THE UNIVERSITY
of EDINBURGH

Phylogeny, taxonomy and biogeography of *Ceiba* Mill. (Malvaceae: Bombacoideae)

Flávia Fonseca Pezzini

Doctor of Philosophy

Royal Botanic Garden Edinburgh

The University of Edinburgh

2019

Phylogeny, taxonomy and biogeography of *Ceiba* Mill. (Malvaceae: Bombacoideae)

Flávia Fonseca Pezzini

The University of Edinburgh

School of Biological Sciences

Institute of Molecular Plant Sciences

Royal Botanic Garden Edinburgh

Supervisors

R. Toby Pennington[1,2], Catherine A. Kidner[2,3], Kyle G. Dexter[3]

[1]University of Exeter

[2]Royal Botanic Garden Edinburgh

[3]The University of Edinburgh

CHORISIA speciosa.

In the previous page: Flora Brasiliensis. Vol. XII, Part III, Fasc. 98, Prancha 40. Publicado em 01-Nov-1886. Família Bombaceae, Tribo Adansonieae Benth., Gênero *Chorisia* Kunth, *Chorisia speciosa* A.St.-Hil.

"Sertão. Sabe o senhor: o sertão é onde o pensamento da gente se forma mais forte do que o poder do lugar."

João Guimarães Rosa

Grande Sertão:Veredas.

Para vô Ismael, com amor.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Flávia Fonseca Pezzini

23rd October 2019

Abstract

The Neotropics is the most species-rich area in the world and the mechanisms that generated and maintain its biodiversity are still debated. This thesis contributes to the debate by investigating the evolutionary and biogeographic history of the genus *Ceiba*. *Ceiba* comprises 18 mostly neotropical species endemic to two major biomes, seasonally dry tropical forests (SDTFs) and rain forests, and therefore represents an ideal case to shed light on patterns of neotropical plant evolution and diversification. Species of *Ceiba*, with their swollen, spiny trunks and large, beautiful flowers are one of the most characteristic elements of neotropical SDTF, one of the most threatened biomes in the tropics. Despite this, *Ceiba* has an historically complex taxonomy with some issues of species delimitation unresolved, especially within a species complex (*Ceiba insignis* agg.).

Initial phylogenetic analyses of DNA sequence data from the nuclear ribosomal internal transcribed spacers (ITS) for 24 accessions representing 14 species of *Ceiba* recovered the genus as monophyletic and showed geographical and ecological structure in three main clades: (i) a humid forest lineage of three accessions of *C. pentandra* sister to the remaining species; (ii) a highly supported clade composed of *C. schottii* and *C. aesculifolia* from Central American and Mexican SDTF plus two accessions of *C. samauma* from inter Andean valleys from Peru; and (iii) a highly supported South American SDTF clade including 10 species showing little sequence variation. Within this South American clade, no species represented by multiple accessions were resolved as monophyletic.

To investigate unresolved species relationships further, next-generation hybrid capture was used to sequence 377 loci for 103 accessions representing all 18 *Ceiba* species. This data set was assembled using different approaches (*de novo* and reference mapping) and with different software and settings to assess their impact in downstream phylogenetic analysis. The 377 loci were concatenated and analysed under the maximum likelihood framework treated as a single partition. The well resolved and sampled NGS phylogenies showed a similar pattern of geographical and ecological structure

as inferred using ITS. The genus *Neobuchia* was recovered within the SDTF Central American and Mexican clade, and should therefore be incorporated within *Ceiba*. In the South American SDTF clade, there were multiple examples where a monophyletic group recognised as a taxonomic species was nested within another, paraphyletic taxonomic species, which suggests recent, ancestor-descendent species relationships. Within this clade, individual gene trees showed high conflict. Coalescent-based species delimitation analysis and morphological data revealed no clear species boundaries between *C. pubiflora* and *C. glaziovii*, and these species should be synonymised.

A subset of 111 loci was used to generate a dated phylogeny based on penalised likelihood analysis using the fossil flower of *Eriotheca prima* from the middle to late Eocene as a primary calibration. The stem node age of *Ceiba* was estimated as 45 Ma. The rain forest species *C. pentandra* and *C. samauma*, and the campos rupestres species *C. jasminodora*, were resolved with long stem lineages and shallow crown groups. Whilst some SDTF species were very old (e.g., *C. trischistandra*) and monophyletic, many South American SDTF species were resolved with short stem lineages and relatively deep crown groups, possibly suggesting low rates of extinction in the large Caatinga SDTF region. In addition, several South American SDTF species were not resolved as monophyletic. Such results of younger, non-monophyletic SDTF species and older, monophyletic rain forest species contrast with recent predictions that rain forest species may, on average, have more recent origins than SDTF species and will more often be non-monophyletic.

Ceiba has different and distinctive phylogenetic patterns that contradict recent theoretical predictions. It demonstrates that studies of other clades sampled densely with multiple accessions of each species using a multi-locus approach are needed if we are to understand the nature of species and their boundaries, and the diversification process in neotropical trees.

Lay summary

The tropics of the New World are the most species-rich area on the planet and the mechanisms that generated and maintain this biodiversity are still debated. This thesis contributes to the debate by investigating the evolutionary and biogeographic history of the genus *Ceiba*. *Ceiba* comprises 18 mostly Latin American species found in two major vegetations, seasonally dry tropical forests (SDTFs) and rain forests and therefore represents an ideal case to shed light on patterns of plant evolution and diversification. Species of *Ceiba*, with their swollen, spiny trunks and large, beautiful flowers are one of the most characteristic elements of SDTF, one of the most threatened biomes in the tropics. Despite this, the species of *Ceiba* are not well understood and require better definition to allow them to be identified with confidence.

An evolutionary (phylogenetic) tree for *Ceiba* was built using DNA sequence data from the nuclear ribosomal internal transcribed spacers (ITS) for 24 accessions representing 14 species of *Ceiba*. This showed Ceiba to be a natural (monophyletic) group. The evolutionary tree showed three main groups, found in different vegetations in different places: (i) a humid forest lineage of three accessions of *C. pentandra*; (ii) a group composed of *C. schottii* and *C. aesculifolia* from Central American and Mexican SDTF plus two accessions of *C. samauma* from inter Andean valleys from Peru; and (iii) a South American SDTF group including 10 species showing little DNA sequence variation. In this South American clade, many accessions representing the same species were not resolved together, raising questions whether they have been defined correctly.

To investigate unresolved species relationships further, a next generation sequencing technique, called hybrid capture, was used to sequence 377 genetic regions for 103 accessions representing all 18 *Ceiba* species. This produced a dataset 1,000 times larger that was analysed using different approaches and with different software and settings in order to assess their impact in inferring phylogenies. The well resolved and sampled NGS phylogenies showed a similar pattern of geography and ecology as inferred using ITS. The genus *Neobuchia* was recovered within the SDTF Central American and Mexican clade, and should therefore be incorporated within *Ceiba*.

In the South American SDTF clade, there were multiple examples where a natural group recognised as a species was nested within accessions representing another species, which suggests recent, ancestor-descendent species relationships. Species delimitation analysis and morphological data revealed no clear boundaries between *C. pubiflora* and *C. glaziovii*, and these two species should be considered as a single species.

A subset of 111 genetic regions was used to generate an evolutionary tree with a time dimension in order to estimate when evolutionary events occurred. The time of origin (stem node) of *Ceiba* was estimated as 45 Ma. The rain forest species *C. pentandra* and *C. samauma*, and *C. jasminodora*, a species from dry, rocky areas in central Brazil, were resolved as very old species (>40 million years). Although some SDTF species were very old (e.g., *C. trischistandra*) and resolved as natural groups, many South American SDTF species were estimated with younger ages. In addition, several South American SDTF species were not resolved as natural groups because other species arose from within them. These results of younger species that are no natural groups in SDTF and older species found in rain forest contrast with recent predictions that rain forest species may, on average, have more recent origins than SDTF species and will more often not be found to be natural groups.

Ceiba has different and distinctive evolutionary and biogegraphic patterns that contradict recent theoretical predictions. It demonstrates that studies of other clades sampled densely with multiple accessions of each species using a next generation DNA sequencing approach are needed if we are to understand the nature of species and their boundaries, and the diversification process in the most species-rich forests of the New World.

Acknowledgements

I am most grateful to the people guiding and inspiring me through this journey. Thank you Toby Pennington. This is the first thank you and the most difficult one. To you my deepest admiration and gratitude. Thank you for receiving me here, for teaching me and for trusting in me. You were an inspiration to me since I started my adventures in the dry forests. I feel lucky to have such a brilliant person around me. Thank you to Vanessa, Alex, Lucas, Marmite, Louis and Scout for receiving me so well here. Thank you Kyle Dexter for your friendship, for sharing science, beer, and laughs. For always being quick and available, and for making things so light and smooth. I am very thankful to Catherine Kidner for all you taught me, for being supportive and always available, and going through together into the crazy bioinformatics pipelines. Thank you James Nicholls for having all the answers to my questions, and having patience with all my questions. I know doing lab work with me was slower. I guess there was a reason for 15 minutes lunch break? I am lucky to have you around.

I am very grateful to the people that helped me during the intense field work expeditions. Chico Diniz, for listening to Marcolino, Inezita Barroso and Fábulas do Carreiro, Pena Branca e Xavantinho over and over again and finally admitting your love for the Sertão; Tati Calaça for showing me a bit of the Sertão and the Agreste lá de cima, passing along do "Riacho do Navio, que corre pro Pajeú, O rio Pajeú vai despejar No São Francisco, O rio São Francisco, Vai bater no mei do mar", and so many other nice memories listening to Fagner, Dominguinhos, Luiz Gonzaga (tengo lengo tengo) e Cantigas de Lampião, with occasional stops for the picolé de limão after lunch. Toby Pennington and Gwil Lewis for a wonderful trip in the matas secas in the north of Minas Gerais and for teaching me a bit about legumes (and also trying to convince me to work with this family, it might work); thank you Toby for coming with me to the Mata Seca state park in Minas Gerais, it is where I started to study the dry forests in 2006, where I started reading your thoughts about this amazing forests and where I often wondered about doing a PhD with you; Moabe Fernandes and Matheus Cota for the perfect combination of field work and sossego in Minas Gerais and Bahia;

Pablo Hendrigo for a long trip sampling the carsts in central Brazil while listening to Elomar and Xangai, especially while passing through Iuiú, Bahia; Felipe Melo for all the help and for a nice visit to Catimbau National Park, sharing stories while Cantoria plays in the background; Antonin Portelli for sharing all the enchantment of being inside a postcard all the time. I am thankful to Jefferson Carvalho-Sobrinho for the nice discussions about the Bombacoideae; Luciano Paganucci for his studies in the Caatinga and for being an inspiration to me, and Teo Nunes for the help with the samples deposited in the Herbário da Universidade Estadual de Feira de Santana, Bahia, Brazil.

A warm thank you to all the people from RBGE. I am especially thankful for the nice moments shared during tea time. Special thanks to Alan Elliott, my official Scottish teacher. Thank you to Jimmy Ratter for being an inspiration, for the nice chats in Portuguese about Januária, Antarctica and isca de peixe na beira do São Francisco. Thank you to everyone that shared laughs during the amazing Pantomime seasons, specially Stephan Helfer. I would like to thank Lesley Scott, Frieda Christie, Deborah Vaile, Alan Sneath, Kirstie Ross and Duncan Reddish for the support in RBGE. Special thank you to the lab faries Michelle Hart, Laura Forrest, Ruth Hollands for the last minute boxes of tips, for always having the door of the office open to me, and for making the lab work easier. Scott McGregor, Agron Shehi, Nicolas Gruter, Yvonne Lockhart, Michael Borland, Terry, Sandra and Marcos for the warm hello everyday, either in the frontdesk, along the corridors or in the canteen.

I am most grateful to my dear friends that shared so many laughs and pints and ceilidhs with me here. The best office I could ask for, the only thing is that you talk too much: Karina Banda, Julieth Serrano, Vanessa Leite, Lucia Campos-Dominguez, Surabhi Ranavat, Yun-Yu Chen, Subhani Ranasinghe, Pakkapol Thaowetsuwan, Thibauld Michel, Andrés Orejuela (you really talk too much!), Mauricio Cano and Andy Griffiths. Special thank you to my dear friends Lucia and Surabhi. All the people that made and make life here in Edinburgh so enjoyable: Julia Weintritt, Pedro Miranda, Victoria Cabrera, Hannah Atkins, Erik Koenen, Maria Fernanda Torres,

James Richardson (special thank you for all the musical advice), Erica Rievrs, Natashi Pilon, Bruno Paganeli, Cynthia Fan, Yu-Hsin Tseng, Domingos Cardoso, Natalia Ortiz, Max Brown, Hannah Wilson, Madhavi Sreenath, Isuru Kariyawasam, Jess Rickenback, Ricardo Segovia, Izabela Barata, Peter Moonlight, Linda Neaves, María Camila Gutiérrez, Ray Considine. Thank you Erik Koenen and Yu-Hsin Tseng for all the help and nice discussions about analysis; and Mafe for sharing nice moments in the lab and questions to James. A special thank you to the quarteto improvável: Karina, Julieth and Vanessa for moments I will always carry with me. To my manauaras-lost-in-the-UK: Laynara Lugli and Fernanda Coelho, with you everything was easy here. Fernanda Costa e Mariana Cassino, saudades, tão longe, mas tão perto. Thank you to the Leeds bunch plus agregados, Julia Tavares, Karina Melgaço, Thaise Emílio, Demétrius Martins, Bruno Ladvocat, Marta Giannichi, Camila Duarte, Adriane Esquivel Muelbert, Gustavo Côrte. Special thank to you Chico Diniz, and of course your pão de queijo com café e conversa fiada. I am most grateful to my two favourite climbing teachers, Mauricio Cano and Andy Griffiths. Thank you for the hard and benevolent training, for all the cycling adventures along the old canal (under the sun or under the rain, and sometimes with the wind against us, it is Scotland afterall), for sport climbing, for pushing me and for not dropping me (Mauricio?), and why not thank you in advance for the new climbing adventures yet to come. For the laughs over teas, pints, drams, and sometimes serious conversations, but only in case of extreme necessity. Mauricio, gracias for being like this, thank you even for the stubbornness. Thank you for placing your desk strategically diagonal to mine, for the singing - especially Sabor a Mi and Quizás, Quizás, Quizás - and for making sure I am always fine. Obrigada Andy for sailing together in bonnie ships through the mayonnaise, for stupid pretty afternoons, and occasional correction of my enigmatic English even with your eyes closed. Also thank you to the nice people I met climbing, especially Berta Rámiro-Sanchez and Dave Cooper (a special thank you for all the trad jamming adventures so far and the ones ahead of us. I am looking forward to see the pretty flowers!).

Thank you Guilherme Braga for sharing 18 years - and counting - of adventures,

of love for the Sertão, and for taking care of me and vice-versa - more versa than vice. And of course for always remebering that there has to be an easier way. Thank you to my dear friends that are far for the moment. Leo Robson, Paulo, Marcell and meus coleguinhas da puc. Thank you Saci, muchiba,,, Thank you Pedro Taucce for all the songs, for the viola caipira, and for understanding. A saudade é uma estrada longa.

Merci Antonin Portelli for being my marmota bem pertinho, for supporting me and for the nice discussions about science with whisky and gü.

Muito obrigada para as pestinhas meus amores Samantha, Gabriela e Gianna. Obrigada Samantha for the help and the re-help with the figure. I owe you many many pizzas. Obrigada com muito amor Mariel e João, KK, Kelli, Dudu, Clarinha, Vô Ismael, Tajo e Tio Afonso, and all my family.

I am very thankful to my two examiners Andrew Hudson and Matt Lavin. Thank you for the precious input to this thesis, and for the pleasant and interesting discussion during my viva. I feel lucky to have had such a wonderful viva, the time passed too fast.

It seems that I arrived yesterday in Scotland. I brought a viola caipira, a typical instrument from the countryside of Brazil, thinking I would have plenty of time to learn it, especially during the cold winters of Edinburgh. Little did I know that the winters here are not so cold, and the viola caipira would spend most of its time on the corner. Not sure if Edinburgh is an incredible place or if I am terrible with musical instruments, or both. Thank you to the Royal Oak and Captain's folk for showing me a bit of the culture of Scotland. Thank you Jim for the book. Thank you to the bagpiper players on the corner of the Royal Mile, it is nice to hear you while cycling to work.

I would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the funding for my PhD, and the Davis Expedition Fund for funding the first field expedition.

Obrigada ao Sertão e ao sertanejo, que é, antes de tudo, um forte.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biogeographic history of the Neotropics

The Neotropics is the most diverse area in the world and the mechanisms that generated and maintain its biodiversity are in constant discussion. Through evolutionary time, the Neotropics experienced intense variation in climate and geology resulting in a great diversity of biomes, from deserts to tropical rain forests (Hughes et al., 2013). In the 1970s, plant distribution patterns were explained mainly in the light of the tectonic and geographic events (Raven and Axelrod, 1974). Later, in the 2000s, dated molecular phylogenies became more available. Those phylogenies helped to investigate the different levels of species diversity across the globe and the evolutionary history of biomes (Webb et al., 2002; Pennington et al., 2006; Donoghue, 2008; Cavender-Bares et al., 2009). The differences found between stem and crown ages of clades and geological events led to the questioning of tectonic events causing vicariance as the main driver of plant distribution patterns (Lavin et al., 2004; Pennington et al., 2006). Instead, dated phylogenies suggested that long-distance dispersal might play an important role in plant community assembly (Lavin et al., 2004; Renner, 2004; Pennington et al., 2006). Furthermore, distinctive plant diversification patterns were found in different biomes, suggesting that the age and ecological difference of the biomes also should be taken into account in biogeographic explanations (Wiens and Donoghue, 2004; Pennington et al., 2006). For example, ecological interactions over time and plant traits such as drought tolerance or the ability to survive fire might determine whether lineages can migrate (niche conservatism) or adapt (niche evolution) (Pennington and Lavin, 2016). In that sense, dated phylogenies of single lineages, considering the ecological part of the evolutionary and biogeographic process may be useful in studying neotropical diversification (Wiens and Donoghue, 2004; Pennington and Lavin, 2016). Endemic clades from two of the major neotropical biomes, seasonally dry tropical forests (SDTFs) and rain forests, show good examples of distinctive diversification history.

SDTFs occur on fertile soils, are characterized by a more or less continuous tree canopy, which becomes more open in the drier sites. They have strong seasonal changes, with plants shedding up to 90–95% of their leaves during the five to six month dry season

(Figure 1.1), and the flora lacks adaptation to fire (Murphy and Lugo, 1986; Pennington et al., 2009). This biome is one of the most threatened and has been one of the least studied ecosystems in the tropics over decades (Miles et al., 2006; DRYFLOR, 2016). It occurs in disjunct areas throughout the Neotropics, has elevated beta-diversity and high species endemism (Pennington et al., 2009; DRYFLOR, 2016) and dates back from the Middle Eocene (Pennington et al., 2009). Leguminosae and Bignoniaceae are dominant families in SDTF, but species from the Bombacoideae clade (Malvaceae), the subject of this thesis, are often common and distinctive.

Figure 1.1: Seasonally dry tropical forests (SDTFs) show striking seasonal changes, losing all the leaves during the dry season from June until October. Pictures from Mata Seca State Park, Minas Gerais, Brazil. Photos: F. Pezzini.

SDTF-confined clades often contain species resolved as monophyletic and with old species stem ages in DNA-sequence-based phylogenies (Pennington and Lavin, 2016). In addition, the geographically structured phylogenetic pattern characteristic of clades in this biome suggests dispersal-limited, old lineages maintained over evolutionary timescales in the stable ecological conditions of the biome (Pennington et al., 2010; Hughes et al., 2013). By contrast, tree clades confined to the Amazon rain forest, the largest tropical forest in the world that dates back to the Paleocene (Burnham and Johnson, 2004), are suggested to more often contain non-monophyletic species, young species stem ages and lack of geographical phylogenetic structure (Dexter et al., 2017). These rain forest patterns might be explained by frequent dispersal and subsequent successful colonization (Pennington and Lavin, 2016). The different and distinctive phylogenetic patterns between SDTF and Amazon rain forest suggests an interaction

of ecology and phylogeny over evolutionary timescales (Pennington et al., 2011; Pennington and Lavin, 2016; Dexter et al., 2017).

The general pattern of species age, monophyly and geographical structure are reported mostly for SDTF species belonging to the Leguminosae family (Pennington and Lavin, 2016), and it is not clear whether the different and distinctive phylogenetic patterns are general across all angiosperm families. To investigate this further, more taxonomic groups require phylogenetic studies. Ideally, phylogenies should be well resolved, highly supported and time-calibrated, including samples of all described species, and with multiple accessions per species.

This thesis describes such a phylogeny for the genus *Ceiba*, which occurs mostly in SDTF but also in the Amazon, as a case study to investigate biome-specific differences in the nature of species and their diversification trajectories in the Neotropics. The genus contains lineages with different geographical and ecological niches, thus representing an ideal case to shed light on patterns of neotropical plant evolution and diversification.

The neotropical genus *Ceiba* Mill. (including *Chorisia* Kunth) is part of Bombacoideae (Malvaceae), which comprises fewer than 250 species that are predominantly tropical trees including genera such as *Scleronema*, *Pseudobombax*, *Eriotheca*, *Cavanillesia*, *Ceiba* as well as the Old World *Adansonia* and *Bombax*. They are thought to have originated in the Neotropics and later dispersed to the Old World (Baum et al., 2004) with a minimum divergence time of 58-60 Ma, although the earliest fossil record occurs in North America (Carvalho et al., 2011).

Ceiba comprises 18 species divided into taxonomic sections Ceiba and Campylanthera (Schott & Endl.) K. Schum. based on morphological characters of pollen and staminal appendages. It is one of the most characteristic elements of many neotropical SDTFs. The genus has an historically complex taxonomy and some issues of species delimitation remain unresolved including a species complex (*Ceiba insignis* agg.) (Gibbs and Semir, 2003) where species boundaries are uncertain.

Ceiba are mostly trees with digitate leaves, with aculeate spines on the trunk and branches, and can vary from canopy emergent in seasonally flooded várzea forests in

Figure 1.2: Flowers of different species of *Ceiba*. a - *Ceiba pubiflora*, b - *Ceiba glaziovii*, c - *Ceiba speciosa*, d - *Ceiba pubiflora*, e - *Ceiba rubriflora*, f - *Ceiba erianthos*, g - *Ceiba jasminodora*, g - *C. samauma*. Photos: F. Pezzini.

the Amazon (*C. pentandra* 30 - 50m) to treelets in rocky outcrops (campos rupestres) in Minas Gerais, Brazil (*C. jasminodora* 1.5 - 2m). In some species (*C. chodatii*, *C. glaziovii*, *C. pubiflora*, *C. speciosa*) the trunk can be ventricose (swollen), justifying its vernacular name in Brazil, barriguda, which means "swollen belly". Most species are deciduous and flower when leafless (Figure 1.3). They occur mostly in SDTF, with the exception of the widespread species *C. samauma*, *C. speciosa* and *C. pentandra* that also occur in more humid environments, and *C. lupuna*, which is the only species with distribution known to be restricted to humid forests (Gibbs and Semir, 2003). Although being generally distinct morphologically, *Ceiba* species are thought to hybridise, especially in the *Ceiba insignis* aggregate where all members are hypothesised to be interfertile (Gibbs and Semir, 2003). *Ceiba* species often occur in sympatry, although with varied anthesis time. Pollinators are mostly bats, but also sphingid moths and

diurnal butterflies (Gibbs and Semir, 2003) (Figure 1.2).

Figure 1.3: *Ceiba rubriflora* flowering during the dry season in Peruaçu National Park, Minas Gerais, Brazil. Photo: F. Pezzini.

1.2 Sanger and Next-generation sequencing

Nuclear ribosomal internal transcribed spacers (ITS) sequences have been widely explored using Sanger sequencing to help elucidate relationships among angiosperm genera and species, including in Bombacoideae (Baum et al., 1998; Duarte et al., 2011; Carvalho-Sobrinho et al., 2016), and even to investigate genetic structure among populations (Dick et al., 2007). In spite of having drawbacks related to paralogous copies (Buckler et al., 1997; Álvarez and Wendel, 2003), ITS can still play an important role in the investigation of species relationships if analysed carefully, for example identifying pseudogenes and assessing orthology in the case of intra-individual polymorphism (Bailey et al., 2003; Feliner and Rosselló, 2007). However, phylogenetic analysis of closely related species belonging to a recently diversified clade might be challenging if based on a single or a few loci. A single locus gives only one gene tree that may not reflect the

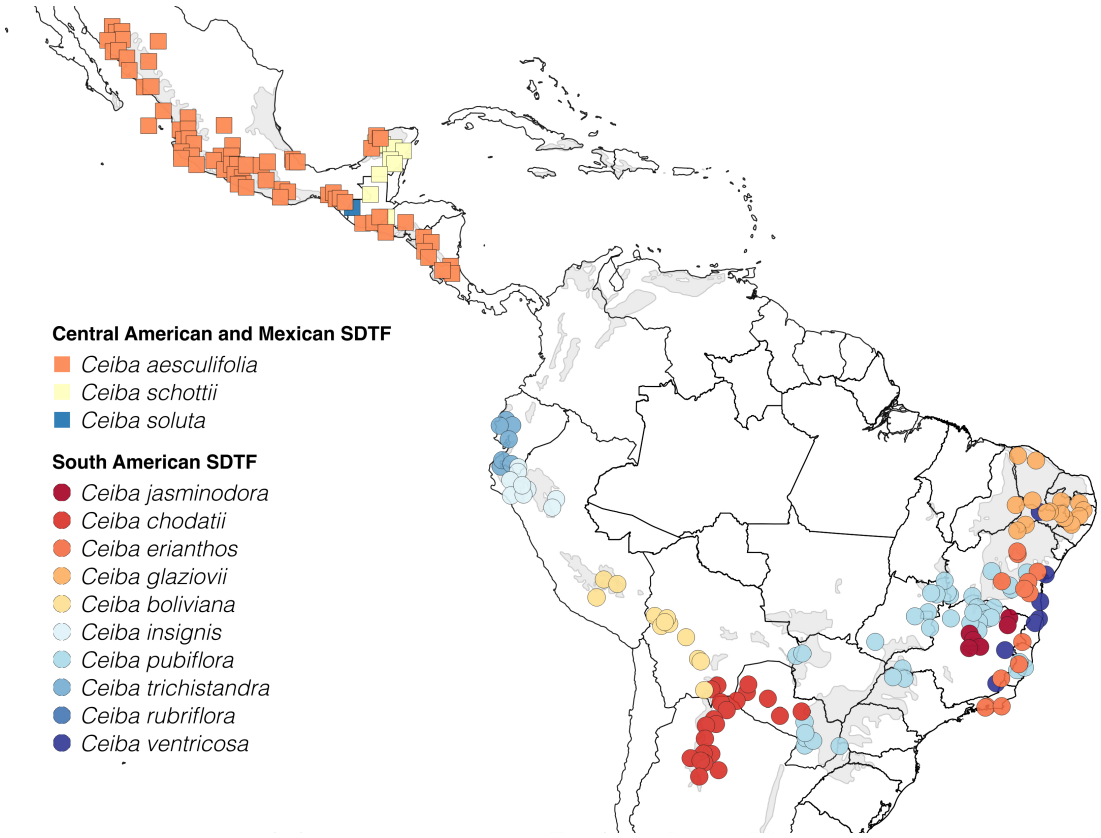


Figure 1.4: Distribution of 13 *Ceiba* species from SDTFs in Central America and North America, and South America. Grey areas represent the distribution of SDTF following (DRYFLOR, 2016). Occurrence records adapted from Gibbs and Semir (2003).

true evolutionary history of a clade. Different loci may show incongruent phylogenetic patterns due to low rates of mutation over short evolutionary time periods since species divergence, retained ancestral polymorphism, hybridisation and/or incomplete lineage sorting (Koopman and Baum, 2010).

New DNA sequencing approaches offer great promise to improve phylogenetic resolution by the use of multiple loci. Next-generation sequencing (NGS) covers a diversity of new techniques that enable the sequencing of hundreds of independent nuclear loci, as well as loci from the chloroplast genome, which together can provide many phylogenetically informative characters. This thesis explores the potential of one of the many next generation sequencing techniques, hybrid capture, for *Ceiba* phylogenetics.

In the hybrid capture technique, libraries containing fragments of genomic DNA (gDNA) are mixed with "baits" representing the target loci in a hybridisation reaction.

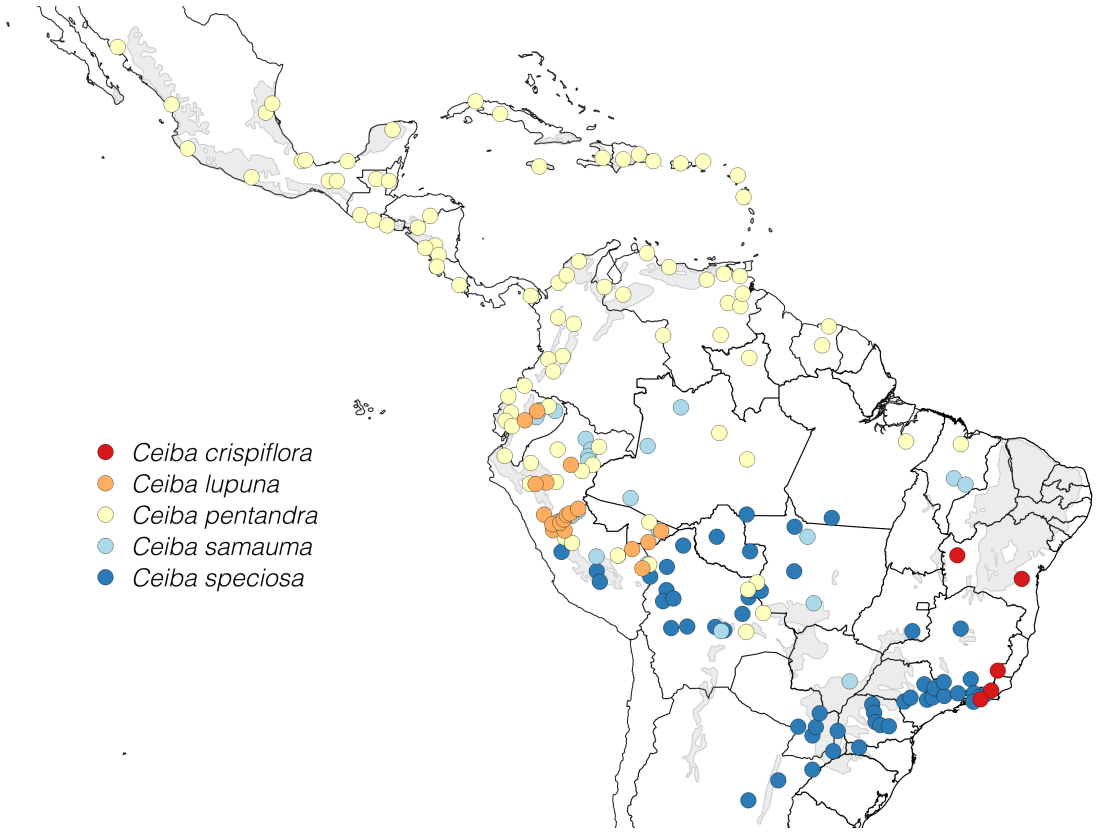


Figure 1.5: Distribution of 5 *Ceiba* species from rain forests from Latin America. Grey areas represent the distribution of SDTF following (DRYFLOR, 2016). Occurrence records adapted from Gibbs and Semir (2003).

After hybridisation, the fragments that did not bind to the baits are washed away and the remaining fragments, which bound to the baits and represent the loci of interest, are sequenced (Figure 1.6).

The target loci have varying length and are normally designed from transcriptomes from one or more chosen species and therefore represent coding regions. The baits are 80-120bp biotinylated RNA fragments. Those fragments are representative sequences covering the entire length of each locus. Normally they are tiled, which means that the bait sequences overlap and therefore that each part of the locus is covered at least twice by the baits. They hybridise with the DNA library fragments at an optimum temperature. The baits are provided in excess to the solution so every possible fragment of library can hybridise with them. The baits allow for mismatches, which is important for capturing flanking regions of the fragments containing introns (Altmann et al.,

2012) that can provide a greater number of informative characters for phylogenetic analysis. A bait set containing intronic regions might improve capture success because the intronic parts of the library fragments might match the bait fragment. Likewise, such a bait set, when used as a reference, might also improve the mapping quality during post-sequence assembly of the data. When baits targeting only coding regions are used as reference to assemble the reads, genomic sequences containing intronic regions might not be considered a good match and would be discarded, representing a waste of sequencing effort.

The choice of the target loci and thus of the bait set is one of the first and most important steps in the target enrichment technique since it influences the hybridisation success directly. For phylogenetics, baits should be designed targeting single-copy, orthologous loci present across the studied taxa. However, if pseudogenes or multiple copy function genes are present, all copies could still be enriched (Fu et al., 2010; Saintenac et al., 2011; Chau et al., 2018) and this problem of paralogy should be dealt with during data analysis.

1.3 Objectives

This thesis aims to investigate the evolutionary history and interspecific relationships within the genus *Ceiba* to gain insights into the evolution of neotropical SDTF, integrating both ecological and molecular approaches. To meet these goals, I will first construct a Sanger-sequence phylogeny for *Ceiba*. This will be used as a basis to evaluate a well-resolved, multi-locus and densely sampled species-level phylogenetic tree for *Ceiba* using the NGS technique hybrid bait capture. The NGS phylogeny will be used to investigate phylogenetic relationships amongst species of *Ceiba*, species boundaries in the genus and diversification events. Specifically, I aim to:

- Chapter 2: Investigate the evolutionary history and interspecific relationships within the genus *Ceiba* using a Sanger-sequence phylogeny of the ITS region, to assess whether the *Ceiba* phylogeny is geographically or ecologically structured and if species confined to SDTFs are resolved differently in the phylogeny as

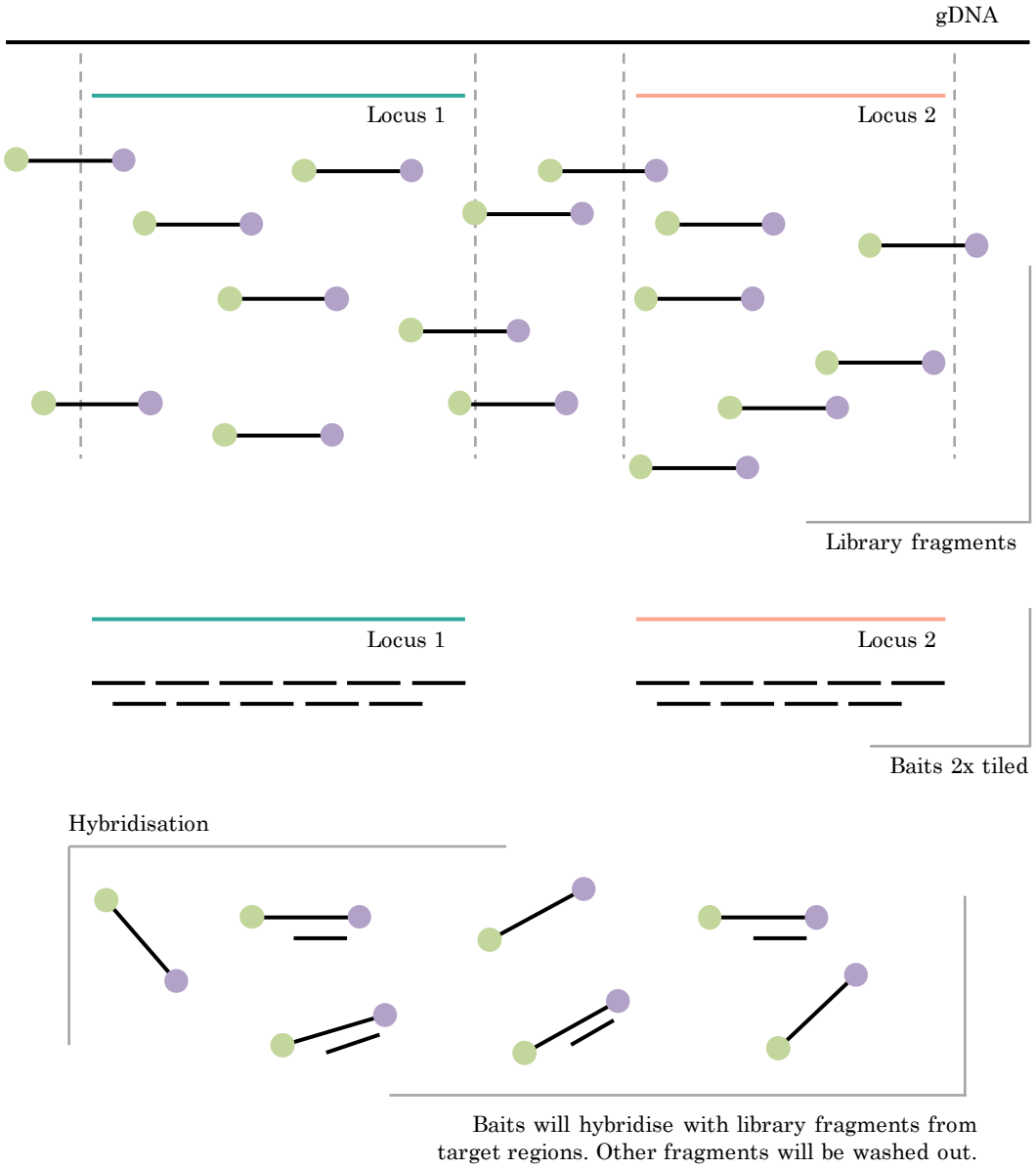


Figure 1.6: Target capture scheme.

compared with rain forest species;

- Chapter 3: Construct a well-resolved, multi-locus and densely sampled species-level phylogenetic tree for *Ceiba* using the NGS technique hybrid bait capture, in which I evaluate the biological implications of different methodologies to assemble NGS data;

- Chapter 4: Investigate possible discordances amongst the individual gene phy-

logenies generated using hybrid bait capture in order to refine the phylogenetic hypothesis for *Ceiba* and to contribute to better species delimitation in the genus;

- Chapter 5: Provide a temporal framework to investigate diversification in *Ceiba*, which I will use to:

  (i) Investigate the nature of species in different biomes e.g., are species confined to SDTFs resolved as monophyletic on long stem lineages contrasting with younger, non-monophyletic rain forest species?

  (ii) Make inferences about the biogeographic history of SDTFs in the neotropical region.

Chapter 2 investigates the evolutionary history and relationships between the species of *Ceiba*, and uses the biogeographic history of the genus to gain insights into the evolution and ecology of neotropical SDTF. Specifically, I aim to assess whether the *Ceiba* phylogeny is geographically or ecologically structured and if species confined to SDTFs are resolved differently in the phylogeny as compared with rain forest species. To address this topic, I use a Sanger-sequence based phylogeny of the ITS region for 24 accessions representing 14 of the 18 species described for *Ceiba*, with multiple individuals per species for five species. The Sanger-sequence phylogeny presented in this chapter also provides a framework to evaluate the results using next-generation hybrid capture sequencing both for phylogenetics (Chapter 3) and for dating diversification events (Chapter 5) in subsequent chapters.

Chapter 3 produces a better resolved and sampled phylogeny for *Ceiba*. The Sanger-sequence phylogeny based on the ITS region from Chapter 2 failed to resolve all the species relationships, especially in the South American SDTF clade. Therefore, I applied a next-generation hybrid capture technique aiming to increase the resolution of the phylogeny. I sequenced 377 nuclear loci for 103 samples representing all species of the genus, with multiple individuals per species for 17 out of the 18 species of *Ceiba*. This chapter focuses on methodological issues to generate and analyse NGS data. I explore different bait sets to generate target capture data and variations in software

and pipelines to assemble the millions of short sequences generated, and the impact of the different results in each step of downstream analysis. I evaluate the consequences of those variations on the final phylogenetic analysis and hence their implications for inferences of comparative biology.

In chapter 4, I investigate the possible reasons for the variation in topology of the phylogenetic trees generated in Chapter 3 and the lack of monophyly of the species in the South American SDTF clade (Chapters 2 and 3). Specifically, I aim to explore the possible incongruence amongst the 377 individual gene trees sequenced with the target capture technique. I then investigate species boundaries in the South American SDTF clade by combining modern species delimitation analysis and morphological data under a coalescent framework, which accounts for discordance amongst individual gene trees. Under the same framework, I build a species tree for *Ceiba* to investigate further the relationship amongst the species.

In chapter 5, I investigate the evolutionary history of *Ceiba*, and use the biogeographic history of the genus to gain insights into the evolution and ecology of neotropical SDTF and whether species of *Ceiba* occurring in different biomes are resolved differently in the phylogeny. This builds of the results of Chapter 2, but uses a well resolved and densely sampled phylogenetic tree, built after a thorough sequence assembly (Chapter 3) and phylogenetic inference under frameworks that allow for gene tree incongruence (Chapter 4).

The final chapter synthesises the main conclusions from the four data chapters of this thesis, discussing whether Sanger sequence analysis is still relevant, the current challenges of analysing NGS data, and the contributions of this study to understanding the biogeographic history of the different major neotropical biomes, especially the threatened Seasonally Dry Tropical Forest.

Bibliography

Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B. 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics*. 131:1541–1554.

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*. 29:417–434.

Bailey CD, Carr TG, Harris SA, Hughes CE. 2003. Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. *Molecular Phylogenetics and Evolution*. 29:435–455.

Baum DA, DeWitt Smith S, Yen A, Alverson WS, Nyffeler R, Whitlock BA, Oldham RL. 2004. Phylogenetic relationships of Malvatheca (Bombacoideae and Malvoideae; Malvaceae sensu lato) as inferred from plastid DNA sequences. *American Journal of Botany*. 91:1863–1871.

Baum DA, Small RL, Wendel JF. 1998. Biogeography and Floral Evolution of Baobabs Adansonia, Bombacaceae as Inferred From Multiple Data Sets. *Systematic Biology*. 47:181–207.

Buckler ESI, Ippolito A, Holtsford TP. 1997. The evolution of plant ribosomal DNA: Divergent paralogues, pseudogenes and phylogenetic implications. *Genetics*. 145:821–832.

Burnham RJ, Johnson KR. 2004. South American palaeobotany and the origins of neotropical rainforests. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. 359:1595–1610.

Carvalho MR, Herrera Fa, Jaramillo Ca, Wing SL, Callejas R. 2011. Paleocene Malvaceae from northern South America and their biogeographical implications. *American Journal of Botany*. 98:1337–1355.

Carvalho-Sobrinho JG, Alverson WS, Alcantara S, Queiroz LP, Mota AC, Baum DA. 2016. Revisiting the phylogeny of Bombacoideae (Malvaceae): Novel relationships, morphologically cohesive clades, and a new tribal classification based on multilocus phylogenetic analyses. *Molecular Phylogenetics and Evolution*. 101:56–74.

Cavender-Bares J, Kozak KH, Fine PVA, Kembel SW. 2009. The merging of community ecology and phylogenetic biology. *Ecology Letters*. 12:693–715.

Chau JH, Rahfeldt WA, Olmstead RG. 2018. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Applications in Plant Sciences*. 6:e1032.

Dexter KG, Lavin M, Torke BM, Twyford AD, Kursar TA, Coley PD, Drake C, Hollands R, Pennington RT. 2017. Dispersal assembly of rain forest tree communities across the Amazon basin. *Proceedings of the National Academy of Sciences*. 114:2645–2650.

Dick CW, Bermingham E, Lemes MR, Gribel R. 2007. Extreme long-distance dispersal of the lowland tropical rainforest tree Ceiba pentandra L. (Malvaceae) in Africa and the Neotropics. *Molecular Ecology*. 16:3039–3049.

Donoghue MJ. 2008. A phylogenetic perspective on the distribution of plant diversity. *Proceedings of the National Academy of Sciences*. 105:11549–11555.

DRYFLOR. 2016. Plant diversity patterns in neotropical dry forests and their conservation implications. *Science*. 353:1383–1387.

Duarte MC, Esteves GL, Salatino MLF, Walsh KC, Baum DA. 2011. Phylogenetic Analyses of Eriotheca and Related Genera (Bombacoideae, Malvaceae). *Systematic Botany*. 36:690–701.

Feliner GN, Rosselló JA. 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*. 44:911–919.

Fu Y, Springer NM, Gerhardt DJ, et al. (18 co-authors). 2010. Repeat subtraction-mediated sequence capture from a complex genome. *The Plant Journal*. 62:898–909.

Gibbs P, Semir J. 2003. A taxonomic revision of the genus Ceiba Mill. (Bombacaceae). *Anales del Jardín Botánico de Madrid*. 60:259–300.

Hughes CE, Pennington RT, Antonelli A. 2013. Neotropical Plant Evolution: Assembling the Big Picture. *Botanical Journal of the Linnean Society*. 171:1–18.

Koopman MM, Baum DA. 2010. Isolating Nuclear Genes and Identifying Lineages withoutMonophyly: An Example of Closely Related Species from Southern Madagascar. *International Journal of Plant Sciences*. 171:761–771.

Lavin M, Schrire BP, Lewis G, Pennington RT, DelgadoSalinas A, Thulin M, Hughes CE, Matos AB, Wojciechowski MF. 2004. Metacommunity process rather than continental tectonic history better explains geographically structured phylogenies in legumes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. 359:1509–1522.

Miles L, Newton AC, DeFries RS, Ravilious C, May I, Blyth S, Kapos V, Gordon JE. 2006. A global overview of the conservation status of tropical dry forests. *Journal of Biogeography*. 33:491–505.

Murphy PG, Lugo AE. 1986. Ecology of Tropical Dry Forest. *Annual Review of Ecology and Systematics*. 17:67–88.

Pennington RT, Daza A, Reynel C, Lavin M. 2011. Poissonia eriantha (Leguminosae) From Cuzco, Peru: An Overlooked Species Underscores a Pattern of Narrow Endemism Common to Seasonally Dry Neotropical Vegetation. *Systematic Botany*. 36:59–68.

Pennington RT, Lavin M. 2016. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytologist*. 210:25–37.

Pennington RT, Lavin M, Oliveira-Filho A. 2009. Woody Plant Diversity, Evolution, and Ecology in the Tropics: Perspectives from Seasonally Dry Tropical Forests. *Annual Review of Ecology, Evolution, and Systematics*. 40:437–457.

Pennington RT, Lavin M, Sarkinen T, Lewis GP, Klitgaard BB, Hughes CE. 2010. Contrasting plant diversification histories within the Andean biodiversity hotspot. *Proceedings of the National Academy of Sciences*. 107:13783–13787.

Pennington RT, Richardson JE, Lavin M. 2006. Insights into the historical construction of species-rich biomes from dated plant phylogenies, neutral ecological theory and phylogenetic community structure. *New Phytologist*. 172:605–616.

Raven PH, Axelrod DI. 1974. Angiosperm Biogeography and Past Continental Movements. *Annals of the Missouri Botanical Garden*. 61:539.

Renner S. 2004. Plant Dispersal across the Tropical Atlantic by Wind and Sea Currents. *International Journal of Plant Sciences*. 165:S23–S33.

Saintenac C, Jiang D, Akhunov ED. 2011. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biology*. 12:R88.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. 2002. Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics*. 33:475–505.

Wiens JJ, Donoghue MJ. 2004. Historical biogeography, ecology and species richness. *Trends in Ecology & Evolution*. 19:639–644.

Chapter 2

Phylogeny and biogeography of *Ceiba* Mill. (Malvaceae, Bombacoideae)

To be submitted to Biotropica as "Phylogeny and biogeography of *Ceiba* Mill. (Malvaceae, Bombacoideae)". Flávia Fonseca Pezzini, Kyle G. Dexter, Jefferson G. de Carvalho-Sobrinho, Catherine A. Kidner, James A. Nicholls, Luciano P. de Queiroz, R. Toby Pennington.

2.1 Introduction

The Neotropics is the most species-rich region in the world and the mechanisms that generated and maintain its biodiversity are under constant discussion. Through evolutionary time, the Neotropics has been climatically and geologically dynamic, resulting in a great diversity of biomes, from deserts to tropical rain forests (Hughes et al., 2013; Rangel et al., 2018). To understand the history and dynamics of those biomes, molecular phylogenetic and phylogeographic approaches have been used, because the relationships of taxa allow inferences to be made of the historical relationships amongst biomes and areas (Pennington et al., 2006). In recent years, the dichotomy regarding the "cradle" vs. "museum" debate (Stebbins, 1974) explaining neotropical diversity has given way to a more nuanced approach, considering plant diversification patterns that may be recent, old, slow and rapid, even within individual clades (Hughes et al., 2013; Koenen et al., 2015). As suggested in the literature more than 10 years ago (Wiens and Donoghue, 2004; Pennington et al., 2006), this heterogeneity in diversification timing and rate within and among clades may be related not only to climatic and geological events, but also to the age and ecological differences of the biomes. For example, geologically old biomes (e.g., rain forest) are likely to have provided lineages that colonised newer biomes (e.g., savannas) and the relative difficulty of evolving adaptations such as drought tolerance or the ability to survive fire might determine whether a lineage can adapt to a new biome (niche evolution; Simon et al. 2009; Pennington and Lavin 2016) or remains confined to the same biome (niche conservatism) over evolutionary timescales (Crisp et al., 2009).

Clades endemic to two of the major neotropical biomes, seasonally dry tropical forests (SDTFs) and rain forests, give good examples of different and distinctive phylogenetic patterns, suggesting an interaction of ecology and phylogeny over evolutionary timescales (Pennington et al., 2011; Pennington and Lavin, 2016; Dexter et al., 2017). SDTFs occur on fertile soils and are characterized by the absence of fire adaptation in the flora and a predominantly continuous tree canopy, which becomes more open in the drier sites, with plants shedding up to 90–95% of their leaves during the five to six

month long dry season (Murphy and Lugo, 1986; Pennington et al., 2009). This biome has been one of the least studied, but is one of the most threatened in the tropics (Miles et al., 2006; DRYFLOR, 2016). It occurs in disjunct areas throughout the Neotropics and has elevated beta-diversity and high plant species endemism (Pennington et al., 2009; DRYFLOR, 2016). Leguminosae and Bignoniaceae are often the most species rich and dominant families in SDTF, but species from the Bombacoideae clade (Malvaceae), the subject of this study, are often common and distinctive.

SDTF-confined clades contain species that often resolve as monophyletic in DNA-sequence-based phylogenies and with old stem ages (Pennington and Lavin, 2016). In addition, the geographically structured phylogenetic pattern characteristic of clades in this biome suggests dispersal-limited, old lineages maintained over evolutionary timescales by the stable ecological conditions of the biome (Pennington et al., 2010; Hughes et al., 2013). By contrast, tree clades confined to the Amazon rain forest, the largest tropical forest in the world, are suggested to contain non-monophyletic species more often, and to have young species stem ages and lack of geographical phylogenetic structure (Dexter et al., 2017). These rain forest patterns might be explained by frequent dispersal and subsequent successful colonization (Pennington and Lavin, 2016) (Figure 2.1).

The neotropical genus *Ceiba* Mill. (Malvaceae: Bombacoideae) comprises 18 species divided into taxonomic sections Ceiba and Campylanthera (Schott & Endl.) K. Schum. based on morphological characters of pollen and staminal appendages. It is one of the most characteristic elements of many neotropical SDTFs. However, it also contains species confined to the Amazon rain forest and is therefore a good case study to investigate biome-specific differences in the nature of species and their diversification trajectories.

Ceiba species have digitate leaves, aculeate spines on the trunk and branches and can vary from 50 m canopy emergents in seasonally flooded vrzea forests in the Amazon (*C. pentandra*) to 2m treelets on rocky outcrops (campos rupestres) in Minas Gerais, Brazil (*C. jasminodora*). In some species (*C. chodatii*, *C. pubiflora*, *C.glaziovii*, *C.*

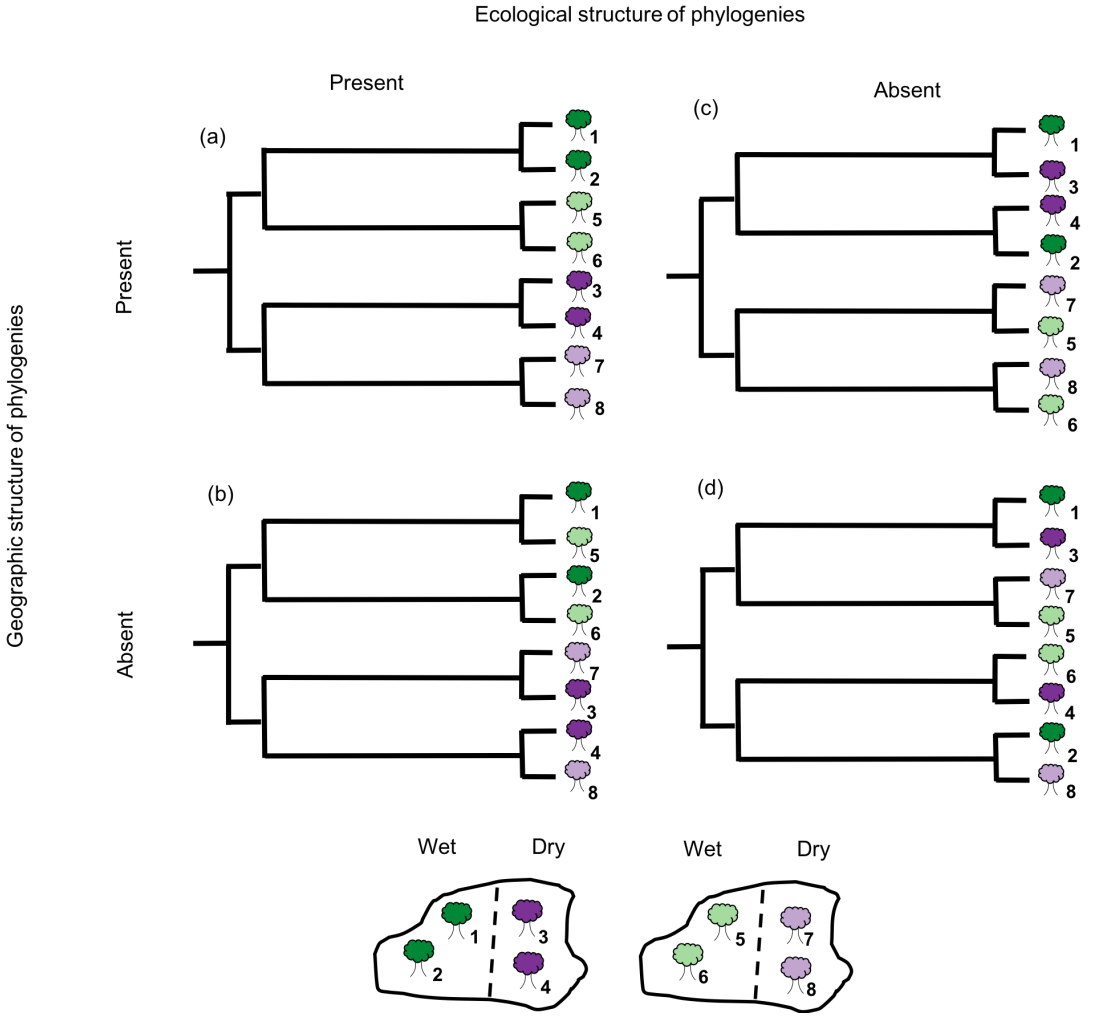


Figure 2.1: Two hypothetical islands, each with an area of seasonally dry tropical forest and rain forest. In total eight different species occur in the two biomes and the different islands and are represented by different colours (dark green species 1, dark green species 2, dark purple 3, etc.). Hypothetical phylogenies showing patterns of presence or absence of geographical and ecological structure. Phylogeny (a) shows ecological (species from same biome are grouped in the same clade) and geographical structure (species from the same island are grouped in the same clade). Phylogeny (b) shows ecological structure since species from SDTFs are grouped together in a clade and rain forest species form a different clade, but no geographical structure because within each clade representing ecological preference, species occurring in different islands are in the same clade. Phylogeny (c) shows no ecological structure since each species from SDTF is recovered as sister to a species from rain forest, but shows geographical structure since within each clade species from the same island are recovered as sisters. Phylogeny (d) shows no geographical or ecological structure. (Modelled after Graham and Fine 2008)

speciosa) the trunk can be ventricose (swollen), explaining its vernacular names barriguda (swollen belly; Brazil) and palo borracho (drunken tree; Peru). Most species are deciduous and flower when leafless. They occur mostly in SDTF (Figure 1.4), with the exception of the widespread *C. samauma*, *C. speciosa* and *C. pentandra* that also occur in more humid environments, and *C. lupuna*, which is the only species restricted to rain forests (Figure 1.5).

Previous Bayesian analyses of DNA sequence data from the nuclear ribosomal internal and external transcribed spacers (ITS and ETS) and plastid markers (matK, trnL-F, trnS-trnG) for 13 species recovered *Ceiba* as monophyletic and sister to *Neobuchia paulinae* (Duarte et al., 2011; Carvalho-Sobrinho et al., 2016). Together with *Spirotheca*, *Pochota fendleri* sensu Alverson and Duarte (2015), and *Pseudobombax*, these taxa form the well supported striated bark clade (Carvalho-Sobrinho et al., 2016). The ITS region was the most informative locus for the genera related to *Ceiba* focused on the studies of Duarte et al. (2011) and Carvalho-Sobrinho et al. (2016). However, relationships within *Ceiba* were poorly resolved and only one individual per species was included in the phylogeny. *Ceiba* has a historically complex taxonomy with species boundaries still confused, which is aggravated by the fact that herbarium specimens are often incomplete because individuals produce flowers and fruits when leafless. Therefore, a well sampled phylogeny with multiple accessions per species could be a useful tool to explore the nature of species in *Ceiba*.

The aim of this chapter is to investigate the evolutionary history and interspecific relationships within the genus *Ceiba*, and use the biogeographic history of the genus to gain insights into the evolution and ecology of neotropical SDTF. I also aim to assess whether the *Ceiba* phylogeny is geographically or ecologically structured and if species confined to SDTFs are resolved differently in the phylogeny as compared with rain forest species (i.e., monophyletic on long stem lineages).

The Sanger-sequence phylogeny of the ITS region presented in this chapter also provides a framework to evaluate the results presented in subsequent chapters using next-generation hybrid capture sequencing both for phylogenetics and for dating diver-

sification events.

2.2 Methods

2.2.1 Taxon sampling

I present the best sampled phylogeny of the genus *Ceiba* to date, covering 24 accessions representing 14 of the 18 species described for the genus. Critically, it samples multiple individuals per species for five species. As outgroups, I included 10 accessions representing species of the closest sister clades (Carvalho-Sobrinho et al., 2016): *Pseudobombax*, *Spirotheca*, *Eriotheca*, and *Pochota fendleri*. The full data set represents a combination of new sequence data from field surveys as well as from herbarium specimens, doubling the number of accessions of *Ceiba* in relation to the previous study by Carvalho-Sobrinho et al. (2016).

2.2.2 DNA sequence data

I used the ITS region to investigate species relationships in *Ceiba*. In Bombacoideae, this region has been widely explored to help elucidate relationships among genera and species (Baum et al., 1998; Duarte et al., 2011; Carvalho-Sobrinho et al., 2016), and to investigate genetic structure among populations (Dick et al., 2007). In spite of having drawbacks related to paralogous copies (Buckler et al., 1997; Álvarez and Wendel, 2003), ITS can still play an important role in the investigation of species relationships if analysed carefully, for example identifying pseudogenes and assessing orthology in the case of intra-individual polymorphism (Bailey et al., 2003; Feliner and Rosselló, 2007).

Genomic DNA extraction was performed for 36 herbarium and silica-gel dried leaf samples using Qiagen DNeasy Plant Mini Kits following the manufacturer's protocol, with the following changes: twice the volume of buffer AP1 in addition to a pinch of PVPP (polyvinyl polypyrrolidone) added at the lyse step followed by an incubation of 30 minutes; addition of 1μL of Riboshredder in the lysate solution followed by incubation at 37°C for 20 minutes; addition of twice the volume of buffer P3; and final

elution in 46μL of EB buffer run through the column twice to increase yield. Each 20μL PCR amplification reaction contained 0.5μL of template, 2μL of dNTPs (2mM), 2μL of 10x reaction buffer, 1μL of $MgCl_2$ (50mM), 0.65μL of both forward primer and reverse primer solutions (10μM), 0.1μL of Taq polymerase, 4μL of CES buffer and 9.1μL of ddH_2O. Amplification followed the same procedure described in Carvalho-Sobrinho et al. (2016). Samples were submitted to the Edinburgh Genomics laboratory at the University of Edinburgh for sequencing. For low quality sequences, we tested variations of the protocol (e.g. diminishing the amount of template in the PCR reaction or varying the sequencing primer). High quality sequences were recovered for 13 out of the 36 samples from which DNA was extracted for the ITS region.

All the inter-accession polymorphisms detected were validated visually by checking the electropherograms. Sequences were edited with Sequencher 5.4.1 (Gene Codes Corp., Ann Arbor, Michigan) and alignments were performed manually in Mesquite (Maddison and Maddison, 2015). I investigated the potential presence of ITS pseudogenes by comparing substitution rates along branch lengths in phylogenies generated using separated matrices representing the 5.8S (conserved region) and the ITS 1 and ITS 2 regions (fast evolving regions), following Bailey et al. (2003).

2.2.3 Phylogenetic analysis and molecular dating

I implemented maximum likelihood (ML) and Bayesian Inference (BI) analysis. I used MrModelTest to determine the best fitting model of sequence evolution. RAxML (Stamatakis, 2014) was used to run the ML with 1,000 bootstrap replicates.

BEAST2 (Bouckaert et al., 2014) was used to perform BI analysis and temporally calibrate the phylogeny. I ran two independent runs of 10 million generations with the following settings: General Time Reversible (GTR) plus Gamma model of sequence evolution; a relaxed lognormal molecular clock, and a Yule Model prior for tree branching. I sampled every 1,000 generations and visually inspected convergence of MCMC and ensured effective sample size $>$ 200 for all parameters of each run using Tracer v1.6. Trees were summarized in TreeAnnotator with a burn-in of 10% of trees for each

run.

I used the fossil flower of *Eriotheca prima* (Duarte, 1974) from the middle to late Eocene (de Lima and Salard-Cheboldaeff, 1981) of Brazil as a primary calibration for our BEAST2 analysis. The flower was identified as *Eriotheca* based on its small size (*Bombacopsis* and *Pachira* have larger flowers) and androecium organisation, which is a synapomorphy for the extant species of the genus (Robyns, 1963; Duarte et al., 2011; Carvalho-Sobrinho et al., 2016). Because the dating of this fossil is imprecise (middle to late Eocene: 33-56 mya, I assigned the approximate mean of these ages (47 Ma) as a minimum age. This calibration was applied to the stem node (Renner, 2005; Pennington et al., 2006) of *Eriotheca* as resolved in the phylogeny of Bombacoideae of Carvalho-Sobrinho et al. (2016), which samples genera related to *Ceiba* thoroughly and is the crown node of the clade comprising *Eriotheca*, *Spirotheca*, *Pseudobombax*, *Pochotoa fendleri* and *Ceiba*. I used a log-normal distribution with a mean of 0.15 and standard deviation of 1.5. In order to explore the effects of using the minimum and maximum ages of the *Eriotheca* fossil on phylogenetic age estimates, I also ran analyses assigning minimum ages of 33 Ma and 56 Ma to the *Eriotheca* stem (Figures 2.4 and 2.5). I followed the dates on the Geologic Time Scale v. 5.0 (Gradstein et al., 2012).

2.2.4 Phylogenetic signal test

I tested for strength of phylogenetic signal for the binary traits related to ecology (rain *versus* dry forests) and geography (Central and North America *versus* South America) using the D value proposed by Fritz and Purvis (2010), and implemented using the Caper package (v. 1.0.1) (Orme, 2013) in R, with 5,000 permutations. Under a null model of Brownian motion evolution of a binary trait, D has an expected value of 0. A negative D value indicates a strongly clustered phylogenetic pattern for a given binary trait (perhaps due to some process of evolutionary constraint), a value of one indicates a completely random pattern with respect to the phylogeny (i.e. no correlation between phylogeny and the trait at all) and values above one indicate an overdispersed phylogenetic pattern (perhaps due to divergent selection).

2.3 Results

The total length of the aligned sequences was 780 nucleotides, of which 263 were variable and 168 (22%) were parsimony-informative characters. The ML and BI trees showed congruent topologies (Figures 2.2 and 2.3). *Ceiba* was strongly supported as monophyletic, with posterior probability (pp) >= 0.95 and bootstrap value = 100 and was recovered as sister to *Pseudobombax* (Figure 2.2). Using the 47 Ma fossil calibration, the stem node age of *Ceiba* is 35.5 (21.547.3 [95% Highest Posterior Density (HPD)]) million years old (Ma) and the crown node age is 19.1 (11.228.4 [95% HPD]) Ma (Figure 2.3).

Ceiba comprises three main clades: (i) a humid forest lineage of the three accessions of *C. pentandra*, which are strongly supported as monophyletic [posterior probability (pp) = 1 and bootstrap value = 100] and sister to the remaining species and with stem node age of 19.1 (11.228.4 [95% HPD]) million years old (Ma) and crown node age of 5.3 (1.710.6 [95% HPD]) Ma.; (ii) a highly supported clade [posterior probability (pp) = 1 and bootstrap value = 100] composed of *C. schottii* and *C. aesculifolia* from Central American and Mexican SDTF plus two accessions of *C. samauma* from inter-Andean valleys in Peru, with stem node age of 17.0 (9.825.0 [95% HPD]) million years old (Ma) and crown age of 10.4 (5.016.7 [95% HPD]) Ma; and (iii) a highly supported [posterior probability (pp) = 1 and bootstrap value = 86] South American SDTF clade including 10 species showing little sequence variation, with stem node age of 17.0 (9.825.0 [95% HPD]) million years old (Ma) and crown node of 12.1 (6.419.0 [95% HPD]) Ma. Within this South American clade, neither *C. rubriflora* nor *C. pubiflora*, which were represented by multiple accessions, were resolved as monophyletic. In contrast, *Ceiba insignis* was resolved as monophyletic in the BI phylogeny, though with low posterior probability. The South American clade contains SDTF species, except for *C. lupuna*, a species with a distribution restricted to rain forest (Figures 2.2 and 2.3).

The D test shows significant phylogenetic signal for both ecological preference (D = -0.01122621, P (D=1) = 0.0098, P (D=0) = 0.4914) and geographical occurrence (D= -0.03625205, P (D=1) = 0.0178, P (D=0) = 0.537). Both D values are statistically

indistinguishable from zero, which indicates that closely related species are more likely to show the same ecological preference or geographical occurrence, as expected under a Brownian model of evolution, whereby there would be a constant rate of state switching over time and any given lineage is more likely to stay within the same biome and geographic region per unit time than to switch to the alternative biome or geographic region.

The phylogeny supports the monophyly of the two sections of the genus, Ceiba and Campylanthera, which are based on pollen and staminal appendages characters. However, it does not support monophyly of the insignis species complex. This species aggregate includes seven species (*C. pubiflora*, *C. chodatii*, *C. insignis*, *C. ventricosa*, *C. lupuna*, *C. speciosa* and *C. crispiflora*, indicated as bold in Figure 2.2) characterized by their entire staminal tube terminating in a collar of anthers, with the exception of *C. pubiflora* which has free stamens.

2.4 Discussion

Taxonomic implications

Our data support: (i) the circumscription of *Chorisia* within *Ceiba*, as proposed by Gibbs et al. (1988); Ravenna (1998); Gibbs and Semir (2003) and confirmed by recent molecular phylogenetic studies (Carvalho-Sobrinho et al., 2016); (ii) the non-monophyly of the *C. insignis* aggregate species proposed by Gibbs and Semir (2003). Our data suggest that *C. boliviana*, *C. erianthos* and *C. rubriflora*, not included by Gibbs and Semir (2003), are also part of this clade. It was suggested that those species are interfertile but also diverge in time of anthesis and pollinator type (Gibbs and Semir, 2003). Five of the seven species within this complex are restricted to the SDTF patches of South America, while *C. speciosa* is widespread and *C. lupuna* occurs in riverine rain forests in the Peruvian and Brazilian Amazon (Figure 1.4); and (iii) the monophyly of the section Campylanthera (Gibbs and Semir, 2003) that includes the Central American species *C. aesculifolia*, *C. schottii* and the widespread *C. samauma*.

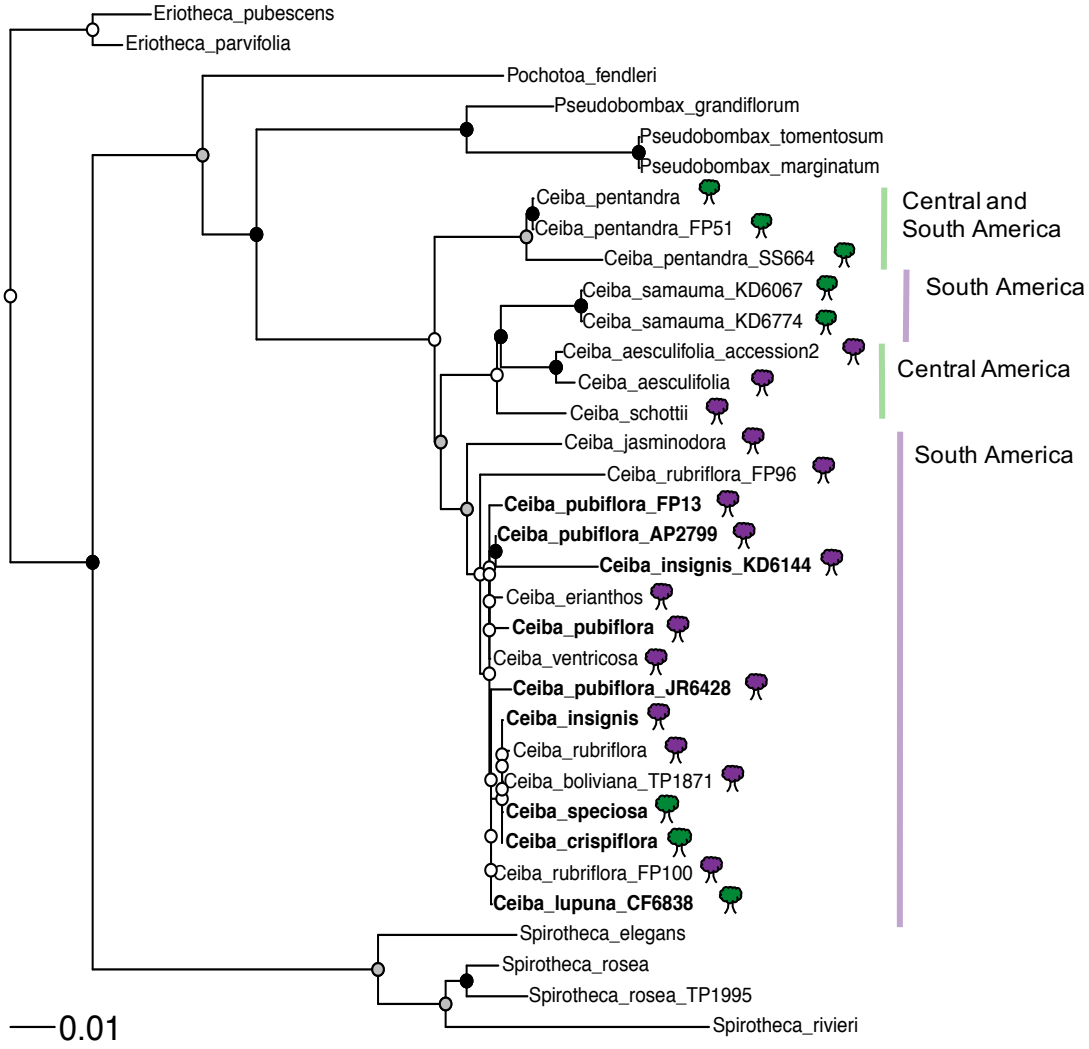


Figure 2.2: Maximum likelihood phylogram derived from analysis of nuclear ribosomal ITS sequence data sets for 14 species of *Ceiba*. Species in bold belong to the *Ceiba insignis* species aggregate. Circles represent bootstrap values for internal nodes: black >= 0.90; grey < 0.90 and >= 0.70, and white < 0.70. Tree symbols in front of accessions represent species occurring in SDTF (purple) and rain forests (green).

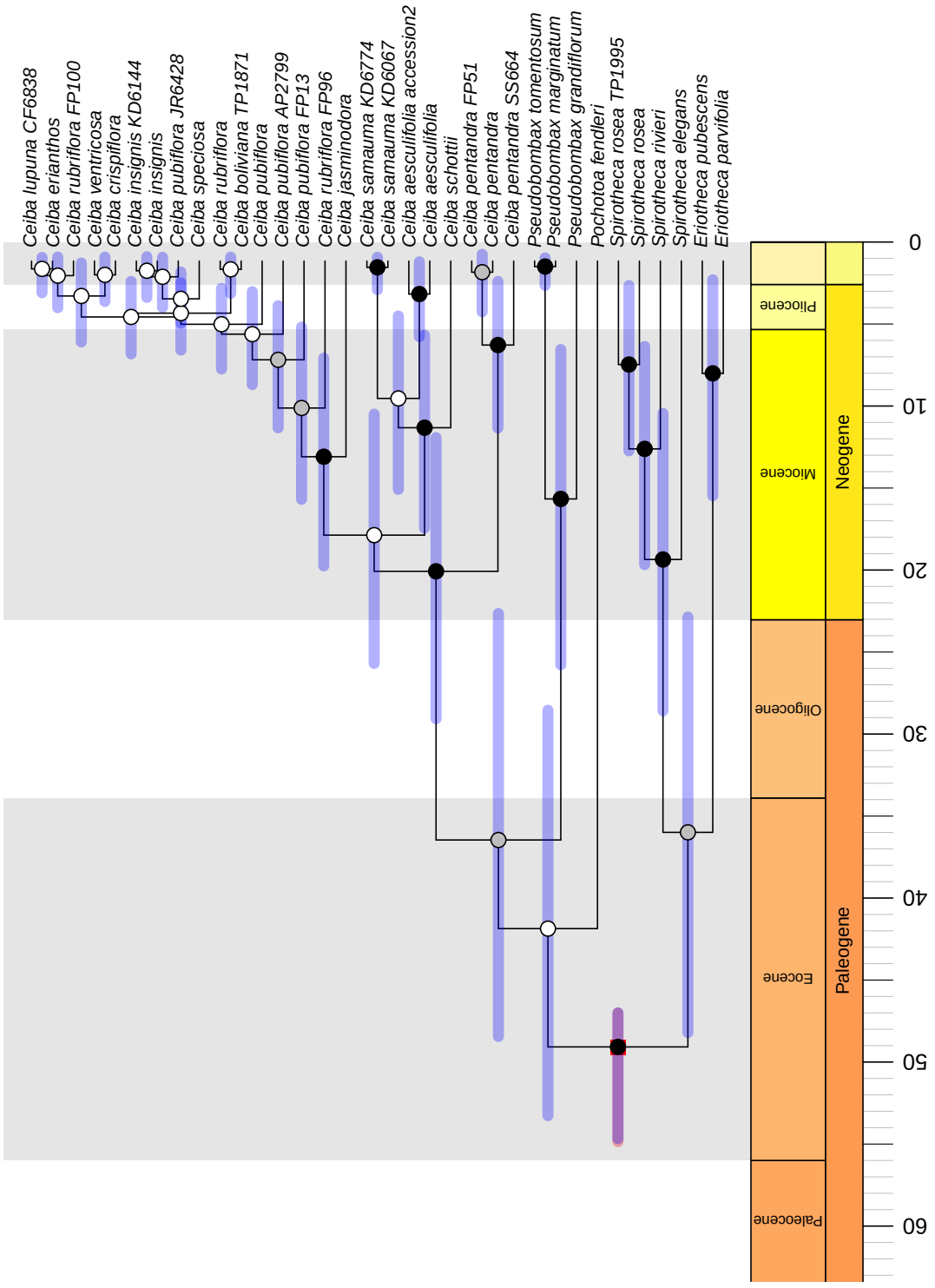


Figure 2.3: Maximum clade credibility tree resulting from BEAST2 analysis of nuclear ribosomal ITS sequence data sets for 14 species of *Ceiba* for a 47mya fossil calibration. Circles represent posterior probabilities for internal nodes: black >= 0.95; grey < 0.95 and >= 0.75, and white < 0.75.

Geographic and ecological structure

Our data suggest multiple shifts from dry to wet forests within *Ceiba* (Figures 2.2 and 2.3) because rain forest species are nested within the two dry forest clades. For example, the two accessions representing *C. samauma*, occurring in rain and riverine forest in South America, are sister to the Central American and Mexican clade and the rain forest species *C. lupuna* and *C. speciosa*, are nested within the South American SDTF clade. The D test shows clear phylogenetic signal for ecological preference and geographic phylogenetic structure (i.e., clear Central and South American clades) in *Ceiba*.

Biome-specific differences in the nature of species and their diversification trajectories

This results show young crown and stem ages for species in the South American SDTF clade, and patterns of long stems with shallow crown groups for rain forest species such as *Ceiba pentandra*. This is in contrast to previous studies of individual SDTF species that showed them to be older, with stem ages of 5-10my (eg. Pennington et al. 2010; de Queiroz and Lavin 2011), and runs contrary to the prediction of Pennington and Lavin (2016) that rain forest species might, on average, tend to have more recent origins.

The stem age of *C. pentandra* is estimated at 19.1 Ma. The long stem and shallow crown suggest this is an old rain forest lineage with more recent origin of extant populations. Likewise, *C. samauma* was recovered as monophyletic, and has a crown node age estimated as 0.6 Ma and a stem node age of 10.4 Ma. Both species therefore contrast with the suggested predominant patterns for rain forest species. Our result, recovering *C. pentandra* as monophyletic, with low sequence divergence amongst accessions, is consistent with that of Dick et al. (2007) who showed *C. pentandra* to have extremely weak phylogeographical structure based on ITS and chloroplast psbB-psbF for 51 individuals. In addition to that, the disjunct distribution of this species in Africa was demonstrated to be due to relatively recent long distance dispersal because of low

genetic divergence of the African populations.

Within the two predominantly SDTF clades, there is little evidence for old lineages with long stems and monophyletic crown groups for morphologically recognized species, as predicted by Pennington and Lavin (2016). The crown age of the South American SDTF clade, containing 10 species, is estimated at 12.1 Ma and the Mexican SDTF clade, containing two species, is estimated at 10.4 Ma with a stem age for both estimated at 19.1 Ma. Only two species from SDTF were recovered as monophyletic, *Ceiba aesculifolia* with a crown age estimated at 2.2 Ma and stem age at 8.6 Ma and *C. insignis* with a crown age estimated at 0.6 Ma and stem age at 1.0 Ma, although only in the BI analysis. Even when assigning a minimum age of 56 Ma to the *Eriotheca* stem, the same pattern is observed. The crown age of the South American SDTF clade is estimated at 14.1 Ma and the Mexican SDTF clade, containing two species, is estimated at 12.0 Ma with a stem age for both estimated at 19.8 Ma. The crown age of *Ceiba aesculifolia* was estimated at 2.4 Ma and stem age at 9.9 Ma (Figure 2.5).

The lack of resolution among the dry forest accessions, with most species being recovered as non-monophyletic, suggests absence of intraspecific coalescence for the ITS locus. Explanations for this include incomplete lineage sorting after speciation events, paralogous gene copies, inaccurate species delimitation and/or hybridisation followed by introgression (Naciri and Linder, 2015; Pennington and Lavin, 2016). We eliminated sequences with possible paralogues by visual inspection of the electropherograms and by comparing substitution rates along branch lengths following Bailey et al. (2003). Some species of *Ceiba* are hypothesised to be interfertile and hybridise (Gibbs and Semir, 2003), especially within the insignis species aggregate. However, it is also suggested that those species diverge in time of anthesis and pollinator type as well, and we have seen no evidence of putative hybrids in the field (Pezzini, pers. obs.). Eight out of the ten species within the South American SDTF clade are from Brazil and of these, four are distributed in the Caatinga, the largest area of SDTF in the Neotropics (700,000km^2) (Silva de Miranda et al., 2018). *Ceiba* species such as *C. pubiflora* are often widespread (Figure 1.4) and abundant (de Lima et al., 2010). Taken together,

this evidence suggests that the non-monophyly of *Ceiba* species found in SDTF such as *C. pubiflora* may be a reflection of large effective population sizes and hence a longer time to coalescence (Naciri and Linder, 2015; Pennington and Lavin, 2016), rather than due to hybridisation or ITS paralogy.

Our study illustrates that the general patterns of species age, monophyly and geographical structure reported for SDTF species belonging to the Leguminosae family (Pennington and Lavin, 2016) are not shared by one of the most characteristic SDTF tree genera and suggests that rather phylogenetic studies of unrelated groups are required.

Bibliography

Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*. 29:417–434.

Alverson WS, Duarte MC. 2015. Hello Again Pochota, Farewell Bombacopsis (Malvaceae). *Novon: A Journal for Botanical Nomenclature*. 24:115–119.

Bailey CD, Carr TG, Harris SA, Hughes CE. 2003. Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. *Molecular Phylogenetics and Evolution*. 29:435–455.

Baum DA, Small RL, Wendel JF. 1998. Biogeography and Floral Evolution of Baobabs Adansonia, Bombacaceae as Inferred From Multiple Data Sets. *Systematic Biology*. 47:181–207.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*. 10:e1003537.

Buckler ESI, Ippolito A, Holtsford TP. 1997. The evolution of plant ribosomal DNA: Divergent paralogues, pseudogenes and phylogenetic implications. *Genetics*. 145:821–832.

Carvalho-Sobrinho JG, Alverson WS, Alcantara S, Queiroz LP, Mota AC, Baum DA. 2016. Revisiting the phylogeny of Bombacoideae (Malvaceae): Novel relationships, morphologically cohesive clades, and a new tribal classification based on multilocus phylogenetic analyses. *Molecular Phylogenetics and Evolution*. 101:56–74.

Crisp MD, Arroyo MTK, Cook LG, Gandolfo MA, Jordan GJ, McGlone MS, Weston PH, Westoby M, Wilf P, Linder HP. 2009. Phylogenetic biome conservatism on a global scale. *Nature*. 458:754–756.

de Lima MR, Salard-Cheboldaeff M. 1981. Palynologie des bassins de Gandarela et Fonseca (Eocene de l'etat de Minas Gerais, Bresil). *Boletim IG*. 12:33–54.

de Lima MS, Damasceno-Júnior GA, Tanaka MO. 2010. Aspectos estruturais da comunidade arbórea em remanescentes de floresta estacional decidual, em Corumbá, MS, Brasil. *Revista Brasileira de Botânica*. 33:437–453.

de Queiroz LP, Lavin M. 2011. Coursetia (Leguminosae) From Eastern Brazil: Nuclear Ribosomal and Chloroplast DNA Sequence Analysis reveal the Monophyly of Three Caatinga-inhabiting Species. *Systematic Botany*. 36:69–79.

Dexter KG, Lavin M, Torke BM, Twyford AD, Kursar TA, Coley PD, Drake C, Hollands R, Pennington RT. 2017. Dispersal assembly of rain forest tree communities across the Amazon basin. *Proceedings of the National Academy of Sciences*. 114:2645–2650.

Dick CW, Bermingham E, Lemes MR, Gribel R. 2007. Extreme long-distance dispersal of the lowland tropical rainforest tree Ceiba pentandra L. (Malvaceae) in Africa and the Neotropics. *Molecular Ecology*. 16:3039–3049.

DRYFLOR. 2016. Plant diversity patterns in neotropical dry forests and their conservation implications. *Science*. 353:1383–1387.

Duarte L. 1974. Sobre uma flor de Bombacaceae da Bacia Terciária de Fonseca, MG. *Anais da Academia Brasileira de Ciências*. 46:407–411.

Duarte MC, Esteves GL, Salatino MLF, Walsh KC, Baum DA. 2011. Phylogenetic Analyses of Eriotheca and Related Genera (Bombacoideae, Malvaceae). *Systematic Botany*. 36:690–701.

Feliner GN, Rosselló JA. 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution*. 44:911–919.

Fritz SA, Purvis A. 2010. Selectivity in Mammalian Extinction Risk and Threat Types: a New Measure of Phylogenetic Signal Strength in Binary Traits. *Conservation Biology*. 24:1042–1051.

Gibbs P, Semir J. 2003. A taxonomic revision of the genus Ceiba Mill. (Bombacaceae). *Anales del Jardín Botánico de Madrid*. 60:259–300.

Gibbs P, Semir J, Da Cruz N. 1988. A proposal to unite the genera Chorisia Knuth with Ceiba Miller (Bombacaceae). *Notes from the Royal Botanic Garden Edinburgh*. 45:125–136.

Gradstein FM, Ogg JG, Schmitz MD, Ogg GM. 2012. The Geologic Time Scale. Elsevier.

Graham CH, Fine PVA. 2008. Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecology Letters*. 11:1265–1277.

Hughes CE, Pennington RT, Antonelli A. 2013. Neotropical Plant Evolution: Assembling the Big Picture. *Botanical Journal of the Linnean Society*. 171:1–18.

Koenen EJM, Clarkson JJ, Pennington TD, Chatrou LW. 2015. Recently evolved diversity and convergent radiations of rainforest mahoganies (Meliaceae) shed new light on the origins of rainforest hyperdiversity. *New Phytologist*. 207:327–339.

Maddison WP, Maddison DR. 2015. Mesquite: A Modular System for Evolutionary Analysis.

Miles L, Newton AC, DeFries RS, Ravilious C, May I, Blyth S, Kapos V, Gordon JE. 2006. A global overview of the conservation status of tropical dry forests. *Journal of Biogeography*. 33:491–505.

Murphy PG, Lugo AE. 1986. Ecology of Tropical Dry Forest. *Annual Review of Ecology and Systematics*. 17:67–88.

Naciri Y, Linder HP. 2015. Species delimitation and relationships: The dance of the seven veils. *Taxon*. 64:3–16.

Orme CDL. 2013. The caper package: comparative analyses in phylogenetics and evolution in R.

Pennington RT, Daza A, Reynel C, Lavin M. 2011. Poissonia eriantha (Leguminosae) From Cuzco, Peru: An Overlooked Species Underscores a Pattern of Narrow Endemism Common to Seasonally Dry Neotropical Vegetation. *Systematic Botany*. 36:59–68.

Pennington RT, Lavin M. 2016. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytologist*. 210:25–37.

Pennington RT, Lavin M, Oliveira-Filho A. 2009. Woody Plant Diversity, Evolution, and Ecology in the Tropics: Perspectives from Seasonally Dry Tropical Forests. *Annual Review of Ecology, Evolution, and Systematics*. 40:437–457.

Pennington RT, Lavin M, Sarkinen T, Lewis GP, Klitgaard BB, Hughes CE. 2010. Contrasting plant diversification histories within the Andean biodiversity hotspot. *Proceedings of the National Academy of Sciences*. 107:13783–13787.

Pennington RT, Richardson JE, Lavin M. 2006. Insights into the historical construction of species-rich biomes from dated plant phylogenies, neutral ecological theory and phylogenetic community structure. *New Phytologist*. 172:605–616.

Rangel TF, Edwards NR, Holden PB, Diniz-Filho JAF, Gosling WD, Coelho MTP, Cassemiro FAS, Rahbek C, Colwell RK. 2018. Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves. *Science*. 361:eaar5452.

Ravenna P. 1998. On the identity, validity, and actual placement in Ceiba of several Chorisia species (Bombacaceae), and description of two new south american species. *Onira*. 3:42–51.

Renner SS. 2005. Relaxed molecular clocks for dating historical plant dispersal events. *Trends in Plant Science*. 10:550–558.

Robyns A. 1963. Essai de monographie du genre Bombax s.l. (Bombacaceae) (Suite). *Bulletin du Jardin botanique de l'État a Bruxelles*. 33:145.

Silva de Miranda PL, Oliveira-Filho AT, Pennington RT, Neves DM, Baker TR, Dexter KG. 2018. Using tree species inventories to map biomes and assess their climatic overlaps in lowland tropical South America. *Global Ecology and Biogeography*. 27:899–912.

Simon MF, Grether R, de Queiroz LP, Skema C, Pennington RT, Hughes CE. 2009. Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proceedings of the National Academy of Sciences*. 106:20359–20364.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–1313.

Stebbins GL. 1974. Flowering plants: evolution above the species level. Cambridge, MA.

Wiens JJ, Donoghue MJ. 2004. Historical biogeography, ecology and species richness. *Trends in Ecology & Evolution*. 19:639–644.

2.A Appendix

Maximum clade credibility tree resulting from BEAST2 analysis of nuclear ribosomal ITS sequence data sets for 14 species of *Ceiba* for two alternative fossil calibrations (33 Ma and 56 Ma).

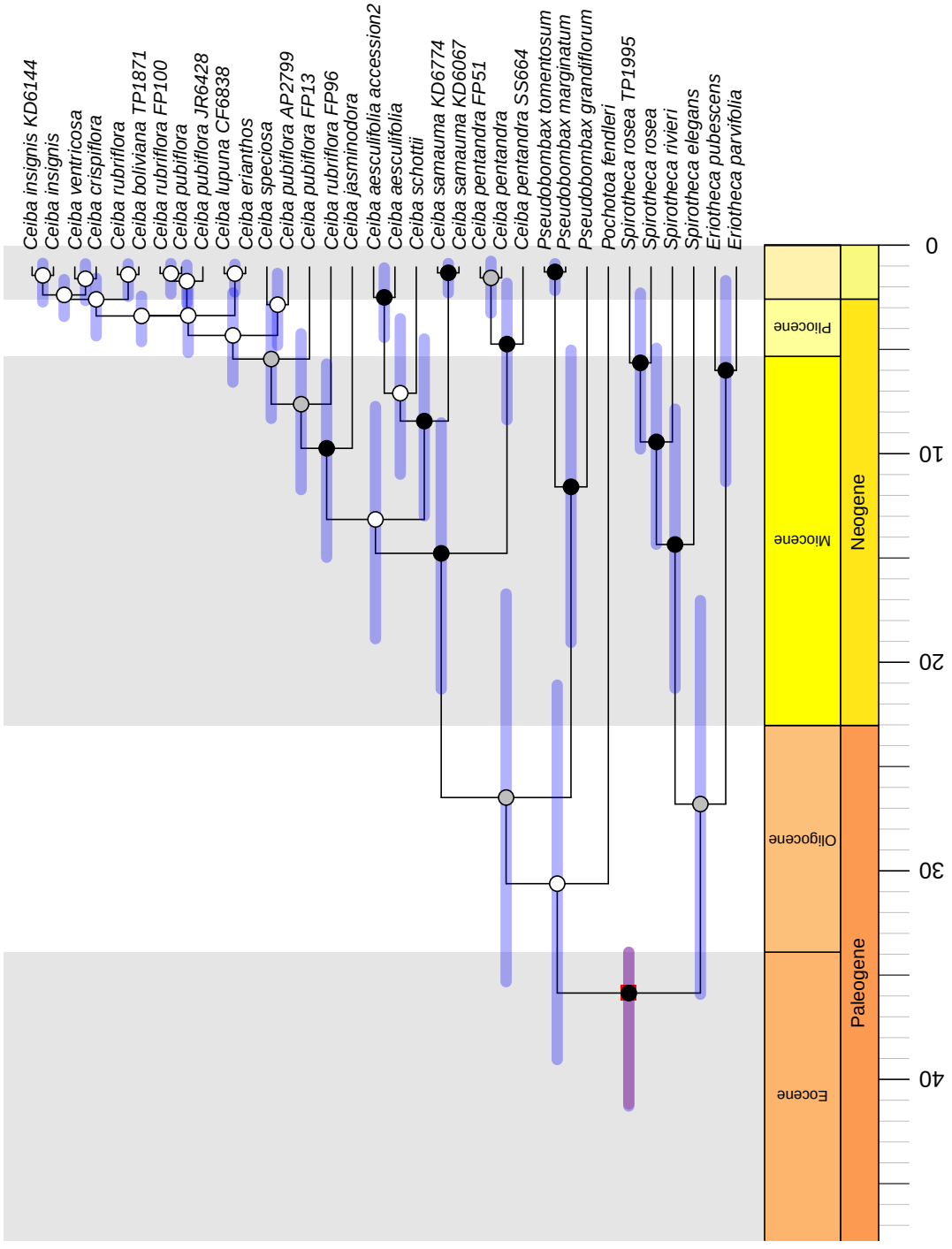


Figure 2.4: Maximum clade credibility tree resulting from BEAST2 analysis of nuclear ribosomal ITS sequence data sets for 14 species of *Ceiba* for a 33mya fossil calibration. Circles represent posterior probabilities for internal nodes: black >= 0.95; grey < 0.95 and >= 0.75, and white < 0.75.

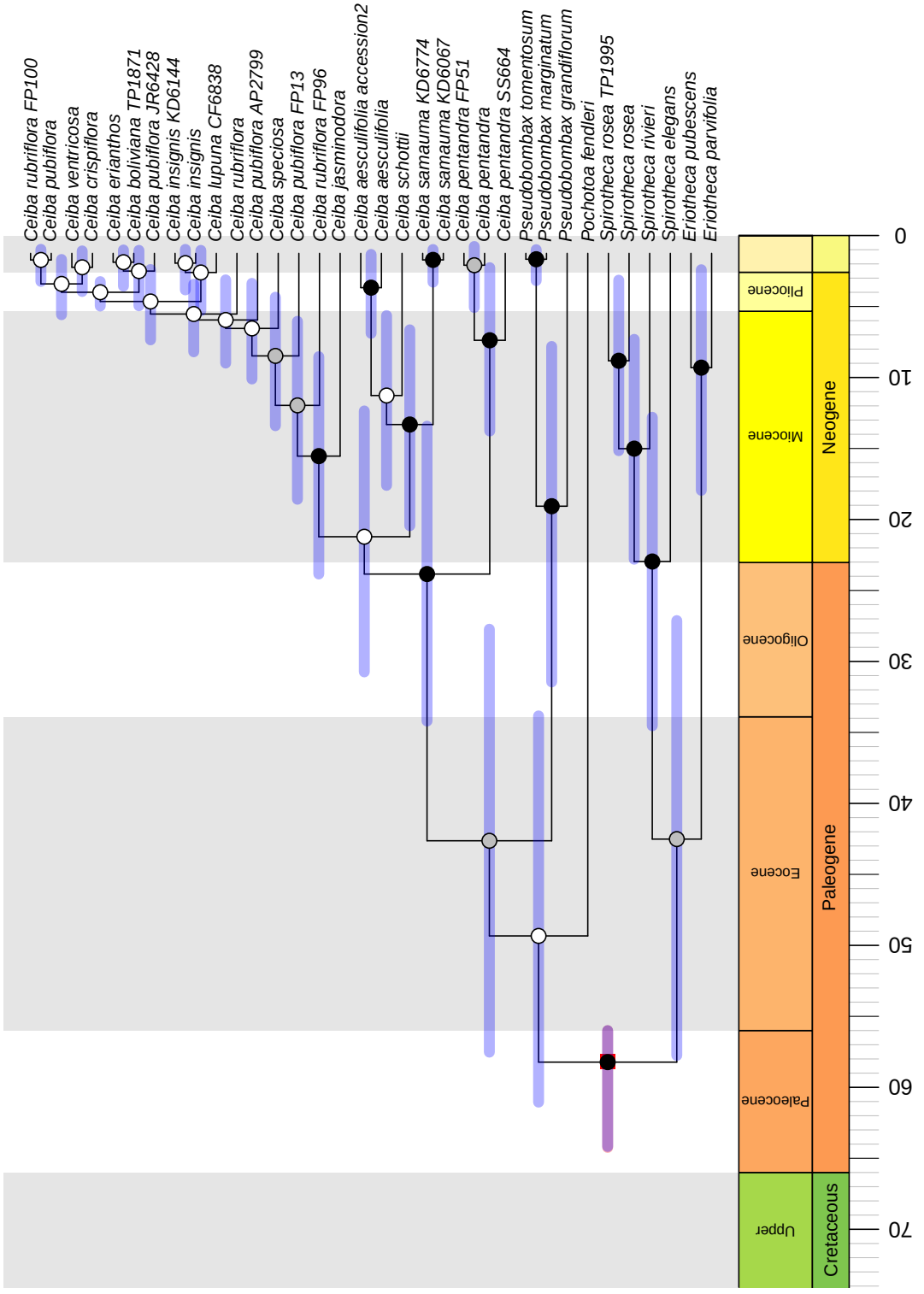


Figure 2.5: Maximum clade credibility tree resulting from BEAST2 analysis of nuclear ribosomal ITS sequence data sets for 14 species of *Ceiba* for a 56mya fossil calibration. Circles represent posterior probabilities for internal nodes: black >= 0.95; grey < 0.95 and >= 0.75, and white < 0.75.

Chapter 3

Phylogeny of *Ceiba* using next-generation targeted enrichment sequencing

3.1 Introduction

The previous chapter presented a phylogeny based on Sanger-sequencing of the ITS region for the genus *Ceiba*. Although the phylogeny provided insights into the diversification and biogeographic history of the genus, it failed to resolve all the species relationships, especially in the South American SDTF clade. The study of Carvalho-Sobrinho et al. (2016) showed that adding a few more chloroplast and nuclear (ETS) loci is unlikely to resolve the relationships amongst recently diverged *Ceiba* species and that much more DNA sequence is required. This chapter explores the potential of next generation hybrid capture sequencing of hundreds of nuclear loci in *Ceiba* phylogenetics.

Over the past 15 years new sequencing technologies marked the beginning of a new era of genetics called next generation sequencing (NGS) or massively parallel sequencing (Mardis, 2008). The main advance was the ability to sequence multiple loci or even the entire genome for several samples at once, in a multiplexing process. The large amount of data now available allow scientists to investigate in depth topics such as whole-genome duplication, transposable elements, gene trees/species tree discordance and the coalescent theory (Degnan and Rosenberg, 2009).

To improve phylogenetic resolution, new sequencing approaches offer great promise. Next-generation sequencing covers a diversity of new techniques that enable the sequencing of hundreds of independent nuclear loci, as well as loci from the chloroplast genome, which together can provide many phylogenetically informative characters. Such techniques include whole-genome sequencing, amplicon sequencing, RAD genotyping, target enrichment and transcriptomics. Apart from whole-genome sequencing, all these techniques are promising for phylogeographic and phylogenetic studies of non-model organisms due to their reduced cost per base pair sequenced (McCormack et al., 2013). NGS differs from Sanger sequencing by producing huge numbers of DNA sequence-reads from single samples, where all the DNA in a sample is sequenced (Harrison and Kidner, 2011). Among the many available approaches, hybrid capture (or target enrichment) has been successfully used to provide large numbers of phylogenetically informative characters (Yu et al., 2018), including for recent evolutionary

radiations in tropical lineages (eg. Nicholls et al. 2015; Carlsen et al. 2018). This technique uses probes (or "baits") designed to capture target loci from fragmented genomic DNA libraries and, in spite of the need of prior knowledge from a draft genome or transcriptomes to design the baits, its moderate cost and ability to use relatively degraded DNA such as from herbarium specimens (Hart et al., 2016) give advantages in non-model organisms (McCormack et al., 2013; Nicholls et al., 2015). By contrast, transcriptome sequencing, for example, requires fresh tissue and RAD genotyping requires high molecular weight DNA. RAD genotyping produces anonymous loci rather than focusing on a particular target of interest. In addition to that, orthology assessment can be challenging in anonymous loci. Hybrid capture allows the selection of many single copy, orthologous loci that give phylogenetic resolution at different taxonomical levels (Nicholls et al., 2015).

Unlike the Sanger sequencing technique, where one can see the input and output data in each step of analysis, NGS techniques often produce large data files that cannot be easily validated or inspected visually. The millions of short reads generated have to be analysed in bioinformatics pipelines, and this is considered the bottleneck for NGS sequencing, especially taking into account the lack of bioinformatics training for biologists. NGS analysis involve many steps with decisions to make on each of them. Raw reads are the base of the analysis and the quality of the base call has direct impact on the outcome (Del Fabbro et al., 2013; Pfeifer, 2017) as can the software and parameters chosen within them (Boussau and Daubin, 2010; Altmann et al., 2012; Yang and Smith, 2013). One key challenge is the assembly of the short reads (Yang and Smith, 2013), not only because of their small size but also because of the sequencing low contiguity (Shendure et al., 2017). Read length can be as small as 35bp (Liu et al., 2012). Paired-end sequencing was an important innovation that improved the contiguity of information because it links the two ends of a pair of raw reads that can be assembled into a longer contig with more confidence (Shendure et al., 2017).

Two main pathways can be followed for assembly. *De novo* assembly was initially the most common approach because it does not require a reference genome. This tech-

nique clusters similar reads and produces contigs that are combined in longer scaffolds, more commonly using *de Bruijn* graphs (Compeau et al., 2011). It can be computationally demanding (McCormack et al., 2013) and produce chimeric contigs from repetitive regions especially in plant genomes. The other approach selects the short raw reads by mapping them back to a reference genome and assembles them into contigs. This technique relies on the similarity of the read and the reference to decide on good matches or bad matches. A read is considered a good match to the reference when it falls above a certain threshold that for most assemblers is calculated taking into account the read length and number of mismatches. However, most non-model organism do not have a reference genome available, or even a sequenced genome from a close relative and the use of a reference from distantly related species could result in fewer reads being mapped back because the sequences are too divergent. Nonetheless, in the target capture technique the baits are normally designed using the transcriptome of one or more species. Thus, the bait sequences represent a subset of the genome and can be used as a reference.

Although the raw reads are the foundation of NGS analysis, few studies have assessed the impact of initial filtering on the final output. Different software for mapping short reads to a reference genome are available and the comparison of their performance is still under debate (Thankaswamy-Kosalai et al., 2017; Schott et al., 2017). Bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2018) and BWA (Burrows-Wheeler Aligner) (Li and Durbin, 2010), for example, use the Burrows-Wheeler transform (BWT) algorithm (Burrows and Wheeler, 1994) to map the reads back to the reference. One of the main differences between BWA and the first version of Bowtie (Langmead et al., 2009) is that the later did not allow gapped alignment and so could overlook true insertions and deletions. Bowtie2 allows the presence of gaps. Other differences between the two software packages are related to the percentage of reads mapped and how fast the analyses run. However, benchmark tests show that no aligner outperforms another and the differences observed might be related to characteristics of the genome and this, along with the nature of the study, should guide the user when

choosing the most suitable software (Hatem et al., 2011; Thankaswamy-Kosalai et al., 2017). For example, BWA might perform better when dealing with longer read lengths (Hatem et al., 2011) and might be faster if dealing with big genomes (Thankaswamy-Kosalai et al., 2017).

Parallel to the development of wet laboratory protocols and software to analyse the new type of data, new metrics were created, not only to automate the data analysis process, but also to reduce human decision making (Ewing et al., 1998; Shendure and Ji, 2008; Shendure et al., 2017) and increase reproducibility. For example, the Phred score was created during the Human Genome Project as a quality metrics of base, read and variants quality (Ewing et al., 1998). In addition to that, several automated pipelines using those software and metrics have been developed to analyse each type of NGS data (eg. Stacks, iPyRAD, HybPiper). The usage of those different pipelines is intensively discussed in the literature (eg. Fér and Schmickl (2018); Herrando-Moraira et al. (2018)). Although they intend to facilitate the analysis, especially for the user not comfortable with the command line, care should be taken with the one-size-fits-all approach. In line with the diverse software available, there is little consensus on which pipeline is the best (Herrando-Moraira et al., 2018).

Likewise, different approaches exist within the hybrid bait technique in the lab. For example, the bait set can be designed to target orthologous loci across taxa at a broad taxonomic scale - "universal" baits; eg. Angiosperms (Johnson et al. 2018) or designed for a specific taxon in a finer scale study (eg. Nicholls et al. 2015; Chau et al. 2018). The advantages and disadvantages of both approaches are still under debate (Kadlec et al., 2017). Baits more similar to the target have a higher hybridisation success (Cronn et al., 2012; Chau et al., 2018) and thus generate more complete data for all loci and all species. By contrast, a more universal bait set would be useful to investigate questions at a broader taxonomic scale and results from different studies would more likely be compatible. Broad taxonomic scale baits have been applied successfully for animals (Schott et al., 2017) but in plants target capture might be more efficient using taxon specific bait sets (Chau et al., 2018; Kadlec et al., 2017) because events of genome

and gene duplication are frequent (McKain et al., 2018) and a universal bait set would be less likely to capture the same loci across different taxa. In addition to that, the presence of paralogues might result in high levels of discordance amongst individual gene trees, a phylogenetic inference that is statistically inconsistent, and with high bootstrap support values for the incorrect tree (Kubatko and Degnan, 2007; Roch and Steel, 2015). A specific bait set could be designed aiming at single copy, orthologous loci to ameliorate this issue.

The lack of consensus on the best software, best settings, best pipeline and best type of bait set allied to the fact that the choice of the best option might depend on characteristics of each genome urges for studies that properly address those issues instead of following the most popular approaches. In addition, because the metrics developed to make such comparisons are often based in factors such as computational performance and volume of data, it would be beneficial to evaluate the consequences of those variations on the final analysis and hence their biological implications.

In this chapter I aim to construct a well-resolved, multi-locus and densely sampled species-level phylogenetic tree for *Ceiba* using the NGS technique hybrid bait capture. Specifically I aim to answer the following questions:

1. What is the impact of data quality filtering in downstream analysis?
2. What is the impact of different genome assembly pipelines on phylogenetic inference?
3. Does a taxon-specific bait set influence sequence capture success?
4. What are the relationship amongst the species of *Ceiba*?
5. Does the NGS analysis provide an improvement over previous phylogenetic studies?

3.2 Methods

3.2.1 Sample selection

The laboratory procedure was done in four different rounds in order to test the baits and guide taxon sampling for the following rounds. The four rounds are represented by the sequencing runs 1,2,3 and 4 in Figures 3.2 to 3.10, 3.12 to 3.16, 3.18 to 3.30 and 3.39 and Table 3.3. For the first round, I aimed to test whether the baits designed using the transcriptome of *Adansonia* and *Bombax* (Karimi et al. in prep) would work well in *Ceiba*. For that, I selected 24 samples both from silica-gel dried and herbarium specimen leaf tissue, representing species from across the genus. For the second round I selected 25 more samples to test whether the newly designed baits based on sequence data from *Ceiba* (see below) would improve capture success and be compatible with the previous data set. For the third and fourth rounds I selected another 54 samples, half coming from silica-gel dried leaf tissue and the other half from herbarium specimens (Figure 3.1 and Table 3.1, see also supplementary Figures 3.40 to 3.42). Genomic DNA (gDNA) from silica-gel dried and herbarium tissue vary in quality, and this can interfere with the capture success and therefore on downstream analysis (Hart et al., 2016).

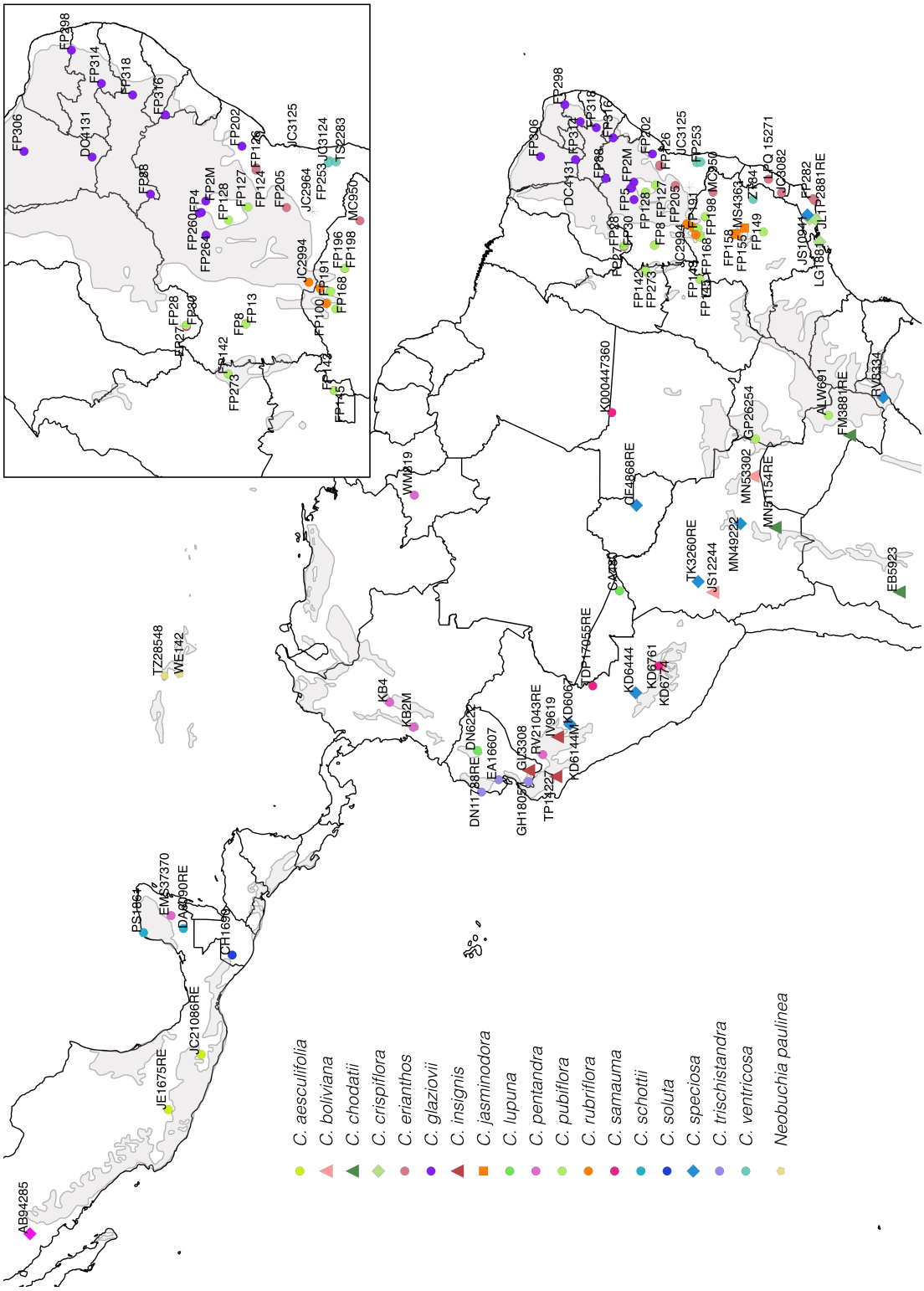


Figure 3.1: Map of the occurrence of the 103 accessions of *Ceiba* and *Neobuchia* sequenced. Grey areas represent SDTF patches according to DRYFLOR (2016).

Table 3.1: Collection details of each of the 103 accessions of *Ceiba* and *Neobuchia* sequenced. Samples indicated with asterisk (*) represent sterile samples.

n	Accession	Species	Locality	Country	Collection year
1	AB94285	*Ceiba aesculifolia aesculifolia*	San Javier	Mexico	1994
2	JC10136	*Ceiba aesculifolia aesculifolia*	Berriozabal, Chiapas	Mexico	1983
3	JE1675RE	*Ceiba aesculifolia aesculifolia*	Pátzcuaro, Michoacan	Mexico	1988
4	CH1806	*Ceiba aesculifolia parvifolia*	San Juan de los Cues	Mexico	1993
5	JC21086RE	*Ceiba aesculifolia parvifolia*	Santiago Juxtlahuaca, Oaxaca	Mexico	1996
6	JS12244	*Ceiba boliviana*	Nor Yungas	Bolivia	1984
7	KD6761	*Ceiba boliviana*	Quillabamba	Peru	2012
8	MN53302	*Ceiba boliviana*	Cordillera, Santa Cruz	Bolivia	2005
9	EB5923	*Ceiba chodatii*	El Potererillo	Argentina	1939
10	FM3881RE	*Ceiba chodatii*	Pazo Correo, Central	Paraguay	1990
11	MN51154RE	*Ceiba chodatii*	Cordillera, Santa Cruz	Bolivia	2000
12	LG1881	*Ceiba crispiflora*	Parati, Rio de Janeiro	Brazil	1995
13	MG743	*Ceiba crispiflora*	Rio de Janeiro	Brazil	1839
14	FP124	*Ceiba erianthos*	Itatim, Bahia	Brazil	2016*
15	FP126	*Ceiba erianthos*	Itatim, Bahia	Brazil	2016*

Table 3.1 continued from previous page

n	Accession	Species	Locality	Country	Collection year
16	FP13	*Ceiba erianthos*	So Desidério, Bahia	Brazil	2014
17	FP205	*Ceiba erianthos*	Feira de Santana, Bahia	Brazil	2016
18	FP282	*Ceiba erianthos*	Búzios, Rio de Janeiro	Brazil	2017
19	JC3082	*Ceiba erianthos*	Alegre, Espirito Santo	Brazil	2011
20	JL1	*Ceiba erianthos*	Rio de Janeiro, Rio de Janeiro	Brazil	2016
21	LPQ15271	*Ceiba erianthos*	Santa Teresa, Espirito Santo	Brazil	2011
22	MC950	*Ceiba erianthos*	Itaobim, Minas Gerais	Brazil	2017
23	P2881RE	*Ceiba erianthos*	Búzios, Rio de Janeiro	Brazil	1993
24	DC4131	*Ceiba glaziovii*	Jardim, Cear	Brazil	2016
25	FP202	*Ceiba glaziovii*	Feira de Santana, Bahia	Brazil	2016
26	FP260	*Ceiba glaziovii*	Umburanas, Bahia	Brazil	2017
27	FP264	*Ceiba glaziovii*	Jussara, Bahia	Brazil	2017
28	FP298	*Ceiba glaziovii*	Araruna, Paraiba	Brazil	2017
29	FP2M	*Ceiba glaziovii*	Jacobina, Bahia	Brazil	2017
30	FP306	*Ceiba glaziovii*	Quixadá, Ceará	Brazil	2017*
31	FP314	*Ceiba glaziovii*	So José dos Cordeiros, Paraba	Brazil	2017

Table 3.1 continued from previous page

n	Accession	Species	Locality	Country	Collection year
32	FP316	*Ceiba glaziovii*	Canindé do São Francisco, Alagoas	Brazil	2017*
33	FP318	*Ceiba glaziovii*	Buíque, Pernambuco	Brazil	2017*
34	FP4	*Ceiba glaziovii*	Ourolândia, Bahia	Brazil	2014
35	FP5	*Ceiba glaziovii*	Morro do Chapéu, Bahia	Brazil	2014
36	FP88	*Ceiba glaziovii*	Juazeiro, Bahia	Brazil	2014
37	GL3308	*Ceiba insignis*	Gonzanama-Catamyo	Ecuador	1997
38	IV9619	*Ceiba insignis*	Elias Soplin Vargas	Peru	1998
39	TP14227	*Ceiba insignis*	Olmos	Peru	1986
40	FP155	*Ceiba jasminodora*	Joaquim Felício, Minas Gerais	Brazil	2016
41	FP158	*Ceiba jasminodora*	Joaquim Felício, Minas Gerais	Brazil	2016
42	MS4363	*Ceiba jasminodora*	Diamantina, Minas Gerais	Brazil	1990
43	CA480	*Ceiba lupuna*	Senador Guiomard, Acre	Brazil	2010
44	DN6222	*Ceiba lupuna*	Napo	Ecuador	1985
45	EMS37370	*Ceiba pentandra*	Jose Maria Morelos	Mexico	2005
46	KB2M	*Ceiba pentandra*	Santiago de Cali, Valle del Cauca	Colombia	2015
47	KB4	*Ceiba pentandra*	Armero, Tolima	Colombia	2015

Table 3.1 continued from previous page

n	Accession	Species	Locality	Country	Collection year
48	RV21043RE	*Ceiba pentandra*	Bagua, Amazonas	Peru	1996
49	WM819RE	*Ceiba pentandra*	Caracarai, Roraima	Brazil	1988
50	ALW691	*Ceiba pubiflora*	San Pedro	Paraguay	1956
51	FP127	*Ceiba pubiflora*	Wagner, Bahia	Brazil	2016*
52	FP128	*Ceiba pubiflora*	Cafarnaum, Bahia	Brazil	2016*
53	FP131	*Ceiba pubiflora*	Novo Jardim, Tocantins	Brazil	2016*
54	FP135	*Ceiba pubiflora*	Novo Jardim, Tocantins	Brazil	2016*
55	FP139	*Ceiba pubiflora*	Novo Jardim, Tocantins	Brazil	2016*
56	FP142	*Ceiba pubiflora*	Novo Jardim, Tocantins	Brazil	2016*
57	FP143	*Ceiba pubiflora*	Formosa, Goiás	Brazil	2016*
58	FP145	*Ceiba pubiflora*	Formosa, Goiás	Brazil	2016*
59	FP149	*Ceiba pubiflora*	Matozinhos, Minas Gerais	Brazil	2016
60	FP168	*Ceiba pubiflora*	Brejo do Amparo, Minas Gerais	Brazil	2016
61	FP170	*Ceiba pubiflora*	Januária, Minas Gerais	Brazil	2016
62	FP173	*Ceiba pubiflora*	Januária, Minas Gerais	Brazil	2016
63	FP175	*Ceiba pubiflora*	Januária, Minas Gerais	Brazil	2016

Table 3.1 continued from previous page

n	Accession	Species	Locality	Country	Collection year
64	FP189	*Ceiba pubiflora*	Jaíba, Minas Gerais	Brazil	2016
65	FP191	*Ceiba pubiflora*	Jaíba, Minas Gerais	Brazil	2016
66	FP196	*Ceiba pubiflora*	Porteirinha, Minas Gerais	Brazil	2016
67	FP198	*Ceiba pubiflora*	Porteirinha, Minas Gerais	Brazil	2016
68	FP273	*Ceiba pubiflora*	Novo Jardim, Tocantins	Brazil	2017
69	FP30	*Ceiba pubiflora*	Corrente, Piauí	Brazil	2017*
70	FP8	*Ceiba pubiflora*	Barreiras, Bahia	Brazil	2014*
71	GP26254RE	*Ceiba pubiflora*	Corumbá, Mato Grossso do Sul	Brazil	1979
72	FP100	*Ceiba rubriflora*	Manga, Minas Gerais	Brazil	2015
73	FP108	*Ceiba rubriflora*	Matias Cardoso, Minas Gerais	Brazil	2016*
74	JC2994	*Ceiba rubriflora*	Iuiú, Bahia	Brazil	2011*
75	FP27	*Ceiba samauma*	Corrente, Piauí	Brazil	2017*
76	K000447360	*Ceiba samauma*	Novo Mundo, Mato Grosso	Brazil	2006
77	KD6067	*Ceiba samauma*	Tarapoto	Peru	2012
78	KD6774	*Ceiba samauma*	Quillabamba	Peru	2012
79	TDP17055RE	*Ceiba samauma*	Ucayali, Pucallpa	Peru	2001

Table 3.1 continued from previous page

n	Accession	Species	Locality	Country	Collection year
80	DA6090RE	*Ceiba schottii*	Calakmul, Campeche	Mexico	2003
81	PS1861	*Ceiba schottii*	Hunucma, Yucatan	Mexico	1995
82	CH1690	*Ceiba soluta*	Malacatancito, Huehuetenango	Guatemala	1992
83	CF4868RE	*Ceiba speciosa*	Presidente Médici, Rondnia	Brazil	1984
84	FP283	*Ceiba speciosa*	Petrópolis, Rio de Janeiro	Brazil	2017
85	JS10941	*Ceiba speciosa*	Volta do Pio	Brazil	1980
86	KD6144M	*Ceiba speciosa*	Tarapoto	Peru	2012
87	KD6444	*Ceiba speciosa*	Puerto Ocopa	Peru	2012
88	MN49222	*Ceiba speciosa*	Andres Ibanez	Bolivia	1998
89	RV3334	*Ceiba speciosa*	Misiones	Argentina	1995
90	TK3260RE	*Ceiba speciosa*	General Ballivian, Beni	Bolivia	1991
91	DN11738RE	*Ceiba trischistandra*	Manta, Manabí	Ecuador	1998
92	EA16607	*Ceiba trischistandra*	Guayaquil	Ecuador	1955
93	GH18057	*Ceiba trischistandra*	Pindal	Ecuador	1980
94	FP253	*Ceiba ventricosa*	Jussari, Bahia	Brazil	2017*
95	JC3124	*Ceiba ventricosa*	Jussari, Bahia	Brazil	2011

Table 3.1 continued from previous page

n	Accession	Species	Locality	Country	Collection year
96	TS2283	*Ceiba ventricosa*	Camacan, Bahia	Brazil	1972
97	ZT841RE	*Ceiba ventricosa*	Governador Valadares, Minas Gerais	Brazil	1964
98	JC3125	*Eriotheca obcordata*	Maraú, Bahia	Brazil	2011
99	TC10218	*Neobuchia paulinae*	Distrito Nacional	Dominican Republic	2017
100	TZ28548	*Neobuchia paulinae*	Presqu'ile du Nord-Ouest	Dominican Republic	1984
101	WE142	*Neobuchia paulinae*	Pointe--Raquette	Haiti	1927
102	FP28	*Pseudobombax marginatum*	Corrente, Piauí	Brazil	2017
103	JC2964	*Spirotheca elegans*	Anagé, Bahia	Brazil	2011

3.2.2 Choice of sequencing machine and library preparation kit

The first step of library preparation was the choice of the sequencing platform, raw read length and library preparation kit. The Illumina TruSeq Nano DNA LT Sample Preparation kit (version from November 2013) was chosen because it has been used successfully in other tropical lineages (eg. Nicholls et al. 2015).

The Illumina TruSeq Nano DNA LT Sample Preparation kit requires either 100ng or 200ng of gDNA as initial input. Because I aimed to sequence a high number of herbarium specimens, from which genomic DNA is frequently degraded and in low quantity (Hart et al., 2016), the minimal initial input of gDNA was set to 100ng to include as many samples as possible. With 100ng input, the insert size required by the Illumina TruSeq Nano DNA LT Sample Preparation kit protocol is 350 base-pairs (bp). In addition to that, I aimed to sequence no more than 27 samples in each round. An Illumina MiSeq run with up to 20 million reads is suitable for sequencing a subset of the genome applying a technique such as the targeted enrichment for a few accessions as it would yield ca. 20x coverage across the samples. Because the insert size was 350bp, the read length was set to 150 bp paired-end so as not to waste sequencing effort. The paired-end sequencing sequences both ends of the same library fragment producing a group of forward reads and another one of reverse reads. A 150bp paired-end read in a 350bp fragment would leave a sequencing gap of 50bp. The information provided by both reads, also called "mates", increases the confidence in the genome assembly. In case one mate matches multiple places on the reference genome, the information on the second mate can aid the confidence in the placement of that read. Therefore, the paired-end sequencing mimics longer read sequencing.

3.2.3 DNA Extraction

I extracted genomic DNA using the Qiagen DNeasy kit (QIAGEN Inc., Valencia, California, USA) protocol with the modifications described in Nicholls et al. (2015). A few samples showed non-consistent behaviour during the sonication process which might indicate carry over of polysaccharides after extraction and interfere with library prepa-

ration. Those samples were re-extracted and treated with two to three Sorbitol buffer washes (100nM Tris-HCL, pH 8.0, 0.35 M sorbitol, 5 mM EDTA, pH 8.0, stored at 4°C, 1% PVP-40 with 1% 2-mercaptoethanol (added just before use)) applied immediately after the tissue disruption step following indications in Souza et al. (2012) and Russell et al. (2010). Genomic DNA quantity was assessed with Qubit dsDNA HS Assay kit (Thermo Fisher Scientific, USA) with two replicates per sample (1μL for each replica) and only samples with minimum 2ng/μL average were used. Genomic DNA integrity was assessed with the Tapestation 2200 (Agilient Technology, USA).

3.2.4 Sonication

Genomic DNA from all selected samples was normalized to 53μL and 100ng because the input size for library preparation was 350bp. The initial volume required by the Illumina TruSeq Nano DNA LT Sample Preparation kit protocol is 50μL, however the extra 3μL were used to access the size distribution of the fragments of gDNA on a second run of the Tapestation 2200 (Agilient Technology, USA) after sonication.

For each sample, the number of cycles of sonication was decided based on the fragment size distribution from the first run on the Tapestation. Samples with the majority of fragments showing a high molecular weight band were sonicated for 5 cycles of 25 seconds ON and 90 seconds OFF. The number of cycles for other samples varied from 4 cycles of 25 seconds ON and 90 seconds OFF to no sonication at all in accordance with the state of degradation of the gDNA. After sonication, samples were bead-cleaned following Illuminas TruSeq Nano DNA LT Sample Preparation kit protocol and then run on the Tapestation one more time to check that the majority of the fragments were within the 350bp insert size value for library preparation.

Fourteen samples were treated with the PreCR end repair kit (New England Biolabs, Inc., USA) following manufacturer's instructions in order to repair the highly fragmented and therefore damaged DNA. Those samples are indicated with RE at the end of the collector's number.

3.2.5 Library preparation

Libraries were prepared following Illuminas TruSeq Nano DNA LT Sample Preparation kit protocol, except by final elution with 37μL of water in order to use the extra μL for quality check. The PCR step to enrich the library fragments was done with 9 cycles. I ran each library on the Tapestation to check the fragment size distribution curve and on the Qubit dsDNA HS Assay kit to check DNA concentration. All libraries were diluted to 10nM.

3.2.6 Sample pooling

Libraries were pooled in groups of eight or nine prior to the capture reaction in equimolar amounts. The criteria to pool them together was first the type of preservation (silica-gel dried or herbarium tissue) followed by closer related species (Nicholls et al., 2015; Johnson et al., 2016; McGee et al., 2016). Fragments of library with better genomic DNA quality may dominate the capture reaction over other samples. By pooling together samples with similar DNA quality I increase the chance that samples with worse DNA quality will produce enough data. Likewise, by pooling closely related species I increased the chance that all samples in that pool produce equivalent amount of data.

3.2.7 Capture reaction

For the capture reaction, I followed the MyBaits version v2.3.1 protocol with 19 hours hybridisation at 65^{o}C and stringent wash afterwards. Each capture reaction sample was quantified on the Qubit dsDNA HS Assay kit, equal amounts pooled together in a 60μL 10nM sample and submitted to a 150bp long, paired-end Illumina MiSeq run. Sequencing was performed by Edinburgh Genomics.

3.2.8 Quality check of raw reads

The raw reads were demultiplexed and run on FastQC v0.11.5[1] for initial quality check. The forward and reverse raw reads were then run on Trimmomatic v0.38 (Bolger et al., 2014) to remove Illumina adapters (TruSeq3-PE.fa) and poor quality bases with varying leading, trailing and sliding window settings (Table 3.2). Trimmomatic applies a window-based algorithm to trim low-quality read regions, which provides better results when compared with software using running sum algorithms (Del Fabbro et al., 2013).

The quality is measured by the Phred score (Ewing et al., 1998) which is calculated as $Q = -10log10(e)$, where e is the estimated probability of the wrong base call[2]. A Phred score of 10 (Q =10) for example has a probability of 1 in 10 of being wrong and therefore 90% base call accuracy. A Phred score of 20 has a probability of 1 in 100 of being wrong and 99% base call accuracy. A higher Phred score indicates more confidence in the call of a certain nucleotide.

The leading setting of Trimmomatic removes bases from the beginning of the read if they fall below the Phred quality threshold set by the user. The read is parsed until a base pair equal or above the quality threshold value is reached. For example, given a raw read where the first four base pairs have quality of 10 and the fifth base pair has a quality score of 20, a leading value of 20 would remove the first four base pairs and stop parsing at the fifth base pair. The trailing setting follows the same logic but parses the read starting from the end.

The sliding window argument includes two values. The first value is the window size and the second value is the average required Phred score. The read is dropped when the average quality of the number of bases represented by window size falls below the quality threshold. For example, a sliding window of 4:15 means that the read is dropped when the average quality of any group of four sequential base pairs is below a Phred score of 15. This setting prevents a good quality read being removed in case of a single poor quality base pair within this read (Bolger et al., 2014). Therefore in the settings I tested (Table 3.2) the leading/trailing 3 and sliding window 4:15 represent

[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[2] http://www.illumina.com/science/education/sequencing-quality-scores.html

a less conservative setting than the leading/trailing 20 and sliding window 4:20, which represents a more stringent filter for base call quality. After Trimmomatic, the results were once more run on FastQC and summarised with MultiQC v1.7[3]. I selected two (3-4:15 and 3-4:20) of the five different settings to use as input for downstream analysis and assess the impact of the different initial quality filter on final results (MacManes, 2014).

Table 3.2: Variation in the leading (removes bases from the beginning of the read if they fall below the Phred quality threshold), trailing (removes bases from the end of the read if they fall below the Phred quality threshold set by the user) and sliding window (removes reads when the average quality of the number of bases represented by the window size (first value) falls below the quality threshold (second value)) settings in Trimmomatic applied to raw reads.

Trimmomatic settings		
Leading	Trailing	Sliding Window
3	3	4:15
3	3	4:20
20	20	4:15
20	20	4:20
3	3	5:20
20	20	5:20

3.2.9 Data assembly

Pipelines tested

In this study, the raw data were assembled using three different pipelines:

1. Reference mapping following Nicholls et al. (2015)

   This pipeline uses the bait sequence as a reference to map back the raw reads. The authors take a conservative approach on orthology assessment by using a stringent value for the mapping similarity. This means that only reads highly similar to the baits would be assembled, and possible more divergent paralogous copies would be discarded. This pipeline uses the software Bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2018) to map back the raw reads.

[3]https://multiqc.info

Additionally, I ran the same pipeline using BWA (Burrows-Wheeler Aligner) (Li and Durbin, 2010) to compared both assemblers, and also the impact of variations in default settings of mapping threshold values for each of them.

For Bowtie2, the alignment score is used to determine if a read is similar to the reference and the alignment threshold is used to decide if a read should be kept as a valid mapped read. The alignment score is calculated by giving a penalty of -6 to each mismatch, -11 for a length-2 gap and +2 bonus for each match on each base pair of a read. The higher the alignment score, more similar a read is to the reference. The minimal alignment score threshold is calculated as $constant + 8 \times \ln(L)$ where L is the length of the read. The higher the constant, more similar a read has to be to be kept. The default constant in the `--local` alignment mode is 20, i.e., in a read of length one, if the alignment score is lower than 20 the read is discarded. I tested variations of this constant from 20 to 240, increasing the value in units of 10 at every step. Each resulting vcf file (Danecek et al., 2011) was then examined in terms of % of reads aligned, number of variants and the standardized quality of variant sites (ratio between average quality of variant/average quality of non-variant) following Nicholls et al. (2015).

For BWA, I used the MEM (maximal exact matches) algorithm which is suitable for longer reads (70bp - 1Mbp). The MEM algorithm uses the maximal exact matches to seed alignments followed by extension of the seed with the affine-gap Smith-Waterman algorithm (SW)[4]. The mismatch penalty threshold (-B argument) is one of the settings used to determine if a read is considered a match or not. This value is calculated using the sequencing error rate that is estimated as $0.75 \times \exp[-\log(4) \times threshold/A]$ where A is the matching score with default value of one. The default value of the mismatch penalty is four. An increase on this value represents an increase in the error rate, and therefore a higher penalty for a mismatch. The higher the threshold, more similar a read has to be to the reference to be considered good enough. I tested variations of this score from 0 to

[4]http://bio-bwa.sourceforge.net/bwa.shtml

50, increasing the value in units of 5 at every step. Each resulting vcf file was then examined in terms of % of reads aligned, number of variants and the standardized quality of variant sites (ratio between average quality of variant/average quality of non-variant) following Nicholls et al. (2015).

2. *de novo* assembly following Yang and Smith (2014)

   This pipeline infers orthologues using a phylogenetic approach (Gabaldón, 2008). After initial quality cleaning (eg. Trimmomatic), it uses a *de novo* assembler (SPAdes - Bankevich et al. (2012)) to create scaffolds for each accession. The scaffolds consist of contigs and gaps and are the sequences used for later analysis. All scaffolds are then blasted (BLASTN) against each other in a all-by-all blast. Homology is then inferred by two different clustering steps. First, the blast results are filtered by hit fraction (E-value). The E-value assigned to each hit represents the number of alternative matches with a similar score expected by chance in a database of that size. The lower the E-value, the smaller the probability of that match being by chance. The minimal value used in the pipeline is 0.3. This step aims to produce clusters with a reasonable size that can be aligned by filtering out hits from conserved motifs for example. The filtered clusters are then subjected to a Markov Cluster Algorithm analysis - MCL (Enright, 2002) to infer final homologue clusters. Given a network of nodes (scaffolds) connected to each other (E-value), the algorithm will break the connections based on their presence or absence and strength. The filtering during the MCL step is done using the inflation value (I) that is set by the user. A smaller value (eg. $I = 1.2$) produces coarser clusters. i.e. clusters that allow less strongly connected nodes. A higher value (eg. $I = 2$) produces fine grained clusters, i.e. tighter clusters that are strongly connected. Therefore homology is inferred in a two-step filtering process. The final homologue clusters are then aligned for tree inference. For each tree representing homologous sequences, orthology is inferred using four different approaches: maximum inclusion, rooted ingroups, monophyletic outgroups and one-to-one ortologues.

3. HypPiper pipeline, which combines mapping back and *de novo* tools following (Johnson et al., 2016)

   The input for the HybPiper is the bait sequence that is used as a target file and the forward and reverse paired reads. Those reads should be in fastq format and cleaned from poor quality base calls and adapters using Trimmomatic, for example. Version 1.2 allows the inclusion of the forward unpaired reads as well, but not the reverse unpaired reads.

   The HybPiper pipeline starts by mapping the raw reads back to the bait sequence (also called target file or reference file) using BWA. This step is done with BWA default values and will include paralogues and/or alleles if present. The second step is a *de novo* assembly of the reads assigned to each locus using SPAdes (Bankevich et al., 2012).

   The HybPiper pipeline also aims to recover contigs including the intronic region adjacent to the coding regions of the target file. The extension of the sequence recovered for each locus might improve informative characters and phylogenetic resolution in shallower phylogenies of recently evolved species. This is done using Exonerate (Slater and Birney, 2005), an heuristic alignment approach which integrates splice site prediction to allow the incorporation of the intronic regions. The recovered contigs from SPAdes are aligned to the reference using the protein2genome alignment model from Exonerate. For this alignment, the reference file is translated to protein sequence using Biopython and the resulting contigs represent the coding sequence. Those contigs are then grouped into scaffolds for each locus (contigs + gaps) generating the final supercontig composed by coding sequence and introns.

   The HybPiper pipeline identifies as potential paralogues cases where more than one contig is assembled to the same region of each target locus and represents more than 85% of the length of that locus. The contigs are sorted by depth of sequence and one of them picked for downstream analysis. The contig that is kept is chosen based on two criteria. Firstly, if it has a coverage depth 10 times

greater than the other(s) and secondly, in case of similar coverage depth, the contig with the highest percentage identity to the reference sequence calculated during the alignment with Exonerate. Those loci are flagged so they can be further investigated, for example to distinguish between paralogues sequences, alleles or contamination.

3.2.10 Phylogenetic inference

Over the past few years there has been a constant increase in the amount of DNA sequence data available for phylogenetic analysis. Concatenation of available loci in one matrix treated as a single partition has been widely used but might not be the best approach because it may hide important information such as conflict among gene trees. Furthermore, with large amounts of data, bootstrap values can be misleading and inflated (Felsenstein, 1985; Kubatko and Degnan, 2007). Ideally, a thorough phylogenetic inference of multi-locus data sets would involve the combination of at least two different approaches: concatenation and species tree/gene trees analysis. Each gene tree can tell a different evolutionary history that could be different from the species tree itself (Maddison, 1997). The hundreds and thousands of loci allow a deeper investigation about the incongruence between gene trees and the species tree and causes of it. Researchers have developed different methods for this approach (Edwards, 2009; Fujita et al., 2012) but they can be computationally demanding, especially in big data sets. Differences between the concatenation approach and species-tree inference, especially under the coalescent model, have been investigated (Lambert et al., 2015; Simmons and Gatesy, 2015; Smith et al., 2015; Guo et al., 2018). Lambert et al. (2015), for example, suggested that the concatenated approach is a reasonable proxy for the species tree, and that some properties of a phylogeny generated using a concatenated matrix might indicate where the discordance with species-tree might occur. Branches that are short, with weak bootstrap support and few concordant individual gene trees, tend to conflict with the species-tree.

3.2.11 Baits design

Two different sets of baits were used during this study: (i) 380 loci derived from the transcriptome of *Adansonia digitata* and *Bombax ceiba* checked against sequences of *Gossypium* (*Adansonia-Bombax* bait set) and (ii) 377 loci designed using *Ceiba* sequence data (*Ceiba* bait set). The loci for the *Adansonia-Bombax* bait set were chosen because they were single copy in the *Gossypium* genome, therefore minimising issues of paralogy. The *Ceiba* bait set was used to test whether baits designed using *Ceiba* sequence data would improve capture success (Chau et al., 2018).

The data from the first sequencing run were generated by hybridisation of library fragments of 24 samples of *Ceiba* with the 380 target loci from the *Adansonia-Bombax* bait set. Because the *Adansonia-Bombax* bait set was generated from the transcriptome of those species (Karimi et. al. in prep.), it contains only coding sequence. However, during capture reaction, fragments of library representing flanking regions to the targets potentially get captured and sequenced (Tsangaras et al., 2014). The off-target sequences might contain intronic regions, which would be useful for phylogenetic analysis, especially amongst recently diverged species, because they are likely to be sequence-variable. I analysed the data from the first sequencing run following the Yang and Smith (2014) pipeline, which produces *de novo* scaffolds containing both coding sequence and possibly introns. I used BLAST to blast the *Adansonia-Bombax* bait set sequence against the *de novo* scaffolds derived from the Yang and Smith (2014) pipeline for all accessions. The top blast hit was chosen as a candidate new bait set (i.e., *Ceiba* bait set). I then used the candidate new bait set as a reference to map back the raw reads from the same first 24 samples following Nicholls et al. (2015) pipeline. With this last step I aimed to assure that a data set generated by hybridisation with baits derived from different genera (first sequencing run or first round) would produce data with enough quality and coverage when mapped back against a new bait set produced using *Ceiba* sequence and therefore allow the data from future captures to be analysed together.

3.3 Results

3.3.1 Final libraries and raw reads

A total of 103 samples were sequenced representing all 18 species of *Ceiba* described so far. Library fragment size varied from 355 bp (accession TS2283) to 686 (accession FP173) (Figures 3.2 to 3.5).

The number of reads varied from 113,386 for accession MS4363 and 1,945,568 for accession IV9619 (Table 3.3, Figure 3.6).

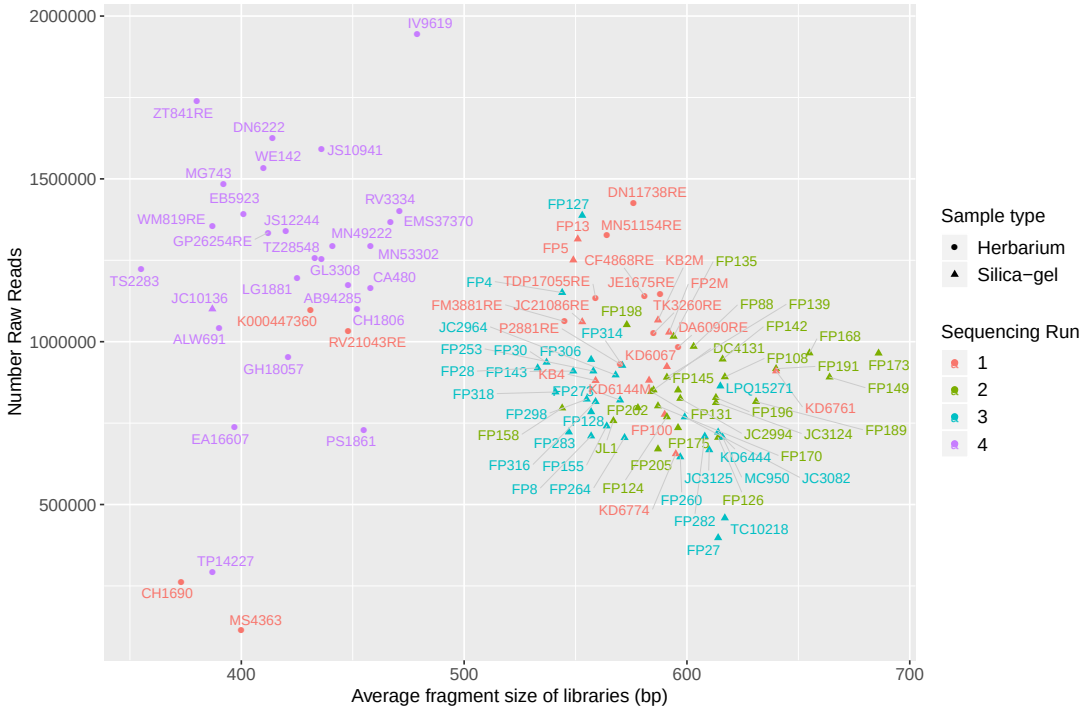


Figure 3.2: Average size of fragments in each library (bp) x number of reads recovered for each library.

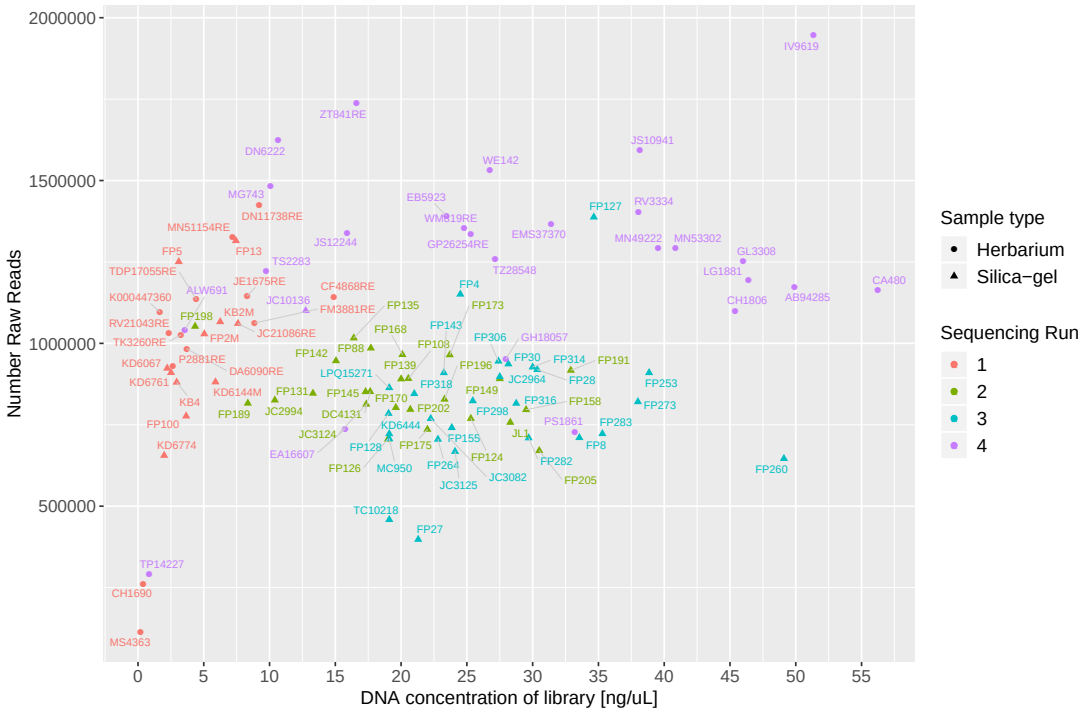


Figure 3.3: Final concentration of libraries after normalization (ng/μL) x number of reads recovered for each library.

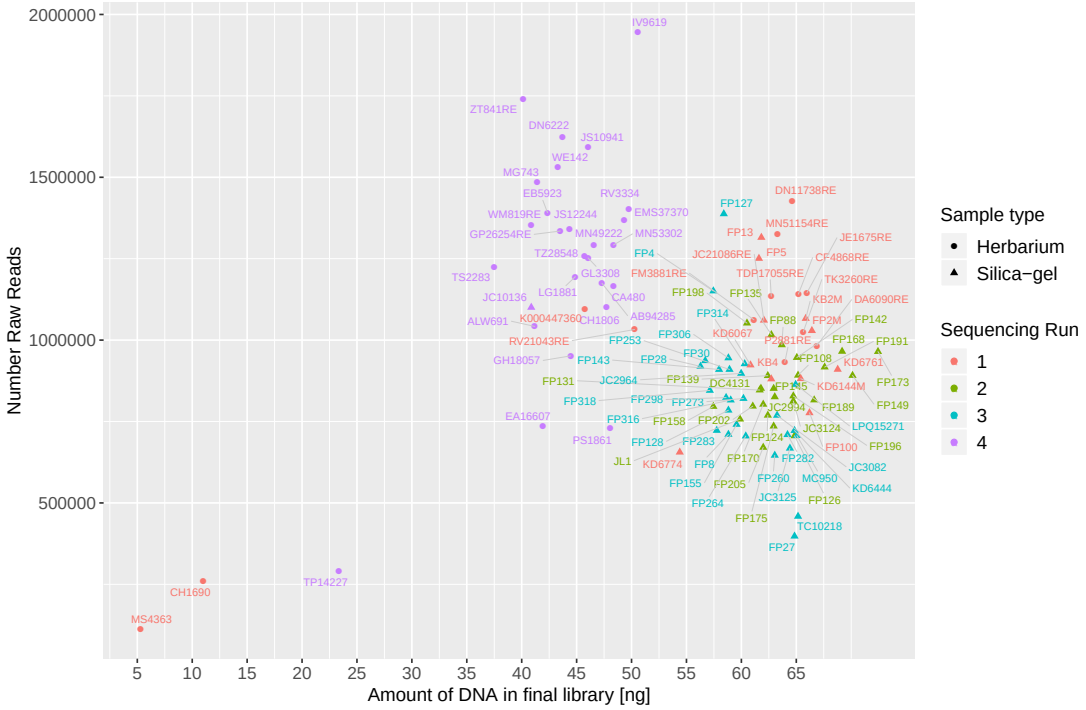


Figure 3.4: Amount of DNA recovered in each final library (ng) x number of reads recovered for each library.

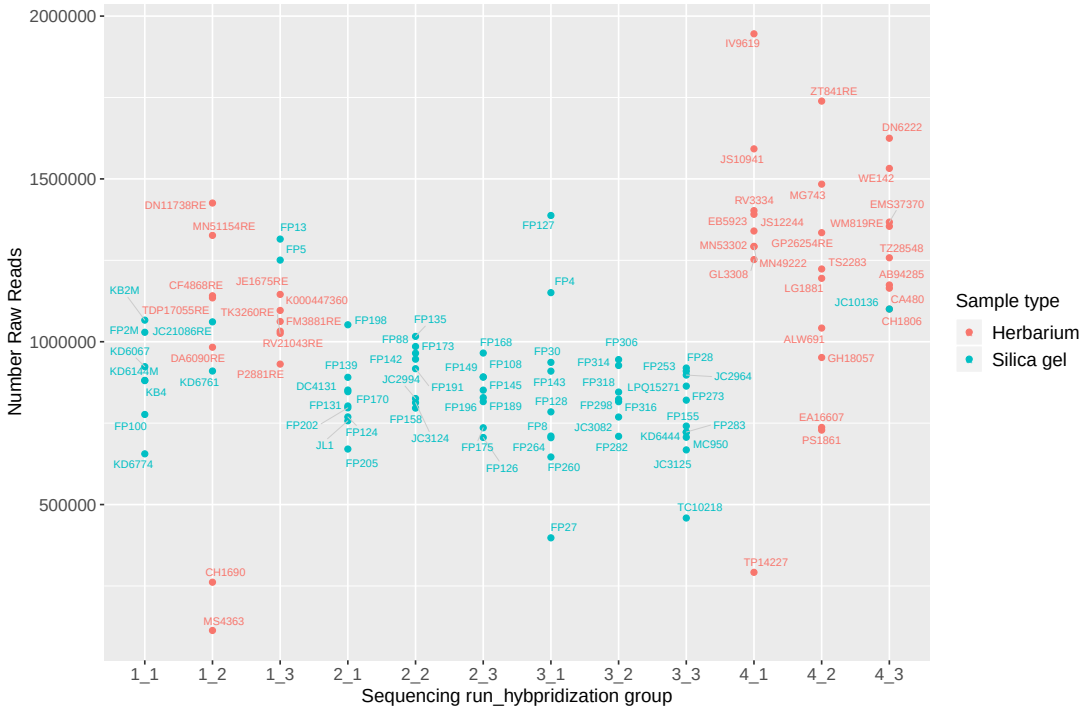


Figure 3.5: Groups of hybridisation of each sequencing run [sequecing run]_[hybridisation pool].

Table 3.3: Library quality and number of raw reads recovered for each sample.

n	Accession	Number of Reads	Sample type	Species	Sequencing Run	Capture Pool	Average size (bp)	ng/uL	Amount DNA (ng)
1	AB94285	1,174,374	herbarium	*Ceiba aesculifolia aesculifolia*	4	3	448	49.9	47.31
2	JC10136	1,100,446	silica gel	*Ceiba aesculifolia aesculifolia*	4	3	387	12.75	40.87
3	JE1675RE	1,145,412	herbarium	*Ceiba aesculifolia aesculifolia*	1	3	588	8.29	65.97
4	CH1806	1,100,314	herbarium	*Ceiba aesculifolia parvifolia*	4	3	452	45.4	47.73
5	JC21086RE	1,061,044	silica gel	*Ceiba aesculifolia parvifolia*	1	2	553	7.6	62.05
6	JS12244	1,340,080	herbarium	*Ceiba boliviana*	4	1	420	15.9	44.35
7	MN53302	1,292,260	herbarium	*Ceiba boliviana*	4	1	458	40.85	48.36
8	KD6761	910,080	silica gel	*Ceiba boliviana*	1	2	640	2.51	68.77
9	EB5923	1,391,354	herbarium	*Ceiba chodatii*	4	1	401	23.45	42.35
10	FM3881RE	1,062,248	herbarium	*Ceiba chodatii*	1	3	545	8.86	61.15

Table 3.3 continued from previous page

n	Accession	Number of Reads	Sample type	Species	Sequencing Run	Capture Pool	Average size (bp)	ng/uL	Amount DNA (ng)
11	MN51154RE	1,326,580	herbarium	*Ceiba chodatii*	1	2	564	7.18	63.28
12	LG1881	1,194,798	herbarium	*Ceiba crispiflora*	4	2	425	46.4	44.88
13	MG743	1,483,950	herbarium	*Ceiba crispiflora*	4	2	392	10.07	41.4
14	FP124	769,158	silica gel	*Ceiba erianthos*	2	1	591	25.3	62.41
15	FP126	706,008	silica gel	*Ceiba erianthos*	2	3	614	19	64.84
16	FP13	1,315,272	silica gel	*Ceiba erianthos*	1	3	551	7.44	61.82
17	FP205	670,518	silica gel	*Ceiba erianthos*	2	1	587	30.5	61.99
18	FP282	709,576	silica gel	*Ceiba erianthos*	3	2	608	29.7	64.2
19	JC3082	768,864	silica gel	*Ceiba erianthos*	3	2	599	22.25	63.25
20	JL1	757,238	silica gel	*Ceiba erianthos*	2	1	567	28.3	59.88
21	LPQ15271	863,770	silica gel	*Ceiba erianthos*	3	3	615	19.1	64.94
22	MC950	706,736	silica gel	*Ceiba erianthos*	3	3	616	19.1	65.05
23	P2881RE	931,332	herbarium	*Ceiba erianthos*	1	3	570	2.64	63.95
24	DC4131	851,734	silica gel	*Ceiba glaziovii*	2	1	585	17.65	61.78
25	FP202	797,012	silica gel	*Ceiba glaziovii*	2	1	578	20.7	61.04

Table 3.3 continued from previous page

n	Accession	Number of Reads	Sample type	Species	Sequencing Run	Capture Pool	Average size (bp)	ng/uL	Amount DNA (ng)
26	FP298	823,432	silica gel	*Ceiba glaziovii*	3	2	555	25.45	58.61
27	FP2M	1,029,054	silica gel	*Ceiba glaziovii*	1	1	592	5.04	66.42
28	FP306	945,268	silica gel	*Ceiba glaziovii*	3	2	557	27.4	58.82
29	FP314	927,156	silica gel	*Ceiba glaziovii*	3	2	571	30	60.3
30	FP316	815,836	silica gel	*Ceiba glaziovii*	3	2	559	28.75	59.03
31	FP318	845,486	silica gel	*Ceiba glaziovii*	3	2	541	21	57.13
32	FP88	985,416	silica gel	*Ceiba glaziovii*	2	2	603	17.7	63.68
33	FP260	646,152	silica gel	*Ceiba glaziovii*	3	1	597	49.1	63.04
34	FP264	705,306	silica gel	*Ceiba glaziovii*	3	1	572	22.8	60.4
35	FP4	1,150,804	silica gel	*Ceiba glaziovii*	3	1	544	24.5	57.45
36	FP5	1,250,764	silica gel	*Ceiba glaziovii*	1	3	549	3.1	61.6
37	GL3308	1,252,336	herbarium	*Ceiba insignis*	4	1	436	46	46.04
38	IV9619	1,945,568	herbarium	*Ceiba insignis*	4	1	479	51.35	50.58
39	TP14227	291,922	herbarium	*Ceiba insignis*	4	1	387	0.85	23.34
40	FP155	741,050	silica gel	*Ceiba jasminodora*	3	3	564	23.85	59.56

Table 3.3 continued from previous page

n	Accession	Number of Reads	Sample type	Species	Sequencing Run	Capture Pool	Average size (bp)	ng/uL	Amount DNA (ng)
41	FP158	796,240	silica gel	*Ceiba jasminodora*	2	2	544	29.5	57.45
42	MS4363	113,386	herbarium	*Ceiba jasminodora*	1	2	400	0.19	5.29
43	CA480	1,164,488	herbarium	*Ceiba lupuna*	4	3	458	56.25	48.36
44	DN6222	1,624,984	herbarium	*Ceiba lupuna*	4	3	414	10.65	43.72
45	EMS37370	1,367,478	herbarium	*Ceiba pentandra*	4	3	467	31.4	49.32
46	KB2M	1,066,350	silica gel	*Ceiba pentandra*	1	1	587	6.25	65.86
47	KB4	880,666	silica gel	*Ceiba pentandra*	1	1	559	2.94	62.72
48	RV21043RE	1,032,488	herbarium	*Ceiba pentandra*	1	3	448	2.35	50.27
49	WM819RE	1,354,440	herbarium	*Ceiba pentandra*	4	3	387	24.8	40.87
50	ALW691	1,042,048	herbarium	*Ceiba pubiflora*	4	2	390	3.57	41.18
51	FP127	1,387,740	silica gel	*Ceiba pubiflora*	3	1	553	34.65	58.4
52	FP128	784,680	silica gel	*Ceiba pubiflora*	3	1	557	19.05	58.82
53	FP135	1,016,438	silica gel	*Ceiba pubiflora*	2	2	594	16.4	62.73
54	FP142	946,458	silica gel	*Ceiba pubiflora*	2	2	616	15.05	65.05
55	FP143	909,780	silica gel	*Ceiba pubiflora*	3	1	549	23.25	57.97

Table 3.3 continued from previous page

n	Accession	Number of Reads	Sample type	Species	Sequencing Run	Capture Pool	Average size (bp)	ng/uL	Amount DNA (ng)
56	FP145	851,442	silica gel	*Ceiba pubiflora*	2	3	596	17.3	62.94
57	FP149	891,324	silica gel	*Ceiba pubiflora*	2	3	664	27.5	70.12
58	FP168	965,340	silica gel	*Ceiba pubiflora*	2	3	655	20.1	69.17
59	FP170	802,810	silica gel	*Ceiba pubiflora*	2	1	587	19.6	61.99
60	FP173	964,768	silica gel	*Ceiba pubiflora*	2	2	686	23.7	72.44
61	FP175	735,844	silica gel	*Ceiba pubiflora*	2	3	596	22	62.94
62	FP189	816,100	silica gel	*Ceiba pubiflora*	2	3	631	8.35	66.63
63	FP191	917,180	silica gel	*Ceiba pubiflora*	2	2	640	32.9	67.58
64	FP196	828,404	silica gel	*Ceiba pubiflora*	2	3	613	23.3	64.73
65	FP198	1,052,140	silica gel	*Ceiba pubiflora*	2	1	573	4.34	60.51
66	FP273	820,664	silica gel	*Ceiba pubiflora*	3	3	570	38	60.19
67	FP30	936,710	silica gel	*Ceiba pubiflora*	3	1	537	28.15	56.71
68	FP8	710,394	silica gel	*Ceiba pubiflora*	3	1	557	33.55	58.82
69	GP26254RE	1,334,956	herbarium	*Ceiba pubiflora*	4	2	412	25.3	43.51
70	FP100	776,694	silica gel	*Ceiba rubriflora*	1	1	590	3.66	66.2

Table 3.3 continued from previous page

n	Accession	Number of Reads	Sample type	Species	Sequencing Run	Capture Pool	Average size (bp)	ng/uL	Amount DNA (ng)
71	FP108	891,998	silica gel	*Ceiba rubriflora*	2	3	617	20.55	65.16
72	JC2994	825,694	silica gel	*Ceiba rubriflora*	2	2	597	10.4	63.04
73	KD6067	923,554	silica gel	*Ceiba samauma*	1	1	591	2.22	60.83
74	KD6774	655,796	silica gel	*Ceiba samauma*	1	1	595	1.99	54.39
75	FP131	846,604	silica gel	*Ceiba samauma*	2	1	584	13.3	61.67
76	FP139	890,804	silica gel	*Ceiba samauma*	2	1	591	20	62.41
77	FP27	397,860	silica gel	*Ceiba samauma*	3	1	614	21.3	64.84
78	K000447360	1,096,108	herbarium	*Ceiba samauma*	1	3	431	1.67	45.76
79	TDP17055RE	1,134,646	herbarium	*Ceiba samauma*	1	2	559	4.43	62.72
80	DA6090RE	983,152	herbarium	*Ceiba schottii*	1	2	596	3.71	66.87
81	PS1861	728,648	herbarium	*Ceiba schottii*	4	2	455	33.2	48.05
82	CH1690	261,562	herbarium	*Ceiba soluta*	1	2	373	0.4	11.01
83	KD6144M	881,404	silica gel	*Ceiba speciosa*	1	1	583	5.89	65.41
84	KD6444	722,256	silica gel	*Ceiba speciosa*	3	3	614	19.1	64.84
85	TK3260RE	1,025,678	herbarium	*Ceiba speciosa*	1	3	585	3.27	65.64

Table 3.3 continued from previous page

n	Accession	Number of Reads	Sample type	Species	Sequencing Run	Capture Pool	Average size (bp)	ng/uL	Amount DNA (ng)
86	CF4868RE	1,141,188	herbarium	*Ceiba speciosa*	1	2	581	14.9	65.19
87	FP283	722,554	silica gel	*Ceiba speciosa*	3	3	547	35.3	57.76
88	JS10941	1,592,314	herbarium	*Ceiba speciosa*	4	1	436	38.15	46.04
89	MN49222	1,292,922	herbarium	*Ceiba speciosa*	4	1	441	39.55	46.57
90	RV3334	1,402,700	herbarium	*Ceiba speciosa*	4	1	471	38.05	49.74
91	DN11738RE	1,425,966	herbarium	*Ceiba trischistandra*	1	2	576	9.2	64.63
92	EA16607	736,922	herbarium	*Ceiba trischistandra*	4	2	397	15.75	41.92
93	GH18057	951,624	herbarium	*Ceiba trischistandra*	4	2	421	27.95	44.46
94	FP253	909,794	silica gel	*Ceiba ventricosa*	3	3	558	38.85	58.92
95	JC3124	812,880	silica gel	*Ceiba ventricosa*	2	2	613	17.35	64.73
96	TS2283	1,223,426	herbarium	*Ceiba ventricosa*	4	2	355	9.74	37.49
97	ZT841RE	1,738,996	herbarium	*Ceiba ventricosa*	4	2	380	16.6	40.13
98	JC3125	667,924	silica gel	*Eriotheca obcordata*	3	3	610	24.1	64.42
99	TC10218	458,714	silica gel	*Neobuchia paulinae*	3	3	617	19.1	65.16
100	TZ28548	1,257,840	herbarium	*Neobuchia paulinae*	4	3	433	27.15	45.72

Table 3.3 continued from previous page

n	Accession	Number of Reads	Sample type	Species	Sequencing Run	Capture Pool	Average size (bp)	ng/uL	Amount DNA (ng)
101	WE142	1,532,348	herbarium	*Neobuchia paulinae*	4	3	410	26.75	43.3
102	FP28	919,332	silica gel	*Pseudobombax marginatum*	3	3	533	30.35	56.28
103	JC2964	897,702	silica gel	*Spirotheca elegans*	3	3	568	27.5	59.98

3.3.2 Quality check

I evaluated the impact of the different Trimmomatic filters applied to the raw reads by comparing the percentage of paired forward and reverse reads surviving (Figure 3.7), only forward reads surviving (i.e. passed the filtering process - Figure 3.8), only reverse reads surviving (Figure 3.9), reads dropped (Figure 3.10), and average Phred quality scores at each base pair of surviving reads (Figure 3.11).

The different Trimommatic filters revealed two main patterns. The settings 3-4:15 and 20-4:15 showed similar percentage of both surviving reads, forward only, reverse only, reads dropped and Phred score for all accessions. Likewise, the settings 3-4:20, 20-4:20, 3-5:20 and 20-5:20 showed similar patterns for all variables measured. The variation in the leading, trailing (3 or 20) and the number of base pairs in the sliding window setting (4 or 5) did not influence the quantity and quality of surviving reads in both directions. The Phred quality threshold value within the sliding window setting (15 or 20) was the main factor influencing the percentage of reads surviving in the two patterns observed.

The percentage of reads recovered for the less conservative filters (3-4:15) varied from 73.1% (accession CH1690) to 99.8% (accession FP260). For the 3-4:20 filter setting, the percentage of both pairs surviving varied from 65.2% (accession CH1690) to 98.1% (accession FP158) (Figure 3.7). The main difference between the filters was observed in the percentage of only reverse reads surviving. For 17 of the samples the percentage of only reverse reads surviving was higher than 1% for the 3-4:20 filter. On the other hand, for the filter 3-4:15 all samples had less than 0.76% of only reverse readings surviving (Figure 3.9).

Likewise, the variation in the Trimmomatic filters revealed two main patterns related to the average Phred score (Figure 3.11). Each accession is represented in the graph by four lines representing the four output group files from Trimmomatic: forward paired, forward unpaired, reverse paired and reverse unpaired. The graph (Figure 3.11) show the average quality Phred score (Y axis) for each base pair position in a 150bp long read (X axis) for each output group. The reverse unpaired reads are represented

in a lighter colour tone. For all Trimmomatic settings the worst quality scores are found at the beginning of the reads. However, the settings 3-4:15 and 20-4:15 showed lower quality values than the other settings with reads scattered along the Y axis of the graph, especially the reverse unpaired reads represented in a lighter colour tone on Figure 3.11.

Because the HybPiper (Johnson et al., 2016) and Yang and Smith (2014) pipelines are not optimized to work with reverse unpaired reads, the amount of reads discarded in the 3-4:20 Trimmomatic filter could potentially represent a waste of sequencing effort. To test the impact of the initial quality filtering on downstream analysis, HybPiper (not using reverse unpaired reads) and Nicholls et al. (2015) pipelines (using reverse unpaired reads) were run using two different sets of trimmed data (3-4:15 and 3-4:20) as input. For the Nicholls et al. (2015) pipeline the difference between the two inputs was accessed in terms of % aligned, number of reads, average quality of reads, average quality of non-variant, average quality of variant, number of variants and number of non-variants of the vcf file for two Bowtie2 alignment scores threshold, 20 and 190. Within each Bowtie2 score threshold both inputs show similar patterns for all variables measured (Figures 3.12 to 3.16, see also Figures 3.22 to 3.24 and appendix Figures 3.43 to 3.48 and 3.49 to 3.51 for comparison among more Bowtie2 thresholds).

For HybPiper, the difference was accessed in terms of percentage of each locus covered by sequence, number of reads mapped, percentage of reads mapped, number of loci with contigs and number of paralogue warnings (Figures 3.17 to 3.21). Likewise, the two different inputs showed similar patterns for all metrics.

All pipelines include further steps of data filtering later in analyses (eg. additional trimming with Cutadapt version 1.15 (Martin, 2011) and variant call quality filtering in the vcf file, both in Nicholls et al. (2015) pipeline). Using the more permissive Trimmomatic filtering (3-4:15), even if a base pair is low quality it would then be filtered out on those data filtering steps without the need to discard the entire read. Thus, the regions of the read with good quality (Figure 3.11) would still be used without compromising the quality of the analysis. Therefore, the 3-4:15 filtering was considered the most

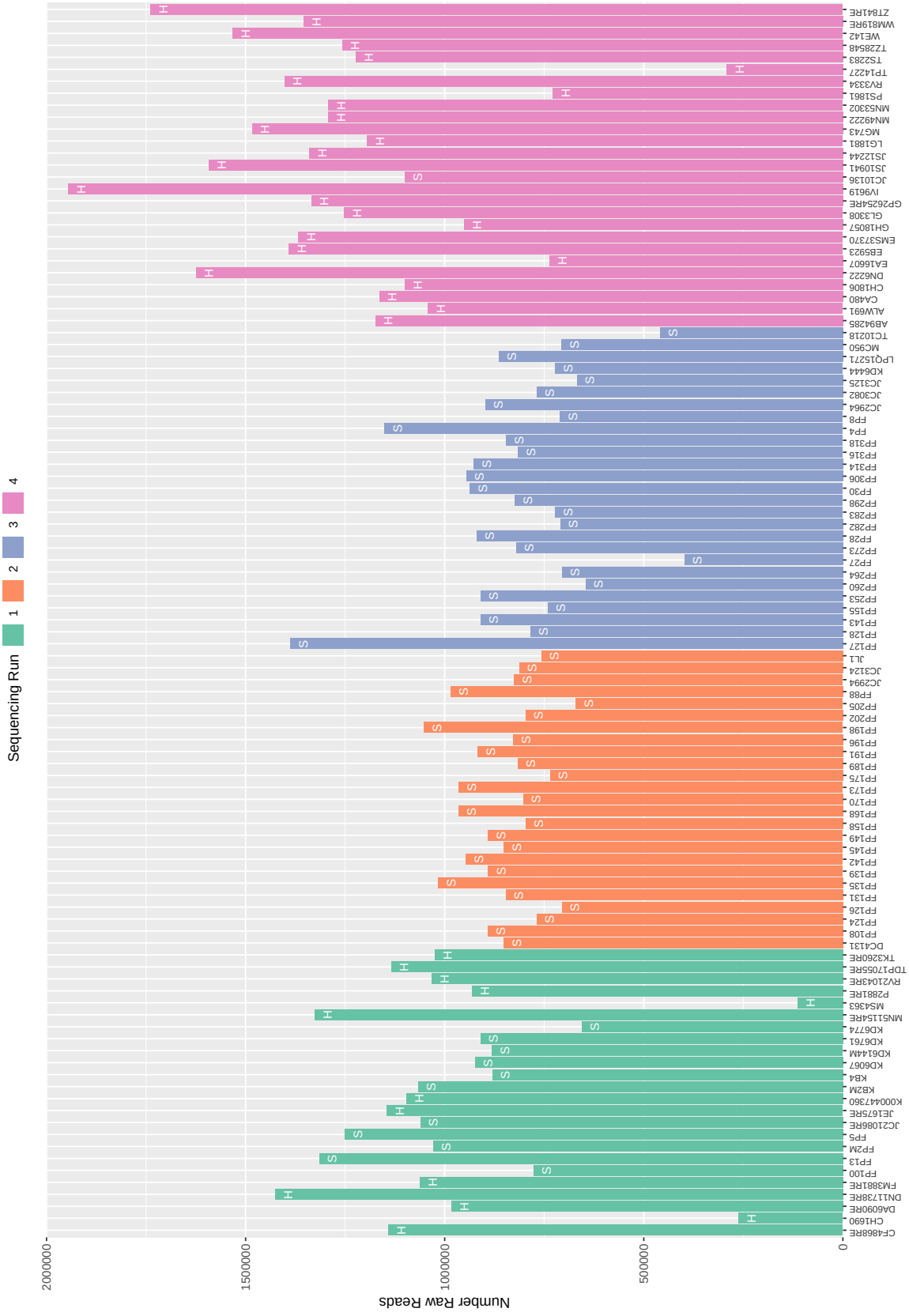


Figure 3.6: Number of raw reads recovered for each sample. Colours represent the sequencing run (1, 2, 3 or 4) and letters on bars represent type of sample (H) herbarium and (S) silica-gel dried.

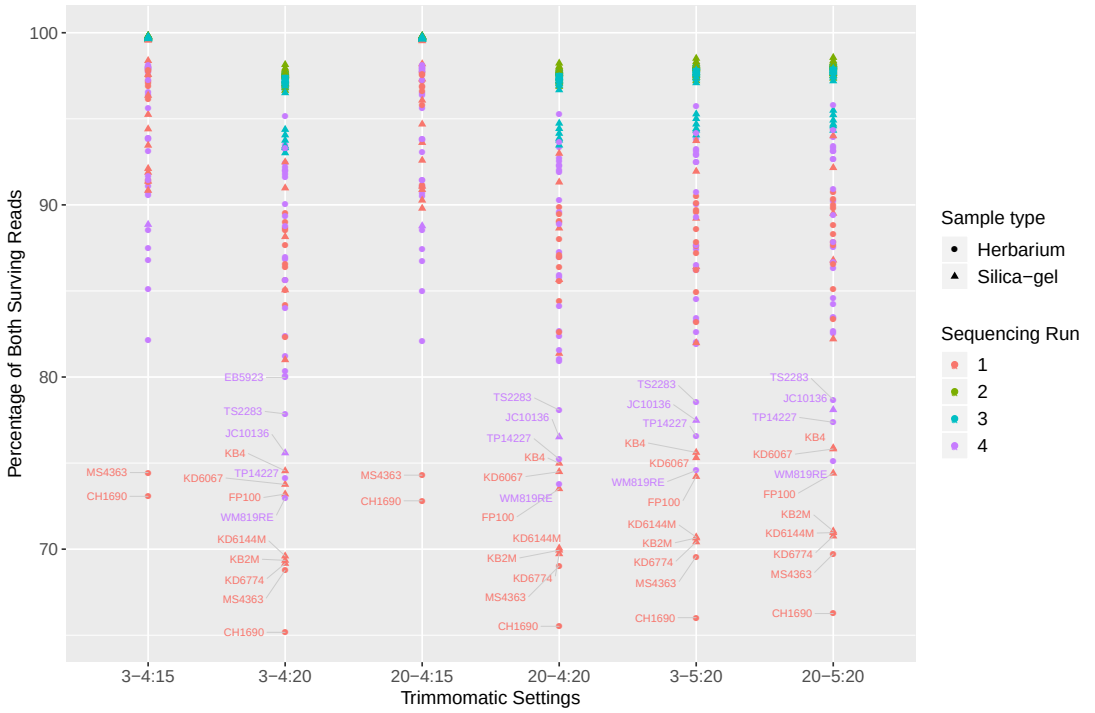


Figure 3.7: Percentage of paired forward and reverse reads surviving after Trimmomatic trimming. Samples with less than 80% of forward and reverse reads surviving are labelled.

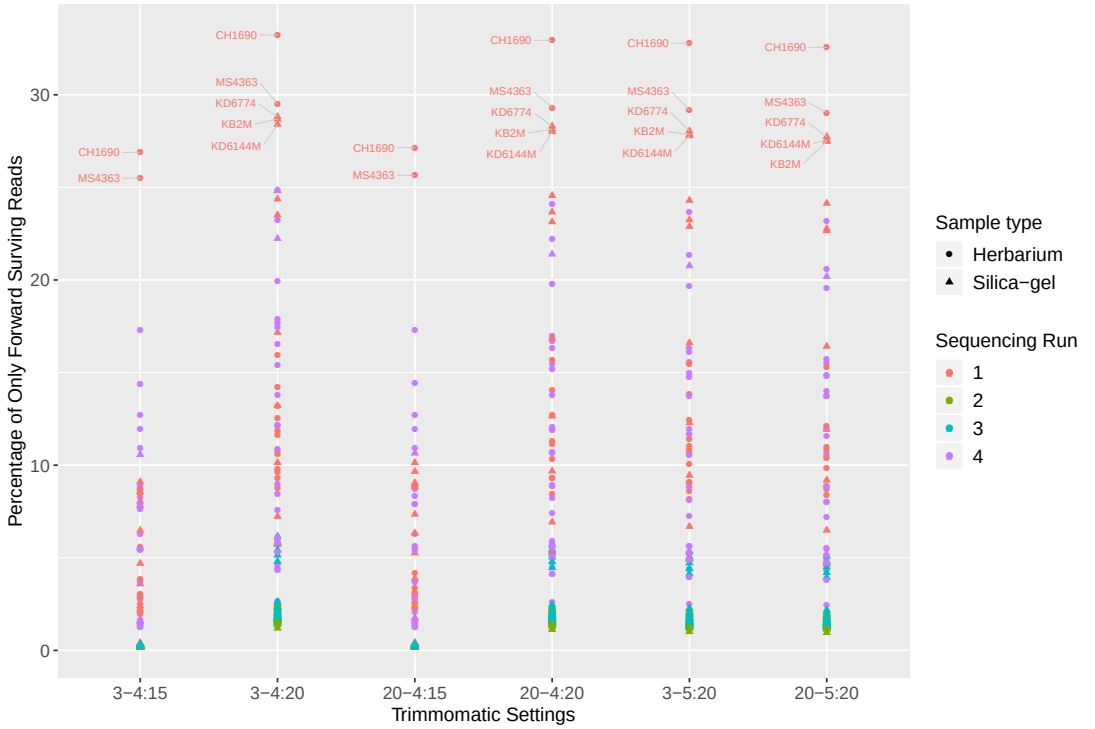


Figure 3.8: Percentage of forward reads surviving after Trimmomatic trimming. Samples with more than 25% of only forward reads surviving are labelled.

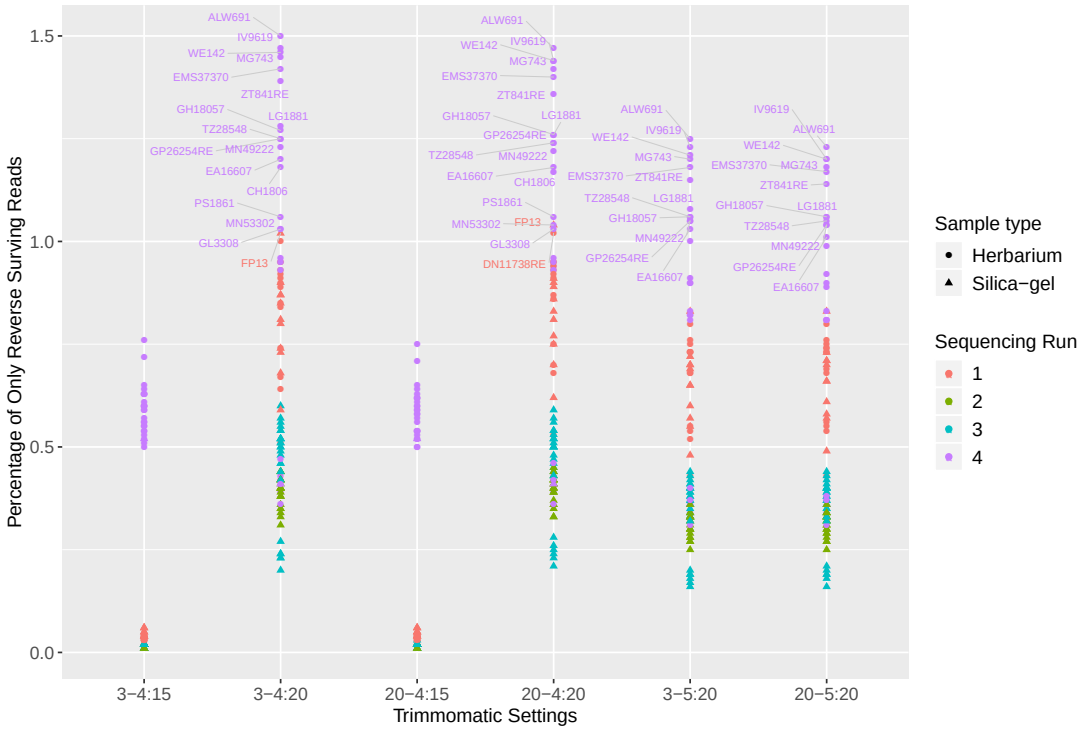


Figure 3.9: Percentage of reverse reads surviving after Trimmomatic trimming. Samples with more than 10% of only reverse reads surviving are labelled.

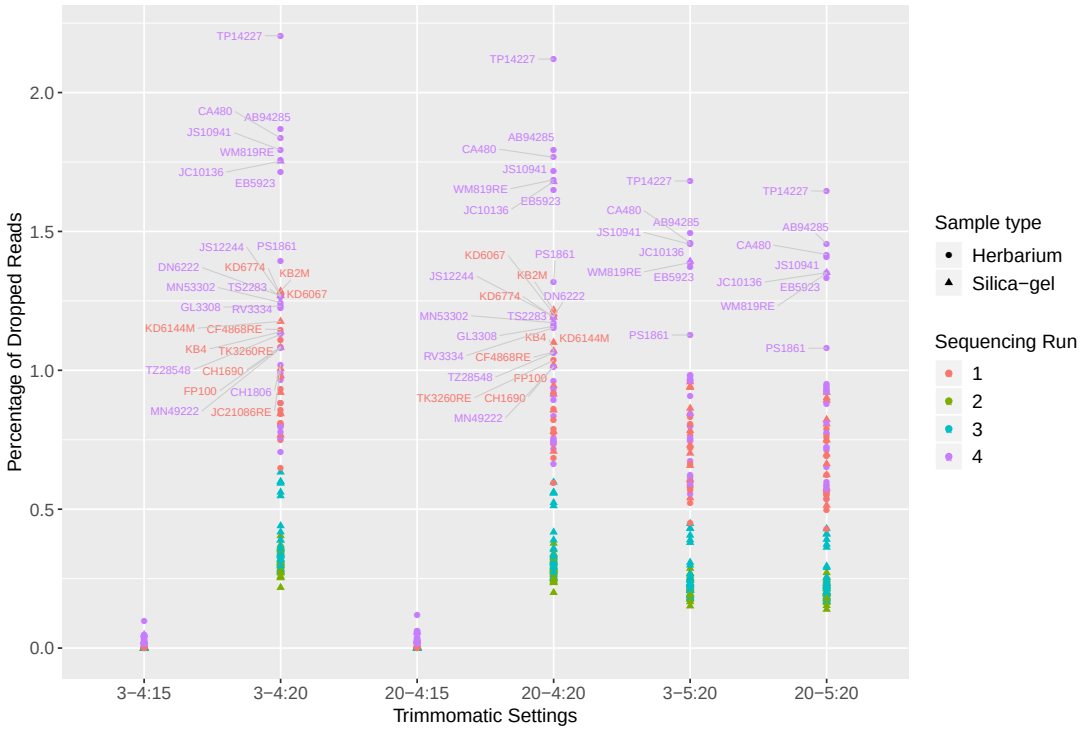


Figure 3.10: Percentage of dropped reads. Samples with more than 1% of reads dropped are labelled.

suitable in this case and Trimmomatic was finally called with the following command: `PE -phred33 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`.

3.3.3 Data assembly

1. Nicholls et al. (2015)

   For Bowtie2, the increase in the constant value in the threshold alignment score had an impact on all the metrics measured on the resulting vcf files (Figures 3.22 to 3.24). The percentage of reads aligned showed an overall constant decrease rate until the constant value of 130 when it started to decrease rapidly, particularly for the three accessions representing the outgroups (JC3125(*Eriotheca*), JC2964 (*Spirotheca*) and FP28 (*Pseudobombax*), Figure 3.22). The number of variants (Figure 3.23) showed an overall decrease until the constant of 210 for all samples except for the three outgroups that showed a rapid decrease from the value of 140. This decrease in variants indicates that fewer reads from possible paralogues are being mapped to a locus. The standardized quality of variant sites showed a peak between 170 and 200 for most accessions (Figure 3.24). The peak for each accession suggests that at those constant values the fewer reads from paralogues being mapped to a locus are impacting positively on the quality of the variant called. This suggests that the optimum constant value would be between 130 and 200. A visual examination of the bam files generated with the constant value of 130 for two accessions (FP13, JL1 - especially for the nodes with paralogue warnings from the HybPiper pipeline) revealed that putative paralogues were still being mapped to the reference indicated by the presence of reads with many variants at a given site. The bam files generated with the constant of 190 did not indicate the presence of paralogues and this value was chosen as the most suitable one for downstream analysis.

   For BWA, the increase in the constant value in the threshold alignment score had an impact on all the metrics measured on the resulting vcf files (Figures 3.25

to 3.27). The percentage of reads aligned showed a decrease rate until it stabilized at the value of 10 (Figure 3.25). The number of variants (Figure 3.26) showed a rapid decrease rate until the value of 20 when it stabilised close to zero for all samples. This decrease in variants indicates that fewer reads from paralogues are being mapped to a locus until only reads similar to the reference are kept. The standardized quality of variant sites showed a peak between 5 and 10 for most accessions (Figure 3.27). A value of 20 could seem a good choice for an optimum threshold value because it would guarantee that the same copy of a given paralogue would be recovered for each locus given the absence of variants between the reads and the reference with minimal impact in the percentage of reads mapped. However, at the value of 20, most accessions show a standardized quality of variant sizes lower than one, indicating that there is a drop on the quality of the non-variant sites. A visual examination of the bam files generated with the threshold value of 10 for two accessions (FP13, JL1 - especially for the nodes with paralogue warnings from the HybPiper pipeline) did not indicate the presence of paralogues and this value was chosen as the most suitable one for downstream analysis.

BWA recovers a higher percentage of reads than Bowtie2 (Figure 3.28), more variants (Figure 3.29) for each accession and a longer final alignment (Table 3.4). However, the standardised variant quality shows similar values for both assemblers (Figure 3.30), which indicates that the extra reads and longer sequences recovered by BWA do not mean improvement of the quality of data. Furthermore, the final alignment had less informative characters for BWA (Table 3.4) than for Bowtie2. A further step to evaluate the impact of the different assemblers on the final alignments would be to compare the variants called by each for the same sites. A visual inspection of the variants called for annotated genes based on the cotton genome showed that for the sequences in common recovered by both assemblers, the same variants are called. Therefore, Bowtie2 was chosen as the most suitable assembler for this data set.

Therefore, the final data set from Nicholls et al. (2015) pipeline was derived from the 3-4:15 Trimmomatic filter, using Bowtie2 as the assembler with threshold value of 190. The percentage of reads mapped backed varied from 17.9% (accession CH1690) to 59.7%(accession FP198) (Figure 3.39).

2. HypPiper (Johnson et al., 2016)

   The HybPiper pipeline produced 69 paralogue warnings with 3-4:15 Trimmomatic filter and 67 warnings for the 3-4:20 Trimmomatic filter. For the 3-4:15 input, the loci 688 and 1331 gave warnings and for the 3-4:20 input the loci 339, 521, 1449 and 2483 gave warnings. All the other loci with warnings were common between both inputs.

   From the 377 target loci, between 366-377 were recovered for most accessions, except for accessions CH1690 where 308 loci were recovered and MS4363 where 198 were recovered (Figure 3.20). These two accessions were the only ones with no paralogue warnings (Figure 3.21). The percentage of reads mapped backed varied from 27.3% (accession CH1690) to 88.5% (accession FP198).

3. Yang and Smith (2014)

   The Yang and Smith (2014) pipeline produced final homologue trees with maximum of 50 accessions per tree, even after seven attempts of running it with different inflation values and minimal taxa. Therefore, the clusters had large amount of missing data and orthology inference was not possible.

3.3.4 Final alignments and phylogenetic inference

The final concatenated alignment length for 377 loci with the 3-4:20 input using Bowtie2 190 was 1,065,476bp with 197,788 variable sites and 115,586 parsimony informative sites (Table 3.4). For the same settings and using the 3:4-15 input, the aligment was 842,001bp long with 128,092 variable sites and 79,086 parsimony informative sites. Using BWA 10 for genome assembly and 3-4:15 as input, the alignment length was 1,072,670bp with 144,205 variable sites and 89,381 parsimony informative sites. The

final alignment from the HybPiper pipeline using the 3-4:15 input was 1,386,782bp long with 254,256 variable sites and 130,591 parsimony informative sites. For the same pipeline and using the 3-4:20 data set as input, the final alignment length was 1,321,238bp with 223,775 variable sites and 114,613 parsimony informative sites.

For each pipeline tested, I concatenated the consensus sequence for all 377 loci in a matrix and analysed under a maximum likelihood (ML) framework implemented on RAxML (Stamatakis, 2014) with 100 rapid bootstrap replicates. Recent research suggest that the time-consuming model selection step might not be essential for phylogenetic inference especially when interested only in the tree topology (Abadi et al., 2019). Although the authors suggest the widely used GTR+Γ+I model (Abadi et al., 2019), adding the proportion of invariable site (I) might cause correlation between alpha and p0 which leads to incorrect estimation of both parameters (Yang, 2014). Therefore, I analysed the concatenated matrix under the GTRGAMMA (GTR+Γ) model, as suggested by developers of RAxML (Stamatakis, 2016).

The concatenated maximum likelihood analysis produced similar topologies for all inputs and all pipelines for the individual species (Figures 3.31 to 3.35). The phylogenies were fully resolved and overall highly supported as measured by 100 bootstrap replicates. However, the relationship among the species was different for the different inputs and pipelines.

Ceiba was resolved in six main clades common to all input and pipelines: (i) *Ceiba trischistandra*, the only SDTF species occurring west of the Andes; (ii) *Ceiba pentandra*, a rain forest species occurring across the Neotropics and reaching Africa; (iii) a Central American and Mexican SDTF clade including three species of *Ceiba* and *Neobuchia paulinae*; (iv) *Ceiba samauma*, a widespread rain and semideciduous forest species from South America; (v) *Ceiba jasminodora*, the only species occuring within the Cerrado biome; and (vi) a South American SDTF clade including 11 species. These clades had high support values, but the relationships amongst them were less well supported.

Among these clades, the Central American and Mexican SDTF and *C. samauma* were always recovered as sister to each other with high support values for all inputs

and pipelines. The main conflict among the different inputs and pipelines was in relation to the placement of the *C. trichstandra*, *C. jasminodora* and *C. pentandra* clades. The Nicholls et al. (2015) pipeline recovered *C. trischistandra* as sister to all remaining species for both inputs and assemblers. Within this clade, *C. pentandra* was recovered as sister to the remaining species. The 3-4:20 Bowtie2 190 (Figure 3.32), the 3-4:15 BWA 10 (Figure 3.33) and the Hybpiper input 3-4:15 (Figure 3.34) recovered *C. jasminodora* as sister to the South American SDTF clade and the 3-4:15 Bowtie2 190 (Figure 3.31) recovered *C. jasminodora* as sister to a clade comprising *C. samauma*, the South American SDTF clade and the Central American and Mexican SDTF clade. The Hybpiper pipeline with the 3-4:20 (Figure 3.35) input recovered the most distinct topology, with *C. jasminodora* recovered as sister to all remaining species.

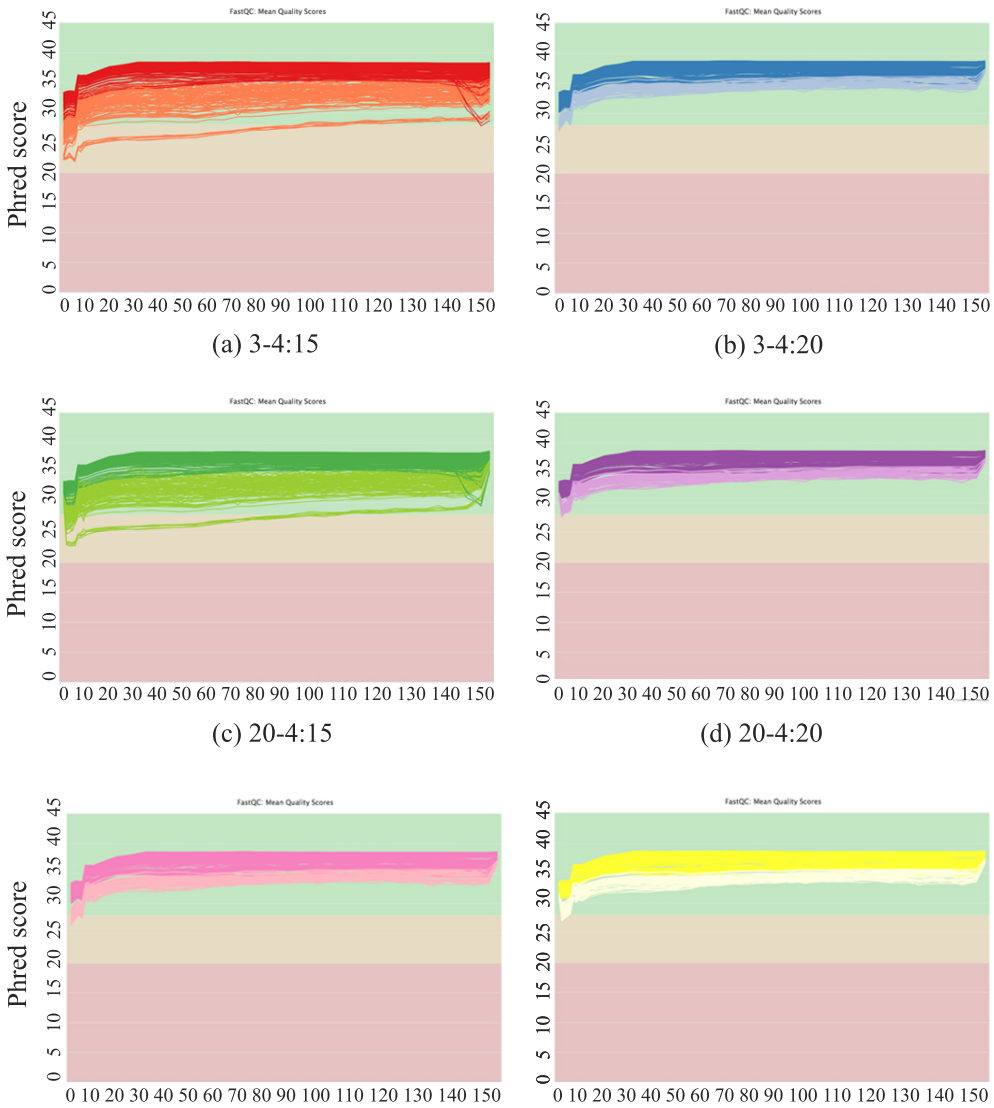


Figure 3.11: Comparison between output of different Trimmomatic filterings. Lighter tones of each colour in each graph represent the reverse unpaired reads, darker tones represent forward paired, reverse paired and forward unpaired reads.

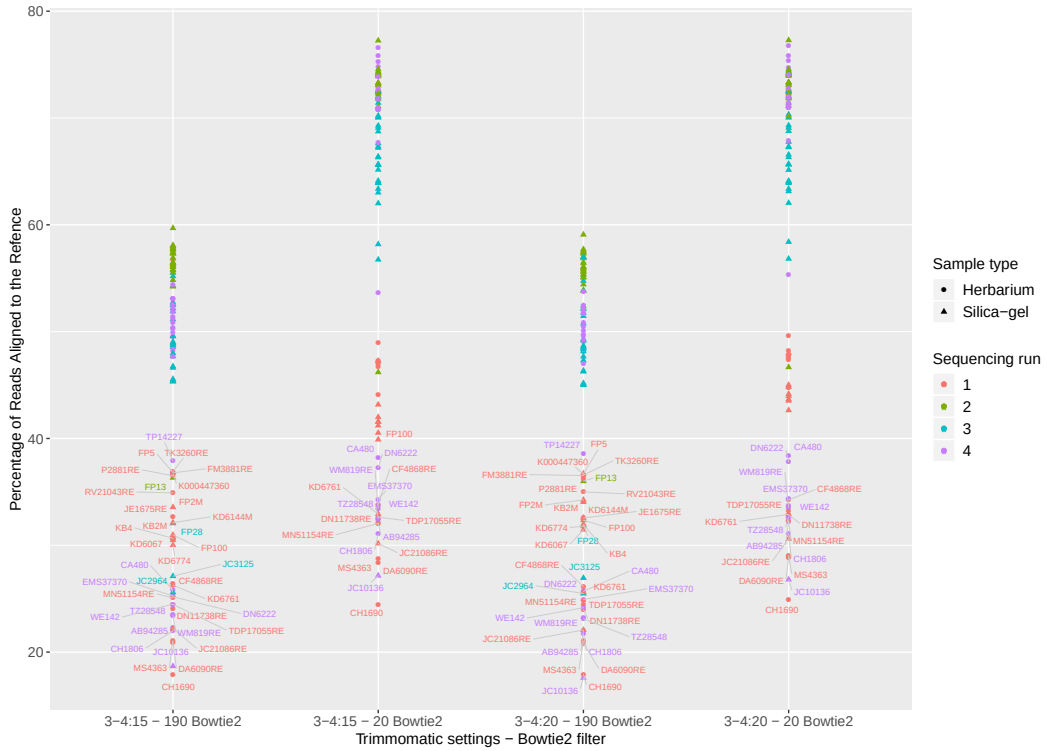


Figure 3.12: Percentage of reads mapped with two Bowtie2 threshold alignment scores (20 and 190) for two input data from Trimmomatic filtering settings (3-4:15 and 3-4:20).

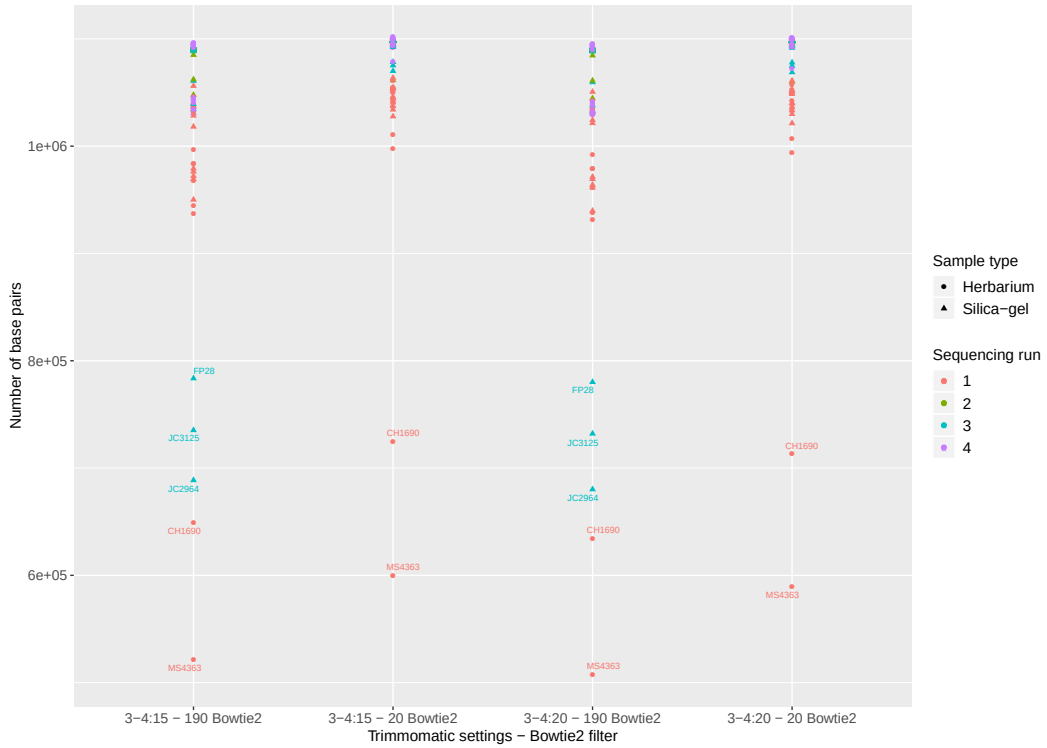


Figure 3.13: Number of base-pairs with two Bowtie2 threshold alignment scores (20 and 190) for two input data from Trimmomatic filtering settings (3-4:15 and 3-4:20).

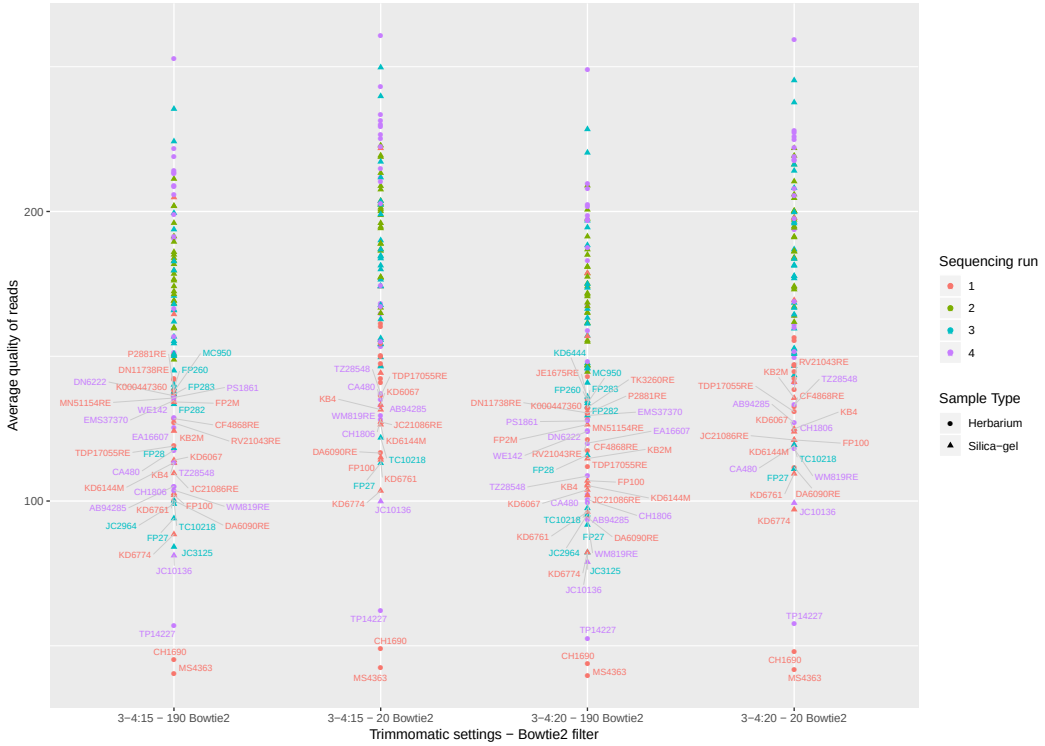


Figure 3.14: Average quality of reads with two Bowtie2 threshold alignment scores (20 and 190) for two input data from Trimmomatic filtering settings (3-4:15 and 3-4:20).

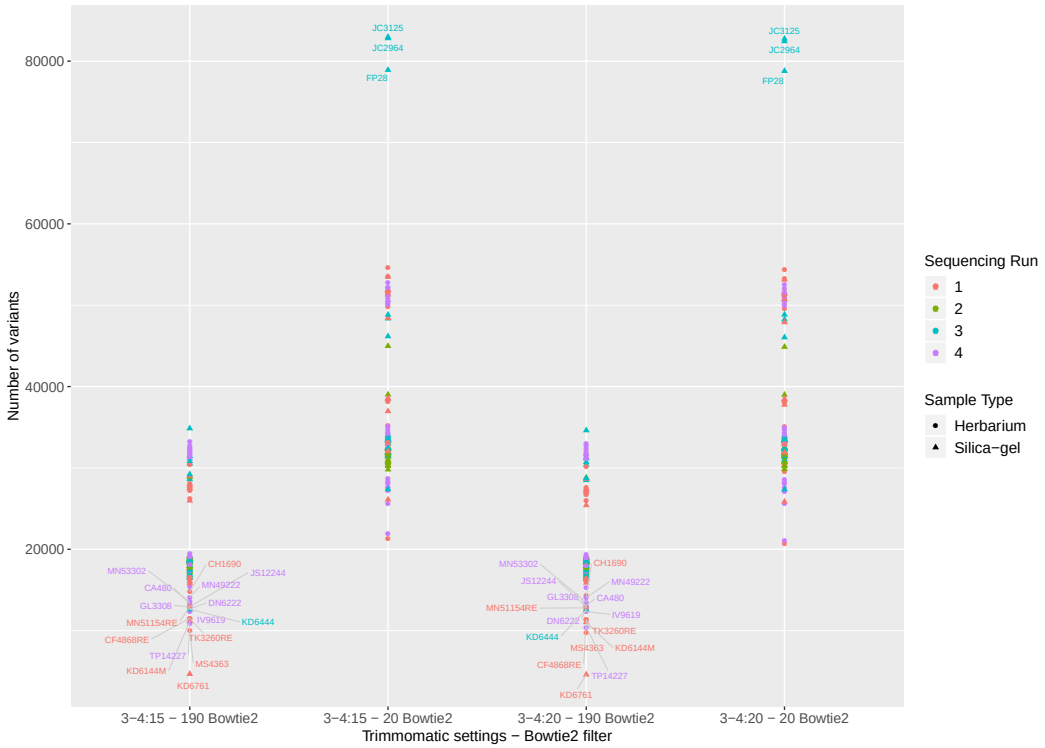


Figure 3.15: Number of variants with two Bowtie2 threshold alignment scores (20 and 190) for two input data from Trimmomatic filtering settings (3-4:15 and 3-4:20).

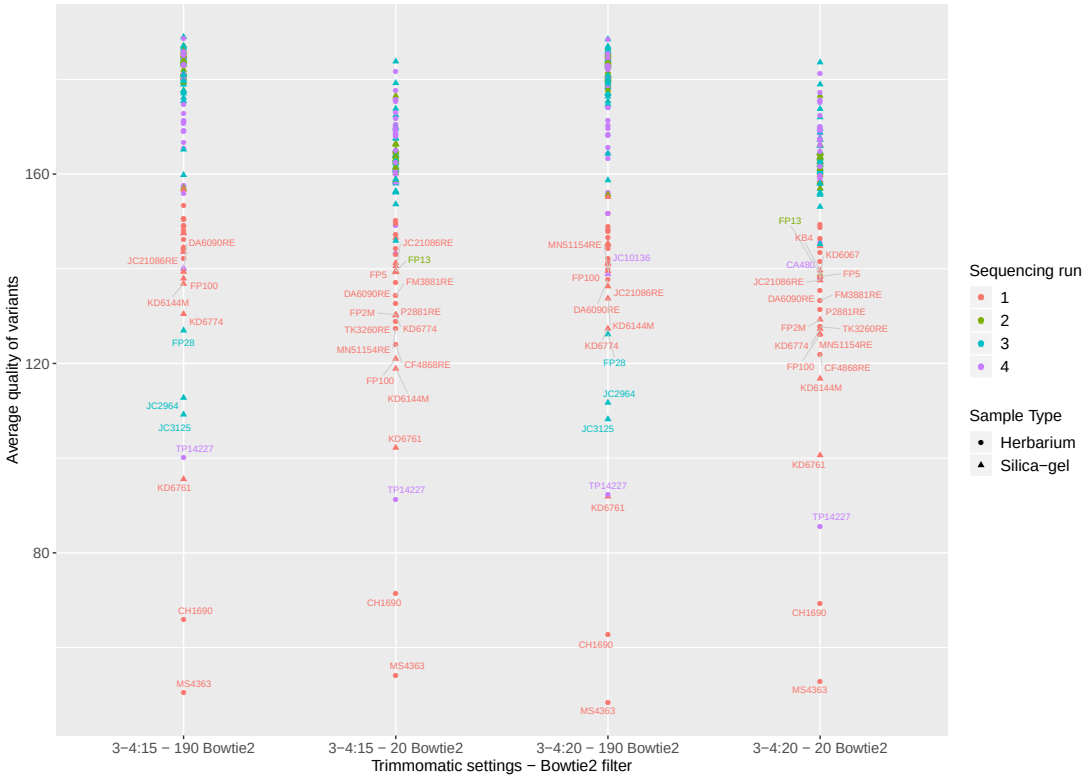
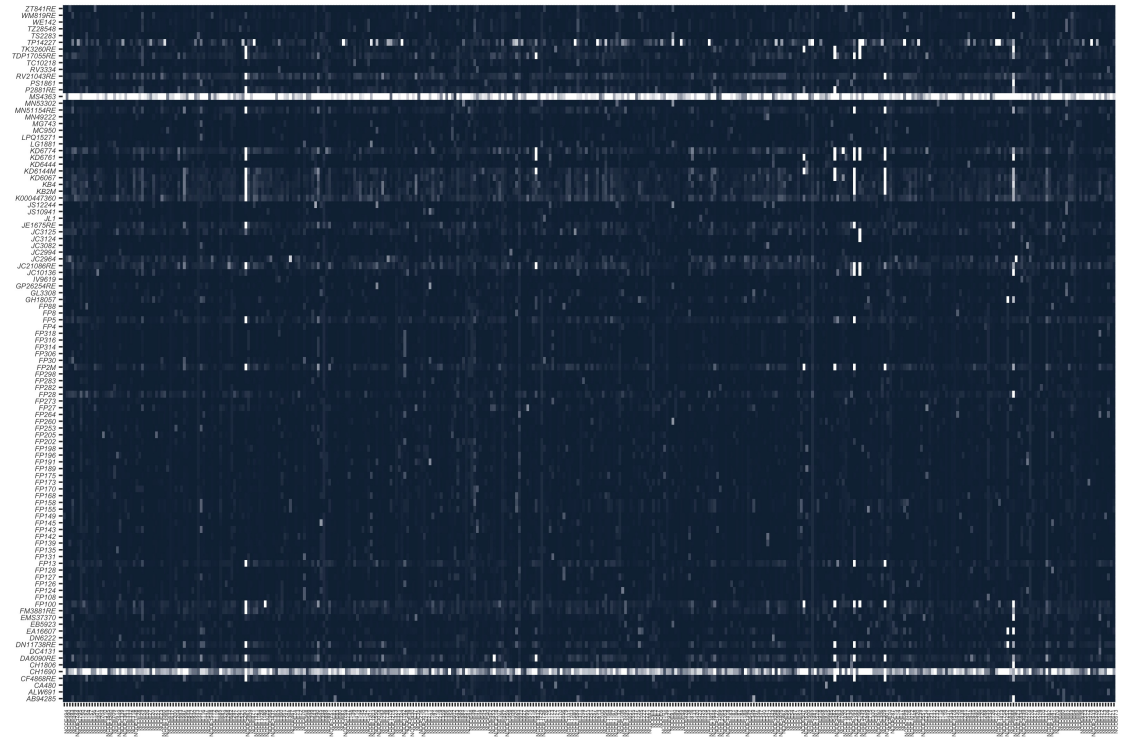
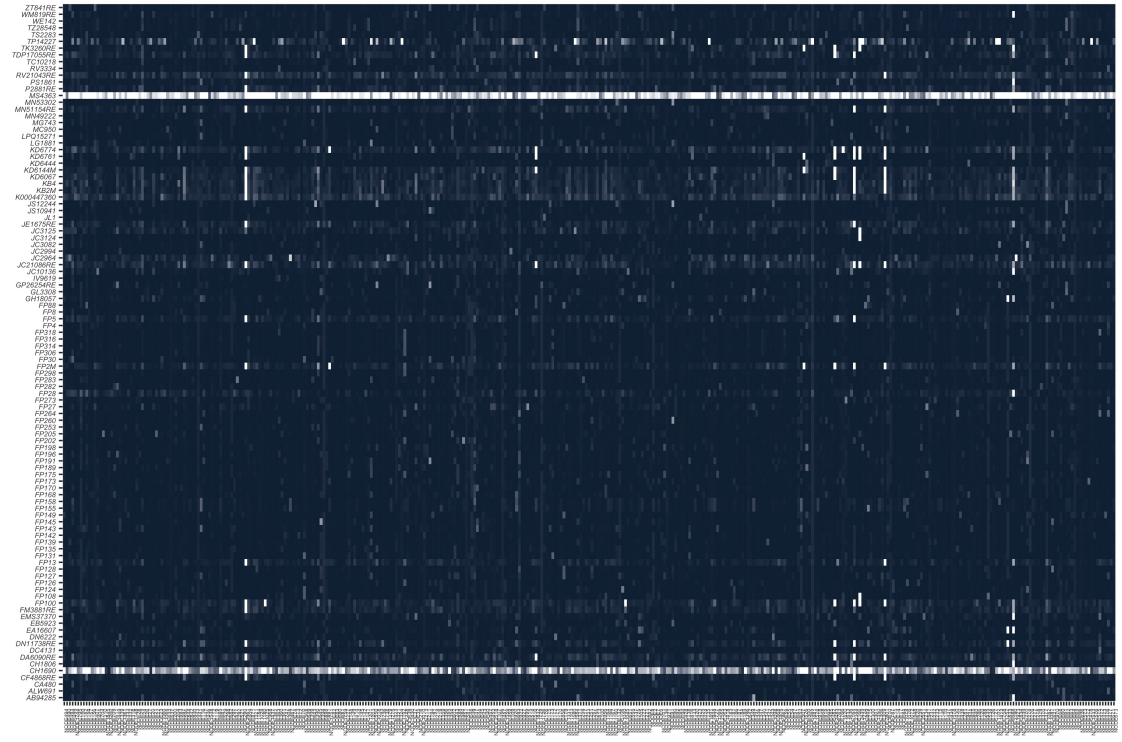


Figure 3.16: Average quality of variants with two Bowtie2 threshold alignment scores (20 and 190) for two input data from Trimmomatic filtering settings (3-4:15 and 3-4:20).



(a) 3-4:20 Trimmomatic filtering as input.



(b) 3-4:15 Trimmomatic filtering as input.

Figure 3.17: Comparison between Heatmap output for the Johnson et al. (2016) pipeline using two input files coming from the two different Trimmomatic filterings. Each one of the 377 loci is represented on the x-axis and each sample on the y-axis. The grey scale represents the percentage in the length of each locus covered by a consensus sequence for each sample.

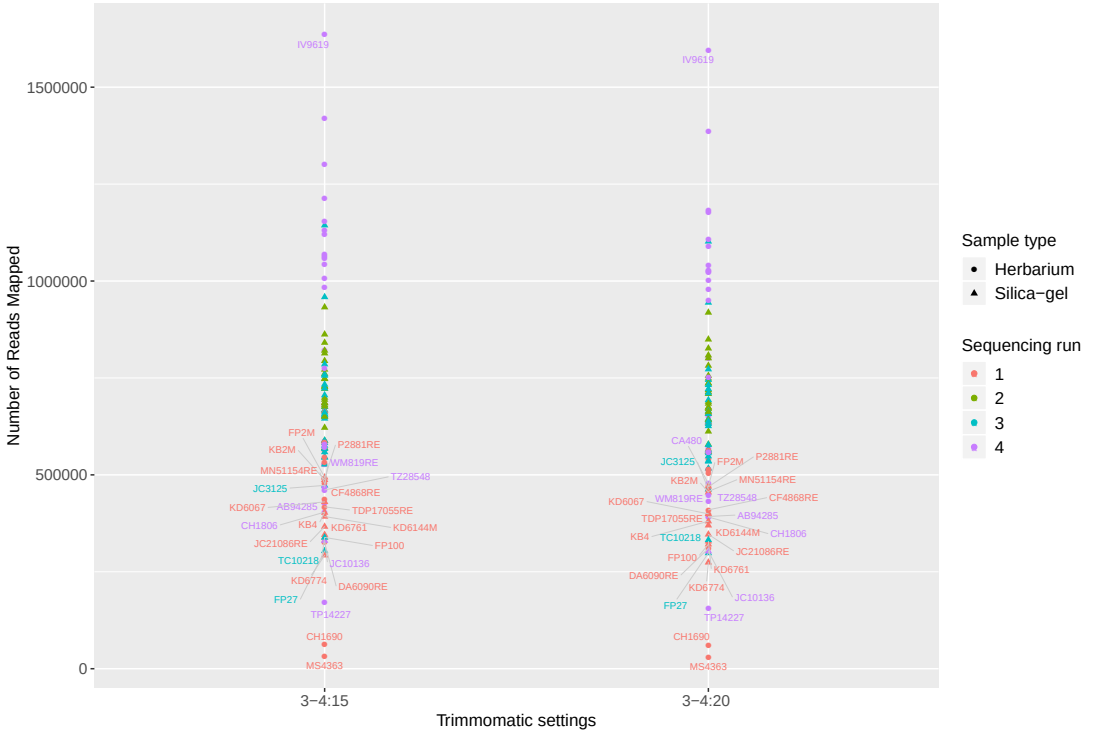


Figure 3.18: Number of reads mapped to the reference using the HybPiper pipeline with two different Trimmomatic filtering data sets as input (3-4:15 and 3-4:20).

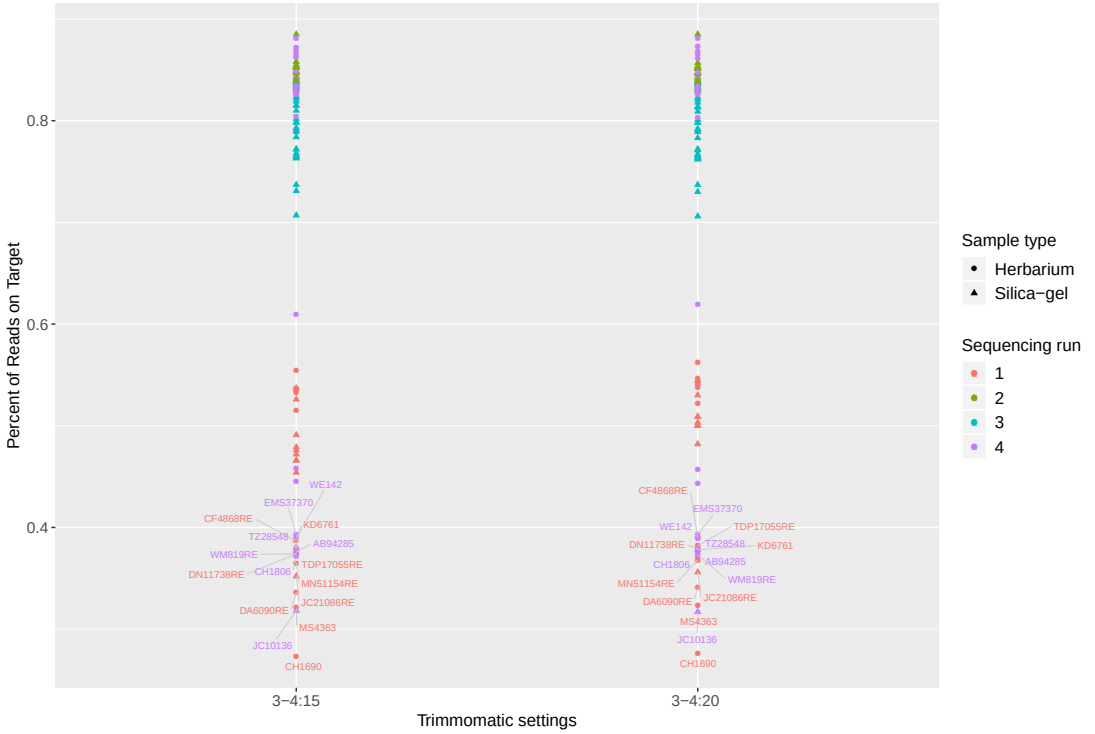


Figure 3.19: Percentage of reads on target using the HybPiper pipeline with two different Trimmomatic filtering data sets as input (3-4:15 and 3-4:20).

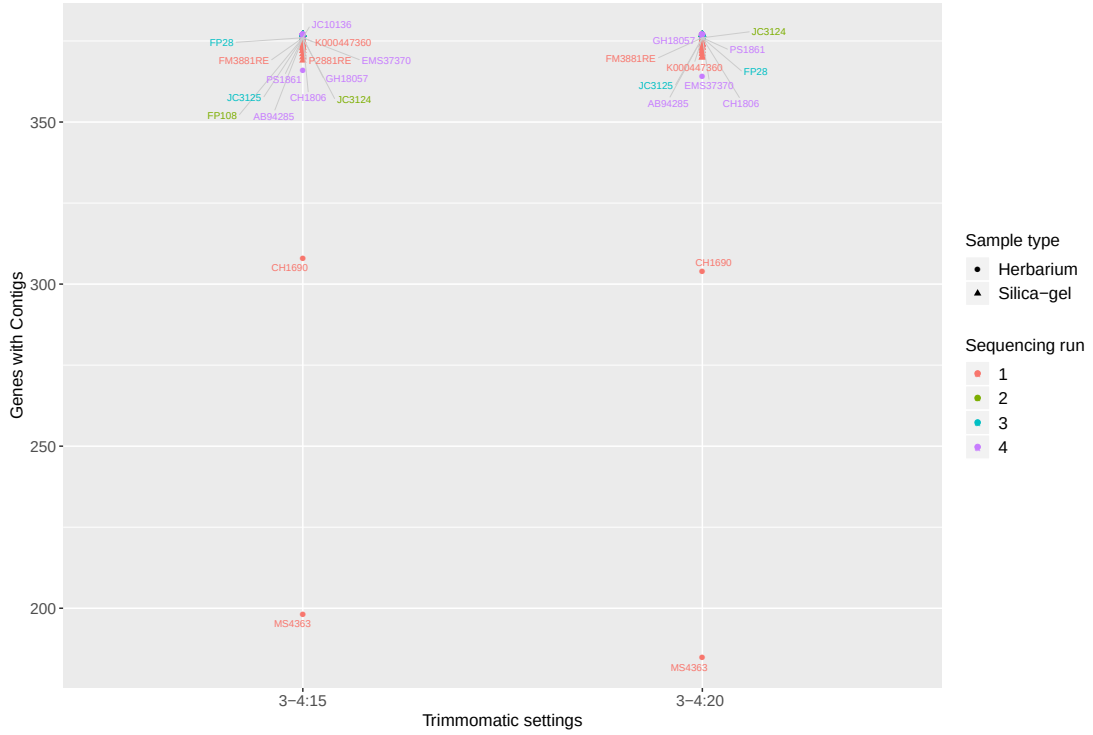


Figure 3.20: Number of genes with contigs for each accession using the HybPiper pipeline with two different Trimmomatic filtering data sets as input (3-4:15 and 3-4:20).

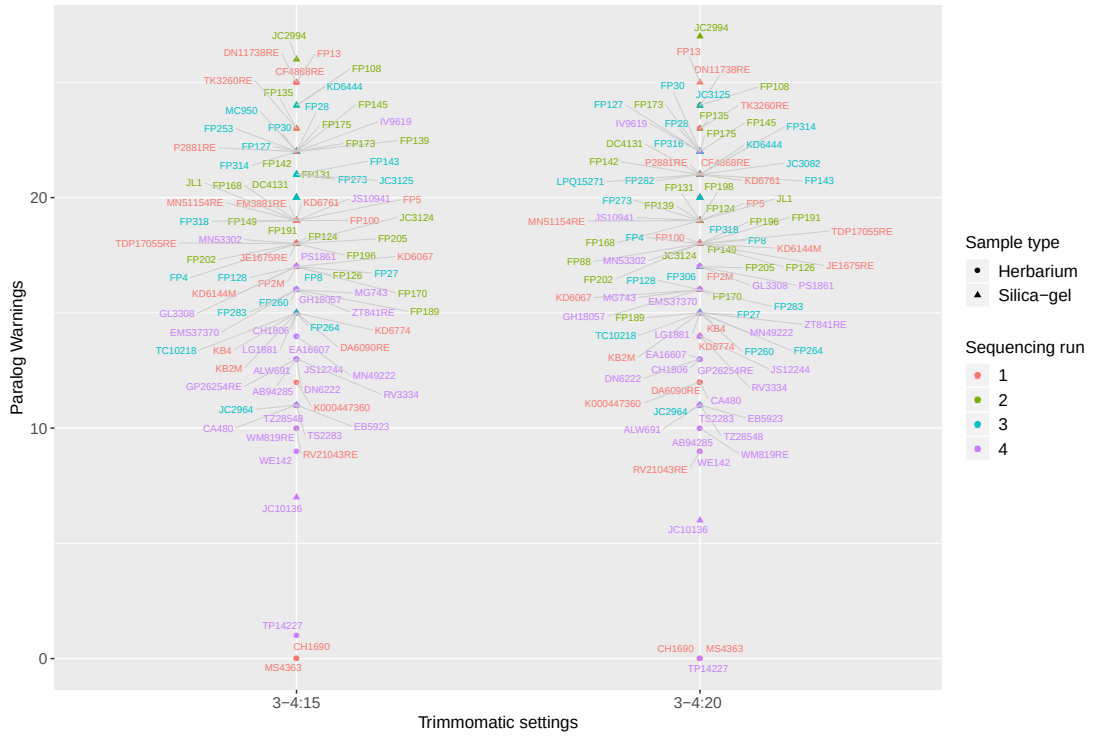


Figure 3.21: Number of paralogues warnings using the HybPiper pipeline with two different Trimmomatic filtering data sets as input (3-4:15 and 3-4:20).

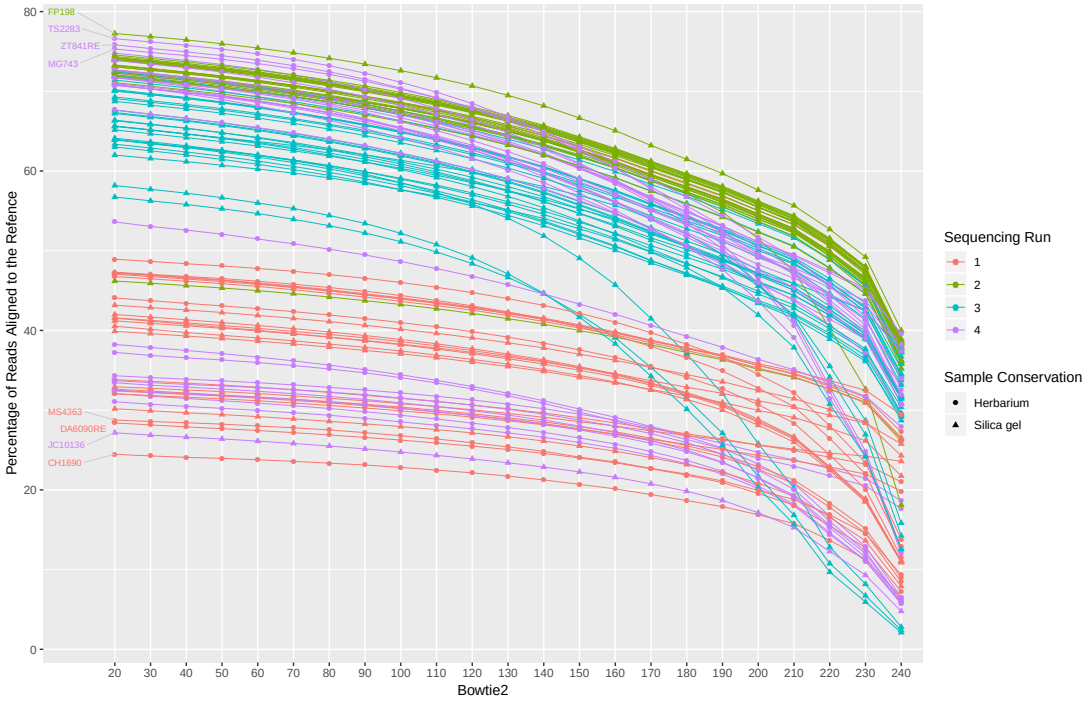


Figure 3.22: Percentage of reads mapped with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

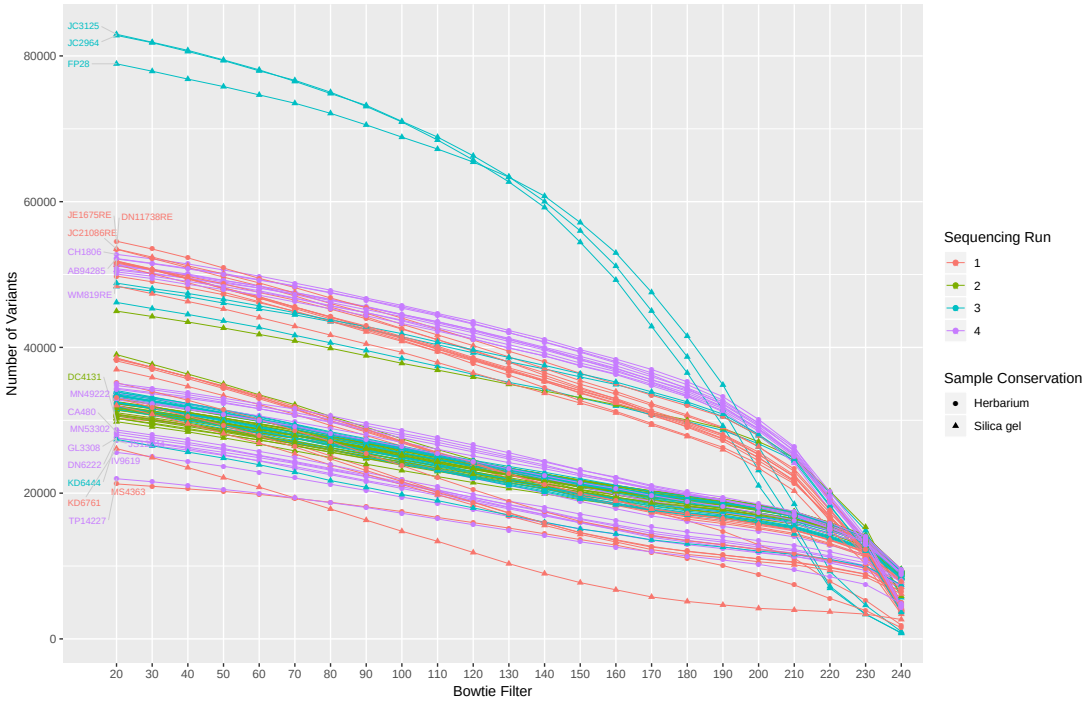


Figure 3.23: Number of variants with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

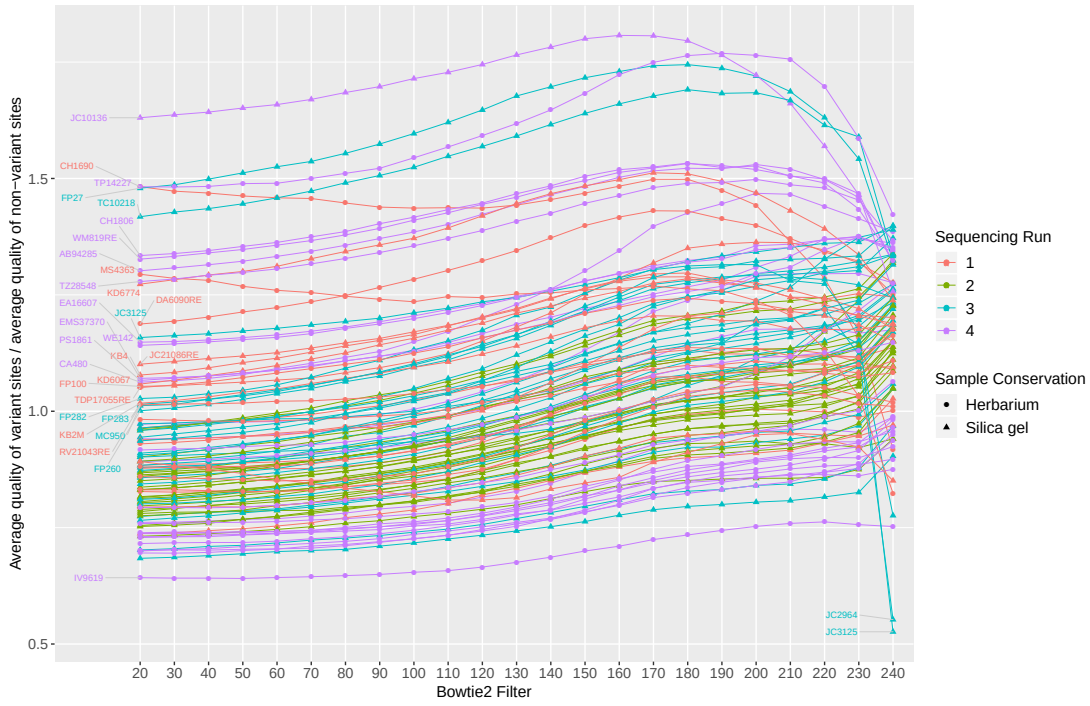


Figure 3.24: Standardized variants quality with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

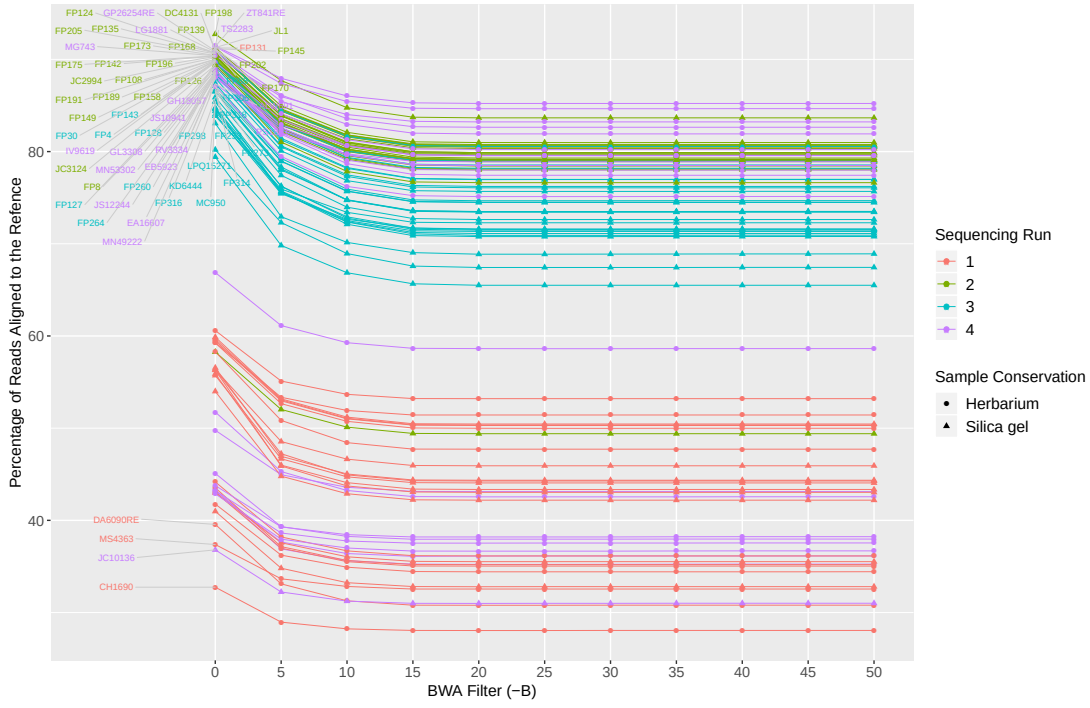


Figure 3.25: Percentage of Reads Mapped with varied BWA mismatch penalty threshold. Input data from Trimmomatic filtering setting 3-4:15.

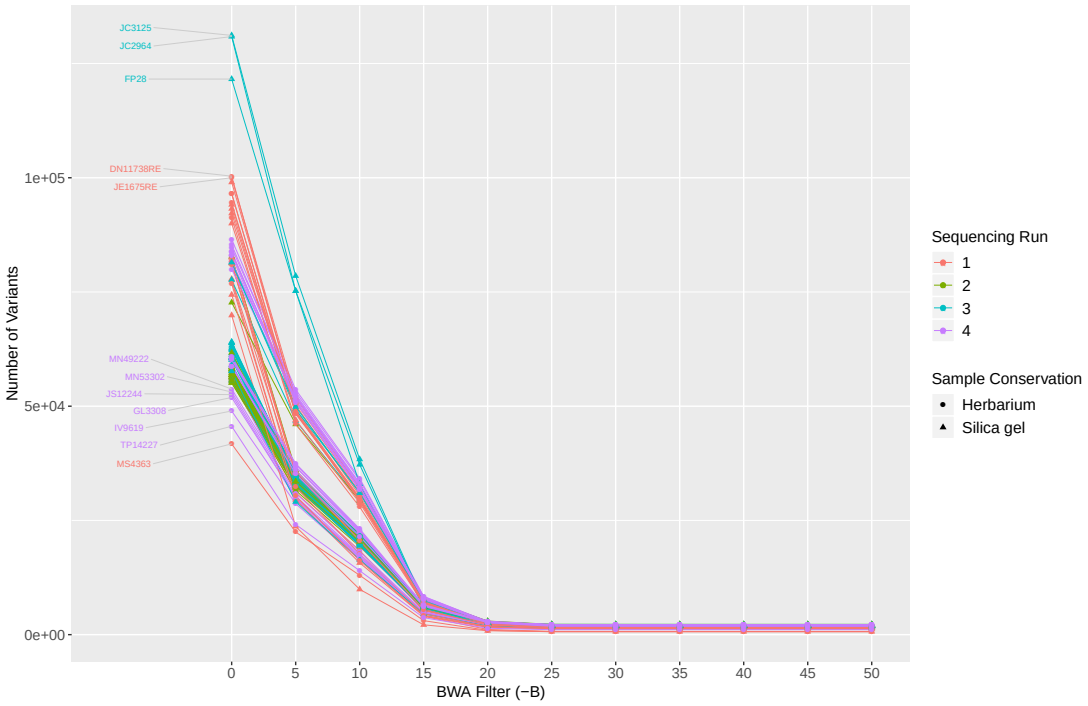


Figure 3.26: Number of variants with varied BWA mismatch penalty threshold. Input data from Trimmomatic filtering setting 3-4:15.

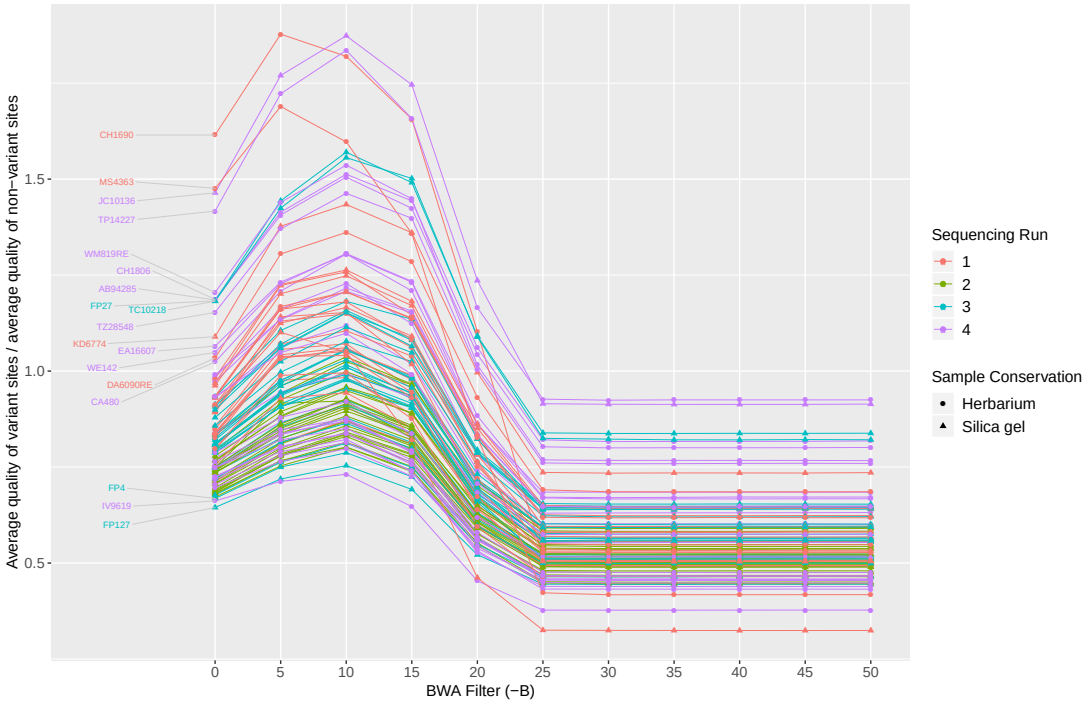


Figure 3.27: Standardized variants quality with varied BWA mismatch penalty thresholds. Input data from Trimmomatic filtering setting 3-4:15.

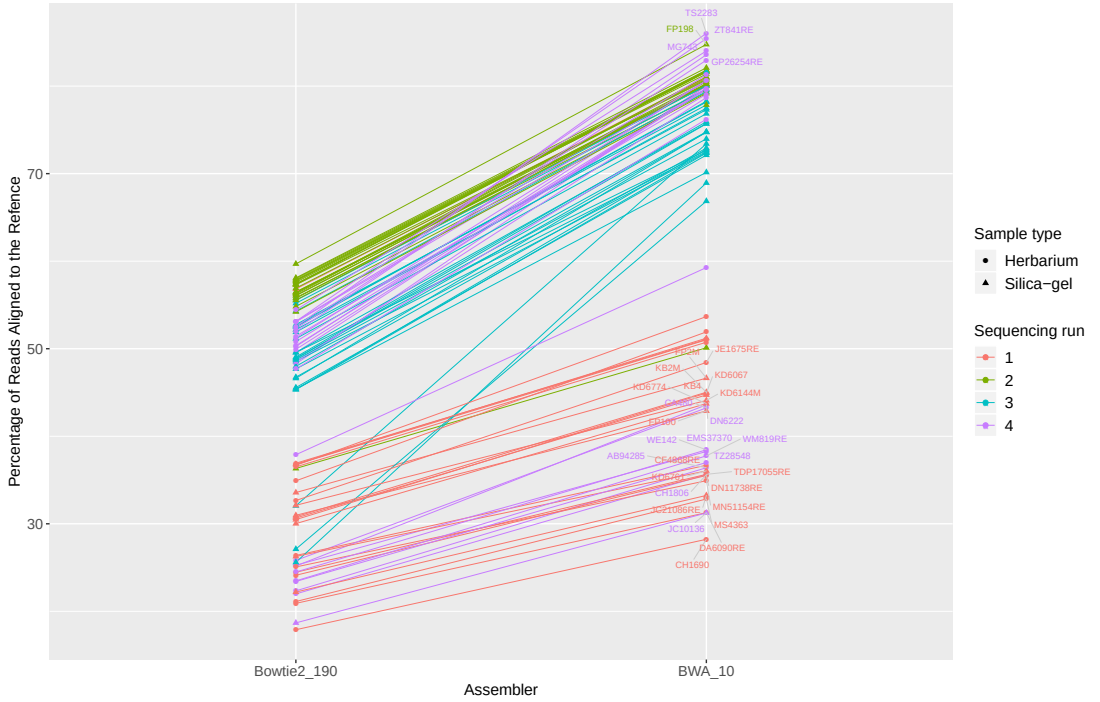


Figure 3.28: Percentage of reads mapped for two assemblers: BWA with mismatch penalty threshold of 10 and Bowtie2 with constant of alignment score of 190. Input data from Trimmomatic filtering setting 3-4:15.

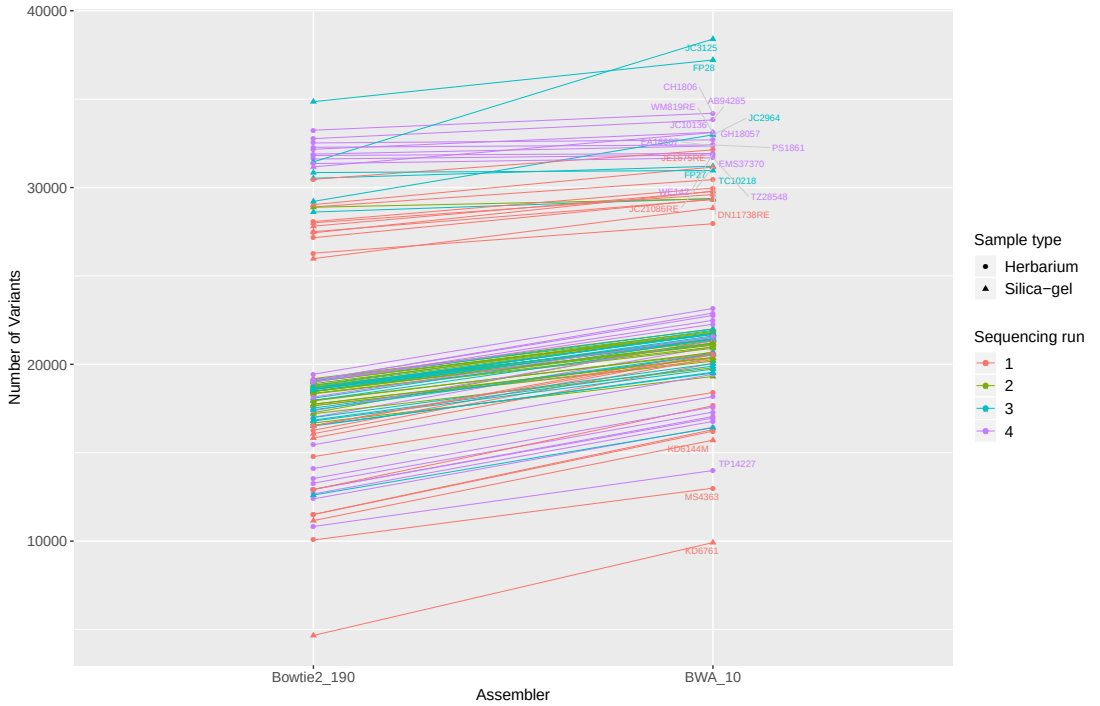


Figure 3.29: Number of variants for two assemblers: BWA with mismatch penalty threshold of 10 and Bowtie2 with constant of alignment score of 190. Input data from Trimmomatic filtering setting 3-4:15.

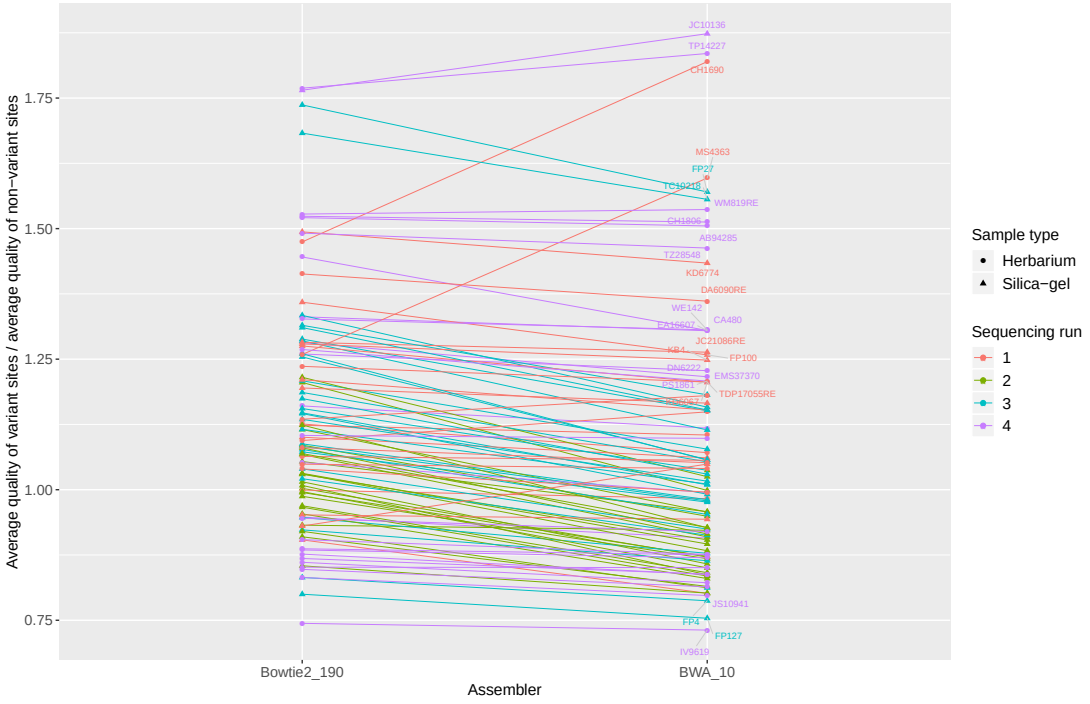


Figure 3.30: Standardized variants quality for two assemblers: BWA with mismatch penalty threshold of 10 and Bowtie2 with constant of alignment score of 190. Input data from Trimmomatic filtering setting 3-4:15.

Table 3.4: Comparison among the final aligments for all 377 loci for the three different pipelines and the two different data sets as input. A = Adenine, C= Cytosine, G = Guanine and T = Thymine

Pipeline / Alignment	**Alignment length**	**Undetermined characters**	**No variable sites**	**Parsimony informative sites**	**A**	**C**	**G**	**T**
Nicholls et al 3-4:20 Bowtie2 190	1065476	7101943	197788	115586	32293596	19126248	19125794	32096447
Nicholls et al 3-4:15 Bowtie2 190	842001	3396100	128092	79086	25571543	15211464	16227921	26319075
Nicholls et al 3-4:15 BWA 10	1072670	7019577	144205	89381	32520956	19307831	19293944	32342702
HybPiper 3-4:15 supercontig	1386782	59772387	254256	130591	26107296	15232314	15459744	26266805
HybPiper 3-4:20 supercontig	1321238	55449210	223775	114613	25421137	14832601	14966892	25417674

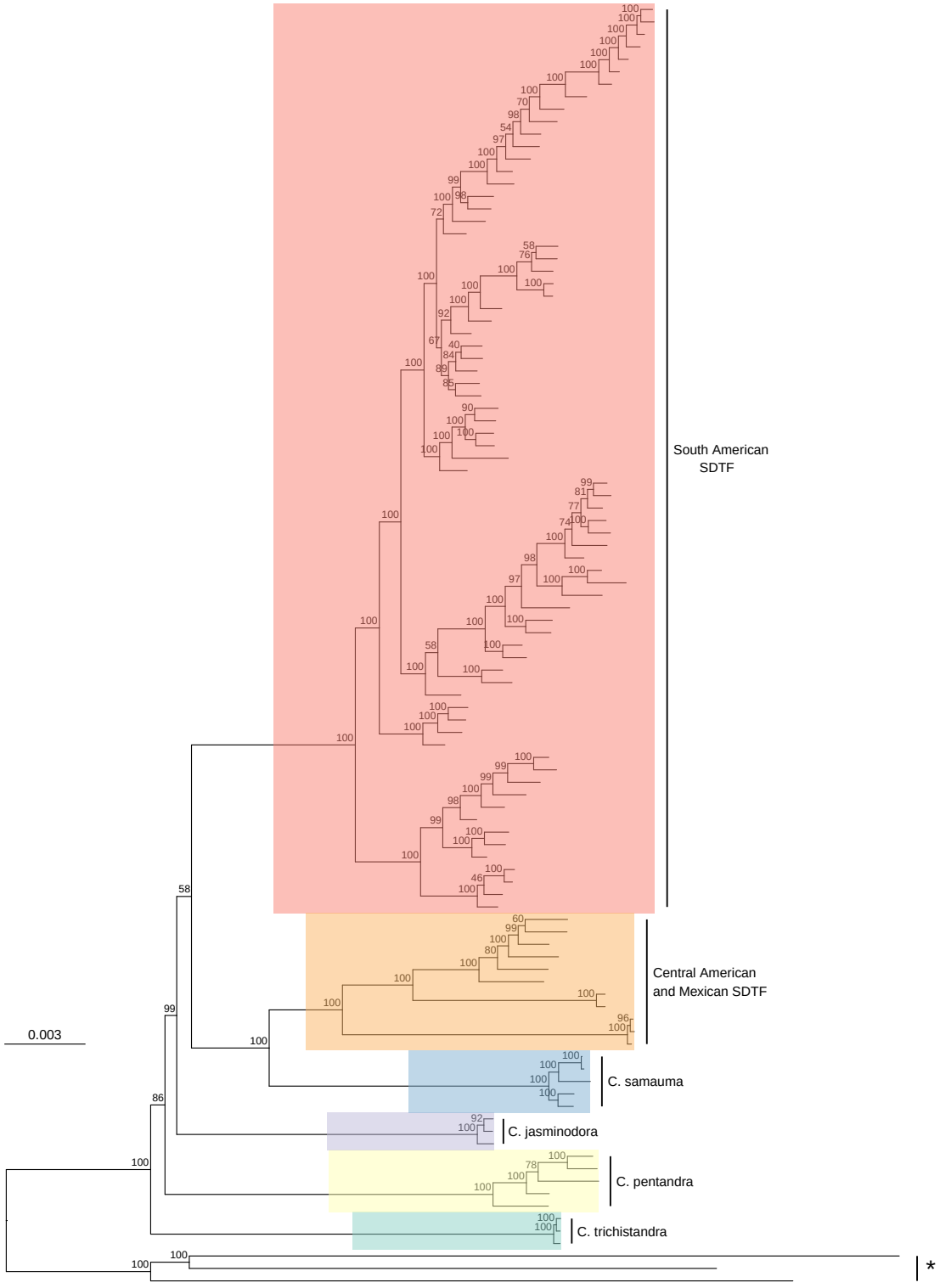


Figure 3.31: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:15 Trimmomatic using 190 Bowtie2 threshold. Numbers above branches represent bootstrap values and clades with an asterisk (*) represent the ougroups.

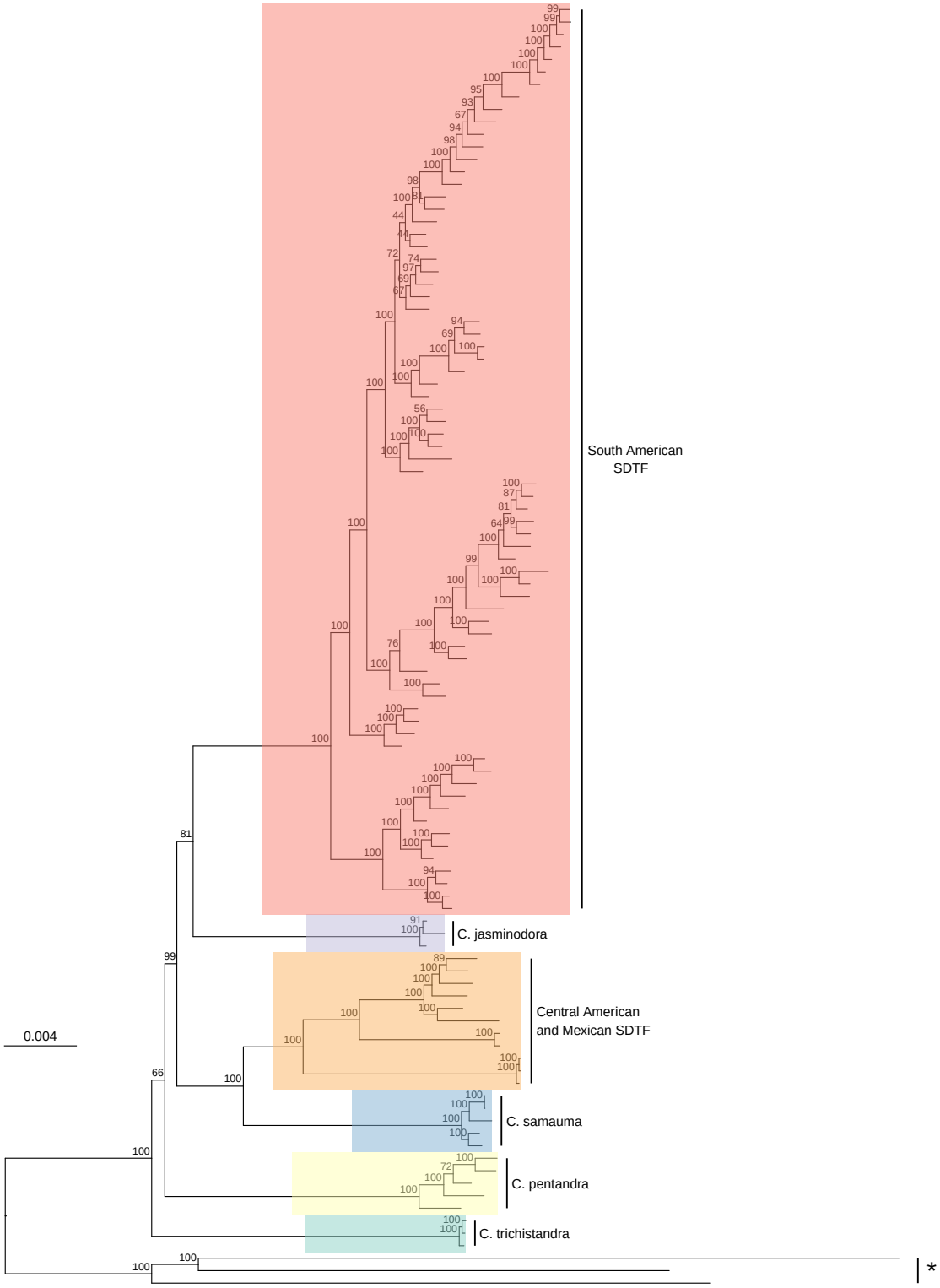


Figure 3.32: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:20 Trimmomatic using 190 Bowtie2 threshold. Numbers above branches represent bootstrap values and clades with an asterisk (*) represent the ougroups.

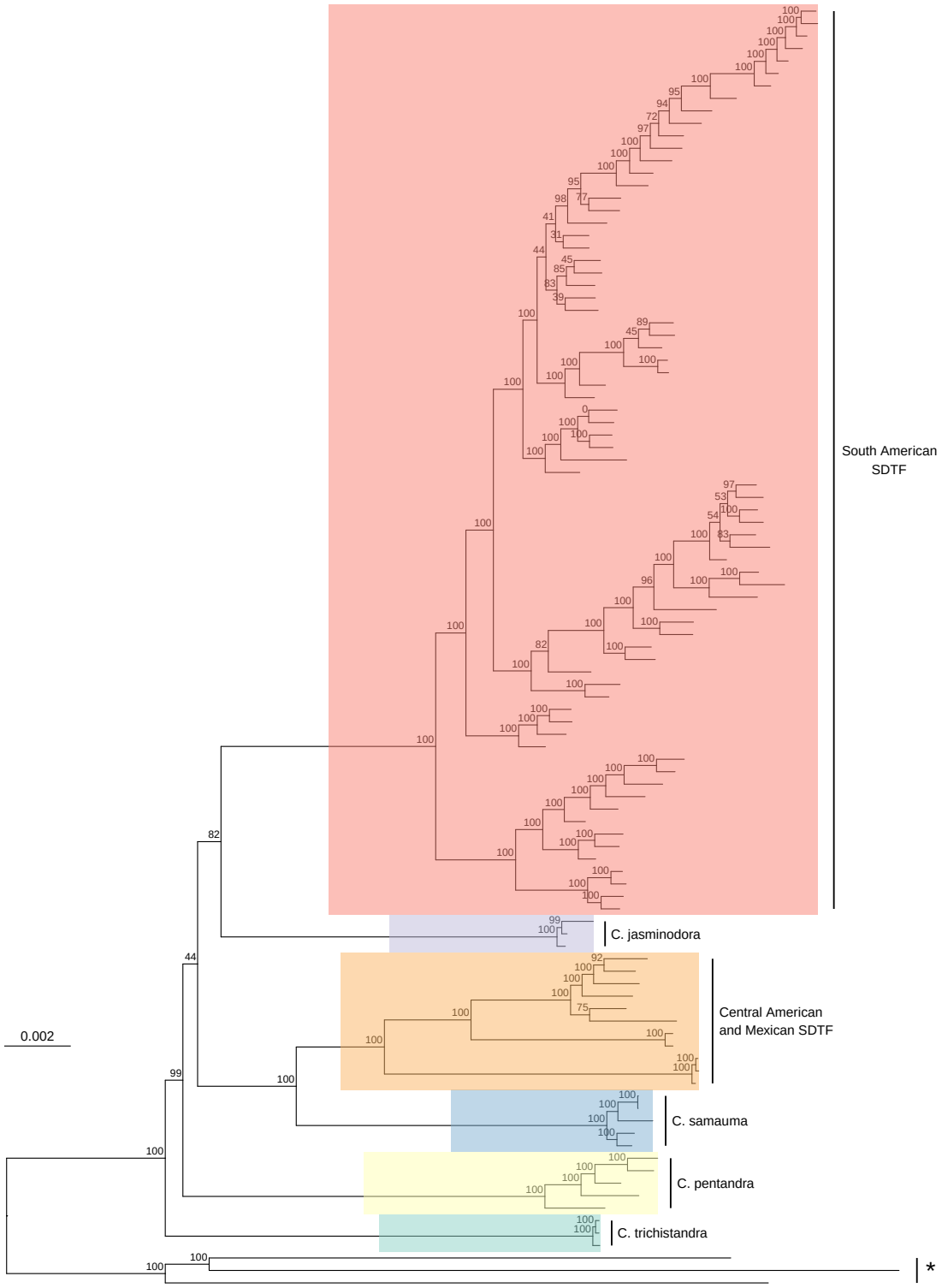


Figure 3.33: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:15 Trimmomatic using 10 BWA threshold. Numbers above branches represent bootstrap values and clades with an asterisk (*) represent the ougroups.

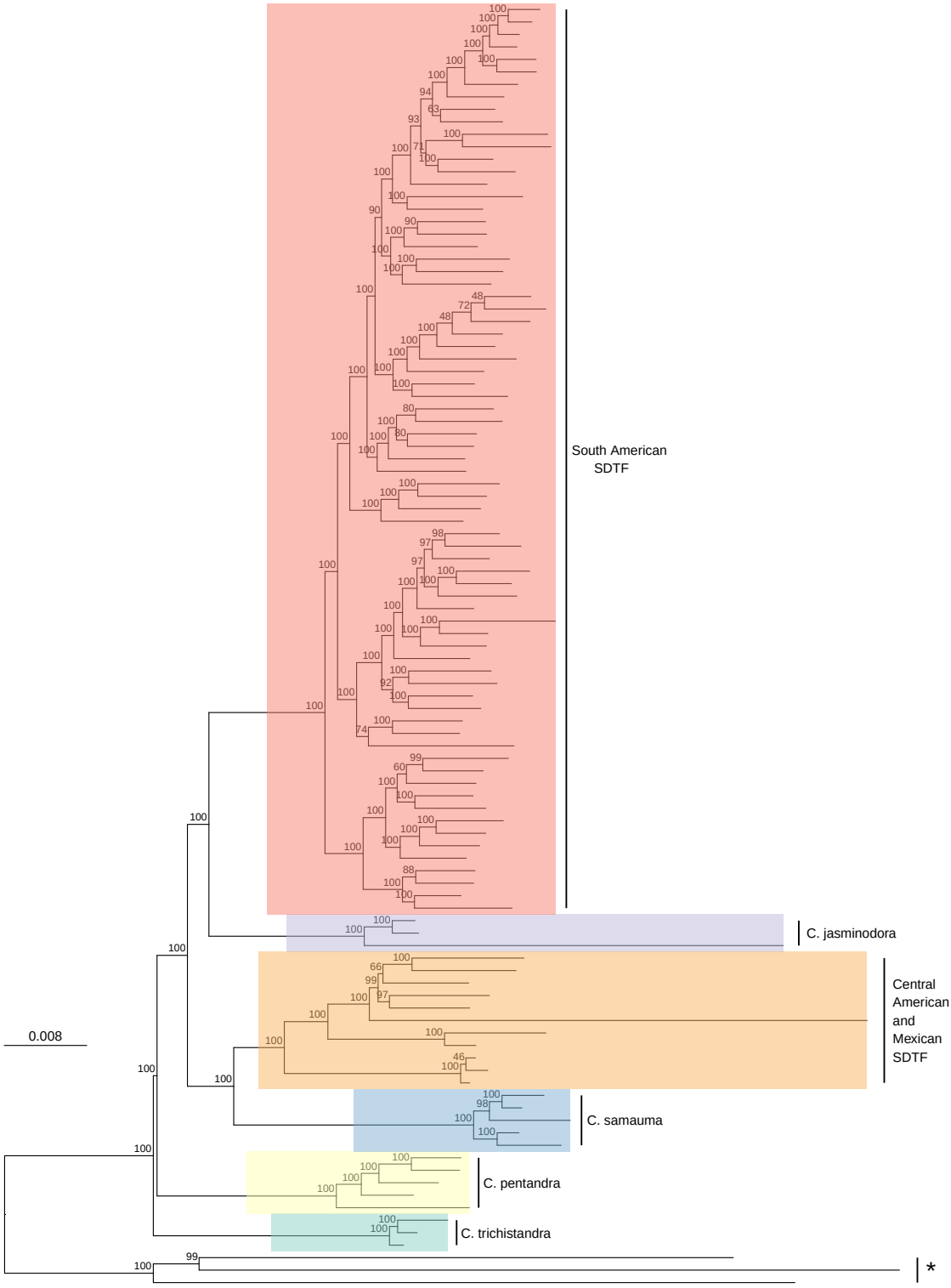


Figure 3.34: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:15 Trimmomatic using the HybPiper pipeline. Numbers above branches represent bootstrap values and clades with an asterisk (*) represent the ougroups.

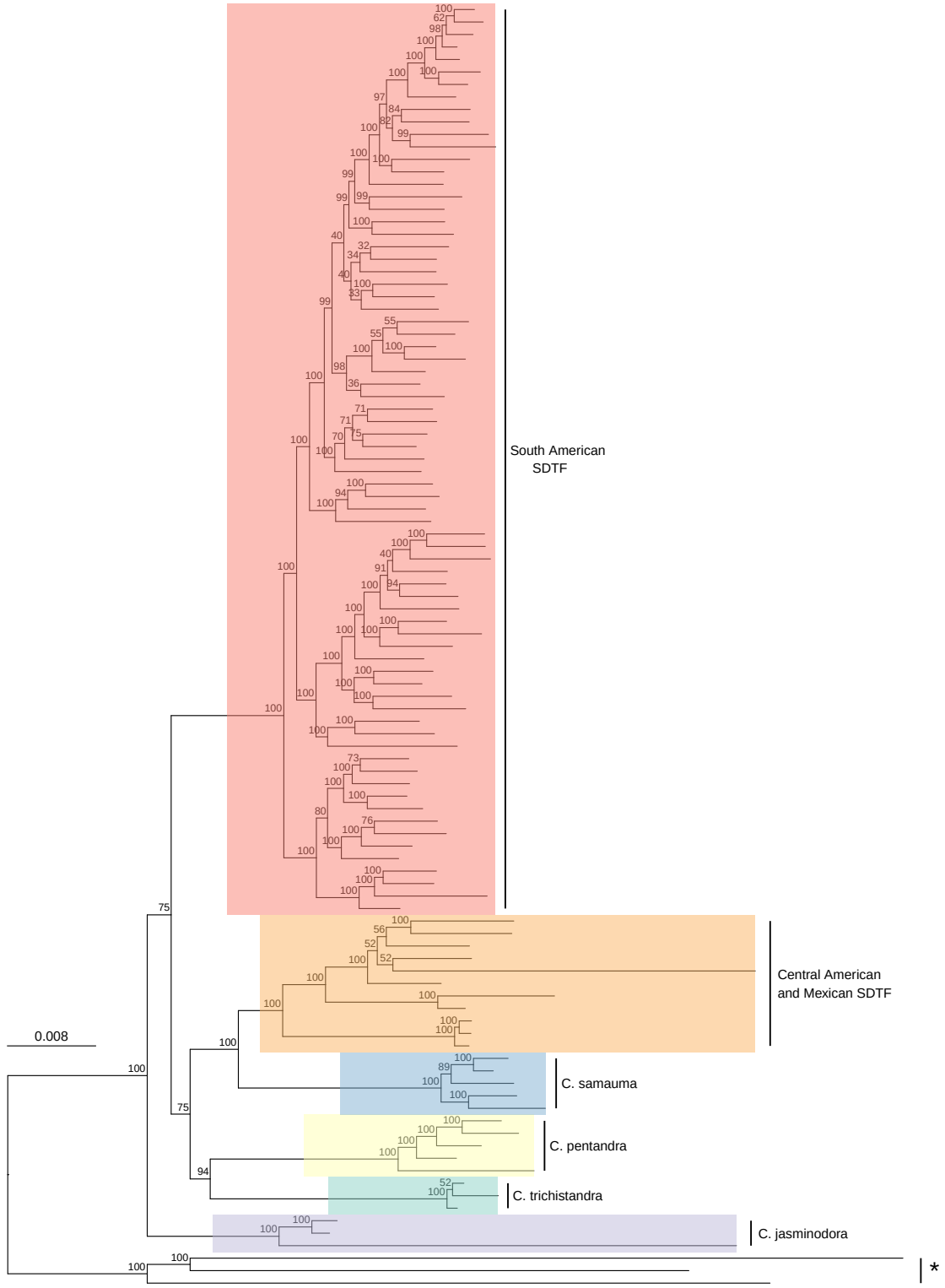


Figure 3.35: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:20 Trimmomatic using the HybPiper pipeline. Numbers above branches represent bootstrap values and clades with an asterisk (*) represent the ougroups.

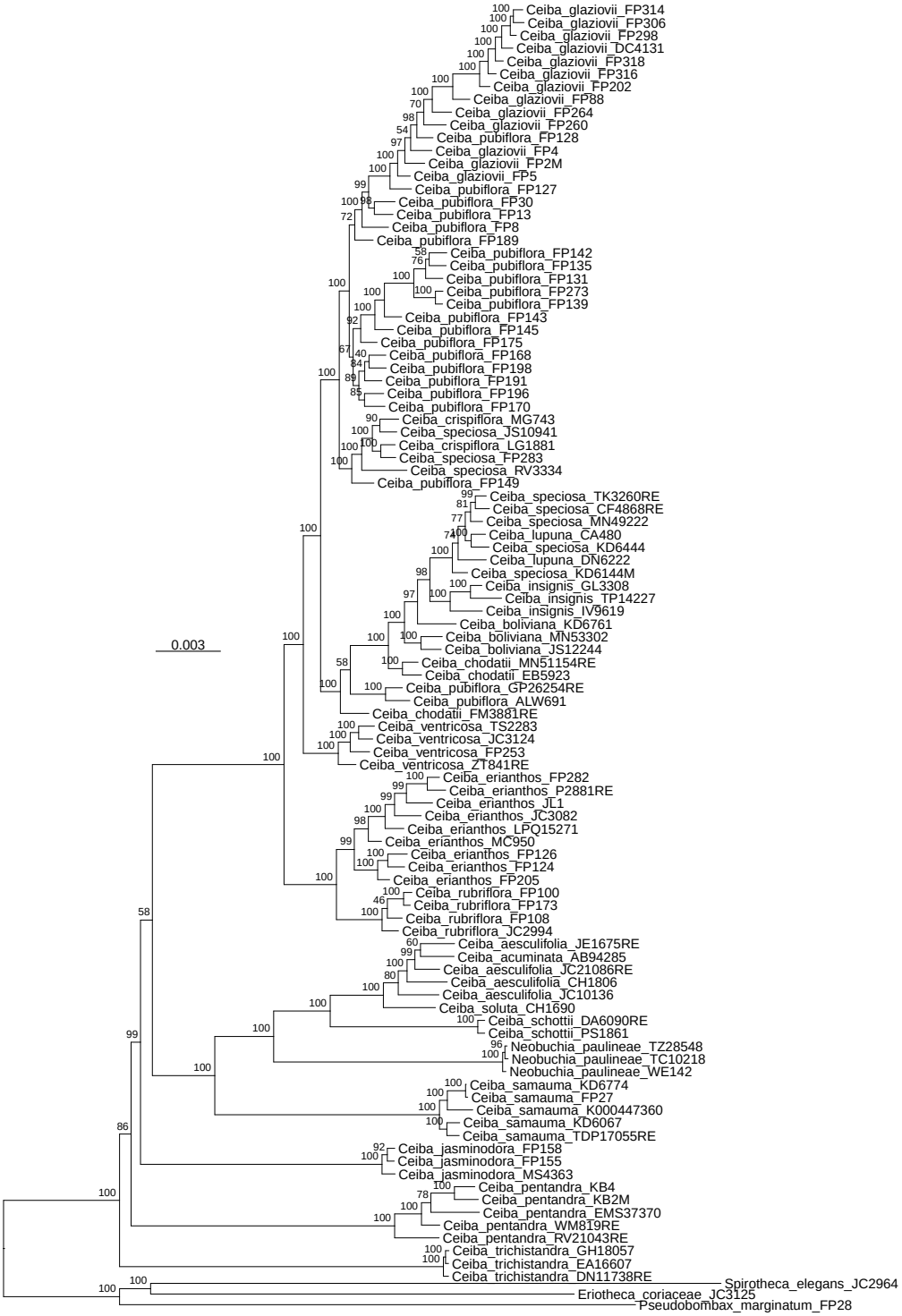


Figure 3.36: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:15 Trimmomatic using 190 Bowtie2 threshold. Numbers above branches represent bootstrap values.

Within the South American SDTF clade, eight of the 11 species were not resolved as monophyletic. For all pipelines and inputs the species were resolved as follows: (i) *C. ventricosa* and (ii) *C. boliviana* both recovered as monophyletic; (iii) *C. rubriflora* and (iv) *C. erianthos*, sister to each other and recovered as monophyletic; (v) a clade including *C. pubiflora* and *C. glaziovii* sister to; (vi) a clade comprising *C. speciosa*, *C. crispiflora* and one accession of *C. pubiflora*; (vii) a single individual of *C. chodatii* from Paraguay sister to (viii) two individuals of *C. pubiflora* from Paraguay and west Brazil; and (ix) a clade containing accessions of *C. insignis*, *C. lupuna*, *C.speciosa* and two individuals of *C. chodatii*, all from west South America.

3.3.5 Baits Design

Among the 24 accessions tested, the accession KD6761 had the highest percentage of top blast hits (i.e, highest percentage of loci with high sequence similarity) when blasted against the *Adansonia - Bombax* bait set indicating that the *de novo* contigs from that accession, if used as the new bait set, would more likely be compatible with the data set generated by hybridisation with the *Adansonia - Bombax* bait set (Figure 3.37). This accession also had the longest sequence recovered when comparing to candidate bait sets from other accessions (Figure 3.38), indicating that more flanking regions were recovered.

When using the *Adansonia - Bombax* bait set for hybridisation and as a reference, the average percentage of reads recovered was 42.1% for the 24 accessions sequenced in the first run. When using the new *Ceiba* bait set for hybridisation and as a reference the average percentage of reads recovered was 48.3% for the 79 accessions sequenced in runs 2, 3 and 4 (Figure 3.39).

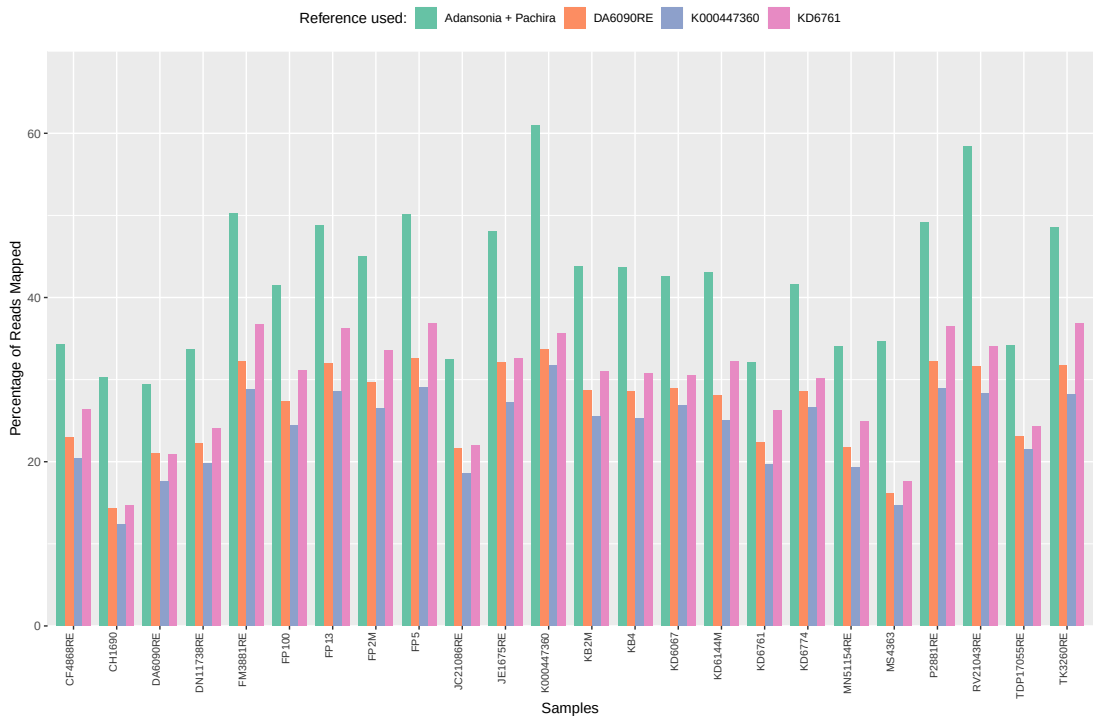


Figure 3.37: Percentage of reads mapped back to different references using Bowtie2 and 190 threshold value: *Adansonia* - *Bombax* bait set, and candidate bait sets based on the accessions DA6090RE, K000447360 and KD6761.

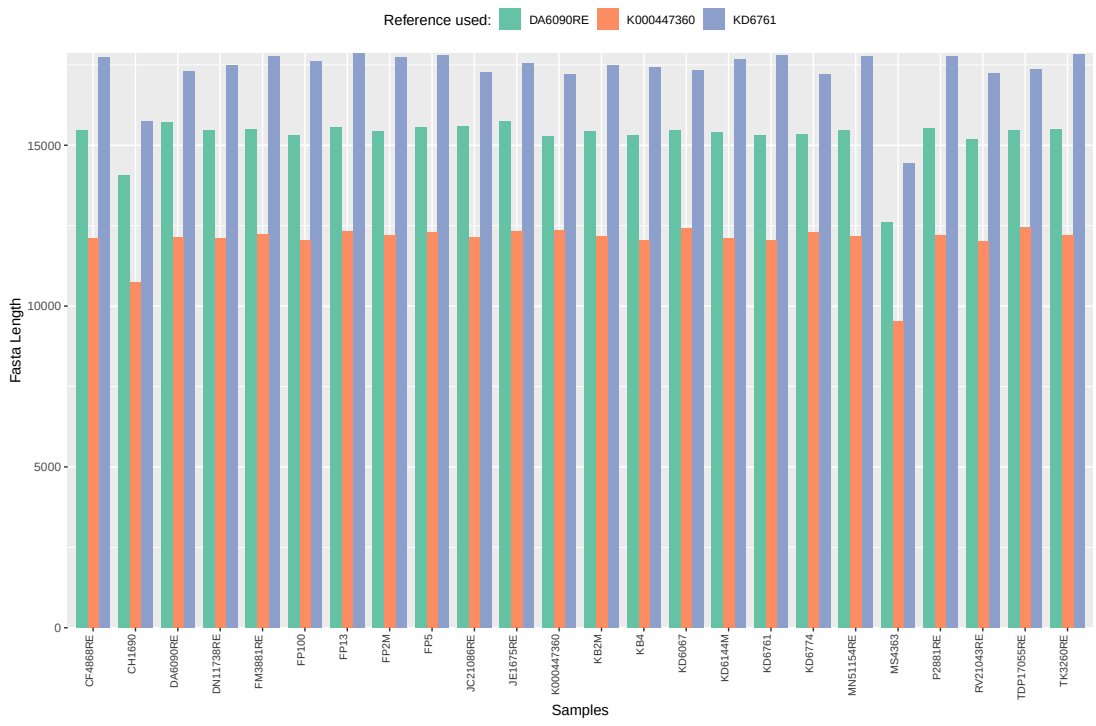


Figure 3.38: Length of final fasta recovered for the candidate bait set using three different accessions: DA6090RE, K000447360 and KD6761.

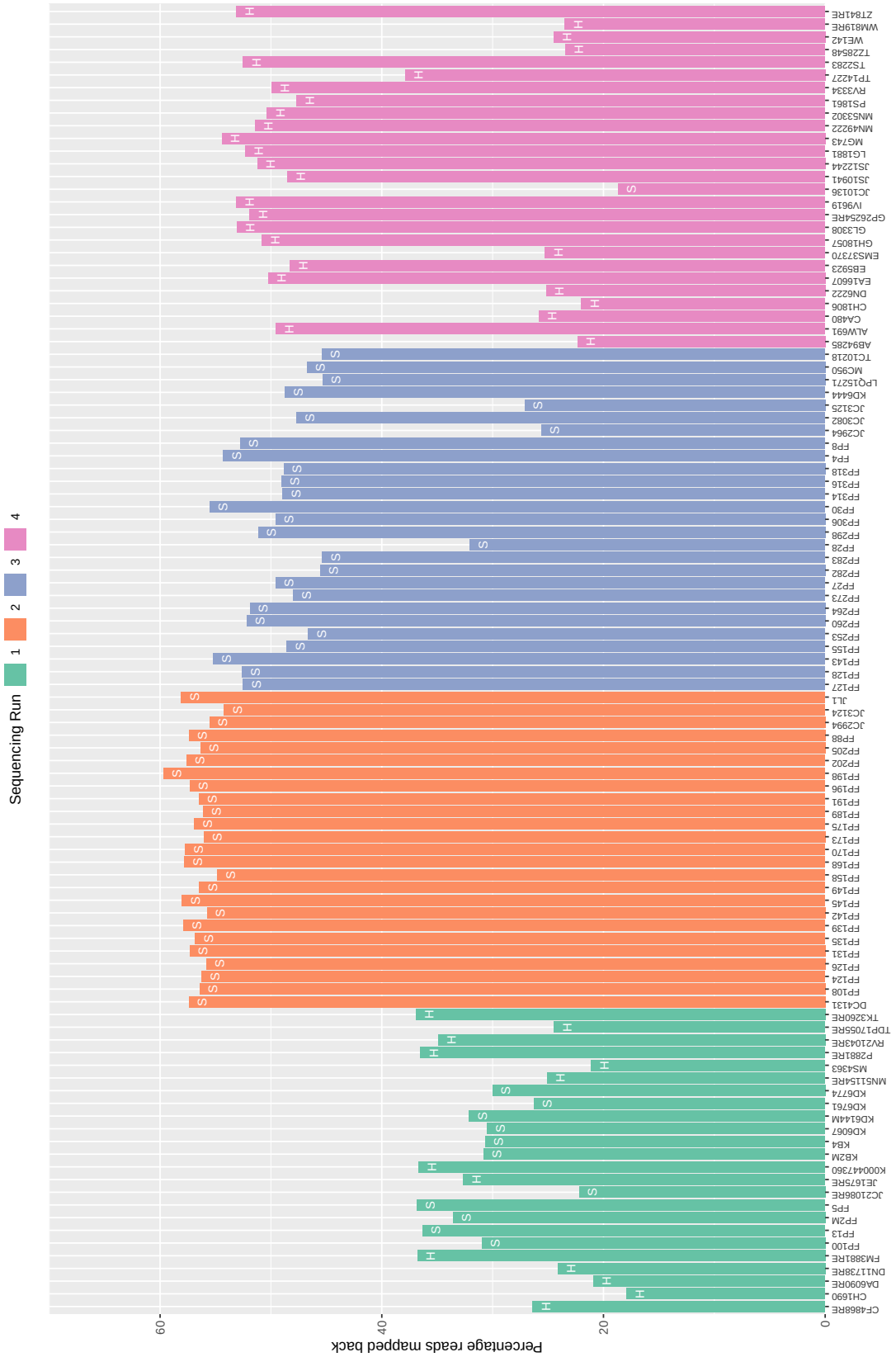


Figure 3.39: Percentage of reads mapped back to the reference using Bowtie2 - 190 for each sample. Colours represent the sequencing run (1, 2, 3 or 4) and letters on bars represent sample type (H) herbarium and (S) silica gel.

3.4 Discussion

The impact of data quality filtering and genome assembly pipelines in downstream analysis

The target enrichment sequencing technique applied to the tropical, non-model lineage *Ceiba* provided 377 independent nuclear loci with enough phylogenetically informative characters to fully resolve the phylogeny of the genus for all 18 species represented by multiple individuals per species. The technique generated data for old herbarium samples (the oldest one being from 1834) and newly collected silica-gel dried leaves from as little as 2 ng/μL of initial genomic DNA input.

The four sequencing runs of 103 accessions produced around 10 GB of raw data and a final alignment of more than 1 million base pairs. This data set was analysed using pipelines following different approaches (*de novo* and reference mapping), with variations in the software used and in their settings. All of the pipelines and settings produced results with desirable data quantity and quality. However, the final output of each was different, which led to different phylogenetic inferences. The conflicting results reinforce the fact that although next generation sequencing techniques represent an important step forward in phylogenetics, more care should be taken when analysing the data with bioinformatics pipelines and interpreting the output.

The initial trimming step of all pipelines filters the raw reads based on quality of base call, and influenced the final phylogeny. Del Fabbro et al. (2013), for example, tested nine different trimming softwares and found that they produce distinct results depending on the settings used. The authors could not answer which trimmer performed best and suggested that this is dependent on the data set characteristics and the type of downstream analysis. They suggested that the choice should then be made by the user based on the trade-off between read loss and desired data quality (Del Fabbro et al., 2013).

The loss of data in the more stringent Trimmomatic filtering (3-4:20) results in a different phylogenetic tree. The main difference between the filters lies in the fact that the 3-4:20 Trimmomatic filter resulted in a higher percentage of reverse unpaired reads

when compared to the more permissive filter, 3-4:15, especially for samples from the fourth run (Figure 3.9). Hence, in the 3-4:15 setting both mates from the paired-end reads passed the filter whereas only the reverse read, i.e, one of the mates, passed the quality filter in the 3-4:20. The information stored in the paired reads is useful to improve the mapping step because if one of the reads cannot be mapped with confidence, its mate often can provide enough information for high confidence mapping (Pfeifer, 2017). Most pipelines use as input the forward paired, reverse paired and forward unpaired reads and do not include the reverse unpaired reads. The different phylogenies resulting from the different pipelines could then be explained by the increase in the reverse unpaired reads in the 3-4:20 filter, which were not used as input in HybPiper and Yang and Smith (2014) pipelines. However, I ran the Nicholls et al. (2015) pipeline with (Figure 3.31) and without (Figure 3.61) the reverse unpaired reads included in the input files and the topology inferred is identical, although with different bootstrap values for some nodes. Because I used a conservative mapping threshold (190) for the Nicholls et al. (2015) pipeline, it is likely that Bowtie2 did not consider the unpaired reads a good enough match to the reference and therefore they were not incorporated in the final contigs. HybPiper, for example, uses the default parameters of its assembler (BWA) and those reads might have been incorporated to the final contigs.

The two different assemblers tested, Bowtie2 and BWA, recovered equally good quality data, but BWA recovered a higher percentage of reads. Studies assessing assemblers often use metrics related to sensitivity and speed and consider a higher percentage of reads mapped and computational efficiency as indicators of better performance (Hatem et al., 2011) in simulated and real data sets. When using simulated data sets, the percentage of correctly versus incorrectly reads can be evaluated (Thankaswamy-Kosalai et al., 2017), but this is not trivial when using real sequencing data. Furthermore, most comparisons of NGS software are between different software packages and not different settings within one package (Hatem et al., 2011). Similarly to the trimming step, there is little consensus on which assembler performs better and which settings are more adequate. In fact, Pfeifer (2017) stated that "determining

the ideal parameter settings is often nontrivial, requiring an in-depth understanding of both the data and the alignment algorithm". In my case, a higher percentage of reads does not imply better data because I aim to filter out possible paralogous copies in the mapping step, and having more reads mapped might increase the chance of mapping different copies.

Does a taxon-specific bait set influence sequence capture success?

The new *Ceiba* bait set was designed from the *de novo* contigs assembled from sequencing data of one of the accessions of *Ceiba*. The taxon-specific *Ceiba* bait set improved capture efficiency, recovering a high percentage of raw reads even from herbarium tissue (Figure 3.6), and an average of 30.6% reads mapped back when using a stringent mapping threshold for genome assembly (Figure 3.39). The hybridisation allows for mismatches and captures off-target sequences, therefore this bait set contains intronic regions. Those intronic regions are likely to provide more phylogenetically informative characters and improve the resolution of a phylogenetic tree at population and species level.

Normally, baits are designed from transcriptomes of one or more species. All the exons are then represented consecutively in the bait sequence even when multiple exons are present in one gene in the genomic DNA. However, because in the genomic DNA the exons are physically intercalated with intronic regions, the consecutive representation of exons in the bait sequence (both actual bait fragments and bait sequence in the computer) might interfere with the hybridisation and sequence assembly (Fér and Schmickl, 2018). Pipelines like Fér and Schmickl (2018) were designed to accommodate intronic regions in the reference bait sequence during assembly. Here, I opted to redesign the baits using sequence data to include the intronic regions in the fragments of baits for hybridisation rather than handling multiple exons in one gene in the analysis step. With this approach, if paralogous loci are present in *Ceiba*, the intronic regions in the bait set would favour the capture of one of the copies, i.e., the copy with higher sequence similarity to that of the intron bait fragment (Fér and Schmickl, 2018),

thus minimising the need to handle paralogy issues during data analysis. Dealing with paralogous loci is one of the bottlenecks of bioinformatics pipelines. For example, differentiating paralogues from alleles is not straightforward (Johnson et al., 2016) as the topology recovered for the different copies depends on the age of the duplication event and the impact of including them in the phylogenetic inference is difficult to access (McKain et al., 2018). Likewise, the intronic bait set used as a reference would recover a higher percentage of reads even when using the conservative mapping approach (eg. Nicholls et al. 2015) because the intronic regions are represented in the reference.

The higher number of raw reads and the number of loci recovered using a specific bait set impact positively the phylogenetic inference using the concatenation approach with thousands of phylogenetic informative characters spread across the matrix, especially for closely related species or populations. Additionally, the presence of the intronic regions in the individual loci might provide more phylogenetic informative characters for the inference of the individual gene trees. Those gene trees can be used as input for modern analysis such as the incongruence among individual gene trees and between gene tree and the species tree to investigate incomplete lineage sorting or species delimitation under the coalescent model (Maddison 1997; Edwards 2009; Fujita et al. 2012; Chapter 4).

What are the relationship amongst the species of *Ceiba*?

The phylogenetic trees generated for all data set inputs and pipelines were fully resolved with high support values. All the 377 loci were concatenated and the final alignment was analysed under the maximum likelihood framework treated as a single partition. The alignments were at least 842,001 bp long with 79,086 parsimony informative characters (Table 3.4). The target enrichment technique is a successfull approach to resolve recent radiation of tropical lineages such as *Ceiba*, unlike the Sanger sequence approach (Chapter 2).

Despite the topological differences observed among the phylogenies from the different pipelines for the deeper nodes, the shallower branches of the trees had the same

phylogenetic arrangement. *Ceiba jasminodora*, *C. trischistandra*, *C. samauma*, *C. pentandra*, *C. shottii*, and *Neobuchia paulinae* were recovered as monophyletic.

The phylogenies inferred based on Sanger-sequencing of the ITS region (Chapter 2) and next generation hybrid capture sequecing for 377 nuclear loci were highly similar. Both phylogenies had four main clades in common: (i) *Ceiba pentandra*; (ii) a Central American and Mexican SDTF clade; sister to (iii) *Ceiba samauma*, and (iv) a South American SDTF clade. *Ceiba jasminodora* and *C. trischistandra* were not sampled in the Sanger-sequencing phylogeny. Similar to the ITS phylogeny, the NGS phylogeny supports the monophyly of the two sections of the genus, Ceiba and Campylanthera, and does not support monophyly of the 'insignis' morphological complex.

Four out of the five assembly methods tested recovered *C. trischistandra* as sister species to the rest of the genus and in the HybPiper 3-4:20 pipeline *C. jasminodora* was recovered as the sister species to the rest of the genus. Similarly, the second outmost species recovered by four of the pipelines was *C. pentandra* and for HybPiper 3-4:20 it was a clade containing *C. trischistandra*, *C. pentandra*, *C. samauma*, the South American SDTF clade, and the Central American and Mexican SDTF clade. *Ceiba trischistandra* is a dry forest species with a narrow distribution in the east of Ecuador (Figure 1.4). *Ceiba jasminodora* occurs in the "campos ruprestres" ("rocky field") habitat on granite outcrops within the Cerrado biome in the north of Minas Gerais state in east of Brazil. The variation observed in the topology at the base of the *Ceiba* tree has important implications for understanding the evolutionary and biogeographic history of the genus, for example the role of biome shifts early in the evolution of *Ceiba*.

Conclusions

This chapter produced a well resolved and highly supported phylogeny, which included 103 samples of all 18 described species for *Ceiba*, with multiple accessions per species. The next-generation hybrid capture sequencing of 377 nuclear loci increased the phylogenetic resolution in comparison with the phylogeny based on Sanger-sequence of the ITS region from Chapter 2. In addition, a taxon-specific bait set improved capture

efficiency, especiallly for recent radiations in tropical plants.

NGS approaches represent an important advance for the field of genetics, and a great promise to help elucidate topics such as the distinctive plant diversification patterns found in different biomes. However, data analysis and interpretation should be done thoroughly. Good practice in the analysis of NGS data involve applying different pipelines with various settings followed by an investigation of their biological implications. It is important to understand the means by which the different software packages work. The assessment of intermediate steps of the analysis is normally made based on metrics related to the amount of data. Although BWA had more reads mapping back to the reference, Bowtie2 performed better in this study. More data does not necessarily mean better data. It is important to inspect visually the intermediate files. For example, the sam/bam (sequence alignment map/binary alignment map) files are generated after mapping the raw reads to the reference. By inspecting those files visually, the user can identify particular regions with variable sites and the frequency of different base pairs in a certain position can be a good indication of the presence of paralogues or alleles. For Bowtie2, the variation in the alignment threshold had a consistent pattern of lowering the percentage of reads mapped as the alignment score increased, as expected. Furthermore, the examination of the bam files showed that fewer potential paralogues were being mapped to the reference. Conversely, BWA did not show a consistent pattern of fewer reads being mapped as I increased the stringency of the mapping threshold. Likewise, the bam files showed an unexpected increase in gappy regions with the variation in the mapping settings. Nonetheless, to exclude possible paralogues, testing a range of alignment scores is important to decide on the best value for a particular data set.

Filtering the raw reads with the default values of Trimmomatic (3-4:15) results in a data set with good balance between data quality and data loss. Any eventual poor quality base can be discarded in subsequent steps of data analysis, for example by applying a conservative mapping threshold.

Amongst the different pipelines I tested, the Nicholls et al. (2015) pipeline was the

most suitable one because it allowed me to conduct each step of the analysis separately, varying the default setting of the software and inspecting visually the intermediate files. Pipelines designed to reduce the manipulation of the data by the user intend to facilitate the analysis, but can create difficulties for the users to vary their settings and inspect the intermediate results. Therefore, I used the phylogeny inferred using the data filtered with Trimmomatic default settings (3-4:15) and analysed using the Nicholls et al. (2015) pipeline with 190 Bowtie2 threshold as input for the analysis in Chapters 4 and 5 (Figures 3.31 and 3.36).

Bibliography

Abadi S, Azouri D, Pupko T, Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*. 10:934.

Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B. 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics*. 131:1541–1554.

Bankevich A, Nurk S, Antipov D, et al. (16 co-authors). 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*. 19:455–477.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 30:2114–2120.

Boussau B, Daubin V. 2010. Genomes as documents of evolutionary history. *Trends in Ecology & Evolution*. 25:224–232.

Burrows M, Wheeler D. 1994. A block-sorting lossless data compression algorithm. Technical report, Technical Report Digital Equipment Corporation, Palo Alto.

Carlsen MM, Fér T, Schmickl R, Leong-Škorničková J, Newman M, Kress WJ. 2018. Resolving the rapid plant radiation of early diverging lineages in the tropical Zingiberales: Pushing the limits of genomic data. *Molecular Phylogenetics and Evolution*. 128:55–68.

Carvalho-Sobrinho JG, Alverson WS, Alcantara S, Queiroz LP, Mota AC, Baum DA. 2016. Revisiting the phylogeny of Bombacoideae (Malvaceae): Novel relationships, morphologically cohesive clades, and a new tribal classification based on multilocus phylogenetic analyses. *Molecular Phylogenetics and Evolution*. 101:56–74.

Chau JH, Rahfeldt WA, Olmstead RG. 2018. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Applications in Plant Sciences*. 6:e1032.

Compeau PEC, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*. 29:987–991.

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*. 99:291–311.

Danecek P, Auton A, Abecasis G, et al. (12 co-authors). 2011. The variant call format and VCFtools. *Bioinformatics*. 27:2156–2158.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*. 24:332–340.

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. 2013. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS ONE*. 8:e85024.

DRYFLOR. 2016. Plant diversity patterns in neotropical dry forests and their conservation implications. *Science*. 353:1383–1387.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution*. 63:1–19.

Enright AJ. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*. 30:1575–1584.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*. 8:175–185.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the boostrap. *Evolution*. 39:783–791.

Fér T, Schmickl RE. 2018. Hybphylomaker: Target enrichment data analysis from raw reads to species trees. *Evolutionary Bioinformatics*. 14.

Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*. 27:480–488.

Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biology*. 9:235.

Guo X, Thomas DC, Saunders RM. 2018. Gene tree discordance and coalescent methods support ancient intergeneric hybridisation between Dasymaschalon and Friesodielsia (Annonaceae). *Molecular Phylogenetics and Evolution*. 127:14–29.

Harrison N, Kidner CA. 2011. Next-generation sequencing and systematics: What can a billion base pairs of DNA sequence data do for you? *Taxon*. 60:1552–1566.

Hart ML, Forrest LL, Nicholls JA, Kidner CA. 2016. Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon*. 65:1081–1092.

Hatem A, Bozdag D, Catalyurek UV. 2011. Benchmarking Short Sequence Mapping Tools. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, volume 33, pp. 109–113.

Herrando-Moraira S, Calleja JA, Carnicero P, et al. (21 co-authors). 2018. Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae). *Molecular Phylogenetics and Evolution*. 128:69–87.

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ. 2016. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. *Applications in Plant Sciences*. 4:1600016.

Johnson MG, Pokorny L, Dodsworth S, et al. (18 co-authors). 2018. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Systematic Biology*. 0:1–36.

Kadlec M, Bellstedt DU, Le Maitre NC, Pirie MD. 2017. Targeted NGS for species level phylogenomics: made to measure or one size fits all? *PeerJ*. 5:e3569.

Kubatko LS, Degnan JH. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology*. 56:17–24.

Lambert SM, Reeder TW, Wiens JJ. 2015. When do species-tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny. *Molecular Phylogenetics and Evolution*. 82:146–155.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9:357–359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 10:R25.

Langmead B, Wilks C, Antonescu V, Charles R. 2018. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*. pp. 1–12.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with BurrowsWheeler transform. *Bioinformatics*. 26:589–595.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*. 2012:1–11.

MacManes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*. 5:1–7.

Maddison WP. 1997. Gene Trees in Species Trees. *Systematic Biology*. 46:523–536.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*. 24:133–141.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 17:10.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*. 66:526–538.

McGee MD, Faircloth BC, Borstein SR, Zheng J, Darrin Hulsey C, Wainwright PC, Alfaro ME. 2016. Replicated divergence in cichlid radiations mirrors a major vertebrate innovation. *Proceedings of the Royal Society B: Biological Sciences*. 283:20151413.

McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences*. 6:e1038.

Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science*. 6:1–20.

Pfeifer SP. 2017. From next-generation resequencing reads to a high-quality variant data set. *Heredity*. 118:111–124.

Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*. 100:56–62.

Russell A, Samuel R, Rupp B, Barfuss MHJ, Safran M, Besendorfer V, Chase MW. 2010. Phylogenetics and cytology of a pantropical orchid genus Polystachya (Polystachyinae, Vandeae, Orchidaceae): evidence from plastid DNA sequence data. *Taxon*. 59:389–404.

Schott RK, Panesar B, Card DC, Preston M, Castoe TA, Chang BS. 2017. Targeted capture of complete coding regions across divergent species. *Genome Biology and Evolution*. 9:evx005.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. *Nature*. 550:345–353.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nature Biotechnology*. 26:1135–1145.

Simmons MP, Gatesy J. 2015. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular Phylogenetics and Evolution*. 91:98–122.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 6.

Smith Sa, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*. 15:150.

Souza HA, Muller LA, Brandão RL, Lovato MB. 2012. Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis* (Leguminosae), a tree from the Brazilian Cerrado. *Genetics and molecular research : GMR*. 11:756–764.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–1313.

Stamatakis A. 2016. The RAxML v8.2.X Manual. Technical Report 1, Heidelberg Institute for Theoretical Studies.

Thankaswamy-Kosalai S, Sen P, Nookaew I. 2017. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*. 109:186–191.

Tsangaras K, Wales N, Sicheritz-Pontén T, Rasmussen S, Michaux J, Ishida Y, Morand S, Kampmann ML, Gilbert MTP, Greenwood AD. 2014. Hybridization Capture Using Short PCR Products Enriches Small Genomes by Capturing Flanking Sequences (CapFlank). *PLoS ONE*. 9:e109101.

Yang Y, Smith Sa. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*. 14:328.

Yang Y, Smith Sa. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution*. 31:3081–3092.

Yang Z. 2014. Molecular Evolution - a statistical approach. Oxford: Oxford University Press.

Yu X, Yang D, Guo C, Gao L. 2018. Plant phylogenomics based on genome-partitioning strategies: Progress and prospects. *Plant Diversity*. 40:158–164.

3.A Appendix

Occurrence of the accessions of *Ceiba* sequenced mapped in relation to the distribution of each of the 18 species.

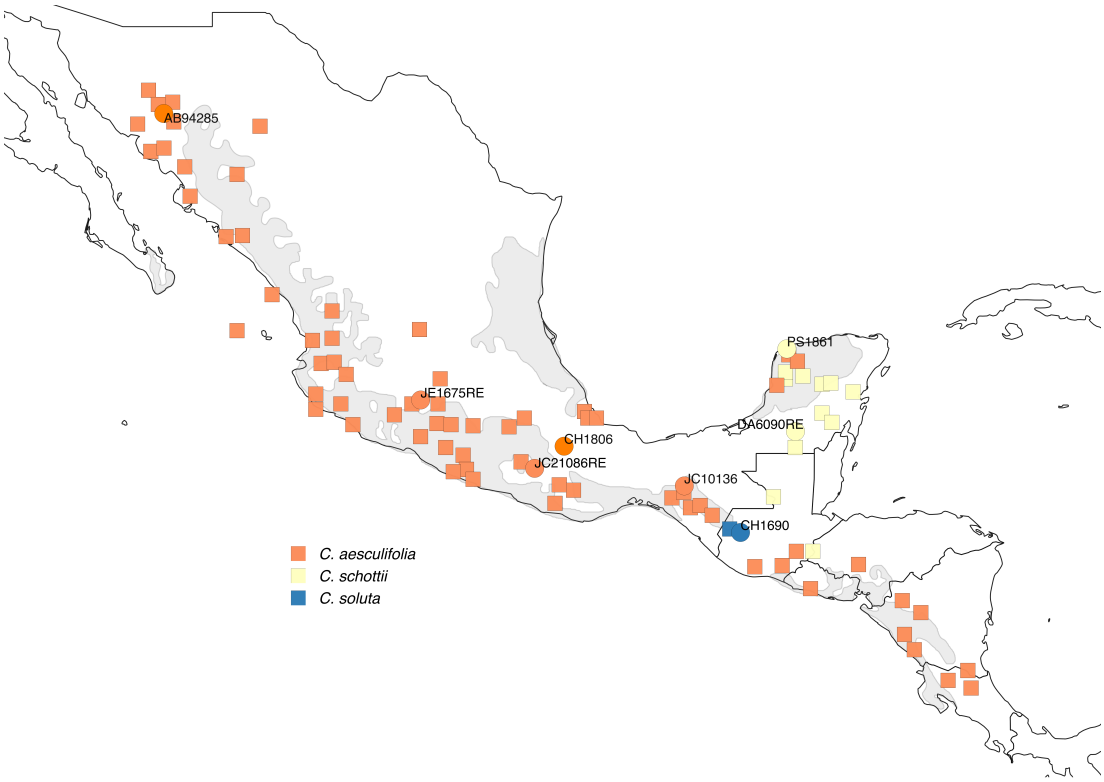


Figure 3.40: Occurrence of the accessions of *Ceiba* from Central American and Mexican SDTF sequenced mapped in relation to the distribution of each species. Different colours represent different species. Circles represent the distribution of each species and squares represent samples sequeced within each species. Grey areas represent SDTF patches according to DRYFLOR (2016).

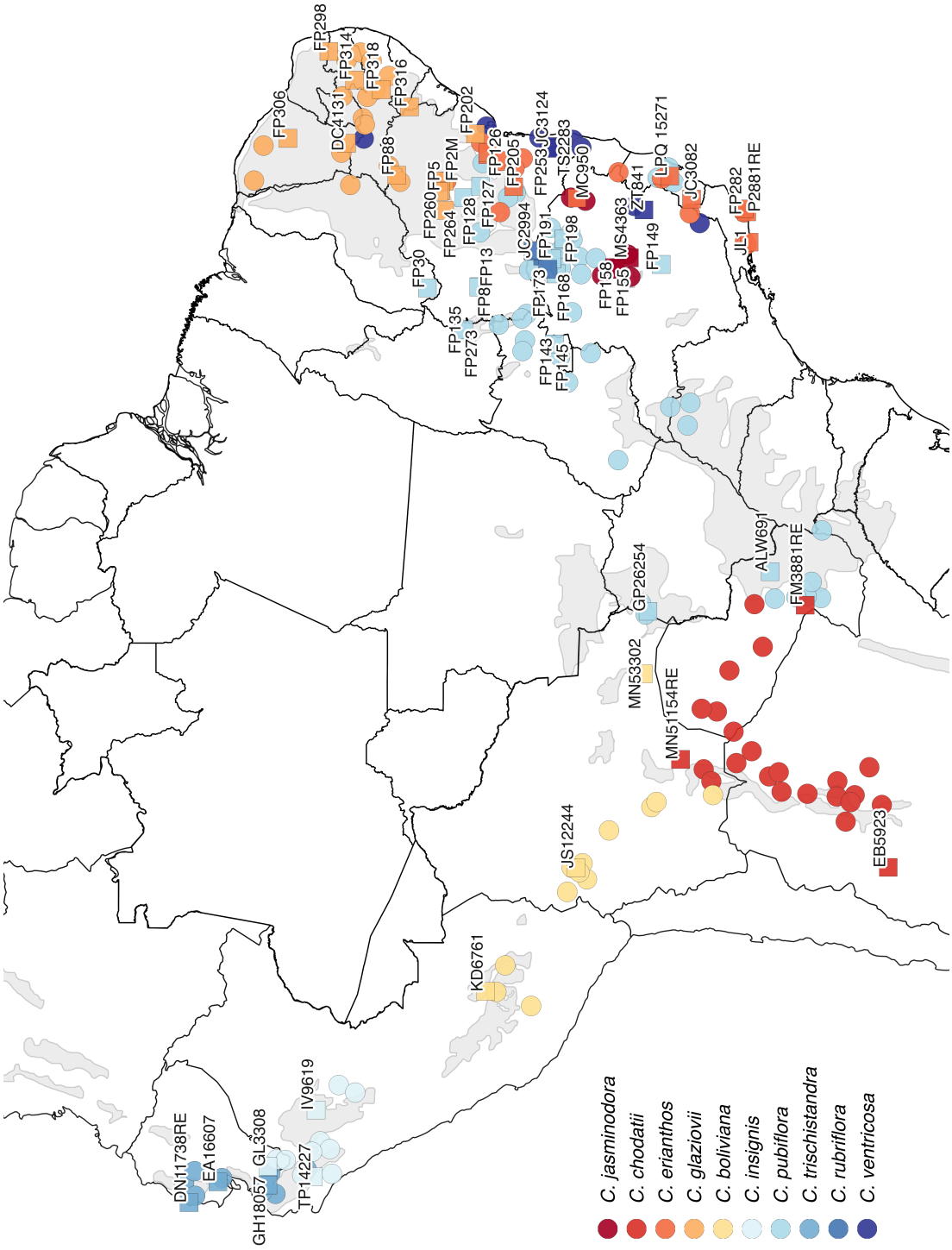


Figure 3.41: Occurrence of the accessions of *Ceiba* from South American SDTF sequenced mapped in relation to the distribution of each species. Different colours represent different species. Circles represent the distribution of each species and squares represent samples sequeced within each species. Grey areas represent SDTF patches according to DRYFLOR (2016).

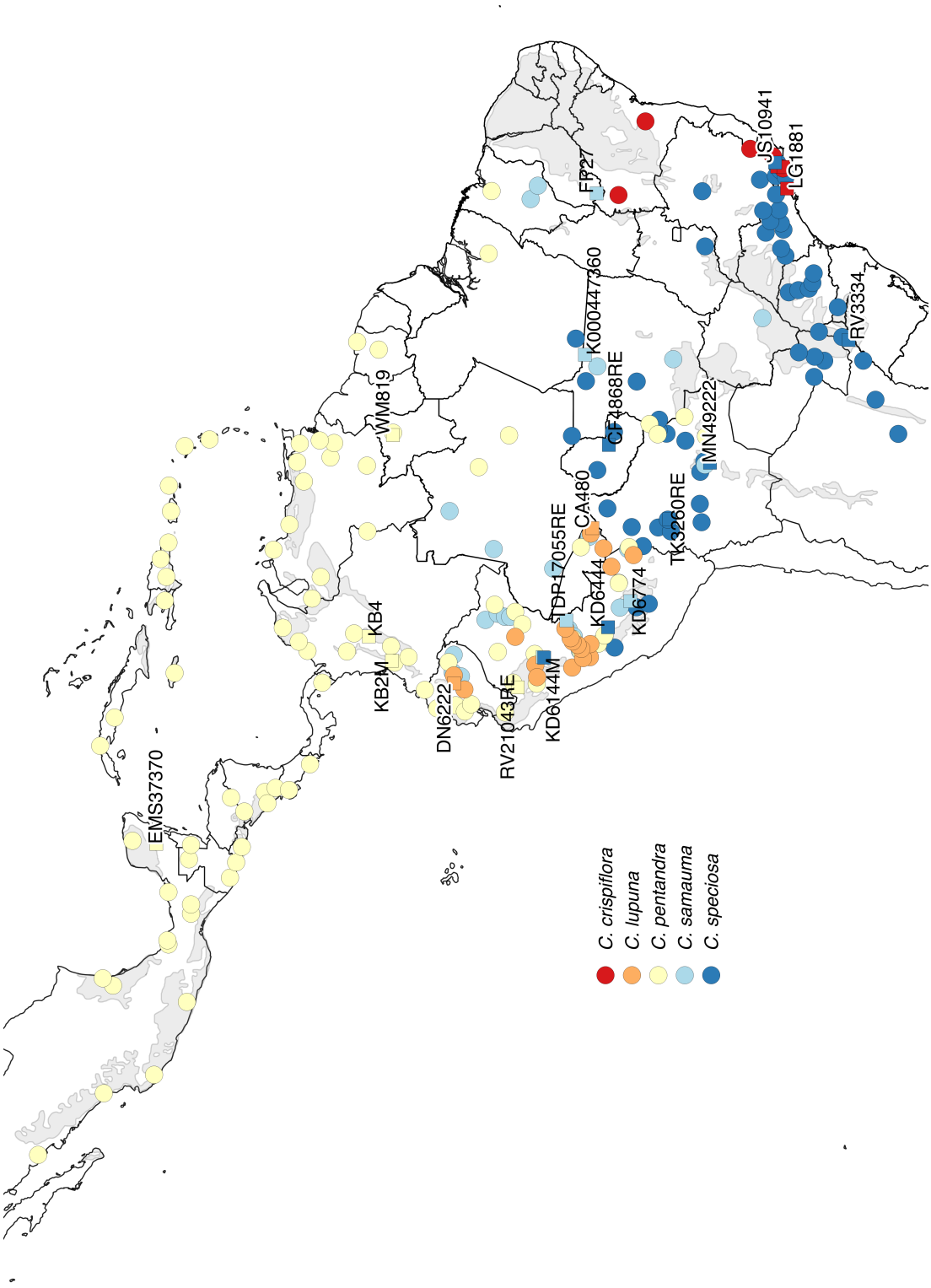


Figure 3.42: Occurrence of the accessions of *Ceiba* from rain forests sequenced mapped in relation to the distribution of each species. Different colours represent different species. Circles represent the distribution of each species and squares represent samples sequeced within each species. Grey areas represent SDTF patches according to DRYFLOR (2016).

Variation in metrics from Nicholls et al. (2015) pipeline for the two inputs 3-4:15 and 3-4:20.

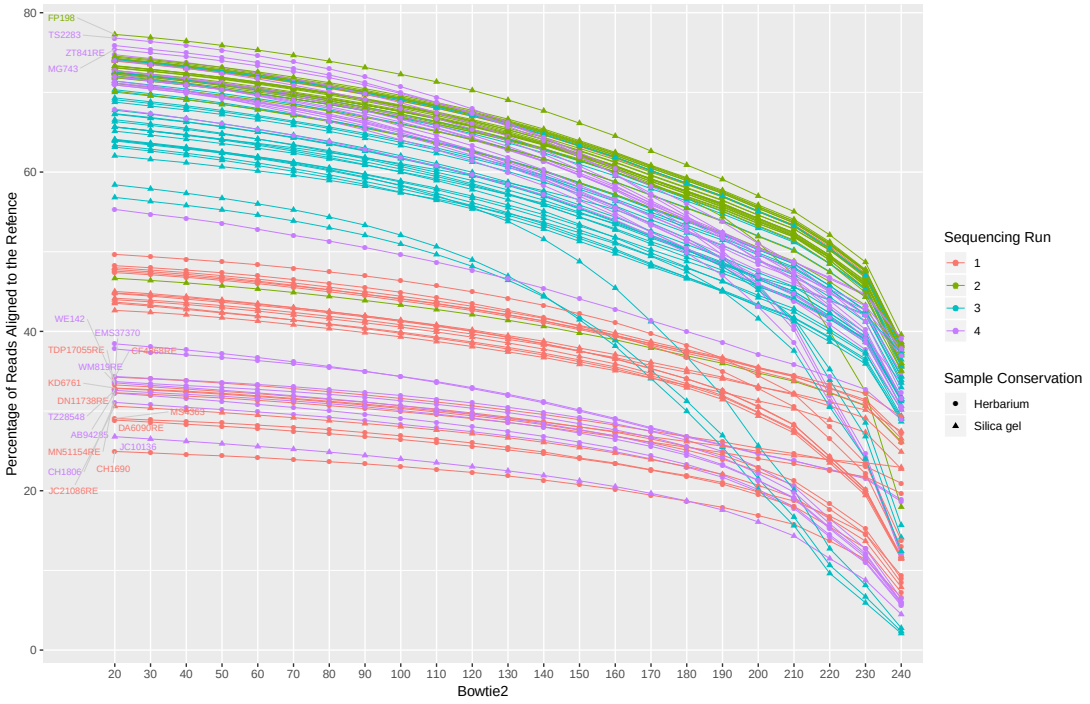


Figure 3.43: Percentage of reads mapped with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:20.

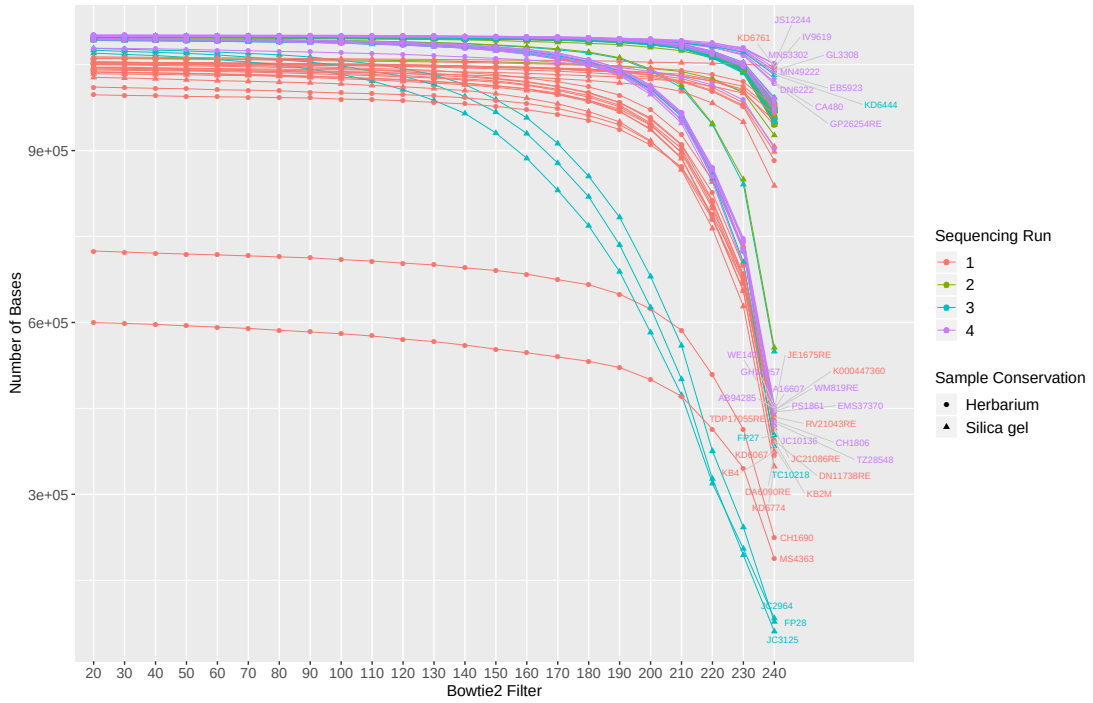


Figure 3.44: Number of base-pairs with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

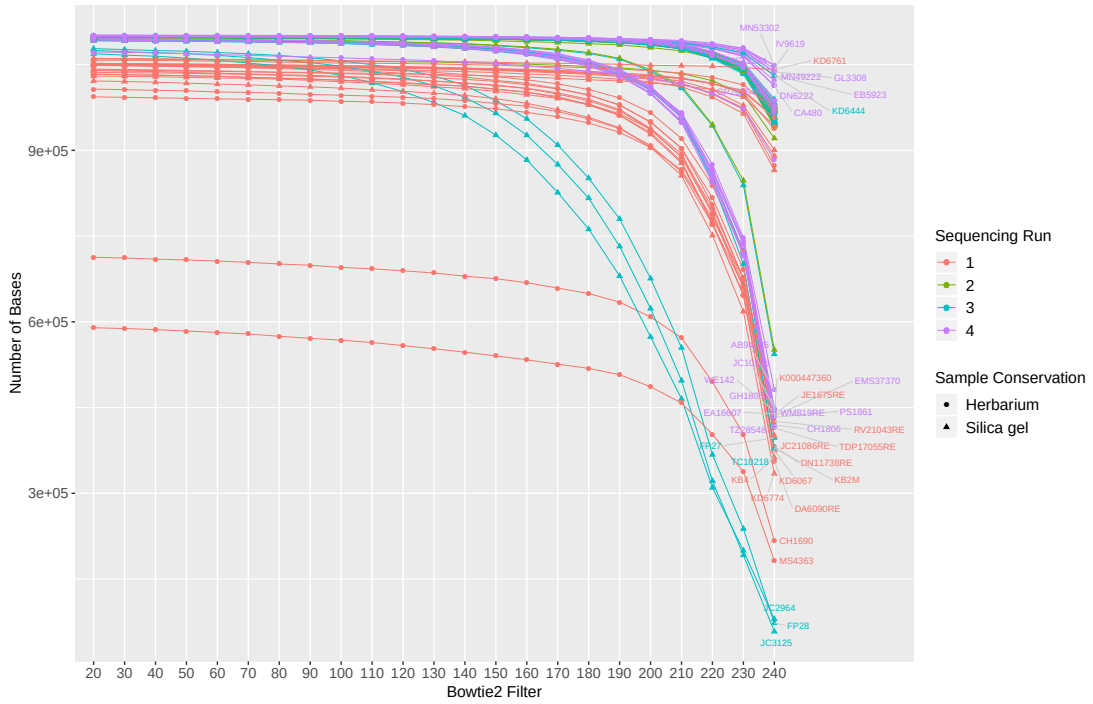


Figure 3.45: Number of base-pairs with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:20.

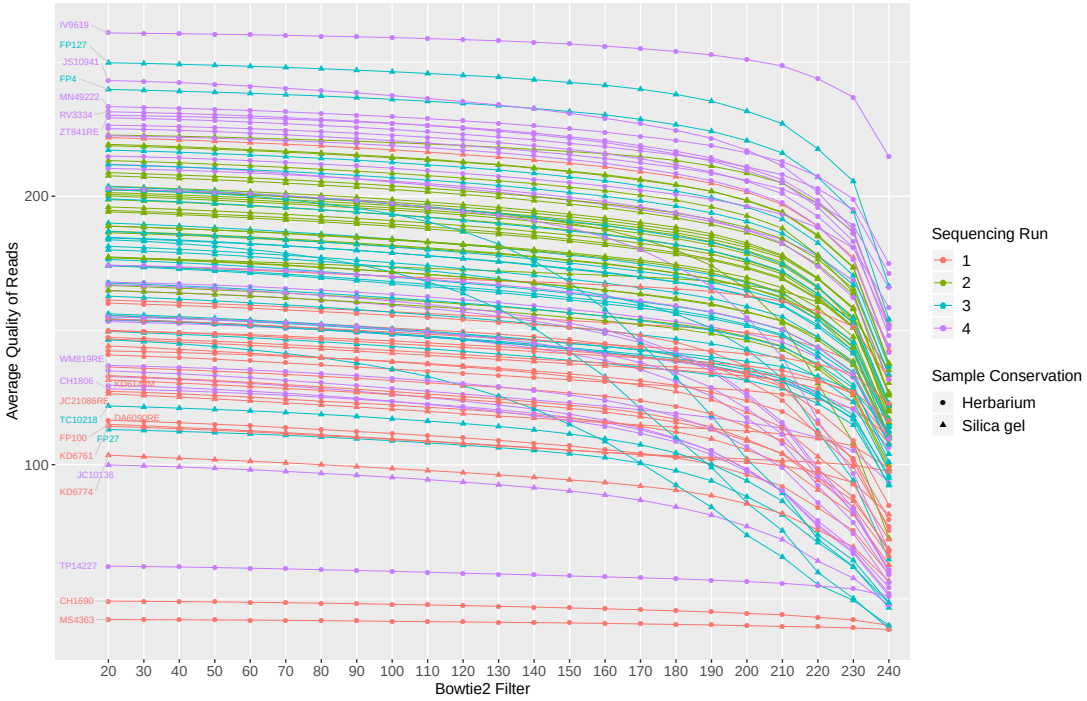


Figure 3.46: Average quality of reads with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

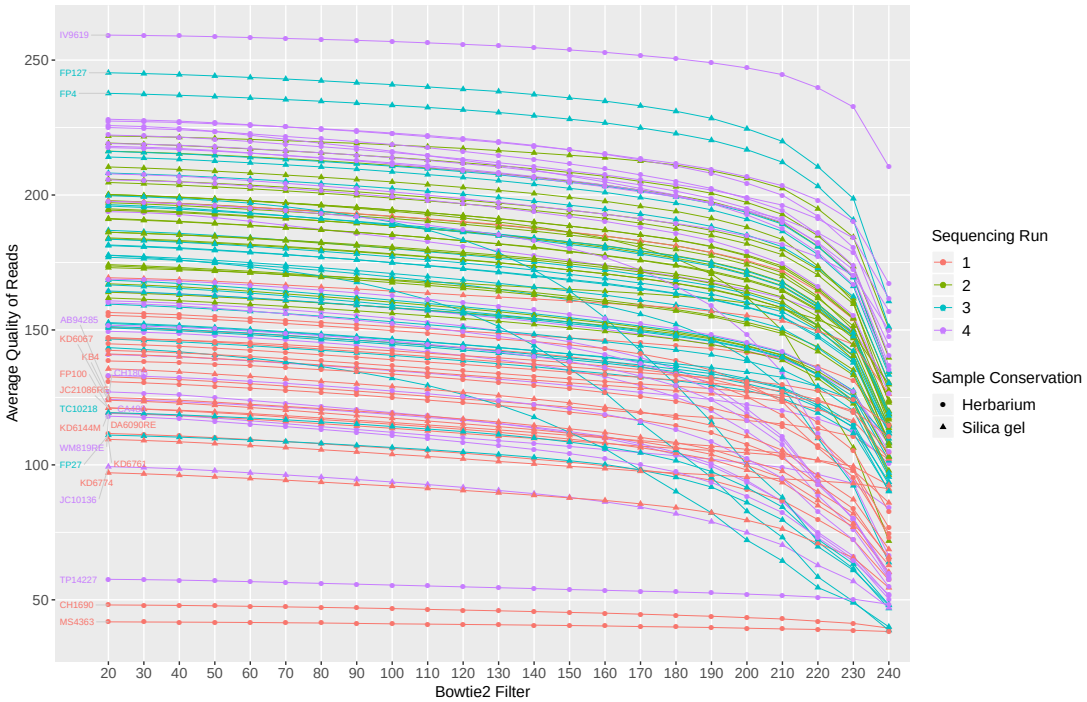


Figure 3.47: Average quality of reads with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:20.

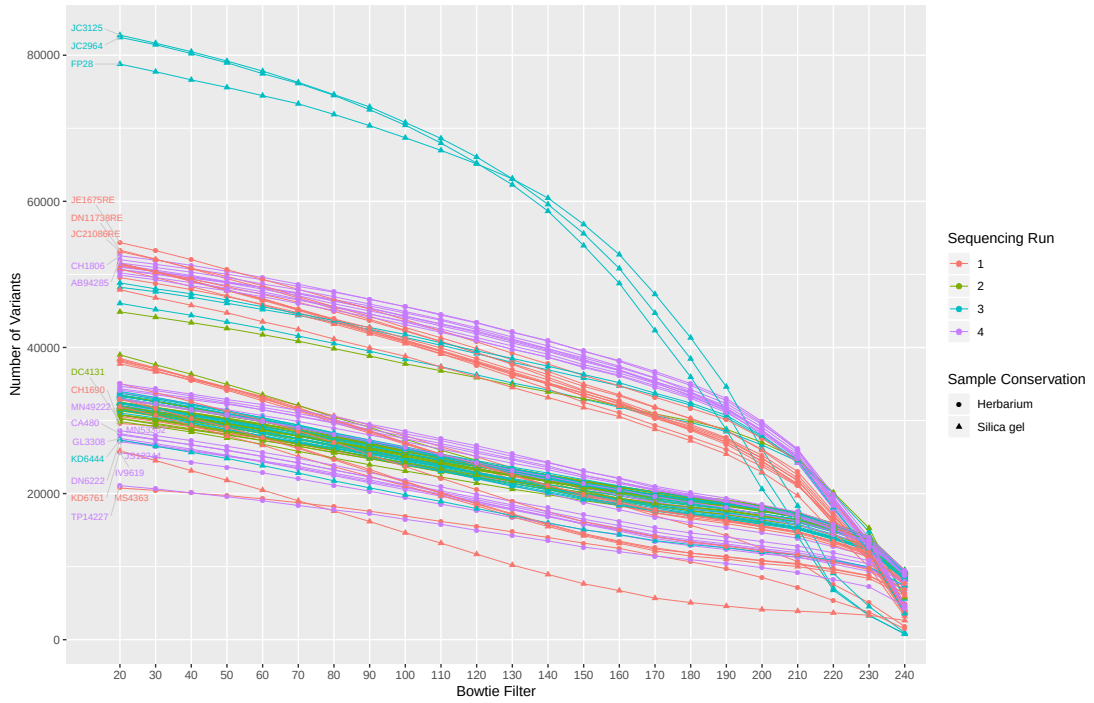


Figure 3.48: Number of variants with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:20.

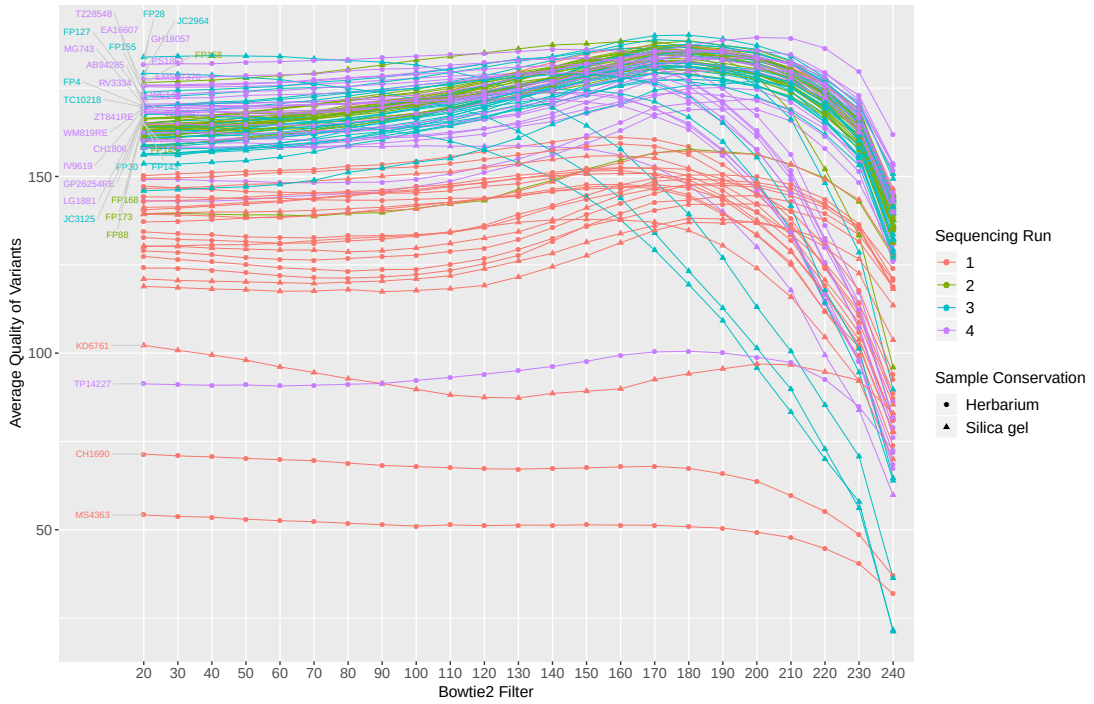


Figure 3.49: Quality of variants with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

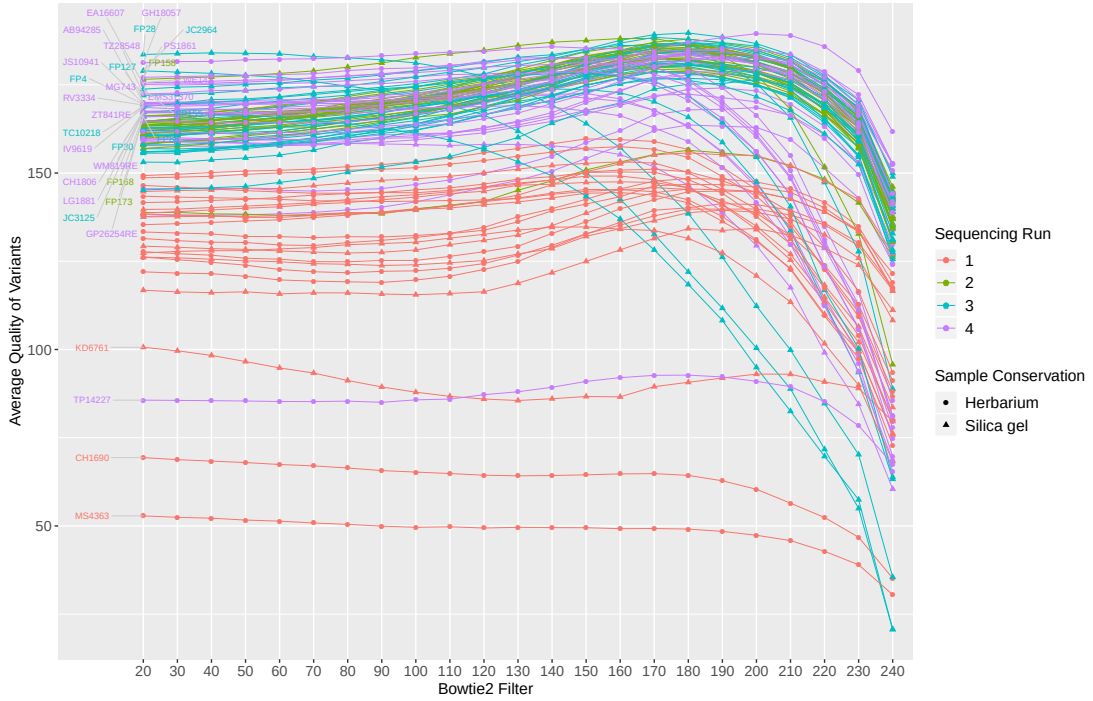


Figure 3.50: Quality of variants with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:20.

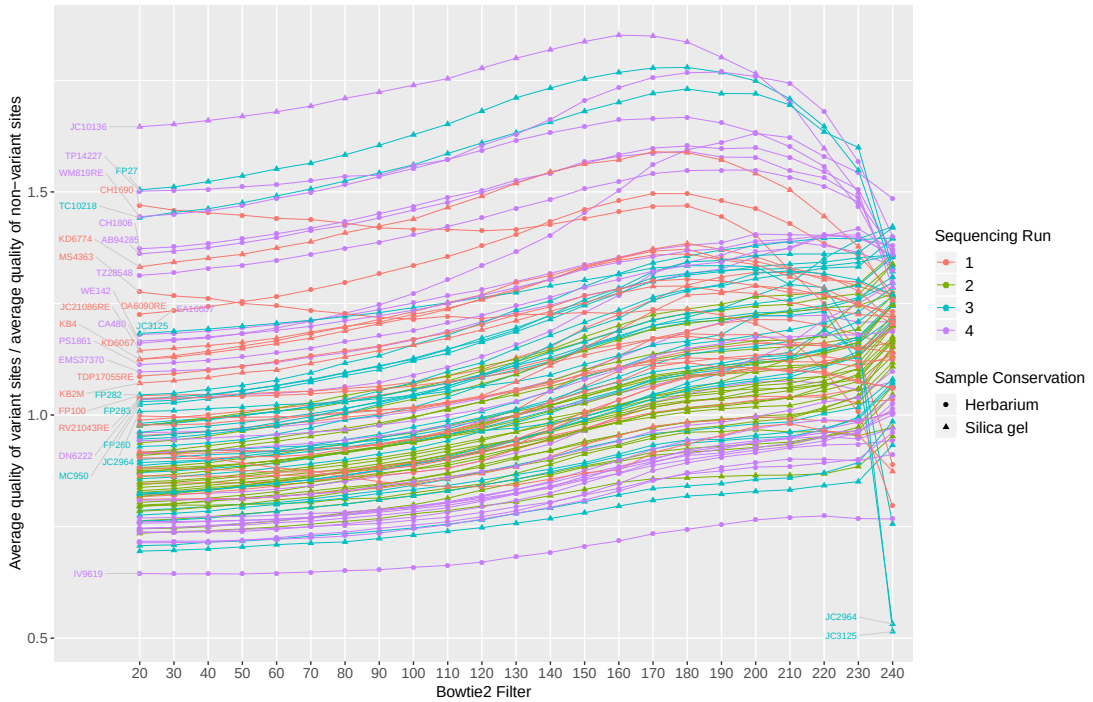


Figure 3.51: Standardized variants quality with varied Bowtie2 threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:20.

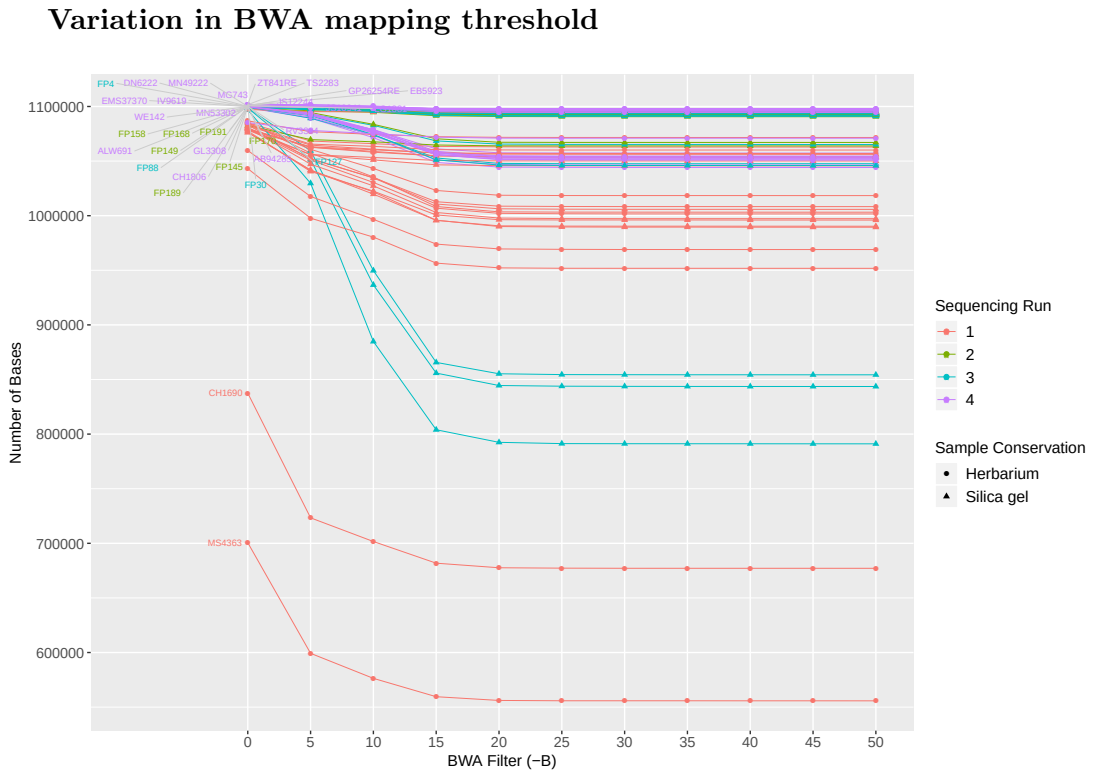


Figure 3.52: Number of base pairs retrieved with varied BWA threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

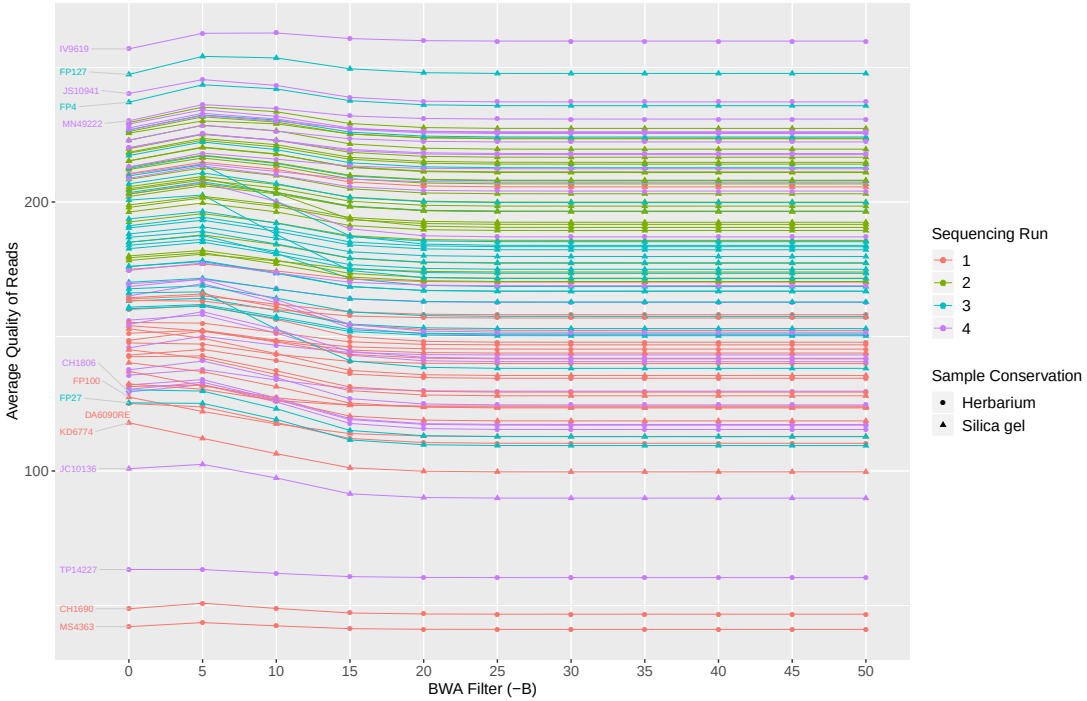


Figure 3.53: Average quality of reads with varied BWA threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

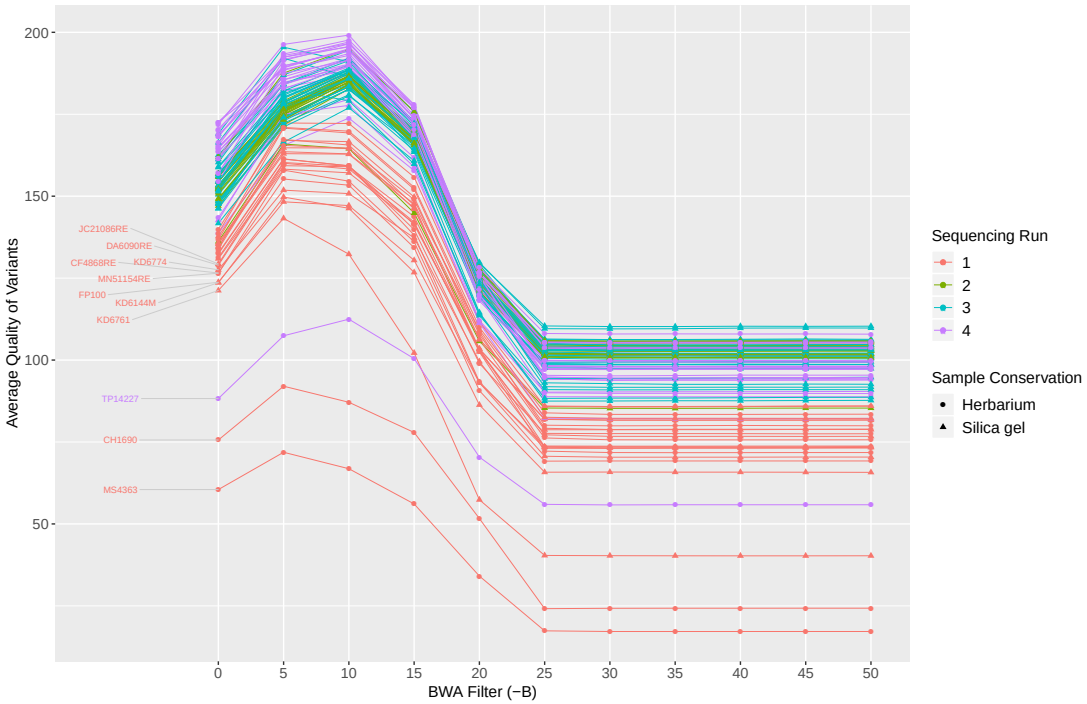


Figure 3.54: Average quality of variants with varied BWA threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

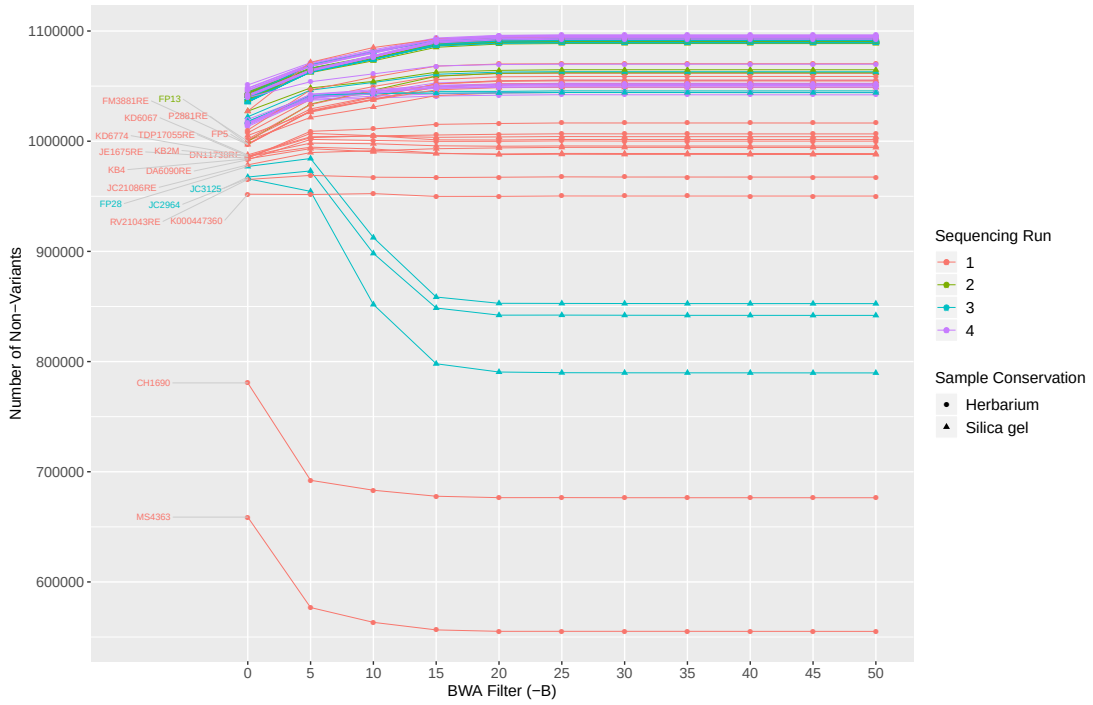


Figure 3.55: Number of non-variants quality with varied BWA threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

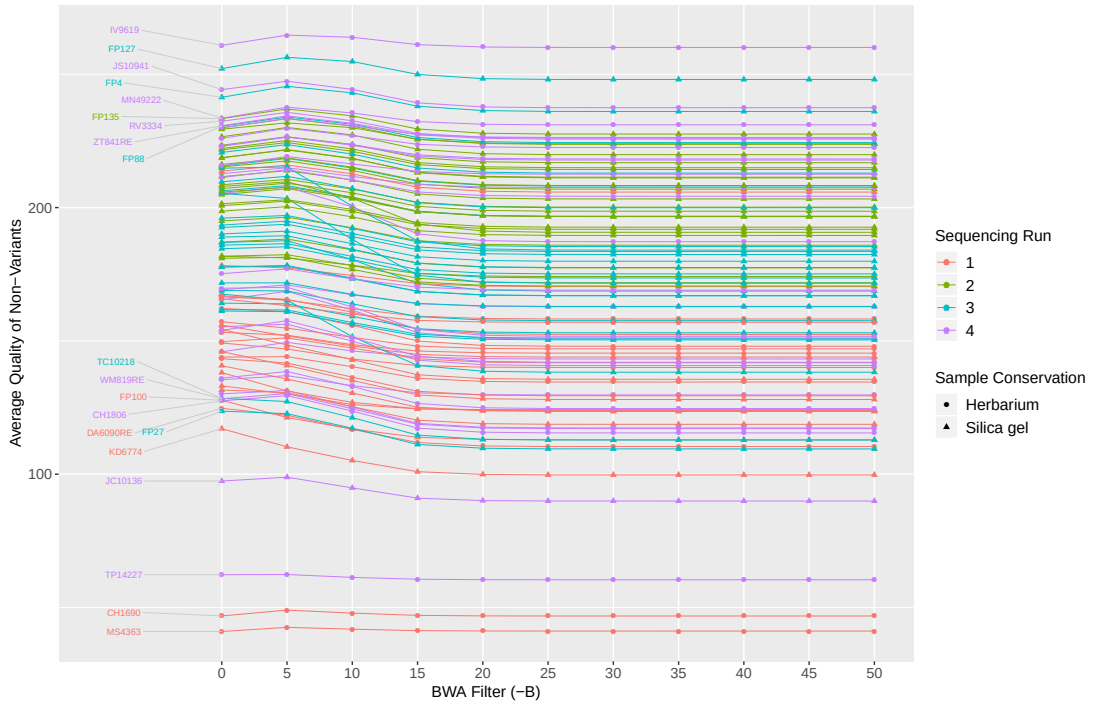


Figure 3.56: Average quality of non-variants with varied BWA threshold alignment scores. Input data from Trimmomatic filtering setting 3-4:15.

Phylogenies from different inputs and different pipelines.

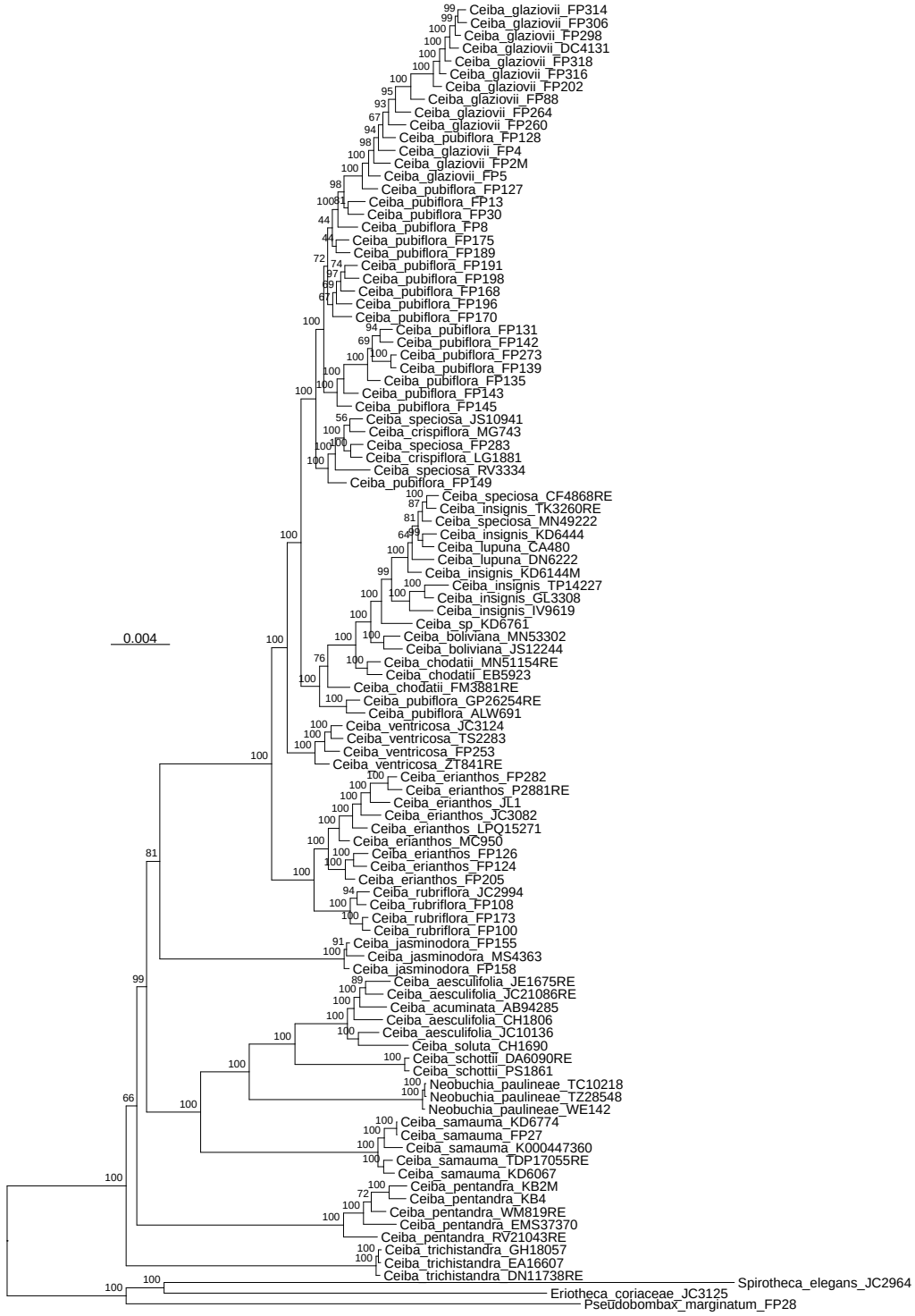


Figure 3.57: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:20 Trimmomatic using 190 Bowtie2 threshold.Numbers above branches represent bootstrap values.

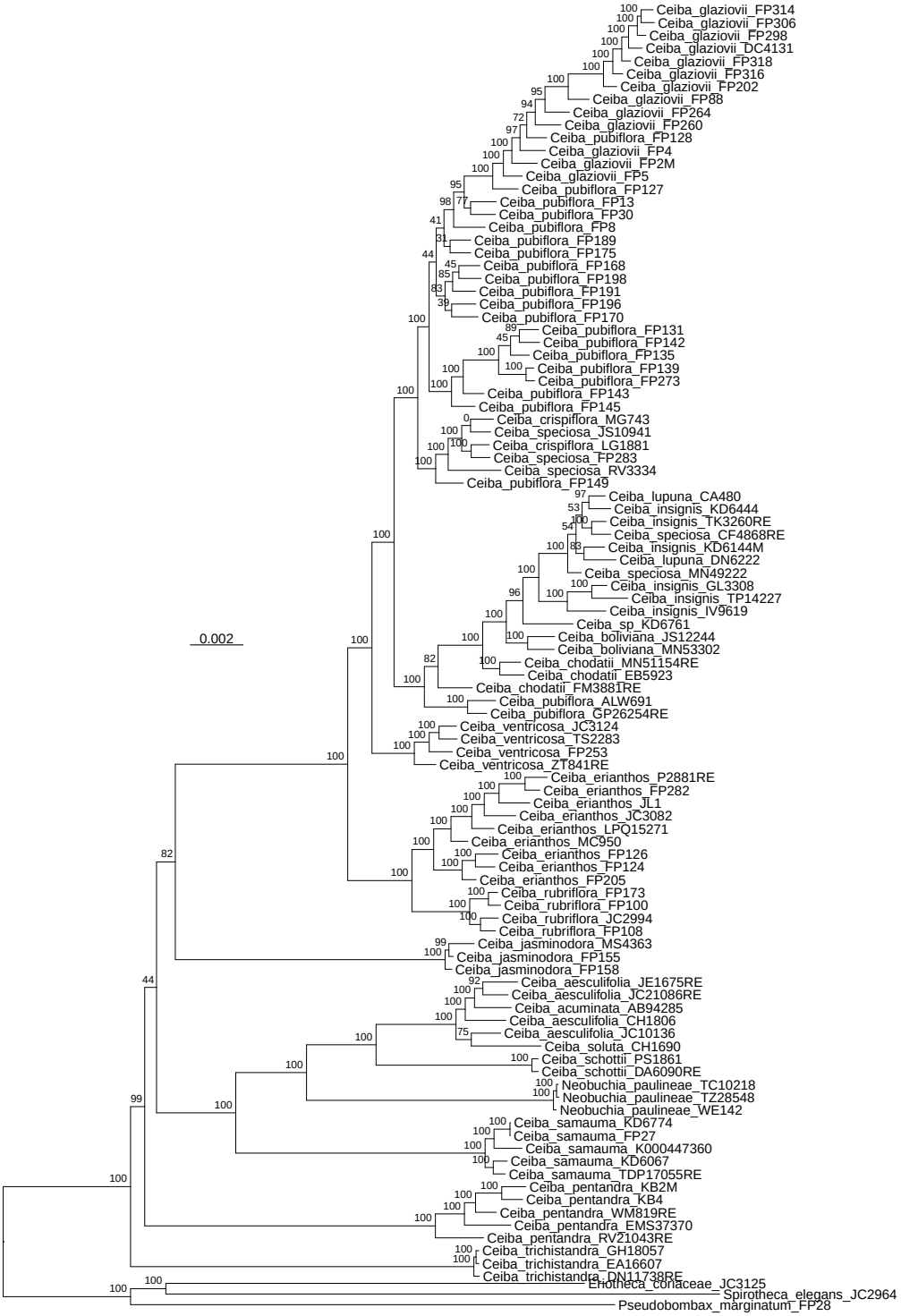


Figure 3.58: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:15 Trimmomatic using 10 BWA threshold. Numbers above branches represent bootstrap values.

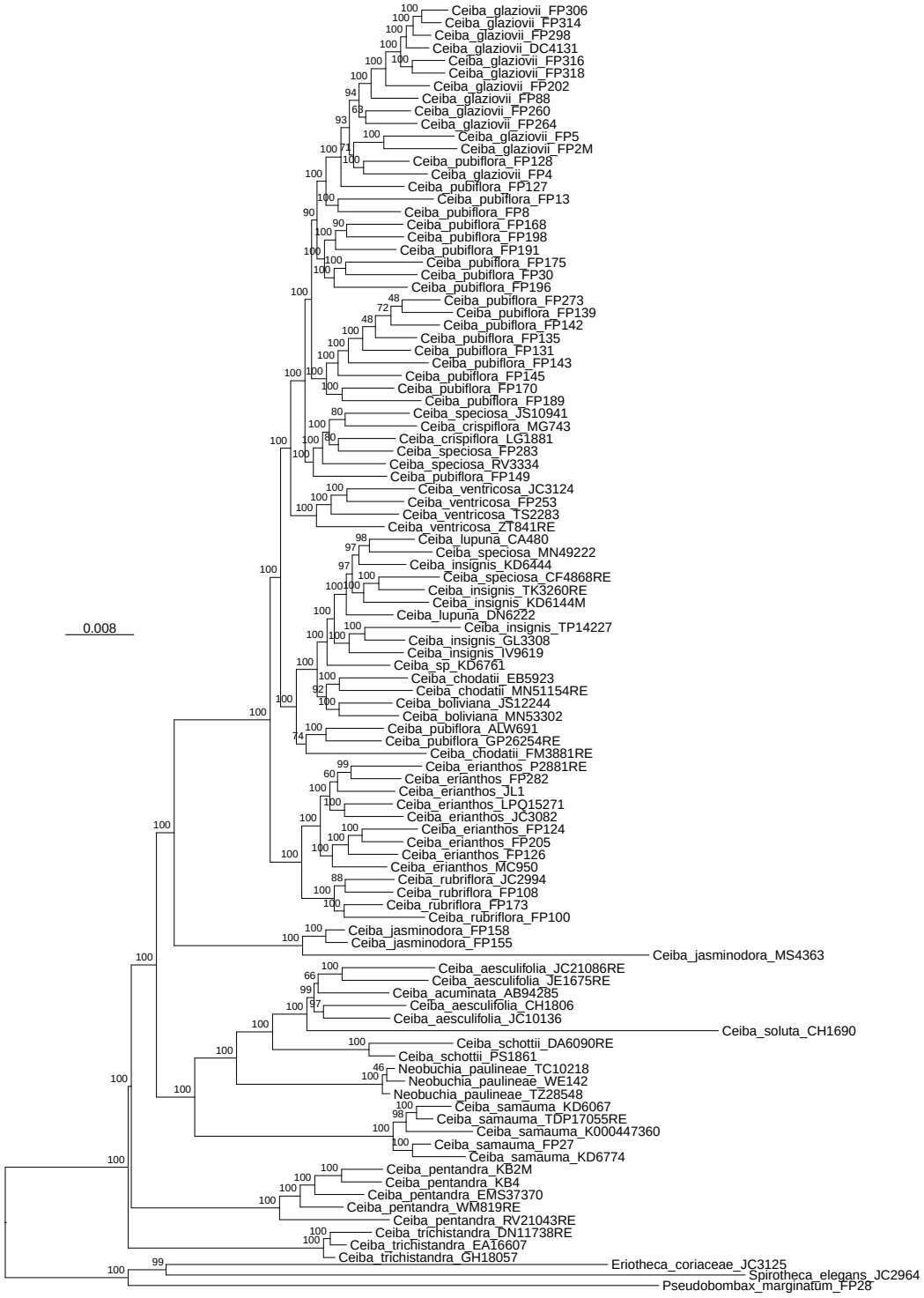


Figure 3.59: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* unsing the concatenated matrix approach. Input data were 3-4:15 Trimmomatic using the HybPiper pipeline. Numbers above branches represent bootstrap values.

Figure 3.60: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* using the concatenated matrix approach. Input data were 3-4:20 Trimmomatic using the HybPiper pipeline. Numbers above branches represent bootstrap values.

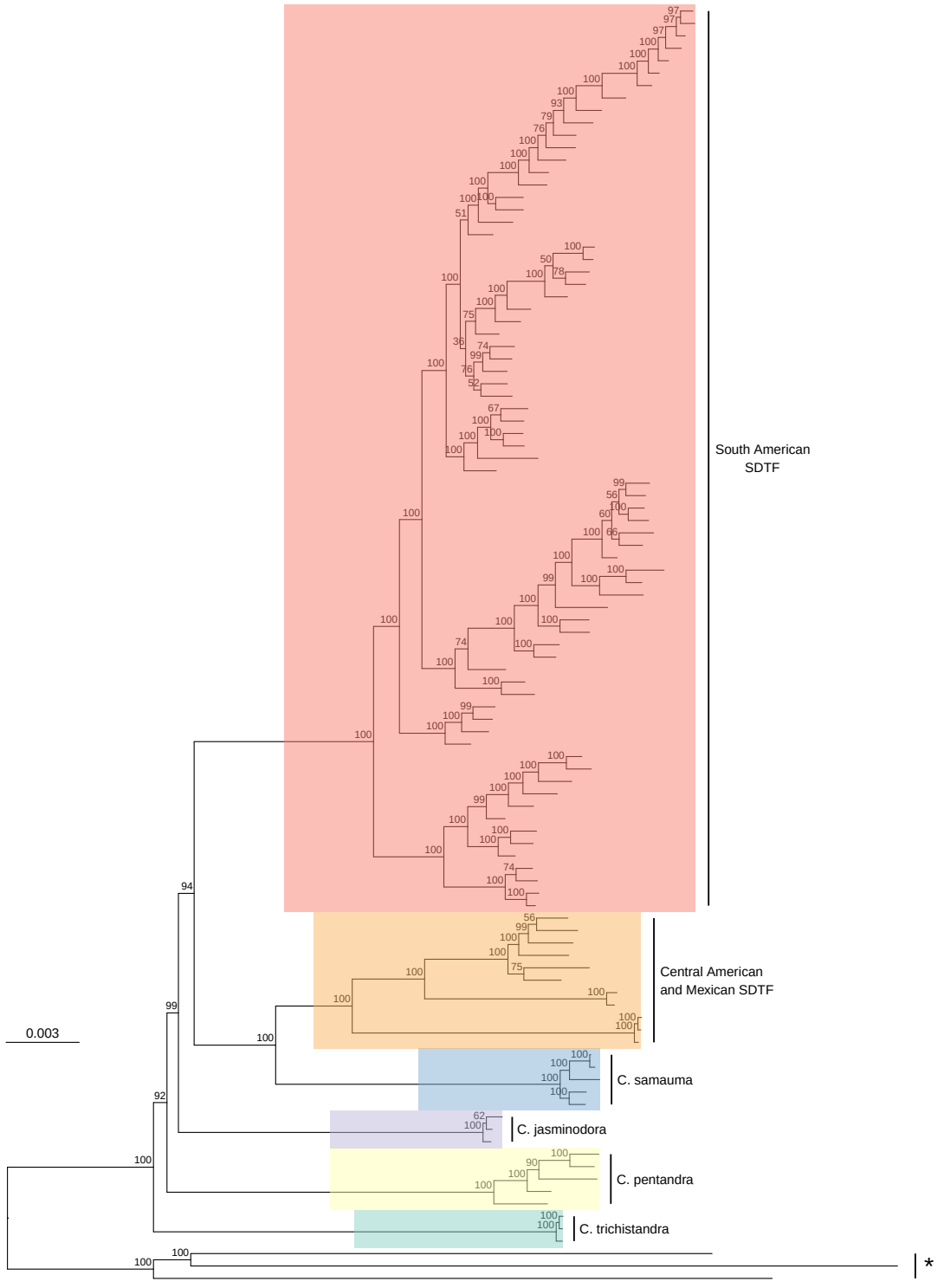


Figure 3.61: Maximum likelihood phylogram derived from analysis of 377 nuclear loci for 18 species of *Ceiba* unsing the concatenated matrix approach. Input data were 3-4:15 Trimmomatic using 190 Bowtie2 threshold with no reverse unpaired reads. Numbers above branches represent bootstrap values and clades with an asterisk (*) represent the ougroups.

Chapter 4

Taxonomy of *Ceiba*: a phylogenetic perspective

4.1 Introduction

The previous chapter presented a phylogeny based on NGS hybrid capture sequencing of 377 nuclear loci for the genus *Ceiba*. The NGS hybrid capture phylogeny is fully resolved, with high support for all branches. However, concatenating all loci ignores potentially conflicting signal amongst different loci (Doyle, 1992; Maddison, 1997). The conflicting signal in the data can be caused either by methodological issues, such as contamination, poor variant calls or presence of paralogues; or by natural process such as incomplete lineage sorting or hybridisation (Edwards, 2009) followed by introgression (Harrison and Larson, 2014). All these processes would result in conflicts in the genealogical history of the loci. However, some properties of phylogenies inferred with concatenated matrices can be used to predict where the species tree and the concatenated tree will disagree (Lambert et al., 2015). An investigation of the incongruence of gene trees, species delimitation and species tree reconstruction will be the focus of this chapter.

The large amount of multi-locus sequence data now available has helped to shed light on important issues in phylogenetics. It has become more evident that phylogenies inferred from matrices of concatenated loci often disagree with the species tree (Lambert et al., 2015; Smith et al., 2015). This is due to the fact that individual gene trees can be discordant amongst each other and with the species tree itself (Doyle, 1992; Maddison, 1997). Therefore, individual gene trees cannot be treated as an accurate estimate of species relationships (Degnan and Rosenberg, 2009). Reasons for gene tree heterogeneity include hybridisation, horizontal gene transfer, incomplete lineage sorting (ILS, also called deep coalescence), and gene duplication (Doyle, 1992; Maddison, 1997; Degnan and Rosenberg, 2009; Edwards, 2009) or even non-biological causes such as sample contamination (Guo et al., 2018). These different processes often result in different inferred species trees (Maddison, 1997).

The inference of the species tree has to account for the process underlying gene tree discordance and distinguishing among those causes is still a challenge (Guo et al., 2018). Gene duplication, for example, can be excluded with proper orthology assign-

ment during sequence assembly (Chapter 3, Smith et al. 2015) and horizontal gene transfer is more often recorded in prokaryotes (Galtier and Daubin, 2008) or in mitochondrial genes for plants (Richardson and Palmer, 2007). However, discriminating between hybridisation and ILS is not straightforward since they show similar phylogenetic patterns (Holder et al., 2001; Buckley et al., 2006; Smith et al., 2015; Guo et al., 2018). In fact, methods developed to infer species trees often assume that only one or few of these causes are the source of gene tree heterogeneity, and they mainly presume ILS is the principal one (Smith et al., 2015).

The complications around inferring a species tree are aggravated by the fact that the species concept itself is debated (Edwards, 2009; Freudenstein et al., 2016), and so as a consequence are the criteria used to delimit species (De Queiroz, 2007). As a result, species delimitation methods often raise the question of whether species or structure are being delimited (Sukumaran and Knowles, 2017; Leaché et al., 2019). The increase in the availability of multi-locus data for multiple individuals per species, allied to the improvement of species delimitation methods, has led to a resurgence of interest in this area (Fujita et al., 2012). These DNA sequence data combined with morphological data are now used to delimit species using an integrative approach (Wiens and Penkrot, 2002; Knowles and Carstens, 2007; Yang and Rannala, 2010; Dexter et al., 2010; Fujita et al., 2012).

Phylogenies built with multiple individuals per species evidenced the discordance between species delimited based on morphological and on genetic data (Wiens and Penkrot, 2002) and revived the discussion about whether species should be monophyletic (Rieseberg and Brouillet, 1994). For example, different modes of speciation such as allopatric, parapatric or sympatric result in different phylogenetic patterns (Rieseberg and Brouillet, 1994; Knowles and Carstens, 2007). Rieseberg and Brouillet (1994) predicted that allopatric speciation is likely to produce monophyletic daughter species, in contrast to parapatric and sympatric speciation that often result in a paraphyletic progenitor and monophyletic derivative species. In this sense, speciation is a process producing a paraphyletic entity (the ancestral species) that only becomes

monophyletic over time. A key aspect is that the time to achieve monophyly of a paraphyletic ancestral species will vary depending on factors such as effective population size, level of gene flow and type of genetic data (Rieseberg and Brouillet, 1994). For example, nuclear genes take longer to coalesce when compared to chloroplast or mitochondrial genes (Rieseberg and Brouillet, 1994; Hudson and Coyne, 2002). Likewise, large census population sizes are more likely to cause longer coalescence times (Hudson and Coyne, 2002) because they reflect large effective population size, which helps the persistence of ancestral polymorphism (Knowles and Carstens, 2007). Therefore, the different phylogenetic patterns of the individual gene trees are an important source of information to investigate the nature of species and their evolutionary history (Knowles and Carstens, 2007).

New methods to infer species trees and delimit species have been developed over the last ten years (Edwards, 2009; Leaché and Rannala, 2011). Recent improvements include the capacity to deal with many loci and many species, and still be computationally efficient. The most successful are constructed under the Coalescent theory (Edwards, 2009; Fujita et al., 2012), which was initially proposed in the early 1980s (Kingman, 1982) as a series of probabilistic models of population genetics and has been developed since (Sigwart, 2009). With the increase of the availability of genetic data, the discordance of gene trees and species trees became more evident (Degnan and Rosenberg, 2009). In this context, the multispecies coalescent model arose as a framework to accommodate the heterogeneity of gene trees while connected by the evolutionary tree (i.e., the species tree; Degnan and Rosenberg (2009)), and therefore to account for natural biological phenomena such as incomplete lineage sorting. This model represents a combination of population genetics, because it encompasses different coalescent process, and phylogenetics because it accounts for the underlying evolutionary history of the species (Rannala and Yang, 2003; Edwards, 2009; Yang, 2015; Flouri et al., 2018). Two main parameters are calculated under the model: species divergence time (τ) (i.e., coalescence time or the time to the most recent common ancestor (MRCA)); and the amount of genetic variation (θ) as a proxy of population size for modern and ancestral

species (Rannala and Yang, 2003). Specifically, θ is calculated as $\theta = 4N\mu$, where μ is the mutation rate per site and N is population size. The τ parameter is calculated using the "expected number of mutations per site from the ancestral node in the species tree to the present time" (Rannala and Yang, 2003).

In this chapter, I aim to explore possible incongruence among the 377 individual gene trees sequenced with the target capture technique, build a species tree for *Ceiba* and investigate the lack of monophyly of the species in the South America SDTF clade (Chapters 2 and 3). Specifically, I aim to answer the following questions:

1. Is there incongruence among the 377 individual gene trees?
2. What is the species tree for *Ceiba*?
3. Are the non-monophyletic *Ceiba* species in the South American STDF clade correctly delimited?

4.2 Methods

4.2.1 Gene tree discordance

To investigate possible discordance among the gene trees and the tree generated using the concatenated alignment I used two different approaches. The first is the measurements of gene and site concordance factors (gCF and sCF respectively) developed by Minh et al. (2018) and implemented in IQ-Tree Beta version 1.7-beta9[1]. The gene concordance factor is a measurement of the proportion of individual gene trees that contain a given branch when compared with the concatenation tree. The site concordance factor is a measurement of the proportion of decisive alignment sites supporting a given branch in the concatenation tree. For the site concordance factor, each individual gene tree is divided in quartets, and their sub-alignment examined. A site is considered decisive if characters are present for each one of the four accessions in the quartet and that site is parsimony informative (Minh et al., 2018).

[1] https://github.com/Cibiv/IQ-TREE/releases

The second approach uses the pipeline developed by Smith et al. (2015) to calculate internode certainty (ICA), which is a measure of the degree of certainty for each internal edge (bipartition) by considering the frequency of all conflicting bipartitions (Salichos et al., 2014).

4.2.2 Loci filtering

Despite the increase in availability of DNA sequence data, individual loci are often filtered (i.e. excluded from analysis by not meeting the criteria of relevant NGS) either because the software cannot deal with the amount of data or because they fail the assumptions of the software (Smith et al., 2018). I filtered the loci using the SortaDate package[2] (Smith et al., 2018). This package uses rooted individual gene trees and a species tree topology as input to calculate three metrics: root-to-tip variance (an indication of clock-likeness), tree length (an indication of discernible information), and bipartition (an indication of concordance with the species tree). These metrics are useful to comply with assumptions of analysis of divergence time estimation especially under the Bayesian framework. Because software capable of dealing with a large amount of genetic data to infer a species tree often uses similar assumptions (eg. clock-like evolution) (Flouri et al., 2018), I used the filtered data set to infer the species tree for *Ceiba*. Therefore, I used the phylogeny generated in Chapter 3 following the concatenation approach as input in SortaDate rather than a species tree built with the complete data set that would likely fail to comply with the assumptions of the analysis. Likewise, software developed to delimit species using the same framework shares similar assumptions (eg. BPP) (Flouri et al., 2018) and thus the same subset of loci was used in the analysis of species delimitation, to infer a species tree and for the dating analysis (Chapter 5).

4.2.3 Coalescent species delimitation analysis

Seven out of the 11 species in the SDTF South American clade were not resolved as monophyletic (Chapter 3). To evaluate the limits of the species I ran a species delimi-

[2]https://github.com/FePhyFoFum/SortaDate

tation analysis under the multispecies coalescence (MSC) model. The MSC model has been used both to infer species trees (Rannala and Yang, 2017) and to delimit species (Yang and Rannala, 2010; Fujita et al., 2012). Different computational methods have been developed to apply MSC to multi-locus genomic sequence data, either based on frequentist or Bayesian frameworks. Many of them assume resolved gene trees without accounting for uncertainties in gene tree inference. However, individual loci often do not contain enough phylogenetically informative characters and gene trees might be poorly resolved. Among these different methods, BPP (Bayesian Phylogenetics and Phylogeography) (Yang, 2015) was considered the most accurate for species delimitation for simulated and empirical data (Camargo et al., 2012). The species delimitation method applied by BPP considers that gene tree conflicts might be the result not only of biological process but also of errors in phylogenetic inference (Yang and Rannala, 2010). BPP implements the MSC model under a Bayesian Markov chain Montecarlo (MCMC) framework and generates posterior probabilities of species delimitation hypotheses. The authors Yang and Rannala (2010) define species as follows:

"Our current implementation considers good biological species only, in which exchange of migrants ceases as soon as species separate, and uses genomic data to examine the evidence concerning competing models of species delimitation given this species definition."

BPP implements four types of analysis: A00, which is the estimation of τ (species divergence time) and θ (population size for modern and ancestral species) for a given species phylogeny; A01, which is the inference of the species tree given the delimited species; A10, which delimits species given a fixed phylogeny; and A11 that is a joint species delimitation and species-tree inference (Yang, 2015). For the A11 analysis, BPP tests different species delimitation models and different species phylogenies, with the individuals fixed to a determined population, i.e., it tries to group different populations into one species and explore the different phylogenetic relationship among them without trying to split different populations into different species (Yang and Rannala, 2014). Theoretically this could be done by assigning each individual to a different population,

but presently the software would not be able to deal with the calculation and the priors currently implemented (0 and 1) are not appropriate for analysis with large number of individuals as they favour a large number of delimited species (Yang and Rannala, 2014). A key feature of the A11 analysis is that the user does not need to know the species tree to delimit the species, which is often the case in recent radiations of non-model organisms. This is possible because the analysis uses a novel approach of the Markov chain Monte Carlo (MCMC) based on the nearest-neighbour interchange (NNI) algorithm to vary the topology of the species tree (Yang and Rannala, 2014). To modify species delimitations the A11 analysis uses the reversible-jump Markov chain Monte Carlo (rjMCMC) (Rannala and Yang, 2013; Yang and Rannala, 2014). The authors developed four priors for models of species delimitation and species phylogeny. Prior 1 gives a higher probability of delimiting a smaller number of species than prior 0 for a large number of populations (Yang and Rannala, 2014). The parameters in the MSC model τs and θs are calculated using sequence distance or expected number of mutations per site (Yang, 2015). Analysis of simulated and empirical data sets suggests that large θs favour fewer species (Yang and Rannala, 2014).

For the analysis in this chapter I ran BPP version 3.4 (March, 2018)[3]. I used the inverse-gamma prior of $\theta \sim$G(14,1000). BPP calculates the inverse-gamma prior as IG(α, β) with the mean to be $\beta/(\alpha - 1)$, thus in this case the mean is equal to 0.014, i.e., 14 substitutions per 1,000 bases. I used the inverse-gamma prior of $\tau \sim$G(2,1000), which corresponds to 0.002 or 0.2% sequence divergence between the root of the species tree and the present time. Those priors fit relatively large population sizes and recent diversification events (Prata et al., 2018), which is appropriate for *Ceiba* given the evidence from Chapter 2, for example the non-monophyly of *Ceiba* species found in the recently diversified South American SDTF clade such as *C. pubiflora* may be a reflection of large effective population sizes and hence a longer time to coalescence (Naciri and Linder, 2015; Pennington and Lavin, 2016).

I conducted the analysis for 1,000,000 MCMC generations, sampling every five iterations and with burn-in of 100,000 generations (Prata et al., 2018) for a combination

[3]https://github.com/bpp/bpp

of algorithms 0 and 1 with priors 0 and 1. Convergence was evaluated by running three independent runs and checking for consistent results (Yang and Rannala, 2014).

Within the South American SDTF group I ran BPP on five clades: (i) a clade comprising 19 accessions currently identified as *C. pubiflora* and 13 accessions of *C. glaziovii*; (ii) a clade comprising clade (i) and its sister clade with two accessions of *C. speciosa*, two of *C. crispiflora* and one of *C. pubiflora*; (iii) a clade comprising 18 accessions, of which two are currently identified as *C. speciosa*, two as *C. lupuna*, two as *C. pubiflora*, three as *C. boliviana*, two as *C. chodatii* and six as *C. insignis*; (iv) a clade comprising clades (ii) and (iii); and (v) finally on a clade including nine accessions of *C. erianthos* and four of *C. rubriflora* (Figure 4.1, which corresponds to a zoom in the South American SDTF clade from Figures 3.31 and 3.36 in Chapter 3). The species in clades i-iv were not resolved as monophyletic on the phylogeny generated with the 377 loci using the concatenated approach. Each of these clades represent a problem of species delimitation because the taxonomic species within them were not resolved as monophyletic and in most cases are morphologically similar. Analyses using larger, more inclusive clades do not make sense in that morphological and genetic differences amongst them are large and they could not be considered as single species. Although *C. erianthos* and *C. rubriflora* were recovered as reciprocally monophyletic, I ran the BPP A11 analysis on them as a control because the accuracy of species delimitation under a MSC model is under debate (Fujita et al., 2012; Carstens et al., 2013; Sukumaran and Knowles, 2017).

4.2.4 Morphological investigation

The clade composed by the currently recognised species *C. pubiflora* and *C. glaziovii* has 32 accessions. These species were not recovered as monophyletic in the phylogeny generated using 377 loci under the concatenated approach (represented by the green circle in Figure 4.1). The separation of *C. pubiflora* and *C. glaziovii* in the field is difficult even when individuals are fertile. The morphological variation in size and colour of flowers, for example, contributes to the unclear boundaries between the species

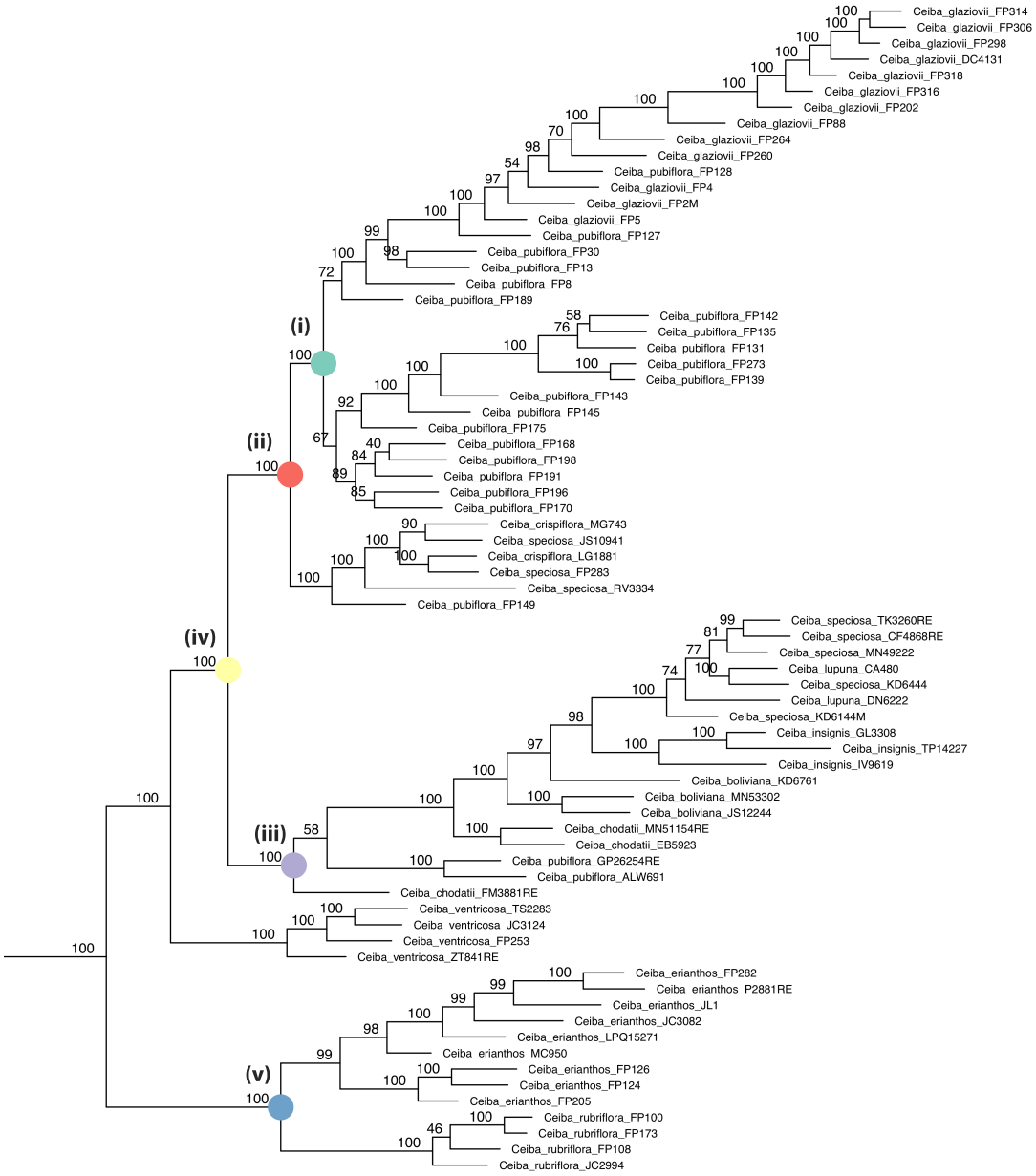


Figure 4.1: SDTF South American clade derived from the ML analysis using 377 loci under the concatenated approach. The five different coloured circles numbered from (i) to (v) represent the different clades where the BPP A11 analysis was run.

(Figure 4.2). The two species have overlapping distributions in the central region of Bahia state in Brazil. In this area, collections of both species are available in herbaria, and often the determination on the label of the specimens has been changed from *C. pubiflora* to *C. glaziovii* and vice-versa by different specialists.

In the recent morphological revision for *Ceiba* (Gibbs and Semir, 2003), the authors separated the species by flower colour, anthesis time, stamen form (re-supinate or spreading) and the presence or absence of hairs in the staminal tube. *Ceiba pubiflora* is also described as having flowers with variable sizes and colours (Gibbs and Semir, 2003).

The complete description of both species from Gibbs and Semir (2003) is as follows:

- *C. pubiflora* (A. St.-Hil.) K. Schum. in Mart, (ed.), Fl. Bras. 12(3): 213 (1886)

  Trees with sometimes ventricose, aculeate trunk. Leaflets somewhat chartaceous, usually serrate. Pedicels 5-10 mm long. Petals 47-85 x 20-25 mm, initially somewhat erect, subsequentially spreading, obovate-oblong, margin somewhat undulate sericeous externally, glabrous internally, uniformly pale pink with sparse dark flecks, or deep pink-lilac with conspicuous carmine striations which may coalesce midlength. Staminal column glabrous, 10-15 mm long; staminal appendages pink-yellowish, glabrous, with five bifid lobes; above the appendages the column divides either immediately, or at c. 5- 10 mm, into 5 usually strongly resupinate,white filaments which have large, sinuous anthers. Stigma white. Fruit a somewhat rotund to ellipsoidal capsule, 10-15x8-10 cm. Flowering February-May. Semi-deciduous woodlands, particularly on calcareous soils. Argentina (Missiones), Paraguay, Centre-West Brazil from Corumbá to NE Minas Gerais, extending to Bahia and Espírito Santo. (...) Flowers rather variable in size, and also include forms ranging from pale pink petal with very few striations, to others flushed dark pink-lilac and with distinct dark, wine-coloured striations which tend to coalesce. *C. pubiflora* has diurnal anthesis. Flowers on trees in Bahia, and also in cultivation in São Paulo, were observed to be frequently visited, and so probably pollinated, by hummingbirds.

- *Ceiba glaziovii* (Kuntze) K. Schum., Just's Bot. Jahresber. 26: 343 (1900)

  Trees usually 10-15 m, with swollen, aculeate trunk. Leaves 4-7 foliolate, petioles 60- 90 mm long; leaflets 50-130 x 30-60 mm, chartaceous, elliptic-oblanceolate, denticulate, especially distally, acuminate, glabrous, with petiolules 2-4 mm long. Flowers in fascicles of three or solitary. Pedicels 6-10 mm long. Calyx 20-28mm, campanulate, glabrous, with 3-5 lobes. Petals c. 65 x 25 mm, spathulate, spreading, white, externally villous, internally glabrous basally, hairy distally,sometimes with magenta striations towards the base. Staminal tube 10-50 mm, 5 entire appendages, all covered with dense hairs; tube continuing above the appendages for 3-20 mm and then dividing into 5 spreading, white filaments which terminate in yellow, sinuous anthers. Ovary subglobose, with the slender style terminating in a white globose stigma a little above the anthers. Fruit elongate to ellipsoidal capsule, c. 8-12x5-9 cm. Flowering July-September. Dry woodlands (Caatinga). NE Brazil (Bahia, Pernambuco, Paraíba, Ceará).

The key for the species of *Ceiba* provided in the revision (Gibbs and Semir, 2003) separates both species as follows:

- Petals pale pinkish, or pink-lilac, distally with sparse to marked dark carmine coloured striations which may fuse below; stamens re-supinate; with diurnal flowering: *Ceiba pubiflora*

- Petals white distally, dark livid towards the base internally; stamens spreading; with nocturnal flowering: *Ceiba glaziovii*

In summary, Gibbs and Semir (2003) differentiate species principally on flower colour and the form of the stamens. Additionally, they use the presence or absence of hairs on the staminal tube. I conducted two expeditions to the Caatinga that allowed me to collect samples and observe the flower colour difference and indumentum in the staminal tube. In addition, in the herbarium, I investigated flower size, indicated to be especially variable in *C. pubiflora* by Gibbs and Semir (2003). I also noted that the leaf size varied across *C. pubiflora* and *C. glaziovii*, so I measured that trait in the herbarium for 142 samples for both species (26 *C. glaziovii* and 115 *C. pubiflora*)

across the eastern distribution of the species, i.e., I did not include samples identified as *C. pubiflora* from Paraguay or Corumbá area in the state of Mato Grosso do Sul in Brazil since the samples I sequenced from those areas were recovered in a different clade (samples ALW691 and GP26254RE - Figure 4.1).

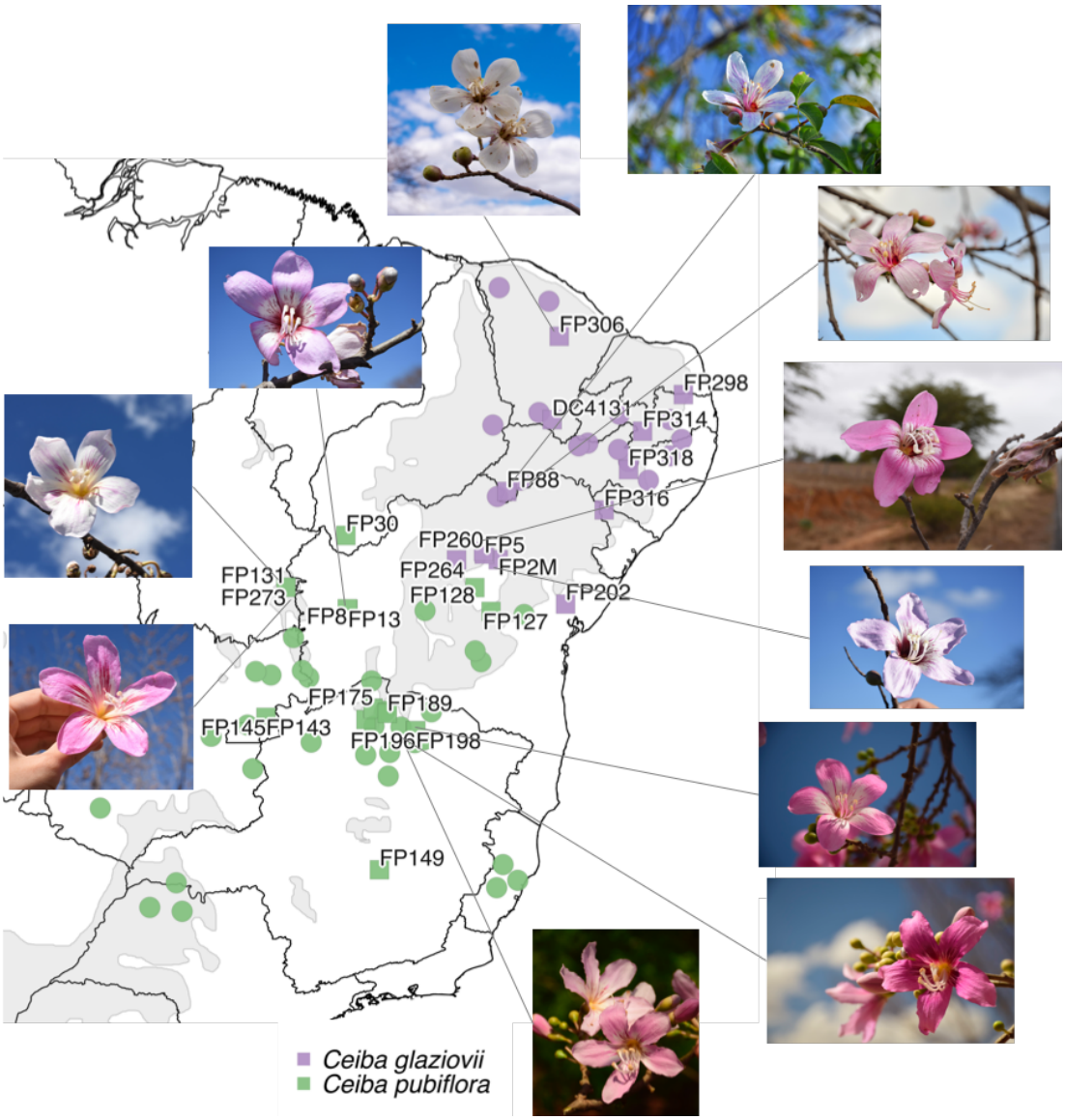


Figure 4.2: Morphological variation in samples *Ceiba pubiflora* (green dots) and *Ceiba glaziovii* (purple dots). Squares represent samples sequenced in this study and circles represent register of occurence according to Gibbs and Semir (2003). Photos: F. Pezzini.

4.2.5 Species tree inference

Species tree estimation followed two MSC models using two different software packages. Astral (version 5.6.3) tries to find the tree that maximizes the number of induced quartet trees in individual gene trees that are shared by the species tree (Mirarab and Warnow, 2015). The induced quartet trees are scored as the percentage of the quartets trees from the individual gene trees that are in the species tree. Therefore, the higher the value, less variation between the gene tree and the species tree. Astral can be run in a multi-individual data set by providing a mapping file assigning each individual to its respective species. Astral will then force species to be monophyletic. I ran the software twice, first including each accession as an independent sample (Astral-ind) and a second run assigning each accession to a species name (Astral-multi). The individual gene trees were inferred using RAxML (version 8.0) (Stamatakis, 2014) with 100 rapid bootstrap replicates for the filtered data set. Inilialy, I attempt to use BPP (described above) to run the A01 analysis that infers the species tree. However, after two weeks, the first run of the analysis was not finished, and was discontinued. Both software packages are considered accurate methods when inferring species trees, especially for recent radiations with high levels of ILS (Simmons and Gatesy, 2015; Shi and Yang, 2018).

4.3 Results

4.3.1 Gene tree discordance

Gene and site concordance factors (gCF and sCF respectively) are shown in Figure 4.3. Ideally a data set would have a similar value of sCF and gCF[45]. When analysing the 377 loci, *Ceiba* had a higher value of sCF than gCF indicating natural causes of discordance such as ILS might not be solely responsible for the conflict. The absence of phylogenetic signal in individual gene trees might be contributing to the differences in values observed between sCF and gCF. The fact that some of the branches show zero

[4]http://www.robertlanfear.com/blog/files/concordance_factors.html
[5]http://www.iqtree.org/doc/Concordance-Factor

values for gCF but still a percentage of sCF suggests that the variant sites supporting a given branch in the concatenated tree are found in just a few loci.

The PhyParts analysis showed a high level of incongruence between the individual gene trees (Figure 4.4). Both analyses showed similar patterns regarding the branches showing high and low gene concordance. The monophyletic species *C. trischistandra*, *C. pentandra*, *C. jasminodora*, *C. samauma*, *Neobuchia paulinae* and *C. schottii* had a high percentage of gene concordance, in contrast to the branches containing the species within the South American SDTF clade. Among the species recovered as monophyletic, *Ceiba trischistandra* was the only that had high concordance among gene trees for its branch. The high percentage of conflict among the gene trees in the branch of *C. pentandra* and the branch of *C. jasminodora* reinforces the uncertainty in the placement of those species.

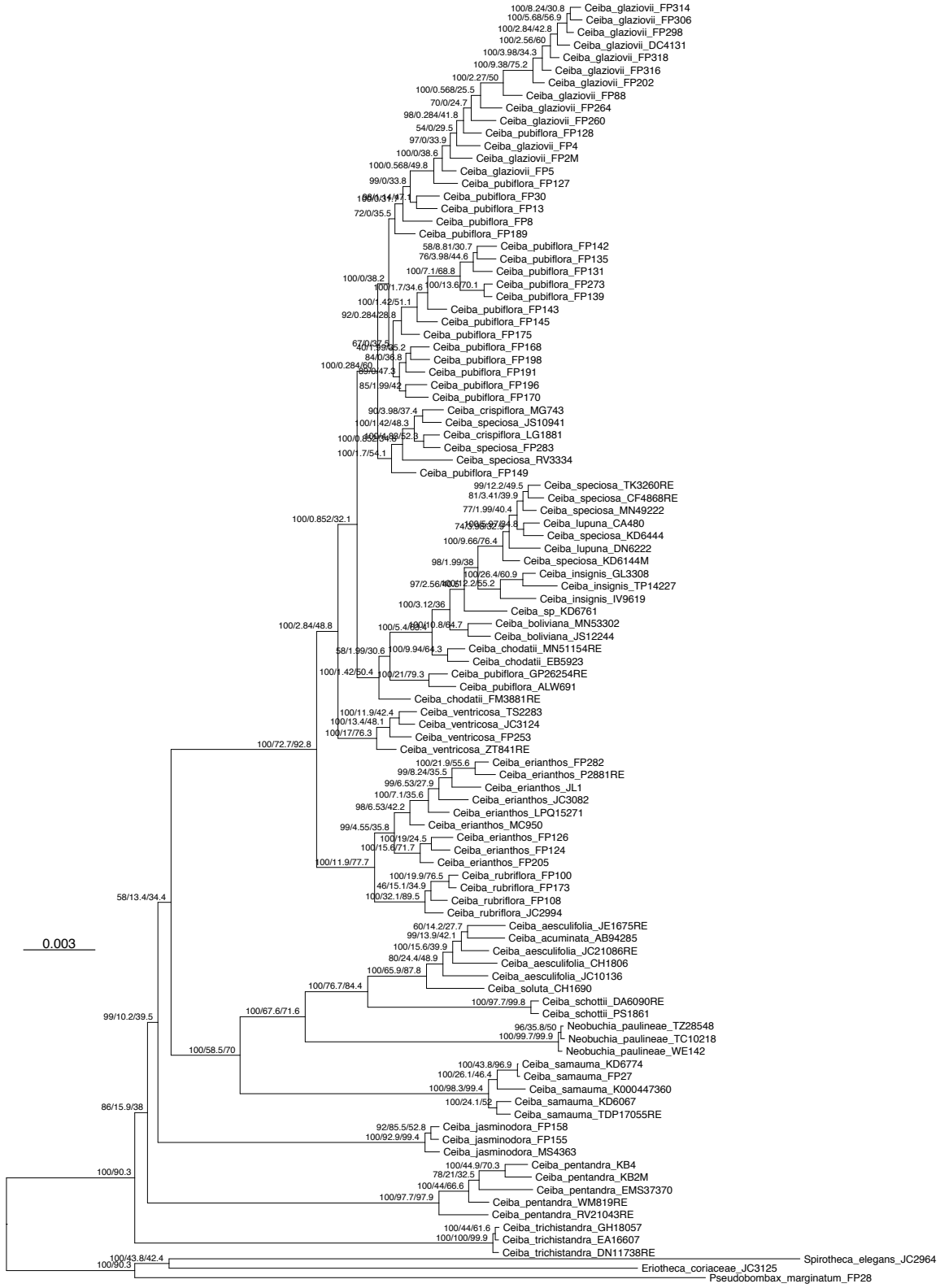


Figure 4.3: Gene and site concordance factors using IQ-Tree plotted on the phylogeny inferred from the concatenated matrix with 377 loci. Each branch is labeled with three numbers: the bootstrap value from the concatenation analysis/gene concordance factor/site concordance factor.

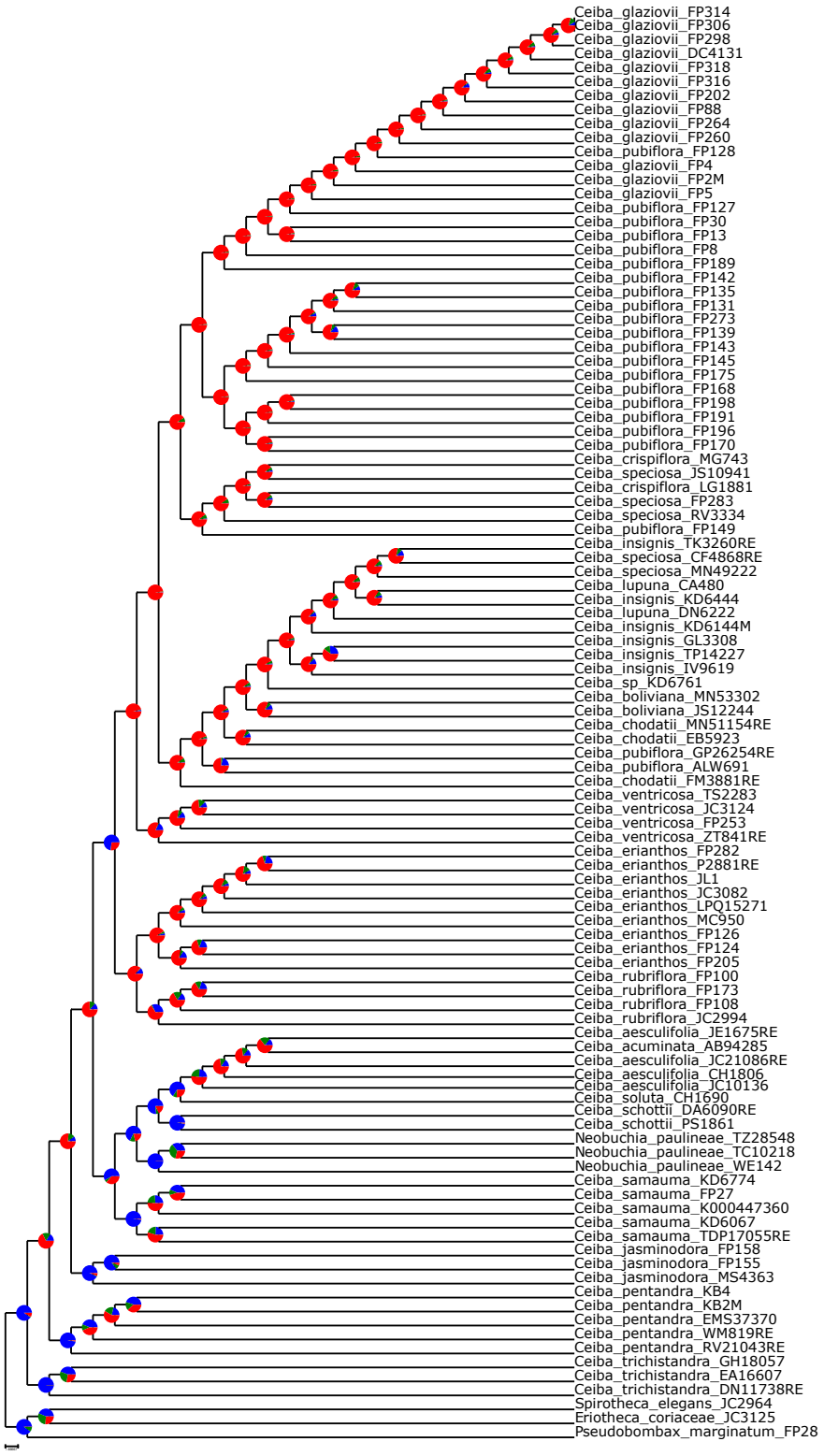


Figure 4.4: Gene tree discordance using PhyParts. The pie charts represent the proportion of gene trees that (i) support the shown topology (blue), (ii) conflict with the shown topology (most common conflicting bipartion) (green); (iii) conflict with the shown topology (all other supported conflicting bipartitions) (red); (iv) have no support for conflicting bipartion (grey).

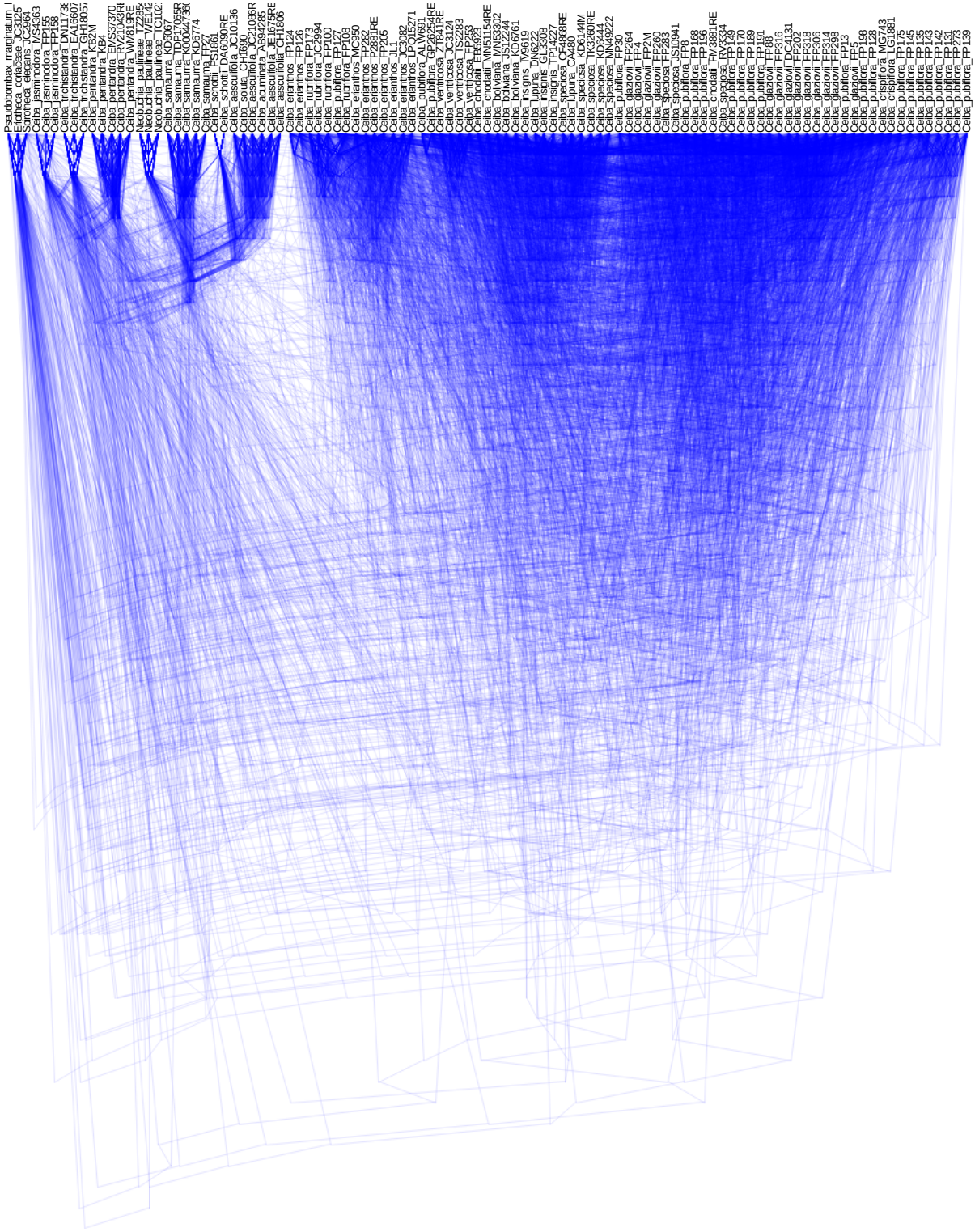


Figure 4.5: Individual gene trees plotted together.

4.3.2 Loci filtering

Root-to-tip variation ranged from 0.000000918 to 0.000258107. Tree length varied from 0.0584237 to 0.381305, and bipartition concordance from 0.059405941 to 0.346534653465 (Figure 4.6). I included in the filtered data all loci that followed all these criteria: root-to-tip variation below 0.000036; tree length above 0.15 and bipartition concordance above 0.15. The first two values were selected based on the minimum amount of remaining loci that provided enough information to run the analysis (breaks indicated by the blue vertical lines in Figure 4.6). I chose a low bipartition concordance threshold because I am also interested in using the filtered data set to infer the species tree. The final data set included 111 loci. This data set was used as input for species delimitation, species tree inference and molecular dating (Chapter 5).

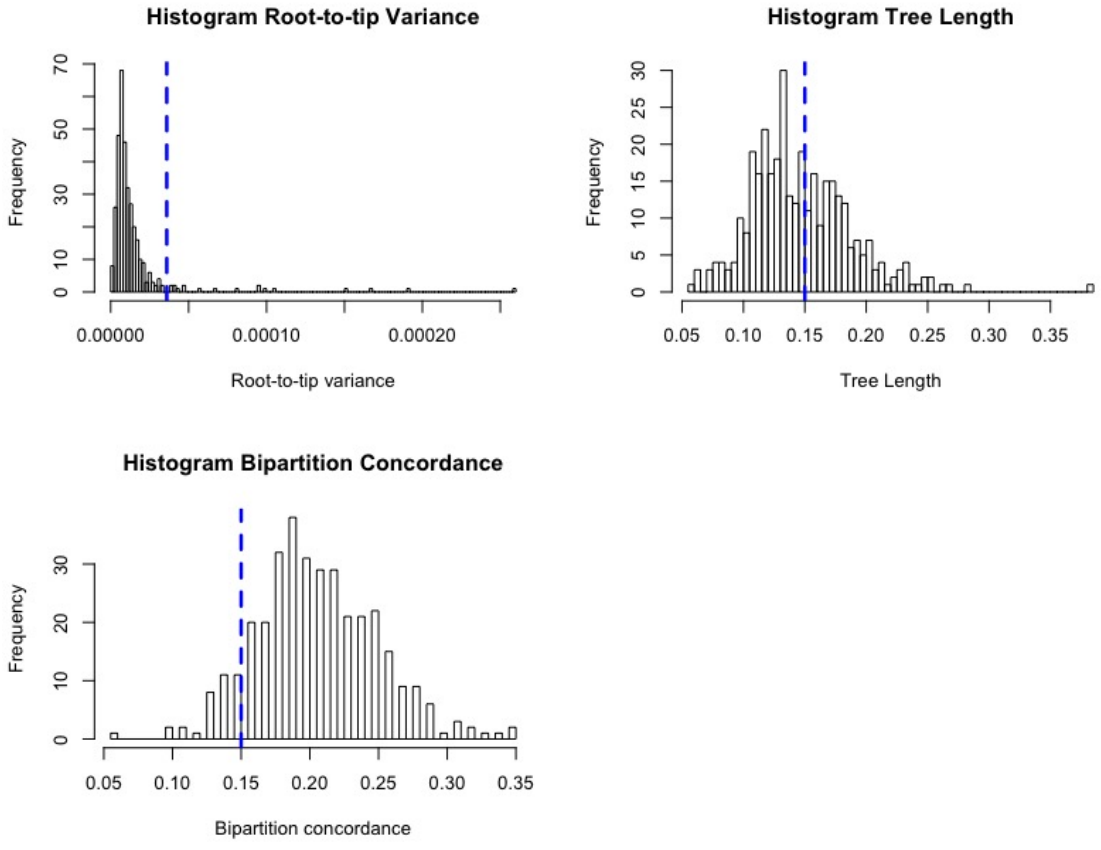


Figure 4.6: Root to tip variance, tree length and bipartition concordance with species tree output for 352 loci analysed using the SortaDate package. Dashed blue lines indicate threshold values for each filter.

4.3.3 Species delimitation

The BPP A11 analysis for *C. erianthos* and *C. rubriflora* for 111 loci indicated that they represent two species with high posterior probability (PP = 1) for all runs (node (v) in Figure 4.1). Two out of the three runs for *C. pubiflora* and *C. glaziovii* indicated that they represent a single species with posterior probability of 1.0 (node (i) in Figure 4.1). For the group including *C. speciosa*, *C. lupuna*, *C. insignis*, *C. boliviana*, *C. chodatii* and the two accessions of *C. pubiflora* from western Brazil and Paraguay (node (iii) in Figure 4.1) each of the three runs recovered a different output. The first run delimited four species with 0.99 posterior probability: *C. speciosa* + *C.lupuna*, *C. boliviana* + *C. chodatii*, *C. insignis* and *C. pubiflora*. Likewise, the second run delimited four species, delimited as *C. speciosa* + *C.lupuna*, *C. insignis*, *C. boliviana*, and *C. pubiflora* + *C. chodatii*. In the clade including *C. pubiflora*, *C. glaziovii* and the eastern accessions of *C. speciosa* and *C. crispiflora*, three species were delimited in all three runs with 0.99 posterior probability, structured as *C. pubiflora*, *C. glaziovii* and *C. speciosa* + *C. crispiflora* (node (ii) in Figure 4.1). The analysis including clades (i), (ii) and (iii) in Figure 4.1 recovered eight species in all runs with posterior probability of 1.0, therefore it identified the same eight species included as the input as separated species.

4.3.4 Morphological investigation

The morphological investigation I conducted for *C. pubiflora* and *C. glaziovii* showed that there was no difference in flower size, leaf length and leaf width between them. Both species showed overlapping values for the three measurements with no clear separation (Figures 4.7 to 4.9).

Individuals from the north of Caatinga have predominantly white flowers and individuals in the center and north of Minas Gerais state have predominantly pink flowers. However, individuals in the central region of Bahia state and southeast of Tocantins state have both pink and white flowers (Figure 4.2). The presence or absence of hairs in the staminal tube does not seem a consistent character from field observations.

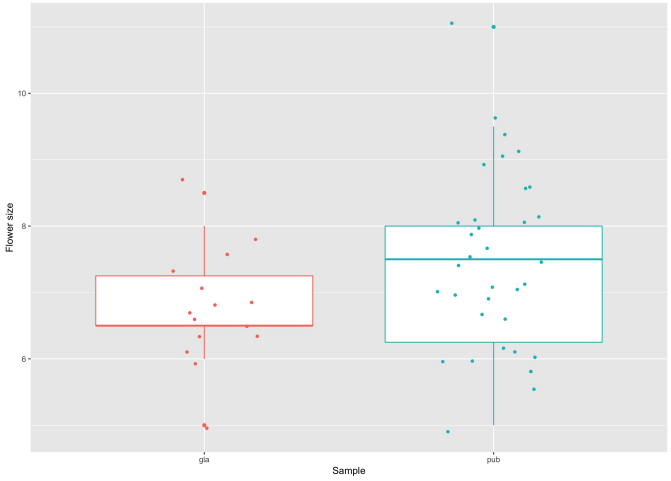


Figure 4.7: Flower size of specimens of *Ceiba glaziovii* (red) and *Ceiba pubiflora* (green).

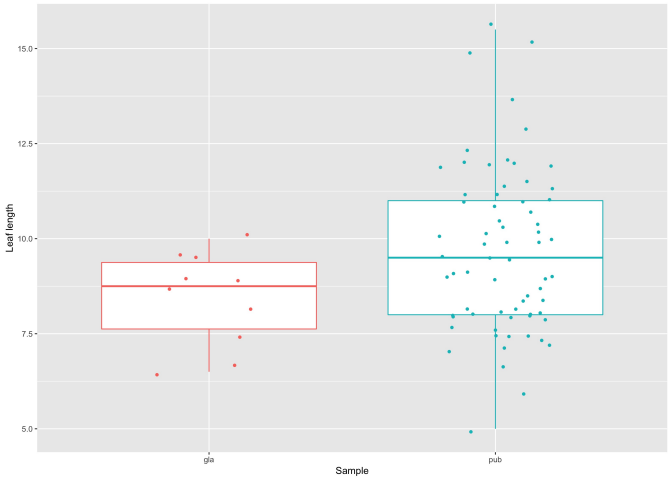


Figure 4.8: Leaf length of specimens of *Ceiba glaziovii* (red) and *Ceiba pubiflora* (green).

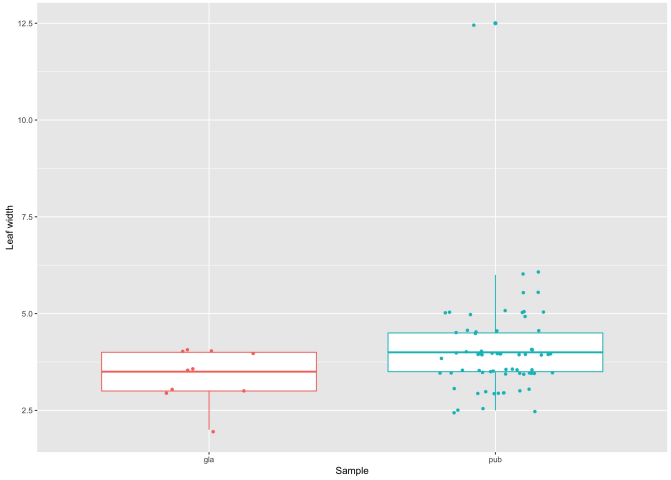


Figure 4.9: Leaf width of specimens of *Ceiba glaziovii* (red) and *Ceiba pubiflora* (green).

4.3.5 Species tree inference

In the Astral analysis including each accession as an independent sample (Astral-ind), *Ceiba* was resolved in six main clades (Figure 4.10): (i) *C. trischistandra*, the only SDTF species occurring west of the Andes; sister to (ii) *C. pentandra*, a rain forest species occurring across the Neotropics and reaching Africa; (iii) a Central American and Mexican SDTF clade including three species of *Ceiba* and *Neobuchia paulinae*; sister to (iv) *C. samauma*, a widespread rain and semideciduous forest species from South America; (v) *C. jasminodora*, the only species occuring within the cerrado biome; and (vi) a South American SDTF clade including 11 species. These clades had high support values, but the relationships amongst them were less well supported.

Within the South American SDTF clade, seven of the 11 species were not resolved as monophyletic. The species were resolved as follows: (i) *C. rubriflora* and (ii) *C. erianthos*, sister to each other and recovered as monophyletic; (iii) *C. ventricosa*, recovered as monophyletic sister to (iv) a clade including *C. pubiflora* and *C. glaziovii*; (v) a clade comprising *C. speciosa*, *C. crispiflora* and one accession of *C. pubiflora* from east Brazil; (vi) two individuals of *C. pubiflora* from Paraguay and west Brazil sister to a clade comprising (vii) a single individual of *C. chodatii* from Paraguay and (viii) a clade containing accessions of *C. insignis*, *C. lupuna*, *C.speciosa*, *C. boliviana* and two individuals of *C. chodatii*, all from west South America.

Overall, the branch support values were high for the monophyletic species and low (< 0.7) for the branches representing relationships between species. In a multi-locus analysis, branch support values represent the local posterior probability and measure the support for a quadripartition (the four clusters around a branch) (Sayyari and Mirarab, 2016) and not the bipartition, as in bootstrap.

Similar to the ITS phylogeny (Chapter 2) and the concatenation phylogeny (Chapter 3), the Astral-ind coalescent analysis supports the monophyly of the two sections of the genus, Ceiba and Campylanthera, and does not support monophyly of the 'insignis' morphological complex.

The species tree generated with Astral (Astral-multi) was fully resolved, although

with low support values for some nodes (Figure 4.11). *C. trischistandra* was recovered as sister to *C. pentandra* (0.91 posterior probability). *Ceiba soluta*, *C. aesculifolia*, *C. schottii*, *Neobuchia paulinae*, and *C. samauma* were recovered in one clade fully resolved with posterior probability of 1 for each node. Within the South American SDTF clade, *C. erianthos* and *C. rubriflora* were recovered as sister to each other with posterior probability of 1. A clade comprising *C. lupuna*, *C. speciosa*, *C. chodatii*, *C. insignis* and *C. boliviana* was resolved with posterior probability of 1, but values lower than 0.9 for the inner nodes. The remaining species recovered within this clade also had low posterior probabilities.

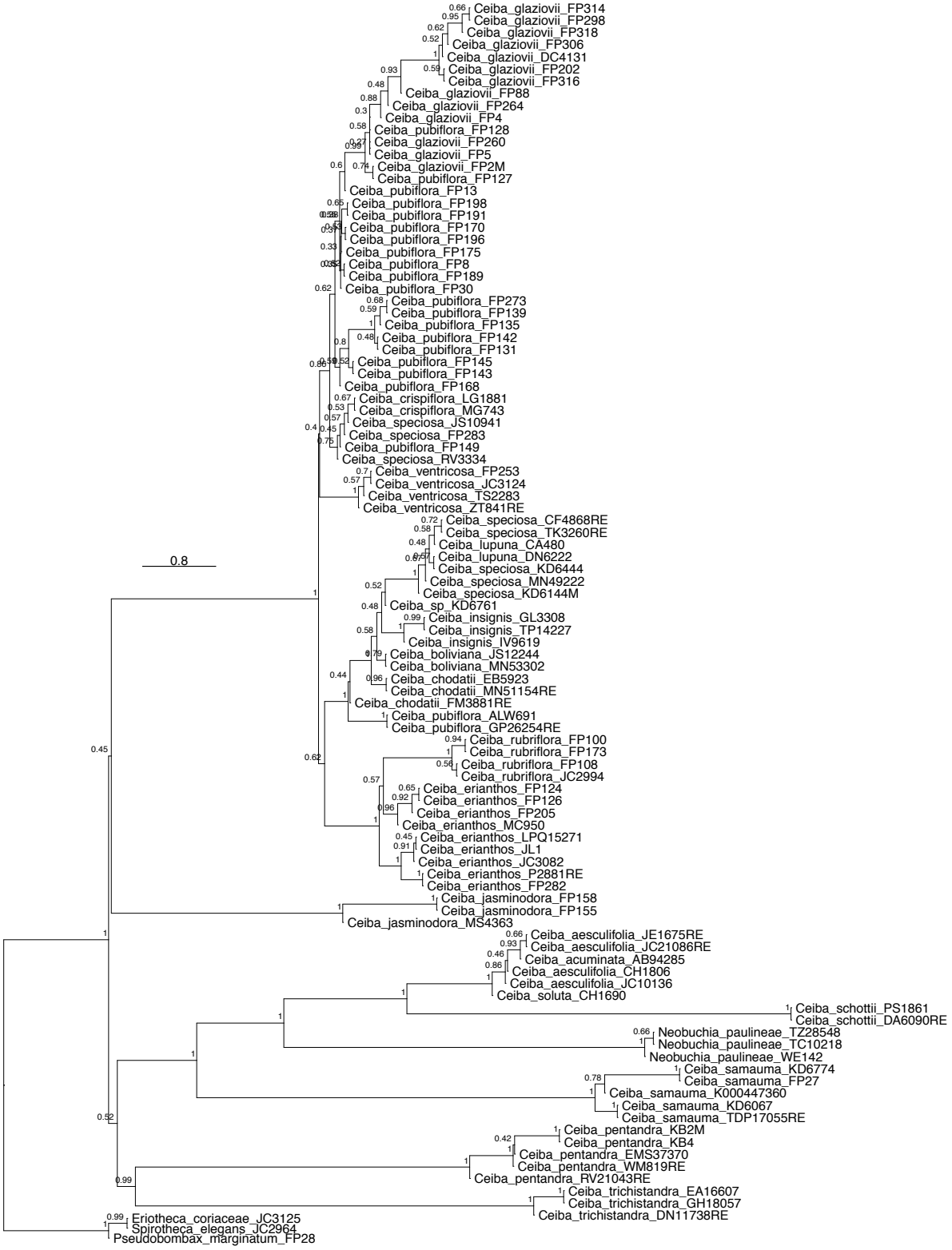


Figure 4.10: Phylogenetic tree inferred by the Astral-ind analysis with 111 loci.

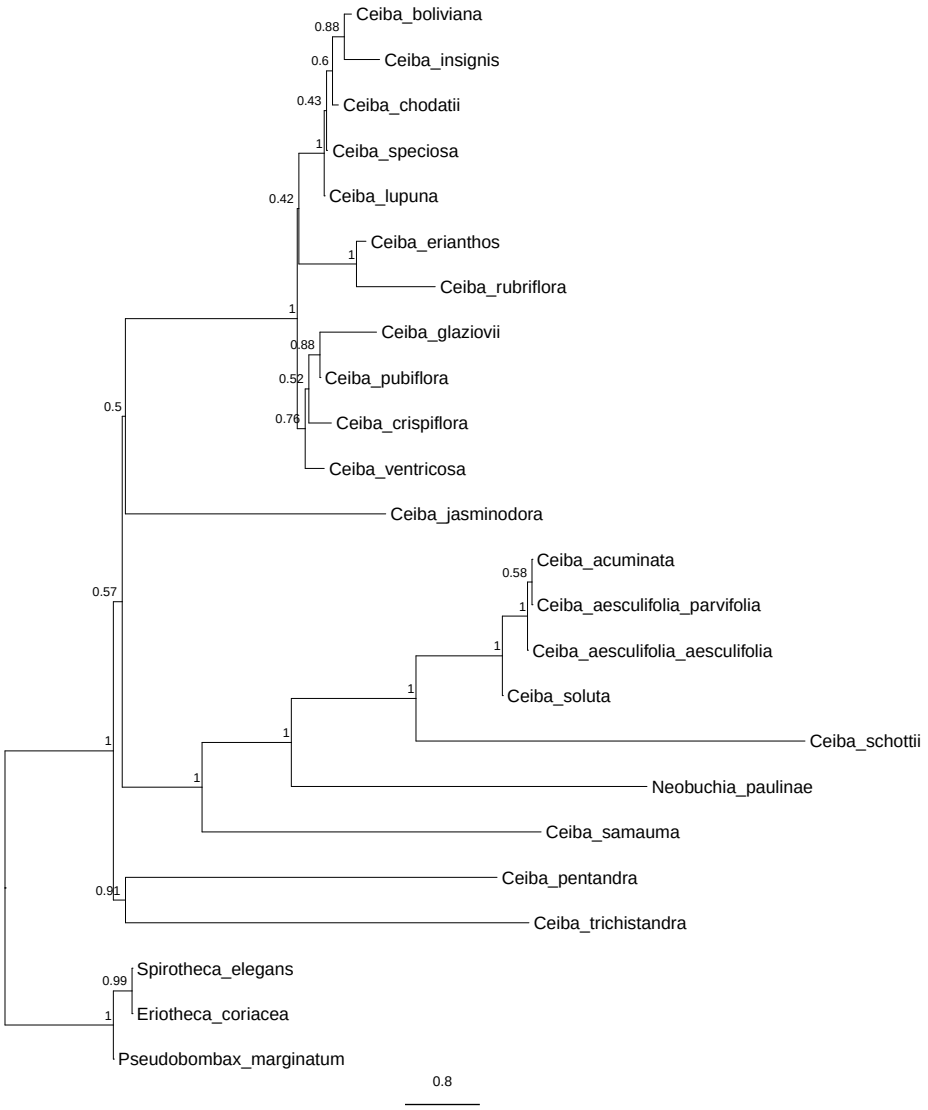


Figure 4.11: Species tree inferred by Astral-multi analysis with 111 loci.

4.4 Discussion

Incongruence among individual gene trees

The analysis of individual gene tree concordance showed that there is a high level of genealogy discordance for the 377 independent nuclear loci recovered for *Ceiba*. A high percentage of discordance is present in the branches representing the relationships among *C. jasminodora*, *C. pentandra*, the SDTF South American clade, and a clade comprising the SDTF Central American and Mexican species, *C. samauma* and *Neobuchia paulinae*. It is interesting to note that there was variation in topology amongst the same nodes in the different pipelines and different inputs reported in Chapter 3 (Figures 3.31 to 3.35).

It has been shown that a phylogeny inferred for a multi-locus data set using the concatenation approach in the presence of high levels of discordance amongst individual gene trees can be statistically inconsistent, and with high bootstrap support values for the incorrect tree (Kubatko and Degnan, 2007; Roch and Steel, 2015). Bootstrap values can be misleading in big data sets because they are sampled from the variance and with large amounts of data, the maximum value is reached faster (Felsenstein, 1985). However, Lambert et al. (2015) showed that short and weakly supported branches of phylogenies generated using concatenation were more likely to conflict with the species tree in the scincid lizard family. For *Ceiba*, the branches showing high levels of discordance were also long and well supported in the concatenation analysis. This is particularly evident in the gene concordance factor analysis implemented in IQTree where branches showing 100% bootstrap often showed less than 20% gene concordance.

The exception to this behaviour is *C. trischistandra* that was recovered as monophyletic, with a long branch and high bootstrap support value, no individual gene tree discordance, and as sister to the rest of the genus in most phylogenies. However, it was recovered with a different placement in one of five combinations of input data and pipelines tested in Chapter 3 (Figure 3.35) and as sister to *C. pentandra* in both Astral analyses (Figures 4.10 and 4.11). Therefore, even with high individual gene tree concordance, *C. trischistandra* cannot be placed with confidence as the sister species

to the rest of the genus.

The nature of the non-monophyletic *Ceiba* species in the South American STDF clade

Within the South American SDTF clade, seven out of the 11 species were not recovered as monophyletic in the concatenation analysis. Among them, *Ceiba pubiflora* and *C. glaziovii* had the best sampling in this study, represented by 32 accessions in the phylogeny. The topology inferred using the concatenation approach had the 13 individuals of *C. glaziovii* nested within the 19 individuals of *C. pubiflora* (Figure 4.1). The morphological investigation showed that there was no clear difference between the two species for the traits measured (Figures 4.7 to 4.9). The species delimitation analysis supported the recognition of one species in two of the three independent runs conducted. Together, this evidence suggests that *C. pubiflora* and *C. glaziovii* are the same species.

The majority of species delimitation studies have been done using morphological characters. Later, phylogenies built including multiple individuals per species have been used to test the morphologically delimited species. Initially, researchers used a few loci (nuclear, chloroplast and mitochondrial) to test whether the morphologically recognised species were recovered as monophyletic. Wiens and Penkrot (2002), for example, compared morphological species delimitation, with phylogenies constructed using mtDNA sequences, and mtDNA plus morphology combined, for the spiny lizards *Sceloporus*. All three methodologies disagreed with each other when delimiting the five species. Several other studies followed investigating conflicting species delimitation using DNA sequences and morphology (eg. Dexter et al. 2010). More recently, the amount of data generated with NGS techniques provided new hope for species delimitation (Leaché and Fujita, 2010; Leaché et al., 2014), especially under the multispecies coalescent model (Fujita et al., 2012). Those methods have improved the accuracy of species delimitation using multi-locus data, but still are under debate as to whether they split populations or species (Sukumaran and Knowles, 2017; Leaché et al., 2019).

Despite criticism, all agree that species delimitation should be conservative and rely on multiple interdisciplinary methods (Dexter et al., 2010; Carstens et al., 2013; Freudenstein et al., 2016; Sukumaran and Knowles, 2017; Leaché et al., 2019). Morphological and ecological information, for example, are important components of an integrative and robust taxonomy (Camargo et al., 2012).

However, many of these methods expect species to be monophyletic. The topology from the phylogenies generated both with the concatenation (Figure 4.1) and the coalescent (Figure 4.10) approaches showing *C. pubiflora* and *C. glaziovii* as paraphyletic, their lack of morphological separation, and the high level of incongruence amongst individual gene trees in the SDTF South American clade (Figures 4.3 to 4.5), confirm the importance of complementary methods when investigating whether species should be regarded as monophyletic. *Ceiba pubiflora* and *C. glaziovii* have broad parapatric distribution (Figure 4.2), and large effective population sizes. The large population sizes mean a longer time to achieve monophyly (Rieseberg and Brouillet, 1994; Hudson and Coyne, 2002). As a consequence, lineages are not completely sorted with the persistence of ancestral polymorphism as indicated by the conflicting individual gene trees. In contrast, *Ceiba erianthos* and *C. rubriflora* were recovered as monophyletic in the concatenation analysis (Figure 4.1) and as monophyletic in the coalescent analysis, although accessions from southeast of Brazil were grouped in a separate clade from the accessions from the northeast of Brazil (Figure 4.10). *Ceiba erianthos* occurs on granite outcrops though the southeast and east of Brazil (Gibbs and Semir, 2003) and *Ceiba rubriflora* on limestone outcrops in the north of Minas Gerais and south of Bahia states (Figure 1.4). The restricted distribution in different ecological niches suggest a case of allopatric speciation and the small effective population size of both species will result in a shorter time to coalescence and sorted lineages.

Simmons and Gatesy (2015) recommend care when dealing with new methods: "We conclude that enthusiastic application of novel tools is not a substitute for rigorous application of first principles, and that trending methods (e.g., shortcut coalescent methods applied to ancient divergences, tree-independent character subsampling), may

be novel sources of previously under-appreciated, systematic errors.". *Ceiba pubiflora* and *C. glaziovii* had extensive sampling of individuals (32 accessions across the geographical distribution) and loci (111), and the data set was analysed with numerous methods (coalescent-based species delimitation analysis, morphological data and field observations), all of which indicated that they are the same species. Whilst monophyly or reciprocal monophyly is not required to define a species, species are entities in the course of change that eventually would achieve reciprocal monophyly. When delimiting species we ask ourselves how much morphological variation is enough? Or how much genetic variation is enough? Although it can be argued that those are arbitrary decisions, the evidence from more than one source is compelling in showing that *C. pubiflora* and *C. glaziovii* are towards the similarity end of that continuum and are best recognised as one species.

The species tree for *Ceiba*

The species tree inferred for *Ceiba* showed low local posterior probability support values especially within the South American SDTF clade. This level of incongruence of the individual gene trees (Figure 4.5) combined with a large volume of data might be too difficult to deal with for the Astral algorithm. In addition, Astral assumes that ILS is the only source of disagreement amongst gene trees. It has been suggested that *Ceiba* species can hybridise (Gibbs and Semir, 2003), although based on evidence from cultivated individuals. Distinguishing hybridisation, or the effects of introgression, from ILS is not straightforward and cannot be excluded as a possible cause of the low support values observed for the species tree of *Ceiba*. However, *C. pentandra*, *C. trischistandra*, *C. samauma*, *Neobuchia paulinae*, *C. schottii*, *C. jasminodora*, *C. ventricosa*, *C. erianthos* and *C. rubriflora* were recovered as monophyletic in all analyses conducted (Figures 3.31 and 4.10).

The Central American species *C. soluta* was nested within *C. aesculifolia* or resolved as sister to the sympatric Mexican and Meso-American species. *Ceiba soluta* was the only species represented by one accession in the phylogeny. This species is also the only

one with 10 to 15 free staminal filaments. Until 1992, *C. soluta* was known from the type represented by flowers collected from the ground (Gibbs and Semir, 2003). Gibbs and Semir (2003) considered synonymising *C. soluta* and *C. aesculifolia* due to the morphological similarity of both species and their overlapping distribution. However, based upon a new collection of *C. soluta* (sample CH1690 sequenced in this study) made in Huehuetenango, Guatemala, in 1992, they decided to maintain *C. soluta* as a separate species. The current phylogeny for the genus is not conclusive about the monophyly or the placement of this species and more collections would be ideal to shed light on its phylogenetic relationship with other *Ceiba* species.

The phylogeny shows geographical structure, with the Central American and Mexican species recovered in one clade sister to a clade containing the South American species. *Ceiba samauma* is an exception. This species occurs in semi-deciduous forests in South America and was recovered as sister to the Central American and Mexican clade. Overall, the phylogeny shows ecological structure, with species ocurring in the Central American and Mexican SDTF recovered in the same clade, sister to the South American SDTF clade that comprises predominantly dry forest species. The wet forest species *C. speciosa*, *C. crispiflora* and *C. lupuna* were nested within the South American SDTF clade. However, it is important to note that *C. speciosa* shows great ecological tolerance, being able to occupy wet and dry areas (Gibbs and Semir, 2003). The phylogenies inferred using ITS in Chapter 2 and using the concatenation approach in Chapter 3 have a similar pattern of geographical and ecological structure. Therefore, closely related species in *Ceiba* are likely to show the same ecological preference or geographical occurrence. Those biogeographical and ecological patterns will be discussed in detail in Chapter 5.

Methodological considerations

The new version of Astral used to conduct the coalescent analysis (Astral-III, v.5.3.6) (Rabiee et al., 2019) implements an improved algorithm capable of dealing with multi-allele data sets. The analysis is run with a mapping file that assigns individuals to

species and estimates the species tree under the Multi-species Coalescent model considering ILS as the main source of individual gene tree disagreement. However, the authors assume that species are correctly delimited (Rabiee et al., 2019), although recognising that this is not straightforward, especially within recent radiations. Like Rabiee et al. (2019), I ran this analysis twice, assigning individuals to species (Astral-multi) and considering each individual as a separate species (Astral-ind). The authors argue that the Astral-ind analysis can be used to assess whether delimiting species increases the accuracy of the Astral-multi analysis. However they consider that an improvement in accuracy is achieved when species are recovered as monophyletic and high support values (local posterior probability) in branches of non-monophyletic species reflects poor species delimitation (Rabiee et al., 2019). Most species in the SDTF South American clade were not recovered as monophyletic in the Astral-ind analysis (Figure 4.10). However, interpreting this pattern as a poor species delimitation misunderstands that species might not be monophyletic (Rieseberg and Brouillet, 1994; Knowles and Carstens, 2007).

Other factors can cause individual gene tree disagreement, such as data contamination, gene duplication, horizontal gene transfer and hybridisation. The first three can be minimised during genome assembly, for example. However, hybridisation and ILS can cause the same pattern and distinguishing between them is not straightforward. Smith et al. (2015) were able to identify conflict among gene tree and species tree for two different data sets representing Caryophyllales and Hymenoptera. They could not, however, conclude whether the source of conflict was due to ILS or absence of phylogenetic signal (i.e. informative characters). The fact that Astral considers only ILS as a source of conflict should be taken into consideration when interpreting the output.

Future work needed in species delimitation in *Ceiba*

Ceiba boliviana, *C. chodatii*, *C. speciosa*, *C. crispiflora* and *C. lupuna* were not resolved as monophyletic in the SDTF South American clade and the species delimitation analysis was not conclusive. More field samples are needed for morphological

and genetic analysis to duplicate the detailed investigation conducted for *C. pubiflora* and *C. glaziovii*. In this section, I highlight some priorities for future work in species delimitation in *Ceiba*.

Ceiba boliviana and *Ceiba chodatii*

Ceiba chodatii and *C. boliviana* are not as widely distributed as *C. pubiflora* and *C. glaziovii*, but still are distributed over large areas of SDTF in Peru, Bolivia, Paraguay and Argentina, with overlapping occurrence in the south of Bolivia and north of Argentina (Figure 1.4). Each species was represented in the phylogeny by three accessions (*C. boliviana*: MN53302, JS12244 and KD6761; *C. chodatii*: MN51154RE, EB5923 and FM3881RE), but neither species is resolved as monophyletic. The species are morphologically distinct based upon my study of herbarium specimens and according to field researchers (e.g. notes on specimen Michael H. Nee - 53186 - *Ceiba chodatii* - determined by D.A. Neill (MO), 2009). Based upon this morphological distinctness, sampling more individuals of each species across their ranges may resolve them as monophyletic in future phylogenies.

Ceiba crispiflora, *Ceiba lupuna* and *Ceiba speciosa*

Ceiba lupuna and *C. crispiflora* were not recovered as monophyletic in the NGS phylogeny, with accessions nested in different positions within *C. speciosa* in two separate clades (Figure 4.1). The lack of monophyly (non-coalescence) of these three wet-inhabiting species initially suggests large effective population sizes and low degree of dispersal limitation.

Ceiba crispiflora has a narrow distribution in the state of Rio de Janeiro, Brazil, and is partially sympatric with *C. speciosa*. Gibbs and Semir (2003) separated both species by the narrower and markedly undulate margined petals in *C. crispiflora*. The species delimitation analysis recovered both as the same species and, although more samples are needed for taxonomical and genetic analysis, the results here point towards the combination of both species.

Ceiba lupuna is the only species with distribution known to be restricted to rain forests (Gibbs and Semir, 2003). As in the NGS phylogeny (Figure 4.1), in the Sanger-sequenced phylogeny from Chapter 2, the accession currently identified as *C. lupuna* (CF6838) was nested within the South American SDTF clade. Further investigation of this specimen revealed distinct determinations in the duplicates found in different herbaria, being identified as *C. lupuna*, but also as *C. insignis*. *C. lupuna* is sympatric with *C. speciosa*, differing mainly with respect to flower colour according to Gibbs and Semir (2003). It was further hypothesized that soil fertility could account for ecological differences between the two species (Gibbs and Semir, 2003). *Ceiba speciosa* is found in the same type of environments, as well as in wet forests, although this species is broadly cultivated as ornamental and natural occurring populations are somewhat difficult to determine.

The flowers of *Ceiba* show colour variation even within the same individual. For example, a recently collected specimen from Reserva Florestal do Humaitá, Acre, Brazil, had flowers with petals light pink with white stripes towards the base and dark pink flowers with yellow stripes towards the base (M. Acosta, per. comm.). Similar colour variation is reported for other species of *Ceiba*, for example in a specimen collected by M. Nee (Michael H. Nee - 52932 - *Ceiba boliviana*) who describes variation from creamy yellow to dark pinkish.

Gibbs and Semir (2003) state that *C. speciosa*, *C. lupuna* and *C. crispiflora* are morphologically similar. Although the species delimitation analysis was inconclusive, the two separated clades of accessions of *C. speciosa* that are mixed with accessions of *C. crispiflora* and *C. lupuna* may represent two different species occurring east and west of the Cerrado biome. This hypothesis can be tested by sequencing more accessions and collecting morphological data on herbarium specimens. Given that the type of *C. speciosa* was collected in 1827 in Minas Gerais, Brazil, and *C. crispiflora* in 1822, it is likely that the eastern populations will retain this name, whilst the western populations will be named *C. speciosa*.

Conclusions

In this chapter, I investigated the possible reasons for the variation in topology of the phylogenetic trees generated in Chapter 3 by exploring possible incongruence amongst the 377 individual gene trees sequenced with the target capture technique. I also investigated the lack of monophyly (non-coalescence) of the species in the South American SDTF clade (Chapters 2 and 3). In the South American SDTF clade, there were multiple examples where a monophyletic group recognised as a taxonomic species was nested within another, paraphyletic taxonomic species, which suggests recent, ancestor-descendent species relationships. I combined modern species delimitation analysis and morphological data under a coalescent framework to investigate the boundaries between *C. pubiflora* and *C. glaziovii*. The data revealed no clear species boundaries between *C. pubiflora* and *C. glaziovii*, and these species should be synonymised. Similar studies can help elucidade the boundaries of other species not recovered as monophyletic in the South American SDTF clade. I built a species tree for *Ceiba* to investigate further the relationship amongst the species under a coalescent framework and showed that species in *Ceiba* are structured ecologically and geographically. This species tree will be used to study species diversification and biogeography in more detail in the subsequent chapter.

Bibliography

Buckley TR, Cordeiro M, Marshall DC, Simon C. 2006. Differentiating between Hypotheses of Lineage Sorting and Introgression in New Zealand Alpine Cicadas (Maoricicada Dugdale). *Systematic Biology*. 55:411–425.

Camargo A, Morando M, Avila LJ, Sites JW. 2012. Species delimitation with ABC and other coalescent- based methods: a test of accuracy with simulations and an em- pirical example with lizards of the Liolaemus darwinii complex (Squamata: Liolaemidae). *Evolution*. 66:2834–2849.

Carstens BC, Pelletier TA, Reid NM, Satler JD. 2013. How to fail at species delimitation. *Molecular Ecology*. 22:4369–4383.

De Queiroz K. 2007. Species Concepts and Species Delimitation. *Systematic Biology*. 56:879–886.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*. 24:332–340.

Dexter KG, Pennington TD, Cunningham CW. 2010. Using DNA to assess errors in tropical tree identifications: How often are ecologists wrong and when does it matter? *Ecological Monographs*. 80:267–286.

Doyle JJ. 1992. Gene Trees and Species Trees: Molecular Systematics as One-Character Taxonomy. *Systematic Botany*. 17:144.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution*. 63:1–19.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the boostrap. *Evolution*. 39:783–791.

Flouri T, Jiao X, Rannala B, Yang Z. 2018. Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution*. 35:2585–2593.

Freudenstein JV, Broe MB, Folk RA, Sinn BT. 2016. Biodiversity and the Species ConceptLineages are not Enough. *Systematic Biology*. 66:syw098.

Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*. 27:480–488.

Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 363:4023–4029.

Gibbs P, Semir J. 2003. A taxonomic revision of the genus Ceiba Mill. (Bombacaceae). *Anales del Jardín Botánico de Madrid*. 60:259–300.

Guo X, Thomas DC, Saunders RM. 2018. Gene tree discordance and coalescent methods support ancient intergeneric hybridisation between Dasymaschalon and Friesodielsia (Annonaceae). *Molecular Phylogenetics and Evolution*. 127:14–29.

Harrison RG, Larson EL. 2014. Hybridization, Introgression, and the Nature of Species Boundaries. *Journal of Heredity*. 105:795–809.

Holder MT, Anderson JA, Holloway AK. 2001. Difficulties in Detecting Hybridization. *Systematic Biology*. 50:978–982.

Hudson RR, Coyne JA. 2002. Mathematical consequences of the genealogical species concept. *Evolution*. 56:1557–1565.

Kingman J. 1982. The coalescent. *Stochastic Processes and their Applications*. 13:235–248.

Knowles LL, Carstens BC. 2007. Delimiting Species without Monophyletic Gene Trees. *Systematic Biology*. 56:887–895.

Kubatko LS, Degnan JH. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology*. 56:17–24.

Lambert SM, Reeder TW, Wiens JJ. 2015. When do species-tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny. *Molecular Phylogenetics and Evolution*. 82:146–155.

Leaché AD, Fujita MK. 2010. Bayesian species delimitation in West African forest geckos (Hemidactylus fasciatus). *Proceedings of the Royal Society B: Biological Sciences*. 277:3071–3077.

Leaché AD, Fujita MK, Minin VN, Bouckaert RR. 2014. Species Delimitation using Genome-Wide SNP Data. *Systematic Biology*. 63:534–542.

Leaché AD, Rannala B. 2011. The Accuracy of Species Tree Estimation under Simulation: A Comparison of Methods. *Systematic Biology*. 60:126–137.

Leaché AD, Zhu T, Rannala B, Yang Z. 2019. The Spectre of Too Many Species. *Systematic Biology*. 68:168–181.

Maddison WP. 1997. Gene Trees in Species Trees. *Systematic Biology*. 46:523–536.

Minh BQ, Hahn MW, Lanfear R. 2018. New methods to calculate concordance factors for phylogenomic datasets. *bioRxiv preprint*. .

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*. 31:i44–i52.

Naciri Y, Linder HP. 2015. Species delimitation and relationships: The dance of the seven veils. *Taxon*. 64:3–16.

Pennington RT, Lavin M. 2016. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytologist*. 210:25–37.

Prata EMB, Sass C, Rodrigues DP, Domingos FMCB, Specht CD, Damasco G, Ribas CC, Fine PVA, Vicentini A. 2018. Towards integrative taxonomy in Neotropical botany: disentangling the Pagamea guianensis species complex (Rubiaceae). *Botanical Journal of the Linnean Society*. 188:213–231.

Rabiee M, Sayyari E, Mirarab S. 2019. Multi-allele species reconstruction using ASTRAL. *Molecular Phylogenetics and Evolution*. 130:286–296.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*. 164:1645–1656.

Rannala B, Yang Z. 2013. Improved Reversible Jump Algorithms for Bayesian Species Delimitation. *Genetics*. 194:245–253.

Rannala B, Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*. 66:syw119.

Richardson AO, Palmer JD. 2007. Horizontal gene transfer in plants. *Journal of Experimental Botany*. 58:1–9.

Rieseberg LH, Brouillet L. 1994. Are Many Plant Species Paraphyletic? *Taxon*. 43:21.

Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*. 100:56–62.

Salichos L, Stamatakis A, Rokas A. 2014. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Molecular Biology and Evolution*. 31:1261–1271.

Sayyari E, Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution*. 33:1654–1668.

Shi CM, Yang Z. 2018. Coalescent-Based Analyses of Genomic Sequence Data Provide a Robust Resolution of Phylogenetic Relationships among Major Groups of Gibbons. *Molecular Biology and Evolution*. 35:159–179.

Sigwart J. 2009. Coalescent Theory: An Introduction. *Systematic Biology*. 58:162–165.

Simmons MP, Gatesy J. 2015. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular Phylogenetics and Evolution*. 91:98–122.

Smith SA, Brown JW, Walker JF. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PLOS ONE*. 13:e0197433.

Smith Sa, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*. 15:150.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–1313.

Sukumaran J, Knowles LL. 2017. Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences*. 114:1607–1612.

Wiens JJ, Penkrot TA. 2002. Delimiting Species Using DNA and Morphological Variation and Discordant Species Limits in Spiny Lizards (Sceloporus). *Systematic Biology*. 51:69–91.

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. *Current Zoology*. 61:854–865.

Yang Z, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*. 107:9264–9269.

Yang Z, Rannala B. 2014. Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci. *Molecular Biology and Evolution*. 31:3125–3135.

4.A Appendix

Gene tree discordance using PhyParts including support values.

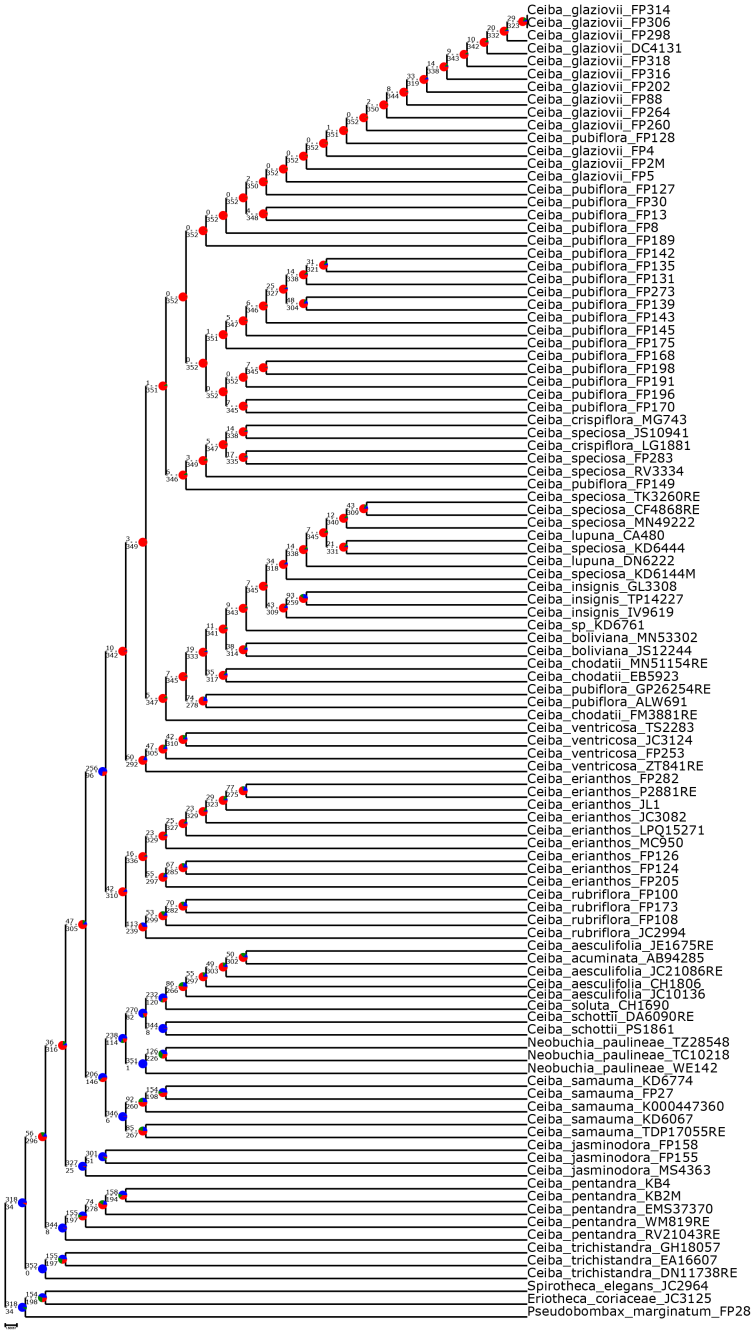


Figure 4.12: Gene tree discordance using PhyParts. The pie charts represent the proportion of gene trees that (i) support the shown topology (blue), (ii) conflict with the shown topology (most common conflicting bipartion) (green); (iii) conflict with the shown topology (all other supported conflicting bipartitions) (red); (iv) have no support for conflicting bipartion (grey).The values above branches represent the number of individual gene trees that support the shown topology and the numbers below the branches represent the number of individual gene trees that conflict with the shown topology.

Chapter 5

Patterns of species diversification and biogeographical history of *Ceiba*

5.1 Introduction

SDTF-confined clades contain species that often resolve as monophyletic in DNA-sequence-based phylogenies and with old stem ages (Pennington and Lavin, 2016). In addition, the geographically structured phylogenetic pattern characteristic of clades in this biome suggests dispersal-limited, old lineages maintained over evolutionary timescales by the stable ecological conditions of the biome (Pennington et al., 2010; Hughes et al., 2013). By contrast, tree clades confined to the Amazon rain forest, the largest tropical forest in the world, are suggested to contain non-monophyletic species more often, and to have young species stem ages and lack of geographical phylogenetic structure (Dexter et al., 2017). These rain forest patterns might be explained by frequent dispersal and subsequent successful colonization (Pennington and Lavin, 2016). These patterns have been reported mostly for legumes (e.g., Pennington et al. 2010, 2011; Särkinen et al. 2012; Lavin et al. 2018), and it is not clear whether they are general.

The results in Chapter 2 show patterns of long stems with shallow crown groups for rain forest species such as *C. pentandra* and runs contrary to the prediction of Pennington and Lavin (2016) that rain forest species might, on average, tend to have more recent origins. Within the two predominantly SDTF clades reported in Figure 2.3, there is little evidence for morphologically recognized species being monophyletic lineages with long stems and shallow crown groups, as predicted by Pennington and Lavin (2016). Those results illustrate that the general patterns of species age, monophyly and geographical structure reported for SDTF species belonging to the Leguminosae family (Pennington and Lavin, 2016) are not shared by one of the most characteristic SDTF tree genera and suggests that phylogenetic studies of other, unrelated groups are required. However, Chapter 2 was based on a single marker (ITS) covering 24 accessions representing 14 of the 18 species described for the genus. To further investigate species diversification and biogeography of *Ceiba*, I used the well sampled and well resolved multi-accession dated phylogeny presented in Chapter 3 and Chapter 4.

In this chapter I aim to investigate the evolutionary history of *Ceiba*, and use the

biogeographical history of the genus to gain insights into the evolution and ecology of neotropical SDTF. I also aim to assess whether the *Ceiba* phylogeny is geographically or ecologically structured and whether species confined to SDTFs are resolved differently in the phylogeny as compared with rain forest species (i.e., monophyletic on long stem lineages). Specifically I aim to answer the following questions:

1. How old are the SDTF that *Ceiba* inhabit?
2. How old are individual species in the genus *Ceiba* and do *Ceiba* species occurring in different biomes show different phylogenetic patterns?

5.2 Methods

5.2.1 Molecular dating

As in Chapter 2, I used the fossil flower of *Eriotheca prima* (Duarte, 1974) from the middle to late Eocene (de Lima and Salard-Cheboldaeff, 1981) of Brazil as a primary temporal calibration. The flower was identified as *Eriotheca* based on its small size (*Bombacopsis* and *Pachira* have larger flowers) and organisation of its androecium, which are apomorphies for the extant species of the genus (Robyns, 1963; Duarte, 1974; Carvalho-Sobrinho et al., 2016). Because the dating of this fossil is imprecise (middle to late Eocene: 33-56 mya), I assigned a minimal age of 33 Ma and maximum age of 56 Ma as a constraint to the age to the stem node of *Eriotheca* (and as explained in Chapter 2, assigned it to the crown node of the clade comprising *Eriotheca*, *Spirotheca*, *Pseudobombax*, and *Ceiba* following relationships in Bombacoideae elucidated by Carvalho-Sobrinho et al. (2016)). I followed the dates on the Geologic Time Scale v. 5.0 (Gradstein et al., 2012).

To run the dating analysis, initially, I attempted using a bayesian framework using BEAST2 for 111 filtered loci and 103 accessions. However, the MCMC did not converge after several attempts with up to 400 million generations, possibly due to the large volume of data.

The dating analysis was then conducted using a penalised likelihood method implemented with treePL, which was developed to deal with large data sets (Smith and O'Meara, 2012). Penalised likelihood is a semiparametric approach to estimate rates of molecular evolution (Sanderson, 2002), allowing different rates on different branches of the phylogenetic tree. It uses a smoothing parameter to determine how much the rate differences over the tree are penalised (Smith and O'Meara, 2012).

As input, I used the phylogeny inferred under the multi-species coalescent model using Astral-ind analysis from Chapter 4. The latest version of Astral (v. 5.6.3) can deal with multi-alleles and the analysis can be run either by assigning individuals to species (Astral-multi) or by considering each individual as a separate sample (Astral-ind). However, it is not clear whether Astral-ind will treat each accession as a separate species and attempts to estimate population sizes for each "species" and thus treat a single true species as multiple single-sample units, which would not be logical. Therefore, to account for possible methodological issues, I also used the concatenated phylogeny following the Nicholls et al. (2015) pipeline built in Chapter 3 for a second dating analysis.

For each of the inputs, I conducted an initial thorough run (option to designate that you want a thorough analysis) with the prime set (option to designate that you want to determine the best optimization parameters) to test different optimisation possibilities and random cross-validation[1]. The optimal smoothing values are the ones with lower cross-validation scores recovered by the prime run. Those values were then included in the control file for a second thorough run with smoothing value of 0.1 for the phylogeny generated using the concatenation approach (Figure 5.1) and 0.02 for the phylogeny generated under the multi-species coalescent model (Figure 5.2).

[1]https://github.com/blackrim/treePL/wiki

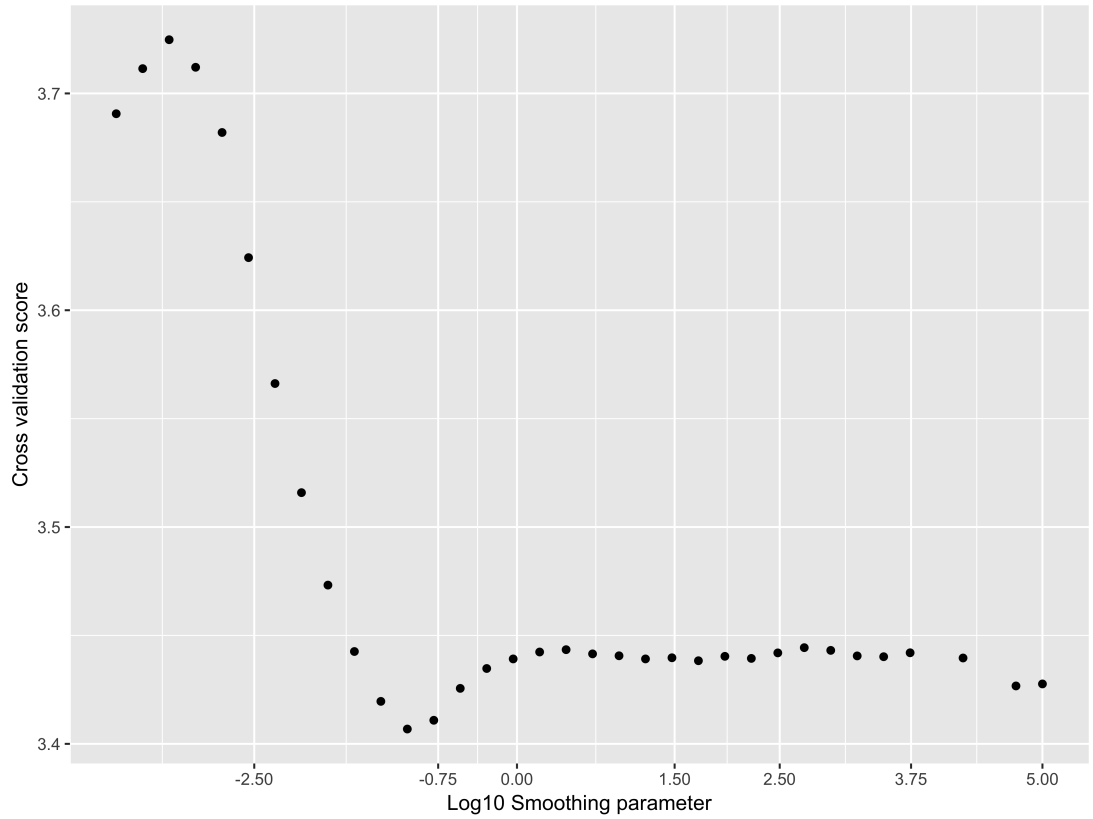


Figure 5.1: Cross validation scores obtained using different smoothing parameters on the prime run of treePL for the concatenated phylogeny inferred following the Nicholls et al. (2015) pipeline (Chapter 3). The lower value indicates the cross-validation score for penalized likelihood under optimal smoothing sensu Sanderson (2002). Axes are represented in logarithmic scale.

5.3 Results

5.3.1 Molecular dating

Concatenated phylogeny

The crown node of *Ceiba* was estimated at 45 Ma for the phylogeny inferred from the concatenated matrix following Nicholls et al. (2015) pipeline and built in Chapter 3 (Figure 5.3). The crown node age of the *C. trischistandra* clade was estimated at 0.7 Ma and the stem node 45 Ma. The *C. pentandra* clade was recovered with a crown node age of 8 Ma and stem node age of 44 Ma. The crown node age of the Central American and Mexican SDTF clade was estimated as 26.5 Ma and the stem node age 34 Ma. Within this clade, *C. schottii* was recovered with a crown node age of 0.9 Ma and stem

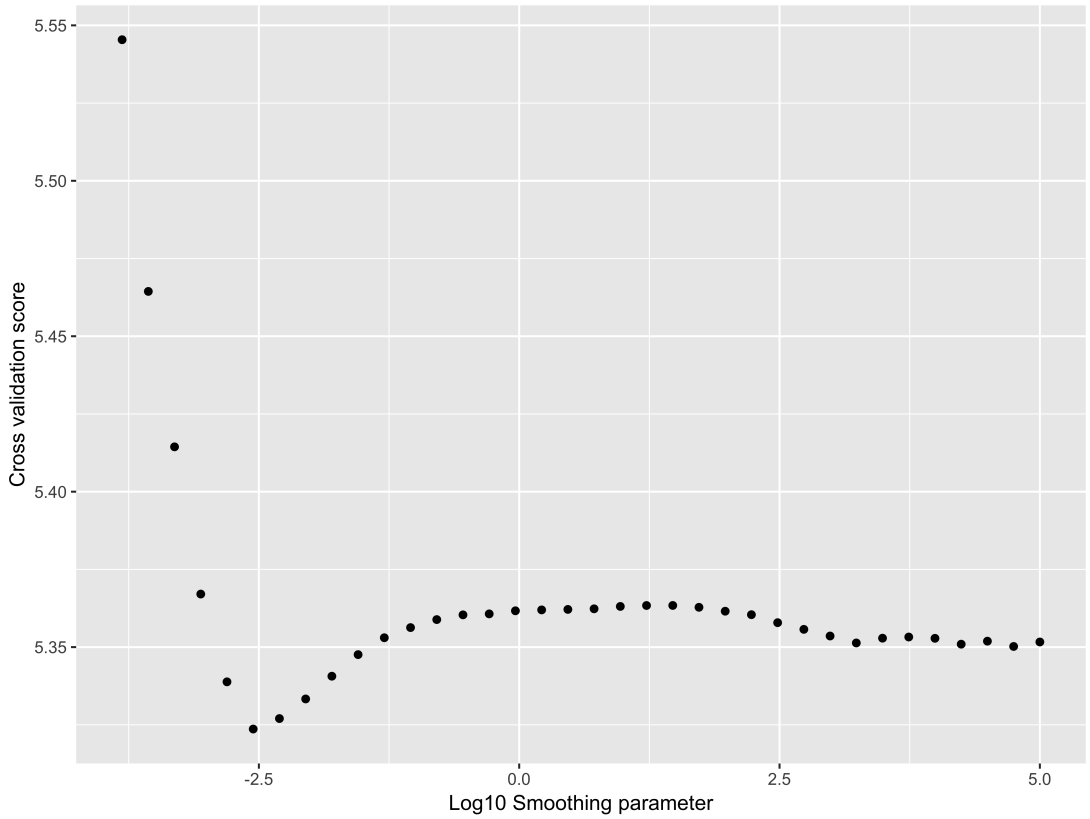


Figure 5.2: Cross validation scores obtained using different smoothing parameters on the prime run of treePL for the phylogeny inferred under the multi-species coalescent model in the Astral-ind analysis (Chapter 4). The lower value indicates the cross-validation score for penalized likelihood under optimal smoothing sensu Sanderson (2002). Axes are represented in logarithmic scale.

node age of 18.5 Ma. The crown node age of the *Neobuchia paulinae* clade was 0.5 Ma and the stem node age was 26.5 Ma. For the *C. samauma* clade, the crown node age was estimated at 3.5 Ma and the stem node age at 34 Ma. The crown node age of the *C. jasminodora* clade was estimated as 2.1 Ma and the stem node age 43 Ma. The South American SDTF clade crown node age was estimated at 25.3 Ma and the stem node age 41.6 Ma. Within this clade, three species were recovered as monophyletic. The *Ceiba erianthos* crown node age was estimated as 10.6 Ma and stem node age 14.4 Ma. The crown node age of the *C. rubriflora* clade was estimated as 4.5 Ma and the stem node age 14.4 Ma. The crown node age of the *C. ventricosa* clade was estimated as 9.05 Ma and the stem node age 22.6 Ma (Figure 5.3).

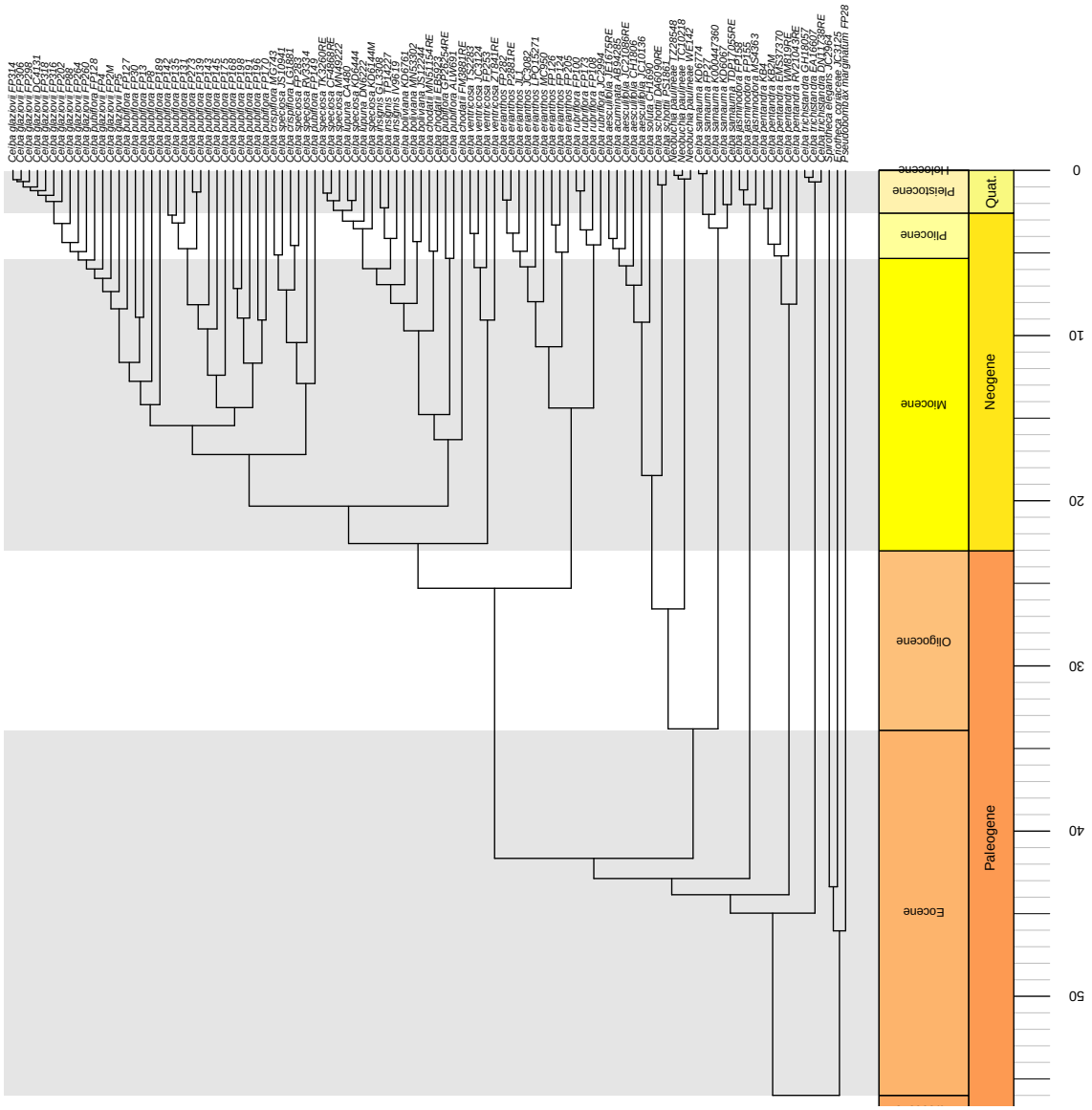


Figure 5.3: Dated phylogeny generated in treePL with smoothing parameter of 0.01, using as input the concatenated phylogeny following the Nicholls et al. (2015) pipeline (Chapter 3).

Coalescent phylogeny

The crown node of *Ceiba* was estimated at 50 Ma for the phylogeny inferred under the multi-species coalescent model using following the Astral-ind analysis from Chapter 4 (Figure 5.4). The crown node age of the *C. trischistandra* clade was estimated as 1.9 Ma and the stem node 47.2 Ma. The *C. pentandra* clade was recovered with a crown node age of 4.6 Ma and stem node age as 47.2 Ma. The crown node age of the Central American and Mexican SDTF clade was estimated as 38 Ma and the stem node age 44.3 Ma. Within this clade, *C. schottii* was recovered with stem node age of 26.7 Ma. The crown node age of the *Neobuchia paulinae* clade was 0.5 Ma and the stem node age was 38 Ma. For the *C. samauma* clade, the crown node age was estimated at 4.7 Ma and the stem node age 44.3 Ma. The crown node age of the *C. jasminodora* clade was estimated as 7.5 Ma and the stem node age 49 Ma. The South American SDTF clade crown node age was estimated as 10.4 Ma and the stem node age as 49 Ma. Within this clade, two species were recovered as monophyletic. The crown node age of the *C. rubriflora* clade was estimated as 0.8 Ma and the stem node age 5.2 Ma. The crown node age of the *C. ventricosa* clade was estimated as 0.8 Ma and the stem node age 9.4 Ma. In contrast to the concatenation analysis, *C. erianthos* was not recovered as monophyletic (Figure 5.4).

5.4 Discussion

Phylogenetic dating

Overall, the dates inferred using the two different input trees were similar. The differences in topology between the two phylogenies might explain the differences in the absolute ages found for the different clades. The largest difference was ca. 15 Ma for the crown node of the SDTF South American clade, which was estimated as 25.3 Ma for the treePL run using as input the phylogeny generated by concatenation analysis and at 10.4 Ma for the treePL run using as input the phylogeny inferred by the Astral-ind analysis. This variation is likely due to the fact that *C. jasminodora*, a species

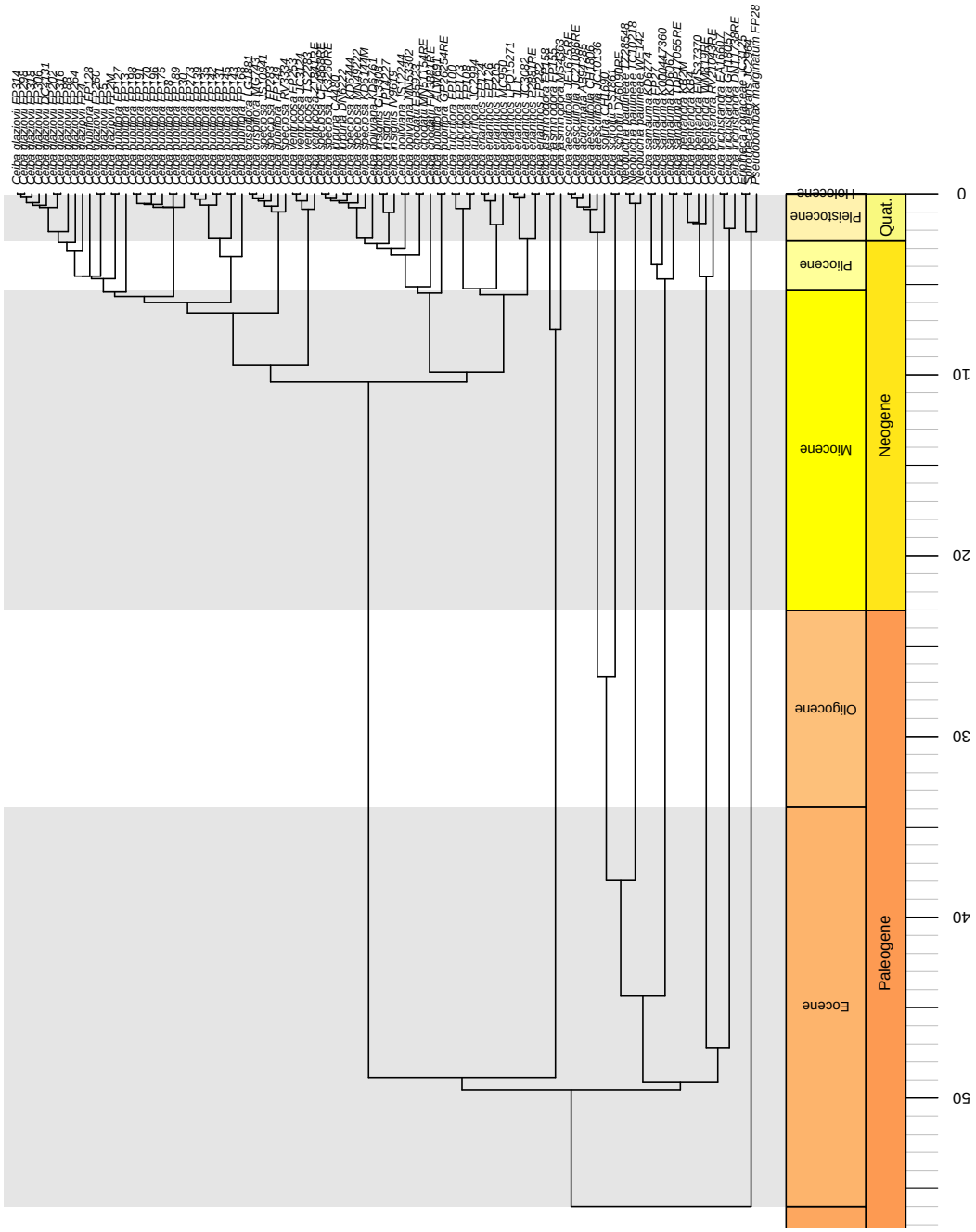


Figure 5.4: Dated phylogeny generated in treePL with smoothing parameter of 0.002, using as input the phylogeny inferred under the multi-species coalescent model following the Astral-ind analysis (Chapter 4).

recovered with old stem age and recent crown age, is recovered as sister to the SDTF South American clade in the Astral-ind analysis. However, because of concerns about the assumptions in the Astral-ind analysis, the ages recovered (in general older) should be interpreted carefully. Nonetheless, the relative ages for each of the clades recovered, with old long stems and shallow young crown for the species recovered as monophyletic and the shorter stem and deeper crown for the species within both SDTF clades were congruent between analyses.

The results from the dating analysis in Chapter 2 inferred with BEAST2 for ITS sequence data (Figure 2.3) and the treePL analysis using both inputs were similar. The results show a pattern of old species-poor lineages maintained over evolutionary time since the late Eocene and more recent, more species-rich clades diversifying from the Miocene.

How old are the SDTF that *Ceiba* inhabit?

The dating analysis recovered the crown node of *Ceiba* at 45 Ma (Figure 5.3) or 50 Ma (Figure 5.4). De-Nova et al. (2012) showed that, for the SDTF Mexican genus *Bursera*, diversification started around the same time (c. 50 Ma), reinforcing the hypothesis that neotropical SDTF arose in the middle Eocene (Pennington et al., 2006, 2009) possibly in North America (Pennington et al., 2009; De-Nova et al., 2012). The lineages of *Ceiba* that diverged in the middle Eocene occur in South America (*C. samauma*, *C. jasminodora* and *C. trischistandra*), with the exception of *Ceiba pentandra* that is widespread in both continents. However, of these, only *C. trischistandra* is a SDTF species. The uncertainty around the relationships of the basally divergent lineages of *Ceiba* (*C. trischistandra*, *C. jasminodora*, *C. pentandra*,*C. samauma*, and the Central American clade; see Chapter 3, Chapter 4) means that it is uncertain whether SDTF would be optimised as the ancestral biome at the *Ceiba* crown node. Hence, an Eocene origin for SDTF in South America as found in Central America cannot be certain. Nonetheless, the crown age of the South American SDTF clade dating from late Oligocene (25.3 Ma) is earlier than previous mid-Miocene estimates (Burnham and

Carranco, 2004). For most *Ceiba* species, especially in the SDTF South American and SDTF Central American and Mexican clades, diversification took place beginning in the late Miocene, similarly to *Bursera* (De-Nova et al., 2012) in North America. Likewise, Pennington et al. (2004) showed that diversification happened beginning in the late Miocene and Pliocene for five lineages highly endemic to SDTF in South America: *Ruprechtia* (Polygonaceae), robinioid legumes (Fabaceae), *Chaetocalyx* and *Nissolia* (Fabaceae), and *Loxopterygium* (Anacardiaceae).

How old are individual species in the genus *Ceiba* and do *Ceiba* species occurring in different biomes show different phylogenetic patterns

Ceiba trischistandra, *C. pentandra*, *C. jasminodora*, *C. samauma*, *C. schottii* and *Neobuchia paulinae* were recovered with young crown ages (between 0.5 and 8.1 Ma) and old stem ages (between 34 and 45 Ma). *Ceiba pentandra* and *C. samauma* occur in wet forests and have widespread distributions. *Ceiba trischistandra*, *C. schottii* and *Neobuchia paulinae* occur in SDTF, and *C. jasminodora* in the granitic *campos rupestres* (rocky upland vegetation) in central Brazil. Although occurring in different biomes and with different distribution ranges, all six species have restricted distributions and have the same pattern of old lineages maintained over evolutionary timescales since the late Eocene or middle Oligocene. Those lineages fit the concept of depauperons sensu Donoghue and Sanderson (2015), which are not uncommon in nature (Donoghue and Sanderson, 2015), though understanding how they are generated and maintained over time is challenging (Donoghue and Sanderson, 2015). They could reflect low net diversification rates over time (i.e., they have always been species-poor), or alternatively they may be the "dying embers" of more species-rich clades (Donoghue and Sanderson, 2015).

Conversely, species-rich clades arise with a high net diversification rate maintained over evolutionary time and, in Angiosperms, those species-rich clades have been correlated with large geographic areas (Donoghue and Sanderson, 2015). The South American SDTF clade, where most species occur in the Caatinga biome, the largest area of

SDTF (Pennington et al., 2006), is the most species-rich clade in *Ceiba*.

Wet Forests

Ceiba pentandra occurs mainly in the vast continuous areas of the Amazon (Gibbs and Semir, 2003; Dick et al., 2007) and, although large geographical areas have been correlated with species-rich clades (Donoghue and Sanderson, 2015), it was recovered as a depauperate lineage. The Amazon was the stage of major landscape transformations such as changes in the course of rivers (Hoorn and Wesselingh, 2009) and those events may have caused extinctions over evolutionary time. In addition to that, Dick et al. (2007) reported that this species has the weakest phylogeographical structure when compared to other widespread rainforest tree species. The weak phylogeographical structure implies a high level of gene flow that would lower the likelihood of speciation events (Coyne and Orr, 2004). Therefore, a low diversification rate combined with high extinction rate might be the cause of the depauperate lineage in *Ceiba pentandra*, which may represent the only surviving lineage of a once more diverse clade.

Ceiba samauma is an interesting case. It is widely and discontinuously distributed in humid and riverine forest from Bolivia and Peru to the Brazilian Amazon (Gibbs and Semir, 2003). Two samples sequenced in this study came from populations occurring in semi-deciduous forests by the slopes of Andean valleys in Peru (KD6774 and KD6067). *C. samauma* was recovered as a depauperate lineage. It is possible that *C. samauma* has effective pollen and seed flow similar to *C. pentandra*, which would make speciation events less likely. However, the literature about this species is scarce and further studies are necessary.

Campos Rupestres

Ceiba jasminodora was recovered with crown node age of 2.1 Ma and stem age of 43 Ma in the phylogeny infered using the concatenated matrix (Figure 5.3) and 7.5 Ma and stem age of 49 Ma in the phylogeny inferred under the coalescent model (Figure 5.4). *Ceiba jasminodora* occurs in the granitic *campos ruprestres* (rocky upland

vegetation) of the southern Serra do Espinhaço in Minas Gerais state in Brazil (Gibbs and Semir, 2003). Campos rupestres are hyperdiverse habitats occuring in isolated patches throughout Brazil (Neves et al., 2018). Although campos rupestres harbour ca. 15% of the Brazilian vascular flora (Neves et al., 2018), little is known about the evolutionary history of these endangered landscapes (Hughes et al., 2013). Campos ruprestres are thought to be old, stable, and to contain a combination of old lineages and young species diversified *in situ* (Inglis and Cavalcanti, 2018), in contrast with the Cerrado biome, which is floristically highly related (Neves et al., 2017), but assembled relatively recently, with most lineages diversifying at 4 Ma or less (Simon et al., 2009). The biomes surrounding campos rupestres such as the Cerrado, Amazon, Atlantic Forest, and SDTF, might be the source of plant lineages that colonised there (Neves et al., 2018). For example, the genus *Calliandra*, common in campos rupestres, seems to have a SDTF origin (de Souza et al., 2013). However, unlike the Cerrado, colonisation of campos rupestres from surrounding biomes may not be evolutionarily recent. The stem nodes of *Calliandra* species occurring in campos rupestres are dated to the Miocene (de Souza et al., 2013). The stem age of *Ceiba jasminodora*, at 43 Ma is far older and represents the oldest lineage ever reported for campos rupestres.

Seasonally Dry Tropical Forests

Ceiba trischistandra is the only species occurring in SDTF patches west of the Andes, in the dry valleys of the Pacific coast in southern Ecuador and northern Peru (Gibbs and Semir, 2003). *Neobuchia paulinae* is endemic to Haiti and associated with calcareous soils (Urban, 1898). The geographically isolated distribution of these two species are likely reflected in the persistence of depauperate lineages over evolutionary time. Allopatric speciation in small geographical areas is less likely and sympatric speciation is controversial and rare cases are reported in nature (Coyne, 2011). Therefore, the pattern shown by *C. trischistandra* and *Neobuchia paulinae* might reflect low speciation rates due to their restricted geographical range, and their evolutionary persistence may be explained by their strong adaptations to seasonal SDTF climates (Pennington

and Lavin, 2015).

Within the SDTF clade, where seven out of 11 species were not recovered as monophyletic, species have varying distribution ranges. Among the 11 species, *Ceiba insignis*, *C. ventricosa* and *C. rubriflora* have the narrower distributions (Figure 1.4). Although the geographical range of *Ceiba erianthos* spans from the southeast of Brazil in Rio de Janeiro state to the centre of Bahia state in the northeast of Brazil, it is restricted to granitic outcrops and therefore has a small total distribution area (Figure 1.4). *Ceiba insignis*, *C. ventricosa*, *C. rubriflora* and *Ceiba erianthos* were recovered as monophyletic with relatively old, deep crown (ages between 4.1 and 10.7 Ma) and short stems (ages between 6 and 22.5 Ma). Pennington and Lavin (2016) predicted old lineages with long stems and shallow crowns recovered as monophyletic for taxonomically recognised SDTF species. *C. insignis*, *C. ventricosa*, *C. rubriflora* and *C. erianthos* fit their prediction because they are relatively old and monophyletic species. Their narrow distribution and consequently small effective population size indicate a shorter time to coalesce (Naciri and Linder, 2015), which reflects in the monophyly of the morphologically recognised species in small SDTF populations as predicted by Pennington and Lavin (2016). However, their crowns are not shallow and the stems are not long, which could indicate a lower extinction rate for these lineages than found in the examples discussed by Pennington and Lavin (2016).

Similarly, the seven species recovered as non-monophyletic, *Ceiba pubiflora*, *C. glaziovii*, *C. boliviana*, *C. chodatii*, *C. speciosa*, *C. crispiflora* and *C. lupuna*, are relatively lineages, and there is little evidence for accessions clustered in shallow crown groups across the South American SDTF clade, which again could indicate low extinction rates. Among these species, *C. pubiflora*, *C. glaziovii*, *C. lupuna* and *C. speciosa* are widely distributed, the first two in dry forests and the last two in wet forests. *Ceiba chodatii* and *C. boliviana* are not as widely distributed as *C. pubiflora* and *C. glaziovii*, but still occur in larger areas when compared to other species in the west of South America such as *C. insignis* and *C. trischistandra*. Although it may seem initially that these seven species have a different phylogenetic pattern from the predicted by Pen-

nington and Lavin (2016), their potentially larger effective population size is reflected in a longer time to coalesce (Naciri and Linder, 2015) and, allied to the possibility of less extinction rates, means that monophyly requires more time.

Ceiba pubiflora and *C. glaziovii* were not recovered as monophyletic in spite of sequencing 32 accessions of both species. Instead, *C. glaziovii* is nested within *C. pubiflora*. Moreover, there is no clear morphological boundary between the two species, and, together with the species delimitation analysis (Chapter 4), there is clear indication that they represent the same species. The crown age of the clade containing the two species dates to the Miocene and therefore represent a relatively old radiation. The pattern for those two species may represent a signature of a successful species that originated late Miocene in central Brazil and has been spreading gradually north into the Caatinga region of northeastern Brazil.

Within the Central American and Mexican SDTF clade all three species were recovered as monophyletic. *Ceiba schottii* had a young crown age (0.9 Ma) and old stem age (18.5 Ma) as expected for a SDTF species (Pennington and Lavin, 2016) with narrow distribution (Figure 1.4). *Ceiba aesculifolia* is widely distributed from Mexico to Costa Rica, and although recovered as monophyletic, the pattern of older crown (7 Ma) and a shorter stem (9.2 Ma) ages may reflect lower extinction rates as discussed above for South American SDTF species.

Conclusions

The stem and crown ages of the South American SDTF clade dated to the late Oligocene reinforces the fact that *Ceiba* has occupied the South American dry forests at least for the last 25 Ma. *Ceiba trischistandra* is recovered outside this clade in all phylogenies inferred, indicating that the origin of South American SDTF might be even older.

Overall, the dated phylogenies in this thesis does not fully sustain the predictions that species confined to SDTFs are resolved differently as compared with rain forest species (Pennington and Lavin, 2016). Although SDTF species in *Ceiba* can be old, in many cases they are not resolved as monophyletic. Rain forest species of *Ceiba*

are resolved as monophyletic, and with old stem ages. The shallow crown and long stem of the wet forest species *C. pentandra* and *C. samauma* possibly indicate high extinction along the stems. The deep crowns of the SDTF South American species possibly indicates lower extinctions along the stem lineages than found in previously published examples in legumes (Pennington et al., 2010, 2011; Särkinen et al., 2012).

Bibliography

Burnham RJ, Carranco NL. 2004. Miocene winged fruits of Loxopterygium (Anacardiaceae) from the Ecuadorian Andes. *American Journal of Botany*. 91:1767–1773.

Carvalho-Sobrinho JG, Alverson WS, Alcantara S, Queiroz LP, Mota AC, Baum DA. 2016. Revisiting the phylogeny of Bombacoideae (Malvaceae): Novel relationships, morphologically cohesive clades, and a new tribal classification based on multilocus phylogenetic analyses. *Molecular Phylogenetics and Evolution*. 101:56–74.

Coyne JA. 2011. Speciation in a small space. *Proceedings of the National Academy of Sciences*. 108:12975–12976.

Coyne JA, Orr HA. 2004. Speciation. Sunderland, Massachusetts: Sinauer & Associates.

de Lima MR, Salard-Cheboldaeff M. 1981. Palynologie des bassins de Gandarela et Fonseca (Eocene de l'etat de Minas Gerais, Bresil). *Boletim IG*. 12:33–54.

De-Nova JA, Medina R, Montero JC, Weeks A, Rosell JA, Olson ME, Eguiarte LE, Magallón S. 2012. Insights into the historical construction of species-rich Mesoamerican seasonally dry tropical forests: the diversification of Bursera (Burseraceae, Sapindales). *New Phytologist*. 193:276–287.

de Souza ÉR, Lewis GP, Forest F, Schnadelbach AS, van den Berg C, de Queiroz LP. 2013. Phylogeny of Calliandra (Leguminosae: Mimosoideae) based on nuclear and plastid molecular markers. *Taxon*. 62:1200–1219.

Dexter KG, Lavin M, Torke BM, Twyford AD, Kursar TA, Coley PD, Drake C, Hollands R, Pennington RT. 2017. Dispersal assembly of rain forest tree communities across the Amazon basin. *Proceedings of the National Academy of Sciences*. 114:2645–2650.

Dick CW, Bermingham E, Lemes MR, Gribel R. 2007. Extreme long-distance dispersal

of the lowland tropical rainforest tree Ceiba pentandra L. (Malvaceae) in Africa and the Neotropics. *Molecular Ecology*. 16:3039–3049.

Donoghue MJ, Sanderson MJ. 2015. Confluence, synnovation, and depauperons in plant diversification. *New Phytologist*. 207:260–274.

Duarte L. 1974. Sobre uma flor de Bombacaceae da Bacia Terciária de Fonseca, MG. *Anais da Academia Brasileira de Ciências*. 46:407–411.

Gibbs P, Semir J. 2003. A taxonomic revision of the genus Ceiba Mill. (Bombacaceae). *Anales del Jardín Botánico de Madrid*. 60:259–300.

Gradstein FM, Ogg JG, Schmitz MD, Ogg GM. 2012. The Geologic Time Scale. Elsevier.

Hoorn C, Wesselingh FP, editors. 2009. Amazonia: Landscape and Species Evolution. Oxford, UK: Wiley-Blackwell Publishing Ltd.

Hughes CE, Pennington RT, Antonelli A. 2013. Neotropical Plant Evolution: Assembling the Big Picture. *Botanical Journal of the Linnean Society*. 171:1–18.

Inglis PW, Cavalcanti TB. 2018. A molecular phylogeny of the genus diplusodon (Lythraceae), endemic to the campos rupestres and cerrados of South America. *Taxon*. 67:66–82.

Lavin M, Pennington RT, Hughes CE, Lewis GP, Delgado-Salinas A, Duno de Stefano R, de Queiroz LP, Cardoso D, Wojciechowski MF. 2018. DNA Sequence Variation among Conspecific Accessions of the Legume Coursetia caribaea Reveals Geographically Localized Clades Here Ranked as Species. *Systematic Botany*. 43:664–675.

Naciri Y, Linder HP. 2015. Species delimitation and relationships: The dance of the seven veils. *Taxon*. 64:3–16.

Neves DM, Dexter KG, Pennington RT, Bueno ML, de Miranda PL, Oliveira-Filho AT. 2018. Lack of floristic identity in campos rupestres A hyperdiverse mosaic of rocky montane savannas in South America. *Flora*. 238:24–31.

Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science*. 6:1–20.

Pennington RT, Daza A, Reynel C, Lavin M. 2011. Poissonia eriantha (Leguminosae) From Cuzco, Peru: An Overlooked Species Underscores a Pattern of Narrow Endemism Common to Seasonally Dry Neotropical Vegetation. *Systematic Botany*. 36:59–68.

Pennington RT, Lavin M. 2016. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytologist*. 210:25–37.

Pennington RT, Lavin M, Oliveira-Filho A. 2009. Woody Plant Diversity, Evolution, and Ecology in the Tropics: Perspectives from Seasonally Dry Tropical Forests. *Annual Review of Ecology, Evolution, and Systematics*. 40:437–457.

Pennington RT, Lavin M, Prado DE, Pendry CA, Pell SK, Butterworth CA. 2004. Historical climate change and speciation: neotropical seasonally dry forest plants show patterns of both Tertiary and Quaternary diversification. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. 359:515–538.

Pennington RT, Lavin M, Sarkinen T, Lewis GP, Klitgaard BB, Hughes CE. 2010. Contrasting plant diversification histories within the Andean biodiversity hotspot. *Proceedings of the National Academy of Sciences*. 107:13783–13787.

Pennington RT, Lewis GP, Ratter JA, editors. 2006. Neotropical Savannas and Seasonally Dry Forests. Boca Raton: CRC Press.

Robyns A. 1963. Essai de monographie du genre Bombax s.l. (Bombacaceae) (Suite). *Bulletin du Jardin botanique de l'État a Bruxelles*. 33:145.

Sanderson MJ. 2002. Estimating Absolute Rates of Molecular Evolution and Divergence

Times: A Penalized Likelihood Approach. *Molecular Biology and Evolution*. 19:101–109.

Särkinen T, Pennington RT, Lavin M, Simon MF, Hughes CE. 2012. Evolutionary islands in the Andes: persistence and isolation explain high endemism in Andean dry tropical forests. *Journal of Biogeography*. 39:884–900.

Simon MF, Grether R, de Queiroz LP, Skema C, Pennington RT, Hughes CE. 2009. Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proceedings of the National Academy of Sciences*. 106:20359–20364.

Smith SA, O'Meara BC. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*. 28:2689–2690.

Urban I. 1898. Symbolae Antillanae, Seu Fundamenta Florae Indiae Occidenttalis/editi Ignatius Urban. February. Berolini, Parisiis: Fratres Borntraeger; Paul Klincksieck.

Chapter 6

Conclusions

6.1 More data does not solve the problem

Sanger sequence and Next Generation Sequence

In Chapter 2 I used a phylogeny based upon Sanger-sequencing of the ITS region to investigate species relationships in *Ceiba*. ITS has been widely explored to help elucidate relationships among genera and species of flowering plants, and even to investigate genetic structure among their populations, for over 30 years (Baldwin et al., 1995). ITS has proven important to complement the information generated from Sanger-sequenced phylogenies based upon cpDNA and it is generally more variable than cpDNA loci. Now, NGS has presented a solution to increase phylogenetic resolution in situations where ITS fails (e.g, recent species radiations), by sequencing multiple loci or even the entire genome. NGS approaches are an important advance for the field of genetics, phylogenetics and evolution. This new perspective leaves us with the question: are Sanger-sequence phylogenies of ITS and other cpDNA and nuclear regions still relevant? My view is that they are.

As with any new technology, there was an initial excitement around the possibilities that NGS data offer. For example, in Chapter 4 I produced a well resolved and highly supported phylogeny, which included 103 samples of all 18 described species for *Ceiba*, with multiple accessions per species. The next-generation hybrid capture sequencing of 377 nuclear loci increased the phylogenetic resolution in comparison with the phylogeny based on Sanger-sequence of the ITS region from Chapter 2. In addition, the taxon-specific bait set improved capture efficiency.

The bottleneck for NGS sequencing is the use of bioinformatics pipelines, especially considering the lack of programming training for biologists. The four sequencing runs of 103 accessions produced around 10 GB of raw data and a final alignment of more than 1 million base pairs. This data set was analysed using pipelines following *de novo* and reference mapping approaches, with variations in the software used and in their settings. The *de novo* assembly does not require a reference genome and might be preferred for non-model organisms. For *Ceiba*, the Yang and Smith (2014) pipeline produced clusters with a large amount of missing data (maximum of 50 accessions per

tree), even after seven attempts of running it with different settings. *De novo* assembly can be computationally demanding (McCormack et al., 2013) and produce chimeric contigs from repetitive regions especially in plant genomes, and was not suitable for the analysis of next-generation target capture data for *Ceiba*.

All of the reference mapping pipelines tested produced results with desirable data quantity and quality. However, the final output of each was different, which led to different phylogenetic inferences, all with high support values. The conflicting results reinforce the fact that although next generation sequencing techniques represent an important step forward in phylogenetics, more care should be taken when analysing the data with bioinformatics pipelines and interpreting the output, and few workers are doing this. Many papers using hybrid capture do not explain their choice of informatics pipelines and barely mention the key issue of paralogy (e.g Muñoz-Rodríguez et al. (2018)).

Data analysis is still not straightforward with several decisions to be made along the steps that require knowledge of the organism, genome and bioinformatics skills, especially for close species relationships (Degnan and Rosenberg, 2009). Good practice for the analysis of NGS data involves applying different pipelines with various settings followed by an investigation of their biological implications. The assessment of intermediate steps of the analysis is normally made based on metrics related to the amount of data. However, it is important to inspect the intermediate files visually, a common practice in Sanger-sequence analysis that is often neglected in NGS studies. In summary, more data does not necessarily mean better data.

Analysing multiple, independent nuclear loci in a phylogenetic context is also not straightforward. One approach is to combine them all in a concatenated alignment (Nicholls et al., 2015; Carlsen et al., 2018). However, because of issues of incomplete lineage sorting, many workers recommend the use of phylogenetic methods drawing on the background on coalescent theory. These coalescent approaches represent a combination of population genetics, and phylogenetics, accounting for natural biological phenomena such as incomplete lineage sorting. However, new software developed to

apply the multispecies coalescent model often make assumptions that most data sets do not meet. For example, both Astral and BPP, the software packagess I used in Chapter 4, assume neutral clock-like evolution, and less than 10% sequence divergence. A solution to achieve loci with these properties is filtering NGS data. Filtering genes is a common practice when analysing NGS data. I reduced my initial 377 loci to 111 loci to conduct the analysis under the multi-species coalescent model and dating analysis. Even with filtered NGS data, current software still require optimisation. For example, for the dating analysis, it is likely that BEAST2 was not able to deal with this amount of data, as shown by the lack of convergence even after 400 M generations. Simmons and Gatesy (2015) recommend care with the enthusiasm caused by new methods, and not forgetting to rigorously apply the basic principles.

The Sanger-sequence phylogeny I generated in Chapter 2 provided a framework to evaluate the results using next-generation hybrid capture sequencing both for phylogenetics (Chapter 3) and for dating diversification events (Chapter 5). Sanger sequence does not need to be replaced by next generation sequencing. Instead, it would be more beneficial for science if they work together, for example NGS being validated by careful phylogenetic analysis of carefully chosen Sanger-sequenced loci. For the moment, NGS relies on Sanger-sequencing, but the opposite is not necessarily true.

Practical recommendations for NGS data analysis

In this thesis I analysed a data set generated using target capture for 377 nuclear loci of *Ceiba*, whose species are in some cases recently derived. I used a specific bait set containing intronic regions to generate a subset of the genome of *Ceiba* to generate a phylogeny to access the relationships amongst the species represented by 103 accessions, with multiple individuals per species. A *de novo* assembly is not adequate for such data sets because the high amount of genetic variation between the individual samples provided by the introns prevent the formation of clusters containing all taxa. Those clusters would generate an alignment with missing data, which makes phylogentic inference imprecise. Instead, a reference based assembly is better for this type of

data set.

The assemblers tested here, Bowtie2 and BWA, allow the user to choose mapping threshold values to determine if a read is similar to the reference or discarded. Although both software packages use the Burrows-Wheeler transform algorithm (Burrows and Wheeler, 1994), Bowtie2 performed better in this study. The variation in the alignment threshold had a consistent pattern of lowering the percentage of reads mapped as the alignment score increased, as expected. Furthermore, the examination of the bam files showed that fewer potential paralogues were being mapped to the reference. Conversely, BWA had no variation in the percentage of reads being mapped for settings other than -B (results not shown), and did not show a consistent pattern of fewer reads being mapped as I increased the stringency of the mapping threshold. Likewise, the bam files showed an unexpected increase in gappy regions with the variation in the mapping settings. Nonetheless, to exclude possible paralogues, testing a range of alignment scores is important to decide on the best value for each data set.

Filtering the raw reads with the default values of Trimmomatic (3-4:15) results in a data set with good balance between data quality and data loss. Any eventual poor quality base can be discarded in subsequent steps of data analysis, for example by applying a conservative mapping threshold.

Amongst the different pipelines I tested, the Nicholls et al. (2015) pipeline was the most suitable one because it allowed me to conduct each step of the analysis separately, varying the default setting of the software and inspecting visually the intermediate files. Pipelines designed to reduce the manipulation of the data by the user intend to facilitate the analysis, but can create difficulties for the users to vary their settings and inspect the intermediate results.

Ideally, a thorough phylogenetic inference of multi-locus data sets would involve the combination of at least two different approaches: concatenation and species tree/gene trees analysis. Concatenation analysis for NGS shows flaws, for example the inflated bootstrap support values, and it does not consider conflicting gene trees. However, the assumptions made in some species-tree approaches, such as species should be mono-

phyletic in ASTRAL-multi, are also questionable. I therefore recommend running both, and comparing topologies, which for *Ceiba*, were similar.

6.2 Using NGS sequencing in a taxonomic context

The use of NGS data sets in phylogenetics has been explored by several groups (Nicholls et al., 2015; Muñoz-Rodríguez et al., 2018; Carlsen et al., 2018). However, the relevance of these data for taxonomy, specifically the exploration of their utility in defining species, has been relatively neglected. In Chapter 4, I explore the use of modern species delimitation analysis and morphological data under a coalescent framework.

Figure 6.1: *Ceiba pubiflora* in a fragment of SDTF in Porteirinha, Minas Gerais, Brazil. Photo: F. Pezzini.

Seven out of the 18 species of *Ceiba* were not recovered as monophyletic, all belonging to the South American SDTF clade. Among them, *Ceiba pubiflora* and *C. glaziovii* had the best sampling in this study (Figure 6.1), represented by 32 accessions. The topology inferred using the concatenation approach had the 13 individuals of *C. glaziovii* nested within the 19 individuals of *C. pubiflora* (Figure 4.1). The morpho-

logical investigation showed that there was no clear difference between the two species (Figures 4.7 to 4.9). The species delimitation analysis supported one delimited species in two of the three independent runs conducted. Together, this evidence suggests that *C. pubiflora* and *C. glaziovii* are the same species. Future work, sampling more individuals of other *Ceiba* species, and careful study of their morphology in the field and herbarium, should also be helpful in resolving issues of species delimitation, as discussed at the end of Chapter 4.

6.3 Biogeographic history of SDTF in South America

The results here show that the expected patterns of species age, monophyly, ecological and geographical structure reported for SDTF species belonging to the Leguminosae family (Pennington and Lavin, 2016) are only partially shared by one of the most characteristic tree genera of neotropical SDTF. One of the main contrasts is the deep crowns and relatively short stems for *Ceiba* species, as well as their lack of monophyly, within the South American SDTF clade. These differences might reflect the large areas where those species occur in the Brazilian Caatinga, the largest area of neotropical SDTF, which might cause lower extinction rates over evolutionary time. In contrast, the predictions made by Pennington and Lavin (2016) were made mostly based on species occurring in the smallest patches of SDTF (DRYFLOR, 2016), in the inter Andean valleys.

The dating analysis in Chapter 5 recovered the crown node of the South American SDTF clade at 25.3 Ma, reinforcing the evidence of the presence of SDTF in South America at least since late Oligocene (ca. 25 Ma), and possibly longer due to the placement of *Ceiba trischistandra* outside of this clade and the very old age (45 Ma) of this species. *Ceiba trischistandra*, *C. schottii* and *Neobuchia paulinae* were recovered with long stems, indicating that those lineages have been isolated over evolutionary time, and thus represent unique evolutionary history. This reinforces the importance of conservation of the small patches of SDTF spread across the Neotropics, which often represent museums of diversity (Pennington et al., 2010).

Similar patterns of old, long stems and shallow crowns were found for the rain forest species *C. pentandra* and *C. samauma*, which contradicts predictions made by Pennington and Lavin (2016). The same pattern found in *C. jasminodora*, a species characteristic of the campos rupestres is fascinating, because it is the oldest lineage yet reported for this biome, dating to 43 Ma (Figure 6.2).

Figure 6.2: *Ceiba jasminodora* in the campos ruprestres of Serra do Cabral State Park, Minas Gerais, Brazil. Photo: F. Pezzini.

The variation in placement of the basally divergent branches of the phylogenies inferred using the different pipelines and settings, which include *C. trischistandra*, complicate the inference of the ancestral biome for *Ceiba*. The ancestral biome could be wet forest, campos rupestres or SDTF, because all are represented amongst these basal lineages. Given the large amount of data deployed here, and thorough phylogenetic inference, these basal relationships in *Ceiba* may never be resolved clearly. However, it would be worth using different software or optimised versions of current software (eg. BPP 4.0 and above) applying coalescent analysis to further investigate the variation in topology and to conduct an ancestral biome reconstruction analysis in the context of a

wider phylogeny that densely samples related genera across Bombacoideae.

Bibliography

Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden*. 82:247.

Burrows M, Wheeler D. 1994. A block-sorting lossless data compression algorithm. Technical report, Technical Report Digital Equipment Corporation, Palo Alto.

Carlsen MM, Fér T, Schmickl R, Leong-Škorničková J, Newman M, Kress WJ. 2018. Resolving the rapid plant radiation of early diverging lineages in the tropical Zingiberales: Pushing the limits of genomic data. *Molecular Phylogenetics and Evolution*. 128:55–68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*. 24:332–340.

DRYFLOR. 2016. Plant diversity patterns in neotropical dry forests and their conservation implications. *Science*. 353:1383–1387.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*. 66:526–538.

Muñoz-Rodríguez P, Carruthers T, Wood JR, et al. (14 co-authors). 2018. Reconciling Conflicting Phylogenies in the Origin of Sweet Potato and Dispersal to Polynesia. *Current Biology*. 28:1246–1256.e12.

Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science*. 6:1–20.

Pennington RT, Lavin M. 2016. The contrasting nature of woody plant species in

different neotropical forest biomes reflects differences in ecological stability. *New Phytologist*. 210:25–37.

Pennington RT, Lavin M, Sarkinen T, Lewis GP, Klitgaard BB, Hughes CE. 2010. Contrasting plant diversification histories within the Andean biodiversity hotspot. *Proceedings of the National Academy of Sciences*. 107:13783–13787.

Simmons MP, Gatesy J. 2015. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular Phylogenetics and Evolution*. 91:98–122.

Yang Y, Smith Sa. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution*. 31:3081–3092.

In the following page: Vol. I, Part I, Fasc. See Urban Prancha 10. Publicado em 1906. Família Palmae (Arecaceae), SubFamília Ceroxylinae, Tribo Cocoineae, SubTribo Attaleeae Drude, Gênero *Cocos* L., Seção Arescatrum Drude, SubSeção Macranthae Drude, *Cocos coronata* Mart.

SILVA AESTU APHYLLA, QUAM DICUNT CAA-TINGA,

IN PROVINCIAE BAHIENSIS DESERTO AUSTRALI.