal configurations by stabilizing mutation recovery.
- Terminal Branch Number * StD ($P < 0.001, \beta = 0.11$): Structured elongation counteracts the dilution effect of increasing branch numbers, maintaining elevated estimated mutation rates.

Interactions with the branching parameter, biasVar, consistently fail to reach significance, including:

- Input Mutation Rate * biasVar ($P = 0.294$)
- Terminal Branch Number * biasVar ($P = 0.399$)
- Topology Branch Ratio * biasVar ($P = 0.05$)

Confirming that branching behaviour does not meaningfully influence mutation rate recovery.

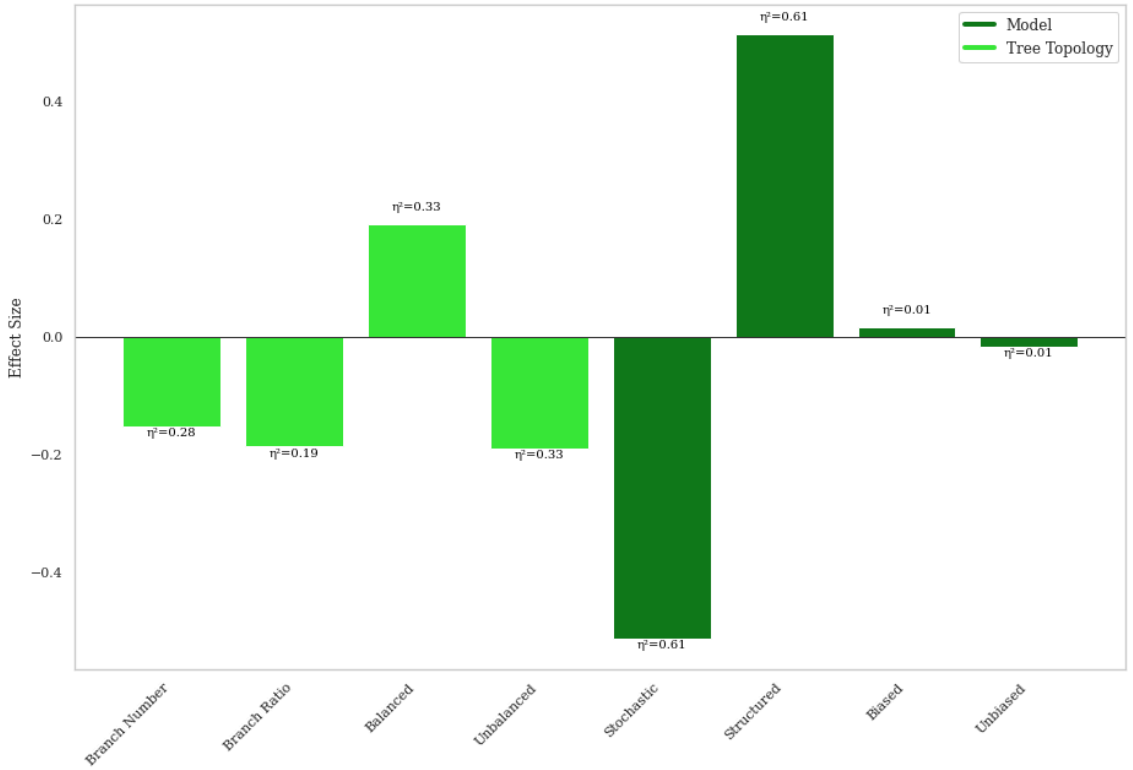


Figure 2.4 Standardised Beta Coefficients (Effect Size = β) and Partial Eta-Squared (η^2) Values for Predictors in the Mutation Rate GLM. Predictors include input tree topology variables (light green: terminal branch number, internal:terminal branch ratio, topology balance) and somatic mutation accumulation model parameters (dark green: elongation and branching as defined by Tomimoto and Satake 2023). Predictors were Z-score standardised, meaning the direction and magnitude of the beta coefficient indicate effect size on estimated mutation rates. Partial-eta squared quantifies the contribution of each predictor to mutation rate bias (overestimation or underestimation, depending on the direction of β).

The Predicted vs. Observed plot for the mutation rate GLM demonstrates a moderate-to-strong correlation between predicted and observed mutation rates (Figure 2.5). While some deviation is evident at higher values, the clustering of points along the diagonal suggests that the model captures key trends in mutation rate variation. The spread at lower mutation rates indicates some underestimation in specific cases, but overall, the model maintains relative accuracy across different configurations.

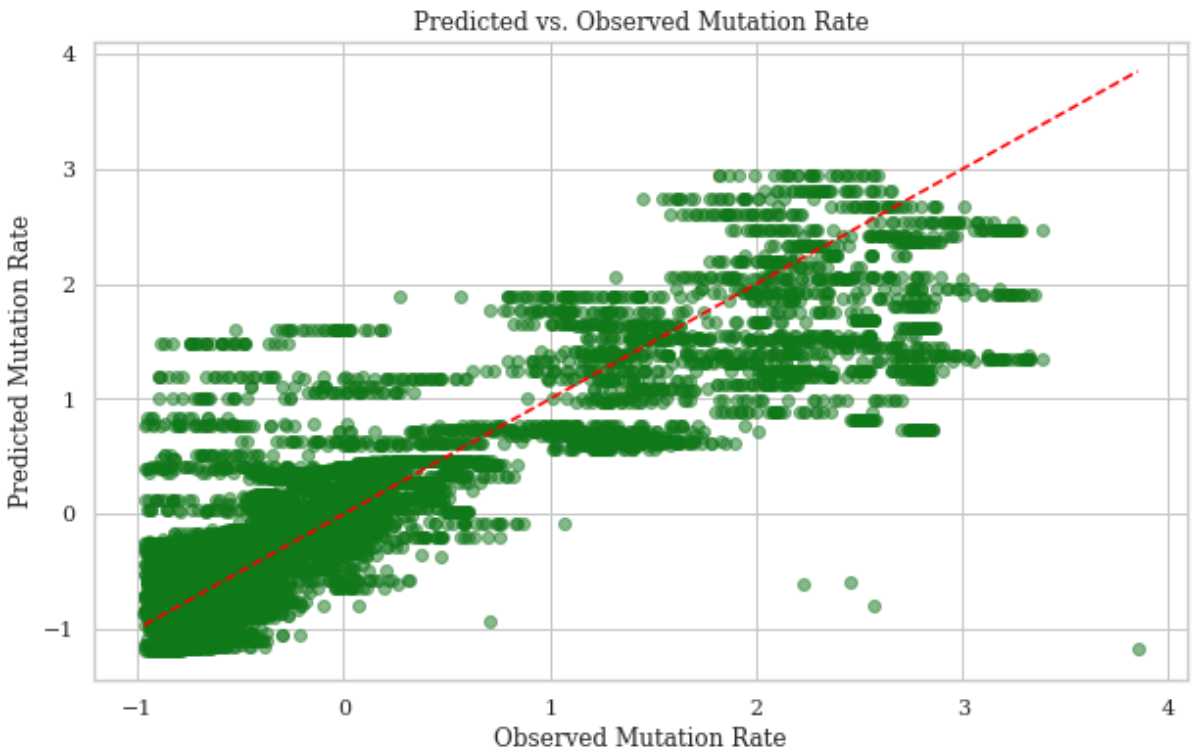


Figure 2.5 The Observed vs. Predicted Estimated Mutation Rates for the Mutation Rate GLM. Clustering of points along the diagonal indicates that the model captures key trends in mutation rate variation, though deviations at higher values and a spread at lower mutation rates suggest some underestimation in specific cases.

2.3.1.3 Topology Recovery

The recovery of tree topologies from simulated somatic mutations was evaluated using the Robison-Foulds (RF) distance metric (Methods, Section 2.2.4). RF distances measure the dissimilarity between reconstructed and input tree topologies, where a value of 0 indicates that the topology of the reconstructed tree is identical to that of the input tree. Topology recovery is a critical test of the phylogenomic method, which relies on the assumption of somatic mutations closely following a tree's branching structure or ontogeny. This assumption underpins the method's ability to use the tree topology as a framework for filtering mutations across branches and ultimately estimating the somatic mutation rate.

Figure 2.6 shows that **standardised** RF distances (ranging from 0 to 1 across all 20 input topologies) increase with terminal branch number, reflecting the increased difficulty in recovering more complex tree structures. Balanced topologies consistently yield lower RF distances than unbalanced topologies with the same number of terminal branches. For instance, the 8-branch balanced topology achieves an average standardised RF distance of 0.5, compared to 0.6 for its unbalanced counterpart. This trend likely reflects the proportional distribution of mutations in balanced topologies, which enhance recovery accuracy compared to uneven mutation signals in unbalanced structures.

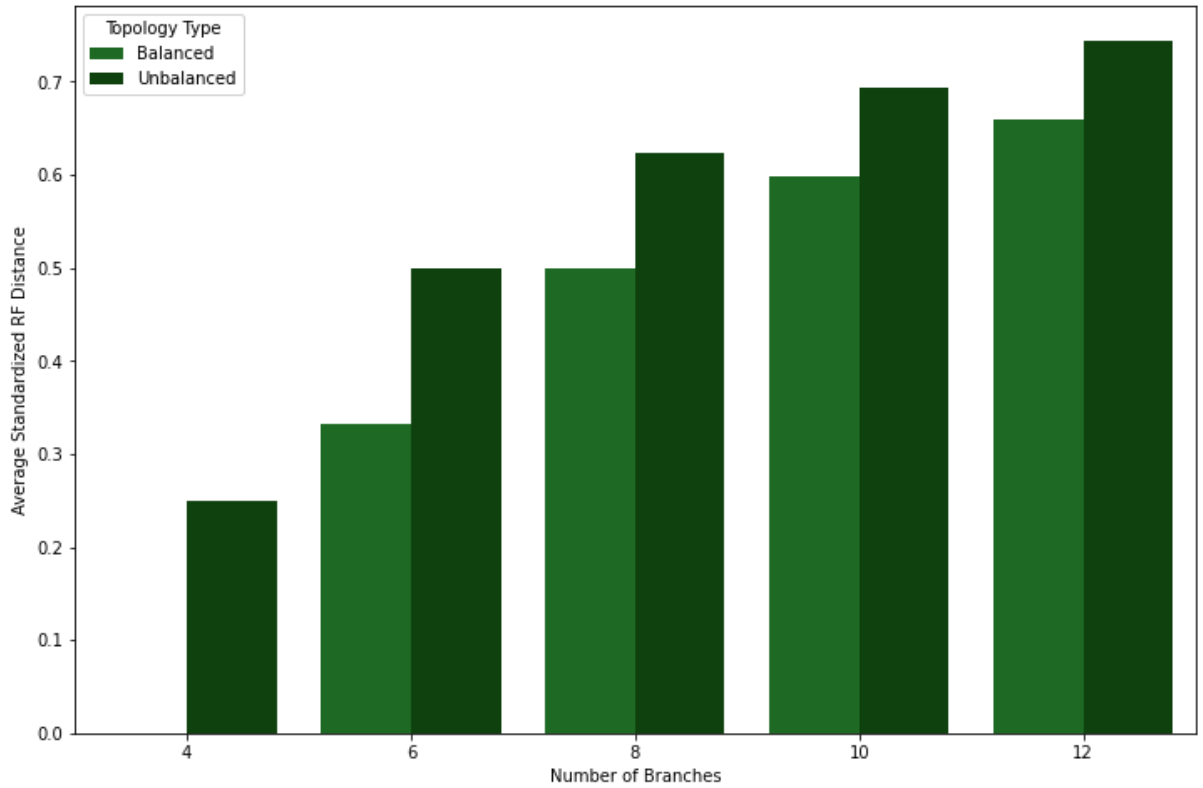


Figure 2.6 Standardised RF Distances for Balanced and Unbalanced Topologies with 4-12 Terminal Branches. Phylogenetic trees were reconstructed across 20 input topologies using simulated mutations via Tomimoto and Satake (2023) models. RF distance counts unique bipartitions present in one tree but absent in the other, measuring phylogenetic dissimilarity. RF distance standardised by dividing raw RF distance by the maximum possible for a given topology, scaling results between 0 and 1. Results indicate that topology recovery is more efficient in balanced topologies across all branch numbers.

To further investigate the effects of tree size and balance on topology recovery, I performed a subsampling test using the real topology of the *Eucalyptus melliodora* individual (Figure 2.7). This test compared a reconstructed four-branch balanced subsample to the original eight-branch unbalanced topology as inputs of the the simulation framework, using Model III (Stochastic Elongation, Biassed Branching) of Tomimoto and Satake's 2023 models for somatic mutation accumulation (refer to section 2.3.3). The recovery accuracy for these trees was assessed using RF distance and mutation rate deviation (estimated mutation rate - input mutation rate).

The four-branch balanced topology exhibited perfect recovery with an RF distance of 0 (Figure 2.14, right). All bipartitions in the reconstructed tree matched the input tree with no structural differences.This topology also had a much lower mutation rate deviation (3.28×10^{-9}), reflecting that a simpler input topology is more easily estimated and may better predict somatic mutation rates under the phylogenomic method.

In contrast, the original eight-branch unbalanced topology achieved a standardised RF distance of approximately 0.25 (non-standardised RF = 6). This value reflects moderate differences between the true tree topology and the reconstructed trees, which may have contributed to the larger mutation rate deviation (4.29×10^{-8}). Specifically, the reconstructed tree showed branch swaps involving

branches 7, 8, and 6, in which branch 8 is incorrectly placed as a sister to branch 6. Interestingly, Orr et al. 2020 also encountered this exact branch misplacement (where Orr et al. 's M8 = b6 and M5= b8 in my topology interpretation) when reconstructing a maximum-likelihood tree from the genome sequence data across branches. A novel grouping of branches 5 and 1 was also introduced but was not present in the real input topology. These discrepancies likely arise from the topology's increased complexity and unbalanced nature, which diminishes mutation signal distribution and recovery.

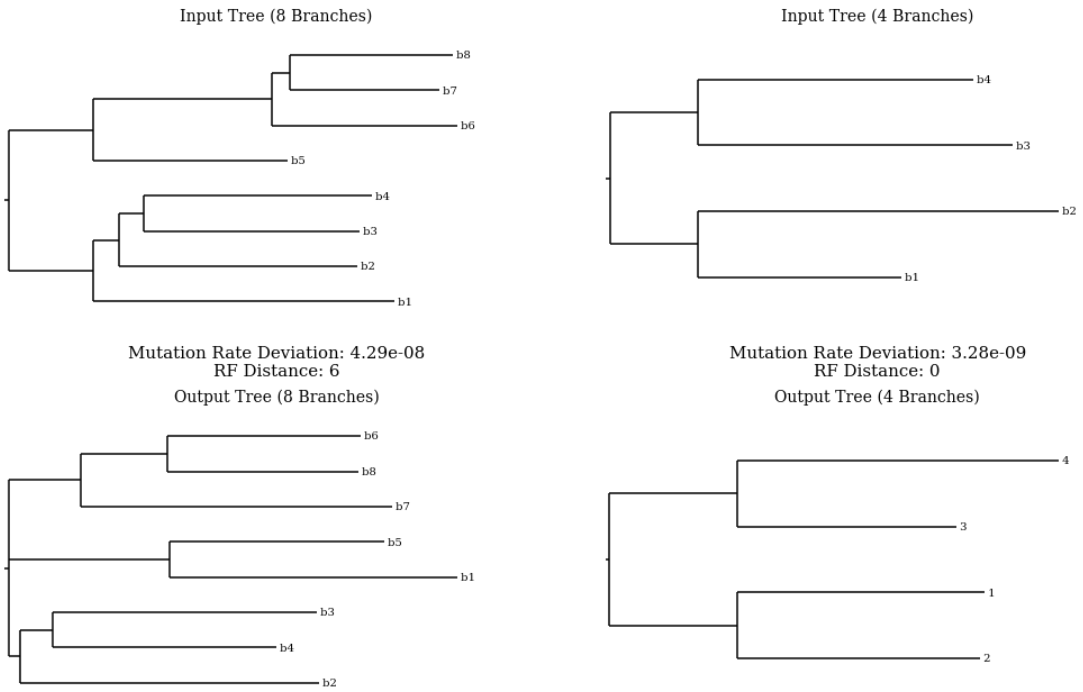


Figure 2.7 Input and Reconstructed Phylogenies for Original 8-Branch *E. melliodora* topology and a Four-Branch Subsampled Topology. The balanced four-branch tree had perfect recovery (RF distance = 0), while the unbalanced eight-branch tree showed reconstruction errors (RF distance = 6) and higher mutation rate deviation.

2.3.1.4 Topology Recovery GLM

To evaluate how topological and model-specific parameters influence topology recovery, I applied a Generalised Linear Model (GLM) to analyse topology recovery, measured in standardised Robison-Foulds (RF) distances. Predictor variables included topological parameters (e.g., topology balance, internal:terminal branch ratio, terminal branch number) and model parameters (e.g., elongation behaviour, branching bias), along with their two-way interactions. Variables were Z-score standardised to ensure comparability, and both beta coefficients (β) and partial eta-squared (η^2) values were assessed to evaluate the contribution of each predictor.

The GLM results (Figure 2.8) reveal that the number of terminal branches remains the strongest predictor of topology recovery, with the largest partial eta-squared ($\eta^2 = 0.982$) and a substantial positive beta coefficient ($\beta = 0.983$). Thus, despite standardising RF distance and Z-scoring the predictors, topological complexity introduced by increasing terminal branch number strongly dictates topology recovery. The addition of terminal branches increases the total number of bipartitions, making topology recovery inherently more complex, and further amplifies the structural noise and mutation signal dilution, contributing to greater RF distances even after accounting for these

adjustments. Topology balance is the second most important predictor ($\eta^2 = 0.63$), with a moderate positive beta coefficient ($\beta = 0.34$) for unbalanced topologies. Thus, I can confirm that unbalanced trees exacerbate recovery challenges, with uneven mutation distributions across branches causing somatic mutation accumulation patterns to differentiate from the tree topology. The branch length ratio, while less influential ($\eta^2 = 0.045$), has a negative beta coefficient ($\beta = -0.022$), indicating that longer branches slightly improve recovery by reducing mutation overlap at shared internal nodes.

Model-specific variables, including elongation (StD) and branching behaviour (biasVar), play minimal roles in topology recovery. Elongation has a minor effect ($\eta^2 = 0.001$), with an insignificant beta coefficient ($\beta = 0.0005$), suggesting structured (or conversely stochastic) elongation neither significantly aids nor hinders topology recovery. Similarly, branching bias (biasVar) shows negligible impact ($\eta^2 = 0.001$) and an insignificant beta coefficient ($\beta = 8.5 \times 10^{-5}$).

Few interactions in the topology recovery GLM are both statistically and practically significant. For example, the interaction between balance and branch length ratio ($P = 0.032$, $\beta = -0.0048$) is statistically significant. However it has negligible effect size, suggesting that unbalanced trees with longer terminal branches offer only marginal improvements in recovery. Similarly, the interaction between branch length ratio and terminal branch number ($P < 0.001$, $\beta = -0.0138$) is 'statistically significant'. However, it has a minimal effect size, reflecting some increased difficulty resolving highly branched trees with longer terminal branches. While some interactions are detectable, their contributions to RF distances are minimal compared to the main effects of tree complexity and balance.

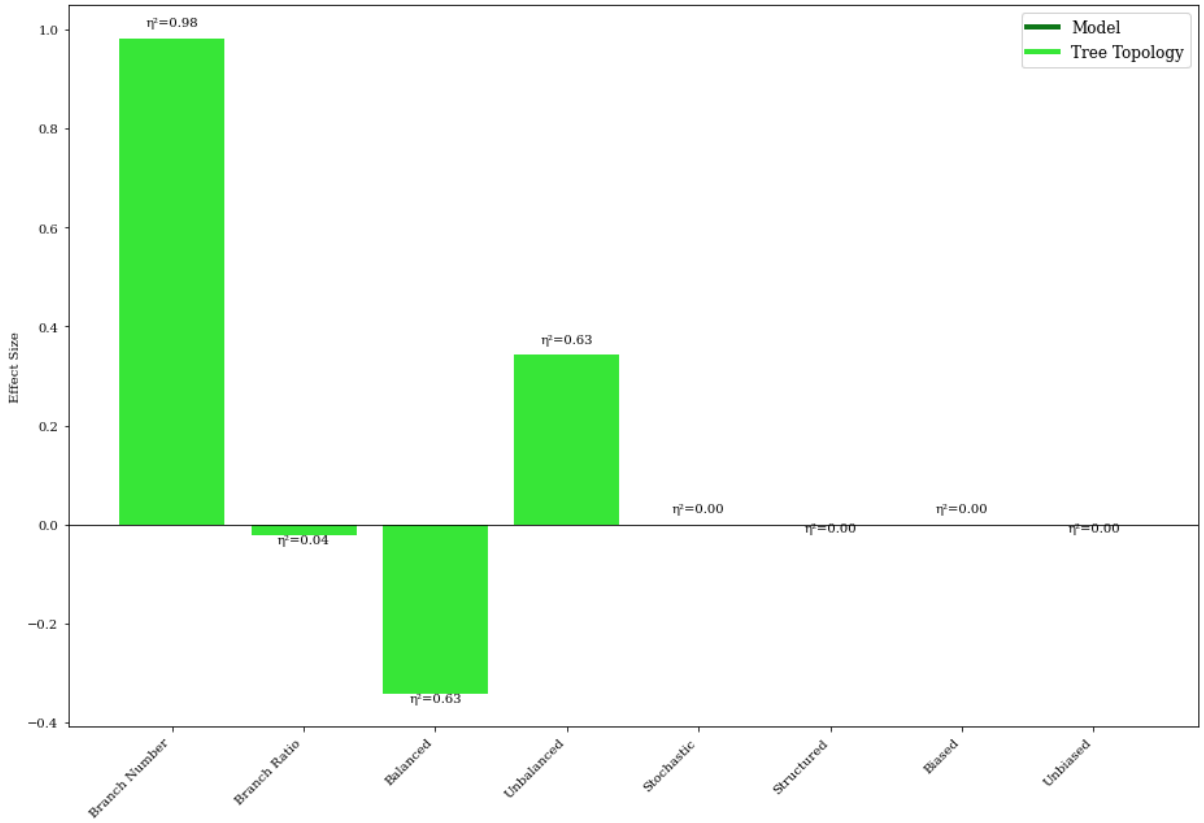


Figure 2.8 Standardised Beta Coefficients (Effect Size = β) and Partial Eta-Squared (η^2) Values for Predictors in the Topology Recovery GLM. Topology recovery was measured using

standardised RF distances between input and topologies reconstructed using simulated mutations. Predictors include input tree topology variables (light green: terminal branch number, internal:terminal branch ratio, topology balance) and somatic mutation accumulation model parameters (dark green: elongation and branching as defined by Tomimoto and Satake 2023). Predictors were Z-score standardised, meaning the direction and magnitude of the beta coefficient indicate effect size on estimated mutation rates. Partial-eta squared quantifies the contribution of each predictor to mutation rate bias (overestimation or underestimation, depending on direction of β**),** with branch number being the most significant predictor ($\eta^2 = 0.982$) overall. Model-specific parameters show little to no significance.

The observed vs. predicted plot supports the model's validity, with most points closely aligned along the diagonal, confirming strong agreement between observed and predicted standardised RF distances (Figure 2.9). Deviations are more noticeable for extreme RF distances, aligning with earlier findings that tree complexity and asymmetry introduce systematic errors in topology recovery.

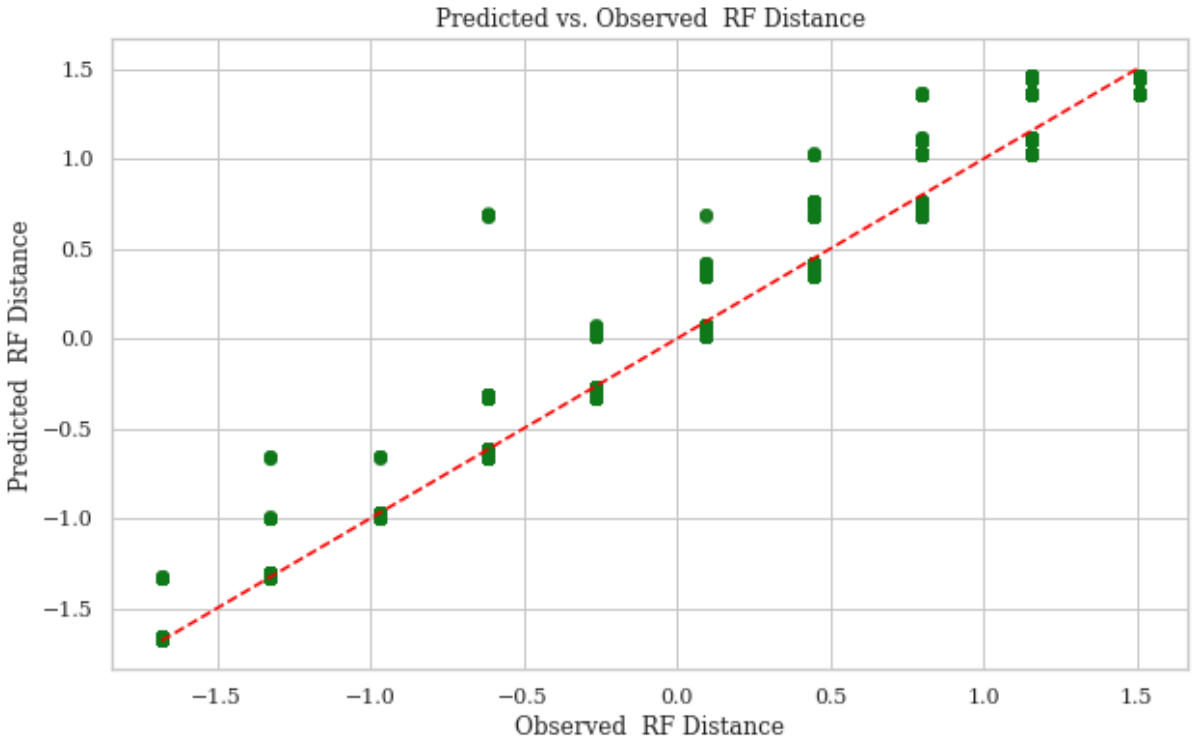


Figure 2.9 The Observed vs. Predicted Robinson-Foulds Distances for the Topology Recovery GLM. Most points align along the diagonal, indicating strong agreement, while deviations at extreme RF distances reflect systematic errors from tree complexity and asymmetry.

2.3.2 Analysis of Simulation and Phylogenomic Method Constraints

2.3.2.1 Back Mutation Analysis

Poisson distributions demonstrated that back mutations remain exceedingly rare across all mutation rates throughout simulations (Figure 2.10). The expected number of back mutations per site (λ) remains close to zero, regardless of mutation rate, models of elongation and branching or tree

branching structure. This analysis confirms that back mutations do not significantly contribute to the observed discrepancies in mutation rate recovery, even across larger mutation rates.

The expected number of back mutations per branch was also analysed across mutation rates (Figure 2.11). Although higher mutation rates slightly increased the total expected back mutations per branch, the values were still consistently negligible across all scenarios.

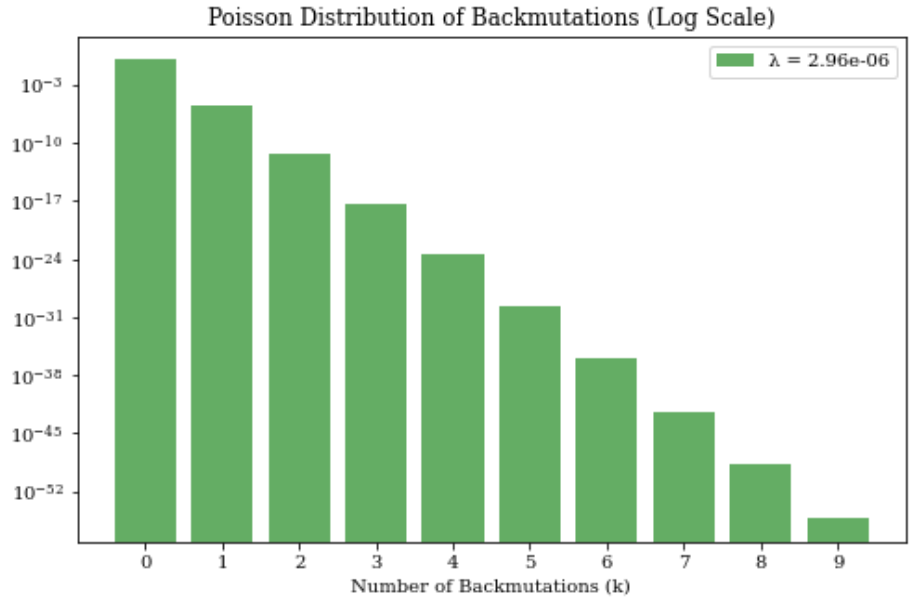


Figure 2.10 Poisson Distribution of Back Mutations.

Probabilities of observing k back mutations per site modelled across mutation rates (4×10^{-10} to 8×10^{-9}) are plotted on a logarithmic scale. The expected number of back mutations per site (λ) remains close to zero, indicating that back mutations are negligible.

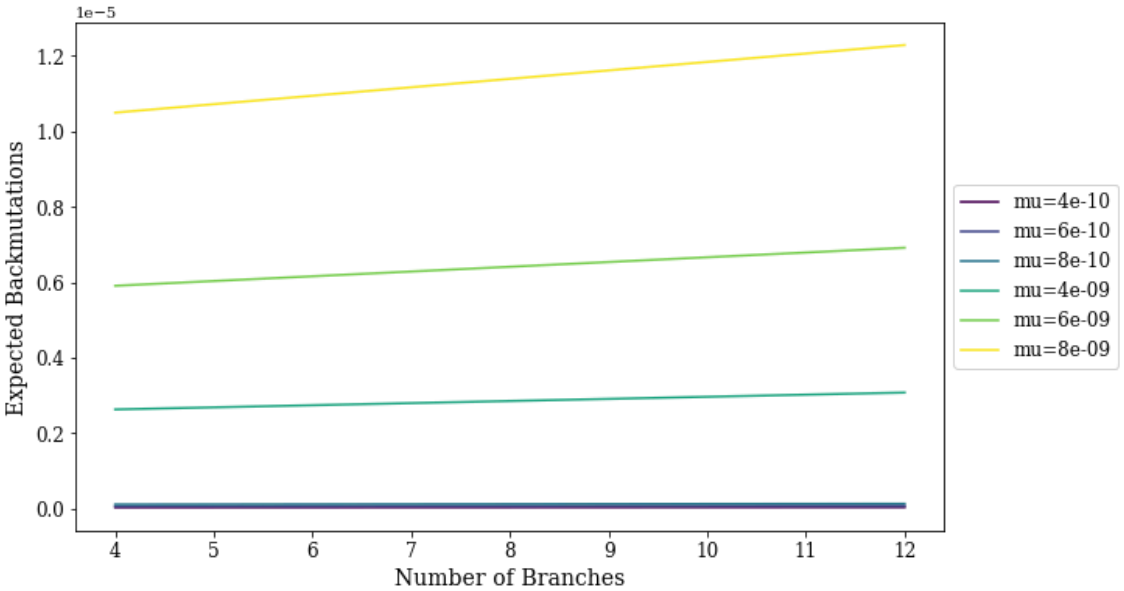


Figure 2.11 Expected Back Mutations Per Branch.

Above is expected number of back mutations per branch according to the Poisson distribution $P(k)$ across mutation rates, which remain negligible throughout simulations.

2.3.2.2 Shared Mutation Homogeneity Analysis

To evaluate how mutation rates affect the distribution of shared mutations across branches, I examined the coefficient of variation (CV) of shared mutations under different mutation rate conditions across simulations. Shared mutations in this context refer to those inherited from common ancestral cells, rather than independent parallel mutations. Since the simulation model developed by Tomimoto and Satake (2023) applies a constant per-division mutation rate, I expect both unique and shared mutations to increase proportionally with mutation rate.

The analysis revealed a substantial increase in the mean number of shared mutations, from approximately 4,504 at low mutation rates to 15,310 at high mutation rates (Figure 2.12). Despite this increase, the CV remained nearly constant (1.70 for small rates and 1.73 for large rates). This stability suggests that the relative dispersion of mutations across branches did not expand proportionally with increasing mutation rates. Instead, the differences in shared mutation counts between branch pairs became less pronounced, indicating greater homogeneity in shared mutations across branches.

This pattern suggests that at high mutation rates, shared mutations become more uniformly distributed across branches, potentially diluting the phylogenetic signal that distinguishes individual branches. As a result, the phylogenomic method struggles to resolve tree topology accurately under high mutation rate conditions. Given that the simulation applies a constant mutation rate, this trend may reflect an increased proportion of shared-to-unique mutations rather than an unexpected breakdown of the model.

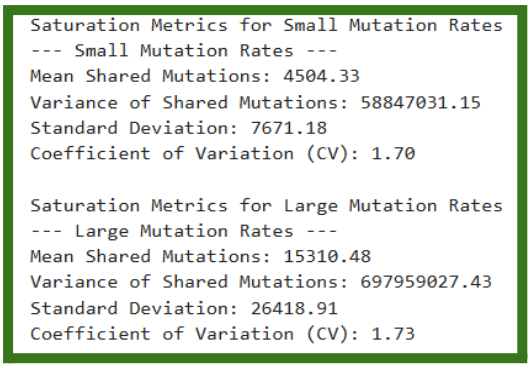


Figure 2.12 Shared Mutation Homogeneity Metrics for Small and Large Mutation Rates.
Mean shared mutations, variance, and coefficient of variation (CV) are compared.
The constant CV value indicates greater homogeneity in shared mutations across branches at larger mutation rates, disrupting the phylogenomic method.

2.3.3 Somatic Mutation Distributions across *E. melliodora* Branches

Tomimoto and Satake (2023) applied their hierarchical modular architecture models to *Populus trichocarpa*, a long-lived poplar tree (330 years old), to simulate the somatic mutation accumulation and expansion *in silico*. Their study revealed that while the phylogenomic method effectively filters mutations which follow the tree topology, it may over-filter mutations, excluding valid mutational patterns that do not strictly align with the topology. In this section, I replicate their approach by applying the same framework to the *Eucalyptus melliodora* individual investigated by Orr et al. 2020. I investigate the 'real' mutation accumulation patterns of *E. melliodora* both before and after filtering with dng according to topology, i.e. before and after applying the phylogenomic method.

Using Tomimoto and Satake's (2023) simulation code, I evaluated the distribution of singleton mutations under four models of somatic mutation accumulation: structured-biassed, structured-unbiased, stochastic-biassed, and stochastic-unbiased. Singleton mutations were the primary focus, as the phylogenomic filtering method in this dataset does not retain shared mutation patterns, unlike the comprehensive mutation datasets available for *P. trichocarpa*. Figures 2.13 and 2.14 illustrate the mutation patterns of pre- and post-phylogenetic filtering, respectively. Figure 2.15 shows a phylogenetic interpretation of the *E. melliodora* topology, showing branch lengths in years (converted using the growth rate of 10 cm per year) to contextualise the distribution of mutations.

The pre-filtered dataset in Figure 2.13 shows a relatively equal distribution of mutations across branches, with no pronounced deviations across branches. Each branch in the *E. melliodora* topology has accumulated approximately 30-40 mutations prior to dng filtering, closely matching the predictions of the stochastic elongation models (40-50 mutations per branch). Structured elongation models overestimate mutation counts, predicting approximately 150-200 mutations per branch, as they assume strict maintenance of cell lineages during elongation and thus experience no mutation loss. This difference is reflected in the NMRSE values: for the pre-filtered dataset, the structured-biassed and structured-unbiased models have an NMRSE of 3.4 and 3.1, respectively, both of which are higher than the stochastic models, which achieve lower NMRSE values, of 0.2.

Post-filtering results (Figure 2.14) demonstrate a sharp reduction in singleton mutations in branches 6 and 5*, while branches 7 and 8 retain higher mutation counts despite their comparable branch lengths. This discrepancy is likely due to the shared internal node connecting branches 5 and 6, which is 77 years old (approximately 770 cm)- almost as old (and long) as the branches themselves, which are 84 and 80 years old, respectively. Mutations that arose in this internal node would be shared by both branches and thus removed during filtering, reducing singleton mutation counts in branches 5 and 6. In contrast, branches 7 and 8 share a much smaller internal node (8 years), resulting in fewer shared mutations being filtered. As Tomimoto and Satake (2023) noted, the phylogenomic method's preferential filtering of mutations that do not strictly follow the tree topology may lead to an underestimation of somatic mutation accumulation, particularly in cases involving mutations arising in branches which share a sizeable internal node. With a constant somatic mutation rate, one should expect a similar unique or singleton mutation in branches 5 and 6 as in branches 7 and 8, regardless of the size of the internal nodes which connect them.

Among the models, the stochastic-unbiassed model best captures the post-filtering mutation distribution of *E. melliodora*, achieving the lowest NMRSE value (NMRSE = 0.8). This model reflects the branch-specific variation introduced by biassed sampling during branching while still accounting for random mutation losses. The structured-biassed model, despite its higher NMRSE value (NMRSE = 3.4), shows greater branch-to-branch variation due to its retention of all mutations, with strict lineage maintenance and back mutations disregarded (across all models, due to the assumed low mutation rate in the model's formulation). While Tomimoto and Satake (2023) recommended the structured-biassed model for the long-lived angiosperm *Populus trichocarpa*, the data for *E. melliodora* seem to favour stochastic models.The preference for stochastic models may be tied to the high-confidence variant counts used in this study, originating from Orr et al. 2020. The variant counts are relatively low before filtering, suggesting a low mutation rate, as previously described by Orr et al. (2020). Alternatively, it may be that Orr et al. overfiltered even before applying the phylogenomic method, compounding the underestimation of somatic mutation counts. Nonetheless, both pre- and post-dng filtered data appear to indicate that a stochastic-biased model of somatic mutation comes close to describing the accumulation of genetic diversity in *E. melliodora*.

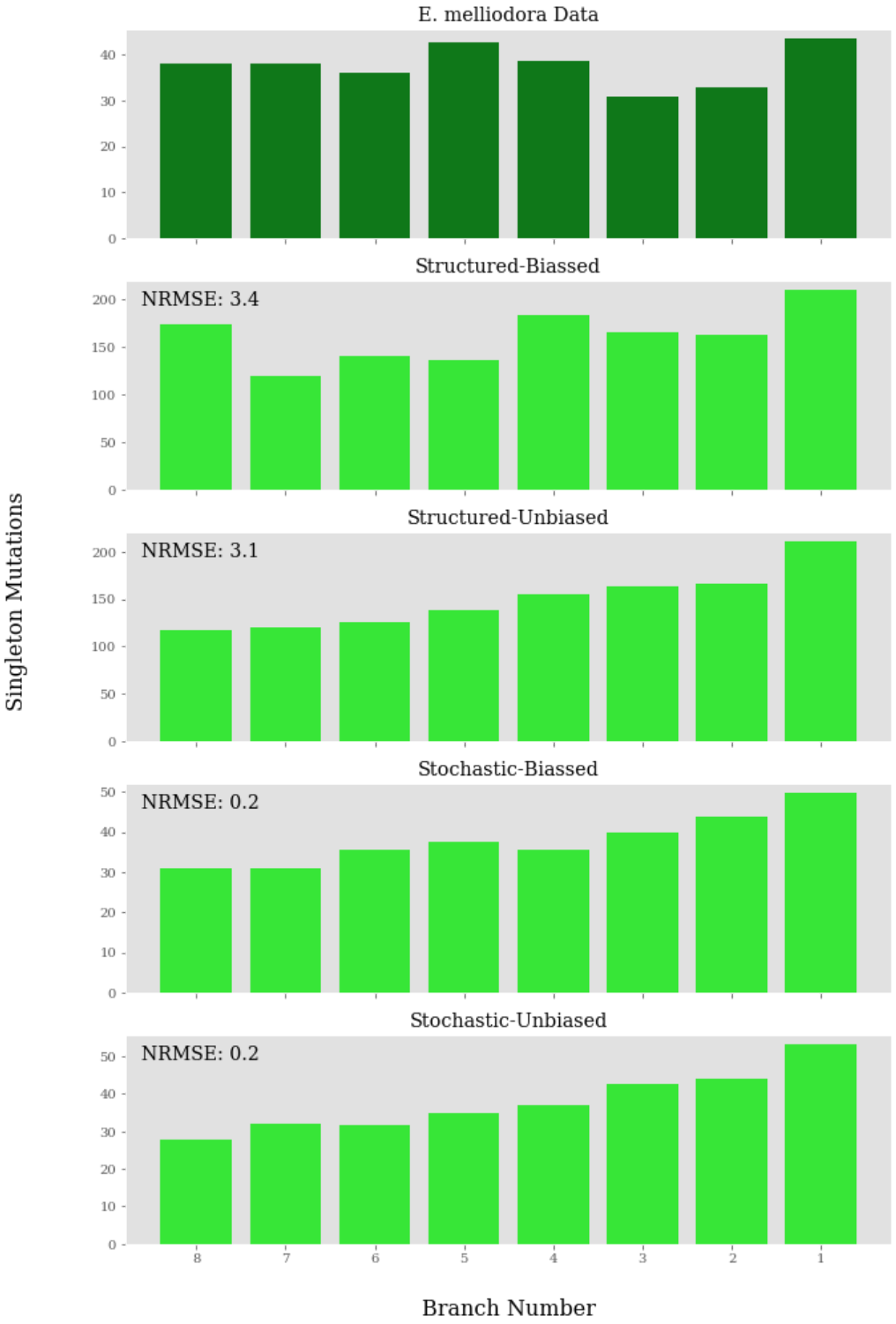


Figure 2.13 Distribution of Singleton Mutations in *E. melliodora* Pre-Phylogenetic Method. The top panel shows the mutation counts from the whole genome reads of *E. melliodora* prior to filtering with dng according to topology (Orr et al. 2020). The Normalised Root Mean Square Error (NMRSE) quantifies the dissimilarity between observed and simulated data, with lower values indicating a better alignment of models to the 'true' data. *Branches numbered 8-1, left to right. Stochastic models of somatic mutation accumulation best align with pre-phylogenetic *E. melliodora* data.

Figure 2.14 Distribution of Singleton Mutations in *E. melliodora* Post-Phylogenetic Method. The top panel shows the mutation counts from the whole genome reads of *E. melliodora* after filtering with dng according to topology (Orr et al., 2020). The Normalised Root Mean Square Error (NMRSE) quantifies the dissimilarity between observed and simulated data, with lower values indicating a better alignment of models to the 'true' data. *Branches numbered 8-1, left to right. Stochastic models of somatic mutation accumulation best align with post-phylogenetic *E. melliodora* data.

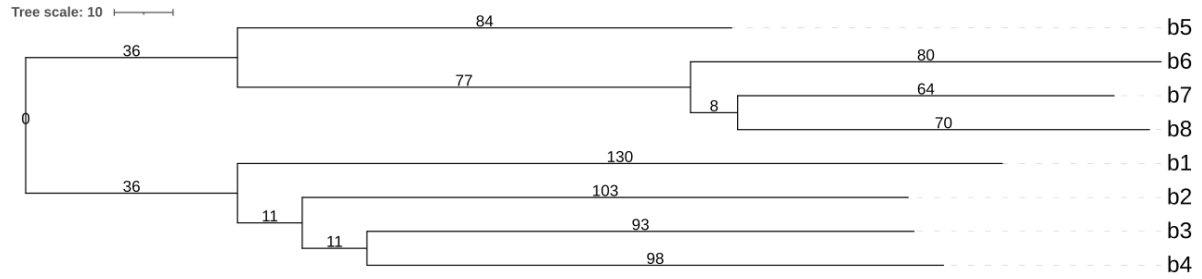


Figure 2.15 Phylogenetic Representation of *E. melliodora* Tree Topology.
Branches are labelled from b8 to b1, representing branches 8 to 1. Branch lengths denoted for the 8 terminal branches and 6 internal branches represent the age of branches in years, estimated using the growth rate of 10 cm per year. The branch lengths (in cm) were measured directly by Orr et al. (2020). For simplicity, the tree has a root age of 1. The tree scale (10 years/100cm) is shown above for reference.

2.3.4 Phylogenomic Method Application

The modified phylogenomic method estimates the somatic mutation rate of a given tree by applying regression models to genetic and physical distance matrices. Designed to address the computational challenges of Orr et al.'s (2020) pipeline, this method acts as a scalable alternative suitable for large-scale simulations. I tested the validity of the modified phylogenomic method on both the 'real' high-confidence variants of *E. melliodora* (the variants remaining after filtering with topology as reported by Orr et al. 2020) and the simulated variants generated using Tomimoto and Satake's (2023) models.

To potentially improve the modified phylogenomic method, I tested the inclusion of branching events- the number of nodes separating branch pairs- as an additional predictor variable alongside physical distance. I hypothesised that branching events might independently contribute to genetic divergence by capturing the influence of shared internal nodes. The resulting multi-linear regression model (Figure 2.16), incorporating both physical distance and branching events as predictors, yielded the equation:

$$y = 2.26 * Physical\,Distance + 0.54 * Branching\,Events - 29.60$$

Where each variable was Z-transformed, allowing the coefficients to represent the change in genetic distance (in standard deviations) per one standard deviation increase in the predictor variable.

This model achieved a high R^2 (0.743), indicating that 74.3% of the variance in genetic distance was explained. Physical distance had a much stronger effect ($\beta = 2.26,\ p < 0.001$) compared to branching events ($\beta = 0.54,\ p = 0.695$), which was not statistically significant. Despite the overall significance of the model ($F = 36.15,\ p < 0.0001$), the inclusion of branching events introduced increased the variance inflation factor (VIF), suggesting multicollinearity issues.

To confirm multicollinearity between physical distance and branching events, I tested physical distance as a predictor of branching events, giving the following regression equation (Figure 2.17):

$$y = 0.19 * Physical\,Distance - 1.24 \quad (R^2 = 0.4617)$$

The moderate R^2 indicates that nearly half of the variance in branching can be explained by physical distance alone, demonstrating correlation. The standardised coefficient for physical distance (=0.19)

further suggests a significant but weak overlap between the two predictors. Given the statistical insignificance of branching events in the multi-linear model-

- ($p = 0.695$, $R^2_{partial} = 0.0016$), the results suggest that branching events provide redundant information already captured by physical distance. I thus excluded branching events in favour of a simpler and more robust model for use with the modified phylogenomic method.

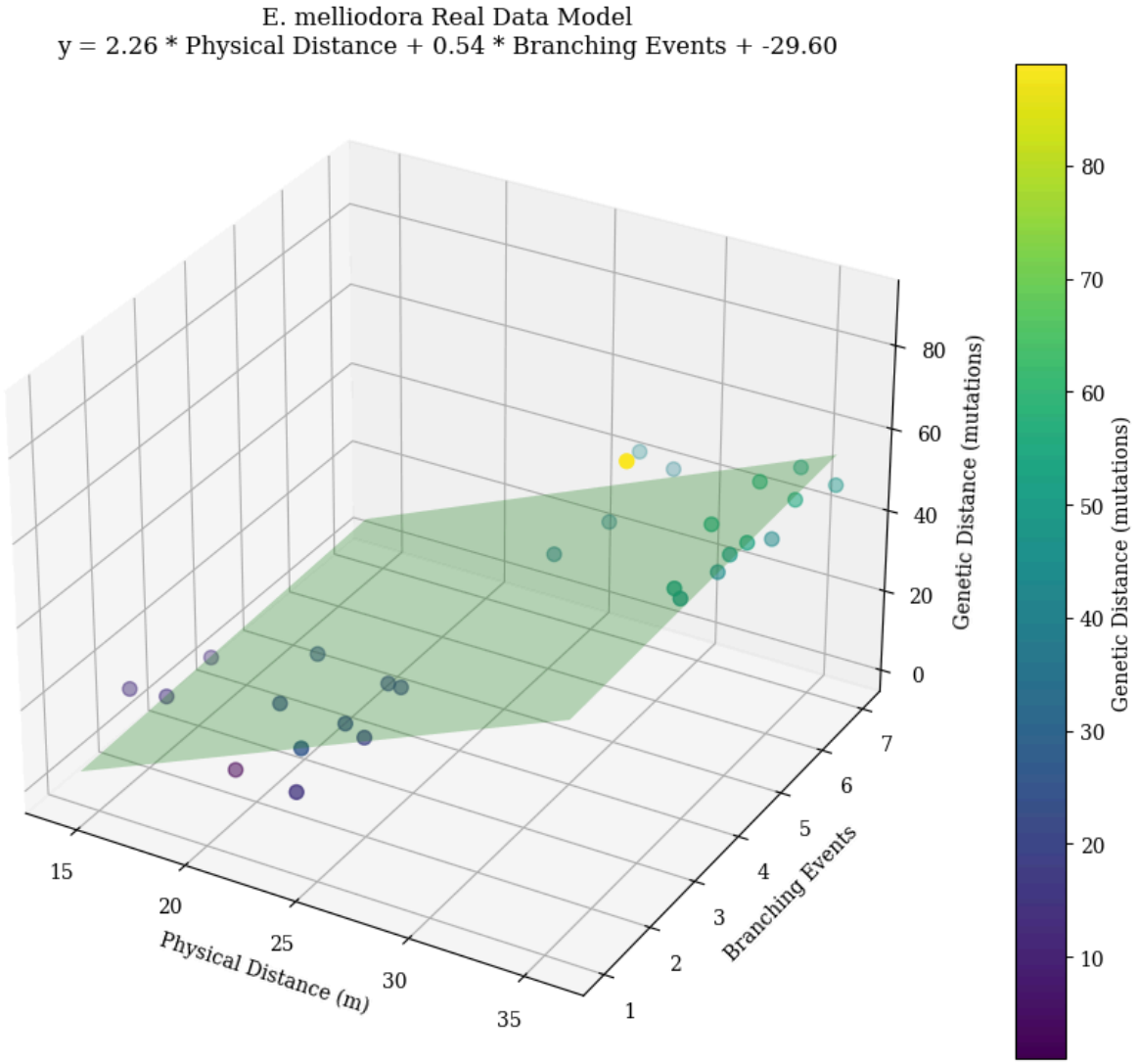


Figure 2.16 Multi-linear Regression Model including Physical Distances and Branching Events as Predictors of Genetic Distance. The model achieved $R^2 = 0.743$, with a significant effect of physical distance ($\beta = 2.26$, $p < 0.001$) but an insignificant effect of branching events ($\beta = 0.54$, $p = 0.695$).

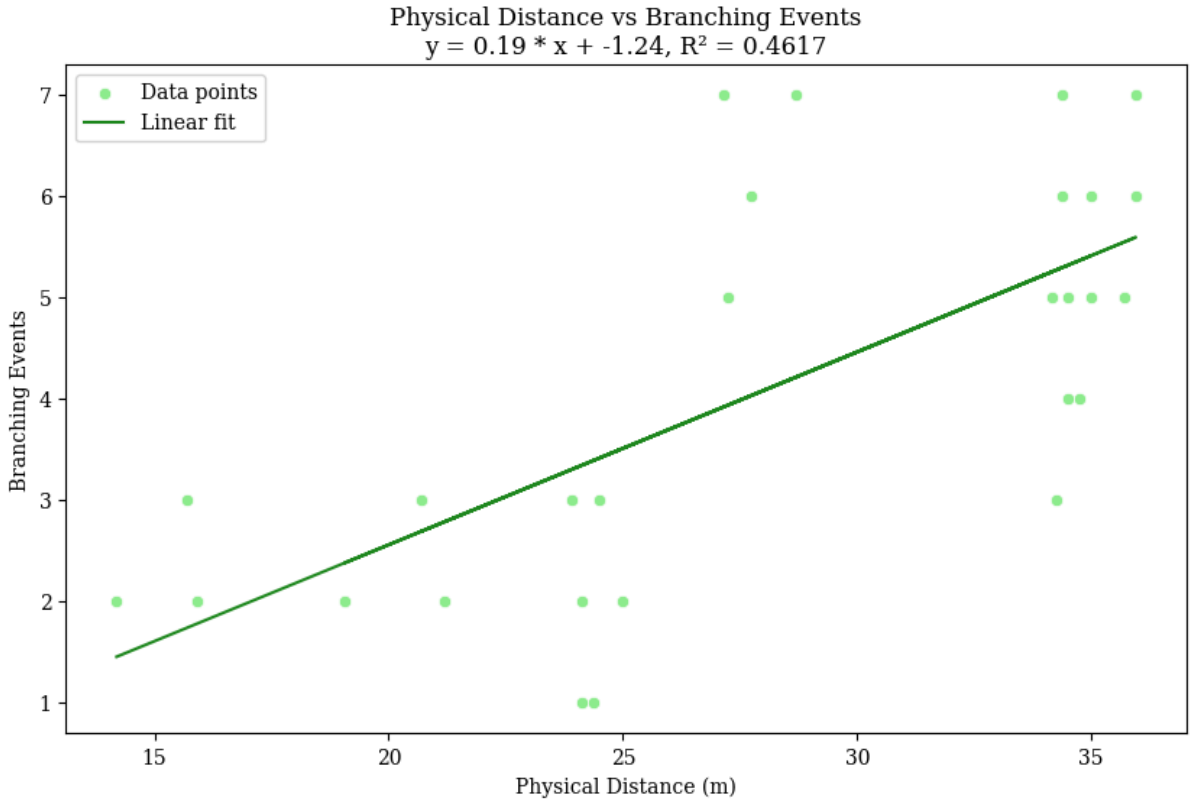


Figure 2.17 Linear Regression of Branching Events vs Physical Distance.

The equation $y = 0.19 * Physical\ Distance - 1.24$ and R-squared value ($R^2 = 0.4617$) indicate a moderate correlation, with nearly half of the variance in branching events explained by physical distance. This result further suggests redundancy between the two predictors, resulting in multicollinearity.

I applied the modified phylogenomic data to the real, post-dng filtered, *Eucalyptus melliodora* data, using the regression equation $y = 2.36x - 30.27$ ($R^2 = 0.7415$) derived from the physical distances measured and the genome wide-variant data computed by Orr et al. 2020 (Figure 2.18). The regression captured 74.15% of the variance in genetic distance, indicating a strong correlation between physical and genetic distances between branches. Using the tree's total physical span (90.1m), a total of 182 variants was initially estimated. I then corrected this estimate using a recovery rate of 29.95% and false-positive rate 0.12% derived directly from Orr et al.'s (2020) empirical data, recovering a corrected number of 610 variants. When normalised by the tree's estimated age (averaged to 125 years, based on Orr et al.'s range of 50-200 years) and genome size (500Mb), the corrected somatic rate was calculated to be 9.76×10^{-11} mutations per site per year.

This rate falls slightly below Orr et al.'s reported range (1.16×10^{-10} *to* 1.12×10^{-09}), derived per site per year per apical meristem. While my estimate captures genome-wide mutation dynamics across the tree, Orr et al.'s range specifically reflects rates associated with individual apical meristems during branch elongation. This distinction should be considered when comparing results.

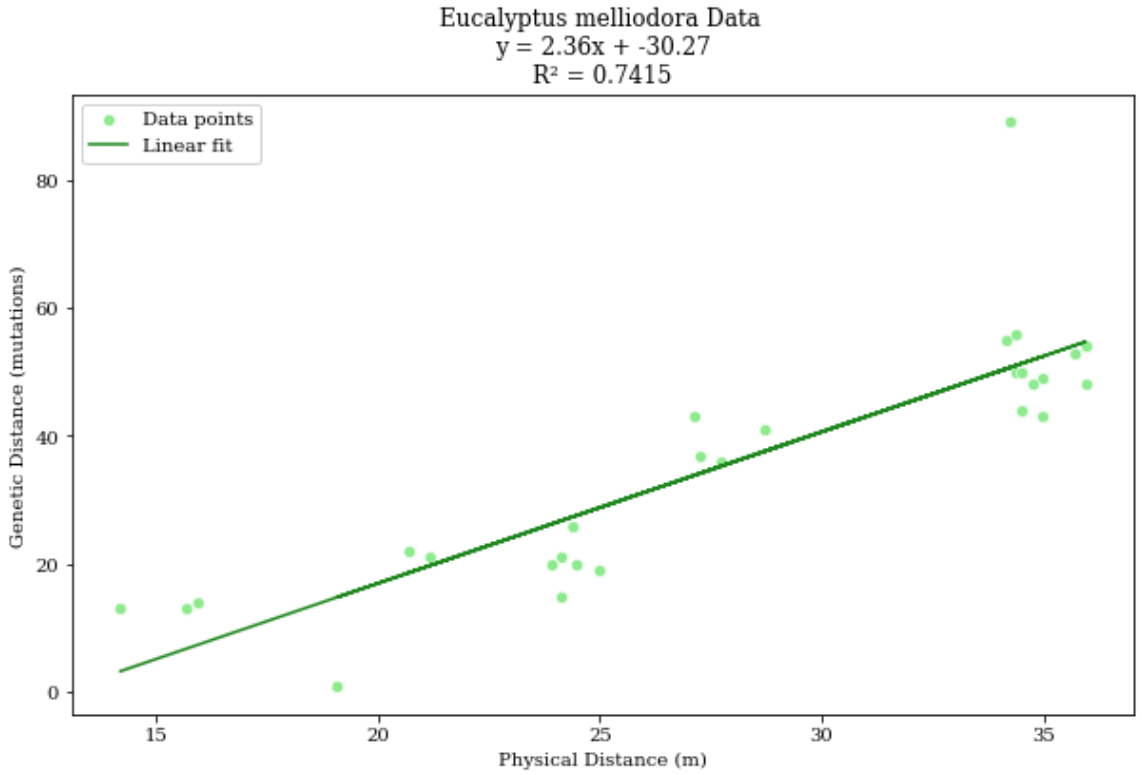


Figure 2.18 Application of the Modified Phylogenomic Method to real *Eucalyptus melliodora* data. The regression equation $y = 2.36x - 30.27$ ($R^2 = 0.7415$) estimates a corrected somatic mutation rate of 9.76×10^{-11} mutations per site per year, falling below Orr et al.'s reported range (1.16×10^{-10} *to* 1.12×10^{-09}).

I then applied the modified phylogenomic method to simulated data generated using the four models of somatic mutation accumulation defined by Tomimoto and Satake (2023) for the *E. melliodora* individual (Figure 2.19). The topology, measured physical distances, estimated age (125) and somatic mutation rate (6.18×10^{-10}, the mean value of Orr et al.'s range) of the *E. melliodora* tree were inputted directly into the simulation code.

The structured-biassed model (Model I) produced the regression equation $y = 5.29x + 310.11$, with a weak fit ($R^2 = 0.2015$). The weak R^2 suggests that physical distance explains little to no variation in genetic distance and may indicate that the modified phylogenomic method struggles to recover mutations under structured-biassed conditions. In Model I, mutations are uniformly maintained during structured elongation, minimizing variation in genetic divergence required for a strong correlation with physical distance. Biased branching further introduces localised effects, where select stem cell lineages disproportionately contribute to new branches. Depending on the mutation history of these lineages, biased branching may amplify or dampen genetic divergence between branches, further weakening the regression. This model estimated 788.38 total mutations, corresponding to a somatic mutation rate of 9.75×10^{-10} per site per year. This estimated rate falls within Orr et al.'s range but is substantially higher than the real data value estimated using the modified phylogenomic method (9.76×10^{-11}).

The structured-unbiased model (Model II) improved moderately ($R^2 = 0.6$), yielding the regression equation $y = 2.99x + 343.59$. The assumption of unbiased branching removes the localised effects introduced by biased lineage contributions, resulting in a stronger correlation between physical and genetic distances. However, the retention of all mutations during structured elongation continues to limit variation in genetic divergence. The estimated mutation rate (7.59×10^{-10}) remains higher than the real data estimate, indicating that structured elongation conditions fail to capture the dynamic observed in *E. melliodora*.

The stochastic-biassed model (Model III) provided a stronger fit ($R^2 = 0.796$) with the regression equation $y = 1.69x + 69.46$. Stochastic elongation introduces mutation loss through random sampling of lineages during cell divergence, increasing variation in genetic divergence as the physical distance grows. Biassed branching adds another layer of complexity, where select lineages disproportionately contribute to branches generating sufficient variation for the phylogenomic method to capture effectively. The estimated mutation rate (2.75×10^{-10}) aligns more closely with the real data estimate, suggesting the stochastic-biased dynamics partially reflect *E. melliodora*'s mutation accumulation processes.

The stochastic-unbiased model (Model IV) produced the best fit ($R^2 = 0.9$) and the regression equation $y = 1.46x + 71.41$. Model IV represents a scenario where random mutation loss during stochastic elongation, combined with unbiased sampling of stem cells during branching, creates a direct and proportional relationship between physical and genetic distances. This dynamic closely aligns with the real data results (Figure 2.18), where the estimated mutation rate (2.51×10^{-10}) is the closest to the real data estimate. The high $R^2$and similar regression equations indicate that stochastic-unbiased dynamics most accurately reflect mutation accumulation in *E. melliodora,* and they perform best under the modified phylogenomic method.

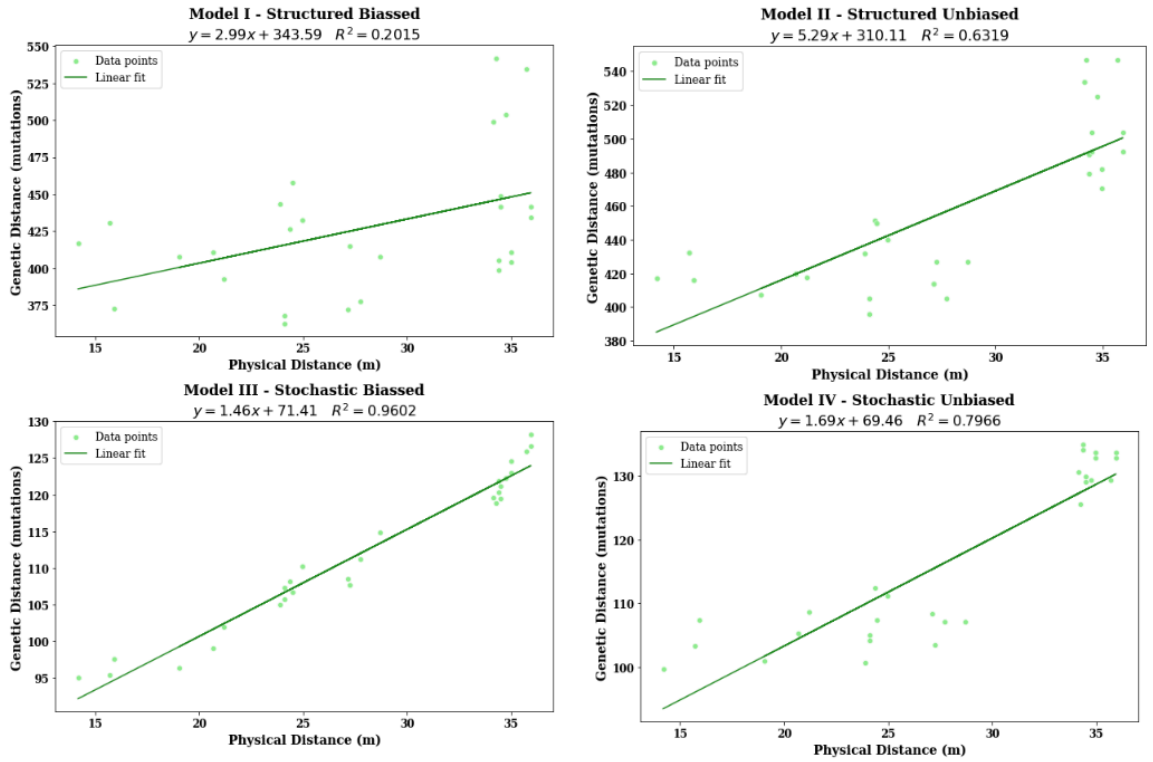


Figure 2.19 Regression Results for Simulated Data using Tomimoto and Satake's (2023) Models. Mutations were simulated under Models I-IV, using the topology, physical distances and age of the *Eucalyptus melliodora* individual described by Orr et al. 2020. Models I-IV vary by elongation behaviour, structured (retaining all cell lineages throughout divisions) or stochastic (random selection of lineages), and branching behaviours, biassed (lineages biasedly sampled in the formation of a new branch) or unbiased (lineages uniformly sampled in the formation of a new branch). Across all models, estimated somatic mutation rates (per site per year) fell in the same order of magnitude (between $2.51\ to\ 9.75 \times 10^{-10}$), with the stochastic-unbiased model providing the best fit ($R^2 = 0.9$).

2.4 Discussion

The findings presented in this chapter highlight the strengths and limitations of the phylogenomic method for estimating somatic mutation rates and its reliance on key assumptions about mutation dynamics and tree topology. While the method demonstrated some success in structured growth systems and balanced topologies, it faced challenges when applied to more complex biological systems, including the real *E. melliodora* individual (Orr et al., 2020).

The pre-dng genomic reads of the *E. melliodor*a individual aligned more closely with the expected distributions predicted by Tomimoto and Satake's (2023) models of somatic mutation accumulation. These results maintained low somatic mutation rates consistent with *E. melliodora* while retaining biologically relevant mutations that do not follow the tree's topology. However, post-topology filtering, designed to refine mutation estimates using tree topology, introduced systematic underestimation of mutation rates, particularly in branches 5 and 6. These branches share a long internal node, and the phylogenomic method disproportionately removed rare mutations that, while biologically meaningful, deviated from the strict topological assumptions of the method. This filtering

step, intended to reduce noise and improve accuracy, inadvertently discarded mutations vital for accurate rate estimation, particularly in low mutation-rate systems.

The challenges posed by topological filtering are consistent with criticisms raised by Tomimoto and Satake (2023), whose model framework demonstrated that mutations deviate from the physical topology of trees and that the method posed by Orr et al. (2020) risks overfiltering. In this study, unbalanced long-terminal topologies provided the most accurate recovery of mutation rates at low mutation rates, with regression coefficients statistically closest to one. Their extended terminal branches accumulated sufficient mutations proportional to physical length, while their asymmetry reduced shared internal mutations, enhancing signal distinctiveness. In contrast, balanced long-terminal topologies systematically underpredicted mutation rates, with regression coefficients dropping to 0.29–0.51 due to diluted signals and shared mutations at internal nodes. Unbalanced short-terminal topologies overestimated mutation rates, with coefficients exceeding 1.5, likely due to amplified signals in short branches and uneven mutation distributions. These findings reinforce that the phylogenomic method relies on strong, distinct mutational signals in terminal branches. While further simulations may refine the precise topological conditions required, it is clear that topologies with greater shared internal nodes and uneven mutation distributions hinder mutation rate recovery, while those with extended terminal branches and minimal shared mutations improve accuracy.

At higher mutation rates, the phylogenomic method struggled across all topologies due to shared mutation homogeneity, which reduced genetic distinctiveness and led to widespread underestimation of mutation rates, with regression slopes approaching zero. Initially, I hypothesized that back mutations might contribute to these distortions, but analyses confirmed they were rare and had no significant impact, affirming the robustness of the simulation framework. Instead, the key issue appears to be that at high mutation rates, shared mutations become more uniformly distributed across branches, diminishing phylogenetic resolution.

Branching bias also had negligible effects on mutation rate recovery and phylogenetic accuracy, with small beta coefficients across simulations. While this suggests that bias in axillary meristem formation may have little biological impact, it remains unclear whether this result reflects a genuine lack of influence or a limitation of the simulation framework itself. Given that the model applies a constant mutation rate, I would expect both unique and shared mutations to increase proportionally, yet the stable coefficient of variation (CV) suggests a shift in the ratio of shared-to-unique mutations. This raises the question of whether the observed trends are true biological properties of somatic mutation accumulation or artifacts of the model's assumptions.

Future work should explore whether these patterns persist in empirical datasets and assess how modifying branching bias, elongation structure, or mutation rate distributions influences shared mutation distributions. Additionally, investigating alternative phylogenetic reconstruction methods that account for mutation saturation and homoplasy may help clarify whether underestimation at high mutation rates is an inherent limitation of the phylogenomic method or an issue arising from the simulation approach.

The application of the method to *Eucalyptus melliodora* underscores its strengths and weaknesses. Stochastic elongation aligned more closely with real-world data for this species, which exhibits low mutation rates likely driven by reduced fixation. However, this alignment was disrupted by the phylogenomic method’s topological filtering, which compounded the issues of mutation rate underestimation. These findings suggest that while stochastic growth can align with real biological

contexts, the method's reliance on strict topological assumptions limits its broader applicability to systems with irregular growth patterns.

Phylogenetic accuracy also varied significantly across topologies. Balanced configurations consistently yielded lower Robinson-Foulds distances compared to unbalanced topologies with the same number of terminal branches, reflecting the proportional distribution of mutations in balanced trees. Subsampling tests using *E. melliodora's* real topology reinforced these observations: a four-branch balanced subsample exhibited perfect recovery (RF distance = 0), whereas the full eight-branch unbalanced topology displayed branch misplacements and novel groupings, reducing recovery accuracy and increasing mutation rate deviations. These results further support Tomimoto and Satake's (2023) findings, where mutation accumulation deviates from the tree topology -specifically in complex modular architecture structures.

In conclusion, while the phylogenomic method has demonstrated some potential in specific scenarios, its broader application is limited by its reliance on regular mutation patterns, topological filtering, and tree balance. Its performance diminishes significantly in systems with high mutation rates, complex branching structures, or shared internal nodes. These findings underscore the need for a new approach that moves beyond the constraints of the phylogenomic method, which is fundamentally dependent on a flawed assumption- mutations accumulate following topology. The complex biological dynamics of somatic mutation accumulation in real long-lived trees do not result in a simple correlation between physical and genetic distances capable of being handled by the phylogenomic method. To address this, in the following chapter, I propose a prototype of an alternative novel mutation rate estimation method utilising approximate Bayesian computation, extending the modelling framework posed by Tomimoto and Satake (2023).

Chapter 3: A Novel ABC Approach For Somatic Mutation Inference in Long-Lived Trees

3.1 Introduction

This chapter introduces a novel Approximate Bayesian Computation (ABC) methodology for estimating the somatic mutation rate of long-lived trees. Like the phylogenomic method explored in Chapter 2, the ABC approach uses observed mutation patterns across branches and tree topology information to estimate mutation rates. However, unlike the phylogenomic method, it does not assume mutations strictly follow the tree's topology. This assumption hindered the phylogenetic method, which performed poorly across complex tree topologies and high mutation rates with decreased genetic signal and mutation overload. To address these limitations, I developed an alternative ABC-based approach utilising Tomimoto and Satake's (2023) models of somatic mutation accumulation and my previously developed simulation program.

Approximate Bayesian Computation (ABC) is a simulation-based framework for parameter estimation that bypasses the need to explicitly evaluate the likelihood function, which can often be analytically intractable or computationally expensive (Sunnåker et al., 2013). Instead, ABC relies on simulations to approximate the posterior distribution of parameters by comparing observed real data to simulated data using a distance metric. In this study, I employ the ABC-Rejection algorithm, the simplest form

of ABC, to estimate somatic mutation rates in long-lived trees (Tavare et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002). The algorithm begins by sampling mutation rates from a biologically informed prior distribution and simulating mutation patterns across the input tree topology. The observed distribution of mutations across branches is compared to the simulated data using Euclidean distance (Gower, 1985). Parameter sets of a simulation are accepted if the discrepancy falls within a pre-specified tolerance (epsilon, ϵ). This process produces a posterior distribution of accepted mutation rates, which reliably approximates the true somatic mutation rate of the tree individual (Turner & Van Zandt, 2012).

In addition to mutation rate recovery, the ABC method estimates developmental parameters such as StD (elongation parameter) and biasVar (branching parameter). The StD parameter reflects the extent of stem cell lineage preservation within meristems throughout elongation growth processes, providing insights into somatic genetic drift, with higher values indicating greater stochasticity or somatic drift. Moreover, biasVar reveals the degree of bias of stem cell lineage selection during axillary meristem formation (branching), with lower values suggesting stronger bias. The posterior distributions of these parameters approximate the developmental behaviour of meristems in long-lived trees, allowing inference of somatic accumulation dynamics alongside a somatic mutation rate.

As a proof of concept, this chapter demonstrates the feasibility and potential of the ABC methodology in capturing somatic mutation dynamics while highlighting areas for refinement in future iterations. Free from the restrictive assumptions and limitations of the phylogenomic method, this approach allows for greater flexibility in exploring a broader range of mutation scenarios across diverse tree topologies and taxa.

3.2 Methodology

3.2.1 ABC-Reject Framework

Approximate Bayesian Computation (ABC) is a Bayesian inference method that replaces explicit likelihood evaluation with simulations, making it ideal for models where the likelihood function is analytically intractable or computationally prohibitive. To understand ABC, one must begin with Bayes' theorem, the foundation of Bayesian inference (Efron, 2013):

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Here:

- $p(\theta|D)$ is the **posterior distribution**, representing the probability of parameter values θ given the observed data D.
- $p(D|\theta)$ is the **likelihood**, quantifying the probability of observing D if the parameters θ are true.
- $p(\theta)$ is the **prior distribution**, reflecting one's prior knowledge or assumptions about the parameters before observing any data.
- $p(D)$ is the **marginal likelihood** (aka. evidence), a normalisation constant ensuring the posterior integrates to 1.

In traditional Bayesian methods, the likelihood $p(D|\theta)$ must be explicitly calculated. However, this is often computationally challenging for complex systems- like the accumulation and distribution of

somatic mutations across tree branches. ABC overcomes this limitation by replacing the likelihood with simulations. Rather than evaluating $p(D|\theta)$ directly, ABC approximates the posterior distribution $p(\theta|D)$ as follows (see Figure 3.1):

1. **Sample Parameter Values**: Parameter values θ are drawn from the prior distribution $p(\theta)$. For my application, this includes biologically relevant priors for the somatic mutation rate as well as the model parameters (StD, biasVar) ranges defined by Tomimoto and Satake (2023) (Section 2.2.1).
2. **Simulate Data**: θ values are used to simulate dataset D', which,in this case, is the distribution of somatic mutations across branches of the input tree topology.
3. **Summarize Data**: Compute summary statistics $S(D')$ for the simulated data D' and compare them to the summary statistics $S(D)$ of the observed data (the real or observed distribution of somatic mutations across a long-lived tree).
4. **Measure Distance**: The distance ρ between the observed and simulated summary statistics is calculated using the Euclidean distance (Gower, 1985):

$$\rho(S(D'), S(D)) = \sqrt{\sum_{i=1}^{n} (S_i(D') - S_i(D))^2}.$$

   Here, S_i represents the i-th summary statistic, and n is the total number of statistics. The distance ρ provides a quantitative measure of how closely the simulated distribution of mutations resembles the observed or true distribution.

5. **Accept of Reject Parameters**: If the distance ρ is less than or equal to a tolerance threshold epsilon (ϵ), the parameter set θ is accepted; otherwise, it is rejected. Mathematically, this is expressed as:

$$\theta \; is \; accepted \; if \; (S(D'), S(D)) \leq \epsilon$$

6. **Posterior Approximation**: After many iterations, the accepted values of θ form a sample from the approximate posterior distribution $p(\theta|D)$.

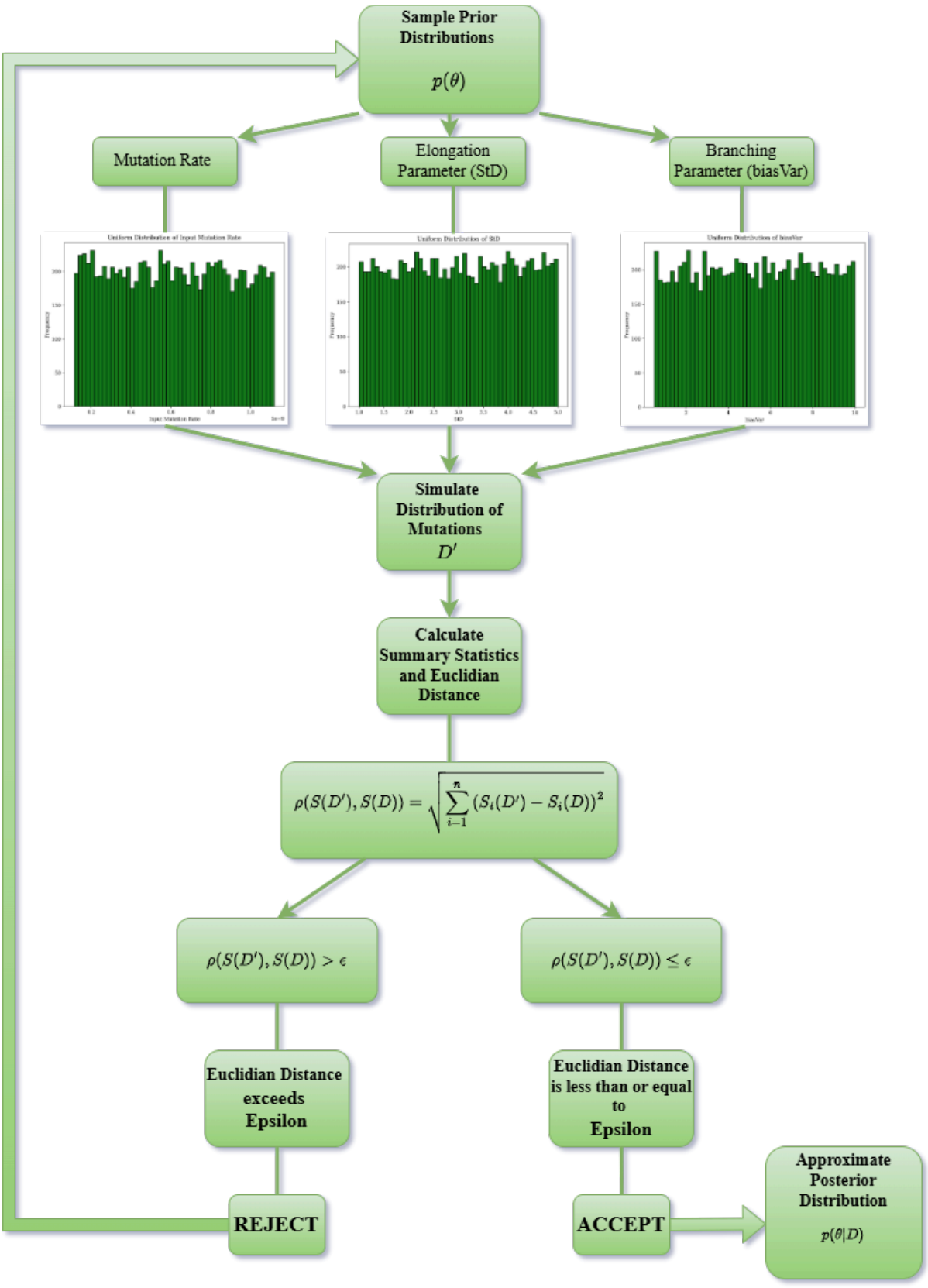


Figure 3.1 Flowchart of the ABC-Reject Framework.

3.2.2 Validation of ABC-Reject Methodology

To validate the accuracy and reliability of my novel ABC-Reject methodology, I used simulated datasets in which the true parameter values and mutation distributions were known. This approach ensured that validation was independent of empirical errors inherent in real genomic data, specifically those reported for *E. melliodora* by Orr et al. (2020). A total of 200 validation samples were generated using the Latin Hypercube Sampling (LHS) framework employed in my prior simulations, ensuring comprehensive coverage of the parameter space. Each validation sample included predefined 'true' values of mutation rate (μ), elongation parameter (StD) and branching bias (biasVar). The StD and biasVar true values were chosen as the lower quartile, median and upper quartile range of their prior distributions (where StD=0 is excluded, see Section 3.2.3.1), while the mutation rate was sampled at two biologically realistic extremes: 6×10^{-9} (large) and 6×10^{-10}(small). Other attributes, such as topology and branch number, were also sampled using the predefined 20 unique topologies designed for the simulations.

Synthetic mutation distributions were simulated using the sampled parameters and the pipeline described in Chapter 2 (Section 2.2.5). These simulated distributions acted as true observed datasets for the validation samples, allowing for a direct and controlled assessment of the ABC-Reject methodologies' ability to recover true parameter values in the approximated posterior distribution. Once the synthetic datasets were generated, the ABC-Reject framework was applied to each validation sample using the final priors and tolerance threshold ($\epsilon = 20$) selected after sensitivity analyses with *E. melliodora* data (see Section 3.2.3.1). For each validation sample, the script ran until 100 accepted samples (where $\rho \leq \epsilon$) were obtained. While most validation runs were completed successfully, computational limitations- such as memory constraints and timeouts- prevented 31 samples from converging, leaving 169 validation samples for analysis.

Posterior distributions of the accepted parameter sets were then analysed across 169 validation samples to assess the framework's accuracy. Highest posterior density (HPD)/credible intervals and summary statistics were computed for each parameter (`input_mut`, `StD` and `biasVar`) using the `arviz` package. For a validation sample to be successful, the true parameter value had to fall within the 95% HPD interval of the posterior distribution. To further evaluate posterior performance, randomly selected validation samples were analysed and visualised using `arviz`, ensuring that posterior approximations accurately captured the true input values.

The validation was implemented in the script valid_new.py, available in the GitHub repository (https://github.com/andreag186/sim_tree_mut). The number of validation samples (`NUM_SAMPLES`), number of accepted samples per validation sample (`NUM_ACCEPTED`), threshold (`EPSILON`) and true values for parameters (`input_mut_values`, `StD_values`, and `biasVar_values`) can be altered for different applications. This script can enable SLURM batch processing with high-performance computing environments (e.g. NeSI) using `{task_id}`.

3.2.3 Application of ABC-Reject to *E. melliodora*

The ABC-Reject framework was applied to *Eucalyptus melliodora* somatic mutation estimates reported by Orr et al. (2020), namely two datasets: the pre-dng dataset, containing 330 high-confidence variants prior to applying the phylogenomic method, and the post-dng dataset representing the 90 mutations which remained after filtering with the tree topology.

The distribution of mutations across the eight branches of the *E. melliodora* tree for each data-set, pre- and post-dng, were both treated as the 'observed' or 'true' data (D) in separate ABC-Reject applications.

3.2.3.1 Sensitivity Analysis

An iterative, rudimentary sensitivity analysis was conducted to refine prior distributions and the acceptance threshold (ϵ) in the application of the ABC framework to *E. melliodora* pre- and post-dng empirical data. Initial priors were drawn from literature estimates, but early trials indicated constraints that limited posterior exploration, necessitating adjustments.

The mutation rate (μ) prior was initially based on Orr et al. (2020) (1.6×10^{-10} to 1.12×10^{-9} per site per year per apical meristem) but proved too restrictive, particularly for post-DNG data which clustered around lowered values. The mutation rate prior was iteratively expanded until there was a visible improvement in posterior distributions. The final selected range (1×10^{-11} to 9×10^{-9}) effectively captured the posterior density while preventing excessive parameter rejection (Table 3.1).

The elongation parameter (StD) prior was initially set to 0 to 5, following the range defined by Tomimoto and Satake (2023). However, preliminary runs revealed that when StD = 0, Tomimoto and Satake's (2023) model effectively reduced to a lower-dimensional parameter space, with only branching bias (biasVar) actively influencing somatic accumulation. In contrast, when $StD > 0$, both elongation and branching played a role in determining mutation patterns. This dimensionality shift led to model-jumping (Hubin & Storvik, 2018), where posterior distributions clustered around StD = 0 and failed to mix effectively across parameter space. To address this, the value of $StD = 0$ was removed, and the final prior was adjusted to 1 to 5, representing a spectrum for partially structured to fully stochastic ($StD = 5$). A summary of initial and prior ranges is provided in Table 3.1.

ABC acceptance thresholds (ϵ) control the trade-off between posterior accuracy and computational efficiency. If ϵ is too large, the posterior closely resembles the prior, reducing inference. If ϵ is too small, excessive parameter sets are rejected, increasing computational costs. Initial trials tested multiple ϵ values using the Euclidean distance metric to compare the observed mutation distribution with simulated distributions. After several refinements, an acceptance threshold of $\epsilon = 20$ was selected as it provided sufficiently strong posterior distributions while maintaining an acceptable rejection rate.

	Mutation Rate (μ)	Elongation Parameter (StD)	Branching Parameter (biasVar)
Initial Prior Range	[1.6E-10, 1.12E-09]	[0,5]	[0.5,10]
Final Prior Range	[1E-11, 9E-09]	[1,5]	[0.5,10] (*unchanged*)

Table 3.1 Initial and Prior Ranges for Parameters

For both the pre-dng and post-dng datasets, the ABC framework was run using 10,000 trials/parameter sets sampled from the final prior distributions. Each sampled set was used to simulation the mutation distribution across the *E. melliodora* tree topology, and simulations were accepted if the Euclidean distance between simulated and observed distributions was within the selected threshold ($\epsilon = 20$).

3.2.3.2 ABC Implementation

The analysis was conducted using the Python scripts `abc_non_dng.py` (pre-dng) and `abc_phylo.py` (post-dng), available on the GitHub repository (https://github.com/andreag186/sim_tree_mut). These scripts implement the ABC-Reject framework described earlier to estimate the somatic mutation rate and developmental parameters (StD and biasVar) for the *Eucalyptus melliodora* investigated by Orr et al. (2020). These scripts differ in the observed mutation distributions they evaluate- post-dng (filtered with topology, post-phylogenomic method) in `abc_phylo.py` and pre-dng (filtered normally with haplotype caller) in `abc_non_dng.py`. Both scripts build on the simulation detailed in Chapter 2 (Section 2.2.5) but introduce new functions and external modules to implement the ABC-Reject framework alongside simulations. For a more detailed analysis

The scripts are written in **Python 3.11.4** and heavily rely on the `arviz` (Kumar et al., 2019) library (not included in `sim_code.py`) for analysis and handling of the approximated posterior distribution. To reproduce results, ensure this dependency and other key libraries are installed (`pip install arviz numpy scipy pandas matplotlib`). A detailed explanation of the ABC code's functions, parameters and caveats can also be found in Appendix C.

3.3 Results

3.3.1 Validation Results

The validation of the ABC framework was assessed using 169 simulated datasets, each with known true parameter values for mutation rate (input_mut, μ), elongation parameter (StD), and branching bias (biasVar). The primary objective was to evaluate the framework's ability to recover these true values across a diverse range of tree topologies and parameter values. The results revealed a high level of accuracy, with **100%** (169 of 169) of the mutation rate and biasVar true values falling within the 95% HPD intervals. For StD, **99.4%** (168 of 169) datasets successfully captured the true value in the 95% HPD of the approximated posterior distribution. The near-flawless validation outcomes affirm the performance of the ABC framework, which successfully produces credible posterior estimates that align with known parameter values across diverse topologies/biological parameters.

To explore the range of outcomes in greater detail, two specific validation samples were selected as illustrative examples; one representing one of the samples with the best distance metric and posterior distribution shapes (the '***BEST***') and the other being the single sample where StD's true value falls outside the HPD interval (the '***WORST***'). These samples provide insight into the framework's performance under optimal and challenging conditions.

The '***best***' sample demonstrated exemplary results, with all true parameter values for input_mut, StD, and biasVar closely aligned with posterior distributions (Figure 3.2). The mutation rate posterior was unimodal and symmetrical, with a mean of 6.01E-10, a median of 5.91E-10, and a 95% HPD interval [2.43E-10, 9.44E-10], closely matching the true value of 6E-10 (Table 3.2). StD exhibited a slight skew, with a mean of 2.66, a median of 2.48, and a 95% HPD of [1.08, 4.52], capturing the true value of 2.50. BiasVar showed a broad yet centred posterior, with a mean of 5.12, a median of 5.03, and a 95% HPD of [0.57, 7.69], including the true value 5.25.

Trace plots for the best sample exhibited moderate mixing, consistent variability and no signs of divergence (Figure 3.4). ESS values were especially strong for the meristem behaviour parameters: 100 for StD and 98.20 for biasVar, indicating efficient exploration across the 100 accepted samples forming the posteriors. The mutation rate, however, had a lower ESS value of 67.06, likely due to its broader posterior distribution and higher sensitivity to variability in the data, which can result in autocorrelation during sampling.

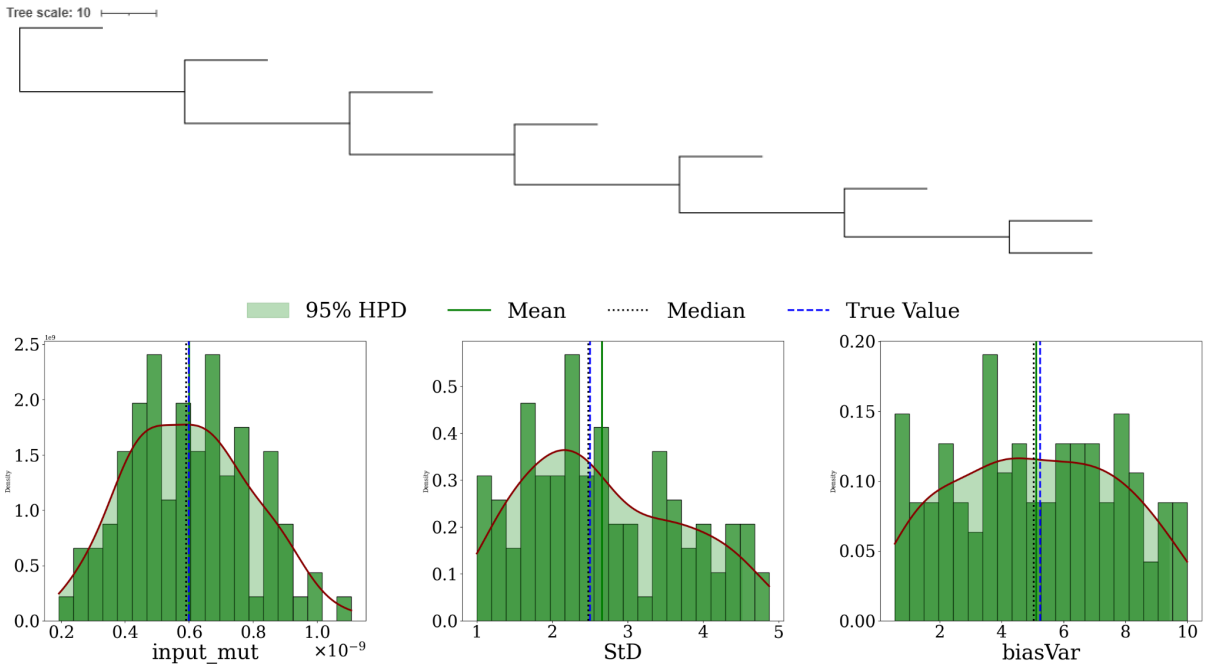


Figure 3.2 Posterior Distributions of Mutation Rate (input_mut), Elongation Parameter (StD) and Branching Bias (biasVar) for the "*BEST*" Validation Sample. The shaded area represents the 95% highest posterior density (HPD) interval, with mean (solid green line), median (dotted green line), and true values (dotted blue line) indicated. The green histograms represent the density distribution of accepted sampled parameter values, while the red line represents the smoothed Kernel Density Estimation KDE curve, providing a continuous estimate of the probability density. The topology of the 'best' sample is plotted above, an unbalanced topology with 8 terminal short branches. The tree scale (10 years/100cm) is shown above for reference.

In contrast, the '***worst***' sample illustrated the potential limitations of posterior approximation within the ABC Framework. For this sample, the true value for StD (1.25) fell outside the 95% HPD interval of [1.27, 4.97], marking the only instance where this occurred among all validation datasets (Figure 3.3). The mutation rate posterior showed a broader distribution, with a mean of 3.67E-10, a median of 3.23E-10, and an HPD of [1.50E-10, 6.97E-10], reflecting greater uncertainty while still capturing the true mutation rate (6.0E-10). BiasVar showed moderate precision, with a mean and median of 4.95 and a 95% HPD interval of [5.05,9,54], successfully including the true value of 5.25 (Table 3.2).

Trace plots for this sample revealed noisier chains with higher autocorrelation and reduced mixing efficiency (Figure 3.4), as reflected in lower ESS values: 67.73 for input_mut, 55.55 for StD, and 57.73 for biasVar (Table 3.2). However, the issues with trace plots across both the best and worst

samples are likely due to the small number of accepted samples (100), which may have constrained the framework's ability to explore the posterior fully.

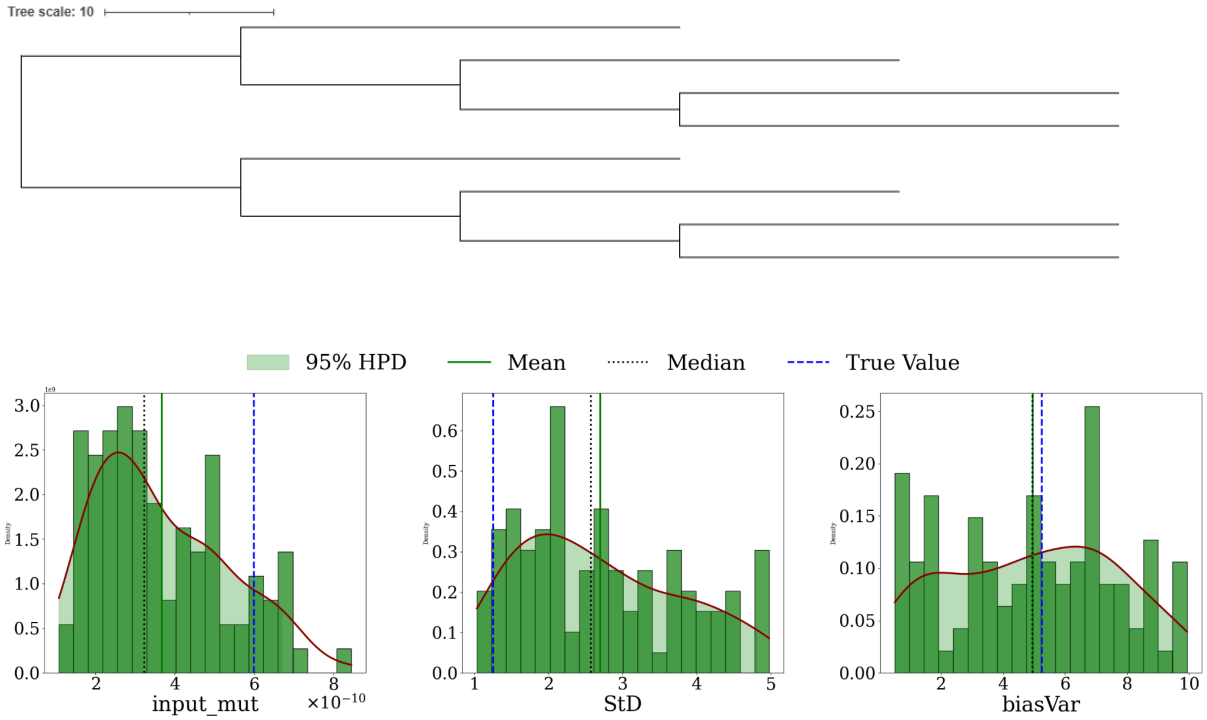


Figure 3.3 Posterior Distributions of Mutation Rate (input_mut), Elongation Parameter (StD) and Branching Bias (biasVar) for the "*WORST*" Validation Sample. The shaded area represents the 95% highest posterior density (HPD) interval, with mean (solid green line), median (dotted green line), and true values (dotted blue line) indicated. The green histograms represent the density distribution of accepted sampled parameter values, while the red line represents the smoothed Kernel Density Estimation KDE curve, providing a continuous estimate of the probability density. The topology of the 'worst' sample is plotted above, a balanced topology with 8 terminal long branches. The tree scale (10 years/100cm) is shown above for reference. The true value of StD (1.25) for the worst sample falls just below the 95% HPD range [1.27, 4.97].

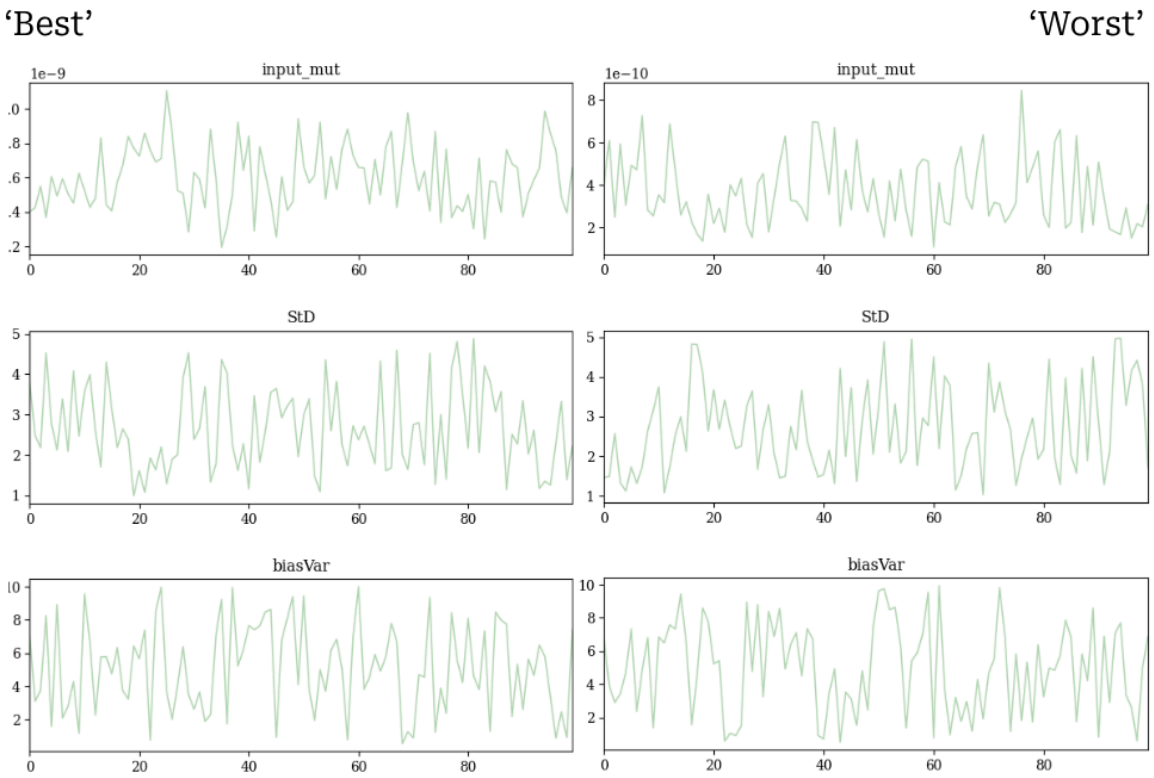


Figure 3.4 Trace Plots for Mutation Rate (input_mut), Elongation Parameter (StD) and Branching Bias (biasVar) for the "*BEST*" (left) and "*WORST*" (right) Validation Samples. Each validation sample has exactly 100 accepted samples per parameter. The best sample shows moderate mixing and high ESS values, while the worst sample exhibits higher autocorrelation and reduced mixing efficiency, likely due to limited posterior exploration (see Table 3.2).

	Mean	**Standard Deviation**	**Median**	**95% HPD**	**True Value**	**ESS**
StD	2.66	1.03	2.48	[1.08, 4.52]	2.50	100.00
biasVar	5.12	2.64	5.03	[7.69, 0.57]	5.25	98.20
input_mut	6.01E-10	1.90E-10	5.91E-10	[2.43E-10, 9.44E-10]	6.00E-10	67.06
StD	2.70	1.09	2.57	[1.27, 4.97]	1.25	57.55
biasVar	4.95	2.70	4.95	[5.05, 9.54]	5.25	57.73
input_mut	3.67E-10	1.64E-10	3.23E-10	[1.50E-10, 6.97E-10]	6.00E-10	67.72

Table 3.2 Summary Statistics for the Posterior Distributions of Mutation Rate (input_mut), Elongation Parameter (StD), and Branching Parameter (biasVar) for the "*BEST*" and "*WORST*" Validation Samples. Best (highlighted green) and worst (highlighted blue) summary statistics were generated using ArviZ (Kumar et al., 2019). Corresponding to Figures 3.2-3.4.

3.3.2 *E. melliodora* ABC Applications

The phylogenomic method of somatic mutation rate estimation, as explored in Chapter 2, revealed significant limitations, mainly due to the overfiltering of mutations inconsistent with the tree topology. To address this, I applied the novel Approximate Bayesian Computation (ABC) framework to the *Eucalyptus melliodora* mutation data (Orr et al., 2020) with the following objectives:

1. To assess how well the ABC framework approximates mutation rates for a given real biological tree, using 'real' mutation datasets provided by Orr et al. 2020
2. To explore whether developmental parameters such as elongation (StD) and branching bias (biasVar) provide biologically meaningful insights under the ABC approach.

Two datasets were analyzed:

- The pre-dng dataset, containing a distribution approximately of 330 high-confidence variants identified before the phylogenomic filtering
- The post-dng dataset, containing a more restrictive dataset of 90 mutations filtered using *E. melliora* 's topology.

The results highlight the strengths and potential limitations of my novel ABC method and further investigates the impact of the phylogenomic method. The posterior distributions for mutation rates (μ, input_mut) and developmental parameters (StD, biasVar) revealed stark differences between the pre-dng and post-dng datasets (Figure 3.5).

3.3.2.1. Elongation Parameter (StD)

The posterior distribution of StD approximated for the pre-dng dataset was relatively flat, with a mean of 2.943 and a standard deviation of 1.174 (Figure 3.5, Table 3.3). The 95% highest posterior density (HPD) interval ranged from 1.002 to 4.760, indicating no strong structured or stochastic elongation signal. In contrast, the post-dng dataset revealed a more skewed posterior, with a mean of 2.36, a narrower standard deviation of 0.980 and an HPD range of 1.033 to 4.26- suggesting a shift towards more structured elongation. Notably, the final prior range ($StD = 1\ to\ 5$) prevents StD from reaching 0, meaning I could not fully assess whether the posterior would approach highly structured elongation for the observed post-dng data. The potential implications of this will be further explored in this chapter's discussion.

Trace plots (Figure 3.6) demonstrated consistent mixing for StD in the pre-dng dataset, with a bulk Effective Sampling Size (ESS) of 914 and a tail Effective Sampling Size (ESS) of 819, supporting robust sampling. Bulk ESS measures how well the entire posterior distribution is sampled, while tail ESS focuses on the precision in the extreme 5% of the posterior distribution (2.5% at each end), ensuring accurate exploration of the tails. In the post-dng dataset, slower mixing and higher autocorrelation were evident, with lower ESS values (bulk ESS = 342, tail ESS = 320). The trace plot further demonstrates autocorrelation with the appearance of streaks of similar values, indicating that successive samples are more dependent on each other, reducing the scope and efficiency of posterior exploration.

3.3.2.2. Branching Bias (biasVar)

BiasVar exhibited broad, primarily flat posterior distributions across both datasets (Figure 3.5). The prior distribution ranged from 0.5 to 10, where 0.5 represents biassed branching and 10 unbiased. For the pre-dng dataset, the mean was 5.307 with a standard deviation of 2.681, and a 95% HPD interval spanning from 0.639 to 9.442 (Table 3.3). Similarly, in the post-dng dataset, the posterior remained flat, with a mean of 5.18 a similarly high standard deviation of 2.730 and an HPD range of 0.991 to 9.939. These means, located near the prior range's centre, suggest some branching bias; however, the flat posterior shape indicates that these values are more likely influenced by the prior distribution rather than the observed datasets. In other words, the posterior closely mirrors the prior.

Trace plots (Figure 3.6) showed moderate mixing for biasVar in the pre-dng dataset, with a bulk ESS of 842 and a tail ESS of 741. The trace plot indicates a reasonable exploration of parameters, with the absence of long streaks indicating sufficient variation across iterations. In the post-dng dataset, slower mixing was evident, with reduced ESS values (bulk ESS = 282, tail ESS= 272). The trace plot highlights this with greater clustering of values, suggesting higher autocorrelation and inefficient sampling.

These findings indicate that biasVar may not be strongly informed by the observed datasets, suggesting limited biological relevance under the current framework. Further work should explore whether this flat distribution reflects true biological significance or limitations in the Tomimoto and Satake (2023) model and/or its implementation.

3.3.2.3. Mutation Rate (μ, input_mut)

The posterior distributions for somatic mutation rate (Figure 3.5) were strikingly different between datasets. For the pre-dng dataset, the posterior was symmetrical and unimodal (albeit with a long left tail between 2.3E-10 to 4E-10), with a mean value of 5.3E-10, a standard deviation of 1E-10, and an HPD of 2.3E-10 to 7.1E-10 (Table 3.3). These values fall mostly within Orr et al.'s reported range (1.16E-10 to 1.12 E-09), derived from post-dng 90 high-confidence variants after extensive transformations, including scaling by recovery rates, branch lengths and age assumptions.

In contrast, the post-DNG dataset exhibited a broader and more complex posterior, with a mean of 1E-10, a standard deviation of 1E-10, and an HPD interval of 4.2E-11 to 2.2E-10 (Table 3.3). Unlike the pre-DNG dataset, which had a single peak, the post-DNG mutation rate posterior displays bimodal characteristics, with one small peak around 2E-10 and a second, more pronounced peak at approximately 1E-10/9E-11 (Figure 3.5). The kernel density estimate (KDE) curve further supports this, showing a clear dip between these two regions, indicating the presence of two distinct groups of accepted value.

To investigate the potential drivers of this bimodal structure, I examined how mutation rate relates to other key parameters. Figure 3.8 presents scatter plots of mutation rate against biasVar and StD, revealing a negative correlation between mutation rate and StD. Mutation rates are highest when StD is low (~1-2) and systematically decrease as StD increases, clustering into two broad groups. In contrast, mutation rate does not show a clear relationship with biasVar, suggesting that the bimodal structure is not influenced by the branching bias.

Trace plots (Figure 3.6) revealed mixed performance in sampling quality for mutation rate. For the pre-dng dataset, the trace plot shows some inconsistency in mixing, with long tails in certain regions

indicating occasional inefficiencies in exploring the extremes of the posterior distribution. Specifically, the long tails likely correlate to the left tail and small peaks observed in the pre-dng posterior. This observation aligns with the lower tail ESS (707) compared to the bulk ESS (970), indicating that the chain may struggle to sample the lower extremes of the mutation rate posterior fully. In the post-dng dataset, the trace plot appears relatively consistent, with fewer streaks of similar values than parameters StD and biasVar. Bulk ESS (263) and tail ESS (283) indicate lower sampling efficiency compared to pre-dng, which can be attributed to the smaller number of accepted samples - 350 out of 10,000 trials, compared to 1,020 accepted samples for pre-dng. This disparity is likely due to the highly-filtered nature of the post-dng data, with branches 5 and 6 of the tree retaining little to no unique mutations after the phylogenomic method was applied.
This reduced number of accepted samples limits the chain's ability to explore the posterior distribution comprehensively. Despite these limitations, the shorter trace length for post-dng shows relatively consistent variability across iterations, suggesting that sampling was reasonably effective given the reduced genetic signal.

Figure 3.7 further illustrates the disparity between the two datasets, visually contrasting the HPD ranges for mutation rates. The pre-dng HPD closely aligns with Orr et al.'s estimated range, despite consisting of a distribution of mutations which occurred prior to phylogenomic topology filtering. In contrast, the post-dng HPD range- in which the phylogenomic method was applied- falls mostly below Orr et al. 's range, emphasizing the impact of overfiltering. While Orr et al.'s method attempts to compensate for these effects through scaling, the ABC framework's direct inference avoids such assumptions, preserving a clearer representation of the genetic signal within the observed data.

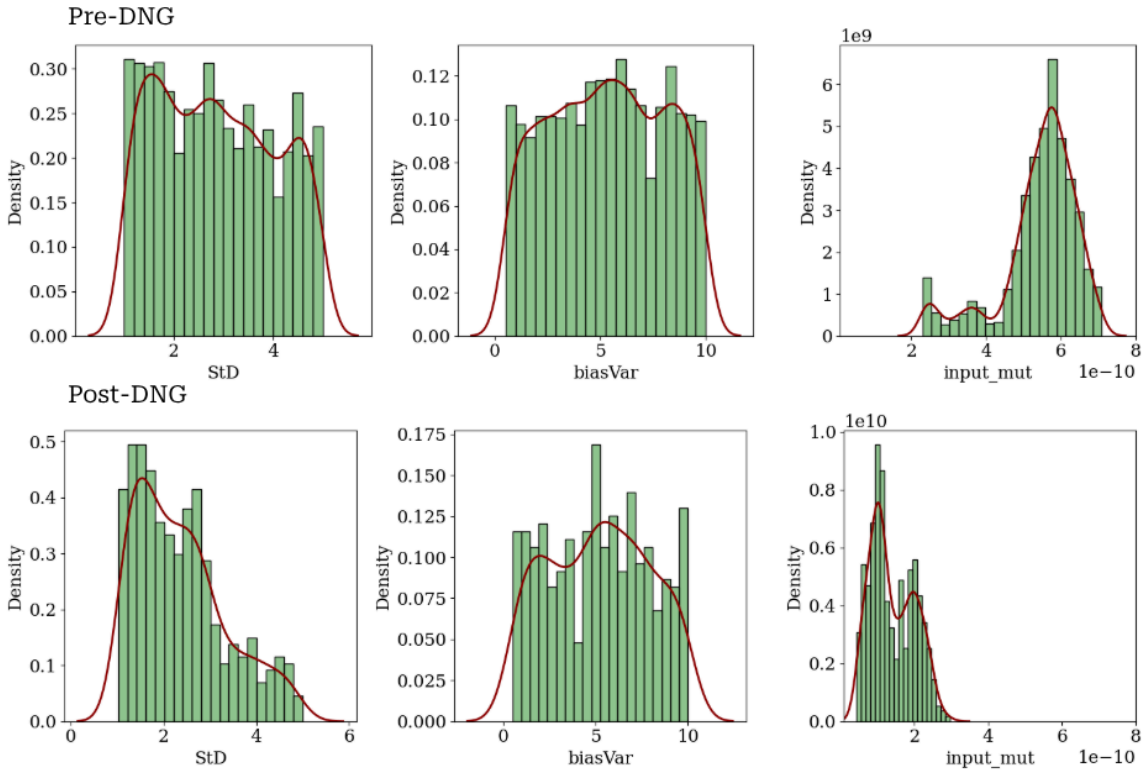


Figure 3.5 Posterior Distributions of Mutation Rate (input_mut), Elongation Parameter (StD), and Branching Parameter (biasVar) for *Eucalyptus melliodora*. Pre-dng (top) and post-dng (bottom) mutation data of *E. melliodora* were used as real data inputs from Orr et al. 2020. The green histograms represent the density distribution of accepted sampled parameter values, while the red line

represents the smoothed Kernel Density Estimation KDE curve, providing a continuous estimate of the probability density. Posterior distributions were generated using ArviZ (Kumar et al., 2019).

	Mean	Standard Deviation	95% HPD Low	95% HPD High	ESS- Bulk	ESS-Tail
StD	2.943	1.174	1.002	4.760	914	819
biasVar	5.307	2.681	0.639	9.442	842	741
input_mut	5.3E-10	1E-10	2.3E-10	7.1E-10	970	707
StD	2.36	0.980	1.033	4.26	342	320
biasVar	5.18	2.730	0.991	9.939	282	272
input_mut	1.0E-10	1E-10	4.2E-11	2.2E-10	263	283

Table 3.3 Summary Statistics for the Posterior Distributions of Mutation Rate (input_mut), Elongation Parameter (StD), and Branching Parameter (biasVar) for *Eucalyptus melliodora*. Pre-dng (highlighted green) and post-dng (highlighted blue) mutation data of *E. melliodora* were used as real data inputs from Orr et al. 2020. Summary statistics were generated using ArviZ (Kumar et al., 2019).

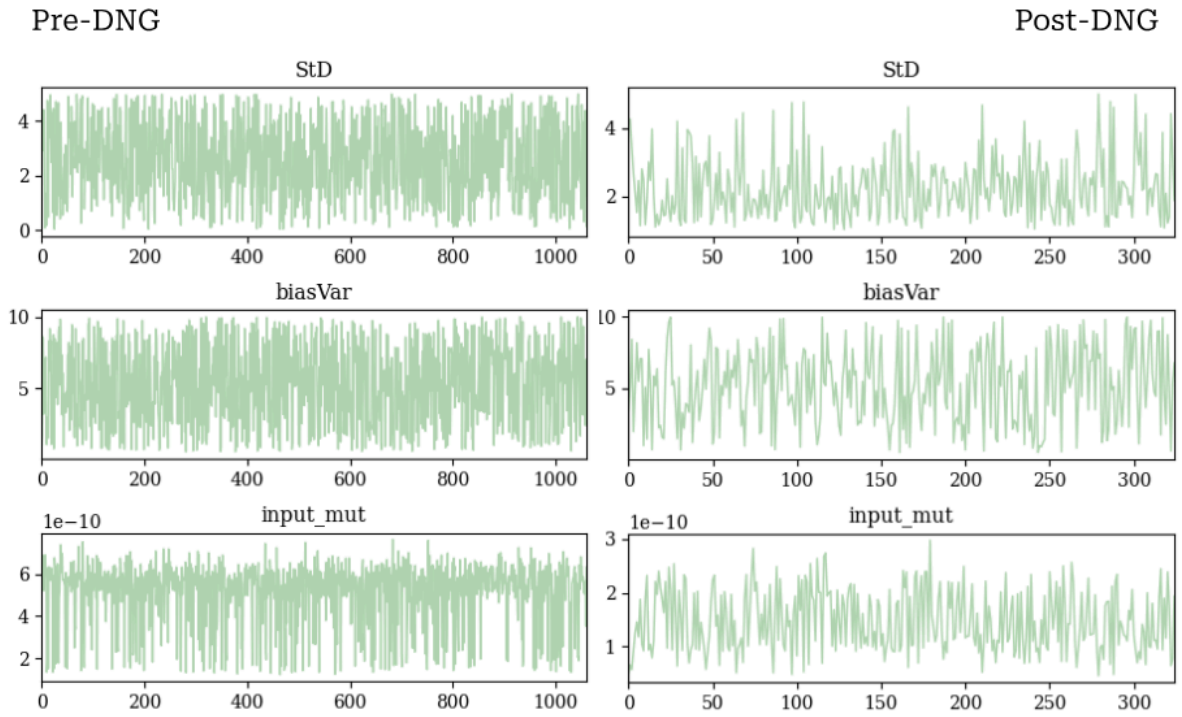


Figure 3.6 Trace Plots of Approximate Bayesian Computation (ABC) Sampling for Mutation Rate (input_mut), Elongation Parameter (StD), and Branching Bias Parameter (biasVar) for *Eucalyptus melliodora*. Pre-dng (left) and post-dng (right) mutation data of *E. melliodora* were used as real data inputs from Orr et al. 2020. Trace plots were generated using ArviZ (Kumar et al., 2019).

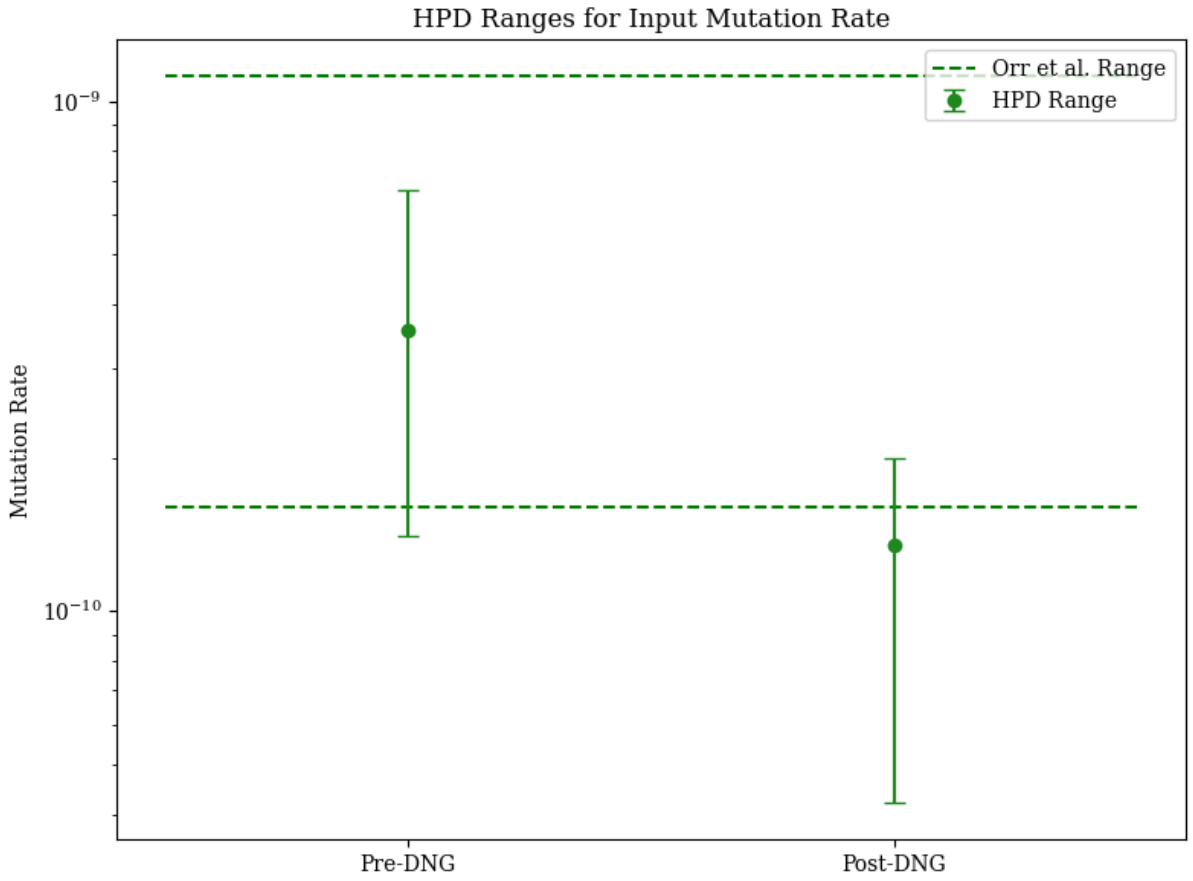


Figure 3.7 Comparison of 95% HPD Intervals for Mutation Rate Posteriors of Pre-Dng and Post-Dng Datasets Against the 'True' Reported Range of *E. melliodora* (Orr et al. 2020). The Post-Dng HPD range falls mostly outside of the reported range, indicating greater overfiltering due to the phylogenomic method than previously reported.

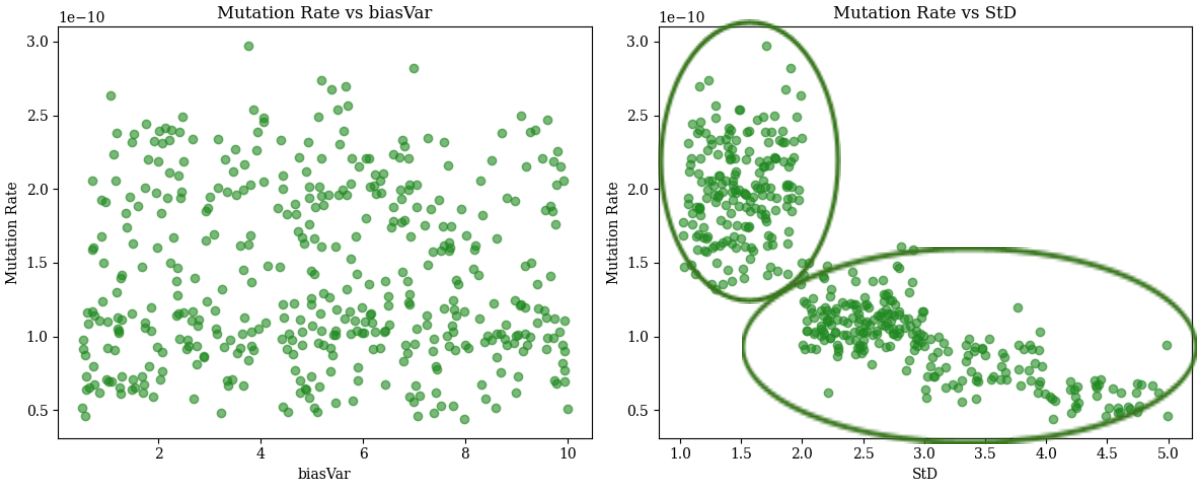


Figure 3.8 Scatter Plots of Posterior Mutation Rate against BiasVar (Branching Bias, Left) and StD (Elongation Parameter, Right) Values for the Post-Dng Dataset of *E. melliodora*.
The interaction effect between StD values and Mutation Rate likely correlates to the bimodal peaks observed in the mutation rate posterior distribution for post-dng.

3.3.3 Model Jumping

A notable discontinuity in mutation rate dynamics was observed when comparing the pre-dng dataset before and after excluding $StD = 0$ (fully structured elongation) from the prior distribution (Figure 3.9). When $StD = 0$ is included (left), mutation rates cluster tightly at the lower end of the distribution, with an abrupt shift in mutation rate as *StD* transitions from 0 to values between 1 and 5. This behaviour suggests that $StD = 0$ represents a structurally different model rather than a continuous extension of $StD > 0$ parameter space.

Unlike $StD = 1\ to\ 5$, where both elongation and branching influence mutation accumulation, $StD = 0$ eliminates elongation as a contributing factor, making only the branching bias parameter (biasVar) active. This effectively shifts the model into a lower-dimensional parameter space, causing a sharp transition in posterior values when moving from $StD = 0$ to $StD \geq 1$. Such abrupt changes are characteristic of model-jumping, a well-documented issue in Bayesian inference where different models occupy non-overlapping parameter spaces, leading to instability in posterior sampling (Hubin & Storvik, 2018).

After removing $StD = 0$ (right panel), mutation rate estimates become more smoothly distributed across $StD = 1\ to\ 5$, eliminating the artificial discontinuity.* This suggests that the observed model-jumping was due to a mismatch in parameter dimensionality rather than biological signal. In Bayesian inference, models with different dimensionalities typically require specialised sampling techniques, such as Reversible Jump MCMC (Green, 1995), which ABC does not inherently support. By restricting the prior to $StD = 1\ to\ 5$, a consistent parameter space is maintained preventing instability while preserving meaningful biological variation.

*A downward trend can be observed in mutation rate estimates across both scatter plots, before and after removing StD =0. This trend is likely linked to long-tailed deviations in the StD trace plot for pre-dng, suggesting occasional inefficiencies in posterior exploration, particularly at the extremes of the parameter space. Furthermore, the left tail of the posterior distribution contains small peaks corresponding to low mutation rates, reinforcing the possibility that poor mixing or ABC rejection artifacts contribute to this trend. No significant interaction between biasVar was observed in relation to this trend.

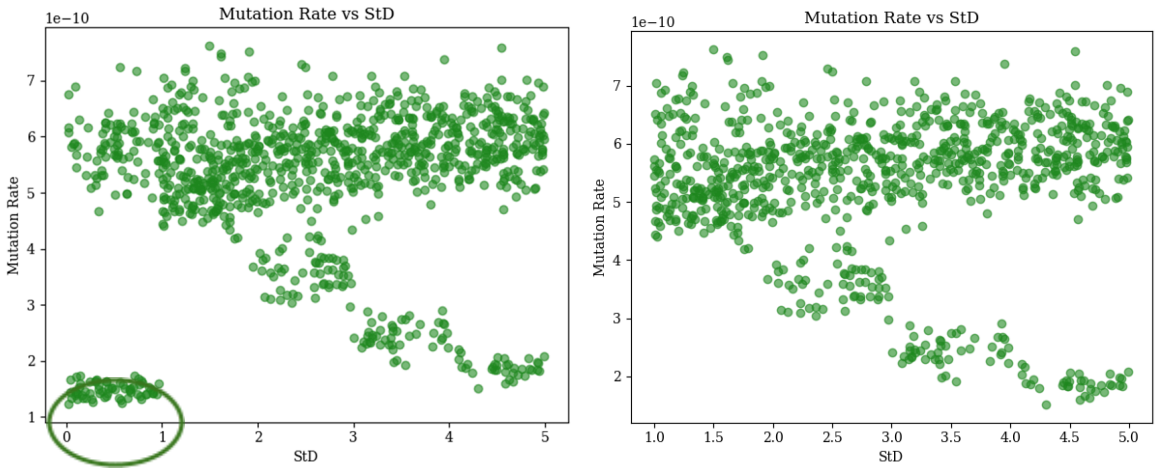


Figure 3.9 Scatter Plots Between Posterior Mutation Rate and StD (Elongation Parameter) Values for the Pre-Dng Dataset of *E. melliodora*. Plots occur before (left) and after (right), excluding StD = 0-1 from the prior distribution of StD due to model-jumping effects. A StD value of 0 is strictly structured, so the resulting prior distribution represents a spectrum of less stochastic (somewhat structured) to fully stochastic (StD = 5).

3.4 Discussion

In this chapter, I introduced and validated a prototype Approximate Bayesian Computation (ABC) framework designed to address limitations in traditional methods for somatic mutation rate estimation of long-lived trees. Validation across 169 simulated datasets demonstrated the framework's robustness, with all mutation rate and branching bias values, and 168 of the elongation parameter values successfully recovered within 95% HPD intervals. These results substantiate my novel ABC framework as a simulation alternative capable of approximating somatic mutation rates of long-lived trees, as well as biologically relevant parameters of meristem behaviour. In the following sections, I explore the theoretical, methodological, and practical implications researchers should consider regarding the ABC framework.

Model-Jumping

The observed discontinuity in mutation rate estimates between StD = 0 and StD = 1-5 raises concerns about the ABC-Reject framework's ability to reconcile distinct parameter regimes. This abrupt shift in posterior mutation rate values suggests a form of model jumping, where the algorithm struggles to accommodate biologically distinct parameter regions within the posterior space. This behaviour is analogous to challenges observed in other Bayesian frameworks, such as reversible jump MCMC, where transitions between models between different dimensions or parameterizations can result in poor exploration of the posterior space (Lee et al., 2014).

Highly structured elongation dynamics, as represented by StD = 0-1, are biologically plausible, particularly in systems where strict lineage maintenance dominates growth. While these dynamics differ markedly from those seen in StD = 1-5, their exclusion from the prior space sacrifices biological relevance for stability and computational convenience. Furthermore, the issue lies not with Tomimoto and Satake's (2023) models themselves- StD = 0 is both biologically and mathematically

valid- but rather with the ability of the current ABC framework to handle the reduction of dimensions that occurs when StD is set to 0.

To overcome this issue of model-jumping and ensure robust inference, future work could draw on strategies from advanced ABC methodologies and related Bayesian frameworks. One promising avenue is the use of Adaptive Sequential Monte Carlo (SMC) ABC, where sampled parameters are propagated through progressively stricter tolerance (epsilon, ϵ) thresholds (Toni, 2011). This gradual refinement could enable smoother transitions between distinct parameter regimes, potentially model-jumping in posterior recovery. Another approach involves reparameterization or the use of shared hyperpriors, which could embed StD = 0 and StD = 1-5 within a unified parameter space (Overcast et al., 2017). This structure could facilitate continuous exploration of biologically distinct elongation dynamics while avoiding model-jumping. Finally, integrating multinomial logistic regression into the calculation of summary statistics offers a better way to capture model indicators within the ABC framework. As demonstrated by Lee et al. (2014), this approach enables discrimination between biologically relevant models and could improve posterior stability. Future work should explore these strategies to refine the ABC-Reject framework to accommodate highly structured biological trees better.

Challenges in Estimating Structured and Stochastic Elongation

The shift towards lower StD values in the post-dng posterior suggests that structured elongation may provide a better fit for the mutations remaining after topology-based filtering in *E. melliodora*. In a completely structured meristem ($StD = 0$), mutations are inherited and preserved within strictly defined cell lineages, leading to more mutations being fixed. As meristems become more stochastic (approaching $StD = 5$), lineage preservation weakens, and mutations are randomly gained or lost across branches. One might initially expect that the post-dng dataset, which contains fewer overall mutations than the pre-dng dataset, would favour a stochastic elongation model, as stochastic meristems do not fix as many mutations. This was previously observed in Chapter 2, where stochastic elongation models produced lower NMRSE (Normalised Mean Root Square Error) values compared to structured models for the post-dng dataset- when simulations were run under a mutation rate of 6.18×10^{-10} per cell division per cell (the median value of Orr et al.'s (2020) estimated range for *E. melliodora*).

However, when mutation rates lower than that of Orr et al's predicted range were allowed in the expanded prior range (1E-11 to 9E-10), the structured model of elongation was able to compensate for its tendency to fix more mutations. In fact, structured elongation aligns more closely with the assumption that mutations strictly follow the topology, as strict lineage inheritance ensures that mutations remain within the defined topology rather than being randomly redistributed. As a result, the post-dng dataset favoured lower, more structured StD values, which better approximated the particular mutation distribution remaining after filtering with *E. melliodora's* topology.

A major limitation, however, is that the current ABC framework does not allow StD to reach 0, despite this potentially being the best fit for post-dng. As discussed in Section 3.3.3, StD = 0 introduced model-jumping artifacts due to a reduction in dimensionality, meaning it cannot be sampled continuously from the prior alongside other values using ABC-Reject. Future work could explore whether fixing StD at 0- rather than attempting to sample it- would result in an even lower posterior for the mutation rate of the post-dng dataset of *E. melliodora*.

The Use of Low Mutation Rates in Tomimoto and Satake's (2023) Simulation Code

Testing revealed that Tomimoto & Satake's (2023) simulation code struggles to capture extremely low mutation rates (<9.0E-12) due to the way mutations are assigned probabilistically. The model determines mutation occurrence by drawing a random number between 0 and 1 and comparing this generated number to the mutation rate itself. If the random number is lower than the mutation rate, a mutation is introduced; if it is higher, no mutation occurs. When mutation rates fall below 9E-12, they are so small that the random number is almost always higher, meaning that mutations are almost never assigned, effectively leading to zero mutations and disrupting downstream functions.

While Tomimoto and Satake's simulation code functions properly under a plausible range of mutation rates for long-lived trees (E-9 to E-11), it likely fails to capture estimates below this range. This could also contribute to the observed compression of lower mutation rates in the post-dng posterior, meaning that a left-tail or lower HPD boundary could occur, but not have been explored due to the breakdown of the simulation framework at such low mutation rates. Future refinements should focus on improving the mutation assignment strategies to better accommodate ultra-low mutation rates, ensuring that values below 9.0E-12 can still be effectively sampled.

Bimodality in the Post-Dng Mutation Rate Posterior

A surprising feature of the post-dng mutation rate posterior was its bimodal shape, with one peak at approximately 2E-10 and another around 1E-10/9E-11. Unlike the pre-dng dataset, which exhibited a smooth, unimodal distribution, the post-dng dataset retained two distinct clusters of values. Bimodal posterior distribution in a Bayesian framework indicates the presence of two competing parameter regions with substantial posterior probability (Firat & Thompson, 1996; Liu & Hodges, 2003). This suggests that the data does not support a single dominant mutation rate but rather two plausible models, potentially arising from different biological processes or filtering effects.

One likely explanation is that elongation mode (StD) strongly influences mutation retention, creating a structured relationship between mutation rate and StD in post-dng. While pre-dng mutation rates were more evenly distributed, post-dng mutations were subject to topology-based filtering, which preferentially retained mutations that aligned with the tree's structure. This process appears to have reinforced the relationship between mutation rate and StD, creating distinct parameter regimes instead of a continuous gradient (Figure 3.8). This effect may have been amplified if $StD = 0$ was included in the prior. A fully structured meristem ($StD = 0$) would retain all generated mutations across strict lineages, strengthening the structured elongation cluster and potentially exaggerating the bimodal separation. The exclusion of $StD = 0$ from the prior may obscure the full extent of this effect, leaving only the observed clustering at low StD values (1-2). Additionally, Tomimoto and Satake's (2023) mutation assignment model may have amplified this pattern by artificially compressing the lower tail of the posterior. Since mutations are assigned probabilistically- occurring only when a randomly drawn number between 0 and 1 is below the mutation rate- mutation rates below 9E-12 rarely generate any mutations, effectively removing ultra-low values from the posterior. This could force more accepted values into the higher peak (~2E-10), further reinforcing the bimodal structure. If the model was adjusted to allow proper sampling of ultra-low mutation rates, one might observe a smoother posterior distribution with a more continuous relationship between StD and mutation rate.

Together, these findings suggest that the bimodal posterior is the result of topology filtering reinforcing elongation-driven mutation retention, further intensified by prior constraints on StD and the limitations of Tomimoto and Satake's mutation assignment process. Future work should explore

whether expanding the StD prior to include 0 and improving mutation rate assignment for ultra-low rates can mitigate these effects and reveal a more accurate representation of mutation dynamics.

Application to Real Data and Practical Considerations

A key feature of the ABC framework is the ability to approximate posterior distributions directly from the observed mutation data in a given tree without relying on topological filtering or scaling assumptions. This distinct feature is particularly evident when compared to Orr et al. 's methodology for *Eucalyptus melliodora*. Orr et al. scaled the 90 high-variants remaining after topological filtering to a final estimate of 300 'true mutations' through a series of transformations, including recovery rate adjustments, branch/tree length scaling, and age assumptions. While these adjustments attempt to compensate for over-filtering, they do so at the cost of obscuring the true extent of genetic signal loss caused by stringent criteria, incorporating no additional knowledge of the mutation distribution across branches. As a result, the final somatic mutation rate range reported by Orr et al. (1.16E-10 to 1.12E-09) was significantly higher than the 95% HPD post-dng posterior approximated via ABC (4.20E-11 to 2.2E-11). This shift indicates that the phylogenomic method of mutation rate estimation not only overfilters mutations across topology but does so to a degree far exceeding Orr et al.'s reported range.

The success of the alternative ABC framework is highly dependent on the observed data, topological information, and priors utilised.

The near-perfect recovery of true values across validation datasets determines the ABC's framework's capability of posterior approximation when true values are known. Even the worst validation sample, in which StD falls outside the 95% HPD by 0.02 points, produced stronger posteriors than both the pre-dng and post-dng real data applications. Validation posteriors exhibited clear peaks and tighter credible intervals, particularly for meristem behaviour parameters (StD and biasVar), despite having far fewer accepted samples (100) in comparison to pre- and post-dng posteriors (1020 and 350, respectively). In contrast, applying the ABC framework to *E. melliodora* observed data failed to produce meaningful posteriors for meristem behaviour parameters, with flat distributions, no discernible peaks, and 95% HPD intervals that essentially mirrored prior ranges. The exception was the post-dng StD posterior, which demonstrated a skew to lower StD values, suggesting structured elongation behaviour resulting from stringent topological filtering. These results highlight how the quality of inputted data critically influences the ability to approximate meaningful posteriors.

The quality of the inputted topology and estimated age for a given tree should also be considered alongside the distribution of mutations across branches (the observed data, D). In my ABC framework, the branching architecture (topology) of the input tree is required as input with branch lengths given in years, alongside a total age estimate. While measuring the topology physically is straightforward, measuring the age of a biological tree and each of its branches/nodes is more difficult. Accurately converting physical branch length into years is essential when applying the framework to real organisms. This step, in combination with challenges introduced by filtering, likely contributed to the limited results observed for *E. melliodora*. Refining mutation detection and filtering strategies and improving the conversion of physical branch lengths into tree age will be essential for recovering biologically meaningful posteriors. Both processes are interdependent in this framework and influence the ability to approximate parameters such as StD and biasVar, where meaningful signals may exist but remain obscured by poor input data. By addressing both challenges, researchers can fully leverage the potential of the ABC framework- generating biologically relevant insights.

In addition to data quality, posterior inference is strongly influenced by priors. For example, in the post-dng analysis, the initially restrictive prior resulted in accepted samples clustering near the lower boundary, limiting posterior exploration. To address this, I conducted a rudimentary sensitivity analysis based on accepted sample numbers to total trials, posterior shapes, and credible intervals, which helped select an alternative prior range. This process improved the posterior distribution and mitigated boundary effects. More advanced approaches, such as the use of perturbation functions to evaluate the impact of prior modifications on posterior recovery, could formalise and extend this analysis. As discussed by Weiss (1996), influence diagnostics offer a method to quantify how individual data points or prior settings disproportionately affect posterior distributions, providing opportunities for targeted refinement. Iterative adjustments to kernel scale parameters and acceptance thresholds, as highlighted by Csilléry et al. (2010), could also enhance alignment between observed and simulated data, ensuring robust posterior recovery. By incorporating these advanced diagnostics and sensitivity analysis techniques, the framework could achieve systemic improvements in posterior inference, particularly in complex real-world applications where prior choice disproportionately affects approximation.

Conclusion

This chapter presents a prototype Approximate Bayesian Computation (ABC) framework for estimating the somatic mutation rates in long-lived trees, offering a proof of concept for an alternative approach to other methods. While the framework successfully recovers mutation rates and key developmental parameters in validation datasets, its application to real data highlights important caveats. The analysis reveals that model-jumping artifacts, the inability to sample StD = 0, the limitation of Tomimoto and Satake's mutation assignment model, and topology-based filtering all influence posterior recovery. However, the strong performance of the ABC framework in simulation, where 169 datasets confirmed reliable posterior recovery of mutation rate and branching bias parameters, demonstrate its potential as a powerful inference tool. With further refinement- such as improved mutation assignment for ultra-low rates, adjustments to prior parameterisation, and strategies to mitigate model jumping- this framework could become a robust and flexible approach for estimating somatic mutation rates across diverse plant systems.

Chapter 4: General Discussions

This thesis introduces a novel, non-phylogenetic framework for estimating somatic mutation rates in long-lived trees, leveraging Tomimoto and Satake's (2023) models of somatic mutation accumulation alongside an ABC-Reject methodology. Initially, my goal was to critically evaluate the phylogenomic method outlined by Orr et al. (2020), testing its performance across a range of tree architectures, meristem behaviors, and mutation rate conditions through simulation. My findings revealed critical limitations, particularly the method's dependence on topological congruence, which resulted in the exclusion of biologically relevant mutations that did not adhere strictly to tree structure. In response, I developed a prototype ABC-Reject framework that circumvents these constraints by avoiding rigid phylogenetic assumptions. Instead, it generates an approximate posterior distribution of somatic mutation rates based on observed mutation frequencies, while simultaneously estimating developmental parameters related to meristem behavior.

4.1 Evaluating the Phylogenomic Method: Strengths, Limitations, and Best Practices

The phylogenomic method introduced by Orr et al. (2020) provides a structured approach for estimating somatic mutation rates, but its effectiveness depends on the biological context. Simulations using Tomimoto and Satake's (2023) models show that unbalanced long-terminal topologies perform best at low mutation rates, as they minimize shared mutations and preserve distinct genetic signals. In contrast, balanced long-terminal topologies underpredict mutation rates due to dilution effects, while unbalanced short-terminal topologies overestimate rates due to uneven mutation distribution. These findings highlight the phylogenomic method's reliance on strong, distinct mutational signals in terminal branches for accurate mutation rate recovery.

Beyond mutation rate recovery, topology recovery also varied significantly across tree structures. Balanced topologies consistently yielded lower Robinson-Foulds (RF) distances, indicating stronger phylogenetic recovery, while unbalanced topologies, despite sometimes recovering mutation rates well, had higher RF distances due to uneven mutation distributions. Subsampling tests using the empirical *Eucalyptus melliodora* topology reinforced these findings—a four-branch balanced subsample achieved perfect recovery (RF distance = 0), while the original eight-branch unbalanced topology resulted in misplaced branches and novel groupings, increasing both RF distance and mutation rate deviations. These results align with the critiques of Orr et al.'s (2020) approach raised by Iwasa et al. (2023) and others, who argue that mutation accumulation often deviates from strict tree topology, especially in complex modular architectures. My findings support this criticism, showing that more complex topologies reduce the method's ability to reconstruct the true branching structure, further limiting its utility in diverse tree forms.

At higher mutation rates, performance declines across all topologies. As mutations accumulate more uniformly across branches, genetic distinctiveness is lost, leading to severe underestimation of mutation rates. While I initially suspected that back mutations could be a contributing factor, my results confirmed that back mutations are negligible. Instead, the major issue is mutation homogeneity—too many shared mutations disrupt the genetic-physical distance relationship, making the method ineffective. Based on this, I strongly recommend applying the phylogenomic method only under low mutation rates (below 1×10^{-10} per site per year), where genetic signals remain distinct enough for reliable inference.

The analysis of empirical *Eucalyptus melliodora* data reinforced these concerns. Before topological filtering, the mutation distribution closely resembled what would be expected under stochastic mutation accumulation. However, after filtering, mutation rates were systematically underestimated, particularly in branches 5 and 6 which are connected by a long internal node. This suggests that the phylogenomic method over-filters mutations that do not conform to strict topological expectations, even when they are biologically meaningful.

Further confirmation came from the Approximate Bayesian Computation (ABC) approach, which estimated a much lower mutation rate after filtering than the range reported by Orr et al. (2020). Orr et al. attempted to adjust for lost mutations by scaling up the remaining 90 high-confidence variants of *E. melliodora*, resulting in a final estimate (1.16E-10 to 1.12E-09) that was significantly higher than the ABC-derived HPD interval (4.20E-11 to 2.2E-11), which accounted for these variants directly. While previous critiques, including those by Tomimoto and Satake (2023) and Iwasa et al. (2023),

argued that the phylogenomic method over-filters mutations, my results demonstrate that the extent of this over-filtering is even more severe than previously assumed. Rather than simply refining mutation estimates, the method systematically eliminates a substantial proportion of biologically meaningful mutations, and Orr et al.'s subsequent transformations appear to mask, rather than correct, this loss. This raises significant concerns about the validity of their final mutation rate estimates and broader conclusions regarding somatic mutation accumulation in long-lived trees.

While my findings provide key recommendations, some limitations of conclusions reached must be acknowledged. First, the conversion of branch lengths to ages relied on a growth rate of 10 cm per year, based on *E. grandis*. While this provided a reasonable approximation, species-specific growth variability could introduce uncertainty in the estimates. Second, the modified phylogenomic method represented an idealized scenario with perfect data, free from sequencing errors. In real-world applications, additional error correction would be necessary, and results might differ in more complex datasets. Lastly, my conclusions depend on the accuracy of Tomimoto and Satake's (2023) model for simulating somatic mutation accumulation. While their framework provides a strong theoretical basis, empirical validation remains limited.

4.2 ABC Prototype: A Flexible Approach for Estimating Somatic Mutation Rates

The Approximate Bayesian Computation (ABC-Reject) framework developed in this thesis provides a flexible alternative for estimating somatic mutation rates in long-lived trees, free from the strict topological assumptions of phylogenomic methods. To apply this approach to a real tree, researchers must define branch ages and architecture, using physical branch lengths to approximate time and inputting the observed distribution of unique mutations across terminal branches as empirical data. The ABC-Reject method samples mutation rates from biologically informed priors, simulates mutation distributions across the input topology, and compares observed and simulated datasets using Euclidean distance. Accepted parameter sets generate a posterior distribution of mutation rate, elongation (StD), and branching bias (biasVar), allowing for inference of both somatic mutation dynamics and meristematic behavior.

To validate the method, it was tested against simulated datasets with known mutation rates and developmental parameters, ensuring that validation was independent of empirical sequencing errors. Using 200 validation samples, the ABC framework demonstrated near-perfect accuracy, with 100% of true mutation rates and biasVar values and 99.4% of true StD values recovered within the 95% highest posterior density (HPD) interval. These results confirm that the ABC-Reject method reliably estimates somatic mutation rates and developmental traits across a range of tree architectures.

The framework was then applied to empirical somatic mutation data from *Eucalyptus melliodora*, using two datasets from Orr et al. (2020): the pre-DNG dataset (330 high-confidence variants before phylogenomic filtering) and the post-DNG dataset (90 variants remaining after filtering). Prior distributions for mutation rate, StD, and biasVar were iteratively refined based on early trial runs, ensuring posterior exploration was not constrained by overly restrictive priors. The mutation rate prior was expanded beyond Orr et al.'s estimates to accommodate lower observed values in the post-DNG dataset, while the elongation parameter (StD) prior was adjusted to exclude StD = 0, preventing the model from collapsing into a lower-dimensional space. The acceptance threshold ($\varepsilon = 20$) was optimized through sensitivity analysis, balancing posterior accuracy with computational efficiency.

The results revealed clear shifts between pre- and post-DNG datasets. The mutation rate posterior was unimodal for pre-DNG data, aligning with prior estimates, but bimodal and significantly lower for post-DNG data, reinforcing that phylogenomic filtering systematically reduces mutation rate estimates. This bimodality likely arises because elongation mode (StD) influences mutation retention, reinforcing a structured relationship between mutation rate and StD in the post-DNG dataset. While pre-DNG mutation rates were more evenly distributed, topological filtering preferentially retained mutations that aligned with tree structure, creating distinct parameter regimes rather than a continuous gradient in the posterior distribution.The elongation parameter (StD) posterior for post-DNG data was more skewed, suggesting a shift toward structured elongation in filtered data, while branching bias (biasVar) showed no significant influence in either dataset. These findings demonstrate that ABC provides a more flexible and biologically realistic estimation of somatic mutation rates, preserving stochastic variation that would otherwise be lost under phylogenomic filtering.

By applying this framework to real tree systems, researchers can infer mutation rates and meristematic behavior directly from empirical mutation distributions, providing a data-driven alternative to traditional phylogenomic approaches. The implications of these findings, including potential limitations of the model, are further explored in the next section.

4.3 Limitations of the ABC Prototype Method

While the ABC-Reject framework provides a flexible alternative for estimating somatic mutation rates, several limitations must be acknowledged. One key issue is the observed discontinuity in mutation rate estimates between StD = 0 and StD = 1–5, raising concerns about the framework's ability to reconcile distinct parameter regimes. The abrupt shift in posterior mutation rate values suggests a form of model-jumping, where the algorithm struggles to accommodate biologically distinct parameter spaces. While StD = 0 represents a highly structured elongation regime that may best fit the post-DNG dataset, the current ABC implementation does not allow StD to reach 0. As discussed in Section 3.3.3, attempts to incorporate StD = 0 led to model-jumping artifacts due to a reduction in dimensionality, making it incompatible with continuous sampling in ABC-Reject. This limitation restricts one's ability to fully explore the potential for highly structured elongation modes in post-filtered datasets, potentially biasing posterior estimates.

Additionally, low mutation rates pose a challenge for Tomimoto & Satake's (2023) simulation framework. The model probabilistically assigns mutations by comparing a randomly generated number (between 0 and 1) to the mutation rate itself, introducing a mutation only when the random number is lower than the mutation rate. At extremely low mutation rates (<9.0E-12), this threshold becomes so restrictive that mutations are almost never assigned, effectively resulting in zero mutations and disrupting downstream functions. While the model functions properly across plausible somatic mutation rates for long-lived trees (E-9 to E-11), its inability to capture values below this range may have compressed the lower tail of the post-DNG posterior, preventing full exploration of mutation rate estimates at the lower boundary.

Another unresolved issue is the lack of significance for branching bias (biasVar) in all posterior distributions. When applied to empirical *E. melliodora* data, biasVar posteriors remained flat, displaying no discernible peaks, with 95% HPD intervals closely mirroring the prior distribution. This result is consistent with the simulations in Chapter 2, where branching bias had negligible effects on mutation rate recovery and phylogenetic accuracy. While this could indicate that axillary meristem

formation has minimal influence on somatic mutation accumulation, it is unclear whether the simulation framework correctly models the biological impact of branching bias. Tomimoto and Satake (2023) also observed a similar trend in their own empirical applications, finding that elongation, rather than branching, played the dominant role in mutation accumulation. Given this, it remains uncertain whether biasVar is truly biologically uninformative or whether its effects are obscured by the limitations of the current modeling approach.

Beyond issues with parameter estimation, the quality of observed data, topological accuracy, and priors all strongly influence posterior inference. The success of the ABC framework depends on accurate input data, particularly in defining the observed mutation distribution (D) and estimating branch age. While tree topology can be measured physically with relative ease, assigning chronological ages to branches is more difficult, particularly when applying the framework to real trees. In this thesis, tree age was converted from physical length using a generalized growth rate of 10 cm per year, but this assumption introduces uncertainty, as growth rates vary due to environmental, species-specific, and life-history factors. Accurately converting branch lengths to chronological time is critical for applying this method to real organisms and likely contributed to the limited resolution observed in the *E. melliodora* results.

Posterior inference is also highly sensitive to prior distributions, particularly in the post-DNG dataset. The initially restrictive mutation rate prior constrained posterior exploration, causing accepted samples to cluster at the lower boundary. To mitigate this, I conducted a rudimentary sensitivity analysis, adjusting prior distributions based on the ratio of accepted samples to total trials, posterior shapes, and credible intervals. While these refinements improved posterior exploration and reduced boundary effects, this approach was iterative rather than formally parameterized, meaning that prior selection remains subjective and requires further optimization.

Finally, a fundamental limitation of the ABC-Reject framework itself is its computational inefficiency. Unlike ABC methods that incorporate iterative refinement, such as Sequential Monte Carlo (ABC-SMC), the rejection algorithm requires a large number of simulations, discarding most sampled parameter sets that fall outside the tolerance threshold. This makes parameter estimation computationally expensive and time-intensive, particularly as the number of parameters increases.

4.4 Future Work

This study highlights key areas where both the phylogenomic method and the ABC-Reject framework could be refined to improve the accuracy and applicability of somatic mutation rate estimation in long-lived trees. Future work should focus on addressing the limitations identified in this study, refining existing methodologies, and expanding their application to a broader range of biological systems.

For the phylogenomic method, further simulations are needed to precisely determine the topological and mutational conditions under which the method remains reliable. My findings suggest that trees with extensive shared internal nodes and uneven mutation distributions hinder mutation rate recovery, but additional trials using a greater diversity of simulated topologies could better define the boundaries of its effectiveness. More critically, applying the phylogenomic method to additional empirical datasets would help assess whether the trends observed in simulations are consistent in real biological systems. Expanding the scope of real-world testing is essential for determining whether the method's limitations are generalizable across tree taxa or specific to certain growth architectures.

Refinements to the ABC-Reject framework should focus on improving parameter inference and computational efficiency. Model-jumping issues, particularly for StD = 0 versus StD > 0, could be mitigated by adopting Adaptive Sequential Monte Carlo (ABC-SMC) methodologies, where progressively stricter tolerance thresholds (ε) improve posterior refinement (Toni et al., 2011). This would allow smoother transitions between biologically distinct parameter spaces, preventing abrupt posterior shifts. Alternative solutions include reparameterization using shared hyperpriors, which could integrate StD = 0 within a continuous parameter space, avoiding the sharp discontinuities observed in the ABC-Reject framework (Overcast et al., 2017). Further, incorporating multinomial logistic regression into summary statistic selection (Lee et al., 2014) could provide a more robust means of distinguishing between elongation modes, improving posterior stability. Future trials should also explore whether fixing StD at 0—rather than attempting to sample it—results in even lower posterior mutation rate estimates in the post-DNG dataset of *E. melliodora*.

To improve inference at ultra-low mutation rates, refinements to mutation assignment strategies in Tomimoto & Satake's (2023) simulation framework are necessary. The current probabilistic approach struggles to assign mutations at rates below 9.0E-12, effectively excluding extremely low mutation rates from the posterior space. Addressing this issue would allow full posterior exploration at the lower boundary, ensuring that low-mutation-rate biological systems can be accurately modeled.

Another area for refinement is how developmental parameters, particularly branching bias (biasVar), are modeled within the framework. While my results suggest that branching bias has a limited role in shaping mutation distributions, it remains unclear whether this reflects a true biological trend or a limitation of the simulation framework. Tomimoto & Satake (2023) also found that elongation -not branching- played a dominant role in mutation accumulation, but future research should explore alternative mathematical representations of axillary meristem behavior to determine whether bias in axillary meristem lineage selection has greater biological relevance than my findings suggest. Further simulation studies may also reveal whether alternative tree growth models yield different results regarding the impact of branching patterns on somatic mutation accumulation.

More rigorous sensitivity analysis techniques should also be implemented to ensure that prior distributions and acceptance thresholds do not artificially constrain posterior exploration. While this study relied on an iterative approach based on accepted sample ratios and posterior shapes, more systematic methods, such as perturbation functions and influence diagnostics (Weiss, 1996; Csilléry et al., 2010), could provide quantitative assessments of prior impact on posterior inference. Additionally, adaptive distance metrics, which assign variable weights to summary statistics based on their contribution to posterior accuracy, could be tested to improve ABC-based inference robustness (Liu & Niranjan, 2021).

A major computational limitation of ABC-Reject is its inefficiency in high-dimensional parameter spaces, as it requires large numbers of simulations with most parameter sets rejected. Future work should explore alternative ABC methodologies, particularly ABC-SMC, which iteratively refines posterior distributions across multiple generations, reducing computational costs while maintaining accuracy (Beaumont et al., 2009; Toni et al., 2009). This adaptive thresholding approach improves efficiency by focusing sampling on regions of high posterior probability, making it better suited for large datasets and complex parameter spaces (Marin et al., 2012). Trialing ABC-SMC on the same dataset would be a logical next step to assess whether computational improvements translate to more accurate biological inference.

Finally, refining mutation detection and filtering strategies and improving tree age estimation techniques will be essential for ensuring biologically meaningful posterior recovery. The ability to accurately convert physical branch length into chronological age remains a major challenge in applying this framework to real trees, and any refinements to this process would directly improve mutation rate inference. Combining improvements in data quality with methodological refinements in ABC inference will be key to fully realizing the potential of this framework for estimating somatic mutation rates in long-lived trees.

References

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L. Y., Fonnesbeck, C., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T. V., & Zinkov, R. (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ*, *9*, e1516–e1516. https://doi.org/10.7717/peerj-cs.1516

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., Heisler, L. E., Beck, T. A., Simpson, J. T., Tonon, L., Sertier, A.-S., Patch, A.-M., Jäger, N., Ginsbach, P., Drews, R., Paramasivam, N., Kabbe, R., Chotewutmontri, S., Diessl, N., & Previti, C. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*(1). https://doi.org/10.1038/ncomms10001

Baker, P. J. (2003). Tree Age Estimation For the Tropics: A Test from the Southern Appalachians. *Ecological Applications*, *13*(6), 1718–1732. https://doi.org/10.1890/02-5025

Ban, S., & Jung, J.-H. (2023). Somatic Mutations in Fruit Trees: Causes, Detection Methods, and Molecular Mechanisms. *Plants*, *12*(6), 1316–1316. https://doi.org/10.3390/plants12061316

Bandelt, H.-J., Kong, Q.-P., Richards, M., & Macaulay, V. (2006). Estimation of Mutation Rates and Coalescence Times: Some Caveats. *Nucleic Acids and Molecular Biology*, *18*, 47–90. https://doi.org/10.1007/3-540-31789-9_4

Barton, M. K. (2010). Twenty years on: The inner workings of the shoot apical meristem, a developmental dynamo. *Developmental Biology*, *341*(1), 95–113. https://doi.org/10.1016/j.ydbio.2009.11.029

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. https://doi.org/10.1098/rstl.1763.0053

Beaumont, M. A., Cornuet, J.-M. ., Marin, J.-M. ., & Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, *96*(4), 983–990. https://doi.org/10.1093/biomet/asp052

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, *162*(4), 2025–2035. https://doi.org/10.1093/genetics/162.4.2025

Bobiwash, K., Schultz, S. T., & Schoen, D. J. (2013). Somatic deleterious mutation rate in a woody plant: estimation from phenotypic data. *Heredity*, *111*(4), 338–344. https://doi.org/10.1038/hdy.2013.57

Böcker, S., Canzar, S., & Klau, G. W. (2013). The Generalized Robinson-Foulds Metric. *Lecture Notes in Computer Science*, 156–169. https://doi.org/10.1007/978-3-642-40453-5_13

Boyko, A., Zemp, F., Filkowski, J., & Kovalchuk, I. (2006). Double-Strand Break Repair in Plants Is Developmentally Regulated. *Plant Physiology*, *141*(2), 488–497. https://doi.org/10.1104/pp.105.074658

Brocchieri, L. (2001). Phylogenetic Inferences from Molecular Sequences: Review and Critique. *Theoretical Population Biology*, *59*(1), 27–40. https://doi.org/10.1006/tpbi.2000.1485

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *47*(1), 69–100. https://doi.org/10.1111/1467-9884.00117

Brown, C. E. (1998). Coefficient of Variation. *Applied Multivariate Statistics in Geohydrology and Related Sciences*, 155–157. https://doi.org/10.1007/978-3-642-80328-4_13

Brunner, H. (1995). Radiation induced mutations for plant selection. *Applied Radiation and Isotopes*, *46*(6-7), 589–594. https://doi.org/10.1016/0969-8043(95)00096-8

Bruno, W. J., & Halpern, A. L. (1999). Topological bias and inconsistency of maximum likelihood using wrong models. *Molecular Biology and Evolution*, *16*(4), 564-566.

Burian, A., Barbier de Reuille, P., & Kuhlemeier, C. (2016). Patterns of Stem Cell Divisions Contribute to Plant Longevity. *Current Biology*, *26*(11), 1385–1394. https://doi.org/10.1016/j.cub.2016.03.067

Cabrera-Ponce, J. L., Valencia-Lozano, E., & Trejo-Saavedra, D. L. (2019). Genetic Modifications of Corn. *Corn*, 43–85. https://doi.org/10.1016/b978-0-12-811971-6.00003-6

Chen, Y., Burian, A., & Johannes, F. (2024). Somatic epigenetic drift during shoot branching: a cell lineage-based model. *Genetics*, *227*(4). https://doi.org/10.1093/genetics/iyae091

Colan, S. D. (2013). The Why and How of Z Scores. *Journal of the American Society of Echocardiography*, *26*(1), 38–40. https://doi.org/10.1016/j.echo.2012.11.005

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., Fenton, J., Fenzl, T., Fish, J., Garcia-Gutierrez, L., Garland, P., Gluck, J., González, I., Guermond, S., Guo, J., & Gupta, A. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research*, *4*, 900. https://doi.org/10.12688/f1000research.6924.1

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Darrah, L. L., McMullen, M. D., & Zuber, M. S. (2019, January 1). *Chapter 2 - Breeding, Genetics and Seed Corn Production* (S. O. Serna-Saldivar, Ed.). ScienceDirect; AACC International Press. https://www.sciencedirect.com/science/article/abs/pii/B9780128119716000024

Davis-Stober, C. P., Morey, R. D., Gretton, M., & Heathcote, A. (2016). Bayes factors for state-trace analysis. *Journal of Mathematical Psychology*, *72*, 116–129. https://doi.org/10.1016/j.jmp.2015.08.004

De Finetti, B (1961). The Bayesian approach to the rejection of outliers. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability 1,* Berkeley, CA, USA: University of California Press.

Dermen, H. (1969). Directional Cell Division in Shoot Apices. *CYTOLOGIA*, *34*(4), 541–558. https://doi.org/10.1508/cytologia.34.541

Domagalska, M. A., & Leyser, O. (2011). Signal integration in the control of shoot branching. *Nature Reviews Molecular Cell Biology*, *12*(4), 211–221. https://doi.org/10.1038/nrm3088

Drovandi, C. C., & Pettitt, A. N. (2013). Bayesian Experimental Design for Models with Intractable Likelihoods. *Biometrics*, *69*(4), 937–948. https://doi.org/10.1111/biom.12081

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214. https://doi.org/10.1186/1471-2148-7-214

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Efron, B. (2013). Bayes' Theorem in the 21st Century. *Science*, *340*(6137), 1177–1178. https://doi.org/10.1126/science.1236536

Emerson, R. A. (1913). The possible origin of mutations in somatic cells. *The American Naturalist*, *47*(558), 375-377.

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 419–474. https://doi.org/10.1111/j.1467-9868.2011.01010.x

Feng, Y., Comes, H. P., Chen, J., Zhu, S., Lu, R., Zhang, X., Li, P., Qiu, J., Olsen, K. M., & Qiu, Y. (2023). Genome sequences and population genomics provide insights into the demographic history, inbreeding, and mutation load of two "living fossil" tree species of Dipteronia. *The Plant Journal*, *117*(1), 177–192. https://doi.org/10.1111/tpj.16486

Fiorillo, C. D. (2012). Beyond Bayes: On the Need for a Unified and Jaynesian Definition of Probability and Information within Neuroscience. *Information*, *3*(2), 175–203. https://doi.org/10.3390/info3020175

Firat, M. Z., & Thompson, R. (1996). Investigation of bimodality in likelihoods and posterior densities. *Journal of Statistical Computation and Simulation*, *54*(4), 379–386. https://doi.org/10.1080/00949659608811741

Forrester, D. I., Medhurst, J. L., Wood, M., Beadle, C. L., & Valencia, J. C. (2010). Growth and physiological responses to silviculture for producing solid-wood products from Eucalyptus plantations: An Australian perspective. *Forest Ecology and Management*, *259*(9), 1819–1835. https://doi.org/10.1016/j.foreco.2009.08.029

Frank, M. H., & Chitwood, D. H. (2016). Plant chimeras: The good, the bad, and the "Bizzaria." *Developmental Biology*, *419*(1), 41–53. https://doi.org/10.1016/j.ydbio.2016.07.003

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series a (Statistics in Society)*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gascuel, O. (2006). Neighbor-Joining Revealed. *Molecular Biology and Evolution*, *23*(11), 1997–2000. https://doi.org/10.1093/molbev/msl072

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*(3), 445–449. https://doi.org/10.1214/08-ba318

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721-741.

Gill, D. E., Chao, L., Perkins, S. L., & Wolf, J. B. (1995). Genetic Mosaicism in Plants and Clonal Animals. *Annual Review of Ecology and Systematics*, *26*, 423–444. http://www.jstor.org/stable/2097214

Golubov, A., Yao, Y., Maheshwari, P., Bilichak, A., Boyko, A., Belzile, F., & Kovalchuk, I. (2010). Microsatellite Instability in Arabidopsis Increases with Plant Development . *Plant Physiology*, *154*(3), 1415–1427. https://doi.org/10.1104/pp.110.162933

Gomez-Ramirez, J., & Sanz, R. (2013). On the limitations of standard statistical modeling in biological systems: A full Bayesian approach for biology. *Progress in Biophysics and Molecular Biology*, *113*(1), 80–91. https://doi.org/10.1016/j.pbiomolbio.2013.03.008

Gower, J. C. (1985). Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and Its Applications*, *67*, 81–97. https://doi.org/10.1016/0024-3795(85)90187-9

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732. https://doi.org/10.1093/biomet/82.4.711

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Herrera, C. M., Bazaga, P., Pérez, R., & Alonso, C. (2021). Lifetime genealogical divergence within plants leads to epigenetic mosaicism in the shrub Lavandula latifolia (Lamiaceae). *New Phytologist*, *231*(5), 2065–2076. https://doi.org/10.1111/nph.17257

Hofmeister, B. T., Denkena, J., Colomé-Tatché, M., Shahryary, Y., Hazarika, R. R., Grimwood, J., Mamidi, S., Jenkins, J., Grabowski, P., Sreedasyam, A., Shu, S., Barry, K., Lail, K., Adam, C., Lipzen, A., Sorek, R., Kudrna, D., Talag, J., Wing, R. A., & Hall, D. W. (2020). A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial Populus trichocarpa. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02162-5

Hubin, A., & Storvik, G. (2018). Mode jumping MCMC for Bayesian variable selection in GLMM. *Computational Statistics & Data Analysis*, *127*, 281–297. https://doi.org/10.1016/j.csda.2018.05.020

Huelsenbeck, J. P., Larget, B., Miller, R. E., & Ronquist, F. (2002). Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. *Systematic Biology*, *51*(5), 673–688. https://doi.org/10.1080/10635150290102366

Imai, R., Fujino, T., Sou Tomimoto, Ohta, K., Na'iem, M., Sapto Indrioko, Widiyatno, Susilo Purnomo, Almudena Mollá–Morales, Viktoria Nizhynska, Tani, N., Suyama, Y., Sasaki, E.,

Kasahara, M., & Satake, A. (2023). Somatic mutation rates scale with time not growth rate in long-lived tropical trees. *ELife*, *12*(RP88456). https://doi.org/10.7554/elife.88456

Iwasa, Y., Tomimoto, S., & Satake, A. (2023). The genetic structure within a single tree is determined by the behavior of the stem cells in the meristem. *Genetics*, *223*(4). https://doi.org/10.1093/genetics/iyad020

Jeffreys, H. (1998). *The theory of probability*. OuP Oxford.

Jin, L., & Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution*, *7*(1). https://doi.org/10.1093/oxfordjournals.molbev.a040588

Joseph, L., Wolfson, D. B., & Berger, R. D. (1995). Sample Size Calculations for Binomial Proportions via Highest Posterior Density Intervals. *The Statistician*, *44*(2), 143. https://doi.org/10.2307/2348439

Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, *21*(7), 428–444. https://doi.org/10.1038/s41576-020-0233-0

Kass, R. E., & Wasserman, L. (1996). Formal rules for selecting prior distributions: A review and annotated bibliography. *Journal of the American Statistical Association*, *435*(1343-1370), 28.

Katoh, K., Asimenos, G., & Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. *Methods in Molecular Biology (Clifton, N.J.)*, *537*, 39–64. https://doi.org/10.1007/978-1-59745-251-9_3

Kennedy, J. J. (1970). The Eta Coefficient in Complex Anova Designs. *Educational and Psychological Measurement*, *30*(4), 885–889. https://doi.org/10.1177/001316447003000409

Khoury, C., Laliberté, B., & Guarino, L. (2010). Trends in ex situ conservation of plant genetic resources: a review of global crop and regional conservation strategies. *Genetic Resources and Crop Evolution*, *57*(4), 625–639. https://doi.org/10.1007/s10722-010-9534-z

Klekowski, E. J. (1984). Mutational Load in Clonal Plants: A Study of Two Fern Species. *Evolution*, *38*(2), 417. https://doi.org/10.2307/2408500

Klekowski, E. J. (1988). *Mutation, Developmental Selection, and Plant Evolution*. Columbia University Press. https://doi.org/10.7312/klek92068

Klekowski, E. J., & Godfrey, P. J. (1989). Ageing and mutation in plants. *Nature*, *340*(6232), 389–391. https://doi.org/10.1038/340389a0

Kojima, M., Fabio Minoru Yamaji, Yamamoto, H., Yoshida, M., & Nakai, T. (2009). Effects of the lateral growth rate on wood quality parameters of Eucalyptus grandis from different latitudes in Brazil and Argentina. *Forest Ecology and Management*, *257*(10), 2175–2181. https://doi.org/10.1016/j.foreco.2009.02.026

Kovalchuk, I. (2000). Genome-wide variation of the somatic mutation frequency in transgenic plants. *The EMBO Journal*, *19*(17), 4431–4438. https://doi.org/10.1093/emboj/19.17.4431

Kuhner, M. K., & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, *11*(3). https://doi.org/10.1093/oxfordjournals.molbev.a040126

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, *4*(33), 1143. https://doi.org/10.21105/joss.01143

Kurland, C. G., Canback, B., & Berg, O. G. (2003). Horizontal gene transfer: A critical view. *Proceedings of the National Academy of Sciences*, *100*(17), 9658–9662. https://doi.org/10.1073/pnas.1632870100

Lanfear, R. (2018). Do plants have a segregated germline? *PLOS Biology*, *16*(5), e2005439. https://doi.org/10.1371/journal.pbio.2005439

Lanfear, R., Ho, S. Y. W., Jonathan Davies, T., Moles, A. T., Aarssen, L., Swenson, N. G., Warman, L., Zanne, A. E., & Allen, A. P. (2013). Taller plants have lower rates of molecular evolution. *Nature Communications*, *4*(1). https://doi.org/10.1038/ncomms2836

Lee, X. J., Drovandi, C. C., & Pettitt, A. N. (2014). Model choice problems using approximate Bayesian computation with applications to pathogen transmission data sets. *Biometrics*, *71*(1), 198–207. https://doi.org/10.1111/biom.12249

Leonard, T. H. (2014). A personal history of Bayesian statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(2), 80–115. https://doi.org/10.1002/wics.1293

Lepage, T., Bryant, D., Philippe, H., & Lartillot, N. (2007). A General Comparison of Relaxed Molecular Clock Models. *Molecular Biology and Evolution*, *24*(12), 2669–2680. https://doi.org/10.1093/molbev/msm193

Lindley, D. V. (2000). The Philosophy of Statistics. *The Statistician*, *49*(3), 293–337. https://doi.org/10.1111/1467-9884.00238

Li Song, Liliana Florea, Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads, *GigaScience*, Volume 4, Issue 1, December 2015, s13742–015–0089–y, https://doi.org/10.1186/s13742-015-0089-y

Liu, J., & Hodges, J. S. (2003). Posterior Bimodality in the Balanced One-Way Random-Effects Model. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *65*(1), 247–255. https://doi.org/10.1111/1467-9868.00384

Liu, X., & Niranjan, M. (2017). Parameter Estimation in Computational Biology by Approximate Bayesian Computation coupled with Sensitivity Analysis. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1704.09021

Loewe, L., & Hill, W. G. (2010). The population genetics of mutations: good, bad and indifferent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1544), 1153–1167. https://doi.org/10.1098/rstb.2009.0317

Loh, W.-L. (1996). On Latin hypercube sampling. *The Annals of Statistics*, *24*(5). https://doi.org/10.1214/aos/1069362310

Lukan, P. (2019). Interpretations of Probability and Bayesian Inference—an Overview. *Acta Analytica*, *35*(1), 129–146. https://doi.org/10.1007/s12136-019-00390-4

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337. https://doi.org/10.1023/a:1008929526011

Machina, M. J., & Schmeidler, D. (1992). A More Robust Definition of Subjective Probability. *Econometrica*, *60*(4), 745. https://doi.org/10.2307/2951565

Maddison, W. P., & Knowles, L. L. (2006). Inferring Phylogeny Despite Incomplete Lineage Sorting. *Systematic Biology*, *55*(1), 21–30. https://doi.org/10.1080/10635150500354928

Maia, J. C. F., & Bonat, W. H. (2024). Multivariate Covariance Generalized Linear Models in Python: The mcglm library. *Journal of Open Source Software*, *9*(98), 6037. https://doi.org/10.21105/joss.06037

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2011). Approximate Bayesian computational methods. *Statistics and Computing*, *22*(6), 1167–1180. https://doi.org/10.1007/s11222-011-9288-2

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *100*(26), 15324–15328. https://doi.org/10.1073/pnas.0306899100

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

Michel, A., Arias, R. S., Scheffler, B. E., Duke, S. O., Netherland, M., & Dayan, F. E. (2004). Somatic mutation-mediated evolution of herbicide resistance in the nonindigenous invasive plant hydrilla (Hydrilla verticillata). *Molecular Ecology*, *13*(10), 3229–3237. https://doi.org/10.1111/j.1365-294x.2004.02280.x

Miryeganeh, M., & Armitage, D. W. (2024). Epigenetic responses of trees to environmental stress in the context of climate change. *Biological Reviews of the Cambridge Philosophical Society*. https://doi.org/10.1111/brv.13132

Moulton, V., Spillner, A., & Wu, T. (2018). UPGMA and the normalized equidistant minimum evolution problem. *Theoretical Computer Science*, *721*, 1–15. https://doi.org/10.1016/j.tcs.2018.01.022

Mueller, R. L. (2006). Evolutionary Rates, Divergence Dates, and the Performance of Mitochondrial Genes in Bayesian Phylogenetic Analysis. *Systematic Biology*, *55*(2), 289–300. https://doi.org/10.1080/10635150500541672

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., Goodstein, D. M., Dubchak, I., Poliakov, A., Mizrachi, E., Kullan, A. R. K., Hussey, S. G., Pinard, D., van der Merwe, K., Singh, P., & van Jaarsveld, I.

(2014). The genome of Eucalyptus grandis. *Nature*, *510*(7505), 356–362. https://doi.org/10.1038/nature13308

Nicholls, A. O. (1989). How to make biological surveys go further with generalised linear models. *Biological Conservation*, *50*(1-4), 51–75. https://doi.org/10.1016/0006-3207(89)90005-0

Nielsen, R. (2002). Mapping Mutations on Phylogenies. *Systematic Biology*, *51*(5), 729–739. https://doi.org/10.1080/10635150290102393

Nikooienejad, A., Wang, W., & Johnson, V. E. (2016). Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics*, *32*(9), 1338–1345. https://doi.org/10.1093/bioinformatics/btv764

Nunes, M. A., & Balding, D. J. (2010). On Optimal Selection of Summary Statistics for Approximate Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology*, *9*(1). https://doi.org/10.2202/1544-6115.1576

Nybom, H. (1991). Applications of DNA Fingerprinting in Plant Breeding. *Experientia Supplementum*, 294–311. https://doi.org/10.1007/978-3-0348-7312-3_21

Nylander, J. A. A., Wilgenbusch, J. C., Warren, D. L., & Swofford, D. L. (2007). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, *24*(4), 581–583. https://doi.org/10.1093/bioinformatics/btm388

O'Hagan, A. (2024). Bayesian statistics: principles and benefits. *Frontis*, 31–45. https://library.wur.nl/ojs/index.php/frontis/article/view/856

Ochatt, S. J. (2008). Flow cytometry in plant breeding. *Cytometry Part A*, *73A*(7), 581–598. https://doi.org/10.1002/cyto.a.20562

OpenAI. (2025). ChatGPT (Version 4) [Large language model]. https://chat.openai.com/

Orr, A. J., Padovan, A., Kainer, D., Carsten Külheim, Lindell Bromham, Bustos-Segura, C., Foley, W., Haff, T., Hsieh, J.-F., Morales-Suarez, A., Cartwright, R. A., & Lanfear, R. (2020). A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. *Proceedings of the Royal Society B Biological Sciences*, *287*(1922), 20192364–20192364. https://doi.org/10.1098/rspb.2019.2364

Otto, S. A. (2019, January 7). *How to normalize the RMSE [Blog Post]*. Marine Data Science. https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/

Otto, S. P., & Orive, M. E. (1995). Evolutionary consequences of mutation and selection within an individual. *Genetics*, *141*(3), 1173–1187. https://doi.org/10.1093/genetics/141.3.1173

Overcast, I., Bagley, J. C., & Hickerson, M. J. (2017). Strategies for improving approximate Bayesian computation tests for synchronous diversification. *BMC Evolutionary Biology*, *17*(1). https://doi.org/10.1186/s12862-017-1052-6

Padovan, A., Lanfear, R., Keszei, A., Foley, W. J., & Külheim, C. (2013). Differences in gene expression within a striking phenotypic mosaic Eucalyptus tree that varies in susceptibility to herbivory.

Pick, J. L., Kasper, C., Allegue, H., Dingemanse, N. J., Dochtermann, N. A., Laskowski, K. L., Lima, M. R., Schielzeth, H., Westneat, D. F., Wright, J., & Araya‐Ajoy, Y. G. (2023). Describing posterior distributions of variance components: Problems and the use of null distributions to aid interpretation. *Methods in Ecology and Evolution*, *14*(10), 2557–2574. https://doi.org/10.1111/2041-210x.14200

Poethig, S. (1989). Genetic mosaics and cell lineage analysis in plants. *Trends in Genetics*, *5*, 273–277. https://doi.org/10.1016/0168-9525(89)90101-7

Popov, V. N., Syromyatnikov, M. Yu., Franceschi, C., Moskalev, A. A., & Krutovsky, K. V. (2022). Genetic mechanisms of aging in plants: What can we learn from them? *Ageing Research Reviews*, *77*, 101601. https://doi.org/10.1016/j.arr.2022.101601

Prach, K., & Pyšek, P. (1994). Clonal plants—What is their role in succession? *Folia Geobotanica et Phytotaxonomica*, *29*(2), 307–320. https://doi.org/10.1007/bf02803803

Prangle, D., Blum, M. G. B., Popovic, G., & Sisson, S. (2014). Diagnostic tools for approximate Bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics*, *56*(4), 309–329. https://doi.org/10.1111/anzs.12087

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091

Qian, S. S., Stow, C. A., & Borsuk, M. E. (2003). On Monte Carlo methods for Bayesian inference. *Ecological Modelling*, *159*(2-3), 269–277. https://doi.org/10.1016/s0304-3800(02)00299-5

Quiroz, D., Lensink, M., Kliebenstein, D. J., & J. Grey Monroe. (2023). Causes of Mutation Rate Variability in Plant Genomes. *Annual Review of Plant Biology*, *74*(1), 751–775. https://doi.org/10.1146/annurev-arplant-070522-054109

Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., & Conrad, D. F. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature Methods*, *10*(10), 985–987. https://doi.org/10.1038/nmeth.2611

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Richter, A., & Singleton, W. R. (1955). The Effect of Chronic Gamma Radiation on the Production of Somatic Mutations In Carnations. *Proceedings of the National Academy of Sciences*, *41*(5), 295–300. https://doi.org/10.1073/pnas.41.5.295

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

Roth, P. L., Bobko, P., Switzer, F. S., & Dean, M. A. (2001). Prior Selection causes Biased Estimates of Standardized Ethnic Group Differences: Simulation and Analysis. *Personnel Psychology*, *54*(3), 591–617. https://doi.org/10.1111/j.1744-6570.2001.tb00224.x

Schmid-Siegert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., Cattaneo, P., Schütz, F., Farinelli, L., Pagni, M., Schneider, M., Voumard, J., Jaboyedoff, M., Fankhauser, C., Hardtke, C. S., Keller, L., Pannell, J. R., Reymond, A., Robinson-Rechavi, M., & Xenarios, I. (2017). Low number of fixed somatic mutations in a long-lived oak tree. *Nature Plants*, *3*(12), 926–929. https://doi.org/10.1038/s41477-017-0066-9

Schmitt, S., Heuret, P., Troispoux, V., Beraud, M., Cazal, J., Chancerel, É., Cravero, C., Guichoux, E., Lepais, O., Loureiro, J., Marande, W., Martin-Ducup, O., Vincent, G., Chave, J., Plomion, C.,

Leroy, T., Heuertz, M., & Tysklind, N. (2024). Low-frequency somatic mutations are heritable in tropical trees *Dicorynia guianensis* and *Sextonia rubra*. *Proceedings of the National Academy of Sciences*, *121*(10). https://doi.org/10.1073/pnas.2313312121

Schoen, D. J., & Schultz, S. T. (2019). Somatic Mutation and Evolution in Plants. *Annual Review of Ecology, Evolution, and Systematics*, *50*(1), 49–73. https://doi.org/10.1146/annurev-ecolsys-110218-024955

Secrier, M., Toni, T., & Stumpf, M. P. (2009). The ABC of reverse engineering biological signalling systems. *Molecular BioSystems*, *5*(12), 1925-1935.

Sedgwick, P. (2012). Multiple significance tests: the Bonferroni correction. *BMJ*, *344*, e509–e509. https://doi.org/10.1136/bmj.e509

Sedlazeck, F. J., Rescheneder, P., & von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, *29*(21), 2790–2791. https://doi.org/10.1093/bioinformatics/btt468

Sekiguchi, F., Yamakawa, K., & Yamaguchi, H. (1971). Radiation damage in shoot apical meristems of Antirrhinum majus and somatic mutations in regenerated buds. *Radiation Botany*, *11*(2), 157–169. https://doi.org/10.1016/s0033-7560(71)90693-4

Sereno, P. C. (2005). The Logical Basis of Phylogenetic Taxonomy. *Systematic Biology*, *54*(4), 595–619. https://doi.org/10.1080/106351591007453

Shao, K.-T., & Sokal, R. R. (1990). Tree Balance. *Systematic Biology*, *39*(3), 266–276. https://doi.org/10.2307/2992186

Shields, M. D., & Zhang, J. (2016). The generalization of Latin hypercube sampling. *Reliability Engineering & System Safety*, *148*, 96–108. https://doi.org/10.1016/j.ress.2015.12.002

Shiraishi, Y., Sato, Y., Chiba, K., Okuno, Y., Nagata, Y., Yoshida, K., Shiba, N., Hayashi, Y., Kume, H., Homma, Y., Sanada, M., Ogawa, S., & Miyano, S. (2013). An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Research*, *41*(7), e89–e89. https://doi.org/10.1093/nar/gkt126

Sparrow, A. H., & Cuany, L. R. (1959). *Radiation-Induced Somatic Mutations In Plants*. Osti.gov. https://www.osti.gov/biblio/4225810

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. *PLoS Computational Biology*, *9*(1), e1002803. https://doi.org/10.1371/journal.pcbi.1002803

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring Coalescence Times From DNA Sequence Data. *Genetics*, *145*(2), 505–518. https://doi.org/10.1093/genetics/145.2.505

Thiébaux, H. J., & Zwiers, F. W. (1984). The Interpretation and Estimation of Effective Sample Size. *Journal of Applied Meteorology and Climatology*, *23*(5), 800–811. https://doi.org/10.1175/1520-0450(1984)023%3C0800:TIAEOE%3E2.0.CO;2

Tilney-Bassett, R. A. (1986). *Plant chimeras* (pp. vi+-199pp).

Tomimoto, S., & Satake, A. (2023). Modelling somatic mutation accumulation and expansion in a long-lived tree with hierarchical modular architecture. *Journal of Theoretical Biology*, *565*, 111465–111465. https://doi.org/10.1016/j.jtbi.2023.111465

Toni, T. (2011). ABC SMC for parameter estimation and model selection with applications in systems biology. *Nature Precedings*. https://doi.org/10.1038/npre.2011.5964.1

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. H. (2008). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, *6*(31), 187–202. https://doi.org/10.1098/rsif.2008.0172

Tooke, F., & Battey, N. (2003). Models of shoot apical meristem function. *New Phytologist*, *159*(1), 37–52. https://doi.org/10.1046/j.1469-8137.2003.00803.x

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69–85. https://doi.org/10.1016/j.jmp.2012.02.005

Van der Auwera GA & O'Connor BD. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition)*. O'Reilly Media.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-Normalization, Folding, and Localization: An Improved R for Assessing Convergence of MCMC. *Bayesian Analysis*. https://doi.org/10.1214/20-ba1221

Vincent, Otto, S. P., & Aitken, S. N. (2019). Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. *Evolution Letters*, *3*(4), 348–358. https://doi.org/10.1002/evl3.121

Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., Zhang, X., Zhao, L., Zhang, Y., Jia, Y., Qin, C., Yu, L., Huang, J., Yang, S., Hurst, L. D., & Tian, D. (2019). The architecture of intra-organism mutation rate variation in plants. *PLOS Biology*, *17*(4), e3000191. https://doi.org/10.1371/journal.pbio.3000191

Wang, Y., & Li, J. (2008). Molecular Basis of Plant Architecture. *Annual Review of Plant Biology*, *59*(1), 253–279. https://doi.org/10.1146/annurev.arplant.59.032607.092902

Weiss, G. (2003). Testing Substitution Models Within a Phylogenetic Tree. *Molecular Biology and Evolution*, *20*(4), 572–578. https://doi.org/10.1093/molbev/msg073

Weissman, A. (1885). Germ-plasm: A theory of heredity. Charles Scribner's Sons., Translated by W. Newton Parker and Harriet Rönnfeldt.

Weiss, R. (1996). An Approach to Bayesian Sensitivity Analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(4), 739–750. https://doi.org/10.1111/j.2517-6161.1996.tb02112.x

Werhli, A. V., Grzegorczyk, M., & Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, *22*(20), 2523–2531. https://doi.org/10.1093/bioinformatics/btl391

Whelan, S., Liò, P., & Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics*, *17*(5), 262–272. https://doi.org/10.1016/s0168-9525(01)02272-7

Whitham, T. G., & N, S. C. (1981). Evolution by Individuals, PlantHerbivore Interactions, and Mosaics of Genetic Variability: The Adaptive Significance of Somatic Mutations in Plants. *Oecologia*, *49*(3), 287–292. JSTOR. https://doi.org/10.2307/4216386

Wilkinson, D. J. (2006). Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics*, *8*(2), 109–116. https://doi.org/10.1093/bib/bbm007

Wilkinson, R. D. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, *12*(2). https://doi.org/10.1515/sagmb-2013-0010

Xia, X., Xie, Z., Salemi, M., Chen, L., & Wang, Y. (2003). An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution*, *26*(1), 1–7. https://doi.org/10.1016/s1055-7903(02)00326-3

Yao, Y., & Kovalchuk, I. (2011). Abiotic stress leads to somatic and heritable changes in homologous recombination frequency, point mutation frequency and microsatellite stability in Arabidopsis plants. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *707*(1-2), 61–66. https://doi.org/10.1016/j.mrfmmm.2010.12.013

Yifan Ren, Zhen He, Pingyu Liu, Brian Traw, Shucun Sun, Dacheng Tian, Sihai Yang, Yanxiao Jia, Long Wang, Somatic Mutation Analysis in *Salix suchowensis* Reveals Early-Segregated Cell Lineages, *Molecular Biology and Evolution*, Volume 38, Issue 12, December 2021, Pages 5292–5308, https://doi.org/10.1093/molbev/msab286

Zahradníková, E., Ficek, A., Mičieta, K., Bresova, B., & Vinar, T. (2019). Mosaicism in old trees and its patterns. *Trees*, *34*(2), 357–370. https://doi.org/10.1007/s00468-019-01921-7

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, *8*(2), 357–366. https://doi.org/10.1016/0022-5193(65)90083-4

Appendix A

Challenges with Implementation of Orr et al. (2020) Pipeline

The pipeline setup requires a Linux system, preferably configured with the Ubuntu operating system (Ubuntu 20.04 LTS), due to its compatibility with bioinformatic tools and libraries. The initial setup includes the installation of the Conda package manager (version 23.1.0, Anaconda), which facilitates the management of software dependencies of the pipeline. The channel Bioconda was utilised as a specialised repository with Conda, hosting bioinformatic-specific software packages unavailable on Conda itself (Grüning et al., 2018). Once Bioconda was initialised, additional tools- Rcorrector, khmer, NextGenMap (NGM), GATK, BCFtools, RAxML and DNG - must be installed. Specific versions of each tool are required, as specified by the repository, to avoid incompatibility. With the environment established, I followed the repositories instructions to download and configure all necessary files. The repositories' makefiles serve as the primary method for executing pipeline steps, running sequentially to process data from raw reads through to variant calling, phylogenetic tree construction and the final estimation of the somatic mutation rate.

Within the first makefile, Rcorrecter (Song & Florea, 2015) and khmer (Crusoe et al., 2015) preprocess raw reads of each branch (with three replicates each) by correcting for errors and filtering repetitive kmers. Reads are then aligned using NextGenMap (Sedlazeck et al., 2013), first to a pseudo-reference genome (*E. grandis*) and later to a consensus sequence derived from the iterative alignments. GATK (Van der Auwera & O'Connor, 2020) performs variant calling on the aligned reads, generating a VCF file of detected variants, which BCFtools (Danecek et al., 2021) subsequently filters for quality control to retain high-confidence mutations. RAxML (Stamatakis, 2014) was utilised to construct a phylogenetic tree from the remaining variants per branch, which was validated using the actual topology of the *E. melliodora* individual as a positive control. At this stage, DNG aka. DeNovoGear (Ramu et al., 2013) is employed to further filter variants by distinguishing 'true mutations' from errors using inheritance patterns and expected variant distributions as defined by the *E. melliodora* topology. This last step explicitly applies Orr et al. 2020's phylogenomic method, with the remaining high-confidence variants used to estimate a final somatic mutation rate.

The pipeline's structure and resource demands make it unsuitable for high-throughput replication, especially in large-scale simulations involving large sample sizes (e.g. 12,000 samples). Each stage in the pipeline must be executed sequentially, with outputs from one step required as inputs for the next, leading to compounding delays, which collectively slow down the overall workflow. Additionally, the makefiles lack parallel processing support, meaning many tasks cannot be distributed across multiple threads to reduce runtime. Several of the pipeline stages are also highly memory intensive, adding to the complications of the pipeline. For instance, the 'create_consensus' script, used to create a consensus *E. melliodora* reference genome, required over 500GB of RAM (random access memory) and took approximately five days to complete while using 48 threads. Replicating this pipeline for simulations is highly impractical, with the resources required to simulate and process whole genomes for each branch of an input tree across a sufficiently large sample size of input trees being infinite.

Appendix B

Simulation Code in Detail

This section details the simulation methodology used within Chapter 2 to assess the applicability and limitations of the phylogenomic method across varying trees. A Python script denoted '`sim_code.py`' is used to execute the simulations and can be found within the open GitHub repository **sim_tree_mut** (https://github.com/andreag186/sim_tree_mut). This script was written in Python and is compatible with version 3.11.4. The simulation code comprises several functions which handle various stages of the pipeline, from parameter sampling to simulation of somatic mutations and application of the modified phylogenomic method, and finally, the estimation of a somatic mutation rate. Here, each function's role is explained in the order in which they are executed.

1. **Latin Hypercube Sampling (LHS) of Parameters** (`latin_hypercube_sampling`)
   LHS generates diverse parameter sets for each sample to explore parameter space comprehensively. LHS is implemented using the `qmc.LatinHypercube` class from `scipy`. The variables which form a parameter set are listed below.
   - `mu_0`: Mutation rate parameter, selected from 6 treatment levels specified: [4, 6, 8, 40, 60, 80] * 1^{-10}
   - `GenSize`: Genome size, fixed for each sample as proportional to `mu_0`
   - `tree_topologies`: A list of samples, through a dictionary of predefined tree topologies and input tree parameters, of 20 treatment levels specified by a unique code.*
   - `StD`: Elongation parameter, 0 = Structured & 5=Stochastic (equal to stem cells)
   - `biasVar`: Branching parameter, 10 = Unbiased & 0.5 = Biassed.
   - `num_params` : The number of parameters used in LHS (= 4, with `GenSize` constant).
   - `num_samples`: The number of samples specified when the function is called

* tree_topologies_dict *is the dictionary that defines each input tree's number of branches, age, and topology. See the example entry of tree 'bS4' below:*

```
"bS4": {
    "numBranch":4,   -The number of terminal branches
    "age": 123,   -The root to tip 'age', not total age(= 300) , equivalent to height in cm
    "s10": 10,    - The age of the trunk/root prior to the first split, 10 years for each tree (1m)
    "b11": 75, "bb11": 38, "b12": 38   -Branches RIGHT of split
    "s40": 75, "b41":38, "s41": 38     -Branches LEFT of split
}
```

Here, `"bS4"` *indicates this topology is balanced (b) with short terminal branches (S) and 4 terminal branches. For branches RIGHT of the split, internal branches are denoted by b11-bxx, with the last 'internal branch' treated as a terminal branch. Terminal branches are denoted by bb11-bbxx, with the number corresponding to the internal branch on which they stem. For branches LEFT of the split, internal branches are denoted by s40-sxx, and terminal branches by b41-bxx. This nomenclature*

structure is preserved from the original Tomimoto & Satake (2023) source code. See Figure 1, Appendix.

2. **Result Compilation and Output to CSV** (`write_samples_to_csv`)
   This function writes all simulation results to a CSV file using `csv.writer` and `pandas` for handling tabular data. The sample number, input parameter values, and all output values are recorded.

3. **'Decode' Tree Topology Configuration** (`create_tree_list_and_dict`):
   This function 'decodes' the input trees defined in the `tree_topologies_dict`, extracting all information needed for future functions. This function allows ANY input tree with any number of branches and topology to be compatible with the simulation, given that the tree is correctly structured in tree_topologies_dict. This function returns the following:
   - `tree_list`: An output list of all branch names and corresponding age (in years)
   - `tree_dict`: An output dictionary classifying branches into right or left and terminal and internal branches for future applications.
   - `numBranch` and `age` are also outputted here for subsequent calculations.

4. **Average Meristem Time Calculation** (`calculate_ave_meristem_time`)
   This function calculates the average time (in years) spent by meristem cells in each branch, using root-to-tip distances of left and right branches. This value is used later to estimate mutation accumulation over time (in `mut_dist_func`)
   - `tree_dict` is inputted specifying the `root`, `right_internal`, `right_terminal`, `left_internal` and `left_terminal values` for each input tree.
   - `ave_meristem_time`: The average meristem time, calculated as the sum of root-to-tip distances divided by `numBranch`

5. **Main Simulation Loop** (`simulate_somatic_mutations`)
   This function iterates through each specified replicate (`NumTime`), generating mutations in unique genome positions across branches based on input parameters (e.g. model). A combined matrix (`br_brmutmatrix`) is created (using the package `numpy`) containing all the mutations accumulated in the tree and their unique site patterns, specifying in which branch or branches each mutation occurs. *****
   - `NumStem` (number of stem cells = 5), `NumTime` (number of simulation replications = 3), `mu_0` (input mutation rate), `GenSize` (genome size = 500Mb), `NumDiv` (branching division constant = 7), `nDiv` (elongation division constant = 1), `StD` & `biasVar` (model parameters): The core input parameters which define mutation simulation.
   - `List10`, `List_br`, and `List_Left`: The data structures storing mutations for the root (growth prior to the first split), right branches, and left branches, respectively.

** The function* `simulate_somatic_mutations`*, calls upon functions 6-9 defined by Tomimoto & Satake (2023) in their source code to simulate somatic mutations by applying their mathematical models. I did not alter these functions, only their application.*

6. **Stem Cell Mutation Simulation** (`mutInStemCells`)
   This function initialises a list of stem cells and simulates mutations across consecutive divisions; in other words, it simulates elongation in the 'stem' of the tree prior to the first split of branches. `copy.deepcopy` ensures each stem cell's mutation history is preserved accurately, with `random.random()` introducing mutations based on the probability `mu_0` (mutation rate). The mutation history of each stem cell is saved in the object `tCells` and updated after each division cycle.

7. **Branch Mutation Simulation** (`mutInBrStemCells`)
   Similar to the function `mutInStemCells`, this function models mutation accumulation in branch-specific cells (axillary meristems formed from the SAM). In other words, it simulates elongation along a branch. The `stemCells` list is passed as input.
   - `stemCells`: A list of the initial set of stem cells with their mutation history, as sampled from the SAM
   - `ccells` & `ccells2`: Temporary lists for the daughter cells produced in each division cycle (during axillary meristem formation, aka. branching AND elongation along a branch)
   - `tCells`: Final list of all mutated cells for the given branch.

8. **Sampling and Mutation Assignment** (`sample_mutations`)
   This function samples cells for the next branch generation using weights defined by `weightList`, which is based on spatial ($x \in [0, 2\pi)$) and bias (σ or biasVar) parameters. In other words, it simulates the branching process. `numpy.random.choice` selects cells for each division.
   - `wList`: A list of weights generated to reflect spatial positioning and bias
   - `NumDiv`: Number of cell divisions for each branching event ($r_b = 7$)
   - `sample`: List of sampled cells for the new meristem, used as input for further elongation simulations along that branch.

9. **Mutation Matrix Generation** (`makeMutMatrix`)
   This function generates a binary mutation matrix, indicating the presence or absence of across stem cells in each branch. The matrix provides a detailed mutation map.
   - `List`: An input list of meristems with mutation histories
   - `mutList`: A list of all unique mutations used to populate the matrix
   - `np_per_mutMatrix`: The final mutation matrix with binary entries, representing mutation presence in each cell of one branch. The final `br_brmutMatrix` output for the whole tree collates and filters `np_per_mutMatrix` objects for each branch.

10. **Mutation Frequency and Distribution Analysis** (`mut_dist_func`)

This function computes mutation distribution patterns, calculating each branch's total, shared and unique mutations. It summarises the mutation frequencies and patterns, capturing distribution across branches.

- `mutShapeTemplate`: A template of all possible mutation combinations across branches (dependent on numBranch)
- `allMutDist`: An array of mutation distribution patterns averaged across simulation runs (as defined by `NumTime`)
- `total_mutations`, `shared_mutations` and `unique_mutations`: Dictionaries storing counts of specified mutations corresponding relevant branch number(s).

11. **Distance Matrix Construction** (`gen_matrices`)
    This function constructs genetic and physical distance matrices to analyse each tree's mutational and spatial distances. The genetic matrix (`genetic_distance_matrix`) captures shared mutations between branches, while the physical matrix (`physical_distance_matrix`) represents distances (in cm) between branches based on root-to-tip calculations (`root_to_tip_distances`).

12. **Mutation Rate and Variant Calculations** (`calc_variants`)
    This function performs a linear regression using `sklearn.linear_model.LinearRegression`. This function estimates the relationship between genetic and physical distances within a tree, thereby estimating a somatic mutation rate using the modified phylogenomic method. The function outputs the estimated mutation rate as well as the regression equation.
    - `genetic_distances` and `physical_distances`: Arrays of distances extracted from the upper triangle of matrices, used as inputs for the regression model.
    - `regressor`: A linear regression model fits the genetic and physical distances.
    - `output_mut_rate`: The estimated mutation rate per nucleotide, derived directly from the regressor. This value is later converted to per site per year by dividing by age (300 years).

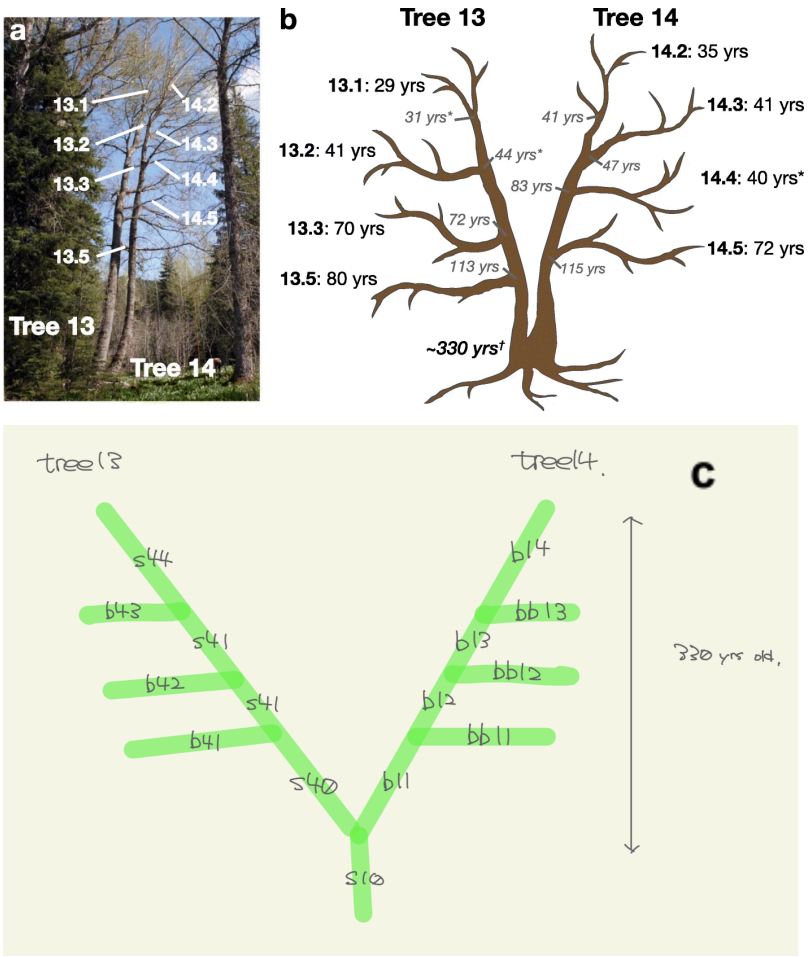


Figure B.1 Combined Illustration of the Branch Nomenclature System used in Tomimoto and Satake's (2023) Simulation Framework, Based on Hoffmeister et al.'s (2020) Study of a Poplar Tree Individual.

(a) Photograph of wild *Populus trichocarpa* trees 13 and 14, located near Mt. Hood, Oregon. These trees emerged following a decapitation event approximately 300 years ago, with Tree 14 re-sprouting from the stump and Tree 13 re-sprouting 80-100 years later. Terminal branches sampled for genomic analysis are labeled (Fig 1(a), Hoffmeister et al. 2020)

(b) Schematic of Trees 13 and 14 showing the age of each branch at its terminal end (black font) and where it joins the main stem (grey italics). Ages marked with * were estimated using branch diameter, while other ages were determined from core samples. Leaf samples from these branches were used for genomic and transcriptomic sequencing (Fig 1(b), Hoffmeister et al. 2020)

(c) Hand-drawn diagram provided by Sou Tomimoto, illustrating the branch nomenclature used in their simulation for the application of their models to *Populus trichocarpa* as studied by Hoffmeister et al. (2020). Internal branches on the right side of the split (Tree 14) are labelled b11-bxx, with the last internal branch treated as a terminal branch, and terminal branches labeled bb11-bbxx, corresponding to their parent internal branch. On the left side (Tree 13), internal branches are labeled s40-sxx, and terminal branches as b41-bxx. This nomenclature, applied to the Poplar tree, reflects the hierarchical branch architecture and is preserved in the simulation framework developed by Tomimoto and Satake (2023).

Appendix C

ABC Code in Detail

Below is a description of the scripts used to apply the Approximate Bayesian Computation (ABC) framework to empirical *E. melliodora* data (pre and post-dng filtering), as well as for validating the ABC framework. The scripts can be found in the open Github repository: https://github.com/andreag186/sim_tree_mut/tree/main
The scripts introduce several new functions that extend the original simulation pipeline:

1. **Parameter Sampling** (`sample_prior`):
   This function samples parameters (θ) values from uniform prior distributions ($p(\theta)$) for mutation rate (`input_mut`), elongation behaviour (`StD`), and branching bias (`biasVar`) using the `scipy.stats.uniform` module.

2. **Simulation Execution** (`run_simulation`):
   The function calls the previously described simulation functions using the sampled parameters (θ) . It returns the simulated mutation distribution across branches (`unique_mutations`) and the estimated phylogenomic method mutation rate (`output_mut_rate`)- this can optionally be saved and analysed alongside the ABC posterior.

3. **Distance Calculation** (`calculate_distance`):
   Using `numpy.linalg.norm`, this function calculates the Euclidean distance (ρ) between the simulated and observed (either pre-dng or post-dng) mutation distributions. This distance determines whether a parameter is accepted or rejected based on the tolerance threshold ($\epsilon = 20$). This threshold can be defined using the `epsilon` variable in the `main` execution function.

4. **Posterior Approximation** (`abc_rejection`):
   The main loop of the ABC-Reject framework, this function iterates over:
   - Sampling parameters from prior distributions (`sample_prior`)
   - Running simulations and calculating the Euclidean distance (run_simulation and calculate_distance)
   - Accepting parameters if the distance satisfies $\rho \leq \epsilon$

   Periodically, the function saves accepted samples to CSV files using `pandas`. The `save_interval` indicates after how many trials the accepted samples are saved (=100), with the total number of samples defined by `num_samples` (=500).

5. **Posterior Analysis** (`arviz`):
   After running the ABC-Reject framework, the accepted parameter sets can be analysed using `arviz`. Although not explicitly part of the scripts, `arviz` enables posterior summary statistics and trace plot visualization for `StD`, `biasVar` and `input_mut`.

Both scripts can be run from the command line or within a Python IDE. SLURM batch processing is also supported for high-performance computing environments, such as NeSI. Unlike other Bayesian methods, the ABC-Reject framework allows for parallelisation, with trials not requiring to be run sequentially. As long as the prior distributions and epsilon remain constant, several trials accepted samples can be collated to form the approximated posterior distribution.