UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

CAROLINA LEMES NASCIMENTO COSTA

SPATIAL MODELS OF SPECIATION

MODELOS ESPACIAIS DE ESPECIAÇÃO

CAMPINAS
2019

CAROLINA LEMES NASCIMENTO COSTA

SPATIAL MODELS OF SPECIATION

MODELOS ESPACIAIS DE ESPECIAÇÃO

Thesis presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Ecology

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutora em Ecologia

Orientador: Marcus Aloizio Martinez de Aguiar

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELA ALUNA CAROLINA LEMES NASCIMENTO COSTA, E ORIENTADA PELO PROF DR. MARCUS ALOIZIO MARTINEZ DE AGUIAR.

CAMPINAS
2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

C823s

Costa, Carolina Lemes Nascimento, 1989-
Spatial models of speciation / Carolina Lemes Nascimento Costa. – Campinas, SP : [s.n.], 2019.

Orientador: Marcus Aloizio Martinez de Aguiar.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Especiação. 2. Radiação adaptativa (Evolução). 3. Modelos biológicos. 4. Padrão espacial. 5. Macroevolução. I. Aguiar, Marcus Aloizio Martinez de, 1960-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Modelos espaciais de especiação
Palavras-chave em inglês:
Speciation
Adaptive radiation (Evolution)
Biological models
Spatial pattern
Macroevolution
Área de concentração: Ecologia
Titulação: Doutora em Ecologia
Banca examinadora:
Marcus Aloizio Martinez de Aguiar [Orientador]
Mathias Mistretta Pires
Sabrina Borges Lino Araujo
Rodrigo André Caetano
Gustavo Burin Ferreira
Data de defesa: 25-02-2019
Programa de Pós-Graduação: Ecologia

Comissão Examinadora:

Prof. Dr. Marcus Aloizio Martinez de Aguiar

Prof. Dr. Mathias Mistretta Pires

Profa. Dra. Sabrina Borges Lino Araujo

Prof. Dr. Rodrigo André Caetano

Prof. Dr. Gustavo Burin Ferreira

Os membros da Comissão Examinadora acima assinaram a Ata de defesa, que se encontra no processo de vida acadêmica do aluno.

Agradecimentos

"A gratidão é o único tesouro dos humildes", disse Shakespeare. Eu acrescentaria que é o mais importante de todos os tesouros. Agradeço a todos que de alguma maneira colaboraram para o desenvolvimento dessa tese. À minha família, sem a qual eu não poderia galgar nenhum dos caminhos que segui durante minha vida até hoje. Ao meu companheiro, que em todos os nossos anos juntos sempre foi meu amigo, me incentivando a me desenvolver. Aos meus amigos, que em momentos descontraídos trouxeram leveza à esses anos de árduo estudo, e em momentos de seriedade me ajudaram a conceber problemas e foram ainda grandes colaboradores nos projetos aqui desenvolvidos. Agradeço aos professores com quem tive oportunidade de estudar, seja assistindo encantada as suas aulas, seja em diálogos enriquecedores a dois, que com suas experiências me ajudaram a desenvolver concepções maduras sobre problemas diversos da Biologia. Agradeço ao meu orientador pelos anos de convivência, que me proporcionaram muito amadurecimento, como profissional e como pessoa. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil, CAPES/PROEX - 0487, processo nº 1456640 - Código de Financiamento 001. Espero que com os conhecimentos aqui adquiridos eu possa retribuir dignamente ao investimento que em mim foi feito.

A todos que participaram de alguma maneira dos meus últimos quatro anos de vida, sem dúvida influenciando no meu desempenho profissional, meu muito obrigada.

Resumo

A impressionante diversidade observada na natureza nos faz pensar quais processos podem ser responsáveis por tamanha variedade. Responder esta questão foi o objetivo de muito biólogos evolutivos, que tentaram descobrir os processos olhando para os padrões que eles poderiam gerar. O desenvolvimento de modelos teóricos, em particular modelos baseados em indivíduo, é indispensável para lidar com esta questão, pois apenas com modelos podemos isolar processos específicos em um ambiente controlado, o que não é completamente possível em experimentos naturais, e em um tempo realizável. Nesta tese eu investiguei quais são os padrões gerados por um modelo de especiação baseado em indivíduo no qual apenas processos neutros e o espaço estão regulando a dinâmica populacional. A população evoluiu sob as influências combinadas de reprodução sexuada, mutação e dispersão. No primeiro capítulo, desenvolvemos um algoritmo que registra as relações de ancestralidade–descendência entre pares de indivíduos da comunidade final, e um algoritmo que registra os tempos exatos de especiação e extinção das espécies. Com ambas as informações foi possível construir genealogias e filogenias, a partir das quais padrões macroevolutivos foram obtidos, servindo como um referencial de evolução neutra. O segundo capítulo foi dedicado a usar esta nova informação filogenética do modelo para investigar se diferentes contextos geográficos de especiação (parapátrica e simpátrica) deixam assinaturas distintas nos padrões macroevolutivos de diversificação, como a simetria de árvores e a velocidade da diversificação. Os resultados das simulações foram comparados com dados empíricos de radiações evolutivas. O terceiro capítulo, por fim, incorporou barreiras espaciais ao modelo anterior, para buscar por possíveis assinaturas deixadas pela especiação alopátrica, com barreiras variando em tamanho e permitindo que indivíduos as cruzassem dependendo de seu tamanho. O modelo foi adaptado ao sistema particular dos macacos Platyrrhini, com o espaço modelado de modo a se ajustar à forma da América do Sul, e as barreiras representando os principais rios da região. O número de gerações foi adaptado a diferentes subfamílias e gêneros dos Platyrrhini, para examinar a "Riverine Hypothesis" com um enfoque de modelagem. Os resultados dos três capítulos mostraram que o espaço possui um papel fundamental na especiação quando processos neutros são os únicos a agir sob as populações, com o contexto geográfico da especiação deixando assinaturas nos padrões macroevolutivos emergentes. A incorporação de processos não neutros e a investigação do papel da extinção em moldar os padrões são possíveis passos seguintes para esta pesquisa.

Abstract

The impressive diversity observed in nature makes us wonder what processes could be responsible for so great variety. The answer to this question has been the goal of many evolutionary biologists, who have tried to discover the processes looking for the patterns they would generate. The development of theoretical models, particularly individual based models, is imperative to address this question, as only with models we can isolate specific processes in a controled environment, something not completely possible in natural experiments, and in a feasible time. In this thesis I investigated what are the patterns generated by an individual based model of speciation in which only neutral processes and the space are regulating the dynamics of the population. The population evolved under the combined influences of sexual reproduction, mutation and dispersal. In the first chapter, we developed an algorithm that records the ancestor-descendant relationships between each pair of individuals of the final community, and an algorithm which records the exact speciation and extinction times of species. With both information was possible to construct genealogies and phylogenies, from which macroevolutionary patterns could be derived, offering a neutral referential of evolution. The second chapter was dedicated to use this new phylogenetic information of the model to investigate if different geographical contexts of speciation (parapatric and sympatric) leave different signatures in the macroevolutionary patterns of diversification, like tree symmetry and the speed of diversification. The simulations results were compared with empirical data about evolutionary radiations. The third chapter, lastly, incorporated spatial barriers to the previous model with the goal of looking for possible signatures left by allopatric speciation, with barriers varying in sizes and allowing the crossing of individuals depending on the individual size. The model adapted to the particular system of Platyrrhini monkeys, with space modeled to fit the shape of South America, and spatial barriers representing the main rivers of the region. The number of generations was adapted to conform different subfamilies and genera of Platyrrhini monkeys, with the aim of examine the Riverine Hypothesis in a modeling approach. All results from the three chapters have showed that the space plays a fundamental role in speciation when neutral processes are the only acting upon populations, with the geographic context of speciation leaving signatures in the macroevolutionary patterns emerged. The incorporation of non neutral processes and the investigation of the role of extinction in shaping the patterns are possible next steps to this research.

Summary

1 Introduction 11

2 Registering the evolutionary history in individual-based models of speciation 17

2.1 Introduction 18

2.2 Registering the history of individuals and species 20

2.2.1 Ancestor-descendant relationships among individuals - MRCAT 20

Asexual models 21

Sexual models 22

Drawing genealogies from MRCAT matrices 24

2.2.2 Recording all speciation and extinction events - SSEE 25

2.2.3 Phylogenies generated by ancestor-descendant relationships (MRCAT) versus trees from speciation and extinction events (SSEE) . 27

2.3 Applications of MRCAT and SSEE algorithms to an individual-based model 28

2.3.1 The speciation model 28

2.3.2 Phylogenies based on genetic distances 30

2.3.3 Statistical indexes to compare phylogenies 31

2.4 Results 33

2.5 Discussion 37

2.6 Conclusions 42

3 Signatures of microevolutionary processes in phylogenetic patterns 43

3.1 Introduction 44

3.2 Materials and Methods 47

3.2.1 Model 47

3.2.2 Rooted bifurcating trees and quantifying indexes 50

3.2.3 Empirical phylogenies 52

3.3 Results 54

3.3.1 Large genome 55

3.3.2 Small mating range 58

3.3.3 Comparison with empirical data 59

3.4 Discussion 60
3.5 Supplementary Material 66
3.6 Funding 66
3.7 Acknowledgments 67

4 Speciation promoted by geographic barriers: modeling the Riverine Hypothesis with an individual-based model 68
4.1 Introduction 69
4.2 Material and Methods 72
4.2.1 The recent Platyrrhini radiation 72
4.2.2 The speciation model 74
4.2.3 Quantification of diversification properties 76
4.2.4 Comparison with empirical platyrrhine data 77
4.3 Results 78
4.4 Discussion 83
4.4.1 Future perspectives 88

5 Discussion 90

Bibliography 94

A Supporting Information - Chapter 1 109
A.1 Properties of MRCAT and SSEE matrices 109
A.1.1 The N-1 distinct times 109
A.1.2 Canonical form of matrices 110
A.2 Constructing the tree with the different methods 112
A.2.1 Asexual, maternal and paternal MRCAT 112
A.2.2 SSEE 114
A.2.3 Hermaphroditic MRCAT and Genetic distances 115
A.3 The γ-statistic 117

B Supporting Information - Chapter 2 118
B.1 Extinction rate 118
B.2 Time to divergence as a function of B and G/B 121
B.3 The Asymmetric Yule model 123
B.4 γ-statistic 126
B.4.1 Pure-birth model 126

B.4.2 γ-statistic for Yule Model 130
B.5 The α value as an acceleration of speciation 131
B.6 Phylogenetic trees estimated from empirical data 134
B.7 Spatial patterns of speciation 140
B.8 Abundance distributions 141
B.9 Parameter inference 142

C Supporting Information - Chapter 3 **144**
C.1 Empirical information about platyrrhine monkeys 144
C.2 Additional results 152

D Bioética e Biossegurança **157**

E Direitos autorais **158**

Chapter 1

Introduction

The diversity of species on earth is one of the most striking features among nature's phenomena. Unconvering the processes responsible for the origin of the incredible diversity of species observed today is a goal of evolutionary biologists. A particularity of speciation process that makes it challenging to understand is its paradoxal nature, with different evolutionary factors (such as mutation, genetic drift, natural selection) acting simultaneously and having often opposite effects on the dynamics of speciation (Gavrilets, 2014). The use of theoretical models is a necessary step to untangle the importance of each evolutionary factor to the emergence of species diversity in a variety of systems. Questions about the conditions for speciation, the probability of speciation (Gavrilets et al., 2000b), the waiting time to speciation (Gavrilets, 2000), the degree of genetic divergence between emerging species (Higgs and Derrida, 1992), can only be answered with the help of theoretical models (Gavrilets, 2014).

Particularly, Individual-Based Models (IBM) – which simulate populations as being composed of discrete agents representing individuals or groups of similar individuals with sets of traits that vary among the agents – are crucial in the development of evolutionary theory, in the sense they allow to include explicitly individual variation in a greater detail than classical analytical models (DeAngelis and Mooij, 2005). The incorporation of microevolutionary processes like mutation, genetic drift, recombination, at the level of individual organisms make IBMs more flexible than classical models, allowing them to mimic real populations with small numbers of individuals, where stochasticity is important (DeAngelis and Mooij, 2005).

Speciation models have been broadly studied in the light of the Neutral Biodiversity Theory (NBT) (Kopp, 2010). The neutral biodiversity theory (Hubbell, 2001) posits that macroecological patterns like the distribution and abundance of species do not depend on adaptive differences between species, but only depend on random fluctuations in population sizes and dispersal (Kopp, 2010). Also, at the population level, differences between individuals of the same species are irrelevant for birth, death and dispersal rates (Hubbell, 2001; Gavrilets, 2014). Modeling neutral evolution, a strand of

NBT, is important to provide a theoretical reference about the behavior of a population evolving without the influence of deterministic processes, such as natural selection. The study of this benchmark enables, by contrast, to unravel the features that are a consequence of natural selection itself. Before going deeper into the speciation models it is helpful to understand how variability arises in neutral models at the population level.

An example of neutral model is the population model of asexually reproducing individuals evolving in a flat fitness landscape proposed by Derrida and Peliti (1991). A fitness landscape is flat when most evolutionary changes at the molecular level are neutral, which means that genetic differences do not imply in changes in individual fitness (Kimura, 1983; Gavrilets, 2004). The stochastic nature of this model establishes an equivalence with dynamical systems, allowing the computation of quantities like genetic variability and genealogy statistics and the prediction of population features (Derrida and Peliti, 1991). The model consists of M individuals represented by their genomes, which reproduce asexually and leave offsprings in their places (non-overlapping generations). The source of variability is the probability μ of point mutations that can happen when the genome of offsprings are constituted. The genetic variability of the population is measured by the statistics of overlap $q^{\alpha,\beta}$ between two individuals (α and β), representing the degree of similarity between its genomes. With this statistic they compute the population average similarity, where the comparison runs over all distinct pairs of individuals in the population. Additionally, they take the population average similarity over a very long time range, obtaining the time average.

The link between this population model and dynamical systems consists of the population average being analogous to the thermal average and the time average being analogous to the average over disorder (Derrida and Peliti, 1991). They observed that the population average fluctuates along time, even for large populations, due to the stochastic nature of the model, which is called in system dynamics “lack of self-averaging”. They show how these properties appear in the explicit solution of the model, inspired in some previous works (Fontana and Schuster, 1987; Amitrano et al., 1989; Zhang et al., 1990). An equivalent population model with sexual reproduction was presented in Serva and Peliti (1991). The main difference between the results of these models is that in the sexual case the genetic distance between individuals does not fluctuate in the infinite-population limit, so for very large populations they behave distinctly.

These population models of genetic variability can be expanded to speciation models that have foundations on Dobzhansky's idea (Dobzhansky, 1936; Dobzhansky, 1937) about the appearance of genetic incompatibilities due to genetic variation and the evolution of reproductive isolation as a side effect of genetic divergence (Gavrilets, 2014). The Bateson- Dobzhansky-Muller (BDM) model of genetic incompatibilities (Bateson, 1909; Dobzhansky, 1937; Muller, 1939; Muller, 1942) posits that the interaction between specific alleles could generate genetic incompatibilities, in the sense that their interaction "produces one of the physiological isolating mechanisms" (Dobzhansky, 1937; Gavrilets, 2014). As a consequence, two isolated subpopulations that produce each one of these alleles when come into contact again become reproductively isolated populations. As the BDM model consider a $2 - locus$, $2 - allele$ diploid population, the interaction of only few alleles brings to the emergence of genetic incompatibilities and reproductive isolation.

For models with multiple *loci* it is necessary to extend this idea to the notion of accumulation of genetic incompatibilities. One prediction of the BDM model is that genetic divergence between separated subpopulations results in a faster than linear growth in the number of genetic incompatibilites – the rapid accumulation of genetic incompatibilies, known as Snowball Effect (Orr, 1995; Orr and Orr, 1996; Gavrilets, 2014). With the incompatibilities accumulating fast, the second prediction is the occurrence of a rapid transition in the degree of reproductive isolation from low to high – the Threshold Effect (Gavrilets, 2004). The threshold effect is only possible because of the non-linearity in the growth of genetic incompatibilities expressed in the snowball effect. These two phenomena together could explain the emergence of reproductive incompatible subpopulations due to the emergence of genetic incompatibilities.

As expressed by Gavrilets (2014), "the threshold effect can be used to justify a simple model of reproductive isolation in which 2 populations are reproductively compatible as long as the genetic distance between them remains below a particular constant K. Once genetic distance reaches K, the populations become reproductively isolated". Models based on the threshold effect assumption have been used for studying different mechanisms of speciation, especially in a geographical framework (Higgs and Derrida, 1991; Gavrilets et al., 1998; Gavrilets, 2000; Gavrilets et al., 2000b; Yamaguchi and Iwasa, 2013). However, only few models incorporate the space explicitly in simulations (spatially explicit models), a topic considered neglected which can give important contributions to the understanding of speciation processes dependent on geography (Hoelzer et al., 2008; De Aguiar et al., 2009; Desjardins-Proulx and Gravel, 2011;

Melián et al., 2012). Earlier work on this topic showed that the interaction between isolation by spatial and genetic distances led to a rapid emergence of species through a neutral process that generates diversity from mutation and genetic drift (Gavrilets et al., 1998; Gavrilets et al., 2000a; Gavrilets et al., 2000b).

Natural systems in which space may have been crucial during speciation processes include, for example, African cichlids (Kocher, 2004), Caribbean anolis lizards (Losos and Thorpe, 2004), Darwin's finches (Grant, 1999), and the New World monkeys (Schneider and Sampaio, 2015). Cichlids inhabiting African lakes exhibit low levels of migration, due to lack of larval or juvenile dispersal and high philopatry in adults. These features lead to a high degree of population structure, promoting parapatric speciation (Kocher, 2004). The Caribbean anole lizards are another group in which space could play a role in speciation. With more than 150 species living in Caribbean islands today, most of them evolved from a single colonizing species over millions of years. The occupation of different geographic regions in these islands have led to changes in traits as the colorful flap of skin under lizards'throats, which in turn cause reproductive isolation of species (Losos and Thorpe, 2004).

The association of mechanistic models which incorporate the threshold effect for reproductive isolation in the context of the neutral biodiversity theory (Hubbell, 2001) confirms the general ability of the neutral theory to predict empirical patterns of biodiversity (Kopp, 2010). The unified neutral biodiversity theory (NBT) (Hubbell, 2001) put in check the question of whether adaptive differences between species are necessary for explaining large-scale patterns of biodiversity. Theory derived macroecological patterns like species-abundance distribution (SAD) and species-area relationship (SAR) but focused only in patterns, neglecting the processes that generate them. Mechanistic models strengthened the NBT and showed that SAD and SAR patterns are in agreement with empirical data. The model proposed by De Aguiar et al. (2009) addresses both the NBT requirements of ecological equivalence assumption and the importance dispersal limitation. The model is also mechanistic in the sense it is an individual-based model (IBM) which is spatially explicit and incorporates a genetic model of reproductive isolation based on the threshold assumption. This model is considered "the first attempt to relate a population genetics speciation model to NBT" (Kopp, 2010).

De Aguiar et al. (2009) performed simulations of a community living on a rectangular grid with $L \times L$ dimensions, representing sites in which individuals are placed. The initial condition is represented by a homogeneous population randomly distributed. Each individual has a genome consisting of B binary *loci*, as explained for the Higgs and Derrida (1991) model above. The De Aguiar et al. (2009) model is very similar to that model, but with sexual reproduction as in Serva and Peliti (1991). The population evolves under the combined effects of sexual reproduction, mutations and dispersal (De Aguiar et al., 2009). Each individual searches for mates inside its mating range, a circular area around its location. Compatible mates are defined by the threshold assumption, with individuals being reproductively compatible only if they have a genetic distance below a threshold G. Once a mate is chosen, reproduction happens, with the offspring genome composed by recombination of parental genomes succeeded by a probability of mutation μ for each *locus*. Finally, the offspring has a probability D to disperse to the neighboring sites in which its expiring parent was located. Under suitable conditions (spatial and genetic restrictions) speciation emerges from neutral processes as mutation and genetic drift combined with isolation-by-distance, without the requirement of geographic barriers or the action of natural selection (De Aguiar et al., 2009; Kopp, 2010).

The patterns of biodiversity resulting from this speciation process show the same patterns predicted by the NBT. The SAD is lognormal-like with an excess of rare species, and the triphasic SAR shows a rapid increase in species richness at small spatial scales, followed by a power law at intermediate scales (De Aguiar et al., 2009, Fig. 3). These patterns provide good fits to empirical data from, among others, British birds, moths and Panamanian trees. In a later study the same authors investigated the effect of barriers on biodiversity patterns by simulating a neutral model of ring species formation, with the expansion of a population around the barrier (Martins et al., 2013). Their results match with empirical patterns of the greenish warblers' complex, an example of extant ring species, regarding the distribution of subspecies, the principal components of genetic diversity, and the linear spatial-genetic correlation of the data (Martins et al., 2013). They also suggested that the warbles'ring will break up into multiple species in 10,000 to 50,000 years. This prediction is an example of the importance that IBMs can exercise to understand and describe empirical complex processes, even without claiming non-neutral forces.

As pointed out by Kopp (2010) and Manceau et al. (2015), the speciation process modeled in these frameworks leave signatures not only on macroecological patterns but also in the structure of phylogenetic trees. However, these signatures have only been investigated in a few models, like the point-mutation model (Hubbell, 2001) and the protracted speciation model (Rosindell et al., 2015; Manceau et al., 2015). The first chapter of this thesis aims to fill this gap using a modified version of De Aguiar et al. (2009) model to register the historical dimension of the evolutionary dynamics in the speciation process. Building algorithms that record the ancestor-descendant relationships and the speciation and extinction events my colleagues and I provide in this neutral spatially explicit model of speciation, based on the threshold assumption, the generation of genealogies and phylogenies resulting from the processes modeled. The generated trees allow us to verify the existence of macroevolutionary patterns that could serve as a referential of neutral evolution.

With the algorithm developed in Chapter One we have the tools to search for possible signatures that different speciation processes can leave in macroevolutionary patterns, which was the scope of Chapter Two. Specifically, evolutionary radiations were simulated and it was observed if different geographical frameworks of speciation (parapatric and sympatric) leave distinct signatures in macroevolutionary patterns exhibited in the phylogenies. Finally, in Chapter Three, to further explore the consequences of breaking the spatial homogeneity I modified the previous models (De Aguiar et al., 2009; De Aguiar, 2017) adding spatial barriers different from that implemented in Martins et al. (2013), simulating rivers in a spatial grid representing South America. This new version of the model was studied in an applied context, wondering if spatial heterogeneity in the shape of rivers can explain alone the distribution and macroevolutionary patterns of subfamilies and genera of the New World monkeys (Parvorder Platyrrhini), taxa which have been hypothesized to speciate from neutral processes. I hope with these three frameworks do contribute to the understanding of the role of neutral processes in shaping the striking diversity scattered upon nature.

Chapter 2

Registering the evolutionary history in individual-based models of speciation

Article published by: Carolina L.N. Costa, Flavia M.D. Marquitti, S. Ivan Perez, David M. Schneider, Marlon F. Ramos, Marcus A.M. de Aguiar. 2018 "Physica A" doi.org/10.1016/j.physa.2018.05.150.

Abstract

Understanding the emergence of biodiversity patterns in nature is a central problem in biology. Theoretical models of speciation have addressed this question in the macroecological scale, but little has been done to connect microevolutionary processes with macroevolutionary patterns. Knowledge of the evolutionary history allows the study of patterns underlying the processes being modeled, revealing their signatures and the role of speciation and extinction in shaping macroevolutionary patterns. In this paper we introduce two algorithms to record the evolutionary history of populations and species in individual-based models of speciation, from which genealogies and phylogenies can be constructed. The first algorithm relies on saving ancestor-descendant relationships, generating a matrix that contains the times to the most recent common ancestor between all pairs of individuals at every generation (the Most Recent Common Ancestor Time matrix, MRCAT). The second algorithm directly records all speciation and extinction events throughout the evolutionary process, generating a matrix with the true phylogeny of species (the Sequential Speciation and Extinction Events, SSEE). We illustrate the use of these algorithms in a spatially explicit individual-based model of speciation. We compare the trees generated via MRCAT and SSEE algorithms with trees inferred by methods that use only genetic distance between individuals of extant species, commonly used in empirical studies and applied here to simulated genetic data. Comparisons between trees are performed with metrics describing the overall topology, branch length distribution and imbalance degree. We observe that both MRCAT and distance-based trees differ from the true phylogeny, with the first being closer to the true tree than the second.

Keywords genealogies of individuals | phylogenies of species | macroevolutionary patterns | distance-based trees | tree statistics

2.1 Introduction

The origin of the patterns of diversity at macroecological scale is a central problem in biology (Coyne and Orr, 2004; De Aguiar et al., 2009; Gavrilets, 2014). In the last decades patterns such as geographical variation in species richness, species abundance distributions and species-area relationships, have been studied from empirical and theoretical perspectives (Turelli et al., 2001; Field et al., 2009; Martins et al., 2013; May et al., 2015; Kopp, 2010). Neutral models of speciation – where differences between individuals are irrelevant for their birth, death, and dispersal rates (Gavrilets, 2014; Hubbell, 2001) – have played a central role in understanding the patterns of diversity at the macroecological scale. With the help of computers, it became possible to test different hypothesis about the mechanisms of speciation, such as sympatric versus allopatric processes, assortative mating and the effect of number of genes (Gavrilets et al., 2000b; Dieckmann and Doebeli, 1999; Rettelbach et al., 2013).

Among the different theoretical approaches designed to quantitatively study speciation (Gavrilets, 2014; Gavrilets, 2003), models that explicitly incorporate space have allowed the study of major macroecological patterns that could be compared with those observed in nature (De Aguiar et al., 2009; May et al., 2015; Campos et al., 2012; Martín and Goldenfeld, 2006). However, these models have given little attention to the historical or evolutionary dimension of the origin of diversity, which is reflected in the macroevolutionary patterns described by phylogenetic trees (Manceau et al., 2015; Pigot et al., 2010; Hagen et al., 2015; Quental and Marshall, 2011). Because of the increased interest in the role of microevolutionary processes on the resulting macroecological patterns, the extension of these approaches to include algorithms that track the branching or phylogenetic divergence process is a next fundamental step to further explore models of speciation using simulations (Manceau et al., 2015; Davies et al., 2011; Rosindell et al., 2015). Individual-based models (IBM) widely used in biology (DeAngelis and Grimm, 2014) have the advantage that can be easily extended to include this historical perspective and to provide a record of the ancestor-descendant relationships among the simulated individuals and/or species. These relationships can be stored in matrices from which individual genealogies and species trees (i.e. phylogenies) may be directly obtained.

In this article we describe two algorithms that save historical information in individual-based models of speciation. The first algorithm focuses on genealogies and the quantity saved is the parenthood of each individual. With parenthood registered, the *time to the most recent common ancestor*, i.e., the number of generations needed to go backward to find a common ancestor of one individual with another individual of the population, can be easily calculated in terms of the common ancestor of the parents. These times are computed at every generation between all pairs of individuals and, at the end of the simulation, are saved in a matrix (the Most Recent Common Ancestor Time matrix - MRCAT). The second algorithm focuses on phylogenies and consists of directly records all speciation and extinction events (the Sequential Speciation and Extinction Events - SSEE) and set a matrix analogous to MRCAT but whose entries are species rather than individuals. The SSEE matrix contains the exact branching times in the simulated clade or community, including all extinct species. The MRCAT and SSEE matrices can be used to draw the exact branching sequence of the simulated individuals and species, respectively. These procedures differ from the inference methods based on phenotypic and genetic traits used to estimate phylogenies in natural studies, because in our model we are looking for the branching process forward in time, while in usual approaches the same process is looked backwards in time. In addition to the presentation of the MRCAT and SSEE algorithms, we compare the trees they generate with those obtained by usual distance-based methods of phylogenetic inference using only genetic data from simulated individuals of the final community. Comparing these inferred phylogenies with those generated by MRCAT or SSEE algorithms might offer a practical way to evaluate the reliability of the estimated trees to recover natural macroevolutionary patterns.

The paper is organized as follows: in section 2.2 we describe the algorithms to record ancestor-descendant relationships (MRCAT, subsection 2.2.1) and speciation/extinction events (SSEE, subsection 2.2.2). In subsection 2.2.3 we compare the true phylogenetic tree obtained from the SSEE algorithm with genealogies of individuals obtained from the MRCAT algorithm considering only one individual per species. In section 2.3 we discuss the applications of the algorithms proposed in section 2.2. First, we present an individual-based model of speciation proposed in (De Aguiar et al., 2009) in which the algorithms regarding the ancestor-descendant relationships and the branching process were incorporated (subsection 2.3.1). We emphasize that the algorithms are quite general and could be implemented in most IBM's. Next, we briefly describe the Unweighted Paired Group Method with Arithmetic mean (UPGMA) (Murtagh, 1984), the Neighbor Joining (NJ) (Saitou and Nei, 1987) and the Minimum Evolution (ME) (Rzhetsky and Nei, 1993) methods, which are based on genetic distances calculated directly from one individual of each species present in the last generation of the simulation (subsection 2.3.2). While closer to what empiricists do, the phylogenies derived from these methods are further from the true phylogeny generated by the SSEE algorithm than is the phylogeny based on the MRCAT algorithm presented here. We end this section presenting the statistical measurements used to compare phylogenies obtained from algorithms proposed here with those estimated by distance-based methods (subsection 2.3.3). The goal is to show that the accuracy of some methods usually employed when the only information available is the data of individuals collected from nature can be evaluated with the help of models. In section 2.4 we present the results regarding the output of simulations and the comparisons of phylogeny summary statistics. Finally, section 2.5 was devoted to discussion and section 2.6 to conclusions.

2.2 Registering the history of individuals and species

In this section we describe two algorithms to record historical information during the evolution of a population. The first algorithm records genealogical relationships between all pairs of individuals at every generation. The second, in turn, registers all the speciation and extinction events that occur along the evolutionary history. These algorithms are general enough to be applied to most individual-based models of speciation.

2.2.1 Ancestor-descendant relationships among individuals - MRCAT

In this subsection we show how the time to the most recent common ancestor between all pairs of individuals can be obtained by keeping track of parental relationships at every generation. We also show how this information can be used to draw the genealogy of individuals of the last simulated generation. We distinguish between asexual and sexual models because of the technical differences in tracking only one or two parents.

Individuals at generation $t+1$	Parent at generation t
1	$P(1) = 4$
2	$P(2) = 8$
3	$P(3) = 1$
4	$P(4) = 4$
...	...
N_{t+1}	$P(N_{t+1}) = 15$

TABLE 2.1: List of individuals (i) at generation $t+1$ and their respective parents ($P(i)$) at generation t in an asexual model. This information is necessary to construct the MRCAT matrix. Parents of each individual must be recorded to track the most recent common ancestor between individuals at the end of a simulation. Note that individuals at generation t are not the same individuals at generation $t+1$ (discrete generations).

Asexual models

Consider a population of N_t asexual individuals at generation t. The population at the next generation, $t+1$, will be comprised of offspring of these individuals and the parent of individual i will be denoted $P(i)$.

An example is shown in Table 2.1, where $P(1) = 4$, $P(2) = 8$, $P(3) = 1$, etc. The MRCAT between individuals i and j is

$$T_{t+1}(i,j) = T_t(P(i), P(j)) + 1. \tag{2.1}$$

which is simply the time to the most recent common ancestor between the parents plus one, since a generation has passed (Higgs and Derrida, 1992). As examples

$$T_{t+1}(1,2) = T_t(4,8) + 1$$

and

$$T_{t+1}(1,4) = T_t(4,4) + 1 = 1.$$

since in this last case they have the same parent. Starting from $T_0(i,j) = 1$ if $i \neq j$ and noting that $T_t(i,i) = 0$ at all times the rule (2.1) allows one to compute the MRCAT matrix for any number of generations. The matrix T is stored only for two times, the past and the present generation, so that the memory cost does not depend on time, only on the (square) size of population. A schematic view of the algorithm is shown in Fig. 2.1, where the genealogical relationships between 9 individuals originated from a single ancestor is represented. In this example the total population size is kept fixed,

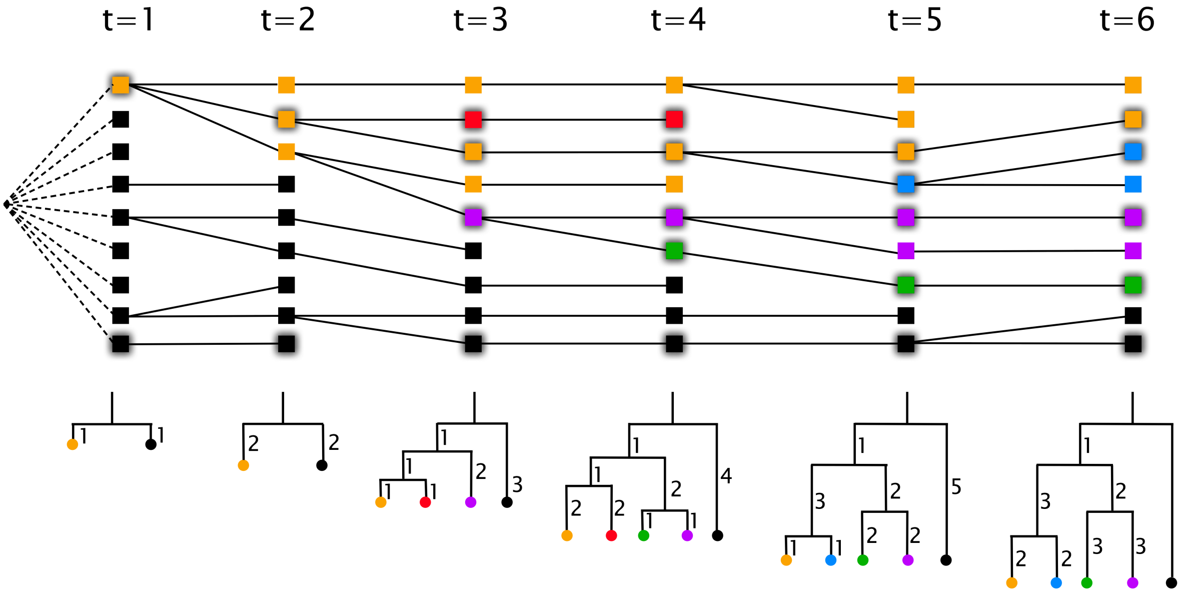


FIGURE 2.1: Illustration of ancestor-descendant relationships for an asexual population with constant size $N = 9$ implemented with MRCAT algorithm. Each square is an individual and colors represent different species. Phylogenetic trees are constructed by selecting one individual per species (shaded squares).

so that the full MRCAT matrix is always 9×9. The phylogeny of the community can be drawn by selecting one individual per species at each moment in time. The corresponding matrices at $t = 3$ and $t = 6$ are given by

$$T_3 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 2 & 3 \\ 2 & 2 & 0 & 3 \\ 3 & 3 & 3 & 0 \end{pmatrix}; \qquad T_6 = \begin{pmatrix} 0 & 2 & 5 & 5 & 6 \\ 2 & 0 & 5 & 5 & 6 \\ 5 & 5 & 0 & 3 & 6 \\ 5 & 5 & 3 & 0 & 6 \\ 6 & 6 & 6 & 6 & 0 \end{pmatrix}. \qquad (2.2)$$

where the selected individuals are shown in shaded colors (from top to bottom) at the corresponding times.

Sexual models

The generation of MRCAT matrices in sexual models is slightly different, since each individual i has two parents, a mother $P_1(i)$ and a father $P_2(i)$. Consider as

Individuals at generation $t+1$	Mother at generation t	Father at generation t
Females		
1	$P_1(1) = 4$	$P_2(1) = 6$
2	$P_1(2) = 3$	$P_2(2) = 7$
3	$P_1(3) = 1$	$P_2(3) = 7$
4	$P_1(4) = 4$	$P_2(4) = 5$
5	$P_1(5) = 2$	$P_2(5) = 6$
Males		
6	$P_1(6) = 1$	$P_2(6) = 5$
7	$P_1(7) = 3$	$P_2(7) = 5$
8	$P_1(8) = 3$	$P_2(8) = 7$

TABLE 2.2: List of individuals (i) at generation $t+1$ and their respective parents ($P_1(i) = mother$ and $P_2(i) = father$) at generation t in a sexual model. In this case each individual has two parents, P_1 and P_2. Notice that the couple 3 and 7 at generation t had two offspring, the individuals 2 and 8 at generation $t+1$, while other couples had only one offspring. Additionally, notice that there were 4 females and 3 males at generation t, while there are 5 females and 3 males at generation $t+1$.

an example a population which has 4 females and 3 males in generation t and gives rise to 5 females and 3 males in generation $t+1$ (Table 2.2). Notice that not only the total number of individuals but also the number of males and females may vary over generations. As the model is sexual, both maternal and paternal lineages can be followed in the simulations, allowing the generation of two different MRCAT matrices and their corresponding trees. A third option is not tracking lineages by sex, but record the most recent common ancestor taking into account both parents, which is the only option if the model considers hermaphroditic individuals.

– *Maternal and paternal lineages.* The maternal lineage of individuals is obtained by computing the time to the most recent common ancestor of their corresponding mothers:

$$T^M_{t+1}(i,j) = T^M_t(P_1(i), P_1(j)) + 1 \tag{2.3}$$

with $T^M_0(i,j) = 1$ if $i \neq j$ and $T^M_t(i,i) = 0$. Similarly, the paternal lineage is computed with

$$T^F_{t+1}(i,j) = T^F_t(P_2(i), P_2(j)) + 1 \tag{2.4}$$

with $T^F_0(i,j) = 1$ if $i \neq j$ and $T^F_t(i,i) = 0$. Both T^M and T^F are computed for all individuals, females and males.

– *Lineages of hermaphroditic individuals.* Many simulations consider, for simplicity, hermaphroditic individuals. In this case, the separation into maternal and paternal lineages does not make sense and the definition of the MRCAT matrix is

$$T_{t+1}(i,j) = \min_{\{k,l\}}\{T_t(P_k(i), P_l(j))\} + 1 \tag{2.5}$$

with $k, l = \{1, 2\}$, $T_0(i,j) = 1$ and $T_t(i,i) = 0$. This considers, literally, the most recent common ancestor of i and j, taking all parental combinations into account. The same definition is applied to sexual models with sex separation when the recorded genealogy does not separate the maternal and paternal lineages. In the case of hermaphroditic model the MRCAT matrix does not determine the tree uniquely. A detailed example of this situation is described in Supporting Information A, section A.1.

Drawing genealogies from MRCAT matrices

At the end of the simulated evolutionary process the MRCAT matrix contains the time to the most recent common ancestor between every pair of individuals of the extant population and this information can be used to draw genealogical trees. Drawing the tree from the MRCAT matrix consists in joining individuals into groups according to their most recent common ancestral (Fig. 2.1). The tree starts with *N units* (the extant individuals) and at each step of the process two of these units are joined together to form a group, so that the number of units decreases by 1. Next, the time to the most recent common ancestral between the newly formed group and the other units of the tree (previously formed groups or extant individuals) are recalculated with a so called *clustering method*. Once the times have been recalculated, the pair of units with the least time is joined into a new group. The process ends when a single unit is left, the root of the tree. As discussed in the SI A, section A.1, a unique tree is generated independently of the clustering method for asexual, maternal or paternal lineages. For hermaphroditic populations or for sex separation but with the MRCA taking into account both parents that is not the case. In these situations more than one tree can be constructed from the same MRCAT matrix using different clustering procedures. In all cases the tips (or leaves) of the tree represent extant individuals whereas internal nodes represent the most recent common ancestor between a pair of individuals. Branch length denote the time in generations between an ancestor and its descendants

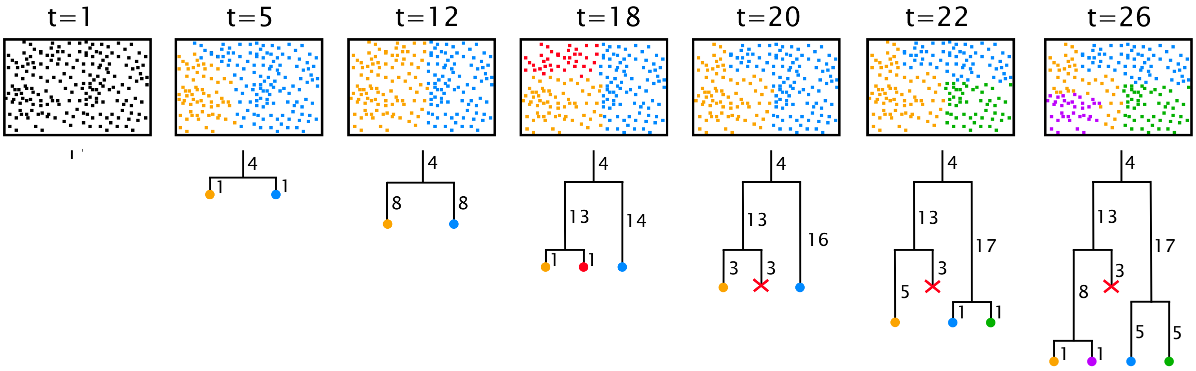


FIGURE 2.2: Illustration of speciation and extinction events implemented with SSEE algorithm and the corresponding phylogenetic trees exhibiting the complete history. Colored squares represent individuals of different species, and colored circles in phylogenies represent each species, with numbers denoting the time to speciation and extinction events.

(see, for instance, Fig. A.1 in the SI A). More information about the drawing of trees is available in Supporting Information A, section A.2.

2.2.2 Recording all speciation and extinction events - SSEE

The algorithm described in subsection 2.2.1 records the ancestor-descendant relationships between all pairs of individuals in the population at a given point in time. This allows the drawing of entire genealogies. However, information about individuals that died without leaving descendants or species that went extinct is totally lost. In this subsection we describe an algorithm that allows the construction of the true phylogenetic tree, retaining information about all species that ever existed during the evolution (Fig. 2.2). [!htpb]

We will use a new matrix S_t (the SSEE matrix) such that $S_t(i,j)$ is the time when species i and j branched off a common ancestral species. Species that go extinct will be kept in the matrix but will be assigned a label to distinguish them from living (extant) species. This label will be stored in a *extinction vector* E_t such that $E_t(i) = 0$ indicates a living species at time t and $E_t(i) = \tau \neq 0$ indicates the moment τ when the species disappeared.

The algorithm is as follows: consider the hypothetical sequence of speciation and extinction events displayed in Fig. 2.2. At time t=18 there are three species that we denote as Orange(18), Red(18) and Blue(18) and the corresponding S matrix and E vector are

$$S_{18} = \begin{pmatrix} 0 & 1 & 14 \\ 1 & 0 & 14 \\ 14 & 14 & 0 \end{pmatrix}; \qquad E_{18} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{2.6}$$

Two generations later, at $t = 20$, one finds only two species, Orange(20) and Blue(20). Notice that names (and colours) are arbitrary and to determine the relation between these species and the ones at the previous time step we need to look at the parents of individuals in each species. Suppose, as illustrated in the figure, that we find that the parents of individuals in Orange(20) belonged to species Orange(18). In this case we draw a link between Orange(18) and Orange(20) and mark Orange(18) as a species that 'survived' that time step, i.e., we set $E_{20}(1) = 0$. Similarly Blue(20) links with Blue(18) and $E_{20}(2) = 0$. Looking at the previous generation we notice that species Red(18) did not leave any descendant species, i.e., it went extinct. In order to keep track of it we create a virtual species Red(20) and set $E_{20}(3) = 20$ as a mark that it is no longer a living species and went extinct at time 20. The SSEE and E vector at time 20 become

$$S_{20} = \begin{pmatrix} 0 & 16 & 3 \\ 16 & 0 & 16 \\ 3 & 16 & 0 \end{pmatrix}; \qquad E_{20} = \begin{pmatrix} 0 \\ 0 \\ 20 \end{pmatrix}. \tag{2.7}$$

Extinct species are, therefore, treated as species that will never again speciate, but will be kept in the matrix. When drawing the corresponding tree its branch will stop at the value $E(i)$. Proceeding in this way, with the living species always filling the first part of the matrix, followed by copies of extinct species, we can draw the complete phylogeny and study extinction dynamics as well. At time $t = 26$ the SSEE matrix and extinction vector E are

$$S_{26} = \begin{pmatrix} 0 & 1 & 22 & 22 & 9 \\ 1 & 0 & 22 & 22 & 9 \\ 22 & 22 & 0 & 5 & 22 \\ 22 & 22 & 5 & 0 & 22 \\ 9 & 9 & 22 & 22 & 0 \end{pmatrix}; \qquad E_{26} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 20 \end{pmatrix}. \tag{2.8}$$

One important case occurs when two species merge into a single species (speciation reversal). This might happen, for instance, when two species that have just become reproductively isolated are able to breed again because of a mutation. The resulting merged species will have individuals with parents in both ancestral species and we need to define which one "survived" and which went extinct. Although this is just a matter of labeling the species, we call the surviving species the one with most parents in the previous generation.

The drawing of species phylogenies for SSEE matrices is almost identical to that for MRCAT matrices. The only differences are that internal nodes represent speciation events, not the time to MRCA, and branches associated to extinct species should not be drawn all the way down to present time, but should stop at the extinction time recorded in the vector E. As in the MRCAT case of separation of lineages by sex, a unique tree is generated independently of the clustering procedure chosen, due to the exact times of speciation and extinction recorded in simulations based on this algorithm.

2.2.3 Phylogenies generated by ancestor-descendant relationships (MRCAT) versus trees from speciation and extinction events (SSEE)

At the end of a simulation the MRCAT matrix contains the exact time to the most recent common ancestor between every pair of individuals in the population. The SSEE matrix contains the equivalent information at the species level, including extinct species. Both matrices can be used to draw phylogenetic trees. To draw a phylogeny of species considering the ancestor-descendant relationships between individuals we can use the MRCAT matrix with the following reasoning: if N_S species exist at time t and $ind(i,j)$ is the j-th individual of the i-th species, a $N_S \times N_S$ sub-matrix of the full MRCAT matrix can be generated considering only one individual per species (Fig. 2.1); a simple choice is to take $ind(i,1)$ for $i = 1, 2, \dots N_S$ so that $T^{phy}_{i,j} \equiv T_{ind(1,i),ind(1,j)}$.

The tree drawn from the SSEE algorithm is the true phylogeny of species, because it records the exact speciation and extinction events, representing the actual branching process. On the other hand, the phylogeny of species drawn from the MRCAT algorithm is different, although similar, from the true phylogeny, because the time to the most recent common ancestor between individuals of different species is only an approximation to the speciation time, since speciation can happen several generations later. Figure 2.3 illustrates this situation: if a population splits into three species in two closely spaced speciation events, it might happen that the first group to speciate,

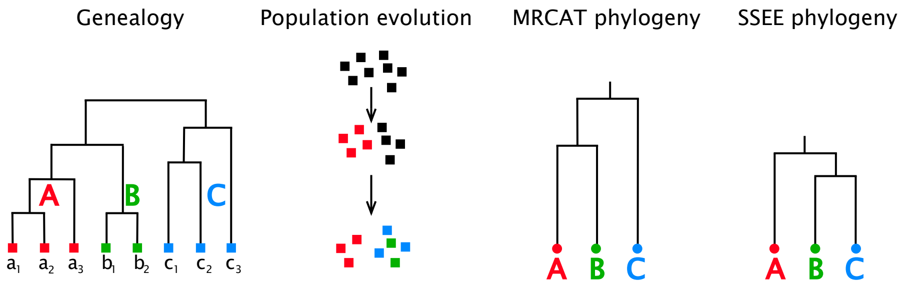


FIGURE 2.3: Illustration of a genealogy recorded with MRCAT and the corresponding population evolution. The phylogenies constructed via MRCAT and SSEE differ in this case because, although individuals from species A and B have a more recent common ancestor than with individuals in C, species A split first, followed by the separation of B and C.

species A in the figure, has a more recent common ancestor with the subgroup B than B with C. During the time when B and C still form a single species reproduction between their individuals might not happen for a while until they split, preserving the long time ancestry. This is more likely to happen in populations with a spatial structure when individuals belonging to the two subpopulations occupy different areas.

2.3 Applications of MRCAT and SSEE algorithms to an individual-based model

2.3.1 The speciation model

The model considered here to exemplify the MRCAT and SSEE algorithms is an extension of the speciation model introduced in (De Aguiar et al., 2009) and adapted in (Baptestini et al., 2013b) to characterize individuals with separated sexes (males and females). The model has already been studied in terms of speciation rates, species-area relationships and species abundance distributions. Here we are adding the historical information generated by MRCAT and SSEE algorithms, i.e., recording the parenthood of individuals from one generation to another (genealogy) as well as the pattern and time of the speciation and extinction events (phylogeny or time tree).

The model describes a population of N haploid individuals that are genetically identical at the beginning of the simulation and are randomly distributed in a $L \times L$ spatial lattice with periodic boundary conditions. More than one individual is allowed in each site of the lattice, but because the density of the population is low, this seldom occurs. The genome of each individual is represented by a sequence of B binary *loci*, with state 0 or 1, where each *locus* plays the role of an independent biallelic gene. Individuals also carry one separate label that specify their sex, male or female. The evolution of the population involves the combined influence of sexual reproduction, mutation and dispersal (De Aguiar et al., 2009).

The reproduction trial starts with individual 1 and goes to individual N, so that all individuals of the population have a chance to reproduce. The individual selected for reproduction, the *focal individual*, searches for potential mates in its *mating range*, a circular area of radius S centered on its spatial location. The focal individual can only reproduce with those within its mating range and if they are genetically compatible, i.e., if the genetic distance between them is below a particular threshold G. Among the compatible individuals within its mating range one of the opposite sex is randomly chosen as mating partner. Individuals whose genetic distance is larger than G are considered reproductively isolated (threshold effect, Gavrilets, 2014). Genetic distances between individuals are calculated as the Hamming distance (Hamming, 1950) between their genetic sequences, i.e., the number of *loci* at which the corresponding alleles are different.

Once the focal individual finds a compatible mate of the opposite sex, reproduction proceeds with the combination of their genetic materials to produce the offspring genome, with each *locus* having an equal probability of being transmitted from mother or father. After combination of parental genomes, each *locus* in the offspring genome can mutate with probability μ. Finally, the offspring replaces the focal reproducing individual. In each reproductive event only one descendant is generated. The offspring is then dispersed with probability D to one of the 20 nearest sites (radius approximately equal to $\sqrt{5} \approx 2.24$) around the expiring focal parent. Conversely, with probability $1 - D$ the offspring will be placed exactly in the same site of its focal expiring parent. Hence, close to the location of every individual of the previous generation there will be an individual in the present generation, keeping the spatial distribution homogeneous. There is a probability Q that the focal individual will die without reproducing. In this case a neighbor is randomly selected from its mating range to reproduce in its place, so that the population size remains constant.

Evolution proceeds in non-overlapping discrete generations such that the entire population is replaced by offspring. Species are defined as groups of individuals connected by gene flow, so that any pair of individuals belonging to different species are reproductively isolated (genetic distance greater than G). However, two individuals belonging to the same species can also be reproductively isolated, as long as they can exchange genes indirectly through other individuals of the species. This model is considered neutral because individuals choose their mates randomly from a mating range, independent of their genetic composition except for the genetic threshold of reproductive compatibility, so differences between individuals are irrelevant for their birth, death, and dispersal rates (Gavrilets, 2014; Hubbell, 2001).

2.3.2 Phylogenies based on genetic distances

As we have described in the previous subsection, the genome of all individuals are identical at the beginning of the simulation but mutations introduce differences and after many generations the population will display a distribution of genomes. Genetic distances can, therefore, be calculated between pairs of individuals and be used as a proxy for ancestry, such that the larger the genetic distance between two individuals the farther back should be their common ancestor. In order to estimate phylogenies by genetic distance, we selected the same individuals per species that were used to draw the phylogeny via MRCAT and computed a matrix of genetic distances. This process mimics the sampling of individuals from a real population and the comparison of their DNA's as a measure of ancestry.

From the genetic distance matrix, we estimated trees from three distance-based methods. Firstly, we used the UPGMA hierarchical clustering method (Murtagh, 1984). In this algorithm two groups of species are clustered based on the average distance between all members of the groups. This method assumes a constant rate of change, generating ultrametric trees in which distances from the root to all tips are equal. Secondly, we used the NJ method (Saitou and Nei, 1987) of phylogenetic inference. In this method the procedure is to find pairs of neighbors in which the total branch length at each stage of the clustering is minimal, starting with a starlike tree. Finally, we used the ME method (Rzhetsky and Nei, 1993), which assumes that the true phylogeny is probably the one with the smallest sum of branch lengths, as in the NJ method. The difference is that in the ME method a NJ tree is constructed first and next tree topologies close to this NJ tree are estimated by certain criteria, with all these trees being examined and the tree with the small sum of branch lengths being chosen. We used the function `hclust` of the `stats` package in R (R Core Team, 2017) to estimate ultrametric trees from the UPGMA method. To estimate trees from the NJ method, we used the `nj` function of the `ape` package in R (Paradis et al., 2004). In this case, the estimated trees are not ultrametric, so we transform then in ultrametric trees using the `chronoMPL` and `multi2di` functions in `ape` package (Paradis et al., 2004; Britton et al., 2002). We used the `Rkitsch` function of the `Rphylip` package in R (Revell and Chamberlain, 2014; Felsenstein, 2005) to estimate ultrametric trees from the ME method assuming an evolutionary clock. The NJ and ME methods are generally considered superior to UPGMA because they optimize a tree according to minimum evolution criteria. Similarly to the UPGMA, the NJ and ME methods are fast and efficient computationally.

2.3.3 Statistical indexes to compare phylogenies

To evaluate the accuracy of the phylogenies generated by the MRCAT algorithm and by the genetic distance methods (UPGMA, FM and ME) in relation to the true phylogeny generated by SSEE we use three statistics: the Robinson and Foulds (RF, Robinson and Foulds, 1981) metric, the gamma statistic (γ, Pybus and Harvey, 2000) and the Sackin's index (I_s, Sackin, 1972; Blum and François, 2005).

The RF metric measures the distance between phylogenetic trees, providing the overall topological resemblance of the phylogenies. Specifically, the RF metric calculates the number of internal branches present in only one of the trees being compared. Given two trees, $T1$ and $T2$, we define

$$RF(T_1, T_2) = \frac{L_1}{L_1'} + \frac{L_2}{L_2'} \tag{2.9}$$

where L_1 and L_2 are the number of branches on T_1 and T_2, respectively. The number of branches shared by T_1 and T_2 are represented by L_1' and L_2'. The RF metric was calculated using the `RF.dis` function of the `phangorn` package in R (Schliep, 2011).

The γ-statistic measures the distribution of branch lengths of a tree and is defined as (Pybus and Harvey, 2000):

$$\gamma = \frac{1}{D}\left[\frac{1}{N_S - 2}\sum_{k=2}^{N_S-1} T(k) - T(N_S)/2\right] \tag{2.10}$$

with

$$T(k) = \sum_{j=2}^{k} j g_j; \tag{2.11}$$

$$D = T(N_S)/\sqrt{12(N_S - 2)} \tag{2.12}$$

where N_S is the number of leaves and g_k is the time interval between speciation events as represented by the nodes of the tree (see Fig. A.4 in section A.3 of the SI A). The γ-statistic was calculated using the `gammaStat` function of the `ape` package in R (Paradis et al., 2004).

The Sackin index measures the degree of imbalance, or asymmetry, of a tree (Sackin, 1972; Blum and François, 2005). It is defined as

$$I_s = \sum_j d_j \tag{2.13}$$

in which d_j is the number of nodes to be traversed between each leaf j and the root, including the root (Dearlove and Frost, 2015). The expected Sackin index under a pure birth process (the Yule model, Yule, 1925) is

$$E(I_s(N_S)) = 2N_S \sum_{k=2}^{N_S} \frac{1}{k} \approx 2N_S \log N_S \tag{2.14}$$

where the approximation holds for N_S large (Blum and François, 2005). Since the expected value of the Sackin index increases with the tree size, a normalized index is defined to compare trees of different sizes:

$$I_s^n = \frac{I_s(N_S) - E(I_s(N_S))}{N_S}. \tag{2.15}$$

Here we used the normalized Sackin index to compare the phylogenies and calculated it using the `sackin` function of the `apTreeshape` package in R (Bortolussi et al., 2012).

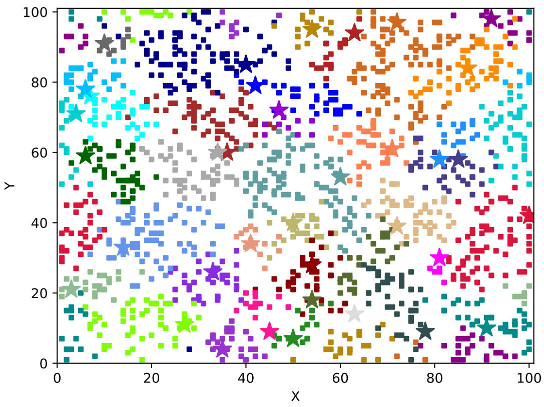


FIGURE 2.4: Spatial distribution of individuals from one simulation based on the model described in section 2.3.1. Individuals are represented by circles, and each color represents a different species. Stars indicate the individuals used to draw the phylogenies shown in figure 2.6.

2.4 Results

We ran simulations of the speciation model described in section 2.3.1 with parameters $N = 1500$, $L = 100$, $B = 150$, $S = 5$, $G = 7$, $\mu = 0.001$, $D = 0.05$, $Q = 0.05$. We start with the results of a single simulation to show examples of phylogenies. Figure 2.4 shows the population after 1000 generations, with squares representing individuals and colors indicating the 36 species generated. Species form spatial clusters, a consequence of the small S value used the simulation.

The true phylogenetic tree of the population, generated using the SSEE algorithm, is shown in Fig. 2.5. Figure 2.5(a) shows the full phylogeny, which includes all speciation and extinction events. The large number of events seen near the root of the tree correspond mostly to unsuccessful or incomplete speciation events, in which a group of individuals momentarily splits in two species but quickly recombines into a single species due to mutations. We distinguish these events from *true extinctions*, which are characterized by the collapse of a long living species by a sharp decline in population size. This phenomenon is very common at the beginning of the speciation process in the model described in section 2.3.1. In Fig. 2.5(b),(c),(d) the full phylogeny was filtered in order to remove speciation reversals and keep only *true* extinction events. In the model, extinctions occur by stochastic fluctuations in the number of individuals of a species, which might become very small and go to zero. Figure 2.5(b) shows the phylogeny filtered by the criterion of population size at the moment of

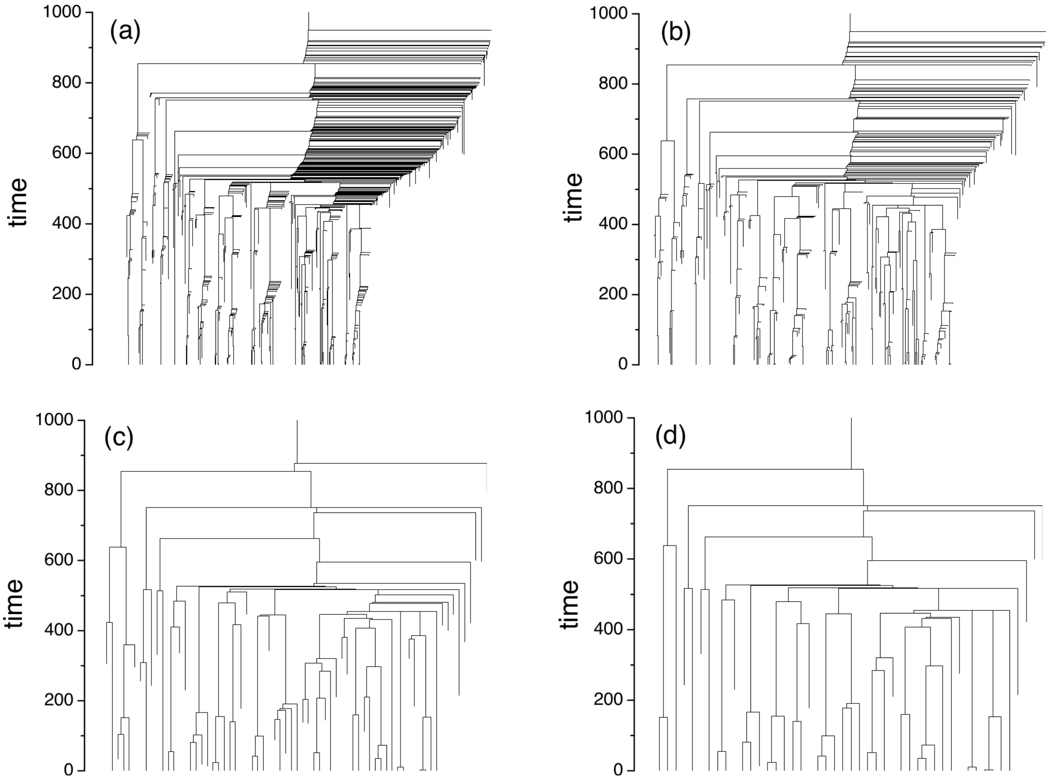


FIGURE 2.5: True phylogenies obtained with the SSEE method. (a) full phylogeny, including all speciation and extinction events; (b) filtered phylogeny, excluding branches (species) which had more than 20 individuals at the moment of extinction; (c) filtered phylogeny, excluding also branches that lasted less than 50 generations and (d) 100 generations.

vanishing: species that disappear with more than 20 individuals were considered speciation reversals and removed from the tree. Figures 2.5(c) and (d) display the same phylogenies but filtered also by the criterion of persistence in time: branches of species that lasted less than 50 generations (c) or 100 generations (d) were also removed.

Phylogenies computed from the SSEE, MRCAT and genetic distance matrices are shown in Fig. 2.6. Panel (a) shows the true SSEE phylogeny, filtered to exhibit only the extant species. Panel (b) was obtained from the MRCAT algorithm, with one individual from each species being selected to represent the species. We showed in section A.2 of the SI A (Fig. A.2) that the choice of the individual for constructing the phylogenetic tree with MRCAT can matter. However, the final structure of the tree will barely vary. Finally, panel (c) shows the phylogeny estimated from the genetic distance matrix of the same individuals used in Fig. 2.6(b) by the UPGMA clustering method. Differences in topology and branch lengths are qualitatively visible between these trees. Maternal and paternal genealogies obtained from the MRCAT algorithm are shown in Fig. A.3 in the SI A.

Statistical comparisons between phylogenies generated by the MRCAT algorithm and by the genetic distance methods (UPGMA, NJ and ME) in relation to the true phylogeny (SSEE) are shown in Fig. 2.7. The first line shows comparisons of topology (RF metric), branch length distribution (γ-statistic) and degree of imbalance (Sackin index) among phylogenies after 500 generations in 50 simulations. The second line shows the same comparisons after 1000 generations for the same 50 simulations. Colors represent the different methods utilized to generate the trees. In the RF scatterplots (Fig. 2.7(a)(b)) the coordinates of each point refer to the normalized topological distance between the tree calculated with the MRCAT matrix (y-axis) or by genetic distance matrix (x-axis) from the true phylogenies generated by the SSEE algorithm. Small values of RF indicate that phylogenies are closer to the true phylogeny (SSEE). The diagonal dotted line defines the condition in which the topology of the phylogenies (RF-value) was equal in trees generated by genealogical relationships (MRCAT trees) and that estimated by genetic distance (UPGMA, NJ and ME methods). The scatterplot for $T = 500$ (Fig. 2.7(a)) shows that phylogenies generated by MRCAT and genetic distance using UPGMA method (orange points) were similar in their RF-values, while trees estimated from NJ and ME methods (yellow and pink) had more different RF-values. For $T = 1000$ (Fig. 2.7(b)) all phylogenies estimated by genetic distance-based methods differ from those obtained by MRCAT. The density distribution of RF values shown above the scatterplots indicates that MRCAT is always closer

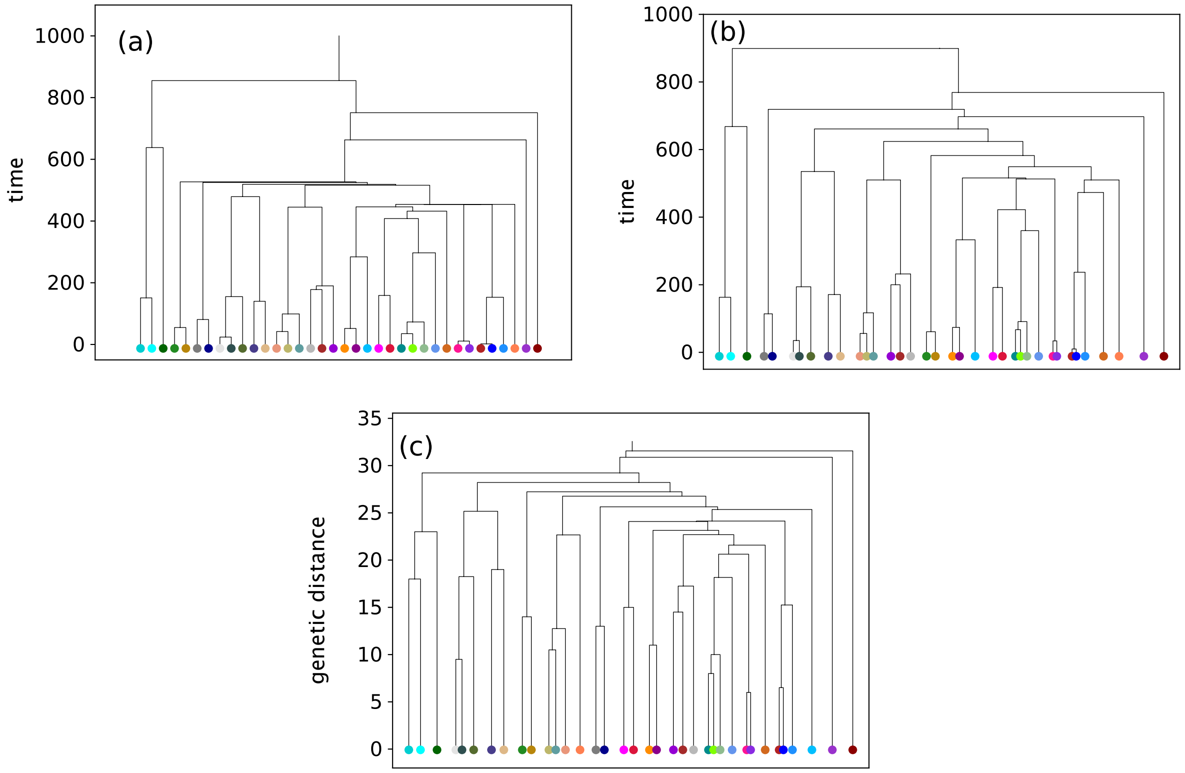


FIGURE 2.6: (a) Extant phylogeny obtained via SSEE (species are separated by one unit on x-axis); (b) via MRCAT; (c) via genetic distance matrix using UPGMA (neighbor species are separated by genetic distances). Colors correspond to species in Fig. 2.4.

to SSEE, especially for $T = 1000$.

Regarding the branch length distribution, the scatterplots (Fig. 2.7(c),(d)) show the difference between γ-values in SSEE phylogenies (y-axis) and MRCAT or genetic distance (UPGMA, NJ or ME) phylogenies (x-axis). The diagonal dotted line defines the condition in which the γ-values of trees generated by genealogical relationships (MRCAT trees) or by genetic distance (by UPGMA, NJ and ME methods) were equal to values of true phylogenies. We observe that for both times (Fig. 2.7(c),(d)) MRCAT trees had γ distributions closer to true phylogenies (SSEE) than all genetic distance-based trees, with a good match for $T = 1000$. Finally, the normalized Sackin index is presented in Fig. (Fig. 2.7(e),(f)). The imbalance of MRCAT phylogenies was closer to the true phylogenies for $T = 500$ (Fig. 2.7(e)). On the other hand, for $T = 1000$ the imbalance was similar for MRCAT and all distance-based methods, except for the NJ. The NJ trees exhibited the most incorrect Sackin index (Fig. 2.7(e)(f)), possibly because NJ trees are not rooted, a necessary condition to compute this index. The rooting procedure chosen can be quite arbitrary, affecting the balance of the trees and consequently the Sackin index. The distributions above all scatterplots show qualitatively the differences in topology (Fig. 2.7(a),(b)), branch length distribution (Fig. 2.7(c),(d)) and degree of imbalance (Fig. 2.7(e),(f)) of phylogenies generated from each algorithm or method in the 50 simulations performed in each time ($t = 500$ or $t = 1000$).

2.5 Discussion

Understanding all the mechanisms that promote speciation is still an open problem in evolutionary biology (Gavrilets, 2014; Kirkpatrick and Ravigné, 2002). Even more challenging is to identify which of these mechanisms were important in a particular case. A large number of mathematical and computational models were developed in the past years to understand different speciation processes, such as neutral (Hoelzer et al., 2008; Desjardins-Proulx and Gravel, 2011; Melián et al., 2012; Baptestini et al., 2013a), sexual (Doorn et al., 2009; Uyeda et al., 2009; M'Gonigle et al., 2012) and ecological selection (Rettelbach et al., 2013; Nosil, 2012). Models have also considered the role of geography in speciation, such as allopatric (Fierst and Hansen, 2010; Gourbiere and Mallet, 2010; Fraïsse et al., 2014; Yamaguchi and Iwasa, 2013), parapatric (Gavrilets et al., 2000b; Bank et al., 2012) and sympatric (M'Gonigle et al., 2012; Rettelbach et al.,

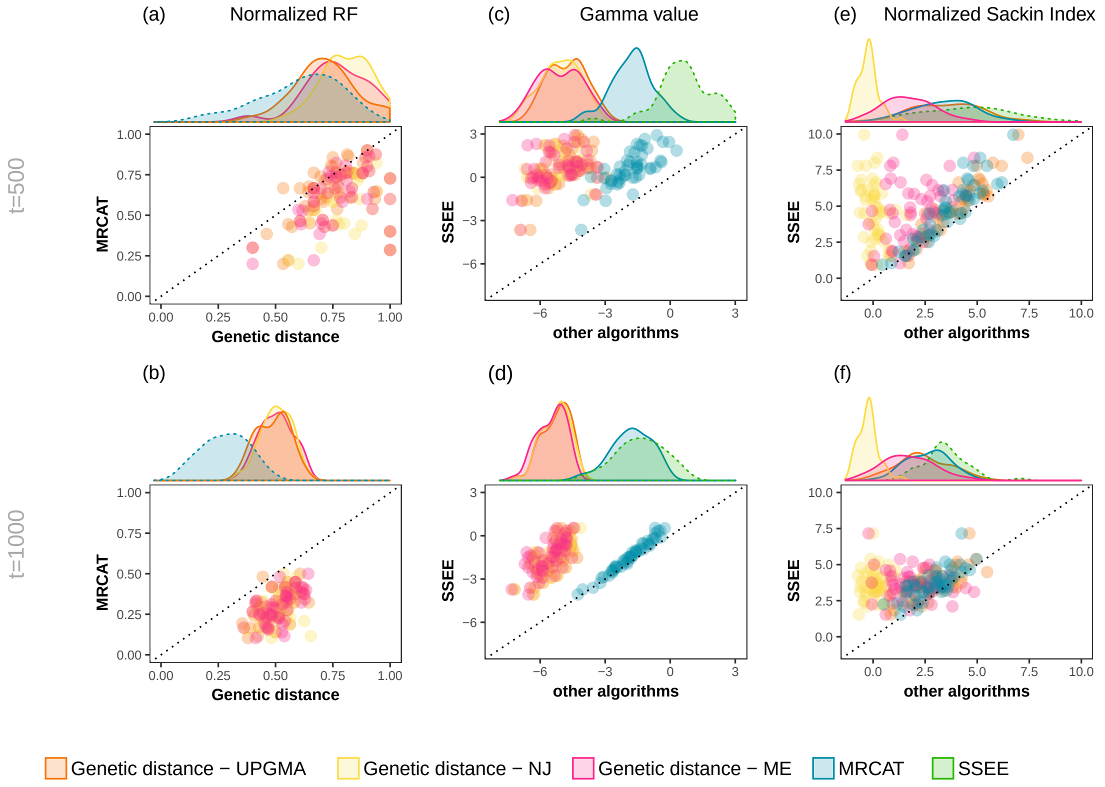


FIGURE 2.7: Comparisons among phylogenies generated by the algorithms proposed here (MRCAT and SSEE) and phylogenies estimated from genetic distance by UPGMA, NJ and ME methods. Lines exhibit the comparisons of RF, gamma and Sackin's metrics of 50 simulations at times 500 (first line) and 1000 (second line) generations. Colors represent the different methods utilized to generate the trees. (a) and (b): difference between RF-values of phylogenies obtained by MRCAT (y-axis) and by genetic distance-based methods (x-axis). Small values of RF indicate that phylogenies are closer to the true phylogeny (SSEE). (c) and (d): difference between branch length distributions (γ) of phylogenies generated by SSEE (y-axis, green distribution) and MRCAT algorithm (blue) or genetic distance-based methods (orange, yellow and pink) (x-axis). (e) and (f): the same as (c) and (d), but considering now the degree of imbalance (Sackin index). Distributions above all scatterplots illustrate qualitatively the differences in topology (a,b), branch length distribution (c,d) and degree of imbalance (e,f) of phylogenies generated from each algorithm or method in the 50 simulations.

2013; Bürger et al., 2006; Pennings et al., 2007) scenarios. The results of models, however, can seldom be compared with real data (Bolnick and Fitzpatrick, 2007; Gavrilets and Losos, 2009). In these cases comparisons are often made in a macroecological scale, including qualitative species abundance and spatial distributions, species-area relationships and genetic or phenotypic distances (De Aguiar et al., 2009; Martins et al., 2013; May et al., 2015; Campos et al., 2012; Martín and Goldenfeld, 2006). Nevertheless, little attention has been given to the evolutionary history of individuals and species, neglecting the macroevolutionary scale underlying the speciation process (Manceau et al., 2015; Rosindell et al., 2015).

In this paper we have described two procedures to register the history of individuals (MRCAT) and species (SSEE) in individual-based models. With the ancestor-descendant relationships or speciation events saved in MRCAT and SSEE matrices we have constructed trees using a clustering algorithm. These trees have properties demonstrated in section A.1 of Supporting Information A. In the MRCAT algorithm, genealogies of individuals and phylogenies of species were obtained, whereas in the SSEE algorithm only phylogenies of species can be accessed. In the SSEE algorithm speciation events are precisely recorded and the resulting phylogenetic tree is the *true* tree of the community, whereas in the MRCAT algorithm the relations among species are recovered from genealogical relationships between individuals of each species. The MRCAT algorithm allows the construction of maternal, paternal and general lineages, the last being analogous to cases with hermaphroditic individuals. We have applied these algorithms to a spatially explicit IBM where individuals are separated into males and females and sexual reproduction is restricted by genetic difference below a threshold and by spatial proximity. We showed that maternal, paternal and general genealogies generated from the MRCAT algorithm are different even if the same individuals are chosen to draw the trees (Supporting Information A, section A.2). Maternal and paternal genealogies (Fig. S3(a),(b)) are different because they were obtained from different MRCAT matrices. In the first case, the MRCAT matrix contains the time to the most recent common $female$ ancestor between each pair of individuals, while in the second case the MRCAT matrix has the time to the most recent common *male* ancestor between the same individuals, which lead to different ancestor times and genealogical relationships. In addition, for the general genealogy - taking the most recent common ancestor among females and males (*i.e.*, disregarding sex) - the resulting MRCAT matrix does not uniquely specify the genealogy (Fig. S3(c)). Regarding the phylogenetic trees, we showed that they may be different if obtained by MRCAT or SSEE algorithm

(Fig. 2.6(a),(b), Fig. 2.7). As discussed in subsection 2.2.3, this mismatch happens because the time to the most recent common ancestor between individuals of different species is only an approximation to the speciation time, since speciation can happen several generations later (Fig. 2.3).

Structural properties of phylogenies, such as the Sackin index and the gamma distribution, obtained from SSEE and MRCAT trees were compared to values calculated in phylogenies estimated from the genetic distance between individuals of extant species by distance-based methods (UPGMA, NJ and ME). The aim of this comparison was to show that the validity of these methods commonly used in empirical studies, where the complete past history is inaccessible, can be assessed with the help of models. Differences in topology and branch length distribution measured by the RF metric and γ-statistic, respectively, revealed that MRCAT trees were closer to the true phylogenies (SSEE) than genetic distance-based trees. The difference between the results of these two methods possibly lies in back mutations that can happen in the genome of individuals, erasing the information needed to uncover the real history among species (Hein et al., 2004). This phenomenon is more likely to happen at long times and for small genome size. Indeed, we observed that in 500 generations (Fig. 2.7(a)(c)) the phylogenies estimated from genetic distance were closer to the ones generated from MRCAT algorithm than in 1000 generations (Fig. 2.7(b)(d)), because in the first case the number of back mutations were probably smaller. Another factor that might explain the difference between genetic distance-based and true phylogenies is the sampling of only one individual to estimate the trees in the first case (Yang and Rannala, 2012). However, phylogenies generated with MRCAT algorithm also used only one individual per species - the same individuals used to compute genetic distance indeed - which suggests that this is not a very important factor (Fig. 2.7(a),(b),(c),(d)). The degree of imbalance showed a different picture, with less differences between MRCAT trees and genetic distance trees. Still, MRCAT trees were closer to the true phylogenies than the others. Trees estimated from genetic information in IBMs should be closer to the true phylogenies for larger genome sizes, where the probability of back mutations is smaller. Individual-based models with large or infinite genome sizes already available (De Aguiar, 2017; Higgs and Derrida, 1992) would provide good tests for measuring the accuracy of trees obtained by distance-based methods.

The better performance of MRCAT algorithm in recover the topology and balance of phylogenetic trees is not surprising, since matrices generated from this algorithm hold the exact times to the most recent common ancestors. However, this type of exact information cannot be recovered from empirical data of contemporary samples. On the other hand, distance-based methods are commonly used for inference of phylogenetic trees from empirical data (Yang and Rannala, 2012). The advantage of these methods, especially the NJ method, is their computational efficiency. Indeed, cluster algorithms are faster than optimality criteria used in character-based methods, like maximum parsimony and maximum likelihood (Yang and Rannala, 2012; Gascuel and Steel, 2006). Distance methods are particularly useful for analysis of data sets containing sequences with low levels of divergence (Yang and Rannala, 2012). However, methods based on genetic distances can perform poorly when the data set contains sequences with high levels of divergence due to greater sampling error in larger genetic distances. As most distance-based methods do not account for the high variances of large distance estimates, the inference of phylogenetic relationships could be impaired when these methods are employed (Yang and Rannala, 2012). In our model, trees generated from genetic distance methods were more different from the true trees (SSEE) than MRCAT phylogenies possibly because of high divergence among simulated genomes. This also could explain the high similarity in tree summary statistics among distance methods (Fig. 2.7). Moreover, the worst performance of NJ method in recover tree balance might be due to the lack of an explicit optimization criterion in the selection of taxon pairs in the original method proposed by Saitou and Nei (Saitou and Nei, 1987) and utilized here (Paradis et al., 2004; Gascuel and Steel, 2006). In addition, the choice of a substitution model to compute the pairwise distance between sequences might be important to determine the efficacy of distance methods (Yang and Rannala, 2012). Here we used the Hamming distance to calculate differences between pairs of sequences, but other methods could yield different results (Jukes and Cantor, 1969; Kimura, 1980; Hasegawa et al., 1985; Yang, 1994).

Modifications of the model to include *loci* not linked to the computation of genetic threshold would be important to understand how phylogenetic trees computed from these *loci* would differ from the ones computed here. Changing parameters values such as genome size and mutation rate could also affect tree estimations from distance-based methods and are a possible direction to future research. Nevertheless, the incorporation of algorithms that record the evolutionary history of individuals and species in an IBM context is an important step to help understanding the patterns left by specific speciation mechanisms at the macroevolutionary level.

2.6 Conclusions

The recent interest in the role of evolutionary history to explain the spatial patterns of abundance and species diversity calls for the incorporation of phylogenetic trees in the speciation modeling approach. Phylogenetic trees are essential tools to understand macroevolutionary patterns of diversity. They reveal how species are related to each other and the times between speciation events. Moreover, topological structure and branch length distribution also contain clues about processes originating a particular group of species. Previous works have already considered this problem for simpler models where each mutation corresponds directly to a new species (Manceau et al., 2015). Our study provides the first general attempt to extend individual-based models by incorporating the branching process using the ancestor-descendant relationships between individuals and species. We believe this methodology will help predict and classify the macroevolutionary branching process, as well as the corresponding macroecological patterns (*e.g.*, species abundance distributions), resulting from different speciation models. The comparison of these results with empirical studies may clarify the role of different processes in generating the patterns observed in nature (Turelli et al., 2001; Field et al., 2009). Finally, the role of extinction in determining macroevolutionary patterns is an open field (Quental and Marshall, 2011) which could be explored by using the full phylogenetic trees generated from the SSEE algorithm introduced here.

Acknowledgments

This research was supported by Sao Paulo Research Foundation (FAPESP), National Council for Scientific and Technological Development (CNPq), and Coordination for the Improvement of Higher Education Personnel (CAPES). We thank P. L. Costa for his constructive comments about the manuscript, and A. B. Martins and L. D. Fernandes, who provided expertise that greatly assisted the research, all contributing with their insights and comments about the research results.

Conflicts of interest: none.

Chapter 3

Signatures of microevolutionary processes in phylogenetic patterns

Article published by: Carolina L.N. Costa, Paula Lemos-Costa, , Flavia M.D. Marquitti, Lucas D. Fernandes, Marlon F. Ramos, David M. Schneider, Ayana B. Martins, Marcus A.M. de Aguiar. 2018 "Syst Biol" doi.org/10.1093/sysbio/syy049.

Abstract

Phylogenetic trees are representations of evolutionary relationships among species and contain signatures of the processes responsible for the speciation events they display. Inferring processes from tree properties, however, is challenging. To address this problem we analysed a spatially-explicit model of speciation where genome size and mating range can be controlled. We simulated parapatric and sympatric (narrow and wide mating range, respectively) radiations and constructed their phylogenetic trees, computing structural properties such as tree balance and speed of diversification. We showed that parapatric and sympatric speciation are well separated by these structural tree properties. Balanced trees with constant rates of diversification only originate in sympatry and genome size affected both the balance and the speed of diversification of the simulated trees. Comparison with empirical data showed that most of the evolutionary radiations considered to have developed in parapatry or sympatry are in good agreement with model predictions. Even though additional forces other than spatial restriction of gene flow, genome size, and genetic incompatibilities, do play a role in the evolution of species formation, the microevolutionary processes modeled here capture signatures of the diversification pattern of evolutionary radiations, regarding the symmetry and speed of diversification of lineages. [speciation; diversification rates; sympatry; parapatry; evolutionary radiation; individual-based model]

3.1 Introduction

Understanding the origin and maintenance of species diversity is a major goal in ecology and evolutionary biology. Processes influencing speciation have been studied in numerous ways, testing how different ecological and non-ecological processes contribute to the generation of species (Nee et al., 1992; Pybus and Harvey, 2000; Rundell and Price, 2009; Butlin et al., 2012; Gavrilets, 2014; Morlon, 2014; Seehausen et al., 2014). Model-based approaches have played an essential role in this avenue (Gavrilets, 2014) and, over the past years, have moved from providing proofs of concept of novel processes (Dieckmann and Doebeli, 1999; De Aguiar et al., 2009; Servedio et al., 2014; Melián et al., 2015) to tools which generate predictions that can be directly contrasted with empirical patterns (McPeek, 2008; Higgs and Derrida, 1991; Pigot et al., 2010; Martins et al., 2013; Gascuel et al., 2015a; Manceau et al., 2015).

An important way to unveil which processes promote speciation is to identify their fingerprints in the macroevolutionary patterns of phylogenetic trees (Nee et al., 1992; Morlon, 2014). Model-based approaches derived from the Neutral Theory of Biodiversity (NTB) have been successful at predicting empirical macroecological patterns, like species abundance distribution and species-area relationship (Etienne and Alonso, 2005; Jabot and Chave, 2009; O'Dwyer and Green, 2010; Hurlbert and Stegen, 2014), but have not provided accurate predictions about macroevolutionary patterns, such as phylogenetic tree shape. Most of these are birth-death models that consider speciation (birth) and extinction (death) as random events, not taking into account microevolutionary processes, as the population dynamics within species, which could explain their deficiency to predict macroevolutionary patterns (Mooers et al., 2007; Davies et al., 2011). Speciation models that incorporate microevolutionary processes along with phylogenetic trees arising from the theoretical simulated populations are key to clarify how population-level processes affect diversification rates and the emerging macroevolutionary patterns, filling the micro-macro gap of the biodiversity theory. These models are rarely utilized because of their complexity, but they describe individual-level processes that can directly scale-up to influence diversification (Rosindell et al., 2015).

The interest in searching for signatures of microevolutionary processes in phylogenetic trees using mechanistic simulation models has increased over the past years (Cabral et al., 2017). Previous studies have explored the phylogenetic signatures of speciation processes that evolves by accumulation of incompatibilities in the context of demographical non-equilibrium in sympatry (Manceau et al., 2015), or of ecological spatial heterogeneity in metacommunities (Gascuel et al., 2015a). These studies provide interesting insights about microevolutionary process driving speciation and their emerging phylogenetic patterns, nevertheless, the role of geographical isolation and other population level ingredients, such as population density, dispersal and number of genes, on the build-up of reproductive isolation have not been investigated in the context of the NTB. Different macroevolutionary patterns, such as branching slowdown, have been attributed to ecological modes of speciation (Gavrilets and Vose, 2005; Gascuel et al., 2015a). However, these patterns can also be predicted by non-ecological speciation processes (Pigot et al., 2010; Wang et al., 2013). Studies focusing primarily on non-ecological processes seldom incorporate microevolutionary processes (but see Manceau et al., 2015), hence the importance of filling such a gap in the literature.

Here we link microevolutionary processes, such as gene flow and the evolution of reproductive isolation, to macroevolutionary patterns displayed by phylogenetic trees. More specifically, we identify the signatures that different geographical modes of speciation and genome size leave on phylogenetic patterns. We use a spatially explicit individual-based model (IBM) of speciation in which reproductively isolated species evolve in response to genetic and spatial restrictions imposed on reproduction (De Aguiar, 2017). The model has the potential to explain macroecological patterns compatible with the NTB (De Aguiar et al., 2009), patterns of genetic differentiation between species (Martins et al., 2013), and diversification on a macroevolutionary scale (Costa et al., 2018a). Our approach differs from previous studies in two important aspects: first, we explore the connection between population genetics, speciation theory and macroevolution by studying the phylogenetic patterns of a mechanistic model for the evolution of reproductive isolation that explicitly simulates sexual reproduction, dispersal, mutation and genetic drift. We model genomes with B biallelic loci and individuals are considered incompatible if the number of loci carrying different alleles becomes larger than a fixed threshold G. Genome size, B, has a key role in defining the possible geographical modes of speciation, even when the proportion G/B of allowed genetic incompatibilities is fixed (De Aguiar, 2017). And second, the model is ecologically neutral and reproductively isolated species evolve in response to genetic and spatial restrictions imposed on reproduction (De Aguiar et al., 2009).

The incorporation of spatial restriction makes this model particularly convenient to study geographical modes of speciation (for a detailed description of the model, see Methods Section). By changing only two model parameters, the degree of restriction to gene flow and the total number of loci, speciation can change from strongly (parapatric) to weakly (sympatric) space dependent (De Aguiar et al., 2009; De Aguiar, 2017). For simplicity, we hereafter refer to these two parameters as the size of mating range and the genome size, respectively. In order to effectively compare results generated by genome sizes that differ in many orders of magnitude, we fixed the proportion G/B between threshold value and genome size in all simulations. Therefore, increasing the total number of loci (B) also increases the absolute number of divergent loci recognized as the compatibility threshold, but not the relative proportion of accepted differences. In each case, we computed the corresponding phylogenetic tree from ancestor-descendant relationship information about extant individuals at the end of the simulation (Costa et al., 2018a) and calculated three indexes assessing tree topology: tree balance (Sackin index), branch sizes distribution (γ-statistic), and acceleration of diversification (α-value). We show that the genome size affects both tree balance and rates of diversification through time. The size of the mating range affects only the acceleration of diversification rates. We compare our results with empirical phylogenies focusing on evolutionary radiations, which are a burst in species diversification in which a lineage of species occupies a large (adaptive) or a minimal (non-adaptive) diversity of ecological roles (Rundell and Price, 2009). Our findings agree with previous results in that sympatric speciation produces more balanced trees than parapatric speciation (Barraclough and Vogler, 2000; Losos and Glor, 2003; Pigot et al., 2010), but we expand these results by linking geographic modes with genome size and rates of speciation. The good agreement between the degree of restriction in gene flow in the model and the corresponding patterns seen in the empirical trees of adaptive radiations suggests that the signatures on the tree balance and on the acceleration of the diversification rate are either direct consequences of the geographic mode of speciation or that the similar patterns are generated by adaptive processes not modeled here. Finally, we show that our model of non-adaptive radiations can display slowdown in diversification rate and the overshooting effect (an early increase in number of species followed by a decline that plateaus (Higgs and Derrida, 1991), patterns usually considered signatures of the adaptive counterparts of the diversification process (Gavrilets and Vose, 2005; Higgs and Derrida, 1991).

3.2 Materials and Methods

3.2.1 Model

The spatial model we discuss here is a modified version of the model proposed by (De Aguiar et al., 2009). We consider a population of $M = 1,000$ haploid individuals randomly distributed in a square lattice of linear size $L = 100$ with periodic boundary conditions (individuals can, by chance, occupy the same lattice site). The choice of haploids is for computational simplicity, since diploid models lead to qualitatively similar results (Schneider et al., 2016). The value of M is kept constant at each time step. Genomes are represented by binary strings of size B, $\{\sigma_1^i, \sigma_2^i, \ldots, \sigma_B^i\}$ for individual i, where each *locus* σ_k^i, can assume the allele values 0 or 1. Here, we refer to B as genome size for simplicity. The genetic distance d between two individuals i and j is the Hamming distance between the corresponding sequences and measures the number of genes bearing different alleles:

$$d^{i,j} = \sum_{k=1}^{B} |\sigma_k^i - \sigma_k^j|. \tag{3.1}$$

Mating is restricted by genetic similarity and by spatial proximity, so that an individual i can only choose as mating partner those inside a circular neighborhood of radius S centered in its spatial location (the mating range), and whose genetic distance satisfies $d^{i,j} \leq G$ (Gavrilets et al., 2000b). The parameter G represents the *genetic threshold* of reproductive isolation, by which individuals find potential compatible mates. Therefore, partner choice is determined solely by the compatibility condition $d^{i,j} \leq G$ and not by minimizing the genetic distance $d^{i,j}$. The parameter S represents the size of the *mating range*. Given the set of compatible individuals within the mating range, the *focal* individual mates randomly with one of the potential compatible partners. In the simulations, we used $G = 0.05B$, which means that genetic threshold will always be proportional to genome size. The scaling of model parameters with population size was discussed in (Baptestini et al., 2013a). The model dynamics depend critically on genome size (De Aguiar, 2017), so that for very large genomes speciation may occur in sympatry ($S \to L$) and also in parapatry ($S \ll L$). However, for smaller genomes, it only happens if mating is restricted to sufficiently close neighbors – parapatry ($S \ll L$). In the limit of infinitely large genomes and $S = L$ the model recovers the Derrida-Higgs dynamics (Higgs and Derrida, 1991; De Aguiar, 2017).

Each one of the M individuals has a chance of reproducing, but there is a probability Q that it will not do so, accounting for the fact that not all individuals in the present generation will be first parents of the next. In case the focal individual does not reproduce, another one from its mating range is randomly chosen to reproduce in its place, which maintains the population size approximately constant. In our simulations, we set $Q = 0.37 \approx e^{-1}$, which corresponds approximately to the probability that an individual is not selected in M trials with replacement, $(1 - 1/M)^M \approx e^{-1}$, in accordance with the Derrida-Higgs model (Higgs and Derrida, 1991; De Aguiar, 2017). The first parent (the $f$ocal individual or a neighbor) chooses a compatible second parent within its mating range of radius S. The number of individuals in the mating range may be close to zero due to fluctuations in the spatial distribution. To avoid this situation, we follow the procedure introduced in (De Aguiar et al., 2009): if the number of compatible mates in the range is smaller than P ($P = 3$ in our simulations), the individual expands the search radius to $S + 1$. If the number of compatible mates is still smaller than P, the process is repeated up to $S + 2$, and if there is still less than P potential mates, another neighbor is randomly selected to reproduce in its place (De Aguiar et al., 2009).

The reproduction is sexual: the offspring inherits, *locus* by *locus*, the allele of either parent with equal probability. The reproductive process is repeated until all M individuals had its chance to reproduce. After reproduction, the resulted offspring is also subjected to a mutation rate μ per *locus*($\mu = 0.00025$ in our simulations). This is a modification with respect to the original model in (De Aguiar et al., 2009), where offspring genomes were generated with a single crossover of the parental genomes. The dynamics is constructed in such a way that offspring is placed close to the location of the original parents and the homogeneous distribution of the population is preserved at all times. In either case, the offspring generated will be positioned exactly at the location of the focal parental individual, or it will disperse with probability D (here we set $D = 0.01$) to one of the 20 nearest neighbors (radius approximately equal to $\sqrt{5} \approx 2.24$). Therefore, close to the location of every individual of the previous generation, there will be an individual in the present generation, keeping the spatial distribution homogeneous and avoiding the formation of spatial inhomogeneities. With this choice, we also avoid a stronger influence of dispersal on gene flow, regarding the latter as a result essentially of mating.

We identify a species as a group of individuals reproductively isolated from all others by the genetic threshold on mating defined by parameter G. Not all members of the group have to be able to mate with each other, but could maintain an indirect gene flow through an intermediary individual. Therefore, if individuals $\mathcal{A}$ and $\mathcal{B}$ are compatible and so are B and $\mathcal{C}$, but $\mathcal{A}$ and $\mathcal{C}$ are reproductively isolated, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ will belong to the same species, owing to the ongoing gene flow among all of them. No condition on spatial proximity is imposed on the members of a species.

Individuals are genetically identical at the beginning of the simulation but, as time proceeds, mutations accumulate and reproductively isolated species branch off the population. After a transient time of increase in species number, speciation events are balanced by extinctions and the number of species remains approximately constant. During the transient, extinctions and incomplete speciation events are also observed (Rosindell et al., 2010) (see Section B.1 in Supporting Information B). In all cases studied here the populations evolved only up to equilibration time, when speciation and extinction balance, so as to describe the process of radiation from a population inhabiting a single area.

In order to evaluate how the size of mating range affects the diversification patterns observed in phylogenetic trees, we vary the parameters S (mating range) and B (genome size) of the model. Fixing a large enough value of genome size, $B = 150,000$, and varying the size of the mating range (S), we control the geographical mode of speciation, ranging from parapatry with low levels of gene flow (small S) to sympatry with high levels of gene flow (large S). The sympatric case is characterized by a value of S that is large enough to allow most individuals in the initial state to potentially interbreed – see "Large genome size" in Results. On the other hand, fixing a small value of S, such as $S = 5$, and increasing B from 150 to $150,000$, we evaluate the role of genome size in parapatric speciation with low levels of gene flow – see "Small mating range" in Results. We have fixed $G/B = 0.05$ in all simulations. Analytical and simulation results for the time to divergence for different values of G/B can be found in Section B.2 of the SI B. We do not test for small B and large S, because speciation does not occur under this combination of parameters (De Aguiar, 2017).

3.2.2 Rooted bifurcating trees and quantifying indexes

Differently from distance-based methods usually employed for species relationship inference, here we constructed the phylogenetic trees based on the ancestor-descendant relationships among individuals. In each generation, we recorded parenthood for the entire population and we registered the time to the Most Recent Common Ancestor (MRCA) between each pair of individuals, tracing their genealogical relationships registered at each time step. In this way, at the end of the simulation we had the time to the MRCA between all individuals of the extant population, which evolved from a single ancestor. We define the branch lengths and the structure of the phylogenetic trees using only one individual of each species, since both are insensitive to individuals' choice in our simulations (Costa et al., 2018a). The most recent common ancestor of all the extant species of a tree was called the tree-MRCA, located in the root node of the tree. Fig. 3.1 describes the most important elements of a rooted bifurcating tree. In order to quantify and compare different trees we use the following indexes:

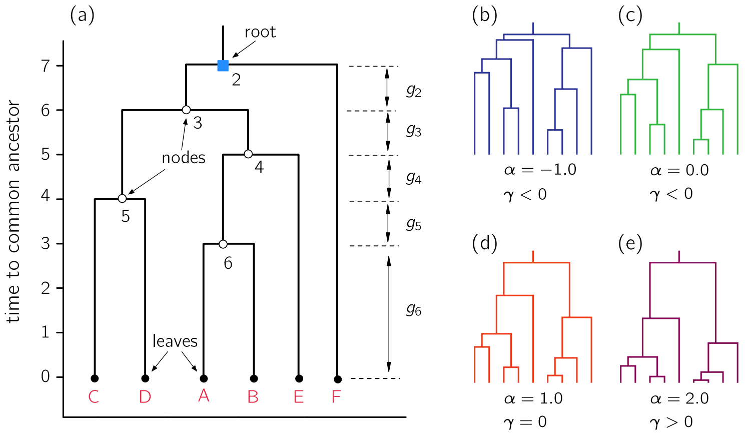


FIGURE 3.1: (a) Example of a phylogenetic tree with 6 species, represented by filled circles as the leaves of the tree. Open circles denote the nodes, which are numbered from 2 (topmost node, also called the root) to N. Letters below the leaves label the species. Times between speciation events, g_k, are used to calculate the γ-statistic. (b-e) Phylogenetic trees showing different distributions of branch length and the relation between statistics γ and α. The tree represented at (c) has a constant bifurcation (speciation) rate per unit of time – Yule model, resulting in constant inter-node distances ($alpha = 0.0$). The tree represented at (d) has a constant bifurcation (speciation) rate per branch – pure-birth model ($gamma = 0.0$).

- **Sackin index ($I(N)$)** - The Sackin index evaluates tree imbalance (Sackin, 1972; Blum and François, 2005; Frost and Volz, 2013) by measuring the distances between the leaves and the root. It is defined as $I(N) = \sum_j d_j$, where d_j is the number of nodes to be traversed between each leaf j and the root, including the root (Dearlove and Frost, 2015). The minimum value occurs for the most symmetric bifurcating tree and it can be shown that $\min(I) \approx N \ln(N)/\ln(2)$. This expression is exact when the tree size N is a power of 2, and it is an approximation in the other cases. The maximum value of the Sackin index is found in trees that are most unbalanced and is given by $max(I) = (N-1)(N+2)/2$. For the tree (a) in Fig. 3.1 $I(6) = 18$, whereas the maximum value for 6 leaves is 20 and the minimum 16. For the Yule model, the average Sackin index for a tree with N leaves is given by $E(I(N)) = 2N\sum_{k=2}^{N} \frac{1}{k}$. Thus, it is possible to define a relative, normalized, Sackin index as $I_n(N) = \frac{I(N)-E(I(N))}{N}$, which we use in this study. Since $E(I(6)) = 17.4$ we find $I_n = 0.1$ for the tree (a) in Fig. 3.1. We present detailed explanation about the Asymmetric Yule model in Section B.3 in SI B. We used the *sackin* function of the apTreeshape package (Bortolussi et al., 2005) to compute the Sackin's index on trees and normalize it.

- **Alpha value (α)** - The distribution of branch lengths along the tree is usually computed with the Gamma statistic, defined as (Pybus and Harvey, 2000):

$$\gamma = \frac{1}{D}\left[\frac{1}{N-2}\sum_{k=2}^{N-1} T(k) - T(N)/2\right] \tag{3.2}$$

where

$$T(k) = \sum_{j=2}^{i} j g_j; \tag{3.3}$$

$$D = T(N)/\sqrt{12(N-2)} \tag{3.4}$$

and g_k is the time interval between speciation events as labeled by the nodes (see Fig. 3.1). The γ-statistic is constructed in such a way that $\gamma = 0$, if $g_k = 1/(bk)$. This corresponds to a continuous time process in which all species bifurcate with fixed rate b, leading to $\langle g_k \rangle = 1/(bk)$, $\langle \gamma \rangle = 0$ and $\langle \gamma^2 \rangle = 1$ (see SI B, Section B.4).

One of the criticisms about γ-statistic is that it depends on the number of leaves in the tree (McPeek, 2008; Phillimore and Price, 2008). Writing the time between speciation events as $g_k = k^{-\alpha}/b$ we obtain $\gamma = \gamma(N, \alpha)$. This relation can be numerically inverted such that for any tree with given γ and N, a unique α-value can be computed, measuring the changes of speciation rate, *i.e.* the acceleration of speciation, along the tree. Constant speciation rates per branch corresponds to $\alpha = 1$, and constant rates of speciation events at any time corresponds to $\alpha = 0$ (trees (d) and (c) in Fig. 3.1 and see Section B.5 in SI B). For the tree in Fig. 3.1(a) $\alpha = -1.58$, which fits into the type described in Fig. 3.1(b).

3.2.3 Empirical phylogenies

We compared our simulations with 16 known cases of evolutionary (adaptive and non-adaptive) radiations in animals and plants (Rundell and Price, 2009; Simoes et al., 2016), listed in Table 3.1 (check the phylogenetic trees in the Section B.6 in SI B). In order to compare the tree statistics of empirical data with the different degrees of gene flow modeled here (though the parameter S), we define qualitatively the level of gene flow during diversification in empirical data as low, intermediate, or high, based on information of system characteristics as dispersal rate and geographic distribution of species.

Within the set of adaptive radiations (points 1 – 9), cichlid radiations in crater lakes were considered to have been subjected to the highest levels of gene flow because the lakes are small and these radiations have been traditionally considered examples of sympatric speciation (Martin et al., 2015). Lake Tanganyika and lake Malawi cichlid radiations were considered to have been subjected to the lowest levels of gene flow because the species in these groups inhabit a great lake and are highly phylopatric, exhibiting low levels of dispersal and migration (Kocher, 2004). The four terrestrial adaptive radiations were classified as intermediate cases. Studies suggest that in all these cases, islands were initially invaded by a single species and the spreading to different islands gave rise to independent small radiations (Rundell and Price, 2009). The analyzed trees encompass all species distributed across these island systems and evolutionary patterns result from the combination between processes occurring between islands with processes occurring within islands. Cichlid radiations in large lakes probably have a comparable history of mixed levels of gene flow (Meier et al., 2017a; Meier et al., 2017b). Nonetheless, given the dispersal potential of these species in relation to the size of lakes Malawi and Tanganyika (orders of magnitude larger than crater

TABLE 3.1: Empirical radiations. Point in this table is used in the Fig. 3.5 (RESULTS section). N_S is the number of species in each empirical tree.

Point	N_s	Species	Gene flow during diversification	Radiation	Reference
1	10	Barombi Mbo cichlids	high	Adaptive	(Martin et al., 2015)
2	9	Bermin cichlids	high	Adaptive	(Martin et al., 2015)
3	35	Hawaiian silversword alliance	intermediate	Adaptive	(Blonder et al., 2016)
4	71	Caribbean *Anole* lizards	intermediate	Adaptive	(Alföldi et al., 2011)
5	25	*Tetragnatha* spiders	intermediate	Adaptive	(Gillespie, 2004)*
6	14	Darwin's finches	intermediate	Adaptive	(Clarke et al., 2017)
7	44	Tanganyika cichlids - 1	low	Adaptive	(Meyer et al., 2015)
8	40	Tanganyika cichlids - 2	low	Adaptive	(McGee et al., 2016)
9	16	Malawi cichlids	low	Adaptive	(McGee et al., 2016)
10	38	Australian *Gehyra* geckos	high	Non-adaptive	(Heinicke et al., 2011)
11	25	Delphinidae	high	Non-adaptive	(Banguera-Hinestroza et al., 2014)
12	74	Philippine sun skinks	high	Non-adaptive	(Barley et al., 2013)
13	68	Rodents *Rattus*	high	Non-adaptive	(Rowe et al., 2011)
14	68	Lichen *Sticta*	intermediate	Non-adaptive	(Simon et al., 2018)
15	46	Caviomorph rodents *Ctenomys*	low	Non-adaptive	(Álvarez et al., 2017)
16	53	New World titi monkeys	low	Non-adaptive	(Byrne et al., 2016)

**personal communication*

lakes), the general assumption that gene flow between lineages has been lower during diversification is warranted.

Regarding the non-adaptive radiations (points 10 – 16), the australian geckos, delphinids, philippine sun skinks and rodents of genus *Rattus* were considered to have high gene flow during diversification due to the high dispersal rates and/or broadly sympatric distributions observed in these groups (Heinicke et al., 2011; Banguera-Hinestroza et al., 2014; Barley et al., 2013; Rowe et al., 2011). The lichen of genus *Sticta*, as the terrestrial adaptive radiations, was classified with intermediate gene flow, because also in this system the radiation was initiated by a single long-distance dispersal event (in the MIOI region in this case) followed by multiple dispersal events between islands (Simon et al., 2018). Finally, the caviomorph rodents and the titi monkeys were considered to have low gene flow during diversification, since the first are strictly territorial and both present restricted dispersal abilities due to their subterranean habits and geographic isolation by rivers, respectively (Álvarez et al., 2017; Byrne et al., 2016). We considered these as non-adaptive radiations because they display conserved ecomorphological traits, even though the initial phase of radiation presents a burst in speciation rate, traditionally associated with adaptive radiations (but see Moen and Morlon, 2014).

3.3 Results

Gene flow has a strong negative effect on final species richness. Indeed, species richness at equilibration is higher when gene flow is restricted and the number of loci is large (small S and large B, blue and red curves in Fig. 3.2), thus facilitating the formation of new species both from the spatial and genetic aspects. Genome size drastically affects the time for equilibration of species number (Fig. 3.2). For small genomes ($B = 150$), equilibration takes about $1,000$ generations, and speciation events occur at an approximately constant rate (see also Fig. B.12 in SI B). For large genomes ($B = 150,000$ and $B = 1,500$), equilibration happens earlier, about 250 generations, and diversification patterns resemble those of adaptive radiations, where many species arise early and species accumulation declines over time (see Fig. B.13 in SI B). For large genomes, and intermediate to large mating range ($B = 150,000$ and $S = 20$ or $S = 40$), there is an overshooting effect (Gavrilets and Vose, 2005) and number of species at equilibrium is lower than the maximum reached during the radiation (Fig. 3.2 and Fig. S10).

Genome and compatibility threshold size have important effects on the radiation patterns. We recall that results are shown for $G/B = 0.05$, so that small(large) genome size implies small(large) compatibility threshold. Radiation is slower for small genomes (Fig. 3.2) with small species branching off the original population and rarely suffering speciation again (see Fig. B.12 in SI B) resulting in unbalanced trees. For large genome sizes, on the other hand, the radiation process is fast, with more abundant species branching off and splitting again before equilibration (see Fig. B.13 in SI B), giving rise to balanced trees. Stochastic fluctuations in mating choice cause the number of individuals per species to change over time, eventually leading species to extinction by ecological drift. It is worth noticing that extinctions are negligible during the radiation for large genome/threshold ($B = 150,000$, $G = 7500$) and occur at a constant rate for small genome/threshold ($B = 150$, $G = 7$) (see Figs B.1 and B.2 in SI B). The extinction pattern however does not appear to be a consequence of the average abundance distribution of individuals per species, as species have a larger population size for small genome ($B = 150$) than for large genome ($B = 150,000$) at equilibration time under the same mating range (see Section B.8 in SI B). In Section B.9 of SI B, we show that model parameters S and B can be inferred from the phylogenies with some accuracy. The power of inference of the model has limitations regarding the parameter space covered, the variance of the distributions (size of ellipses) and the number of

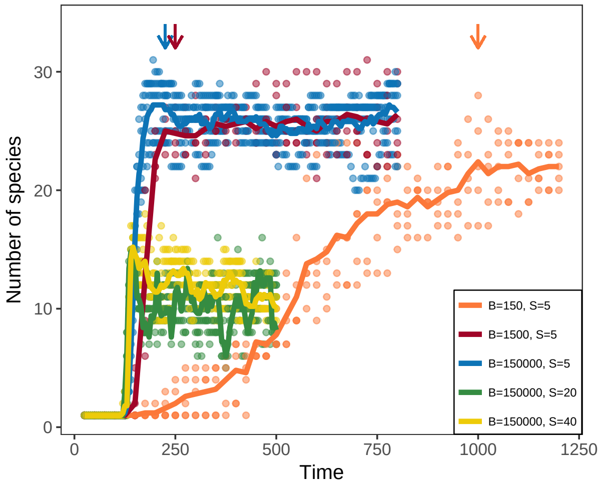


FIGURE 3.2: Number of species as a function of time for different combinations of mating range, S and genome size, B. Points show results of 5 realizations for each value of B at each time step, darker points depict overlapping values. Solid lines show the average values. Arrows indicate the simulation time when phylogenetic trees were computed for $B = 150,000$ (225 generations, blue, all values of S), $B = 1,500$ (250 generations, red) and $B = 150$ (1,000 generations, orange). For $B = 150,000$ and $S = 20$ and $S = 40$ the number of species reaches a maximum before equilibration (see Fig. B.10 in SI B).

simulations.

3.3.1 Large genome

Figure 3.3 shows the effect of the geography of speciation for large genome size ($B = 150,000$, $G = 7,500$) with typical species distributions in space (first row) and the corresponding phylogenetic trees (second row). Columns correspond to increasing values of S, *i.e.*, low ($S = 5$), intermediate ($S = 20$) and high ($S = 40$) levels of gene flow during diversification. Gene flow has a key role in the spatial organization of the populations, going from clustered to totally mixed in space. For small S, speciation is preceded by the accumulation of local genetic variation within species. As a result, genetic and physical spaces become coupled by the dynamics and species form in well defined clusters. As S increases this correlation is lost and species mix in space (see first row in Fig. 3.3 from left to right). The time to the most recent common ancestor of all the extant species of the tree (tree-MRCA) is similar for all values of S (see the time of first bifurcations in the phylogenies in Fig. 3.3), which is around 100 generations.

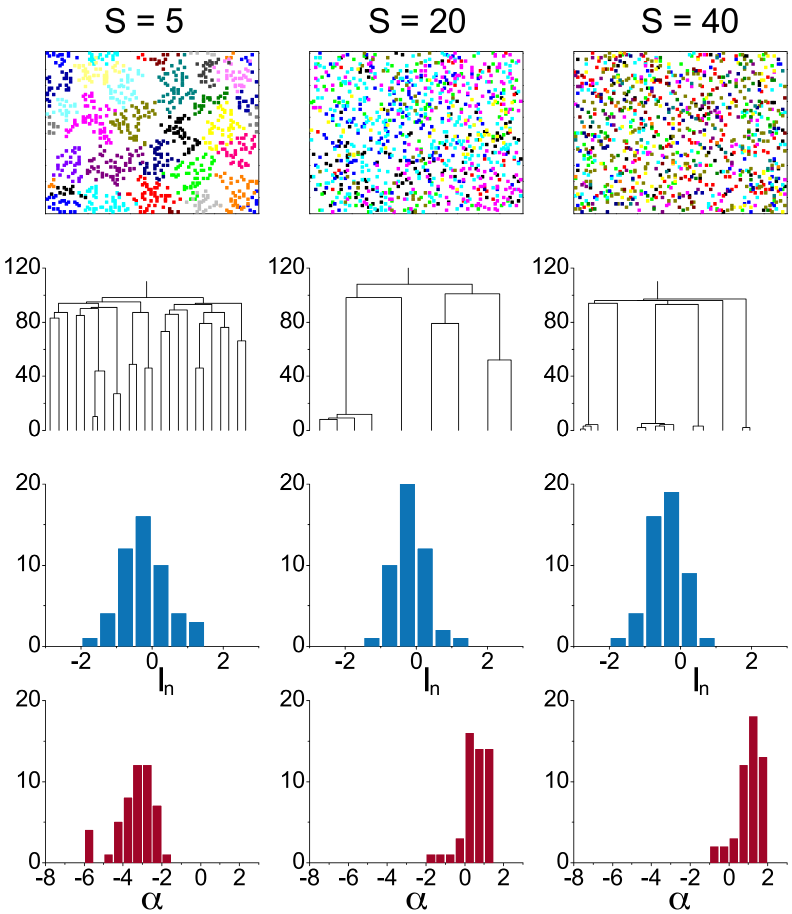


FIGURE 3.3: Simulations for $B = 150,000$ and different values of S. From top to bottom rows: spatial distribution of the populations (different colors show different species); examples of corresponding phylogenetic trees; normalized Sackin index distribution (I_n); α-value distribution (α). The distributions were computed for 50 realizations of the process.

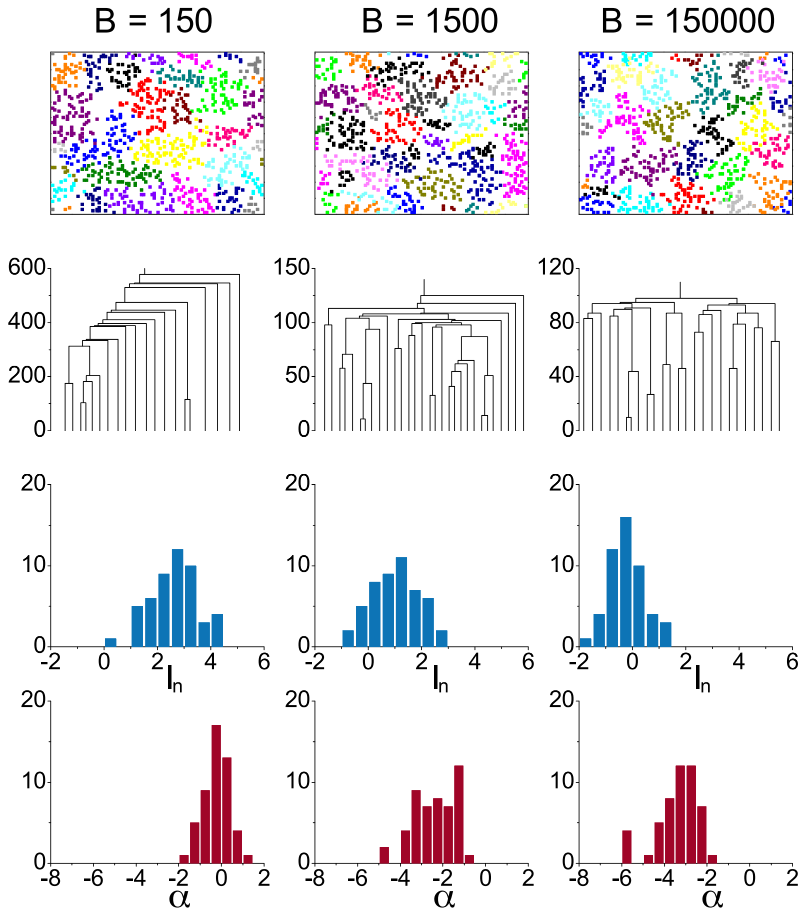


FIGURE 3.4: Simulations for $S = 5$ and different values of B. From top to bottom rows: spatial distribution of the populations (different colors show different species); examples of corresponding phylogenetic trees; normalized Sackin index distribution (I_n); α-value distribution (α). The distributions were computed for 50 realizations of the process.

The last two rows show the distribution of normalized Sackin index and α-value. The normalized Sackin index distributions (third row) are largely insensitive to S and indicate balanced trees, as predicted by the Yule model (Yule, 1925). Distribution centers and standard deviations are -0.1 ± 0.6, -0.2 ± 0.4 and -0.4 ± 0.5 for $S = 5$, 20 and 40 respectively. In Fig. B.7 in SI B we superimposed the simulated histograms with the predictions of an asymmetric Yule model (see Section B.3 in SI B for more information). Alternatively, the acceleration of the diversification rate, here characterized by the α-value, is highly influenced by gene flow: with stemmy trees when gene flow is restricted and tippy trees when gene flow is high (see second row in Fig. 3.3). Distribution centers and standard deviation of alpha values are: -3.4 ± 0.9 ($S = 5$), 0.6 ± 0.6 ($S = 20$), and 1.0 ± 0.7 ($S = 40$).

3.3.2 Small mating range

Figure 3.4 shows the effect of genome size (for fixed $G/B = 0.05$) on phylogenetic trees when speciation is strongly dependent on space (parapatry, $S = 5$) with spatial distributions of species (first row), and the corresponding phylogenetic trees (second row). Now, columns correspond to increasing values of genome size (B) and compatibility threshold (G). In all cases, species display strong correlation between genetic and physical spaces, as expected for a small S. However, the time to the most recent common ancestor of all the extant species of the tree (tree-MRCA) depends critically on B and G. For $B = 150$, $G = 7$ the time to the tree-MRCA is around 600 generations (see the time of first bifurcation in the left phylogeny in Fig. 3.4), which differs from the time species started to be formed (around 200 generations, see the time the orange curve detaches from the time axis on Fig. 3.2), *i*.e., species formed at the initial radiation (between 200 and 600 generations) went extinct. As B increases, the time of the first speciation event decreases and the equilibration time occurs earlier (Fig. 3.2).

The distribution of normalized Sackin index (third row) are highly unbalanced for small B ($I_n = 2.6 \pm 0.9$ for $B = 150$, $G = 7$). As B increases the center of the distribution shifts monotonically to the left: $I_n = 1.0 \pm 0.9$ for $B = 1,500$, $G = 75$ and $I_n = -0.1 \pm 0.6$ for $B = 150,000$, $G = 7,500$. In Fig. B.7 in SI B, we superimposed the predictions of the asymmetric Yule model for the histograms, with $\delta = 0.3$ for $B = 150$, $G = 7$, $\delta = 0.65$ for $B = 1,500$, $G = 75$ and $\delta = 0.8$ for $B = 150,000$, $G = 7,500$. The

α-value distributions (last row) also display strong dependence on genome/threshold size, with speciation rate decelerating as genome size increases ($\alpha = -0.3 \pm 0.6$ for $B = 150$, $G = 7$, $\alpha = -2.4 \pm 1.0$ for $B = 1,500$, $G = 75$ and $\alpha = -3.4 \pm 0.9$ for $B = 150,000$, $G = 7500$).

3.3.3 Comparison with empirical data

Figure 3.5 shows how the acceleration of diversification (α-value) and tree balance (normalized Sackin index, I_n) are influenced by mating range and genome size in the simulated model (colored dots and ellipses). The mating range affects specially the acceleration of diversification ($F = 508.18$, $\eta^2 = 0.87$), but the balance of the trees are only marginally affected by this parameter ($F = 3.39$, $\eta^2 = 0.04$). Genome size, for fixed G/B, affects both the acceleration of diversification ($F = 126.08$, $\eta^2 = 0.70$) and tree balance ($F = 151.53$, $\eta^2 = 0.67$). The numbered symbols in Fig. 3.5 show indexes of phylogenetic trees estimated from the empirical radiations listed in Table 3.1 (see also the phylogenetic trees in Figs. B.11(a)-(p) in SI B).

Most empirical points fall in two regions, corresponding to: (i) $B = 150,000$ (large genome size) with $S = 20$ or $S = 40$ (moderate to high gene flow) and balanced trees – green and yellow regions (1, 2, 3, 4, 5, 10, 11, 12, and 13), and (ii) $B = 150$ (small genome size) with $S = 5$ (low gene flow) and unbalanced trees – orange ellipse (points 6, 7, 8, and 15). The Malawi cichlids (point 9) fall in the region of low gene flow with large genome. The lichen *Sticta* and the New World titi monkeys (points 14 and 16) fall between the low/high gene flow and small/large genome size regions (orange/green-yellow ellipses). The α-value close to 1, corresponding to constant speciation rate per branch, was characteristic of speciation with moderate or high levels of gene flow ($S = 20$ and 40, sympatry). Simulations of low levels of gene flow ($S = 5$, parapatry), on the other hand, displayed average values of α-value smaller or equal to 0, associated with a slow down of diversification rate.

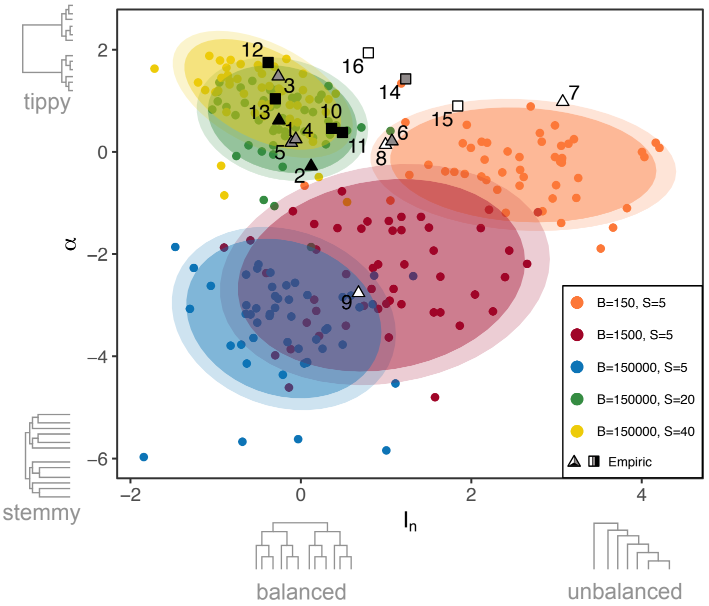


FIGURE 3.5: Distribution of α-value and normalized Sackin index I_n for different model parameters. Populations evolved up to species number equilibration time (50 replicates). Ellipses comprehend 90% (darker) and 95% (lighter) of the simulated replicates for each parameter combination. Numbered triangle and square symbols represent phylogenetic tree indexes estimated from empirical data as listed on the Table 3.1 in the MATERIALS AND METHODS section. Triangles represent adaptive and squares represent non-adaptive radiations. The colors of triangles and squares are related to the gene flow during diversification as defined on Table 3.1: white = low, gray = intermediate, black = high.

3.4 Discussion

In this paper we investigated how different geographical modes of speciation leave signatures on phylogenetic trees using a neutral spatially-explicit IBM. The origin of new species resulted from spatial and genetic thresholds, denoting sympatric and parapatric speciation under certain parameter values. Phylogenies were constructed by tracking ancestor-descendant relationships between individuals (Costa et al., 2018a). We simulated evolutionary radiations and calculated two indexes assessing tree topology: tree balance (Sackin index) and acceleration of diversification ($alpha$-value). We compared the indexes obtained from simulated results with those found in trees estimated from empirical data of adaptive and non-adaptive radiations. We found signatures of the geographical mode of speciation displayed as the relationship between the acceleration of diversification (y-axis in Fig. 3.5), and the balance of the trees (x-axis in Fig. 3.5). Adaptive and non-adaptive empirical radiations exhibit macroevolutionary patterns that are consistent with gene flow simulated for the neutral model (squares and triangle symbols in Fig. 3.5).

Evolutionary radiations are often thought to be characterized by either a slowing down in diversification rates, associated with ecological speciation and the process of niche filling (negative diversity-dependence) (Phillimore and Price, 2008; Rabosky and Lovette, 2008; Higgs and Derrida, 1991; Gascuel et al., 2015a), or by a speeding up in diversification associated with coexistence of newly branched species driven by ecological interactions (positive diversity-dependence often referred as "diversity begets diversity") (Emerson and Kolm, 2005; McPeek, 2008; Gascuel et al., 2015a; Burin et al., 2018). In our model, the acceleration of diversification (α-value) is an important measure of the overall speeding-up or slowing down of speciation rate during the radiation (see Fig.B.9 in SI B). We specifically refer to speciation rates because extinctions are either irrelevant (for large genomes) or constant (for small genomes) during the radiation (see Section B.1 in SI B). Speciation rate has an accelerated initial phase followed by a slowing down before equilibration, which results in an S-shaped curve of number of species through time (Fig. 3.2 and Fig. B.10 in SI B). For small genomes, however, this pattern is attenuated. The geography of the diversification process affects diversification slowdown (Moen and Morlon, 2014; Pigot et al., 2010). Stable geographic ranges have been associated with slowdown in diversification rates (Pigot et al., 2010), which in our model encompasses parapatric radiations with large genomes and resulting in slowdown of diversification rates (negative α-value). Alternatively, sped up diversification rates are characteristic of sympatric radiations with large genome or parapatric radiations with small genome (positive α-value). We emphasize that we have fixed the ratio between threshold value and genome size (G/B) in all simulations, so that small(large) genome sizes always implies small(large) compatibility threshold. Peripatric speciation produces highly unbalanced trees (Pigot et al., 2010), which, here, were associated with parapatric speciation (positive I_n). The distinction between sympatric and parapatric radiations with nearly constant diversification rates is in the tree balance. We reinforce the conjecture that when considering diversification alone, similar patterns of diversification rate in time can result from different processes (McPeek, 2008; Yoder et al., 2010). In our results this is reflected in the size of the elipses representing the different scenarios simulated here. The stochastic nature of the model implies that different trees can be generated with the same parameter values and, more importantly, different parameters can produce similar trees.

Although the 'low resolution' manifested by the size of the elipses makes the connection between trees and processes not so acurate, sympatry, represented by the green and yellow elipses (balanced and tippy trees) can still be clearly distinguished in the two-dimensional tree topology space from parapatry, blue and orange elipses for large and small genomes/thresholds, going from balanced stemmy to unbalanced tippy (Lewitus and Morlon, 2016).

The spatial scale relevant to the speciation process depends on the intensity of gene flow, so that species with lower levels of gene flow require smaller areas to speciate (Kisel and Barraclough, 2010). In island radiations, reduced gene flow increases the probability of speciation (Kisel and Barraclough, 2010). We report a similar pattern, as for reduced gene flow (low S), speciation can occur regardless of the genome/threshold size considered. In models of adaptive speciation, however, the source of genetic divergence emerges from competition and gene flow promotes speciation (Dieckmann and Doebeli, 1999). The connection between the spatial component of the speciation process and the degree of gene flow brings up the question of how the underlying process driving genetic divergence affects diversification in the macroevolutionary scale. In our case, genetic divergence evolves as a consequence of the emergence of reproductive isolation, arising directly from microevolutionary processes incorporated in the model, and the degree of gene flow can be directly related to the size of the mating range and, consequently, to the geography of speciation. When the genome is large, mutations spread over loci uniformly, so that the average genetic distance between any two individuals is about the same and increases monotonically with time (see Section B.2 in SI B). When the average distance reaches the threshold G the population splits very quickly into several species (see Section B.7 in SI B) giving rise to stemmy trees. However, because there is potential gene flow between all individuals until very close to speciation, species form with relatively large populations and split again later, producing symmetric trees. For small genomes speciation is slower, mutations accumulate closer to each other and species arise one by one from the main population, giving rise to asymmetric trees (Fig. 3.4). Mating range affects gene flow directly, so that large values of S slows down speciation, leading to tippy trees. Because speciation for large S only occurs for large B, the trees will also be symmetric.

An additional component contributing to the evolution of genetic divergence in our model is related to the combined effect of genome and compatibility threshold size with large values facilitating speciation. Interestingly, this is also a dynamic feature found in adaptive evolution models (Flaxman et al., 2014). Studies investigating the relationship between genome size and diversification rates using empirical data show that, in most cases, small-genome taxa present higher diversification rates across different species groups (Kraaijeveld, 2010). This relation is grounded in the assumption that the smaller the genome, the quicker the genomic rearrangements, and the greater the rate in which mutations can accumulate (Kraaijeveld, 2010). This body of assumptions are not met by our modeling approach, as we do not incorporate genomic rearrangements and their consequences to organism function. In our model, mutations accumulate at the same rate despite genome size and population divergence appears to be faster as genome size increases. In fact, in angiosperms, there is strong evidence for a positive relation between the rate of genome size evolution and speciation rate, where genome size evolution is mainly related to genome duplications (as polyploidy) (Puttick et al., 2015). Also contributing to the debate between genome size and diversification rates, and in agreement with our results, actinopterygiian fishes present a positive correlation between genome size and species richness among genera (Mank and Avise, 2006). Even though there are evidences supporting the relationship between genome size and rate of diversification, we do recognize they are not striking and the debate is still open (Kraaijeveld, 2010). Additionally, we are aware that genome size or the number of genes involved in speciation are parameters hard to correlate and estimate in empirical systems. The signatures left by the effect of genome size on the speciation process may be smaller or greater than the actual measurable genome metrics (for instance, C-value and G-value). This effect can be even more pronounced given nonlinearities of genetic architecture. The part of the genome under selection, the amount of non-coding DNA, genetic interactions (pleiotropy and epistasis) and splicing processes can alter the effective size of the genome (Taft et al., 2007; Sanjuán and Elena, 2006; Wagner et al., 2007; Brett et al., 2002).

Empirical data display structural phylogenetic patterns that can be associated with different geographic modes of speciation in our model. Most of the data presented here have tree balance and speed of diversification statistics falling into model predicted regions that are in accordance with the hypothesized degree of gene flow during the radiation of these systems (Table 3.1 and Fig. 3.5). Groups with intermediate gene flow (Hawaiian silversword alliance, *Anole* lizards, *Tetragnatha* spiders, Darwin's finches, and lichen *Sticta* – points 3, 4, 5, 6 and 14, respectively) were located in varied regions, with some falling into either sympatric (points 3, 4, and 5) or parapatric (point

6) regions, and one (point 14) located in the transition between these regions. This variation could be attributed to the myriad of factors promoting speciation in these groups as they all belong to island systems, in which radiations are especially prone to occur with parapatric and sympatric components playing preeminent roles during diversification (Rundell and Price, 2009; Simoes et al., 2016). The phylogenetic tree of the New World titi monkeys (point 15), hypothesized to have radiated with low gene flow, does not resemble the trees generated with low gene flow by our model, which might signalize that rivers are not effective barriers to gene flow as previously suggested for the Amazonian monkey groups (Santorelli et al., 2018). Most of empirical trees are tippy, indicating that this pattern might not only be attributed to non-ecological modes of speciation but could also be a prevailing pattern in evolutionary radiations independently of the nature of processes driving diversification (McPeek, 2008).

The empirical adaptive (non-neutral) radiations also fell in structural phylogenetic regions simulated by our neutral model. This suggests that the geographic mode of speciation can leave signatures in the balance and acceleration of diversification of phylogenetic trees, irrespective of the process being adaptive or non-adaptive. On the other hand, it is also possible that adaptive processes generate phylogenetic patterns overlapping with those predicted by our neutral model. Further studies including adaptive processes are needed to resolve this matter. A prolonged geographical isolation in the early history of a system resulting in multiple reproductive isolated species, characterizing a non-adaptive radiation, could also be confounded with an adaptive one, as species may diversify ecologically when conditions change (Rundell and Price, 2009). Even in cases traditionally considered as adaptive radiations, such as Darwin's finches and African cichlids, previous studies (Grant and Grant, 2011; Losos and Ricklefs, 2009; Seehausen, 2015; Moen and Morlon, 2014; Simoes et al., 2016) show that geographical processes seem to be an important promoter of diversification patterns along with competition for resources and niche filling. Other examples of this misdiagnose can be found in systems with many species within a radiating clade with allopatric distributions, such as the *Ctenomys* caviomorph rodents, *Anoles* lizards, *Tetragnatha* spiders, Darwin's finches, and Hawaiian silversword alliance displayed here (Simoes et al., 2016). Although some macroevolutionary patterns – such as an early burst in diversification and overshooting effect – have been considered signatures of adaptive radiations (*i.e.* fueled by adaptation to distinct niches), there is a debate about whether non-ecological processes may result in similar patterns (McPeek, 2008; Yoder et al., 2010; Gascuel et al., 2015a). These findings might explain the matching between the patterns observed in empirical adaptive radiations and our simulated neutral radiations, although only the inclusion of non-neutral processes in the model (as in Gascuel et al., 2015a) would provide clearer interpretations.

Observed association between ecological divergence and diversification does not necessarily indicate that speciation was caused by adaptation to different environments, because ecological differences can evolve after diversification (Rundell and Price, 2009). Moreover, niche conservatism followed by allopatric divergence may lead to similar patterns of early diversification (Rundell and Price, 2009). Other non-ecological models of rapid radiations have predicted dispersal limitation and isolation of species without environmental differentiation as potential drivers of rapid diversification (Wiens, 2004; Rundell and Price, 2009). Factors associated with non-adaptive (neutral) and adaptive (non-neutral) radiations could operate synergistically (Simoes et al., 2016), being difficult in empirical studies to isolate the relative importance of each process through the phylogenetic summary statistics employed here. The *Tetragnatha* spiders are an illustration of this mixture of processes, as they contain strongly ecologically differentiated sympatric species, as well as ecologically similar allopatric or parapatric species (Gillespie, 2004; Rundell and Price, 2009). Additionally, some signatures could be left during the radiation specifically by adaptive process that might be better detected by other measures of phylogenetic structure, such as phylogenetic diversity.

Our results contribute to bridge the gap between microevolutionary processes and macroevolutionary patterns. We incorporated a broad range of speciation mechanisms in which the microevolutionary forces at the individual level, such as gene flow and genomic architecture, play important roles in the diversification. We have shown that speciation based on genetic and spatial restrictions imposed on reproduction predict clear macroevolutionary patterns in phylogenetic trees. Our neutral model recovers a wide range of patterns associated to evolutionary radiations, including the speeding up, followed by the slowing down of diversification rates and even the overshooting effect, which are considered a signature of adaptive radiations (Higgs and Derrida, 1991). Our findings also corroborate previous results in that the geographic mode of speciaton is important to the tree balance: sympatric speciation produces more balanced trees than parapatric speciation (Barraclough and Vogler, 2000; Losos and Glor, 2003; Pigot et al., 2010). We have shown these effects for neutral simulated radiations and observed similar patterns in both adaptive and non-adaptive empirical sets. We hypothesize that the observed signatures in Sackin index and alpha-value in evolutionary radiations are possibly related to gene flow and genome size. Confirmation of this hypothesis would require further investigations with models of adaptive radiations. We further hypothesize that adaptive/ecological components leave signatures in other phylogenetic tree statistics. For instance phylogenetic diversity (Faith's PD), and the biogeographically weighted evolutionary distinctiveness (BED), (Tucker and Cadotte, 2013) might summarize more accurately the role played by the adaptive components of evolutionary radiations. Expanding our analysis to include adaptive speciation processes will be important to improve our understanding on the phylogenetic signatures. Even within the framework of our model other microevolutionary features, such as population density (M/L^2), compatibility threshold (G/B), dispersal probability, evolution of genome size and mating range (variation of B and S) and mutation rate (μ) can be related to realistic features and might also be explored to understand the importance of other microevolutionary processes on the macroevolutionary patterns.

3.5 Supplementary Material

Supplementary material, including data files and online-only appendices, can be found in the Dryad data repository `https://datadryad.org/review?doi=doi:10.5061/dryad.4kn5j5d`.

3.6 Funding

This work was partly supported by the Brazilian agencies: Capes, grant #302049/2015-0 (CNPq), and grants #2016/06054-3 São Paulo Research Foundation (FAPESP) and 2016/01343-7 (ICTP-SAIFR FAPESP). FMDM is supported by grant #2015/11985-3 (FAPESP), ABM is supported by grant #2014/10470-7 (FAPESP), LDF is supported by grant #2015/26989-4 (FAPESP), MFR is supported by grant #152885/2016-1 (CNPq).

3.7 Acknowledgments

We are very grateful to Prof. Rosemary Gillespie, Dr Eulalia Banguera-Hinestroza, Dr Hazel Byrne, Dr Alicia Álvarez, for the support on the dataset of the *Tetragnatha*, Delphinidae, New World titi monkeys, and Caviomorph rodents *Ctenomys* trees, respectively. We thank Gustavo Burin, Laura Alencar, Nathália Caldeira and Tiago Quental for the inspirational comments on the manuscript. Author contributions: MAMA designed research; MAMA, CLNC, PLC and FMDM performed the research; LDF, MFR, DMS led the numerical analysis and theoretical model; CLNC, PLC, FMDM, ABM, and MAMA contributed substantially to the analysis and writing.

Chapter 4

Speciation promoted by geographic barriers: modeling the Riverine Hypothesis with an individual-based model

Manuscript made by: Carolina L.N. Costa, S. Ivan Perez, Marcus A.M. de Aguiar. 2018 Target journal: Evolution.

Abstract

The development of speciation models to envision patterns of diversification emerging from specific mechanisms is a powerful tool in theoretical biology. Individual-based models are crucial to understand how processes acting on the population level leave signatures on large-scale diversification patterns. Here we used a spatially explicit individual-based model to investigate the role of geographic barriers in shaping macroevolutionary and distributional patterns of an initially homogeneous population which evolve to multiple species. The population evolves under the combined effects of sexual reproduction, mutation and dispersal. Reproductive isolation was based on a multilocus generalization of the Bateson-Dobzhansky-Muller model, with individuals more distant than a genetic threshold being reproductive isolated. The speciation model was adapted to a particular system, subfamilies and genera of the New World monkeys, with the spatial lattice designed with the South America shape, and barriers representing the main rivers of the region. We test if rivers act as effective geographic barriers to monkeys living in these areas, promoting allopatric speciation as proposed by the Riverine Hypothesis. Macroevolutionary patterns analyzed were the Sackin index and the α-value, which measure tree symmetry and speed of diversification, respectively. The distribution of species richness in Operative Region Units was also analyzed. Comparisons with empirical data of Platyrrhini showed a great correspondence between the patterns of simulations and empirical information, suggesting that neutral processes associated with spatial barriers can explain some patterns of diversity of Platyrrhini monkeys, giving plausibility to the Riverine Hypothesis.

Keywords spatial models | neutral processes | spatial barriers | macroevolutionary patterns | distribution patterns | New World monkeys

4.1 Introduction

A major aim in evolutionary biology is to understand what processes are responsible for the diversification of clades. The resulting branching pattern of this species diversification have been extensively studied using phylogenetic (Yang and Rannala, 2012), and biogeographic approaches (Cox et al., 2016). Studies using these approaches have mainly discussed the pattern of species diversity in living clades in the frame of adaptive radiation models (Simpson, 1955; Schluter, 2000), where the branching process and speciation are linked to niche differentiation and selection (Losos and Mahler, 2010; Beaudrot and Marshall, 2011).

One group which has been extensively studied in the context of niche differentiation and selection (for a review, see Schneider and Sampaio, 2015) is the New World monkeys (Parvorder Platyrrhini), one of the three major monophyletic groups of living and fossil primates. In this clade the diversification process occurred during a large temporal scale (*i.e.*, 20-40 million years ago or megannums [Ma] in Central and South America), displays a remarkable variation in the number of species among clades (Fabre et al., 2009; Perelman et al., 2011; Aristide et al., 2015) and is associated to a great phenotypic variation (e.g., a body mass spanning two orders of magnitude, from 0.1 to more than 10 kg; Rosenberger, 1992; Aristide et al., 2015). Previous phylogenetic and morphological studies have suggested that the early species diversification in platyrrhine was mainly linked to selection (Rosenberger, 1992; Rosenberger et al., 2009; Beaudrot and Marshall, 2011; Aristide et al., 2015). Particularly, it was suggested that a large number of species originated quickly during the early stages of the platyrrhine evolutionary history related to the ecological niche differentiation or adaptive radiation of the clade (Hodgson et al., 2009; Rosenberger et al., 2009; Wildman et al., 2009; Aristide et al., 2015). These speciation events among the early platyrrhines possibly generate the main clades of species that conform the extant platyrrhine families (Aristide et al., 2015).

Alternatively, the more recent pattern of origination and diversity of platyrrhine species, which conforms the species diversity within genera or subfamilies, has been considered the effect of neutral forces as important promoters of speciation (Cropp et al., 1999; Chiou et al., 2011; Lynch-Alfaro et al., 2012; Boubli et al., 2015). The non-selective mechanisms involved in the origin of this pattern of diversity are long unknown and several hypotheses have been discussed. Among them, the Riverine Hypothesis, firstly formulated by Wallace (Wallace, 1854), postulates that major Amazonian rivers act as barriers to gene flow between monkey populations inhabiting opposite sides of river banks, promoting speciation by vicariance (see also Ayres and Clutton-Brock, 1992; Boubli et al., 2015). Several empirical studies have supported the Riverine Hypothesis, showing that for a variety of clades – mainly within primates, but also for lizards, anurans, butterflies, birds – the rivers possibly act as geographic barriers to species dispersal and hence promote speciation without necessarily the presence of selective forces (Ayres and Clutton-Brock, 1992; Gascon et al., 2000; Hall and Harvey, 2002; Butlin et al., 2009; Pomara et al., 2014; Rabelo et al., 2014; Boubli et al., 2015; Byrne et al., 2016; Moraes et al., 2016).

However, this hypothesis has been questioned for the same and other phylogenetic clades, since species distributions are not limited by Amazonian rivers or because the time scale of river formation and species diversification does not coincide (Boubli et al., 2015; Santorelli et al., 2018). These authors have suggested that rivers could be important delimiters of population or species distribution without necessarily causing speciation (Boubli et al., 2015; Santorelli et al., 2018). Alternative mechanisms like sympatric speciation combined with dispersal limitation could produce the same patterns as that observed in studies supporting the Riverine Hypothesis (Beaudrot and Marshall, 2011; Santorelli et al., 2018). Moreover, recent works suggest that geographic isolation, together with diffusion processes, climate change and/or Pleistocene refugia, could be a central cause of the pattern of species diversity within several platyrrhine sub-families (Cropp et al., 1999; Chiou et al., 2011; Lynch-Alfaro et al., 2012).

Despite the great advances in the description of phylogenetic relationships among platyrrhine species (Fabre et al., 2009; Perelman et al., 2011; Aristide et al., 2015), the great knowledge of their geographical distribution (Van Roosmalen et al., 2002; Rylands et al., 2009; Rylands et al., 2016) and the large number of contrasting speciation hypotheses (Lynch-Alfaro et al., 2012; Boubli et al., 2015; Santorelli et al., 2018), the question about the main factors responsible for the recent platyrrhine diversification remaining unanswered (Schneider and Sampaio, 2015). Here, we use a theoretical approach to help to solve the contradictions observed in empirical studies regarding the Riverine Hypothesis.

In our theoretical approach the mechanisms responsible for the diversification pattern in an evolutionary scale, such as that observed in platyrrhine subfamilies and genera, are neutral. The neutral theory, in which the origin of diversity is mainly related to non-selective factors, has been remarkably useful to explain observed patterns of species diversity (Hubbell, 2001; Rosindell et al., 2010; De Aguiar et al., 2009; O'Dwyer and Green, 2010; Martins et al., 2013; Manceau et al., 2015). Particularly, neutral models place the emphasis on dispersal limitation, by geographical barriers or isolation by distance, to explain the patterns of species diversity (De Aguiar et al., 2009; Martins et al., 2013), which is consistent with the Riverine Hypothesis (Beaudrot and Marshall, 2011).

Specifically, our aim is to test if geographical barriers (e.g. amazonian rivers) have played a central role in diversification in the recent platyrrhine divergence (*i.e.*, at the genera and subfamily scales). To do this we used a neutral model of speciation based on de Aguiar et al. model (De Aguiar et al., 2009; De Aguiar, 2017). The de Aguiar et al. model has provided good fits to empirical macroecological and macroevolutionary patterns of diversity in previous studies (De Aguiar et al., 2009; Martins et al., 2013; Costa et al., 2018b), being a helpful tool to investigate the mechanisms behind the platyrrhine diversification. In this paper, we expand the spatially explicit de Aguiar et al. model, in which species evolve in response to genetic and spatial restrictions imposed on reproduction, to include geographic barriers and individual's limitation to cross these barriers, allowing the test of the Riverine Hypothesis.

The spatial lattice is a representation of part of South America, and the geographic barriers have the same shape of main rivers of the region. We registered the evolutionary history of species using an algorithm that records all speciation and extinction events throughout the evolutionary process – the Sequential Speciation and Extinction Events (SSEE) algorithm, included in the de Aguiar model in (Costa et al., 2018a). In our expansion we included a mitochondrial genome to each individual besides the nuclear genome previously incorporated. With both nuclear and mitochondrial genomes we estimated phylogenetic trees from distance-based methods and compared them to the true trees generated by the SSEE algorithm. Finally, we record the spatial coordinates of all individuals to know the spatial distribution of species in the riverine lattice. We compare the phylogenetic and distribution patterns arising from the neutral model – with different levels of geographic barriers – with the diversity empirically observed in phylogenetic trees and geographical distribution of plathyrrhine. The structural properties of phylogenetic trees were compared employing Sackin and alpha statistics (Sackin, 1972; Costa et al., 2018b), the second being a normalization of gamma statistics (Pybus and Harvey, 2000). The geographic distribution was analysed employing species richness calculations from biogeographic regions knowed as Operative Geographic Units (OGUs).

4.2 Material and Methods

4.2.1 The recent Platyrrhini radiation

To understand the processes responsible for the diversification patterns observed in genera and subfamilies of platyrrhines we obtained background information about the system, further using these informations in a neutral model of speciation to test the Riverine Hypothesis.

First, we measure river width separating extant species for each genera and subfamily (Table C.1 in Supporting Information C). The river width information was obtained from satelite images by measuring three distinct points of river: the source, the middle course and the mouth (Table C.2). We also collect species range and body mass data for these species and estimate the correlation between the mean values for these variables and mean river width for the genera (Table C.1). Geographic distribution was obtained from (IUCN, 2015), and body size from (Aristide et al., 2015). We observed that there is a correlation between body mass and river width (Figs. C.1, C.2), with small species distribution being delimitated to narrower rivers than large species, which could implies that smaller species cannot cross larger rivers. Second, we reconstructed maps describing the distribution for all extant platyrrhine species. We observed that genera with differences in body size displayed different patterns of geographic distribution (Fig. C.3), reinforcing the hypothesis that rivers act as barriers to platyrrhine species.

Because body size could be related to the chance that rivers will be a geographic barrier for platyrrhine species (Figs. C.1, C.2, C.3) (Ayres and Clutton-Brock, 1992; Gascon et al., 2000), we incorporated this variable in our model. We divided platyrrhine in three categories: small monkeys (Callithrichinae subfamily, *Saguinus* and *Callicebus* genera), medium monkeys (Pitheciidae family and *Cebus* genera) and large monkeys (Atelidae family). We did not consider the *Saimiri* and *Aotus* genera because of their intermediate body size between small and medium, difficulting to categorize these monkeys. The parameter related to body size was the radius of mating range, S, which has greater values for larger monkeys, assuming that larger monkeys have a greater range to search for mates and, consequently, can cross more rives than small monkeys. To assess if rivers are the main factor shaping geographic distribution and diversity patterns of species in the model we did simulations under three spatial scenarios: 1) South America lattice without rivers, 2) South America lattice with rivers, all with equal width and 3) South America lattice with river sizes proportional to real river widths (Fig. C.4). In scenario 3, the proportional size of rivers was stablished using empirical data (Table C.2).

As the mean generation time vary for each family/subfamily, we defined the number of generations used to run simulations using empirical information about divergence time, generation time and mutation rate for each group, to give a more realist ground for our simulations. First, with empirical divergence and generation times (Aristide et al., 2015; Mittermeier et al., 2013) we calculated the number of generations that have occurred until the present for each group (small, medium and large platyrrhine monkeys). Second, with an empirical mutation rate estimated for great apes ($1.36x10^{-8}$ mutations/site/generation, Scally and Durbin, 2012) we calculated the number of mutations that have happened until the present in the three groups. Finally, with the overall number of mutations estimated, we calculated the number of generations T that should be run in our simulations, given a fixed simulated mutation rate, to reach the same overall number of mutations estimated empirically. These informations are summarized in Table 4.1.

TABLE 4.1: Empirical information about mutation rate, divergence time and generation time for each platyrrhine group utilized in to define model parameters.

Group	Size category	Empirical mutation rate (mut/sit/gen)	Divergence time (Mya)	Mean generation time (yr)	Empirical number of generations	Simulated mutation rate (mut/sit/gen)	Simulated number of generations T
Callithrichinae	Small	$1.36x10^{-8}$	15	6	$2.5x10^{6}$	$0.5x10^{-4}$	680
Pitheciidae	Medium	$1.36x10^{-8}$	12	10	$1.2x10^{6}$	$0.5x10^{-4}$	320
Atelidae	Large	$1.36x10^{-8}$	17	15	$1.13x10^{6}$	$0.5x10^{-4}$	300

4.2.2 The speciation model

The model we discuss here is adapted from (De Aguiar et al., 2009) and (Costa et al., 2018a). Here we are adding spatial barriers to the model to give a more realistic representation of platyrrhine diversification context during the speciation of subfamilies and genera, hence allowing to test the Riverine Hypothesis. We also included a mitochondrial genome to the model to increase the possible comparisons of phylogenetic patterns resulted from simulations with empirical data.

The model considers a population of N haploid individuals randomly distributed in a lattice $L_1 x L_2$ with forbiden sites considered as rivers. The nuclear genome of each individual is represented by a binary string of size B, $\{\sigma_1^i, \sigma_2^i, \ldots, \sigma_B^i\}$ for individual i, where each *locus* σ_k^i can assume the allele values -1 or 1. As in the (Higgs and Derrida, 1991) model, population is characterized by a MxM matrix q (the overlap matrix), that measures the degree of genetic similarity between pairs of individuals:

$$q^{\alpha\beta} = \frac{1}{B} \sum_{i=1}^{B} S_i^{\alpha} S_i^{\beta}. \tag{4.1}$$

If individual α and β have identical nuclear genomes, $q^{\alpha\beta} = 1$. The initial condition of the simulation is all individuals genetically identical, which means the matrix q is composed only by the value 1.

Each generation is constructed from the previous one, with a reproduction trial that starts with individual 1 and goes to individual N, so all individuals have a chance to reproduce. The individual selected to reproduce, the first parent P_1, searches for potential mates in its mating range, a circular area of radius S centered on the location of P_1. The first parent can only reproduce with an individual, the second parent P_2, if they are compatible, *i.e.*, their degree of similarity satisfies $q^{P_1,P_2} \geq q_{min}$. The minimal similarity necessary to mate, q_{min}, represents a threshold below which individuals do not recognize each other as belonging to the same species (Higgs and Derrida, 1991; Gavrilets, 2014). The second parent P_2 is randomly chosen from the mating range of the first parent P_1. If P_1 do not have compatible mates in its mating range, another first parent is selected in its place, which keeps the population size aproximately constant. Once the two parents are chosen, reproduction happens with the generation of an offspring, which inherits, gene by gene, the allele of either parent with equal probability. After combination of parental genomes, each *locus* in offspring genome has a probability μ of mutate. We considered the limit $B \to \infty$, which allows the direct manipulation of the overlap matrix $q^{\alpha\beta}$ rather than storing all the genome sequences (Higgs and Derrida, 1991). In the limit $B \to \infty$, each offspring inherits half the alleles from each parent, and the similarity matrix q between any individuals α and β is updated as follows

$$q^{\alpha\beta} = \frac{e^{-4\mu}}{4}(q^{P_1(\alpha)P_1(\beta)} + q^{P_2(\alpha)P_1(\beta)} + q^{P_1(\alpha)P_2(\beta)} + q^{P_2(\alpha)P_2(\beta)}) \tag{4.2}$$

in which $q^{\alpha\alpha} \equiv 1$. Therefore, in the limit of infinitely genes, the dynamics can be obtained by simply updating the similarity matrix according to equation 4.2.

Additionally to the inheritance of nuclear genome, the offspring also acquire the mitocondrial genome from one of the parents with equal probability, since individuals are hermaphroditic. The mtDNA is characterized by a string of size MT, $\{\sigma_1^i, \sigma_2^i, \ldots, \sigma_{MT}^i\}$ for individual i, where each *locus* σ_k^i can assume the allele values $0, 1, 2$ or 3, representing the nitrogenous bases A, T, C and G. Different from nuclear genome, the mtDNA is finite and it is not considered in the search of potential mates, *i.e.*, there is not a threshold in mtDNA similarity below which two individuals cannot reproduce. The mtDNA is only transmitted from one generation to another. The mtDNA is passed to the offspring without recombination, having only the probability μ of mutate. After the acquisition of both nuclear and mitochondrial genomes, the offspring is dispersed with probability D to one of the sites in the mating range of the expiring parent P_1.

The evolutionary dynamics of the population follows in non-overlapping generations, with the emergence of species through microevolutionary process as mutation, gene flow and genetic drift. Species are defined as groups of individuals connected by gene flow and reproductively isolated from all others by the threshold q_{min}. Therefore, two individuals belonging to the same species can be reproductively incompatible ($q^{\alpha,\beta} < q_{min}$), but if they exchange genes indirectly through an intermediate ($q^{\alpha,\delta} \geq q_{min}$ and $q^{\beta,\delta} \geq q_{min}$) they are in the same species (Costa et al., 2018b). The model presented here is considered neutral because differences between individuals are irrelevant for their birth, death and dispersal rates (Hubbell, 2001).

The evolutionary history of the simulated species was got by an algorithm that directly records all speciation and extinction events throughout the evolutionary process, generating a matrix with the true phylogeny of species (the Sequential Speciation and Extinction Events, SSEE) (Costa et al., 2018a). The SSEE matrix contains the exact speciation and extinction times that had happened along the evolution of the initial homogeneous population. The SSEE matrices are used to draw the exact branching process of the simulated species, giving the true phylogeny of each simulation. This algorithm is advantageous because it eliminates bias regarding estimation of branching times, as usually happens when we deal with phenotypic and genetic traits used to estimate phylogenies in natural studies, in which inferential methods are needed (Costa et al., 2018a). The phylogenies derived from simulations were one of the structures utilized to test if our neutral model of speciation account for the species diversity pattern in the recent platyrrhine divergence.

4.2.3 Quantification of diversification properties

We measured two types of diversification properties: macroevolutionary properties, specifically the tree balance and the branch length distribution, and spatial structure of species distribution.

To quantify the tree balance we utilized the Sackin index ($I(N)$), which measures the degree of imbalance of a tree by calculating the distances between the leaves and the root (Sackin, 1972; Blum and François, 2005):

$$I_s = \sum_j d_j \tag{4.3}$$

where d_j is the number of nodes between each leave to the root. The normalized Sackin index, utilized to compare trees with different number of leaves, is defined as (Blum and François, 2005)

$$I_s^n = \frac{I_s(N_S) - E(I_s(N_S))}{N_S}, \tag{4.4}$$

in which N_S is the number of leaves in a tree, and $E(I_s(N_S))$ is the expected Sackin index under a pure birth process (Yule model) (Yule, 1925). We used the normalized Sackin index to assess tree balance of simulated trees.

To evaluate the speed of diversification we utilized the alpha value (α), which is derived from the γ-statistic (Pybus and Harvey, 2000). The γ-statistic measures the distribution of branch lengths along a tree, with a dependence on the number of leaves (McPeek, 2008; Phillimore and Price, 2008). The α-value allows that for any tree with a given γ and N we have an unique value (α) computing the acceleration of speciation along the tree (Costa et al., 2018b). Constant speciation rates per branch corresponds to $\alpha = 1$, and constant rates of speciation events at any time corresponds to $\alpha = 0$. Negative values of α represents a desacceleration of diversification (stemmy trees), and positive values greater than 1 symbolizes accelerated rates of diversification along the tree (tippy trees) (Costa et al., 2018b).

We computed the Sackin index and the α-value for the true phylogenies recorded from all simulations. We also explored the role of the method from which phylogenies were generated in the resulting values of Sackin index and the α-value. We compared true phylogenies (generated from SSEE algorithm) with phylogenies estimated from distance-based methods, using both nuclear (nucDNA) and mitochondrial DNA (mtDNA) information from all species in each simulation to construct these distance-based phylogenies. The goal was to observe if phylogenies generated from nucDNA and mtDNA get structural properties similar to true simulated SSEE phylogenies in each simulated scenario.

The spatial patterns of species distribution were determined by computing the number of species in regions along South America defined as Operative Geographic Units (OGUs) (Rosen, 1988). We used the OGUs defined in (Goldani et al., 2006), which utilized the interfluvial regions delimited by the main rivers and the Panama Canal to stablish 21 OGUs (Fig C.5). We do not have the regions I and II of Goldadani et al. definition because these regions are not in our lattice, and we prefered keep the notation utilized by these authors, therefore, we had the OGUs $III - XXI$ in our simulations. We calculated the species richness in each OGU of all simulations and considered the species richness per OGUs a measure of spatial structure.

4.2.4 Comparison with empirical platyrrhine data

The phylogenetic and spatial properties computed respectively by Sackin and α indexes and by species richness per OGUs were compared to empirical data of platyrrhines. We collected phylogenetic trees from each platyrrhine family/subfamily, comparing the structural tree properties of these trees to simulated trees in each scenario and monkey size category (Table C.3). We utilized the `sackin` function of the `apTreeshape` package in R (Bortolussi et al., 2012) to compute the Sackin index of phylogenies, and the function `gammaStat` of the `ape` package in R (Paradis et al., 2004) to calculate the γ-statistic from which we compute the α-value. The empirical species distribution maps were reconstructed using geographic distributional data available from (IUCN, 2015) (Fig. C.3).

4.3 Results

We ran 50 simulations for each scenario (without rivers, with equal river widths and with proportional river widths) described in subsection 4.2.1, for small, medium and large monkeys, totalling 450 simulations. The parameter values used in the model were $N = 5000$, $L_1 = 276$, $L_2 = 211$, $q_{min} = 0.95$, $MT = 15000$, $\mu = 0.5x10^{-4}$, $D = 0.05$, $Q = 0.37$. We varied the parameters S and T according to the size category of each family/subfamily of platyrrhine, with S values of 5, 10, 18, and T values of 680, 320 and 300 for small, medium and large monkeys, respectively (Tables 4.1, C.2). Fig. 4.1 shows one simulation for each scenario modeled for small ($S = 5$, $T = 680$), medium ($S = 10$, $T = 320$), and large monkeys ($S = 18$, $T = 300$), with squares representing individuals and colors representing the species generated.

The speed of diversification measured by the α-value differed significantly among the scenarios simulated for each monkey size (Fig. C.6(a), Fig. 4.2, Table C.4). For all monkeys, differences in α-value were significant between scenarios with proportional rivers and equal rivers ($p = 0.89x10^{-3}$ – small, $p = 5.8x10^{-9}$ – medium, $p = 5.6x10^{-11}$ – large, Kruskal-Wallis test) and between scenarios with proportional rivers and without rivers ($p = 2.4x10^{-14}$ – small, $p = 2.6x10^{-11}$ – medium, $1.1x10^{-11}$ – large). Simulations with equal rivers and without rivers differed in α-value for small mokeys ($p = 0.22x10^{-3}$), but did not differ for medium and large monkeys (respectively, $p = 0.71$, $p = 0.97$). Comparisons with empirical data showed that only proportional rivers were not significantly different from empirical data in medium and large monkeys (respectively, $p = 0.2$, $p = 0.57$, Table C.5), which means that this scenario produced a pattern in the speed of diversification in agreement with empirical estimations. For small monkeys, the α-value was significantly different among empirical data and all scenarios (Table C.5).

The tree balance measured by the Sackin index also differed significantly among scenarios for all monkey sizes (Fig. C.6(b), Fig. 4.2, Table C.4). The Sackin index was different for small and large monkeys between scenarios with proportional and equal rivers ($p = 1.1x10^{-13}$ – small, $p = 8.6x10^{-12}$ – large) and between scenarios

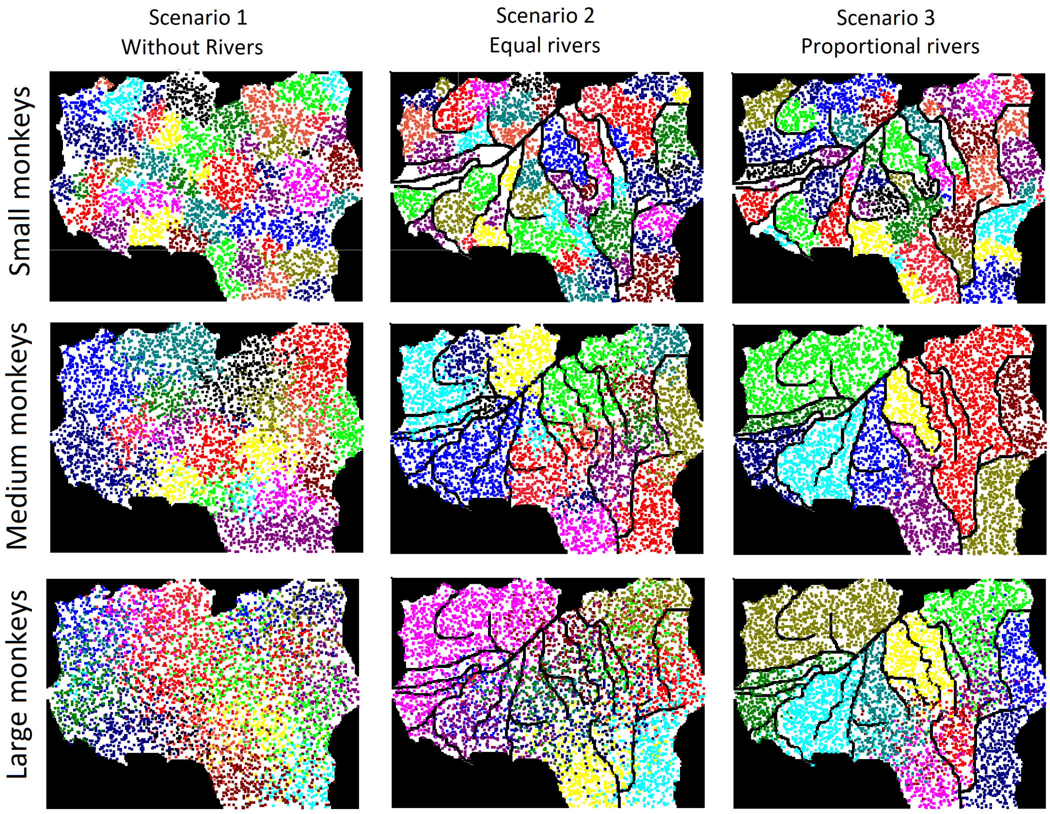


FIGURE 4.1: Spatial distribution of individuals from nine simulations based on the model described in subsection 4.2.2. Individuals are represented by squares, and each color represents a different species. On the left column, the scenario without rivers, only with the South America delimitation. In the middle column, the scenario with main rivers, all with equal width. On the right column, the scenario with the same main rivers, but with river widths proportional to real values (proportions are not visible to the eyes). All scenarios for small, medium, and large monkeys, meaning, respectively, radius of mating range $S = 5$, $S = 10$, $S = 18$ and number of generations $T = 680$, $T = 320$, $T = 300$.

with proportional rivers and without rivers ($p = 2.2x10^{-16}$ – small, $p = 5.2x10^{-9}$ – large). The difference between scenarios with equal rivers and without rivers was significant only for medium monkeys ($p = 0.27x10^{-2}$). Comparisons with empirical data showed that Sackin index did not differ significantly from any scenario in all monkey sizes, with exception only of the proportional rivers for small monkeys ($p = 0.099x10^{-2}$) (Table C.5), meaning that for almost all scenarios and monkey sizes the tree balance pattern are in agreement with that observed in empirical estimations. Both the results showed for α-value and Sackin index reveals that in general the scenario with proportional rivers, which is the more realistic scenario simulated here, generated patterns closer to empirical data than less realistic scenarios (without rivers and equal rivers scenarios).

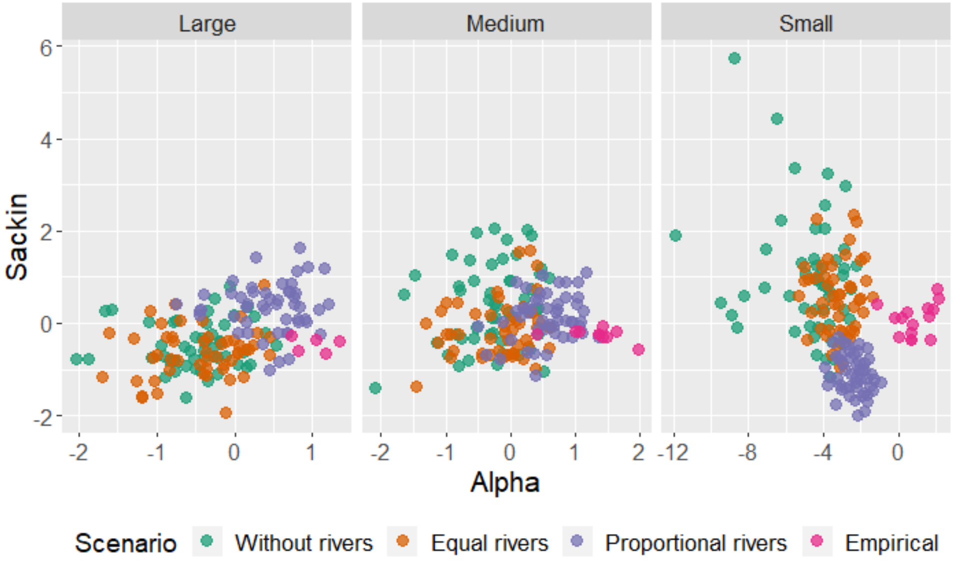


FIGURE 4.2: Scatterplots for both structural properties of true simulated trees (SSEE) measured by α-value and Sackin index of all simulated scenarios and empirical data for large, medium and small monkeys. The α-value calculated from trees generated in the scenario with proportional rivers are closer to empirical data (not statistically different) than α-values from other scenarios for medium and large monkeys. The Sackin index did not differ significantly from any scenario in all monkey sizes, with exception only of the proportional rivers for small monkeys. These results show that in general the scenario with proportional rivers, the more realistic simulated scenario, gave patterns closer to empirical data than less realistic scenarios.

Comparisons among true phylogenies and phylogenies estimated from nucDNA and mtDNA by distance-based methods (UPGMA, ME and NJ) show that these methods yield trees with different Sackin and α values (Fig. 4.3). For all distance-based methods employed and all monkey sizes the statistics were farther to empirical values than true phylogenies. For all monkey sizes, all methods employed in nucDNA and mtDNA yield statistically different α values ($p < 0.05$) in comparison with SSEE phylogenies. In contrast, the Sackin index of nucDNA and mtDNA was not significantly different from SSEE trees in the majority of cases, with mtDNA trees estimated by UPGMA and ME methods and nucDNA trees estimated by NJ being the only exceptions

($p < 0.05$). The general outcome shows that true SSEE phylogenies are closer to empirical patterns of balance and speed of diversification than nucDNA and mtDNA trees estimated from distance-based methods (Fig. 4.3). Separate comparisons between true phylogenies and nucDNA or mtDNA-based phylogenies are displayed in Figs. C.7, C.8.

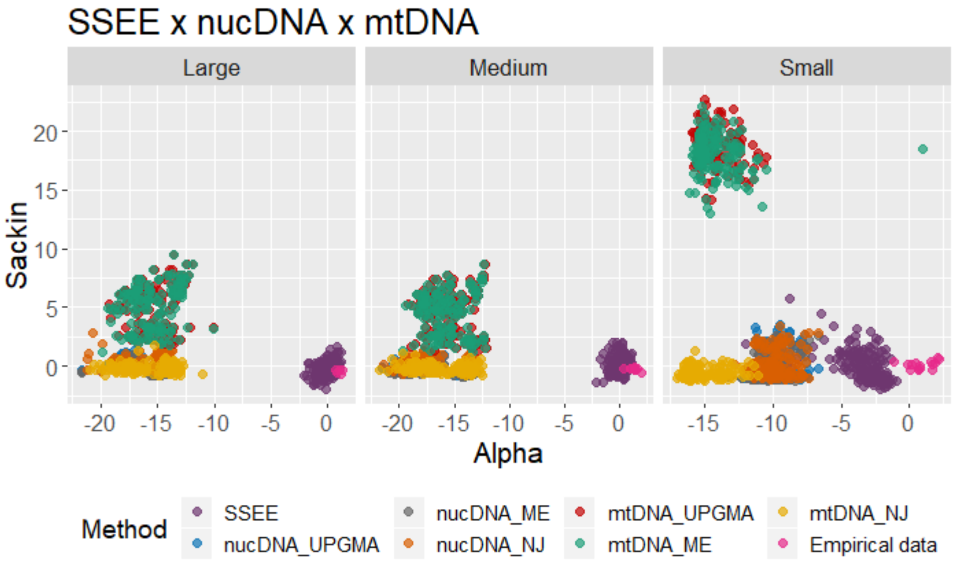


FIGURE 4.3: Scatterplots for both structural properties (α-value and Sackin index) of true simulated trees (SSEE) and trees estimated from nucDNA and mtDNA of all simulated scenarios and empirical data for large, medium and small monkeys. Each color represents a method (UPGMA, ME or NJ) to generate phylogenies for nucDNA and mtDNA, without separating the scenarios in each case.

The spatial distribution of species richness per region and per monkey size is displayed in Fig. 4.4. In general, species richness in each OGUs was smaller in the scenario with proportional river widths, meaning that overall species richness was smaller in this scenario. Moreover, the species richness per OGUs in the proportional rivers scenario was closer to empirical values for most regions in all monkey sizes (Fig. 4.4). The median of species richness considering all scenarios and all regions together were 4 (Q1=3, Q3=6), 3 (Q1=2, Q3=5), and 6 (Q1=4, Q3=9), for small, medium and large monkeys, respectively. The median of species richness for each scenario in each monkey size is presented in Table C.6.

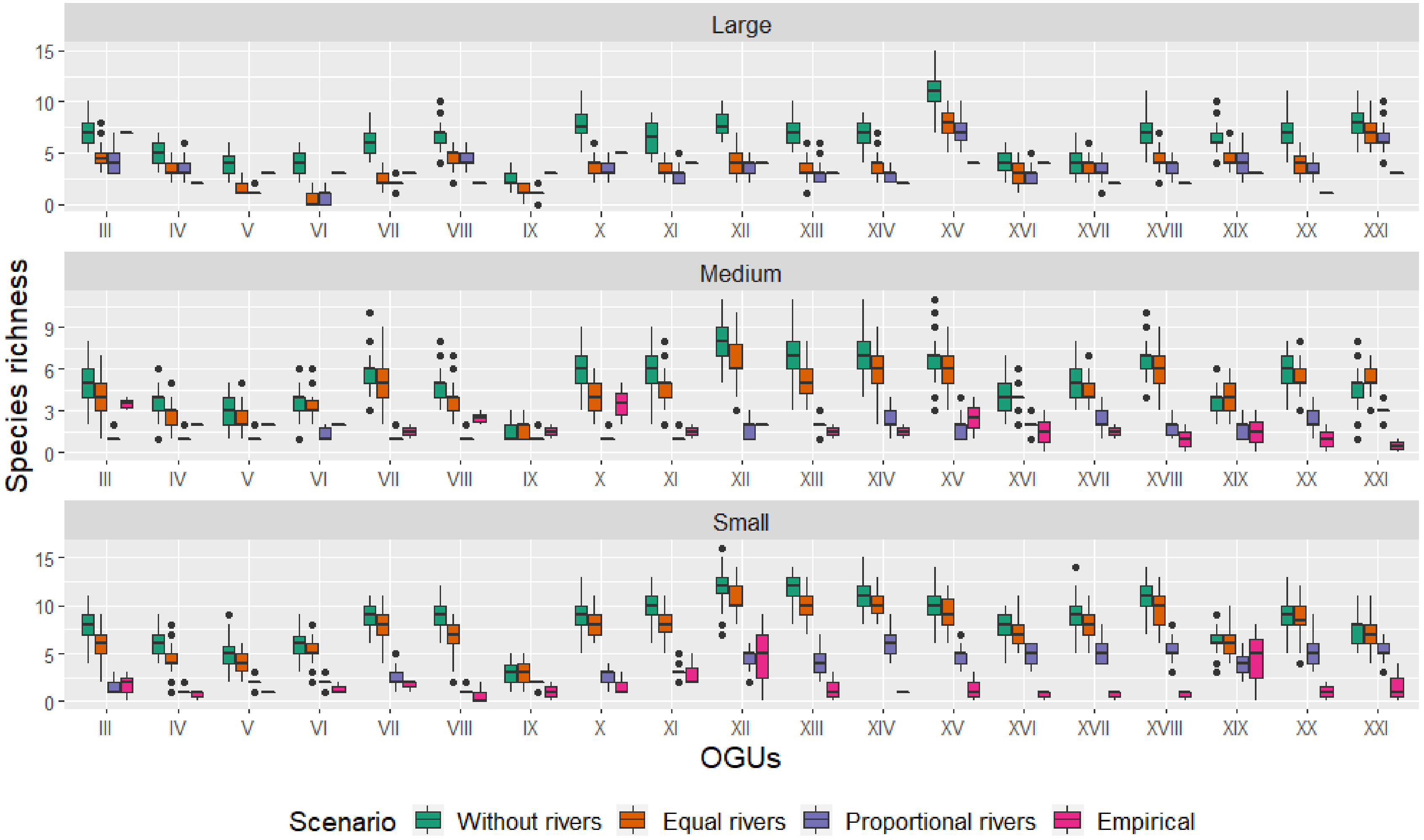


FIGURE 4.4: Boxplots of species richness in each OGUs (III-XXI) for 450 simulations of the three scenarios (without rivers, equal rivers and proportional rivers) and empirical data for large, medium and small monkeys. Each color represents a scenario and the empirical data.

4.4 Discussion

As we can observe in Figs. 4.2, 4.4, the addition of geographical barriers changed macroevolutionary and distributional patterns of species in the model presented here. In the scenario designed for large monkeys (large S, $T = 300$), we observed that heterogeneous spatial conditions, with rivers serving as barriers with different sizes, increases the speed of diversification (α-value) and slightly increases the balance of the trees (Sackin index). For medium monkeys (medium S, $T = 320$), the spatial heterogeneity caused an increase in the speed of diversification, but did not affect the balance of trees. Lastly, for small monkeys (small S, $T = 680$), the spatial proportional barriers led to an increase of the speed of diversification and to a decrease in the balance of trees, the later as opposed to the observed for large monkeys. Regarding the distributional pattern of species richness, we observed for all sizes of monkeys that in the scenario with spatial heterogeneity the number of species per region (OGU) was always smaller than for the scenario with spatial homogeneity. These results show that adding barriers affected significantly the patterns generated in this spatially explicit model in a novel way regarding speciation modelling.

The patterns obtained in this paper show firstly that, as pointed by Kopp (2010), the mode of speciation can leave signatures in the structure of phylogenetic trees. In our case, the scenario with spatial homogeneity, in which only isolation-by-distance operates together with neutral processes like mutation, recombination and genetic drift, leaves a signature of negative α and Sackin values for large and medium monkeys, and a signature of even more negative α values and positive Sackin values for small monkeys (Fig. 4.2). In contrast, the scenario with spatial heterogeneity (with barriers of varied sizes) has the action of the same neutral processes added to isolation-by-distance, associated now to an additional level of complexity which is the spatial structure, that leaves other signatures in the structures of phylogenetic trees, with positive α and Sackin values for large and medium monkeys and negative α and Sackin values for small monkeys (Fig. 4.2). The creation of evolutionary models which explicitly incorporates population dynamics and generate phylogenetic trees are scarce (Hubbell, 2001; McPeek, 2008; Pigot et al., 2010; Gascuel et al., 2015b) and important because they give a more realistic context to the occurrence of speciation (Manceau et al., 2015).

Figures 4.2,4.4 show a greater correspondence between the scenario of spatial heterogeneity (with proportional river widths) and the macroevolutionary and distributional patterns estimated for empirical data. In Fig. 4.2, this can be observed by the closeness (corroborated by statistical analysis) of purple points (scenario with proportional rivers) with pink points (empirical data) for large and medium monkeys. For small monkeys, there was no proximity between any scenario with empirical data 4.2. In Fig. 4.4, the greater correspondence of distributional patterns of the scenario with proportional rivers and the empirical data is observed in the proximity of purple and pink boxplots for most of the regions (OGUs) in large, medium and small monkeys. The fact that the more realistic scenario, which incorporates rivers in a more representative way, matches with empirical data provides evidence that rivers are possibly critical factors in the speciation of the Platyrrhini group, mainly for medium and large monkeys. Therefore, if only neutral processes associated with river barriers and isolation-by-distance recovered some macroevolutionary and distributional patterns observed in subfamilies and genera of Platyrrhinithe monkeys, we can conclude that this model gives plausibility to the Riverine Hypothesis, in which rivers are the main factor promoting speciation.

One of the first attempts to test the Riverine Hypothesis in a non verbal way was the study of Ayres and Clutton-Brock (1992). The authors have investigated the role of rivers in delineating the range of Amazonian primate species, and inquired if the size of rivers is an important factor to delimitate the range of species, and if the size of primates is important to constrain its range in one side of the river. For the 14 rivers analyzed they observed that the similarity index, which measures the similarity of species identity between opposite river banks, ranged from 30 to 100 (where 100 means identical species and 0 means completely different species), and was negatively correlated with river width, corroborating their hypothesis that similarity of primate faunas on opposite banks of rivers declines with the size of the river. They also observed that geographical range size was positively correlated with body weight across species, and provided a direct evidence that body size affects the ability of primate species to disperse across river barriers showing that larger primates have fewer subspecies relative to their range size than smaller ones, presumably because they are better able to cross environmental barriers.

If we look at Fig. 4.1 we can observe the rivers are better delimiters to species range in the scenario with proportional river widths than in the scenario with equal river widths, and we see a pattern of range distribution not consistent with rivers tributaries in the scenario without rivers, supporting the Ayres and Clutton-Brock (1992) results about the importance of rivers in delimitating species ranges. These authors also observed associations between river size and the similarity of primate communities on opposite banks occurred within rivers. For example, the similarity of primate communities on the opposite banks of the Amazon river decreased with its distance from its headwaters, but at its mouth, where the decrease in water speed and associated sedimentation produces numerous islands, the similarity of primate faunas showed an increase (Ayres and Clutton-Brock, 1992). We did not add this level of complexity in our model, of varying the width along a river, but with these results observed by Ayres and Clutton-Brock (1992) we can hipothesize that the addition of this level of variation could yield different patterns of species distribution and macroevolution.

Previous studies have looked for specific primates taxa searching for mechanisms that explain the diversity patterns observed for the monkeys. Boubli et al. (2015) have tested the Riverine Hypothesis for diurnal primates of Rio Negro and Rio Branco, rivers on the Amazon region, by the comparison of spatial and temporal patterns of diversification for these primates. They observed that divergence times for *Cacajao* (Pitheciidae), *Callicebus* and *Cebus* coincided with the divergence times of other nine vertebrate genera living in this region (primates and birds), which offers a strong support to the hypothesis that rivers play a critical role for the diversification of these primates. In contrast, they concluded that rivers are physical barriers but not vicariant agents (*i.e.*, do not promote speciation, only dispersal limitation) for *Ateles*, *Alouatta* (both Atelidae), and *Saguinus* monkeys. In comparison with our results, there was a mismatch between all scenarios and empirical data of small monkeys (*e.g.*, *Saguinus*, Fig. 4.2), possibly reflecting the less importance of rivers for smaller species. The match of simulated large monkeys with empirical data show a different result of Boubli et al. (2015) otherwise, with members of Atelidae family holding a major importance of rivers in the diversification process considering our results.

Empirical studies of various Platyrrhini clades, based on phylogenetic estimation from mtDNA and nucDNA, have shown different supports to the Riverine Hypothesis (Collins and Dubach, 2000; Van Roosmalen et al., 2002; Cortés-Ortiz et al., 2003; Perez-Sweeney et al., 2008; Chiou et al., 2011; Byrne et al., 2016; Lima et al., 2017). Studies of spider monkeys (*Ateles*) have concluded that riverine barriers did not interrupt gene flow significantly among these monkeys, with large-scale geographic changes associated with the rise of the Andes cordilleras being most important causes of speciation in this group (Collins and Dubach, 2000). Conversely, the howler monkey (*Alouatta*) diversification was initiated by Andean vicariance, but the posterior diversification of cis-Andes howlers was probably caused by the formation of modern Amazon river, suggesting the Riverine Hypothesis migth be the plausible in this case (Cortés-Ortiz et al., 2003). For large monkeys, therefore, there is both supporting and rejection of the Riverine Hypothesis in previous literature, with our results giving support to the hypothesis (Fig. 4.2).

In studies of medium monkeys, authors have observed that titi monkeys (*Callicebus*) exhibited a distributional pattern that coincide with rivers which not changed their courses over considerable geological times, suggesting that rivers acted as isolating barriers and may have promoted speciation (Van Roosmalen et al., 2002; Byrne et al., 2016). Lima et al. (2017) have found that the most plausible mechanism explaining the diversification of capuchin (*Cebus* and *Sapajus*) monkeys is the isolation of ancestral populations in the Atlantic Forest versus Amazonian habitats, with subsequent expansion of *Sapajus* to the Amazonian *Cebus* habitat. This hypothesis is an alternative to the Riverine Hypothesis, therefore, rivers were considered less important in the speciation of these monkeys. For medium monkeys we also observed constrasting conclusions about the importance of rivers in the speciation process, with our results giving support to the Riverine Hypothesis (Fig. 4.2).

For small monkeys, some researches suggest that lion tamarin (*Leontopithecus*) diversified as a consequence of allopatric isolation due to Quaternary shifts in climate and habitat distribution, but other hypothesis is that they differentiated through isolation-by-distance (Perez-Sweeney et al., 2008). The small size of these monkeys make the isolation-by-distance a very likely process to drive speciation, with rivers acting as barriers only when the distribution of these species are close to river systems. Accordingly, for small monkeys, the Riverine Hypothesis seems not to be the most likely cause of speciation, which corroborates our findings, with simulations implementing river barriers not exhibiting a good match with empirical data (Fig. 4.2). In general, for all size categories of monkeys there are studies which strengthen or weaken the Riverine Hypothesis, showing the importance of climatic and habitat changes, and the uplifiting of Andean cordilleras as alternative factors promoting speciation for Platyrrhini monkeys.

A recent study have found that the Riverine Hypothesis is innapropriate for most of species inhabiting areas around the Madeira river, an Amazonian river (Santorelli et al., 2018), totalling the analysis of 1952 species in 14 taxonomic groups (*e.g.*, Hymenoptera, Coleoptera, Lepidoptera, Snakes, Lizards, Anura, Chiroptera, Aves, Primates, Small and Large mammals). The authors estimated the proportion of species in different taxonomic groups that hava their distributions limited by a river and the number of species for which there is evidence that river promoted speciation by vicariance (the existence of endemic areas based on large rivers). The results showed that only 0.10% of the 1952 species (primates and aves) had their distributions limited by the Madeira river. They have argued that rivers can function as species delimiters without representing barriers responsible for speciation by vicariance. Alternatively, mechanisms like sympatric speciation via sexual selection, environmental heterogeneity, ecological interactions, combined with dispersal limitation and competition could produce the same patterns of distribution as the hipothesized by the Riverine Hypothesis. Also, although the rivers possibly played an important role in the speciation processes of Platyrrhini, the changes in the environment during the formation of rivers created diversity of habitats and niches, which can be responsible for the speciation of these groups likewise. The interaction of non neutral forces with the neutral ones, associated with geographic barriers performed by rivers, were likely important factors to promote speciation in this riverine region, with a complex interaction played by each component (Schneider and Sampaio, 2015). The incorporation of non neutral processes in the model could help to figure out the relevance of each factor in the speciation process, and the creation of a dynamic environment, with rivers being formed at different time steps, also could yield interestingly patterns which would provide an even more realistic scenario for the speciation of the New World monkeys.

In the model presented here we simulated a scenario without rivers, in which only isolation-by-distance could play a role in the speciation process, and we simulated a scenario with river barriers, with the same isolation-by-distance acting but now associated with spatial barriers. The results showed in Fig 4.2 reveal that the addition of river barriers caused a change in the macroevolutionary pattern of communities, indicating that despite we cannot isolate the role of dispersal limitation and river barriers in the latter scenario we can conclude the spatial barriers play a role apart from the isolation-by-distance mechanism. The complex interaction between different processes, with rivers, Andean cordilleras, climate changes, habitat heterogeneity, and sexual selection playing each a significant role in the speciation process of Platyrrhini monkeys is the most probable explanation for the striking diversity of the New World monkeys. The results presented here just reinforce the relative importance that rivers employed as isolating barriers to promote speciation by vicariance, maintaining the Riverine Hypothesis a plausible explanation.

4.4.1 Future perspectives

The model presented in this paper address the question if spatial barriers, represented by South American rivers could play a critical role in the speciation process of species living in this area. We observed that river barriers changed macroevolutionary and distributional patterns of species in comparison to barrier-free spaces, with a great agreement with empirical data of Platyrrhini subfamilies and genera. To ensure that neutral processes associated with river barriers are the only factors to explain the diversity patterns of Platyrrhini species it is necessary to incorporate non neutral components to the model, in order to verify if the patterns obtained here would change or if they remain the same. The addition of non neutral forces are needed to solve if the metrics chosen here are really informative regarding the mechanisms responsible for the speciation process. In other words, if these metrics leave signatures about the mechanisms driving speciation, we expect that the addition of non neutral forces would yield different patterns of diversification, which could only be assured with new simulations. The incorporation of sex separation allows the modeling of sexual selection, an example of non neutral process. Also, the addition of a fitness associated with particular genotypes, increasing its probability to reproduce, could represent the implementation of natural selection. With respect to the spatial barriers, the variation of river widths along different parts of the river (head, mouth) could also improve the reality of the geographic context experienced by species living in these areas. These are some possible general ways to advance in theoretical research about the importance of spatial barriers in the diversification processes, and, in particular, to clarify the relevance of the Riverine Hypothesis for a diversity of taxa.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq) - grant #302049/2015-0, São Paulo Research Foundation (FAPESP) grants #2016/06054-3 and 2016/01343-7 (ICTP-SAIFR FAPESP).

Chapter 5

Discussion

The outcomes of the different modifications made here in the De Aguiar et al. (2009) model have brought interesting novelties in the speciation theory research. The algorithms introduced in the first chapter (MRCAT and SSEE) allows the evolutionary history of population to be registered and analyzed in IBM's, a novel procedure that has not been explored before. The record of the most recent common ancestor of each pair of individuals of the simulation, allowing the building of genealogies and phylogenies, was an innovation in comparison with trees derived from previous models which did not model individuals explicitly with a mechanism of reproductive isolation (Manceau et al., 2015; Rosindell et al., 2015). The recording of speciation and extinction events, regarding the dynamics at the individual level, is also new and allow the construction of phylogenetic trees of a population evolving gradually, different from point-mutation speciation models, in which an individual changes from one species to another instantaneously (Hubbell, 2001).

The comparison between these trees and those generated by well-know inference methods based on genetic distances (UPGMA, ME, NJ) showed that the trees have different structural properties, like tree shape and balance and branch length distribution. This indicates that the method of tree building is important in IBMs, with different methods leaving distinct patterns, and with the algorithm SSEE bringing the exact representation of the simulated branching process. Also, previous models which incorporates genealogies or phylogenies did not include space explicitly (Derrida and Peliti, 1991; Serva and Peliti, 1991), not allowing an investigation in a biogeography framework, possible with the tools provided in the Chapter One. Possible directions for future research are to include *loci* not linked to the computation of genetic threshold, to investigate if phylogenetic trees computed from *loci* linked to the threshold would differ from not linked *loci*; choose different substitution models to compute the pairwise distance between sequences to look if they affect the patterns recovered from genetic distance; and increase genome size to be possible estimate trees from character-based methods, or bayesian methods, which requires larger genomes than the modeled here. These variations could improve the inference power of the model and makes it more realistic. The results achieved in the first chapter were more technical but crucial to develop the investigation of signatures of speciation processes in the next chapters.

The importance of neutral processes in shaping biodiversity patterns was supported by the results obtained in this thesis. The second chapter has associated the geographic context of speciation with the macroevolutionary patterns displayed by phylogenetic trees in simulations, without the need of implementing non neutral processes to relate geography and tree pattern for a variety of empirical evolutionary radiations used in comparison with simulation results. The modeling employed in this chapter contribute to fill the gap between microevolutionary processes and macroevolutionary patterns present in the literature (Manceau et al., 2015; Gascuel et al., 2015b; Cabral et al., 2017), mainly relating different geographic contexts (sympatric and parapatric) to specific signatures imprinted in phylogenetic trees, as investigated previously (Pigot et al., 2010; Moen and Morlon, 2014; Lewitus and Morlon, 2016). The comparison between macroevolutionary patterns of simulated neutral radiations and the patterns of empirical neutral and non neutral radiations also revealed the importance of geography in shaping the structural tree patterns in a range of systems. The great match between expected degrees of gene flow during speciation of empirical radiations with the simulated ones support the idea of geography playing a critical role in the speciation processes of empirical groups.

The results obtained in Chapter Two show that, despite the observed patterns cannot be exclusively attributed to neutral models, complex patterns can emerge from neutral assumptions, revealing that the simple observation of macroevolutionary patterns generally associated with adaptive radiations could not be enough to infer adaptive processes (McPeek, 2008; Rundell and Price, 2009; Simoes et al., 2016). Li et al. (2018) have showed with a protracted speciation modeling framework that distinct microevolutionary scenarios can generate similar patterns of biodiversity, with macroevolutionary models which do not take individual level processes into account not being able to distinguish between the scenarios. The authors also pointed out that is necessary to be cautious in inferring causal relationships between ecological factors and macroevolutionary patterns, which is also needed when we deal with neutral processes as modeled in Chapter Two. Future studies in this field should include non neutral processes in the modeling approach and the evaluation of other phylogenetic tree statistics, like phylogenetic diversity (Faith's PD, Faith, 1992) and the biogeographically weighted evolutionary distinctiveness (BED, Tucker and Cadotte, 2013) that might signal the different roles played by neutral and non neutral processes in evolutionary radiations.

The Chapter Three also expressed the importance of neutral processes in promoting speciation, but now in a particular context with spatial barriers simulating rivers of South America. The addition of spatial barriers in a more realistic way was a novelty of this work in comparison with previous studies of spatially explicit models (Gavrilets et al., 1998; Gavrilets et al., 2000a; Gavrilets et al., 2000b; Hoelzer et al., 2008; De Aguiar et al., 2009; Martins et al., 2013; Desjardins-Proulx and Gravel, 2011; Melián et al., 2012). The parameters choice to conform biological features of the New World monkeys (Platyrrhini) promote three distinct scenarios for three different size of monkeys, revealing that for most monkeys the more realistic scenario, with river widths proportional to real values, fits better the simulated macroevolutionary and distributional patterns with empirical data. The Riverine Hypothesis, therefore, a long time debated, was supported by the results obtained in the third chapter. In the other hand, small monkeys presented simulated patterns in which rivers appear not to be so important, possibly with isolation by distance being a more critical factor, a finding compatible with previous literature (Perez-Sweeney et al., 2008).

The incorporation of non neutral processes in this model is also a necessary step to improve the understanding of the importance of river barriers in driving speciation of Platyrrhini monkeys. Also, an even more realistic spatial context is possible, with rivers presenting different widths at different points (head, middle course, mouth), an important feature analyzed by Ayres and Clutton-Brock (1992). As in the Chapter Two, the exploration of other tree statistics could improve the understanding and separatation of the roles of geographic barriers and non neutral processes in the evolution of the New World monkeys, and of other species that have evolved in this same spatial context (Gascon et al., 1998; Gascon et al., 2000). The implementation of sex separation could also allow to verify the importance of processes as sexual selection in the speciation of these monkeys, a factor previously suggested by Santorelli et al. (2018). Regarding the methods of tree construction, the same suggestions made for the first chapter are valuable here, like the use of distinct substitution models for calculate genetic distances and the use of character-based and bayesian methods for estimate trees from genome data.

A critical perspective not approached here is the relevance of extinctions in the diversification dynamics. The fossil record show that many clades are in decline –

e.g., Cetacea, perissodactyl mammals, lungfish, brachiopods, gymnosperms, sphenophytes – despite the fact that current methods that infer diversification dynamics from molecular phylogenies ignore this phenomenon (Foote et al., 2007; Quental and Marshall, 2011). Quental and Marshall (2011) have showed that clades in decline leave signatures in macroevolutionary patterns, specifically in the branch lenght distribution measured by the γ statistic. They observed that the molecular signature of clades in decline could be the same as that produced by diversity dependent diversification, which bring dificulties to unveil the main causes of the patterns recovered in phylogenies. The modeling used here could contribute to this field, with individual-based simulations introducing a new approach to analize diversity trajectories. The SSEE algorithm developed in the first chapter records the speciation and the extinction events, which allows the investigation of the dynamics of extinction along the evolutionary history of populations. The computation of tree statistics like γ, α and Sackin would yield different patterns from that observed in trees with only extant species. The investigation of the role of extinction in evolutionary radiations is also paramount, and a possible extension of the second chapter. Extinctions could also play an important role in the diversification of the New World monkeys, and could be taken into account in future investigations. Therefore, the investigation of extinction dynamics is a fundamental further step to the topics approached in this thesis.

Bibliography

Alföldi, J., F. Di Palma, M. Grabherr, C. Williams, L. Kong, E. Mauceli, P. Russell, C. B. Lowe, R. Glor, J. D. Jaffe, et al. (2011). “The genome of the green anole lizard and a comparative analysis with birds and mammals”. In: *Nature* 477.7366, p. 587.

Álvarez, A., R. L. M. Arévalo, and D. H. Verzi (2017). “Diversification patterns and size evolution in caviomorph rodents”. In: *Biological Journal of the Linnean Society* 121.4, pp. 907–922.

Amitrano, C., L. Peliti, and M Saber (1989). “Population dynamics in a spin-glass model of chemical evolution”. In: *Journal of Molecular Evolution* 29.6, pp. 513–525.

Aristide, L., A. L. Rosenberger, M. F. Tejedor, and S. I. Perez (2015). “Modeling lineage and phenotypic diversification in the New World monkey (Platyrrhini, Primates) radiation”. In: *Molecular Phylogenetics and Evolution* 82, pp. 375–385.

Arnold, C., L. J. Matthews, and C. L. Nunn (2010). “The 10kTrees website: a new online resource for primate phylogeny”. In: *Evolutionary Anthropology* 19.3, pp. 114–118.

Ayres, J. and T. Clutton-Brock (1992). “River boundaries and species range size in Amazonian primates”. In: *The American Naturalist* 140.3, pp. 531–537.

Ballard, J. W. O. and M. C. Whitlock (2004). “The incomplete natural history of mitochondria”. In: *Molecular Ecology* 13.4, pp. 729–744.

Banguera-Hinestroza, E., A. Hayano, E. Crespo, and A. R. Hoelzel (2014). “Delphinid systematics and biogeography with a focus on the current genus Lagenorhynchus: Multiple pathways for antitropical and trans-oceanic radiation”. In: *Molecular Phylogenetics and Evolution* 80, pp. 217–230.

Bank, C., R. Bürger, and J. Hermisson (2012). “The limits to parapatric speciation: Dobzhansky–Muller incompatibilities in a continent–island model”. In: *Genetics* 191.3, pp. 845–863.

Baptestini, E. M., M. A. de Aguiar, and Y. Bar-Yam (2013a). “Conditions for neutral speciation via isolation by distance”. In: *Journal of Theoretical Biology* 335, pp. 51–56.

– (2013b). “The role of sex separation in neutral speciation”. In: *Theoretical Ecology* 6.2, pp. 213–223.

Barley, A. J., J. White, A. C. Diesmos, and R. M. Brown (2013). “The challenge of species delimitation at the extremes: diversification without morphological change in Philippine sun skinks”. In: *Evolution* 67.12, pp. 3556–3572.

Barraclough, T. G. and A. P. Vogler (2000). "Detecting the geographical pattern of speciation from species-level phylogenies". In: *The American Naturalist* 155.4, pp. 419–434.

Bateson, W (1909). "Heredity and variation in modern lights, pp. 85–101 in Darwin and Modern Science, edited by Seward AC". In: *Cambridge University Press, Cambridge, UK*.

Beaudrot, L. H. and A. J. Marshall (2011). "Primate communities are structured more by dispersal limitation than by niches". In: *Journal of Animal Ecology* 80.2, pp. 332–341.

Blonder, B., B. G. Baldwin, B. J. Enquist, and R. H. Robichaux (2016). "Variation and macroevolution in leaf functional traits in the Hawaiian silversword alliance (Asteraceae)". In: *Journal of Ecology* 104.1, pp. 219–228.

Blum, M. G. and O. François (2005). "On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited". In: *Mathematical Biosciences* 195.2, pp. 141–153.

Bolnick, D. I. and B. M. Fitzpatrick (2007). "Sympatric speciation: models and empirical evidence". In: *Annual Review of Ecology, Evolution, and Systematics* 38, pp. 459–487.

Bortolussi, N., E. Durand, M. Blum, and O. François (2005). "apTreeshape: statistical analysis of phylogenetic tree shape". In: *Bioinformatics* 22.3, pp. 363–364.

Boubli, J. P., C. Ribas, J. W. L. Alfaro, M. E. Alfaro, M. N. F. da Silva, G. M. Pinho, and I. P. Farias (2015). "Spatial and temporal patterns of diversification on the Amazon: a test of the riverine hypothesis for all diurnal primates of Rio Negro and Rio Branco in Brazil". In: *Molecular Phylogenetics and Evolution* 82, pp. 400–412.

Brett, D., H. Pospisil, J. Valcárcel, J. Reich, and P. Bork (2002). "Alternative splicing and genome complexity". In: *Nature Genetics* 30.1, pp. 29–30.

Britton, T., B. Oxelman, A. Vinnersten, and K. Bremer (2002). "Phylogenetic dating with confidence intervals using mean path lengths". In: *Molecular Phylogenetics and Evolution* 24.1, pp. 58–65.

Bürger, R., K. A. Schneider, M. Willensdorfer, and S Otto (2006). "The conditions for speciation through intraspecific competition". In: *Evolution* 60.11, pp. 2185–2206.

Burin, G., L. R.V. d. Alencar, J. Chang, M. E. Alfaro, and T. B. Quental (2018). "How well can we estimate diversity dynamics for clades in diversity decline?" In: *Systematic Biology*, syy037.

Butlin, R., J. Bridle, and D. Schluter (2009). "Speciation and patterns of diversity". In:

Butlin, R. et al. (2012). "What do we need to know about speciation?" In: *Trends in Ecology & Evolution* 27.1, pp. 27–39.

Byrne, H., A. B. Rylands, J. C. Carneiro, J. W. L. Alfaro, F. Bertuol, M. N. da Silva, M. Messias, C. P. Groves, R. A. Mittermeier, I. Farias, et al. (2016). "Phylogenetic relationships of the New World titi monkeys (*Callicebus*): first appraisal of taxonomy based on molecular evidence". In: *Frontiers in Zoology* 13.1, p. 10.

Cabral, J. S., L. Valente, and F. Hartig (2017). "Mechanistic simulation models in macroecology and biogeography: state-of-art and prospects". In: *Ecography* 40.2, pp. 267–280.

Campos, P. R. A., E. D. C. Neto, V. M. d. Oliveira, and M. A. F. Gomes (2012). "Neutral communities in fragmented landscapes". In: *Oikos* 121.11, pp. 1737–1748.

Chiou, K. L., L. Pozzi, J. W. L. Alfaro, and A. Di Fiore (2011). "Pleistocene diversification of living squirrel monkeys (*Saimiri* spp.) inferred from complete mitochondrial genome sequences". In: *Molecular Phylogenetics and Evolution* 59.3, pp. 736–745.

Clarke, M., G. H. Thomas, and R. P. Freckleton (2017). "Trait evolution in adaptive radiations: modeling and measuring interspecific competition on phylogenies". In: *The American Naturalist* 189.2, pp. 121–137.

Collins, A. and J. Dubach (2000). "Biogeographic and ecological forces responsible for speciation in *Ateles*". In: *International Journal of Primatology* 21.3, pp. 421–444.

Cortés-Ortiz, L, E. Bermingham, C. Rico, E Rodrıguez-Luna, I Sampaio, and M Ruiz-Garcıa (2003). "Molecular systematics and biogeography of the Neotropical monkey genus, *Alouatta*". In: *Molecular Phylogenetics and Evolution* 26.1, pp. 64–81.

Costa, C. L., F. M. Marquitti, S. I. Perez, D. M. Schneider, M. F. Ramos, and M. A. de Aguiar (2018a). "Registering the evolutionary history in individual-based models of speciation". In: *Physica A: Statistical Mechanics and its Applications* 510, pp. 1–14.

Costa, C. L., P. Lemos-Costa, F. Marquitti, L. D. Fernandes, M. F. Ramos, D. M. Schneider, A. B. Martins, and M. A. Aguiar (2018b). "Signatures of microevolutionary processes in phylogenetic patterns". In: *Systematic Biology*.

Cox, C. B., P. D. Moore, and R. Ladle (2016). "Biogeography: an ecological and evolutionary approach". In: *John Wiley & Sons*.

Coyne, J. and H. Orr (2004). "Speciation". In: *Sinauer Associates, Sunderland, MA*.

Cropp, S. J., A. Larson, and J. M. Cheverud (1999). "Historical biogeography of tamarins, genus *Saguinus*: the molecular phylogenetic evidence". In: *American Journal of Physical Anthropology* 108.1, pp. 65–89.

Davies, T. J., A. P. Allen, L. Borda-de Água, J. Regetz, and C. J. Melián (2011). "Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification". In: *Evolution* 65.7, pp. 1841–1850.

De Aguiar, M. A. M., M. Baranger, E. Baptestini, L Kaufman, and Y Bar-Yam (2009). "Global patterns of speciation and diversity". In: *Nature* 460.7253, pp. 384–387.

De Aguiar, M. A. (2017). "Speciation in the Derrida–Higgs model with finite genomes and spatial populations". In: *Journal of Physics A: Mathematical and Theoretical* 50.8, p. 085602.

DeAngelis, D. L. and V. Grimm (2014). "Individual-based models in ecology after four decades". In: *F1000prime Reports* 6.

DeAngelis, D. L. and W. M. Mooij (2005). "Individual-based modeling of ecological and evolutionary processes". In: *Annual Review of Ecology, Evolution, and Systematics* 36, pp. 147–168.

Dearlove, B. L. and S. D. Frost (2015). "Measuring asymmetry in time-stamped phylogenies". In: *PLoS Computational Biology* 11.7, e1004312.

Degnan, J. H. and N. A. Rosenberg (2009). "Gene tree discordance, phylogenetic inference and the multispecies coalescent". In: *Trends in ecology & evolution* 24.6, pp. 332–340.

Derrida, B. and L. Peliti (1991). "Evolution in a flat fitness landscape". In: *Bulletin of Mathematical Biology* 53.3, pp. 355–382.

Desjardins-Proulx, P. and D. Gravel (2011). "How likely is speciation in neutral ecology?" In: *The American Naturalist* 179.1, pp. 137–144.

Dieckmann, U and M Doebeli (1999). "On the origin of species by sympatric speciation." In: *Nature* 400.6742, pp. 354–7.

Dobzhansky, T. (1936). "Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids". In: *Genetics* 21.2, p. 113.

– (1937). "Genetics and the Origin of Species". In: *Columbia University Press, New York*.

Doorn, G. S. van, P. Edelaar, and F. J. Weissing (2009). "On the origin of species by natural and sexual selection". In: *Science* 326.5960, pp. 1704–1707.

Emerson, B. C. and N. Kolm (2005). "Species diversity can drive speciation". In: *Nature* 434.7036, p. 1015.

Etienne, R. S. and D. Alonso (2005). "A dispersal-limited sampling theory for species and alleles". In: *Ecology letters* 8.11, pp. 1147–1156.

Fabre, P., A Rodrigues, and E. Douzery (2009). "Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA". In: *Molecular Phylogenetics and Evolution* 53.3, pp. 808–825.

Faith, D. P. (1992). "Conservation evaluation and phylogenetic diversity". In: *Biological conservation* 61.1, pp. 1–10.

Felsenstein, J. (2004). "Inferring phylogenies". In: *Sinauer associates Sunderland* 2.

Field, R., B. A. Hawkins, H. V. Cornell, D. J. Currie, J. A. F. Diniz-Filho, J.-F. Guégan, D. M. Kaufman, J. T. Kerr, G. G. Mittelbach, T. Oberdorff, et al. (2009). "Spatial species-richness gradients across scales: a meta-analysis". In: *Journal of Biogeography* 36.1, pp. 132–147.

Fierst, J. L. and T. F. Hansen (2010). "Genetic architecture and postzygotic reproductive isolation: evolution of Bateson–Dobzhansky–Muller incompatibilities in a polygenic model". In: *Evolution* 64.3, pp. 675–693.

Flaxman, S. M., A. C. Wacholder, J. L. Feder, and P. Nosil (2014). "Theoretical models of the influence of genomic architecture on the dynamics of speciation". In: *Molecular Ecology* 23.16, pp. 4074–4088.

Fontana, W. and P. Schuster (1987). "A computer model of evolutionary optimization". In: *Biophysical Chemistry* 26.2-3, pp. 123–147.

Foote, M., A. I. Miller, D. M. Raup, and S. M. Stanley (2007). "Principles of paleontology". In: *Macmillan*.

Fraïsse, C, J. Elderfield, and J. Welch (2014). "The genetics of speciation: are complex incompatibilities easier to evolve?" In: *Journal of Evolutionary Biology* 27.4, pp. 688–699.

Frost, S. D. and E. M. Volz (2013). "Modelling tree shape and structure in viral phylodynamics". In: *Philosophical Transactions of the Royal Society B* 368.1614, p. 20120208.

Gascon, C., S. C. Lougheed, and J. P. Bogart (1998). "Patterns of Genetic Population Differentiation in Four Species of Amazonian Frogs: A Test of the Riverine Barrier Hypothesis 1". In: *Biotropica* 30.1, pp. 104–119.

Gascon, C., J. R. Malcolm, J. L. Patton, M. N. da Silva, J. P. Bogart, S. C. Lougheed, C. A. Peres, S. Neckel, and P. T. Boag (2000). "Riverine barriers and the geographic distribution of Amazonian species". In: *Proceedings of the National Academy of Sciences* 97.25, pp. 13672–13677.

Gascuel, F., R. Ferrière, R. Aguilée, and A. Lambert (2015a). "How ecology and landscape dynamics shape phylogenetic trees". In: *Systematic Biology* 64.4, pp. 590–607. ISSN: 1076836X.

Gascuel, F., R. Ferrière, R. Aguilée, and A. Lambert (2015b). "How ecology and landscape dynamics shape phylogenetic trees". In: *Systematic Biology* 64.4, pp. 590–607.

Gascuel, O. and M. Steel (2006). "Neighbor-joining revealed". In: *Molecular Biology and Evolution* 23.11, pp. 1997–2000.

Gavrilets, S (2004). "Fitness Landscapes and the Origin of Species". In: *Princeton University Press, Princeton, NJ*.

Gavrilets, S and A Vose (2005). "Dynamic patterns of adaptive radiation". In: *Proceedings of the National Academy of Sciences* 102.50, pp. 18040–18045.

Gavrilets, S. (2000). "Waiting time to parapatric speciation". In: *Proceedings of the Royal Society of London B: Biological Sciences* 267.1461, pp. 2483–2492.

– (2003). "Perspective: models of speciation: what have we learned in 40 years?" In: *Evolution* 57.10, pp. 2197–2215.

– (2014). "Models of speciation: Where are we now?" In: *Journal of Heredity* 105.S1, pp. 743–755.

Gavrilets, S. and J. B. Losos (2009). "Adaptive radiation: contrasting theory with data". In: *Science* 323.5915, pp. 732–737.

Gavrilets, S., L. Hai, and M. D. Vose (1998). "Rapid parapatric speciation on holey adaptive landscapes". In: *Proceedings of the Royal Society of London B: Biological Sciences* 265.1405, pp. 1483–1489.

Gavrilets, S., R. Acton, and J. Gravner (2000a). "Dynamics of speciation and diversification in a metapopulation". In: *Evolution* 54.5, pp. 1493–1501.

Gavrilets, S., H. Li, and M. D. Vose (2000b). "Patterns of parapatric speciation". In: *Evolution* 54.4, pp. 1126–1134.

Gillespie, R. (2004). "Community assembly through adaptive radiation in Hawaiian spiders". In: *Science* 303.5656, pp. 356–359.

Goldani, A, G. Carvalho, and J. Bicca-Marques (2006). "Distribution patterns of Neotropical primates (Platyrrhini) based on Parsimony Analysis of Endemicity". In: *Brazilian Journal of Biology* 66.1A, pp. 61–74.

Gourbiere, S. and J. Mallet (2010). "Are species real? The shape of the species boundary with exponential failure, reinforcement, and the "missing snowball"". In: *Evolution* 64.1, pp. 1–24.

Grant, P. R. (1999). "Ecology and evolution of Darwin's finches". In: *Princeton University Press*.

Hagen, O., K. Hartmann, M. Steel, and T. Stadler (2015). "Age-dependent speciation can explain the shape of empirical phylogenies". In: *Systematic Biology* 64.3, pp. 432–440.

Hall, J. P. and D. J. Harvey (2002). "The phylogeography of Amazonia revisited: new evidence from riodinid butterflies". In: *Evolution* 56.7, pp. 1489–1497.

Hamming, R. W. (1950). "Error detecting and error correcting codes". In: *Bell Labs Technical Journal* 29.2, pp. 147–160.

Hasegawa, M., H. Kishino, and T.-a. Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *Journal of Molecular Evolution* 22.2, pp. 160–174.

Hein, J., M. Schierup, and C. Wiuf (2004). "Gene genealogies, variation and evolution: a primer in coalescent theory". In: *Oxford University Press, USA*.

Heinicke, M. P., E. Greenbaum, T. R. Jackman, and A. M. Bauer (2011). "Phylogeny of a trans-Wallacean radiation (Squamata, Gekkonidae, *Gehyra*) supports a single early colonization of Australia". In: *Zoologica Scripta* 40.6, pp. 584–602.

Higgs, P. G. and B. Derrida (1991). "Stochastic models for species formation in evolving populations". In: *Journal of Physics A: Mathematical and General* 24.17, p. L985.

– (1992). "Genetic distance and species formation in evolving populations". In: *Journal of Molecular Evolution* 35.5, pp. 454–465.

Hodgson, J. A., K. N. Sterner, L. J. Matthews, A. S. Burrell, R. A. Jani, R. L. Raaum, C.-B. Stewart, and T. R. Disotell (2009). "Successive radiations, not stasis, in the South American primate fauna". In: *Proceedings of the National Academy of Sciences* 106.14, pp. 5534–5539.

Hoelzer, G. A., R. Drewes, J. Meier, and R. Doursat (2008). "Isolation-by-distance and outbreeding depression are sufficient to drive parapatric speciation in the absence of environmental influences". In: *PLoS Computational Biology* 4.7, e1000126.

Hoorn, C et al. (2010). "Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity." In: *Science (New York, N.Y.)* 330.6006, pp. 927–31.

Hubbell, S. P. (2001). "The Unified Neutral Theory of Biodiversity and Biogeography". In: *Princeton University Press, Princeton, NJ*.

Hurlbert, A. H. and J. C. Stegen (2014). "When should species richness be energy limited, and how would we know?" In: *Ecology Letters* 17.4, pp. 401–413.

IUCN (2015). "The IUCN Red List of Threatened Species". In: *Version 2015-1*.

Jabot, F. and J. Chave (2009). "Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests". In: *Ecology Letters* 12.3, pp. 239–248.

Jukes, T. H., C. R. Cantor, et al. (1969). "Evolution of protein molecules". In: *Mammalian Protein Metabolism* 3.21, p. 132.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". In: *Journal of Molecular Evolution* 16.2, pp. 111–120.

Kirkpatrick, M. and V. Ravigné (2002). "Speciation by natural and sexual selection: models and experiments". In: *The American Naturalist* 159.S3, S22–S35.

Kirkpatrick, M. and M. Slatkin (1993). "Searching for evolutionary patterns in the shape of a phylogenetic tree". In: *Evolution* 47.4, pp. 1171–1181.

Kisel, Y. and T. G. Barraclough (2010). "Speciation has a spatial scale that depends on levels of gene flow". In: *The American Naturalist* 175.3, pp. 316–334.

Kocher, T. D. (2004). "Adaptive evolution and explosive speciation: the cichlid fish model". In: *Nature Reviews Genetics* 5.4, p. 288.

Kopp, M. (2010). "Speciation and the neutral theory of biodiversity: Modes of speciation affect patterns of biodiversity in neutral communities." In: *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 32.7, pp. 564–70.

Kraaijeveld, K. (2010). "Genome size and species diversification". In: *Evolutionary Biology* 37.4, pp. 227–233.

Lewitus, E. and H. Morlon (2016). "Natural constraints to species diversification". In: *PLoS Biology* 14.8, e1002532.

Li, J., J.-P. Huang, J. Sukumaran, and L. L. Knowles (2018). "Microevolutionary processes impact macroevolutionary patterns". In: *BMC evolutionary biology* 18.1, p. 123.

Lima, M. G., J. C. Buckner, J. d.S. e. Silva-Júnior, A. Aleixo, A. B. Martins, J. P. Boubli, A. Link, I. P. Farias, M. N. da Silva, F. Röhe, et al. (2017). "Capuchin monkey biogeography: understanding *Sapajus* Pleistocene range expansion and the current sympatry between *Cebus* and *Sapajus*". In: *Journal of Biogeography* 44.4, pp. 810–820.

Losos, J. and R. Thorpe (2004). "Evolutionary diversification of Caribbean Anolis lizards". In: *Adaptive speciation. Cambridge Univ. Press, Cambridge, England*, pp. 322–324.

Losos, J. B. and R. E. Glor (2003). "Phylogenetic comparative methods and the geography of speciation". In: *Trends in Ecology & Evolution* 18.5, pp. 220–227.

Losos, J. B. and D. L. Mahler (2010). "Adaptive radiation: the interaction of ecological diversification appears to be characterized by opportunity, adaptation, and speciation". In: *Evolution since Darwin: The First 150 Years*, pp. 381–420.

Losos, J. B. and R. E. Ricklefs (2009). "Adaptation and diversification on islands". In: *Nature* 457.7231, p. 830.

Lynch-Alfaro, J. W., J. P. Boubli, L. E. Olson, A. Di Fiore, B. Wilson, G. A. Gutiérrez-Espeleta, K. L. Chiou, M. Schulte, S. Neitzel, V. Ross, et al. (2012). "Explosive Pleistocene range expansion leads to widespread Amazonian sympatry between robust and gracile capuchin monkeys". In: *Journal of Biogeography* 39.2, pp. 272–288.

Manceau, M., A. Lambert, and H. Morlon (2015). "Phylogenies support out-of-equilibrium models of biodiversity". In: *Ecology Letters* 18.4, pp. 347–356.

Mank, J. E. and J. C. Avise (2006). "Cladogenetic correlates of genomic expansions in the recent evolution of actinopterygiian fishes". In: *Proceedings of the Royal Society of London B: Biological Sciences* 273.1582, pp. 33–38.

Manzo, F and L Peliti (1994). "Geographic speciation in the Derrida-Higgs model of species formation". In: *Journal of Physics A: Mathematical and General* 27.21, p. 7079.

Martin, C. H., J. S. Cutler, J. P. Friel, C. Dening Touokong, G. Coop, and P. C. Wainwright (2015). "Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on One of the clearest examples of sympatric speciation". In: *Evolution* 69.6, pp. 1406–1422.

Martín, H. G. and N. Goldenfeld (2006). "On the origin and robustness of power-law species–area relationships in ecology". In: *Proceedings of the National Academy of Sciences* 103.27, pp. 10310–10315.

Martins, A. B., M. A. de Aguiar, and Y. Bar-Yam (2013). "Evolution and stability of ring species". In: *Proceedings of the National Academy of Sciences* 110.13, pp. 5080–5084.

Martins, A. d. B. and M. A. M. de Aguiar (2017). "Barriers to gene flow and ring species formation". In: *Evolution* 71.2, pp. 442–448.

May, F., A. Huth, and T. Wiegand (2015). "Moving beyond abundance distributions: neutral theory and spatial patterns in a tropical forest". In: *Proceedings of the Royal Society B: Biological Sciences* 282.1802, p. 20141657.

McGee, M. D., B. C. Faircloth, S. R. Borstein, J. Zheng, C. D. Hulsey, P. C. Wainwright, and M. E. Alfaro (2016). "Replicated divergence in cichlid radiations mirrors a major vertebrate innovation". In: *Proceedings of the Royal Society B: Biological Sciences* 283.1822, p. 20151413.

McPeek, M. A. (2008). "The ecological dynamics of clade diversification and community assembly". In: *The American Naturalist* 172.6, E270–E284.

Meier, J. I., D. A. Marques, S. Mwaiko, C. E. Wagner, L. Excoffier, and O. Seehausen (2017a). "Ancient hybridization fuels rapid cichlid fish adaptive radiations". In: *Nature Communications* 8, p. 14363.

Meier, J. I., V. C. Sousa, D. A. Marques, O. M. Selz, C. E. Wagner, L. Excoffier, and O. Seehausen (2017b). "Demographic modelling with whole-genome data reveals parallel origin of similar *Pundamilia* cichlid species after hybridization". In: *Molecular Ecology* 26.1, pp. 123–141.

Melián, C. J., D. Alonso, S. Allesina, R. S. Condit, and R. S. Etienne (2012). "Does sex speed up evolutionary rate and increase biodiversity?" In: *PLoS Computational Biology* 8.3, e1002414.

Melián, C. J., O. Seehausen, V. M. Eguíluz, M. A. Fortuna, and K. Deiner (2015). "Diversification and biodiversity dynamics of hot and cold spots". In: *Ecography* 38.4, pp. 393–401.

Meyer, B. S., M. Matschiner, and W. Salzburger (2015). "A tribal level phylogeny of Lake Tanganyika cichlid fishes based on a genomic multi-marker approach". In: *Molecular Phylogenetics and Evolution* 83, pp. 56–71.

M'Gonigle, L. K., R. Mazzucco, S. P. Otto, and U. Dieckmann (2012). "Sexual selection enables long-term coexistence despite ecological equivalence". In: *Nature* 484.7395, pp. 506–509.

Missa, O., C. Dytham, and H. Morlon (2016). "Understanding how biodiversity unfolds through time under neutral theory". In: *Philosophical Transactions of the Royal Society of London B* 371.1691, p. 20150226.

Mittermeier, R. A., D. E. Wilson, and A. B. Rylands (2013). "Handbook of the mammals of the world: primates". In: *Lynx Edicions*.

Moen, D. and H. Morlon (2014). "Why does diversification slow down?" In: *Trends in Ecology & Evolution* 29.4, pp. 190–197.

Mooers, A., L. J. Harmon, M. G. Blum, D. H. Wong, and S. B. Heard (2007). "Some models of phylogenetic tree shape". In: *Reconstructing Evolution: New Mathematical and Computational Advances*, pp. 149–170.

Mooers, A. O. and S. B. Heard (1997). "Inferring evolutionary process from phylogenetic tree shape". In: *The Quarterly Review of Biology* 72.1, pp. 31–54.

Moraes, L, D Pavan, M. Barros, and C Ribas (2016). "Combined influence of riverine barriers and flooding gradient on biogeographical patterns of amphibians and squamates in South–eastern Amazonia". In: *Journal of Biogeography* 43, pp. 2113–2124.

Morlon, H. (2014). "Phylogenetic approaches for studying diversification". In: *Ecology Letters* 17.4, pp. 508–525.

Muller, H. J. (1939). "Reversibility in evolution considered from the standpoint of genetics". In: *Biological Reviews* 14.3, pp. 261–280.

Murtagh, F. (1984). "Complexities of hierarchic clustering algorithms: State of the art". In: *Computational Statistics Quarterly* 1.2, pp. 101–113.

Nee, S., A. O. Mooers, and P. H. Harvey (1992). "Tempo and mode of evolution revealed from molecular phylogenies". In: *Proceedings of the National Academy of Sciences* 89.17, pp. 8322–8326.

Nosil, P. (2012). "Ecological speciation". In: *Oxford University Press*.

O'Dwyer, J. P. and J. L. Green (2010). "Field theory for biogeography: a spatially explicit model for predicting patterns of biodiversity." In: *Ecology letters* 13.1, pp. 87–95.

Orr, H. A. (1995). "The population genetics of speciation: the evolution of hybrid incompatibilities." In: *Genetics* 139.4, pp. 1805–1813.

Orr, H. A. and L. H. Orr (1996). "Waiting for speciation: the effect of population subdivision on the time to speciation". In: *Evolution* 50.5, pp. 1742–1749.

Paradis, E., J. Claude, and K. Strimmer (2004). "APE: analyses of phylogenetics and evolution in R language". In: *Bioinformatics* 20.2, pp. 289–290.

Pennings, P. S., M. Kopp, G. Meszéna, U. Dieckmann, and J. Hermisson (2007). "An analytically tractable model for competitive speciation". In: *The American Naturalist* 171.1, E44–E71.

Perelman, P. et al. (2011). "A molecular phylogeny of living primates". In: *PLoS Genetics* 7.3, e1001342.

Perez-Sweeney, B. M., C. Valladares-Padua, C. S. Martins, J. C. Morales, and D. J. Melnick (2008). "Examination of the taxonomy and diversification of *Leontopithecus* using the mitochondrial control region". In: *International Journal of Primatology* 29.1, pp. 245–263.

Phillimore, A. B. and T. D. Price (2008). "Density-dependent cladogenesis in birds". In: *PLoS Biology* 6.3, e71.

Pigot, A. L., A. B. Phillimore, I. P. Owens, and C. D. L. Orme (2010). "The shape and temporal dynamics of phylogenetic trees arising from geographic speciation". In: *Systematic Biology* 59.6, pp. 660–673.

Pomara, L. Y., K. Ruokolainen, and K. R. Young (2014). "Avian species composition across the Amazon R iver: the roles of dispersal limitation and environmental heterogeneity". In: *Journal of Biogeography* 41.4, pp. 784–796.

Puttick, M. N., J. Clark, and P. C. Donoghue (2015). "Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms". In: *Proceedings of the Royal Society of London B: Biological Sciences* 282.1820, p. 20152289.

Pybus, O. G. and P. H. Harvey (2000). "Testing macro–evolutionary models using incomplete molecular phylogenies". In: *Proceedings of the Royal Society of London B: Biological Sciences* 267.1459, pp. 2267–2272.

Quental, T. B. and C. R. Marshall (2011). "The molecular phylogenetic signature of clades in decline". In: *PloS One* 6.10, e25780.

Rabelo, R. M., F. E. Silva, T. Vieira, J. Ferreira-Ferreira, F. P. Paim, W. Dutra, J. Valsecchi, et al. (2014). "Extension of the geographic range of *Ateles chamek* (Primates, Atelidae): evidence of river-barrier crossing by an amazonian primate". In: *Primates* 55.2, pp. 167–171.

Rabosky, D. L. and I. J. Lovette (2008). "Explosive evolutionary radiations: decreasing speciation or increasing extinction through time?" In: *Evolution* 62.8, pp. 1866–1875.

Rettelbach, A., M. Kopp, U. Dieckmann, and J. Hermisson (2013). "Three modes of adaptive speciation in spatially structured populations". In: *The American Naturalist* 182.6, E215–E234.

Revell, L. J. and S. A. Chamberlain (2014). "Rphylip: an R interface for PHYLIP". In: *Methods in Ecology and Evolution* 5.9, pp. 976–981.

Robinson, D. F. and L. R. Foulds (1981). "Comparison of phylogenetic trees". In: *Mathematical Biosciences* 53.1-2, pp. 131–147.

Rosenberg, N. A. and M. Nordborg (2002). "Genealogical trees, coalescent theory and the analysis of genetic polymorphisms". In: *Nature reviews genetics* 3.5, p. 380.

Rosenberger, A. L. (1992). "Evolution of feeding niches in New World monkeys". In: *American Journal of Physical Anthropology* 88.4, pp. 525–562.

Rosindell, J., S. J. Cornell, S. P. Hubbell, and R. S. Etienne (2010). "Protracted speciation revitalizes the neutral theory of biodiversity". In: *Ecology Letters* 13.6, pp. 716–727.

Rosindell, J., L. J. Harmon, and R. S. Etienne (2015). "Unifying ecology and macroevolution with individual-based theory". In: *Ecology Letters* 18.5, pp. 472–482.

Rowe, K. C., K. P. Aplin, P. R. Baverstock, and C. Moritz (2011). "Recent and rapid speciation with limited morphological disparity in the genus *Rattus*". In: *Systematic Biology* 60.2, pp. 188–203.

Rubinoff, D. and B. S. Holland (2005). "Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference". In: *Systematic Biology* 54.6, pp. 952–961.

Rundell, R. J. and T. D. Price (2009). "Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation". In: *Trends in Ecology & Evolution* 24.7, pp. 394–399.

Rylands, A. B., E. W. Heymann, J. Lynch Alfaro, J. C. Buckner, C. Roos, C. Matauschek, J. P. Boubli, R. Sampaio, and R. A. Mittermeier (2016). "Taxonomic review of the new world tamarins (primates: Callitrichidae)". In: *Zoological Journal of the Linnean Society* 177.4, pp. 1003–1028.

Rzhetsky, A. and M. Nei (1993). "Theoretical foundation of the minimum-evolution method of phylogenetic inference." In: *Molecular Biology and Evolution* 10.5, pp. 1073–1095.

Sackin, M. (1972). ""Good" and "bad" phenograms". In: *Systematic Biology* 21.2, pp. 225–226.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Molecular Biology and Evolution* 4.4, pp. 406–425.

Sanjuán, R. and S. F. Elena (2006). "Epistasis correlates to genomic complexity". In: *Proceedings of the National Academy of Sciences* 103.39, pp. 14402–14405.

Santorelli, S., W. E. Magnusson, and C. P. Deus (2018). "Most species are not limited by an Amazonian river postulated to be a border between endemism areas". In: *Scientific Reports* 8.1, p. 2294.

Scally, A. and R. Durbin (2012). "Revising the human mutation rate: implications for understanding human evolution". In: *Nature Reviews Genetics* 13.10, p. 745.

Schliep, K. P. (2011). "phangorn: phylogenetic analysis in R". In: *Bioinformatics* 27.4, pp. 592–593.

Schluter, D. (2000). "The ecology of adaptive radiation". In: *OUP Oxford*.

Schneider, D. M., E. M. Baptestini, and M. A. de Aguiar (2016). "Diploid versus haploid models of neutral speciation". In: *Journal of Biological Physics* 42.2, pp. 235–245.

Schneider, H. and I. Sampaio (2015). "The systematics and evolution of New World primates–A review". In: *Molecular Phylogenetics and Evolution* 82, pp. 348–357.

Seehausen, O. (2015). "Process and pattern in cichlid radiations–inferences for understanding unusually high rates of evolutionary diversification". In: *New Phytologist* 207.2, pp. 304–312.

Seehausen, O. et al. (2014). "Genomics and the origin of species". In: *Nature Reviews Genetics* 15.3, pp. 176–192.

Serva, M. and L Peliti (1991). "A statistical model of an evolving population with sexual reproduction". In: *Journal of Physics A: Mathematical and General* 24.13, p. L705.

Servedio, M. R., Y. Brandvain, S. Dhole, C. L. Fitzpatrick, E. E. Goldberg, C. A. Stern, J. Van Cleve, and D. J. Yeh (2014). "Not just a theory – the utility of mathematical models in evolutionary biology". In: *PLoS Biology* 12.12, e1002017.

Shaw, K. L. (2002). "Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets". In: *Proceedings of the National Academy of Sciences* 99.25, pp. 16122–16127.

Simoes, M, L Breitkreuz, M Alvarado, S Baca, J. Cooper, L Heins, K Herzog, and B. Lieberman (2016). "The evolving theory of evolutionary radiations". In: *Trends in Ecology & Evolution* 31.1, pp. 27–34.

Simon, A., B. Goffinet, N. Magain, and E. Sérusiaux (2018). "High diversity, high insular endemism and recent origin in the lichen genus *Sticta* (lichenized Ascomycota, Peltigerales) in Madagascar and the Mascarenes". In: *Molecular Phylogenetics and Evolution* 122, pp. 15–28.

Simpson, G. G. (1955). "Major features of evolution". In: *Columbia University Press: New York*.

Springer, M. S., R. W. Meredith, J. Gatesy, C. A. Emerling, J. Park, D. L. Rabosky, T. Stadler, C. Steiner, O. A. Ryder, J. E. Janečka, et al. (2012). "Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix". In: *PloS One* 7.11, e49521.

Taft, R. J., M. Pheasant, and J. S. Mattick (2007). "The relationship between non-protein-coding DNA and eukaryotic complexity". In: *Bioessays* 29.3, pp. 288–299.

Tucker, C. M. and M. W. Cadotte (2013). "Unifying measures of biodiversity: understanding when richness and phylogenetic diversity should be congruent". In: *Diversity and Distributions* 19.7, pp. 845–854.

Turelli, M., N. H. Barton, and J. A. Coyne (2001). "Theory and speciation". In: *Trends in Ecology & Evolution* 16.7, pp. 330–343.

Uyeda, J. C., S. J. Arnold, P. A. Hohenlohe, and L. S. Mead (2009). "Drift promotes speciation by sexual selection". In: *Evolution* 63.3, pp. 583–594.

Van Roosmalen, M. G., T. Van Roosmalen, and R. A. Mittermeier (2002). "A taxonomic review of the titi monkeys, genus *Callicebus* Thomas, 1903, with the description of two new species, *Callicebus bernhardi* and *Callicebus stephennashi*, from Brazilian Amazonia". In: *Neotropical Primates* 10.supplement, pp. 1–52.

Wagner, G. P., M. Pavlicev, and J. M. Cheverud (2007). "The road to modularity". In: *Nature Reviews Genetics* 8.12, p. 921.

Wallace, A. R. (1854). "On the monkeys of the Amazon". In: *Annals and Magazine of Natural History* 14.84, pp. 451–454.

Wang, S., A. Chen, J. Fang, and S. W. Pacala (2013). "Speciation rates decline through time in individual-based models of speciation and extinction". In: *The American Naturalist* 182.3, E83–E93.

Wiens, J. J. (2004). "Speciation and ecology revisited: phylogenetic niche conservatism and the origin of species". In: *Evolution* 58.1, pp. 193–197.

Wildman, D. E., N. M. Jameson, J. C. Opazo, and V. Y. Soojin (2009). "A fully resolved genus level phylogeny of neotropical primates (Platyrrhini)". In: *Molecular Phylogenetics and Evolution* 53.3, pp. 694–702.

Yamaguchi, R. and Y. Iwasa (2013). "First passage time to allopatric speciation". In: *Interface Focus* 3.6, p. 20130026.

Yang, Z. (1994). “Estimating the pattern of nucleotide substitution”. In: *Journal of Molecular Evolution* 39.1, pp. 105–111.

Yang, Z. and B. Rannala (2012). “Molecular phylogenetics: principles and practice”. In: *Nature Reviews Genetics* 13.5, p. 303.

Yoder, J. B. et al. (2010). “Ecological opportunity and the origin of adaptive radiations”. In: *Journal of Evolutionary Biology* 23.8, pp. 1581–1596.

Yule, G. U. (1925). “A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS”. In: *Philosophical Transactions of the Royal Society of London B* 213, pp. 21–87.

Zhang, Y.-C., M. Serva, and M. Polikarpov (1990). “Diffusion reproduction processes”. In: *Journal of Statistical Physics* 58.5-6, pp. 849–861.

Appendix A

Supporting Information - Chapter 1

A.1 Properties of MRCAT and SSEE matrices

A.1.1 The N-1 distinct times

MRCAT matrices corresponding to asexual reproduction, maternal and paternal matrices in sexual populations and SSEE matrices have at most $N-1$ distinct positive entries plus the entries 0 at the main diagonal. As an example, also to be used in section A.2, consider the following MRCAT matrix

$$T = \begin{pmatrix} 0 & 3 & 6 & 6 & 5 & 7 \\ 3 & 0 & 6 & 6 & 5 & 7 \\ 6 & 6 & 0 & 4 & 6 & 7 \\ 6 & 6 & 4 & 0 & 6 & 7 \\ 5 & 5 & 6 & 6 & 0 & 7 \\ 7 & 7 & 7 & 7 & 7 & 0 \end{pmatrix} . \tag{A.1}$$

with lines 1 to 6 corresponding to six individuals labeled A to F, respectively. Here only the 5 numbers $3, 4, 5, 6$ and 7 appear. If the time to the most recent common ancestral between individuals C and D were also 3 the matrix would have only 4 different entries, $3, 5, 6$ and 7. However, in any of the matrices listed above there is never more than $N-1$ different positive numbers.

To prove this statement in general we shall consider only the case of asexual reproduction, since the proof for maternal or paternal lineages is very similar. Suppose that at generation $t-1$ T_{t-1} has $k_{t-1} \leq N_{t-1} - 1$ different positive entries $a_1, a_2, ..., a_{k_{t-1}}$ plus the entries 0 at the main diagonal. Then, according to the update rule given by Eq. (1) of the main text the values of $T_t(i,j) = T_{t-1}(P(i), P(j)) + 1$ will assume the values $a_i + 1$ ($i = 1, 2, \ldots k_{t-1}$) only if every individual of generation $t-1$ had exactly one offspring. In this case $N_t = N_{t-1}$ and T_t would have the same number of distinct entries as T_{t-1}, i.e., $k_t = k_{t-1} \leq N_t - 1$.

If, however, D individuals did not reproduce and M individuals had more than one offspring, then

$$k_t = \begin{cases} k_{t-1} - D + 1 & \leq N_{t-1} - D \qquad \text{if} \quad M \neq 0 \\ k_{t-1} - D & \leq N_{t-1} - D - 1 \quad \text{if} \quad M = 0 \end{cases} \tag{A.2}$$

since multiple offspring only contribute with $T_t(i, j) = 1$ whenever i and j shared the same parent. The number of individuals, on the other hand, is $N_t = N_{t-1} - D + (M_o - M)$, where M_o is the total number of offspring of all the M individuals that had more then one offspring. If $M \neq 0$ then $M_o - M \geq 1$ and $N_t \geq N_{t-1} - D + 1$. If $M = M_o = 0$ then $N_t = N_{t-1} - D$. Therefore

$$k_t \leq \begin{cases} N_{t-1} - D & \leq N_t - 1 \quad \text{if} \quad M \neq 0 \\ N_{t-1} - D - 1 & \leq N_t - 1 \quad \text{if} \quad M = 0 \end{cases} \tag{A.3}$$

which implies $k_t < N_t - 1$ in both cases. Since $k_0 = 0$ the relation holds for any t.

For hermaphroditic populations or general genealogies the above proof is not valid, since the update rule in this case is more complicated, given by Eq.(4) of the paper, and there might indeed be more than $N - 1$ distinct times, as discussed in Supporting Information, section I.

Since there are exactly $N - 1$ internal nodes in the tree and since each node indicates a time to the most recent common ancestor between a pair of individuals, the distinct times correspond exactly to the times indicated by the nodes. Notice that these times are the times to the most recent common ancestor between adjacent neighbors in the base of the tree. In matrix (A.1) and Fig. 3 of the paper, for example, $T_{C,D} = 4$, $T_{D,A} = 6$, $T_{A,B} = 3$, $T_{B,E} = 5$ and $T_{E,F} = 7$.

A.1.2 Canonical form of matrices

For the cases considered in the previous subsection, i.e., asexual reproduction, maternal and paternal matrices in sexual populations and SSEE, all the information in the matrix is contained in their $N - 1$ distinct times and all remaining entries of the matrix can be obtained from them. In order to do that it is important to rearrange the matrix by re-labeling individuals by the order they appear in the base of the tree. In the example described in Fig. A.1 this corresponds to rename

$$\begin{aligned} C &\to 1 \\ D &\to 2 \\ A &\to 3 \\ B &\to 4 \\ E &\to 5 \\ F &\to 6. \end{aligned}$$

In this new order the $N-1$ distinct times are given by $T_{k,k+1}$ for $k = 1, \ldots, N-1$, corresponding to the first diagonal of the matrix (right above the main diagonal). In this order the matrix is in the *canonical form* and all other elements can be obtained immediately. The time to common ancestor between second neighbors, individuals k and $k+2$ is the maximum between $T_{k,k+1}$ and $T_{k+1,k+2}$:

$$T_{k,k+2} = \max\{T_{k,k+1}, T_{k+1,k+2}\} \tag{A.4}$$

for $k = 1, \ldots, N-2$. Analogously,

$$T_{k,k+3} = \max\{T_{k,k+2}, T_{k+1,k+3}\} \tag{A.5}$$

for $k = 1, \ldots, N-3$ and so on. In general

$$T_{k,k+m} = \max\{T_{k,k+m-1}, T_{k+1,k+m}\} \tag{A.6}$$

for $m = 2, \ldots, N-1$ and $k = 1, \ldots, N-m$.

For the matrix (A.1) we obtain

$$T = \begin{pmatrix} 0 & \mathbf{4} & 6 & 6 & 6 & 7 \\ 4 & 0 & \mathbf{6} & 6 & 6 & 7 \\ 6 & 6 & 0 & \mathbf{3} & 5 & 7 \\ 6 & 6 & 3 & 0 & \mathbf{5} & 7 \\ 6 & 6 & 5 & 5 & 0 & \mathbf{7} \\ 7 & 7 & 7 & 7 & 7 & 0 \end{pmatrix} \tag{A.7}$$

with the distinct times displayed in bold face.

A.2 Constructing the tree with the different methods

A.2.1 Asexual, maternal and paternal MRCAT

A tree is a graph in which external nodes (or tips) represent extant species (phylogeny) or individuals (genealogy) and are called leaves. Internal nodes are the most common ancestors between a pair of species or individuals and branches reflect the time between an ancestor and its descendants (figure A.1). Building a tree requires the joining of individuals into groups, from the leaves all the way to the root. Three different types of grouping can occur in this process: (1) a pair of individuals can be joined together (a structure called *cherry*); (2) an individual can join an existing group or; (3) two groups can be joined together. At each step of the construction one of these processes occur, always starting with the pair of individuals having the most recent common ancestor, which is necessarily of type (1).

In order to illustrate the building of trees we will consider the MRCAT matrix given by matrix (A.1). To this matrix there corresponds a unique tree, shown in Fig. A.1, and the algorithm we describe here works only in this situation. At the end of this section we comment on the modifications needed to draw trees in other situations. If the individuals belong to the same species the tree will be a genealogy. If each individual is taken from a different species the tree is the corresponding phylogeny.

From matrix (A.1) we see that individuals A and B share a common ancestor only 3 generations ago and they form the first group, $g_1 = \{A, B\}$, since $T_{AB} = 3$ is the smallest entry on the matrix. The second group is formed by individuals C and D ($g_2 = \{C, D\}$), corresponding to $T_{CD} = 4$, the second smallest entry in the MRCAT matrix. The next smallest time, $T_{BE} = 5$, involves an individual already grouped in a previous step (individual B, belonging to g_1), so the grouping in this case is of type (2). Accordingly, the new group will be formed by the juxtaposition of this previous group and the new individual ($g_3 = \{g_1, E\} = \{\{A, B\}, E\}$). Continuing this way, we find next $T_{DE} = 6$, a situation in which both individuals already belong to a previously formed group. In this case the new group is formed by joining the corresponding groups ($g_4 = \{\{g_2\}, \{g_3\}\} = \{C, D, A, B, E\}$). Finally, the highest entry in the matrix corresponds to $T_{EF} = 7$, forming the last group containing all six individuals: $g_5 = \{\{g_4\}, F\} = \{C, D, A, B, E, F\}$. This group gives precisely the order of individuals in the base of the tree, as depicted in figure A.1.

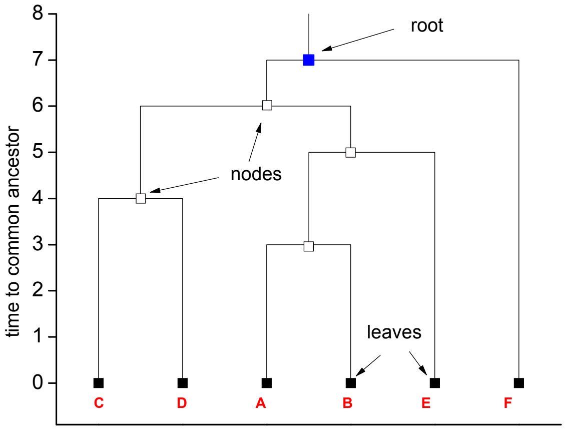


FIGURE A.1: Example of a tree with six individuals, represented by filled squares as the leaves of the tree. Open squares denote the nodes, or the most recent common ancestors between each pair of individuals, and the topmost node is the tree root, or the most recent common ancestor among all six individuals. Letters in red below the leaves are individual labels.

The x-axis of the tree can represent different units of measure, like genetic or phenotypic distance between individuals. The x coordinate of a group is defined as the average of the x value of its members. In figure A.1, the individuals are separated by one unit in x-axis for simplicity. In the example $x_A = 3$, $x_B = 4$ and $x_{g_1} = 3.5$.

Building MRCAT trees with different individuals

The choice of the individual for constructing the phylogenetic tree with MRCAT can matter. However, the final structure of the tree will barely vary for the parameters used in our simulations. Right after a speciation event, for instance from species A to species $A1$ and $A2$, the time to the most recent common ancestor between one individual belonging to $A1$ and one belonging to $A2$ will certainly vary. For instance, let individuals A, B and E in Figure A.1 belong to species $A1$, individuals C and D to species $A2$ (and individual F to another species). If we choose individuals B and C to compute the distance to the common ancestor in the phylogenetic tree between species $A1$ and A2, the distance will be different if we choose individuals E and C. However, after a certain number of generations (of the order of the size of the species population), all individuals of $A1$ will have a single common ancestor in $A1$, and similarly

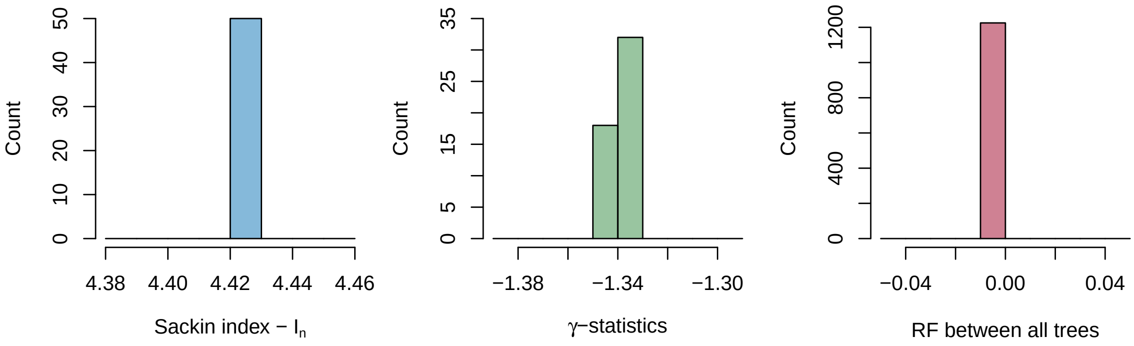


FIGURE A.2: Histograms of γ-statistics, Sackin index (I_n) and the structural comparison between trees thought the Robison-Foulds metric made for 50 phylogenetic trees built by the MRCAT methodology, using different sets of individuals from the same genealogy.

for $A2$. From this moment on, the branch length to the common ancestor of any pair of individuals of different species is the same. The only situation when this is not true is the one depicted in the given example, i.e., when several speciation events occur concomitantly and recently. Even in these cases, however, the variation in the branch lengths distribution is very small, and the balance and structure of the trees are not expected to vary too much. Figure A.2 shows the distributions of γ-statistics, Sackin index (I_n), and RF between every pair of trees (B=150 and S=5) for 50 trees built from a single simulation using random choice of individuals. This means that from the same genealogy we chose a random set of individuals of each species to make the phylogenetic tree. In this case, no structural and balancing differences were found ($\overline{I}_n = 4.424$, $\sigma_{I_n} = 0.000$; $\overline{RF} = 0.000$, $\sigma_{RF} = 0.000$). Only the branch lengths had a small variation ($\overline{\gamma} = -1.342$, $\sigma_\gamma = 0.003$).

A.2.2 SSEE

For SSEE matrices the process is almost identical to that for MRCAT matrices. The only differences are that nodes represent species (not individuals) and branches associated to extinct species should not be draw all the way down to present time, but should stop at the extinction time recorded in the vector E.

A.2.3 Hermaphroditic MRCAT and Genetic distances

In the cases of hermaphroditic simulations, inclusion of sex separation but with the MRCA taking into account both parents and genetic distances between individuals used to estimate trees, there are usually much more than $N-1$ distinct entries and the only property that still holds is that the matrix is symmetric, i.e., $T_{ij} = T_{ji}$. Clustering methods deal with this feature by recalculating the times (or genetic distances) every time a group is formed. We illustrate the procedure here for the case of the UPGMA method (Felsenstein, 2004) and a MRCAT matrix. After finding the two elements i and j with the smallest time to the most recent common ancestral (which might be two individuals in the case of a cherry, an individual and a group or two groups) we remove j as before as leave i as representing the pair ij. Next, the time of all remaining elements to ij is recalculated as

$$T_{i,k} = \frac{n_i}{n_i + n_j} T_{i,k} + \frac{n_j}{n_i + n_j} T_{j,k} \tag{A.8}$$

for all $k \neq i, j$ where n_i and n_j are the size of elements i and j respectively (the size is 1 if the element is an individual or, if the element is a groups, the number of individuals in the group). All other steps of the algorithm remain the same. Figure A.3 illustrates the differences in genealogies drawn from general (without separating the sex), maternal, paternal MRCAT matrices.

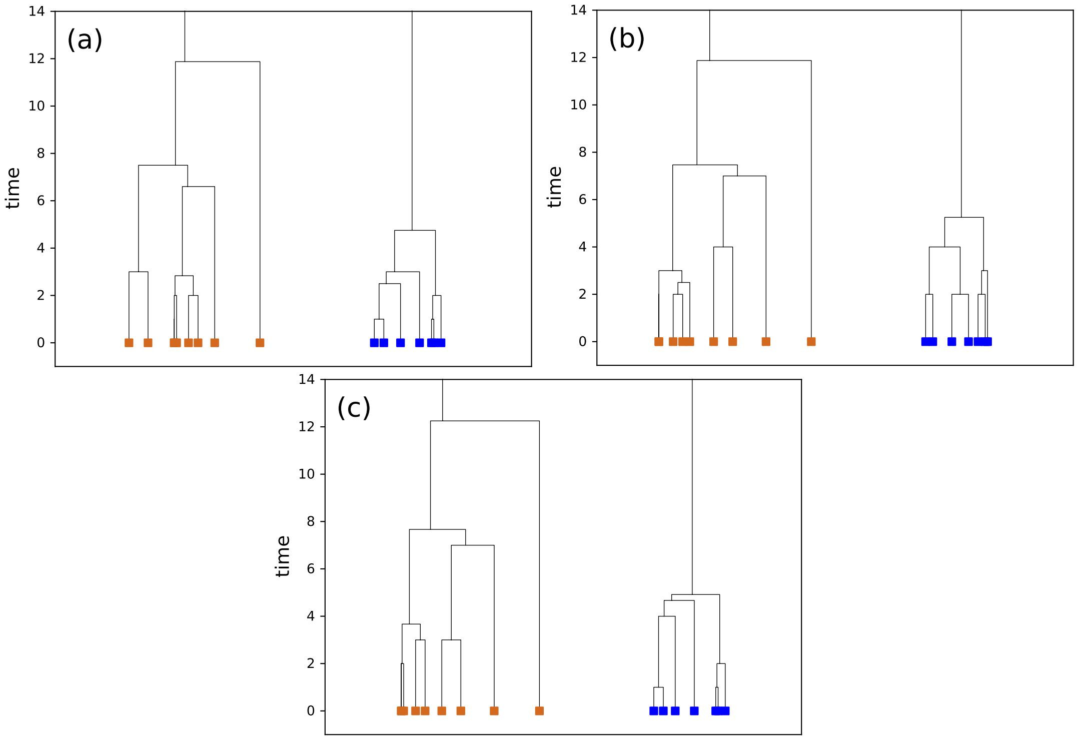


FIGURE A.3: Genealogies of sampled individuals for two species obtained via MRCAT matrices: (a) general; (b) paternal; (c) maternal. Species correspond to the 4th and 6th in Fig.5(b), counting from the right. The number of individuals sampled were 8 and 7 respectively.

A.3 The γ-statistic

Figure A.4 shows the internode distance g_k for a schematic phylogenetic tree. This distance is used to calculate the γ-statistic, as defined on Eq. 10 in the main text. From this figure is straightforward to see that, for $k < NS$, $T(k)$ is the sum of the size of clades up to an speciation event k. If $k = NS$, $T(k)$ is the sum of the size of all clades.

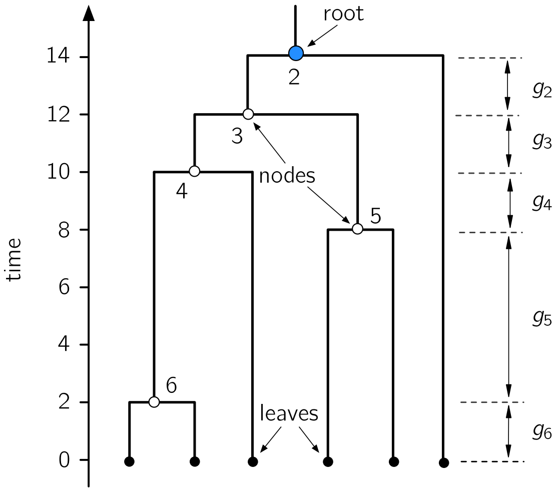


FIGURE A.4: Example of phylogenetic tree with 6 species, represented by filled circles as the leaves of the tree. Open circles denote the nodes, which are numbered from 2 (topmost node, also called the root) to N. Times between speciation events, g_k, are used to calculate the γ-statistic.

Under a pure birth process with constant speciation rate per branch, γ-values follow a standard normal distribution centered on $\gamma = 0$ with unit standard deviation. If $\gamma > 0$ the internal nodes are closer to the tips and if $\gamma < 0$ they are closer to the root, as compared with the case of constant speciation rate.

Appendix B

Supporting Information - Chapter 2

B.1 Extinction rate

The diversification rate is composed by speciation and extinction events. The number of extinction events during the radiation can be computed by following all speciation and extinction events during the evolution (Costa et al., 2018a). We show below that under the conditions of this model, extinction is rare for the large genomes ($B = 1,500$ and $B = 150,000$) and it happens in a constant rate for small genome ($B = 150$). In the Fig. B.1 we show all the speciation and extinction events. We filtered this data to accomplish only the events that we considered true speciation events, by defining two thresholds (i) a time range, τ, and (ii) a population size, η. We filtered the branch lengths of species that thrived for $t < \tau$ generations (here $\tau = 10$ generations). We also considered that species that had a population size $N > \eta$ during the extinction event were also filtered, i.e. we did not considered it as a true extinction event (here $\eta = 10$ individuals). The filtered branches were considered incipient speciation events, which were likely a result of mutations that reversed individuals back to their ancestral species or joined two species in the only one. In the Fig B.2 we show that for the small genome ($B = 150$) the extinctions happen in a constant rate over time ($slope = 0.024$ extinctions/generation, $R^2 = 0.95$).

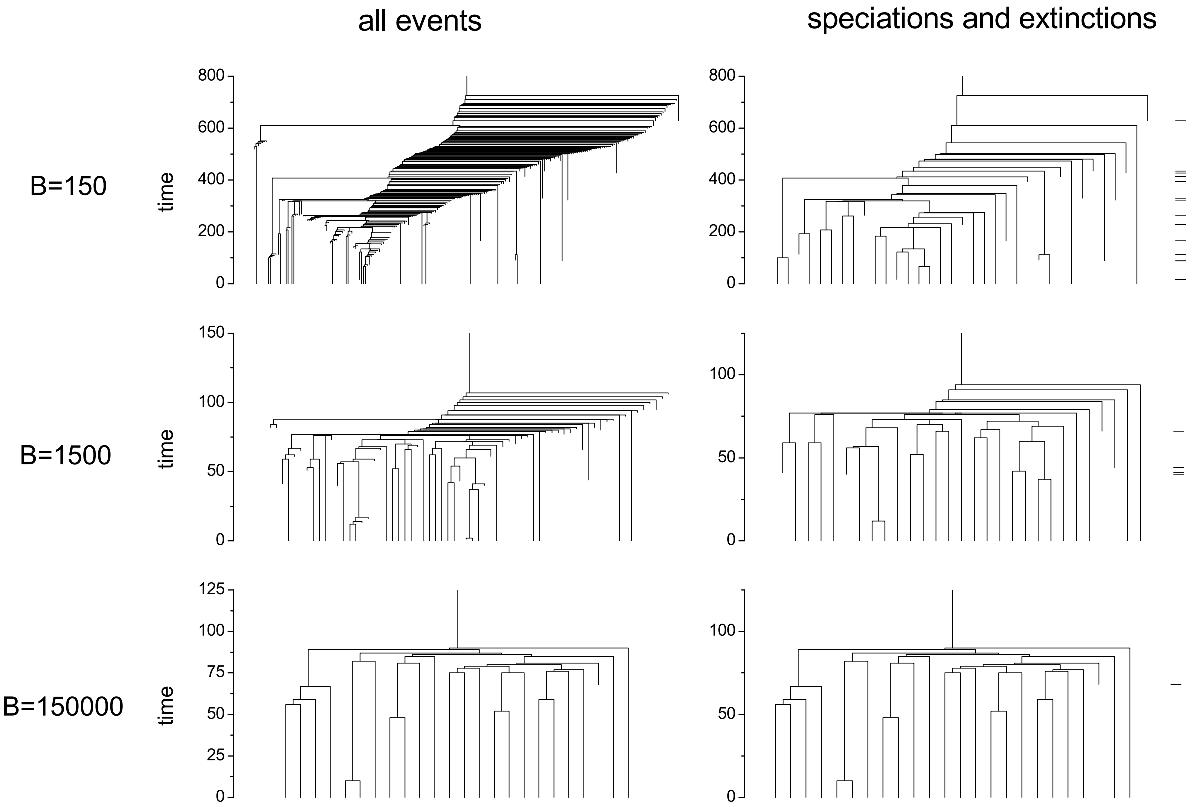


FIGURE B.1: At the left side, we show all the speciation and extinction events for the three lengths of genome. At the right side, the incipient events were filtered, remaining only the true extinction and and true speciation events. We filtered every branch that thrived for less than 10 generations, but we kept those with a population size greater than 10 individuals. The little traces at the most right side mark the true extinctions.

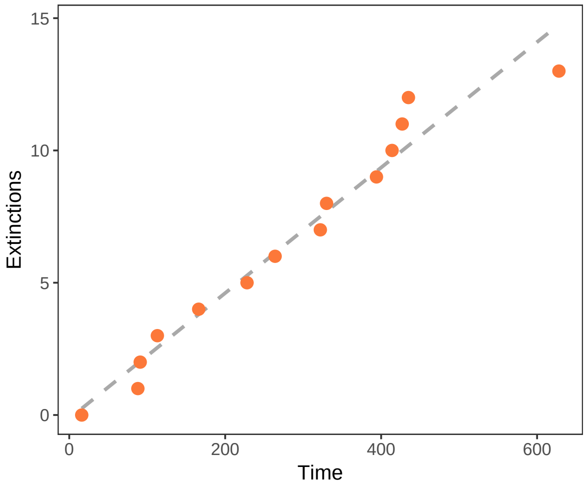


FIGURE B.2: For the case of more true extinction events ($B = 150$), we plot the number of extinctions over time. The linear regression gives: $Extinctions = -0.13 + 0.024 \cdot Time$. Therefore the rate of extinctions per unit of time is constant and equal to 0.024 extinctions/generation ($R^2 = 0.95$).

B.2 Time to divergence as a function of B and G/B

All simulations in the paper used the fixed value $G/B = 0.05$ for all values of S and B. Here we present analytical results (for $B \to \infty$) and simulations (for finite B) for the time to divergence (the waiting time to speciation) as a function of G/B for fixed B and as a function of B for fixed G/B.

For infinitely large genomes the system can be described in terms of the genetic similarity between individuals i and j, $q^{ij} = 1 - 2d^{ij}$ where d^{ij} is their relative genetic distance (see Eq. 1 in the main text). In Derrida-Higgs it has been shown that the distribution of similarities is very narrow for short times and its center (q) follows the dynamic equation (Yule, 1925; De Aguiar, 2017):

$$q' = \left(1 - \frac{4}{M}\right) q e^{-4\mu} + \frac{4}{M}\frac{e^{-4\mu}}{4}(1 + 3q) \tag{B.1}$$

This equation may also be approximated by the continuous version

$$\dot{q} = \frac{1}{M}\left[1 - \frac{q}{\bar{q}}\right]. \tag{B.2}$$

whose solution is

$$q(t) = \bar{q} + (1 - \bar{q})e^{-t/M\bar{q}} \tag{B.3}$$

where $\bar{q} \equiv \frac{1}{1+4\mu M}$. Since mating is restricted by similarity, individuals can mate only if $q > q_{min} = 1 - 2G/B$, then the time τ it takes for the average similarity to go from 1 to q_{min} can be obtained by setting $q(\tau) \to q_{min}$:

$$\tau = M\bar{q}\ln\left(\frac{1 - \bar{q}}{q_{min} - \bar{q}}\right) \tag{B.4}$$

where τ is the time it takes for speciation to start in the populations (hereafter, *divergence time*). Setting $q_{min} \to \bar{q}$ gives the maximum value of G/B under which speciation is possible:

$$[G/B]_{max} = \frac{4\mu M}{2(1 + 4\mu M)} \tag{B.5}$$

In our simulations we fixed $4\mu M = 1$ (M=1000,$\mu = 0.00025$) so that $[G/B]_{max} = 1/4$. Fig. B.3 shows the theoretical curve for $B \to \infty$ (continuous grey curve) and the result of simulations for $B = 150$ (orange). The vertical dotted line shows $G/B = 1/4$

where the time diverges and beyond which speciation does not happen anymore. Notice that the behaviour for $B = 150$ is very similar, although the divergence times are larger – the curve is shifted up. Fixing the ratio G/B in 0.05, we observe that the larger the genome size, the shorter the divergence time, which is approximately 113 units of time for for $B = \infty$ (see Fig. B.4). Once the distribution of similarities reaches q_{min}, the populations breaks into species and the dynamic equation above, Eq. B.2, cannot describe the behaviour of speciation events anymore. At the moment no theory is available to describe the equilibration time for the number of species.

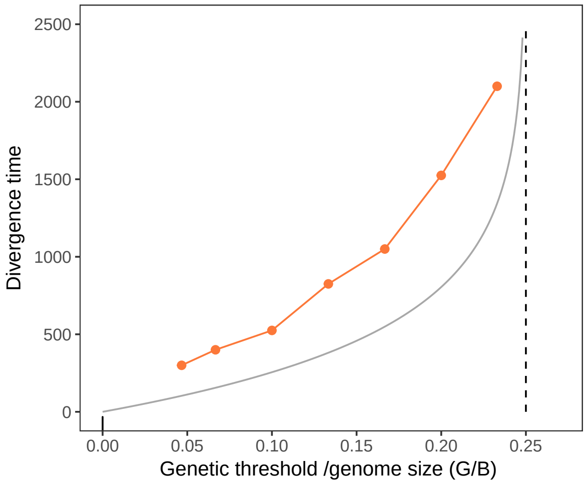


FIGURE B.3: Divergence times, or time to speciation – τ, as a function of G/B for $B = 150$ (orange) and $B = \infty$ (continuous curve in grey). The vertical black dotted line is the limit $G/B = 1/4$ calculated from Eq. B.5.

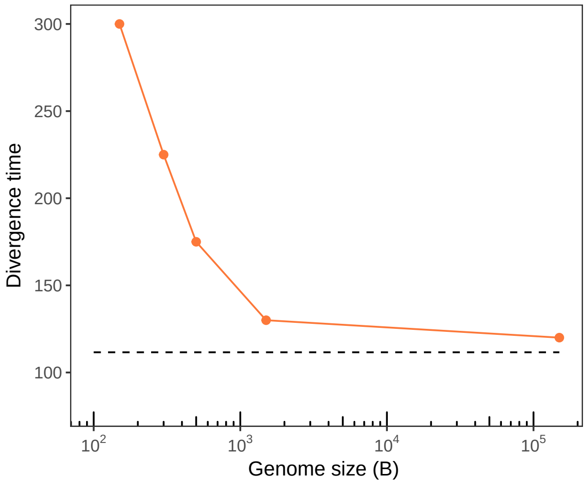


FIGURE B.4: Divergence times, or time to speciation – τ, as a function of B for $G/B = 0.05$ (orange). The dotted black line is the asymptotic limit for $B = \infty$.

B.3 The Asymmetric Yule model

As for the Yule model the distribution of Sackin indexes is always symmetric relative to the mean, in order to assess the degree of imbalance in simulated or empirical phylogenetic trees, we can define a one-parameter family of asymmetric Yule models. We introduce an asymmetry on the probability of bifurcation for each of the branches of the tree at a given time. Starting with the root node and its two branches, the probability of bifurcation of the left branch will be given by $\delta/(1+\delta)$, while the probability of bifurcation of the right branch will be $1/(1+\delta)$ (we assume $0 < \delta < 1$). On the next step, the probability of bifurcation of each of the three branches, from left to right, will be, respectively, $\delta^2/(1+\delta+\delta^2)$, $\delta/(1+\delta+\delta^2)$ and $1/(1+\delta+\delta^2)$. The process continues until the tree reaches N leaves. Both the average Sackin index and the Sackin index probability distribution can be obtained for the asymmetric Yule model.

While in the standard Yule model all the trees with N leaves have the same probability of occurring, in the asymmetric model we propose, the probability of occurrence of a given tree is dependent on the branching history. Starting with the root node and two branches, the branching history will be given by a finite sequence $\{h_n\}$, where h_n gives the index of the branch that will be divided in the n-th step of the branching process. We choose to exclude h_0, which can be regarded as the division of the root node. The sequence then start with h_1 and goes until h_{N-2} for a tree with N leaves. Branches indexes go from 0, on the rightmost branch, until n, on the leftmost. Fig. B.5(a) shows an example of the indexation of branches and the corresponding history $\{h_n\}$ of the branching process.

In the asymmetric model, the probability that a given branch in the n-th step is chosen to divide is given by δ^i/C_n, where $0 < \delta < 1$ is the model parameter, i is the index of the branch and $C_n = \sum_{k=0}^{n} \delta^k$ is the normalization constant for the n-th step. With the branching history $\{h_n\}$ of a tree with N leaves, we can easily calculate the probability of its occurrence in the asymmetric model as:

$$P_{\{h_n\}} = \prod_{n=1}^{N-2} \frac{\delta^{h_n}}{C_n}. \tag{B.6}$$

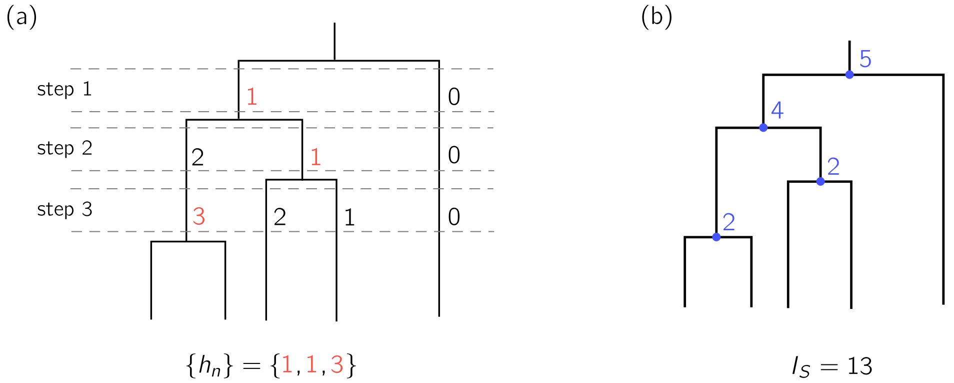


FIGURE B.5: (a) Illustration of a tree with $N = 5$ leaves and 3 steps in the branching process. The selected nodes to divide in each are bold red and the sequence of selected nodes constitutes the branching history of this tree. (b) Loads for each node in the tree shown in panel (a). The sum of the loads for all the nodes is equal to the Sackin index.

For the tree shown in Fig. B.5(a), the probability of occurrence is then given by:

$$P_{\{1,1,3\}} = \left(\frac{\delta}{1+\delta}\right)\left(\frac{\delta}{1+\delta+\delta^2}\right)\left(\frac{\delta^3}{1+\delta+\delta^2+\delta^3}\right). \tag{B.7}$$

To obtain the distribution of Sackin indexes in the asymmetric model, we first note that there is an alternative way of obtaining the Sackin index for a given tree. Instead of counting the numbers of nodes until the root node from each leaf and then summing to obtain the Sackin index for that tree, we can obtain the *load* of each node, where we define the *load* as the number of leaves below each node. The Sackin index will then be the sum of loads for all the nodes, including the root node. The loads for the tree shown in Fig. B.5(a), with the corresponding Sackin index, I_S, is shown in Fig. B.5(b).

The probability distribution can be obtained if for every possible tree we are able to associate a Sackin index and a probability of occurrence for that tree. The probability of occurrence of a tree with a given branching history $\{h_n\}$ is given by Eq. B.6. The corresponding Sackin index can be obtained by a backwards reconstruction of the tree with the branching history sequence. The idea is to start with the final N leaves and cluster them at the nodes, as we follow the history backwards, in order to obtain the nodes' loads. Finally, we sum the obtained loads and calculate the Sackin index.

Figure B.6 ilustrates the iterative process applied to the tree shown in Fig. B.5(a).

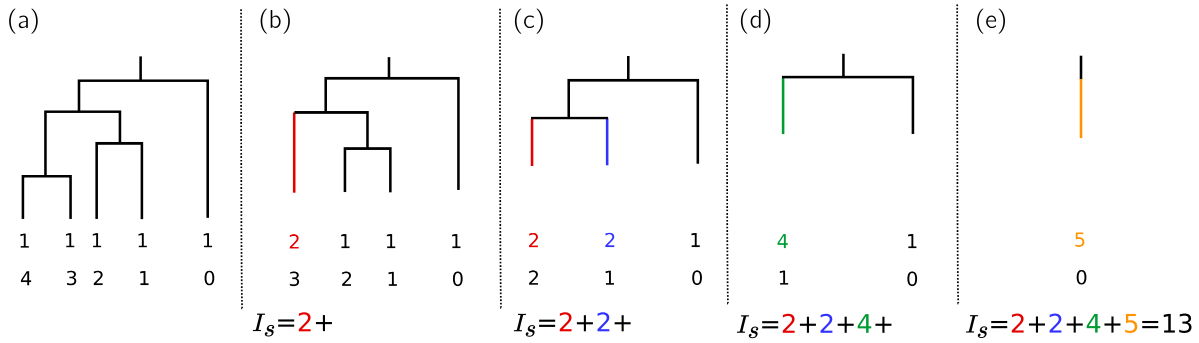


FIGURE B.6: Representation of the iteration process with 4 steps ($k = 4$) to obtain the Sackin index for the tree presented in Fig. B.5(a). The loads are obtained clustering leaves after following the history sequence backwards. The first row of numbers below each tree refers to elements (values) of vector $\mathbf{v}^{(\mathbf{k})}$, whereas the second row refers to the indexes (positions) of the same vector. The Sackin index for the tree is then given by the sum of the loads for all the nodes.

Generally, for a N-leaves tree, we start with a vector $\mathbf{v}^{(\mathbf{0})}$, indexed from 0 to $N-1$, with each element equal to 1 (Fig. B.6(a)). The last entry of the history sequence, h_{N-2}, indicates which elements of the vector $\mathbf{v}^{(0)}$ have to be clustered in the first iteration. These will be the elements h_{N-2} and $h_{N-2}+1$. The next vector, $\mathbf{v}^{(\mathbf{1})}$, will be indexed from 0 to $N-2$, with the element $\mathbf{v}^{(\mathbf{1})}_{\mathbf{h_{N-2}}} = \mathbf{v}^{(\mathbf{0})}_{\mathbf{h_{N-2}}} + \mathbf{v}^{(\mathbf{0})}_{\mathbf{h_{N-2}+1}}$ (all the other elements remain the same. See Fig. B.6(b)). For the next iteration, the clustered values of the vector $\mathbf{v}^{(\mathbf{1})}$ will be elements h_{N-3} and $h_{N-3}+1$, hence $\mathbf{v}^{(\mathbf{2})}_{\mathbf{h_{N-3}}} = \mathbf{v}^{(\mathbf{1})}_{\mathbf{h_{N-3}}} + \mathbf{v}^{(\mathbf{1})}_{\mathbf{h_{N-3}+1}}$. The iteration process continues until the vector $\mathbf{v}^{(\mathbf{k})}$, obtained after k iterations, has size one, when its only element will be $\mathbf{v}^{(\mathbf{k})}_{\mathbf{0}} = N$, corresponding to the load of the root node (Fig. B.6(c)-(e)). The results for to the clustering of the elements in each iteration step are then summed to obtain the Sackin index of the tree. Fig. B.6 ilustrates the iterative process applied to the tree shown in Fig. B.5(a).

The distribution of Sackin indexes in the asymmetric model can be obtained calculating the indexes for each possible history sequence and summing the probabilities corresponding to the same index. Although relatively simple to implement, this algorithm can be very time consuming for increasing tree sizes, as for a tree with N leaves, the number of possible history sequences is $(N-1)!$.

Figure B.7 shows a qualitative fit of the Sackin index distributions by the asymmetric model. The value of δ in each case (corresponding to the Sackin index distributions in Figures 2 and 3 in the main text) gives an idea of how asymmetric the trees in the distribution are.

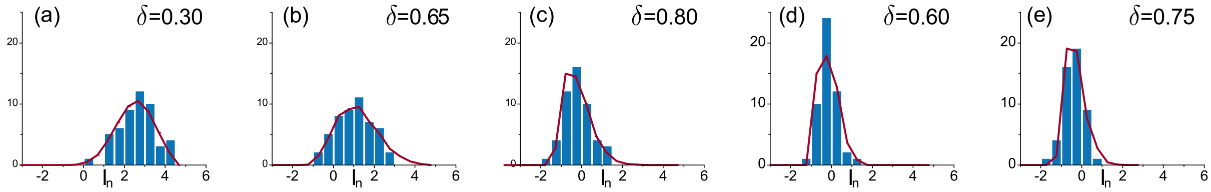


FIGURE B.7: Sackin index distributions for (a) $S = 5$, $B = 150$; (b) $S = 5$, $B = 1500$; (c) $S = 5$, $B = 150000$; (d) $S = 20$, $B = 150000$; (e) $S = 40$, $B = 150000$. The distributions were computed for 50 realizations of the process. Red curves show the distribution for the asymmetric Yule model computed with the average number of species in each set of simulations: (a) $\delta = 0.30$, $N = 20$; (b) $\delta = 0.65$, $N = 27$; (c) $\delta = 0.8$, $N = 27$; (d) $\delta = 0.6$, $N = 11$; (e) $\delta = 0.75$, $N = 13$.

B.4 γ-statistic

The γ statistic is defined as

$$\gamma = \frac{1}{D}\left[\frac{1}{N-2}\sum_{k=2}^{N-1} T(k) - \frac{T(N)}{2}\right] \tag{B.8}$$

where

$$V = \frac{T(N)}{\sqrt{12(N-2)}}, \qquad T(k) = \sum_{j=2}^{k} j g_j \tag{B.9}$$

and g_k is the time between two ramifications events (with the root corresponding to $k = 2$). Each factor in Eq. B.8 can be interpreted as follow:

- $T(k)$ is the sum of the branches sizes up to time $k + 1$.
- $\frac{1}{N-2}\sum_{k=2}^{N-1} T(k)$ is the average of T from the root to 1 step before the end.
- $T(N)$ is the total time length of all branches.

B.4.1 Pure-birth model

In the birth-death process, each branch goes extinct in a time interval Δt with probability $d(t)\Delta t$ and speciate with probability $b(t)\Delta t$. We will consider as a reference the pure birth process, where $d(t) = 0$ and $b(t) = b$. For the sake of completeness we will show that definition (B.8) implies $\langle \gamma \rangle = 0$ and $\langle \gamma^2 \rangle \approx 1$ for the pure birth process.

Consider a tree with k branches at time $t = 0$ and let $P(t)$ be the probability that a new branch happens in the time interval between 0 and t. Then the probability $\rho(t)$ of a branching event in the tree in the time interval $(t, t + dt)$ is $\rho(t) = dP/dt$ where:

$$\frac{dP}{dt} = kb(1 - P(t)) \tag{B.10}$$

and $(1 - P(t))$ is the probability of *not* branching, for each of the k branches, until time t. The solution of Eq. B.10 is

$$P(t) = 1 - \mathrm{e}^{-kbt} \tag{B.11}$$

and then

$$\rho(t) = kb\mathrm{e}^{-kbt} \tag{B.12}$$

With Eq. B.12, we can calculate the average size of branch g_k, and other high order moments:

$$\begin{aligned} \langle g_k \rangle &= \int_0^\infty t\rho(t)dt = \frac{1}{kb} \\ \left\langle g_k^2 \right\rangle &= \int_0^\infty t^2\rho(t)dt = \frac{2}{k^2b^2} \\ \Delta g_k^2 &= \left\langle g_k^2 \right\rangle - \langle g_k \rangle^2 = \frac{1}{k^2b^2} \end{aligned} \tag{B.13}$$

We now define

$$U \equiv \frac{1}{N-2}\sum_{k=2}^{N-1} T(k) \quad \text{and} \quad \Gamma \equiv U - T(N)/2 = \gamma V \tag{B.14}$$

Calculating the average value of these quantities we have:

$$\langle \Gamma \rangle = \langle U \rangle - \langle T(N) \rangle /2, \tag{B.15}$$

$$\langle U \rangle = \frac{1}{N-2}\sum_{k=2}^{N-1}\sum_{j=2}^{k} j \langle g_j \rangle = \frac{1}{2b}(N-1) \tag{B.16}$$

and

$$\langle T(N)\rangle = \sum_{j=2}^{N} j\,\langle g_j\rangle = \sum_{j=2}^{N} \frac{1}{b} = \frac{N-1}{b} \tag{B.17}$$

Therefore

$$\langle \Gamma\rangle = \frac{N-1}{2b} - \frac{N-1}{2b} = 0 \tag{B.18}$$

For the second moment of Γ we need

$$\left\langle \Gamma^2\right\rangle = \left\langle U^2 - UT(N) + \frac{1}{4}T^2(N)\right\rangle \tag{B.19}$$

To calculate U^2 we first write

$$\begin{aligned} U &= \frac{1}{N-2}\left[T(2)+T(3)+\ldots+T(N-1)\right] \\ &= \frac{1}{N-2}\left[(2g_2)+(2g_2+3g_3)+(2g_2+3g_3+4g_4)+\ldots\right. \\ &= \frac{1}{N-2}\left[(2g_2)(N-2)+(3g_2)(N-3)+\ldots+(N-1)g_{N-1}\right] \\ &= \frac{1}{N-2}\sum_{k=2}^{N-1} kg_k(N-k), \end{aligned} \tag{B.20}$$

therefore,

$$U^2 = \frac{1}{(N-2)^2}\sum_{k=2}^{N-1}\sum_{k'=2}^{N-1} kk'(N-k)(N-k')g_k g_{k'} \tag{B.21}$$

We will need to calculate the correlation of g_k

$$\langle g_k g_{k'}\rangle = \begin{cases} \dfrac{1}{b^2kk'} & \text{if} \quad k \neq k' \\ \dfrac{1}{b^2k^2} & \text{if} \quad k = k', \end{cases}$$

if $\langle g_k g_{k'}\rangle = \langle g_k\rangle\langle g_{k'}\rangle$ for all k, k', then $\left\langle U^2\right\rangle = \langle U\rangle^2$, but when $k = k'$ there is a factor 2. Hence,

$$\begin{aligned} \left\langle U^2\right\rangle &= \langle U\rangle^2 + \frac{1}{(N-2)^2}\sum_{k=2}^{N-1}\frac{1}{b^2}(N-k)^2 \\ &= \langle U\rangle^2 + \frac{(N-1)(2N-3)}{6b^2(N-2)} \end{aligned} \tag{B.22}$$

The cross term is (we will use $T = T(N)$ for simplicity)

$$\begin{aligned} UT &= \frac{1}{N-2}\left[\sum_{k=2}^{N-1} k g_k (N-k)\right]\left[\sum_{k'=2}^{N} k' g_{k'}\right] \\ &= \frac{1}{N-2}\sum_{k=2}^{N-1}\sum_{k'=2}^{N-1} k k'(N-k) g_k g_{k'} + \frac{N g_N}{N-2}\sum_{k=2}^{N-1} k g_k (N-k), \end{aligned} \tag{B.23}$$

once again $\langle UT \rangle \neq \langle U \rangle \langle T \rangle$ because of the term $k = k'$ in the first line that has a factor 2, then:

$$\langle UT \rangle = \langle U \rangle \langle T \rangle + \frac{1}{N-2}\sum_{k=2}^{N-1} \frac{1}{b^2}(N-k) = \langle U \rangle \langle T \rangle + \frac{N-1}{b^2}. \tag{B.24}$$

Lastly, we need the average of T^2:

$$T^2 = \sum_{k,k'=2}^{N} k k' g_k g_{k'} \quad \Rightarrow \quad \left\langle T^2 \right\rangle = \langle T \rangle^2 + \sum_{k=2}^{N} \frac{1}{b^2} = \langle T \rangle^2 + \frac{N-1}{b^2} \tag{B.25}$$

With these results, the second ordem moment for Γ (Eq. B.19), can be written as

$$\begin{aligned} \left\langle \Gamma^2 \right\rangle &= \left\langle U^2 \right\rangle + \frac{(N-1)(2N-3)}{6b^2(N-2)} - \langle U \rangle \langle T \rangle - \frac{(N-1)}{2b^2} + \frac{1}{4}\left\langle T^2 \right\rangle + \frac{(N-1)}{4b^2} \\ &= \left(\langle U \rangle - \frac{1}{2}\langle T \rangle\right)^2 + \frac{N-1}{12b^2(N-2)}\left[2(2N-3) - 6(N-2) + 3(N-2)\right] \\ &= \frac{N-1}{12b^2(N-2)}\left[4N - 6 - 3N + 6\right] \\ &= \frac{N(N-1)}{12b^2(N-2)}. \end{aligned} \tag{B.26}$$

Using

$$\left\langle T^2 \right\rangle = \langle T \rangle^2 + \frac{N-1}{b^2} = \frac{(N-1)^2}{b^2} + \frac{N-1}{b^2} = \frac{N(N-1)}{b^2} \tag{B.27}$$

we have

$$\left\langle \Gamma^2 \right\rangle = \frac{\langle T^2 \rangle}{12(N-2)}. \tag{B.28}$$

In order to normalize γ to have unit variance we should choose $V = \sqrt{\langle \Gamma^2 \rangle}$. However, because γ should be defined for each individual tree, averages such as $\left\langle T^2 \right\rangle$ should be avoided. Therefore, assuming that T and γ are approximately independent variables, so that $\left\langle T^2\gamma^2\right\rangle \approx \left\langle \gamma^2\right\rangle \left\langle T^2\right\rangle$ we set

$$V = \frac{T}{\sqrt{12(N-2)}} \tag{B.29}$$

to get $\left\langle \gamma^2\right\rangle \approx 1$ for the null model.

B.4.2 γ-statistic for Yule Model

Measures of tree imbalance are usually made by comparing an actual shape with the outcome of a null model (Kirkpatrick and Slatkin, 1993; Mooers and Heard, 1997). The Yule model (Yule, 1925) assumes that bifurcating events occur at constant time intervals, with equal probability of bifurcating for each of the branches of the tree at a given time. The construction of trees under the Yule model is as follows: start with the root node and its two branches. One of the branches is then randomly chosen to bifurcate with uniform probability 1/2, generating a tree with three branches. One of the three branches is then randomly chosen to bifurcate, with probability 1/3 generating a tree with four branches and so on.

As the bifurcation happens at constant time intervals in the Yule model, we set $g_j = g$. Then

$$T(k) = \sum_{j=2}^{k} j g_j = g \sum_{j=2}^{k} j = \frac{g}{2}\left(k^2 + k - 2\right) \tag{B.30}$$

Also

$$\sum_{k=2}^{N-1} T(k) = \frac{g}{2} \sum_{k=2}^{N-1} k^2 + k - 2 = \frac{g}{6}\left(N^3 - 7N + 6\right) \tag{B.31}$$

Replacing the sums in Eq. B.8, we obtain

$$\gamma = \sqrt{12(N-2)} \left[\frac{g/6\ (N^3 - 7N + 6)}{g/2\ (N^2 + N - 2)\ (N-2)} - \frac{1}{2}\right] = \frac{-N\sqrt{12(N-2)}}{6(N+2)}. \tag{B.32}$$

Eq. B.32 is depicted in Fig. B.8 (green triangles).

B.5 The α value as an acceleration of speciation

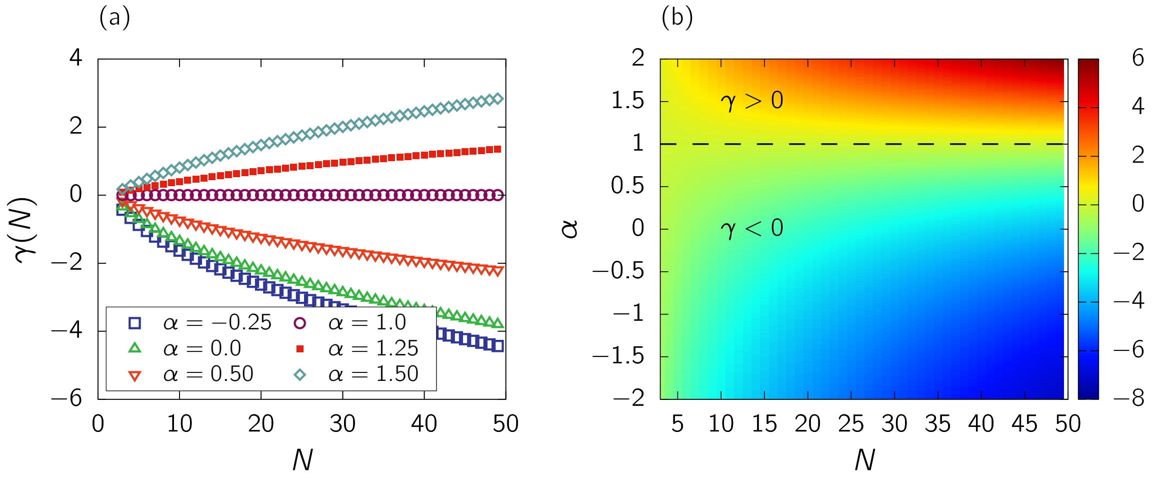


FIGURE B.8: (a) γ as function of N for fixed values of α. (b) Color map for $\gamma(\alpha, N)$. Although N only take discrete values, we interpolate the color map to get a better visualisation.

Let us consider a more general case where

$$g_k = \frac{k^{-\alpha}}{b}. \tag{B.33}$$

Note that $\alpha = 0$ corresponds to the Yule model, while $\alpha = 1$ results in $g_k = 1/(kb)$ (the same as the average g_k for the pure-birth model), so we have:

$$\begin{aligned} \alpha &= 0 \quad \text{(Yule model)} \quad \Rightarrow \quad \gamma < 0 \quad \text{(Eq. } B.32) \\ \alpha &= 1 \qquad\qquad\qquad\quad\;\; \Rightarrow \quad \gamma = 0. \end{aligned} \tag{B.34}$$

For $\alpha = 0$, the internode distances are equal, while $\alpha < 0$ results in a slowing down of speciation rates. Finally, if $\alpha > 0$, the bifurcations happen more frequently at the base of the tree.

The sums involved in the calculation of γ can be easily done numerically. Fig. B.8(b) shows a colormap for γ as an function of N and α. The dotted line separates positive and negatives values of γ. For a fixed α, γ is a monotonic function of N, as showed in Fig. B.8(a). The value $\alpha = 0.0$ in this panel corresponds the Eq. B.32.

The total time t_s of the process is proportional to the length ℓ_s of the tree. Then

$$t_s \sim \ell_s = \sum_{k=2}^{N} \frac{k^{-\alpha}}{b} \Rightarrow t_s = A \sum_{k=2}^{N} \frac{k^{-\alpha}}{b} \tag{B.35}$$

where A is a constant related to velocity of speciation. Let's suppose that in the time t_s^* the number of species is the same, i.e, $N = N^*$, independent of the α value, so the constant A should be adjusted as:

$$A = \frac{t_s^*}{\sum_{k=2}^{N^*} \frac{k^{-\alpha}}{b}}. \tag{B.36}$$

After t_s^* the process of speciation stopped and the curves $N \times t_s$ attained the saturation value N^*, as illustrated in Fig. B.9.

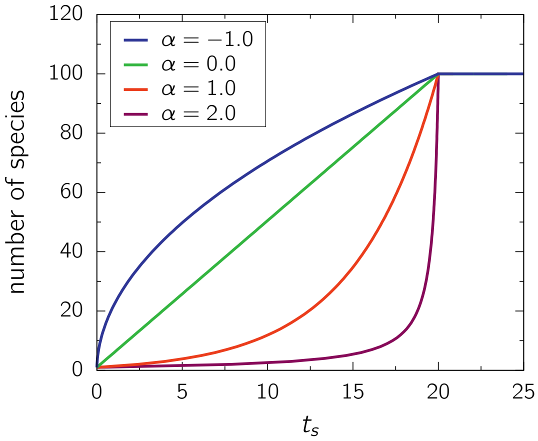


FIGURE B.9: Time evolution of the number of species for different values of the metric α-value.

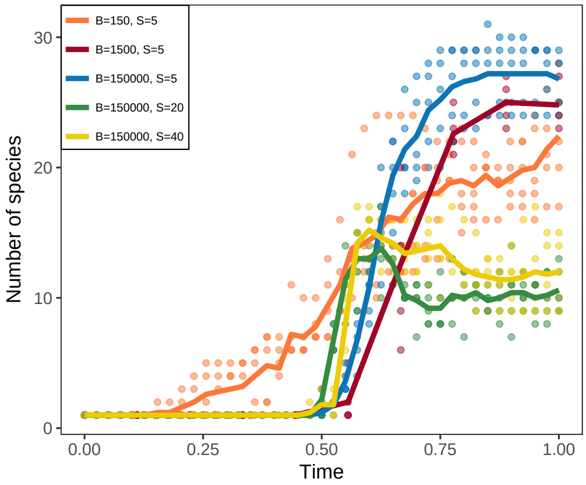
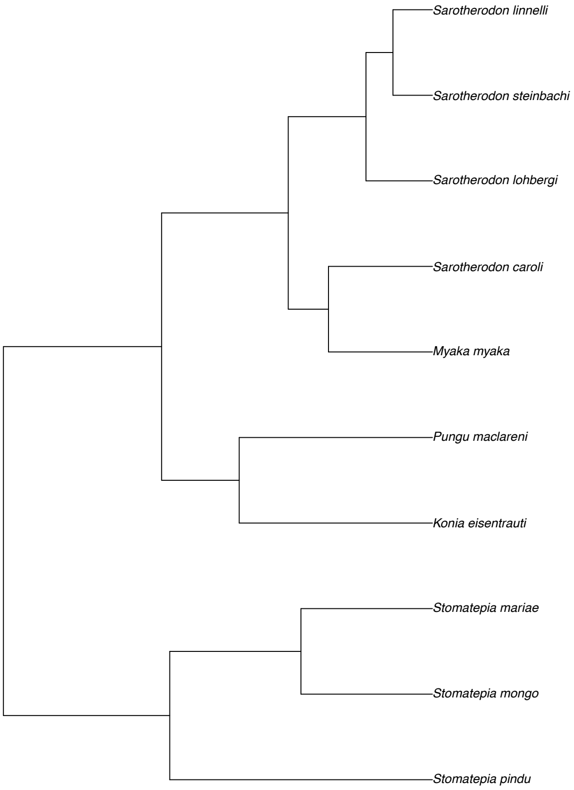


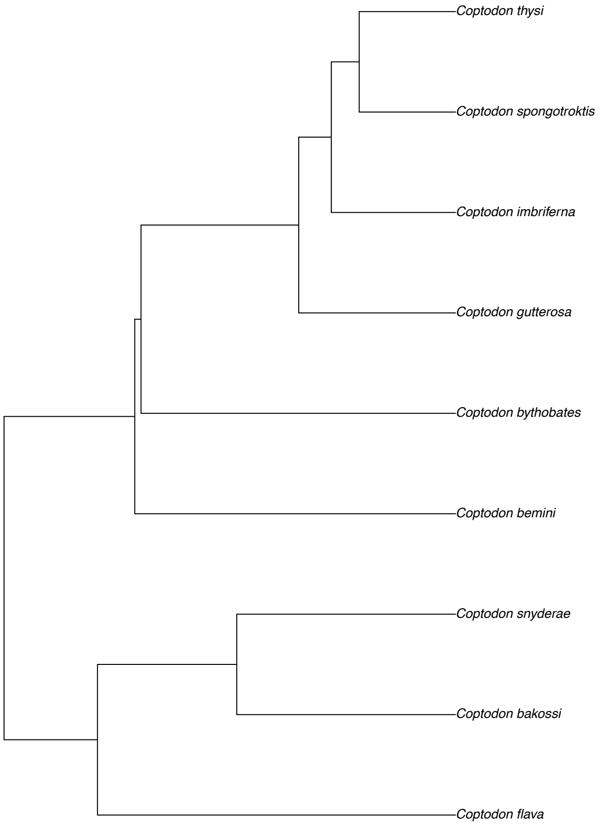
FIGURE B.10: Time evolution of the number of species for different combinations of mating radius, S and genome size, B of the individual-based model. The data plotted here is the same shown in Figure 2 in the main text. In order to allow for a better comparison of the pattern of species accumulation in the course diversification, here the time has been rescaled from the start of the simulation to the time to equilibration of number of species in each case (225 generations for $B = 150,000$, 250 generations for $B = 1,500$ and 1000 generations for $B = 150$). As in Figure 1, points show results of 5 realizations for each value of B at each time step, darker points depict overlapping values. Solid lines show the average values. α.

B.6 Phylogenetic trees estimated from empirical data

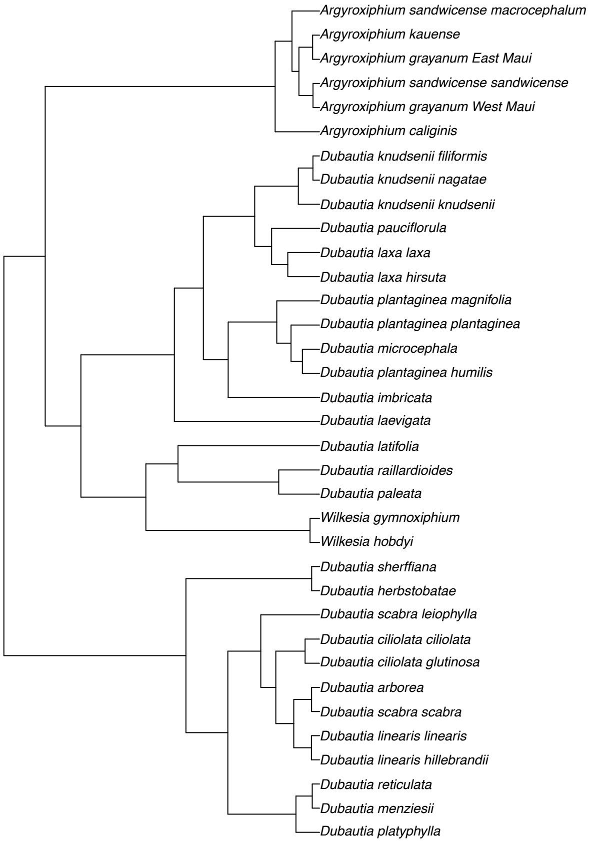
Fig. B.11(a)-(p) show the trees corresponding to empirical data used to generate points 1 to 16 in Figure 5 in the main text.



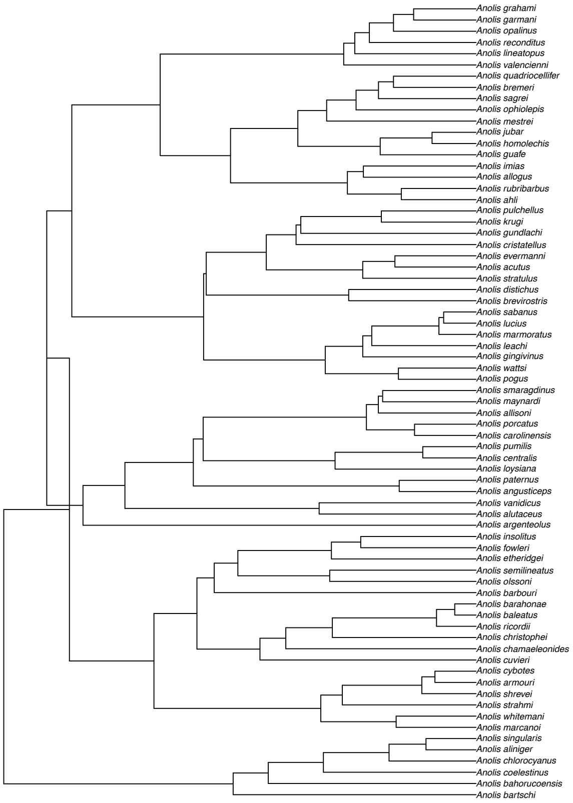
(a) Barombi Mbo crater lake cichlids



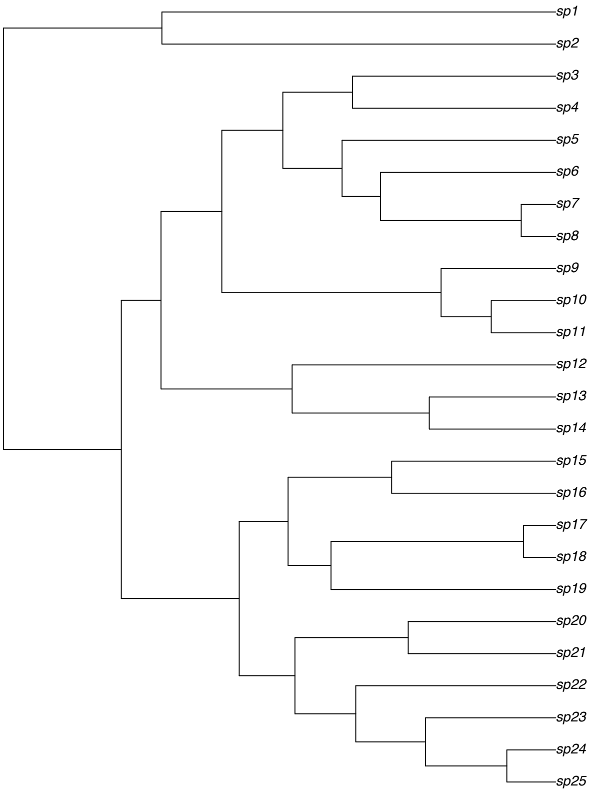
(b) Bermin crater lake cichlids


Argyroxiphium sandwicense macrocephalum
Argyroxiphium kauense
Argyroxiphium grayanum East Maui
Argyroxiphium sandwicense sandwicense
Argyroxiphium grayanum West Maui
Argyroxiphium caliginis
Dubautia knudsenii filiformis
Dubautia knudsenii nagatae
Dubautia knudsenii knudsenii
Dubautia pauciflorula
Dubautia laxa laxa
Dubautia laxa hirsuta
Dubautia plantaginea magnifolia
Dubautia plantaginea plantaginea
Dubautia microcephala
Dubautia plantaginea humilis
Dubautia imbricata
Dubautia laevigata
Dubautia latifolia
Dubautia raillardioides
Dubautia paleata
Wilkesia gymnoxiphium
Wilkesia hobdyi
Dubautia sherffiana
Dubautia herbstobatae
Dubautia scabra leiophylla
Dubautia ciliolata ciliolata
Dubautia ciliolata glutinosa
Dubautia arborea
Dubautia scabra scabra
Dubautia linearis linearis
Dubautia linearis hillebrandii
Dubautia reticulata
Dubautia menziesii
Dubautia platyphylla

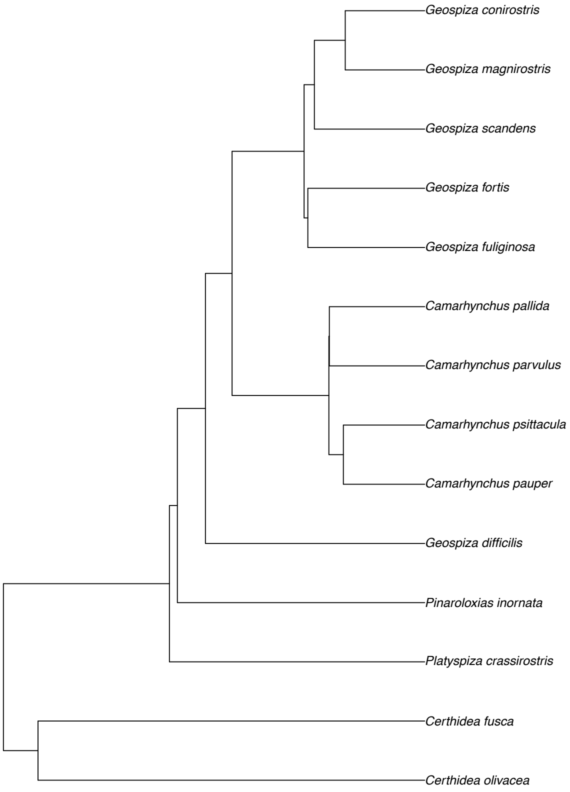
(c) Hawaiian silversword alliance



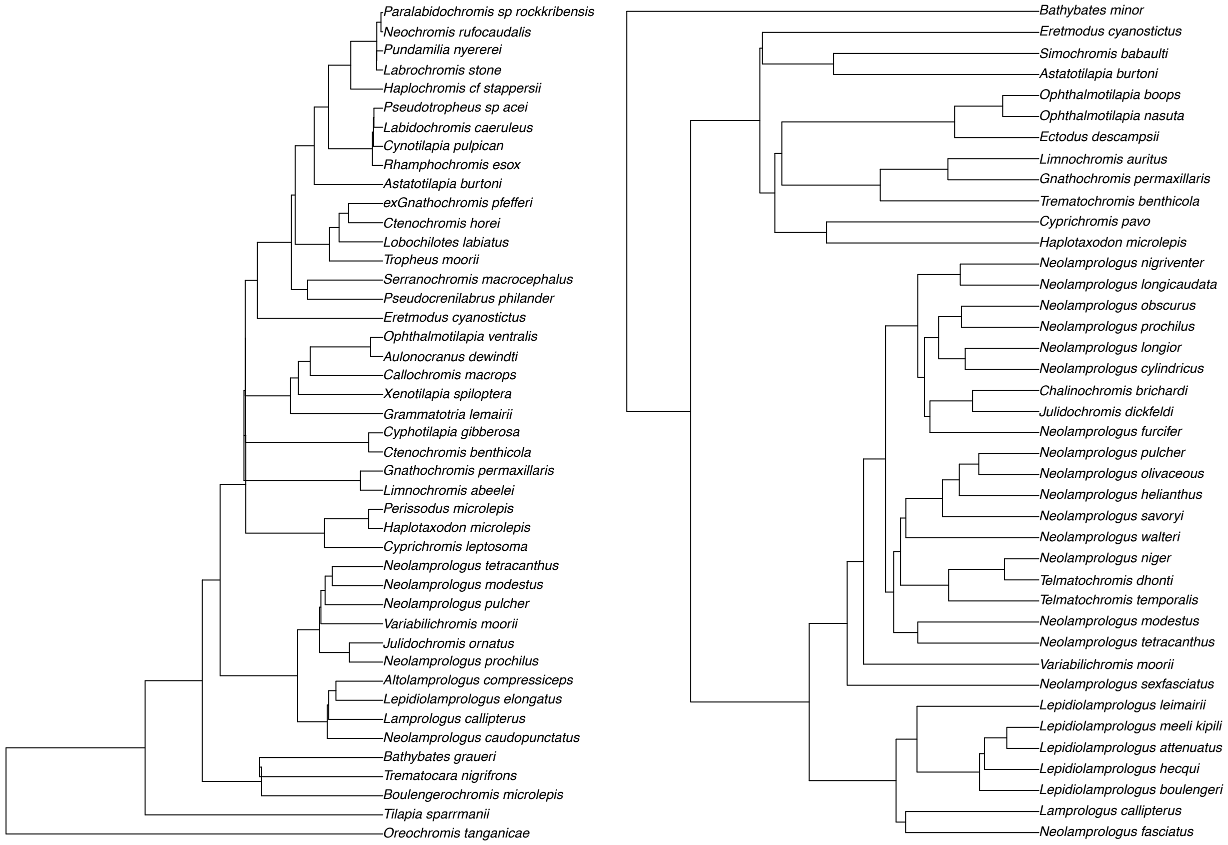
(d) Caribbean Anole lizards



(e) *Tetragnatha* spiders

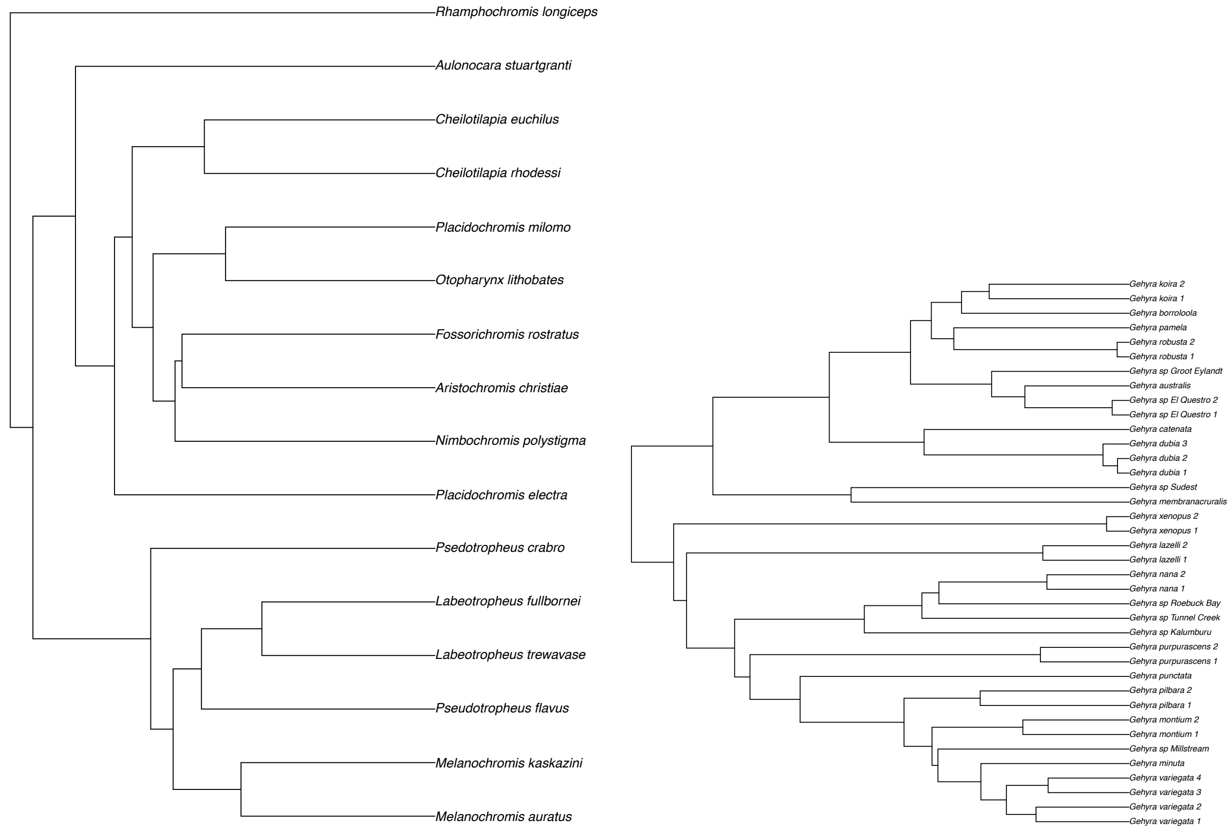


(f) Darwin's finches



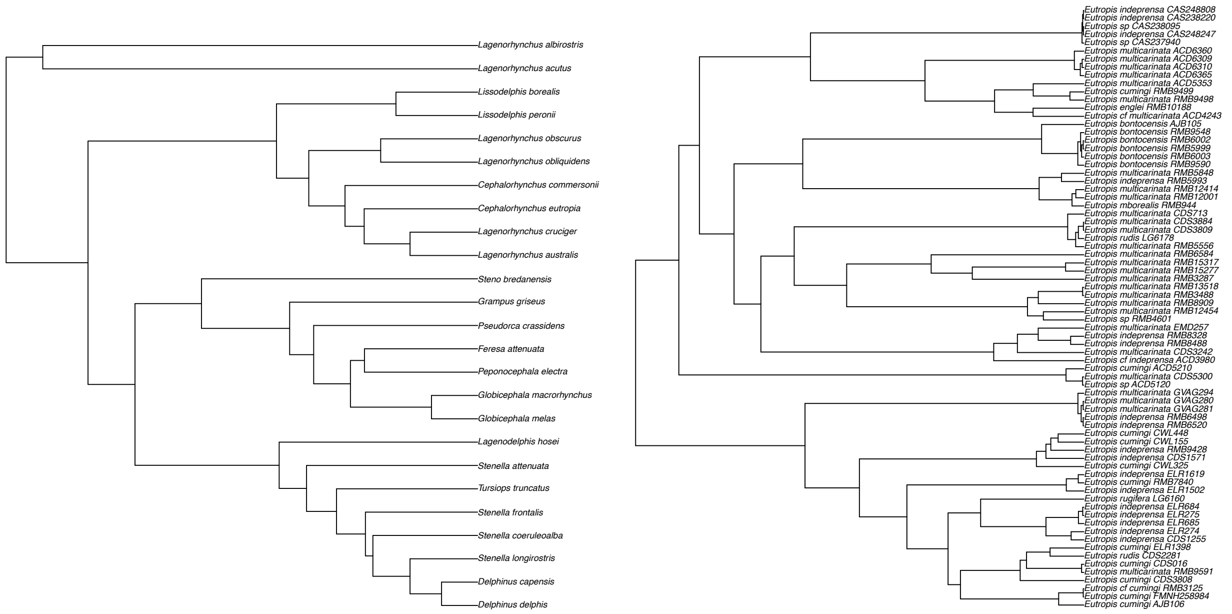
(g) Tanganyika cichlids - 1

(h) Tanganyika cichlids - 2



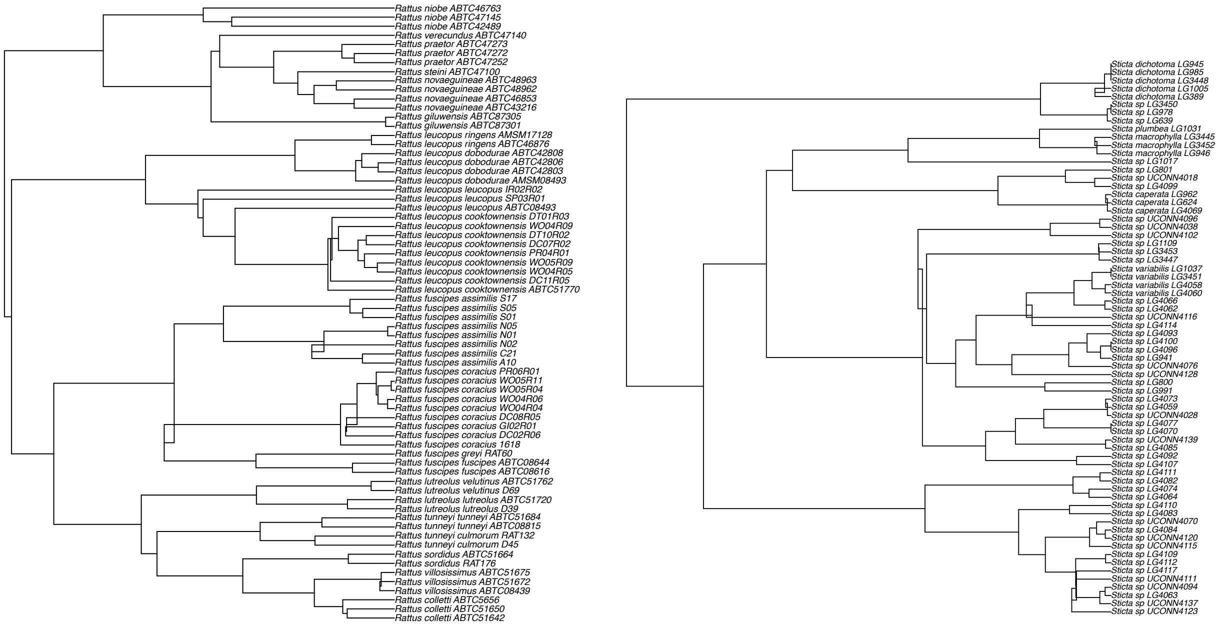
(i) Malawi cichlids

(j) Australian *Gehyra* geckos


Lagenorhynchus albirostris
Lagenorhynchus acutus
Lissodelphis borealis
Lissodelphis peronii
Lagenorhynchus obscurus
Lagenorhynchus obliquidens
Cephalorhynchus commersonii
Cephalorhynchus eutropia
Lagenorhynchus cruciger
Lagenorhynchus australis
Steno bredanensis
Grampus griseus
Pseudorca crassidens
Feresa attenuata
Peponocephala electra
Globicephala macrorhynchus
Globicephala melas
Lagenodelphis hosei
Stenella attenuata
Tursiops truncatus
Stenella frontalis
Stenella coeruleoalba
Stenella longirostris
Delphinus capensis
Delphinus delphis
Eutropis indeprensa CAS248808
Eutropis indeprensa CAS238220
Eutropis sp CAS238095
Eutropis indeprensa CAS248247
Eutropis sp CAS237940
Eutropis multicarinata ACD6360
Eutropis multicarinata ACD6309
Eutropis multicarinata ACD6310
Eutropis multicarinata ACD6365
Eutropis multicarinata ACD5353
Eutropis cumingi RMB9499
Eutropis multicarinata RMB9498
Eutropis englei RMB10188
Eutropis cf multicarinata ACD4243
Eutropis bontocensis AJB105
Eutropis bontocensis RMB9548
Eutropis bontocensis RMB6002
Eutropis bontocensis RMB5999
Eutropis bontocensis RMB6003
Eutropis bontocensis RMB9590
Eutropis multicarinata RMB5848
Eutropis indeprensa RMB5993
Eutropis multicarinata RMB12414
Eutropis multicarinata RMB12001
Eutropis mborealis RMB944
Eutropis multicarinata CDS713
Eutropis multicarinata CDS3884
Eutropis multicarinata CDS3809
Eutropis rudis LG6178
Eutropis multicarinata RMB5556
Eutropis multicarinata RMB6584
Eutropis multicarinata RMB15317
Eutropis multicarinata RMB15277
Eutropis multicarinata RMB3287
Eutropis multicarinata RMB13518
Eutropis multicarinata RMB3488
Eutropis multicarinata RMB8909
Eutropis multicarinata RMB12454
Eutropis sp RMB4601
Eutropis multicarinata EMD257
Eutropis indeprensa RMB8328
Eutropis indeprensa RMB8488
Eutropis multicarinata CDS3242
Eutropis cf indeprensa ACD3980
Eutropis cumingi ACD5210
Eutropis multicarinata CDS5300
Eutropis sp ACD5120
Eutropis multicarinata GVAG294
Eutropis multicarinata GVAG280
Eutropis multicarinata GVAG281
Eutropis indeprensa RMB6498
Eutropis indeprensa RMB6520
Eutropis cumingi CWL448
Eutropis cumingi CWL155
Eutropis indeprensa RMB9428
Eutropis indeprensa CDS1571
Eutropis cumingi CWL325
Eutropis indeprensa ELR1619
Eutropis cumingi RMB7840
Eutropis indeprensa ELR1502
Eutropis rugifera LG6160
Eutropis indeprensa ELR684
Eutropis indeprensa ELR275
Eutropis indeprensa ELR685
Eutropis indeprensa ELR274
Eutropis indeprensa CDS1255
Eutropis cumingi ELR1398
Eutropis rudis CDS2281
Eutropis cumingi CDS016
Eutropis multicarinata RMB9591
Eutropis cumingi CDS3808
Eutropis cf cumingi RMB3125
Eutropis cumingi FMNH258984
Eutropis cumingi AJB106

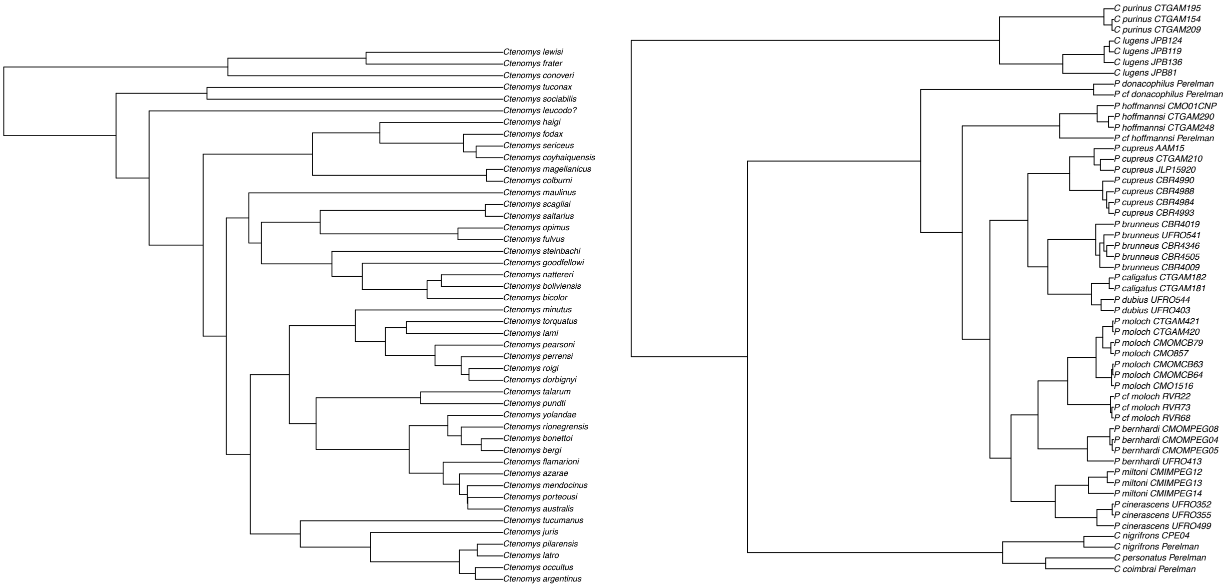
(k) Delphinids

(l) Philippine sun skinks


Rattus niobe ABTC46763
Rattus niobe ABTC47145
Rattus niobe ABTC42489
Rattus verecundus ABTC47140
Rattus praetor ABTC47273
Rattus praetor ABTC47272
Rattus praetor ABTC47252
Rattus steini ABTC47100
Rattus novaeguineae ABTC48963
Rattus novaeguineae ABTC48962
Rattus novaeguineae ABTC46853
Rattus novaeguineae ABTC43216
Rattus giluwensis ABTC87305
Rattus giluwensis ABTC87301
Rattus leucopus ringens AMSM17128
Rattus leucopus ringens ABTC46876
Rattus leucopus dobodurae ABTC42808
Rattus leucopus dobodurae ABTC42806
Rattus leucopus dobodurae ABTC42803
Rattus leucopus dobodurae AMSM08493
Rattus leucopus leucopus IR02R02
Rattus leucopus leucopus SP03R01
Rattus leucopus leucopus ABTC08493
Rattus leucopus cooktownensis DT01R03
Rattus leucopus cooktownensis WO04R09
Rattus leucopus cooktownensis DT10R02
Rattus leucopus cooktownensis DC07R02
Rattus leucopus cooktownensis PR04R01
Rattus leucopus cooktownensis WO05R09
Rattus leucopus cooktownensis WO04R05
Rattus leucopus cooktownensis DC11R05
Rattus leucopus cooktownensis ABTC51770
Rattus fuscipes assimilis S17
Rattus fuscipes assimilis S05
Rattus fuscipes assimilis S01
Rattus fuscipes assimilis N05
Rattus fuscipes assimilis N01
Rattus fuscipes assimilis N02
Rattus fuscipes assimilis C21
Rattus fuscipes assimilis A10
Rattus fuscipes coracius PR06R01
Rattus fuscipes coracius WO05R11
Rattus fuscipes coracius WO05R04
Rattus fuscipes coracius WO04R06
Rattus fuscipes coracius WO04R04
Rattus fuscipes coracius DC08R05
Rattus fuscipes coracius GI02R01
Rattus fuscipes coracius DC02R06
Rattus fuscipes coracius 1618
Rattus fuscipes greyi RAT60
Rattus fuscipes fuscipes ABTC08644
Rattus fuscipes fuscipes ABTC08616
Rattus lutreolus velutinus ABTC51762
Rattus lutreolus velutinus D69
Rattus lutreolus lutreolus ABTC51720
Rattus lutreolus lutreolus D39
Rattus tunneyi tunneyi ABTC51684
Rattus tunneyi tunneyi ABTC08815
Rattus tunneyi culmorum RAT132
Rattus tunneyi culmorum D45
Rattus sordidus ABTC51664
Rattus sordidus RAT176
Rattus villosissimus ABTC51675
Rattus villosissimus ABTC51672
Rattus villosissimus ABTC08439
Rattus colletti ABTC5656
Rattus colletti ABTC51650
Rattus colletti ABTC51642
Sticta dichotoma LG945
Sticta dichotoma LG985
Sticta dichotoma LG3448
Sticta dichotoma LG1005
Sticta dichotoma LG389
Sticta sp LG3450
Sticta sp LG978
Sticta sp LG639
Sticta plumbea LG1031
Sticta macrophylla LG3445
Sticta macrophylla LG3452
Sticta macrophylla LG946
Sticta sp LG1017
Sticta sp LG801
Sticta sp UCONN4018
Sticta sp LG4099
Sticta caperata LG962
Sticta caperata LG624
Sticta caperata LG4069
Sticta sp UCONN4096
Sticta sp UCONN4038
Sticta sp UCONN4102
Sticta sp LG1109
Sticta sp LG3453
Sticta sp LG3447
Sticta variabilis LG1037
Sticta variabilis LG3451
Sticta variabilis LG4058
Sticta variabilis LG4060
Sticta sp LG4066
Sticta sp LG4062
Sticta sp UCONN4116
Sticta sp LG4114
Sticta sp LG4093
Sticta sp LG4100
Sticta sp LG4096
Sticta sp LG941
Sticta sp UCONN4076
Sticta sp UCONN4128
Sticta sp LG800
Sticta sp LG991
Sticta sp LG4073
Sticta sp LG4059
Sticta sp UCONN4028
Sticta sp LG4077
Sticta sp LG4070
Sticta sp UCONN4139
Sticta sp LG4085
Sticta sp LG4092
Sticta sp LG4107
Sticta sp LG4111
Sticta sp LG4082
Sticta sp LG4074
Sticta sp LG4064
Sticta sp LG4110
Sticta sp LG4083
Sticta sp UCONN4070
Sticta sp LG4084
Sticta sp UCONN4120
Sticta sp UCONN4115
Sticta sp LG4109
Sticta sp LG4112
Sticta sp LG4117
Sticta sp UCONN4111
Sticta sp UCONN4094
Sticta sp LG4063
Sticta sp UCONN4137
Sticta sp UCONN4123

(m) Rodents *Rattus*

(n) Lichen *Sticta*



(o) Caviomorph rodents *Ctenomys*

(p) New World titi monkeys

FIGURE B.11: Phylogenetic trees corresponding to points 1-16 of Table 1 on the main text.

B.7 Spatial patterns of speciation

In this section we show the time evolution of the spatial patterns of speciation (relative to Fig.1 in the main text). Figs. B.12 and B.13 illustrate the equilibration times and rates of speciation.

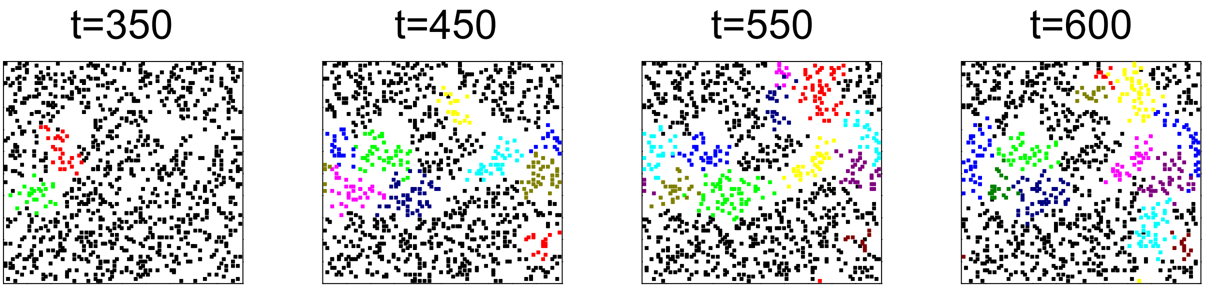


FIGURE B.12: Time evolution of the population for $B = 150$ and $S = 5$. Speciation events occur gradually as small species separate from the original population. Since species originate with small number of individuals, extinctions are more likely to occur during the radiation. Average species size at t=600 is 50, which is already the average size at equilibration.

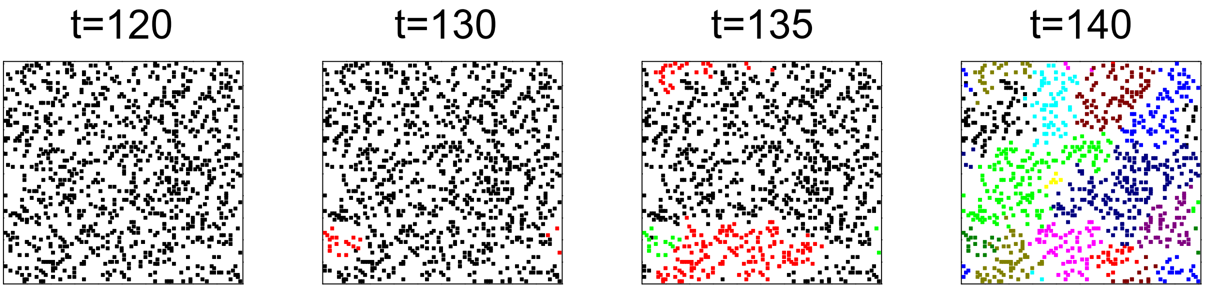


FIGURE B.13: Time evolution of the population for $B = 150,000$ and $S = 5$. Several speciation events happen in a very short time scale. The first species to branch off the main population are large and will speciate again, fragmenting into smaller species. During the radiation, extinction by stochastic fluctuations (ecological drift) is not likely, since species are large. Average species size at t=140 is 83, whereas the average size at equilibration is 37.

B.8 Abundance distributions

Figure B.14 shows abundance distributions calculated at equilibration times for three extreme cases: (a) small S, small B ($S = 5$, $B = 150$); (b) small S, large B ($S = 5$, $B = 150,000$) and; (c) large S, large B ($S = 40$, $B = 150,000$). The distributions are very well fit by a lognormal curve (**DeAguiar2009**). In the first case, the mean population size is 51 individuals, distributed among 20 ± 3 species. In the second case, the mean of population size is smaller than in the previous case: 36 individuals distributed among 27 ± 3 species. Finally, in the third case, there are 80 individuals per species in average, distributed among 13 ± 2 species. The log-scale abundance histograms were draw by using geometrically increasing bin widths, so that the width of bin k was set to $7 \times 1.3^{k-1}$.

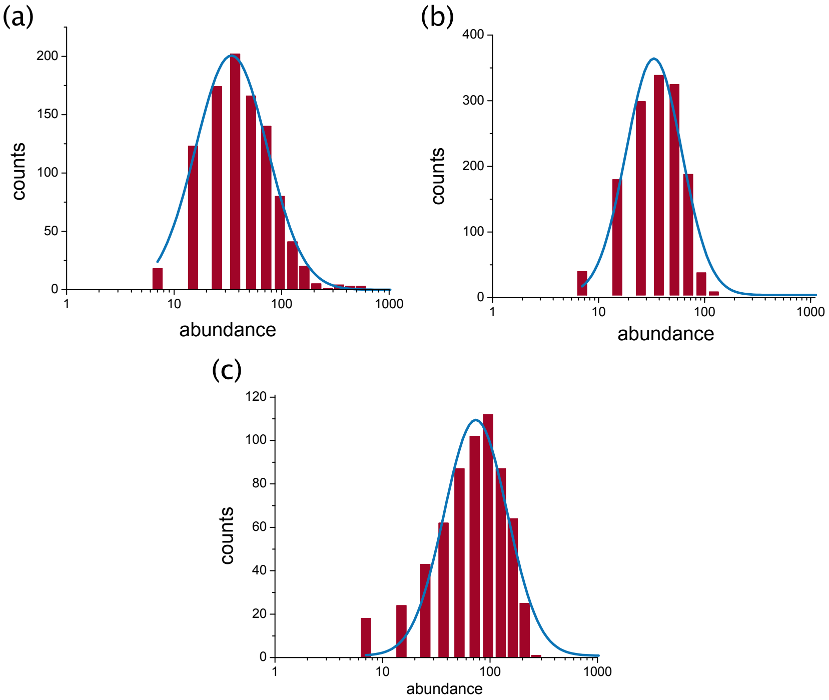


FIGURE B.14: Abundance distribution at equilibration times in log-scale for different values of B and S, obtained with 50 replicas. The solid blue curve shows a fit by the lognormal distribution function centered at the scale parameter μ and with the shape parameter σ. (a) $B = 150$, $S = 5$ ($\mu = 61.6$, $\sigma = 0.77$, $R^2 = 0.993$); (b) $B = 150,000$, $S = 5$ ($\mu = 48$, $\sigma = 0.606$, $R^2 = 0.97$); and (c) $B = 150,000$, $S = 40$ ($\mu = 116$, $\sigma = 0.67$, $R^2 = 0.95$).

B.9 Parameter inference

The model parameters S and B can be inferred from the phylogenies with some accuracy. To make the inference, we used 40 replicates to build theoretic non-linear curves for S and B from data to then estimate the theoretic values of S and B based on the statistics of the trees. Therefore, to first estimate the theoretic values of S, we calculated the regression of alpha-value and of Sackin index of 40 replicates for $B = 150,000$ (top of Fig. B.15). Through the inverse function fitted by the regression, we calculated the theoretic value of S using the alpha-value and Sackin index of the trees of 10 replicates. We plotted the estimated values against the simulated values to visualize the inference made (bottom of Fig. B.15). We used only the data with mating range up to $S = 20$ because of the impossibility to distinguish $S = 20$ and $S = 40$ on the regression.

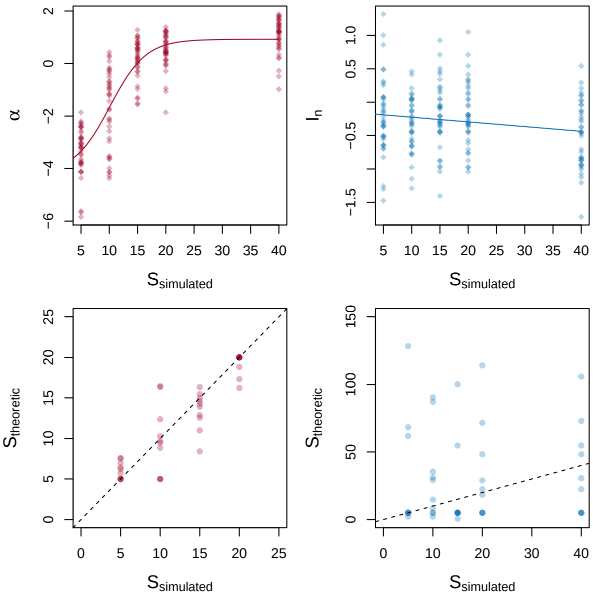


FIGURE B.15: In the first row are the estimated regression curves of the 40 replicates of alpha-value (α, at the left) and Sackin index (I_n, at the right) against the simulated values of S. In the second row are the theoretic values against the simulated values of S. The theoretic values of S were calculated based on the inverse curves of the top and the from the statistics of 10 trees. The dotted black curve represents the identity 1:1 of theoretic and simulated values.

To estimate the theoretic values of B, we calculated the regression of alpha-value and of Sackin index of 40 replicates for $S = 5$ (top of Fig. B.16), using the logarithmic values of B (base 10). Through the inverse function fitted by the regression, we calculated the theoretic value of B (also in logarithmic scale) using the alpha-value and Sackin index of the trees of 10 replicates. We plotted the estimated values against the simulated values to visualize the inference made (bottom of Fig. B.16).

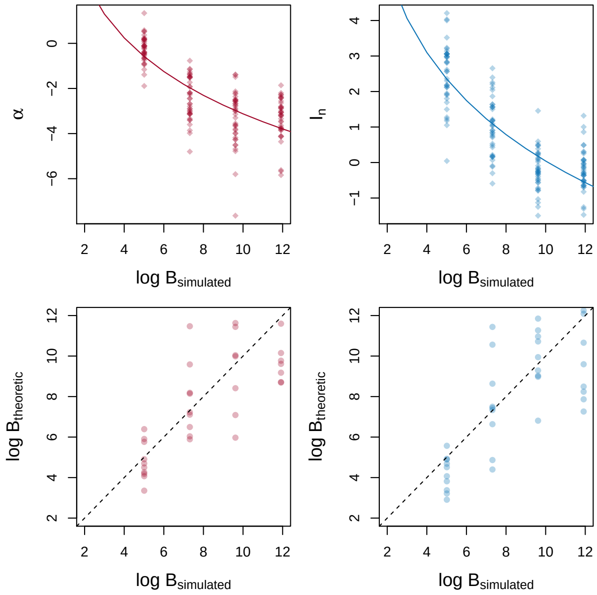


FIGURE B.16: On the top are the estimated regression curves of the 40 replicates of alpha-value (α, at the left) and Sackin index (I_n, at the right) against the simulated values of B. On the bottom are the theoretic values against the simulated values of B. The theoretic values of B were calculated based on the inverse curves of the top and the from the statistics of 10 trees. The dotted black curve represents the identity 1:1 of theoretic and simulated values. Notice the log scale for B values.

In conclusion, the parameter of mating range (S) can be better inferred by the alpha-value. The inference through the Sackin index is not good because of the flatness of the regression curve. The parameter of genome size (B) can be inferred similarly by both alpha-value and Sackin index visualizing the plot of $B_{theroretic} \times B_{simulated}$, even though the effect size of alpha-value is greater than the effect size of the Sackin index ($\eta_{\alpha}^2 = 0.70$ and $\eta_{I_n}^2 = 0.67$).

Appendix C

Supporting Information - Chapter 3

C.1 Empirical information about platyrrhine monkeys

TABLE C.1: Empirical information about species range, body mass and river middle width separating extant species for each platyrrhine genera and subfamily.

Family	Genera	Home range (ha)	Body mass (kg)	River middle width (km)
Atelidae	*Alouatta*	39.74	6.221	2.40
Atelidae	*Ateles*	311.93	8.608	1.10
Atelidae	*Lagothrix*	589.50	8.356	1.90
Cebinae	*Cebus*	329.08	2.847	1.75
Callitrichinae	*Saguinus*	48.73	0.475	1.00
Callitrichinae	*Leontopithecus*	117.75	0.593	0.57
Callitrichinae	*Callithrix*	10.92	0.296	0.70
Callitrichinae	*Callimico*	90.00	0.484	0.80
	Callicebus	15.81	1.107	0.90
Pithecidae	*Pithecia*	18.70	2.093	2.00
Pithecidae	*Cacajao*	531.25	2.960	1.35

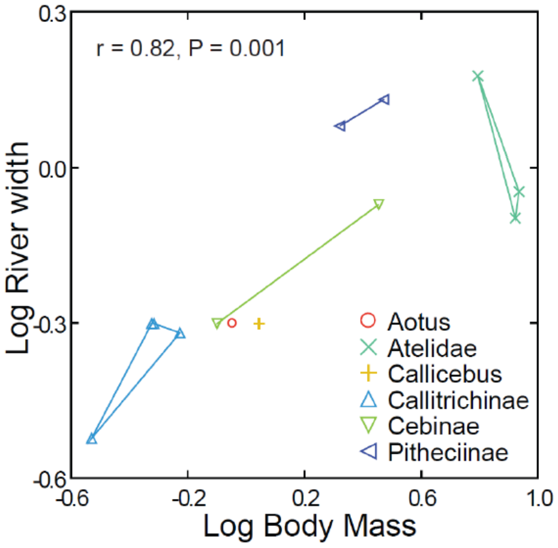


FIGURE C.1: Relationship between body mass and river width in different clades of Platyrrhini.

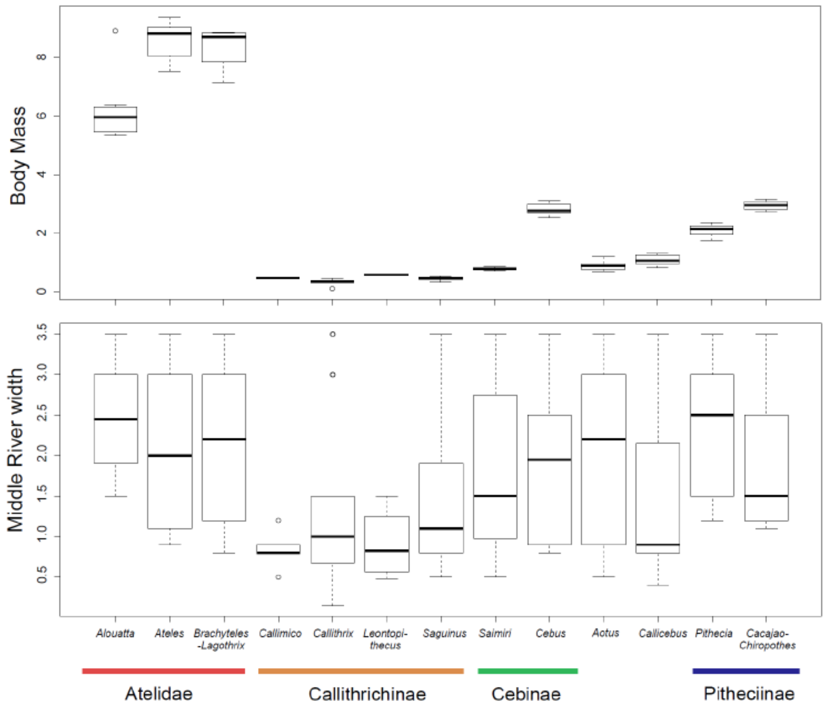
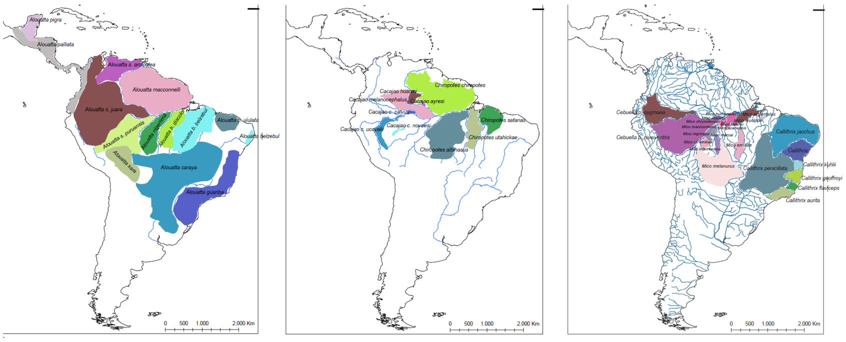
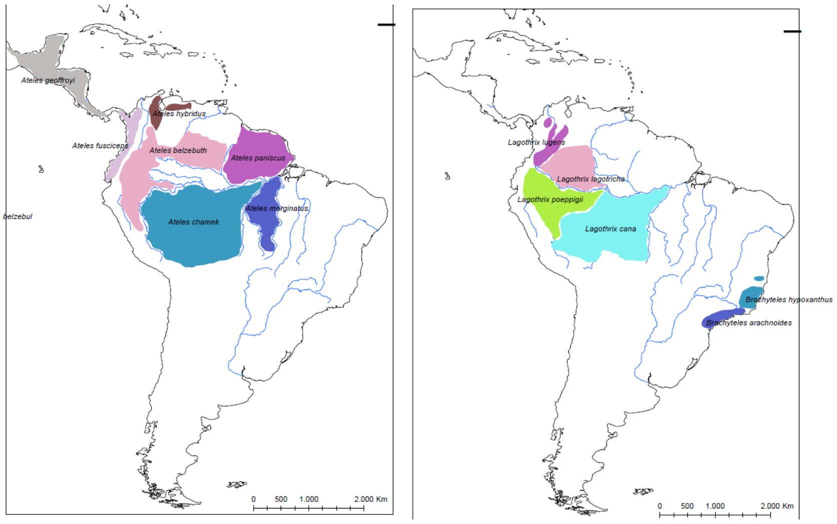


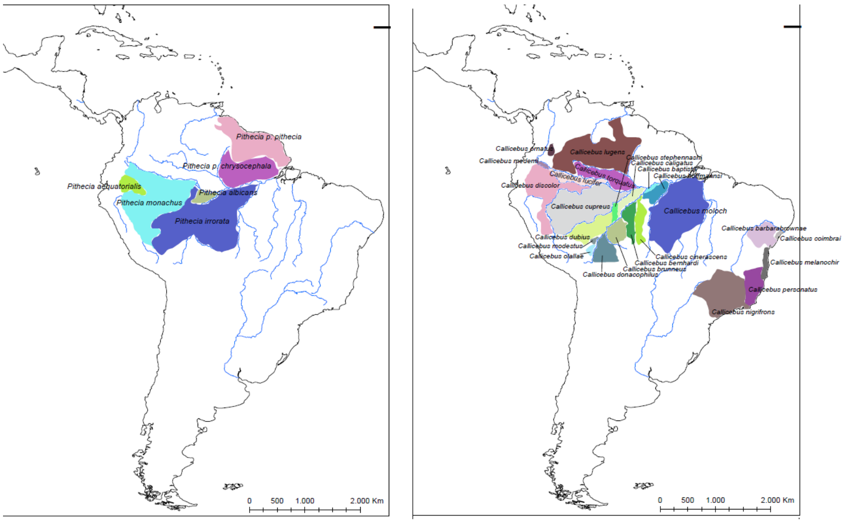
FIGURE C.2: Box plot of body mass and middle river width delimitating the range of each platyrrhine family and subfamily.



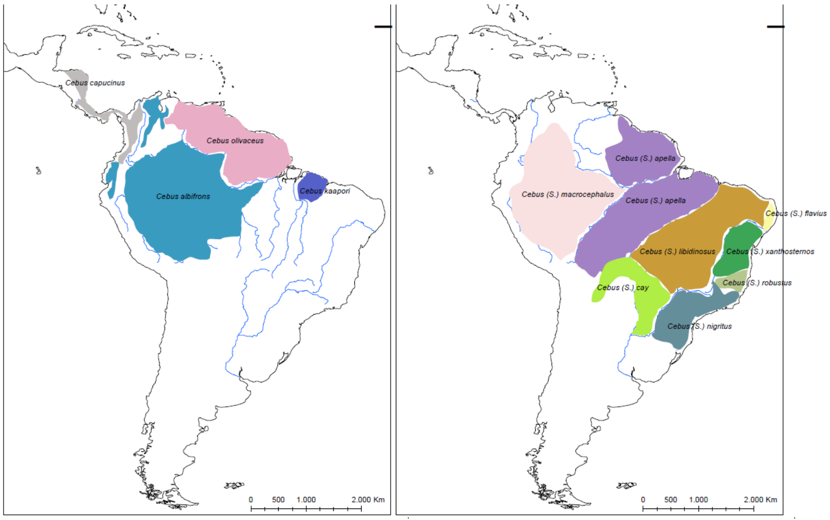
(a) Distribution of (Left): *Alouatta* species (Atelidae family), (Middle): *Chiropotes* species (Pithecidae family), and (Right): *Callithrix*, *Cebuella* and *Mico* species (Callithrichinae subfamily).



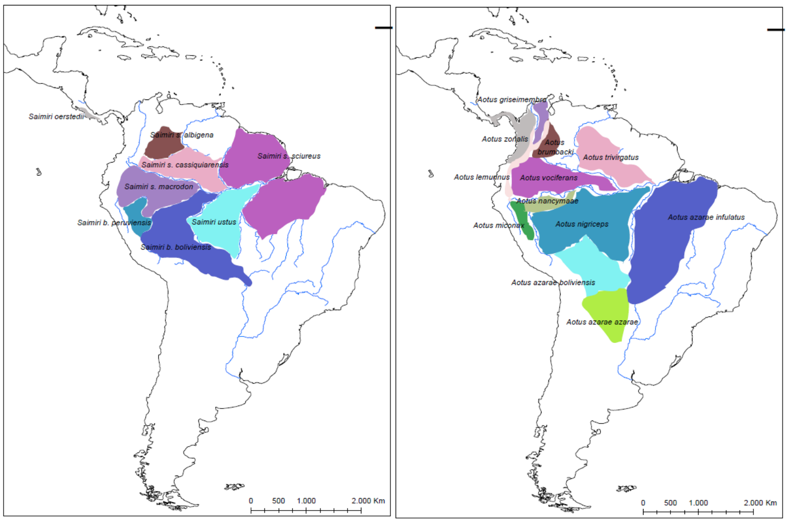
(b) Distribution of *Ateles* and *Lagothrix* species (Atelidae family).



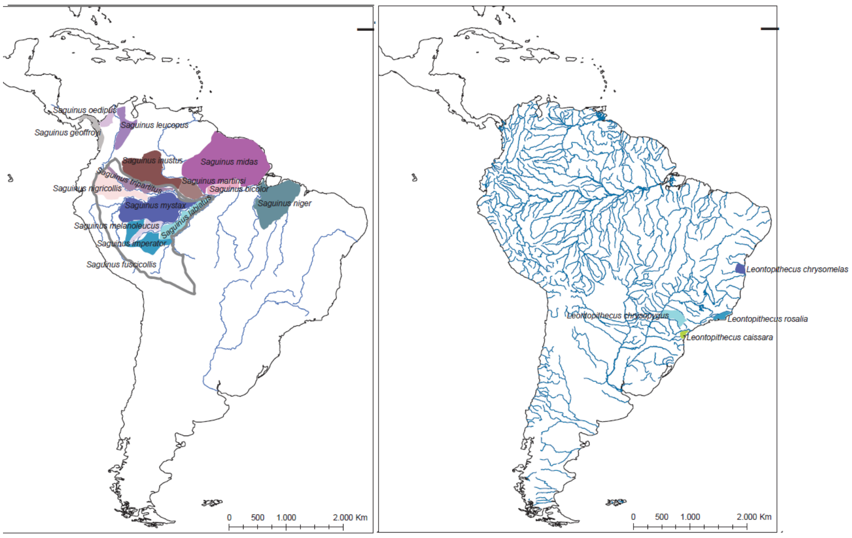
(c) Distribution of *Pithecia* (Pithecidae family) and *Callicebus* species.



(d) Distribution of *Cebus* species (Cebinae subfamily).



(e) Distribution of *Saimiri* (Cebinae subfamily) and *Aotus* species.



(f) Distribution of *Saguinus* and *Leontopithecus* species (Callitrichinae subfamily).

FIGURE C.3: Geographic distribution of platyrrhine families and subfamilies.

TABLE C.2: Empirical information about width of rivers utilized in the South America lattice and minimum values of radius S parameter in which monkeys could cross the rivers.

River name	River source width (km)	River middle width (km)	River mouth width (km)	River size category	Simulated river width	Radius S
Guapore	0.4	0.5	0.5	a	3	5
Purus	0.3	0.5	0.8	a	3	5
Parnaiba	0.3	0.4	1.5	a	3	4
Marañon	0.1	0.5	1.0	a	3	4
Ucayali	0.6	0.9	0.9	b	6	7
Putumayo	0.5	0.8	0.9	b	6	7
Madre de Dios	0.6	0.8	1.2	b	6	7
Teles Pires	0.6	0.75	0.85	b	6	7
Tocantis	0.8	0.9	1.2	b	6	7
São Franciso	0.6	0.8	0.98	b	6	7
Paraguay	0.35	0.75	0.7	b	6	7
Japurá	0.7	1.2	1.5	c	9	10
Rio Branco	1.3	1.5	1.0	c	9	10
Madeira	1.1	1.5	2.0	c	9	10
Xingu	1.0	1.1	1.0	c	9	10
Araguaia	1.0	1.5	1.1	c	9	10
Orinoco	0.9	1.9	3.0	d	15	16
Rio Negro	3.0	2.5	6.0	d	15	16
Paraná	1.5	2.4	1.5	d	15	16
Tapajós	1.3	3.0	11.0	e	18	19
Amazonas	2.0	3.5	6.0	e	18	19

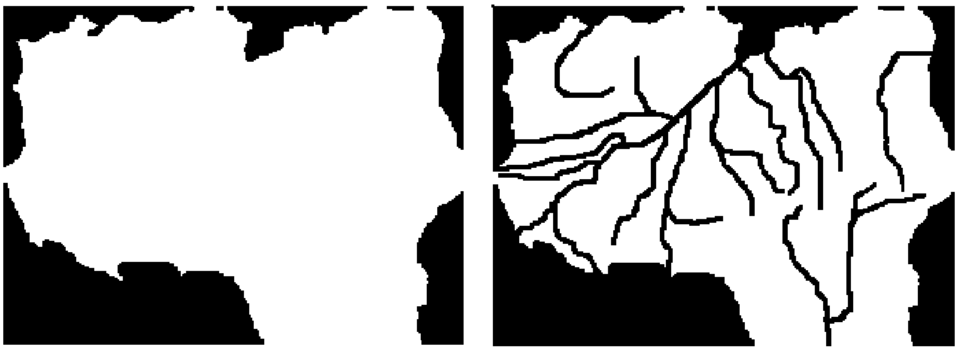


FIGURE C.4: South America lattices utilized to simulate the platyrrine speciation process. On the left, the lattice without rivers, in which individuals can move freely through space. On the right, the lattice with main rivers of the region, acting as barriers to individuals movement. In the second figure, rivers can have the same width (scenario 2) or different widths, proportional to real river widths (scenario 3) as stablished in Table C.2.

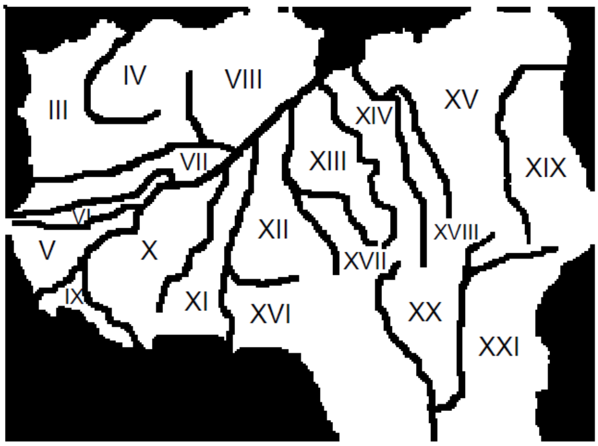


FIGURE C.5: Operative Geographic Units (OGUs) utilized to measure the spatial distribution pattern of platyrrhine species, based on (Goldani et al., 2006) definition.

TABLE C.3: Phylogenetic trees estimated from empirical data for each platyrrhine family and subfamily utilized to compare phylogenetic and spatial patterns obtained with simulations of each size category of monkeys. The number of species in each group varies according to study.

Group	Size category	Number of species	Sackin	Gamma	Alpha	References
Callitrichinae	Small	25	0.728	3.914	2.091	(Aristide et al., 2015)
		11	0.324	0.945	1.533	Personal communication
		14	0.283	1.851	1.832	(Arnold et al., 2010)
		11	0.415	-2.546	-1.122	(Perelman et al., 2011)
		11	0.140	1.446	1.654	(Springer et al., 2012)
Callicebus	Small	24	-0.344	2.324	1.675	(Aristide et al., 2015)
		11	-0.040	-0.405	0.754	Personal communication
		5	0.233	-0.405	0.451	(Arnold et al., 2010)
		8	-0.311	-1.106	0.026	(Perelman et al., 2011)
		9	0.120	-1.440	-0.185	(Springer et al., 2012)
Saguinus	Small	12	-0.373	-0.599	0.657	(Aristide et al., 2015)
		12	-0.373	-0.595	0.659	Personal communication
		11	-0.222	-0.478	0.708	(Arnold et al., 2010)
		9	0.120	-1.069	0.165	(Perelman et al., 2011)
		10	0.542	1.826	2.120	(Springer et al., 2012)
Pithecidae	Medium	14	-0.575	2.169	1.974	(Aristide et al., 2015)
		8	-0.311	0.535	1.412	Personal communication
		5	-0.167	0.040	1.052	(Arnold et al., 2010)
		6	-0.233	0.343	1.360	(Perelman et al., 2011)
		6	-0.233	0.344	1.361	(Springer et al., 2012)
Cebus	Medium	12	-0.290	0.919	1.481	(Aristide et al., 2015)
		8	-0.061	0.546	1.421	Personal communication
		5	-0.167	0.487	1.633	(Arnold et al., 2010)
		6	-0.233	-0.539	0.401	(Perelman et al., 2011)
		7	-0.186	-0.001	0.998	(Springer et al., 2012)
Atelidae	Large	22	-0.654	0.537	1.177	(Aristide et al., 2015)
		18	-0.601	-0.461	0.812	Personal communication
		13	-0.283	-0.516	0.726	(Arnold et al., 2010)
		14	-0.360	0.087	1.041	(Perelman et al., 2011)
		16	-0.386	0.820	1.339	(Springer et al., 2012)

C.2 Additional results

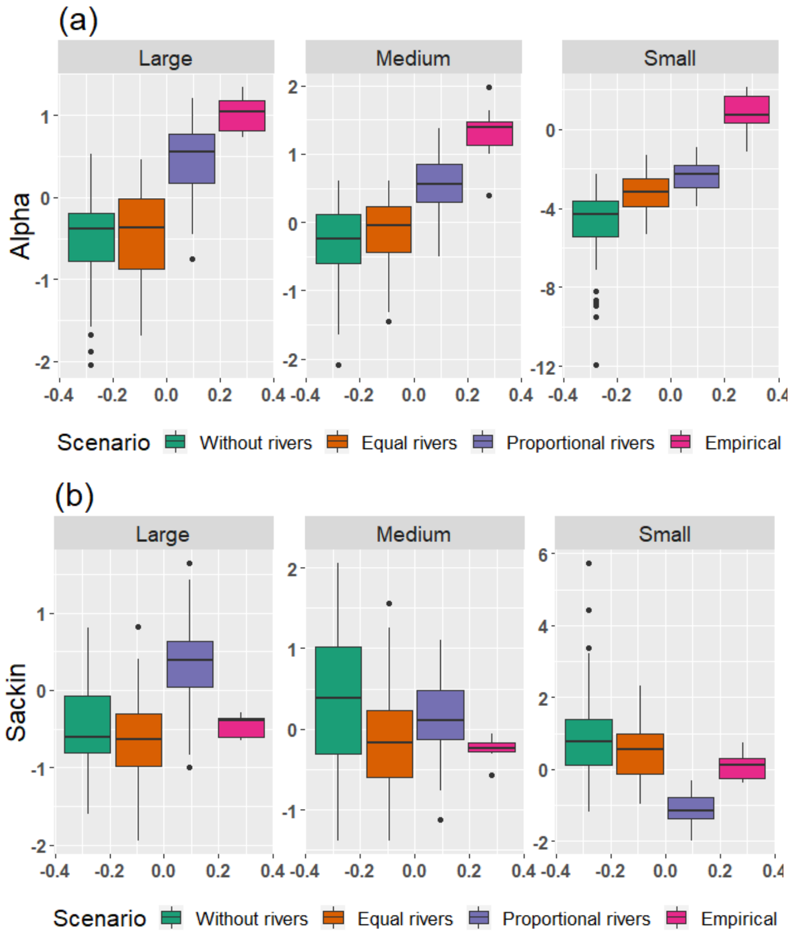


FIGURE C.6: Boxplots for structural properties of true simulated trees generated from the SSEE algorithm. (a) Boxplots for α-values of all simulated scenarios and empirical data for large, medium and small monkeys. (b) Boxplots for sackin values of all simulated scenarios and empirical data for large, medium and small monkeys.

TABLE C.4: Statistical analysis of scenarios in each monkey size. The Kruskal-Wallis rank sum test was used to compute differences between scenarios with Proportional rives (P), Equal rivers (E) or Without rivers (W). Comparisons between each pair of scenarios was computed with the post-hoc Nemenyi test, with p-values showed for each pairwise comparison (P-E, P-W, E-W).

	Small monkeys			
	Kruskal-Wallis	**post-hoc Nemenyi**		
		P-E	P-W	E-W
Alpha	$2.4x10^{-14}*$	$0.68x10^{-3}*$	$2.4x10^{-14}*$	$0.22x10^{-3}*$
Sackin	$< 2.2x10^{-16}*$	$1.1x10^{-13}*$	$< 2.2x10^{-16}*$	0.61

	Medium monkeys			
	Kruskal-Wallis	**post-hoc Nemenyi**		
		P-E	P-W	E-W
Alpha	$2.3x10^{-13}*$	$5.8x10^{-9}*$	$2.6x10^{-11}*$	0.71
Sackin	$0.2x10^{-2}*$	0.072	0.522	$0.27x10^{-2}*$

	Large monkeys			
	Kruskal-Wallis	**post-hoc Nemenyi**		
		P-E	P-W	E-W
Alpha	$7.3x10^{-15}*$	$5.6x10^{-11}*$	$1.1x10^{-11}*$	0.97
Sackin	$9.3x10^{-14}*$	$8.6x10^{-12}*$	$5.2x10^{-9}*$	0.63

$^*p < 0.001$

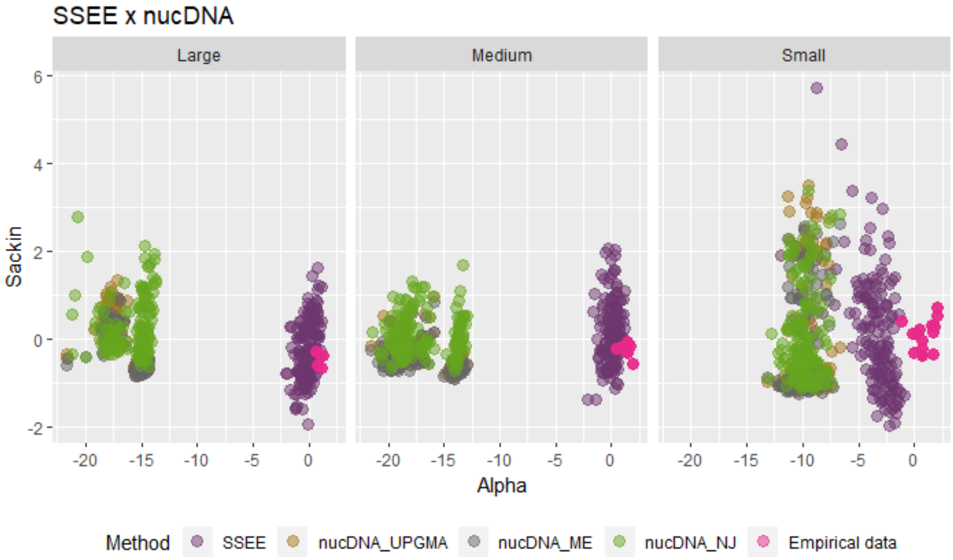


FIGURE C.7: Scatterplots for both structural properties (α-value and Sackin index) of true simulated trees (SSEE) and trees estimated from nucDNA of all simulated scenarios and empirical data for large, medium and small monkeys. Each color represents a method to generate phylogenies for nucDNA, without separating the scenarios in each case.

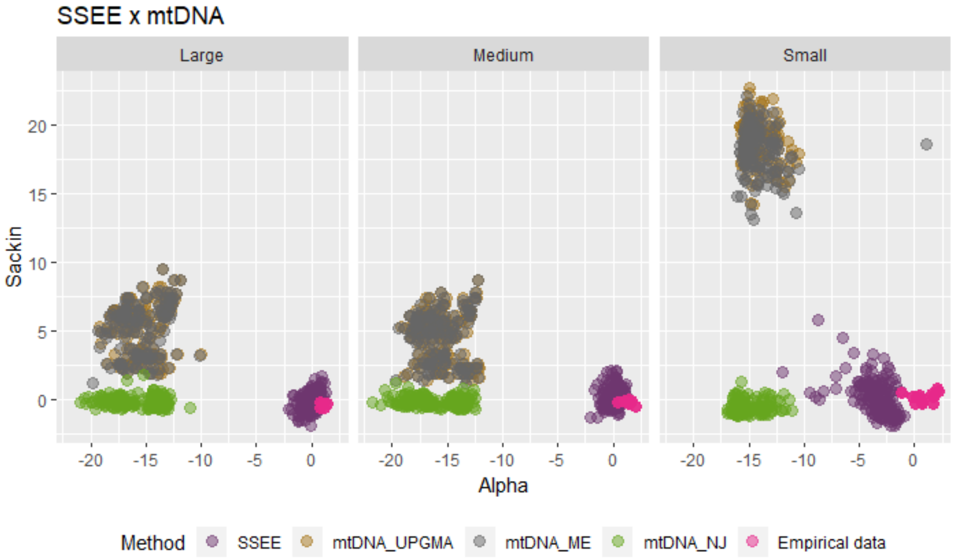


FIGURE C.8: Scatterplots for both structural properties (α-value and Sackin index) of true simulated trees (SSEE) and trees estimated from mtDNA of all simulated scenarios and empirical data for large, medium and small monkeys. Each color represents a method to generate phylogenies for mtDNA, without separating the scenarios in each case.

TABLE C.5: Statistical analysis of scenarios in comparison with empirical data for each monkey size. The Kruskal-Wallis rank sum test was used to compute differences between each scenario and between empirical data (Emp) and scenarios Without rivers (W), with Equal rivers (E) or Proportional rives (P). Comparisons between each pair of combination was computed with the post-hoc Nemenyi test, with p-values showed for each pairwise comparison (Emp-W, Emp-E, Emp-P). Only comparisons between empirical data and scenario are showed, the comparisons between scenarios are presented in Table C.4.

	Small monkeys			
	Kruskal-Wallis	**post-hoc Nemenyi**		
		Emp-W	Emp-E	Emp-P
Alpha	$< 2.2x10^{-16}*$	$5.4x10^{-15}*$	$2.3x10^{-7}*$	0.0079
Sackin	$< 2.2x10^{-16}*$	0.20429	0.53006	$0.099x10-2$*

	Medium monkeys			
	Kruskal-Wallis	**post-hoc Nemenyi**		
		Emp-W	Emp-E	Emp-P
Alpha	$< 2.2x10^{-16}*$	$7.8x10^{-8}*$	$9.5x10^{-7}*$	0.2
Sackin	0.001347*	0.0837	0.9431	0.2519

	Large monkeys			
	Kruskal-Wallis	**post-hoc Nemenyi**		
		Emp-W	Emp-E	Emp-P
Alpha	$5.023x10^{-16}*$	$0.25x10^{-3}*$	$0.36x10^{-3}*$	0.57
Sackin	$2.381x10^{-13}*$	0.99	0.93	0.12

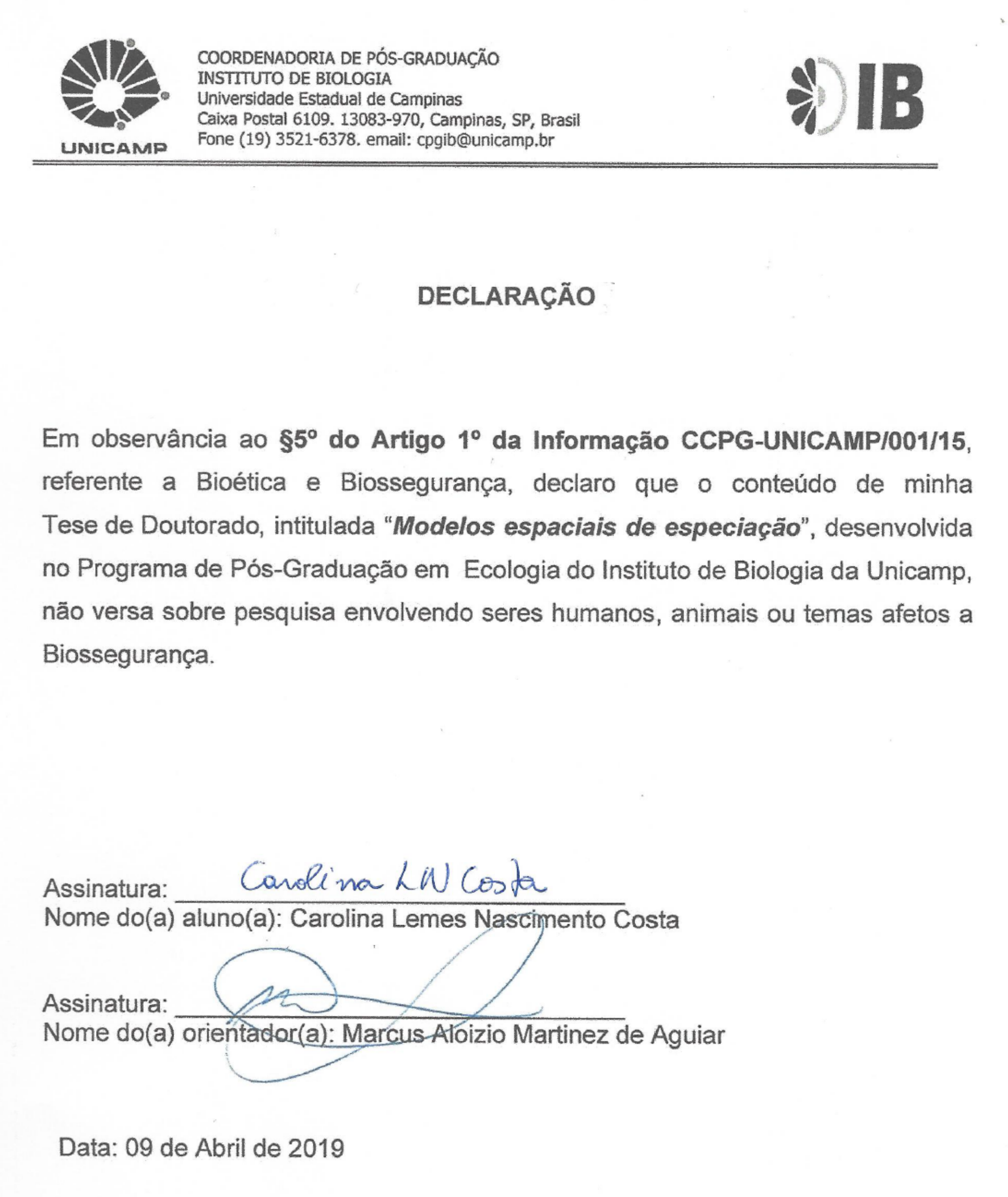
$^{*}p < 0.001$

TABLE C.6: Median of the number of species and the first and third quartiles (*Q*1 and *Q*3, respectively) for all OGUS together in each scenario and monkey size. Scenarios: Without rivers (W), with Equal rivers (E) or Proportional rives (P).

	Small monkeys			**Medium monkeys**			**Large monkeys**		
	W	E	P	W	E	P	W	E	P
Median	6	4	3	5	4	1	9	7	3
Q1	5	2	2	4	3	1	6	6	2
Q2	8	5	4	7	5	2	10	9	5

Appendix D

Bioética e Biossegurança



UNICAMP
COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br
IB

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15**, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado, intitulada "***Modelos espaciais de especiação***", desenvolvida no Programa de Pós-Graduação em Ecologia do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

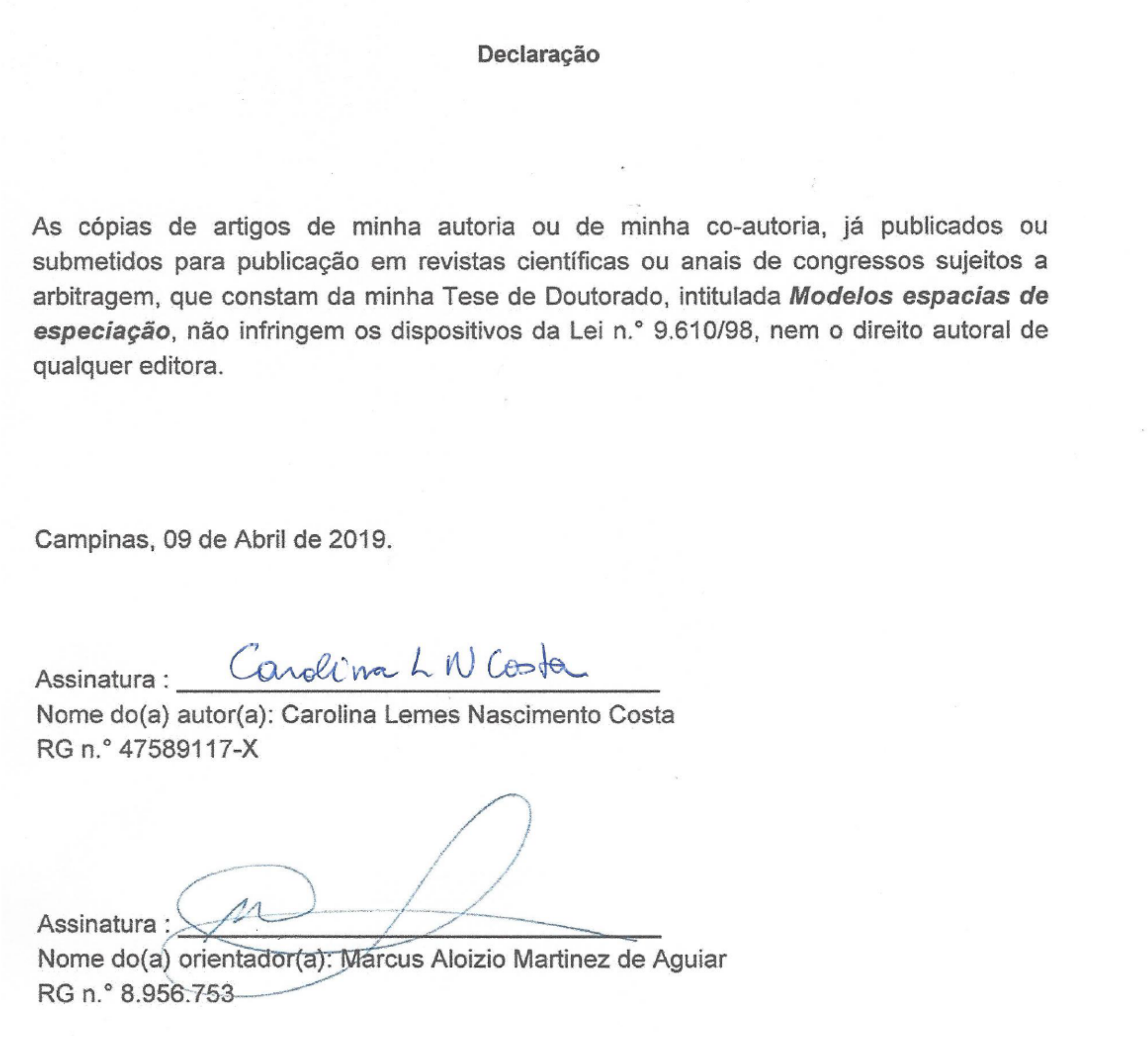
Assinatura: Carolina LN Costa
Nome do(a) aluno(a): Carolina Lemes Nascimento Costa

Assinatura: ______________________
Nome do(a) orientador(a): Marcus Aloizio Martinez de Aguiar

Data: 09 de Abril de 2019

Appendix E

Direitos autorais



Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Tese de Doutorado, intitulada ***Modelos espacias de especiação***, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 09 de Abril de 2019.

Assinatura : Carolina L N Costa
Nome do(a) autor(a): Carolina Lemes Nascimento Costa
RG n.° 47589117-X

Assinatura : ____________________________________
Nome do(a) orientador(a): Marcus Aloizio Martinez de Aguiar
RG n.° 8.956.753