Title

Studies on the biology and evolution of liverworts: from genomes to biogeography

Author

Gonzalez Ramirez, Ixchel Sarahi

Publication Date

2024

Peer reviewed|Thesis/dissertation

Studies on the biology and evolution of liverworts: from genomes to biogeography

by

Ixchel Sarahí González Ramírez

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Brent Mishler, Chair
Professor Carl Rothfels
Professor David Ackerly

Spring 2024

Studies on the biology and evolution of liverworts: from genomes to biogeography

Abstract

Studies on the biology and evolution of liverworts: from genomes to biogeography

by

Ixchel Sarahí González Ramírez

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Brent Mishler, Chair

Most of our understanding about plant evolution comes from the study of angiosperms. Nevertheless, flower-free plants display an impressive diversity of morphologies, physiologies, and life-strategies that are not present in angiosperms. Bryophytes, in particular, differ in fundamental ways from angiosperms; notably they experience life as haploids, match their water content to that of the environment, and disperse through unprotected spores—like most ferns but in great contrast with seed plants. A deeper understanding of '*the bryophyte way*' of being a plant and its evolutionary implications will boost our ability to understand, model, and interpret diversity patterns across all land plants. In this dissertation, I explore different aspects of liverwort evolution, arguably the least studied among the three lineages of bryophytes.

In Chapter 1, I report the *de novo* assembly of a nuclear reference genome for *Calasterella californica*, a West Coast endemic liverwort. The assembly consists of 820 contigs with a coverage estimated at $41x$. The BUSCO score was estimated at 95% which indicates a good completeness. This assembly constitutes only the fifth nuclear reference genome for liverworts, a lineage of $\sim$ 7200 species. The genome of *C. californica* is 520 Mbp long, almost double the size of *M. polymorpha*'s genome, but similar in size to *Lunularia cruciata*'s. A comparison of *C. californica*'s genome to *M. polymorpha*'s genome suggests that *C. californica* has eight autosomes and one sexual chromosome as previously suggested. The reference genome was obtained from a female liverwort, thus carrying a U chromosome.

In Chapter 2, I report the assembly of the chloroplast genome of *C. californica* and use a dataset on the arrangement of coding regions in the chloroplast of liverworts to explore the potential of using genomic arrangement information for phylogenetic inference. The final chloroplast of *C. californica* is 122,592 bp long, and the assembly had a coverage of 1978 reads per bp. This plastome contains a total of 129 genes, including rRNAs and tRNAs. Overall, the chloroplast gene arrangement is very conserved across the 41 species of liverworts compared in this study. The most common differences observed, were the absence of genes

or introns. There were no gene order rearrangements in this dataset. When using the structural data to infer phylogenetic relationships, I obtained topologies that, regardless of the inference method, on one hand recover some large groups like complex thalloid liverworts, but on the other hand disagree with our current understanding of liverwort phylogenetics for other clades. And finally, I found that structural changes (*i.e.*, gene or intron loss) occur at a slower rate than nucleotide substitutions. But other structural changes, like changes on the type of tRNAs, have rates of change comparable to that of nucleotides.

In Chapter 3, taking advantage of the newly generated reference genome, I use a genomic dataset to study the genetic diversity and structure of *C. californica* across its distribution range in California. *C. californica* occurs across the state and it is notably absent in the Central Valley; in fact the Coastal individuals are genetically differentiated from the individuals occurring in Sierra Nevada ranges. Additionally, the individuals occurring in the Sonora Desert appear to be the most genetically differentiated group. Overall, the Southern part of California is the region with the most diverse genetic pool.

Finally, in Chapter 4, I develop a biogeographic model of evolution, PAw GeoSSE + J, that accounts for many of the challenges associated with studying an old, widespread, and vagile lineage like many liverworts. I apply this new model on a dataset of Aytoniaceae, the family that *C. californica* belongs to. Overall, the precision with which we can reconstruct the biogeographic history of Aytoniaceae is limited. But my results suggest that Aytoniaceae most likely originated in North America and became widespread during the Cretaceous, a time when continent configuration facilitated dispersal. The analysis suggests that during the Mesozoic, lineages of Aytoniaceae occupied Antarctica despite the fact that no extant Aytoniaceae occurs there. This inference is derived completely from the connectivity matrices that reflect the reachability of Antarctica during that period of time. Furthermore the parameter estimates obtained from the PAw GeoSSE + J analysis are biologically meaninful, suggesting for example that the rate of dispersal to adjacent regions was 10 times higher than the rate of dispersal to distant regions, and the rate of cladogenetic events during long distant dispersal was three times higher than during adjacent-dispersal.

A mi mamá y mi papá,

y a todos los micromundos de plantitas pequeñas.

Contents

List of Figures

List of Tables

Acknowledgments

The work that I did during my PhD program was supported by multiple people and institutions. First, I want to thank my advisor, Brent Mishler. Thank you for believing in me, allowing me to pursue my academic interests, and supporting me during this academic journey. I also want to thank the professors that were part of my Qualification Exam and Dissertation committees: David Ackerly, Bruce Baldwin, Ben Blackman, Cindy Looy, Brent Mishler, and Carl Rothfels. Having the opportunity to talk to you about science was among the most valuable experiences of my program.

Beyond the formal mentoring structure of the program, my thinking has been heavily shaped by everyday conversations and interactions. Thank you to my labmates Jenna Ekwealor and Javi Jauregui-Lazo, for being role models and friends. Daniel LaTorre, Tanner Frank, Isaac Krone, Cindy Looy and Ivo Duijnstee, thank you for thinking deep-time with me, and keeping my inner paleontologist alive. To the "Botany crew": Carrie Tribble, Jenna Ekwealor, Isaac Mark, Mick Song, Maryam Sedaghatpour, and Forrest Freund, for sharing your passion about the wonders of the plant world. Thank you to Mike May for teaching me phylogenetics from zero, to the point that I can equally enjoy and be frustrated by models and MCsMC; you have been one of my greatest mentors! Thank you to the Rothfels lab (especially Carl) for giving me a second academic family. Thank you to Sonia Nosratinia for all the help in the lab, the nuts and bolts of the dissertation worked thanks to you! Carl, Cathy, Carrie, Isaac, Jenna, and Mike—hard to explain my appreciation for the innumerable slack hours discussing anything from figures display, to social justice, to deep phylogenetic rabbit holes. Thank you to Atra, Lan, and Riya for trusting me as a mentor and sharing my interest on liverwort questions! Monica Albe—thank you so much for all your help navigating paperwork, you made me feel supported. Thank you to David Long for initially introducing me to *C. californica*, and sharing his taxonomic expertise with me. I appreciate the excellent feedback from Michael Landis on my biogeographic model chapter.

Technical and field support. Thank you to all the people that joined me during my fieldwork: Debora Brandt, Daniel LaTorre, Tanner Frank, Devon Comito, Isaac Krone, Mike May, Rodrigo Monjaras-Rueda and John McLaughlin–You made fieldwork (much of it during difficult pandemic times) so enjoyable! Thank you to Forrest Freund, Isaac Krone, John McLaughlin, Brent Mishler, and Keir Wefferling for providing collections of *C. californica*. Huge thanks to Diane Ikeda, who facilitated obtaining collecting permits for the National Forests of California.

Sonia Nosratinia and Forrest Freund—thank you for training me on DNA extraction and quantification protocols. Thank you to Carl Rothfels and Klara Scharnagl for allowing me to use your lab space to perform some of my labwork. Thank you to José Vásquez Medina and Kaitlin Allen for providing liquid nitrogen for my lab work. The sequencing of my samples was carried out at the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01.

My research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer). Additionally, thank you to Will Freyman for allowing me to use his cluster "Io" for conducting long runs of my biogeographic model.

Thank you to Erik Enbody, Merly Escalona, Courtney Miller, Daniel Oliveira, Brad Schaffer, Erin Toffelmier, and the rest of the CCGP team for their roles in processing the samples of *C. californica*. In particular, thank you to Merly Escalona for her time and expertise during the assembly of the reference genome of *C. californica*. Additionally, thank you to Erik Enbody for running the CCGP SNP calling pipeline on my dataset.

Funding sources. I was supported by a doctoral fellowship (beca num 709967) from UC-Mexus and CONAHCYT. I received the 2022 Plant Science Fellowship by Oak Spring Garden Foundation (OSGF), and wrote the fourth chapter of my dissertation during a residency at OSGF. I received the 2022-2023 Philomathia Graduate Fellowship in Environmental Sciences, at UC Berkeley. This fellowship allowed me to focus on my dissertation for a semester. The field- and molecular work of this dissertation was supported by a California Conservation Genomics Project (CCGP) Grant. CCGP is funded by the State of California, led by the UCLA/ La Kretz Center for California Conservation Science. I received a Dissertation Completion Award by the Integrative Biology department during Spring 2024. This award was essential to completing this dissertation's writing.

Friends and family. A few words aren't enough to express how thankful I am for having such wonderful friends and family. And if you know me, you know we risk this section to become three pages long, and accompanied by tears flooding, so let's not! But seriously—I trust that you feel my love!

Mamá, Papá, y Alma—No hay palabras que alcancen para agradecer su infinito apoyo y amor... ¡Los amo!

Daniel and Debora [and Raulzihno!], thank you for being my home far from home! Carrie, thanks for believing in me all along! Mike, not gonna write anything cause you would hate it, but you know! Emily Lam, you are such a loving friend, thank you for everything! Devon—the best outdoors adventures are always with you! Laura—¡Gracias por siempre estar! Tanner, Kat, Isaac, Kaitlin, Mat, Rachael, Jenn, Potions 8, Ryan, Leah, and David—I have made wonderful memories with you, thank you! Mis latinas de IB, ¡gracias por los apapachos! To the coolest cohort of IB—thank you for six years of adventures! Emily Chen—OK, fish are pretty cool! Cin and Ivo—you are probably the coolest academics I know! Xoch, Diego, peque, Torres, Isa, Natha—cambian los tiempos y lugares, pero los sigo queriendo! Connor and Ben—we're pretty good housemates, thank you for our supportive

house! My basketball crew (B4all and ABCD)—you don't even know, but I'm sane thanks to you!

Patata—I am truly the luckiest human in the world, you make me so happy! And although you will never read this, I hope all the pets, treats, adventures, and attention make you feel loved!

Introduction

More than any other group, photosynthetic organisms have shaped the evolution of life on Earth (Beerling, 2017). Around 2.7 billion years ago (Sessions et al., 2009), cyanobacteria transformed Earth's atmosphere from reductive to oxidant, causing a wave of extinction across anaerobic lineages of bacteria. In terrestrial ecosystems, land plants paved the way for the evolution of terrestrial animals around 500 Mya. Since then, plants have been a resource and a habitat for other organisms. And when woody plants took to the skies, they created a third dimension of habitat for other forms of life, including other plants. Today, there are close to 400,000 accepted species of plants. Of those, almost 300,000 are flowering plants (Christenhusz and Byng, 2016). Because flowering plants are undoubtedly dominant in most ecosystems, a lot of our sophisticated understanding of plant biology and evolution (*e.g.*, genomics, population genetics, regulatory mechanisms, models of selection, models of evolution) is based on the study of angiosperms. Nevertheless, flower-free plants display an impressive diversity of morphologies, physiologies, and life-strategies that are not present in angiosperms. And if we consider that during 3/4 of land plant history angiosperms didn't exist, it becomes clear that in order to understand plant evolutionary history, we need to better understand the 'flower-free way' of being a plant.

Liverworts (Fig. 0.1) are one of the three lineages of bryophytes, and potentially, the sister group of all the remaining land plants (Cox, 2018a; Puttick et al., 2018; Rensing, 2018; Finet et al., 2010; Renzaglia et al., 2000). Like other bryophytes, they are comparatively small, with individuals growing usually up to a couple centimeters at the most. Although small, liverworts have important ecological roles, from being abundant elements of epiphytic cover in forests—and thus having a key role in water cycling—to being part of the biocrust that facilitates soil formation in dry ecosystems, like shrublands and deserts. The 7,000 described species of liverworts (Söderström et al., 2016) occur in almost all terrestrial ecosystems, from the coast of Antarctica, to the tundra of the Northern hemisphere. And many liverwort species have remarkably wide geographic distributions, often covering several continents (Laenen et al., 2016; Shaw, 2001), raising the question of how such small organisms can maintain such broad distributions, and whether species concepts are comparable with those in vascular plants.

Not only are liverworts diverse and widespread, but they have multiple unique features which, together with the haploid-dominant life cycle characteristic of bryophytes, makes the 'liverwort way' of being a plant unique. For example, liverworts possess exclusive membrane

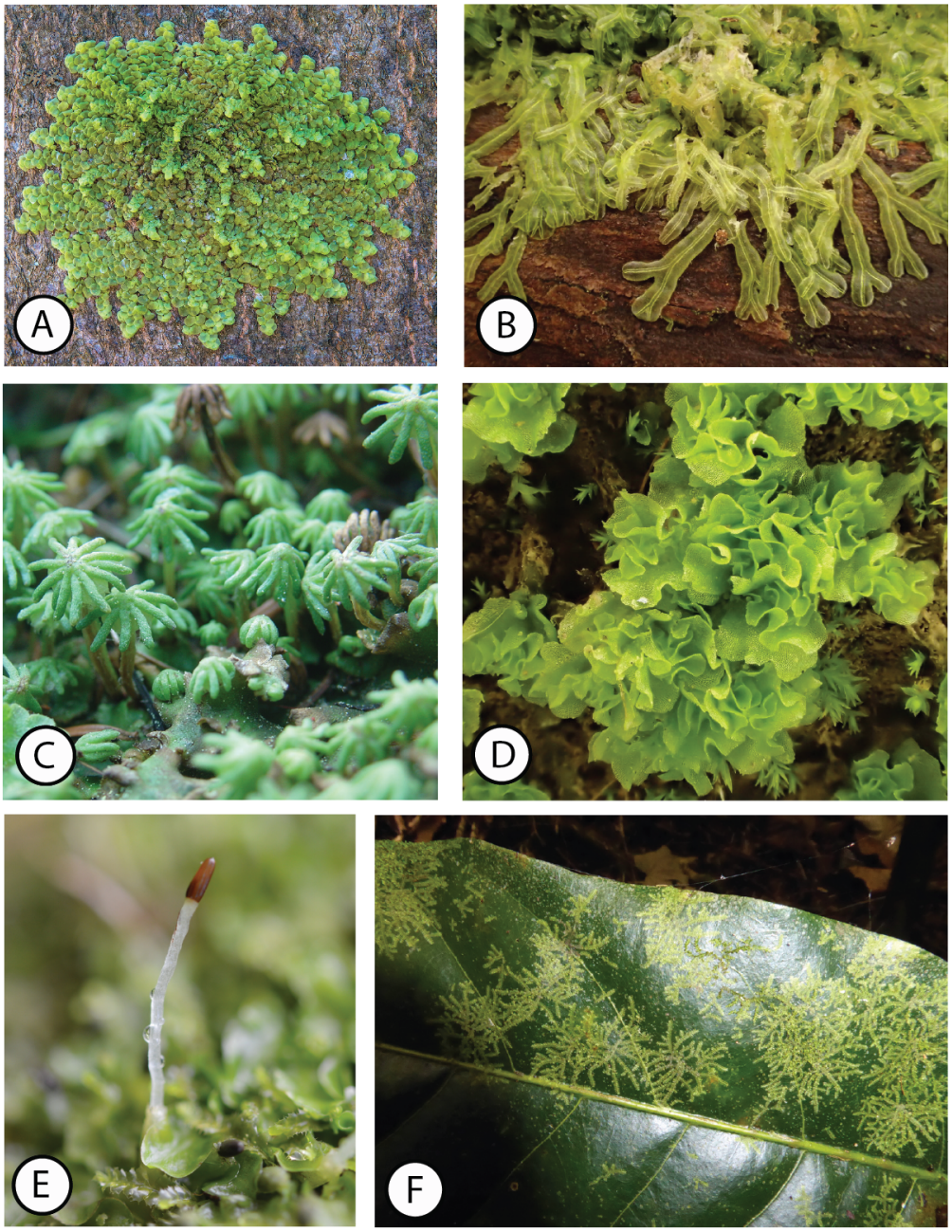


Figure 0.1: Liverworts display unique morphologies among plants. **A**. *Radula complanata* living on the bark of a tree. **B**. *Metzgeria furcata*, a simple thalloid liverwort living on a tree trunk. **C**. *Marchantia polymorpha*'s umbrella-shaped reproductive structures. **D**. *Fossombronia longiseta* on soil. **E**. A reproductive *Monoclea gottschei*, the delicate reproductive structures are characteristics of liverworts. **F**. *Lejeuna epiphylla* living on a leaf. Image credits: A—Vladimir Bryukhov, B—Penelope Noel Gillete, C—Sergio Díaz-Martínez, D—Ken Kellman, E—Daniela Perez Orellana, and F—Peter de Lange.

bounded organelles that store essential oils. And although the function of these organelles is unclear, it has been hypothesized that they play a role in deterring herbivores, which is consistent with the unusually low evidence of herbivory in this group. Like other bryophytes, liverworts experience most of their life as haploid organisms, reducing the opportunity for masking fitness effects of alleles, and potentially increasing the efficacy of natural selection. Additionally, unlike plants with pollen—whose sperm may be transported long distance—liverwort sperm disperses only as far as it can swim through water. Conversely, liverwort spores are comparatively much smaller than most seeds, potentially increasing dispersal ability through wind. While we have reasons to believe that 'the liverwort way' of being a plant affects their diversity patterns and evolutionary mechanisms, historically, there have been very few studies addressing these questions.

In recent years, liverwort diversity (Söderström et al., 2016), phylogenetics (Forrest and Crandall-Stotler, 2004; Schill et al., 2010), evolution (Villarreal A. et al., 2016; Heinrichs et al., 2007) and paleobiology (Hernick et al., 2008; Labandeira et al., 2014; Heinrichs et al., 2018) have received more attention. With this increase in available data for liverworts, and the constant development of new methods to study evolution, there are increasingly more options to study *if* and *how* 'the liverwort way' of being a plant has ecological and evolutionary implications. In this dissertation, I study different aspects of the biology of liverworts, from the molecular level of genomes, to the deep-time biogeographic evolution of the more than 100 My old family Aytoniaceae. First, in Chapter 1, I describe the assembly of a new nuclear reference genome for *Calasterella californica*, a West Coast endemic liverwort, which was essential for the next two chapters. In Chapter 2, I use the data generated in Chapter 1 to assemble the chloroplast genome of *C. californica* and explore the patterns of structural changes in the plastome of liverworts and whether or not they could be used for phylogenetic inference. Next, in Chapter 3, I use a whole genome approach to study the genetic structure of the *C. californica* populations by sampling 110 individuals of this liverwort across California. And finally, in Chapter 4, I zoom out (in space, time, and phylogenetically) to study the biogeographic history of the family Aytoniaceae, the old and widespread lineage to which *C. californica* belongs. This last chapter involved the development of a new phylogenetic model that combines geological and biological information to study groups that have historically remained challenging for biogeographic reconstruction (*e.g.*, groups with widespread, ancient, and very vagile species).

Chapter 1

A reference genome for *Calasterella californica*

1.1 Introduction

Liverworts diverged from the rest of the land plants more than 400 mya. While they retain some general ancestral features (*e.g.*, their water relationships and life cycle), they have evolved a unique way of being a land plant, with many novel features such as a unique organelle (oil bodies), reproductive structures (*e.g.*, carpocephalla and elaters), and vegetative structures (*e.g.*, pores and ventral scales). Although they have undergone more recent periods of diversification (Laenen et al., 2014), many of their major morphological features have remained virtually unaltered since the Mesozoic according to the fossil record (Hernick et al., 2008). Despite many studies documenting structural features of liverworts, their genetic basis remains understudied. Up till now there are only four liverwort species with available nuclear reference genomes: *Marchantia polymorpha*, *Marchantia paleacea*, *Marchantia inflexa*, and *Lunularia cruciata*. In this chapter, I report the assembly of the nuclear genome of *Calasterella californica* (Hampe ex Austin) D.G. Long & T.X. Zheng, which represents the first reference genome within the family Aytoniaceae, and only the third genus (and 5th species) of liverwort to have a reference genome.

Calasterella californica is the sole species in the recently described genus *Calasterella* (Long and Zheng, 2023) and is endemic to the California Floristic Province on the west coast of North America, from southern Oregon to the northwest of Baja California (Mexico), including Guadalupe, Cedros, and the Channel islands (Fig. 1.1). It is one of the most common liverworts within this region, occurring in outcrops and cliffs, at low to moderate elevations from the Coast Ranges and Sierra Nevada, to the Sonoran and Mojave Deserts. *Calasterella californica* shows evidence of adaptation to desiccation (*e.g.*, a specialized folding mechanism and scales to protect tissues when dry) that might be associated with its ecological success in such a diverse array of ecosystems. *Calasterella californica*—together with other biotic soil crust organisms like bacteria, lichens, and mosses—plays an important role in

soil formation by retaining and compacting sediment particles, as well as providing a seed bed for flowering plants and a microhabitat for small animals. *Calasterella californica* is the only bryophyte, a group of very diverse and ecologically important lineages, studied in the California Conservation Genomics Project (Shaffer et al., 2022). *Calasterella* is on its own deeply diverging phylogenetic branch, diverging from extant relatives approximately 98 mya (as shown in Chapter 4), which is deeper than the age of the coast redwood-giant sequoia split (about 72 mya; Liu et al. (2022)), thus this species certainly qualifies as a paleoendemic in the California Floristic Province along with those iconic trees. The availability of this reference genome will facilitate the study of the unique features of *C. californica* and other liverworts, and pave the way towards a comparative understanding of liverwort genomes.

1.2 Methods

Biological materials. Given its small size, clustered growth pattern, and tight contact with the soil, I decided to grow individuals of *C. californica* in culture to avoid soil contamination and allow the use of material from a single genotype. I isolated spores from a single sporangium from an individual of *C. californica* from the collection IGR150. That collection was made on January 29th 2021 in Los Padres National Forest, along Jesusita trail (Coordinates: 34.4722 N, 119.707367 W) under collection permit Bot-4-2021; voucher specimen is deposited in the University Herbarium at UC Berkeley. The spores were placed in sterile petri dishes with Hoagland's nutrient agar (modified from Hatcher 1965), and kept in a Conviron 3244 growing chamber at 12°C. After germination, I separated individual thalli into 10 individual petri dishes; thus each dish corresponded to a single unique genotype, coming from a single spore. Because in this species thallus fragments have the ability to regenerate new thalli, as soon as each thallus reached 4 mm in size, they were cut in smaller pieces in order to increase the amount of tissue for library preparation. For the HiFi library, we used 260 mg of young thallus tissue from the genotype G1 (sample IGR150-G1).

DNA extraction, sequencing and genome assembly. The high molecular weight DNA extraction was performed by the CCGP team following the protocol by Inglis et al. (2018). Then, they were sequenced using a PacBio HiFi long read sequencer and the assembly was done by the CCGP team according to the standard CCGP protocol described in `https://github.com/ccgproject/ccgp_assembly` using as input 2.1 million PacBio reads. A summary snail plot (Fig. 1.2), which visually displays the cumulative scaffold length and assembly statistics, was generated with bloobtools (`https://blobtoolkit.genomehubs.org`). And finally, to compare *C. californica* assembly to the high quality chromosome level assembly of *Marchantia polymorpha*, I aligned the *C. californica*'s scaffolds against *M. polymorpha* genome using the dotplot online visualizer (`https://dot.sandbox.bio`; Fig. 1.2).

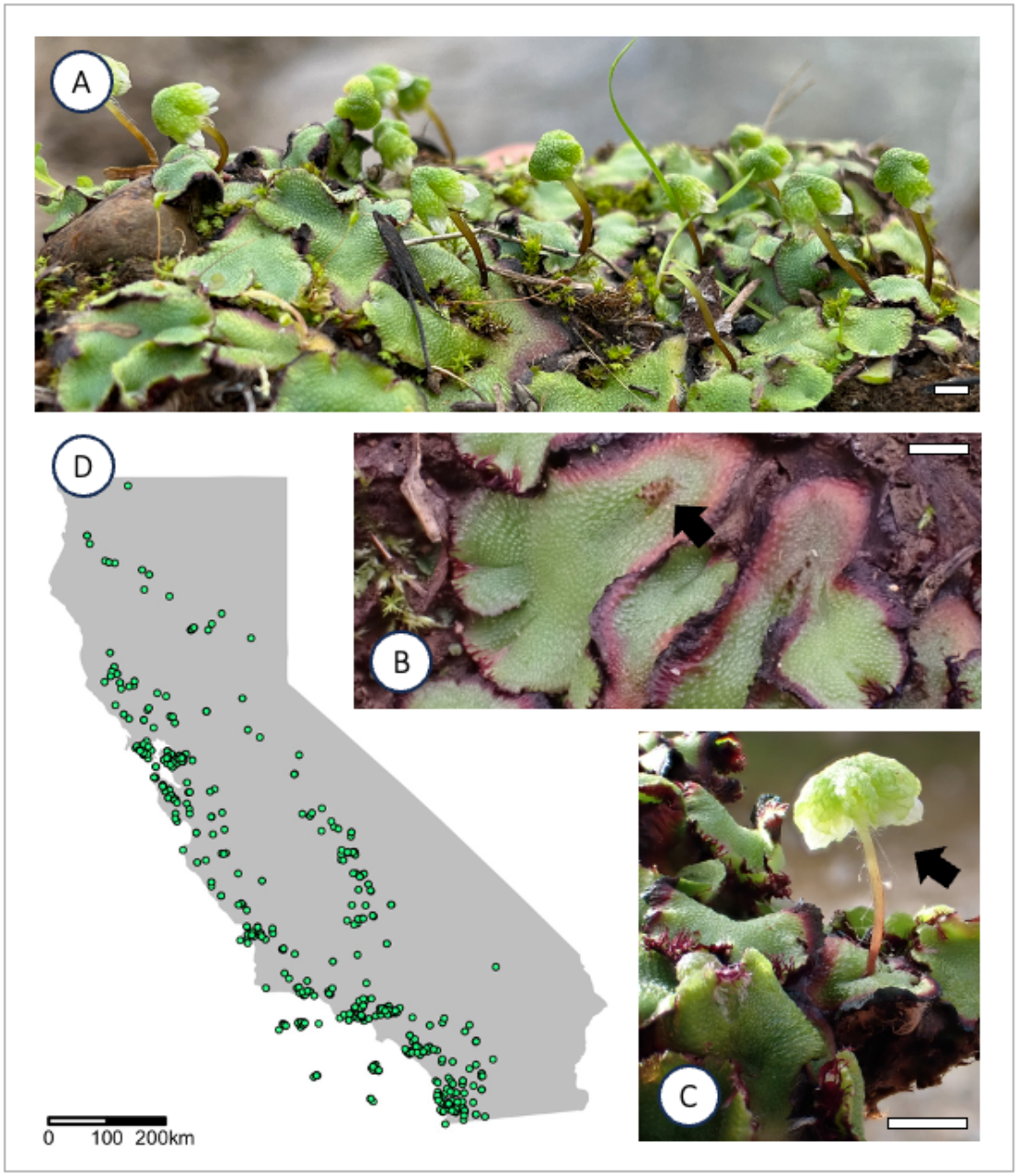


Figure 1.1: ***Calasterella californica.*** **(A)** *C. californica* grows in clumps on soil. It is a dioicous species with **(B)** male plants that produce sperm in antheridia (black dots), and **(C)** female plants that produce elevated archegoniophores (umbrella-like structures) that contain the eggs and eventual sporophytes. **(D)** Map showing all the herbarium and iNaturalist research grade records for *C. californica* in California. The range of the species slightly extends to Oregon and Baja California as well. Scale bars in pictures correspond to 5 mm.

1.3 Results

The haploid nuclear genome assembly of *Calasterella californica* (cmAstCali1) consists of 820 contigs, with a total length of 520 Mb, and a contig $N50 = 32Mb$ (Fig. 1.2). The assembly is available at NCBI at https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036924285.1. The read coverage was calculated as 41-fold based on the genome size of 703 Mbp reported by Bainard et al. (2013). And according to the Benchmarking Universal Single-Copy Orthologs (BUSCO) score—which compares the set of genes in the assembly with a database of very conserved ortholog genes—this assembly has a completeness score of 95.1%, which is considered good.

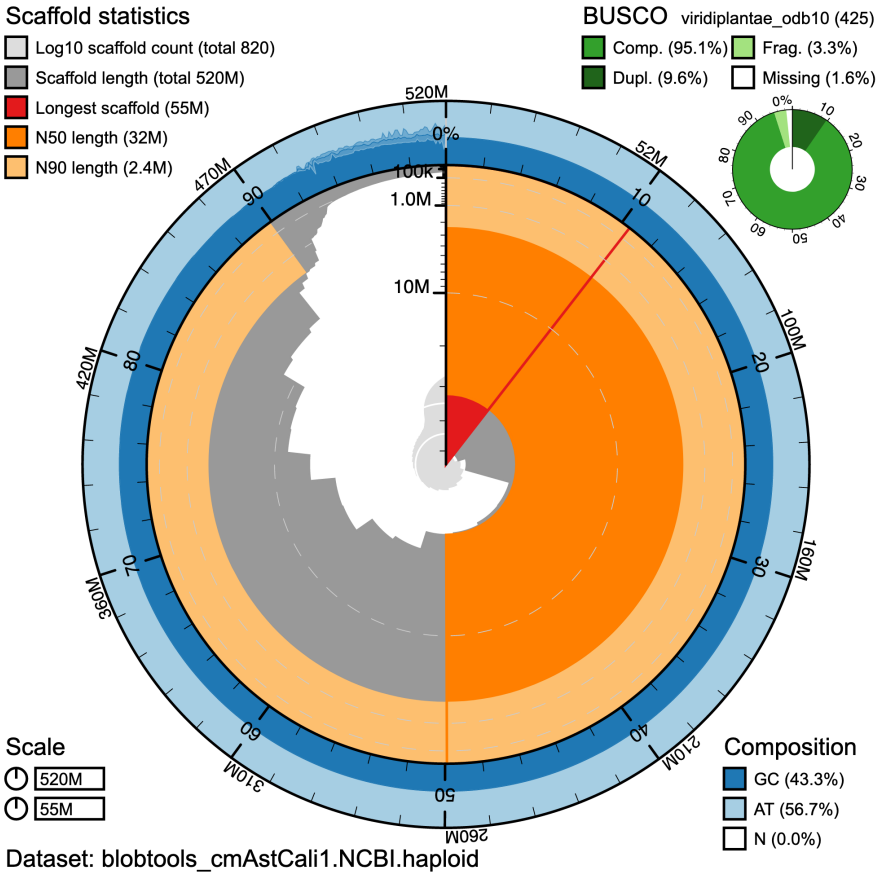


Figure 1.2: ***C. californica* nuclear genome assembly.** The main plot shows the scaffold statistics of the nuclear assembly. The grey internal circle represents the cumulative length of scaffolds, and the orange layer marks the N50 and N90 length of the assembly. The green circle in the top right represents the BUSCO completeness score based in a comparison to the Virdiplantae dataset.

Based on the alignment of *C. californica* scaffolds against the chromosome level assembly of *Marchantia polymorpha* (Fig. 1.3), *C. californica* also appears to have nine chromosomes (eight autosomes and one sex chromosome). Additionally, based on the poor alignment of the *C. californica*'s sex chromosome to the V sex chromosome of *M. polymorpha*, we can infer that we sequenced a female individual carrying a U chromosome.

1.4 Discussion

This assembly constitutes only the fifth reference genome for liverworts, a lineage of 7200 species (Söderström et al., 2016). It is also only the third genus of liverwort that has a reference genome, after *Marchantia* and *Lunularia*. While the expected size according to flow cytometry was 704 Mbp (Bainard et al., 2013), our assembly was shorter (520.2 Mbp). When compared to the size of other assemblies, *C. californica*'s genome size is similar to that of *Lunularia cruciata* (580 Mbp), while both of these genomes are almost double the size of genomes assembled for *Marchantia polymorpha* and *Marchantia paleacea*, which range from 223 to 238 Mbp in size (Bowman et al., 2017; Linde et al., 2023). In the case of *L. cruciata*, the larger genome size was explained by a high content of transposable elements (Linde et al., 2023), so it would be interesting to further investigate the proportion of these elements in *C. californica*.

Even if this *C. californica* reference genome is not a chromosome level assembly, it is good quality and was essential for aligning the short-reads of more than 100 samples of *C. californica* in order to study the population structure of this species (Chapter 3). Furthermore, since *C. californica* is the only desiccation tolerant liverwort with a reference genome, the door is open to study the genetic basis of desiccation tolerance in this group.

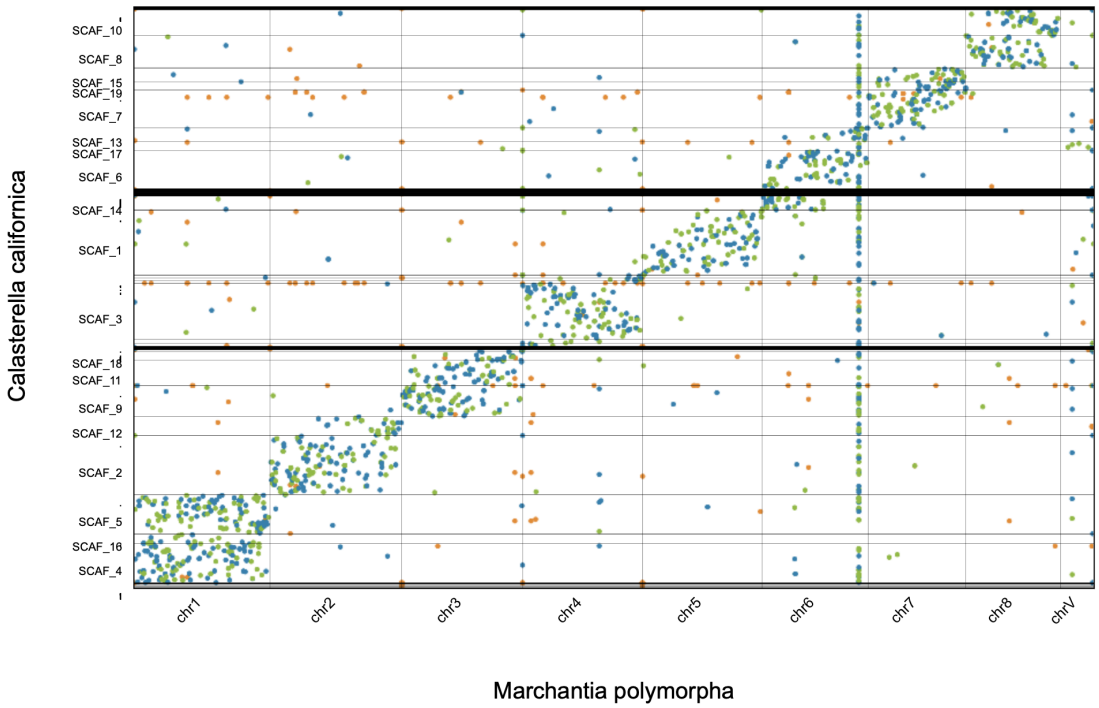


Figure 1.3: **Alignment of *C. californica* scaffolds against *Marchantia polymorpha* chromosome level assembly.**

Chapter 2

The chloroplast genome of *Calasterella californica* in relation to structural genomic evolution in liverworts

2.1 Introduction

Most models of molecular evolution focus on the evolution of DNA (or amino acid) sequences through changes in single units (nucleotides or amino acids), *i.e.*, the gain, loss or substitution of units. Nevertheless, when looking at large sequences of DNA, it is clear that there are some structural changes that involved the gain, loss, or inversion of one or multiple genes. These structural changes are often cited as diagnostic of large clades, for example, in the chloroplast of land plants there is an inversion of the *ycf2-psbM* region in the euphyllophytes (ferns + seed plants), an inversion in the *trnG-trnT* region of the ferns, and independently evolved expansion of the IR in angiosperms and gymnosperms (Mower and Vickrey, 2018). As these structural changes often seem to be synapomorphies for major clades, Kelch et al. (2004) coded a total of 40 genome structural changes for 18 species of green plants, and conducted a parsimony-based phylogenetic inference. Their results suggested that at this broad taxonomic scale, structural changes observed in the chloroplast can be used to make phylogenetic inferences.

The number of organellar genomes available has increased recently due in part to the development of new sequencing tools like Illumina or PacBio that facilitate the sequencing of genomic DNA extractions, in which organellar genomes are over-represented in comparison to the nuclear genome—since a single cell usually contains many organelles. For liverworts, this means that the number of whole chloroplast genomes available has gone from 5 to ~ 80 in the last five years, thanks to recent work by Yu et al. (2019), Dong et al. (2021), and Xiang et al. (2022). Compared to the nuclear genome, organellar genomes are small, circular,

uni-parentally inherited, and overall maintain a relatively conserved architecture and gene content. The chloroplast genome of land plants has a highly conserved circular structure that comprises four regions: two single copy regions—the large single copy (LSC) and the short single copy (SSC)—and two identical copies of an inverted repeat (IRA and IRB) that separate the single copies (Mower and Vickrey, 2018).

With the increase of liverwort chloroplast genomes available, and because the number of genes in the chloroplast is only about 150, the chloroplast offers a great opportunity to carefully examine the evolution of gene arrangement across a large taxonomic group. A better understanding of the structural differences among lineages and the evolutionary rates of structural rates in comparison to nucleotide evolution will help to assess the value of using these genomic features as phylogenetic characters, and has the potential to inform better models for the evolution of genome structure. In this chapter, I (1) report the *de novo* assembly of the chloroplast genome of *Calasterella californica* from HiFi long reads obtained as part of the ongoing California Conservation Genomics Project (CCGP, Shaffer et al., 2022); (2) compare it to other known liverwort chloroplast genomes; and (3) study the pattern and rates of evolution of structural changes of liverwort chloroplast genomes.

2.2 Methods

De novo assembly of *Calasterella californica* chloroplast genome

Biological materials and DNA extraction. To ensure that the tissue used for DNA extraction was as clean as possible, *Calasterella californica* thalli were grown from spores collected in the field. I isolated spores from specimen IGR150, collected in Los Padres National Forest, California (Coordinates: 34.4722 N, 119.707367 W; voucher deposited in the UC Herbarium), under permit Bot-4-2021. Once the spores germinated, I isolated an individual spore and increased the amount of tissue by manually dividing the thalli; thus the sequence data were from a single genetic individual. The DNA extraction was made from 260 mg of the G1 genotype according to the CCGP protocol.

Assembly of the chloroplast genome. The chloroplast assembly was done using the `Oatk` pipeline (`https://github.com/c-zhou/oatk`) for *de novo* assembly of organelle genomes from PacBio HiFi data. It was annotated using the online version of `GeSeq` (Tillich et al., 2017). I then visualized the draft plastome in Geneious and identified two misassembled regions, located at each extreme of the inverted repeat (Appendix A.1). Each error consisted of the duplication of a small region of the inverted repeat. I report in Appendix A.1 the sequences that were deleted. The final assembly was re-annotated using `GeSeq` and visualized using the online version of `OGDRAW` (Greiner et al., 2019).

Structural evolution of liverwort plastomes

Assessing structural variation. To compare the structure of the chloroplast genome of *C. californica* to other liverwort plastomes, I sampled plastid sequences across liverworts from GenBank. For the family Aytoniaceae I included all the plastomes available. For the rest of the liverworts, I selected one of each of the plastid structural variants identified by Dong et al. (2021, fig. 1). Most of the chloroplast genomes for Aytoniaceae were generated by Xiang et al. (2022). Although these authors reported Genbank accession numbers, they were not yet available in the NCBI database. Fortunately, the authors provided the raw sequences of their assemblies as supplemental materials, so I used those sequences and re-annotated them. In Appendix A.2, I report the 41 species that I used, and whether the sequence data was obtained from Genbank or from Xiang et al. (2022). Next, I used `GeSeq` (Tillich et al., 2017) to annotate these plastomes, using the *Marchantia polymorpha* (`NC_042505.1`) chloroplast genome as the reference sequence. Once annotated, I visualized each plastome in `Geneious` and recorded (1) gene, rRNA, and tRNA presence or absence, (2) intron presence or absence, and (3) region size and GC content.

Phylogenetic analyses of liverworts using structural changes in the chloroplast. I used the gene and intron presence/absence data to infer the phylogeny of liverworts. Structural variation in the plastome resembles morphological data in that it can potentially be multistate (*i.e.*, multiple variants of a tRNA across the species, or a gene being present, absent or in a pseudogene state). But unlike morphological data, these structural changes do not have an ascertainment bias, since I scored all the genes present in the chloroplast genome. Therefore, I decided to compare the results of two commonly used approaches for this type of data: parsimony-based inference and Bayesian inference. The parsimony analysis was performed using the ratchet algorithm in `R`, as implemented in the package `phangorn` (Schliep, 2011). The Bayesian analysis was implemented in `RevBayes` (Höhna et al., 2016), where I implemented a model of trait evolution partitioned by type of data (gene presence/absence *vs* intron presence/absence) and by number of states. I allowed asymmetry in the stationary frequencies of binary states (*i.e.*, an F81 model) to account for the rarity of re-gaining a coding region that has been lost, and a standard MK model for the multistate partitions. The final analysis had a total of six partitions. I assessed the convergence of the analyses using `Tracer` (Rambaut et al., 2018). The plots were produced in `R` (R Core Team et al., 2013) and required the libraries `ape` (Paradis et al., 2004), `dplyr` (Wickham et al., 2019), `ggplot2` (Wickham and Wickham, 2016), `ggdist` (Kay, 2023), `ggpubr` (Kassambara, 2018), `phytools` (Revell, 2012), and `tidyverse` (Wickham et al., 2019).

2.3 Results

The chloroplast genome of *C. californica*

General features. The final chloroplast assembly of *C. californica* (Fig. 2.1) is 122,592 bp long, with a mean coverage of 1978 reads per bp (Appendix A.3), and 28.8% GC content. The large single copy region (LSC) is 82,287 bp long, the small single copy region (SSC) is 20,030 bp long, and the inverted repeat is 10,137 bp long. This plastome contains a total of 129 genes including rRNAs and tRNAs (not counting the duplication of the IR). The LSC contains 73 protein coding genes and 29 tRNAs, while the SSC contains 16 protein coding genes and two tRNAs. The inverted repeat contains five tRNA and four rRNA genes in the following order: *trnV*, *rrn16*, *trnI*, *trnA*, *rrn23*, *rrn4.5*, *rrn5*, *trnR*, *trnN*.

Comparison with other liverwort chloroplasts. In terms of size, the chloroplast of *C. californica* falls within the size range reported for other members of the family Aytoniaceae. In comparison with other liverworts the overall plastome size, SSC, and IR regions of Aytoniaceae are relatively large (Appendix A.4). The chloroplast genome structure of *C. californica* is almost identical to the chloroplasts of other Aytoniaceae (Fig. 2.2). In fact, only the presence/absence of *ndhF*, *trnK*, *ndhB*, and a pseudogenized copy of *psbD* in the LSC region vary among members of Aytoniaceae (Fig. 2.2). But when compared to a broader phylogenetic sample of liverworts, out of the 140 genes potentially present in the chloroplast for this group, there are 47 with a structural variant (*i.e.*, absent, pseudogenized, or transformed into a different tRNA; Fig. 2.2). Most of the variation in gene content occurs in the LSC and the SSC, while the IR remains very conserved; this pattern coincides with the larger size variation observed in the LSC and SSC compared to the IR (Fig. 2.2) and with the reduced substitution rates in the IR (Li et al., 2016). While the most common change is the complete absence of individual genes, the second most common type of variant is a change in the identity of tRNAs. In fact, some tRNAs have up to six different variants at the same position (Fig. 2.2). Among the sampled liverworts, *Treubia lacunosa* and *Aneura mirabilis* stand out for having the smallest number of genes in the chloroplast (Fig. 2.2). In addition to variation in gene content, I identified 30 introns, of which 10 are present in all the species sampled (Fig. 2.3). Many introns were present in tRNAs, and some genes, like *pafI*, *rpoC1*, *rpoC2*, and *ycf2*, contained more than one intron (Fig. 2.3).

Phylogenetic inference of liverworts using structural changes

The final structural dataset used for phylogenetic inference was a matrix of 170 characters (140 genes + 30 introns). Fig. 2.4 contrasts the most parsimonious tree with the Bayesian maximum clade credibility tree. Overall, the results of the two analyses are highly congruent, but do not completely coincide with our current understanding of liverwort phylogenetics. For example, groups that are monophyletic on nucleotide-evolution based inferences are not

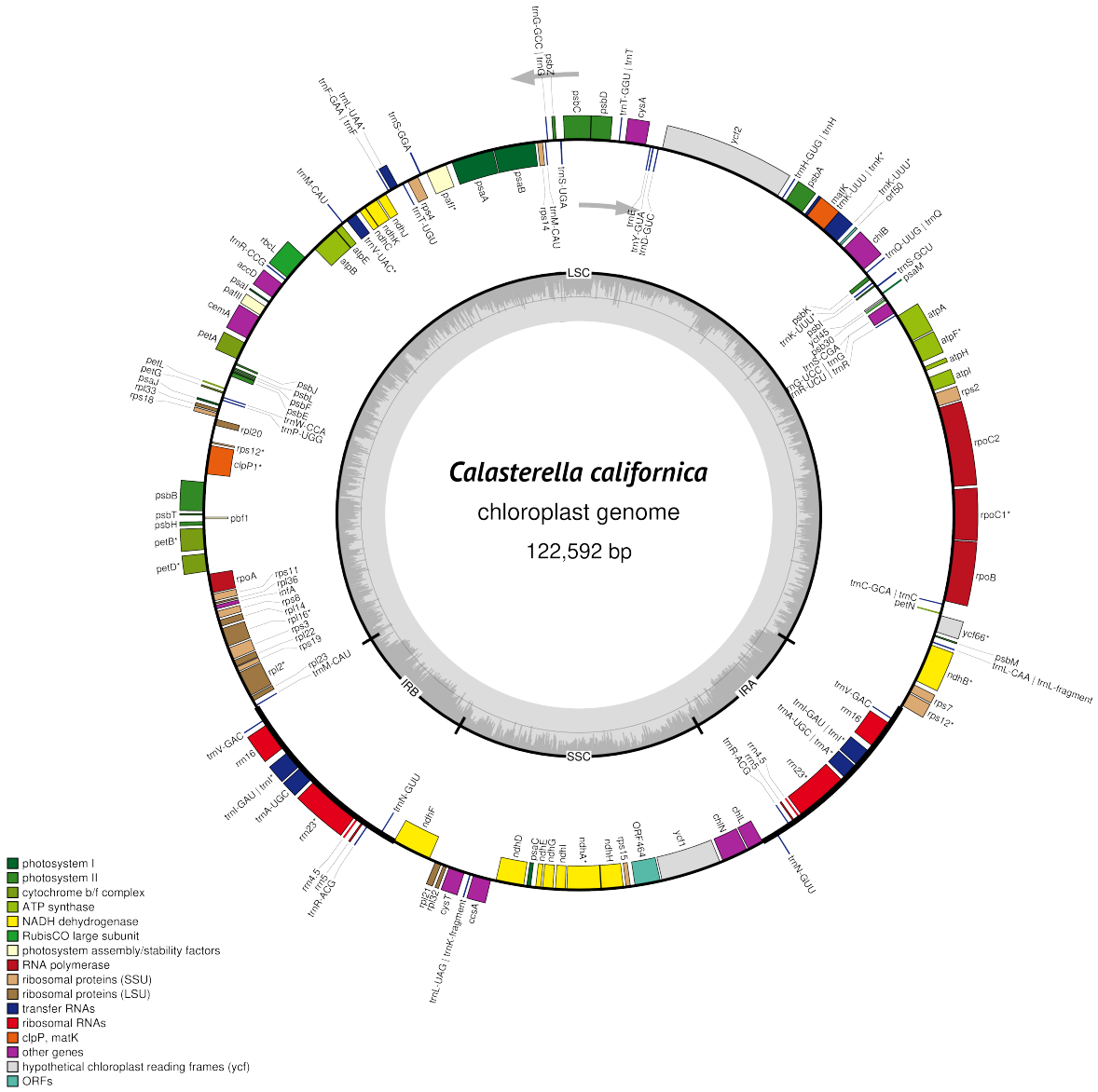


Figure 2.1: **Circular gene map of the chloroplast genome of *Calasterella californica*.** The inner circle indicates the GC content, and identifies the different regions of the genome: the large-single copy (LSC), the short single-copy (SSC) and the two inverted repeats (IRA and IRB). The external circle indicates the genes color coded by function. Genes inside the circle and outside the circle are transcribed in different directions (see arrows). Genes that contain introns are marked with an asterisk (*). This figure was produced with the software OGDRAW (Greiner et al., 2019).

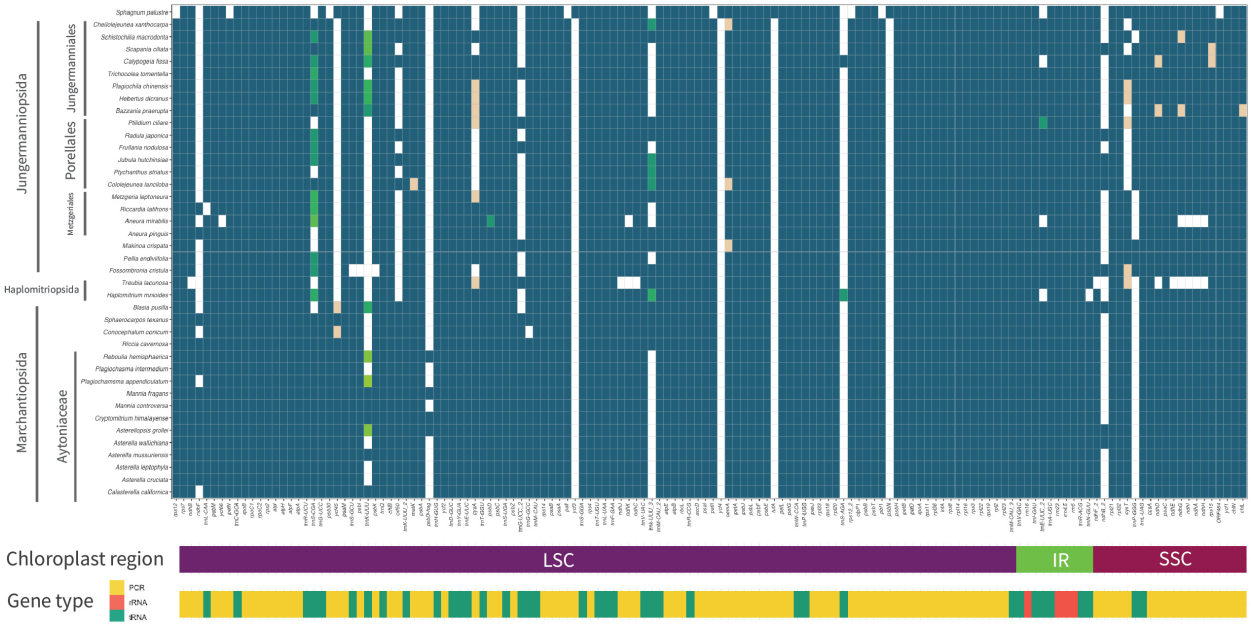


Figure 2.2: **Comparison of the coding regions in the chloroplast of liverworts.** Cells in blue indicate the presence of a coding region, while cells in white indicate the absence of it. Green cells indicate a variation on the tRNA produced by that coding region. Yellow cells indicate pseudogenization.

recovered as monophyletic in this study, notably the leafy liverworts (Jungermanniopsida), and *Treubia* + *Haplomitrium*.

The evolution of structural changes in chloroplast liverworts

The ancestral state reconstructions (Appendices A.6 and A.7) show that while there are multiple genes (*e.g.*, *petN*, *psb30*, *pafI*, *rps12*, *pbf1*, *ycf3*, *ycf4*, *tufA*, *psbN*) and introns (*e.g.*, *trnL*-intron, *trnV*-intron, *slP1*-intron, *petD*-intron, rpl2-intron, pafI-intron1) that are present in the chloroplast of all liverworts, the history of genes that get lost from the plastome (or otherwise change via pseudogenization, or change of identity in the case of transfer RNAs) varies in complexity. For example, there are gene losses that occur only once in the history of liverworts (*e.g. rps15*, *trnS*-GCU, *psbI*, *psbK*, *ycf66*, *matK*, *psbD*, *chlL*, *pafI*-intron2, *ycf2*-intron4). But there are also multiple examples of genes and introns that have been lost multiple times (*e.g. rpoC1*-intron 2, *cysT*, *ndhD*, *ndhG*, *cemA*). Interestingly, some of the more complicated histories occur in tRNAs, where we observe that different species have a tRNA in the same position, but encoding for up to six different types of tRNAs (Fig. 2.2).

The rates of evolution of gene gain and loss (structural partition 1 in Fig. 2.5) are lower

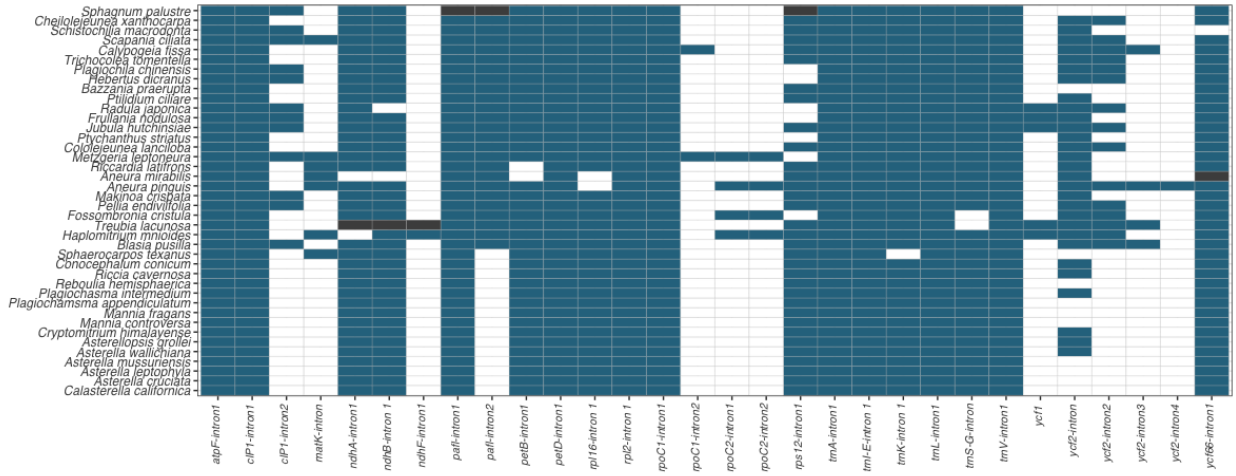


Figure 2.3: **Intron content in the chloroplast of multiple species of liverworts.** Blue cells indicate the presence of the intron, while white cells indicate the intron is absent. Black cells indicate that the coding region is absent, so the coding of introns is not applicable.

in comparison to the rest of structural partitions and nucleotide evolution. Overall, the higher the number of possible states of a gene (this is almost exclusively for tRNA genes), the higher the rate of evolutionary changes (Fig. 2.5).

2.4 Discussion

Even as new sequencing tools have increased the number of plastid assemblies across plants, it is important to continue with the effort to increase the representation of the vast diversity of plants. For instance, for $\sim 7,300$ species or liverworts described, there are only ~ 80 species with plastome data available. The chloroplast of *C. californica* is similar to the chloroplast of the rest of members of Aytoniaceae. Overall, the structure of the chloroplast in this family is extremely conserved, with only three genes and one intron varying across Aytoniaceae. This pattern extends to the rest of liverworts; most of the variation observed occurs in specific genes, particularly tRNAs, and in specific taxa that seem to have a tendency of gene loss (*e.g. Fossombronia cristula*, *Aneura mirabilis*, and *Treubia lacunosa*).

When using the structural data to infer phylogenetic relationships, I obtained topologies that, regardless of the inference method, on one hand recover some large groups like complex thalloid liverworts, but on the other hand disagree with our current understanding of liverwort phylogenetics for other clades (*e.g.*, phylogenies of Bechteler et al. (2023) or Dong et al. (2021)). For example, the family Aytoniaceae and most of the complex thalloid clade (Marchantiopsida) are recovered as monophyletic in both analyses, and most of the

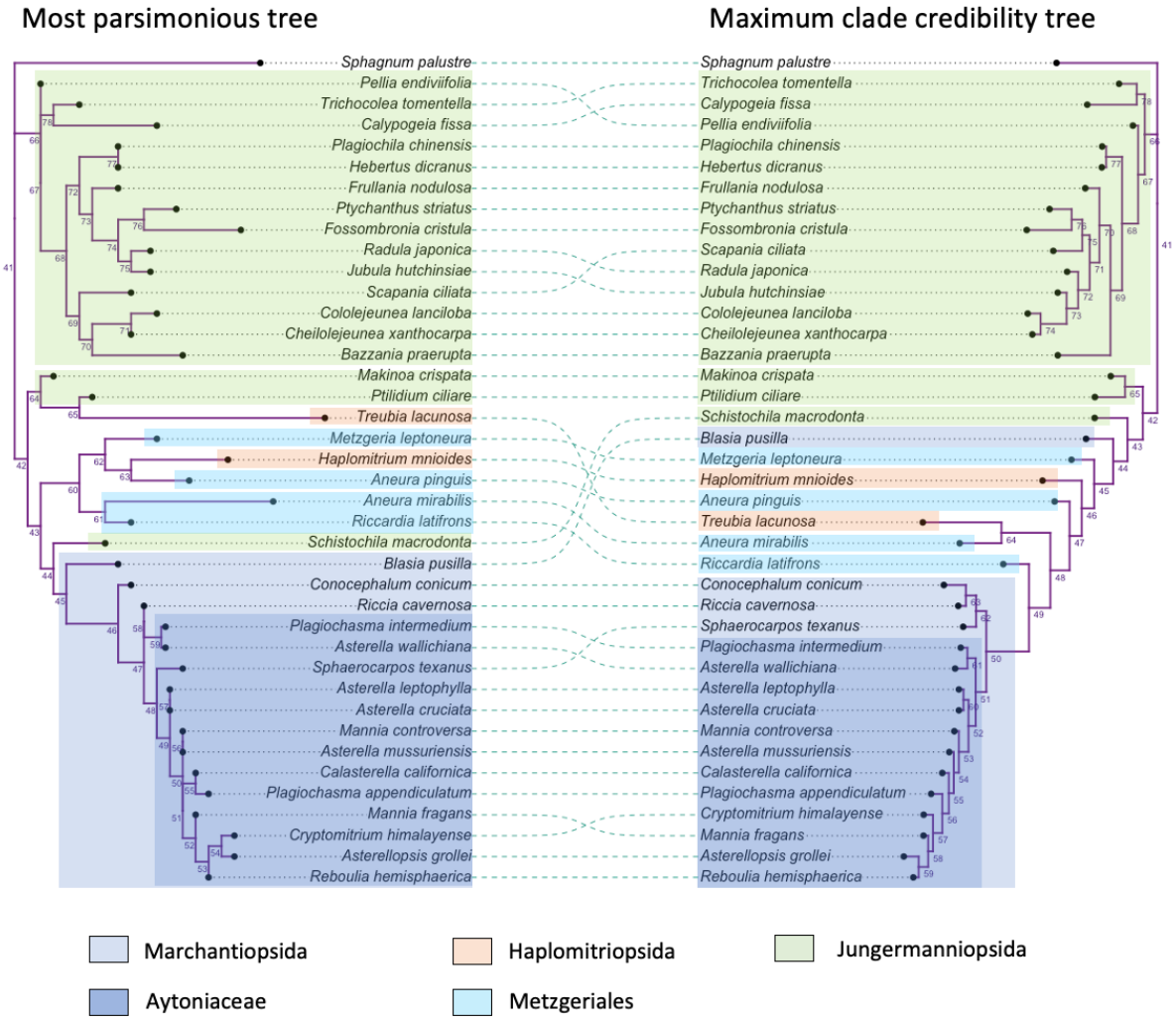


Figure 2.4: **Liverwort phylogenetics based on structural data.** Most parsimonious (left) and maximum clade credibility (right) trees obtained from a parsimony and Bayesian analysis respectively. Both analyses used the presence/absence and modification of genes and introns in the chloroplast as characters. The colors highlight traditionally accepted clades of liverworts (*e.g.* Bechteler et al., 2023).

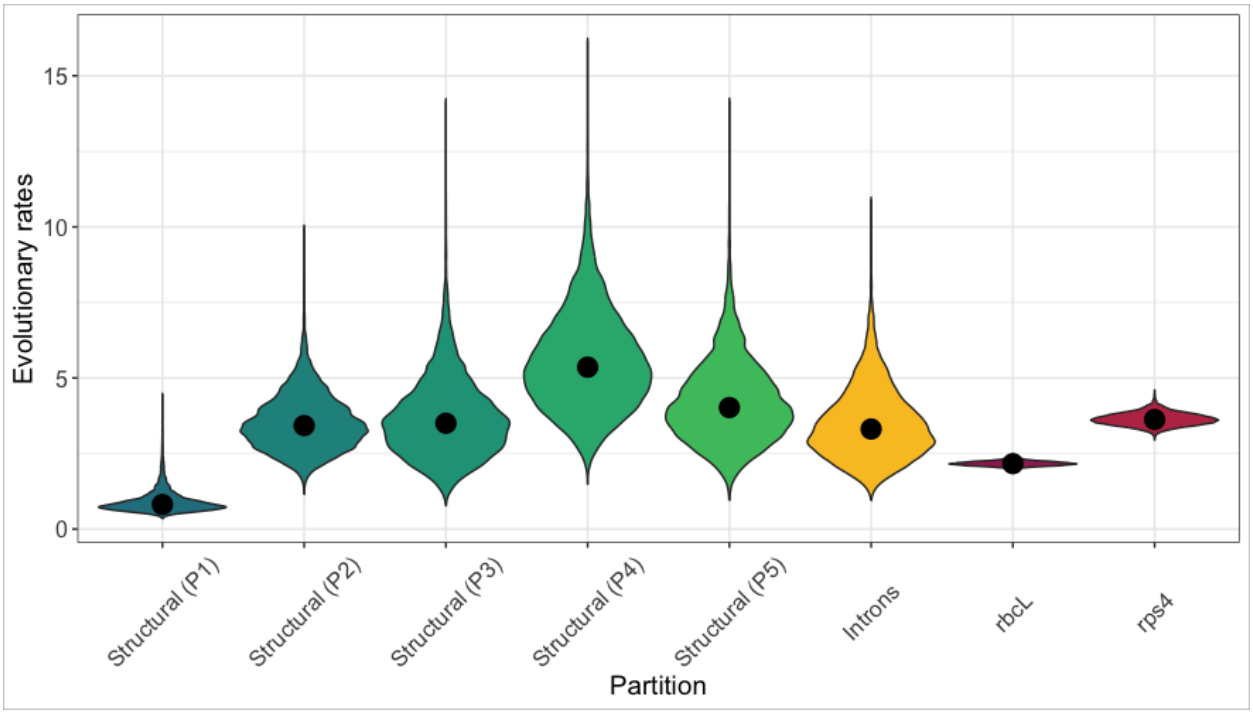


Figure 2.5: **Differences in evolutionary rates between structural and DNA sequence data.** The rates are expressed as the expected number of changes per character (gene or nucleotide respectively). The five structural partitions correspond to the different number of states as follows: P1: 2 states, P2: 3 states, P3: 4 states, P4: 6 states, and P5: 8 states.

members of the clade Jungermanniidae are recovered as monophyletic, with the exception of *Schistochilla macrodonta*, *Makinoa crispata* and *Ptilidium ciliare*. But, in both analyses, the commonly recognized clades Metzgeriales and Haplomitriales appear as polyphyletic. These disagreements seem to be particularly associated with taxa that have lost multiple genes like *Fossombronia cristula*, *Aneura mirabilis*, and *Treubia lacunosa*. For instance, according to our ancestral state reconstructions *Treubia lacunosa* and *Aneura mirabilis*, independently lost the set of *ndh* genes (*ndhK*, *ndhI*, *ndhA*, and *ndnH*). In fact, these genes have independently been lost in multiple groups of land plants, particularly among parasitic (such as *Aneura mirabilis*) and epiphytic plants (Sabater, 2021). Thus, this independent loss of a set of genes seems to be having a strong effect on the phylogenetic inference using structural data. In this sense, structural data, as well as DNA sequence and morphological data, seems to be sensitive to the non-independence of certain characters. When inferring the phylogenetic relationships using sequence data (*rbcL* and *rps4*, Appendix A.5) we obtained a phylogeny that is highly congruent with our current understanding of liverwort phylogenetics that is based on sequence data. When adding the structural data to the molecular data

(Appendix A.8), the overall support values were higher, and the only topological difference was the position of *Pellia endiviifolia*. In fact, the relationship of the order Pelliales to other liverworts remains inconsistent across analyses, and is retrieved with high uncertainty as the sister group of simple thalloids + leafy liverworts in a comprehensive multilocus analysis by Bechteler et al. (2023).

Finally, relative to sequence evolution, gene loss occurs at a slower rate in the chloroplast of liverworts. Nevertheless, there are other structural changes that have rates of change comparable to that of nucleotides. Notably, transfer RNAs can have up to six different states across liverworts. This may be because in tRNAs there are three nucleotides at the end of the "acceptor arm" that define what codon the tRNA will attach to. Functionally, a substitution in these very specific positions of a tRNA equates to a change in the identity of the tRNA; therefore, it is not surprising that the evolutionary rates of variable tRNAs are comparable to the evolutionary rates of nucleotides of coding regions. While these observations suggest that structural evolution in the chloroplast is more complex than initially thought, our results also suggest that structural changes can be used as characters for phylogenetic inference. Further studies using structural changes for phylogenetic inference need to account for correlations among gene losses (as shown in the example of the *ndh* set of genes) and different rates of evolution, particularly when including tRNA genes. In this study we coded the presence-absence of all genes, but a different approach could exclude some genes (*e.g.*, tRNAs) or group sets of genes into single characters. Finally, it would be interesting to study the evolution of structural changes in the chloroplast, across a broader range of land plants.

Chapter 3

The population structure of *Calasterella californica*

3.1 Introduction

Calasterella californica (Fig. 3.1) is a species of complex thalloid liverwort in the family Aytoniaceae. It is a west coast endemic with a geographic range that extends from Oregon to Baja California. In California, this species occurs in most ecoregions, from the northern redwood forests to the Sonoran Desert. *C. californica* is particularly common along the coastal ranges, but it is also found in the Sierra Nevada, San Gabriel, and San Bernardino Mountains at elevations up to 2000 m above sea level. The only ecoregions from where *C. californica* appears to be absent is the Central Valley, the Eastern Cascades and the Mojave Basin.

Unlike the majority of the family Aytoniaceae, *C. californica* is dioicous, *i.e.*, it has male and female individuals (Fig. 3.1), thus requiring that the motile sperm swim through a film of water to fertilize the egg. In the field, it is common to observe sporophytes and the characteristic bright yellow spores of *C. californica*, which is evidence of common sexual reproduction happening in this species. Additionally, while fragments of the thallus of *C. californica* can produce new individuals, *C. californica* doesn't have specialized asexual reproductive structures, unlike many other species of liverwort (*e.g.*, *Marchantia polymorpha*, *Lunularia cruciata*, and multiple leafy liverwort species that have propagula), which might indicate a heavier reliance on sexual reproduction and individual longevity to maintain its populations.

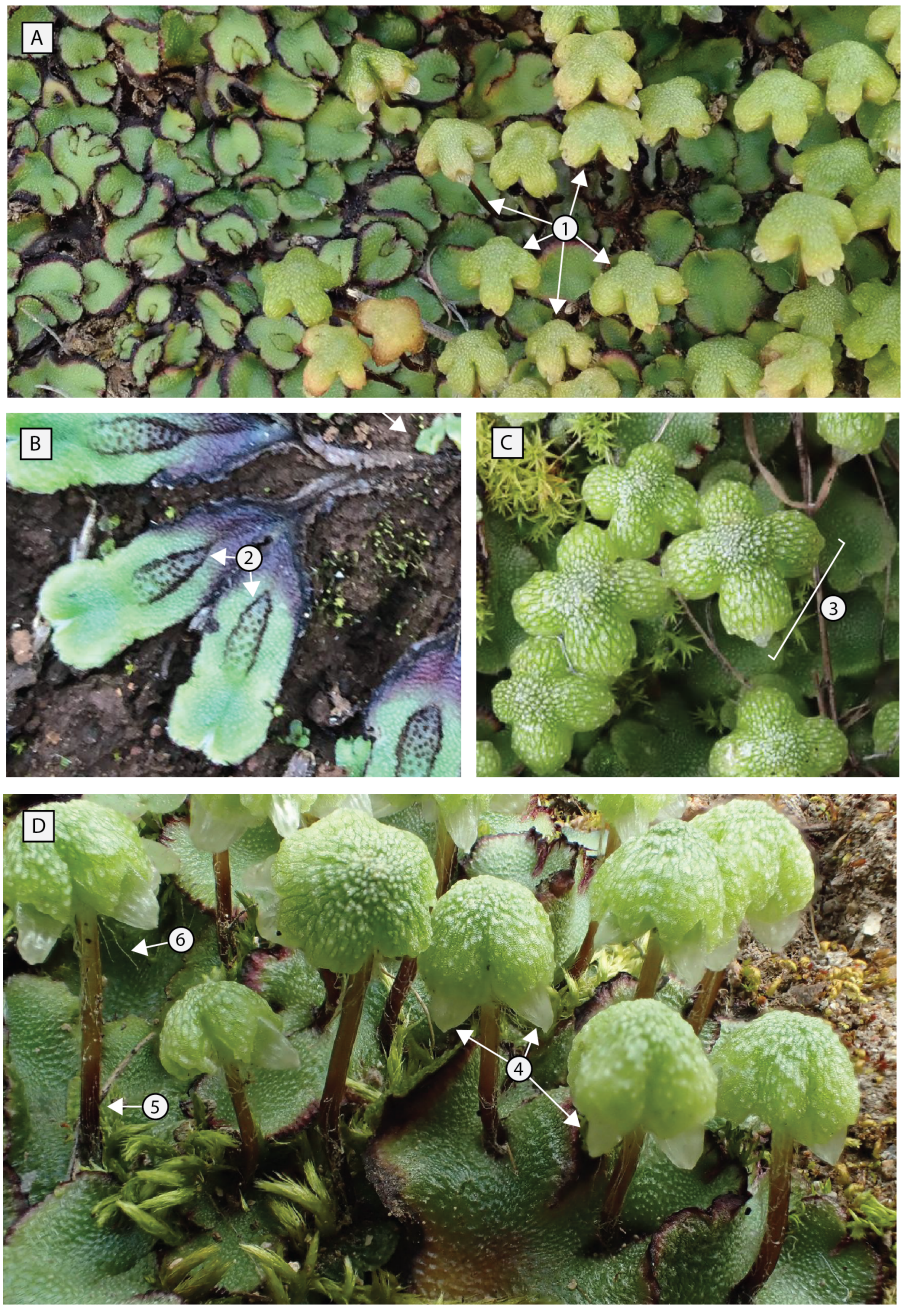


Figure 3.1: **The distinctive characteristics of *Calasterella californica*. A**. *C. californica* is dioicious, with male (left) and female (right) plants. Female plants produce eggs and subsequent sporophytes in elevated structures called archegoniophores (1) **B**. Male plants produce sperm in purplish antheridia clusters on the surface of the thallus (2). **C**. *C. californica* is characterized by having carpocephalla with 4 (3-5) very conspicuous lobes (3). **D**. The carpocephalla have persistent hyaline pseudo-perianths that protect the sporophytes (4), stalks with dark-purplish bases (5) and filiform hyaline scales near the apex (6). Image credits as follows: A: John Garrett (iNaturalist: 21898117), B: Arvel Hernandez (iNaturalist: 38506014), C: Aaron Echols (iNaturalist: 21041398), and D: Ken Kellman (iNaturalist: 21633535).

A continuous production of spores in *C. californica* might also contribute to the widespread range of this species, since it has been hypothesized that spore size (comparatively smaller than wind dispersed seeds) facilitate long distance dispersal, but it is unclear to what extent distant populations of *C. californica* maintain gene flow. And given the breadth—and sometimes harshness—of the ecosystems in which this species occurs, we might expect natural selection to play an important role in maintaining locally adapted populations. This is perhaps especially true in haploid organisms, where detrimental alleles can not 'hide' in the population through masking, potentially increasing the efficacy of natural selection. In this chapter, I constructed a genomic dataset for 95 individuals of *C. californica* to characterize the genetic diversity of this species across California. I interpreted the geographic structure of this diversity by correlating it to environmental variables in order to identify factors that might be driving the genetic diversity of this species.

3.2 Methods

Collection of *Calasterella californica*. I collected 110 samples of *C. californica* across California. Detailed information about collection points can be found in Appendix B.1. The collection effort focused on sampling across the geographic range of the species in California as well as sampling across the diversity of ecosystems in which *C. californica* occurs. I guided my fieldwork by looking at previous herbarium and iNaturalist records of this species (for a map, see `https://rpubs.com/Ixchel/664552` and Fig. 1.1). Most collections were made within National Forests and California State Parks. A list of the parks sampled and the information about corresponding collection permits can be found in Appendix B.2. Because *C. californica* lives in aggregated clumps of thalli, each sample consisted of 15 to 30 thalli—roughly covering a square area of 15 by 15 cm. For each sample, photographs of the plants *in situ* were made, and the clump plus a thin layer of soil was transported in a hard plastic container to be processed in the lab.

Sample processing. The samples were processed immediately upon returning to the lab. For each sample, I (1) isolated material for DNA extraction, (2) made permanent slides, and (3) prepared a voucher herbarium specimen. The largest thallus of the sample was isolated for DNA extraction, but because *C. californica* individuals are very small, up to five thalli were often included to ensure enough material for extractions. Each thallus was manually cleaned with tweezers and de-ionized water under the dissecting scope to remove other organisms and soil. After cleaning, the thalli for DNA extraction were dried in silica. When *C. californica* was collected dry, I did not use water for cleaning, to avoid hydration and dehydration, which could potentially damage the tissues.

The permanent slides contained (1) a cross section of the thallus, (2) three to six isolated and complete ventral scales, (3) and if the plant was a reproductive female, a cross-section of the archegoniophore. When the plant was a reproductive male, the thallus cross section was made such that it would intersect the antheridia. Cross-sections were made using razor

blades and then stained with Toluidine blue for 30 seconds, rinsed with de-ionized water, and finally prepared using Hoyer's solution as the mounting medium. The rest of the material of each sample was used for making a voucher herbarium specimen that, with the slides, will be deposited in the University Herbarium, at UC Berkeley.

DNA extraction and sequencing. To maximize the amount of DNA extracted from each sample I used two different extraction methods: a modification of CTAB (detailed in Appendix B.3), or the Quiagen DNeasy Plant Pro kit. The library prep and Illumina sequencing were done by the UC Davis Genome Center using a NovaSeqX 300 (PE150) sequencer.

Reads alignment and variant calling. The Illumina reads of the 110 samples were processed using the CCGP workflow (`https://github.com/ccgproject/ccgpWorkflow/`). This workflow is based on SNPArcher (Mirchandani et al., 2024), which is an aligning and variant-calling workflow optimized for genomics of non-model organisms. The original pool of single nucleotide polymorphisms (SNPs) detected by the workflow (35,448,107 SNPs) were filtered to remove all indels, non-biallelic SNPs, SNPs with an allele frequency < 0.01, SNPs with $> 75\%$ missing data, and samples with $< 2x$ sequencing depth. And finally, a sliding window for SNP selection was applied in order to avoid linkage by proximity. The variant calling analyses and filtering steps are computationally intensive and were run by the CCGP bioinformatics team using Google Cloud. After applying the processing workflow, the resulting sample of 100,881 SNPs for 95 individuals was used to conduct the population structure analyses.

Genetic similarity across *C. californica* individuals. In order to visualize the genetic similarity of *C. californica* individuals, I conducted a cluster analysis, using only the subset of SNPs with no missing data (*i.e.*, a total of 23,180 SNPs). The first step was to obtain a distance matrix that counted the number of differences in the bi-allelic SNPs between pairs of samples. This matrix was used to compute a UPGMA clustering analysis as implemented in the R package `phagorn` (Schliep, 2011).

Genetic composition analysis. I used the software `ADMIXTURE` (Alexander et al., 2009) on the 100k SNPs dataset. This software estimates the 'global ancestry' of each individual in the sample, *i.e.*, it allocates individual's genetic variation among a predefined number of hypothetical source populations (Alexander et al., 2009). Since this algorithm requires the number of ancestral populations (K) to be set *a priori*, the `ADMIXTURE` cross-validation procedure was used to select the K-value that yields the lowest error, which for this dataset was K=4, with tested K values from 1 through 7 (Appendix B.4). The ancestry proportions obtained from this analysis were used to produce a plot in `R` (R Core Team et al., 2013), using the package `ggplot2` (Wickham, 2006).

Principal component analysis. Using the SNP dataset without missing data (23,180 SNPs), a principal component analysis (PCA) was carried out to reduce the multidimensional variation of the dataset to a manageable number of dimensions. Both bi-dimensional and tri-dimensional plots were examined, and in order to identify whether or not the grouping of individuals was driven by the environment in which each liverwort occurred, I color coded the samples by each of the 15 WorldClim environmental variables (Fick and Hijmans, 2017) and ecoregions of California (Griffith et al., 2016). These analysis were performed in R (R Core Team et al., 2013), and used the libraries `raster` (Hijmans et al., 2015), `sf` (Pebesma et al., 2018), and `ggplot2` Wickham (2006).

Testing for geographic and genetic distance correlation. In order to explore whether or not individuals closer in space tend to be more similar genetically, I computed the geographic and genetic distance between each pair of *C. californica* samples. For each pair of samples, the R package `geosphere` (Hijmans et al., 2017) was used to obtain the geographic distance, and the genetic distance was expressed as the total number of differences in the SNPs of each pair. Finally, I used a Mantel test to asses potential correlations between geographical and genetic distance.

3.3 Results

Collections of *C. californica*. The sampled localities span the geographic breadth of *C. californica*'s range (Fig. 3.2), and cover the range of environmental variables that this species occupies across California (right panels of Fig. 3.2).

Alignment and variant calling metrics. After aligning each sample to the reference genome of *C. californica*, 95 samples passed the threshold of 2x read depth. Among the samples that were not used, IGR204_ext_IG043 potentially belongs to another species of *Asterella* based on the extremely low mapping of the reads.

***C. californica* population structure.** The approximate nucleotide diversity of this *C. californica* sample was 0.4% according to the Watterson estimator. Based on the clustering analysis based on SNPs (Fig. 3.3), *C. californica* individuals form two main clusters (yellow and brown, versus the rest, Fig. 3.3). When placed on a map of California, this division roughly corresponds to an imperfect north vs south division of the state.

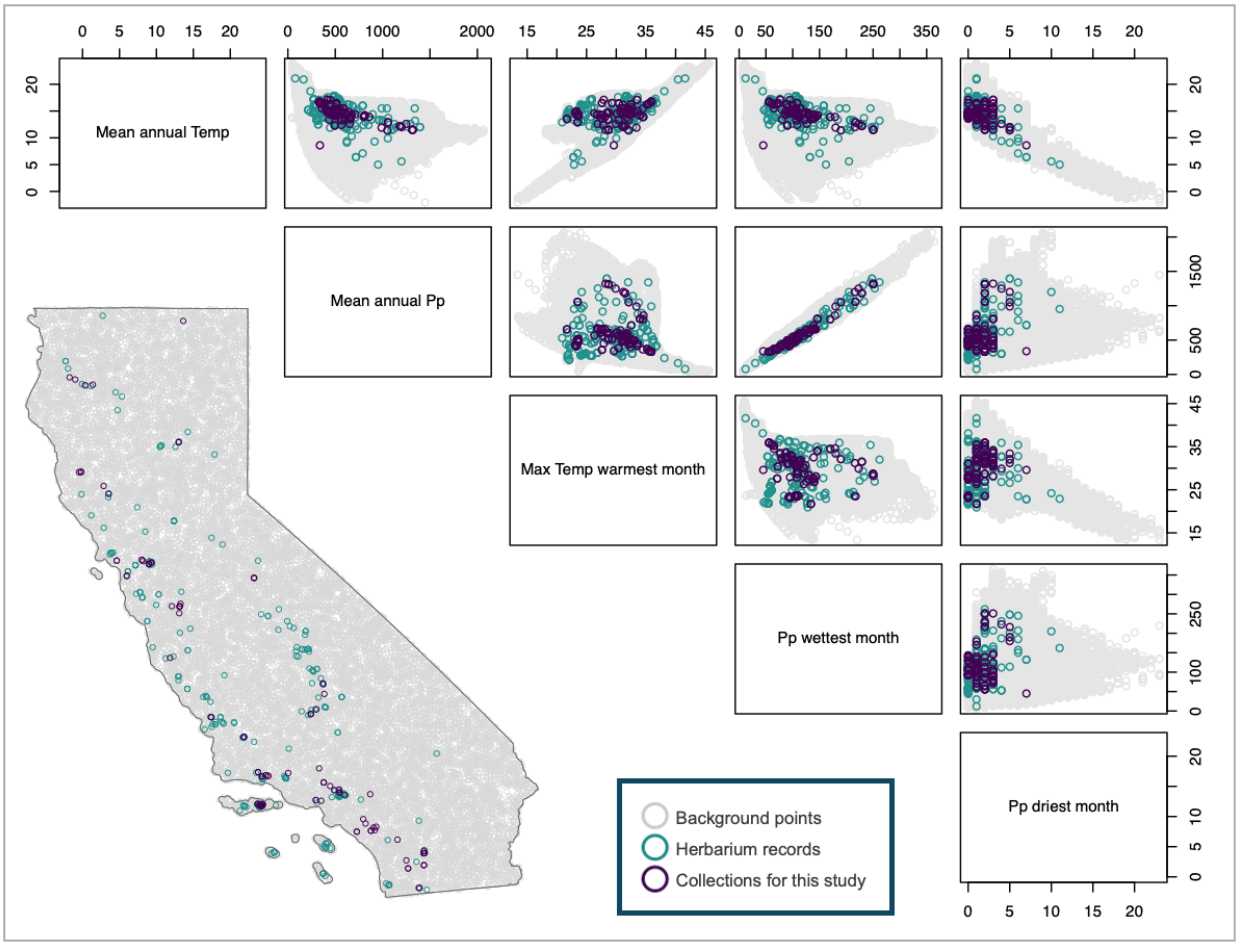


Figure 3.2: ***C. californica* collection points across California.** The collection points for this project are indicated in purple, which are contrasted with all the herbarium records for *C. californica*, in blue. The panels on the right indicate the range of environmental variables available in California (grey), of the localities for the herbarium records of *C. californica* (blue), and of the localities where the samples for this study were collected (purple).

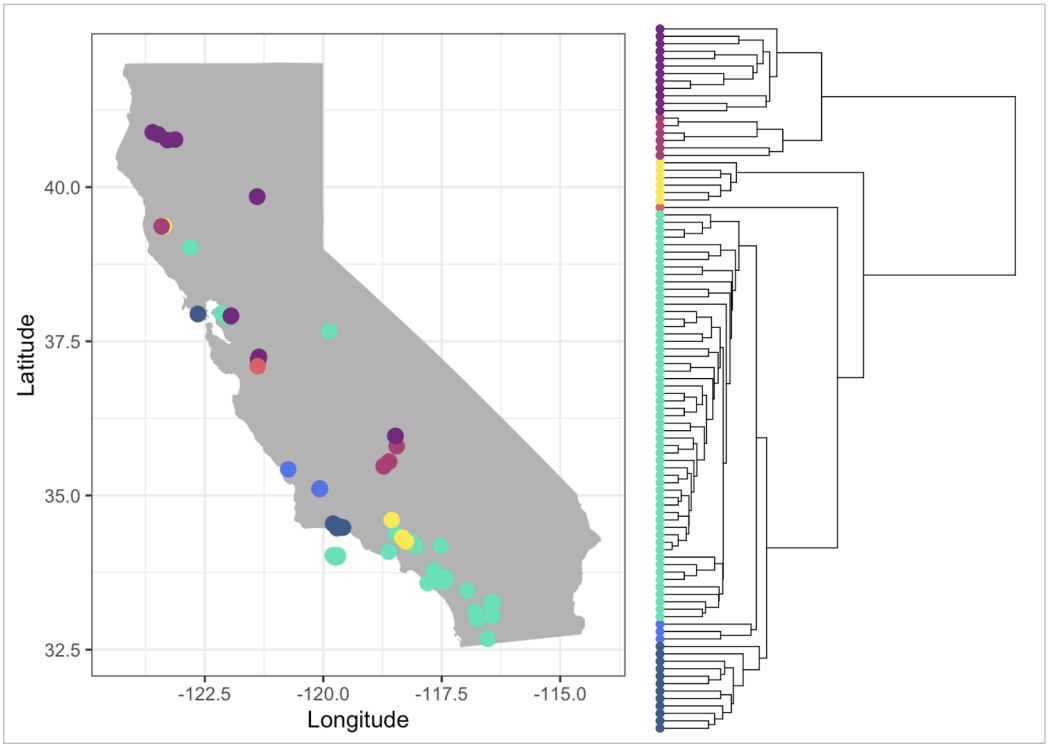


Figure 3.3: **UPGMA cluster of *C. californica* individuals.** On the right, UPGMA dendrogram inferred from the SNP genomic dataset of 95 samples of *C. californica*. The branch lengths are proportional to the number of differences between samples. Each tip in the cluster is represented in the map. Groups of samples are color coded in the map and cluster to facilitate visualization.

The `ADMIXTURE` cross-validation analysis suggested that the number of theoretical ancestral populations that best fit the data was $K = 4$ (Appendix B.4). In Fig. 3.4, we observe the ancestry composition of each individual, as well as a map where each individual is placed in the coordinates where it was collected. Assuming four theoretical ancestral populations, the majority of individuals have contributions from a single theoretical population, *i.e.*, they are single colored. When observed in a geographic context, the individuals in the Sierra Nevada have exclusively blue ancestry, while most of the individuals in the coastal ranges have purple ancestry. Individuals with only yellow and only green ancestry are restricted to the region south of the Transverse Range. In particular, individuals with only yellow ancestry occur in the Sonoran Desert. Additionally, the individuals with mixed ancestry are mostly found in the southwestern region of the state, particularly along the Transverse and Peninsular Ranges (Fig. 3.4).

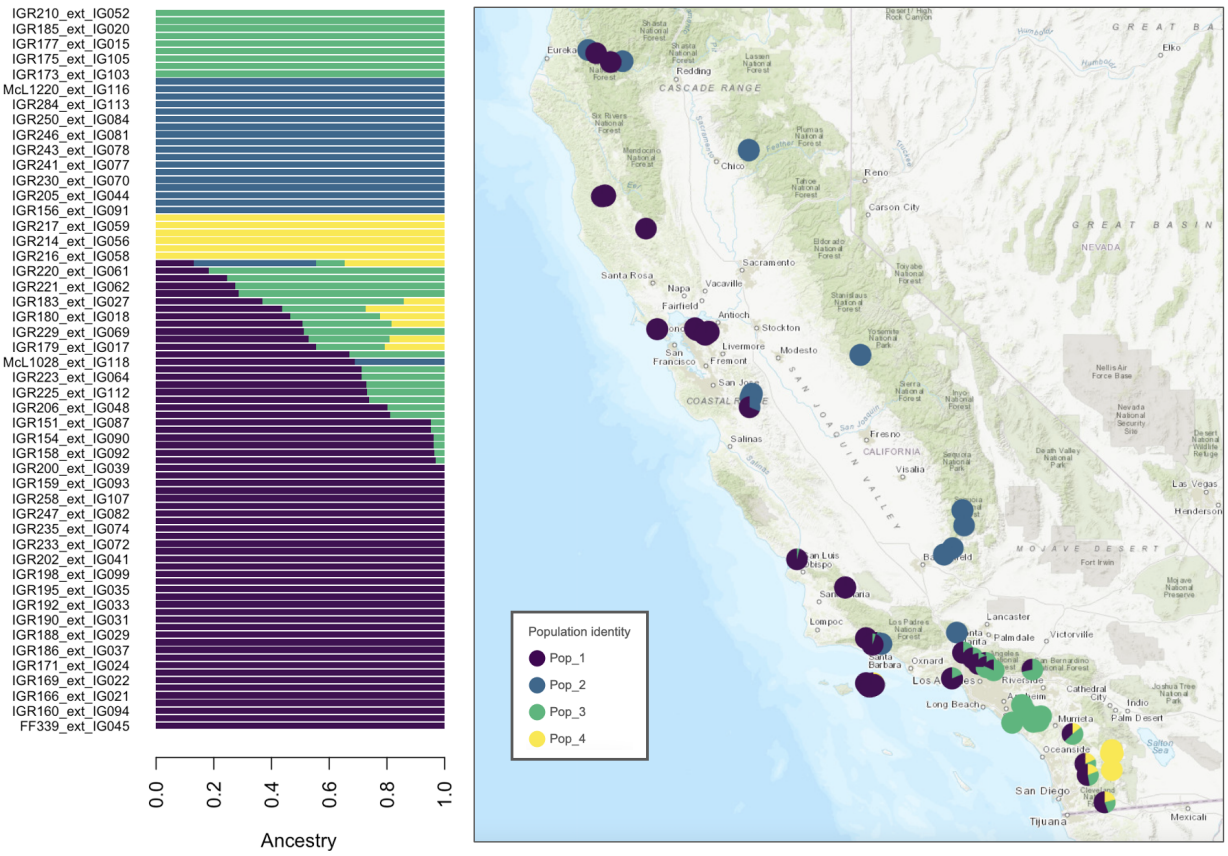


Figure 3.4: **Ancestry of *C. californica* individuals as inferred by `ADMIXTURE`.** The plot to the left represents, for each individual, the proportion of SNPs that belong to each of the four theoretical ancestral populations of *C. californica*. In the map, each individual is colored using the same colors and proportions as on the left.

Fig. 3.5, shows the two largest principal components obtained from the PCA analysis that grouped *C. californica* individuals by similarity based on SNPs. Together, PC1 and PC2 explain 52.8% of the variation in the SNP dataset, and the samples were color coded by different WorldClim climatic layers (Appendix B.5), by the ecoregion in which they were collected and by the genetic composition according to the `ADMIXTURE` analysis (Fig. 3.5). There are three evident groups of samples. The first cluster exclusively contains all the samples collected in the Sonoran Desert (group I), which also have a characteristic genetic composition according to ADMIXTURE (Fig. 3.5). Group III contains that share the blue genetic composition and they occur along the Sierra Nevada, and Southern California Mountains ecoregion. And group II contains most individuals of the 'Southern California/ Northern Baja Coast' and Coast range; these group appears to be the one with the most

diverse genetic pool according to the admixture analysis.

When coloring the PCA by WorldClim environmental variables (Appendix B.5), group 1 occurs in drier and warmer environments compared to group II and III. And although the differences between cluster II and III are less clear, they appear to be correlated with higher temperature seasonality for group III. Interestingly, there is one sample that seems to be genetically different from the rest (middle of the plot Fig. 3.5). This sample (IGR201_ext_IG040) was obtained in Santa Cruz island, and corresponds to a specimen that was collected in an isolated canyon 100 m from the ocean, in an environment uncommon for *C. californica*.

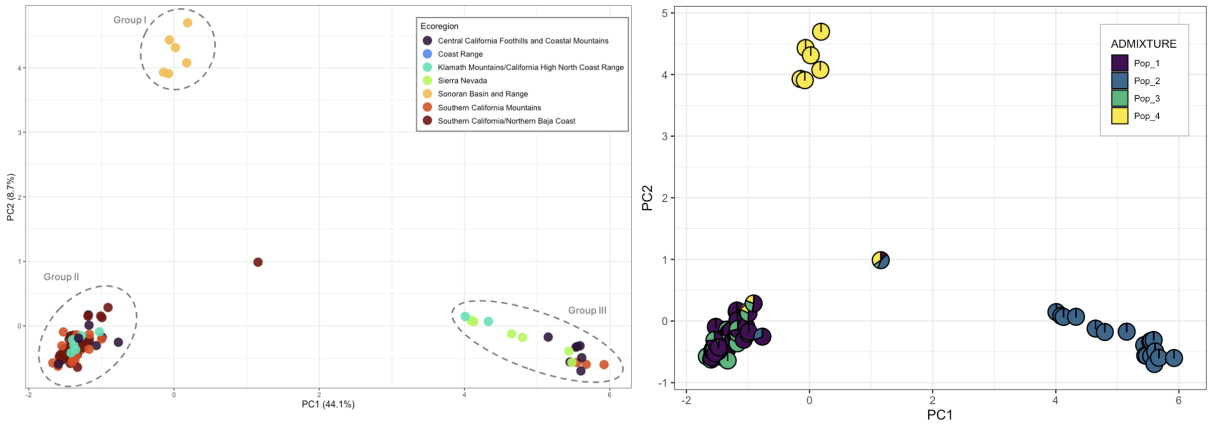


Figure 3.5: **Principal components analysis of *C. californica* individuals based on SNP data.** PC1 and PC2 which together explain $\sim 52\%$ of the variation in the SNP dataset. In the left plot the samples are colored by Califonia ecoregion. In the right panel, the samples represent the genetic contribution as inferred by `ADMIXTURE`.

Finally, of the four subpopulations investigated (as inferred by `ADMIXTURE`, three show a positive correlation (coastal population does not); in addition, the global pool of samples analyzed together shows no indication of a correlation (Fig. 3.6).

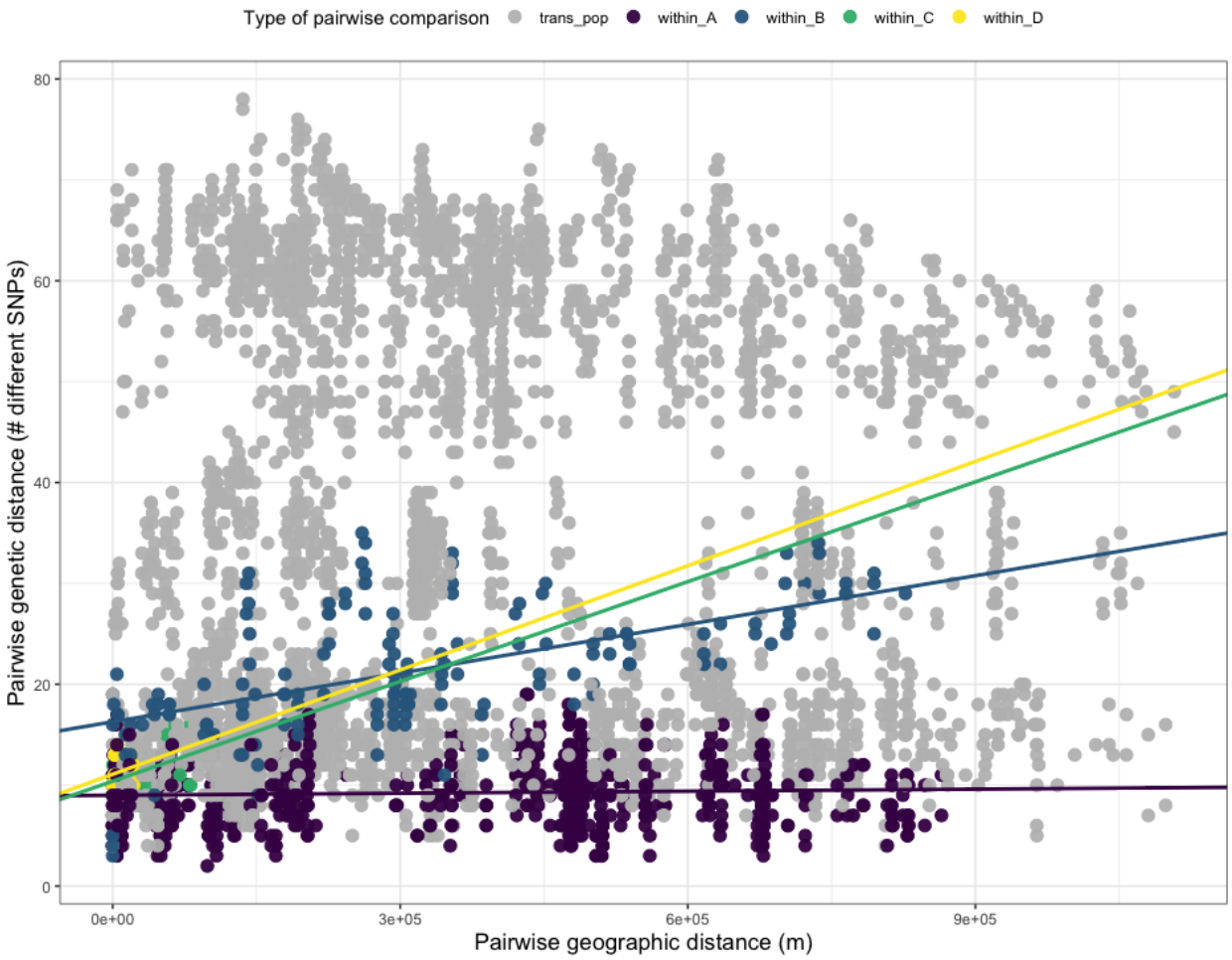


Figure 3.6: **Pairwise comparison of the genetic and geographic distance between *C. californica* samples.** Each dot represents a pairwise comparison of two samples of *C. californica*. Pairwise comparisons among individuals of same ancestry (as defined in Fig. 3.4) shared the same color (purple, blue, yellow, or green). Pairwise comparisons involving individuals of mixed ancestry, or between individuals of different ancestries are colored as grey.

3.4 Discussion

Field observations on *C. californica*'s biology

While collecting this plant across California during a two year period, I made multiple observations that help to uncover its life history, and to better interpret the genetic patterns found in this species. First, *C. californica* very often co-occurs with a group of xerophytic

ferns: *Pellea andromedifolia*, *Pellea mucronata*, *Pentagramma triangularis*, and to a lesser extent, the lycophyte *Selaginella sp.* Additionally, *C. californica* also tends to occur in the same environments as *Targionia hypophylla*, a complex thalloid liverwort in the family Targionaceae that shares with *C. californica* the ability to 'roll-up' during dry periods. Interestingly, but not surprisingly, according to the California Consortium Herbaria (https://ucjeps.berkeley.edu/consortium/), these xerophytic ferns, *Targionia hypophylla*, and *C. californica* all have almost identical distributions in California. Notably, all of them are absent from the Central Valley.

These xerophytic flower-free species are common on rocky outcrops, especially if they have a high sun-exposure. But when rocky outcrops are shaded, this group of species tends to be replaced by species of *Polypodium* or grass. In general, it appears that *C. californica* does not occur in the presence of fast growers (*e.g.*, grass), or among plants that produce a lot of leaf litter. Notably, it did not occur in stream banks dominated by *Acer*, or forest dominated by pines. Although this hypothesis requires more study, I think that the reason *C. californica* does not occur among grass or deciduous trees is because it can not survive being heavily shaded. As I observed in the lab while growing *C. californica* tissue from spores, the growth rate of this plant is extremely slow—*e.g.*, it took six months of growth in ideal conditions to obtain $0.26g$ of tissue—so being covered by other plants likely results in the death of the plant by lack of sunlight.

Another interesting observation is that *C. californica* plants seem to live multiple years. *C. californica* thalli bifurcate as they grow, and based on my observations, I estimate that a single thallus grows at a maximum rate of one bifurcation per growing season (*i.e.*, one thallus bifurcation per year). When observing larger patches of this species, it is possible to trace back the 'path' that a thallus has grown, which often indicates that a single individual can be at least five years old. The longevity of this species is probably facilitated by the fact that it can survive California's highly seasonal climate. During the dry season, *C. californica* looks very different, as it is covered by its ventral scales during a radical morphological shift that I refer to as 'rolling up'. As the thallus dries and shrinks, the large purple ventral scales fold over the thallus, forming a barrier around the desiccated thallus tissue (Fig. 3.7).

The *roll up* morphology has been mentioned in the past in taxonomic treatments, but its potential ecological implications have not received enough attention. This feature is not exclusive of *C. californica*. Other xerophytic liverworts in different families, like *Targionia hypophylla*, *Reboulia hemisphaerica*, and some *Riccia*, have similar strategies. In fact, the ventral scales of liverworts are understudied structures that might be performing a variety of functions from external water conductance, substrate attachment, mechanical protection, to UV protection (*in prep.*), thus warranting more physiological and comparative work.

The genetic structure of *C. californica* across California

According to the genetic ancestry, Fst values, and PCA analyses, *C. californica* populations are genetically differentiated across the species range in California. First, there is a general pattern of differentiation between coastal and inland populations in central and northern

Figure 3.7: **Hydrated and dry thalli of *C. californica*.** On the left there are some hydrated male *C. californica* thalli. On the right picture, we can observe how dry thalli have *rolled up* and the dark ventral scales are surrounding the thalli. Arrows point to ventral scales.

California (populations 1 and 2 in Fig. 3.4 and clusters II and III in Fig. 3.5) that seems to be driven by the apparent inhospitably of the Central Valley, supported by the fact that no individuals of *C. californica* have been collected there, according to the herbarium records for the species (Fig. 3.2). It is unclear why *C. californica* is absent from California's Central Valley, since there are individuals of *C. californica* that inhabit regions with more extreme temperature and precipitation conditions (*e.g.*, the Sonoran Desert), suggesting that the climatic variables might not be the drivers of this pattern. One hypothesis to explain this would be that the slow-growing strategy of *C. californica* makes it a very bad competitor in the very dynamic—now agricultural but deltaic in the past—lands of the Central Valley.

In addition to the coastal/interior differentiation, *C. californica* individuals that occur in the Sonora Desert are inferred as highly differentiated in all the analyses. Furthermore, the PCA analysis suggests that the conditions in which this group occurs are drier and warmer (Y-axis of Fig. 3.5, and Appendix B.5) than the conditions that the rest of the individuals experience. In this regard, it would be interesting to study whether or not the genetic differentiation among these populations is linked to traits that are adaptive to live in the desert.

And finally, the genetic ancestry plot identifies a fourth ancestral genetic group (green, Fig. 3.4), that is present in individuals occurring in the Southwestern California region.

Overall, the region South of the Transverse Range and east of the San Bernardino mountains contains most of the individuals inferred as having 'mixed-ancestry', which indicates that the populations of these regions have overall a more diverse gene pool, compared to the other regions that have unique ancestries (*e.g.*, Sierra Nevada and Coastal). The higher diversity in the Southern region is probably related to the fact that three otherwise isolated genetic pools (*i.e.*, coastal, Sierra Nevada, and desert) converge in this geographic region. Nevertheless, the methodological approaches that I have explored so far do not allow an analysis of the direction of this pattern, *e.g.*, was the diverse gene pool of Southern California ancestral and then subsets of this population spread across the state, or did different genotypes evolved in different regions and then only recently converged in Southern California.

Lastly, the pairwise distance plot (Fig. 3.6) suggests that within subpopulation, with the exception of the Coastal population, we can expect that there is a correlation between genetic and geographic distance. Nevertheless, when comparing across all the samples, the correlation between geographic and genetic distance dilutes. At this point, there is not enough information to address the question of whether the populations are differentiated enough that they deserve to be treated as different taxonomic units. More information will come from future examination of gene trees and morphological comparisons among individuals of the clusters (notably variation in the spore and scale morphology), as well as genetic associations with traits and environment, might help to better understand whether or not the genetically different populations of *C. californica* are phylogenetically coherent, locally adapted, and interbreeding.

Chapter 4

A new paleogeographically informed model to infer the history of Aytoniaceae

4.1 Introduction

Studying how lineages move in space through evolutionary time not only allows us to reconstruct the history of a group, but it also provides an opportunity to understand the processes that have shaped lineage distributions. In the last two decades, scientists have developed phylogenetic methods to study geographic range evolution, as a special case of trait evolution. Earlier approaches used a parsimony approach to infer the relationship between areas (area cladogram) based on the study of organismal cladograms, and are collectively known as "cladistic biogeography". One of the most used parsimony based methods is the dispersal-vicariance analysis (DIVA, Ronquist, 1997)), which is a parsimony character optimization method, that infers the areas of ancestral nodes by selecting the states that minimize the number of dispersal, extinction and vicariance events. DIVA became a very popular method because unlike its predecessors it accounts for the possibility of dispersal-extinction and vicariance events to occur on the same tree and it does not require previous knowledge about the history of areas (Nylander et al., 2008).

More recently, scientists have tackled the challenge of using models to understand the evolutionary history of groups. Taking a probabilistic approach has multiple advantages, from allowing the estimation of biologically relevant parameters (*e.g.*, dispersal, extinction, and cladogenesis), providing a framework for testing hypothesis, to explicitly accounting for other variables (*e.g.*, area features and connectivity, trait-dependency) that might affect range evolution (**?**).

One special feature of range evolution is that we expect that changes in range distribution often occur in association with speciation events (*e.g.*, allopatric speciation or jump-dispersal speciation events). Therefore we usually model range evolution as evolving by anagenetic

events occurring along branches (dispersal and extinction) as well as cladogenetic events at nodes. Phylogenetic models of range evolution can accommodate multiple scenarios. For example, intrinsic differences between regions, such as spatial heterogeneity or size, might affect biogeographic rates, as might trait-dependence of dispersal abilities or differences in region's connectivity. One scenario that remains particularly challenging to study from a phylogenetic perspective is the range evolution of an ancient and widespread lineage because there is an increased probability of unobserved cladogenetic range changes due to extinction, and because we must account for changes in the configuration of Earth's landmasses over geological time.

Life evolves on a changing Earth. For example, many terrestrial organisms today would be unlikely to disperse from South America to Africa. However, during the Triassic these two landmasses were contiguous and dispersal would have been comparatively much more likely. Accounting for this ever-changing scenario makes models more realistic and improves our biogeographic inferences. Additionally, old lineages have experienced more extinction in comparison to younger clades, and therefore the number of unobserved cladogenetic events is higher. If we are interested in disentangling the effects of different biogeographic processes (*e.g* cladogenetic vs. anagenetic), then it is necessary to account for all those missing cladogenetic events to avoid the underestimation of range evolution due to cladogenesis. For example, the GeoSSE framework (Goldberg et al., 2011) is able to account for unobserved cladogenetic events through time. Using a biogeographic model that combines GeoSSE with a time-stratification of region connectivity can accommodate the processes necessary for studying the range evolution of ancient and widespread clades.

Liverworts are a particularly interesting group because of their potential phylogenetic position as the sister group of all extant land plants (Cox, 2018b; Puttick et al., 2018; Rensing, 2018; Finet et al., 2010; Renzaglia et al., 2000), the uniqueness of some of their morphological traits (Shaw et al., 2011), and their apparent morphological stability through geological time (Hernick et al., 2008). Liverworts occur in almost all terrestrial ecosystems and some liverwort species have remarkably wide geographic distributions, often covering several continents (Laenen et al., 2016; Shaw, 2001). These broad distributions have been attributed at least in part to high dispersal abilities compared to seed plants. Members of the family Aytoniaceae, a clade of complex thalloid liverworts, are outstanding in the variation of their geographic ranges, with some species—like *Reboulia hemisphaerica* (L.) Raddi and *Plagiochasma rupestre* (G.Forst.) Stephani—occurring simultaneously in places as distant as North America and Australia, while other species—like *Asterella innovans* (Austin) H.A. Mill. and *Mannia capensis* (Stephani) S.W.Arnell—are narrow-range endemics (in this case, to Hawai'i and southern South Africa, respectively). Aytoniaceae is also thought to be ancient, with estimates of having originated ∼111 Mya in the Cretaceous (Villarreal A. et al., 2016). The age of the crown group of this family of liverworts is comparable to that of the crown group of angiosperms, but unlike them, this family contains only ∼92 named species (Söderström et al., 2016). The history of Aytoniaceae is most likely shaped by the interplay of long-distance dispersal abilities, extinction, and a changing Earth. In this chapter, I estimated a time-calibrated phylogeny of Aytoniaceae. Then, I developed a new

PAw GeoSSE + J model and applied it to infer the biogeographic history of Aytoniaceae and better understand the role of dispersal in shaping the history of this family of liverworts.

The Paleo-geographically aware (PAw) GeoSSE + J model of range evolution

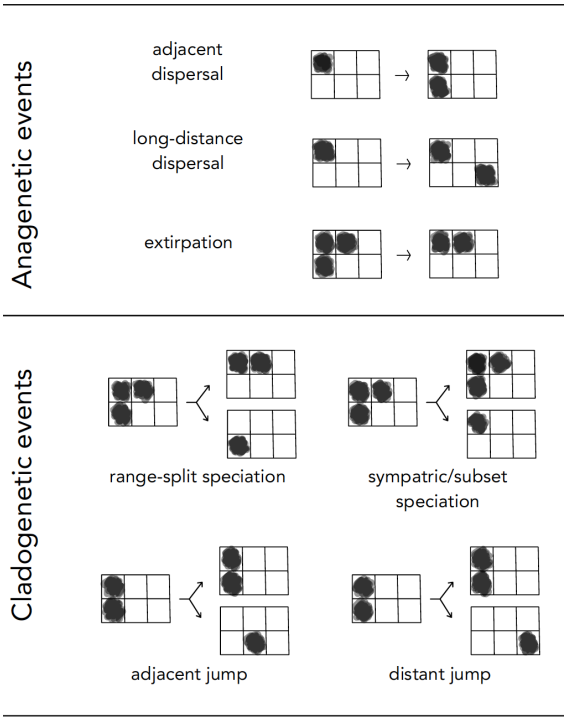


Figure 4.1: **Events of range evolution in the PAw GeoSSE + J model.** The events that are associated with range evolution in this model can be anagenetic or cladogenetic.

This model is an extension to the original GeoSSE model by Goldberg et al. (2011), where the range-occupancy of a lineage depends on dispersal, extirpation, and cladogenetic events. There are three main differences between this model and the original model. First, I include a +J parameter, often referred as jump dispersal. This parameter allows cladogenetic events to co-occur with dispersal events, i.e. an event where one of the descendant lineages disperses to a region that wasn't previously occupied by the ancestor lineage. Second, this model is time-stratified, so it incorporates information on the relative position of the regions through geological time using a time-dependent set of connectivity matrices. And third, I distinguish between adjacent and distant dispersal because I consider it likely that these types of events occur at different rates.

Region considerations

This model assumes that each region is a unit with negligible within-region variation (*e.g.*, in climatic or topographic features), which is equivalent to assuming that a lineage existing in one region occupies the entire region. Therefore it is not possible to make any inferences about within-region processes. The model allows lineages to occupy multiple regions, but it assumes that events always happen from/to an individual region. Thus, the events in widespread lineages are the sum of the events of the regions they occupy.

Events of range evolution

It is important to clearly distinguish between the predefined regions and the range of a lineage. While regions are set by the researcher and remain constant through time, the range of a lineage is the set of regions it occupies a given point in time and it evolves stepwise within the context of the fixed regions. A lineage can occur in a single or any combination of multiple regions at the time. Therefore, for N regions, there are 2^N possible range states. In this model, the range of a lineage evolves along a time-calibrated tree following a continuous-time Markov chain (CTMC) and according to seven global parameters (Fig. 4.1) that are the rates at which different range evolution events occur. These events occur according to exponentially distributed waiting times. The scale parameter of the waiting time distributions corresponds to the mean rate of each type of range evolution event. There are two categories of range evolution events: anagenetic and cladogenetic (Fig. 4.1). Figure 4.2 is a simple example of the possible range-states (boxes), and biogeographic events (arrows) when there are three regions: two adjacent (A and B), and a third non-adjacent (C).

Anagenetic events. These events occur along branches when the range of a lineage increases or decreases one region at the time. I denote anagenetic changes as "initial state → end state". For example, a change from state A to state AB is noted as $A \rightarrow AB$. In this model, there are three global range-evolution anagenetic events: adjacent-dispersal, distant-dispersal, and extirpation.

Adjacent-dispersal happens at a rate δ when a lineage increases its range by one region through dispersal into a previously unoccupied adjacent region. In Figure 4.2 for example, adjacent dispersal happens when a lineage goes from state $A \rightarrow AB$, or from $B \rightarrow AB$. Similarly, **distant-dispersal** happens when a lineage disperses into a distant region, increasing its range by one, and this events happen at a rate Δ. In Figure 4.2, distant dispersal occurs when a lineage goes from state $A \rightarrow AC$, or $B \rightarrow BC$. It is important to note that a widespread lineage (*i.e.* occupying more than one region) can increase its range through dispersal from any of the regions that it occupies, and thus the type of dispersal (*i.e* adjacent vs distant) depends on the relationship of the source region with the newly occupied region. For example, in Figure 4.2, the dispersal from state AC to ABC (AC → ABC) can occur in two different ways: through adjacent dispersal from A to B, or through long distance dispersal from C to B. Therefore, for lineages that occupy more than one region, likelihood calculations during inference involve the enumeration of all the possible ways in which a change can occur. The third anagenetic event is **extirpation**, happening at a rate ϵ, when the range of a lineage decreases by one region. For widespread lineages, extirpation translates into a decrease in the geographic range, for example $AB \rightarrow B$, $AB \rightarrow A$, or $ABC \rightarrow AB$ (Fig. 4.2). For lineages occurring in only a single region, the extirpation event causes the extinction of the lineage, $A \rightarrow \emptyset$ or $B \rightarrow \emptyset$ (Fig. 4.2).

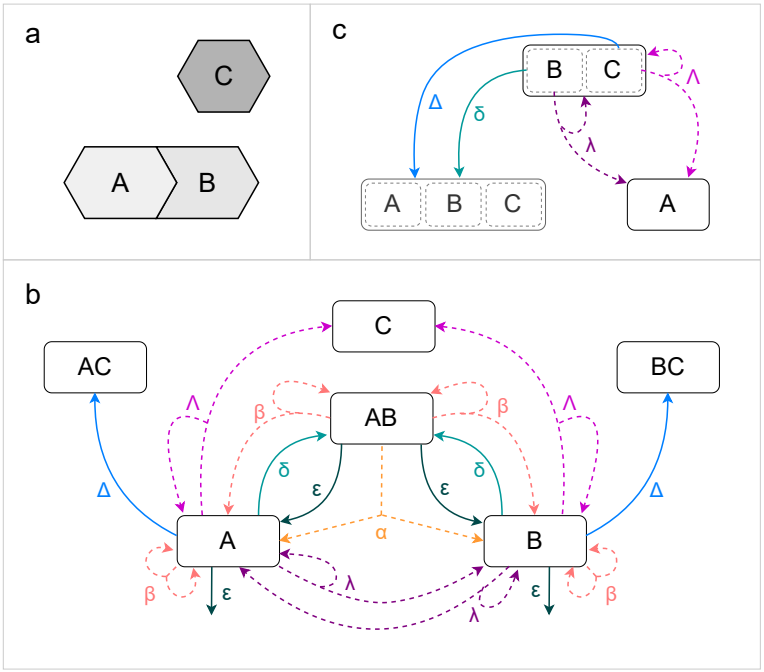


Figure 4.2: **An example of range-states and biogeographic events with a GeoSSE +J model. a**. In this scenario there are two adjacent regions, A and B, and one non-adjacent region, C. **b**. Examples of biogeographic events (not all the possible events are showed). The possible states are showed in boxes, and the biogeographic transition events from one range state to the other are shown with arrows. Cladogenetic events are shown in dashed lines, while anagenetic events are shown in continuous lines. Greek characters accompany the arrows to indicate the type of event. The model allows three types of anagenetic events: adjacent dispersal (δ), distant dispersal (Δ), and extirpation (ϵ). There are two types of speciation events that occur within the range of the ancestor lineage: range split speciation (α), and sympatric/subset speciation (β). Two types of speciation events involve a dispersal event from one of the daughter lineages: adjacent-jump speciation (λ), and distant-jump speciation (Λ). **c**. Events occurring in multi-region range states (BC for example) occur from single regions. In this example, a dispersal from $BC \rightarrow ABC$ can occur from B or from C. Because B and A are adjacent regions, a dispersal from b is an adjacent-dispersal(δ), while a dispersal from C is a distant-dispersal (Δ). Similarly, a cladogenetic jump, $BC \rightarrow BC, A$, can occur from B as an adjacent-jump speciation (λ) or from C as a distant-jump speciation (Λ).

Cladogenetic events. These events involve simultaneously an event of range evolution (e.g., a change in range state) and a cladogenetic event, *i.e*, the separation of an ancestral lineage into two descendant lineages. Therefore the notation of cladogenetic events needs to consider the range of the ancestor lineage and its two descendants as follows: "ancestor's range state $\rightarrow$ first descendant's range, second descendant's range". The PAw GeoSSE + J model accounts for four different cladogenetic events: range-split speciation, sympatric/subset speciation, adjacent-jump speciation, and distant-jump speciation.

First, **range-split speciation**, occurs at a rate α when the range of the ancestral lineage is split between the daughter lineages. For example, $AB \rightarrow A, B$ (Fig. 4.2). Because range-split speciation requires the splitting of the ancestor's range, it can only occur in lineages whose range includes more than two regions. Second, in a **sympatric/subset speciation** event one of the descendants inherits the full range of the ancestor, while the other lineage inherits a one-region subset of it. For example, a lineage with range AB going through sympatric/subset speciation will have one descendant with the same range AB, and another lineage with range A or range B; $AB \rightarrow AB, A$ or $AB \rightarrow AB, B$ (Fig. 4.2). A special case of this event is when the ancestor lineage occupies a single-region range. In this case, both lineages inherit the same region (*e.g.* $A \rightarrow A, A$, or $B \rightarrow B, B$; Fig. 4.2). Sympatric/subset speciation events occur at a rate (β).

The last two cladogenetic events involve a range-expansion in comparison to the ancestor's range. This happens when one of the descendants disperses to a new region while the other descendant maintains the ancestor's geographic range. These two events together are analougous to the +J parameter in the context of DEC models, commonly named "jump speciation". The fourth cladogenetic event is **adjacent jump-speciation**. It happens at a rate λ when one of the descendant lineages inherits the ancestor's range, while the other descendant disperses into an adjacent region. For example, $A \rightarrow A, B$ or $B \rightarrow A, B$ (Fig.4.2). Similarly, **distant-jump speciation** events happens when one of the descendant lineages disperses to a distant region, while the other descendant keeps the ancestor's range (*e.g.* $A \rightarrow A, C$ or $B \rightarrow B, C$; Fig. 4.2). Distant-jump speciation events occur at a rate Λ. As I noted for anagenetic dispersal events, when the ancestor is widespread, the jump-dispersal speciation events can occur from any of the regions in the ancestor's range.

Time stratification

The time stratification component of the model is set up through the usage of connectivity matrices that describe the relationship between the regions. At any given point in time, two regions are adjacent to each other if they are in contact through a distance that would facilitate dispersal. Otherwise the regions are non-adjacent, so the definition of adjacent vs non-adjacent requires the consideration of both the geological history of the regions and the dispersal biology of the study group, since barriers to dispersal are lineage-dependant.

The adjacency relationship among regions is scored in two complementary matrices: the adjacent-regions matrix and the non-adjacent regions matrix. In the adjacent-region matrix, two regions that are adjacent are scored as "1", while two regions that are not adjacent

are scored as "0". The opposite is true for the non-adjacent regions matrix. Because the adjacency of regions is expected to change through time, it is possible to assign different adjacency-matrices to different time intervals, and there is no limit on the number of time intervals that can be included. Through this coding system, it is possible to include information about the non-existence/unavailability of certain regions at a given time. Notably, oceanic islands have short geological life-span. By coding the relationship of a region as "0" in both the adjacent and non-adjacent matrices, this region becomes unavailable, reflecting the special geologic context of the region. Other instances in which this coding can be helpful are scenarios of ocean-level fluctuations that make regions unavailable for terrestrial organisms.

During the CTMC simulation, at a time τ, an event that involves adjacent dispersal (*i.e* adjacent-dispersal or adjacent-jump speciation) is possible only between regions scored as 1 in the "adjacent matrix" of the time interval that includes τ. On the other side, for the same time interval, distant dispersal events (*i.e.* distant-dispersal or distant-jump speciation) can occur only between regions scored as 1 in the "non-adjacency matrix".

4.2 Methods

In order to infer the biogeographic history of Aytoniaceae, I assembled a molecular dataset, then used it to infer a time calibrated phylogeny of Aytoniaceae. I also inferred divergence times using three different clock models and compared them using Bayes factors. Finally, I used PAw GeoSSE + J to infer the biogeographic history of Aytoniaceae.

Molecular data-set assembly

I used `MatrixMaker` to search and download the available *rbcL*, *matk*, and *trnL* sequences for members of the family Aytoniaceae. I screened those sequences to avoid the use of contaminants or mislabeled accessions. The accession numbers of the sequences used in this study are listed in Appendix C.1. I aligned the sequences using `mafft` (Katoh and Standley, 2013) including the `--auto` and `--adjustdirection` settings. I visualized the alignments in `aliview` (Larsson, 2014) and excluded ambiguously aligned regions from downstream analyses. The final dataset contained three markers for 40 species out of 92 known species in Aytoniaceae.

Estimating a time-calibrated phylogeny of Aytoniaceae

Gene trees and concatenated analysis. To explore the potential conflict between gene trees, I inferred gene trees for all the three markers (*rbcl*, *matK*, and *trnL*) using maximum likelihood. For these analyses, I included outgroups of Aytoniaceae (*Marchantia polymorpha*, *Dumortiera hirsuta*, *Targionia hypophylla*, *Conocephalum conicum*, and *Riccia fluitans*). For each gene, I did five independent runs. I used a GTR + I + Γ model and partitioned by

codon-position in `IQtree` (Minh et al., 2020). And I used 1000 rounds of ultra-fast bootstrap to assess the node support. I also inferred a tree using the concatenated dataset of the three markers. Similarly to the gene trees analyses, I ran five replicates of the analysis using a GTR + I + Γ model, partitioned by gene and codon position.

Time-divergence and topology co-estimation. Next, I co-estimated the topology and divergence times of the Aytoniaceae dataset in `RevBayes` (Höhna et al., 2016). My phylogenetic model had three main components: a tree model, a substitution model, and a clock model. For the tree model, I used a birth-death model. I used a GTR + Γ substitution model. Because the clock model is expected to have a big effect on divergence-time estimates, I used three different clock models: strict clock (Zuckerkandl and Pauling, 1965), uncorrelated exponential (UCE), and uncorrelated log-normal (UCLN) models. I partitioned the molecular dataset by marker and codon position, for a total of nine partitions. To avoid the use of outgroups—which could potentially violate the specified sampling parameter for the ingroup ($\rho = 0.44$) and ultimately biasing the age estimates—I used a combination of clade-constraints and a four-tip backbone tree to specify the root. Due to the absence of a fossil record of this group, I used a single secondary time constraint on the root of Aytoniaceae according to the estimates of Villarreal A. et al. (2016) who inferred an age of 111.3 My for this node with a 95% HPD between 80.2 My and 152.5 My. Therefore, I assigned a normal prior distribution on the root age with mean = 111.3 and standard deviation = 18, which matched the 95% HPD of Villarreal A. et al. (2016). I ran four MCMCs for each of the three models (that differed only on the clock model used). I visually checked for MCMC convergence and ESS values > 200 in `Tracer` (Rambaut et al., 2018) and I assessed the topological convergence of the analyses using the `R` package `RWTY` (Warren et al., 2017). All these analyses were performed in RevBayes (Höhna et al., 2016) and the phylogeny was plotted using RevGadgets (Tribble et al., 2022).

Clock-model comparison. In order to select the best clock model for the Aytoniaceae dataset, I compared them using Bayes factors, which require estimates of the marginal likelihoods of each model. For each clock model I ran a stepping stone analysis with 40 stones for 5000 generations. I estimated the marginal likelihood of each model using the path-sample and the stepping stone algorithms.

Inferring the biogeographic history of Aytoniaceae

Connectivity matrices and distribution data In order to include the paleo-biogeographic information, I identified 11 regions that are relevant for Aytoniaceae species and behave as geographical units—*i.e.*, they are contiguous and share a geological history through the last 200 my. I used the software `gplates` (Müller et al., 2018) to visualize the relative position of these regions through geological time, which helped both in the definition of the regions, and on the scoring of their relative position through time. More details about

the definition of the regions can be found in Appendix C.2. The relationship between pairs of regions was scored as either "adjacent" or "non-adjacent". Two regions were scored as adjacent if they were in contact, or if they shared a substantial length that would easily allow liverwort dispersal.

I translated the region relationships to a set of 16 connectivity matrices, each corresponding to a different time interval with a unique configuration of region relationships. Finally, I gathered information on the distribution of the 41 extant species of Aytoniaceae that were included in the phylogeny and coded their distribution in a presence-absence matrix for the 11 regions that I defined (Appendix C.3).

The analysis I implemented the PAw GeoSSE + J model in `RevBayes` (Höhna et al., 2016). Due to the large state space of the analysis—the state space is $2^N = 2048$—the analysis required the package `Tensorphylo` (May and Meyer, 2022) for likelihood calculations. I ran the analysis on a fixed topology: the MAP tree from the timetree inference analysis. I ran two MCMCs and assessed convergence using `Tracer`.

Post-processing I calculated the probability of certain events from the posterior sample. First, I calculated the per lineage rate of adjacent- and distant-dispersal as the sum of the anagenetic and cladogenetic events per type of dispersal ($\delta+\lambda$ and $\Delta+\Lambda$, respectively). Then, I calculated the per-lineage probability of speciation during adjacent- and distant-dispersal as follows:

$$P(\text{speciation during adjacent dispersal}) = \frac{\lambda}{\delta + \lambda} \tag{4.1}$$

$$P(\text{speciation during distant dispersal}) = \frac{\Lambda}{\Delta + \Lambda} \tag{4.2}$$

To visualize the results of the ancestral state reconstructions, for each node, I calculated the marginal probabilities of the ancestor occurring at each region. These marginal probabilities were then represented as histograms. Additionally, I plotted the maximum a posteriori ancestral range estimate. Due to the uncertainty of the ancestral state estimations, I computed the ratio of the state with the highest PP over the state with the second highest PP to reflect the preference for the MAP state (Fig. 4.6). Using this metric, a value of 10 would indicate that the MAP state is 10 times more preferred over the second most probable state.

4.3 Results

Time calibrated Aytoniaceae phylogeny

The exploratory ML inferences of gene trees and the concatenated molecular data produced topologies that only slightly disagree between closely related species (Appendix C.4). These

preliminary analyses, that included outgroups, resolved Aytoniaceae as monophyletic and agreed on the position of the root. When co-estimating the topology and divergence times, I conducted four independent runs of each model (three different clock models) and checked for convergence of the posterior sample. The preferred clock model was UCLN, which outperformed the strict clock model (2lnBF = 6.43) and the UCE clock model (2lnBF = 36.73) (Appendix C.5).

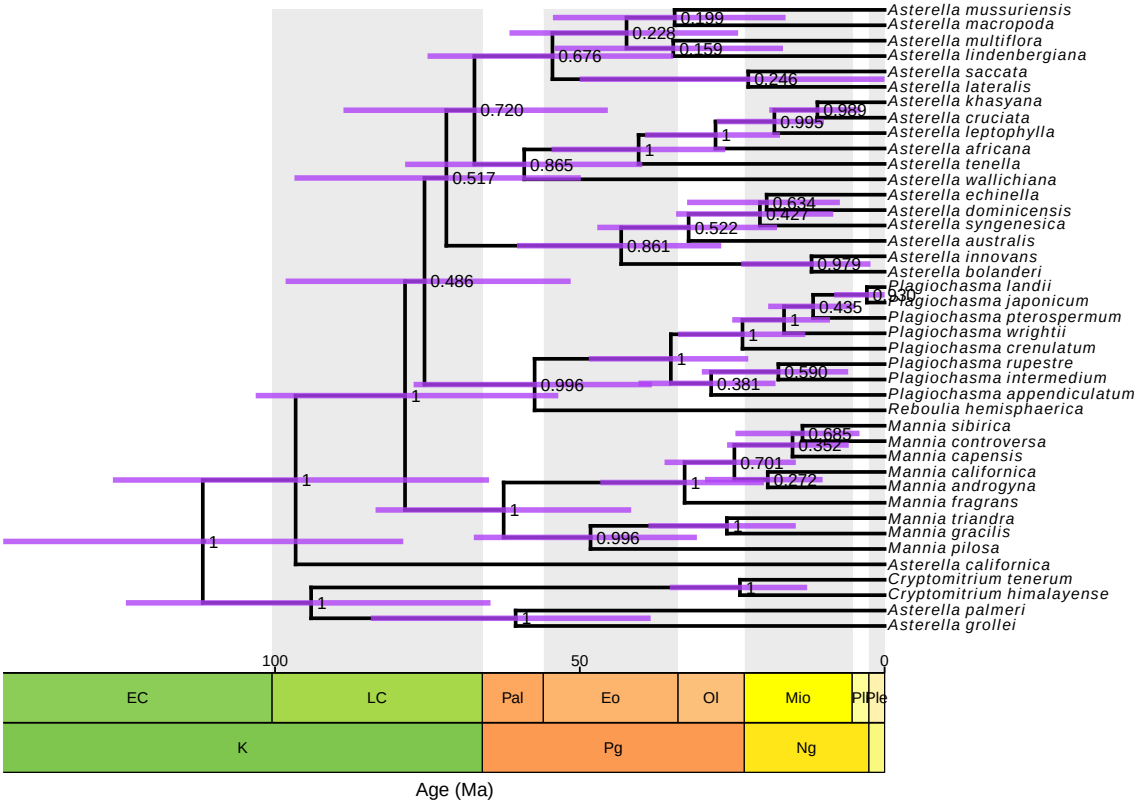


Figure 4.3: **Time calibrated phylogeny of Aytoniaceae.** Maximum *a posteriori* tree of the joint divergence-times and topology estimation using a UCLN clock model. The values at the nodes are the posterior probabilities of the nodes. The pink bars represent the 95% HPD of the ages of the nodes.

I summarize the results of the topology and time divergence estimation using the maximum a posteriori (MAP) tree of the preferred clock model in Fig. 4.3. The inferred phylogenetic relationships are consistent with previous analyses (*e.g.*, Bechteler et al., 2023; Villarreal A. et al., 2016). Except for *Asterella*, all the genera are recovered as monophyletic

with posterior probabilities of 1. The species assigned to the genus *Asterella* are recovered as polyphyletic and belong to three different lineages. The first lineage, *Asterella palmeri* + *Asterella grollei*, is retrieved as the sister group of the genus *Cryptomitrium* (PP = 1). The second lineage corresponds to *Calasterella californica*, which is inferred as the sister group of a large clade that includes the species in the genera *Mannia*, *Reboulia*, *Plagiochasma*, and the rest of the species in *Asterella*. The third lineage of *Asterella* species was inferred as the sister group of the genus *Plagiochasma* with a low support value (PP = 0.48). Aytoniaceae's crown group is estimated to have originated in the Cretaceous $\sim$ 109 Mya (with a HPD of 77-142 My). *Asterella californica* and *Reboulia hemisphaerica* are two lineages whose most recent common ancestor with their closest extant sister group are very old, $\sim$98 and $\sim$ 59 Myo respectively, so these two species are subtended by very long branches.

The biogeographic history of Aytoniaceae

In Fig. 4.4, I report the estimates of the anagenetic (Fig. 4.4a) and cladogenetic (Fig. 4.4b) parameters from the biogeographic model. My inferences show that for this group, range evolution occurs more often along the branches than at cladogenetic events, *e.g.*, the most common biogeographic event is adjacent dispersal (δ). However, given a dispersal event, cladogeneis is more likely if the event is a distant- rather than adjacent-dispersal, by a factor of more than three Increases in range occur more often through adjacent-dispersal, which is almost ten times more common than distant-dispersal (mean rates 0.02812 and 0.00306 respectively; Fig. 4.4c). However, given a disperal event, cladogeneis is more likely if the event is a distant- rather than adjacent-dispersal, but a factor of more than three (probabilities 0.294 and 0.086 respectively Fig. 4.4d).

The ancestral range reconstructions of Aytoniaceae are highly uncertain. In Fig. 4.5 I summarize the marginal probabilities of region occupancy at each node. The marginal probabilities are calculated from the range state probabilities, *i.e.*, the marginal probability of occurring in one region accounts for the probability of occurring in any range that contains that region. The height of the bar is the probability of being in each region, which explains why tips have probabilities of 1. Nodes closer to the root have more probability of being widespread than nodes closer to the tips.

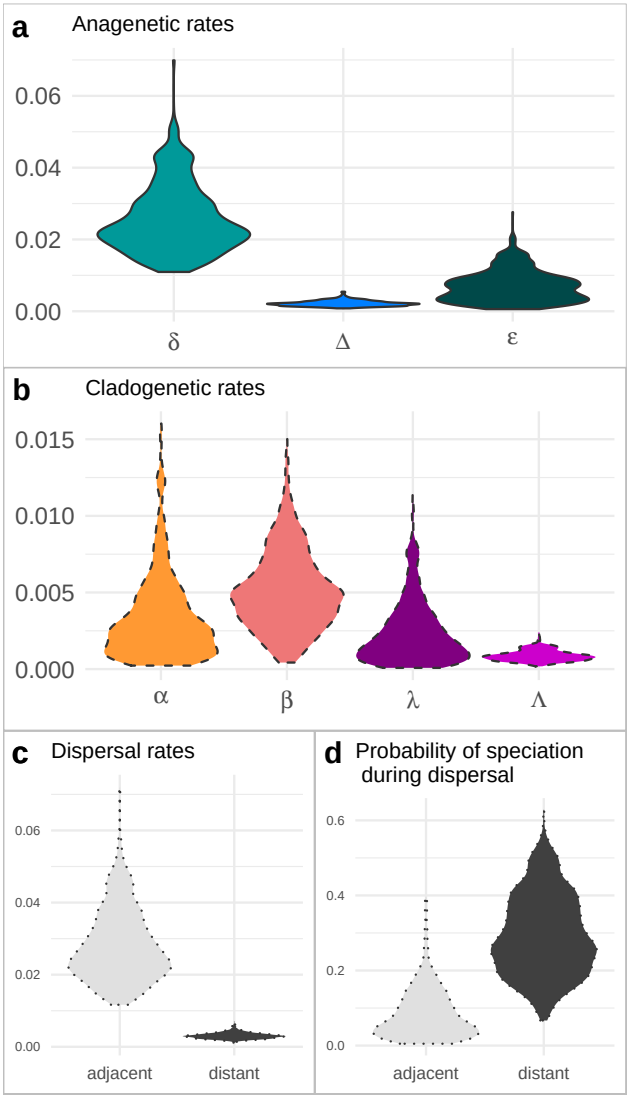


Figure 4.4: **Parameter estimates of the time-stratified GeoSSE + J biogeographic model applied to Aytoniaceae**. **a**. Estimates of the anagenetic parameters. **b**. Estimates of the cladogenetic events. **c**. Dispersal rates calculated as $adjacent\ dispersal = \delta + \lambda$ and $distant\ dispersal = \Delta + \Lambda$. **d**. Probability of speciation during each type of dispersal, calculated as $P(speciation\ during\ adjacent\ dispersal) = \frac{\lambda}{\delta+\lambda}$ and $P(speciation\ during\ distant\ dispersal) = \frac{\Lambda}{\Delta+\Lambda}$.

The MAP ancestral-range reconstruction of Aytoniaceae (Fig. 4.6) favors ancestral states with single region occupancy. The ranges are inferred with low posterior probabilities—mostly below 0.30 PP—and the state with the highest posterior probability is usually only slightly preferred over the state with the second highest posterior probability. According to

this inference, Aytoniaceae originated in North America and subsequently extended to Europe, India, and Southeast Asia. *Mannia* has primarily evolved in Europe and the Mediterranean region, while *Asterella*'s largest clade evolved in India with dispersal events to North America through Asia.

4.4 Discussion

Aytoniaceae phylogeny and divergence times

My crown-age estimate for Aytoniaceae is consistent with Villarreal A. et al. (2016), and the intergeneric splits are inferred as older than 50 mya. These divergence time estimates are outstandingly old compared to other groups of extant lineages of plants. As an extreme example, *C. californica* lies on a depauperate branch that diverged from its closest extant relative more than 100 mya (Fig. 4.3). The history of this lineage is most likely marked by extinction, and perhaps a slow speciation rate. Not surprisingly, the age estimates of divergences become very uncertain as we go back in time. This behaviour is expected given the little amount of information that we have on the ages, that are informed only by a secondary calibration at the node. A group like Aytoniaceae, with an nonexistent unambiguous fossil record would enormously benefit from a biogeographically informed dating analysis, where the biogeographic model informs the divergence-time estimates. The analysis presented in this paper is a step forward on that direction, but the implementation of a joint estimation of topology and biogeographic history remains a methodological and computational challenge.

The circumscription of *Asterella*

In both the preliminary ML analyses and the Bayesian timetree inference (wich co-estimated topology and divergence times) the genus *Asterella* was retrieved as polyphyletic, indicating that some species will need to be assigned to new genera. The largest clade of *Asterella* contains 18 of the 21 species used in the analysis, including the type species *Asterella tenella*; therefore, this group will remain *Asterella*. The second clade of *Asterella* (*Asterella palmeri* + *Asterella grollei*) diverged from its sister group *Cryptomitrium* more than 75 Mya. The last clade has been recently recognized as a new genus, *Calasterella* (Long and Zheng, 2023); it is inferred as the sister group of *Mannia* + *Plagiochasma* + the main clade of*Asterella* (Fig. 4.3). A proper taxonomic revision, including a full morphological re-evaluation, is needed to resolve the genus-level taxonomy of Aytoniaceae.

The biogeographic history of Aytoniaceae

The ancestral range reconstructions of Aytoniaceae are highly uncertain, *i.e.*, the posterior probability of the MAP ancestral state reconstructions tend to be low ($< 50\%$ for all the nodes; Fig. 4.6). Given this result, point estimates of the biogeographic history of Aytoniaceae need to be taken with caution as they will fail to capture the uncertainty of the

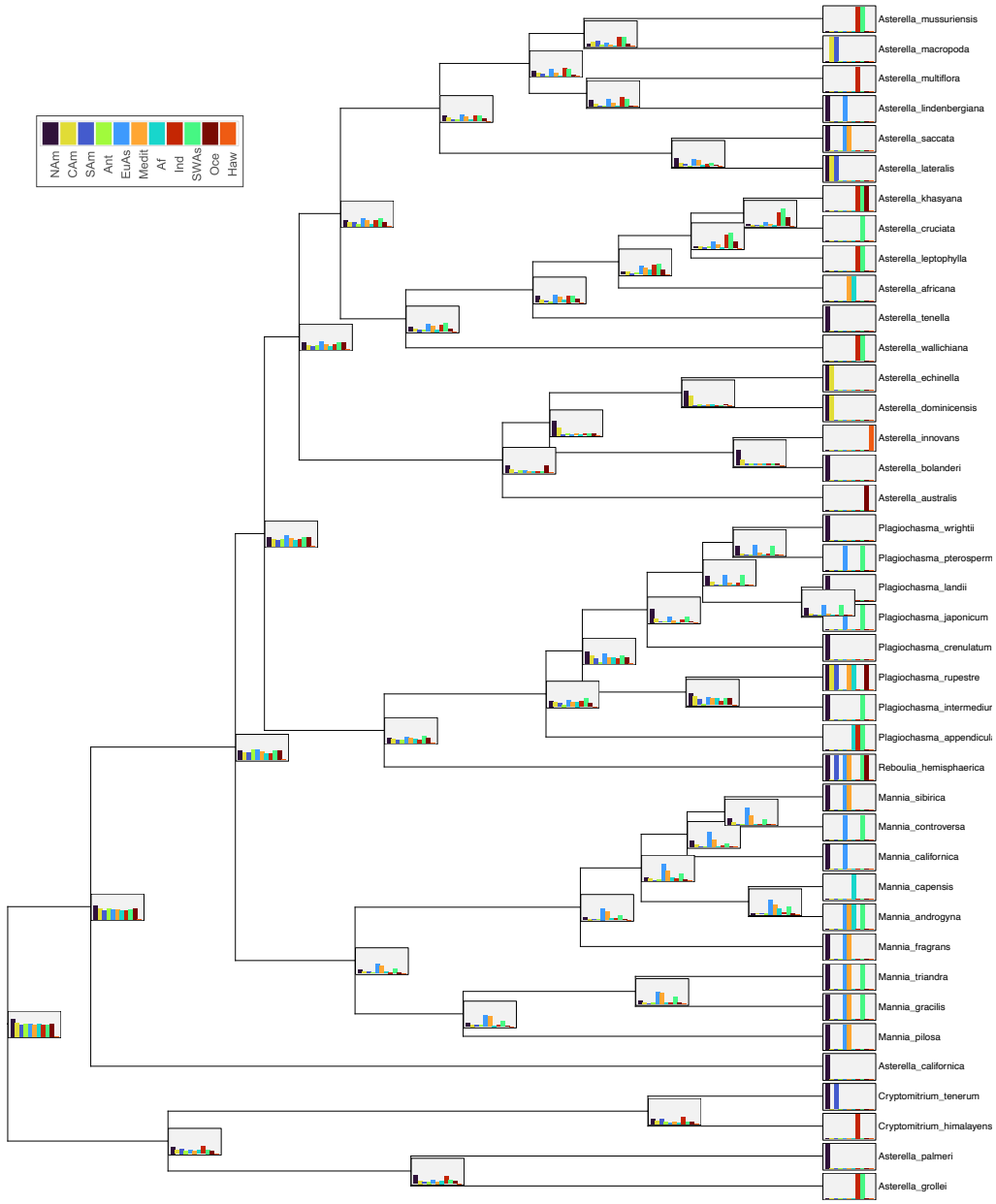


Figure 4.5: **Marginal probabilities of the region occupation.** The marginal probability of occupancy of each of the eleven regions is showed at each node.

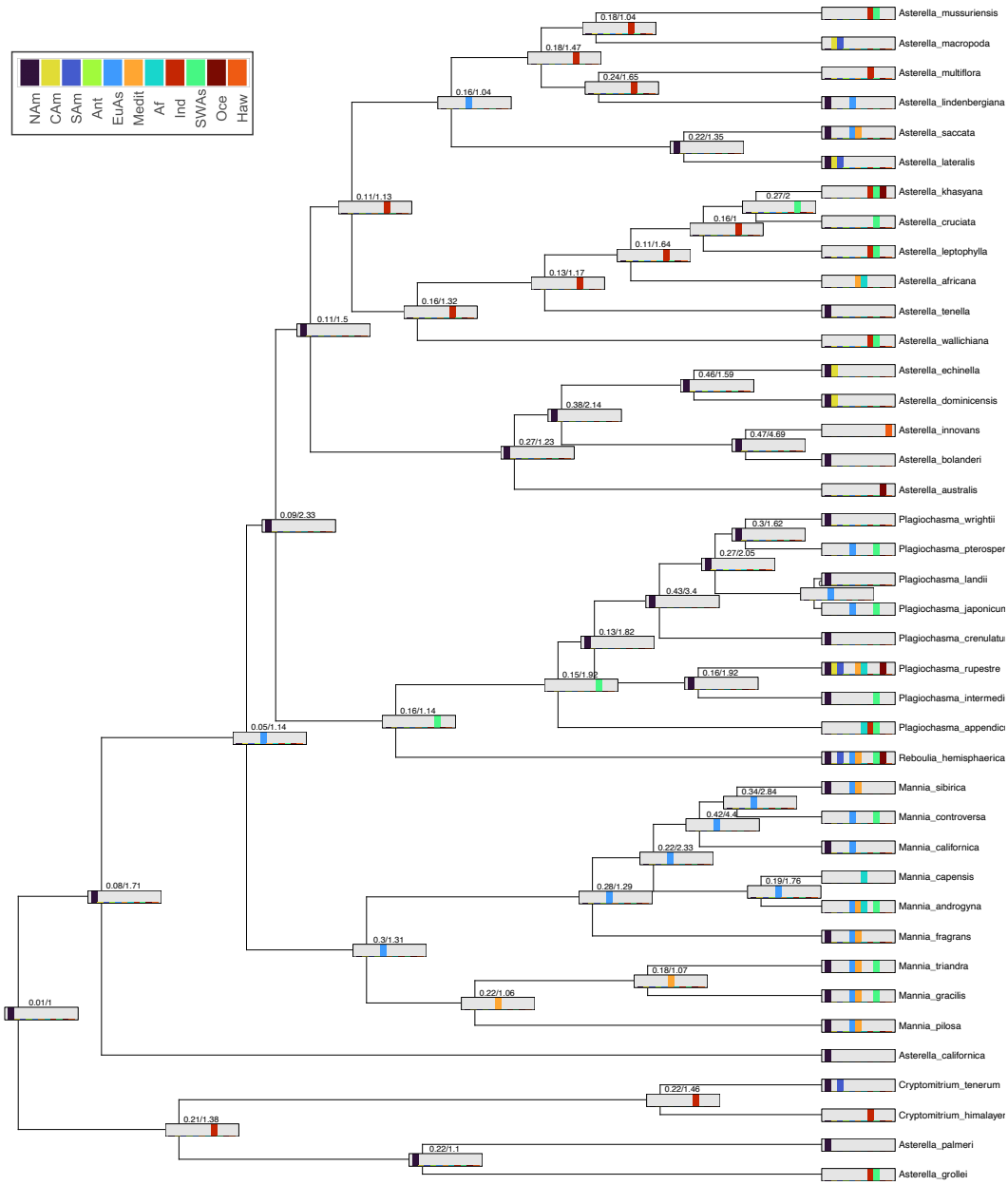


Figure 4.6: **Maximum a posteriori (MAP) ancestral-range reconstruction of Aytoniaceae.** At each node we show to the left of the "/" the posterior probability of the range, and to the right the preference for the MAP state computed as the ratio of the MAP state over the state with the second highest posterior probability.

inference. In particular, the MAP estimate seems to be biased towards single area ranges (Fig. 4.6). This feature of the inference deserves to be explored with a different dataset to explore whether this is a bias of the model or a feature of the dataset.

The marginal likelihoods of area occupancy for nodes close to the root tend to have multiple areas with high marginal probabilities which could be interpreted as a higher probability of widespread ranges. One of the interesting observations is that for Mesozoic nodes, the marginal probability of occupying Antarctica is comparable with the probability of occupying regions like Europe or Centroamerica, despite the fact that no extant Aytoniaceae lineage occurs in Antarctica. This inference—of considerable probably of an ancestral state that is not present in the tips of the phylogeny—derived completely from the connectivity matrices that reflect the reachability/availability of Antarctica during that period of time.

Overall, the precision with which we can reconstruct the biogeographic history of Aytoniaceae is limited. But some of the features that are consistent among the MAP estimate and the marginal likelihood estimates per area is that the crown Aytoniaceae most likely originated in North America and became widespread during the Cretaceous, a time when continent configuration facilitated dispersal. There is significant uncertainty on the states of intermediate nodes, but relatively simultaneously in the early Cenozoic *Mannia* occupied the Eurasian and Mediterranean regions while *Plagiochasma* occurred primarily in Northamerica and Southeast Asia.

Regardless of the uncertainty in the biogeographic inferences, the fact that the model produces biologically coherent parameter estimates (as discussed in the following section) is a positive outcome. Since range evolution is a very complex phenomenon, and the models we are currently using do not capture the complexity of this process, it is expected that estimates will have a high degree of uncertainty.

Drivers of the range evolution in Aytoniaceae

The parameter estimates obtained from the PAw GeoSSE + J analysis suggest that adjacent- and distant-dispersal events occur at different rates, and that each type of dispersal has a different probability of producing speciation. These rates are probably capturing the biological phenomenon of isolation by distance. The farther the distance between two individuals or populations, the less likely they are to be able to maintain gene flow, and the higher the chances of speciation. So even if the rate of dispersal to adjacent regions was 10 times higher than the rate of dispersal to distant regions, the rate of cladogenetic events during long distant dispersal was three times higher than during adjacent-dispersal (Fig. 4.4). This result shows that, despite their potential for long-range dispersal through spores and the extraordinarily broad range of some extant species, these liverworts are at least somewhat dispersal-limited: distances matters! As a consequence, changes in continental configurations have the potential to affect the evolution of organisms with relatively high vagility, and its is important to consider it when making biogeographic inferences.

Bibliography

D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

J. D. Bainard, L. L. Forrest, B. Goffinet, and S. G. Newmaster. Nuclear dna content variation and evolution in liverworts. *Molecular Phylogenetics and Evolution*, 68(3):619–627, 2013.

J. Bechteler, G. Peñaloza-Bojacá, D. Bell, J. Gordon Burleigh, S. F. McDaniel, E. Christine Davis, E. B. Sessa, A. Bippus, D. Christine Cargill, S. Chantanoarrapint, et al. Comprehensive phylogenomic time tree of bryophytes reveals deep relationships and uncovers gene incongruences in the last 500 million years of diversification. *American Journal of Botany*, 110(11):e16249, 2023.

D. Beerling. *The emerald planet: how plants changed Earth's history*. Oxford University Press, 2017.

J. L. Bowman, T. Kohchi, K. T. Yamato, J. Jenkins, S. Shu, K. Ishizaki, S. Yamaoka, R. Nishihama, Y. Nakamura, F. Berger, et al. Insights into land plant evolution garnered from the marchantia polymorpha genome. *Cell*, 171(2):287–304, 2017.

M. J. Christenhusz and J. W. Byng. The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3):201–217, 2016.

C. J. Cox. Land Plant Molecular Phylogenetics: A Review with Comments on Evaluating Incongruence Among Phylogenies. *Critical Reviews in Plant Sciences*, 37(2-3):113–127, may 2018a. ISSN 0735-2689. doi: 10.1080/07352689.2018.1482443. URL https://doi.org/10.1080/07352689.2018.1482443https://www.tandfonline.com/doi/full/10.1080/07352689.2018.1482443.

C. J. Cox. Land plant molecular phylogenetics: a review with comments on evaluating incongruence among phylogenies. *Critical Reviews in Plant Sciences*, 37(2-3):113–127, 2018b.

S. Dong, S. Zhang, L. Zhang, H. Wu, B. Goffinet, and Y. Liu. Plastid genomes and phylogenomics of liverworts (marchantiophyta): conserved genome structure but highest relative plastid substitution rate in land plants. *Molecular Phylogenetics and Evolution*, 161: 107171, 2021.

S. E. Fick and R. J. Hijmans. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12):4302–4315, 2017.

C. Finet, R. E. Timme, C. F. Delwiche, and F. Marlétaz. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology*, 20(24): 2217–2222, 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.11.035.

L. L. Forrest and B. J. Crandall-Stotler. A phylogeny of the simple thalloid liverworts (Jungermanniopsida, Metzgeriidae) as inferred from five chloroplast genes. (98):390–411, 2004.

E. E. Goldberg, L. T. Lancaster, and R. H. Ree. Phylogenetic Inference of Reciprocal Effects between Geographic Range Evolution and Diversification. *Systematic Biology*, 60 (4):451–465, 05 2011. ISSN 1063-5157. doi: 10.1093/sysbio/syr046. URL https://doi.org/10.1093/sysbio/syr046.

S. Greiner, P. Lehwark, and R. Bock. Organellargenomedraw (ogdraw) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic acids research*, 47(W1):W59–W64, 2019.

G. E. Griffith, J. M. Omernik, D. W. Smith, T. D. Cook, E. Tallyn, K. Moseley, and C. B. Johnson. Ecoregions of california. *US Geological Survey open-file report*, 1021:1–45, 2016.

J. Heinrichs, J. Hentschel, R. Wilson, K. Feldberg, and H. Schneider. Evolution of leafy liverworts (Jungermanniidae, Marchantiophyta): estimating divergence times from chloroplast DNA sequences using penalized likelihood with integrated fossil evidence. *Taxon*, 59(1): 31–44, 2007.

J. Heinrichs, K. Feldberg, J. Bechteler, L. Regalado, M. A. Renner, A. Schäfer-Verwimp, C. Gröhn, P. Müller, H. Schneider, and M. Krings. A Comprehensive Assessment of the Fossil Record of Liverworts in Amber. In *Transformative Paleobotany*, pages 213–252. Elsevier, 2018. ISBN 9780128130124. doi: 10.1016/B978-0-12-813012-4.00012-7. URL https://linkinghub.elsevier.com/retrieve/pii/B9780128130124000127.

L. V. A. Hernick, E. Landing, and K. E. Bartowski. Earth's oldest liverworts-*Metzgeriothallus sharonae* sp. nov. from the Middle Devonian (Givetian) of eastern New York, USA. *Review of Palaeobotany and Palynology*, 148(2-4):154–162, 2008. ISSN 00346667. doi: 10.1016/j.revpalbo.2007.09.002.

R. J. Hijmans, J. Van Etten, J. Cheng, M. Mattiuzzi, M. Sumner, J. A. Greenberg, O. P. Lamigueiro, A. Bevan, E. B. Racine, A. Shortridge, et al. Package 'raster'. *R package*, 734:473, 2015.

R. J. Hijmans, E. Williams, C. Vennes, and M. R. J. Hijmans. Package 'geosphere'. *Spherical trigonometry*, 1(7):1–45, 2017.

S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, 65(4):726–736, 2016.

P. W. Inglis, M. d. C. R. Pappas, L. V. Resende, and D. Grattapaglia. Fast and inexpensive protocols for consistent extraction of high quality dna and rna from challenging plant and fungal samples for high-throughput snp genotyping and sequencing applications. *PloS one*, 13(10):e0206085, 2018.

A. Kassambara. ggpubr:'ggplot2'based publication ready plots. *R package version*, page 2, 2018.

K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

M. Kay. ggdist: Visualizations of distributions and uncertainty in the grammar of graphics. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

D. Kelch, A. Driskell, and B. Mishler. Inferring phylogeny using genomic characters: a case study using land plant plastomes. *Molecular systematics of bryophytes [Monographs in systematic botany 98]*, 3:12, 2004.

C. C. Labandeira, S. L. Tremblay, K. E. Bartowski, and L. Vanaller Hernick. Middle Devonian liverwort herbivory and antiherbivore defence. *New Phytologist*, 202(1):247–258, 2014. ISSN 0028646X. doi: 10.1111/nph.12643.

B. Laenen, B. Shaw, H. Schneider, B. Goffinet, E. Paradis, A. Désamoré, J. Heinrichs, J. Villarreal, S. Gradstein, S. McDaniel, et al. Extant diversity of bryophytes emerged from successive post-mesozoic diversification bursts. *Nature communications*, 5(1):5134, 2014.

B. Laenen, A. Machac, S. R. Gradstein, B. Shaw, J. Patiño, A. Désamoré, B. Goffinet, C. J. Cox, J. Shaw, and A. Vanderpoorten. Geographical range in liverworts: Does sex really matter? *Journal of Biogeography*, 43(3):627–635, 2016. ISSN 13652699. doi: 10.1111/jbi.12661.

A. Larsson. Aliview: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278, 2014.

F.-W. Li, L.-Y. Kuo, K. M. Pryer, and C. J. Rothfels. Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated gc content. *Genome biology and evolution*, 8(8):2452–2458, 2016.

A.-M. Linde, S. Singh, J. L. Bowman, M. Eklund, N. Cronberg, and U. Lagercrantz. Genome evolution in plants: complex thalloid liverworts (marchantiopsida). *Genome Biology and Evolution*, 15(3):evad014, 2023.

G.-Q. Liu, L. Lian, and W. Wang. The molecular phylogeny of land plants: progress and future prospects. *Diversity*, 14(10):782, 2022.

D. G. Long and T.-X. Zheng. A new subfamily calasterelloideae and new genus calasterella for a phylogenetically and morphologically distinct member of the aytoniaceae. *Phytotaxa*, 606(3):225–230, 2023.

M. R. May and X. Meyer. TensorPhylo RevBayes plugin. 2022. https://bitbucket.org/mrmay/tensorphylo/src/master.

B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020.

C. D. Mirchandani, A. J. Shultz, G. W. Thomas, S. J. Smith, M. Baylis, B. Arnold, R. Corbett-Detig, E. Enbody, and T. B. Sackton. A fast, reproducible, high-throughput variant calling workflow for population genomics. *Molecular Biology and Evolution*, 41(1): msad270, 2024.

J. P. Mower and T. L. Vickrey. Structural diversity among plastid genomes of land plants. *Advances in botanical research*, 85:263–292, 2018.

R. D. Müller, J. Cannon, X. Qin, R. J. Watson, M. Gurnis, S. Williams, T. Pfaffelmoser, M. Seton, S. H. Russell, and S. Zahirovic. Gplates: building a virtual earth through deep time. *Geochemistry, Geophysics, Geosystems*, 19(7):2243–2261, 2018.

J. A. Nylander, U. Olsson, P. Alström, and I. SANMARTin. Accounting for phylogenetic uncertainty in biogeography: a bayesian approach to dispersal-vicariance analysis of the thrushes (aves: Turdus). *Systematic Biology*, 57(2):257–268, 2008.

E. Paradis, J. Claude, and K. Strimmer. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.

E. J. Pebesma et al. Simple features for r: standardized support for spatial vector data. *R J.*, 10(1):439, 2018.

M. N. Puttick, J. L. Morris, T. A. Williams, C. J. Cox, D. Edwards, P. Kenrick, S. Pressel, C. H. Wellman, H. Schneider, D. Pisani, and P. C. Donoghue. The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Current Biology*, 28(5): 733–745.e2, 2018. ISSN 09609822. doi: 10.1016/j.cub.2018.01.063. URL https://doi.org/10.1016/j.cub.2018.01.063.

R. R Core Team et al. R: A language and environment for statistical computing. 2013.

A. Rambaut, A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic biology*, 67(5):901–904, 2018.

S. A. Rensing. Plant Evolution: Phylogenetic Relationships between the Earliest Land Plants. *Current Biology*, 28(5):R210–R213, 2018. ISSN 09609822. doi: 10.1016/j.cub.2018.01.034. URL https://doi.org/10.1016/j.cub.2018.01.034.

K. S. Renzaglia, R. J. Du, D. L. Nickrent, and D. J. Garbary. Vegetative and reproductive innovations of early land plants: implications for a unified phylogeny. *Philosophical Transactions of the Royal Society B: Biological Sciences*, (355):769–793, 2000.

L. J. Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, (2):217–223, 2012.

F. Ronquist. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. *Systematic biology*, 46(1):195–203, 1997.

B. Sabater. On the edge of dispensability, the chloroplast ndh genes. *International Journal of Molecular Sciences*, 22(22):12505, 2021.

D. B. Schill, D. G. Long, and L. L. Forrest. A molecular phylogenetic study of Mannia (Marchantiophyta, Aytoniaceae) using chloroplast and nuclear markers. *The Bryologist*, 113(1):164–179, 2010. ISSN 0007-2745. doi: 10.1639/0007-2745-113.1.164.

K. P. Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011.

A. L. Sessions, D. M. Doughty, P. V. Welander, R. E. Summons, and D. K. Newman. The continuing puzzle of the great oxidation event. *Current biology*, 19(14):R567–R574, 2009.

H. B. Shaffer, E. Toffelmier, R. B. Corbett-Detig, M. Escalona, B. Erickson, P. Fiedler, M. Gold, R. J. Harrigan, S. Hodges, T. K. Luckau, et al. Landscape genomics to enable conservation actions: the california conservation genomics project. *Journal of Heredity*, 113(6):577–588, 2022.

J. A. Shaw. Biogeographic patterns and cryptic speciation in bryophytes. *Journal of Biogeography*, 28(2):253–261, 2001. ISSN 03050270. doi: 10.1046/j.1365-2699.2001.00530.x.

J. A. Shaw, P. Szövényi, and B. Shaw. Bryophyte diversity and evolution: Windows into the early evolution of land plants. *American Journal of Botany*, 98(3):352–369, 2011. doi: 10.3732/ajb.1000316.

L. Söderström, A. Hagborg, M. Von Konrat, S. Bartholomew-Began, D. Bell, L. Briscoe, E. Brown, D. C. Cargill, D. P. Costa, B. J. Crandall-Stotler, et al. World checklist of hornworts and liverworts. *PhytoKeys*, 59:1, 2016.

M. Tillich, P. Lehwark, T. Pellizzer, E. S. Ulbricht-Jones, A. Fischer, R. Bock, and S. Greiner. Geseq–versatile and accurate annotation of organelle genomes. *Nucleic acids research*, 45 (W1):W6–W11, 2017.

C. M. Tribble, W. A. Freyman, M. J. Landis, J. Y. Lim, J. Barido-Sottani, B. T. Kopperud, S. Hhna, and M. R. May. Revgadgets: An r package for visualizing bayesian phylogenetic analyses from revbayes. *Methods in Ecology and Evolution*, 13(2):314–323, 2022.

J. C. Villarreal A., B. J. Crandall-Stotler, M. L. Hart, D. G. Long, and L. L. Forrest. Divergence times and the evolution of morphological complexity in an early land plant lineage (marchantiopsida) with a slow molecular rate. *New Phytologist*, 209(4):1734–1746, 2016.

D. L. Warren, A. J. Geneva, and R. Lanfear. Rwty (r we there yet): an r package for examining convergence of bayesian phylogenetic analyses. *Molecular Biology and Evolution*, 34 (4):1016–1020, 2017.

H. Wickham. An introduction to ggplot: An implementation of the grammar of graphics in r. *Statistics*, pages 1–8, 2006.

H. Wickham and H. Wickham. *Data analysis*. Springer, 2016.

H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, et al. Welcome to the tidyverse. *Journal of open source software*, 4(43):1686, 2019.

Y.-L. Xiang, X.-J. Jin, C. Shen, X.-F. Cheng, L. Shu, and R.-L. Zhu. New insights into the phylogeny of the complex thalloid liverworts (marchantiopsida) based on chloroplast genomes. *Cladistics*, 38(6):649–662, 2022.

Y. Yu, H.-M. Liu, J.-B. Yang, W.-Z. Ma, S. Pressel, Y.-H. Wu, and H. Schneider. Exploring the plastid genome disparity of liverworts. *Journal of Systematics and Evolution*, 57(4): 382–394, 2019.

E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins*, pages 97–166. Elsevier, 1965.

Appendix A

Appendix for Chapter 2

A.1 Draft assembly errors

Below I show the sequences that were identified as artificially duplicated in the draft assembly of the *C. californica* plastome. These sequences were manually removed from the draft assembly prior to further analyses.

Inverted repeat A:
Error 1

>TTTGGATTGGTAAGGTGATAATTTATTTATATAAAATAATTAATGGTTGCATAGGTTTATTCAATTTTT
GAGATTTCAGTTTCAAAAATTTGAATGAAAATAGAAAAAATATATTATTGAATATATTATTTATAGATAAT
TAGTTGGGTTTTTTTTTTTTAAGTCAGATTTGAAGTTATTTTTTTTTTTCTGAGAGTAGGGATATAACTCA
GCGGTAGAGTATCACCTTGACGTGGTGGAAGTCATCAGTTCGAACCTGATTATCCCTAAAAATATTTGAAA
CATTTTGATTTGTTGCTTTCTTGTTGTTGAAAGAGGCTTGTGGGATTGACATAATAGGGTAGGTATGGGTA
TACTAGAAATGAGCTTCAAGCTAATATGAAGCGAATGAAAAATAAACATAAGTTATCTATCTCTTAGGAGG
GAAGACGATTTGAAATCTGCTTTGTTTACGAAGAAGGAAGCTATAAGTAAAACTAATATAACTATGAATCT
CATGGAGAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCATGCTTAACACATGCAAGTCGTACGGGAAG
CATCCTAGTGGTGTTTCCAGTGGCGGACGGGTGAGTAACGCGTAAGAACCTGTCCTTGGGAGGGGGACAAC
AGCTGGAAACGGTTGCTAATACCCCATAGGCTGAGGAGCAAAAGGAGGAATCCNNNNNNNNNNAGTTTTTT
TTTG

Error 2

>TCTCCTCAGGTAGGACTTGAACCTACGACCAATCGGTTAACAGCCGACCGCTCTACCACTGAGCTACT
GAGGAACAATGAGTTCAATCCTAAAAACATTCAAAAACTTTTCAACTTAGAATTAGCCCATAAACTGTTCA
AAGAACCAAAAAATTATTGGATTAAGAATGGAAATAACTTTCAGTACACCCTACCCTATTTTATTATATGG
AAAAAAGATAACGATAGCAATCCCCCTAAACTCTCCATCGAAAATTTTTGAGACAAGGGGGAGGCGGTCAA
CCATCACTATGATCNNNNNNNNNNACTAAGTAGGTTTATTTAGAAAACGCAATGAAATTGTGGCAATAACA
TCTCACGAAATTGGATGACTGCATTTTTGAGAAATTGCTCGTACATTATTTGAGTAAGTTATCTATTAGTT
AGTTGAAAATAACTAATTAGAGCTATATCAATATTCAATTTATAGCCAAATTGATATTGGTACATCCTAGT

AATTCTATTCTCAATGCTTTTCAAGCAAAAAAATGACATTGTTACTAGTAAATAACAAGTCTAAAACGAAT
AGACACTTTCACTCAAATCAGGCTCTGATACTGTCATCATAAAAAAAAGAATCCATTTGCTTTTCTTAGTT
TTTCAATACTCCACATAGATCATTGTTTCCATTTTTTTAGATTACTGATAATCATCAGTACATTTTTATTT
TTGTTGCAATTTGACATTAGTATAAACAATGAAAAAGAACACCTAGTTCTAAAAGTAAAAACAAGCA

Inverted repeat B
Error 1

>CTCTCAGAAAAAAAAAAATAACTTCAAATCTGACTTAAAAAAAAAAAACCCAACTAATTATCTATAAATAA
TATATTCAATAATATATTTTTTCTATTTTCATTCAAATTTTTGAAACTGAAATCTCAAAAATTGAATAAACC
TATGCAACCATTAATTATTTTATATAAATAAATTATCACCTTACCAATCCAAACAAAAAAAAACTCTACCTC
CACGCGGCATTGCTCCGTCAGGCTTTCGCCCATTGCGGAAAATTCCCCACTGCTGCCTCCCGTAAGAGTCTG
GGCCGTGTCTCAGTCCCAGTGTGGCTGATCATCCTCTCAGACCAGCTACTGATCGTCGCCTTGGTAAGCTAT
TACCTCACCAACTAGCTAATCAGACGCAAGCCCCTCCTTAGGCGGATTCCTCCTTTTGCTCCTCAGCCTATG
GGGTATTAGCAACCGTTTCCAGCTGTTGTCCCCCTCCCAAGGACAGGTTCTTACGCGTTACTCACCCGTCCG
CCACTGGAAACACCACTAGGATGCTTCCCGTACGACTTGCATGTGTTAAGCATGCCGCCAGCGTTCATCCTG
AGCCAGGATCAAACTCTCCATGAGATTCATAGTTATATTAGTTTTACTTATAGCTTCCTTCTTCGTAAACAA
AGCAGATTTCAAATCGTCTTCCCTCCTAAGAGATAGATAACTTATGTTTATTTTTCATTCGCTTCATATTAG
CTTGAAGCTCATTTCTAGTATACCCATACCTACCCTATTATGTCAATCCCACAAGCCTCTTTCAACAACAAG
AAAGCAACAAATCAAAATGTTTCAAATATTTTTAGGGATAATCAGGTTCGAACTGATGACTTCCACCACGTC
AAGGTGATACTCTACCGCTGAGTTATATCCCTA

Error 2

>TGCTTGTTTTTACTTTTAGAACTAGGTGTTCTTTTTCATTGTTTATACTAATGTCAAATTGCAACAAAAAT
AAAAATGTACTGATGATTATCAGTAATCTAAAAAAATGGAAACAATGATCTATGTGGAGTATTGAAAAACTA
AGAAAAGCAAATGGATTCTTTTTTTTTATGATGACAGTATCAGAGCCTGATTTGAGTGAAAGTGTCTATTCGT
TTTAGACTTGTTATTTACTAGTAACAATGTCATTTTTTTGCTTGAAAAGCATTGAGAATAGAATTACTAGGA
TGTACCAATATCAATTTGGCTATAAATTGAATATTGATATAGCTCTAATTAGTTATTTTCAACTAACTAATA
GATAACTTACTCAAATAATGTACGAGCAATTTCTCAAAAATGCAGTCATCCAATTTCGTGAGATGTTATTGC
CACAATTTCATTNNNNNNNNNNGATCATAGTGATGGTTGACCGCCTCCCCCTTGTCTCAAAAATTTTCGATG
GAGAGTTTAGGGGGATTGCTATCGTTATCTTTTTTCCATATAATAAAATAGGGTAGGGTGTACTGAAAGTTA
TTTCCATTCTTAATCCAATAATTTTTTGGTTCTTTGAACAGTTTATGGGCTAATTCTAAGTTGAAAAGTTTT
TGAATGTTTTTAGGATTGAACTCATTGTTCCTCAGTAGCTCAGTGGTAGAGCGGTCGGCTGTTAACCGATTG
GTCGTAGGTTCAAGTCCTACCTGAGGAGA

A.2 Liverwort chloroplast genomes information

The plastome sequences used in this work were obtained from different sources. Those obtained from Xiang et al. (2022) were not available on Genbank and were instead downloaded directly from their Supplemental Materials.

Order	Family	Species	Source	Acc. num.
Marchantiales	Aytoniaceae	*Asterella cruciata*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Asterella leptophylla*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Asterella mussuriensis*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Asterella wallichiana*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Asterellopsis grollei*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Cryptomitrium himalayense*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Mannia controversa*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Mannia fragrans*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Plagiochasma appendiculatum*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Plagiochasma intermedium*	Xiang et al. (2022)	NA
Marchantiales	Aytoniaceae	*Reboulia hemisphaerica*	Genbank	MK477551
Jungermaniales	Lepidoziaceae	*Bazzania praerupta*	Genbank	MH064512.1
Jungermaniales	Herbertaceae	*Hebertus dicranus*	Genbank; Dong et al. (2021)	MK645822
Jungermaniales	Plagiochilaceae	*Plagiochila chinensis*	Genbank	MH064511
Jungermaniales	Trichocoleaceae	*Trichocolea tomentella*	Genbank; Dong et al. (2021)	MK645849
Jungermaniales	Calypogeiaceae	*Calypogeia fissa*	Genbank	MH064514
Jungermaniales	Scapaniaceae	*Scapania ciliata*	Genbank	MH064513
Jungermaniales	Schistochilaceae	*Schistochilia macrodonta*	Genbank	MH064506
Porellales	Lejeuneaceae	*Cheilolejeunea xanthocarpa*	Genbank	MH064504
Porellales	Lejeuneaceae	*Cololejeunea lanciloba*	Genbank	MH064505
Porellales	Lejeuneaceae	*Ptychanthus striatus*	Genbank; Dong et al. (2021)	MK645842
Porellales	Jubulaceae	*Jubula hutchinsiae*	Genbank	MH064509
Porellales	Frullaniaceae	*Frullania nodulosa*	Genbank	MH064510
Porellales	Radulaceae	*Radula japonica*	Genbank; Dong et al. (2021)	MK645843
Ptilidiales	Ptilidiaceae	*Ptilidium ciliare*	Genbank; Dong et al. (2021)	MK645841
Metzgeriales	Aneuraceae	*Aneura pinguis*	Genbank; Dong et al. (2021)	MK645811
Metzgeriales	Aneuraceae	*Aneura mirabilis*	Genbank	NC010359
Metzgeriales	Aneuraceae	*Riccardia latifrons*	Genbank; Dong et al. (2021)	MK645844
Metzgeriales	Metzgeriaceae	*Metzgeria leptoneura*	Genbank; Dong et al. (2021)	MK645830
Fossombroniales	Makinoaceae	*Makinoa crispata*	Genbank; Dong et al. (2021)	MK645828
Fossombroniales	Fossombroniaceae	*Fossombronia cristula*	Genbank; Dong et al. (2021)	MK645818
Pelliales	Pelliaceae	*Pellia endiviifolia*	Genbank	NC019628
Marchantiales	Ricciaceae	*Riccia cavernosa*	Genbank; Dong et al. (2021)	MK645845
Marchantiales	Conocephalaceae	*Conocephalum conicum*	Genbank; Dong et al. (2021)	MK645816
Blasiales	Blasiaceae	*Blasia pusilla*	Genbank; Dong et al. (2021)	MK645815
Haplomitriales	Haplomitriaceae	*Haplomitrium mnioides*	Genbank; Dong et al. (2021)	MK645820
Treubiales	Treubiaceae	*Treubia lacunosa*	Genbank; Dong et al. (2021)	MK645848
Sphaerocarpales	Sphaerocarpaceae	*Sphaerocarpos texanus*	Genbank	MT023020.1
Sphagnales	Sphagnaceae	*Sphagnum palustre*	Genbank	NC_030198.1

Table A.1: Species of liverworts used in this work, with the source of the plastome data.

A.3 Read coverage of the chloroplast assembly

The figure below shows the read coverage across the length of the plastome of *C. californica*. The mean coverage is 2000 reads per position.

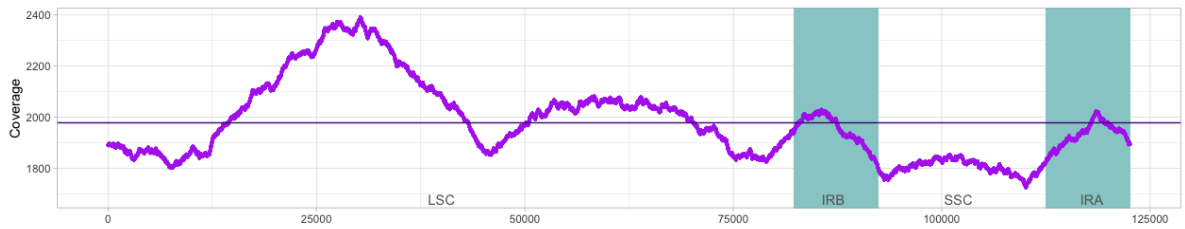


Figure A.1: Read coverage of the *de novo* chloroplast genome assembly of *Calasterella californica*. The horizontal line indicates the mean coverage. The vertical blue bands indicate the two copies of the inverted repeat region.

A.4 *C. californica*'s chloroplast size comparison

The figure below is a visual comparison of the newly assembled chloroplast genome of *C. californica* with other liverwort plastomes.

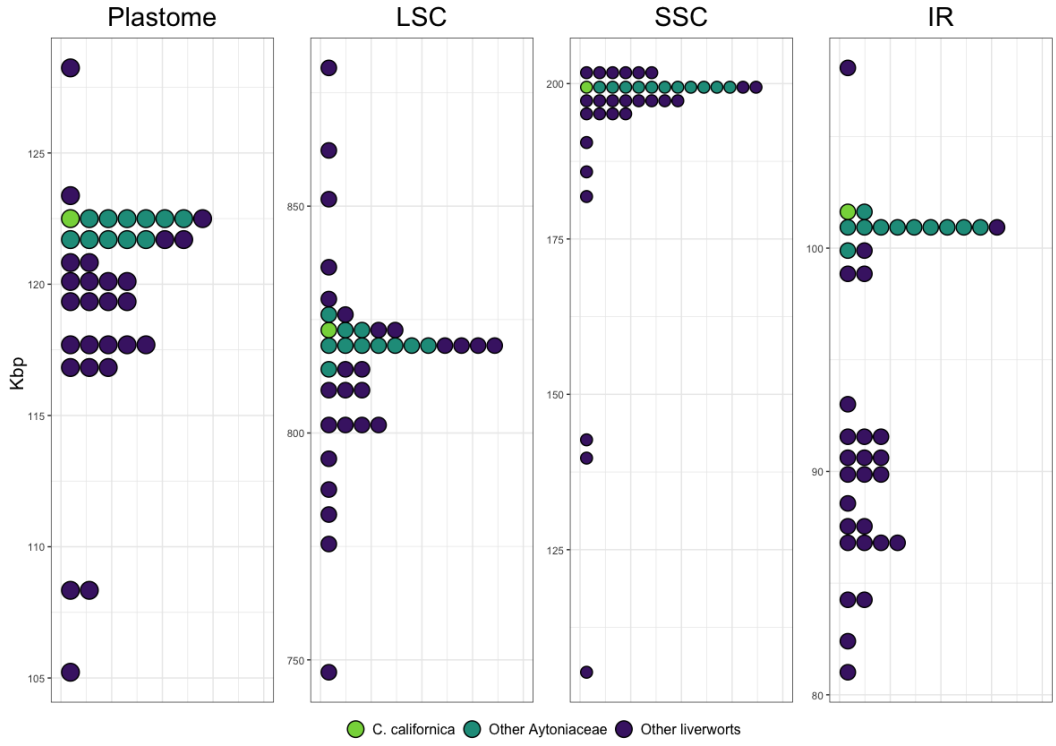


Figure A.2: Comparison of *C. californica*'s plastome region size (light green) with the plastomes of other Aytoniaceae liverworts (green) and other liverwort species (purple).

Table A.2: Detailed plastome size and GC content data for species in Aytoniaceae.

Species	plastome genome	GC content	IRA length	IRB length	LSC length	SSC length
Asterella_cruciata	122355	28.7	10112	10112	82190	19941
Asterella_leptophyla	122166	28.7	10103	10103	82002	19958
Asterella_mussuriensis	121798	28.8	9996	9996	81872	19934
Asterella_wallichiana	122281	28.7	10111	10111	82072	19987
Asterellopsis_grollei	121981	28.9	10101	10101	81869	19910
Cryptomitrium_himalayense	121635	28.7	10141	10141	81354	19999
Mannia_controversa	122043	28.7	10120	10120	81873	19930
Mannia_fragans	122176	28.8	10121	10121	81926	20008
Plagiochamsma_appendiculatum	122831	28.6	10081	10081	82656	20013
Plagiochasma_intermedium	121866	28.8	10067	10067	81817	19915
Reboulia_hemisphaerica	122596	28.8	10112	10112	82421	19951
Calasterella_californica	122592	28.8	10137	10137	82287	20030

A.5 Liverwort relationships from sequence data

I inferred the relationships among the liverworts used in this chapter using a standard DNA sequence evolution model. For this, I first extracted the sequence of rbcL and rps4—two chloroplast markers commonly used for phylogenetic inference—for each species from the plastomic dataset. Then, I aligned each marker using `MAFFT`, I concatenated the two markers in a single dataset and used `IQtree` to do a standard inference under a GTR + G model of molecular evolution.

In fig A.3, I show the ML tree inferred with bootstrap support values. The relationships obtained are congruent with out current understanding of liverwort relationships. This tree was used as a base to conduct analyses about the evolution of structural changes.

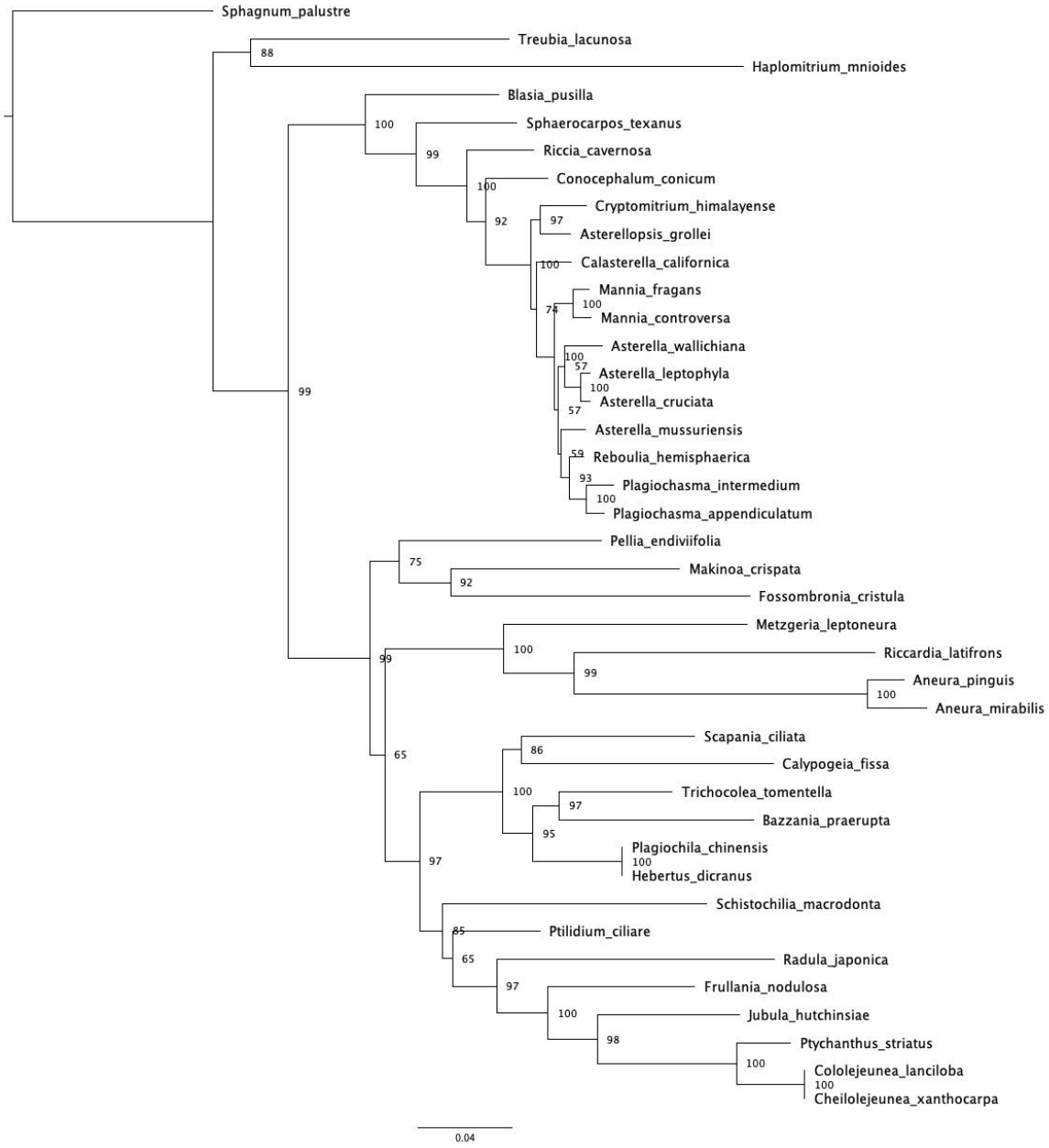


Figure A.3: Phylogenetic relationships of the liverworts studied in this chapter based on sequence data from two plastid markers. The support values indicate bootstrap support.

A.6 Ancestral state reconstruction for gene presence-absence in the chloroplast of liverworts

Figure A.4: **Ancestral state reconstruction for gene presence-absence (part 1).** The genes above each tree share the history represented by each tree. Yellow denotes presence of the gene, green denotes absence, and purple denotes alternative versions.

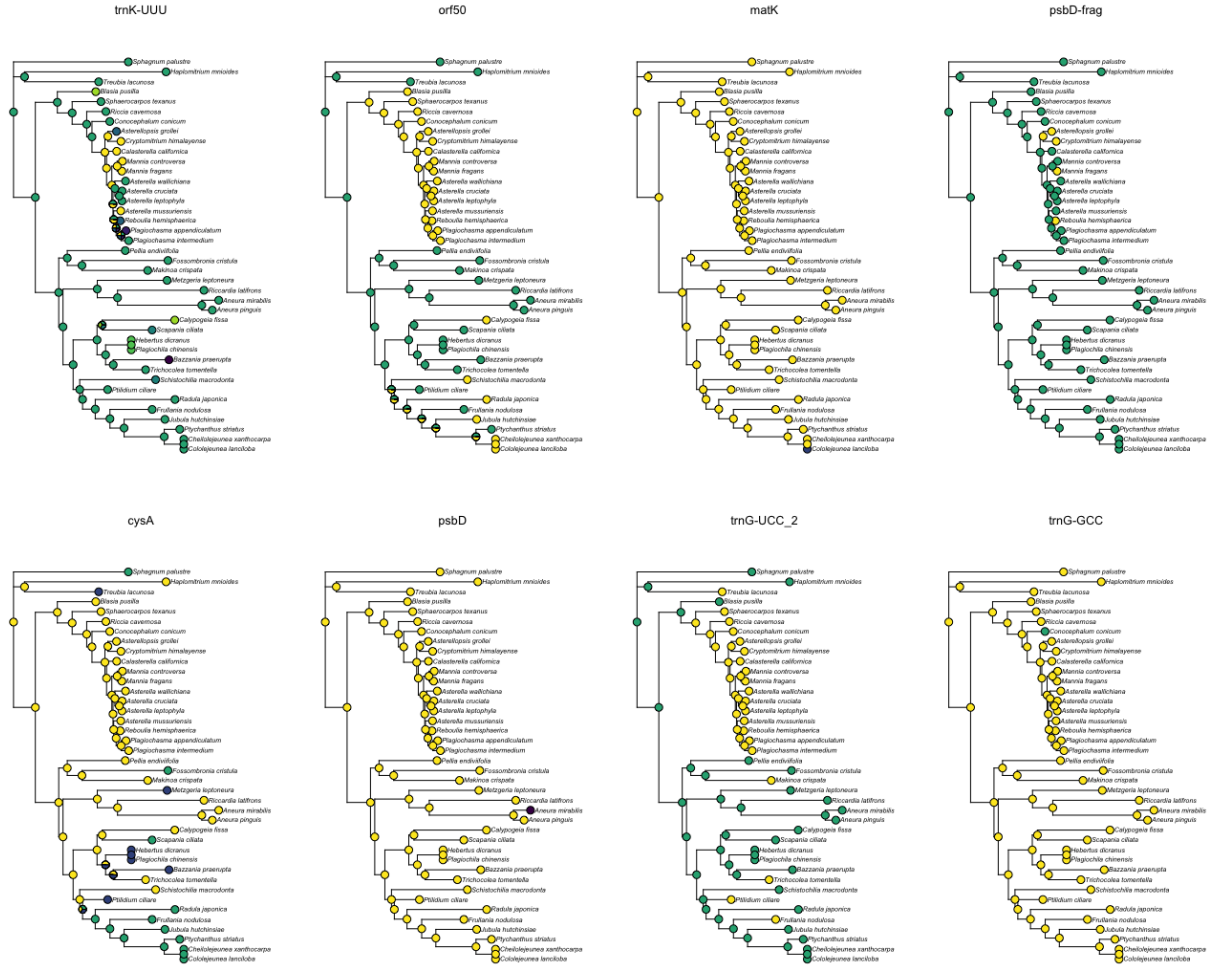


Figure A.5: **Ancestral state reconstruction for gene presence-absence (part 2).** The genes above each tree share the history represented by each tree. Yellow denotes presence of the gene, green denotes absence, and purple denotes alternative versions.

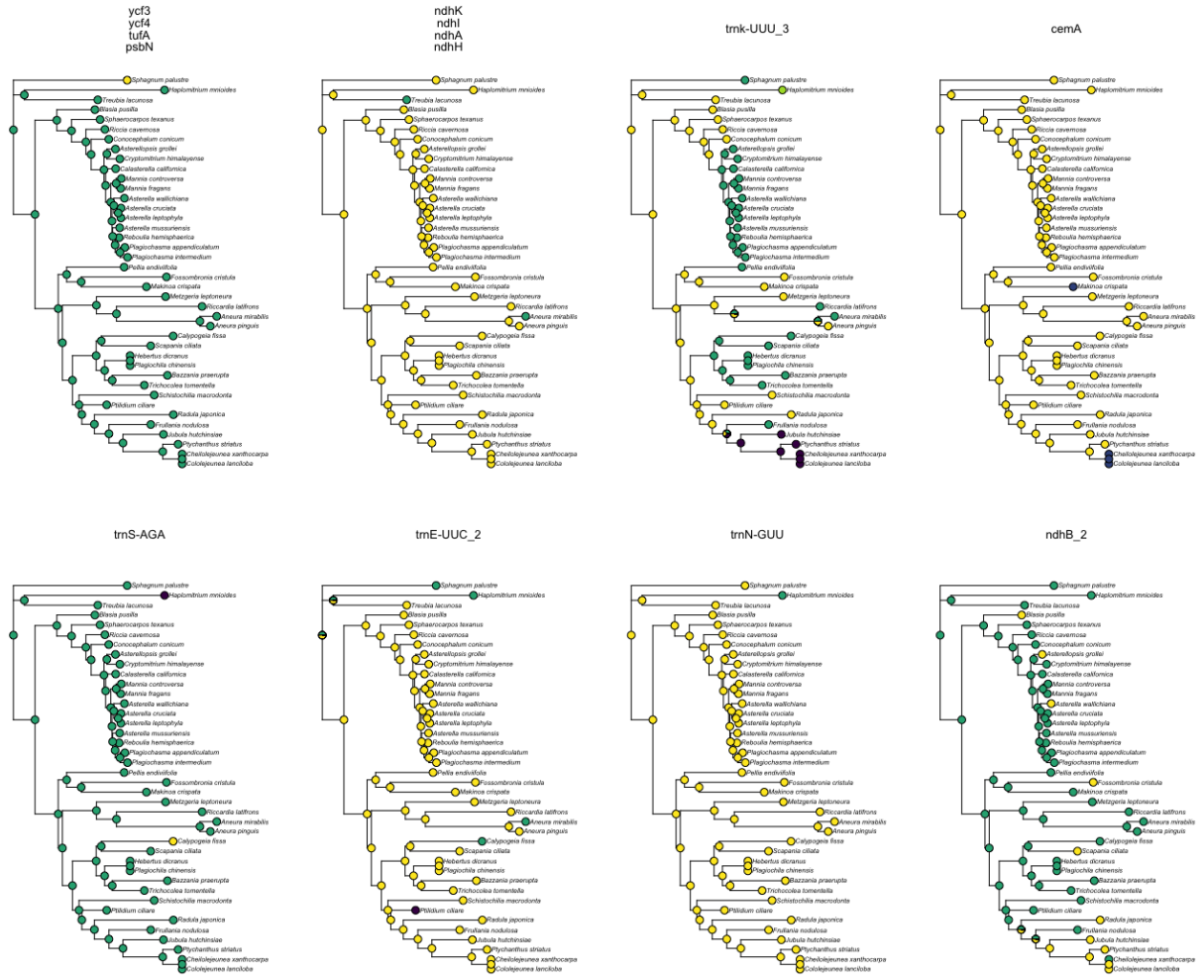


Figure A.6: **Ancestral state reconstruction for gene presence-absence (part 3).** The genes above each tree share the history represented by each tree. Yellow denotes presence of the gene, green denotes absence, and purple denotes alternative versions.

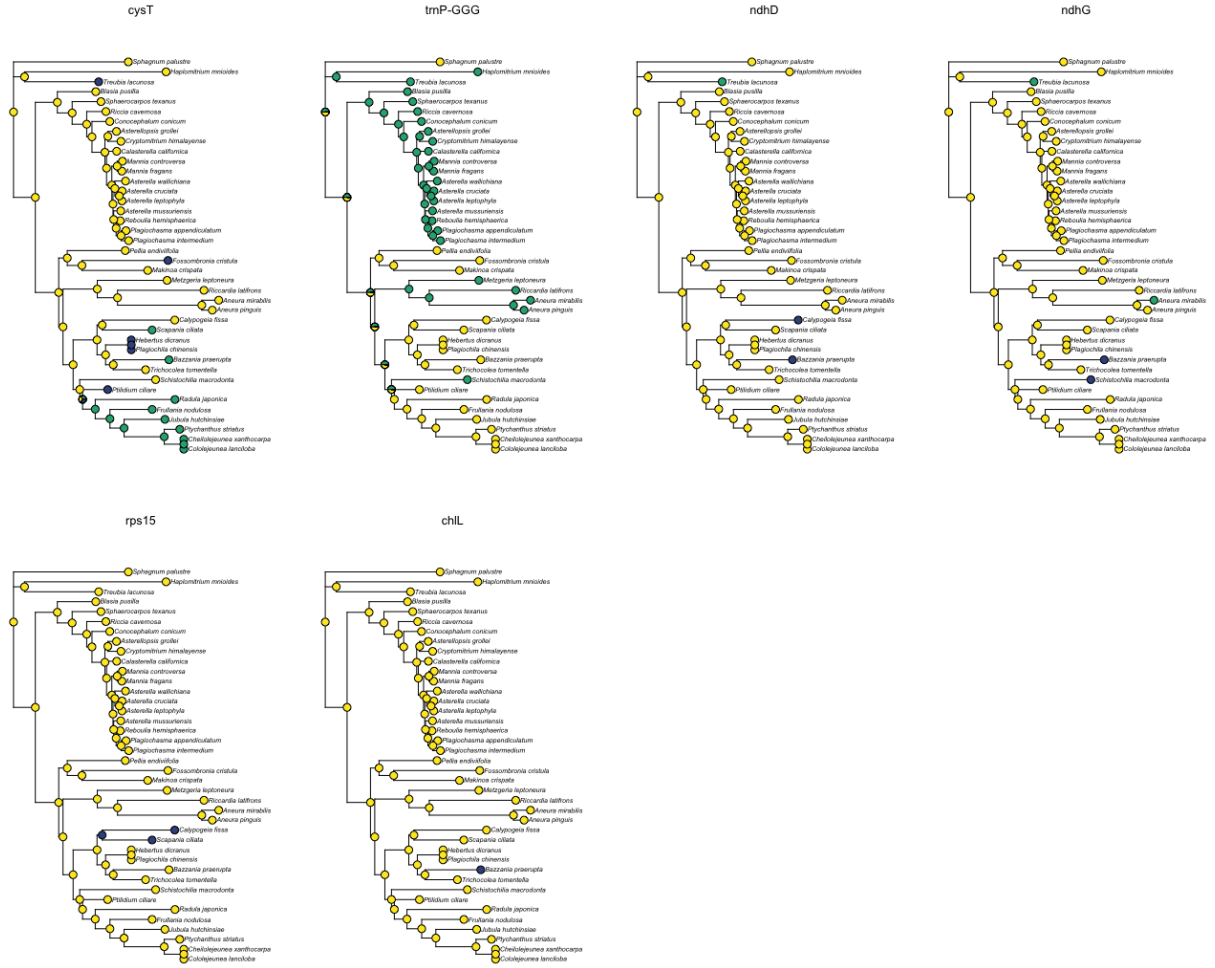


Figure A.7: **Ancestral state reconstruction for gene presence-absence (part 4).** The genes above each tree share the history represented by each tree. Yellow denotes presence of the gene, green denotes absence, and purple denotes alternative versions.

A.7 Ancestral state reconstruction for intron presence-absence in the chloroplast of liverworts

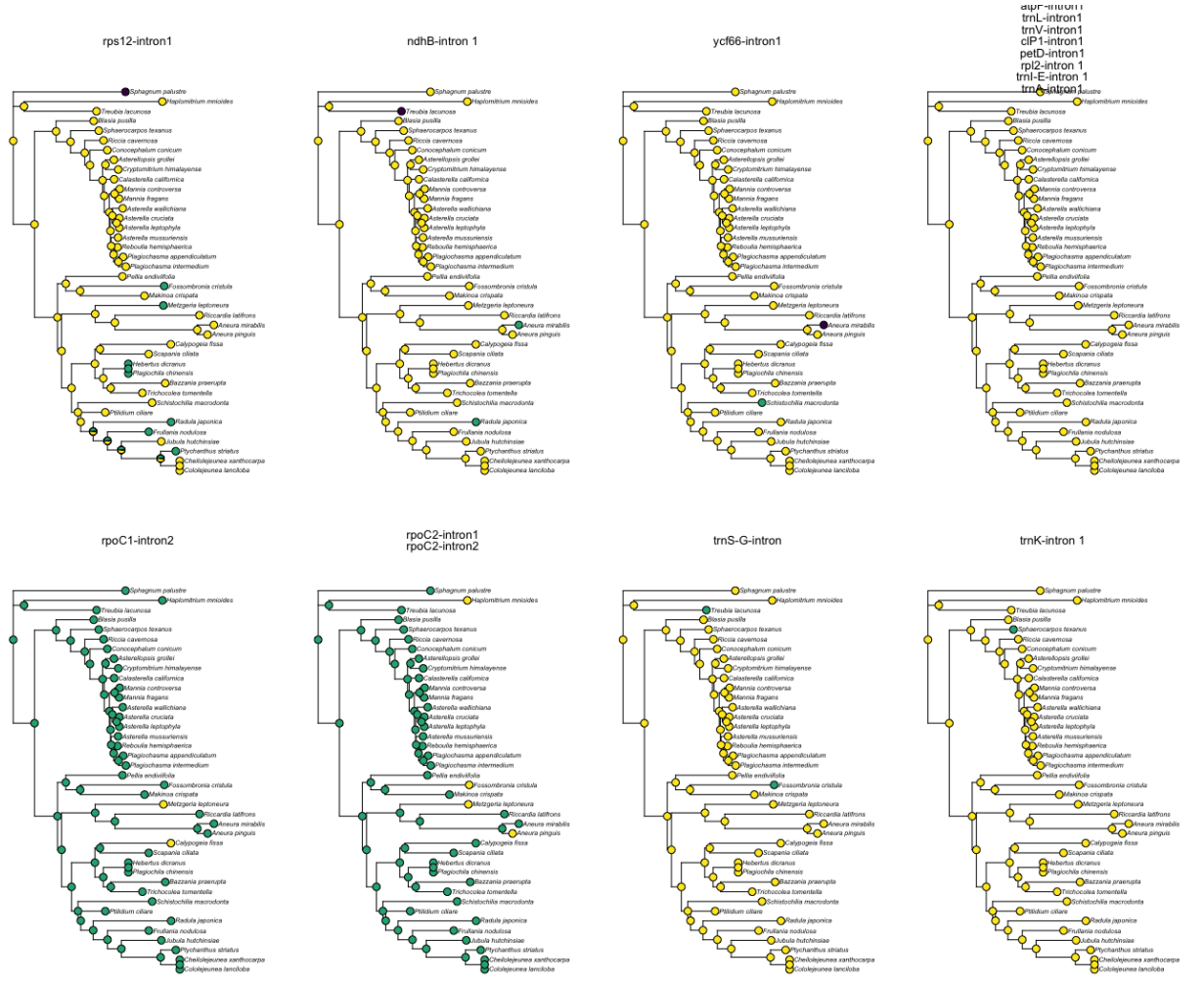


Figure A.8: **Ancestral state reconstruction or intron presence-absence (part 1).** The genes above each tree share the history represented by each tree. Yellow denotes presence of the gene, green denotes absence of the introns.

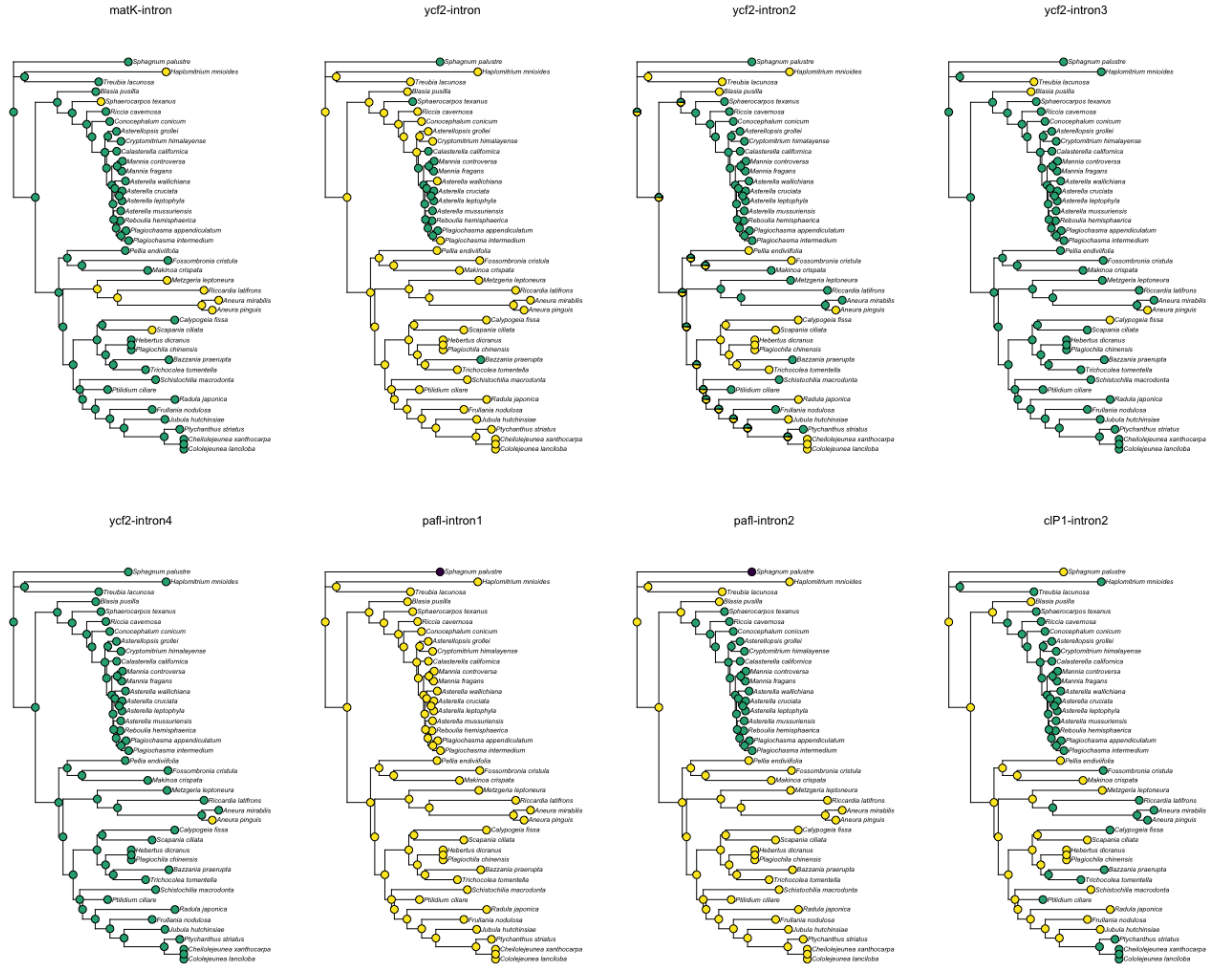


Figure A.9: **Ancestral state reconstruction or intron presence-absence (part 2).** The genes above each tree share the history represented by each tree. Yellow denotes presence of the gene, green denotes absence of the introns.

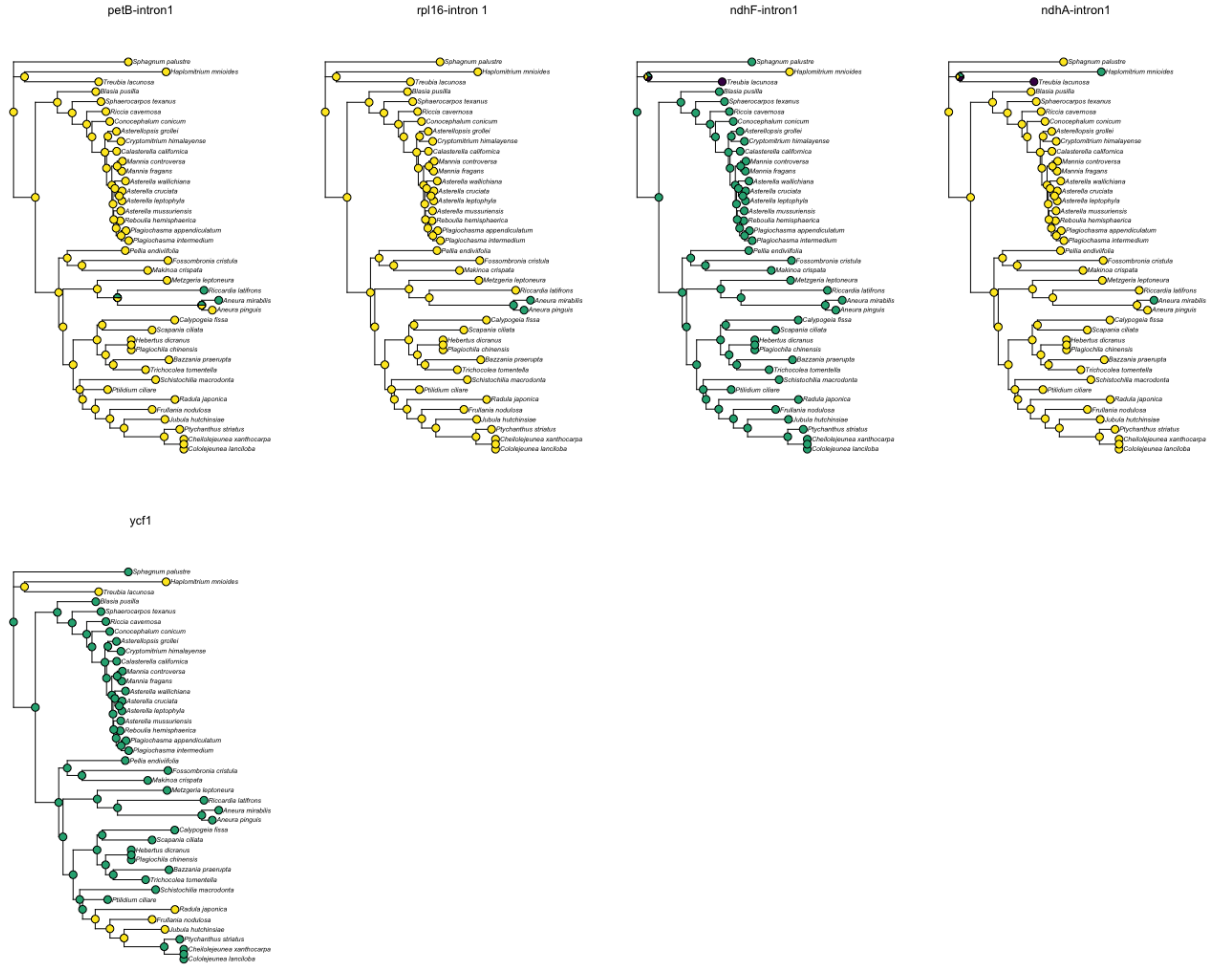


Figure A.10: **Ancestral state reconstruction or intron presence-absence (part 3).** The genes above each tree share the history represented by each tree. Yellow denotes presence of the gene, green denotes absence of the introns.

A.8 Phylogenetic inference of liverworts combining structural and sequence data.

The combined analysis was set up in RevBayes as described in methods. The tree is the maximum clade credibility tree, and the support values are posterior probabilities.

Figure A.11: Maximum Clade Credibility tree inferred from the combined analysis of two markers (*rbcL* and *rps4*) and the structural data of liverwort chloroplasts

Appendix B

Appendix for Chapter 3

B.1 Information about collections of *C. californica*

Table B.1: Collections of *C. californica* samples across California.

Sample name	**Locality**	**Latitude**	**Longitude**
BM001_ext_IG109	Southern California	39.14322	-122.91029
FF339_ext_IG045	Mt. Tamalpais	37.94286	-122.64694
FF340_ext_IG046	Mt. Tamalpais	37.94286	-122.64694
IGR150_ext_IG111	Los Padres National Forest	34.4722	-119.70737
IGR151_ext_IG087	Los Padres National Forest	34.4728833	-119.71172
IGR152_ext_IG088	Los Padres National Forest	34.4743667	-119.71995
IGR153_ext_IG089	Los Padres National Forest	34.4907167	-119.63587
IGR154_ext_IG090	Los Padres National Forest	34.4901667	-119.63598
IGR156_ext_IG091	Los Padres National Forest	34.4800833	-119.5865
IGR157_ext_IG101	Los Padres National Forest	34.4797833	-119.58597
IGR158_ext_IG092	Los Padres National Forest	34.5417667	-119.79088
IGR159_ext_IG093	Los Padres National Forest	34.5417667	-119.79088
IGR160_ext_IG094	Los Padres National Forest	35.10895	-120.0761
IGR161_ext_IG095	Los Padres National Forest	35.1088333	-120.0752
IGR162_ext_IG096	Los Padres National Forest	35.1088333	-120.0752
IGR163_ext_IG097	Los Padres National Forest	35.4238333	-120.73
IGR164_ext_IG102	Los Padres National Forest	35.4248	-120.74017
IGR166_ext_IG021	El Diablo Hills	37.8921	-121.9992
IGR167_ext_IG013	El Diablo Hills	37.8910333	-121.99918
IGR168_ext_IG098	El Diablo Hills	37.8865	-121.99453
IGR169_ext_IG022	El Diablo Hills	37.8862667	-121.9932
IGR170_ext_IG023	Briones Regional Park	37.9346167	-122.11523

Table B.1 continued from previous page

Sample name	Locality	Latitude	Longitude
IGR171_ext_IG024	Briones Regional Park	37.9547167	-122.1396
IGR172_ext_IG025	Briones Regional Park	37.9556683	-122.1399
IGR173_ext_IG103	Cleveland National Forest	33.60021	-117.50813
IGR174_ext_IG104	Cleveland National Forest	33.60092	-117.51035
IGR175_ext_IG105	Cleveland National Forest	33.60984	-117.42995
IGR176_ext_IG014	Cleveland National Forest	33.66291	-117.40418
IGR177_ext_IG015	Cleveland National Forest	33.64905	-117.45667
IGR178_ext_IG016	Cleveland National Forest	32.67966	-116.52902
IGR179_ext_IG017	Cleveland National Forest	32.67966	-116.52902
IGR180_ext_IG018	Cleveland National Forest	32.99369	-116.75946
IGR181_ext_IG026	Cleveland National Forest	32.99345	-116.75747
IGR182_ext_IG106	Cleveland National Forest	33.12284	-116.79305
IGR183_ext_IG027	Cleveland National Forest	33.45579	-116.97228
IGR184_ext_IG019	Cleveland National Forest	33.71511	-117.61733
IGR185_ext_IG020	Cleveland National Forest	33.78634	-117.6647
IGR186_ext_IG037	Santa Cruz Island	34.0009	-119.7506
IGR187_ext_IG028	Santa Cruz Island	34.0009	-119.7506
IGR188_ext_IG029	Santa Cruz Island	34.0009	-119.7506
IGR189_ext_IG030	Santa Cruz Island	33.99689	-119.72771
IGR190_ext_IG031	Santa Cruz Island	34.00419	-119.70552
IGR191_ext_IG110	Santa Cruz Island	34.00419	-119.70552
IGR192_ext_IG033	Santa Cruz Island	34.00419	-119.70552
IGR194_ext_IG034	Santa Cruz Island	34.00634	-119.74715
IGR195_ext_IG035	Santa Cruz Island	34.00634	-119.74715
IGR196_ext_IG036	Santa Cruz Island	34.00634	-119.74715
IGR198_ext_IG099	Santa Cruz Island	34.03072	-119.79944
IGR199_ext_IG038	Santa Cruz Island	34.03072	-119.79944
IGR200_ext_IG039	Santa Cruz Island	34.0086	-119.77123
IGR201_ext_IG040	Santa Cruz Island	34.03201	-119.70087
IGR202_ext_IG041	Santa Cruz Island	34.01979	-119.6899
IGR203_ext_IG042	Clearlake State Park	39.01742	-122.81048
IGR204_ext_IG043	Sierra National Forest	37.663056	-119.87389
IGR205_ext_IG044	Sierra National Forest	37.663889	-119.87056
IGR205_ext_IG047	Sierra National Forest	37.663889	-119.87056
IGR206_ext_IG048	Topanga State Park	34.09158	-118.62363
IGR207_ext_IG049	Topanga State Park	34.09207	-118.62168
IGR208_ext_IG050	Topanga State Park	34.08086	-118.51395
IGR209_ext_IG051	Crystal Cove State Park	33.58415	-117.79495
IGR210_ext_IG052	Crystal Cove State Park	33.58337	-117.79583

Table B.1 continued from previous page

Sample name	Locality	Latitude	Longitude
IGR211_ext_IG053	Anza-Borrego State Park	33.04584	-116.43847
IGR212_ext_IG054	Anza-Borrego State Park	33.04584	-116.43847
IGR213_ext_IG055	Anza-Borrego State Park	33.24238	-116.42376
IGR214_ext_IG056	Anza-Borrego State Park	33.24238	-116.42376
IGR215_ext_IG057	Anza-Borrego State Park	33.23648	-116.44297
IGR216_ext_IG058	Anza-Borrego State Park	33.23648	-116.44297
IGR217_ext_IG059	Anza-Borrego State Park	33.27844	-116.43058
IGR218_ext_IG060	San Bernardino National Forest	34.18655	-117.52338
IGR220_ext_IG061	Angeles National Forest	34.18007	-118.0467
IGR221_ext_IG062	Angeles National Forest	34.21557	-118.1446
IGR222_ext_IG063	Angeles National Forest	34.21804	-118.14352
IGR223_ext_IG064	Angeles National Forest	34.26104	-118.15262
IGR224_ext_IG065	Angeles National Forest	34.25779	-118.15536
IGR225_ext_IG112	Angeles National Forest	34.37624	-118.46304
IGR226_ext_IG066	Angeles National Forest	34.37624	-118.46304
IGR227_ext_IG067	Angeles National Forest	34.37624	-118.46304
IGR228_ext_IG068	Angeles National Forest	34.31837	-118.33321
IGR229_ext_IG069	Angeles National Forest	34.25422	-118.25288
IGR230_ext_IG070	Angeles National Forest	34.6025	-118.55639
IGR231_ext_IG071	Los Padres National Forest	34.52773	-119.18104
IGR233_ext_IG072	Fawn Lily Ranch	39.36841	-123.36649
IGR234_ext_IG073	Fawn Lily Ranch	39.36833	-123.36819
IGR235_ext_IG074	Fawn Lily Ranch	39.36833	-123.36819
IGR237_ext_IG075	Fawn Lily Ranch	39.36343	-123.41087
IGR240_ext_IG076	Sequoia National Forest	35.47513	-118.72465
IGR241_ext_IG077	Sequoia National Forest	35.47513	-118.72465
IGR242_ext_IG012	Sequoia National Forest	35.55136	-118.61015
IGR243_ext_IG078	Sequoia National Forest	35.79916	-118.45097
IGR244_ext_IG079	Sequoia National Forest	35.96612	-118.47743
IGR245_ext_IG080	Sequoia National Forest	35.96612	-118.47743
IGR246_ext_IG081	Six Rivers National Forest	40.88955	-123.60063
IGR247_ext_IG082	Six Rivers National Forest	40.85484	-123.48461
IGR248_ext_IG083	Shasta-Trinity National Forest	40.76234	-123.28562
IGR249_ext_IG011	Shasta-Trinity National Forest	40.77171	-123.12917
IGR250_ext_IG084	Plumas National Forest	39.8468	-121.39388
IGR251_ext_IG085	Plumas National Forest	39.84623	-121.39406
IGR258_ext_IG107	Mount Diablo State Park	37.915278	-121.9475
IGR259_ext_IG108	Mount Diablo State Park	37.90707	-121.94952
IGR284_ext_IG113	Henry Coe State Park	37.210833	-121.37028

Table B.1 continued from previous page

Sample name	Locality	Latitude	Longitude
IGR285_ext_IG114	Henry Coe State Park	37.201944	-121.37694
IGR286_ext_IG115	Henry Coe State Park	37.201111	-121.37861
McL1220_ext_IG116	Henry Coe State Park	37.248889	-121.35778
McL777_ext_IG117	Henry Coe State Park	37.210278	-121.53083
McL1028_ext_IG118	Henry Coe State Park	37.09702	-121.3834
IGR292_ext_IG119	San Bruno Mountain State Park	37.69715	-122.44398
IGR293_ext_IGIG120	San Bruno Mountain State Park	37.69724	-122.44372
IK001_ext_IG121	Hastings Natural History reservation	36.3782877	-121.56813
BM002_ext_IG122	Sierra National Forest	41.79574	-121.30162
KMW496_ext_IG123	Santa Cruz Island	33.999495	-119.7109

B.2 Collection permits information

Table B.2: Collection permits to sample *C. californica* in National Forests and State Parks.

Collection permits]tab:permits

Park	Permit
San Bernardino National Forest	2021 Ramirez-UCB
Cleveland National Forest	3914
East Bay Regional Park District	*21-1083*
El Dorado National Forest	31909
Los Padres National Forest	Bot-4-2021
Plumas National Forest	48174
Shasta-Trinity National Forest	56072
Angeles, Sequoia, Sierra, Cleveland, Los Padres, San Bernardino, Six Rivers, Shasta-Trinity NFs	R5.2021.05 (Gonzalez Ramirez)
Anza Borrego State Park, Mount Diablo State Park, and Mount Tamalpais State Park	20-820-01

B.3 Methods of DNA extraction

The DNA extraction was conducted using two different methods depending on the amount of material. First, the DNeasy Plant Pro Kit from Quiagen was used in samples with enough material. This procedure yields very clean DNA. I followed the protocol described by Quiagen at `https://www.qiagen.com/us/resources`.

The second protocol maximized DNA amount and was used for most of the samples. This protocol is adapted from a protocol facilitated by the Rothfels lab.

1. Add grinding ball to each 2 mL tube with tissue.

2. Grind dry at 1700 rpm for 2 minutes.

3. Add 500 uL C-TAB-BME to tube and vortex.
4. Centrifuge for 5 min at 13,000 rpm.
5. After centrifugation resuspend the clump tissue pelleted during the centrifuge spin using a vortex. Incubate in 65C water-bath for 1hr.
6. centrifuge for 1 min at 13,000 rpm.
7. Add 900 uL of chloroform and vortex.
8. Centrifuge for 10 min at 13,000 rpm. While the samples are in the centrifuge prepare/label new tubes.
9. Transfer 800uL of the aqueous supernatant (top layer) to a new labeled 2mL tube. Be absolutely sure not to get any of the middle or bottom layer.
10. To the aqueous supernatant add an equal volume of ice-cold isopropanol. Invert tubes once and incubate at -20C for 1h to over night.
11. Centrifuge for 20 min at 14,000 rpm
12. Pour off the isopropanol from each tube, being careful the pellet does not become dislodged, and add 1 mL of ice cold 70% ethanol.
13. 13) Centrifuge for 5 min at 14,000 rpm.
14. Pour off the ethanol, again being careful not to dislodge the pellet at the bottom of the tube, and leave at room temperature overnight to dry.
15. Once the pellet is dry, add 30 uL of TE buffer.
16. Transfer to a storage plate, label, and store at -20C.

B.4 Admixture analysis

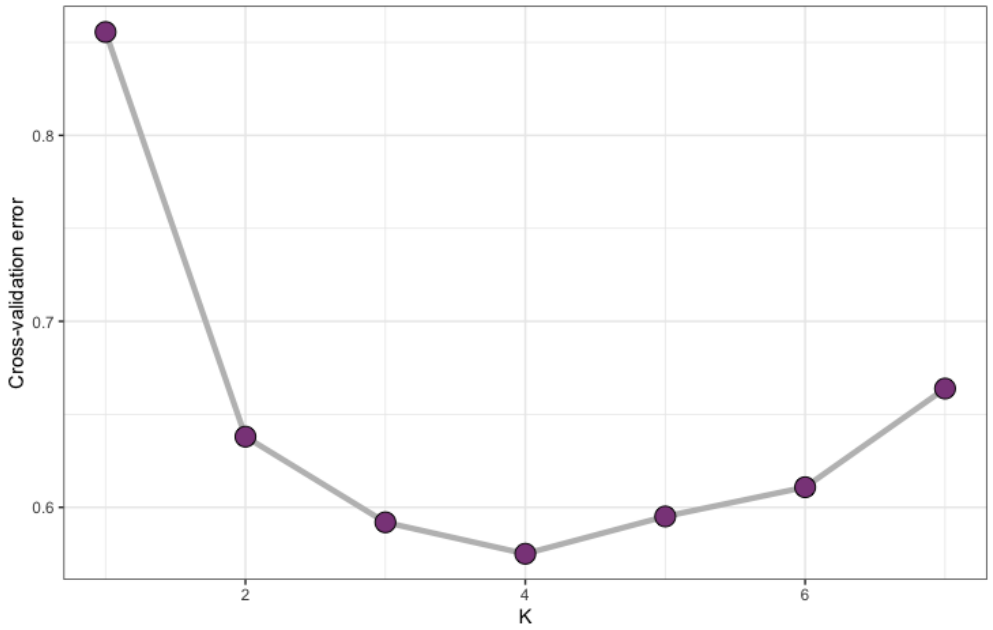


Figure B.1: **Cross-validation error values for different values of K.** The cross validation error was calculated for K ranging from 1 to 7 using the software `ADMIXTURE`

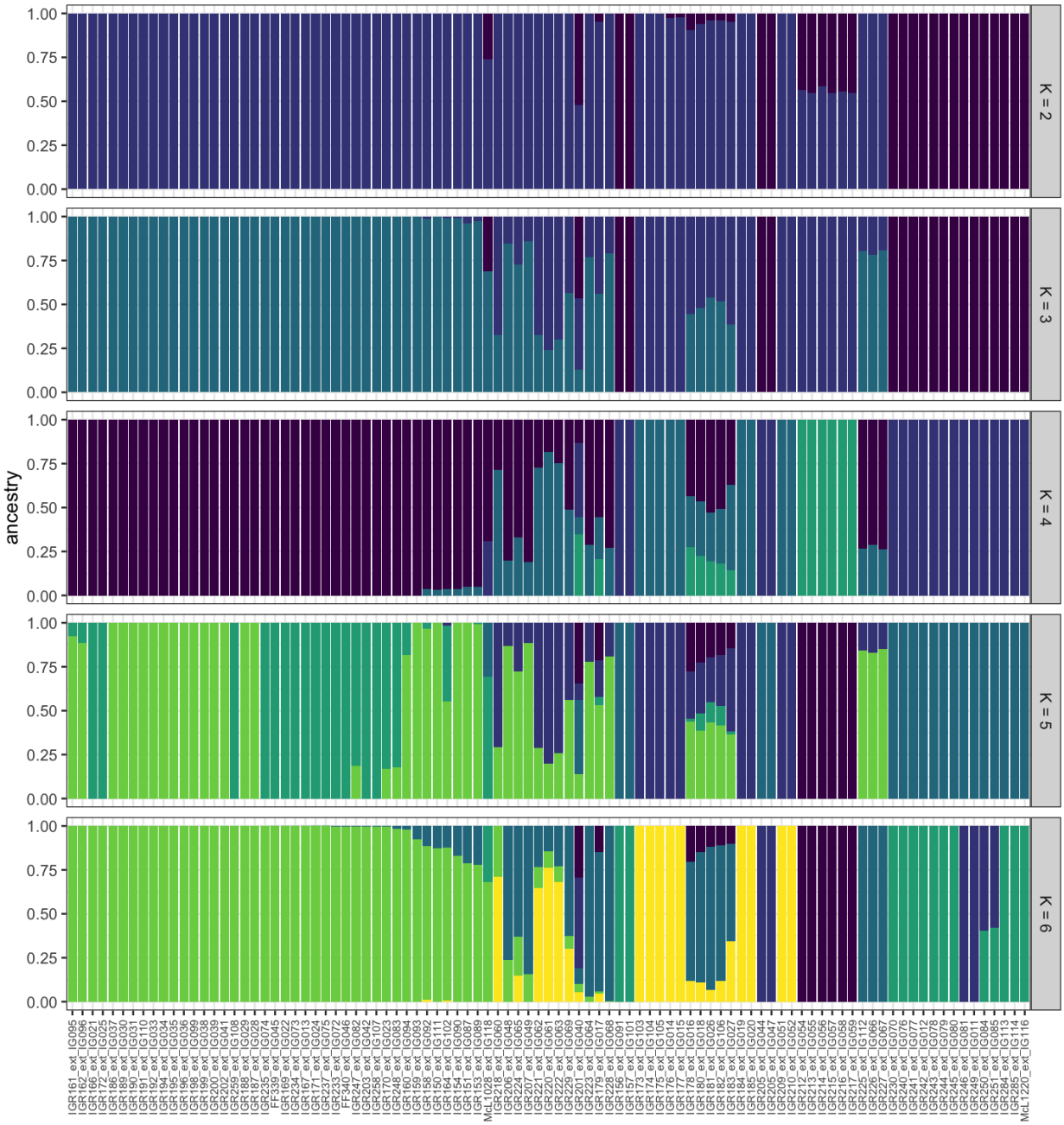


Figure B.2: **Ancestry contributions for individuals of *C. californica* asuuming different K-values.** Inferred ancestral genetic contributions for individuals of *C. californica* assuming different number of ancestral populations.

Table B.3: Fst values among *C. californica* populations as inferred by the software `ADMIXTURE`.

	Pop0	**Pop1**	**Pop2**
Pop0			
Pop1	0.640		
Pop2	0.430	0.649	
Pop3	0.587	0.694	0.571

B.5 Principal components analysis

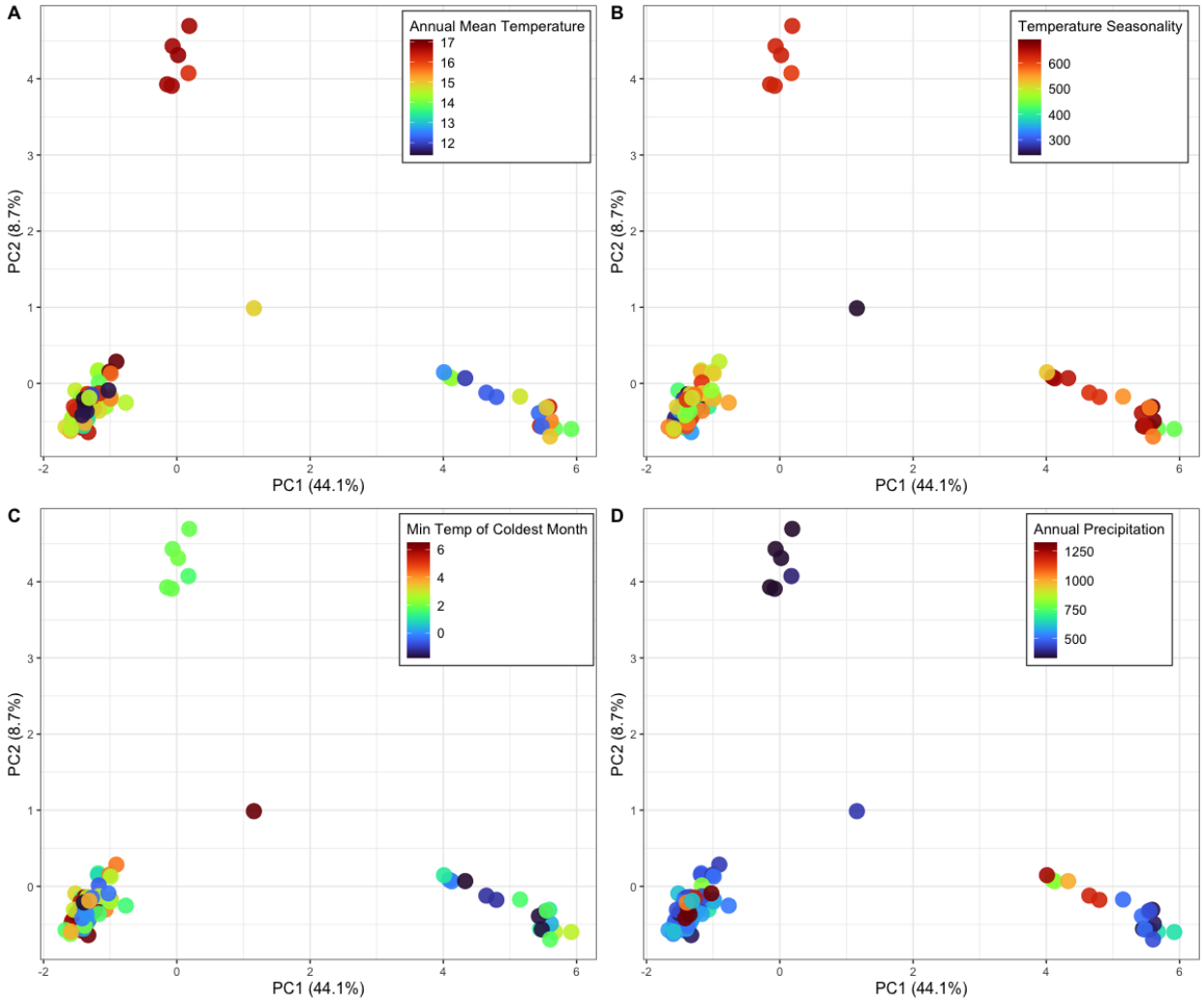


Figure B.3: **PCA of *C. californica* individuals based on SNP data colored by climatic variables.** PC1 and PC2 of the SNP data analysis color coded by different climatic variables obtained from WordClim database

Appendix C

Appendix for Chapter 4

C.1 Accession numbers sequences used

Table C.1: Genbank accession numbers for the Aytoniaceae species used in this work.

Species	**matK**	**rbcL**	**trnL-trnF**
Mannia controversa			GQ910689.1
Mannia triandra		FJ173714.1	GQ910697.1
Mannia sibirica			GQ910695.1
Mannia pilosa		KT793572.1	GQ910694.1
Mannia androgyna		FJ173699.1	GQ910679.1
Mannia fragrans		DQ286013.1	GQ910690.1
Mannia capensis	AF264673.1		
Mannia californica	AF264672.1	KJ590913.1	GQ910683.1
Mannia gracilis	AF264674.1	KT793551.1	GQ910668.1
Cryptomitrium himalayense		DQ286007.1	GQ910678.1
Cryptomitrium tenerum	AF264668.1	KT356972.1	GQ910677.1
Asterella echinella		KT793550.1	
Asterella lateralis			GQ910671.1
Asterella cruciata		FJ173679.1	KR024222.1
Asterella saccata		FJ173689.1	GQ910672.1
Asterella mussuriensis		FJ173688.1	
Asterella dominicensis		KC305694.1	
Asterella africana		DQ285999.1	GQ910666.1
Asterella multiflora		AM920285.1	
Asterella macropoda		FJ173687.1	
Asterella lindenbergiana		FJ173684.1	KR024216.1
Asterella leptophylla		KT793552.1	KR024221.1

Table C.1 continued from previous page

Species	**matK**	**rbcL**	**trnL-trnF**
Asterella innovans		AM920307.1	
Asterella wallichiana	AF264675.1	DQ286001.1	GQ910674.1
Asterella syngenesica	AF264679.1		
Asterella palmeri	AF264669.1	KT793554.1	
Asterella khasyana	AF264677.1	FJ173682.1	
Asterella grollei	AF264670.1	DQ286000.1	GQ910670.1
Asterella australis	AF264680.1	AM920300.1	
Asterella californica		KT793549.1	GQ910667.1
Asterella bolanderi	AF264678.1	KT793548.1	
Asterella tenella	AF264676.1	U87064.1	GQ910673.1
Plagiochasma intermedium		AM920293.1	
Plagiochasma pterospermum		AM920282.1	KR024210.1
Plagiochasma landii		FJ173717.1	
Plagiochasma japonicum		FJ173716.1	GQ910699.1
Plagiochasma crenulatum		AM920273.1	
Plagiochasma wrightii		DQ286021.1	GQ910701.1
Plagiochasma appendiculatum	AF264682.1	KT356976.1	
Plagiochasma rupestre	AF264681.1	KJ590915.1	GQ910700.1
Reboulia hemisphaerica	AF264671.1	KT793580.1	KR024214.1

C.2 Area definition for the biogeographic model

For this study, I defined 11 areas that remain units through time (Fig.C.1). It is worth noting that the number of areas that are tractable under this model is small, since the number of states increases exponentially. For a number of areas n, a lineage can occupy 2^n area combinations. In our case, the model has 11 areas, so it accounts for $2^{11} = 2048$ possible states. Therefore, there is currently an inherent compromise between the level of detail and feasibility in this type of model.

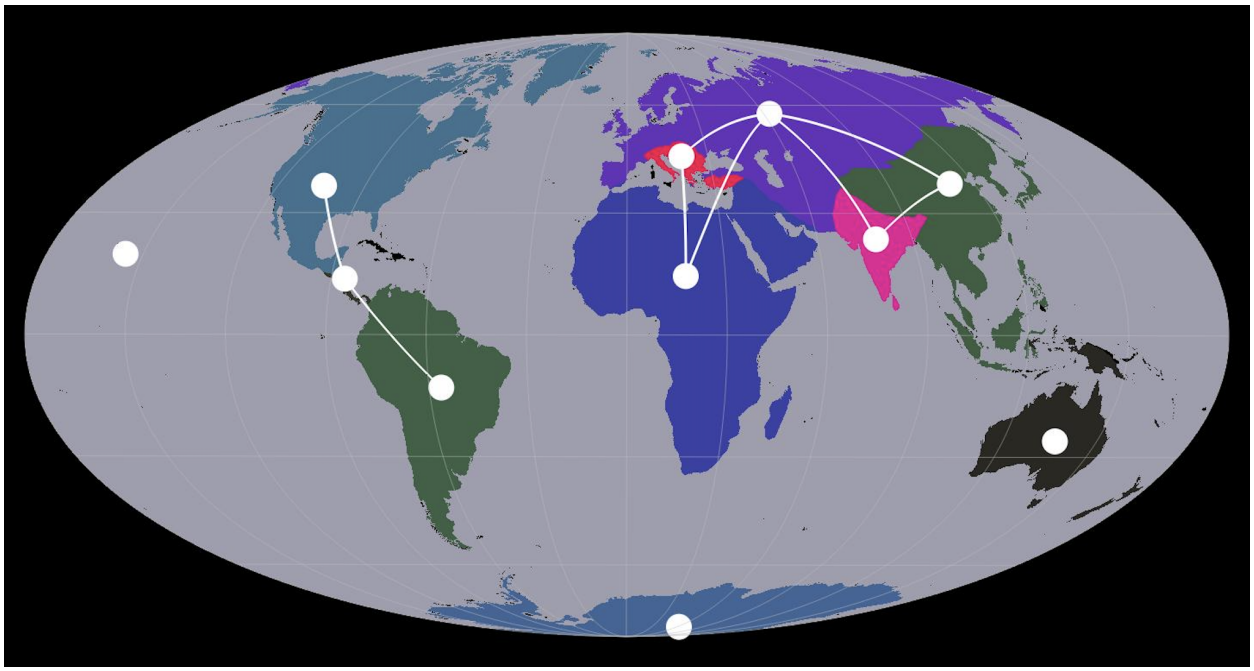


Figure C.1: **Geographic areas used in this study.** Different colors represent the 11 different areas that were used as units in this study

.

The area definitions took into account the movement of areas through time—what pieces of emerged land remain relatively connected throughout the ∼130 million years of crown Aytoniaceae history—and the relevance of these areas in the context of Aytoniaceae biology and current distributions. On the American continent, there are three areas, (1) North America, NAm, in blue, (2) South America, SAm, in green, and (3) Central America, CAm, in black (Fig. C.1). Central America in this study includes the Antilles. I acknowledge that several of these islands have different origin times (in particular the Greater Antilles vs the Lesser Antilles), but for the temporal and spatial resolution of this project, and the biology of the family Aytoniaceae, Central America is the best grouping for the Antilles. On the other side of the Atlantic, in navy blue, (4) Africa + the Arabian peninsula, Af, constitute an area. Even if nowadays these two areas are almost completely separated by the Red sea, this isolation occurred only ∼ 5 Mya. Therefore, the Arabian peninsula is most related to the African continent. The fifth area is (5) the Mediterranean, Med, showed in pink. It includes the Italian and Greek peninsulas as well as a portion of Turkey. These areas were islands separated from Europe and Africa for almost 30 My during the late Cretaceous-early Cenozoic. Furthermore, species like *Asterella africana* and *Asterella macropoda* occur only in this region within Europe. The northern portion of (6) Europe and Asia, EuAs, colored in purple is the sixth area. India is the (7) seventh area and includes all the continental territory

of the India Plate, until the subction zone at along the Himalayas. The (8) eigth area is called Southeastern Asia and its colored in green, it splits from India at the Himalayas, and extends to the northern border of the Tibetian Plateau.

C.3 Occurrence of extant species of Aytoniaceae in the 11 areas

Table C.2: Matrix of presence-absence of the 40 species of Aytoniaceae in the 11 regions defined for the phylogenetic analysis.

	NAm	CAm	SAm	Ant	EuAs	Med	Af	Ind	SAs	Oce	Hw
Asterella_africana	0	0	0	0	0	1	1	0	0	0	0
Asterella_australis	0	0	0	0	0	0	0	0	0	1	0
Asterella_bolanderi	1	0	0	0	0	0	0	0	0	0	0
Asterella_californica	1	0	0	0	0	0	0	0	0	0	0
Asterella_cruciata	0	0	0	0	0	0	0	0	1	0	0
Asterella_dominicensis	1	1	0	0	0	0	0	0	0	0	0
Asterella_echinella	1	1	0	0	0	0	0	0	0	0	0
Asterella_grollei	0	0	0	0	0	0	0	1	1	0	0
Asterella_innovans	0	0	0	0	0	0	0	0	0	0	1
Asterella_khasyana	0	0	0	0	0	0	0	1	1	1	0
Asterella_lateralis	1	1	1	0	0	0	0	0	0	0	0
Asterella_leptophylla	0	0	0	0	0	0	0	1	1	0	0
Asterella_lindenbergiana	1	0	0	0	1	0	0	0	0	0	0
Asterella_macropoda	0	1	1	0	0	0	0	0	0	0	0
Asterella_multiflora	0	0	0	0	0	0	0	1	0	0	0
Asterella_mussuriensis	0	0	0	0	0	0	0	1	1	0	0
Asterella_palmeri	1	0	0	0	0	0	0	0	0	0	0
Asterella_saccata	1	0	0	0	1	1	0	0	0	0	0
Asterella_tenella	1	0	0	0	0	0	0	0	0	0	0
Asterella_wallichiana	0	0	0	0	0	0	0	1	1	0	0
Cryptomitrium_himalayense	0	0	0	0	0	0	0	1	0	0	0
Cryptomitrium_tenerum	1	0	1	0	0	0	0	0	0	0	0
Mannia_androgyna	0	0	0	0	1	1	1	0	1	0	0
Mannia_californica	1	0	0	0	1	0	0	0	0	0	0
Mannia_capensis	0	0	0	0	0	0	1	0	0	0	0
Mannia_controversa	0	0	0	0	1	0	0	0	1	0	0
Mannia_fragrans	1	0	0	0	1	1	0	0	0	0	0
Mannia_gracilis	1	0	0	0	1	1	0	0	1	0	0
Mannia_pilosa	1	0	0	0	1	1	0	0	0	0	0
Mannia_sibirica	1	0	0	0	1	1	0	0	0	0	0
Mannia_triandra	1	0	0	0	1	1	0	0	1	0	0
Plagiochasma_appendiculatum	0	0	0	0	0	0	1	1	1	0	0
Plagiochasma_crenulatum	1	0	0	0	0	0	0	0	0	0	0
Plagiochasma_intermedium	1	0	0	0	0	0	0	0	1	0	0
Plagiochasma_japonicum	0	0	0	0	1	0	0	0	1	0	0
Plagiochasma_landii	1	0	0	0	0	0	0	0	0	0	0
Plagiochasma_pterospermum	0	0	0	0	1	0	0	0	1	0	0
Plagiochasma_rupestre	1	1	1	0	0	1	1	0	0	1	0
Plagiochasma_wrightii	1	0	0	0	0	0	0	0	0	0	0
Reboulia_hemisphaerica	1	0	1	0	1	1	0	0	1	1	0

C.4 Maximum likelihood inferences of gene trees and concatenated dataset for Aytoniaceae

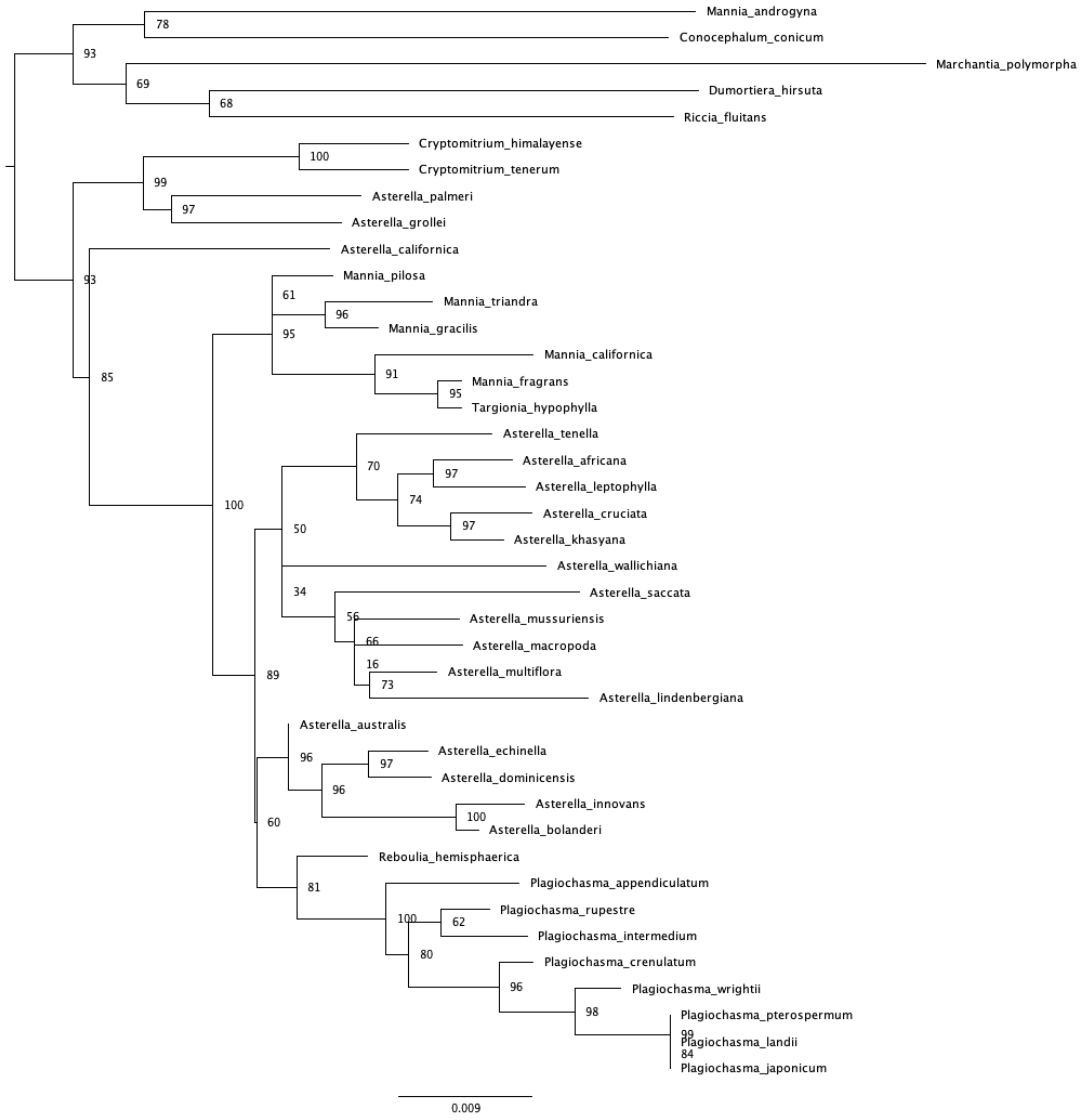


Figure C.2: ***rbcL* gene tree for Aytoniaceae.** Gene tree inferred with four independent runs of IQtree (Minh et al., 2020) under a GTR + I + Γ substitution model. Support values were calculated with 1000 rounds of ultra-fast bootstrapping.

Figure C.3: ***matK* gene tree for Aytoniaceae.** Gene tree inferred with four independent runs of IQtree (Minh et al., 2020) under a GTR + I + Γ substitution model. Support values were calculated with 1000 rounds of ultra-fast bootstrapping.

Figure C.4: ***trnL* gene tree for Aytoniaceae.** Gene tree inferred with four independent runs of IQtree (Minh et al., 2020) under a GTR + I + Γ substitution model. Support values were calculated with 1000 rounds of ultra-fast boostrapping.

Figure C.5: **Concatenated tree for Aytoniaceae.** Concatenated tree ($rbcL + matK + trnL$) inferred with four independent runs of IQtree (Minh et al., 2020) under a GTR + I + Γ substitution model. Support values were calculated with 1000 rounds of ultra-fast bootstrapping.

C.5 Marginal likelihoods for clock model selection

Because the clock model can have a strong effect on the age estimates, I tried three different clock models: strict clock, UCED, and UCLD. I selected among these three different clock models using the marginal likelihood estimates obtained from two different algorithms: path

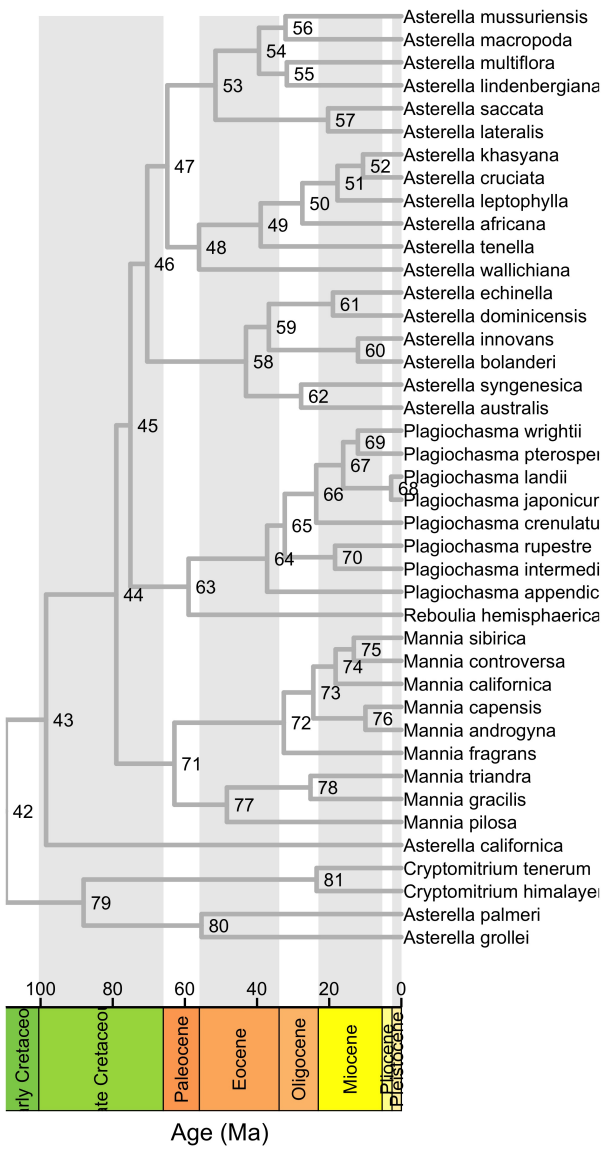
sampling and stepping stone. The results are showed below. Both algorithms preferred the UCLD over the strict clock model (2lnBF = 6.43) and the UCE clock model (2lnBF = 36.73).

Table C.3: Marginal likelihood estimates of different clock models for inferring the timetree of Aytoniaceae. Each analysis used 40 stones.

Path sampling algorithm					
model	run 1	run 2	run 3	run 4	Average
strict	-8693.167	-8691.938	-8693.69	-8691.789	-8692.646
UCED	-8707.136	-8707.424	-8707.631	-8709.001	-8707.798
UCLD	-8689.723	-8688.675	-8689.056	-8690.264	-8689.4295
Stepping stone algorithm					
model	run 1	run 2	run 3	run 4	Average
strict	-8691.994	-8690.559	-8692.781	-8690.721	-8691.5138
UCED	-8706.048	-8706.23	-8706.084	-8707.28	-8706.4105
UCLD	-8689.014	-8687.291	-8687.095	-8689.349	-8688.1873

C.6 Time-divergence age estimates for Aytoniaceae under the preferred clock model

In the following table I report the time divergence estimates for Aytoniaceae according to my results. On the table I reference each of the nodes labeled in the phylogeny (on the left) and report the age value for the MAP tree as well as the minimum and maximum values of the 95 % highest posterior density



node	age	minHPD	maxHPD
42	109.652	77.024500	142.52100
43	98.480	66.875800	130.79000
44	79.012	52.547800	105.90600
45	75.102	49.459500	100.69000
46	70.441	45.540600	95.03600
47	64.876	42.369600	88.60460
48	56.085	35.665700	77.09650
49	38.994	23.389100	54.51560
50	27.443	15.827600	40.52140
51	17.708	9.308370	28.21860
52	10.631	3.663470	19.17360
53	51.524	30.834300	74.05200
54	39.461	20.464500	59.93460
55	31.753	11.968100	51.76080
56	32.109	12.773600	52.43630
57	20.320	0.011007	46.76720
58	43.071	25.139600	63.05600
59	36.732	20.053500	54.07960
60	12.028	2.113820	24.04940
61	18.968	6.414540	33.17390
62	27.787	12.103400	44.76990
63	58.985	37.702600	81.84740
64	37.211	22.509100	53.25910
65	32.323	19.291300	46.94630
66	23.600	12.613200	35.87190
67	16.123	7.875320	25.36650
68	2.834	0.000044	8.08459
69	12.052	4.870090	19.94000
70	18.384	4.868090	32.70960
71	62.908	39.594000	87.65270
72	32.586	17.731300	48.81480
73	24.399	12.434000	36.77660
74	18.223	8.670120	28.85680
75	13.177	3.609990	25.24060
76	9.956	0.000240	23.01750
77	48.364	27.850700	70.26410
78	25.180	11.467700	39.90420
79	88.084	56.726200	120.02500
80	55.399	30.376000	82.54920
81	23.499	10.515300	38.68690

Figure C.6: **Divergence-time estimates for Aytoniaceae under the UCLD clock model** On the left, time calibrated phylogeny of Aytoniaceae, the numbers depict the node identifier numbers. The table on the right shows the MAP age estimate for each node, as well as the minimum and maximum values for the 95% HPD age range.