UNIVERSITY OF CALIFORNIA
RIVERSIDE

Genomic Architecture, Habitat Diversification, and Genetic Differentiation of Manzanitas

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Plant Biology

by

Yi Huang

June 2022

Dissertation Committee:
Dr. Amy Litt, Chairperson
Dr. Janet Franklin
Dr. Jason Stajich

The Dissertation of Yi Huang is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

Through my Ph.D. journey, there are so many people to whom I should dedicate my acknowledgment. Listing all of them here seems impossible, but I will try.

First of all, I would like to express my deepest appreciation to my advisor, Amy Litt. She has provided me with invaluable scientific training and also life support. As a fresh Ph.D. student, I was afraid of talking to any faculty member or graduate student. With her mentoring and help, I not only overcame my shyness, but also participated in and enjoyed many collaborations with other groups of diverse backgrounds. Without her efforts, I would never be able to stand where I am and become who I am now. Her patience, kindness, and encouragement mean a lot to my Ph.D. training and have made me a better scientist with critical thinking and research independence.

I would also like to extend my deepest gratitude to my other committee members. I am incredibly grateful to Prof. Norman Ellstrand, Prof. Janet Franklin, and Prof. Jason Stajich for their invaluable guidance and advice on my research. They not only helped with my projects but also served as career models in the past years. I greatly admire their profound knowledge and scientific insight, and it’s my great fortune to have them as my dissertation committee members. In addition, I also want to thank Prof. Edith Allen, Prof. Zhenyu Jia, and Prof. John Heraty, who have provided me with so much help in my early years at UCR. Without them, many things would not be as good as they are now to me!

Next, I would like to thank my colleagues and friends in the department. I want to thank Dr. Alex Rajweski for his help and support in the past years. He is the dearest

brother that I never had, and I am really gratefμL for having him in my Ph.D. life. I also want to thank Glen Morrison for being the best collaborator and trip partner I can ever ask for. I feel so lucky to know a labmate with super talent in R programming and data visualization. In addition, I would like to thank many other helpful collaborators and lab members, including Brooke Rose, Santiago Velazco, Tom Parker, Jon Keeley, Andrew Sanders, Angela Buehlman, Tito Abbo, Natalie Saavedra, Dinusha Maheepala, and Elizabeth MacCarthy, etc. I also want to thank Laura McGeehan and Fidel Rivas for their help and support to our graduate students.

I want to thank my friends outside of the department. I want to thank Ziqi, Ziting, and Qianyi for being such great friends to my family and me. I will never forget their kindness and help at my lowest point. I want to thank Dr. Zhonghan Li and Prof. Min Xue for so many fun conservations in the past years, which provide essential mental support in my Ph.D. journey.

At last, I would like to thank my dearest family: my parents, husband, and daughter for all their love and support.

ABSTRACT OF THE DISSERTATION

Genomic Architecture, Habitat Diversification, and Genetic Differentiation of Manzanitas

by

Yi Huang

Doctor of Philosophy, Graduate Program in Plant Biology
University of California, Riverside, June 2022
Dr. Amy Litt, Chairperson

Manzanitas (*Arctostaphylos*, Ericaceae) are shrub species found in the California Floristic Province (CFP), a biodiversity hotspot of western North America. These plants are adapted to the summer dry period and fire disturbance of the CFP, and form the most diverse woody genus in the CFP flora. Among over 100 currently recognized manzanita species and subspecies, many are considered rare and endangered. The current understanding of manzanita adaptation and diversification is poor, limiting our ability to carry out ecological, evolutionary, and conservation studies on these plants.

A comprehensive understanding of genomic composition can advance knowledge of the genetic basis underlying the fire- and drought adaptation of manzanitas. We annotated the first manzanita genome assembly to provide genomic resources for downstream studies of adaptation and diversification. Our analyses indicate that our manzanita genome is well-assembled and annotated. It is enriched with

terpenoid genes, which may play essential roles in the fire and drought adaptations of manzanitas.

Understanding the ecological diversification of manzanitas can benefit the identification of species with distinct habitats and help researchers to derive effective conservation strategies. We used machine learning algorithms to conduct a quantitative study of niche differentiation among manzanita species. Although we did not identify any species with habitat distinctiveness within the genus, we determined that soil and climatic data can distinguish some species from other species in the same geographic region.

Next-generation sequencing data can provide invaluable insight into species and subspecies boundaries, and facilitate the identification of taxa with unique genotypes that require conservation attention. We applied reduced-representation genomic sequencing technology to test the hypothesis that Eastwood Manzanita (*Arctostaphylos glandulosa*) subspecies are genetically differentiated. We observe that genetic structure within Eastwood manzanita does not correspond to current subspecies circumscriptions, but rather reflects geographic distribution. In addition, only one of two subspecies of conservation concern appeared to be genetically distinct.

Our findings that resulted from these genomic, genetic, and ecological studies advance our knowledge of manzanita adaptation and diversification and form the basis for better conservation strategies for these important species.

Table of Contents

List of Figures

List of Tables

1 Chapter 1 Introduction

1.1 Importance of plant conservation

Plants are the primary producers in ecosystems and provide food, oxygen, and habitat for other organisms. Plant diversity is closely related to the sustainability of ecosystems and the survival of all life, including humans. Currently, this diversity is under pressure, as approximately one-third of all land plants are threatened with extinction, including many that are not well-understood by science (Corlett 2016). Efforts aimed at conserving this diversity are therefore critical to all life on earth.

Genomic, genetic, and ecological information can facilitate plant conservation. Reference genomes can serve as valuable resources for developing conservation strategies for plants: they can significantly improve the analysis of genetic variation, facilitating studies that are essential for species conservation, including population genetics, landscape genomics, and phylogenetics (Brandies et al. 2019). Understanding the genetic and environmental distinctiveness of rare and endangered plants helps people make better decisions about their conservation. Preservation of both unique genotypes and habitats is critical for plant conservation, therefore, effective conservation management plans require both genomic and ecological studies.

1.2 In the California Floristic Province, manzanitas are of conservation importance

Manzanitas (*Arctostaphylos*, Ericaceae) are shrub and tree species with red, twisting branches, evergreen leaves, and clusters of urn-shaped flowers (Kauffmann et al. 2015). They are conspicuous and dominant woody plants in the chaparral habitat of

the California Floristic Province (CFP), a biodiversity hotspot characterized by the Mediterranean-type climate with dry, hot summers and cool, wet winters (Burge et al. 2016). These plants comprise the most diverse woody genus in the CFP (Minnich and Howard 1984; Baldwin, Goldman, Keil, Patterson, Rosatti, et al. 2012), and their diversity has long fascinated (and perplexed) taxonomists. Like other CFP native plants, manzanitas are adapted to the harsh CFP environment with recurring fire disturbance and summer drought (Keeley 1991; Vasey, Loik, and Parker 2012) and serve many essential roles in their native ecosystems, including altering soil for the establishment of Douglas fir, a dominant native conifer, providing food for fruit-eating animals and pollinators, and releasing chemicals to inhibit the growth of herbaceous plants (Horton, Bruns, and Parker 1999; Chou and Muller 1972; Kauffmann et al. 2015). Manzanitas are also culturally important to indigenous people of California (Anderson 2005). In addition, over half of the more than 100 morphologically defined manzanita species and subspecies are narrow endemics with highly restricted distributions and are considered rare and/or endangered (Baldwin et al. 2012; Kauffmann et al. 2015; https://www.rareplants.cnps.org/). The threat to these taxa is increasing due to climate change and human activities (Halsey and Keeley 2016).

1.3 Genomic architecture and species/subspecies distinction of manzanitas are poorly understood

In contrast to their importance in ecology, evolution, and conservation studies, manzanita taxa are poorly understood in terms of genetic and ecological differentiation. To facilitate studies on manzanita evolution and conservation, my dissertation addresses three knowledge gaps: (1) genomic architecture of manzanitas, (2) ecological

diversification among manzanita species, and (3) genetic differentiation of manzanita subspecies.

(1) Until now, genomic resources for manzanitas have been nearly nonexistent, consisting only of investigations into karyotypes of diploid (2n = 2x = 26) and tetraploid (2n = 4x = 48) species (Wells 1968). We recently reported the first manzanita genome assembly for a widespread diploid species, Big Berry Manzanita (*Arctostaphylos glauca*) (Huang et al. 2021), a potential ancestral species of many putative *Arctostaphylos* hybrids (Parker 2007). However, further investigation into the genomic content had not been conducted, limiting the ability to study adaptation and diversification. In my second chapter, I annotated and analyzed a recently reported Big Berry Manzanita genome and compared it with the genomes of other related species that are not known to be adapted to fire and drought. I hypothesized that the manzanita genome contains lineage-specific genes that might contribute to their adaptation.

(2) Our current understanding of ecological differentiation among manzanita species is based on the description of habitats (Kauffmann et al. 2015) but has not been tested quantitatively to determine if habitats really are distinct from each other. In the third chapter, I investigated environmental diversification among 49 narrowly-distributed manzanita species endemic to the CFP to identify ones with distinct ecological niches. I hypothesized that at least some of these manzanita species live in unique habitats and require habitat preservation for their conservation.

(3) Among the currently recognized manzanita species and subspecies, many are quite similar morphologically and differentiated by a few traits such as glandular hairs, leaf color, and fruit shape (Kauffmann et al. 2015; Baldwin, Goldman, Keil, Patterson,

and Rosatti 2012). However, because of the lack of genetic studies, whether these morphologically defined manzanita taxa are genetically distinct remains unknown. In the fourth chapter, I investigated genetic diversification of multiple subspecies of Eastwood Manzanita (*Arctostaphylos glandulosa*), including two subspecies that are state and federally listed as rare and endangered (Kajtaniak and Easterbrook 2019; Smith 2020). I hypothesized that these two subspecies are genetically distinct from the other subspecies, and that targeted conservation efforts are appropriate.

1.4 gReferences

Anderson, M Kat. 2005. 'Tending the wild.' in, *Tending the Wild* (University of California Press).

Baldwin, Bruce G, Douglas H Goldman, David J Keil, Robert Patterson, Thomas J Rosatti, and Linda Ann Vorobik. 2012. *The Jepson manual: vascular plants of California* (Univ of California Press).

Baldwin, Bruce G., Douglas H. Goldman, David J. Keil, Robert Patterson, and Thomas J. Rosatti. 2012. *The Jepson Manual: Vascular Plants of California* (University of California Press).

Brandies, Parice, Emma Peel, Carolyn J Hogg, and Katherine Belov. 2019. 'The value of reference genomes in the conservation of threatened species', *Genes*, 10: 846.

Burge, Dylan O., James H. Thorne, Susan P. Harrison, Bart C. O'Brien, Jon P. Rebman, James R. Shevock, Edward R. Alverson, Linda K. Hardison, José Delgadillo RodrÍguez, Steven A. Junak, and Others. 2016. 'Plant diversity and endemism in the California Floristic Province', *Madroño*: 3-206.

Chou, Chang-Hung, and Cornelius H. Muller. 1972. 'Allelopathic Mechanisms of Arctostaphylos glandulosa var. zacaensis', *Am. Midl. Nat.*, 88: 324-47.

Corlett, Richard T. 2016. 'Plant diversity in a changing world: status, trends, and conservation needs', *Plant diversity*, 38: 10-16.

Halsey, Richard W, and Jon E Keeley. 2016. 'Conservation issues: California chaparral'.

Horton, Thomas R., Thomas D. Bruns, and V. Thomas Parker. 1999. 'Ectomycorrhizal fungi associated with Arctostaphylos contribute to Pseudotsuga menziesii establishment', *Can. J. Bot.*, 77: 93-102.

Huang, Yi, Merly Escalona, Glen Morrison, Mohan P. A. Marimuthu, Oanh Nguyen, Erin Toffelmier, H. Bradley Shaffer, and Amy Litt. 2021. 'Reference genome assembly of the big berry Manzanita (Arctostaphylos glauca)', *J. Hered.*

Kajtaniak, David, and Nicholas Easterbrook. 2019. 'California Department of Fish and Wildlife'.

Kauffmann, Michael Edward, Tom Parker, Michael Vasey, and Jeff Bisbee. 2015. *Field Guide to Manzanitas* (Backcountry Press).

Keeley, Jon E. 1991. 'Seed germination and life history syndromes in the California chaparral', *Bot. Rev.*, 57: 81-116.

Minnich, R., and L. Howard. 1984. 'Shrublands in California: Literature Review and Research Needed for Management: Biogeography and Prehistory of Shrublands'.

Parker, V. Thomas. 2007. 'Diversity and Evolution of Arctostaphylos and Ceanothus', *Fremontia*: 8.

Smith, James P. 2020. 'A list of the rare, endangered, & threatened vascular plants of California'.

Vasey, Michael C., Michael E. Loik, and V. Thomas Parker. 2012. 'Influence of summer marine fog and low cloud stratus on water relations of evergreen woody shrubs (Arctostaphylos: Ericaceae) in the chaparral of central California', *Oecologia*, 170: 325-37.

Wells, Philip V. 1968. 'New taxa, combinations, and chromosome numbers in Arctostaphylos (Ericaceae)', *Madroño*, 19: 193-210.

2 Chapter 2 Chromosome-level Genome Assembly and Annotation Reveals the Enrichment of Terpenoid Biosynthetic and Metabolic Genes in the Big Berry Manzanita

2.1 Introduction

The California Floristic Province (CFP) is a worldwide biodiversity hotspot, located on the west coast of North America. Estimates of plant diversity in the CFP vary from 3000 to over 6000 species, 60% of which are endemic to the CFP (Burge et al. 2016; Baldwin 2014; Myers 1990; Raven and Axelrod 1978). The CFP is a fire-prone region with a Mediterranean-type climate (MTC) characterized by hot, dry summers and cool, wet winters. The high plant diversity and endemism are thought to be the result of many factors including diversity of soil types, climatic variation from the coast to inland, a range of elevations, and the historical shift to an MTC (Raven and Axelrod 1978; Baldwin 2014). In addition, a recent study pointed out that recurrent fire also has likely played a critical role in assembling the modern flora of the CFP (Rundel et al. 2018). Chaparral species are classic components of CFP flora: they are well adapted to the summer drought and fire disturbance, and form one of the dominant vegetation types in the CFP. Specifically, in the largest political region of the CFP, the state of California, 73% of the vegetation is chaparral (Bolsinger 1989). Currently, the chaparral habitat is decreasing and under threat due to land development, competition with invasive species, climate change, and increased fire frequency (Halsey and Keeley 2016a). Consequently, preservation of chaparral communities is a critical component of conservation efforts in the CFP.

Arctostaphylos (Ericaceae) species, commonly known as manzanitas, are iconic shrubs of the CFP chaparral, characterized by evergreen leaves, red and twisting branches, and clusters of urn-shaped flowers (Kauffmann et al. 2015; Baldwin et al. 2012). This genus is composed of 105 currently recognized species and subspecies, 104 of which have some or all of their distribution in the CFP (Kauffmann et al. 2015) making *Arctostaphylos* the most diverse woody genus in the CFP (Keddy 2017). Within the genus, a large number of species and subspecies are restricted to unique soil types that only occupy small geographic areas (Kauffmann et al. 2015; Parker 2007). A majority of these edaphic endemic taxa only consist of one or two wild populations, and therefore are classified as rare and /or endangered (https://www.rareplants.cnps.org/). Because of the high species richness and endemism, *Arctostaphylos* is a good system for investigation on evolution and diversification of CFP plants. Moreover, as this genus includes so many rare and endangered species and subspecies, manzanitas are a critical component of conservation management in the CFP flora (Gluesenkamp et al. 2011; Burge et al. 2018; Halsey and Keeley 2016b).

Reference genome sequences, with gene annotation, can serve as valuable resources for developing conservation strategies for threatened species: they can significantly improve the analysis of genetic variation, facilitating many studies that are essential for species conservation, including population genetics, landscape genomics, and phylogenetics (Brandies et al. 2019). As part of our ongoing studies of manzanita diversity, we recently published a brief report of a high-quality reference genome assembly for one of the widely distributed species of *Arctostaphylos*, the Big Berry Manzanita (*Arctostaphylos glauca*) (Huang et al. 2021). The assembly indicates a genome size of 547 Mb, indicating moderate size, and the assembly shows 98.2%

BUSCO completeness, suggesting one of the highest quality assemblies available among the Ericaceae family (Huang et al. 2021).

In this study, we report a more in-depth analysis of genomic contents to identify genomic signatures that might underlie the adaptation of manzanitas to the drought- and fire-mediated chaparral. We annotated the *A. glauca* genome and compared it with other members of the Ericales and Ericaceae clades to identify elements of genomic structure and gene contents that are specific to *A. glauca*. We found that our reference genome is composed of 13 long scaffolds, consistent with the haploid number of 13 in the genus. These pseudo chromosomes contain a majority of the annotated gene models. In addition, we also found that terpenoid-related genes, which have been implicated in drought- and fire-adaptation, are enriched in the *A. glauca* genome, suggesting one element of adaptation to the chaparral habitat.

2.2 Materials and Method

Plant material, DNA extraction, library construction, sequencing methods, assembly, and quality assessment methods for the nuclear and organellar genomes, as well as the annotation of the organellar genomes, are reported in Huang et al. 2021.

2.2.1 Biological Materials for RNA-seq

We collected tissues from two Big Berry Manzanita plants in the San Gabriel Mountains, Angeles National Forest, Los Angeles County, California on January 19th, 2021. Voucher specimens were deposited at the herbarium of UCR (UCR). We collected the inflorescence twigs and floral buds from one plant (Z. Guo 04; UCR ACC. # 292565), and sampled young leaf tissues from the other (Z. Guo 01; UCR ACC. #292552). We

immediately froze the tissue in liquid nitrogen and stored at −80 °C before RNA extraction.

2.2.2 RNA isolation, library preparation, and sequencing

We ground a small amount of tissue (two buds, a dime size leaf fragment, or two ~1cm young twigs) in an Omni Bead Ruptor Elite with a liquid nitrogen feed in NEB RNA/DNA protection regent and isolated the RNA using Monarch® Total RNA Miniprep Kit (New England Biolabs, MA, USA) following the manufacturer's protocol. To prepare the RNA-Seq libraries, we used ~500mg RNA from each sample as input and used the NEB Ultra II Directional RNA Library kit (NEB, E7765), with polyA mRNA isolation (NEB, E7490). We followed the manufacturer's protocol with these modifications: (1) after 2nd strand synthesis, we used 0.8X instead of 1.8X bead to purify the product; (2) after the ligation step, we added an extra clean-up step using 0.7X beads; (3) in the PCR enrichment of adaptor-ligated DNA, we set the number of amplification cycles to 11; (5) at the final clean up step, we changed the single bead size selection to dual bead size selection to further refine the final library pool for sequencing. We pooled the 3 samples in equimolar quantities and sequenced on NovaSeq 150PE S4 flow cell (Illumina, San Diego, CA) at UCR Genomics Core.

2.2.3 Nuclear Genome Annotation

To identify repetitive elements,and soft mask them in the genome, we used RepeatModeler v. 2.0.1 (Flynn et al. 2020) and RepeatMasker 187 v. 4.1.1 (Smit, Hubley, and Green 2015).

With the masked genome, we conducted structural annotation and functional annotation using the Funannote pipeline v. 1.8.4 (Palmer and Stajich 2017). To train the gene models with external transcriptome evidence, we used Trinity v. 2.11.0 and PASA v. 2.4.1 to assemble the transcript and align the assembled transcriptome to the genome (Haas et al. 2003; Grabherr et al. 2011). Next, we used the genome sequence and PASA output to perform gene prediction using software including Augustus v. 3.3.3, GeneMark-ETS v. 4.62, GlimmerHMM v. 3.0.4, and SNAP v 2013_11_29 (Korf 2004; Majoros, Pertea, and Salzberg 2004; Stanke et al. 2006; Ter-Hovhannisyan et al. 2008). Following that, we applied EVidenceModeler v. 1.1.1 (Haas et al. 2008) to combine gene predictions and transcript alignments into weighted consensus gene models, and used tRNAscan-SE v. 1.3.1 (Lowe and Eddy 1997) to annotate tRNAs. The RNA-seq training data from PASA were then used to add untranslated regions (UTR) to refine the gene models. We used BUSCO v 3.0.2 (Simão et al. 2015) to assess the completeness of these structural genome annotations with the embryophyta_odb9 lineage data set and default BUSCO parameters under the transcriptome mode.

To assign names and functions to the predicted genes, we used several curated databases including Pfam (Finn et al. 2014), CAZyme domains (Lombard et al. 2014; Huang et al. 2018), MEROPS (Rawlings, Barrett, and Bateman 2014), eggNOG v. 2.1.0 (Huerta-Cepas et al. 2016), InterProScan v. 212 5.47-82.0 (Jones et al. 2014), and Swiss-Prot (Boutet et al. 2016). Additionally, we used Phobius v. 1.01 (Käll, Krogh, and Sonnhammer 2004) to predict transmembrane proteins, and SignalP v. 5.0b (Almagro Armenteros et al. 2019) to predict secreted proteins.

To visualize the gene and repeat contents of the 13 pseudo-chromosomes, we used bedtools (Quinlan and Hall 2010; Quinlan 2014) to divide each chromosome into 1000 equal-size windows and counted the number of annotated genes, and repetitive elements within each window. We represented the data in Circo plots using OmicStudio online tools (https://www.omicstudio.cn/tool/50).

2.2.4 Comparative Genomic Analysis

To investigate chromosomal evolution and whole genome duplications (WGD) in the *Arctostaphylos* genome, we conducted synteny analyses using the protein coding sequence of the *A. glauca* genome and seven other Ericales' genomes that are of chromosome-level assembly (Table 2.1). We used *Vitis vinifera* as a reference for comparison to explore the number of WGD that the Ericales have experienced because this species is known to have experienced only the core eudicot-specific whole genome triplication (WGT) (Jaillon et al. 2007). We used the Python version of the MCScan toolkit (Tang et al. 2008) with the default setting to conduct synteny analysis on *A. glauca* versus *A. glauca, A. glauca* versus every other Ericales species and *V. vinifera* versus every other species. We filtered the all-against-all LAST (Kiełbasa et al. 2011) hits to find the best 1:1 syntenic blocks for these intraspecific and interspecific analyses. We produced macrosynteny dotplots, synteny depth, and karyotype figures using the "dotplot", "synteny" and "karyotype" functions.

Species	Family	Order	Publication	Assembly	Synteny Analysis	Orthology Analysis
Arctostaphylos glauca	Ericaceae	Ericales	Huang et al., 2021	Chromosome	Yes	Yes
Rhododendron delavayi	Ericaceae	Ericales	Zhang et al., 2017	Scaffold	No	Yes
Rhododendron griersonianum	Ericaceae	Ericales	Ma et al., 2021	Chromosome	No	Yes
Rhododendron kiyosumense	Ericaceae	Ericales	Shirasawa et al. 2021	Scaffold	No	Yes
Rhododendron ovatum	Ericaceae	Ericales	Wang et al., 2021	Chromosome	Yes	Yes
Rhododendron simsii	Ericaceae	Ericales	Yang et al., 2020	Chromosome	Yes	Yes
Rhododendron williamsianum	Ericaceae	Ericales	Soza et al., 2019	Chromosome	No	Yes
Vaccinium corymbosum (tetraploid)	Ericaceae	Ericales	Colle et al., 2019	Chromosome	Yes	Yes
Vaccinium corymbosum (diploid)	Ericaceae	Ericales	Gupta et al, 2015	Scaffold	Yes	Yes
Vaccinium myrtillus	Ericaceae	Ericales	Wu et al., 2022	Scaffold	Yes	Yes
Actinidia chinensis	Actinidiaceae	Ericales	Pilkington et al., 2018	Chromosome	No	Yes
Actinidia eriantha	Actinidiaceae	Ericales	Tang et al., 2019	Chromosome	Yes	Yes
Camellia sinensis	Theaceae	Ericales	Xia et al., 2020	Chromosome	Yes	Yes
Diospyros oleifera	Ebenaceae	Ericales	Suo et al., 2020	Chromosome	Yes	Yes
Primula vulgaris	Primulaceae	Ericales	Cocker et al., 2018	Scaffold	No	Yes
Primula veris	Primulaceae	Ericales	Nowak et al., 2015	Scaffold	No	Yes
Arabidopsis thaliana	Brassicaceae	Brassicales	Michael et al., 2018	Chromosome	No	Yes

Table 2.1 taxonomy, relevant publication, assembly information, and analysis types of 17 plant species that we included in the comparative genomic analysis

To identify genomic features that might contribute to the adaptation of *A. glauca* to the harsh MTC conditions, we used OrthoFinder2 (Emms and Kelly 2019) to compare the protein sequences of *A. glauca* with other Ericales species and the model plant species *A. thaliana*. Aside from *A. glauca*, the other plant species included here are not known to be drought- and fire-adapted. OrthoFinder2 can process fragmented, low-coverage genome assemblies, therefore we were able to add additional species for the synteny analysis for a total of 17 species (Table 2.1) (Huang et al. 2021; Zhang et al. 2017; Ma et al. 2021; Shirasawa et al. 2021; Wang et al. 2021; Yang et al. 2020; Soza et al. 2019; Colle et al. 2019; Gupta et al. 2015; Wu et al. 2022; Pilkington et al. 2018; Tang et al. 2019; Xia et al. 2020; Suo et al. 2020; Cocker et al. 2018; Nowak et al. 2015; Michael et al. 2018). Orthofinder2 clustered the protein sequences of the 17 assemblies into orthogroups, reconstructed species tree, and inferred gene trees. We retrieved orthogroups that only consist of Big Berry Manzanita genes, and orthogroups where the number of *A. glauca* genes is larger than the other species, and conducted a Gene Ontology (GO) enrichment analysis on these orthogroups using a customized R script (Rajewski et al. 2021). We rooted the inferred species tree using *A. thaliana* as an outgroup, and plotted the tree using the phytools and ggtree R packages (Revell 2012).

In response to fire disturbance, many manzanita species including *A. glauca* rely on smoke-induced seed germination to replace the populations that are killed by the fires (Keeley 1991). Karrikin signaling pathways have been known to play important roles in such smoke-induced seed germination (Flematti et al. 2004; Flematti et al. 2009; Van Staden et al. 2006). To test whether there was any indication in the history of karrikin-related genes suggesting an adaptive role in manzanitas, we investigated gene family expansion and species-specific gene duplication of these genes across the Ericaceae

clade. We specifically restricted the analysis to the Ericaceae family because we want to restrict the difference of manzanitas relatively to the other plants so that the observed changes of genes could be more confidently linked to the specific fire-adaptation of manzanitas. For such an investigation, we selected four key genes involved in the karrikin signaling pathway: KARRIKIN INSENSITIVE2 (KAI2), Delta Like Non−Canonical Notch Ligand 2 (DLK2), MORE AXILLARY BRANCHES2 (MAX2), SUPPRESSOR OF MAX2 (SMAX1) (Nelson et al. 2011; Conn and Nelson 2015). Using phytools and ggtree R packages, we plot the gene trees of the orthogroups of these candidate genes, and trimmed the tips to retain only the outgroup species *A. thaliana* and six species representing the Ericaceae genera, *Arctostaphylos*, *Vaccinium* and *Rhododendron*.

2.3 Results

2.3.1 Genome Annotation

A previous karyotype study revealed that the Big Berry Manzanita contains 26 chromosomes ((2n = 2x = 26)) (Wells 1968). Among the 271 scaffolds of our final assembly (Huang et al. 2021), 13 of them were noticeably longer than the others, ranging from 29 Mb to 45 Mb. These 13 long scaffolds constituted 440 Mb out of the 547 Mb total genome assembly. Given these results, we sorted these 13 scaffolds according to length and referred to them as pseudo chromosomes numbering from largest (AG1) to smallest (AG13).

To determine gene content and infer gene models for the assembly, we first masked repetitive elements, which constituted 57.71% of the genome. Following that, we annotated 40,204 protein-coding genes from the total assembly and localized 36,665 of them to the 13 pseudo chromosomes. In addition, we annotated 453 tRNA genes, 389

of which aligned to the 13 pseudo chromosomes. BUSCO assessment revealed 85.2% completeness of the structural annotation (Table 2.2). Although this is lower than the 98.25% BUSCO completeness of the nucleotide sequence, such a reduction completeness in annotation compared to the original assembly is common (Seppey, Manni, and Zdobnov 2019; Soza et al. 2019).

Description	Total	Percentage
Complete BUSCOs	1227	85.20%
Complete single-copy BUSCOs	1037	72.00%
Complete duplicated BUSCOs	190	13.20%
Fragmented BUSCOs	115	8.00%
Missing BUSCOs	98	6.80%

Table 2.2 BUSCO assessment for the annotation of *A. glauca* genome

A total of 35,518 genes were assigned putative functions using at least one of the curated databases. The number of genes annotated with a Pfam domain and CAZymes domains are 23,114 and 1,322 respectively. 34,000 genes and 28,975 genes have a match hit for the eggNOG and InterProScan database. Total 31,606 genes were assigned functions using Clusters of Orthologous Groups (COG) databases while a large proportion of them are characterized as functionally unknown (Figure 2.1). We also assigned functional annotation to 21,188 genes using Gene Ontology (GO) database, and such an annotation was also used in the GO Enrichment Analysis. In addition, MEROPS added 1,127 proteases, and Phobius and SignalP together added 3,064 secretome and 9,042 transmembrane domains for the final annotation.

We used annotation edit distance (AED) to evaluate the congruence between the annotation and its supporting evidence such as transcriptome or protein sequences from the UniProtKb/SwissProt database. AED value ranges from 0 to 1 and an AED value of 0 represents a perfect annotation with the great support of the evidence (Eilbeck et al. 2009). Our final annotation was with an AED of 0.006 for protein-coding sequences and 0.077 for mRNA, demonstrating the high quality of the annotation.

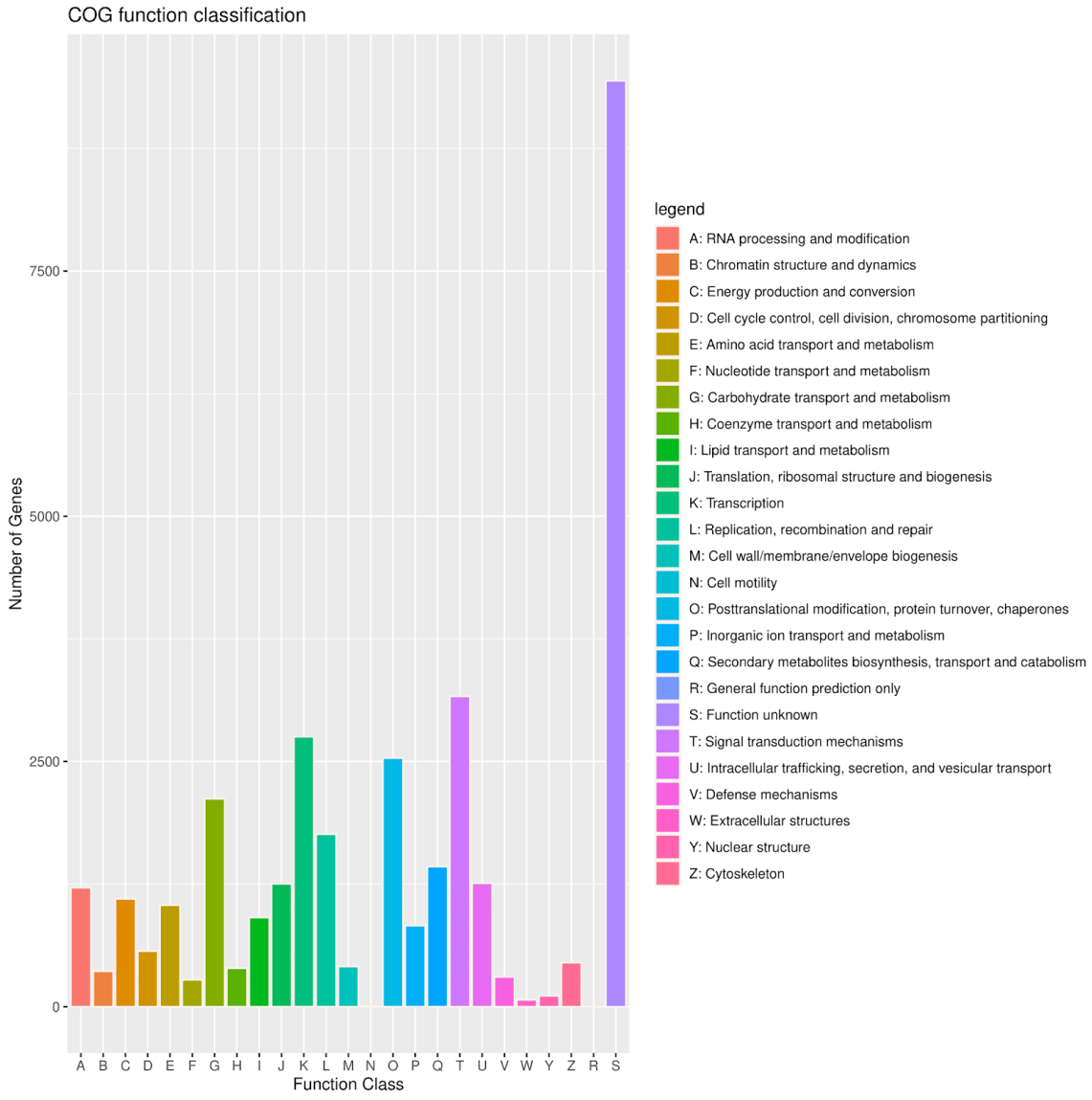


Figure 2.1 COG functional classification of protein-coding genes shows a large number of genes of unknown function in the *A. glauca* annotation. The X-axis shows the COG categories and the Y-axis represents the number of protein-coding genes classified in each category.

2.3.2 Genomic contents in the 13 pseudo chromosomes

We used the circo plot generated by the OmicStudio online tools (https://www.omicstudio.cn/tool/50) to visualize the distribution of genomic elements across these 13 pseudo chromosomes. The plots show high gene density near the ends of the scaffolds and the lowest gene density somewhere in the middle, suggesting the approximate position of the centromeres (Figure 2.2). One segment (~7Mb) of AG5 contains no genes but some repetitive elements. The overall gene density in AG4 was noticeably lower than the others, and there are multiple gaps where no genes, repeats or tRNAs have been identified (Figure 2.2). These regions of chromosomes AG4 without annotated genes correspond to areas of the assembly with a limited number of anchored contigs in the Hi-C heatmap (Huang et al. 2021), suggesting the sequences of the pseudo chromosome is incomplete.

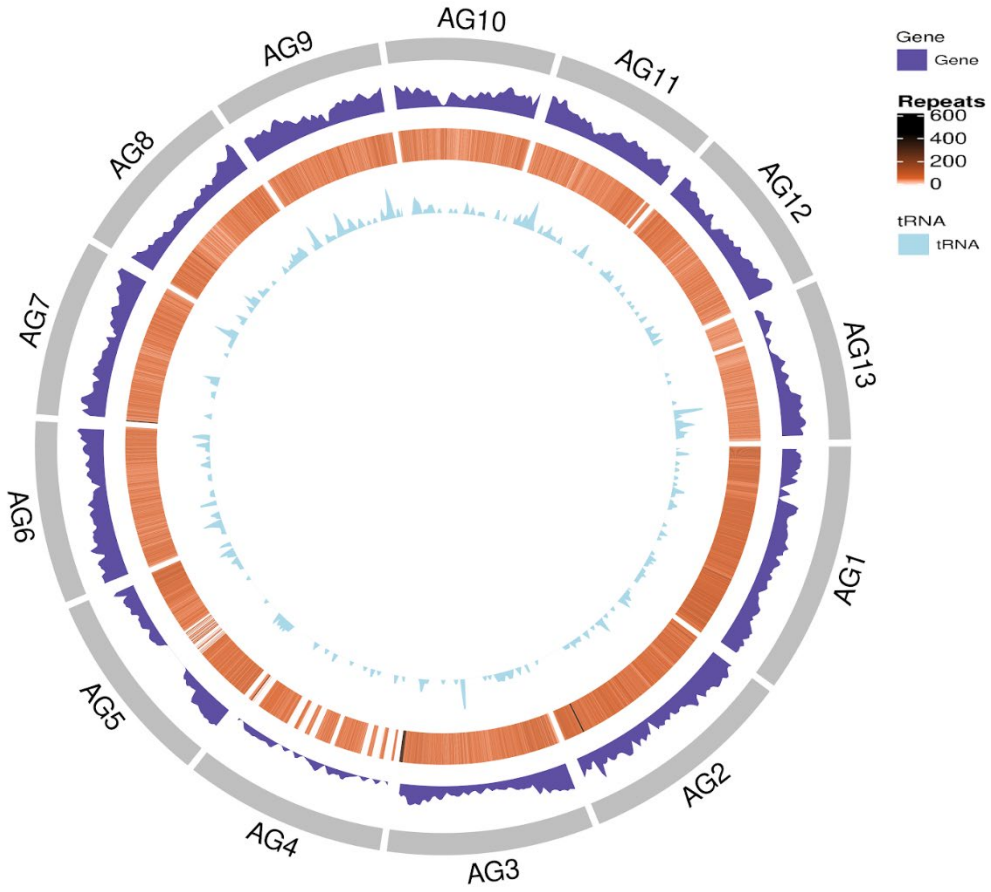


Figure 2.2 Genome features across 13 pseudo chromosomes show a dip in gene density and a peak in repeats toward the chromosomal centers. The four circles show, from outside to inside, the 13 pseudo chromosomes (as gray bars), gene density (in purple), repeat counts (with color gradient of orange to black representing the increasing number of repeats), and tRNA (light blue) respectively.

2.3.3 Evidence from synteny analyses supports independent WGDs in the different Ericales lineages

Through intragenomic comparison using the protein coding sequence of the *A. glauca* genome, we identified 1,441 collinear and duplicated genes and used these to infer the syntenic relationships among *A. glauca* pseudo chromosomes (Figure 2.3). Most pseudo chromosomes consist of large segments that are orthologous to segments of a second *A. glauca* pseudo chromosome, except for AG4 and AG8, which do not appear to share collinearity with other chromosomes. Different segments of pseudo chromosomes AG6 and AG7 share collinearity with three different *A. glauca* chromosomes. This observation of many collinear segments shared with other pseudo chromosomes suggests that *A. glauca* went through an ancient WGD event (Roelofs et al. 2020).

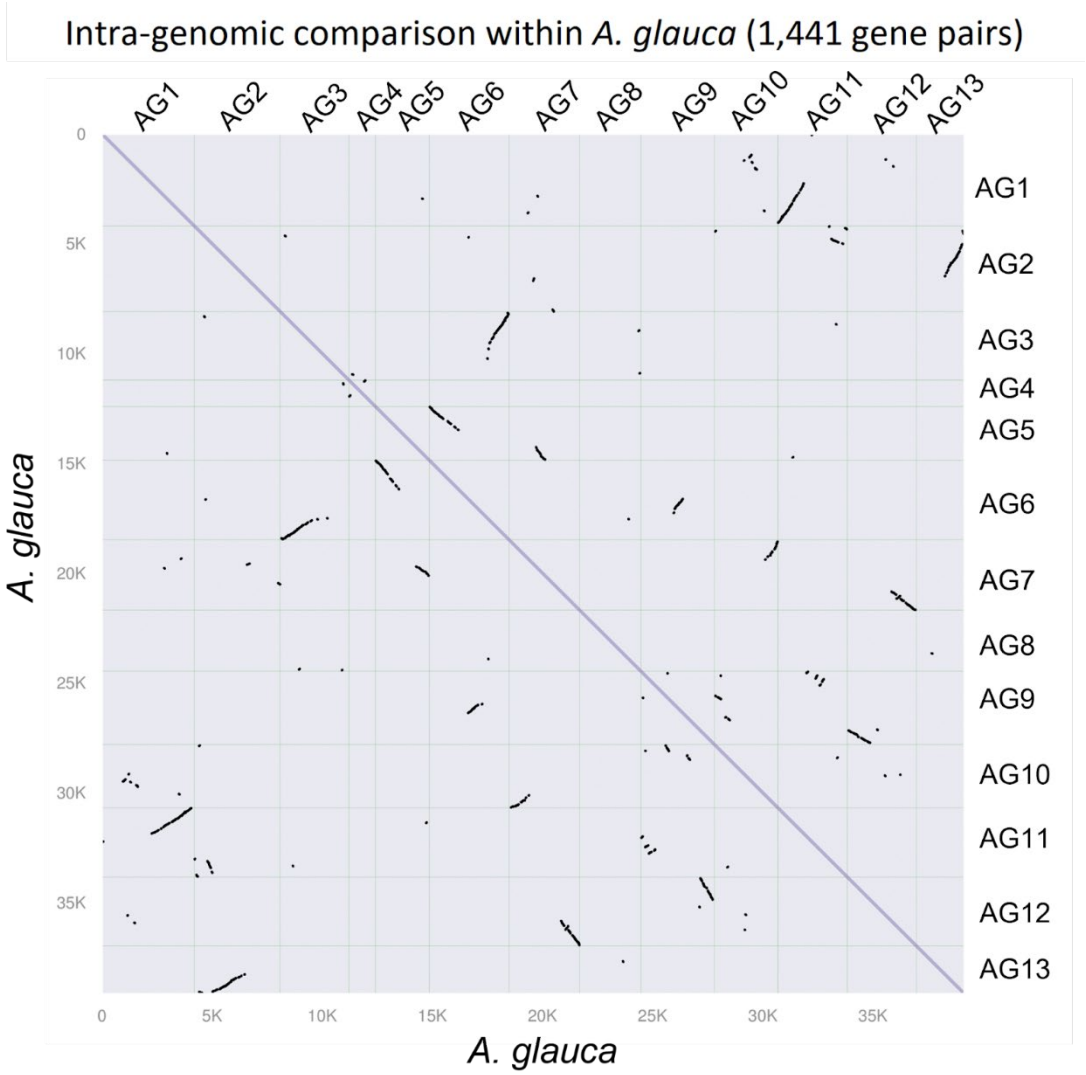
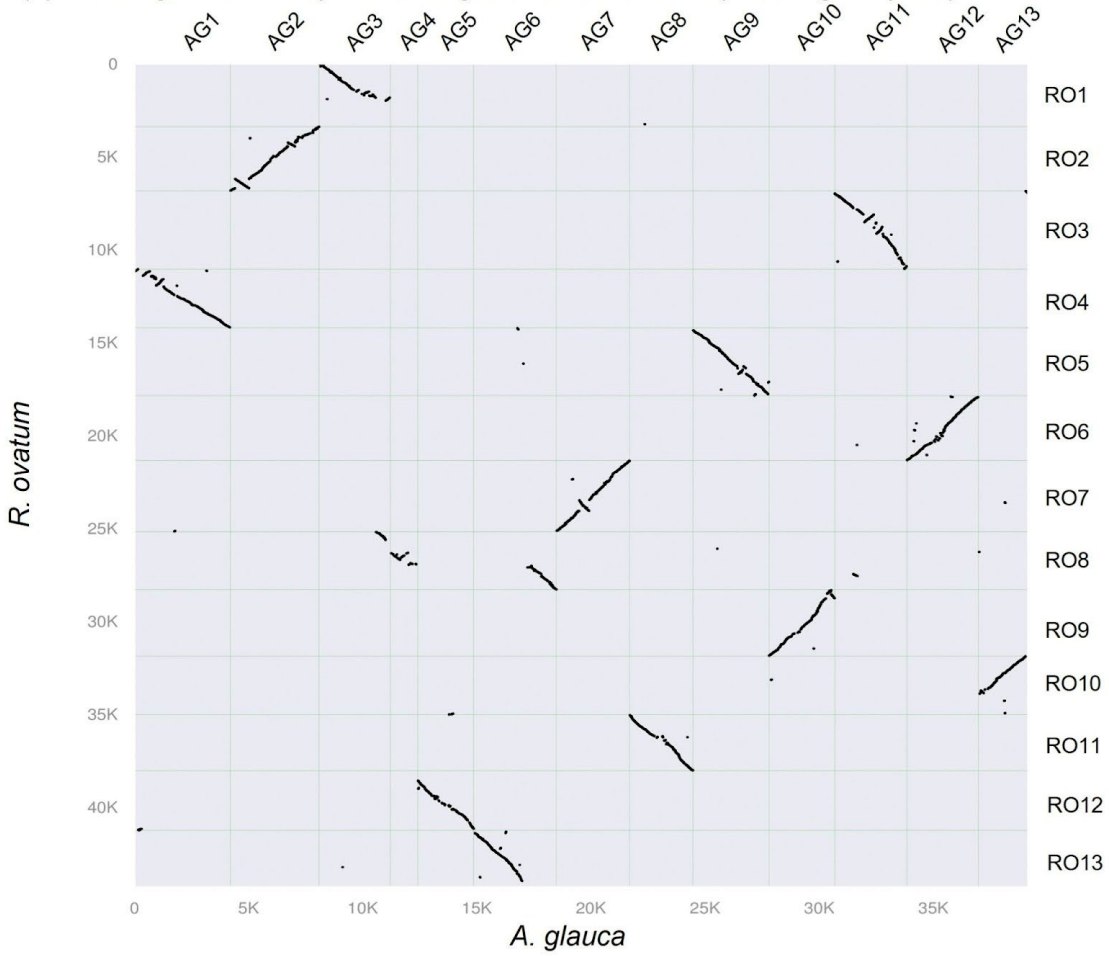

Intra-genomic comparison within A. glauca (1,441 gene pairs)
AG1
AG2
AG3
AG4
AG5
AG6
AG7
AG8
AG9
AG10
AG11
AG12
AG13
0
5K
10K
15K
20K
25K
30K
35K
A. glauca
A. glauca

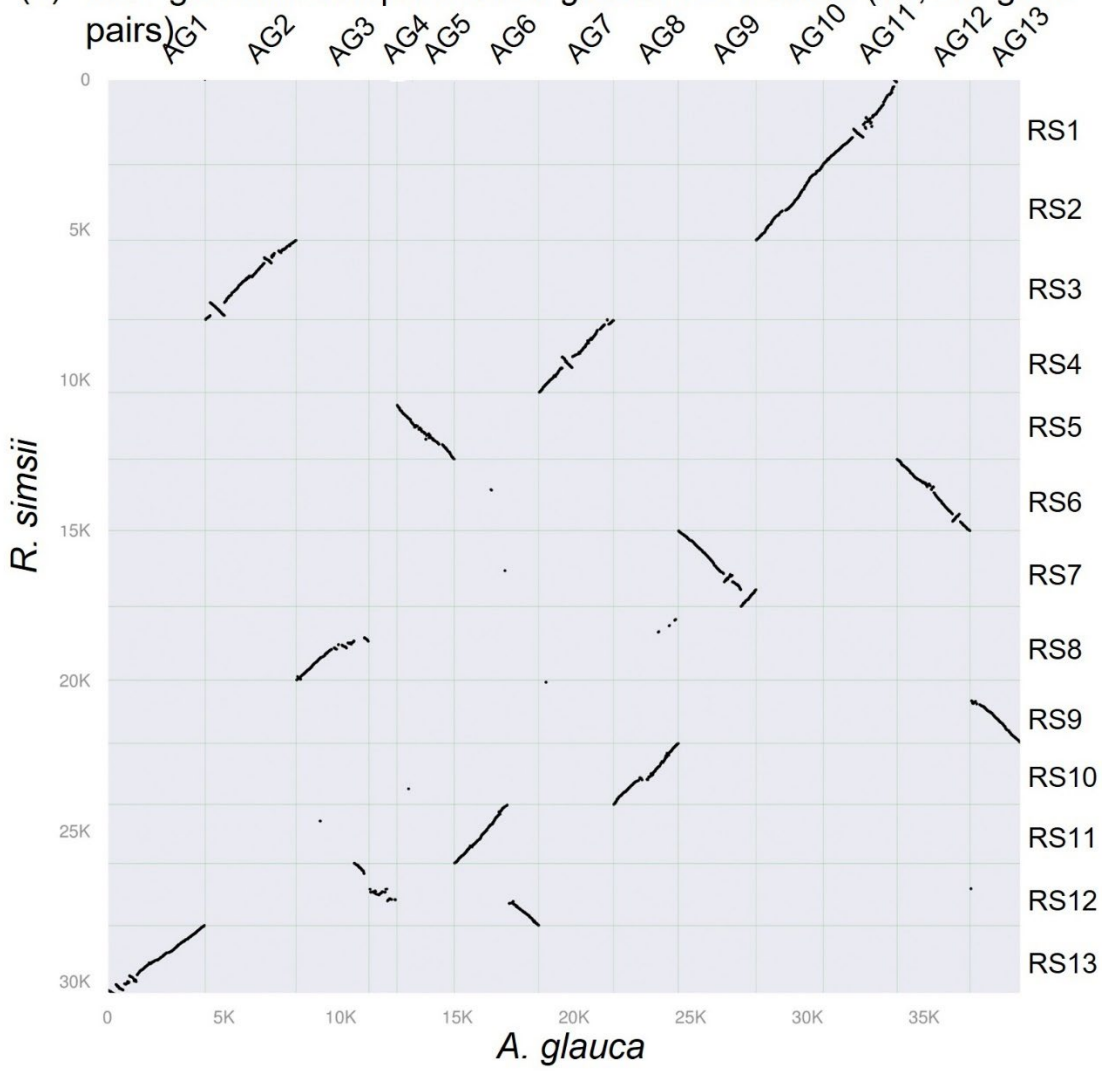
Figure 2.3 Dotplot visualization of pairwise synteny analysis using protein coding sequence as input shows likely whole genome duplication within this species. The y axis represents the reference genome and x axis represents the query genome. Both x axis and y axis show genomic location numbered continuously across all chromosomes, and are divided into blocks corresponding to the chromosomes. The chromosome names of the reference genome and query genome are found in the right and top side. The diagonal is where every locus is mapped and compared to itself. The black dots represent genes that are found in two different positions in the *A. glauca* genome, providing evidence of duplication. Lines of dots indicate a series of collinear loci found in two genomic locations. Lines slanting down from left to right indicate loci in the same order in both locations; lines slanting up indicate loci that are inverted in one location relative to the other.

To investigate chromosomal evolution and organization of *Arctostaphylos*, we conducted interspecific synteny analysis, comparing the protein coding sequence of the *A. glauca* genome with four plant species from two other genera of Ericaceae, *Vaccinium* and *Rhododendron*. Of the 13 total *A. glauca* pseudo chromosomes, 12 had a perfect one-to-one orthologous match with the *Rhododendron* chromosomes. The 13th, AG6, appeared to be orthologous to two *Rhododendron* chromosomes (Figure 2.4; Appendix S2.1). In contrast, only six chromosomes of *A. glauca* have a one-to-one match with the *Vaccinium* species. The remaining seven chromosomes shared collinearity with two different chromosomes of *Vaccinium*, with one segment of those *A. glauca* pseudo chromosomes mapping to one *Vaccinium* chromosome and a second segment mapping to a second *Vaccinium* chromosome (Figure 2.4; Appendix S2.1). These results suggested that *Arctostaphylos* and *Rhododendron* are more similar to each other in terms of chromosomal organization. This contrasts with current hypotheses regarding the phylogenetic relationships of these three genera, which infer that *Rhododendron* and *Vaccinium* form a sister clade to the *Arctostaphylos* lineage (Rose et al. 2018). This suggests that *Vaccinium* has undergone genome rearrangement since its diversification from the common ancestor of the three genera.

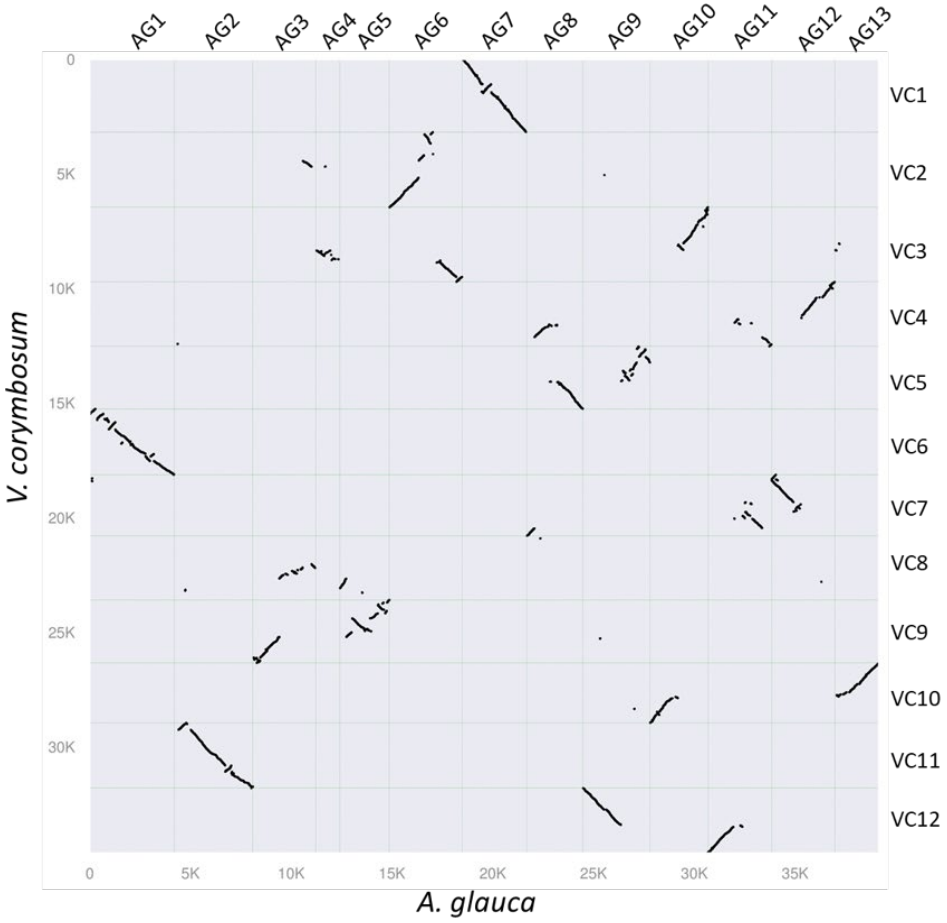
(a) Inter-genomic comparison: *A. glauca* vs *R. ovatum* (14,800 gene pairs)



(b) Inter-genomic comparison: *A. glauca* vs *R. simsii* (13,142 gene pairs)



(c) Inter-genomic comparison: *A. glauca* vs *V. corymbosum* (13,414 gene pairs)



(d) Inter-genomic comparison: *A. glauca* vs *V. myrtillus* (13,308 gene pairs)

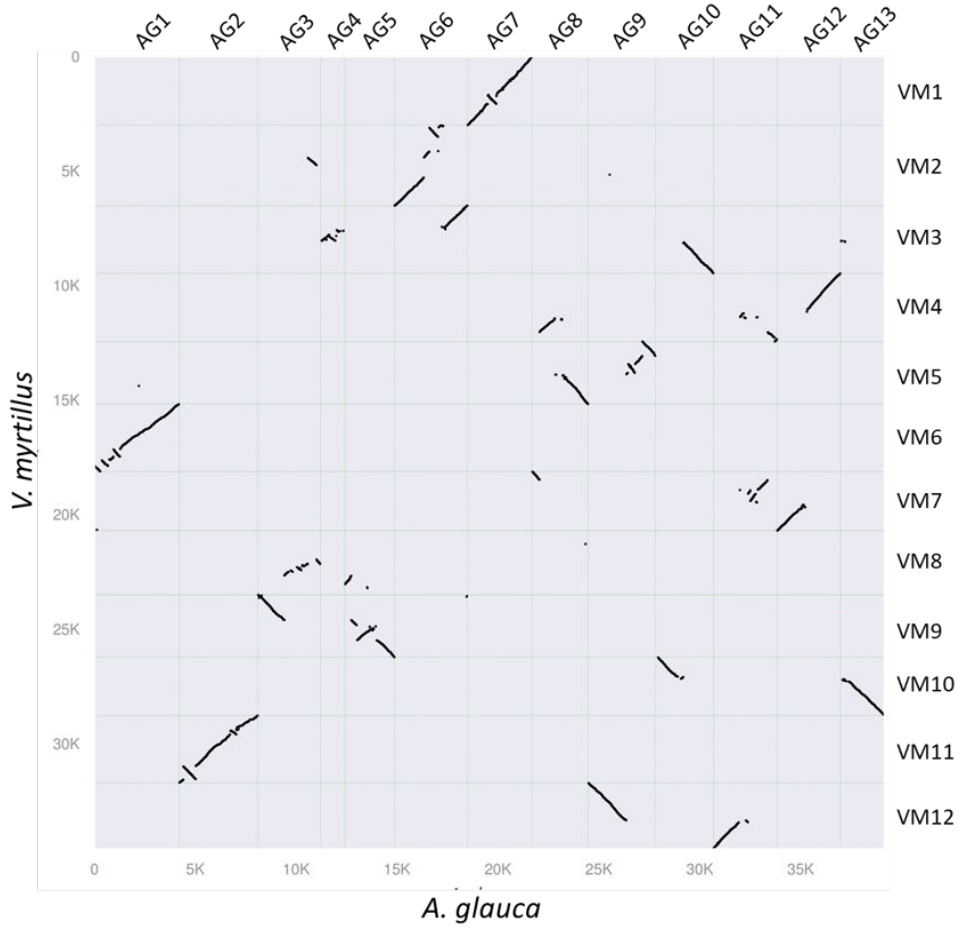


Figure 2.4 Synteny analyses comparing the protein coding sequence of *A. glauca* to four other members of the Ericaceae shows that the chromosomal organization of *Arctostaphylos* is more similar to *Rhododendron* species rather than *Vaccinium* ones. In these dotplots, the y axis represents the reference genome and x axis represents the query genome. In all four panels (a-d), the *A. glauca* assembly is the query genome. The reference genome assemblies are: (a) *Rhododendron ovatum*, (b) *R. simsii*, (c) *Vaccinium corymbosum* and (d) *V. myrtillus*. Both the x axes and y axes show genomic location numbered continuously across all chromosomes, and are divided into blocks corresponding to the chromosomes. The chromosome names of the reference genome and query genome are indicated on the right and top sides. Note that in publications, the chromosomes of the different species have not been labeled in the same order. The black dots show the locations in each genome of orthologous genes in the two species. Lines of dots indicate a series of collinear loci found in the two genomes. Lines slanting down from left to right indicate loci in the same order in both genomes; lines slanting up indicate loci that are inverted in one genome relative to the other.

To investigate the number of WGD events that Ericales has gone through, we compared the synteny ratio between *V. vinifera* and eight species of Ericales including *A. glauca* (Table 2.1). We identified a one-to-two syntenic depth ratio for the five Ericaceae species (*A. glauca*, *Rhododendron ovatum*, *R. simsii*, *Vaccinium corymbosum*, and *V. myrtilus*), *Diospyros oleifera* and *Camellia sinensis*, suggesting one additional WGD event in these species lineages in addition to the core eudicot-specific whole genome triplication (WGT) shared with grape (Appendix S2.2). The synetnic depth ratio between *V. vinifera* and *Actinidia eriantha* is 1:4, indicating that *A. eriantha* has undergone two WGD events after WGT. To investigate the WGD history of *A. glauca* in the context of the Ericales, we compared the synteny depth ratio between the genomes of *A. glauca* and other members of the Ericales that have chromosome-level assemblies. We found a one-to-one syntenic depth ratio with the four Ericaceae species (*R. ovatum*, *R. simsii*, *V. corymbosum*, and *V. myrtilus*), two-to-two with *D. oleifera* and *C. sinensis*, and one-to-two with *A. eriantha* (Appendix S2.3). These together suggest that *Arctostaphylos* shares a WGD with the other Ericaceae. It also confirms previous observations that *A. eriantha* shared one WGD with the Ericaceae and then experienced a subsequent lineage-specific WGD after their divergence (Huang et al. 2013; Shi, Huang, and Barker 2010; Wu et al. 2019). In addition, it also supports previous hypothesis that the WGDs in the Ericaceae + Actinidiaceae clade, the Ebenaceae clade, the Theaceae clade are independent events.

2.3.4 Comparative genomic statistics of the Ericales revealed a high proportion of lineage-specific genes in *A. glauca* genome

Orthofinder 2 assigned 91.5% of the 866,243 protein-coding genes of 17 Ericales species into 70,752 orthogroups, sets of genes derived from common ancestral genes (Emms and Kelly 2019; Emms and Kelly 2015), and reconstructed a species tree (Figure 2.5). This species tree was consistent with previous hypotheses regarding the phylogeny of the Ericales, providing good support for the recognition of orthogroups (Emms and Kelly 2015). Of these 70,752 orthogroups, approximately 52.4% (153,559 genes) were species-specific. In contrast, only 0.8% of the orthogroups included genes from all species.

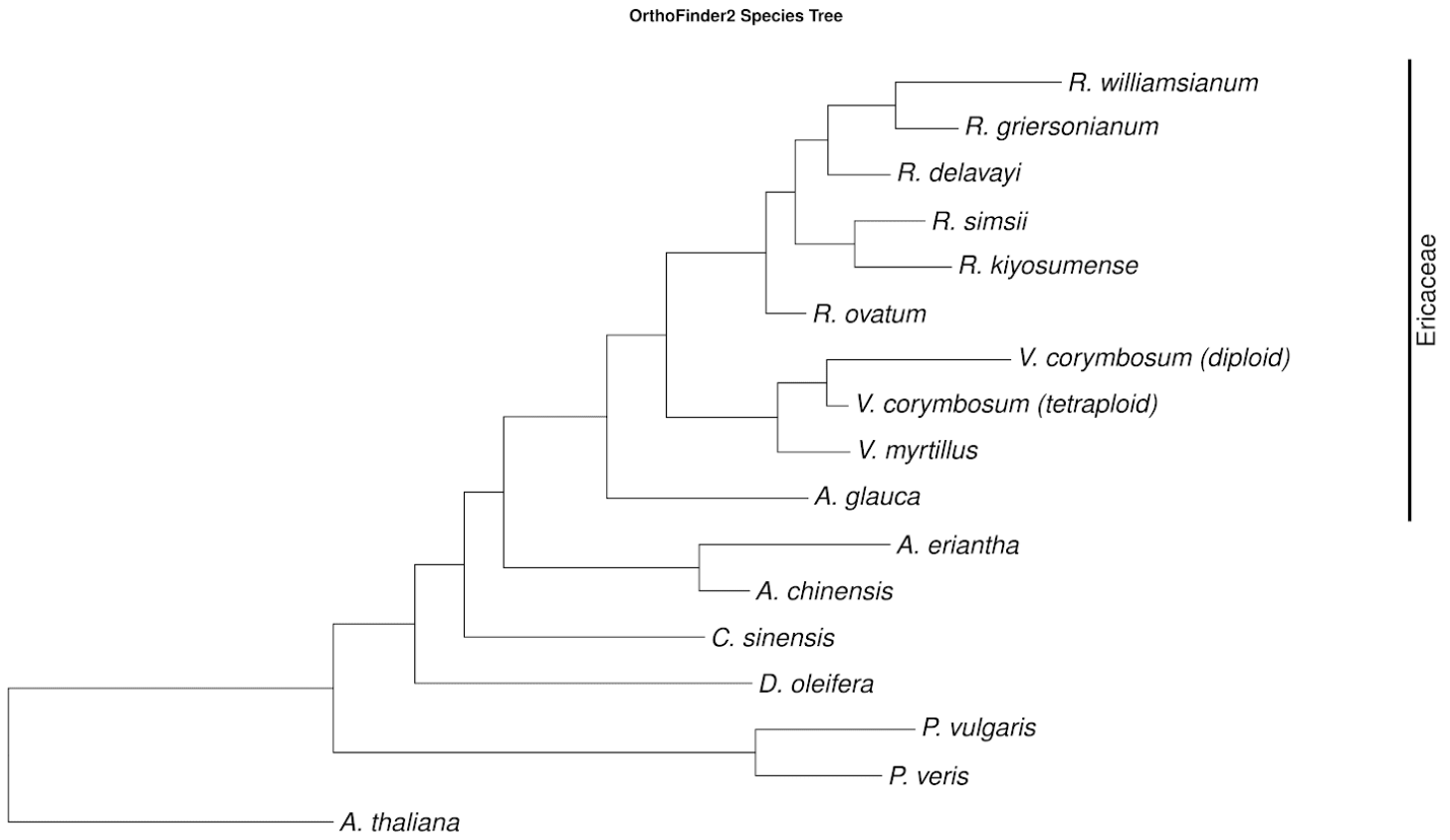


Figure 2.5 Species relationships inferred by OrthoFinder2 are consistent with the results of published phylogenetic analyses. The Ericaceae species form a monophyletic clade that is highlighted by the black bar.

A.glauca had a moderate number of total protein-coding genes when compared to other well-assembled Ericales genomes, which range from 23,548 to 128,559 genes (Table 2.3). Approximately 87.6% of the *A. glauca* genes were assigned to 35,046 orthogroups, and 996 of these were species-specific. These species-specific orthogroups were composed of 4,931 genes, which made up 12.93% of the total number of *A. glauca* genes. Among the well-assembled genomes of the Ericales, the proportion of lineage-specific genes range from 0.1% to 16.1%, thus A. glauca has one of the highest proportions (Table 2.3).

Using individual gene trees based on each orthogroup, and the species tree, Orthofinder identified the number of species-specific gene duplication events for every species. The number of such events ranged from 699 to 22,032 in the Ericales. Among the total 17 species, *A. glauca* ranked in the middle with 8,637 lineage-specific gene duplication events (Table 2.3).

Species	Number of protein-coding genes	Lineage-specific gene duplication events	Orthofinder Genes		
			Assigned Orthogroups	Unassigned	Lineage-specific
A. chinensis	33,115	5,391	32,607 (98.5%)	508 (1.5%)	66 (0.2%)
A. eriantha	42,988	6,299	38,530 (89.6%)	4,458 (10.4%)	1,558 (3.6%)
A. glauca	40,024	8,672	35,046 (87.6%)	4,978 (12.4%)	4,931 (12.3%)
A. thaliana	27,416	9,157	25,172 (91.80%)	2,244 (8.2%)	3,674 (13.4%)
C. sinensis	50,525	22,032	48,399 (95.8%)	2,126 (4.2%)	8,145 (16.1%)
D. oleifera	30,530	6,436	28,001 (91.7%)	2,529 (8.3%)	1,564 (5.1%)
P. veris	18,301	1,007	17,962 (98.1%)	339 (1.9%)	74 (0.4%)
P. vulgaris	174,522	84,256	151,764 (87.0%)	22,758 (13.0%)	112,684 (64.6%)
R. delavayi	32,938	2,206	31,881 (96.8%)	1,057 (3.2%)	245 (0.7%)
R. griersonianum	51,870	16,771	50,587 (97.5%)	1,283 (2.5%)	1,748 (3.4%)
R. kiyosumense	34,606	5,547	34,047 (98.4%)	559 (1.6%)	295 (0.9%)
R. ovatum	44,657	10,482	42,816 (95.9%)	1,841 (4.1%)	999 (2.2%)
R. simsii	32,999	3,982	32,548 (98.6%)	451 (1.4%)	163 (0.5%)
R. williamsianum	23,548	699	23,270 (98.8%)	278 (1.2%)	22 (0.1%)
V. corymbosum (diploid)	60,754	7,309	42,610 (70.1%)	18,144 (29.9%)	3,261 (5.4%)
V. corymbosum (tetraploid)	128,559	70,649	120,914 (94.1%)	7,645 (5.9%)	12,783 (9.9%)
V. myrtillus	38,891	6,487	36,850 (94.8%)	2,041 (5.2%)	1,347 (3.5%)

Table 2.3 Summary statistics of Orthologs analysis for the 16 Ericales species and *Arabidopsis thaliana*.

2.3.5 The manzanita genome is enriched with genes involved in terpenoid biosynthesis and metabolism

To identify *A. glauc*a gene content that might reflect adaptation to the CFP climate, we applied GO enrichment analysis (Figure 2.6; Appendix S2.4) and correlated resulting functions with factors known to be involved in drought- or fire-adaptation. We identified 1,376 orthogroups (7,217 genes) in which the number of manzanita genes appears to be larger than the other species. A GO enrichment analysis on genes from these expanded gene families showed that many GO terms were relevant to constitutive functions such as mitotic cell cycle or actin polymerization that are difficult to correlate to specific adaptation (Appendix S2.4). In addition, we identified 997 orthogroups that only contain *A. glauca* genes (4,931 genes) and used these as an input for GO enrichment analysis. We identified four enriched GO terms that were related to terpenoid and diterpenoid biosynthesis and metabolism. Terpenoids have been shown to play a role in fire and drought adaptation in MTC plants, thus their over-representation suggests they might also be implicated in the adaptation of manzanitas to the CFP environment.

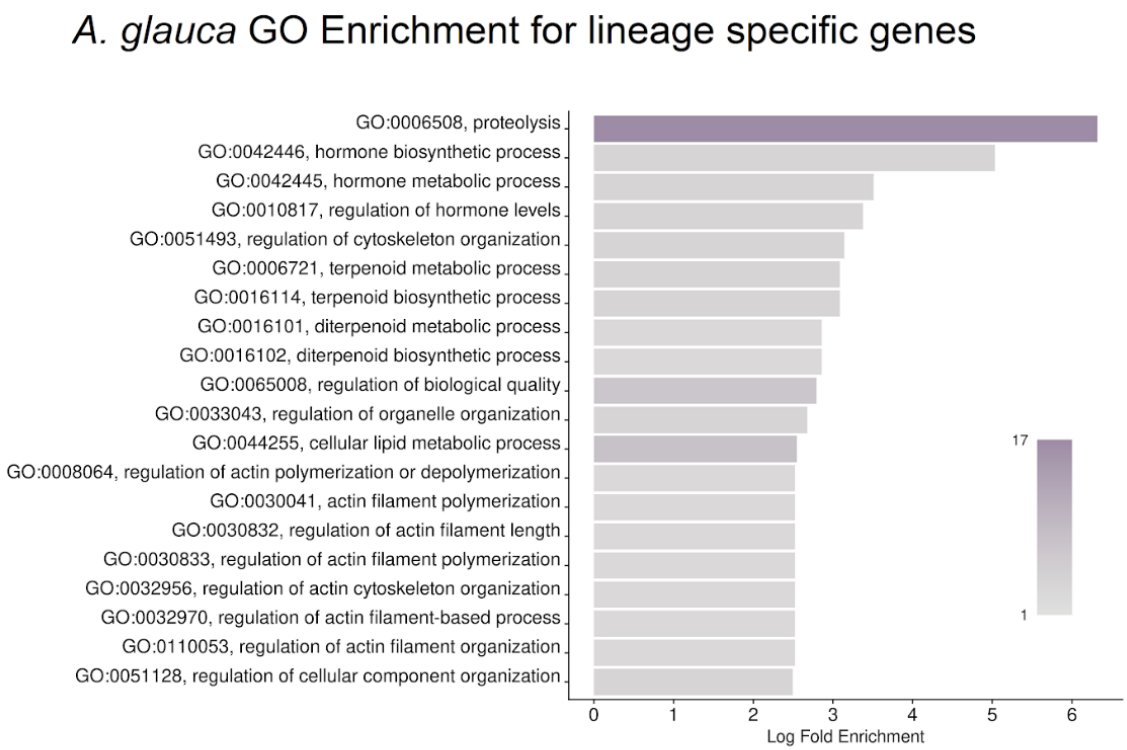


Figure 2.6 GO term enrichment for species-specific genes of *A. glauca* shows over-representation of terpenoid- and diterpenoid-pathway genes. GO term names are listed on the y axis. Bar colors correspond to the number of genes assigned to the given GO term and the color scale is shown in the lower right of each plot.

2.3.6 The manzanita genome shows no evidence of gene family expansion in karrikin signaling pathway gene families

To test whether gene family expansion and species-specific duplication events of karrikin-related genes, which are implicated in fire-mediated seed germination, are

relevant to manzanita's fire-adaptation, we investigated the change in copy number of KAI2, MAX2, SMAX2 and DLK2 genes in the Ericaceae clade. We retrieved four orthogroups containing KAI2, MAX2, SMAX2 and DLK2 respectively and counted the number of orthologous genes for every plant species within each orthogroup. For these four orthogroups, we found that *A. glauca* did not contain more gene copies than the other Ericaceae species, which are not known to be fire-adapted. This indicates that these genes have not undergone gene family expansion in manzanitas.

In addition, we constructed gene trees for each of these four orthogroups to visualize species-specific duplication events across the Ericaceae clade. The topology of the KAI2, MAX2, SMAX2 and DLK2 gene trees (Appendix S2.5) differed from each other. No species-specific gene duplication events were observed in DLK2 phylogeny. In the other three gene trees, species-specific gene duplication events were observed not just in *A. glauca*, but in multiple Ericaceae species. Therefore, we conclude that gene family expansion and lineage-specific gene duplication events of karrikin-related genes are not relevant to manzanitas's drought-tolerance and fire-adaptation.

2.4 Conclusion

In this study, we annotated the genome and found 40,204 protein-coding genes and 453 tRNAs in this widespread and ecologically important manzanita species, *A. glauca*. This annotation will facilitate the identification and interpretation of genetic variants in phylogenetic and population genetics studies of manzanitas, and provide an important tool for their conservation management, especially the many rare and endangered species. In addition, it will serve as a valuable reference for studying the diversification and evolution of a highly complex and diverse fire- and drought-adapted

woody plant genus, which may shed light on aspects of diversification in other complex groups with similar adaptations to the Mediterranean climate of the CFP.

Our comparative genomic analysis provides insight into the evolution and adaptation of manzanitas. Synteny analysis confirmed that Ericaceae do not share WGD found in the Theaceae, and Ebenaceae, supporting the hypothesis that there are multiple independent WGD events in the Ericales clade. Drought tolerance and fire resilience are ecological hallmarks of manzanitas, and these traits are of growing importance in the context of the increased drought and fire frequency and intensity that are occurring in CFP as a result of climate change. Our analyses reveal that the *A. glauca* genome is enriched with genes that are related to terpenoid biosynthesis and metabolism. Terpenoids have been implicated in the response of some Mediterranean plants to MTC, and our findings suggest that they potentially contribute to the response of manzanitas to the MTC of the California Floristic Province. Manzanita species occupy diverse habitats in the CFP, therefore our results are an initial step towards understanding the role of these secondary metabolites in manzanita adaptation. Further investigation into terpenoid-pathway gene family expansion in other manzanita species, and terpenoid composition in relationship to habitat, might provide support for the hypothesis that diversification of this pathway is important in the adaptation of manzanitas to the CFP environment.

2.5 References

Almagro Armenteros, José Juan, Konstantinos D Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019. 'SignalP 5.0 improves signal peptide predictions using deep neural networks', *Nature biotechnology*, 37: 420-23.

Baldwin, Bruce G. 2014. 'Origins of Plant Diversity in the California Floristic Province', *Annu. Rev. Ecol. Evol. Syst.*, 45: 347-69.

Baldwin, Bruce G., Douglas H. Goldman, David J. Keil, Robert Patterson, and Thomas J. Rosatti. 2012. *The Jepson Manual: Vascular Plants of California* (University of California Press).

Bolsinger, Charles L. 1989. *Shrubs of California's Chaparral, Timberland, and Woodland: Area, Ownership, and Stand Characteristics* (U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station).

Boutet, Emmanuel, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. 2016. 'UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view.' in, *Plant Bioinformatics* (Springer).

Brandies, Parice, Emma Peel, Carolyn J Hogg, and Katherine Belov. 2019. 'The value of reference genomes in the conservation of threatened species', *Genes*, 10: 846.

Burge, Dylan O., V. Thomas Parker, Margaret Mulligan, and César García Valderamma. 2018. 'Conservation Genetics of the Endangered Del Mar Manzanita (Arctostaphylos glandulosa subsp. Crassifolia) Based On Rad Sequencing Data', *Madroño*, 65: 117-30.

Burge, Dylan O., James H. Thorne, Susan P. Harrison, Bart C. O'Brien, Jon P. Rebman, James R. Shevock, Edward R. Alverson, Linda K. Hardison, José Delgadillo RodrÍguez, Steven A. Junak, and Others. 2016. 'Plant diversity and endemism in the California Floristic Province', *Madroño*: 3-206.

Cocker, Jonathan M, Jonathan Wright, Jinhong Li, David Swarbreck, Sarah Dyer, Mario Caccamo, and Philip M Gilmartin. 2018. 'Primula vulgaris (primrose) genome assembly, annotation and gene expression, with comparative genomics on the heterostyly supergene', *Scientific reports*, 8: 1-13.

Colle, Marivi, Courtney P Leisner, Ching Man Wai, Shujun Ou, Kevin A Bird, Jie Wang, Jennifer H Wisecaver, Alan E Yocca, Elizabeth I Alger, and Haibao Tang. 2019. 'Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry', *GigaScience*, 8: giz012.

Conn, Caitlin E., and David C. Nelson. 2015. 'Evidence that KARRIKIN-INSENSITIVE2 (KAI2) Receptors may Perceive an Unknown Signal that is not Karrikin or Strigolactone', *Front. Plant Sci.*, 6: 1219.

Eilbeck, Karen, Barry Moore, Carson Holt, and Mark Yandell. 2009. 'Quantitative measures for the management and comparison of annotated genomes', *BMC bioinformatics*, 10: 1-15.

Emms, David M, and Steven Kelly. 2019. 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome biology*, 20: 1-14.

Emms, David M., and Steven Kelly. 2015. 'OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy', *Genome Biol.*, 16: 157.

Finn, Robert D, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, and Jaina Mistry. 2014. 'Pfam: the protein families database', *Nucleic acids research*, 42: D222-D30.

Flematti, Gavin R, Emilio L Ghisalberti, Kingsley W Dixon, and Robert D Trengove. 2009. 'Identification of alkyl substituted 2 H-furo [2, 3-c] pyran-2-ones as germination stimulants present in smoke', *Journal of Agricultural and Food Chemistry*, 57: 9475-80.

Flematti, Gavin R., Emilio L. Ghisalberti, Kingsley W. Dixon, and Robert D. Trengove. 2004. 'A compound from smoke that promotes seed germination', *Science*, 305: 977.

Flynn, Jullien M, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G Clark, Cédric Feschotte, and Arian F Smit. 2020. 'RepeatModeler2 for automated genomic discovery of transposable element families', *Proceedings of the National Academy of Sciences*, 117: 9451-57.

Gluesenkamp, Daniel, Michael Chassé, Mark Frey, V Thomas Parker, M Vasey, and Betty Young. 2011. 'Back from the brink: A second chance at discovery and conservation of the Franciscan Manzanita', *Fremontia*, 38: 3-17.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, and Qiandong Zeng. 2011. 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature biotechnology*, 29: 644-52.

Gupta, Vikas, April D Estrada, Ivory Blakley, Rob Reid, Ketan Patel, Mason D Meyer, Stig Uggerhøj Andersen, Allan F Brown, Mary Ann Lila, and Ann E Loraine. 2015. 'RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive

compounds, and stage-specific alternative splicing', *GigaScience*, 4: s13742-015-0046-9.

Haas, Brian J., Arthur L. Delcher, Stephen M. Mount, Jennifer R. Wortman, Roger K. Smith, Jr., Linda I. Hannick, Rama Maiti, Catherine M. Ronning, Douglas B. Rusch, Christopher D. Town, Steven L. Salzberg, and Owen White. 2003. 'Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies', *Nucleic Acids Res.*, 31: 5654-66.

Haas, Brian J., Steven L. Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E. Allen, Joshua Orvis, Owen White, C. Robin Buell, and Jennifer R. Wortman. 2008. 'Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments', *Genome Biol.*, 9: R7.

Halsey, R. W., and J. E. Keeley. 2016a. 'Conservation Issues: California Chaparral.' in, *Reference Module in Earth Systems and Environmental Sciences* (Elsevier).

Halsey, Richard W, and Jon E Keeley. 2016b. 'Conservation issues: California chaparral'.

Huang, Le, Han Zhang, Peizhi Wu, Sarah Entwistle, Xueqiong Li, Tanner Yohe, Haidong Yi, Zhenglu Yang, and Yanbin Yin. 2018. 'dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation', *Nucleic acids research*, 46: D516-D21.

Huang, Shengxiong, Jian Ding, Dejing Deng, Wei Tang, Honghe Sun, Dongyuan Liu, Lei Zhang, Xiangli Niu, Xia Zhang, Meng Meng, Jinde Yu, Jia Liu, Yi Han, Wei Shi, Danfeng Zhang, Shuqing Cao, Zhaojun Wei, Yongliang Cui, Yanhua Xia, Huaping Zeng, Kan Bao, Lin Lin, Ya Min, Hua Zhang, Min Miao, Xiaofeng Tang, Yunye Zhu, Yuan Sui, Guangwei Li, Hanju Sun, Junyang Yue, Jiaqi Sun, Fangfang Liu, Liangqiang Zhou, Lin Lei, Xiaoqin Zheng, Ming Liu, Long Huang, Jun Song, Chunhua Xu, Jiewei Li, Kaiyu Ye, Silin Zhong, Bao-Rong Lu, Guanghua He, Fangming Xiao, Hui-Li Wang, Hongkun Zheng, Zhangjun Fei, and Yongsheng Liu. 2013. 'Draft genome of the kiwifruit Actinidia chinensis', *Nat. Commun.*, 4: 2640.

Huang, Yi, Merly Escalona, Glen Morrison, Mohan P. A. Marimuthu, Oanh Nguyen, Erin Toffelmier, H. Bradley Shaffer, and Amy Litt. 2021. 'Reference genome assembly of the big berry Manzanita (Arctostaphylos glauca)', *J. Hered.*

Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C Walter, Thomas Rattei, Daniel R Mende, Shinichi Sunagawa, and Michael Kuhn. 2016. 'eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences', *Nucleic acids research*, 44: D286-D93.

Jaillon, Olivier, Jean-Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Cassagrande, Nathalie Choisne, Sébastien Aubourg, Nicola Vitulo, and

Claire Jubin. 2007. 'The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla', *nature*, 449: 463-7.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. 2014. 'InterProScan 5: genome-scale protein function classification', *Bioinformatics*, 30: 1236-40.

Käll, Lukas, Anders Krogh, and Erik L. L. Sonnhammer. 2004. 'A combined transmembrane topology and signal peptide prediction method', *J. Mol. Biol.*, 338: 1027-36.

Kauffmann, Michael Edward, Tom Parker, Michael Vasey, and Jeff Bisbee. 2015. *Field Guide to Manzanitas* (Backcountry Press).

Keddy, PaµL A. 2017. *Plant Ecology* (Cambridge University Press).

Keeley, Jon E. 1991. 'Seed germination and life history syndromes in the California chaparral', *Bot. Rev.*, 57: 81-116.

Kiełbasa, Szymon M., Raymond Wan, Kengo Sato, PaµL Horton, and Martin C. Frith. 2011. 'Adaptive seeds tame genomic sequence comparison', *Genome Res.*, 21: 487-93.

Korf, Ian. 2004. 'Gene finding in novel genomes', *BMC bioinformatics*, 5: 59.

Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat. 2014. 'The carbohydrate-active enzymes database (CAZy) in 2013', *Nucleic Acids Res.*, 42: D490-5.

Lowe, T. M., and S. R. Eddy. 1997. 'tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence', *Nucleic Acids Res.*, 25: 955-64.

Ma, Hong, Yongbo Liu, Detuan Liu, Weibang Sun, Xiongfang Liu, Youming Wan, Xiujiao Zhang, Rengang Zhang, Quanzheng Yun, Jihua Wang, Zhenghong Li, and Yongpeng Ma. 2021. 'Chromosome-level genome assembly and population genetic analysis of a critically endangered rhododendron provide insights into its conservation', *Plant J.*, 107: 1533-45.

Majoros, W. H., M. Pertea, and S. L. Salzberg. 2004. 'TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders', *Bioinformatics*, 20: 2878-79.

Michael, Todd P., Florian Jupe, Felix Bemm, S. Timothy Motley, Justin P. Sandoval, Christa Lanz, Olivier Loudet, Detlef Weigel, and Joseph R. Ecker. 2018. 'High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell', *Nat. Commun.*, 9: 541.

Myers, N. 1990. 'The biodiversity challenge: expanded hot-spots analysis', *Environmentalist*, 10: 243-56.

Nelson, David C., Adrian Scaffidi, Elizabeth A. Dun, Mark T. Waters, Gavin R. Flematti, Kingsley W. Dixon, Christine A. Beveridge, Emilio L. Ghisalberti, and Steven M. Smith. 2011. 'F-box protein MAX2 has dual roles in karrikin and strigolactone signaling in Arabidopsis thaliana', *Proc. Natl. Acad. Sci. U. S. A.*, 108: 8897-902.

Nowak, Michael D., Giancarlo Russo, Ralph Schlapbach, Cuong Nguyen Huu, Michael Lenhard, and Elena Conti. 2015. 'The draft genome of Primula veris yields insights into the molecular basis of heterostyly', *Genome Biol.*, 16: 12.

Palmer, J., and J. E. Stajich. 2017. 'Funannotate: eukaryotic genome annotation pipeline', *Published online. https://github. com/nextgenusfs/funannotate*.

Parker, V. Thomas. 2007. 'Diversity and Evolution of Arctostaphylos and Ceanothus', *Fremontia*: 8.

Pilkington, Sarah M., Ross Crowhurst, Elena Hilario, Simona Nardozza, Lena Fraser, Yongyan Peng, Kularajathevan Gunaseelan, Robert Simpson, Jibran Tahir, Simon C. Deroles, Kerry Templeton, Zhiwei Luo, Marcus Davy, Canhong Cheng, Mark McNeilage, Davide Scaglione, Yifei Liu, Qiong Zhang, PaµL Datson, Nihal De Silva, Susan E. Gardiner, Heather Bassett, David Chagné, John McCallum, Helge Dzierzon, Cecilia Deng, Yen-Yi Wang, Lorna Barron, Kelvina Manako, Judith Bowen, Toshi M. Foster, Zoe A. Erridge, Heather Tiffin, Chethi N. Waite, Kevin M. Davies, Ella P. Grierson, William A. Laing, Rebecca Kirk, Xiuyin Chen, Marion Wood, Mirco Montefiori, David A. Brummell, Kathy E. Schwinn, Andrew Catanach, Christina Fullerton, Dawei Li, Sathiyamoorthy Meiyalaghan, Niels Nieuwenhuizen, Nicola Read, Roneel Prakash, Don Hunter, Huaibi Zhang, Marian McKenzie, Mareike Knäbel, Alastair Harris, Andrew C. Allan, Andrew Gleave, Angela Chen, Bart J. Janssen, Blue Plunkett, Charles Ampomah-Dwamena, Charlotte Voogd, Davin Leif, Declan Lafferty, Edwige J. F. Souleyre, Erika Varkonyi-Gasic, Francesco Gambi, Jenny Hanley, Jia-Long Yao, Joey Cheung, Karine M. David, Ben Warren, Ken Marsh, Kimberley C. Snowden, Kui Lin-Wang, Lara Brian, Marcela Martinez-Sanchez, Mindy Wang, Nadeesha Ileperuma, Nikolai Macnee, Robert Campin, Peter McAtee, Revel S. M. Drummond, Richard V. Espley, Hilary S. Ireland, Rongmei Wu, Ross G. Atkinson, Sakuntala Karunairetnam, Sean Bulley, Shayhan Chunkath, Zac Hanley, Roy Storey, Amali H. Thrimawithana, Susan Thomson, Charles David, Raffaele Testolin, Hongwen Huang, Roger P. Hellens, and Robert J. Schaffer. 2018. 'A manually annotated Actinidia chinensis var. chinensis (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants', *BMC Genomics*, 19: 257.

Quinlan, Aaron R. 2014. 'BEDTools: The Swiss-army tool for genome feature analysis', *Curr. Protoc. Bioinformatics*, 47: 11.12.1-34.

Quinlan, Aaron R., and Ira M. Hall. 2010. 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26: 841-42.

Rajewski, Alex, Derreck Carter-House, Jason Stajich, and Amy Litt. 2021. 'Datura genome reveals duplications of psychoactive alkaloid biosynthetic genes and high mutation rate following tissue culture', *BMC Genomics*, 22: 201.

Raven, Peter H., and Daniel I. Axelrod. 1978. *Origin and Relationships of the California Flora* (University of California Press).

Rawlings, Neil D., Alan J. Barrett, and Alex Bateman. 2014. 'Using the MEROPS Database for Proteolytic Enzymes and Their Inhibitors and Substrates', *Curr. Protoc. Bioinformatics*, 48: 1.25.1-33.

Revell, Liam J. 2012. 'phytools: an R package for phylogenetic comparative biology (and other things)', *Methods Ecol. Evol.*, 3: 217-23.

Roelofs, Dick, Arthur Zwaenepoel, Tom Sistermans, Joey Nap, Andries A. Kampfraath, Yves Van de Peer, Jacintha Ellers, and Ken Kraaijeveld. 2020. 'Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution', *BMC Biol.*, 18: 57.

Rose, Jeffrey P., Thomas J. Kleist, Stefan D. Löfstrand, Bryan T. Drew, Jürg Schönenberger, and Kenneth J. Sytsma. 2018. 'Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections', *Mol. Phylogenet. Evol.*, 122: 59-79.

Rundel, Philip W., Mary T. K. Arroyo, Richard M. Cowling, Jon E. Keeley, Byron B. Lamont, Juli G. Pausas, and Pablo Vargas. 2018. 'Fire and Plant Diversification in Mediterranean-Climate Regions', *Front. Plant Sci.*, 9: 851.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2019. 'BUSCO: Assessing Genome Assembly and Annotation Completeness', *Methods Mol. Biol.*, 1962: 227-45.

Shi, Tao, Hongwen Huang, and Michael S. Barker. 2010. 'Ancient genome duplications during the evolution of kiwifruit (Actinidia) and related Ericales', *Ann. Bot.*, 106: 497-504.

Shirasawa, Kenta, Nobuo Kobayashi, Akira Nakatsuka, Hideya Ohta, and Sachiko Isobe. 2021. 'Whole-genome sequencing and analysis of two azaleas, Rhododendron ripense and Rhododendron kiyosumense', *DNA Res.*, 28.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31: 3210-12.

Smit, A. F. A., R. Hubley, and P. Green. 2015. 'RepeatMasker Open-4.0. 2013--2015'.

Soza, Valerie L., Dale Lindsley, Adam Waalkes, Elizabeth Ramage, Rupali P. Patwardhan, Joshua N. Burton, Andrew Adey, Akash Kumar, Ruolan Qiu, Jay Shendure, and Benjamin Hall. 2019. 'The Rhododendron Genome and Chromosomal Organization Provide Insight into Shared Whole-Genome Duplications across the Heath Family (Ericaceae)', *Genome Biol. Evol.*, 11: 3353-71.

Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. 2006. 'AUGUSTUS: ab initio prediction of alternative transcripts', *Nucleic Acids Res.*, 34: W435-9.

Suo, Yujing, Peng Sun, Huihui Cheng, Weijuan Han, Songfeng Diao, Huawei Li, Yini Mai, Xing Zhao, Fangdong Li, and Jianmin Fu. 2020. 'A high-quality chromosomal genome assembly of Diospyros oleifera Cheng', *GigaScience*, 9.

Tang, Haibao, Xiyin Wang, John E. Bowers, Ray Ming, MaqsudμL Alam, and Andrew H. Paterson. 2008. 'Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps', *Genome Res.*, 18: 1944-54.

Tang, Wei, Xuepeng Sun, Junyang Yue, Xiaofeng Tang, Chen Jiao, Ying Yang, Xiangli Niu, Min Miao, Danfeng Zhang, Shengxiong Huang, Wei Shi, Mingzhang Li, Congbing Fang, Zhangjun Fei, and Yongsheng Liu. 2019. 'Chromosome-scale genome assembly of kiwifruit Actinidia eriantha with single-molecule sequencing and chromatin interaction mapping', *GigaScience*, 8.

Ter-Hovhannisyan, Vardges, Alexandre Lomsadze, Yury O. Chernoff, and Mark Borodovsky. 2008. 'Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training', *Genome Res.*, 18: 1979-90.

Van Staden, Johannes, Shane G. Sparg, Manoj G. Kulkarni, and Marnie E. Light. 2006. 'Post-germination effects of the smoke-derived compound 3-methyl-2H-furo [2, 3-c] pyran-2-one, and its potential as a preconditioning agent', *Field Crops Res.*, 98: 98-105.

Wang, Xiuyun, Yuan Gao, Xiaopei Wu, Xiaohui Wen, Danqing Li, Hong Zhou, Zheng Li, Bing Liu, Jianfen Wei, Fei Chen, Feng Chen, Chengjun Zhang, Liangsheng Zhang, and Yiping Xia. 2021. 'High-quality evergreen azalea genome reveals tandem duplication-facilitated low-altitude adaptability and floral scent evolution', *Plant Biotechnol. J.*, 19: 2544-60.

Wells, Philip V. 1968. 'New taxa, combinations, and chromosome numbers in Arctostaphylos (Ericaceae)', *Madroño*, 19: 193-210.

Wu, Chen, Cecilia Deng, Elena Hilario, Nick W. Albert, Declan Lafferty, Ella R. P. Grierson, Blue J. Plunkett, Caitlin Elborough, Ali Saei, Catrin S. Günther, Hilary Ireland, Alan Yocca, Patrick P. Edger, Laura Jaakola, Katja Karppinen, Adrian

Grande, Ritva Kylli, Veli-Pekka Lehtola, Andrew C. Allan, Richard V. Espley, and David Chagné. 2022. 'A chromosome-scale assembly of the bilberry genome identifies a complex locus controlling berry anthocyanin composition', *Mol. Ecol. Resour.*, 22: 345-60.

Wu, Haolin, Tao Ma, Minghui Kang, Fandi Ai, Junlin Zhang, Guanyong Dong, and Jianquan Liu. 2019. 'A high-quality Actinidia chinensis (kiwifruit) genome', *Hortic Res*, 6: 117.

Xia, Enhua, Wei Tong, Yan Hou, Yanlin An, Linbo Chen, Qiong Wu, Yunlong Liu, Jie Yu, Fangdong Li, Ruopei Li, Penghui Li, Huijuan Zhao, Ruoheng Ge, Jin Huang, Ali Inayat Mallano, Yanrui Zhang, Shengrui Liu, Weiwei Deng, Chuankui Song, Zhaoliang Zhang, Jian Zhao, Shu Wei, Zhengzhu Zhang, Tao Xia, Chaoling Wei, and Xiaochun Wan. 2020. 'The Reference Genome of Tea Plant and Resequencing of 81 Diverse Accessions Provide Insights into Its Genome Evolution and Adaptation', *Mol. Plant*, 13: 1013-26.

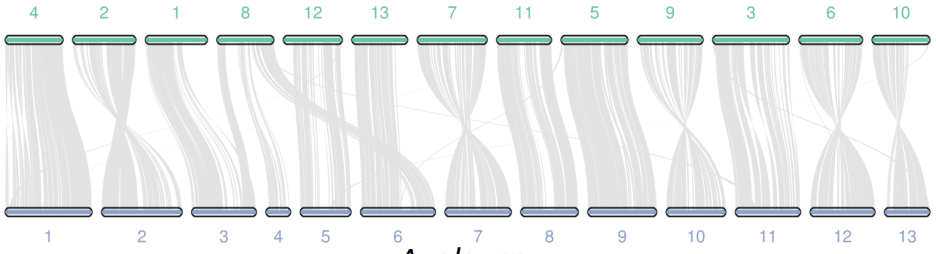
Yang, Fu-Sheng, Shuai Nie, Hui Liu, Tian-Le Shi, Xue-Chan Tian, Shan-Shan Zhou, Yu-Tao Bao, Kai-Hua Jia, Jing-Fang Guo, Wei Zhao, Na An, Ren-Gang Zhang, Quan-Zheng Yun, Xin-Zhu Wang, Chanaka Mannapperuma, Ilga Porth, Yousry Aly El-Kassaby, Nathaniel Robert Street, Xiao-Ru Wang, Yves Van de Peer, and Jian-Feng Mao. 2020. 'Chromosome-level genome assembly of a parent species of widely cultivated azaleas', *Nat. Commun.*, 11: 5269.

Zhang, Lu, Pengwei Xu, Yanfei Cai, Lulin Ma, Shifeng Li, Shufa Li, Weijia Xie, Jie Song, Lvchun Peng, Huijun Yan, Ling Zou, Yongpeng Ma, Chengjun Zhang, Qiang Gao, and Jihua Wang. 2017. 'The draft genome assembly of Rhododendron delavayi Franch. var. delavayi', *GigaScience*, 6: 1-11.

2.6 Appendix

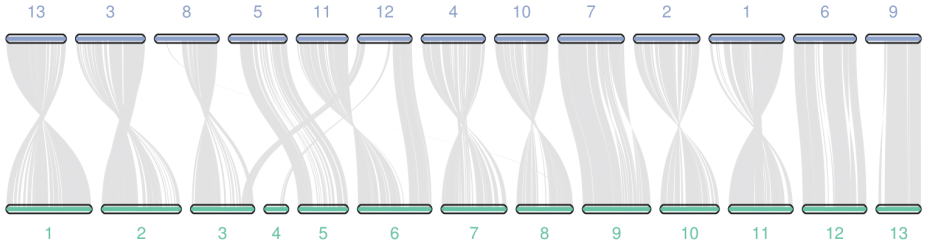
Appendix S2.1: Pseudo-chromosome-scale syntenic relationship between *A. glauca* and four other Ericaceae species suggests that the chromosomal organization of *Arctostaphylos* is more similar to *Rhododendron* than *Vaccinium*. The horizontal colored bars represent the chromosomes of each species. The gray lines connecting pseudo-chromosomes represent syntenic blocks between species. The greater number of such between the pseudo-chromosomes of *A. glauca* and the *Rhododendron* species, than between *A. glauca* and the *Vaccinium* species, indicates greater synteny with the *Rhododendron* species.

(a) *R. ovatum*



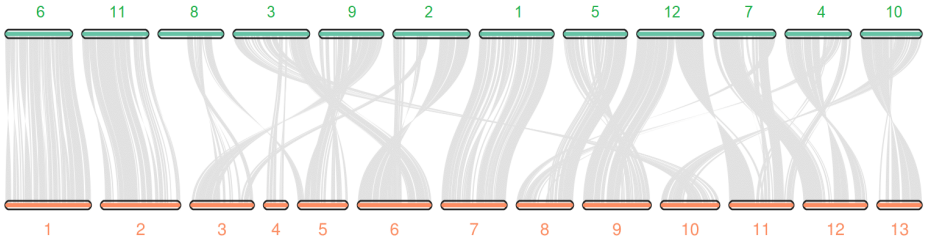
A. glauca

(b) *R. simsii*



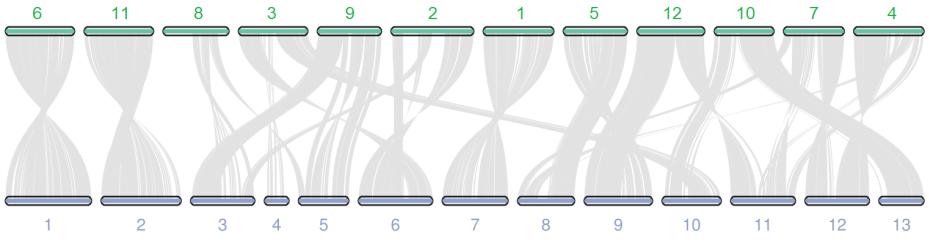
A. glauca

(c) *V. corymbosum*



A. glauca

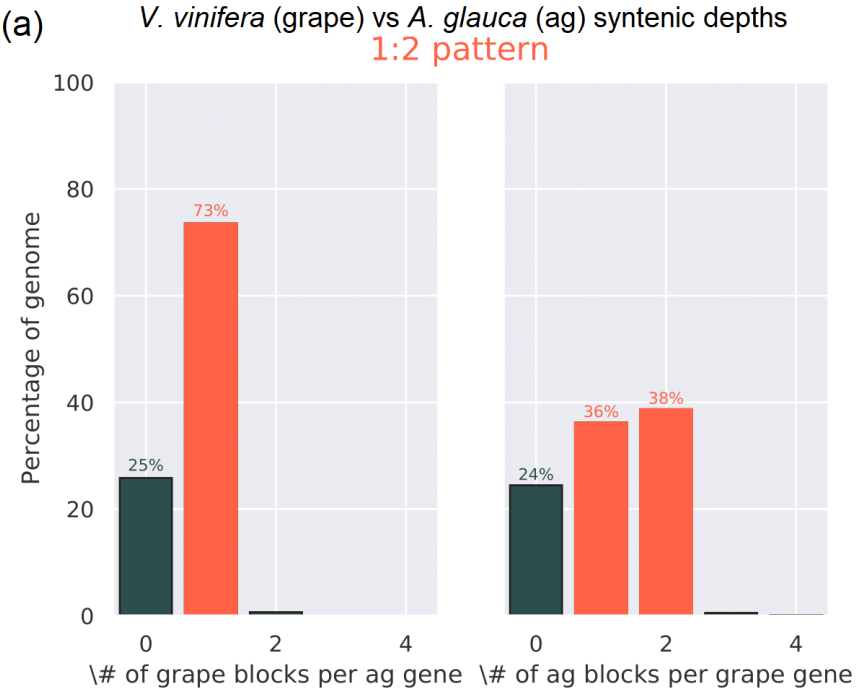
(d) *V. myrtillus*



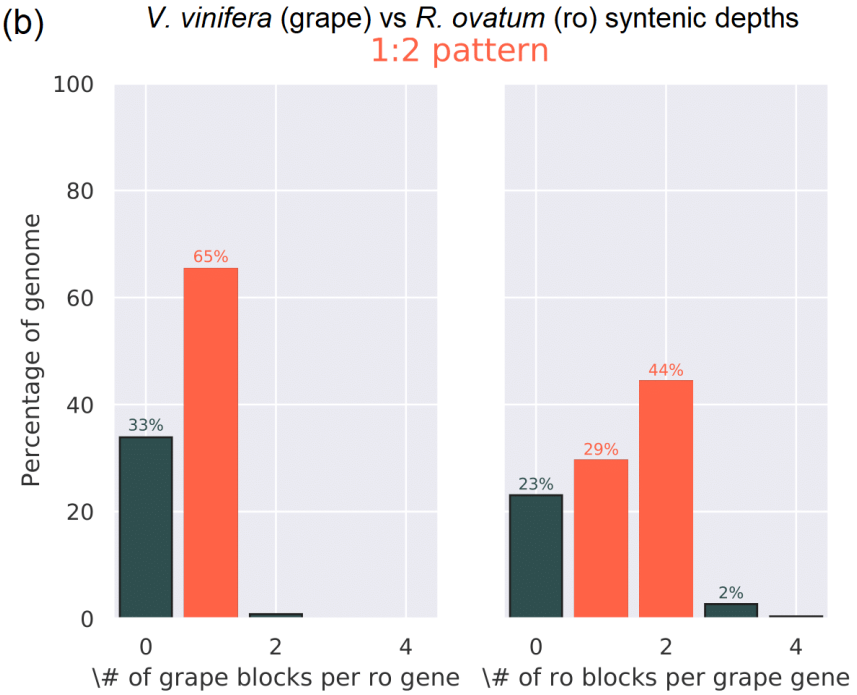
A. glauca

Appendix S2: The syntenic depth ratio between *V. vinifera* (grape) and Ericales species suggests that all of the Ericales species have undergone at least one additional WGD event after WGT (a-e, g-h) with the exception of the *Actinidia eriantha* [ae], which has undergone two (f). Species included in this analysis belong to the Ericaceae (*Arctostaphylos glauca* [ag], *Rhododendron ovatum* [ro], *Rhododendron simsii* [rs], *Vaccinium corymbosum* [vc], *Vaccinium myrtillus* [vmy]), Theaceae (*Camellia sinensis* [cs]), Ebenaceae (*Diospyros oleifera* [do]), and Actinidiaceae (*Actinidia eriantha* [ae]). In the left-hand graph in each panel, *V. vinifera* is the reference genome, and in the right-hand graph, it is the query genome. The x axis is the synteny depth, which refers to the number of syntenic regions (blocks) identified in a reference genome for a given query gene. The y axis is the percentage of the query genome with query genes that are covered in 1-, 2-, to x -fold syntenic regions (blocks). The inferred synteny depth ratio, indicated in orange at the top of the plot, is based on the highest synteny depth with a sufficient percentage of the genome to be considered enough based on empirical information. The bars from 1 to this depth are shown in red. A synteny depth ratio of 1:2 indicates that species have undergone one additional WGD event (a-e, g-h) compared to *V. vinifera*. A synteny depth ratio of 1:4 indicates that *A. eriantha* has undergone two WGD events since its divergence from *V. vinifera* (f).

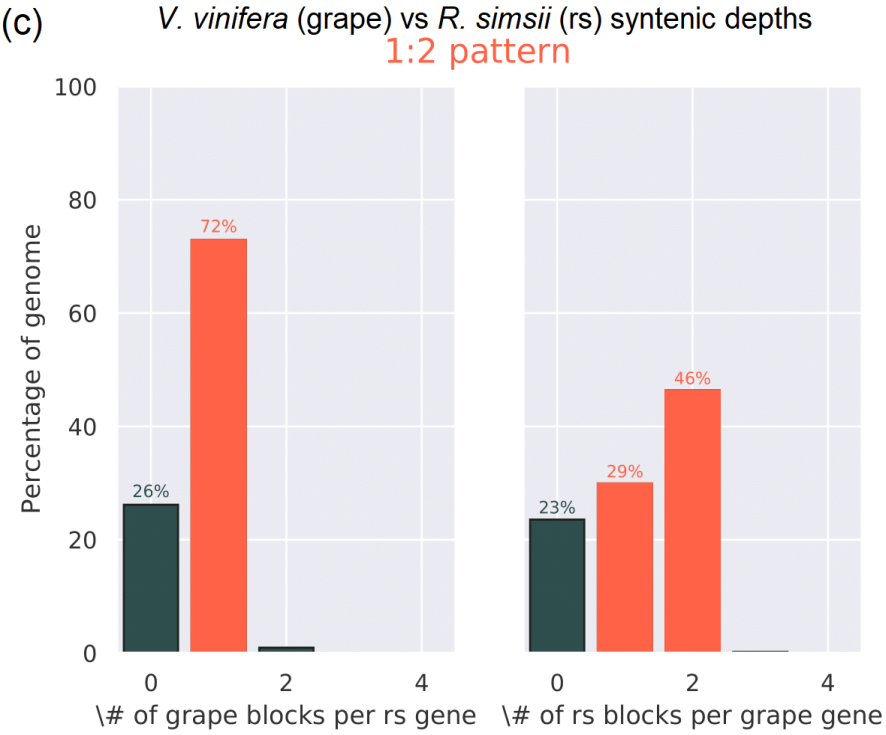
(a)


V. vinifera (grape) vs A. glauca (ag) syntenic depths
1:2 pattern
Percentage of genome
100
80
60
40
20
0
73%
25%
36%
38%
24%
0
2
4
\# of grape blocks per ag gene
\# of ag blocks per grape gene

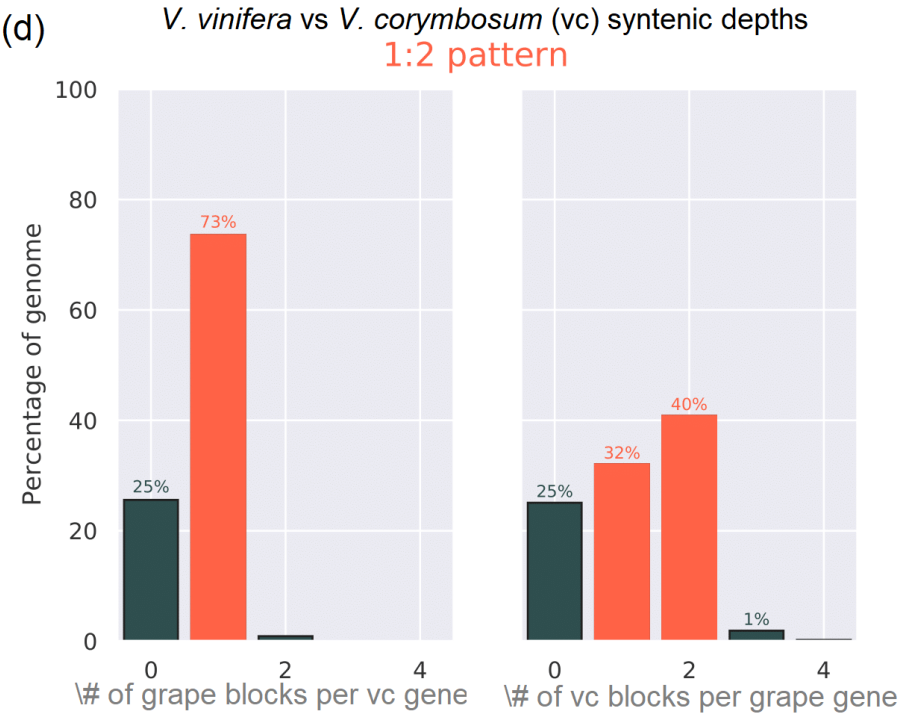
(b)


V. vinifera (grape) vs R. ovatum (ro) syntenic depths
1:2 pattern
Percentage of genome
100
80
60
40
20
0
65%
33%
44%
29%
23%
2%
0
2
4
\# of grape blocks per ro gene
\# of ro blocks per grape gene

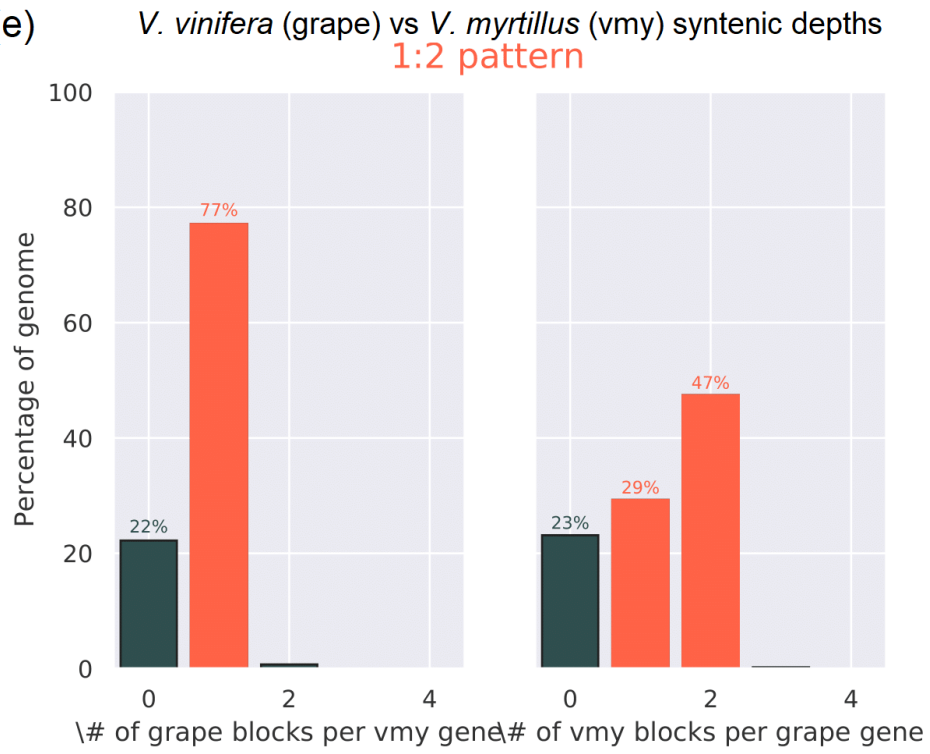
(c)


V. vinifera (grape) vs R. simsii (rs) syntenic depths
1:2 pattern
Percentage of genome
100
80
60
40
20
0
26%
72%
23%
29%
46%
0
2
4
\# of grape blocks per rs gene
\# of rs blocks per grape gene

(d)

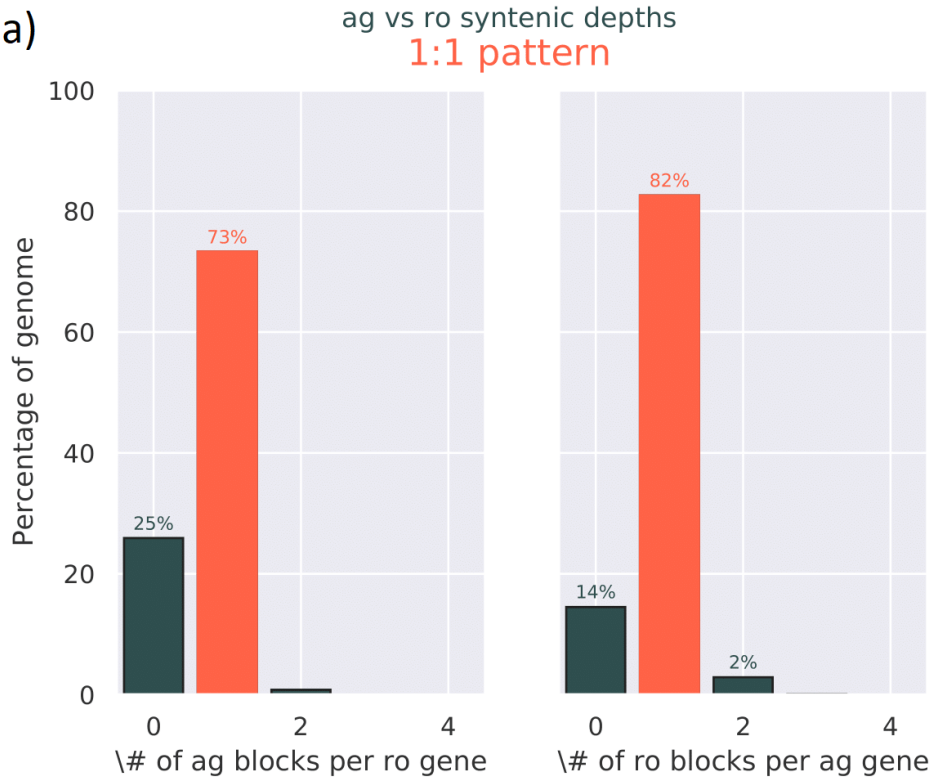

V. vinifera vs V. corymbosum (vc) syntenic depths
1:2 pattern
Percentage of genome
100
80
60
40
20
0
25%
73%
25%
32%
40%
1%
0
2
4
\# of grape blocks per vc gene
\# of vc blocks per grape gene

(e)

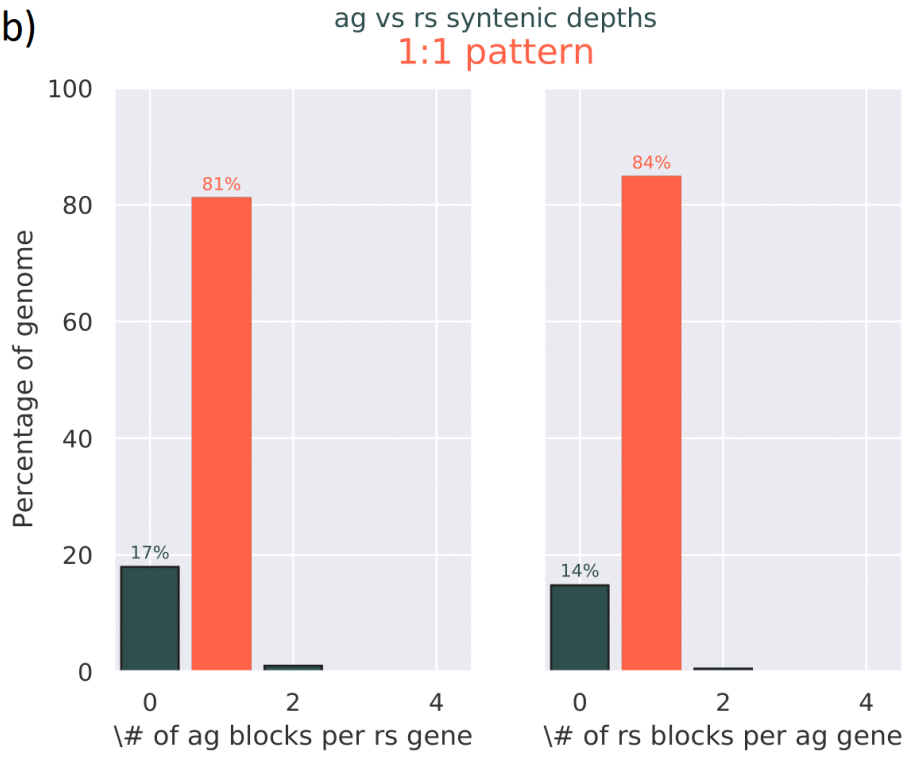

V. vinifera (grape) vs V. myrtillus (vmy) syntenic depths
1:2 pattern
Percentage of genome
100
80
60
40
20
0
22%
77%
23%
29%
47%
0
2
4
0
2
4
\# of grape blocks per vmy gene
\# of vmy blocks per grape gene

Appendix S2.3: The syntenic depth ratio between the Big Berry Manzanita and other Ericales (a-g) suggests at least four independent WGD events (h). Species included in this analysis belong to the Ericaceae (*Arctostaphylos glauca* [ag], *Rhododendron ovatum* [ro], *Rhododendron simsii* [rs]*, Vaccinium corymbosum* [vc]*, Vaccinium myrtillus* [vmy]), Theaceae (*Camellia sinensis* [cs]), Ebenaceae (*Diospyros oleifera* [do]), and Actinidiaceae (*Actinidia eriantha* [ae]). For (a-g), in the left-hand graph in each panel, *A. glauca* is the reference genome, and in the right-hand graph, it is the query genome. The x axis is the synteny depth, which refers to the number of syntenic regions (blocks) identified in a reference genome for a given query gene. The y axis is the percentage of the query genome with query genes that are covered in 1-, 2-, to x -fold syntenic regions (blocks). The inferred synteny depth ratio, indicated in orange at the top of the plot, is based on the highest synteny depth with a sufficient percentage of the genome to be considered enough. The bars from 1 to this depth are shown in red. The pattern of 1:1 between *A. glauca* and other Ericaceae species (a-d) means the *A. glauca* shares the same WGD history with the other Ericaceae members. The pattern of 1:2 for *A. glauca* and *A. eriantha* (g) means that *A. eriantha* has gone through an additional WGD. The 2:2 ratio suggests that since their divergence, the two species have gone through different WGD events (e-f). The (h) is the phylogeny of the multiple families of Ericales based on previous phylogenetic studies (Emms and Kelly 2015). The blue stars represent independent WGD events and the red triangle represents the WGT.

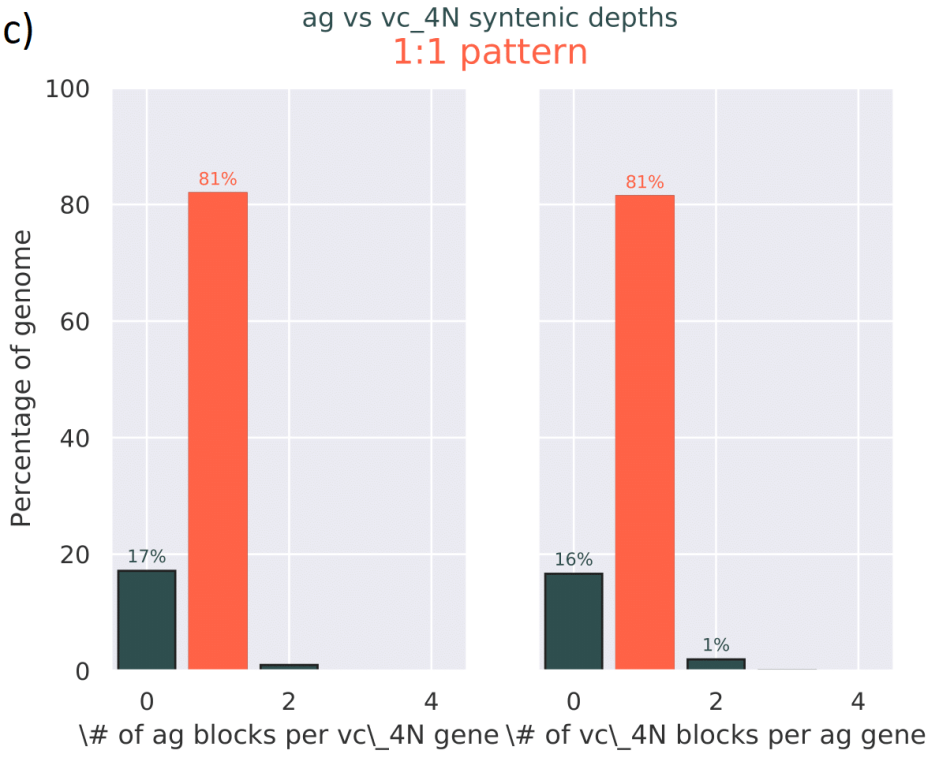
(a)


ag vs ro syntenic depths
1:1 pattern
Percentage of genome
100
80
60
40
20
0
25%
73%
14%
82%
2%
0
2
4
\# of ag blocks per ro gene
\# of ro blocks per ag gene

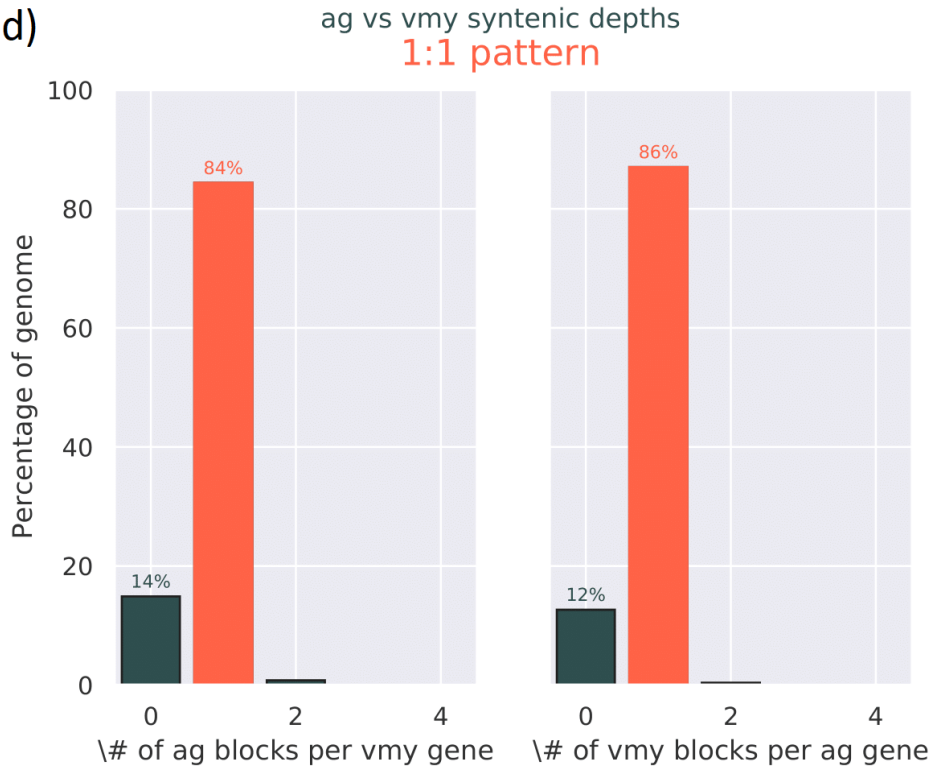
(b)


ag vs rs syntenic depths
1:1 pattern
Percentage of genome
100
80
60
40
20
0
17%
81%
14%
84%
0
2
4
\# of ag blocks per rs gene
\# of rs blocks per ag gene

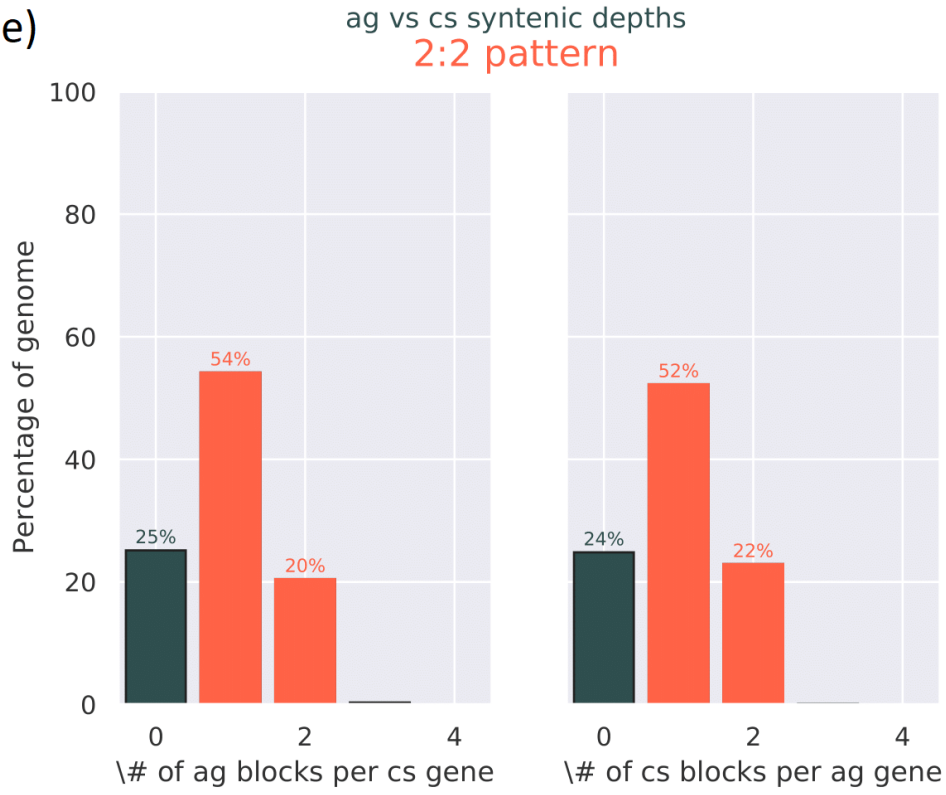
(c)


ag vs vc_4N syntenic depths
1:1 pattern
Percentage of genome
100
80
60
40
20
0
81%
17%
81%
16%
1%
0
2
4
0
2
4
\# of ag blocks per vc_4N gene
\# of vc_4N blocks per ag gene

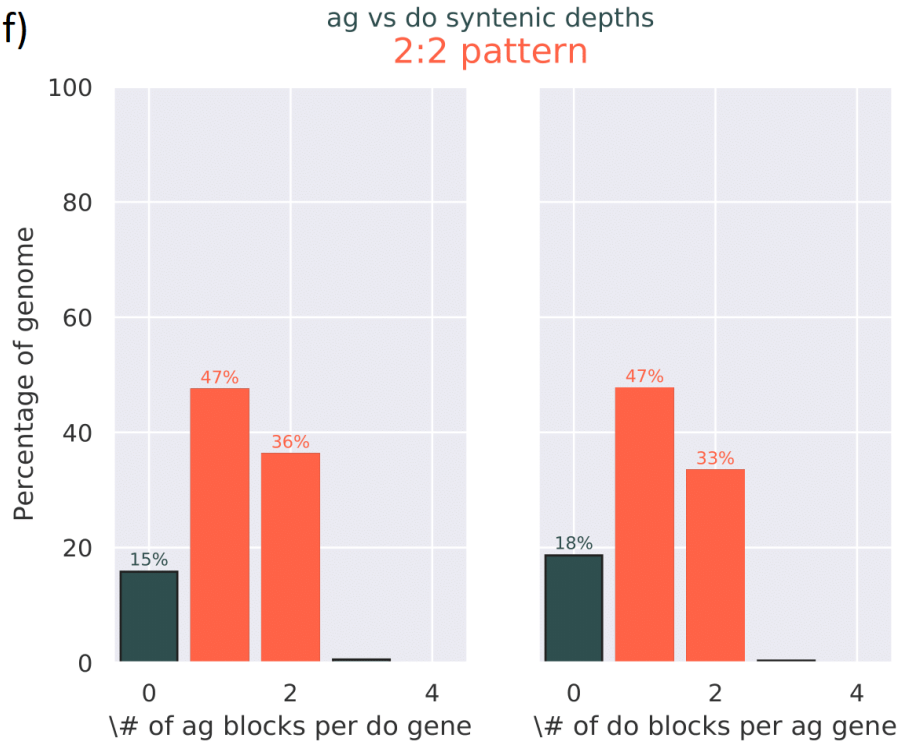
(d)


ag vs vmy syntenic depths
1:1 pattern
Percentage of genome
100
80
60
40
20
0
84%
14%
86%
12%
0
2
4
0
2
4
\# of ag blocks per vmy gene
\# of vmy blocks per ag gene

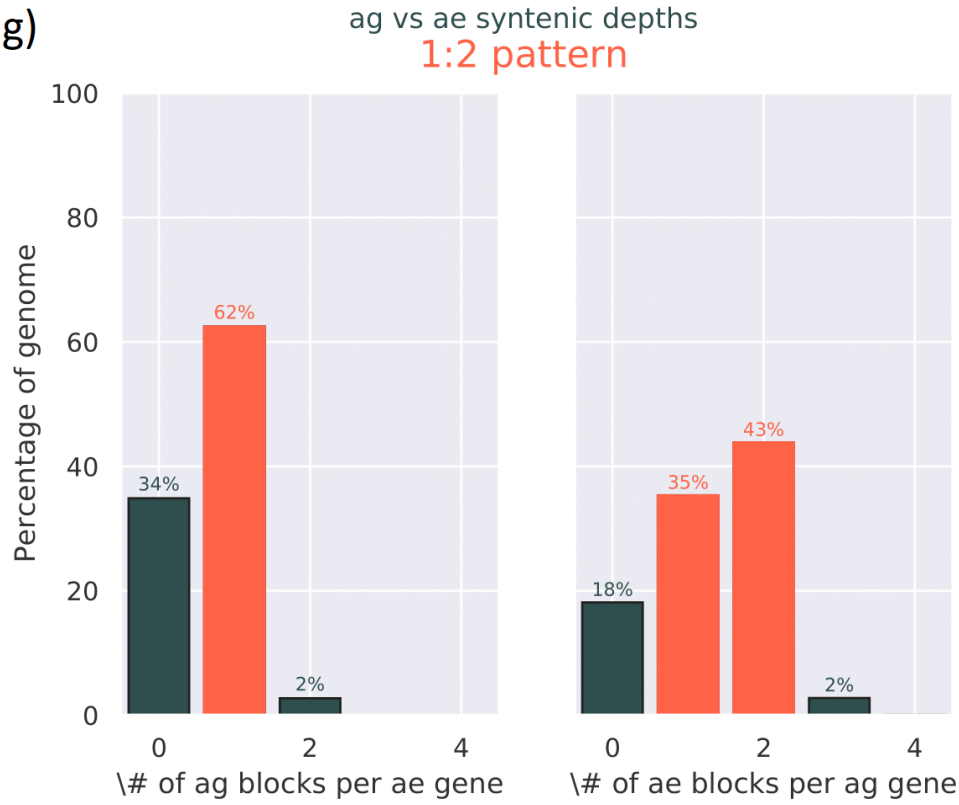
(e)


ag vs cs syntenic depths
2:2 pattern
Percentage of genome
100
80
60
40
20
0
25%
54%
20%
24%
52%
22%
0
2
4
\# of ag blocks per cs gene
\# of cs blocks per ag gene

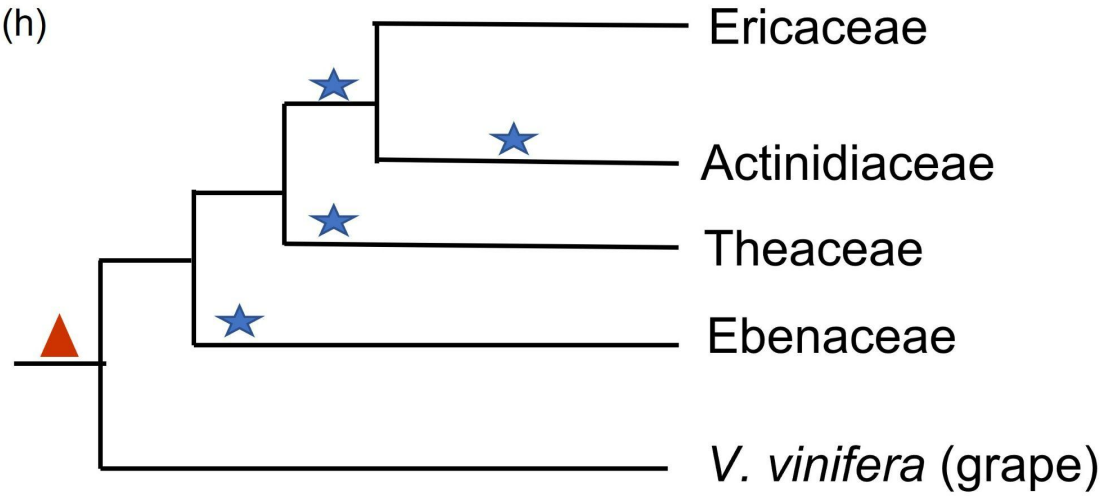
(f)


ag vs do syntenic depths
2:2 pattern
Percentage of genome
100
80
60
40
20
0
15%
47%
36%
18%
47%
33%
0
2
4
\# of ag blocks per do gene
\# of do blocks per ag gene

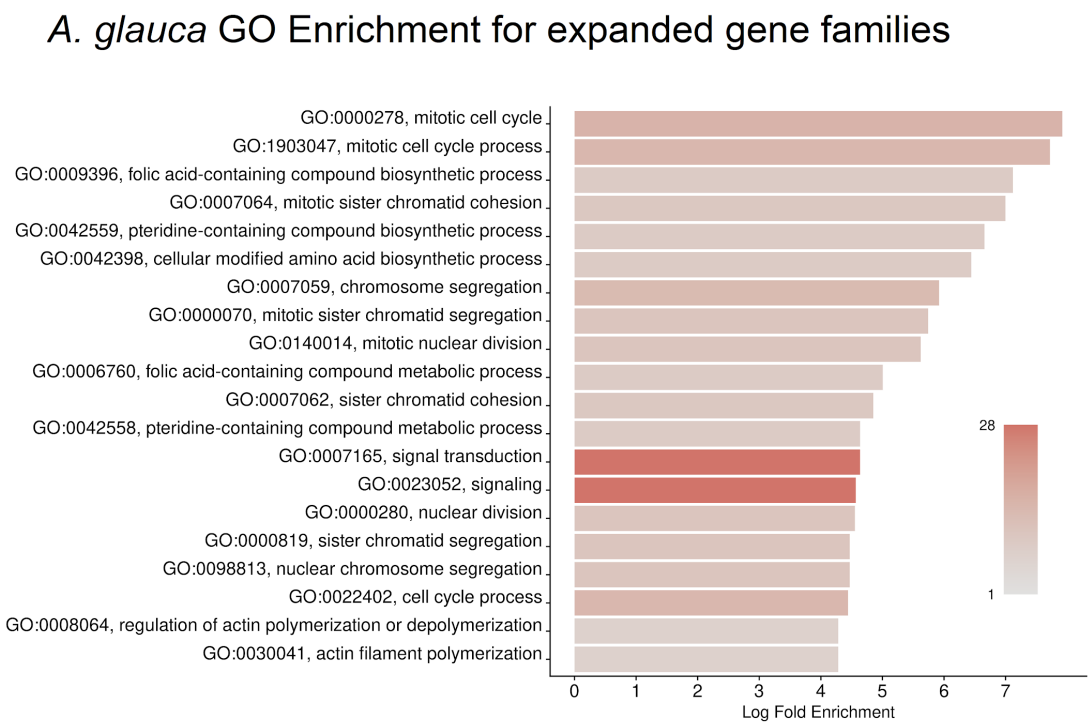
(g)


ag vs ae syntenic depths
1:2 pattern
Percentage of genome
100
80
60
40
20
0
34%
62%
2%
18%
35%
43%
2%
0
2
4
\# of ag blocks per ae gene
\# of ae blocks per ag gene

(h)

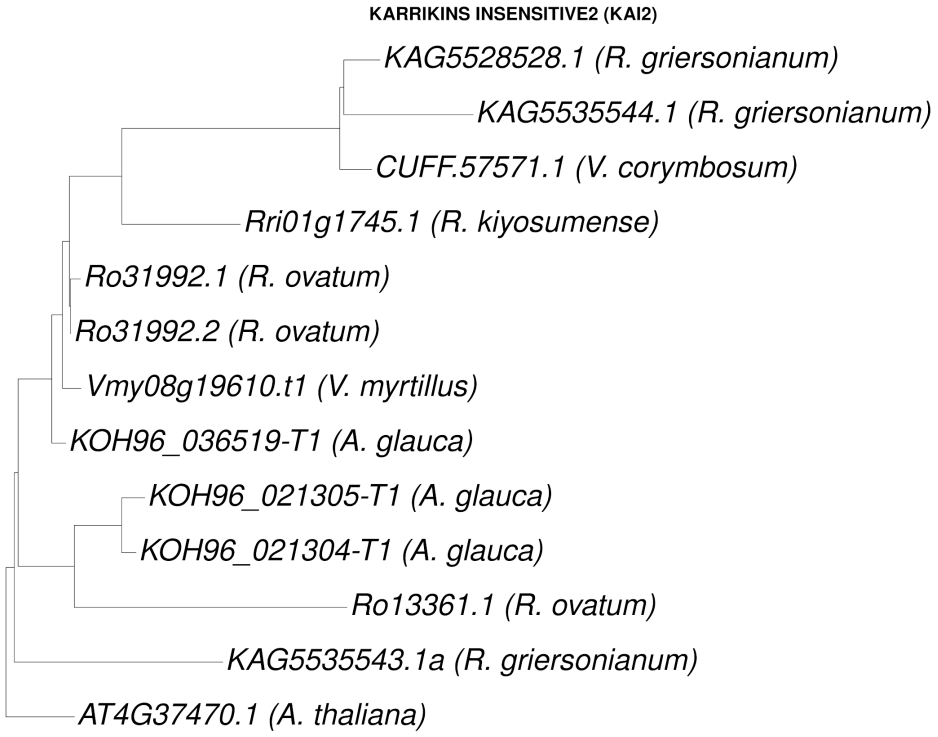

Ericaceae
Actinidiaceae
Theaceae
Ebenaceae
V. vinifera (grape)

Appendix S2.4: GO term enrichment for gene families that are enriched in *A. glauca* compared to other Ericales shows that many are relevant to constitutive functions. GO term names are listed on the y axis. Bar colors correspond to the number of genes assigned to the given GO term and the color scale is shown in the lower right of each plot.

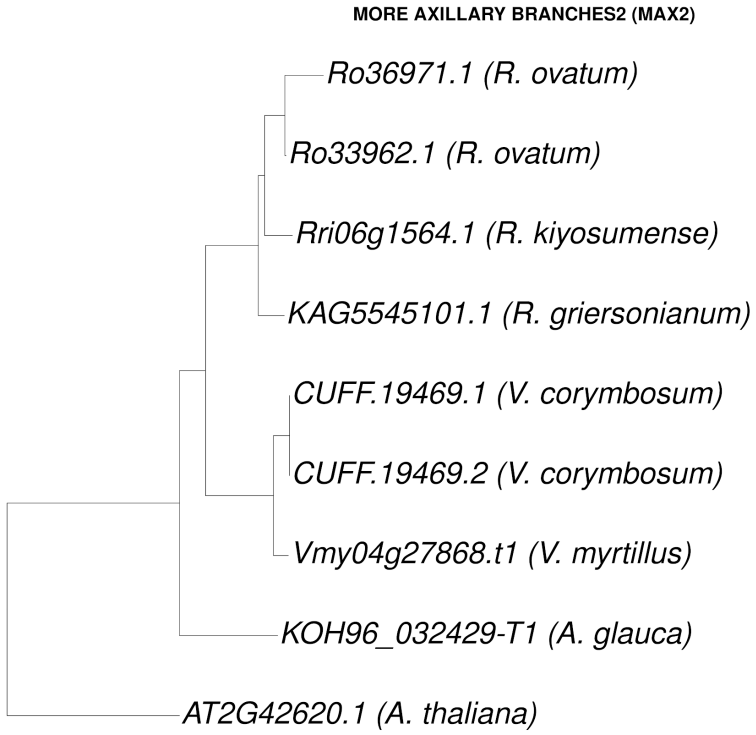


Appendix S2.5: Phylogenies of selected genes involved in the karrikin signaling pathway indicate that the genome of *A. glauca* does not contain more copies of (a) KAI2 , (b) MAX2, (c) SMAX2, and (d) DL2 genes. The tip labels are the gene ID from the annotation files provided by the relevant genome publications (Tabele 2.1), and followed by the species name in parentheses.

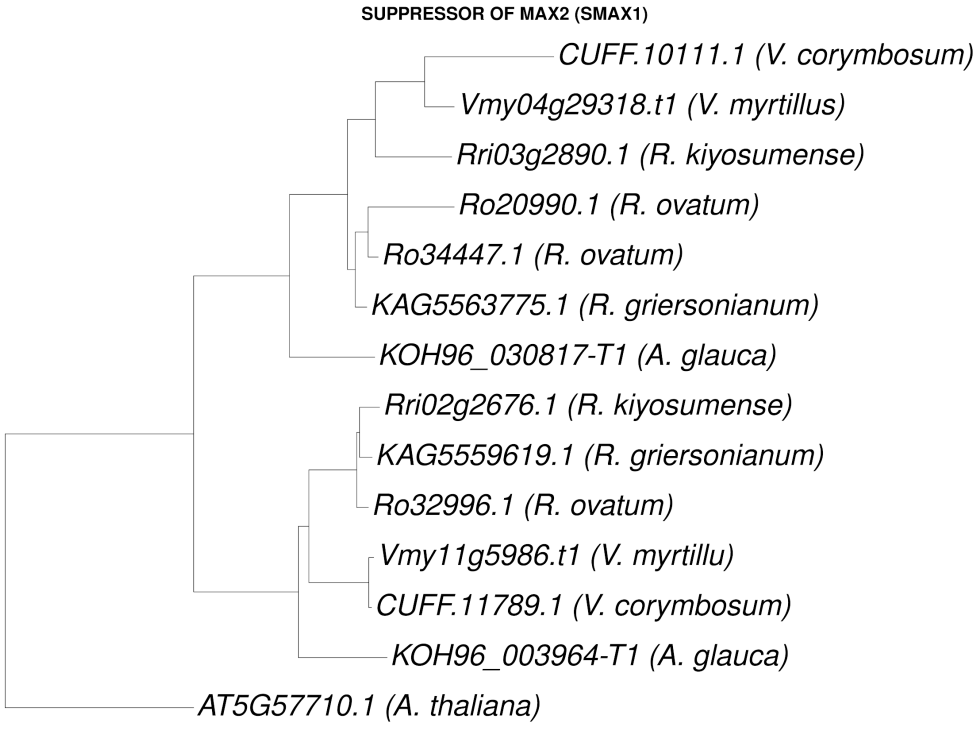
(a)


KARRIKINS INSENSITIVE2 (KAI2)
KAG5528528.1 (R. griersonianum)
KAG5535544.1 (R. griersonianum)
CUFF.57571.1 (V. corymbosum)
Rri01g1745.1 (R. kiyosumense)
Ro31992.1 (R. ovatum)
Ro31992.2 (R. ovatum)
Vmy08g19610.t1 (V. myrtillus)
KOH96_036519-T1 (A. glauca)
KOH96_021305-T1 (A. glauca)
KOH96_021304-T1 (A. glauca)
Ro13361.1 (R. ovatum)
KAG5535543.1a (R. griersonianum)
AT4G37470.1 (A. thaliana)

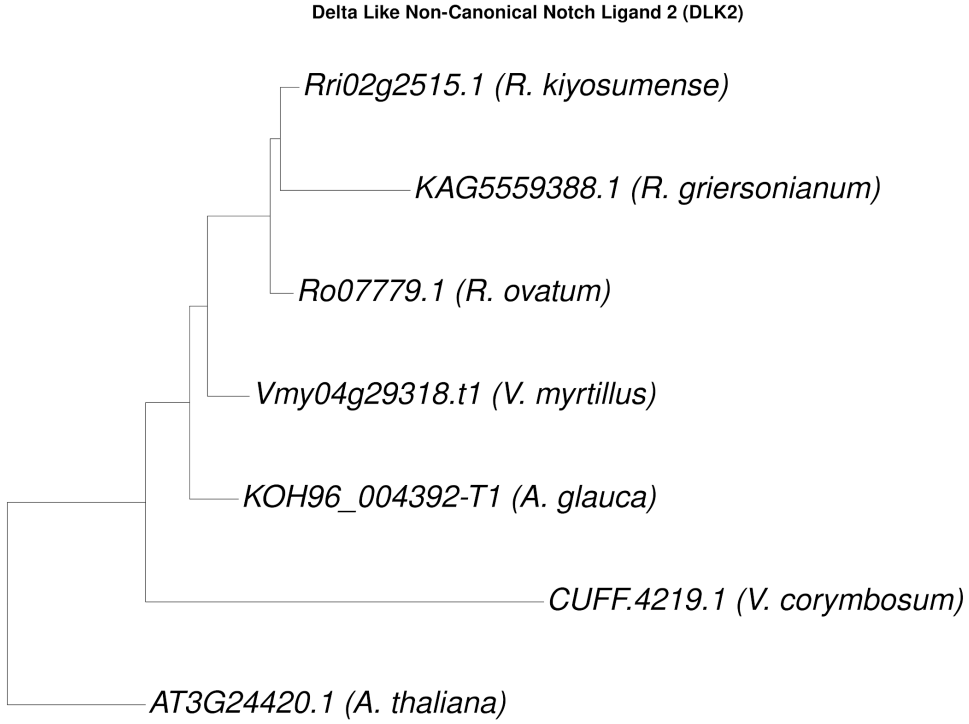
(b)


MORE AXILLARY BRANCHES2 (MAX2)
Ro36971.1 (R. ovatum)
Ro33962.1 (R. ovatum)
Rri06g1564.1 (R. kiyosumense)
KAG5545101.1 (R. griersonianum)
CUFF.19469.1 (V. corymbosum)
CUFF.19469.2 (V. corymbosum)
Vmy04g27868.t1 (V. myrtillus)
KOH96_032429-T1 (A. glauca)
AT2G42620.1 (A. thaliana)

(c)


SUPPRESSOR OF MAX2 (SMAX1)
CUFF.10111.1 (V. corymbosum)
Vmy04g29318.t1 (V. myrtillus)
Rri03g2890.1 (R. kiyosumense)
Ro20990.1 (R. ovatum)
Ro34447.1 (R. ovatum)
KAG5563775.1 (R. griersonianum)
KOH96_030817-T1 (A. glauca)
Rri02g2676.1 (R. kiyosumense)
KAG5559619.1 (R. griersonianum)
Ro32996.1 (R. ovatum)
Vmy11g5986.t1 (V. myrtillu)
CUFF.11789.1 (V. corymbosum)
KOH96_003964-T1 (A. glauca)
AT5G57710.1 (A. thaliana)

(d)


Delta Like Non-Canonical Notch Ligand 2 (DLK2)
Rri02g2515.1 (R. kiyosumense)
KAG5559388.1 (R. griersonianum)
Ro07779.1 (R. ovatum)
Vmy04g29318.t1 (V. myrtillus)
KOH96_004392-T1 (A. glauca)
CUFF.4219.1 (V. corymbosum)
AT3G24420.1 (A. thaliana)

3 Chapter 3 Niche Differentiation among Manzanita Species

3.1 Introduction:

The California Floristic Province (CFP) is located along North America's Pacific Coast (Howell 1957; Raven and Axelrod 1978) (Figure 3.1), from northern Baja CA to southwestern Oregon. Ais a biodiversity hotspot, the region is characterized by a Mediterranean climate, with hot, dry summers and colder, wetter winters. It is enriched with vascular plant species, the estimated number of which varies from 3000 to over 6000 (Burge et al. 2016; Baldwin 2014; Myers 1990; Raven and Axelrod 1978; Myers et al. 2000). More than half of these species can only be found in the CFP (Burge et al. 2016; Baldwin 2014), and around 60% of these endemic species have a geographic range that is sufficiently restricted that they are considered of conservation concern (Thorne et al. 2009; Smith and York 1984). These restricted geographic ranges are hypothesized to be the result of unique niches defined by distinct soil types or climatic conditions (Stebbins and Major 1965; Kruckeberg 1986; Kruckeberg and Rabinowitz 1985; Kraft, Baldwin, and Ackerly 2010). The diversification of endemic plants in the CFP is thought to be related to the historical formation of Mediterranean climate as well as episodes of geological activity (Raven and Axelrod 1978; Baldwin 2014; Axelrod 1981). The uplift of mountains exposed diverse soil types and introduced a steep elevation gradient across which climatic variables such as temperature, precipitation, and solar radiation can vary within a small geographic range (Stebbins and Major 1965; Raven and Axelrod 1978).

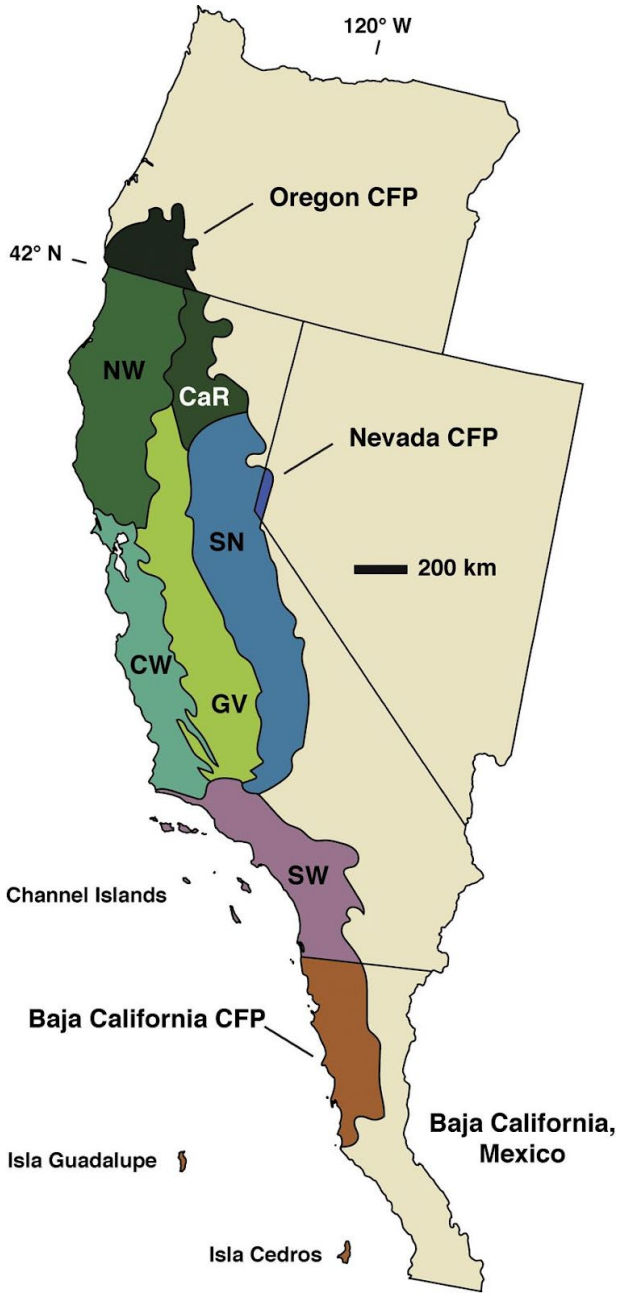


Figure 3.1 Map of the California Floristic Province (CFP) and its ecoregions, from Burge et al. 2016. NW: Northwestern California, CaR: Cascade Ranges, SN: Sierra Nevada, GV: Great Valley, CW: Central Western California, and SW: Southwestern California.

Among the CFP flora, *Arctostaphylos* species, also known as manzanitas, make up a diverse woody genus that is enriched with endemic taxa (Parker, Vasey, and Keeley 2007; Kauffmann et al. 2015). This genus includes over 100 species and subspecies, a majority of which can only be found in the CFP (Kauffmann et al. 2015; Baldwin et al. 2012). Around 49 of the 60 manzanita species occupy narrow geographic ranges ('The Jepson Manual' ; Kauffmann et al. 2015), and 44 are considered threatened or endangered by the California Native Plant Society (Smith and York 1984), making them crucial components in the conservation management of the CFP (Gluesenkamp et al. 2011; Burge et al. 2018; Halsey and Keeley 2016). As is true of many CFP groups, the hypothesized rapid diversification of *Arctostaphylos* is thought to have been a response to aridification and development of heterogeneous edaphic environments in the CFP (Parker 2007; Boykin et al. 2005; Wells 1969).

Because local adaptation is hypothesized to be responsible for the high degree of endemism among CFP species, including *Arctostaphylos* species (Wells 1969; Baldwin 2014; Raven and Axelrod 1978), investigation of niche differentiation is important to understanding the evolution of this highly endemic genus. In general, manzanitas can be said to live in relatively similar habitats: they all have at least some of their distribution in the CFP and thus are exposed to the drought- and heat-mediated Mediterranean climate (Kauffmann et al. 2015; Baldwin et al. 2012). In addition, most manzanita species can be found in climatically related chaparral communities (Kauffmann et al. 2015). Chaparral is a unique CFP biome found at elevations of 300-1800 m between the coastal sage scrub and oak and pine forest communities (Schoenherr 2017). However, within the genus, habitat features have been used to

distinguish species (Ball et al. 1983; Wieslander and Schreiber 1939; Kauffmann et al. 2015). These habitat characteristics are often descriptive, and have not been evaluated quantitatively. Therefore, the extent of niche overlap between manzanita species, and whether ecological factors can distinguish species, remain unknown.

In this study, we used climatic and edaphic variables to estimate the extent of niche overlap among narrowly distributed manzanita species, and to test whether these factors can distinguish species or partition them into ecologically distinct groups. Using species distribution modeling, we found niche overlap between all possible narrowly-distributed species pairs, making it impossible to divide those species into ecologically distinct groups based on climatic and edaphic data. However, our analyses confirmed that individual climatic and edaphic variables can be usefµL in distinguishing some species in the same geographic region. Moreover, our analyses suggest the potential importance of including soil factors in studies designed to evaluate conservation status. In addition, we found that eleven species have predicted ranges that meet the threshold of critically endangered species as recognized by the International Union for Conservation of Nature (IUCN). Only three of these eleven species are currently listed as threatened by the state of CA or the federal government (Kajtaniak and Easterbrook 2019; Smith 2020), therefore our results suggest at least 8 additional species should be examined for protected status.

3.2 Methods

3.2.1 Study area and environmental data

The California Floristic Province (CFP) includes part of southwestern Oregon, the non-desert parts of California, a corner of western Nevada, and northern Baja California

(Howell 1957; Raven and Axelrod 1978; Burge et al. 2016) (Figure 3.1). We used the CFP for a comprehensive study of niche differentiation among species within *Arctostaphylos*. Finer resolution geospatial data is available for California than for the entire CFP, therefore, we also conducted a more in-depth investigation into ecological differentiation among manzanita species within the state.

For analyses using the entire CFP, we used data with ~1 km resolution For the analysis restricted to California, we used data with a resolution of~ 270 m. The environmental variables in both geospatial datasets are climatic, hydrologic, terrain, and soil factors that are associated with plant distributions but the specific factors differ between the two datasets (Vasey, Loik, and Parker 2012; Franklin 1998) (Table 3.1; Table 3.2).

Variables	Unit	Included in building SDMs	Included in the MDS and PCA analyses	Link
BIO1 Annual Mean Temperature	°C × 10	No	No	http://worldclim.org/version2
BIO2 Mean Diurnal Range (Mean of monthly (max temp - min temp))	°C × 10	Yes	Yes	
BIO3 Isothermality (BIO2/BIO7) (* 100)	N/A	No	Yes	
BIO4 Temperature Seasonality (standard deviation *100)	°C × 10	Yes	No	
BIO5 Max Temperature of Warmest Month	°C ×11	Yes	No	
BIO6 Min Temperature of Coldest Month	°C × 12	Yes	No	
BIO7 Temperature Annual Range (BIO5-BIO6)	°C × 13	No	No	
BIO8 Mean Temperature of Wettest Quarter	°C × 14	No	No	
BIO9 Mean Temperature of Driest Quarter	°C × 15	No	No	
BIO10 Mean Temperature of Warmest Quarter	°C × 16	No	No	
BIO11 Mean Temperature of Coldest Quarter	°C × 17	No	No	
BIO12 Annual Precipitation	mm	No	No	
BIO13 Precipitation of Wettest Month	mm	No	No	
BIO14 Precipitation of Driest Month	mm	No	No	
BIO15 Precipitation Seasonality (Coefficient of Variation)	mm	No	Yes	
BIO16 Precipitation of Wettest Quarter	mm	No	No	
BIO17 Precipitation of Driest Quarter	mm	No	No	
BIO18 Precipitation of Warmest Quarter	mm	Yes	No	
BIO19 Precipitation of Coldest Quarter	mm	No	No	
Solar Radiation	kJ m^{-2} day^{-1}	Yes	No	

Bulk density of the fine earth fraction	kg/dm³	No	No	https://soilgrids.org
Cation Exchange Capacity of the soil	cmol(c)/kg	Yes	Yes	
Volumetric fraction of coarse fragments (> 2 mm)	$cm^3/100cm^3$ (vol%)	Yes	Yes	
Proportion of clay particles (< 0.002 mm) in the fine earth fraction	g/100g (%)	No	Yes	
Total nitrogen (N)	g/kg	No	No	
Soil pH	pH	No	No	
Proportion of sand particles (> 0.05 mm) in the fine earth fraction	g/100g (%)	No	No	
Proportion of silt particles (≥ 0.002 mm and ≤ 0.05 mm) in the fine earth fraction	g/100g (%)	Yes	No	
Soil organic carbon content in the fine earth fraction	g/kg	No	No	
Organic carbon density	kg/m³	Yes	Yes	
Organic carbon stocks	kg/m²	No	No	

Table 3.1 Environmental variables of the 1 km dataset and their inclusion in building the Species Distribution Models (SDMs) and niche differentiation analysis, including the multidimensional scaling analysis (MDS) and principal component analysis (PCA).

Variables	Unit	Description	Included in building SDMs	Included in the MDS and PCA analyses	Link
Actual evapotranspiration (aet)	mm	Water variable between wilting point and field capacity	Yes	Yes	Basin Characterization Model
April 1 snowpack (aprpck)	mm	Snow water equivalent in March equivalent to April 1st	Yes	No	
Climatic water deficit (cwd)	mm	Potential minus actual evapotranspiration	No	No	
Potential evapotranspiration (pet)	mm	Water that could evaporate or transpire from plants if available	No	No	
Total precipitation (ppt)	mm	Precipitation	No	No	
Recharge (rch)	mm	Amount of water that penetrates below the root zone	Yes	Yes	
Runoff (run)	mm	Amount of water that becomes stream flow	No	Yes	
Minimum monthly temperature (tmn)	°C	Minimum Monthly Temperature	Yes	Yes	
Maximum monthly temperature (tmx)	°C	Maximum monthly temperature	No	No	

Available Water Capacity (awc)	mm	A common depth of plant rooting (where 80 percent of the roots occur)	Yes	Yes	Gridded National Soil Survey Geographic Database
Calcium Carbonate (cc)	%	The percent of carbonates, by weight, in the fraction of the soil less than 2 millimeters in size	Yes	No	
Cation Exchange Capacity (cec)	Milli-equivalents/ 100 g soil	A general indicator of productivity potential for a soil	Yes	No	
Organic Matter (og)	%	Fraction of the soil composed of anything that once lived	Yes	Yes	
Soil pH (ph)	NA	The degree of soil acidity or alkalinity	Yes	Yes	
Sodium Absorption Ratio (sar)	milli-equivalents/ L	The ratio of the Na concentration divided by the square root of one-half of the Ca + Mg concentration	Yes	No	
Soil Surface Texture (st)	NA	Physical property of soil defined by the composition of sand, clay and silt	Yes	No	

Table 3.2 Environmental variables of the 270 m dataset and their inclusion in building the Species Distribution Models (SDMs) and niche differentiation analysis, including the multidimensional scaling analysis (MDS) and principal component analysis (PCA).

For the 1 km geospatial dataset, we downloaded the 19 bioclimatic variables and the solar radiation variable from WorldClim (http://www.worldclim.com/version2), and the 12 soil variables from the SoilGrid (https://www.isric.org/explore/soilgrids) database (Table 1). We cropped these environmental layers to match the boundaries of the CFP as defined by the CFP polygon file developed by Burge et al. 2016. Because the cropped layers have minor differences in resolution, we adjusted them to the same resolution (~1 km) using the raster package in R (Hijmans 2017).

For the 270 m geospatial dataset, we downloaded the seven climatic and hydrologic variables from the United States Geologic Survey Basin Characterization Model (USGSBCM) (Flint et al. 2013), and nine soil variables from Soil Survey Staff Gridded National Soil Survey Geographic (gNATSGO) Database for California (USDA Natural Resources Conservation Service, 2020) (Staff 2019) (Table 3.2). We cropped these environmental layers to the boundaries of the state of California and adjusted them to the same resolution (~270 m) using the raster package in R (Hijmans 2017).

For both the coarse- (1 km) and fine-resolution (270 m) datasets, we calculated Pearson's correlation coefficient (r) to evaluate pairwise correlation between environmental variables. If two variables were highly correlated (Pearson's correlation coefficient, $|r| > 0.7$), we retained only one of them (Green 1979) (Table 3.1; Table 3.2). After the elimination, we were left with ten variables in the 1 km dataset: (1) BIO2, mean diurnal range (mean of monthly (max temp - min temp)), (2) BIO4, temperature seasonality (standard deviation *100), (3) BIO5, max temperature of warmest Month, (4) BIO6, min temperature of coldest month, (5) BIO18, precipitation of warmest Quarter, (6) solar radiation, (7) cation exchange capacity of the soil (cec), (8) volumetric fraction of

coarse fragments (> 2 mm) (cfvo), (9) organic carbon density (ocd), and (10) proportion of silt particles (≥ 0.002 mm and ≤ 0.05 mm) in the fine earth fraction (silt) (Table 3.1). After the elimination of redundant variables, 11 predictors were retained in the 270 m dataset: (1) actual evapotranspiration (aet), (2) April 1 snowpack (aprpck), (3) recharge (rch), (4) minimum monthly temperature (tmn), (5) available water capacity (awc), (6) calcium carbonate (cc), (7) cation exchange capacity (cec), (8) organic matter (og), (9) soil pH (ph), (10) sodium absorption ratio (sar) and (11) soil surface texture (Table 3.2).

3.2.2 Species records and data cleaning

We eliminated widespread *Arctostaphylos* species from our analyses because their extensive ranges mean that they have broad niches that overlap with those of many other species. We were left with 49 of the original 60 species for the assessment of ecological differentiation in the CFP (Table 3.3). For these CFP-wide analyses, geospatial data are available only at a 1 km scale. Finer-resolution geospatial data (270 m) are available for the state of California, therefore for 44 species that have distributions restricted to California, we also carried out studies with this second data set. (Table 3.3).

Species	Predicted range <100km^2 in the 1 km analyses	Predicted range <100km^2 in the 270 m analyses	Conservation status according to the CA or federal government
A. andersonii	No	No	
A. auriculata	No	No	
A. australis	No	NA	
A. bakeri	No	Yes	ST
A. catalinae	Yes	Yes	
A. confertiflora	Yes	No	FE
A. cruzensis	No	Yes	
A. edmundsii	Yes	Yes	ST
A. gabilanensis #	Yes	NA	
A. glutinosa	Yes	No	
A. hookeri	No	No	subspecies *hearstiorum*: SE; subspecies *ravenii*: FE
A. hooveri	No	No	
A. incognita	No	NA	
A. insularis	Yes	No	
A. klamathensis	No	No	
A. luciana	Yes	Yes	
A. malloryi	Yes	No	
A. montana	Yes	Yes	
A. montaraensis	Yes	Yes	
A. montereyensis	No	Yes	
A. moranii	No	NA	
A. morroensis	Yes	Yes	FT
A. myrtifolia	No	Yes	FT
A. nissenana	No	No	
A. nortensis	No	No	
A. obispoensis	No	No	
A. ohloneana	Yes	Yes	
A. osoensis	Yes	Yes	
A. otayensis	Yes	No	
A. pajaroensis	No	No	

A. pallida	Yes	Yes	SE; FT
A. pechoensis	No	Yes	
A. peninsularis	No	NA	
A. pilosula	No	No	
A. pumila	No	Yes	
A. purissima	No	No	
A. rainbowensis	No	No	
A. refugioensis	Yes	No	
A. regismontana	Yes	Yes	
A. rudis	No	No	
A. sensitiva	No	No	
A. silvicola	No	No	
A. virgata	No	No	
A. viridissima	Yes	Yes	
A. bolensis *	NA	NA	
A. densiflora * #	NA	NA	SE
A. franciscana * #	NA	NA	FE
A. pacifica * #	NA	NA	SE
A. imbricata *	NA	Yes	SE

Table 3.3 Conservation status of 49 manzanita species in the California Floristic Province (CFP). The five species distributed outside the state of California are underlined. Species that were eliminated in the construction of SDMs derived from the 1 km dataset are marked by stars. Species that were eliminated in the construction of SDMs derived from the 270 m dataset are marked by number signs. The 11 species that have a predicted range of less than 100km2 in both the 1km and 270 m analyses are in bold. In the column labeled conservation status, FT and FE indicate US federally threatened and federally endangered species respectively, and ST and SE represent California state threatened and state endangered species respectively.

In order to document the distributional range of each species, we used herbarium specimen data that included the longitude and latitude of the locality where the specimen was collected (occurrence data). For species that have all of their distribution in California, we downloaded occurrence data from the online database of the Consortium of California Herbaria 2 (CCH2) (https://www.cch2.org/portal/). For species with occurrences outside of California, we downloaded occurrence data from the Global Biodiversity Information Facility (GBIF) (https://www.gbif.org/). After the download, we filtered out records that were duplicated or with geographic coordinate uncertainties larger than 1 km.

The identification of manzanita species is challenging for even experienced taxonomists (Keeley, Thomas Parker, and Vasey 2017). To confirm the species identification of herbarium specimens, we used a customized pipeline that applied collection localities, online specimen label data, and images to eliminate records with a high probability of incorrect identification. We used ArcGIS to map the filtered records (Esri 2011). Following that, we adopted two different strategies. For records within the range of a species as described in Kauffmann et al. (2021), we eliminated any with images of, or label data describing, morphological traits that were in conflict with the current species description (Baldwin et al. 2012; Kauffmann et al. 2015). For records outside of the range, we kept those with images of, or label data describing, morphological traits that are diagnostic for that species. Applying this mapping and checking operation to every available specimen is very time-consuming. Furthermore, to verify that a species occurs at a given location, it is not necessary to verify every

specimen - confirmation of only one is needed. Therefore, to speed up the process, we selectively checked the records within every pixel at the finest resolution (~270 m). As long as one record was confirmed to be identified correctly, we retained this record as positive occurrence for that pixel. We further retained all records for that species in that pixel as occurrences, but did not check each one. Even using this generous method, we had fewer than 50 occurrences for most species.

3.2.3 Species distribution models (SDMs) and species distribution maps for species with ≥10 collection records

We constructed SDMs to obtain species distribution maps, which include where the species has been found as well as environmentally similar regions where it might be found. We constructed the models and generated these maps for species with more than 10 collection records using the R package ENMTML (Andrade, Velazco, and De Marco Júnior 2020) as follows. After cleaning the species records data, most manzanita species had fewer than 50 occurrences, which is too few for some analytical packages. Because we wanted to apply the same algorithm to construct the SDMs for all species, we applied the Maximum Entropy algorithm because it is suitable for analysis of species with limited collection records (Hernandez et al. 2006). For each species, we used the Jepson ecoregions (Baldwin et al. 2012; Burge et al. 2016) in which the occurrence data of each species falls as the species accessible area, the broader area in which the species might occur, for model fitting. In this method, every combination of longitude and latitude coordinates represents one point of the species accessible area. For every species, we randomly sampled 10,000 points in the species accessible area to generate the background points, which are used to define the available environment for model

construction. For species with <20 records, we used random bootstrap to partition the occurrence data for the purpose of model evaluation. For species with ≥20 collection records, we used spatial block cross-validation frameworks to partition the data.

Many evaluation metrics measuring the model performance of SDMs depend on the species prevalence, which is defined as the proportion of the species accessible area in which the focal species occurs. The true prevalence is not known for most manzanita species because they may occur in areas where they have not been collected. Therefore, we assessed the predictive accuracy of the SDMs using two evaluation metrics that are independent of prevalence, the true skill statistic (TSS) (Allouche, Tsoar, and Kadmon 2006) and the area under the receiver operating characteristic curve (AUC) (Manel, Williams, and Ormerod 2001). The value of AUC ranges from 0.5 to 1, and a higher AUC value indicates better performance of the SDM. The value of TSS ranges from -1 to +1, and +1 indicates perfect model performance (Swets 1988; Manel, Williams, and Ormerod 2001).

To obtain a binary distribution map, we set the threshold of the presence-absence prediction to be the one at which the sum of the sensitivity and specificity is the highest. To overcome the problem of overprediction, we applied the posterior SDM correction method to restrict the species distribution maps to the patches with confirmed occurrence data (Mendes et al. 2020; Velazco et al. 2020).

3.2.4 SDMs and species distribution map for species with <10 collection records

When species have few collection records relative to the number of environmental variables, the performance of SDMs is usually poor (Franklin 2010). To

obtain a more accurate prediction of distributions for species with fewer than 10 collection records, we applied the Ensemble of Small Models (ESMs) approach using the R package ecospat (Di Cola et al. 2017). The ESM approach creates bivariate models determined by all possible pairs of environmental predictor combinations, removes models with an AUC that is smaller than 0.8, and then assembles all these models into a single model, weighting each based on their AUC value (Breiner et al. 2015). For each small model, we used the same algorithm, the same method to generate background points, and the same approach to set the threshold to produce binary maps as we did for the SDMs of species with ≥10 records. For these species, restricting the distribution to the patches with confirmed occurrence data was ineffective at correcting the overprediction, and still led to species distribution maps that were far beyond the documented range of these species. Therefore, we chose a different posterior SDM correction method and constrained the species distribution maps by drawing a minimum convex hull polygon (MCP) containing all the occurrence data and excluding suitable cells outside the MCP (Kremen et al. 2008; Velazco et al. 2020).

3.2.5 Niche differentiation within the genus

For every species, using the binary distribution map, we extracted the environmental data for each pixel in which the species is predicted to occur using the raster package in R (Hijmans and van Etten 2012). We used the 1 km data set for all species in the analysis, and the 270 m data set for species restricted to CA. Following that, we applied the hypervolume method (Blonder 2014) to construct the environmental space for each species and then calculated the Jaccard similarity coefficient to represent the niche overlap between every two species (Mammola 2019). The Jaccard similarity

coefficient is defined as the size of the intersection between the hypervolumes of each of the pair of species divided by the size of the union of the two hypervolumes (Tanimoto 1958).

With more environmental variables/dimensions, more species occurrence data points are needed to construct the hypervolume, therefore it is necessary to include the lowest number of dimensions that satisfies the analysis goal of quantifying the niche (Blonder 2014; Blonder et al. 2014). Although the 11 uncorrelated variables were suitable for the construction of SDMs, they introduced too many dimensions for the hypervolume analysis. Therefore, we reduced the environmental dimensions by decreasing the threshold of the Pearson coefficient to 0.65 (Blonder 2014), which is the default threshold for the hypervolume method . This eliminated more environmental variables while still ensuring that the datasets included both climatic and edaphic variables. After elimination, the final seven environmental variables in the 1 km datasets were (1) BIO2, mean diurnal range (mean of monthly (max temp - min temp)), (2) BIO3, Isothermality, (3) BIO15 Precipitation Seasonality (Coefficient of Variation), (4) cation exchange capacity of the soil (cec), (5) volumetric fraction of coarse fragments (> 2 mm) (cfvo), (6) proportion of clay particles (< 0.002 mm) in the fine earth fraction and (7) organic carbon density (ocd) (Table 3.1). The hypervolume method can only be applied to continuous variables, therefore we removed soil surface texture (st), a categorical variable, from the 270 m dataset in addition to removing redundant variables based on the new Pearson coefficient threshold. The seven environmental variables in the final 270 m dataset were (1) actual evapotranspiration (aet), (2) Recharge (rch), (3) Runoff (run), (4) minimum monthly temperature (tmn), (5) available Water Capacity (awc), (6) Organic Matter (og), and (7) Soil pH (ph) (Table 3.2).

After calculating the niche overlap for every pair of species, we obtained a niche similarity matrix for CFP species using the 1 km dataset (Table 3.1) and another for CA species using the 270 m dataset (Table 3.2). Applying the formula 1-N (N = original value of the cell) to every cell of the matrices, we transformed the matrices into niche distance matrices. To test whether the two distance matrices are correlated with each other, we trimmed the 1 km matrix to retain only the species that are included in the 270 m matrix and conducted a Mantel test (Mantel 1967) to compare the two matrices.

We used each of the distance matrices as input and applied multidimensional scaling analysis (MDS) using the R package MASS (Ripley et al. 2013) to visualize the ecological distance among species of the CFP and CA. We optimized the parameters of the MDS model, including the number of iterations from 20 to 100, the tolerance threshold for stopping from 0.05 to 0.1, and the number of dimensions from two to three, to obtain optimized models with the lowest stress value, which represents the measure of goodness of fit between the MDS models and the input distance matrix (Wilkinson and Others 2002).

To investigate whether manzanitas form ecologically distinct groups, we applied *k*-means clustering analysis (MacQueen and Others 1967; Forgey 1965) on the two optimized MDS models using the R package MASS (Ripley et al. 2013). We calculated the within-group sum-of-squares (Forgey 1965) to determine the optimal number of clusters. For a given optimal number of clusters, we ran the partition three times to observe how the assignment of species into the clusters changed.

To explore whether patterns of ecological differentiation correspond to other attributes of manzanitas, we color-coded the species in the MDS plots based on their

geographic regions, phylogenetic positions, number of chromosome sets, and adaptation to serpentine soil. We used the manzanita field guide (Kauffmann et al. 2015) to assign species to six different geographic regions: Klamath Mountains, San Francisco Bay, Central Coast, Sierra Nevada, Southern California, and Baja California. Previous phylogenetic analyses revealed two lineages within the genus, referred as the “big clade” and “small clade” (Boykin et al. 2005; Wahlert, Parker, and Vasey 2009). We color-coded the species according to their clade assignment and removed the species with unknown assignments. In addition, we also color-coded species as diploid or tetraploid (Baldwin et al. 2012). However, many polyploid species are widespread and therefore had been eliminated from the analysis. Thus, our analyses only had three polyploid species.

3.2.6 Niche differentiation of manzanita species within the Central Coast and SoCal-Baja CA region

Because we did not initially find a clear clustering pattern when using either the total CFP species or Californian species, we investigated niche differentiation based on subsets of species from different geographic regions. We focused on two regions, the Central Coast of California, and southern California/Baja California Mexico (SoCal-Baja CA). The central coast of California, ranging from Monterey County in the north to Santa Barbara County in the south, is the area of the CFP with the highest number of endemic manzanita species (Baldwin et al. 2012; Kauffmann et al. 2015; Parker 2007). In contrast, far fewer endemic manzanita species are present in the southern California and Baja CA region (Baldwin et al. 2012; Kauffmann et al. 2015).

We conducted principal components analyses (PCA), using the R package factoextra (Kassambara and Mundt 2017), to estimate ecological differentiation among 19 species present in the Central Coast region, and 8 species present the SoCal-Baja region respectively. We used the same environmental variables that were used to quantify the niches and calculate niche overlap. Both the 1 km and 270 m datasets were used in the analysis of the Central Coast species. However, the 270 m geospatial data are not available in the Baja CA region, therefore we only used the 1 km dataset for the analysis of species in the SoCal-Baja CA region.

3.3 Results

3.3.1 A 270 m geospatial dataset produced SDMs with higher accuracy than a 1 km dataset

The goal of this project was to investigate ecological differentiation of manzanita species. There are 60 species in the genus *Arctostaphylos*, but we eliminated 11 widespread species because of their extensive habitat overlap. This left 49 species in this study. We evaluated niche differentiation using two different datasets, one at a scale of 1 km, which included all of the California Floristic Province and 49 species, and one at 270 m, which included only the state of California and 44 species.

Using the 1 km dataset, we constructed SDMs of 44 of the total 49 CFP manzanita species (Table 3.3). Five manzanita species, *A. bolensis*, *A. densiflora*, *A. franciscana*, *A. imbricata*, and *A. pacifica,* were omitted because many of the bivariate models for these species had low AUC value (<0.8) and were eliminated. This left insufficient models to be assembled into final SDMs. Using the 270 m dataset, we performed analyses for the 44 species found only in the state of California. We obtained

SDMs of 40 of these species (Table 3.3). The SDMs of four species (*A. densiflora*, *A. franciscana*, *A. pacifica*, and *A. gabilanensis)* were omitted due to low AUC values.

For the 44 SDMs produced using the 1 km dataset, the average value of the AUC was 0.67 with a standard deviation of 0.17. AUC values of 0.5-0.7 are considered low (Swets 1988; Manel, Williams, and Ormerod 2001), therefore our result of 0.67 indicates poor model performance. The average value of TSS was 0.46 with a standard deviation of 0.23, suggesting that the performance of the SDMs using the coarse-resolution data is good (Landis and Koch 1977). The SDMs using the 270 m dataset performed better, with an average AUC of 0.87 (standard derivation of 0.09), and an average TSS of 0.71 (standard derivation 0.14).

With the binary species distribution maps generated using the 1 km dataset, we found that the predicted range of CFP species varied from 4 km2 to 8,046 km2. Almost all of the species considered except *A. peninsularis* had a geographic distribution restricted to an area less than 5000km2, which is the threshold of the geographical range for endangered species according to the International Union for Conservation of Nature (IUCN) (Iucn 2001; Gaston and Fuller 2009). Eighteen had an estimated range that was less than 100 km2, which is the threshold for IUCN critically endangered species. The analysis using the 270 m datasets indicated that the predicted geographic range of CA species varies from 0.365 km2 to 1346 km2, suggesting that all of the species would qualify as endangered according to IUCN criteria (< 5000km^2), and 18 as critically endangered (Iucn 2001; Gaston and Fuller 2009). In both the 1 km and 270 m analyses, 11 species met the threshold of critically endangered species. Another 12 species qualify as critically endangered in one analysis, but not in the other. In addition,

two species, *A. gabilanensis* and *A. montereyensis*, were suggested to be critically endangered in one analysis but were eliminated from the other analysis due to the limited number of collections (Table 3.3).

3.3.2 The niche similarity matrices derived from the coarse-resolution and fine-resolution datasets suggest different patterns of niche differentiation among manzanita species

We calculated the Jaccard similarity coefficient (Tanimoto 1958; Jaccard 1912) to evaluate niche overlap between every pair of species that we analyzed. The Jaccard similarity coefficient ranges from 0 to 1, where higher values represent more niche overlap(Tanimoto 1958; Jaccard 1912). For the analysis of CFP species using the 1 km dataset, the median value of niche similarity was 0.268 (Figure 3.2). The majority (75%) of species pairs had a niche overlap value ranging from to 0.17 to 0.3 (Figure 3.2). Five pairs of species had substantially higher niche similarity values: *A. purissima* vs *A. rudis*, *A. moranii* vs *A. incognita*, *A. andersonii* vs *A. nortensis*, *A. andersonii* vs *A. silvicola*, and *A. sensitiva* vs *A. silvicola*. However, there were no pairs with substantially lower overlap values. For the analysis of the California species using 270 m dataset, the median value of the pairwise niche similarity was 0.227 (Figure 3.2). The majority (75%) of species pairs had niche overlap values ranging from 0.08 to 0.27 (Figure 3.2). Similar to the result of the analysis using 1 km dataset, we did not find any outlier points with extremely small values (with little or no overlap). There were nine species pairs that had exceptionally high niche overlap values: *A. andersonii* vs *A. sensitiva*, *A. andersonii* vs *A. silvicola*, *A. confertiflora* vs *A. auricula*, *A. edmundsii* vs *A. myrtifolia*, *A. osoensis* vs *A. edmundsii*, *A. hooveri* vs *A. klamathensis*, *A. hooveri* vs *A. refugioensis*, *A. pallida* vs

A. osoensis, and *A. sensitiva* vs *A. silvicola*. Only two pairs, *A. sensitiva* vs *A. silvicola* and *A. andersonii* vs *A. silvicola*, were also high-overlap pairs in the analysis using the 1 km dataset.

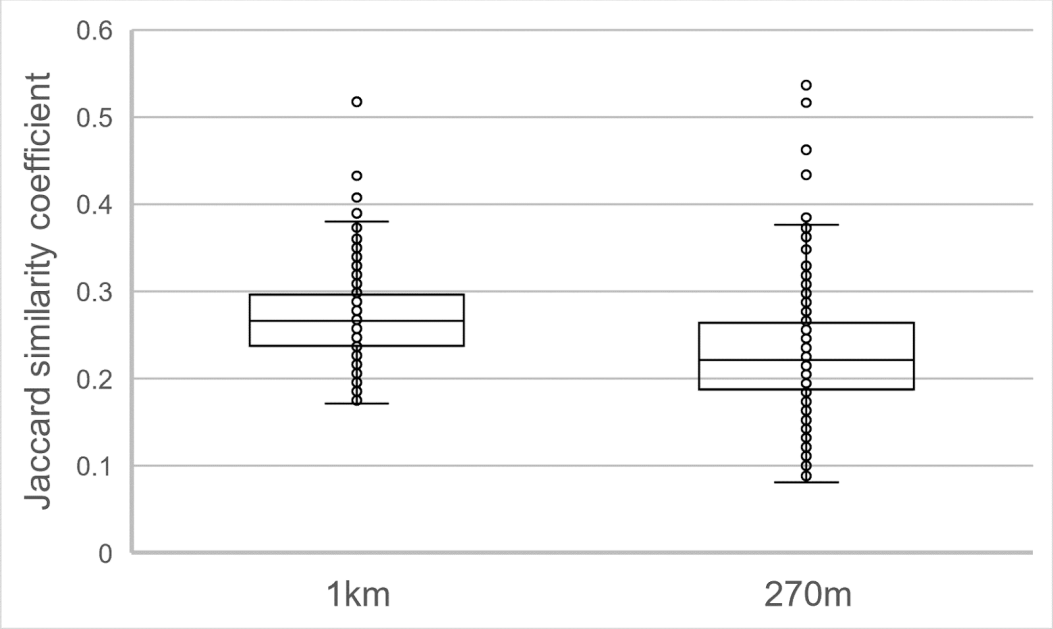


Figure 3.2 Box-and-whisker plots showing the distribution of niche overlap values including the species pairs that have exceptionally high values in the 1 km and 270 m analyses. The Y-axis is the value of the Jaccard similarity coefficient representing the niche overlap of species pairs. Each point represents one species pair. The upper and lower box outlines indicate the third (Q3) and first quartile (Q1) values. The distance between the third and first quartile is interquartile range (IQR). The inner bar of the box indicates the median value. The horizontal lines above and below the box represents the upper extreme value (Q3 + 1.5*IQR) and lower extreme value (Q1-1.5 * IQR). Points that lie either below the lower extreme line or above the upper extreme line are considered outliers.

To test whether the calculated pairwise niche overlap was consistent between the analyses using the coarse- and fine-resolution dataset, we conducted a Mantel test (Mantel 1967) to compare the two niche overlap matrices. Although the median values and upper extreme values were similar between the two datasets, the *p-value* of the Mantel test between the two matrices was 0.56, supporting the null hypothesis they are not correlated with each other. This suggests that the pattern of niche differentiation using the fine-resolution data differs from the one using the coarse-resolution data, and indicates that the choice of datasets affects the results of ecological differentiation analyses.

3.3.3 Both the coarse-resolution and fine-resolution geospatial data failed to cluster species into ecologically distinct groups

To visualize the ecological distance between manzanita species, we applied MDS analysis to the 1 km and 270 m niche matrices. For the optimized MDS models based on the 1 km dataset, the number of dimensions was three. We have presented the visualization of the MDS models through all two-way combinations of MDS1, MDS2, and MDS3. CFP species are evenly distributed and no clusters are visually apparent (Figure 3.3). We applied k-means clustering method to the niche similarity matrix, which determined the optimal numbers of clusters to be 6, 7, or 8 (Appendix S3.1). Because *k-means* clustering can produce different solutions if done multiple times (Li and Wu 2012), we performed the analysis three times for each of the three optimal numbers of clusters. Each time, the assignment of species to the clusters changed. The analyses produced unstable cluster assignment regardless of the optimal number of clusters used in the analysis (Appendix S3.2). This suggests that there is no strong clustering signal in

the data, and that the CFP species cannot be assigned to ecologically distinct groups (subgroups of species with high niche overlap) based on the 1 km data.

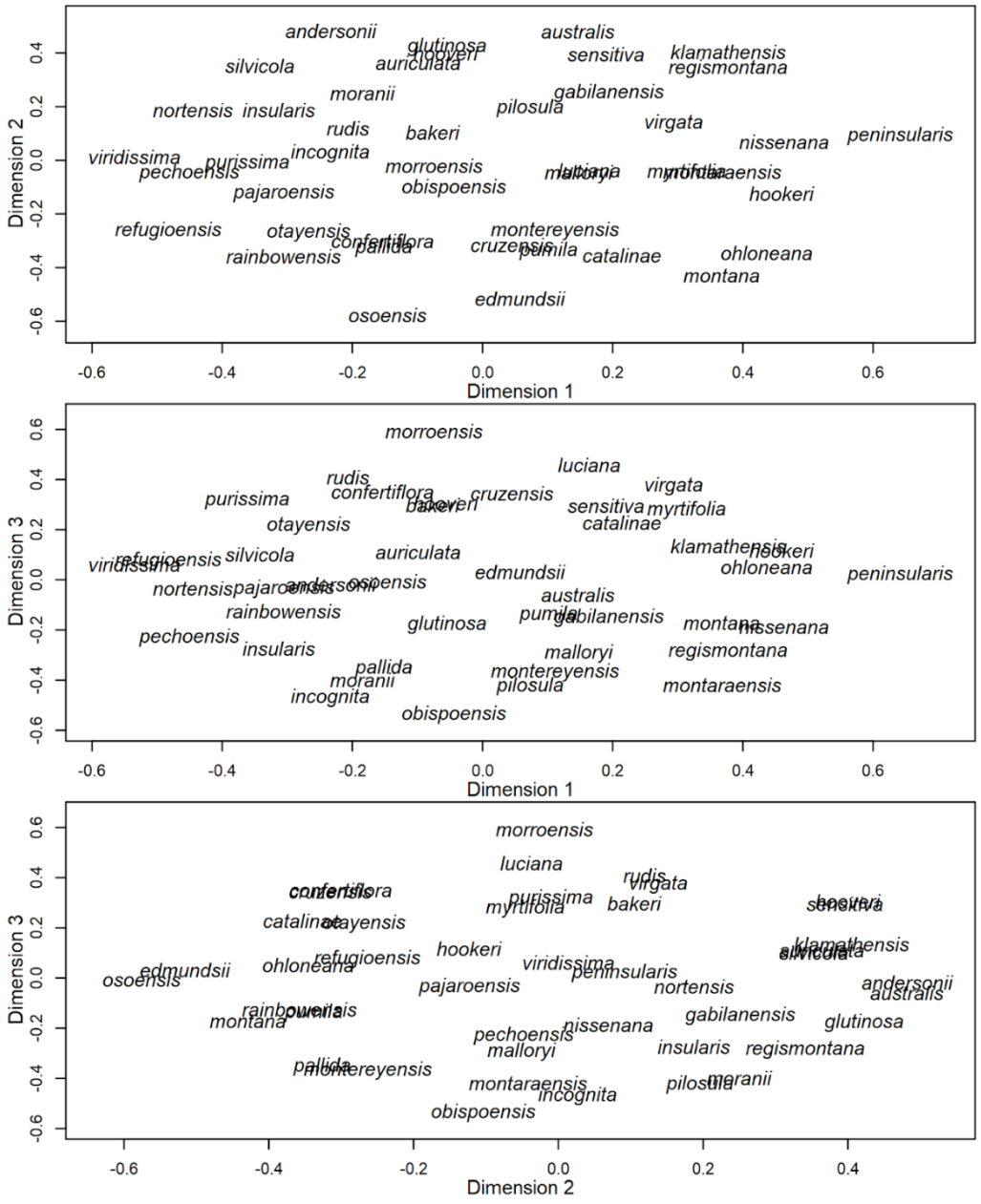


Figure 3.3 Three plots of the three-dimensional MDS models based on the 1 km dataset, each plot showing two dimensions. No clustering pattern is apparent in the plots. The distance between the species labels indicates the ecological distance between the species.

We obtained the same result for the analysis comparing CA species using the 270 m dataset. The optimized number of dimensions in the final MDS model was two in this analysis. As in the previous analysis, the results show the CA species do not form clusters (Figure 3.4). A *k*-means clustering analysis indicated the optimal numbers of clusters were again 6, 7, or 8 (Appendix S3.1). Also, as in the previous analysis, the assignment of species to clusters differed with each repetition of the analysis at a given number of clusters (Appendix S3.3). Thus, dividing manzanita species into ecologically distinct groups was not possible with coarse- or fine-resolution geospatial data.

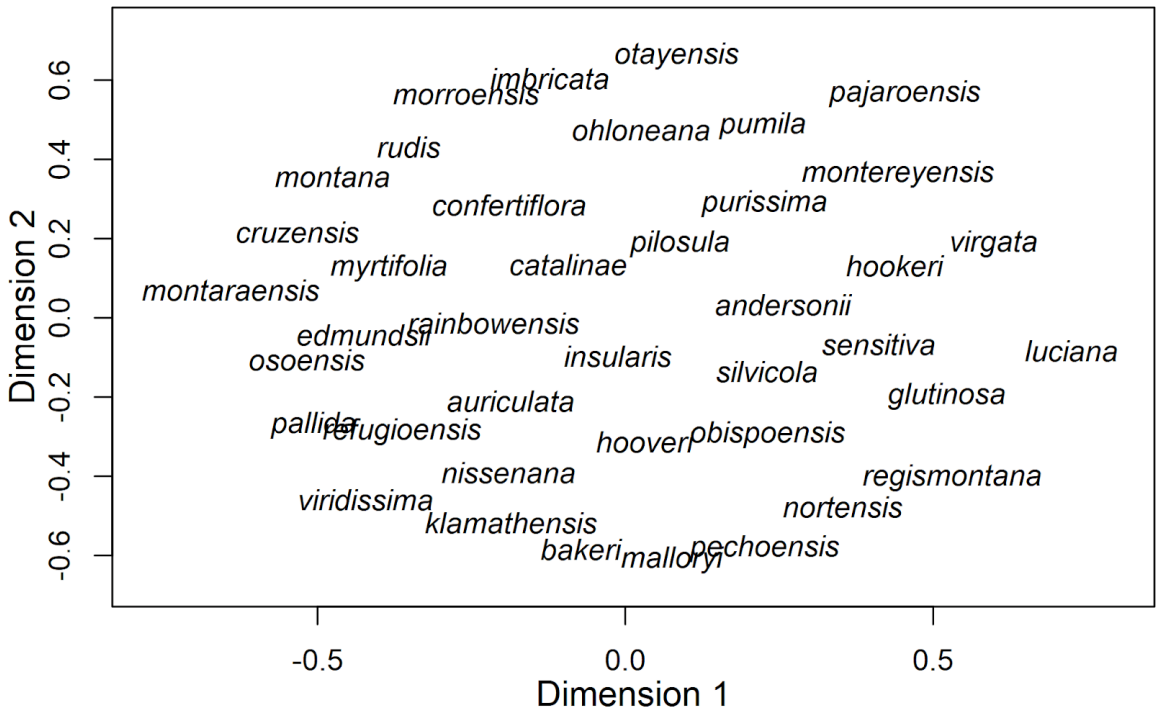


Figure 3.4 Plot of the two-dimensional MDS models based on the 270 m dataset. No clustering pattern is apparent in the plot. The distance between the species labels indicates the ecological distance between the species

Because our analyses found no strong clustering pattern, we tested for other signal in the data. We evaluated correlation between ecological differentiation and geographic region, phylogenetic assignment, ploidy level, and the presence/absence of serpentine soil in the habitat. However, we found no correlation with any of these features (Appendix S3.4; Appendix S3.5). Species from the same geographic region were spread throughout the MDS plot, suggesting that ecological variation did not correspond to geographic distribution (Appendix S3.4; Appendix S3.5). Our results showed the same lack of correlation between ecological differentiation and the other three features, regardless of which environmental dataset was used.

3.3.4 Two of the eight manzanita species in the Southern California-Baja CA region have distinct niches using the 1 km dataset

Because the pattern of niche differentiation using all possible species showed that every species had some overlap with the other species, we narrowed our focus to smaller geographic regions to determine if there were clear patterns of niche differentiation at a more local scale. The focal regions included one area enriched with manzanita species, the Central Coast, and one area with relatively poor diversity of manzanita species, the Southern California-Baja CA Mexico (SoCal-Baja CA) region. We choose PCA for this analysis because it would eliminate the effect of some invariant environmental factors.

To test whether the habitats of the eight manzanita species of the SoCal-Baja CA region can be distinguished by climatic and edaphic factors, we extracted environmental data from the pixels where these species were predicted to be present, and used these data as the input in the PCA analysis. Due to the limits of data available for the Baja CA

region, we used only the 1 km coarse-resolution data, which consisted of seven climatic and edaphic variables. Principal component 1 (PC1) and PC2 respectively explained 41.6% and 34.2% of the variation. PC1 and PC2 were most heavily weighted by four environmental variables ranked by their contribution: (1) BIO3, isothermality, (2) organic carbon density, (3) BIO15, precipitation seasonality, and (4) BIO2, mean diurnal range (Figure 3.5). These four variables were highly correlated with some of the variables eliminated from the analysis due to high correlation: (1) BIO3 isothermality was highly correlated with BIO4 temperature seasonality; (2) organic carbon density was highly correlated with BIO12 annual precipitation, BIO13 precipitation of wettest month, BIO19 precipitation of coldest quarter, total nitrogen, and organic carbon stock and solar radiation; (3) BIO15 precipitation seasonality was highly correlated with BIO11 mean temperature of coldest quarter, BIO14 precipitation of driest month, BIO17 precipitation of driest month, BIO18 precipitation of warmest quarter, and BIO19 precipitation of coldest quarter; (4) BIO2 was highly correlated with BIO5 max temperature of warmest month. In the PCA plot, most species overlapped with each other, indicating overlapping niches. Two species were exceptions (Figure 3.5). Samples of *A. catalinae* formed a distinct group that showed no overlap with the other species. The dominant explanatory factor for this distinction was differences in organic carbon density. Among all eight species, *A. peninsularis* occupied the largest environmental space. A large proportion of this space was unique to *A. peninsularis* and did not overlap with the other manzanita species, suggesting some degree of niche differentiation. The dominant explanatory factor for this distinction was differences in PC2, which corresponds to the precipitation seasonality.

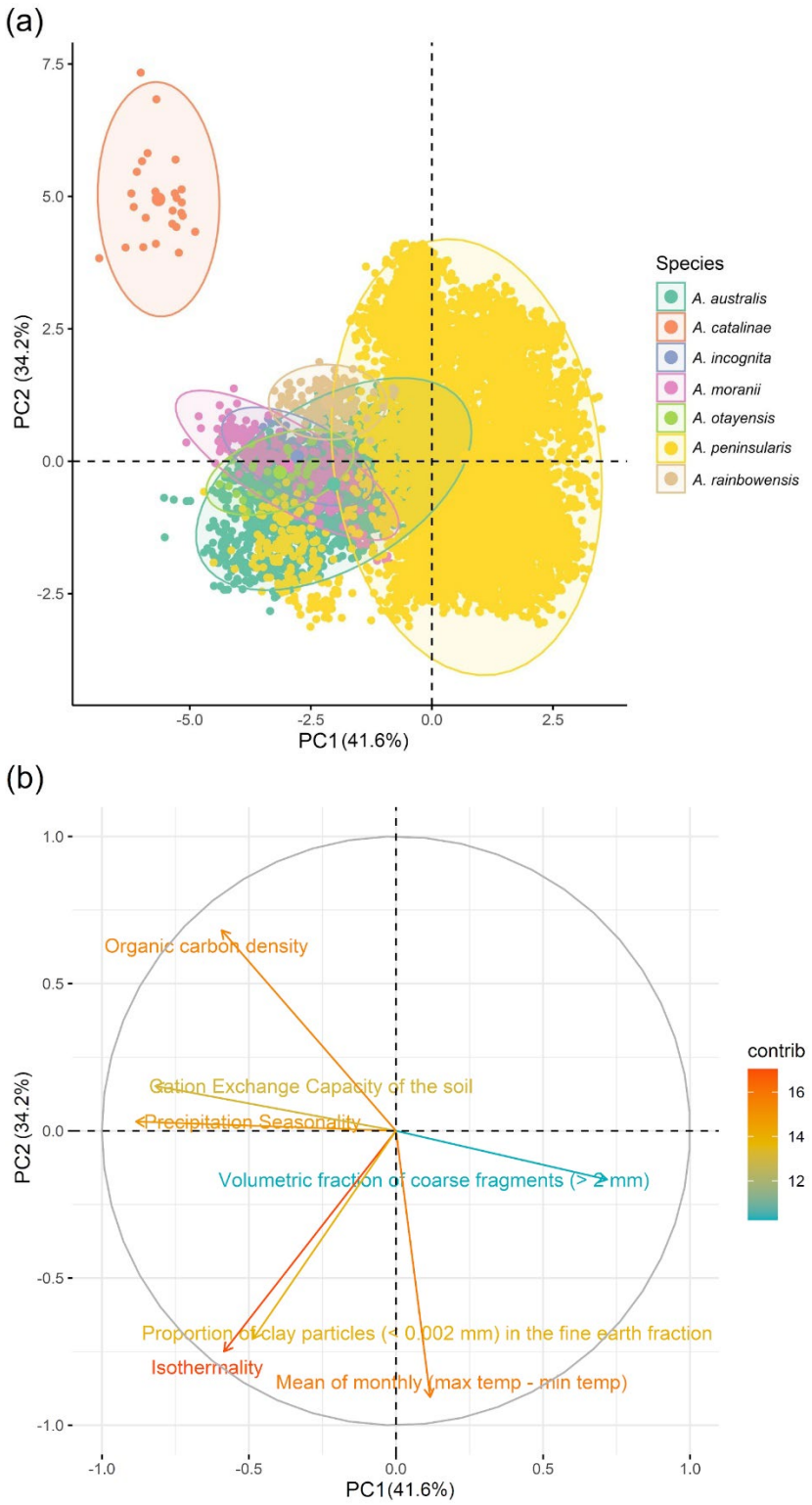
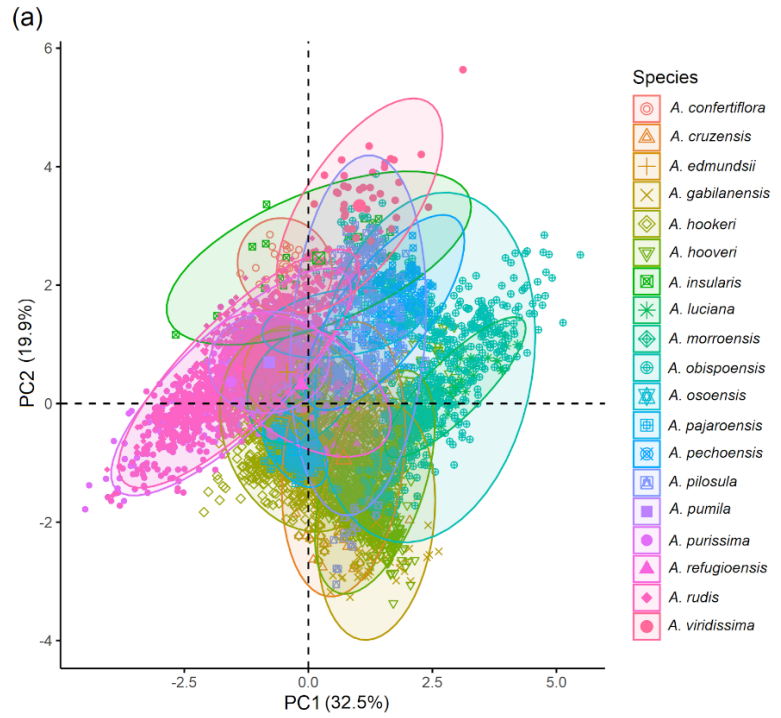

(a)
7.5
5.0
2.5
0.0
-2.5
PC2 (34.2%)
-5.0
-2.5
0.0
2.5
PC1(41.6%)
Species
A. australis
A. catalinae
A. incognita
A. moranii
A. otayensis
A. peninsularis
A. rainbowensis
(b)
1.0
0.5
0.0
-0.5
-1.0
PC2 (34.2%)
Organic carbon density
Cation Exchange Capacity of the soil
Precipitation Seasonality
Volumetric fraction of coarse fragments (> 2 mm)
Proportion of clay particles (< 0.002 mm) in the fine earth fraction
Isothermality
Mean of monthly (max temp - min temp)
-1.0
-0.5
0.0
0.5
1.0
PC1(41.6%)
contrib
16
14
12

Figure 3.5 PCA using the 1 km environmental dataset for eight Southern California-Baja CA species. Two of the eight species, A. catalinae and A. peninsularis, have distinct niches. (a) Points represent pixels in which the species are present in the binary distribution maps, and the position of these points in the plot are determined by the values of the environmental variables of the pixels. Ellipses enclose 95 % of the data points. Both the points and ellipses are color-coded according to species. (b) Arrows represent environmental variables, which are presented as vectors. Arrows are color-coded according to their contribution (contrib) to the separation of samples.

3.3.5 Regardless of resolution, climatic and edaphic variables failed to distinguish the habitats of manzanita species in the Central Coast area

We used both the coarse-resolution and fine-resolution datasets to test whether manzanita species of the Central Coast area, from Monterey County to Santa Barbara county inclusive, could be distinguished by environmental variables (Figure 3.6). For the PCA analysis using the 1 km dataset, the percentages of variation explained by PC1 and PC2 were 32.5% and 19.9% respectively (Figure 3.6). The environmental variables that made major contributions were (1) cation exchange capacity of the soil, and (2) proportion of clay particles (< 0.002 mm) in the fine earth fraction (Figure 3.6). No eliminated variable was highly correlated with cation exchange capacity of the soil but the proportion of clay particles in the fine earth fraction is highly correlated with another edaphic variable, the proportion of sand particles (> 0.05 mm) in the fine earth fraction. The range of environmental variability differed among species, but most species overlapped with others, suggesting species cannot be distinguished on the basis of the 1 km dataset.


(a)
PC2 (19.9%)
PC1 (32.5%)
6
4
2
0
-2
-4
-2.5
0.0
2.5
5.0
Species
A. confertiflora
A. cruzensis
A. edmundsii
A. gabilanensis
A. hookeri
A. hooveri
A. insularis
A. luciana
A. morroensis
A. obispoensis
A. osoensis
A. pajaroensis
A. pechoensis
A. pilosula
A. pumila
A. purissima
A. refugioensis
A. rudis
A. viridissima

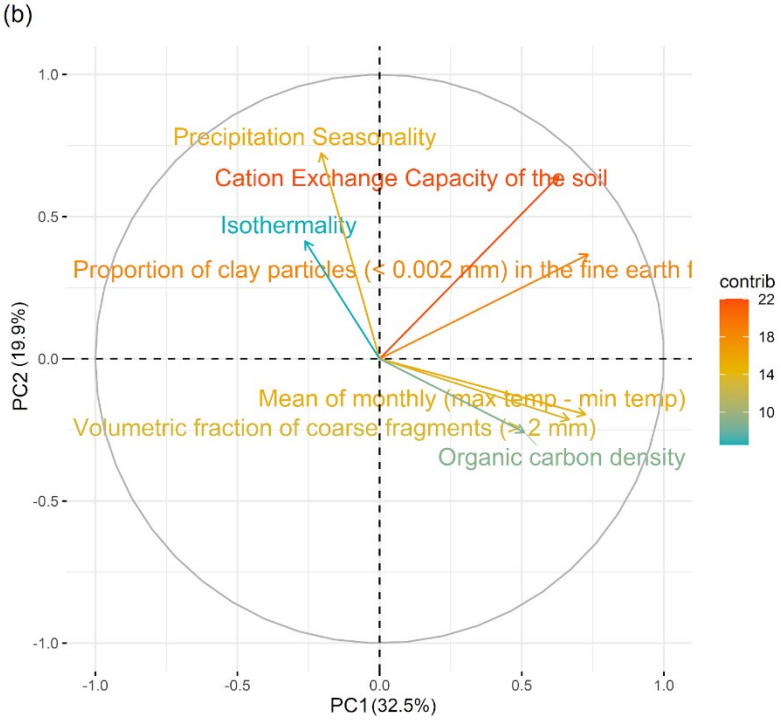

(b)
PC2 (19.9%)
PC1 (32.5%)
1.0
0.5
0.0
-0.5
-1.0
Precipitation Seasonality
Cation Exchange Capacity of the soil
Isothermality
Proportion of clay particles (< 0.002 mm) in the fine earth f
Mean of monthly (max temp - min temp)
Volumetric fraction of coarse fragments (> 2 mm)
Organic carbon density
contrib
22
18
14
10

Figure 3.6 PCA using the 1 km environmental dataset for 19 Central Coast species. The analysis failed to identify any ecologically distinct species.(a) Points represent pixels in which the species are present in the binary distribution maps, and the position of these points in the plot are determined by the values of the environmental variables of the pixels. Points of different species are distinguished by colors and shapes. Ellipses enclose 95 % of the data points and are color-coded according to species. (b) Arrows represent environmental variables, which are presented as vectors. Arrows are color-coded according to their contribution (contrib) to the separation of samples.

In the analysis using the 270 m dataset, the percentages of variation explained by PC1 and PC2 were 20.3% and 16.6% (Figure 3.7). The environmental variables making major contributions toward PC1 and PC2 were (1) soil pH, (2) available water capacity, and (3) the actual evapotranspiration (Figure 3.7). No eliminated variable was highly correlated with these variables. The plot revealed a similar pattern of ecological differentiation as with the coarse-resolution dataset: the environmental space of different species overlapped with each other, suggesting that climatic and edaphic factors cannot distinguish these species.

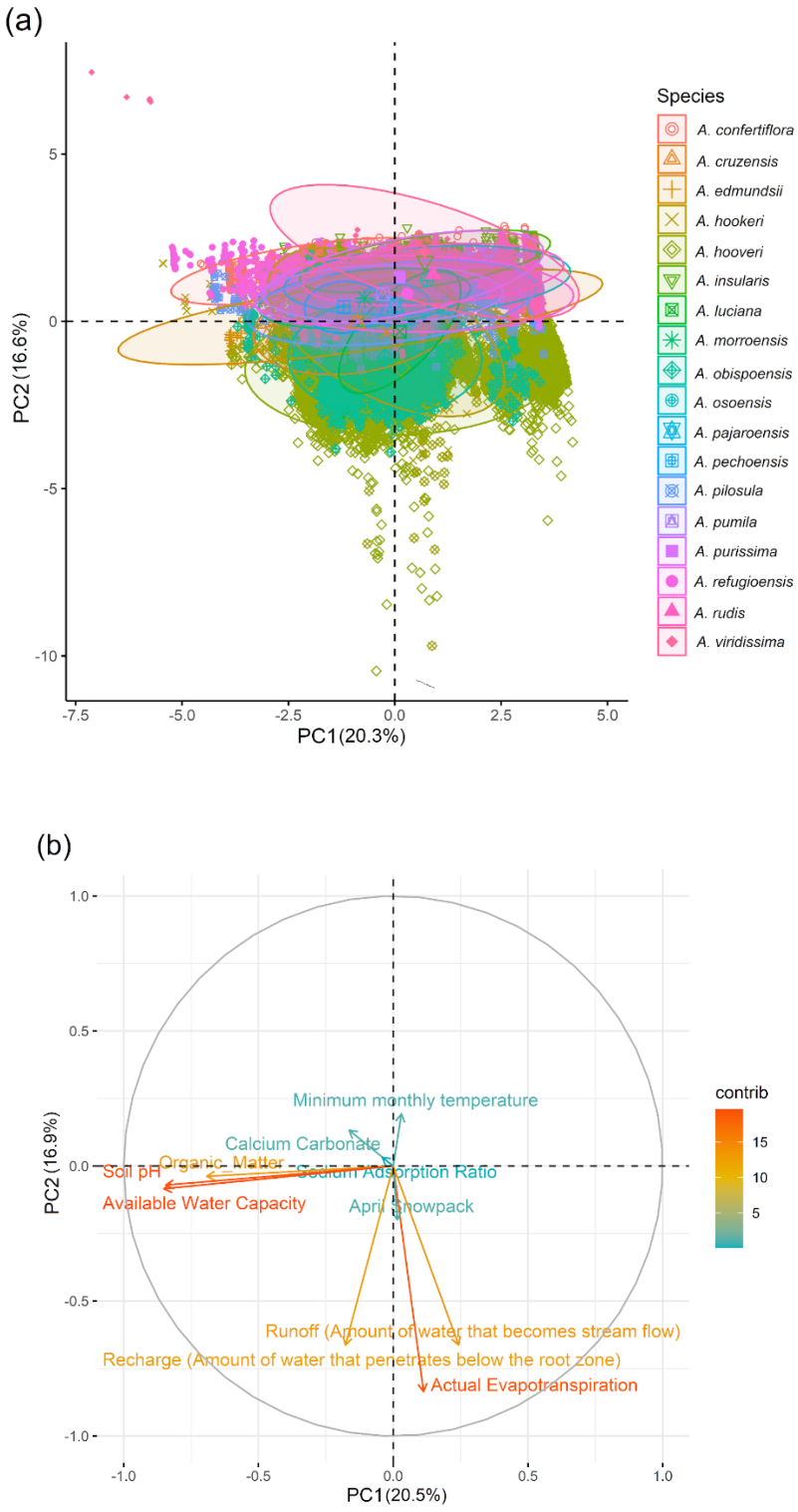

(a)
Species
A. confertiflora
A. cruzensis
A. edmundsii
A. hookeri
A. hooveri
A. insularis
A. luciana
A. morroensis
A. obispoensis
A. osoensis
A. pajaroensis
A. pechoensis
A. pilosula
A. pumila
A. purissima
A. refugioensis
A. rudis
A. viridissima
PC2 (16.6%)
PC1 (20.3%)
5
0
-5
-10
-7.5
-5.0
-2.5
0.0
2.5
5.0
(b)
contrib
15
10
5
Minimum monthly temperature
Calcium Carbonate
Organic Matter
Soil pH
Sodium Adsorption Ratio
Available Water Capacity
April Snowpack
Runoff (Amount of water that becomes stream flow)
Recharge (Amount of water that penetrates below the root zone)
Actual Evapotranspiration
PC2 (16.9%)
PC1 (20.5%)
1.0
0.5
0.0
-0.5
-1.0

Figure 3.7 PCA using the 270 m environmental dataset for 19 Central Coast species. The analysis failed to identify any ecologically distinct species from the Central Coast region. (a) Points represent pixels in which the species are present in the binary distribution maps, and the position of these points in the plot are determined by the values of the environmental variables of the pixels. Points of different species are distinguished by colors and shapes. Ellipses enclose 95 % of the data points and are color-coded according to species. (b) Arrows represent environmental variables, which are presented as vectors. Arrows are color-coded according to their contribution (contrib) to the separation of samples.

3.4 Discussion

3.4.1 The choice of 1 km or 270 m environmental datasets affected the construction of SDMs, the calculation of niche similarity, and the pattern of niche differentiation

On the basis of AUC and TSS evaluation, we found that the analysis using the 270 m dataset produced more accurate SDMs than the analysis using the 1 km dataset. This corresponds to previous reports that increasing the resolution of environmental data can increase the accuracy of species distribution modeling (Connor et al. 2019).

The Mantel test revealed the analyses based on the two datasets generated different patterns of niche differentiation among manzanita species. Although all of the environmental variables in both datasets were climatic and edaphic, both the resolution and specific factors were different. There were more edaphic factors in the 270 m dataset than the 1 km dataset, which included more climatic variables related to temperature and precipitation. Although resolution has been shown to influence the inference of patterns in ecological research (Turner and Gardner 2015), we cannot eliminate the possibility that the difference in environmental factors also played a role in the different results from the two datasets.

3.4.2 Species distribution modeling identified 11 manzanita species that can be considered critically endangered

The geographic range of a species is an important criterion in evaluating their conservation status (Gaston and Fuller 2009). The 1 km analysis and the 270 m analysis both found 18 manzanita species, 11 of which were in common, that were restricted to a

geographic area that was less than the threshold (100km2) used by the IUCN to define critically endangered species (Iucn 2001). The predicted range of species derived from SDMs tends to be larger than their actual range (Franklin 2010), suggesting additional species beyond those identified in the analyses might fall within the limit of critically endangered. However, only three of these 11 species, *A. edmundsii*, *A. morroensis* and *A. pallida*, are currently listed as threatened by the state of CA or the federal government. In addition, several species were eliminated from the analyses because there were too few collection records for the analyses to perform properly; those species are also likely to meet the criterion for critically endangered. Four of these eliminated species, *A. densiflora*, *A. franciscana*, *A. imbricata* and *A. pacifica*, are currently listed as threatened or endangered by the state of California or federal government. However, two, *A. bakeri* and *A. confertiflora,* which are considered rare by the state and endangered by the federal government respectively, are predicted to be critically endangered in the 270 m analysis but not in the 1 km analysis. The difference between our results and recognition by state and federal government is likely due to the governments using stricter criteria and a more complex process to identify endangered species, as well as potential political concerns.

3.4.3 Manzanita species occupy habitats with overlapping environmental features

Based on the Jaccard coefficient index, pairwise niche overlap is universal across the genus, even when the most widespread, generalist species were eliminated. In the analyses using the 1 km and 270 m datasets, the median values of niche overlap are 0.268 and 0.222 respectively. It is not possible to interpret these numbers as low or high in context, because few studies have used the hypervolume method to calculate

niche overlap among a large number of species. We did not find any species pairs with exceptionally low niche similarity values, suggesting the failure in identifying two species that are very ecologically distinct from each other (in the context of the environmental data examined). The observations of common niche overlap of all species pairs are consistent with our findings that no strong clustering pattern was present in the MDS plot, suggesting that manzanita species cannot be divided into ecologically distinct groups. Although manzanitas are found in diverse habitats such as conifer forests, rocky slopes, and sandstone outcrops, most species have at least some populations in the chaparral community, which may explain the extensive overlap (Baldwin et al. 2012; Kauffmann et al. 2015; Parker et al. 2020). In addition, in many instances, multiple manzanita species can be found in the same location(Parker et al. 2020), further suggesting shared habitat parameters.

We found five and nine pairs of species, using the 1 km and 270 m data sets respectively, that have exceptionally high calculated niche similarity values. In each analysis, there were three species pairs in which the two species had extensively overlapping ranges. For the remaining species pairs, most are geographically close to each other and the climatic factors are likely to be similar. The members of one pair, *A. andersonii* and *A. nortensis* are geographically distant: *A. andersonii* is located in Santa Cruz County while *A. nortensis* is located in the Klamath Mountains. Their large niche overlap value might reflect similarity of their environmental conditions despite this distance, or might result from inaccurate niche quantification related to the limited number of data points for each of these.

3.4.4 The limits of the data and analytical methods have an important influence on the evaluation of niche overlap

Many studies in plant species and populations have shown a correlation between ecological variation and the variation of geography, phylogenetic position, and ploidy levels among species (Grant 1981; Anacker and Strauss 2014; Baniaga et al. 2020). In our study, we found no such correlation. In addition, serpentine soil is known to play an important role in the diversification and endemism of CFP plants (Anacker et al. 2011). However, the species living in serpentine soils were spread throughout the MDS plot.

Some limits of our analyses of niche quantification and overlap might explain the inconsistency between our results and previous studies. The original 1 km dataset included 31 variables, and the 270 m dataset included 16 variables. After the elimination of redundant variables, both datasets had seven edaphic and climatic variables as input for the hypervolume method, although the specific factors differed. However, some manzanita species had a limited number of data points. When there are too many environmental variables relative to the number of data points, niche quantification is less accurate, which can affect the calculation of niche overlap (Blonder et al. 2014). Although we reduced the number of variables to account for this, some species may still have had too few data points for accurate quantification.

Ideally, we would like the result of these analyses of niche differentiation to be biologically meaningfµL. However, our method of calculating niche overlap weights every environmental variable equally. This makes the calculated niche similarities of each species pair comparable, and allows visualization of ecological variation at the genus level. However, the environmental variables are probably not all equally critical to the

survival and success of manzanitas, and relative importance might vary from species to species. Therefore, assigning equal weight to every environmental variable for every species pair may diminish signals of ecological differentiation between two species, leading to an overestimation of their niche overlap. Because we have no a priori criteria to weight variables, the null hypothesis of equal weights is necessary.

In addition, data points falling into different ranges are treated equally in the calculations. For example, a temperature difference between 0 and 2 °C would be treated the same as the difference between 28 and 30 °C. However, these different two-degree temperature differences might not be of equal biological importance. The optimal range of ambient temperatures for photosynthesis in Mediterranean woody plants is usually around 25–30 °C (Flexas et al. 2014). However, the difference between 0 °C, which is freezing, and 2 °C can be the difference between freezing and surviving. Thus, the failure to recognize the biological importance of the small numeric difference between 0 and 2 °C may lead to an underestimation of niche overlap. Conversely, a range from 26 to 29 °C would be treated as overlapping but different from a range from 27-30 °C. However, it is plausible that this difference would not be significant to many species, which might experience fluctuations up to 30 °C or down to the 26 °C regardless of their described temperature ranges. In this case, the overlap would be underestimated. Thus, our inability to recognize the biological meaning of the numeric data might lead to an underestimation or overestimation of niche overlap in our analyses.

In previous studies, niche differentiation has been identified among sympatric species using microenvironmental data for individual plants (Savolainen et al. 2006;

Laport, Minckley, and Ramsey 2016; Schönswetter et al. 2007). However, the finest resolution of our geospatial data was 270 m, which may be too coarse to determine habitat variables accurately and therefore to identify niche differentiation of sympatric manzanita species that may inhabit different microclimates. The inclusion of microclimate and soil data of finer resolution might provide us with new insight into ecological differentiation within *Arctostaphylos*.

3.4.5 Restricting analyses to narrow geographic regions can facilitate distinguishing some manzanita species

In addition to using all manzanita species of the CFP or CA in our analyses, we performed PCA analyses on a subset of species from the SoCal-Baja CA and Central Coast regions. All analyses suggested that both edaphic and climatic factors play important roles in explaining ecological variation of manzanita species, regardless of the focal geographic region or the environmental datasets. It is notable that in the analysis of Central Coast species, the environmental variables making major contributions to PC1 and PC2 are all edaphic factors that are not highly correlated with any of the climatic factors. Many ecological and evolutionary studies use geospatial climatic data in their analyses for purposes including predicting suitable habitat in the future, identifying genetic signals associated with environmental adaptation, and evaluating species delimitation (Hannah et al. 2012; Sork et al. 2016; Alvarado-Sizzo et al. 2018). These studies can facilitate plant conservation including determining the conservation units (species), identifying vulnerable populations, and prioritizing conservation area etc (Anacker et al. 2013; Sork et al. 2016; Alvarado-Sizzo et al. 2018). In contrast to climatic factors, edaphic factors were commonly excluded from those studies. However, our

analyses indicate the potential importance of edaphic factors in plant conservation, and suggest that including them will produce more accurate results and provide for more effective conservation strategies.

We only included eight species and used the 1 km environmental dataset in the analysis of SoCal-Baja CA region. However, we found that two species, *A. catalinae* and *A. peninsularis*, were ecologically distinct from their sister species. The dominant explanatory factor for the distinction of *A. catalinae* was differences in organic carbon density, an edaphic variable that was highly correlated with many other climatic and edaphic variables that were eliminated because of the high correlation. The habitat of *A. catalinae* is volcanic outcrops, which is unique among the species included in the analysis. In addition, *A. catalinae* was the only island species in this analysis, and it is also the only manzanita species on Catalina Island, which is geographically isolated from the mainland (Barbour, Keeler-Wolf, and Schoenherr 2007; Baldwin et al. 2012; Kauffmann et al. 2015). Its ecological distinction might also reflect differences between the island and mainland in precipitation, solar radiation, or organic carbon stock and nitrogen of the soil, all of which were highly correlated with organic carbon density and were therefore not directly included in the analyses.

The PCA plot showed that *A. peninsularis* has a wide range of environmental variability, with only a small proportion overlapping with the other species. The geographic range of *A. peninsularis* is relatively broad compared to the other species in the SoCal-Baja CA region, but does not overlap extensively with the ranges of the other species. The broad range may explain the wide range of environmental variability, whereas the lack of range overlap may explain the relative lack of overlap of the

environmental variables. These results suggest that specific edaphic and climatic factors can distinguish some manzanita species, at least in regions with fewer species.

In contrast, the analyses of the Central Coast region included more (18) species, and both the 1 km and 270 m datasets were used. However, we did not find any ecologically distinct species in these analyses. Our results support the hypothesis that niche overlap among manzanita species is prevalent, and therefore the inclusion of more species in any given analysis may lead to more niche overlap and diminished the ability of ecological factors to distinguish species. Among the 18 Central Coast species, *A. confertiflora* is endemic from Santa Rosa Island (Baldwin et al. 2012; Kauffmann et al. 2015; Barbour, Keeler-Wolf, and Schoenherr 2007). Unlike *A. catalinae, A. confertiflora* did not form an ecologically distinct group in the PCA plot. This suggests that unlike the differentiation of environmental conditions on Santa Catalina in comparison to the mainland of the SoCal-Baja CA region, the environmental conditions of Santa Rosa Island might be similar to the mainland part of the Central Coast area.

3.5 Conclusion

Our quantitative analyses of niche differentiation supply critical information for conserving narrowly-distributed manzanita species. Using species distribution modeling, we identified 11 manzanita species with restricted geographic distributions that may qualify as threatened and therefore require additional assessment of their conservation status. In our analyses using all possible species and weighting all soil and climatic variables equally, no narrowly distributed manzanita species had a distinct habitat. However, our in-depth investigation of species in circumscribed geographic regions indicates that *A. catalinae* occupies a geographically and environmentally isolated area

relative to other Southern California-Baja CA species. This finding implies that habitat preservation might be appropriate for conservation of individual species that occupy distinct habitats in specific geographic regions.

3.6 References

Allouche, Omri, Asaf Tsoar, and Ronen Kadmon. 2006. 'Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)', *J. Appl. Ecol.*, 43: 1223-32.

Alvarado-Sizzo, Hernán, Alejandro Casas, Fabiola Parra, Hilda Julieta Arreola-Nava, Teresa Terrazas, and Cristian Sánchez. 2018. 'Species delimitation in the Stenocereus griseus (Cactaceae) species complex reveals a new species, S. huastecorum', *PLoS One*, 13: e0190385.

Anacker, Brian L., Melanie Gogol-Prokurat, Krystal Leidholm, and Steve Schoenig. 2013. 'Climate Change Vulnerability Assessment of Rare Plants in California', *Madroño*, 60: 193-210.

Anacker, Brian L., and Sharon Y. Strauss. 2014. 'The geography and ecology of plant speciation: range overlap and niche divergence in sister species', *Proc. Biol. Sci.*, 281: 20132980.

Anacker, Brian L., Justen B. Whittall, Emma E. Goldberg, and Susan P. Harrison. 2011. 'Origins and consequences of serpentine endemism in the California flora', *Evolution*, 65: 365-76.

Andrade, André Felipe Alves de, Santiago José Elías Velazco, and Paulo De Marco Júnior. 2020. 'ENMTML: An R package for a straightforward construction of complex ecological niche models', *Environmental Modelling & Software*, 125: 104615.

Axelrod, Daniel I. 1981. 'Holocene Climatic Changes in Relation to Vegetation Disjunction and Speciation', *Am. Nat.*, 117: 847-70.

Baldwin, Bruce G. 2014. 'Origins of plant diversity in the California Floristic Province', *Annual Review of Ecology, Evolution, and Systematics*, 45: 347-69.

Baldwin, Bruce G, Douglas H Goldman, David J Keil, Robert Patterson, Thomas J Rosatti, and Linda Ann Vorobik. 2012. *The Jepson manual: vascular plants of California* (Univ of California Press).

Ball, Charles T., Jon Keeley, Harold Mooney, Jeffery Seemann, and William Winner. 1983. 'Relationship between form, function, and distribution of two Arctostaphylos species(Ericaceae) and their putative hybrids', *ACTA OECOL. , OECOL. PLANT.*, 4: 153-64.

Baniaga, Anthony E., Hannah E. Marx, Nils Arrigo, and Michael S. Barker. 2020. 'Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives', *Ecol. Lett.*, 23: 68-78.

Barbour, Michael, Todd Keeler-Wolf, and Allan A. Schoenherr. 2007. *Terrestrial Vegetation of California, 3rd Edition* (University of California Press).

Blonder, Benjamin. 2014. 'Frequently asked questions (FAQ) for hypervolume R package'.

Blonder, Benjamin, Christine Lamanna, Cyrille Violle, and Brian J. Enquist. 2014. 'The n-dimensional hypervolume', *Glob. Ecol. Biogeogr.*, 23: 595-609.

Boykin, Laura M., Michael C. Vasey, V. Thomas Parker, and Robert Patterson. 2005. 'Two lineages of Arctostaphylos (Ericaceae) identified using the internal transcribed spacer (ITS) region of the nuclear genome', *Madroño*, 52: 139-47.

Breiner, Frank T., Antoine Guisan, Ariel Bergamini, and Michael P. Nobis. 2015. 'Overcoming limitations of modelling rare species by using ensembles of small models', *Methods Ecol. Evol.*, 6: 1210-18.

Burge, Dylan O, V Thomas Parker, Margaret Mulligan, and César García Valderamma. 2018. 'Conservation genetics of the endangered Del Mar manzanita (Arctostaphylos glandulosa subsp. crassifolia) based on RAD sequencing data', *Madroño*, 65: 117-30.

Burge, Dylan O., James H. Thorne, Susan P. Harrison, Bart C. O'Brien, Jon P. Rebman, James R. Shevock, Edward R. Alverson, Linda K. Hardison, José Delgadillo RodrÍguez, Steven A. Junak, and Others. 2016. 'Plant diversity and endemism in the California Floristic Province', *Madroño*: 3-206.

Connor, Thomas, Andrés Viña, Julie A. Winkler, Vanessa Hull, Ying Tang, Ashton Shortridge, Hongbo Yang, Zhiqiang Zhao, Fang Wang, Jindong Zhang, Zejun Zhang, Caiquan Zhou, Wenke Bai, and Jianguo Liu. 2019. 'Interactive spatial scale effects on species distribution modeling: The case of the giant panda', *Sci. Rep.*, 9: 14563.

Di Cola, Valeria, Olivier Broennimann, Blaise Petitpierre, Frank T. Breiner, Manuela D'Amen, Christophe Randin, Robin Engler, Julien Pottier, Dorothea Pio, Anne Dubuis, Loic Pellissier, Rubén G. Mateo, Wim Hordijk, Nicolas Salamin, and Antoine Guisan. 2017. 'ecospat: an R package to support spatial analyses and modeling of species niches and distributions', *Ecography*, 40: 774-87.

Esri, Redlands. 2011. 'ArcGIS desktop: release 10', *Environmental Systems Research Institute, CA*.

Flexas, J., A. Diaz-Espejo, J. Gago, A. Gallé, J. Galmés, J. Gulías, and H. Medrano. 2014. 'Photosynthetic limitations in Mediterranean plants: A review', *Environ. Exp. Bot.*, 103: 12-23.

Flint, Lorraine E., Alan L. Flint, James H. Thorne, and Ryan Boynton. 2013. 'Fine-scale hydrologic modeling for regional landscape applications: the California Basin

Characterization Model development and performance', *Ecological Processes*, 2: 1-21.

Forgey, Edward. 1965. 'Cluster analysis of multivariate data: Efficiency vs. interpretability of classification', *Biometrics*, 21: 768-69.

Franklin, Janet. 1998. 'Predicting the distribution of shrub species in southern California from climate and terrain-derived variables', *J. Veg. Sci.*, 9: 733-48.

———. 2010. *Mapping Species Distributions: Spatial Inference and Prediction* (Cambridge University Press).

Gaston, Kevin J., and Richard A. Fuller. 2009. 'The sizes of species' geographic ranges', *J. Appl. Ecol.*, 46: 1-9.

Gluesenkamp, Daniel, Michael Chassé, Mark Frey, V Thomas Parker, M Vasey, and Betty Young. 2011. 'Back from the brink: A second chance at discovery and conservation of the Franciscan Manzanita', *Fremontia*, 38: 3-17.

Grant, Verne. 1981. *Plant speciation* (New York: Columbia University Press xii, 563p.- illus., maps, chrom. nos.. En 2nd edition. Maps, Chromosome numbers. General (KR, 198300748)).

Green, Roger H. 1979. *Sampling Design and Statistical Methods for Environmental Biologists* (John Wiley & Sons).

Halsey, Richard W, and Jon E Keeley. 2016. 'Conservation issues: California chaparral'.

Hannah, L., M. R. Shaw, P. Roehrdanz, M. Ikegami, O. Soong, and J. Thorne. 2012. 'Consequences of Climate Change for Native Plants and Conservation'.

Hernandez, Pilar A., Catherine H. Graham, Lawrence L. Master, and Deborah L. Albert. 2006. 'The effect of sample size and species characteristics on performance of different species distribution modeling methods', *Ecography*, 29: 773-85.

Hijmans, R. J. 2017. 'Geographic data analysis and modeling. R package ver. 2.6-7'.

Hijmans, Robert J., and Jacob van Etten. 2012. 'raster: Geographic analysis and modeling with raster data. R package version 2.0-12'.

Howell, John Thomas. 1957. 'The California flora and its province', *Leaf. West. Bot*, 8: 133–38.

Iucn. 2001. *IUCN Red List Categories and Criteria* (IUCN).

Jaccard, PaμL. 1912. 'The distribution of the flora in the alpine zone.1', *New Phytol.*, 11: 37-50.

'The Jepson Manual'.

Kajtaniak, David, and Nicholas Easterbrook. 2019. 'California Department of Fish and Wildlife'.

Kassambara, Alboukadel, and Fabian Mundt. 2017. 'Package ‘factoextra’', *Extract and visualize the results of multivariate data analyses*, 76.

Kauffmann, Michael Edward, Tom Parker, Michael Vasey, and Jeff Bisbee. 2015. *Field Guide to Manzanitas* (Backcountry Press).

Keeley, Jon E., V. Thomas Parker, and Michael C. Vasey. 2017. 'Characters in Arctostaphylos Taxonomy', *Madroño*, 64: 138-53.

Kraft, Nathan J. B., Bruce G. Baldwin, and David D. Ackerly. 2010. 'Range size, taxon age and hotspots of neoendemism in the California flora', *Divers. Distrib.*, 16: 403-13.

Kremen, C., A. Cameron, A. Moilanen, S. J. Phillips, C. D. Thomas, H. Beentje, J. Dransfield, B. L. Fisher, F. Glaw, T. C. Good, G. J. Harper, R. J. Hijmans, D. C. Lees, E. Louis, Jr., R. A. Nussbaum, C. J. Raxworthy, A. Razafimpahanana, G. E. Schatz, M. Vences, D. R. Vieites, P. C. Wright, and M. L. Zjhra. 2008. 'Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools', *Science*, 320: 222-26.

Kruckeberg, A. R., and D. Rabinowitz. 1985. 'Biological Aspects of Endemism in Higher Plants', *Annu. Rev. Ecol. Syst.*, 16: 447-79.

Kruckeberg, Arthur R. 1986. 'An essay: The stimulus of unusual geologies for plant speciation', *Syst. Bot.*, 11: 455.

Landis, J. R., and G. G. Koch. 1977. 'The measurement of observer agreement for categorical data', *Biometrics*, 33: 159-74.

Laport, Robert G., Robert L. Minckley, and Justin Ramsey. 2016. 'Ecological distributions, phenological isolation, and genetic structure in sympatric and parapatric populations of the Larrea tridentata polyploid complex', *Am. J. Bot.*, 103: 1358-74.

Li, Youguo, and Haiyan Wu. 2012. 'A Clustering Method Based on K-Means Algorithm', *Phys. Procedia*, 25: 1104-09.

MacQueen, James, and Others. 1967. "Some methods for classification and analysis of multivariate observations." In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 281-97. books.google.com.

Mammola, Stefano. 2019. 'Assessing similarity of n- dimensional hypervolumes: Which metric to use?', *J. Biogeogr.*, 46: 2012-23.

Manel, Stéphanie, H. Ceri Williams, and S. J. Ormerod. 2001. 'Evaluating presence-absence models in ecology: the need to account for prevalence', *J. Appl. Ecol.*, 38: 921-31.

Mantel, N. 1967. 'The detection of disease clustering and a generalized regression approach', *Cancer Res.*, 27: 209-20.

Mendes, Poliana, Santiago José Elías Velazco, André Felipe Alves de Andrade, and Paulo De Marco. 2020. 'Dealing with overprediction in species distribution models: How adding distance constraints can improve model accuracy', *Ecol. Modell.*, 431: 109180.

Myers, N. 1990. 'The biodiversity challenge: expanded hot-spots analysis', *Environmentalist*, 10: 243-56.

Myers, N., R. A. Mittermeier, C. G. Mittermeier, G. A. da Fonseca, and J. Kent. 2000. 'Biodiversity hotspots for conservation priorities', *nature*, 403: 853-58.

Parker, V. Thomas. 2007. 'Diversity and Evolution of Arctostaphylos and Ceanothus', *Fremontia*: 8.

Parker, V. Thomas, Christina Y. Rodriguez, Gail Wechsler, and Michael C. Vasey. 2020. 'Allopatry, hybridization, and reproductive isolation in Arctostaphylos', *Am. J. Bot.*, 107: 1798-814.

Parker, V. Thomas, Michael C. Vasey, and Jon E. Keeley. 2007. 'Taxonomic revisions in the genus Arctostaphylos (Ericaceae)', *Madroño*, 54: 148-56.

Raven, Peter H., and Daniel I. Axelrod. 1978. *Origin and Relationships of the California Flora* (University of California Press).

Ripley, Brian, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, and Maintainer Brian Ripley. 2013. 'Package ‘mass’', *Cran r*, 538: 113-20.

Savolainen, Vincent, Marie-Charlotte Anstett, Christian Lexer, Ian Hutton, James J. Clarkson, Maria V. Norup, Martyn P. Powell, David Springate, Nicolas Salamin, and William J. Baker. 2006. 'Sympatric speciation in palms on an oceanic island', *nature*, 441: 210-13.

Schoenherr, Allan A. 2017. *A Natural History of California: Second Edition* (Univ of California Press).

Schönswetter, Peter, Margarita Lachmayer, Christian Lettner, David Prehsler, Stefanie Rechnitzer, Dieter S. Reich, Michaela Sonnleitner, Iris Wagner, Karl Hülber, Gerald M. Schneeweiss, Pavel Trávnícek, and Jan Suda. 2007. 'Sympatric diploid and hexaploid cytotypes of Senecio carniolicus (Asteraceae) in the Eastern Alps are separated along an altitudinal gradient', *J. Plant Res.*, 120: 721-25.

Smith, James P. 2020. 'A list of the rare, endangered, & threatened vascular plants of California'.

Smith, James Payne, and Richard York. 1984. *Inventory of rare and endangered vascular plants of California* (California Native Plant Society).

Sork, Victoria L., Kevin Squire, PaµL F. Gugger, Stephanie E. Steele, Eric D. Levy, and Andrew J. Eckert. 2016. 'Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, Quercus lobata', *Am. J. Bot.*, 103: 33-46.

Staff, Soil Survey. 2019. 'Gridded National Soil Survey Geographic (gNATSGO) Database for the Conterminous United States'.

Stebbins, G. Ledyard, and Jack Major. 1965. 'Endemism and Speciation in the California Flora', *Ecol. Monogr.*, 35: 1-35.

Swets, J. A. 1988. 'Measuring the accuracy of diagnostic systems', *Science*, 240: 1285-93.

Tanimoto, Taffee T. 1958. 'Elementary mathematical theory of classification and prediction'.

Thorne, J. H., J. H. Viers, J. Price, and D. M. Stoms. 2009. 'Spatial Patterns of Endemic Plants in California', *naar*, 29: 344-66.

Turner, Monica G., and Robert H. Gardner. 2015. *Landscape Ecology in Theory and Practice: Pattern and Process* (Springer, New York, NY).

Vasey, Michael C., Michael E. Loik, and V. Thomas Parker. 2012. 'Influence of summer marine fog and low cloud stratus on water relations of evergreen woody shrubs (Arctostaphylos: Ericaceae) in the chaparral of central California', *Oecologia*, 170: 325-37.

Velazco, Santiago José Elías, Bruno R. Ribeiro, Livia Maira Orlandi Laureto, and Paulo De Marco Júnior. 2020. 'Overprediction of species distribution models in conservation planning: A still neglected issue with strong effects', *Biol. Conserv.*, 252: 108822.

Wahlert, Gregory A., V. Thomas Parker, and Michael C. Vasey. 2009. 'A phylogeny of Arctostaphylos (Ericaceae) inferred from nuclear ribosomal ITS sequences', *J. Bot. Res. Inst. Tex.*, 3: 673-82.

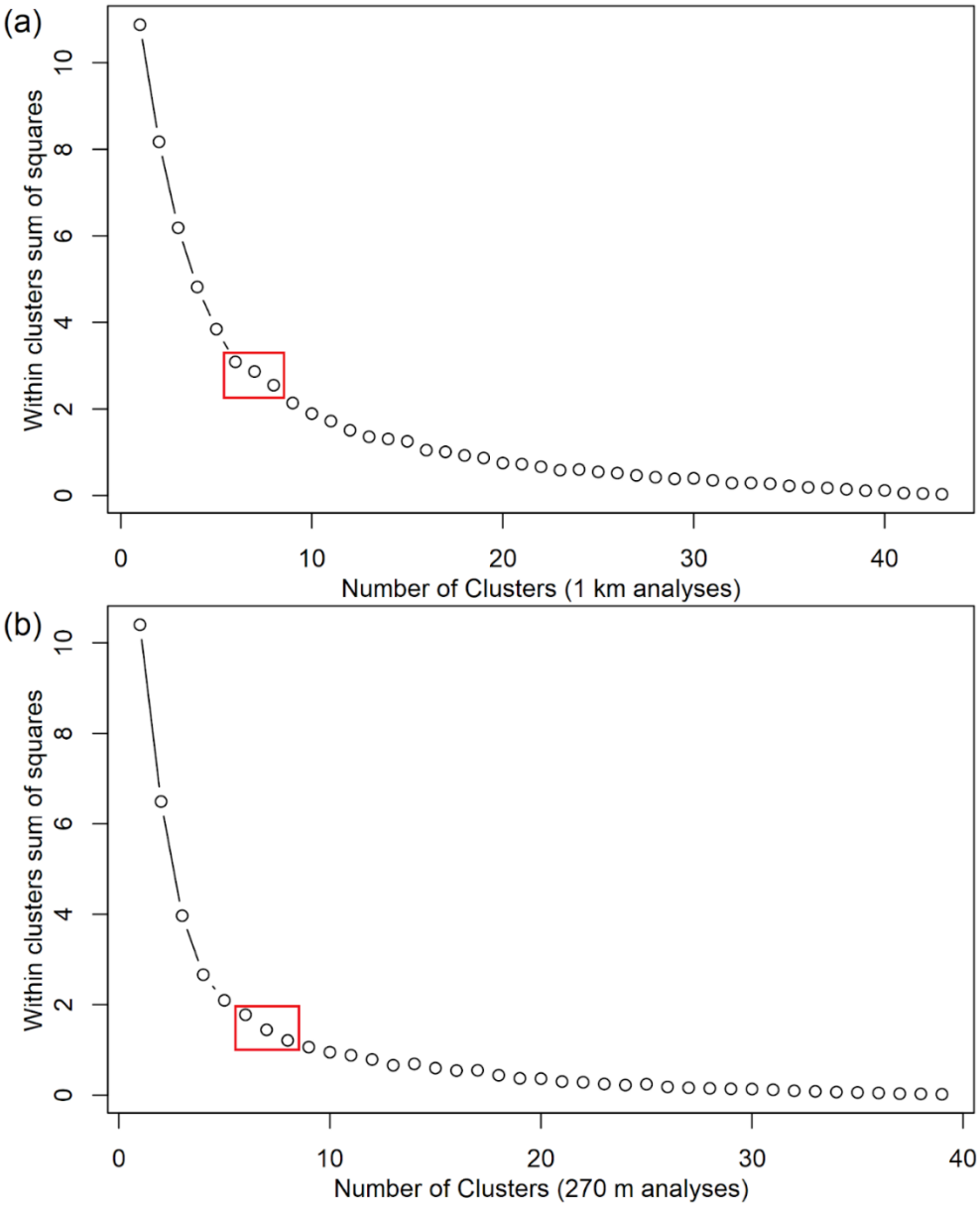
Wells, Philip V. 1969. 'The relation between mode of reproduction and extent of speciation in woody genera of the California chaparral', *Evolution*, 23: 264-67.

Wieslander, A. E., and Beryl O. Schreiber. 1939. 'Notes on the genus Arctostaphylos', *Madroño*, 5: 38-47.

Wilkinson, Leland, and Others. 2002. 'Multidimensional scaling', *Systat*, 10: 119-45.

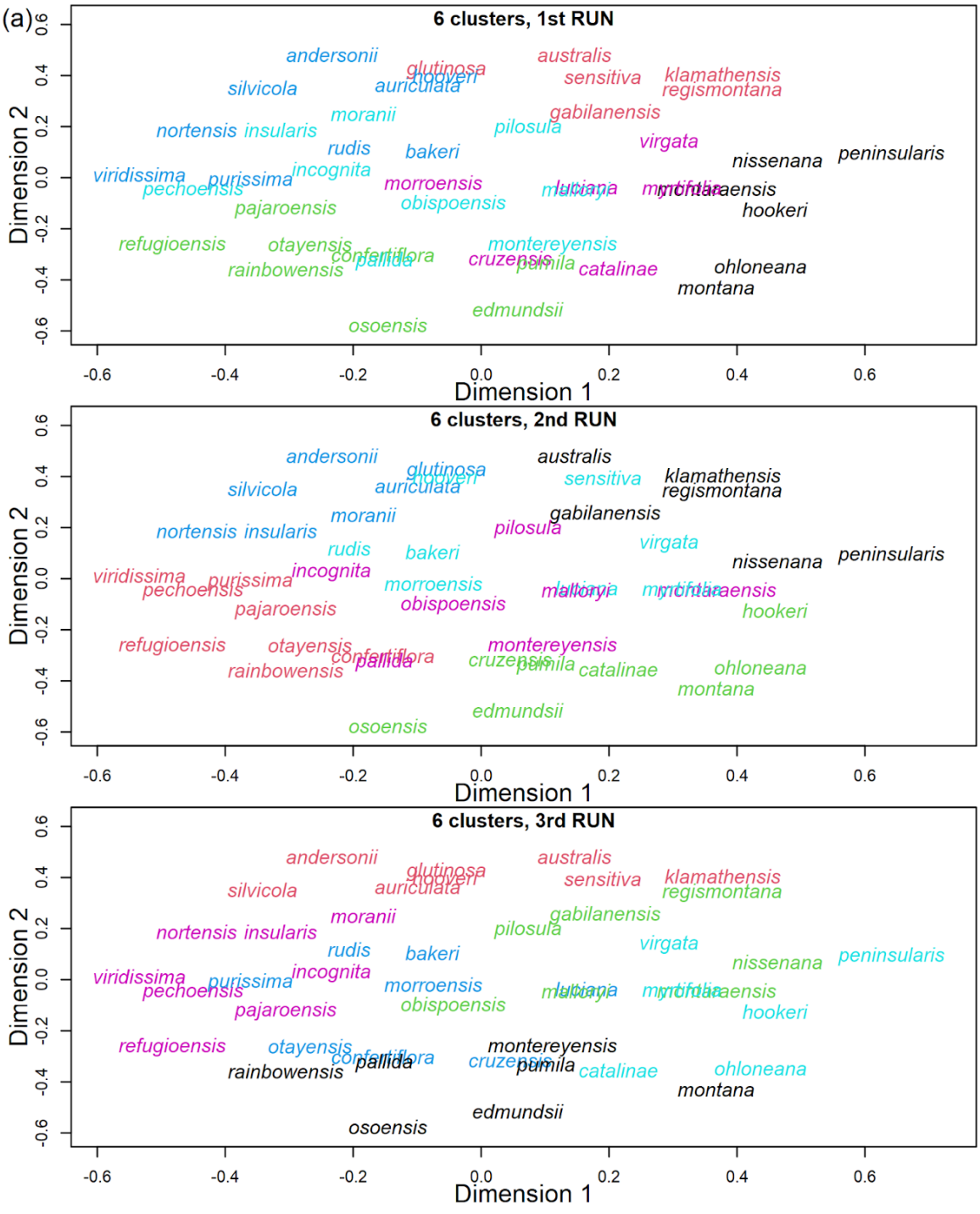
3.7 Appendix

Appendix S3.1 Visual assessment of both (a) 1 km and (b) 270 m analyses indicate the same three possible values (6, 7, and 8) as the optimal number of clusters in the k-means clustering analyses. These optimal numbers of clusters are selected because they are the inflection points of the curve and are framed in a red rectangle in each plot.

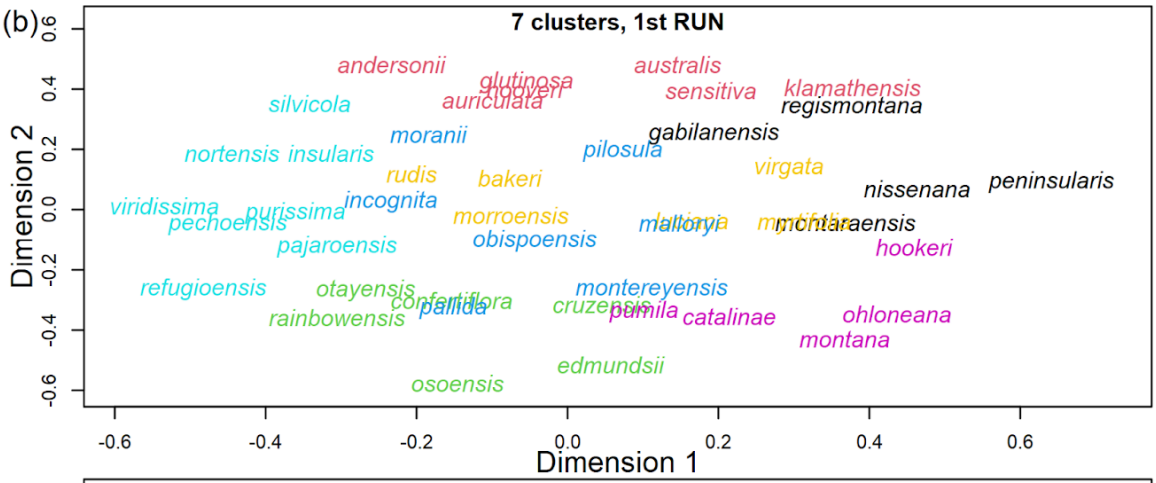


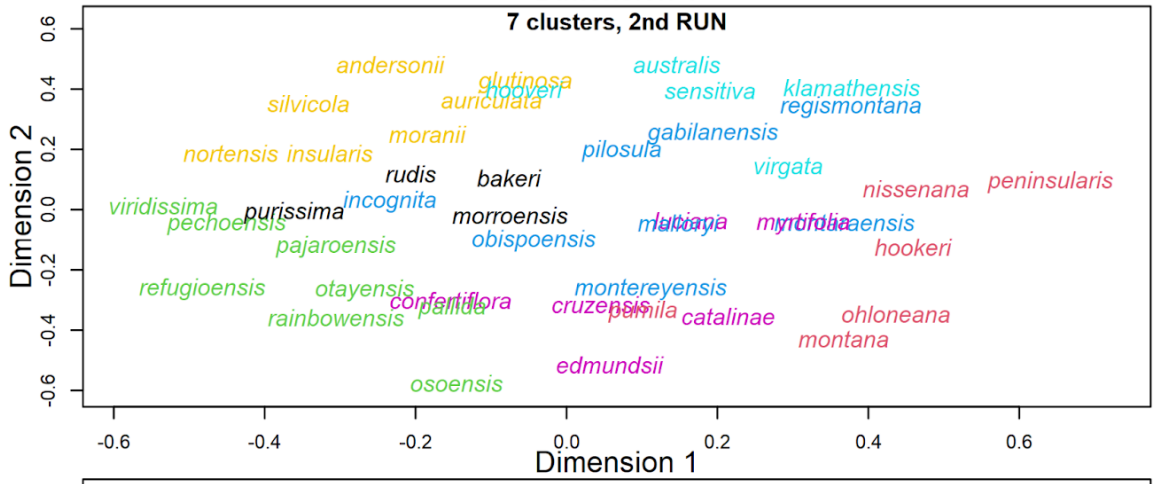
Appendix S3.2 Results of the k-means clustering analyses using the 1 km dataset. Analyses were run with k's of 6 (a), 7 (b), and 8 (c), determined to be the optimal number of clusters, and three analyses were run at each k. Regardless of the optimal number of clusters used, the assignment of species to clusters changed in different attempts. Species are colored by cluster assignment.

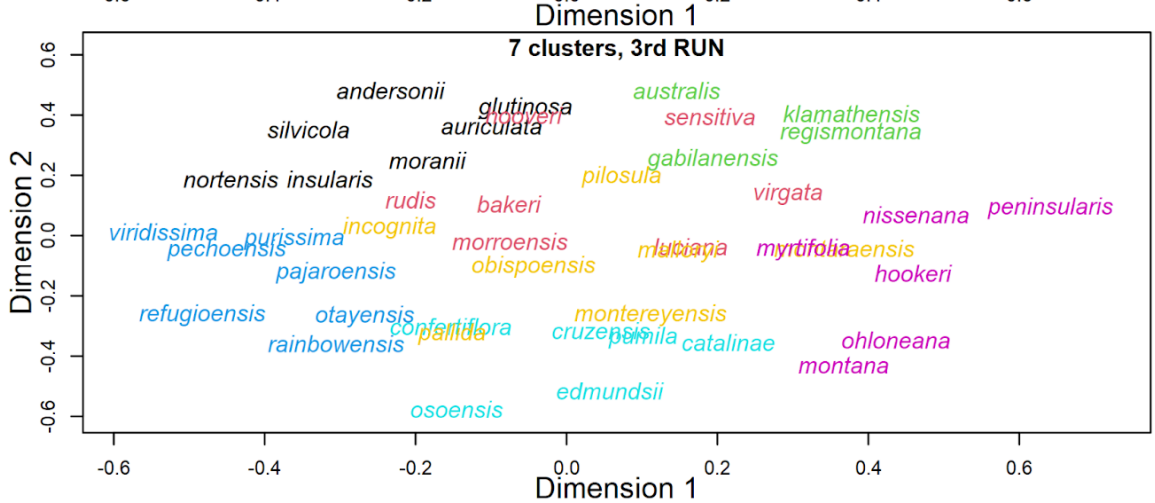
(a)


6 clusters, 1st RUN
andersonii
australis
glutinosa
hoopesii
sensitiva
klamathensis
silvicola
auriculata
regismontana
moranii
gabilanensis
nortensis
insularis
pilosula
virgata
rudis
bakeri
nissenana
peninsularis
viridissima
purissima
incognita
pechoensis
morroensis
luciana
montana
hookeri
pajaroensis
obispoensis
refugioensis
otayensis
montereyensis
confertiflora
cruzensis
pumila
catalinae
ohloneana
rainbowensis
pallida
edmundsii
osoensis
Dimension 1
Dimension 2
6 clusters, 2nd RUN
6 clusters, 3rd RUN

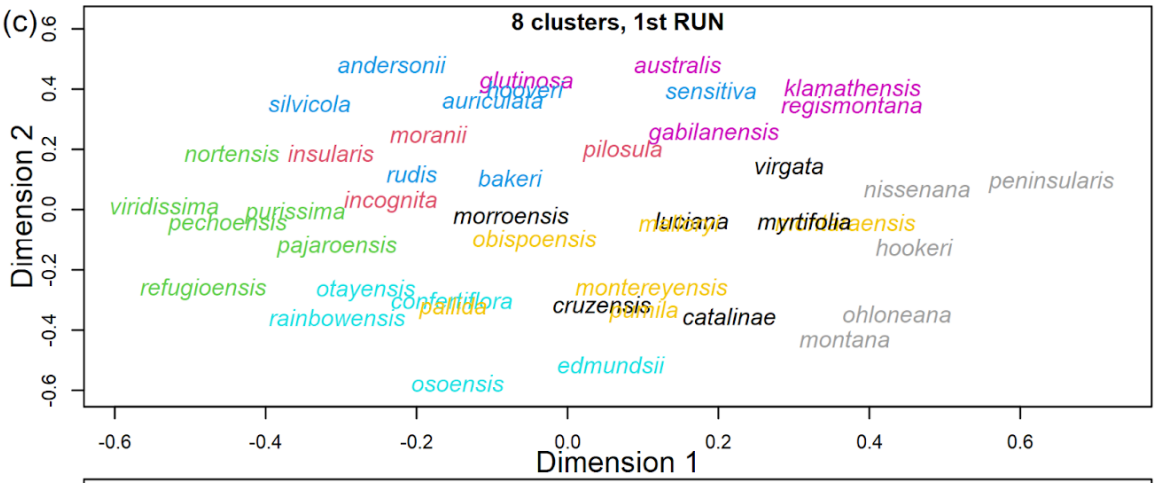
(b)

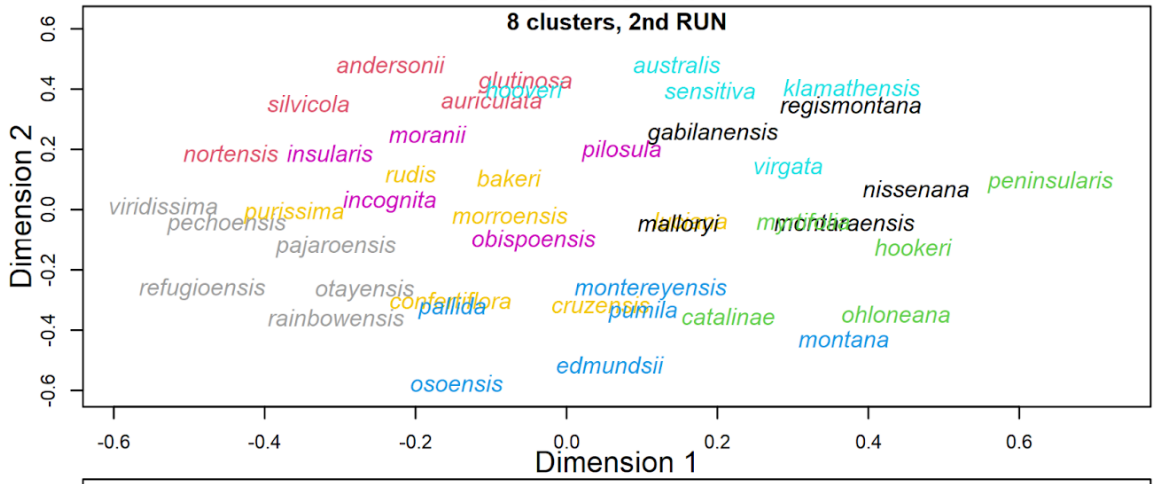

7 clusters, 1st RUN
Dimension 2
Dimension 1
andersonii
glutinosa
hooveri
australis
sensitiva
klamathensis
regismontana
silvicola
auriculata
moranii
gabilanensis
nortensis insularis
pilosula
virgata
rudis
bakeri
nissenana
peninsularis
viridissima
purissima
incognita
morroensis
pechoensis
mallory
michtaraensis
pajaroensis
obispoensis
hookeri
refugioensis
otayensis
montereyensis
confertiflora
cruzensis
pallida
pumila
catalinae
ohloneana
rainbowensis
montana
edmundsii
osoensis

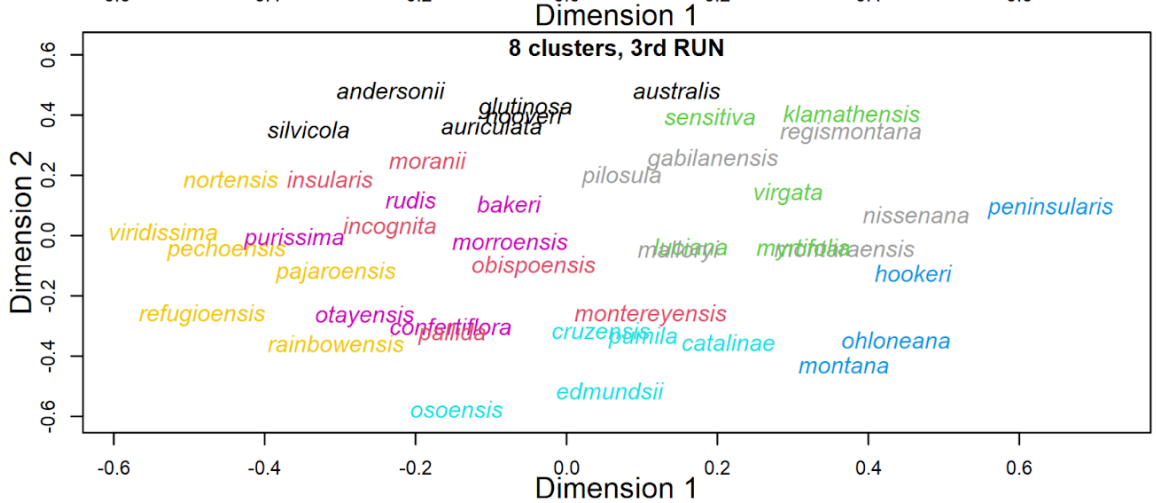

7 clusters, 2nd RUN
Dimension 2
Dimension 1
andersonii
glutinosa
hooveri
australis
sensitiva
klamathensis
regismontana
silvicola
auriculata
moranii
gabilanensis
nortensis insularis
pilosula
virgata
rudis
bakeri
nissenana
peninsularis
viridissima
purissima
incognita
morroensis
pechoensis
mallory
michtaraensis
pajaroensis
obispoensis
hookeri
refugioensis
otayensis
montereyensis
confertiflora
cruzensis
pallida
pumila
catalinae
ohloneana
rainbowensis
montana
edmundsii
osoensis


7 clusters, 3rd RUN
Dimension 2
Dimension 1
andersonii
glutinosa
hooveri
australis
sensitiva
klamathensis
regismontana
silvicola
auriculata
moranii
gabilanensis
nortensis insularis
pilosula
virgata
rudis
bakeri
nissenana
peninsularis
viridissima
purissima
incognita
morroensis
pechoensis
mallory
michtaraensis
pajaroensis
obispoensis
hookeri
refugioensis
otayensis
montereyensis
confertiflora
cruzensis
pallida
pumila
catalinae
ohloneana
rainbowensis
montana
edmundsii
osoensis

(c)

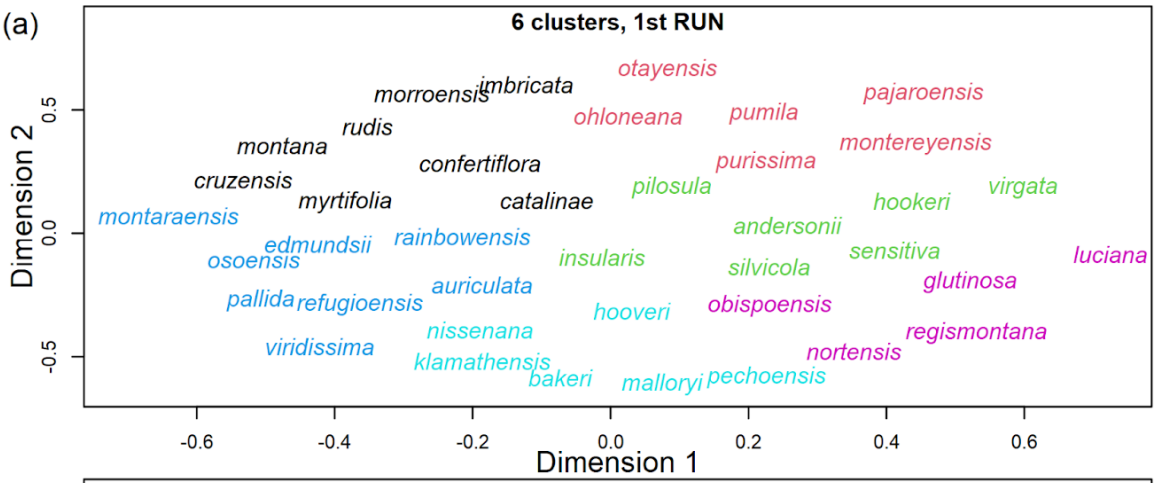

8 clusters, 1st RUN
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
andersonii
australis
glutinosa
hooveri
sensitiva
klamathensis
silvicola
auriculata
regismontana
moranii
gabilanensis
nortensis
insularis
pilosula
rudis
bakeri
virgata
nissenana
peninsularis
viridissima
purissima
incognita
morroensis
luciana
myrtifolia
pechoensis
obispoensis
hookeri
pajaroensis
refugioensis
otayensis
montereyensis
confertiflora
pallida
cruzensis
pumila
catalinae
ohloneana
rainbowensis
montana
edmundsii
osoensis

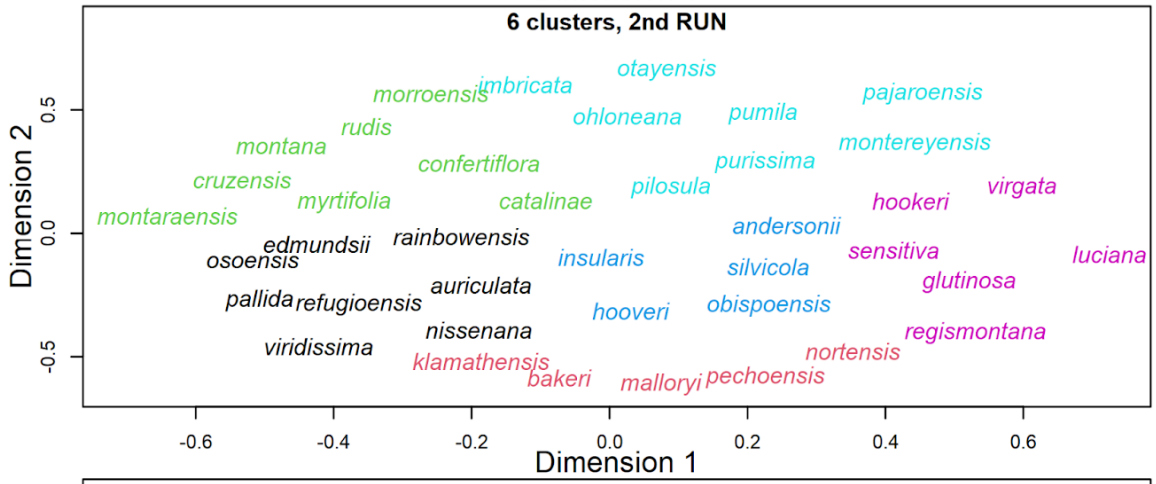

8 clusters, 2nd RUN
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
andersonii
australis
glutinosa
hooveri
sensitiva
klamathensis
silvicola
auriculata
regismontana
moranii
gabilanensis
nortensis
insularis
pilosula
rudis
bakeri
virgata
nissenana
peninsularis
viridissima
purissima
incognita
morroensis
malloryi
montaraensis
pechoensis
obispoensis
hookeri
pajaroensis
refugioensis
otayensis
montereyensis
confertiflora
pallida
cruzensis
pumila
catalinae
ohloneana
rainbowensis
montana
edmundsii
osoensis

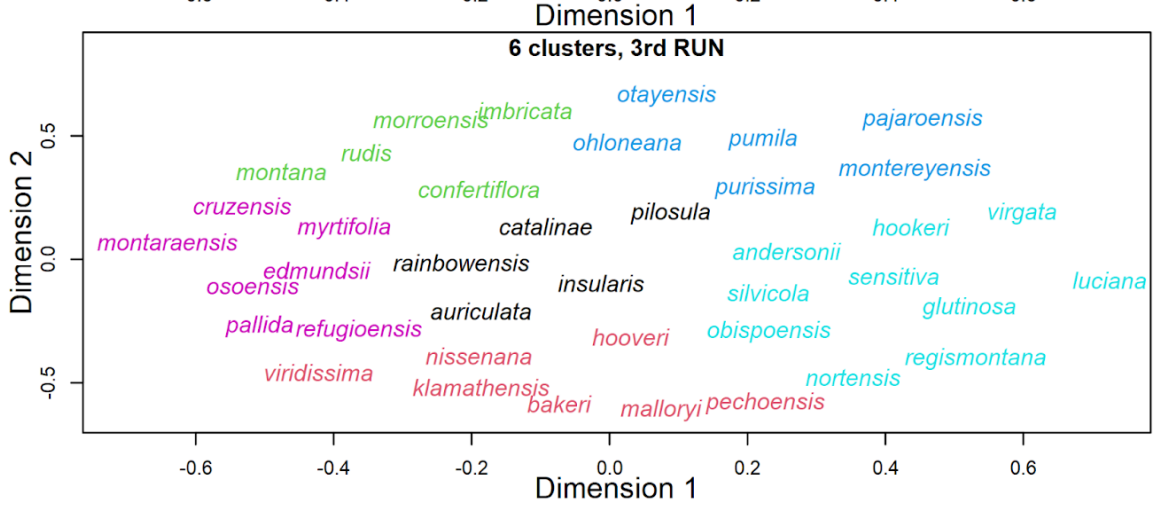

8 clusters, 3rd RUN
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
andersonii
australis
glutinosa
hooveri
sensitiva
klamathensis
silvicola
auriculata
regismontana
moranii
gabilanensis
nortensis
insularis
pilosula
rudis
bakeri
virgata
nissenana
peninsularis
viridissima
purissima
incognita
morroensis
malloryi
myrtifolia
pechoensis
obispoensis
hookeri
pajaroensis
refugioensis
otayensis
montereyensis
confertiflora
pallida
cruzensis
pumila
catalinae
ohloneana
rainbowensis
montana
edmundsii
osoensis

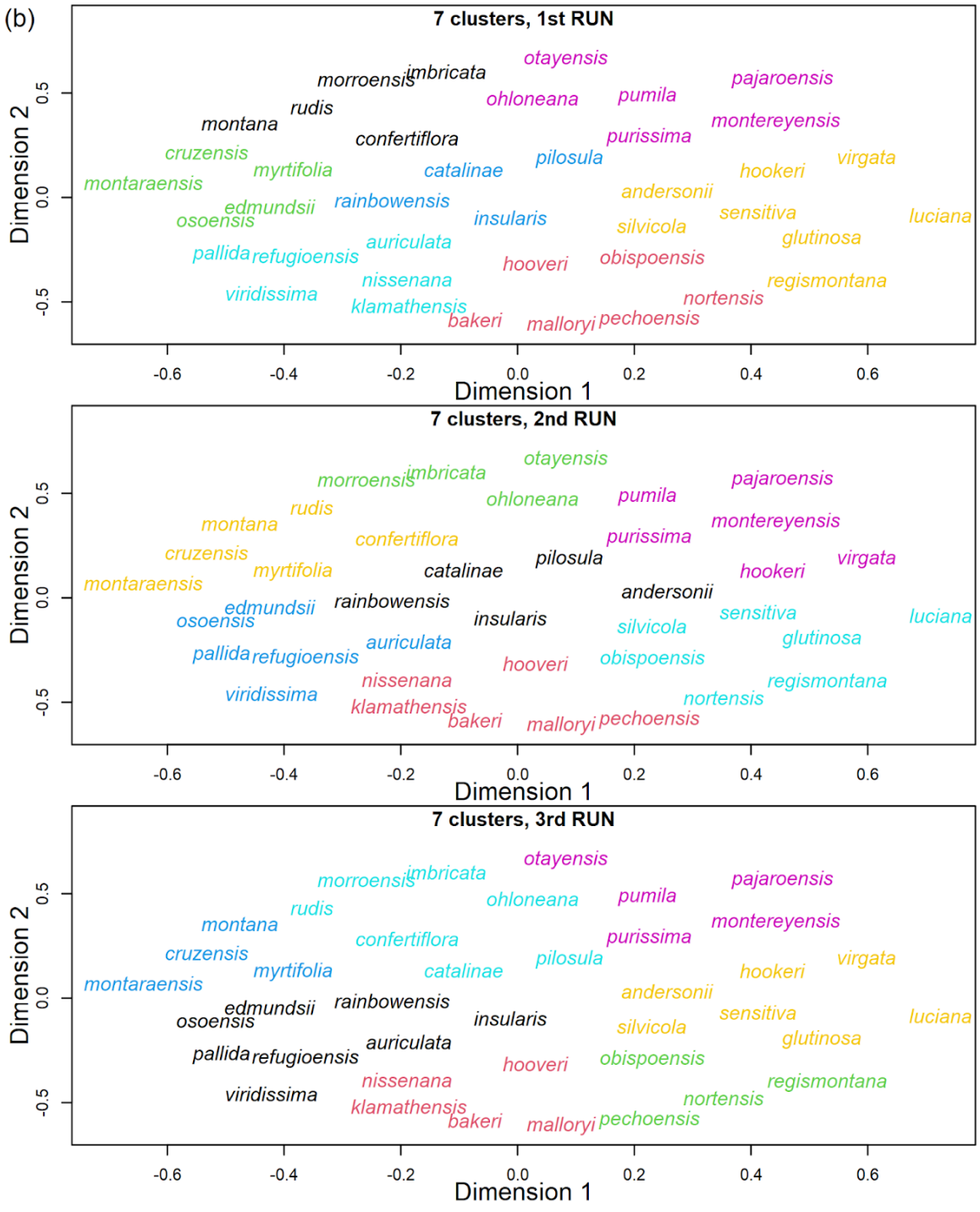
Appendix S3.3 Results of the k-means clustering analyses using the 270 m dataset. Analyses were run with k's of 6 (a), 7 (b), and 8 (c), determined to be the optimal number of clusters, and three analyses were run at each k. Regardless of the optimal number of clusters used, the assignment of species to clusters changed in different attempts. Species are colored by cluster assignment.

(a)

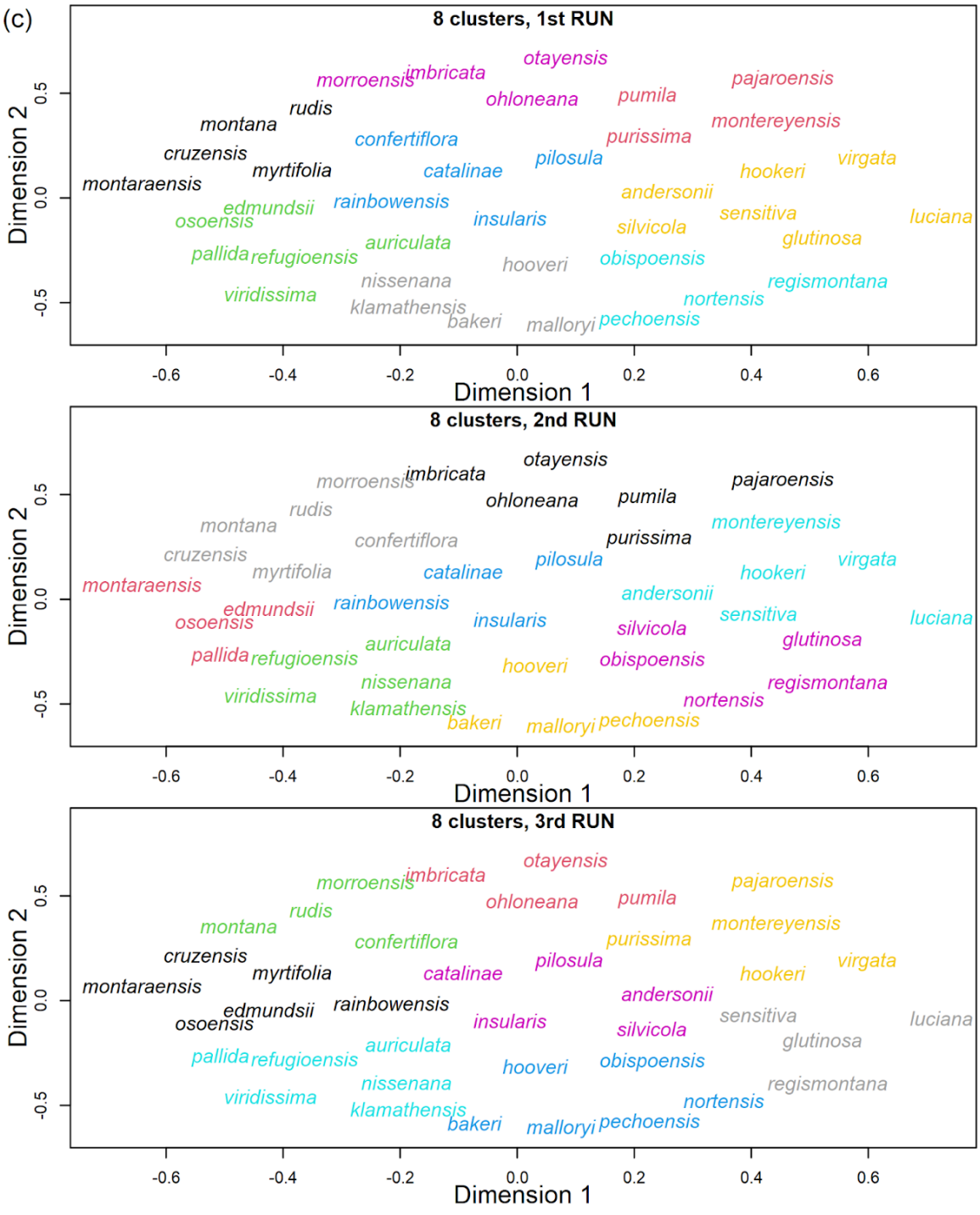

6 clusters, 1st RUN
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.5
0.0
-0.5
otayensis
imbricata
morroensis
pajaroensis
ohloneana
pumila
rudis
montereyensis
montana
purissima
confertiflora
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
edmundsii
rainbowensis
sensitiva
luciana
osoensis
insularis
silvicola
glutinosa
auriculata
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
nortensis
klamathensis
bakeri
malloryi
pechoensis


6 clusters, 2nd RUN
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.5
0.0
-0.5
otayensis
imbricata
morroensis
pajaroensis
ohloneana
pumila
rudis
montereyensis
montana
purissima
confertiflora
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
edmundsii
rainbowensis
sensitiva
luciana
osoensis
insularis
silvicola
glutinosa
auriculata
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
nortensis
klamathensis
bakeri
malloryi
pechoensis

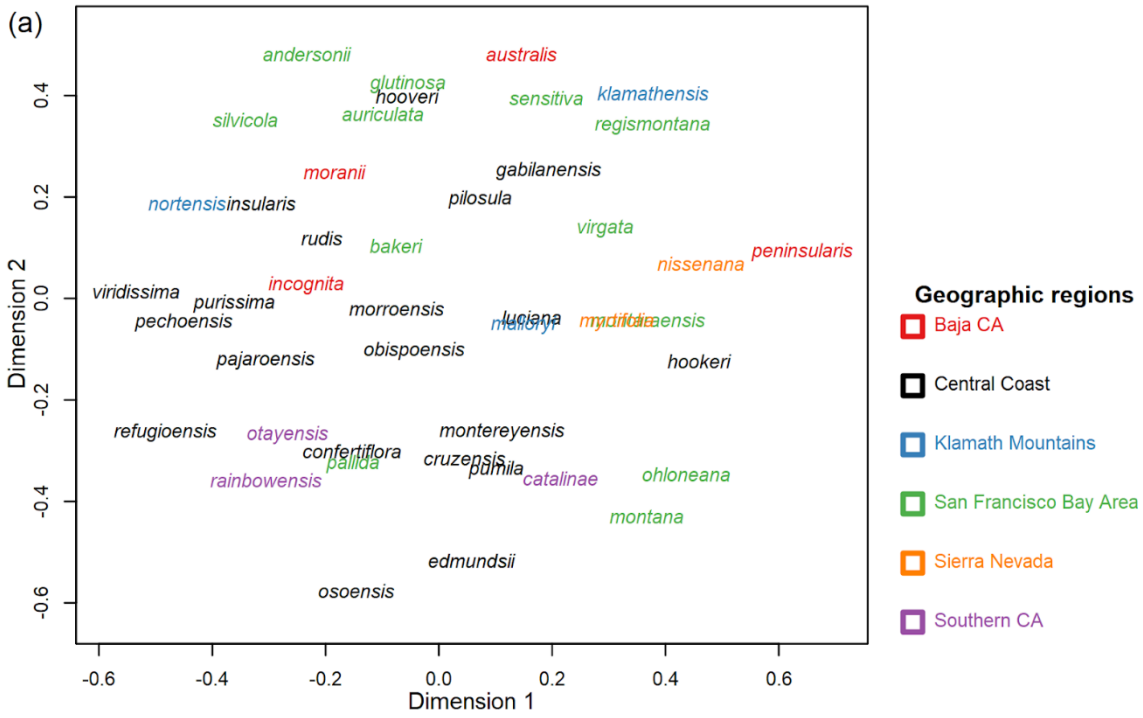

6 clusters, 3rd RUN
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.5
0.0
-0.5
otayensis
imbricata
morroensis
pajaroensis
ohloneana
pumila
rudis
montereyensis
montana
purissima
confertiflora
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
edmundsii
rainbowensis
sensitiva
luciana
osoensis
insularis
silvicola
glutinosa
auriculata
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
nortensis
klamathensis
bakeri
malloryi
pechoensis

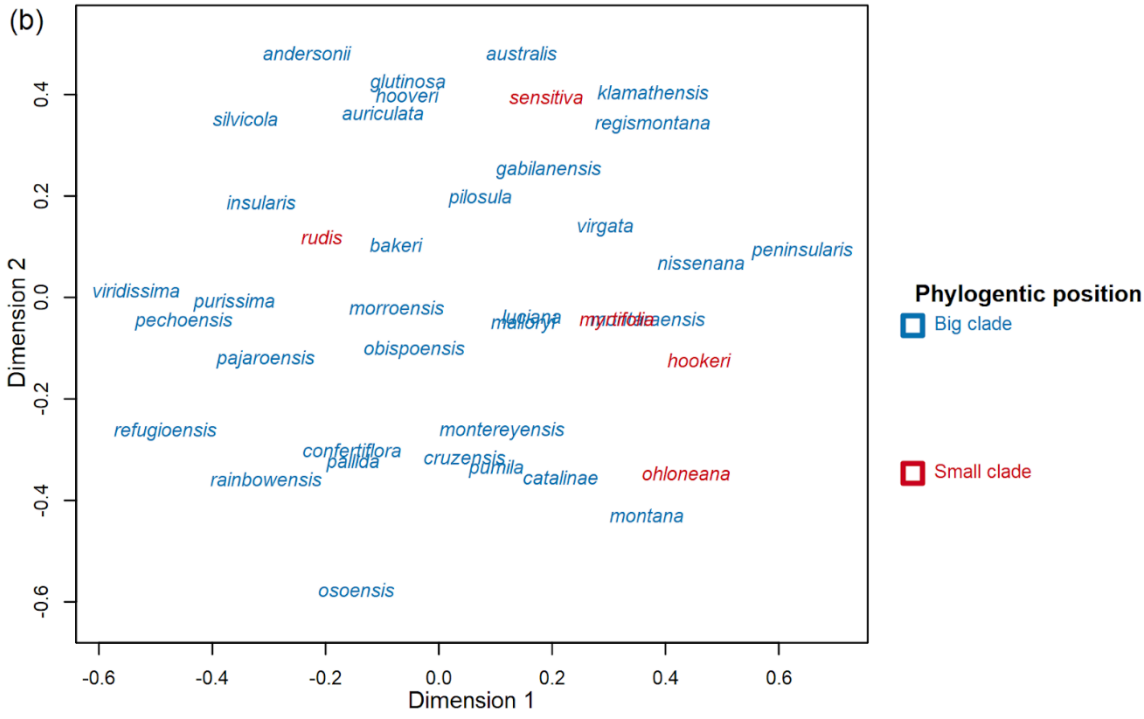

(b)
7 clusters, 1st RUN
otayensis
morroensis
imbricata
pajaroensis
ohloneana
pumila
rudis
montana
montereyensis
confertiflora
purissima
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
edmundsii
rainbowensis
sensitiva
luciana
osoensis
insularis
silvicola
glutinosa
auriculata
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
klamathensis
nortensis
bakeri
malloryi
pechoensis
Dimension 2
Dimension 1
7 clusters, 2nd RUN
7 clusters, 3rd RUN

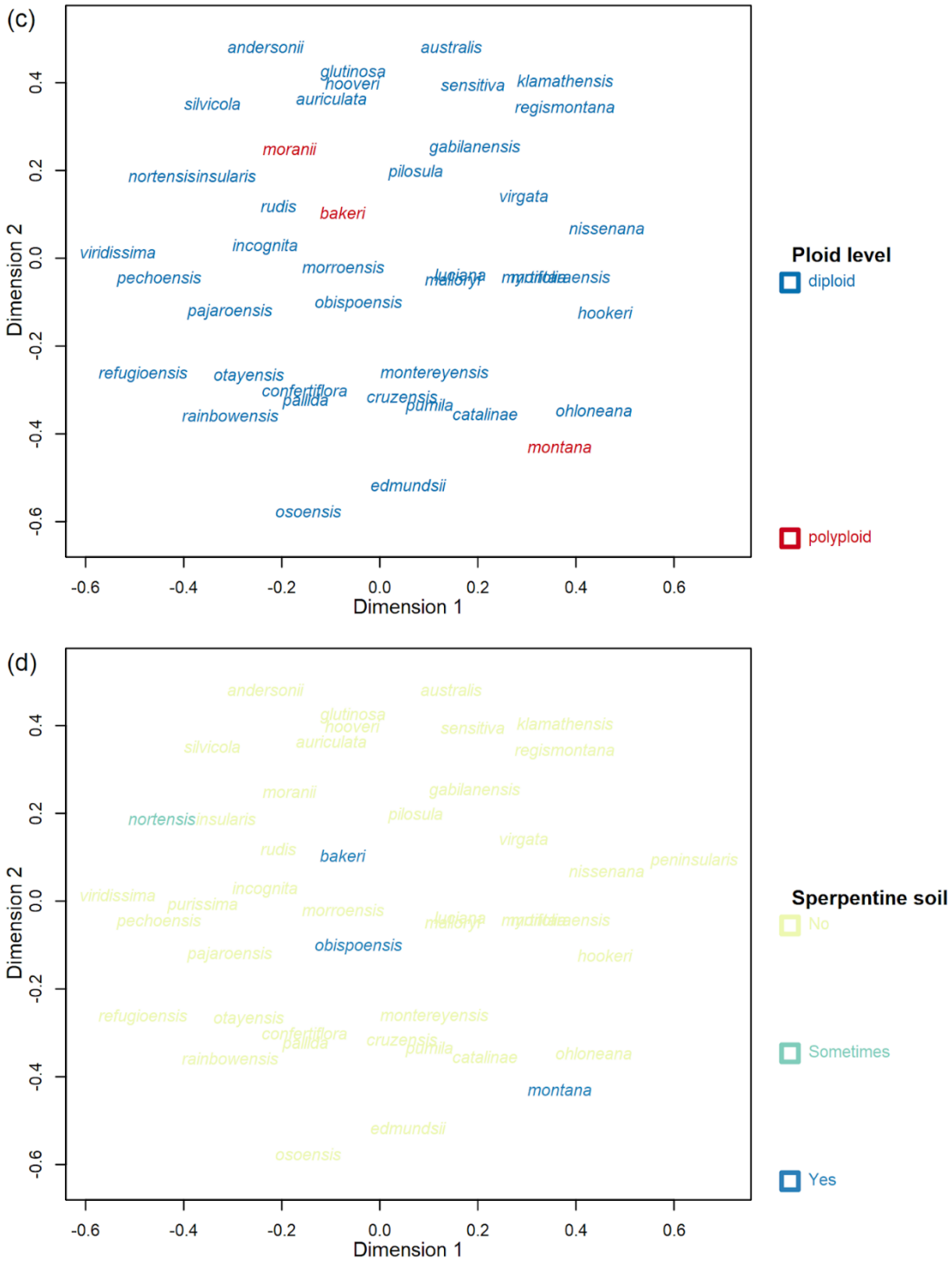
(c)


8 clusters, 1st RUN
otayensis
morroensis
imbricata
pajaroensis
ohloneana
pumila
rudis
montana
montereyensis
confertiflora
purissima
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
edmundsii
rainbowensis
sensitiva
luciana
osoensis
insularis
silvicola
glutinosa
auriculata
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
nortensis
klamathensis
bakeri
malloryi
pechoensis
Dimension 2
Dimension 1
8 clusters, 2nd RUN
8 clusters, 3rd RUN

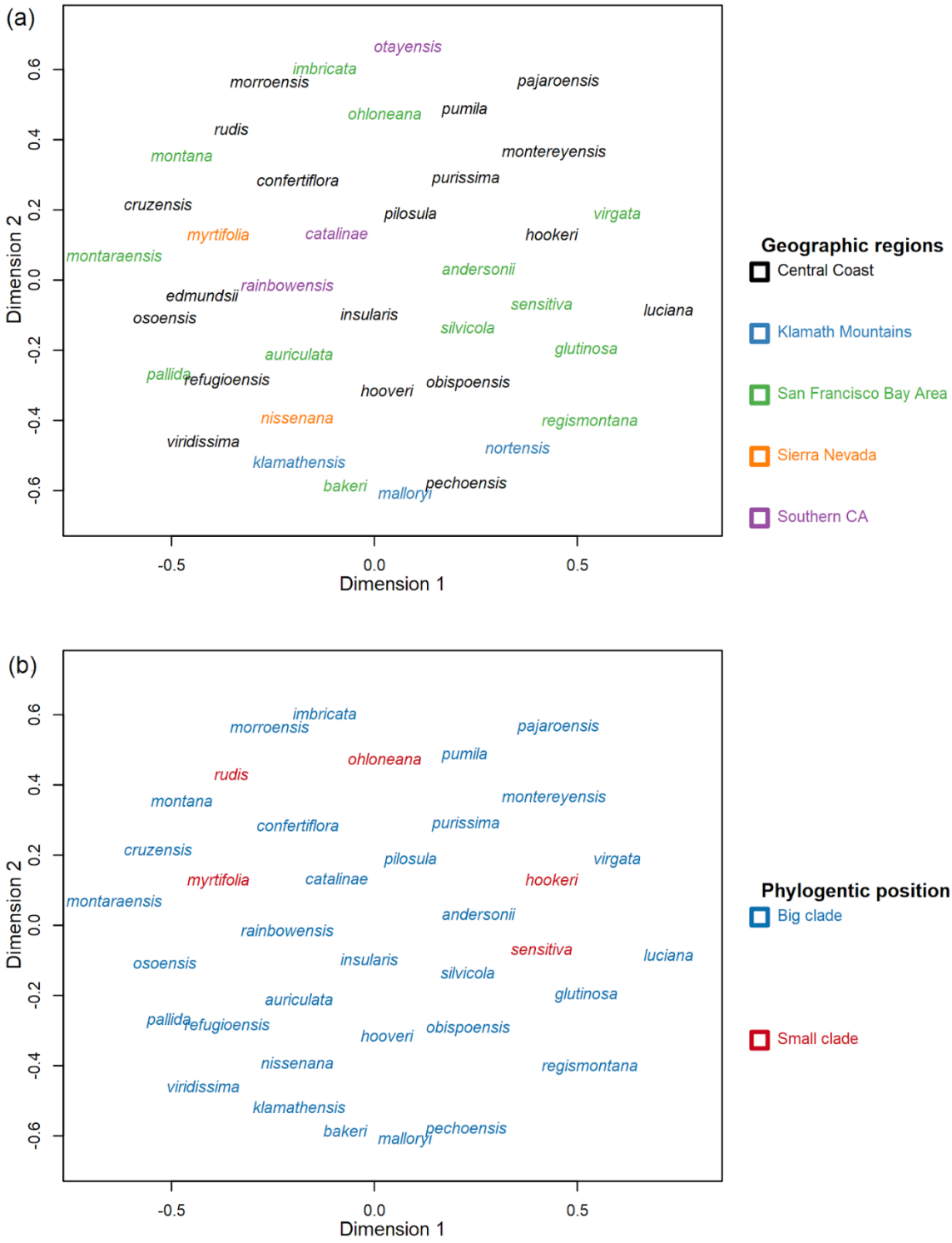
Appendix S3.4: Mapping species attributes to the MDS plots using the 1 km dataset shows no correlation between (a) geographic region, (b) phylogenetic assignment, (c) ploidy level, and (d) the presence/absence of serpentine soil in the habitat. For all plots, the distance between the species labels represents the ecological distance between the species based on MDS1 and MDS2. MDS3 is not depicted but did not change the interpretation of clustering patterns. The species are color-coded according to their attributes. Species of unknown phylogenetic position, chromosome number, or soil type are colored in white to make them invisible.

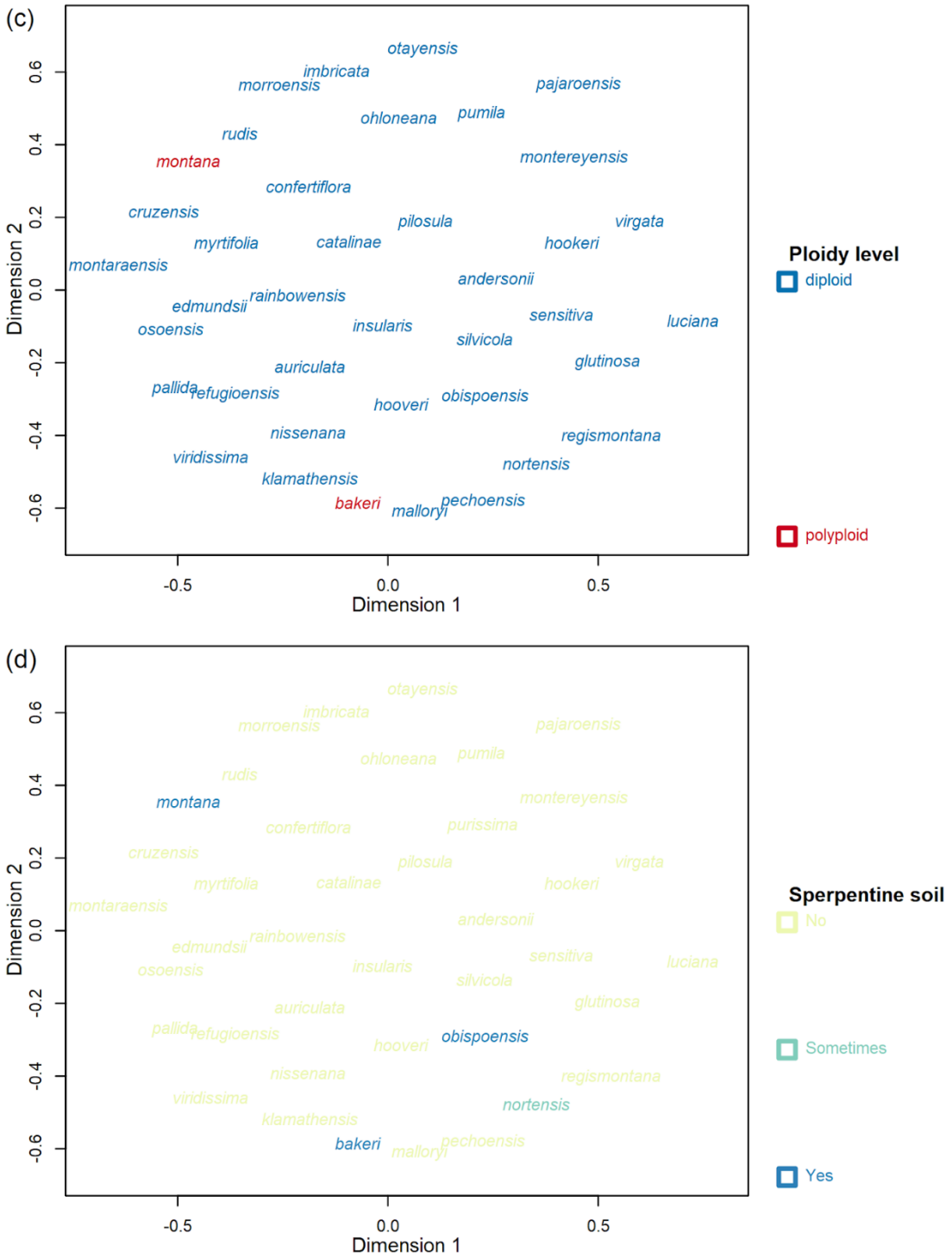

(a)
andersonii
australis
glutinosa
hooveri
sensitiva
klamathensis
silvicola
auriculata
regismontana
moranii
gabilanensis
nortensis
insularis
pilosula
virgata
rudis
bakeri
peninsularis
nissenana
incognita
viridissima
purissima
morroensis
luciana
pechoensis
mariolyi
pajaroensis
obispoensis
hookeri
refugioensis
otayensis
montereyensis
confertiflora
cruzensis
palilda
pumila
catalinae
ohloneana
rainbowensis
montana
edmundsii
osoensis
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Geographic regions
Baja CA
Central Coast
Klamath Mountains
San Francisco Bay Area
Sierra Nevada
Southern CA


(b)
andersonii
australis
glutinosa
hooveri
sensitiva
klamathensis
silvicola
auriculata
regismontana
gabilanensis
insularis
pilosula
virgata
rudis
bakeri
peninsularis
nissenana
viridissima
purissima
morroensis
luciana
pechoensis
mariolyi
pajaroensis
obispoensis
hookeri
refugioensis
montereyensis
confertiflora
cruzensis
palilda
pumila
catalinae
ohloneana
rainbowensis
montana
osoensis
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Phylogentic position
Big clade
Small clade


(c)
andersonii
australis
glutinosa
hooveri
sensitiva
klamathensis
silvicola
auriculata
regismontana
moranii
gabilanensis
nortensisinsularis
pilosula
virgata
rudis
bakeri
nissenana
viridissima
incognita
morroensis
pechoensis
pajaroensis
obispoensis
hookeri
refugioensis
otayensis
montereyensis
confertiflora
pallida
cruzensis
pumila
catalinae
ohloneana
rainbowensis
montana
edmundsii
osoensis
Dimension 2
Dimension 1
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Ploid level
diploid
polyploid
(d)
Sperpentine soil
No
Sometimes
Yes

Appendix S3.5 Mapping species attributes to the MDS plots using the 1 km dataset shows no correlation between (a) geographic region, (b) phylogenetic assignment, (c) ploidy level, and (d) the presence/absence of serpentine soil in the habitat. For all plots, the distance between the species labels represents the ecological distance between the species based on MDS1 and MDS2. MDS3 is not depicted but did not change the interpretation of clustering patterns. The species are color-coded according to their attributes. Species of unknown phylogenetic position, chromosome number, or soil type are colored in white to make them invisible.


(a)
otayensis
imbricata
morroensis
pajaroensis
ohloneana
pumila
rudis
montana
montereyensis
confertiflora
purissima
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
rainbowensis
edmundsii
sensitiva
luciana
osoensis
insularis
silvicola
auriculata
glutinosa
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
nortensis
klamathensis
bakeri
mallory
pechoensis
Dimension 2
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.5
0.0
0.5
Dimension 1
Geographic regions
Central Coast
Klamath Mountains
San Francisco Bay Area
Sierra Nevada
Southern CA
(b)
imbricata
morroensis
pajaroensis
ohloneana
pumila
rudis
montana
montereyensis
confertiflora
purissima
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
rainbowensis
sensitiva
luciana
osoensis
insularis
silvicola
auriculata
glutinosa
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
klamathensis
bakeri
mallory
pechoensis
Dimension 2
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.5
0.0
0.5
Dimension 1
Phylogentic position
Big clade
Small clade


(c)
otayensis
imbricata
morroensis
pajaroensis
ohloneana
pumila
rudis
montana
montereyensis
confertiflora
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
rainbowensis
edmundsii
sensitiva
luciana
osoensis
insularis
silvicola
glutinosa
auriculata
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
nortensis
klamathensis
bakeri
mallory
pechoensis
Dimension 2
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.5
0.0
0.5
Dimension 1
Ploidy level
diploid
polyploid
(d)
otayensis
imbricata
morroensis
pajaroensis
ohloneana
pumila
rudis
montana
montereyensis
confertiflora
purissima
cruzensis
pilosula
virgata
myrtifolia
catalinae
hookeri
montaraensis
andersonii
rainbowensis
edmundsii
sensitiva
luciana
osoensis
insularis
silvicola
glutinosa
auriculata
pallida
refugioensis
hooveri
obispoensis
nissenana
regismontana
viridissima
nortensis
klamathensis
bakeri
mallory
pechoensis
Dimension 2
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.5
0.0
0.5
Dimension 1
Serpentine soil
No
Sometimes
Yes

4 Chapter 4 Subspecies differentiation in an enigmatic chaparral shrub species

4.1 Introduction

A comprehensive understanding of biodiversity is crucial for ecologists, conservationists, land managers, policy makers, and others whose work depends on the accurate recognition of biodiversity units (Regan, Colyvan, and Burgman 2002; Mace 2004; Keller et al. 2011; Renwick et al. 2017). Given current rates of extinction in plants, discovering, identifying, and delimiting plant biodiversity units is more critical than ever (Thomas et al. 2004; Pimm et al. 2014; Grooten and Almond 2018). However, drawing boundaries around taxonomic units is difficult in groups in which populations are poorly differentiated and/or vary along a continuous cline (Carstens et al. 2013; Jörger and Schrödl 2013; Razkin et al. 2017; Bradburd, Coop, and Ralph 2018). A number of factors can blur boundaries, including hybridization, introgression, local adaptation, and phenotypic plasticity (Grant 1981a; Harrison and Larson 2014). These factors can result in highly variable populations and species complexes, which include a range of phenotypes that cannot be easily divided into distinct groups using standard genetic, morphological, or ecological criteria.

Eastwood manzanita (*Arctostaphylos glandulosa* Eastw., Ericaceae), a widespread tetraploid shrub found in the chaparral of southern Oregon, California, and the Baja California (MX) peninsula, is a phenotypically complex system in which delimiting subspecies has been challenging (Keeley, Vasey, and Parker 2007; Baldwin et al. 2012; Kauffmann et al. 2015). Typical of manzanitas, Eastwood manzanita has twisting branches covered with red bark, small simple drought-adapted leaves, and clusters of white-to-pink urn-shaped flowers (Figure 4.1). Plants of this species produce a burl—a large woody structure that develops where roots and stem meet and that

contains dormant buds (Jepson 1916; Wieslander and Schreiber 1939). This allows the plant to resprout and persist through numerous wildfires. Seed germination is fire-dependent. Thus, a population of Eastwood manzanita may include some individuals that are potentially hundreds or thousands of years old, having resprouted after multiple fires, as well as younger individuals that have grown from seed produced by individuals from the original population or brought in by dispersers from other populations (Keeley and Hays 1976; Moore and Vander Wall 2015; Parker 2015). Currently, 10 subspecies are recognized based on traits related to hair density and morphology, leaf color, inflorescence characters, and seed fusion, which have been defined in previous morphometric studies (Figure 4.1) (Keeley, Vasey, and Parker 2007; Baldwin et al. 2012). However, the morphological boundaries among subspecies can be indistinct, with intermediate phenotypes and individuals that do not conform to any one subspecies. Moreover, most subspecies overlap in their geographic range (Kauffmann et al. 2015), and multiple subspecies may be found in the same population, raising the need for a clearer understanding of the relationship between genetic, morphological, and ecological patterns among these subspecies.

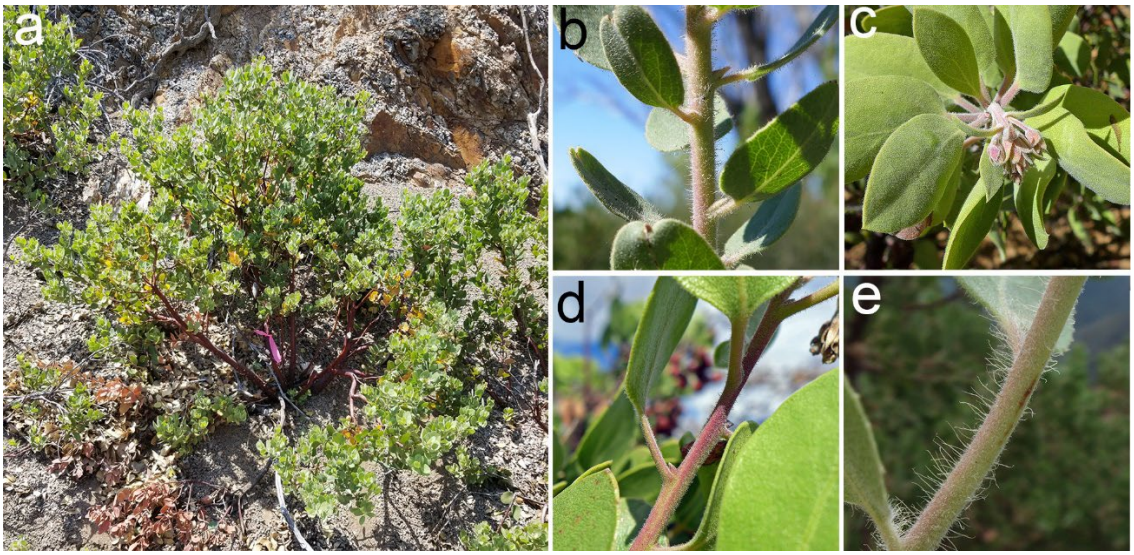


Figure 4.1 Variation in hair traits in subspecies of Eastwood manzanita. a. Eastwood manzanita, Santa Barbara County. b. subsp. *glandulosa*, short- and medium-length hairs, long hairs with terminal glands. c. subsp. *cushingiana*, dense short hairs lacking glands. d. subsp. *gabrielensis*, relatively sparse, short hairs lacking glands. e. subsp. *mollis*, short and very long wavy non-glandular hairs. Photo credits: a. A. Litt. b-d Neil Kramer. e. Michael Charters, CalFlora.net.

Characterizing subspecies differentiation in Eastwood manzanita is particularly critical because two of the currently recognized subspecies are narrow endemics of conservation concern (Figure 4.2). *Arctostaphylos glandulosa* subsp. *crassifolia* (Del Mar manzanita) is federally listed as endangered (https://ecos.fws.gov/), and this subspecies, along with *A. glandulosa* subsp. *gabrielensis* (San Gabriel manzanita), is listed as rare in the California Native Plant Society Inventory of Rare Plants (http://www.rareplants.cnps.org/). Therefore, the ability to distinguish these two subspecies is required for conservation management.

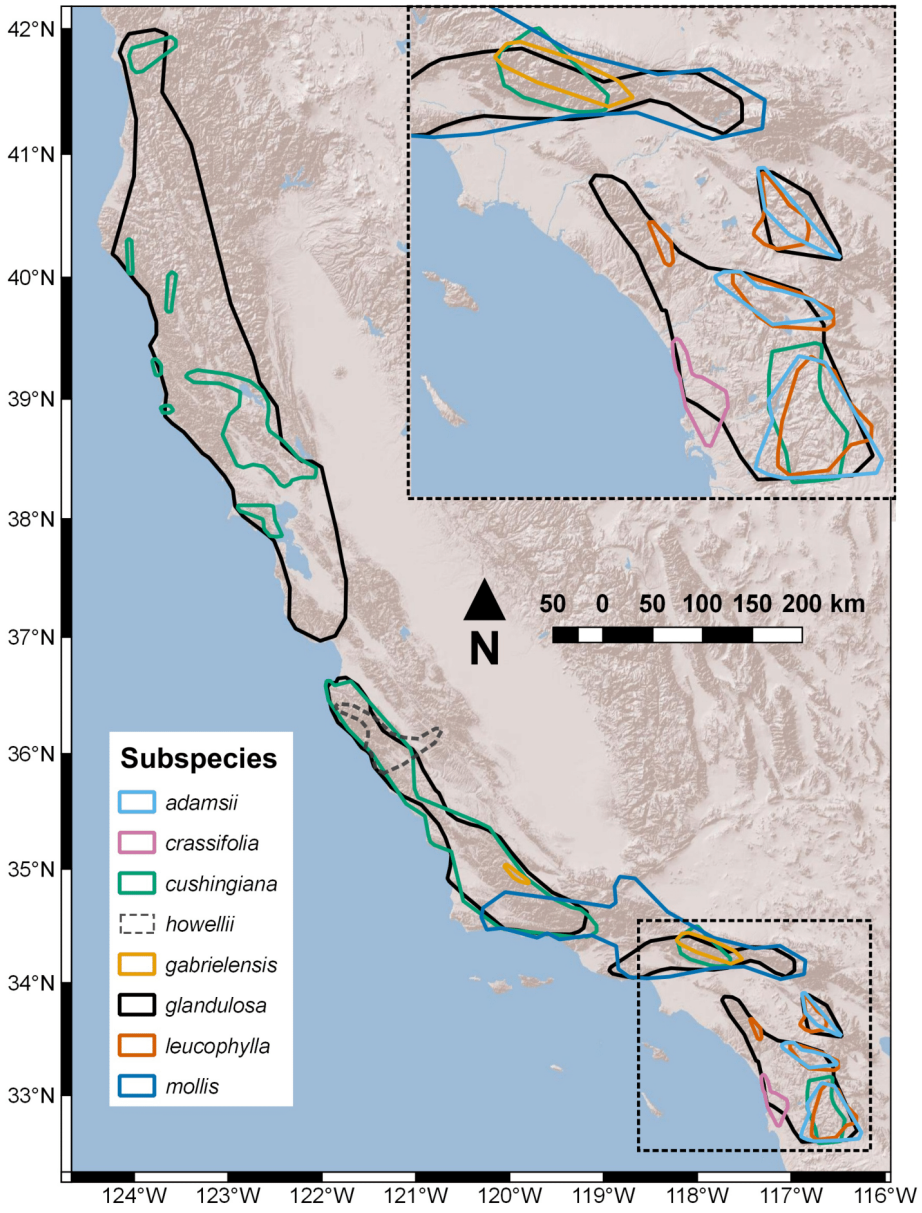


Figure 4.2 Map of California with the ranges of the 8 subspecies of Eastwood manzanita found in California. The two Mexican subspecies are not well-databased and therefore not included. Inset map at top-right shows southern California region, marked with a dotted box on the main map.

There is currently no agreed-upon definition for a subspecies, and relatively few authors have directly considered the concept of subspecies. Most authors define subspecies as conspecific groups of one or more populations that have evolutionary meaning (Patten and Unitt 2002; Patten 2015). Although various authors define "evolutionary meaning" differently, the phrase implies some level of genetic differentiation among subspecies because evolution requires the inheritance of allelic differences. However, previous methods for evaluating genetic variation have not always been capable of detecting differentiation, because of the limited nature of such data (Martien et al. 2017). Modern genomic techniques provide greater power to estimate genetic differentiation than previous methods, allowing us to better investigate evolutionary units (Harrison and Kidner 2011; Andrews et al. 2016).

In this study, we define subspecies as genetically differentiated populations within a species that have unique morphology or demonstrate a difference in adaptation to the local environment (Haig et al. 2006). Because previous work on Eastwood manzanita (Keeley, Vasey, and Parker 2007) indicated that in some populations morphological diagnosis of subspecies can be difficult to apply, we used next-generation sequencing data and online map-based resources to ask whether currently recognized subspecies are (1) genetically differentiated, and/or (2) environmentally differentiated, or if other genetic or environmental structures can be detected within this enigmatic species.

4.2 Materials and Method

4.2.1 Sampling

We collected 137 accessions from seven subspecies of Eastwood manzanita in Southern California (Figure 4.3). An additional three samples from coastal San Diego County, California and two samples of an eighth subspecies collected in northern Baja California (Burge et al. 2018) were included in this study. In addition to Eastwood manzanita, we included 17 samples of five diploid manzanita species collected from this same sampling area. This multispecies sample set was used for comparison of genetic differentiation among species-level taxa. Because a previous study that performed chromosome counts for many manzanita taxa reported no variation in ploidy within *A. glandulosa* (Wells 1968), we did not perform independent checks for ploidy on our *A. glandulosa* samples.

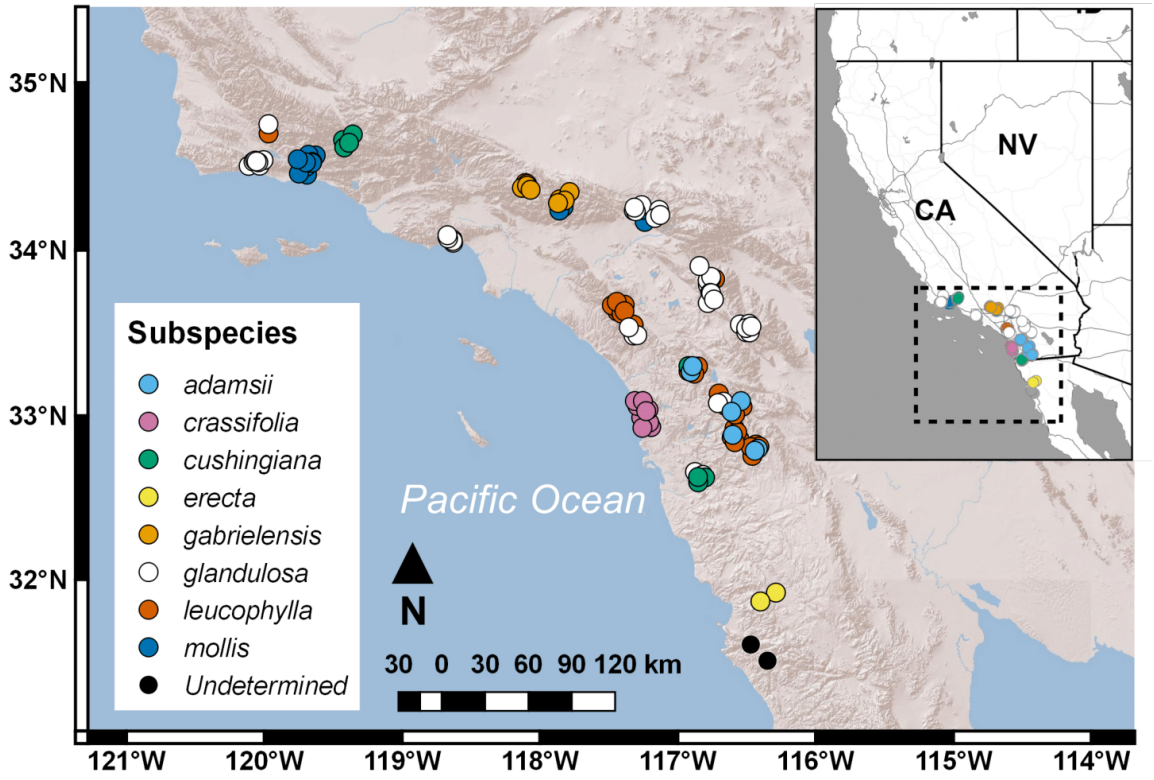


Figure 4.3 Map of collection localities for samples included in genetic analyses. Colors indicate subspecies identification. Because some samples would overlap on the map, a random "jitter" value between -0.15 and 0.15 degrees longitude and latitude was applied to each point. Inset shows the area of focus relative to the state of California.

4.2.2 Identification of samples

We identified samples to species and subspecies using the dichotomous key in the Field Guide to Manzanitas: California, North America and Mexico (Kauffmann et al. 2015), and, when identification was still unclear, we cross referenced with the dichotomous key in Keeley et al. (2007). Identifications were confirmed by V.T.P. and J.E.K. Subspecies identifications were thus in line with the most up-to-date taxonomy based on morphometric study by Keeley et al. (2007). In the case of two Eastwood

manzanita DNA samples from Burge et al. (2018), subspecies identification was not recorded; two others were identified as subspecies erecta. Voucher specimens were not available for those four collections and for three samples of the diploid species.

4.2.3 DNA extraction and quality control

We ground 150–200 mg frozen floral bud, young leaf, or flower tissue in liquid nitrogen, and used the Qiagen DNEasy Plant Mini Kit (Qiagen: Hilden, North Rhine-Westphalia, Germany) to extract DNA, modifying the protocol as described in Appendix S4.1. We used a Qubit 2.0 fluorometer (Invitrogen: Carlsbad, California, USA) to quantify the concentration, and checked for the presence of high molecular weight DNA by running extracts on a 1% agarose tris acetate ethylenediaminetetraacetic acid (TAE) gel (Invitrogen: Carlsbad, California, USA) and imaging on a Bio-Rad Gel Doc (Bio-Rad Laboratories: Hercules, California, USA).

4.2.4 Double digest restriction-site associated DNA sequencing (ddRAD-seq) library preparation and sequencing

One set of libraries (96 samples) was prepared and sequenced at LGC Genomics (Berlin, Germany) (Appendix S4.2). We prepared another set containing the remaining 58 samples at the University of California, Riverside (UCR), and sequenced them at the UCR Genomics Core Facility. The protocol used for library preparation at UCR (Appendix S4.3) was based on the protocol developed by Brelsford (Brelsford, Dufresnes, and Perrin 2016) [which was, in turn, based on the protocol outlined in Parchman (Parchman et al. 2012) and Peterson (Peterson et al. 2012)]. We modified this protocol to use the same enzymes as LGC and to incorporate the normalization step in the LGC protocol. Both libraries were sequenced using paired 150 bp reads on an

Illumina NextSeq 500 V2 (San Diego, California, USA), which was configured to provide ~1.5 million reads per sample. Tetraploid and diploid samples were sequenced for the same target number of reads and in the same sequencing lanes.

4.2.5 Sequence data processing

All data processing was done on the High Performance Computer Cluster at UCR. We removed adapter sequences, and verified that cut sites for both PstI and MspI were present for each read pair. To demultiplex sequences, we separated samples based on their indexed Illumina adapter sequence, and then by their inline barcode sequence using the process_radtags program in Stacks V. 2.1 (Catchen et al. 2013). We used the resulting sequence files to construct two sets of sequence data: one containing the 137 Eastwood manzanita samples ("Eastwood manzanita data set"), and one containing the five other species ("multispecies data set").

We called single nucleotide polymorphisms (SNPs) using freebayes V. 1.3.1 (Garrison and Marth 2012), a software tool that can call diploid or polyploid genotypes. Freebayes calls genotypes at a given locus by modeling all possible genotypes under a multinomial model, and treating reads as a random sample from each genotype to find the genotype call with the greatest probability (Garrison and Marth 2012). Because freebayes requires a reference genome, and no assembly is currently available for Arctostaphylos, we constructed a ddRAD reference file by (1) merging read pairs with PEAR V. 0.9.10 (Zhang et al. 2014), (2) sampling 200,000 reads from each merged sequence file, and (3) clustering these reads by 95% sequence similarity to form contigs using CD-HIT-EST V. 3.1.1 (Li and Godzik 2006). We constructed separate reference files for the Eastwood manzanita and multispecies data sets, and then aligned the

individual sequence files to each respective reference file using BWA-MEM in bwa V. 0.7.12 using default parameters (Li 2013). We then called variants from the aligned reads using the freebayes-parallel script provided in freebayes (Garrison and Marth 2012). We ran freebayes three separate times: (1) on the Eastwood manzanita data set assuming diploidy, (2) on the Eastwood manzanita data set assuming tetraploidy, and (3) on the multispecies data set assuming diploidy. Because the diploid and tetraploid samples were sequenced to the same target read number, we make the assumption that coverage per genome copy is greater for the diploid samples than for the tetraploid samples. However, the method we used to make our reference files is based on simple clustering of pooled sequences by similarity, so we make the assumption that differences in coverage per genome copy should not affect the reference construction.

We filtered the resulting variant call format (VCF) files to impose quality and variant type controls using custom bash and R scripts (R Core Team 2018). This filtering removed indels, multinucleotide polymorphisms (SNPs at successive nucleotides that may be linked), and combinations of different types of variants, and left only simple SNPs. We eliminated the other variants because they could not be analyzed using the methods we used. Additionally, we removed loci with greater than 20% missing data across samples, and loci that had a minor allele recorded for fewer than three samples. We refer to this tetraploid SNP data set as the “4N” data set. We also generated two data sets in which genotypes were called as diploid. For the “2N” SNP data set, we allowed up to two alleles per individual at any given locus, but allowed more than two alleles across all samples. Additionally, because many analyses require SNPs to be biallelic (having no more than two alleles across all samples at a given locus), we created a third SNP data set (“2N-biallelic”), by removing loci from the 2N SNP data set

that had more than two alleles total across all samples. The multispecies data set was processed the same as the 2N data set. After quality filtering, the 4N, 2N, 2N-biallelic, and multispecies SNP data sets contained 4,018, 3,395, 3,337, and 21,660 SNPs, respectively. We removed 11 Eastwood manzanita individuals that had ≥50% missing data, leaving 126 individuals for analyses.

4.2.6 SNP data processing

For downstream analyses, we converted the filtered VCF files to nexus format alignment files with custom scripts in R V. 3.6.0 (R Core Team 2018), using standard International Union of Pure and Applied Chemistry (IUPAC) ambiguity codes to represent heterozygosity while retaining information for each allele. For STRUCTURE analyses (Pritchard, Stephens, and Donnelly 2000), which require unlinked loci, we used custom R scripts to select only the first SNP from each unique RAD fragment.

4.2.7 Genetic distance analyses

We analyzed all data sets in SplitsTree4 V. 4.15.1 (Huson and Bryant 2006), using the uncorrected p measure of genetic distance, defined as the number of nucleotide differences between two sequences divided by the length of the sequences (Nei and Kumar 2000). We retained information for heterozygous genotypes by setting SplitsTree4 to average ambiguous states in the calculation of uncorrected p. We computed a network visualization for each data set using the NeighborNetwork method (Bryant and Moulton 2002). To visualize patterns across these networks, we used the phangorn package (Schliep et al. 2016) in R. Additionally, we used multidimensional scaling analysis (MDS) (Gower 1966), calculated using the cmdscale function in R, to visualize genetic distances. We investigated the results of MDS models calculated with

as many as four dimensions, but the results yielded no additional geographic or taxonomic pattern beyond those of the two-dimensional MDS. We therefore calculated our MDS analyses with two dimensions. To search for patterns of clustering within the MDS results, we performed *k*-means clustering (Forgy 1965; MacQueen and Others 1967) using the kmeans function in R (R Core Team 2018). We used three statistical methods, implemented in the factoextra R package (Kassambara and Mundt 2016), to determine the optimal number of clusters: the within-group sum-of-squares, average silhouette width, and gap statistic methods (Forgy 1965; Rousseeuw 1987; Tibshirani, Walther, and Hastie 2001). To test whether genetic differentiation among samples is associated with geographic distance, we performed a simple Mantel test (Sokal 1979) in the R package ade4 (Dray, Dufour, and Others 2007).

In addition to MDS, we performed nonmetric multidimensional scaling (NMDS) using the prabclust function in the R package prablus (Hennig and Hausdorf 2020). This package calculates pairwise shared allele distances, defined as one minus the proportion of alleles shared between two samples (ignoring loci with missing data for one or both samples) (Bowcock et al. 1994), and performs NMDS on this distance matrix. We selected nine dimensions for the NMDS, because this was the least number of dimensions that yielded a stress value of less than 0.1 (Clarke 1993). Because of the high dimensionality of the NMDS analysis, we could not visually inspect the result so we instead performed Gaussian clustering implemented in the prabclust function, to sort samples into clusters. We tested clustering results ranging from two to four clusters (k = 2 to 4), and used leave-one-out cross-validation (Friedman, Hastie, and Tibshirani 2001) to evaluate whether the clusters were distinct at each value of k. We then compared clustering results with taxonomic determination and geographic pattern. Because the

clustering results at k = 4 provided no additional geographic or taxonomic insight when compared with the result at k = 3, we did not perform clustering for higher values of k. As the NMDS implemented in prabclust can only use diploid, biallelic SNPs (Hennig and Hausdorf 2020), we used the 2N-biallelic SNP data for this analysis, and could not compare results of Gaussian clustering among the three Eastwood manzanita SNP data sets.

4.2.8 STRUCTURE analysis

For each Eastwood manzanita data set, we used the ParallelStructure package (Besnier and Glover 2013) in R to implement computation of STRUCTURE analyses (Pritchard, Stephens, and Donnelly 2000) for fifteen separate calculations of each value of k (the number of genetic clusters) ranging from k = 1 to k = 9. We ran each independent STRUCTURE calculation for 1,100,000 Markov chain Monte Carlo (MCMC) generations, discarding the first 100,000 generations as burn-in. To infer the best-supported value of k, we used the Evanno et al. Δk method (Evanno, Regnaut, and Goudet 2005), implemented on the STRUCTURE Harvester website (Earl and VonHoldt 2012). We created STRUCTURE-style bar charts using a custom R plotting function.

4.2.9 Principal Components Analysis

We conducted principal component analyses (PCA) to estimate genetic differentiation, as an alternative method of ordination (based on similarities) to MDS (based on distances). For the 2N-biallelic data set, we used VCFtools v0.1.13 (Danecek et al. 2011) to convert the VCF file into a numeric genotype matrix. We used the scale2 function in R package flashpcaR (Abraham et al., 2017) to scale the numeric genotypes and conducted a PCA in R using factoextra (Kassambara and Mundt 2017).

For the multispecies data set, doing a similar analysis in R is computationally heavy and prohibitively slow, due to the large number of SNPs. Thus, we used the option "--pca" in the PLINK (Purcell et al. 2007) package to perform principal component analysis to estimate genetic differentiation in the multispecies data set.

4.2.10 Ecological differentiation analysis

To obtain a sufficient number of samples to test whether Eastwood manzanita subspecies are differentiated by habitat, we used geo-referenced herbarium collection records (n = 1648) of the seven California Eastwood manzanita subspecies included in this study from the Consortium of California Herbaria (http://ucjeps.berkeley.edu/consortium/), and cleaned the data by removing duplicates and updating taxonomic names (Appendix S4.4). We used environmental variables, from publicly available sources, which have been suggested to be correlated with the distribution of manzanitas and other chaparral shrub species (Franklin, 1998), including soil pH, downloaded from SoilGrid (https://soilgrids.org/, ~0.25 km^2 resolution), and the 19 Bioclimatic variables, along with solar radiation, sourced from Worldclim (http://worldclim.org/version2, ~1 km^2 resolution) (Appendix S4.5). We used ArcGIS v10.2.2, to extract the environmental values for the coordinates of the specimens and performed principal component analysis using the R package factoextra (Kassambara and Mundt 2017).

4.2.11 Environment-genotype association analysis

To determine if Eastwood manzanita subspecies are genetically differentiated at loci that are potentially linked to local environmental adaptation, we generated the "environment-associated SNP data set" data set. We used the environmental data and

Pearson's correlation coefficient (*r*) to calculate the pairwise correlation between environmental variables and eliminated those that were highly correlated ($|r| > 0.7$) (Appendix S4.6), leaving seven variables: BIO3 Isothermality, BIO5 Max Temperature of Warmest Month, BIO9 Mean Temperature of Driest Quarter, BIO12 Annual Precipitation, BIO14 Precipitation of Driest Month, Solar Radiation and Soil pH (Appendix S4.7). We also used the scale2 function in R package flashpcaR to scale the 2N-biallelic data set and then used the latent factor mixed model implemented in R package LFMM (Frichot et al. 2013) to find SNPs that are highly associated with the environmental variables ($P < 1 \times 10-5$ for a z-test). We set the number of latent factors (*K*) to two in accordance with the results of the STRUCTURE analysis and p-value histogram (Appendix S4.8) as recommended by Frichot et al. (2013).

To determine which genes contain the environment-associated SNPs, we identified the contigs containing the environment-associated SNPs and BLASTed (AltschμL et al. 1997) them against GenBank (https://blast.ncbi.nlm.nih.gov/). We used the default setting and chose megablast. If no significantly similar genes were recovered using megablast, we repeated the search using discontiguous megablast and blastn.

4.3 Results

4.3.1 San Gabriel manzanita subspecies alone is supported as genetically distinct in some analyses

To evaluate whether subspecies of Eastwood manzanita are genetically distinct, we analyzed the 4N SNP data set using MDS, NeighborNetwork, and STRUCTURE analyses. The results of all three analyses suggest that there is no correspondence between the structure of genetic variation within Eastwood manzanita and the

subspecies taxonomy (Figs. 4.4, 4.5). In the MDS analysis (Fig. 4.4A), samples of most subspecies overlap widely. Subspecies *gabrielensis* (San Gabriel manzanita) is an exception to this pattern, because it has almost no overlap with other subspecies.

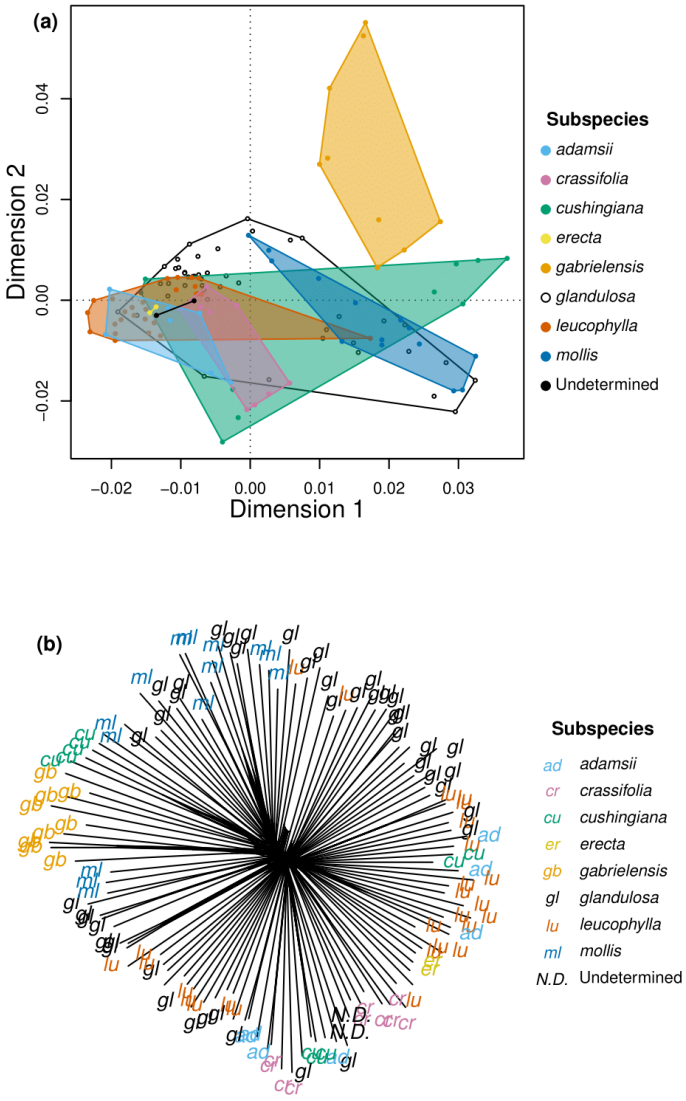


Figure 4.4 MDS analysis (a) and NeighborNetwork (b) for 4N data set. (a) Two dimensional representation of genetic distance among Eastwood manzanita samples. Points and polygons are colored by subspecies identification. Polygons are minimum areas that enclose all samples of each subspecies. (b) Tips are labelled and colored by subspecies identification.

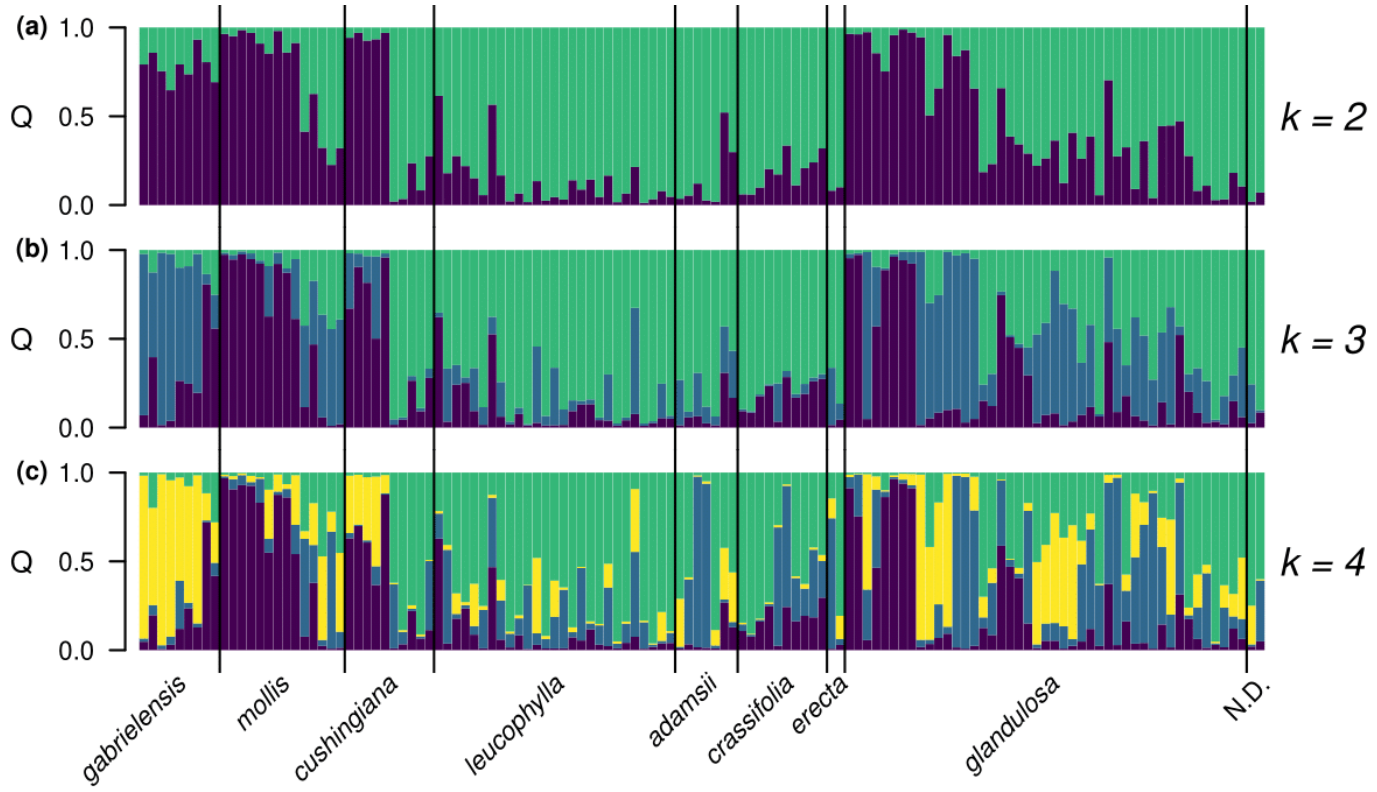


Figure 4.5 STRUCTURE results for *k* = 2 to *k* = 4 (top to bottom) for the 4N data set. Vertical bars represent individuals. Colors within bars represent ancestral clusters of differing genotypes. The proportion of each color in each bar represents the probability of assignment (Q) to each cluster. Individuals are sorted along the x-axis by subspecies, then by latitude of collection.

To determine if there are genetic clusters that do not correspond to currently recognized subspecies, we performed k-means clustering on all three genetic data sets (Appendix S4.9). We used three statistical methods: the within-cluster sum of squares, silhouette, and gap statistic methods (Forgy 1965; Rousseeuw 1987; Tibshirani, Walther, and Hastie 2001), but they did not provide a consistent optimal number of clusters (*k*) (Appendix S4.10), suggesting there is no single optimal number of clusters for these MDS results. At all values of *k*, the clusters do not show any pattern correlated with subspecies identity, but rather, roughly correspond to geographic areas within the range of our sampling (Appendices S4.11–S4.13). One exception is that at $k = 3$, the five samples of San Gabriel manzanita from the type locality (Mill Creek Summit in the San Gabriel Mountains) are identified as a distinct cluster (Figure 4.6, Appendix S4.12).

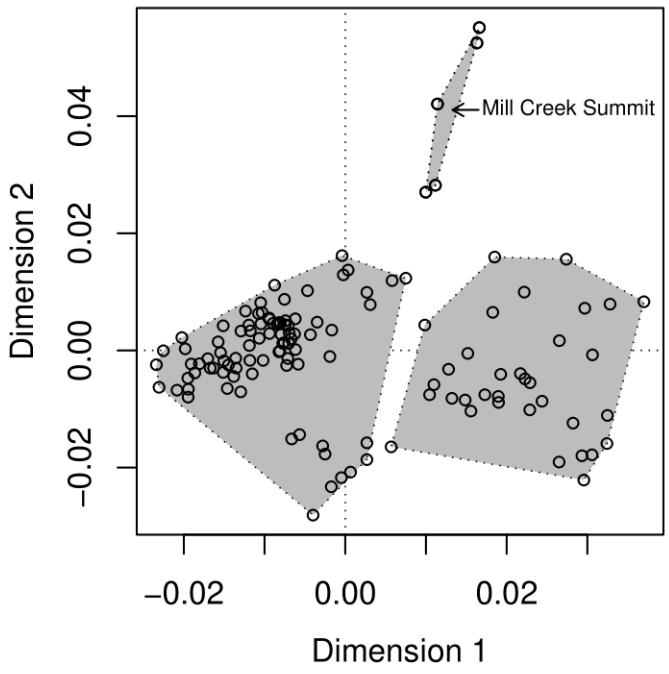


Figure 4.6 Results of k-means clustering, for k = 3, on the MDS of the 4N SNP data set. Points represent individual samples and polygons mark the boundaries of the k-means clusters. The labelled cluster is composed of the five samples from the type locality of *A. glandulosa* subsp. *gabrielensis* at Mill Creek Summit in the San Gabriel Mountains.

In addition to MDS, we assessed whether there were any clusters that do not correspond to subspecies by performing an NMDS analysis paired with Gaussian clustering. NMDS has the same analytical goal as MDS, but uses different underlying mathematics and can thus produce different patterns than MDS. A linear discriminant analysis with leave-one-out cross validation showed no clear separation of subspecies in the NMDS clusters (Appendices S4.14–S4.16). The results of Gaussian clustering on the NMDS show a geographic pattern that is similar to that of the k-means results on the MDS (Appendices S4.11–S4.13). The results of the Gaussian clustering, however, differ

in identifying three of the five Mill Creek Summit samples as a distinct cluster at *k* = 2 (Appendix S4.14).

In the NeighborNetwork analysis (Figure 4.4B), samples from most subspecies are intermingled across the network. There are samples from individual subspecies that group together closely, but other samples from the same subspecies fall elsewhere in the network. Two subspecies, *gabrielensis* and *erecta*, form exclusive groups in the network, however we only have two samples of the latter. Most tips in the network are very long, and show little shared edge length with adjacent tips, however, some groups of tips show more shared edge length, indicating shared genetic variation among these samples. Groups of samples showing these longer shared edges include subsets of subspecies *gabrielensis*, *cushingiana*, and *glandulosa*, as well as some clusters of mixed subspecies identity. The five *A. glandulosa* subsp. *gabrielensis* from the type locality, which were identified as a distinct cluster in both the k-means and NMDS analyses, show considerable shared edge length, supporting their distinctness from other *A. glandulosa* (Appendices S4.13, S4.15–S4.17).

The STRUCTURE analysis (Figure 4.5) shows strongest support for *k* = 2 (Appendix S4.18). Most samples of subspecies *leucophylla*, *adamsii*, and *erecta* show assignment to a single cluster (Figure 4.5A). Individuals of other subspecies show assignment to one of the two clusters or to a combination of the two. At *k* = 2, subspecies *gabrielensis* does not appear to be distinct from other subspecies. At *k* = 3 or 4 (Figure 4.5B, C), there is increased variation within subspecies, but no difference among any subspecies. STRUCTURE analyses of the 4N data set do not support the

distinct identity of subspecies *gabrielensis*, in contrast to the MDS and NeighborNetwork analyses.

4.3.2 Results of analyses of genetic structure are similar when assuming diploidy

Because some methods of analysis are only available for diploid genetic data, many authors have analyzed sequence data from polyploid species as diploid (Rodzen, Famula, and May 2004; Lachmuth, Durka, and Schurr 2010; Burge et al. 2018; Stobie et al. 2018). To evaluate the effect this assumption may have on results, we compared the results of analyses based on the 4N data set to the 2N and 2N-biallelic data sets. The MDS and NeighborNetwork analyses based on the 2N data set (Figure 4.7) yielded a similar pattern of clustering, with almost no correspondence between subspecies and genetic structure. As with the 4N data set, subspecies gabrielensis and the two subspecies *erecta* samples form groups in the NeighborNetwork analysis (Figure 4.7B). In addition, all samples of subspecies *crassifolia* also group together in this analysis, in contrast to the 4N analysis. All other subspecies are highly interspersed with each other. Results of the STRUCTURE analyses based on the 2N data set (Figure 4.8) are also largely consistent with those of the 4N analysis. At $k = 2$, the most strongly supported value, there is no differentiation among subspecies. However, subspecies leucophylla and adamsii, but not erecta in this case, share a genotype that is found less frequently in other subspecies, similar to the results of the analysis with the 4N data set. In contrast to the 4N analysis, though, at $k = 3$ and $k = 4$, the analysis based on the 2N data set shows most individuals of subspecies *gabrielensis* sharing a genotype that is largely genetically distinct from other subspecies.

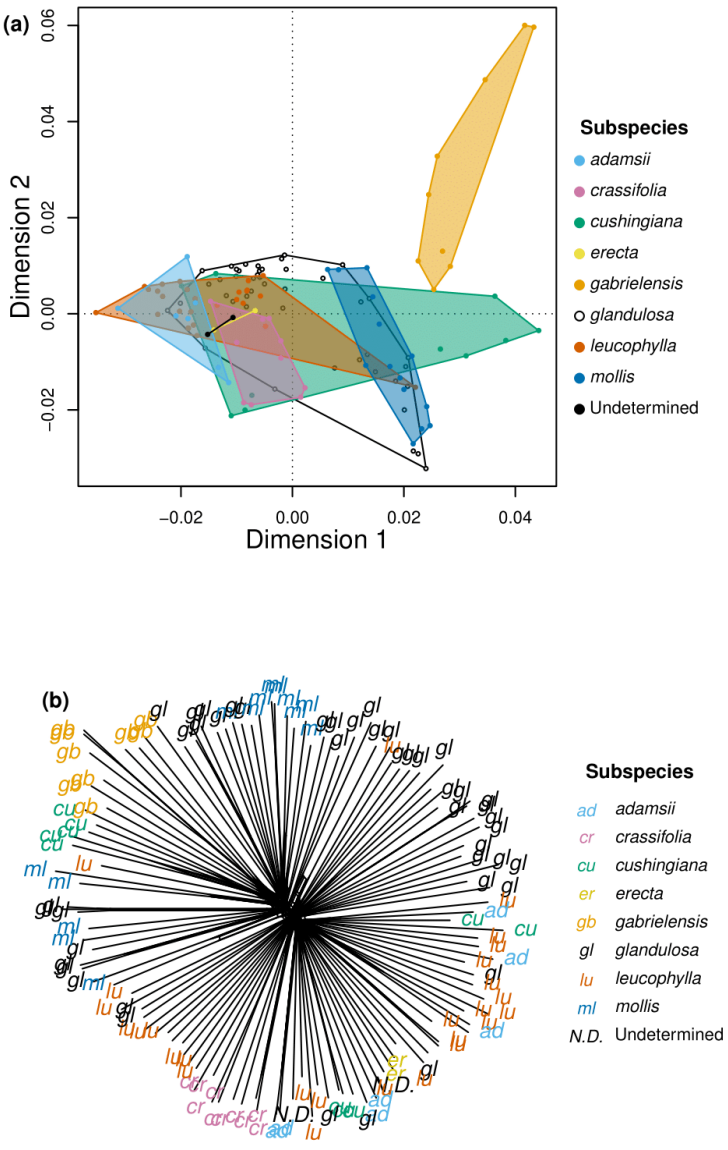


Figure 4.7 MDS analysis (a) and NeighborNetwork (b) for 2N data set. Graphics and colors as in Fig. 4.4.

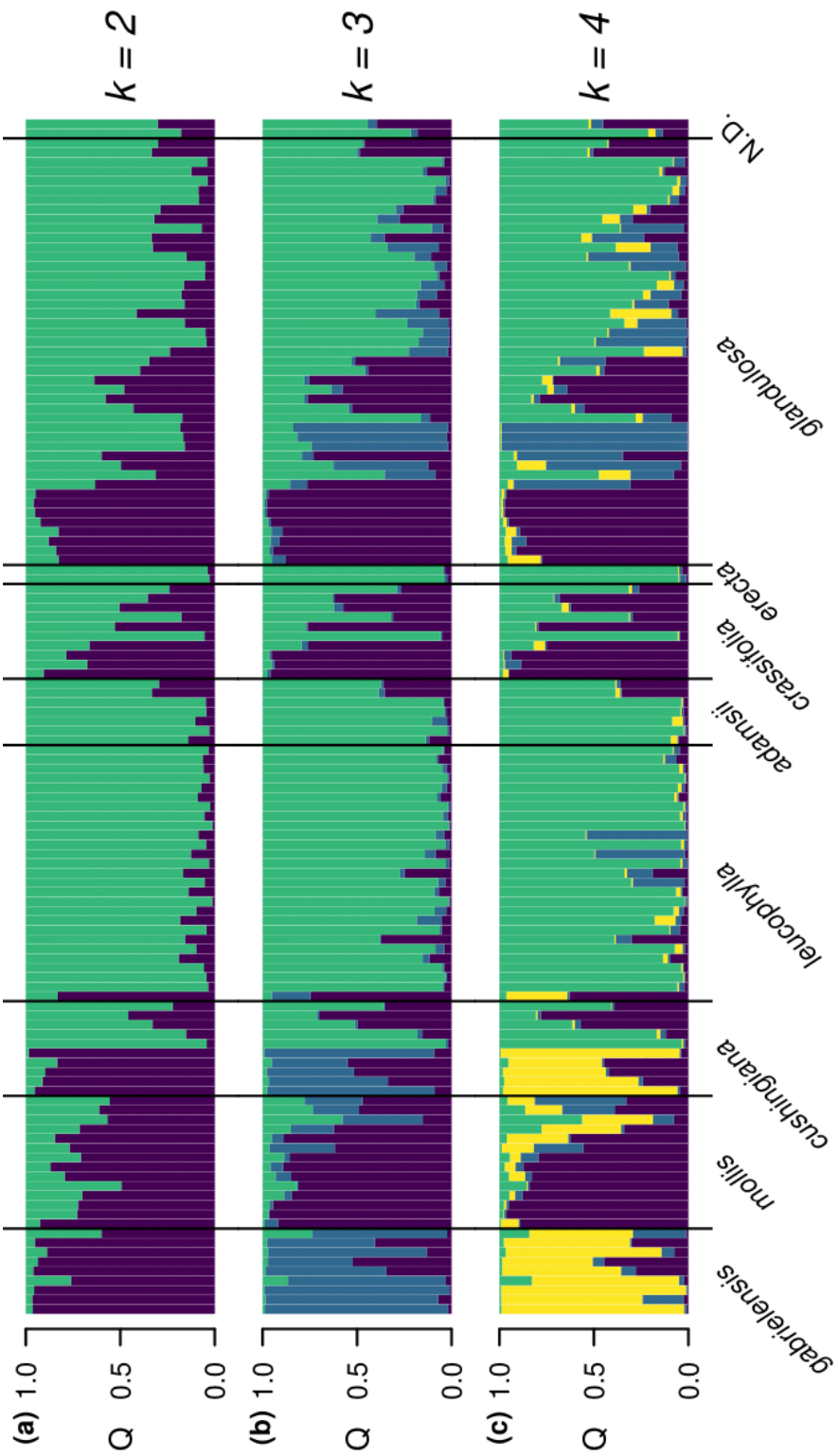


Figure 4.8 STRUCTURE results for $k = 2$ to $k = 4$, for the 2N data set. Graphics and colors as in Fig. 4.5.

The results of the MDS and NeighborNetwork analyses based on the 2N-biallelic (Appendices S4.19, S4.20) data set are nearly identical to those based on the 2N data set. The results of the STRUCTURE analysis on the 2N-biallelic data set (Appendix S4.21) differ slightly from the analysis based on the 2N data set, with more individuals assigned to a single cluster and fewer assigned to multiple clusters.

Because principal components analysis (PCA) is commonly used to analyze large-scale genetic data sets, we performed PCA using the 2N-biallelic data set (Appendix S4.22). Although the percentage of variation explained by PC1 and PC2 is very low (<2%), this analysis reveals a similar pattern of genetic differentiation as the MDS analysis. The combined results of our analyses suggest distinctness of only the San Gabriel manzanita subspecies.

4.3.3 Genetic variation in Eastwood manzanita corresponds to a north-south gradient

A previous study of genetic structure in Eastwood manzanita based on a smaller sample set suggested genetic variation along a north-south transect (Burge et al. 2018), therefore we tested this hypothesis with our data set. Using the 4N data set, we sorted the results of the STRUCTURE analysis by latitude and found a gradient of genotype change from north to south (Figure 4.9, Appendix S4.23). Plotting samples coded by genotype on a map of Southern California shows this north-south gradient (Figure 4.10). Coding the samples in the MDS and NeighborNetwork analyses as north (Transverse Ranges and north) vs. south (south of the Transverse Ranges) rather than by subspecies shows non-overlap of the two groups, although they do not form separated clusters (Appendix S4.24). Moreover, using the Mantel test, we detected a significant

correlation between genetic distance and geographic distance among pairwise samples (Mantel r = 0.27, P < 0.0001) (Appendix S4.25). Analyses using the 2N and 2N-biallelic data sets gave indistinguishable results (Appendices S4.26, S4.27).

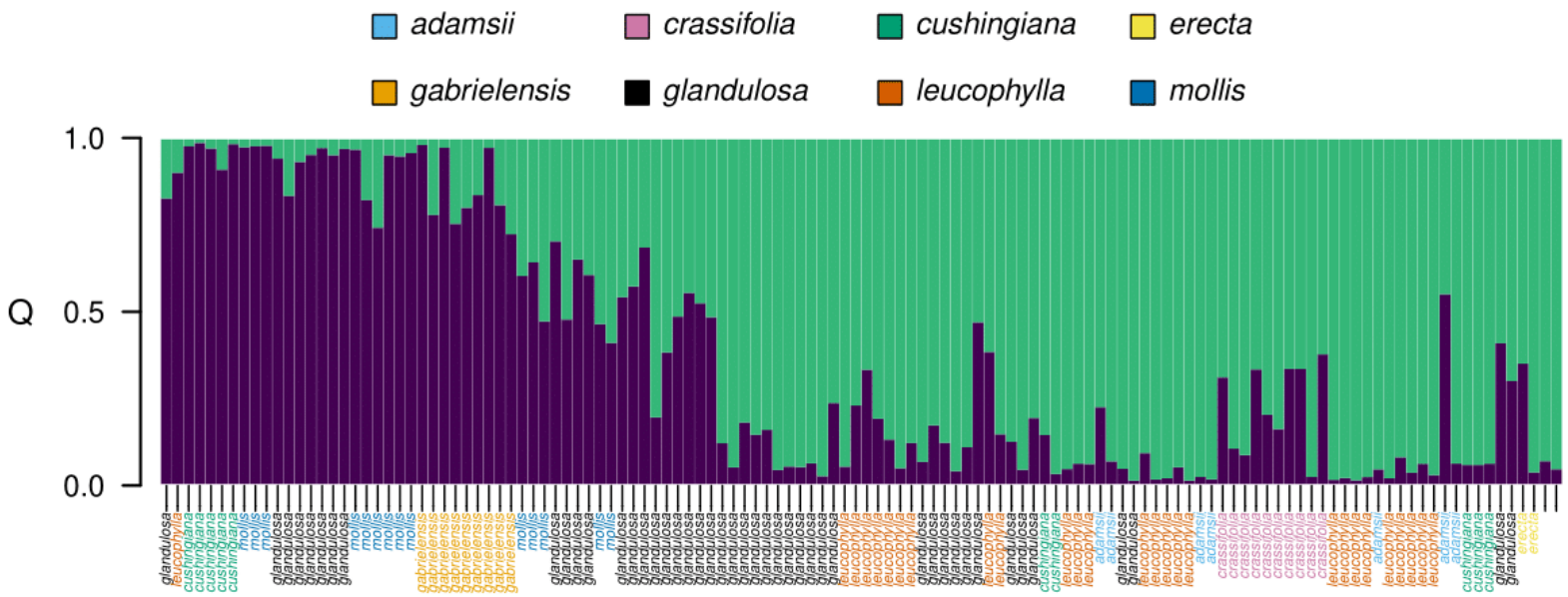


Figure 4.9 STRUCTURE results for *k* = 2, for the 4N data set, sorted by latitudeg Graphics and colors as in Fig. 4.5, except samples sorted by latitude first, and then by subspecies. Two samples undetermined to subspecies (farthest right in graph) are not labeled.

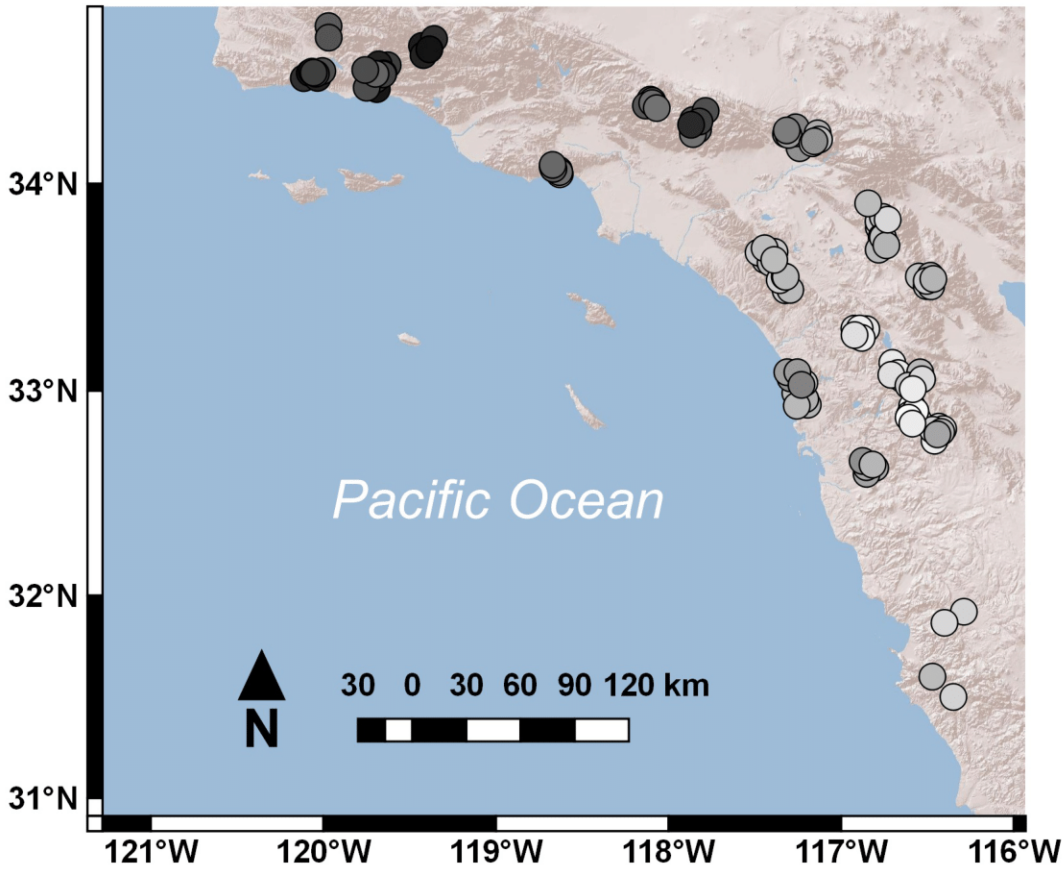


Figure 4.10 Map showing continuous geographic genetic structure within A. glandulosa. Points are shaded in proportion to their value for the first dimension of the MDS on the 4N data set. A random "jitter" value was applied as in Fig. 4.3.

Sorting the samples in the STRUCTURE analysis by latitude (Figure 4.9) exposes a pattern in subspecies *cushingiana* in which the samples of this subspecies fall largely into two discrete groups, one consisting of a northern genotype and one a southern genotype. This is consistent with the localities of our collections, which come from two discrete locations, one northern and one southern (Figure 4.3). This pattern

can also be seen in MDS plots, in which the samples of subspecies *cushingiana* form two clusters, one overlapping with southern samples, and the other forming a cluster close to subspecies *gabrielensis* and other northern samples (Figure 4.5A, 4.7A). Subspecies *glandulosa* shows a similar, but not as clear-cut, pattern in the STRUCTURE results (Figure 4.9). These subspecies both have wide ranges, and span the latitudinal distribution of our samples, thus it is not surprising to see north-south variation within each of them.

4.3.4 Broad-scale environmental data fail to distinguish Eastwood manzanita subspecies

To test whether Eastwood manzanita subspecies can be distinguished by broad-scale topoclimatic and edaphic factors, we performed PCA with environmental data extracted from online mapping resources. We used data from herbarium specimens to provide statistical power. The range of environmental variability captured in this data set differs among subspecies, but samples of most subspecies overlap (Figure 4.11). Principal component 1 (PC1) and PC2, which respectively explain 73% and 25% of the variation, are most heavily weighted by four environmental variables: solar radiation, BIO 4 Temperature Seasonality (standard deviation × 100), BIO 12 Annual Precipitation and BIO16 Precipitation of Wettest Quarter (Appendix S4.28). Subspecies *glandulosa* has the widest range of variation along the niche dimensions considered, and occupies an area on the plot that encompasses the ranges of all other subspecies. Subspecies *cushingiana* occupies the second largest space, although the number of samples classified as subspecies *cushingiana* (n = 221) is far fewer than subspecies *glandulosa* (n = 830).

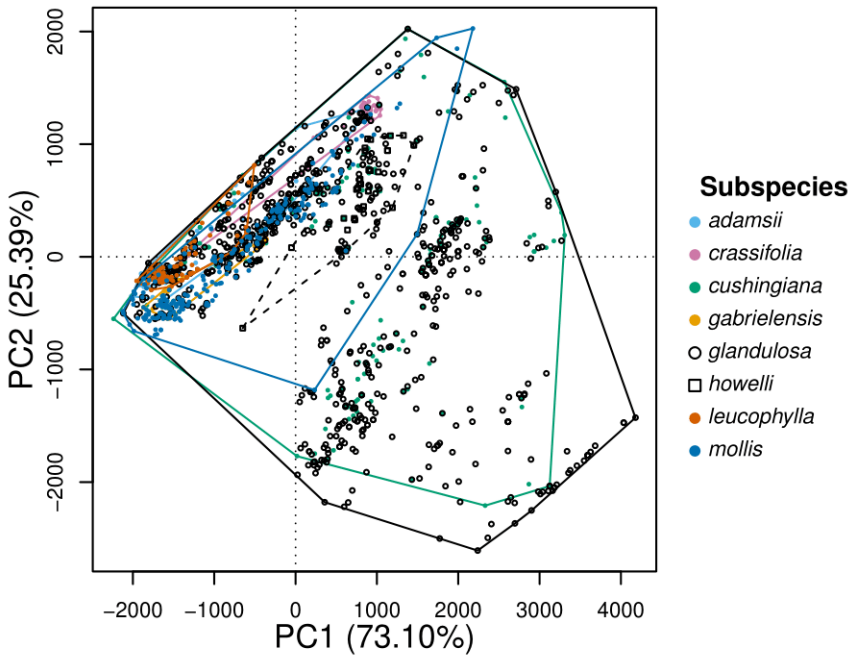


Figure 4.11 PCA using environmental data for herbarium records for Eastwood manzanita. Points represent individual collection records (n = 1648). Subspecies are distinguished by color and shape. Polygons represent minimum areas that enclose all samples of a given subspecies.

4.3.5 Analyses using only environment-associated SNPs suggests subspecies *cushingiana* is also, in part, genetically distinct

To determine if subspecies might be distinguished by differences in genes that potentially play a role in local adaptation, we identified SNPs correlated with variation in environmental variables, and then identified likely genes containing those SNPs. Using the same environmental variables as in the previous analysis, and the 2N-biallelic data set, we identified 73 SNPs that are highly associated with seven of the environmental variables ($P < 1 \times 10{-}5$) (Figure 4.12, Appendix S4.29). Because some SNPs were

correlated with more than one environmental variable, we identified 50 unique SNPs that we included in the environment-associated SNP data set.

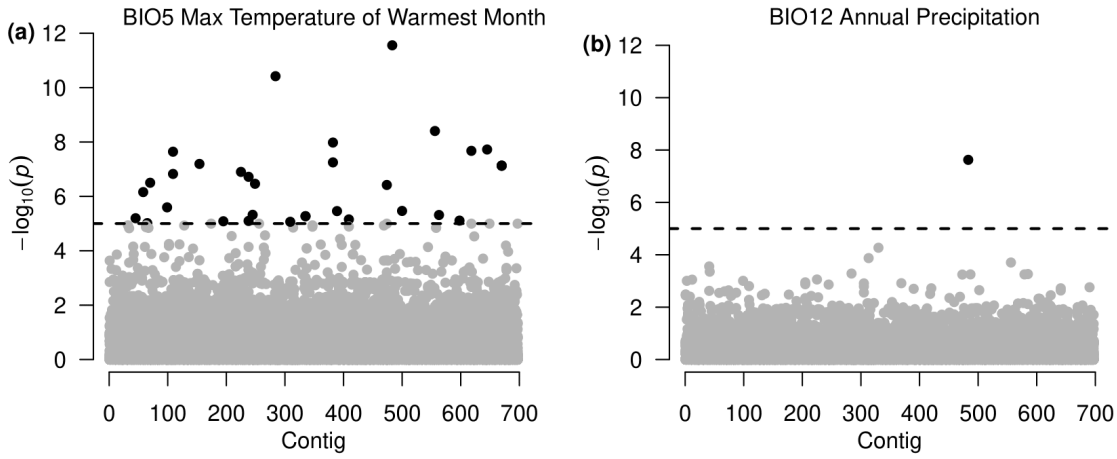


Figure 4.12 SNPs associated with climatic variables. Points represent SNPs. The dashed line represents the threshold for statistical significance at P = 1* 10-5 for the association of SNP and the environmental variable. Solid points show significant association.

We performed PCA, MDS, and STRUCTURE analyses to evaluate whether subspecies of Eastwood manzanita are genetically distinguishable at loci potentially important in environmental adaptation (Figure 4.13, 4.14). The percentage of variation explained by PC1 and PC2 increased greatly compared to the result using the full 2N-biallelic data set (Figure 4.13A, Appendix S4.22), presumably the result of a considerably smaller data set. Plots of PC3 and greater showed no additional pattern of subspecies differentiation. Samples of most subspecies show a higher degree of overlap, although subspecies *gabrielensis* still forms a largely distinct group. However, a subset of the subspecies *cushingiana* samples also form a discrete group in this analysis. The MDS analysis suggests a similar result (Figure 4.13B).

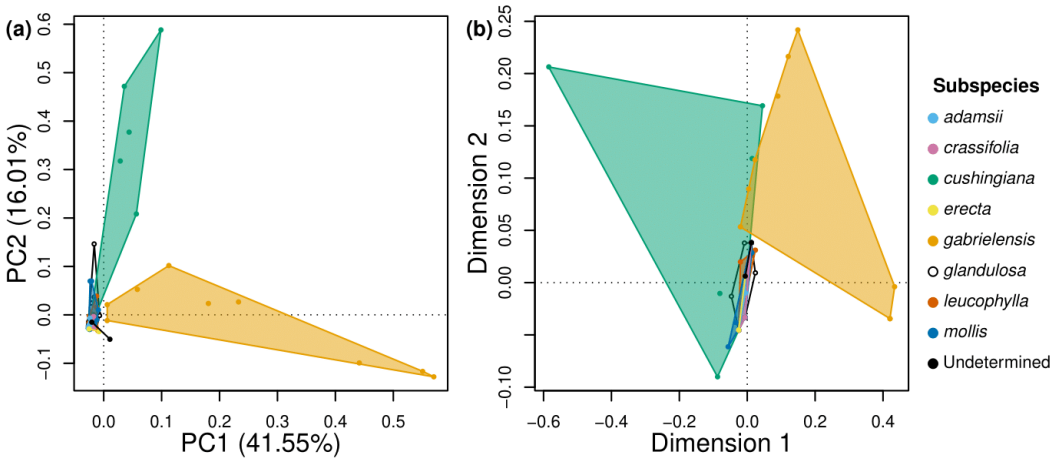


Figure 4.13 PCA (a) and MDS (b) using the environment-associated SNP data set. Graphics and colors as in Fig. 4.4a.

The STRUCTURE analysis using the environment-associated SNP data set shows the greatest support for *k* = 6 (Figure 4.14A). Although the value of k is large, the percentage of assignment to each cluster in most samples is similar, suggesting low levels of differentiation across the subspecies. However, a subset of individuals from subspecies *cushingiana* and all samples of subspecies *gabrielensis* share a genotype that is rare in the remaining samples, consistent with results from the 2N-biallelic data set (Appendix S4.21). Because *k* = 2 was the most highly supported value for the 2N-biallelic data set, we analyzed the environment-associated SNP data set using *k* = 2 as well (Figure 4.14B, C). The results are similar to those with *k* = 6. The assignment to clusters is similar among most samples, with subspecies *gabrielensis* and a subset of subspecies *cushingiana* samples sharing a genotype that is found scattered in a few

samples of other subspecies (Figure 4.14A, B). However, using the reduced data set, we found fewer individuals from other subspecies sharing this genotype (Figure 4.14B, C). Thus, the environment-associated SNP data set does not differentiate most subspecies, but the signal differentiating some samples of subspecies *cushingiana* and subspecies *gabrielensis* is still detected. In contrast to the analyses with the full data set, however, we do not see a north-south gradient of genetic differentiation with this data set (Appendix S4.30). Although subspecies *gabrielensis* and the northern samples of subspecies *cushingiana* still show a distinct genotype, many of the individuals from other subspecies that shared this "northern" genotype in the analyses of the full data set do not share it in the reduced data set, thereby obscuring the north-south gradient.

In the LFMM analysis, we found that the SNPs identified assuming $K = 1$, 3, and 4 almost entirely overlap with the SNPs based on the assumption of $K = 2$ (Appendix S.31). Therefore, we used $K = 2$ in the analysis and traced the 50 environment-associated SNPs to 44 unique contigs, the length of which varied from 130 bp to 260 bp. A BLAST search (AltschμL et al. 1997) found a match for 41 of these contigs (Appendix S.32). The majority of these genes are predicted to play a role in functions such as cell division, protein elongation, cytoskeleton-related processes, and transcription; therefore, relationships with adaptation to local environmental conditions are difficult to establish.

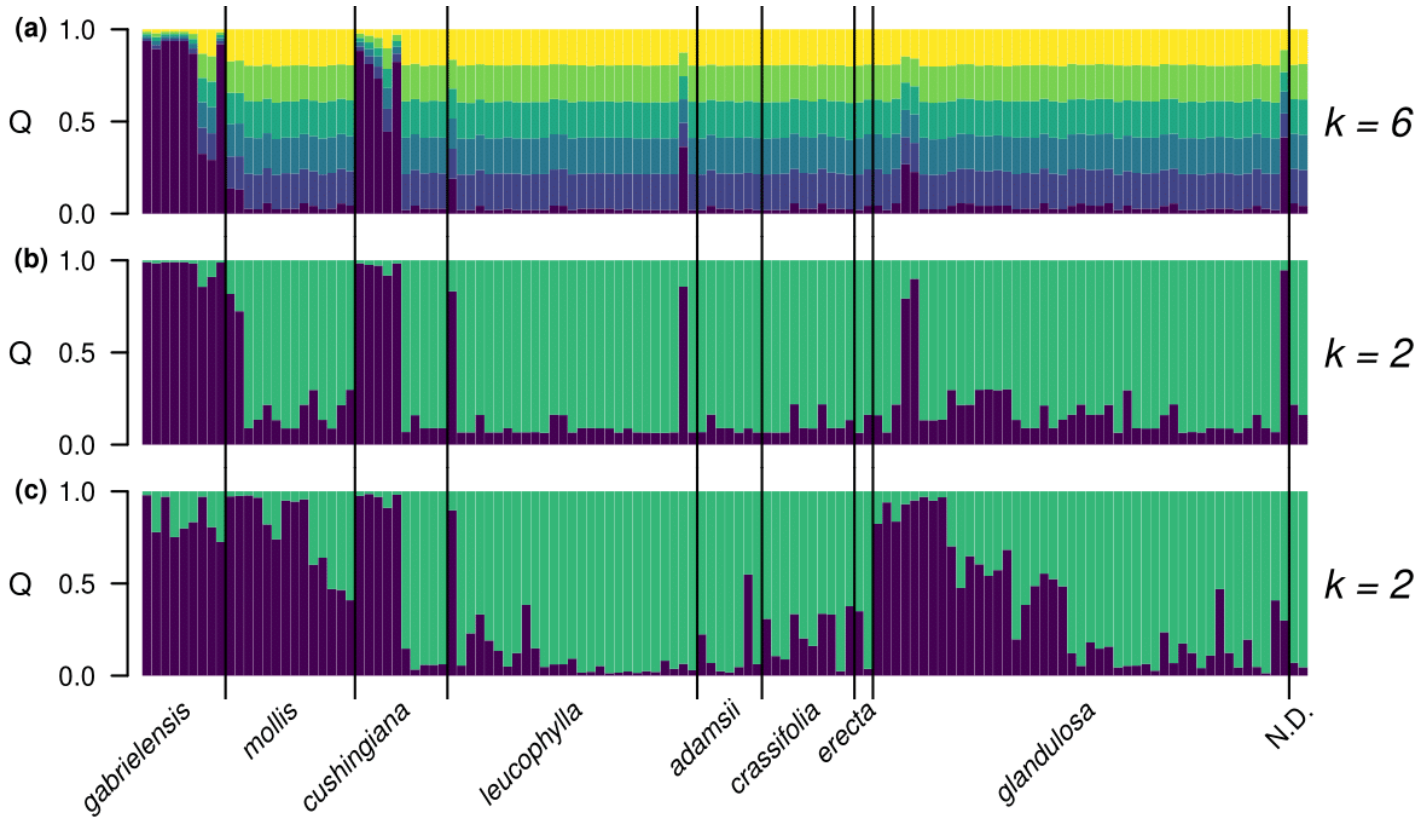


Figure 4.14 STRUCTURE results for k = 6 (a) and k = 2 (b) for the environment-associated SNP data set in comparison to the STRUCTURE result for k = 2 for the 2N-biallelic data set (c). Graphics and colors as in Fig. 4.5.

4.4 Discussion

4.4.1 Most Eastwood manzanita subspecies are not differentiated by reduced-representation genomic sequence data or broad-scale environmental data

Our analyses were unable to detect differentiation among most Eastwood manzanita subspecies on the basis of genomic data, coarse-scale environmental variables, or environment-associated genetic variants. These results are consistent with the morphological variability seen across the species. Instead of a correspondence to taxonomy, we see across our genetic analyses that genetic structure within *A. glandulosa* shows a geographic pattern.

As in the genetic analyses, our analyses of environmental variables found overlap among all subspecies. However, although the geographic distribution of the herbarium specimens we used largely matches the described ranges of subspecies, subspecies taxonomy has changed over the decades, and older specimens may have been classified using a currently outdated system. Furthermore, we are limited in the conclusions we can draw by the coarse resolution of the environmental data and the limited number of ecological factors considered. As is always the case in such analyses, taxa may be separated on niche axes not captured by the environmental data considered, at the scale of the analysis (Fletcher et al. 2013). Therefore, these results should be seen as preliminary and a source of hypotheses for additional testing with updated specimen identification, additional habitat information such as soil water potential, mineral differences, and vapor pressure deficit, as well as finer-scaled data describing these factors. Although it is not surprising to find genetic and habitat similarity among members of the same species, some level of divergence in these factors would

be expected among subspecies as evolutionary units, which may be identified with further study.

Our results did show that genetic structure within Eastwood manzanita reflects geographic distribution, confirming the findings from Burge (Burge et al. 2018). The MDS and NeighborNetwork analyses show that the samples can be divided into northern and southern groups, however, these groups are not clearly separated, and the overall pattern is thus better described as a gradient of genetic variation (Appendix S4.24). The STRUCTURE results show a transition from a predominantly northern genotype to predominantly southern at the Transverse Ranges of Southern California (Figure 4.9). This suggests a pattern of genetic divergence by geographic distance, confirmed by the significant correlation between genetic and geographic distance (Appendices S4.25–S4.27). Such a north-south pattern of genetic variation has also been observed in other studies of the California biota (Zink, Lott, and Anderson 1987; Burge et al. 2011; Sork et al. 2016; Schierenbeck 2017) and the Transverse Ranges have been suggested as a barrier to genetic continuity (Calsbeek, Thompson, and Richardson 2003; Sgariglia and Burns 2003; Forister, Fordyce, and Shapiro 2004; Chatzimanolis and Caterino 2007). Our analyses do not indicate a barrier at the Transverse Ranges, but rather suggest a continuum (Figure 4.9, 4.10).

Our analyses using the 4N, 2N, and 2N-biallelic data sets generate largely similar results (Figure 4.4, 4.5, 4.7, 4.8, S4.1–S4.3), suggesting that the loss of information caused by assuming diploidy in a tetraploid sample set may not prevent detection of genetic structure. In fact, the genetic patterns appear most clear in the 2N-biallelic and least clear in the 4N analyses, suggesting that the biallelic SNPs may hold the strongest

evolutionary signal. This is supported by the observation that most of the loci (~98%) in the data set are biallelic. Overall, our results suggest that the easier and more rapid analyses based on the use of biallelic genetic data can give a good approximation of the genetic structure. Testing in additional systems is needed to determine if this is broadly true.

4.4.2 Analyses support distinction of San Gabriel manzanita, but not Del Mar manzanita

Two subspecies of Eastwood manzanita are considered rare or threatened: San Gabriel manzanita (*A. glandulosa* subsp. *gabrielensis*), found in the San Gabriel and Sierra Madre mountains of California; and Del Mar manzanita (*A. glandulosa* subsp. *crassifolia*), found along the coast of San Diego County (Figure 4.2). A previous study of Del Mar manzanita, based on RAD-Seq data and morphometric analyses, concluded that circumscription of this subspecies based on vegetative morphology was ineffective (Burge et al. 2018). An earlier study found that fruit shape was distinctive in subspecies *crassifolia* (Keeley, Vasey, and Parker 2007), but fruits were not included in the study by Burge (Burge et al. 2018). The Burge et al. study was unable to reach a conclusion regarding genetic boundaries of the subspecies, largely due to insufficient sampling of other subspecies. Our results, with broader sampling across the species, suggest that Del Mar manzanita is not genetically distinct from other subspecies (Figure 4.4, 4.5). The only exception is the 2N and 2N-biallelic NeighborNetwork analyses, in which the Del Mar manzanita samples cluster together (Figure. 4.7, S4.1–S4.3). However, neither the MDS nor the STRUCTURE analyses of the diploid data sets show any differentiation of subspecies *crassifolia*. These results, along with inconclusive morphological distinction and lack of environmentally associated

genetic differentiation along the niche dimensions considered, suggest that the recognition of Del Mar manzanita as a distinct subspecies should be reconsidered.

San Gabriel manzanita (*A. glandulosa* subsp. *gabrielensis*) was historically treated as a separate species, *A. gabrielensis* (Wells 2000, 1992), but was later transferred into *A. glandulosa* as a subspecies (Keeley, Vasey, and Parker 2007). The PCA, MDS, and NeighborNetwork analyses using all three genetic data sets indicate that this subspecies is genetically distinct from the others (Figure 4.4, 4.7, S4.1–S4.4). To determine if San Gabriel manzanita is as distinct as other species are from each other, we analyzed genetic differentiation of five diploid species that are morphologically both distinct and consistent (Appendices S4.33–S4.35). The results showed that all species formed discrete clusters, and occupied well-separated regions of the MDS and PCA plots. In contrast, samples of San Gabriel manzanita fall relatively close to the other subspecies (Figure 4.4, 4.6, Appendices S4.19 and S4.22). The results of the NeighborNetwork analysis of the multispecies data set show genetic clusters completely coincide with species identity, with samples from each species grouping together, and separated from other species (Appendix S4.35). In contrast to this, NeighborNetwork analyses of Eastwood manzanita subspecies show that although San Gabriel manzanita samples form a unique cluster, they are not separated from the rest of the network (Figure 4.4, 4.6, Appendix S4.20). Although these results cannot be compared directly, the lack of discrete separation of the San Gabriel manzanita samples from the other subspecies suggests a closer relationship than seen in the multispecies analyses. However, we found that samples of San Gabriel manzanita from the type locality (Mill Creek Summit in the San Gabriel Mountains) showed greater differentiation from other samples of *A. glandulosa* than did the samples of San Gabriel manzanita from other

localities (Appendix S4.36). The Mill Creek samples were also identified as a distinct cluster in the k-means and Gaussian clustering analyses (Figure 4.6, Appendices S4.12, S4.13, S4.15, S4.16).

Our results may support the original status of San Gabriel manzanita as a distinct species, although it would possibly need to be circumscribed more narrowly as just the plants from Mill Creek Summit. It has been hypothesized that San Gabriel manzanita may be a hybrid of *A. glandulosa* and *A. parryana* (Keeley, Vasey, and Parker 2007; Kauffmann et al. 2015). The latter is a broadly sympatric species that has noticeable morphological similarities with San Gabriel manzanita, including shiny bright green leaves and fusion of the nutlets in the fruits (Kauffmann et al. 2015). The samples of San Gabriel manzanita from other localities, which have typical morphology for San Gabriel manzanita and are genetically intermediate between the Mill Creek Summit cluster and other samples, may be hybrids between *A. glandulosa* and the putatively distinct species *A. gabrielensis*. A definitive evaluation of this hypothesis, however, requires further sampling, including additional San Gabriel manzanita populations and *A. parryana*.

Unlike PCA, MDS, and NeighborNetwork analyses, STRUCTURE analyses, with all three genetic data sets at most values of *k*, do not indicate that the genotype of San Gabriel manzanita is unique, although it is not common in other subspecies (Figure 4.5, 4.7, Appendix S4.21). PCA, MDS, and NeighborNetwork analyses use either allele frequencies (PCA) or distance measures (MDS and NeighborNetwork) to characterize genetic variability or relatedness in a given population of samples (Bryant and Moulton 2004; Jombart, Pontier, and Dufour 2009). These analyses invoke few

assumptions regarding the structure of genetic diversity in the sample population. In contrast, STRUCTURE is based on the assumption of Hardy-Weinberg (HW) equilibrium, and constructs a genetic model that assigns samples to clusters that minimize HW disequilibrium (Pritchard, Stephens, and Donnelly 2000; Falush, Stephens, and Pritchard 2003). The evolutionary assumptions implemented to construct complex models in STRUCTURE analyses likely explain the differences in results obtained in these analyses. Moreover, the formula STRUCTURE uses to calculate expected genotype frequencies differs for diploid and polyploid samples, which may explain the differences in the results obtained using the 4N vs. the 2N or 2N-biallelic data sets (Dufresne et al. 2014).

Taken together, our results indicate that San Gabriel manzanita shows a degree of genetic differentiation from the other subspecies and suggests it should remain of conservation concern, however a final determination regarding its taxonomic rank cannot be made without further sampling, including a disjunct population reported from Santa Barbara County.

4.4.3 Using genetic loci potentially related to environmental adaptation produces a similar result to the full SNP data set

Subspecies are often considered to arise through local adaptation (Grant 1981b; Patten and Unitt 2002; Haig et al. 2006; Walsh et al. 2017). This suggests that subspecies should differ genetically at loci related to responses to environmental factors. Although we found that Eastwood manzanita subspecies are not differentiated on the basis of genome-wide genetic data, we investigated whether an analysis of environmentally linked genetic loci might uncover differences among subspecies. Both

PCA and MDS using the 50 2N-biallelic SNPs that varied in correlation with environmental variables showed a largely similar pattern to the analyses of the full SNP data set (Figure 4.13, Appendices S4.19, S4.22), with perhaps even more pronounced results showing tight clustering of most samples and divergence of samples of subspecies *gabrielensis*. However, with this strongly reduced genetic data set, subspecies *cushingiana* also is found to be partly distinct. Some samples of this subspecies cluster with the other subspecies, but five samples, corresponding to those from the northern part of our sampling range (Figure 4.3), are divergent. This suggests that local adaptation may play a role in genetic differentiation of subspecies *gabrielensis* and *cushingiana*, which is consistent with the habitat differentiation found in a previous study (Keeley, Vasey, and Parker 2007). The partial genetic distinction within subspecies *cushingiana* is also supported by other analyses that suggest that the subspecies can be subdivided into a north and a south component (e.g., Figure 4.9). This reflects our sampling from two geographically separated regions (Figure 4.2, 4.3). Further investigation, including sampling across the range, is needed to elucidate the patterns of genetic variation of subspecies *cushingiana*.

A goal of molecular analyses is to draw direct lines from genetic variants to phenotypes. We evaluated the genes containing environmentally correlated SNPs to determine if we could identify such connections between genotype and environmental adaptation. However, identification of genes containing environmentally associated SNPs did not reveal any genes that can readily be related to adaptation to habitat differences. Many of these genes function in multiple biological processes, making connection with specific environmental adaptations impossible. Establishing such a link would take carefµL analysis of gene function, which is not possible at this time because

of limits posed by the plants themselves, including difficult culturing and long generation times.

4.4.4 Subspecies recognition in Eastwood manzanita

Although Eastwood manzanita has been divided into multiple subspecies, our analyses suggest that with the exception of San Gabriel manzanita, the eight subspecies we sampled are not well differentiated genetically. This is consistent with the overlapping morphological boundaries among the subspecies. Within the ranges of many subspecies there are populations that are fairly uniform phenotypically, and that represent the archetype for that subspecies. However, in between those populations are heterogeneous populations that obscure much of those distinctions. Recognition of the individual phenotypically uniform populations as subspecies leaves the plants found in much of the range of the species as unclassifiable. Furthermore, such population-level variation may allow recognition of different phenotypes, but currently we do not have data to suggest any degree of genetic differentiation, which would be required for subspecies to be evolutionarily significant units, according to our definition. Given that some populations of *A. glandulosa* show consistent morphologies that can be identified as a single subspecies, but our analyses detected no genetic differentiation, the complex may be something akin to the syngameon concept sometimes applied to groups of woody plants species that show extensive interspecific gene flow (Grant 1981b; Cavender-Bares 2018).

One hypothesis that could explain the existence of populations with mixed phenotypes in these long-lived perennial shrubs is that populations that had uniform phenotypes when they were established may have become more variable over time due

to subsequent gene flow from other populations. This is consistent with the predicted very long lifespan of these plants and the fact that recruitment is tied to infrequent fires. These factors may create populations composed of plants of varying ages with genetic admixture from multiple genetic lineages within the species. Studies have shown that the seeds are dispersed by mammals, including large mammals, which may allow them to be moved over large distances (Keeley and Hays 1976; Parker 2015). Although nothing is known about the distance that pollen travels in this species, solitary bees, honey bees, and bumblebees, which have been documented pollinating manzanitas including *A. glandulosa* subsp. *mollis*, can travel up to several miles (Fulton and Lynn Carpenter 1979; Osborne et al. 2007; Zurbuchen et al. 2010; Hagler et al. 2011). Such mixing of genetic material, leading to diverse phenotypes, is therefore possible, however little is known about the demographic structure of Eastwood manzanita populations—an area that needs further investigation.

Another hypothesis to explain the morphological variability in this species is that Eastwood manzanita may be a heterogeneous assemblage of tetraploid hybrids that arose multiple times, possibly from a number of different progenitor species pairs. An additional factor contributing to the variability might be introgression into these tetraploid populations from diploid species via unreduced gametes (Ramsey and Schemske 1998). This is consistent with the regionally localized ranges of many of the subspecies (Figure 4.2). Much more in-depth sampling across the ranges of these subspecies, and inclusion of potential progenitor species, will be needed to evaluate this hypothesis. This would include more northern populations, including *A. glandulosa* subsp. *howellii*, which is an unsampled subspecies from the central coast of California, as well as expanded

sampling of Mexican subspecies. These analyses will benefit from the sequencing of a manzanita genome, which is underway.

Subspecies concepts have not historically focused on plants, particularly not on long-lived perennial plants. It is possible that because of the differences in population structure that result from the long lifespan, dynamics of gene flow, and immobility of individuals, subspecific differentiation in these species may not be well described by available concepts. Further consideration of subspecies concepts, and aspects of practical application, are needed to encompass a wider variety of organisms with diverse life histories. As we now have greater power than ever to assess genetic structure within a species, we have an opportunity to evaluate diversity in an ever-greater array of groups, and must ensure our conceptual framework keeps pace accordingly.

4.5 Conclusion

Our results show that genetic structure within Eastwood Manzanita does not correspond to current subspecies circumscriptions, but rather reflects geographic distribution. We also found that of the two subspecies recognized by state and federal authorities as rare or endangered, only *A. glandulosa* subsp. *gabrielensis* appears to be genetically distinct, and not *A. glandulosa* subsp. *crassifolia*. This implies that genotype preservation is important in the conservation of the former subspecies but that additional data are needed to fully evaluate the genetic distinctiveness of *A. glandulosa* subsp. *crassifolia.* Our study suggests that next-generation sequencing data may provide novel insight into diversification among morphologically defined manzanita taxa.

4.6 References

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res.*, 25: 3389-402.

Andrews, Kimberly R, Jeffrey M Good, Michael R Miller, Gordon Luikart, and PaμL A Hohenlohe. 2016. 'Harnessing the power of RADseq for ecological and evolutionary genomics', *Nature Reviews Genetics*, 17: 81-92.

Baldwin, Bruce G., Douglas H. Goldman, David J. Keil, Robert Patterson, and Thomas J. Rosatti. 2012. *The Jepson Manual: Vascular Plants of California* (University of California Press).

Besnier, Francois, and Kevin A Glover. 2013. 'ParallelStructure: AR package to distribute parallel runs of the population genetics program STRUCTURE on multi-core computers', *PLoS One*, 8: e70651.

Bowcock, Anne M, Andres Ruiz-Linares, James Tomfohrde, Eric Minch, Judith R Kidd, and L Luca Cavalli-Sforza. 1994. 'High resolution of human evolutionary trees with polymorphic microsatellites', *nature*, 368: 455-57.

Bradburd, Gideon S., Graham M. Coop, and Peter L. Ralph. 2018. 'Inferring Continuous and Discrete Population Genetic Structure Across Space', *Genetics*, 210: 33-52.

Brelsford, A., C. Dufresnes, and N. Perrin. 2016. 'High-density sex-specific linkage maps of a European tree frog (Hyla arborea) identify the sex chromosome without information on offspring sex', *Heredity*, 116: 177-81.

Bryant, David, and Vincent Moulton. 2002. "NeighborNet: An agglomerative method for the construction of planar phylogenetic networks." In *International workshop on algorithms in bioinformatics*, 375-91. Springer.

———. 2004. 'Neighbor-net: an agglomerative method for the construction of phylogenetic networks', *Molecular biology and evolution*, 21: 255-65.

Burge, Dylan O, V Thomas Parker, Margaret Mulligan, and César García Valderamma. 2018. 'Conservation genetics of the endangered Del Mar manzanita (Arctostaphylos glandulosa subsp. crassifolia) based on RAD sequencing data', *Madroño*, 65: 117-30.

Burge, Dylan O., Diane M. Erwin, Melissa B. Islam, Jürgen Kellermann, Steven W. Kembel, Dieter H. Wilken, and PaμL S. Manos. 2011. 'Diversification of

Ceanothus (Rhamnaceae) in the California Floristic Province', *Int. J. Plant Sci.*, 172: 1137-64.

Calsbeek, Ryan, John N. Thompson, and James E. Richardson. 2003. 'Patterns of molecular evolution and diversification in a biodiversity hotspot: the California Floristic Province', *Mol. Ecol.*, 12: 1021-29.

Carstens, Bryan C., Tara A. Pelletier, Noah M. Reid, and Jordan D. Satler. 2013. 'How to fail at species delimitation', *Mol. Ecol.*, 22: 4369-83.

Catchen, Julian, PaµL A. Hohenlohe, Susan Bassham, Angel Amores, and William A. Cresko. 2013. 'Stacks: an analysis tool set for population genomics', *Mol. Ecol.*, 22: 3124-40.

Cavender-Bares, Jeannine. 2018. 'Diversification, adaptation, and community assembly of the American oaks (Quercus), a model clade for integrating ecology and evolution', *The New phytologist*, 221: 669-92.

Chatzimanolis, S., and M. S. Caterino. 2007. 'Toward a better understanding of the “Transverse Range Break”: lineage diversification in southern California', *Evolution*.

Clarke, K Robert. 1993. 'Non-parametric multivariate analyses of changes in community structure', *Australian journal of ecology*, 18: 117-43.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and Group Genomes Project Analysis. 2011. 'The variant call format and VCFtools', *Bioinformatics*, 27: 2156-58.

Dray, Stéphane, Anne-Béatrice Dufour, and Others. 2007. 'The ade4 package: implementing the duality diagram for ecologists', *J. Stat. Softw.*, 22: 1-20.

Dufresne, France, Marc Stift, Roland Vergilino, and Barbara K. Mable. 2014. 'Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools', *Mol. Ecol.*, 23: 40-69.

Earl, Dent A, and Bridgett M VonHoldt. 2012. 'STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method', *Conservation genetics resources*, 4: 359-61.

Evanno, Guillaume, Sebastien Regnaut, and Jérôme Goudet. 2005. 'Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study', *Molecular ecology*, 14: 2611-20.

Falush, Daniel, Matthew Stephens, and Jonathan K. Pritchard. 2003. 'Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies', *Genetics*, 164: 1567-87.

Fletcher, Robert J., Andre Revell, Brian E. Reichert, Wiley M. Kitchens, Jeremy D. Dixon, and James D. Austin. 2013. 'Network modularity reveals critical scales for connectivity in ecology and evolution', *Nature communications*, 4: 2572-72.

Forgy, Edward W. 1965. 'Cluster analysis of multivariate data: efficiency versus interpretability of classifications', *Biometrics*, 21: 768-69.

Forister, M. L., J. A. Fordyce, and A. M. Shapiro. 2004. 'Geological barriers and restricted gene flow in the holarctic skipper Hesperia comma (Hesperiidae)', *Mol. Ecol.*, 13: 3489-99.

Frichot, Eric, Sean D. Schoville, Guillaume Bouchard, and Olivier François. 2013. 'Testing for associations between loci and environmental gradients using latent factor mixed models', *Mol. Biol. Evol.*, 30: 1687-99.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. 'The elements of statistical learning. vol. 1 Springer series in statistics', *New York*.

Fulton, Robert E., and F. Lynn Carpenter. 1979. 'Pollination, reproduction, and fire in California Arctostaphylos', *Oecologia*, 38: 147-57.

Garrison, Erik, and Gabor Marth. 2012. 'Haplotype-based variant detection from short-read sequencing', *arXiv [q-bio.GN]*.

Gower, John C. 1966. 'Some distance properties of latent root and vector methods used in multivariate analysis', *Biometrika*, 53: 325-38.

Grant, Verne. 1981a. 'Plant speciation.' in, *Plant Speciation* (Columbia university press).

———. 1981b. *Plant speciation* (New York: Columbia University Press xii, 563p.-illus., maps, chrom. nos.. En 2nd edition. Maps, Chromosome numbers. General (KR, 198300748)).

Grooten, Monique, and Rosamunde EA Almond. 2018. *Living planet report-2018: aiming higher* (WWF international).

Hagler, James R., Shannon C. Mueller, Larry R. Teuber, Scott A. Machtley, and Allen Van Deynze. 2011. 'Foraging range of honey bees, Apis mellifera, in alfalfa seed production fields', *Journal of insect science (Online)*, 11: 144-44.

Haig, Susan M., Erik A. Beever, Steven M. Chambers, Hope M. Draheim, Bruce D. Dugger, Susie Dunham, Elise Elliott-Smith, Joseph B. Fontaine, Dylan C. Kesler, Brian J. Knaus, and Others. 2006. 'Taxonomic considerations in listing subspecies under the US Endangered Species Act', *Conserv. Biol.*, 20: 1584-94.

Harrison, Nicola, and Catherine Anne Kidner. 2011. 'Next--generation sequencing and systematics: What can a billion base pairs of DNA sequence data do for you?', *TAXON*, 60: 1552-66.

Harrison, Richard G., and Erica L. Larson. 2014. 'Hybridization, introgression, and the nature of species boundaries', *J. Hered.*, 105 Suppl 1: 795-809.

Hennig, C, and B Hausdorf. 2020. "Prabclus: functions for clustering and testing of presence-absence, abundance and multilocus genetic data." In.

Huson, Daniel H, and David Bryant. 2006. 'Application of phylogenetic networks in evolutionary studies', *Molecular biology and evolution*, 23: 254-67.

Jepson, Willis L. 1916. 'Regeneration in manzanita', *Madroño*, 1: 3-12.

Jombart, T., D. Pontier, and A. B. Dufour. 2009. 'Genetic markers in the playground of multivariate analysis', *Heredity*, 102: 330-41.

Jörger, Katharina M., and Michael Schrödl. 2013. 'How to describe a cryptic species? Practical challenges of molecular taxonomy', *Front. Zool.*, 10: 59.

Kassambara, Alboukadel, and Fabian Mundt. 2016. 'Factoextra: extract and visualize the results of multivariate data analyses', *R package version*, 1: 2016.

———. 2017. 'Factoextra: extract and visualize the results of multivariate data analyses', *R package version*, 1: 337-54.

Kauffmann, Michael Edward, Tom Parker, Michael Vasey, and Jeff Bisbee. 2015. *Field Guide to Manzanitas* (Backcountry Press).

Keeley, Jon E, Michael C Vasey, and V Thomas Parker. 2007. 'Subspecific variation in the widespread burl-forming Arctostaphylos glandulosa', *Madroño*, 54: 42-62.

Keeley, Jon E., and Robert L. Hays. 1976. 'Differential seed predation on two species of Arctostaphylos (Ericaceae)', *Oecologia*, 24: 71-81.

Keller, Reuben P., Juergen Geist, Jonathan M. Jeschke, and Ingolf Kühn. 2011. 'Invasive species in Europe: ecology, status, and policy', *Environmental Sciences Europe*, 23: 23.

Lachmuth, Susanne, Walter Durka, and Frank M. Schurr. 2010. 'The making of a rapid plant invader: genetic diversity and differentiation in the native and invaded range of Senecio inaequidens', *Mol. Ecol.*, 19: 3952-67.

Li, Heng. 2013. 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', *arXiv [q-bio.GN]*.

Li, Weizhong, and Adam Godzik. 2006. 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, 22: 1658-59.

Mace, Georgina M. 2004. 'The role of taxonomy in species conservation', *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 359: 711-19.

MacQueen, James, and Others. 1967. "Some methods for classification and analysis of multivariate observations." In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 281-97. books.google.com.

Martien, Karen K, Matthew S Leslie, Barbara L Taylor, Phillip A Morin, Frederick I Archer, Brittany L Hancock-Hanser, Patricia E Rosel, Nicole L Vollmer, Amélia Viricel, and Frank Cipriano. 2017. 'Analytical approaches to subspecies delimitation with genetic data', *Marine Mammal Science*, 33: 27-55.

Moore, Christopher M., and Stephen B. Vander Wall. 2015. 'Scatter-hoarding rodents disperse seeds to safe sites in a fire-prone ecosystem', *Plant Ecol.*, 216: 1137-53.

Nei, Masatoshi, and Sudhir Kumar. 2000. *Molecular evolution and phylogenetics* (Oxford University Press, USA).

Osborne, Juliet L., Andrew P. Martin, Norman L. Carreck, Jennifer L. Swain, Mairi E. Knight, Dave Goulson, Roddy J. Hale, and Roy A. Sanderson. 2007. 'Bumblebee flight distances in relation to the forage landscape', *The Journal of animal ecology*, 77: 406-15.

Parchman, Thomas L., Zachariah Gompert, Joann Mudge, Faye D. Schilkey, Craig W. Benkman, and C. Alex Buerkle. 2012. 'Genome-wide association genetics of an adaptive trait in lodgepole pine', *Mol. Ecol.*, 21: 2991-3005.

Parker, V Thomas. 2015. 'Dispersal mutualism incorporated into large-scale, infrequent disturbances', *PLoS One*, 10: e0132625.

Patten, Michael A. 2015. 'Subspecies and the philosophy of science', *The Auk: Ornithological Advances*, 132: 481-85.

Patten, Michael A., and Philip Unitt. 2002. 'Diagnosability Versus Mean Differences of Sage Sparrow Subspecies', *The Auk*, 119: 26-35.

Peterson, Brant K., Jesse N. Weber, Emily H. Kay, Heidi S. Fisher, and Hopi E. Hoekstra. 2012. 'Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species', *PLoS One*, 7: e37135-NA.

Pimm, S. L., C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, L. N. Joppa, P. H. Raven, C. M. Roberts, and J. O. Sexton. 2014. 'The biodiversity of species and their rates of extinction, distribution, and protection', *Science*, 344: 1246752.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. 'Inference of population structure using multilocus genotype data', *Genetics*, 155: 945-59.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, PaμL I. W. de Bakker, Mark J. Daly, and Pak C. Sham. 2007. 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *Am. J. Hum. Genet.*, 81: 559-75.

R Core Team. 2018. "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012." In.

Ramsey, Justin, and Douglas W. Schemske. 1998. 'PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS', *Annu. Rev. Ecol. Syst.*, 29: 467-501.

Razkin, Oihana, Benjamín J. Gómez-Moliner, Katerina Vardinoyannis, Alberto Martínez-Ortí, and María J. Madeira. 2017. 'Species delimitation for cryptic species complexes: case study of Pyramidula (Gastropoda, Pulmonata)', *Zool. Scr.*, 46: 55-72.

Regan, Helen M., Mark Colyvan, and Mark A. Burgman. 2002. 'A taxonomy and treatment of uncertainty for ecology and conservation biology', *Ecol. Appl.*, 12: 618-28.

Renwick, Anna R., Catherine J. Robinson, Stephen T. Garnett, Ian Leiper, Hugh P. Possingham, and Josie Carwardine. 2017. 'Mapping Indigenous land management for threatened species conservation: An Australian case-study', *PLoS One*, 12: e0173876.

Rodzen, Jeff A., Thomas R. Famula, and Bernie May. 2004. 'Estimation of parentage and relatedness in the polyploid white sturgeon (Acipenser transmontanus) using a dominant marker approach for duplicated microsatellite loci', *Aquaculture*, 232: 165-82.

Rousseeuw, Peter J. 1987. 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics*, 20: 53-65.

Schierenbeck, Kristina A. 2017. 'Population-level genetic variation and climate change in a biodiversity hotspot', *Ann. Bot.*, 119: 215-28.

Schliep, Klaus, Alastair Alastair Potts, David A. Morrison, and Guido W. Grimm. 2016. "Intertwining phylogenetic trees and networks." In.: PeerJ Preprints.

Sgariglia, Erik A., and Kevin J. Burns. 2003. 'Phylogeography of the California Thrasher (Toxostoma Redivivum) Based on Nested-Clade Analysis of Mitochondrial-DNA Variation', *Auk*, 120: 346-61.

Sokal, Robert R. 1979. 'Testing Statistical Significance of Geographic Variation Patterns', *Syst. Zool.*, 28: 227-32.

Sork, Victoria L., Kevin Squire, PaµL F. Gugger, Stephanie E. Steele, Eric D. Levy, and Andrew J. Eckert. 2016. 'Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, Quercus lobata', *Am. J. Bot.*, 103: 33-46.

Stobie, Connor Seamus, Carel J. Oosthuizen, Michael J. Cunningham, and Paulette Bloomer. 2018. 'Exploring the phylogeography of a hexaploid freshwater fish by RAD sequencing', *Ecology and evolution*, 8: 2326-42.

Thomas, J. A., M. G. Telfer, D. B. Roy, C. D. Preston, J. J. D. Greenwood, J. Asher, R. Fox, R. T. Clarke, and J. H. Lawton. 2004. 'Comparative losses of British butterflies, birds, and plants and the global extinction crisis', *Science*, 303: 1879-81.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63: 411-23.

Walsh, Jennifer, Irby J. Lovette, Virginia L. Winder, Chris S. Elphick, Brian J. Olsen, W. Gregory Shriver, and Adrienne I. Kovach. 2017. 'Subspecies delineation amid phenotypic, geographic, and genetic discordance in a songbird', *Molecular ecology*, 26: 1242-55.

Wells, P. V. 1992. 'Four new species of Arctostaphylos from southern California and Baja California', *Four Seasons*, 9: 44-53.

———. 2000. *The Manzanitas of California: Also of Mexico and the World* (P.V. Wells).

Wells, Philip V. 1968. 'New taxa, combinations, and chromosome numbers in Arctostaphylos (Ericaceae)', *Madroño*, 19: 193-210.

Wieslander, A. E., and Beryl O. Schreiber. 1939. 'Notes on the genus Arctostaphylos', *Madroño*, 5: 38-47.

Zhang, Jiajie, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. 2014. 'PEAR: a fast and accurate Illumina Paired-End reAd mergeR', *Bioinformatics*, 30: 614-20.

Zink, Robert M., Dale F. Lott, and Daniel W. Anderson. 1987. 'Genetic Variation, Population Structure, and Evolution of California Quail', *Condor*, 89: 395-405.

Zurbuchen, Antonia, Lisa Landert, Jeannine Klaiber, Andreas Müller, Silke Hein, and Silvia Dorn. 2010. 'Maximum foraging ranges in solitary bees: only few individuals have the capability to cover long foraging distances', *Biological Conservation*, 143: 669-76.

4.7 Appendix

Appendix S4.1 Modification of DNA extraction using the Qiagen DNEasy Plant Mini Kit (Qiagen: Hilden, Germany). We modified the manufacturer's protocol in the following ways: (1) the lysis buffer was heated to 60° C before mixing with the tissue sample, (2) TE buffer was preheated to 60° C before using for elution, and (3) the extraction column was eluted only once to improve concentration.

Appendix S4.2 LGC ddRAD-seq protocol:ddRAD library construction on 96 samples *Actostaphylos*;PstI-MspI (provided by LGC, Berlin, Germany)

1. Restriction digest: 100-200 ng of genomic DNA were digested with 2 Unit each MspI and PstI-HF(NEB) in 1 times Cutsmart buffer in 20µl volume for 2 hours at 37°C. The restriction enzymes were heat inactivated by incubation at 80°C for 20 min.

2. ddRAD library construction:

   - Ligation Reaction: 10 µl of each restriction digest were transferred to a new 96-well PCR plate, mixed on ice first with 1.5 µl of one of 96 inline-barcoded forward PstI Adaptors (pre-hybridized, concentration 5 pM/µl), followed by addition of 20µl Ligation master mix (contains: 15 µl NEB Quick ligation buffer, 0.4 µl NEB Quick Ligase, 5 pM prehybridized common reverse MspI Adaptor). Ligation reactions were incubated for 1h at RT, followed by heat inactivation for 10 min at 65°C.

   - Library purification: all reactions were diluted with 30 µl TE 10/50 (10mM Tris/HCl, 50mM EDTA, pH:8.0) and mixed with 50 µl Agencourt XP beads, incubated for 10 min at RT and placed for 5 min on a magnet to collect the beads. The supernatant was discarded and the beads were washed two times with 200 µl 80% Ethanol. Beads were air dried for 10 min and libraries were eluted in 20 µl Tris Buffer (5 mM Tris/HCl pH:9).

   - Library amplification: 10 µl of each of the 96 Libraries were separately amplified in 20 µl PCR reactions using MyTaq (Bioline) and standard Illumina TrueSeq amplification primers. Cycle number was limited to 14 Cycles.

3. Pooling and clean up of ddRAD libraries: 5 µl from each of the 96 amplified libraries were pooled. PCR primer and small amplicons were removed by Agencourt XP bead purification using 1 Volume of beads. The PCR enzyme was removed by an additional purification on Qiagen MinElute Columns. The pooled library was eluted in a final volume of 20µl Tris Buffer (5 mM Tris/HCl pH:9).

4. Normalisation: Normalisation was done using Trimmer Kit (Evrogen). 1 µg pooled ddRAD library in 12 µl was mixed with 4 µl 4x hybridization buffer, denatured for 3 min at 98°C and incubated for 5 hours at 68°C to allow reassociation of DNA fragments. 20 µl of 2x DSN master buffer was added and the samples were incubated for 10 min at 68°C. One Unit of DSN enzyme (1U/µl) was added and the reaction was incubated for another 30 min. Reaction was terminated by the addition of 20µl DSN Stop Solution, purified on a Qiagen MinElute Column and eluted in 10µl Tris Buffer (5 mM Tris/HCl pH:9).

5. Reamplification: The normalized library pool was re-amplified in 100µl PCR reactions using MyTaq (Bioline). An i5-Adaptor primer was used to include an i5-Index into the library, allowing parallel sequencing of multiple libraries on the Illumina NextSeq 500 sequencer. Cycle number was limited to 14 cycles.

6. Size selection: The ddRAD library was size selected on Blue Pippin, followed by a second size selection on a LMP-Agarose gel, removing fragments smaller than 200 bp and those larger than 500 bp.

7. Sequencing: Sequencing was done on an Illumina NextSeq 500 using V2 Chemistry (300 cycles)

Appendix S4.3 UCR ddRAD-seq protocol

1. **Primer and adapter sequences (similar to Peterson et al. 2012):**

   PstI adapter p1.1

   ACACTCTTTCCCTACACGACGCTCTTCCGATCTnnnnnnTGCA

   PstI adapter p1.2

   NNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

   where “NNNNNNN” is a unique 4-7 bp sequence for each of 96 barcoded adapters. If you plan to use PstI, avoid ending the barcode with “C.”

   MspI adapter 1:

   GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

   MspI adapter 2:

   CGAGATCGGAAGAGCGAGAACAA

   ILLPCR1 primer:

   AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG

   Indexed PCR2 Primers: each unique pcr2 primer lets you reuse the 96 barcodes in the same lane. E.g., if you plan to run 192 samples/lane, order two pcr2 primers. If using two indexed primers, Illumina recommends 6 and 12. If three primers, use 4, 16, 12. If six primers, 2,4,5,6,7,12.

   ILLPCR2-06

   CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGC

   ILLPCR2-12

   CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGC

2. **Preparation of specialized reagents:**

   - Barcoded SbfI/PstI adapter combinations:

PstI adapter p1.1(100 µM stock)	5 uL
PstI adapter p1.2(100 µM stock)	5 uL
water	90uL

Anneal oligo pairs by mixing 5 µL of each oligo in a pair (100 µM stock) with 90 µL of water to make 100 µL of 5 pmole/µL (5 uM) of annealed, doubled stranded adaptor stock. Heat to 95° C for 5 minutes and slowly cool to room temperature. Keep the set of adaptors organized in plate format that is convenient for later use in setting up reactions.

- MspI adapter:

Mix 100 µL of the MspI-adap1 and MspI-adap2 oligos (100 µM stock) with 800 µL of water to make 1000 µL of 10 pmole/µL (10 µM) stock.

MspI adapter p1.1(100 µM stock)

MspI adapter p1.1(100 µM stock)	100 uL
MspI adapter p1.2(100 µM stock)	100 uL
water	800 uL

Heat to 95° C for 5 minutes and slowly cool (0.1 C/s) to room temp (18 C) to anneal µoligos into double-stranded adaptor.

- PCR primers:

Mix 50 µL of the Illpcr1 and Illpcr2 oligos with 900 µL of water to make a working solution (5 µM of each oligo). The dual-indexing barcode is incorporated in the Illpcr2-## oligo, so this step must be repeated for each dual-indexing barcode (mixing each uniquely indexed version of Illpcr2 with Illpcr1, which will be the same oligo in all working solutions).

ILLPCR1 primer	50uL
ILLPCR2-06	50uL
water	900uL

ILLPCR1 primer	50uL
ILLPCR2-12	50uL
water	900uL

3. **Restriction Digest**

Restriction Master Mix

Reagent	1x	1.1x	96x
Cutsmart buffer (10x)	2 uL	2.2	211.2
PstI-HF	0.1 µL (= 2 units)	0.11	10.56
MspI	0.1 µL (= 2 units)	0.11	10.56

- Prepare the 96 well plate and load 100-200ng gDNA into each well and add specific amount of water to make the total volume to be 17.8µL.
- Add 2.2 µL Restriction Master Mix into every well.
- Incubate at 37°C for 3 hours.

4. **Adaptor Ligation**

Reagents	Volume for 1 sample	Volume for 9 samples (* 1.1)
10x Cutsmart	0.26uL	2.57uL
100 mM ATP	0.12uL	1.19uL
MspI adaptor (10pM/uL)	1uL	9.90uL
Water	0.05uL	0.50uL
T4 DNA Ligase	0.17uL	1.68uL

- Ligation master mix
- Mix 9 µL restriction digest with 1 µL of PstI adaptors (5pM/uL) on ice

- Add 1.6 µL ligation master mi
- The total reaction volume should now be 11.6 µL. Cover, seal and centrifuge the plate then incubate in thermocycler according to the following program:

  16°C for 3 hours (cover temp: 70°C, reaction volume: 12 µL)

  12°C indefinitely

5. Purification (short fragment removal) using Agencourt AMPure

- Mix the reaction mixture from step 4 with 39µL water
- Add 41.13µL Agencourt XP beads. Pipet 10 times. Incubate the mixture for 5 min at Room Temperature.
- Place the plate in magnet for 10min.
- Still on the magnet plate, discard the supernatant.
- Still on the magnet plate, add 190 µL 70% ethanol (freshly prepared same day
- from absolute ethanol) to each well.
- Incubate at room temperature for 1 min.
- Still on the magnet plate, aspirate out the ethanol and discard
- Still on the magnet plate, repeat the wash steps.
- Still on the magnet plate, let the plate at room temperature for 10 minutes to remove all traces of ethanol. Not too long; avoid over-drying the beads.
- Remove the plate from the magnet plate. Add 20 ìL ddH2O by pipetting the mix 10-30 times.
- Put the plate on the magnet plate again for 1 minute to separate the beads from the solution.
- Still on the magnet plate, transfer the eluate with DNA into a new plate. Be carefµL not to move the plate or move/touch the beads.

6. PCR Amplification

- This PCR step uses the Illumina PCR primers to amplify fragments that have our adapters + barcodes ligated onto the ends. To ameliorate stochastic differences in PCR production of fragments in reactions, we run four separate 5 µL reactions per restriction-ligation product, and later combine them.
- Prepare master mix III in a 2mL tube per the recipe below, vortex and centrifuge. Be sure to prepare separate master mixes for samples to be indexed with different Illumina index sequences- these will each require a different primer mix

Reagent	**1x (uL)**	**1.1x (uL)**	**1 Plate (uL)**
ddH_20	5.06	5.566	**534.336**
Q5 Buffer	4	4.4	**422.4**
dNTP (10mM)	0.4	0.44	**42.24**
PCR Primer Mix	1.34	1.474	**141.5**
Q5 High GC Enhancer	4	4.4	**422.4**
Q5 Hot Start DNA Polymerase	0.2	0.22	**21.12**

- Add 15 µL of the combined Master Mix III to the first set of 8 wells of a new 384 well plate, using a multichannel.
- Add 5µL of the diluted purified restriction-ligation DNA. Mix well by pipetting 10-20 times.
- Distribute 5µL into the 3 adjacent wells for each sample for a total of four 5µL reactions per restriction-ligation product.
- Run the reaction in a thermocycler according to the following PCR profile: 98°C for 30 sec (cover temp: 105°C, reaction volume: 5ìL); 20 cycles of: 98°C for 20 sec, 60°C for 30 sec, 72°C for 40 sec; 72°C for 2 min; 12°C indefinitely

- Pool the four PCR product of each sample together. Run >=3 uL(adjust if necessary) of PCR product in a gel to find out which samples successfully amplified.
- Pool all samples that successfully amplified, using 5 µL from each amplified sample.

7. **Clean-up of pooled dd-RAD library:** Use 1 Volume of Agencourt XP beads to purify the library and use 20µL Tris Buffer(5mM Tris/HCl pH: 9) to elute
8. **Normalization: Trimmer Kit (Evrogen):**
    - Mix 1 µg pooled ddRAD library in 12 µl total volume with 4 µl 4x hybridization buffer. This may require concentrating the library. Buffer will be the same as elution buffer in the bead purification step.
    - Denature for 3 min at 98°C.
    - Incubate for 5 hours at 68°C to allow reassociation of DNA fragments.
    - Add 20 µl of 2x DSN master buffer to the samples.
    - Incubate for 10 min at 68°C.
    - Add one Unit of DSN enzyme (1U/µl) and incubate the reaction for another 30 min.
    - Terminate the reaction by the addition of 20µl DSN Stop Solution.
    - Purify on a Qiagen MinElute Column and elute in 10µl Tris Buffer (5 mM Tris/HCl pH:9).
9. **Reamplification: 100 µL PCR system**
    - Prepare master mix III, again, in a 1mL/0.5mL tube per the recipe below, vortex and centrifuge. Remember to make separate master mixes if using indexing primers.

PCR (Master Mix III)

Reagent	1x (uL)	1.1x (uL)	2 sample (uL)
ddH_20	3.66	4.026	**8.052**
Q5 Buffer	4	4.4	**8.8**
dNTP (2.5mM)	1.6	1.76	**3.52**
PCR Primer Mix	1.34	1.474	**2.948**
Q5 High GC Enhancer	4	4.4	**8.8**
Q5 Hot Start DNA Polymerase	0.4	0.44	**0.88**

- Add 10 µL of the normalized library and 30 µL of Master Mix III to a new PCR tube, and distribute this volume across 8 wells so that you have 4 separate replicates of 10 µL each. Mix well by pipetting 10-20 times.
- Run the reaction in a thermocycler according to the following PCR profile: 98°C for 30 sec (cover temp: 105°C, reaction volume: 10iL); 14 cycles of: 98°C for 20 sec, 60°C for 30 sec, 72°C for 40 sec; 72°C for 2 min; 12°C indefinitely
- Pool the four replicates (10 µL PCR product each) together for a total volume of 40 µL.
- Prepare Master Mix IV (see below, 2 µL per sample), remember to account for dual-indexing primers; they need to be prepared in separate mixes. In order to make the ingredient volumes tractable and avoid error associated with pipetting small volumes (< 1uL), prepare the volumes given below which are calculated for ten samples.

PCR (Master Mix IV)

Reagent	1x (uL)	1.1x (uL)	10 samples 1x (uL)
ddH_20	0.1	0.11	**1**
Q5 Buffer	0.4	0.44	**4**
dNTP (25mM)	0.16	0.176	**1.6**
PCR Primer Mix	1.34	1.474	**13.4**

- Add 4μL Master Mix IV to the pooled 40 μL PCR product in step
- Run the reaction in a thermocycler according to the following PCR profile: 98°C for 3 min (cover temp: 105°C, reaction volume: 44μL); 60°C for 2 min; 72°C for 12 min; 12°C indefinitely

10. Gel Purification and Size Selection:

- Fill a gel rig with new, clean TBE buffer. Run the PCR product on 2.5% agarose gel at 100 volts for 2.5 hours. Include a good ladder on multiple gel lanes so that a clear line can be visualized across the gel. Ethidium bromide in the gel will not interfere after gel purification. The best approach is to tape two gel combs together to allow for larger wells, and to load 50-80 μL of sample into 8-12 lanes.
- Cut the 200 bp to 500 bp region out of the gel using a scalpel. Minimize gel exposure to UV by turning off UV table after each gel extraction. When cutting the gel try to choose the region by avoiding clear bands resulting from repetitive elements that might appear on the DNA distribution.
- Purify the excised gel punches using MinElute gel extraction kit or similar gel purification kit. One spin column can be used for two gel punches.
- Pool all the gel extracted products, quantify them using the Qubit, and perform an AMPure purification step using a 1:1 ratio as described in step 3. Elution volume should be chosen according to the concentration found by Qubit quantification. A 15% loss should be expected while calculating the optimal elution volume. Volumes lower than 40 μL should be avoided in order to get a good elution on a 1.5ml tube AMPure purification.

11. Preparing final template for Illumina sequencing: Use the Qubit to measure DNA concentration of the prepared library. Also run 2 to 5 μL of library on agarose gel to

verify size and concentration. A total concentration of >5 ng/μL is ideal for Illumina sequencing, but we can go as low as 2 ng/μL.

Appendix S4.4: Number of herbarium records for each Eastwood manzanita subspecies included in the analyses of subspecies habitat differentiation.

Subspecies	**Number of herbarium records used in habitat differentiation analysis**
A. glandulosa subsp. adamsii	112
A. glandulosa subsp. crassifolia	66
A. glandulosa subsp. cushingiana	221
A. glandulosa subsp. gabrielensis	35
A. glandulosa subsp. glandulosa	830
A. glandulosa subsp. howellii	32
A. glandulosa subsp. leucophylla	66
A. glandulosa subsp. mollis	266

Appendix S4.5: URL, units, and resolution of environmental variables

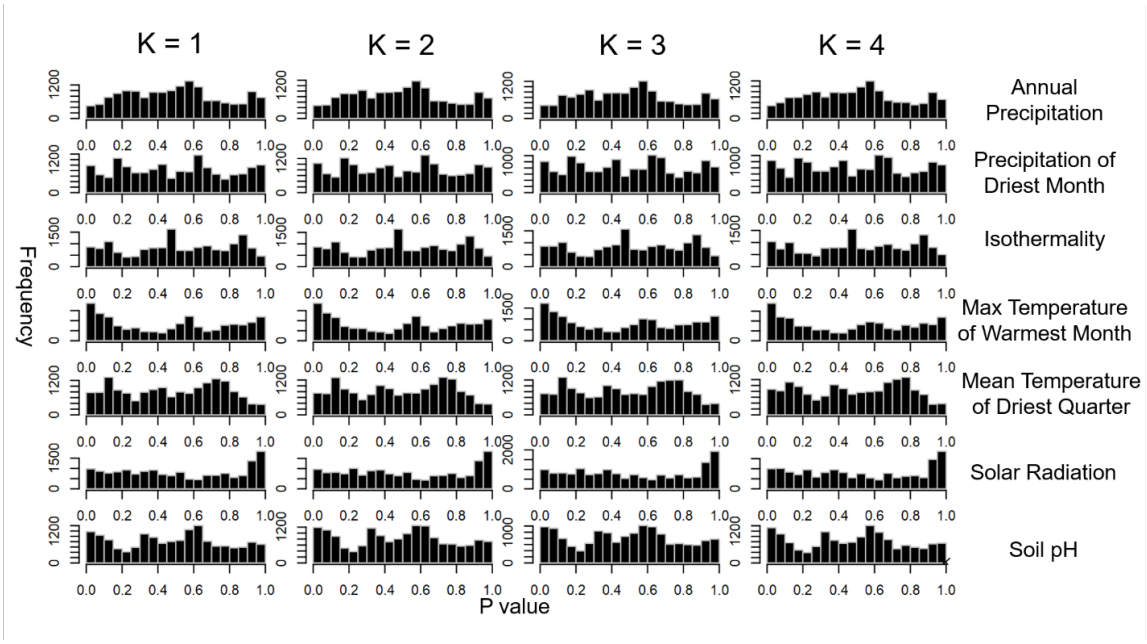
Variables	Link	Unit	Resolution
BIO1 Annual Mean Temperature	http://worldclim.org/version2	°C × 10	~1 km^2
BIO2 Mean Diurnal Range (Mean of monthly (max temp - min temp))	http://worldclim.org/version2	°C × 10	~1 km^2
BIO3 Isothermality (BIO2/BIO7) (* 100)	http://worldclim.org/version2	NA	~1 km^2
BIO4 Temperature Seasonality (standard deviation *100)	http://worldclim.org/version2	°C × 10	~1 km^2
BIO5 Max Temperature of Warmest Month	http://worldclim.org/version2	°C × 10	~1 km^2
BIO6 Min Temperature of Coldest Month	http://worldclim.org/version2	°C × 10	~1 km^2
BIO7 Temperature Annual Range (BIO5-BIO6)	http://worldclim.org/version2	°C × 10	~1 km^2
BIO8 Mean Temperature of Wettest Quarter	http://worldclim.org/version2	°C × 10	~1 km^2
BIO9 Mean Temperature of Driest Quarter	http://worldclim.org/version2	°C × 10	~1 km^2
BIO10 Mean Temperature of Warmest Quarter	http://worldclim.org/version2	°C × 10	~1 km^2
BIO11 Mean Temperature of Coldest Quarter	http://worldclim.org/version2	°C × 10	~1 km^2
BIO12 Annual Precipitation	http://worldclim.org/version2	mm	~1 km^2
BIO13 Precipitation of Wettest Month	http://worldclim.org/version2	mm	~1 km^2
BIO14 Precipitation of Driest Month	http://worldclim.org/version2	mm	~1 km^2
BIO15 Precipitation Seasonality (Coefficient of Variation)	http://worldclim.org/version2	mm	~1 km^2
BIO16 Precipitation of Wettest Quarter	http://worldclim.org/version2	mm	~1 km^2
BIO17 Precipitation of Driest Quarter	http://worldclim.org/version2	mm	~1 km^2
BIO18 Precipitation of Warmest Quarter	http://worldclim.org/version2	mm	~1 km^2
BIO19 Precipitation of Coldest Quarter	http://worldclim.org/version2	mm	~1 km^2
Solar Radiation	http://worldclim.org/version2	kJ m^{-2} day^{-1}	~1 km^2
Soil pH	https://soilgrids.org/	NA	~0.25 km^2

Appendix S4.6: Correlation of environmental variables evaluated for the analysis of subspecies habitat differentiation. Highly correlated variables were excluded. Variables included in the habitat differentiation analysis are marked with an asterisk.

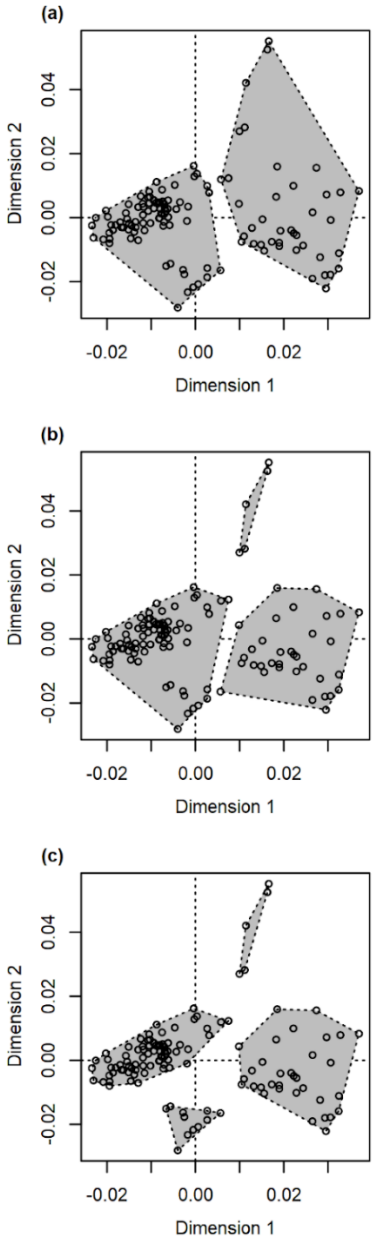
BIO1	BIO2	BIO3 *	BIO4	BIO5 *	BIO6	BIO7	BIO8	BIO9 *	Solar Radiation*	BIO 10	BIO 11	BIO 12*	BIO 13	BIO 14*	BIO 15	BIO 16	BIO 17	BIO 18	Soil pH*	BIO 19	
1	-0.32	0.72	-0.68	0.02	0.86	-0.56	0.96	0.59	-0.55	0.66	0.94	-0.8	-0.71	-0.29	0.28	-0.73	-0.44	-0.56	0.57	1	BIO1
-0.32	1	-0.18	0.78	0.88	-0.7	0.91	-0.52	0.03	0.85	0.36	-0.56	0.32	0.16	0.64	-0.64	0.21	0.67	0.73	-0.14	-0.32	BIO2
0.72	-0.18	1	-0.76	-0.18	0.72	-0.57	0.79	0.09	-0.39	0.2	0.8	-0.67	-0.61	-0.36	0.2	-0.62	-0.37	-0.45	0.34	0.72	BIO3*
-0.68	0.78	-0.76	1	0.68	-0.94	0.97	-0.86	-0.08	0.84	0.09	-0.89	0.64	0.48	0.65	-0.58	0.52	0.71	0.8	-0.28	-0.68	BIO4
0.02	0.88	-0.18	0.68	1	-0.46	0.81	-0.24	0.41	0.64	0.72	-0.31	0.13	0.01	0.62	-0.53	0.04	0.54	0.55	0.03	0.02	BIO5*
0.86	-0.7	0.72	-0.94	-0.46	1	-0.89	0.96	0.33	-0.84	0.22	0.98	-0.73	-0.59	-0.55	0.53	-0.63	-0.67	-0.78	0.39	0.86	BIO6
-0.56	0.91	-0.57	0.97	0.81	-0.89	1	-0.76	-0.01	0.88	0.23	-0.8	0.55	0.39	0.68	-0.62	0.43	0.72	0.8	-0.24	-0.56	BIO7
0.96	-0.52	0.79	-0.86	-0.24	0.96	-0.76	1	0.45	-0.71	0.43	1	-0.8	-0.68	-0.44	0.42	-0.71	-0.58	-0.69	0.49	0.96	BIO8
0.59	0.03	0.09	-0.08	0.41	0.33	-0.01	0.45	1	-0.29	0.74	0.4	-0.2	-0.13	-0.03	0.33	-0.14	-0.39	-0.4	0.31	0.59	BIO9*
-0.55	0.85	-0.39	0.84	0.64	-0.84	0.88	-0.71	-0.29	1	0.1	-0.74	0.32	0.13	0.65	-0.77	0.18	0.78	0.84	-0.08	-0.55	Solar Radiation*
0.66	0.36	0.2	0.09	0.72	0.22	0.23	0.43	0.74	0.1	1	0.36	-0.44	-0.47	0.29	-0.21	-0.46	0.11	0.04	0.46	0.66	BIO10
0.94	-0.56	0.8	-0.89	-0.31	0.98	-0.8	1	0.4	-0.74	0.36	1	-0.79	-0.66	-0.48	0.45	-0.69	-0.61	-0.72	0.47	0.94	BIO11
-0.8	0.32	-0.67	0.64	0.13	-0.73	0.55	-0.8	-0.2	0.32	-0.44	-0.79	1	0.97	0.23	0.04	0.98	0.25	0.38	-0.68	-0.8	BIO12*
-0.71	0.16	-0.61	0.48	0.01	-0.59	0.39	-0.68	-0.13	0.13	-0.47	-0.66	0.97	1	0.03	0.26	1	0.04	0.17	-0.7	-0.71	BIO13
-0.29	0.64	-0.36	0.65	0.62	-0.55	0.68	-0.44	-0.03	0.65	0.29	-0.48	0.23	0.03	1	-0.8	0.08	0.82	0.78	0.06	-0.29	BIO14*
0.28	-0.64	0.2	-0.58	-0.53	0.53	-0.62	0.42	0.33	-0.77	-0.21	0.45	0.04	0.26	-0.8	1	0.21	-0.92	-0.86	-0.19	0.28	BIO15
-0.73	0.21	-0.62	0.52	0.04	-0.63	0.43	-0.71	-0.14	0.18	-0.46	-0.69	0.98	1	0.08	0.21	1	0.09	0.22	-0.69	-0.73	BIO16
-0.44	0.67	-0.37	0.71	0.54	-0.67	0.72	-0.58	-0.39	0.78	0.11	-0.61	0.25	0.04	0.82	-0.92	0.09	1	0.97	-0.04	-0.44	BIO17
-0.56	0.73	-0.45	0.8	0.55	-0.78	0.8	-0.69	-0.4	0.84	0.04	-0.72	0.38	0.17	0.78	-0.86	0.22	0.97	1	-0.15	-0.56	BIO18
1	-0.32	0.72	-0.68	0.02	0.86	-0.56	0.96	0.59	-0.55	0.66	0.94	-0.8	-0.71	-0.29	0.28	-0.73	-0.44	-0.56	0.57	1	BIO19
0.57	-0.14	0.34	-0.28	0.03	0.39	-0.24	0.49	0.31	-0.08	0.46	0.47	-0.68	-0.7	0.06	-0.19	-0.69	-0.04	-0.15	1	0.57	Soil pH*

Appendix S4.7 Subspecies identification and environmental data for Eastwood manzanita herbarium records included in analysis of subspecies habitat differentiation.

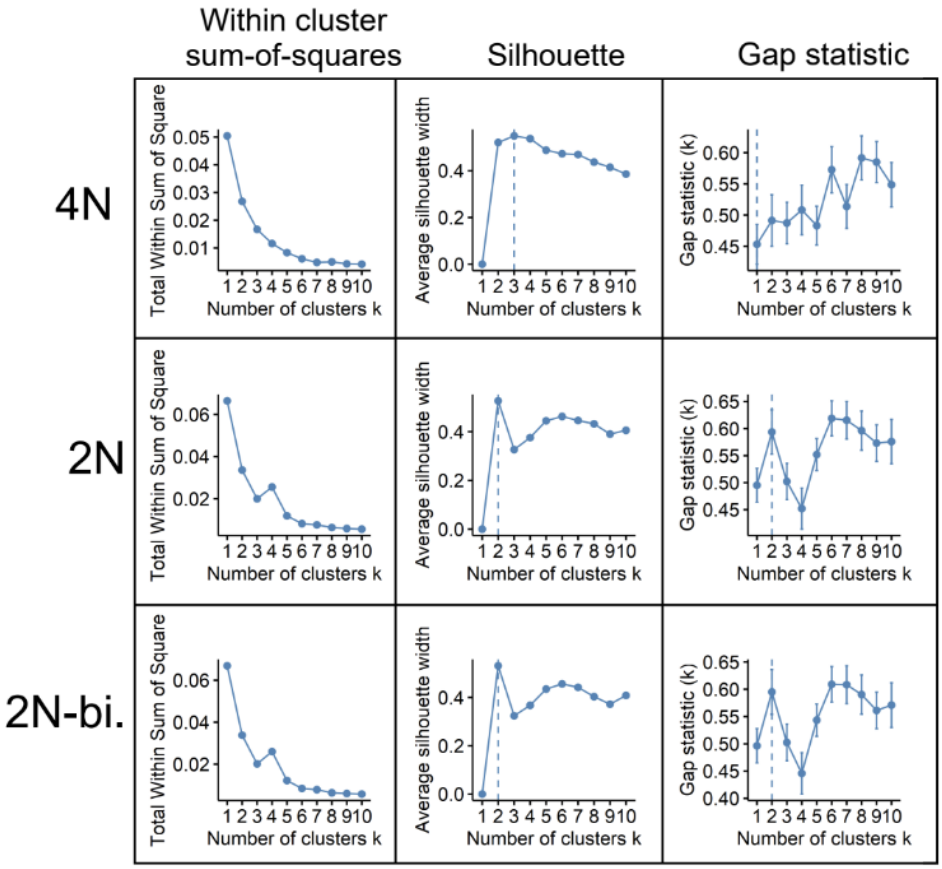
Appendix S4.8 Histograms of adjusted p-values of LFMM analyses for seven climatic variables when K=1, 2, 3, and 4.



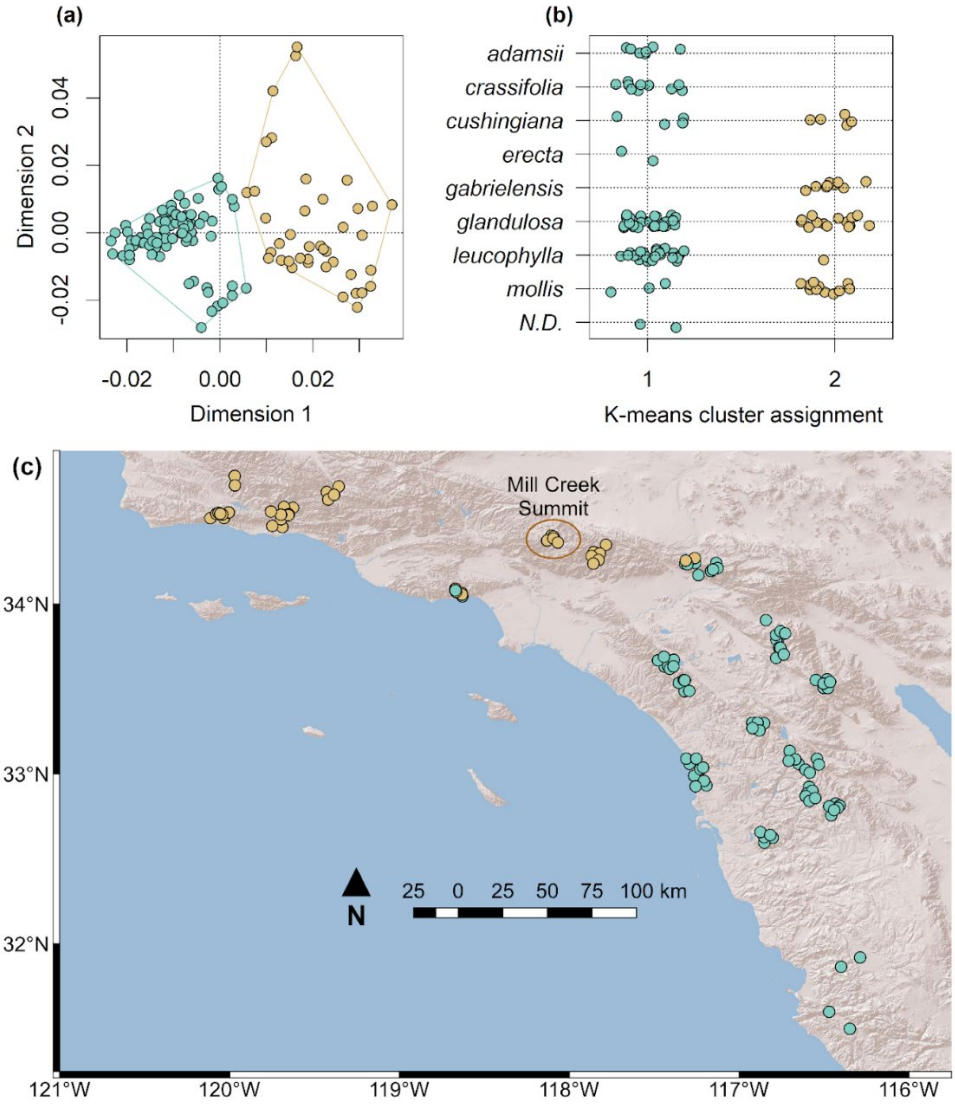
Appendix S4.9 K-means clustering results for *k* = 2 to *k* = 4 on a two-dimensional MDS of the 4N data set. Polygons mark the boundaries of the resulting clusters of samples at each k value. Results for the 2N and 2N-biallelic data sets differed in details but, like the 4N results, did not produce conclusive results and are not shown.



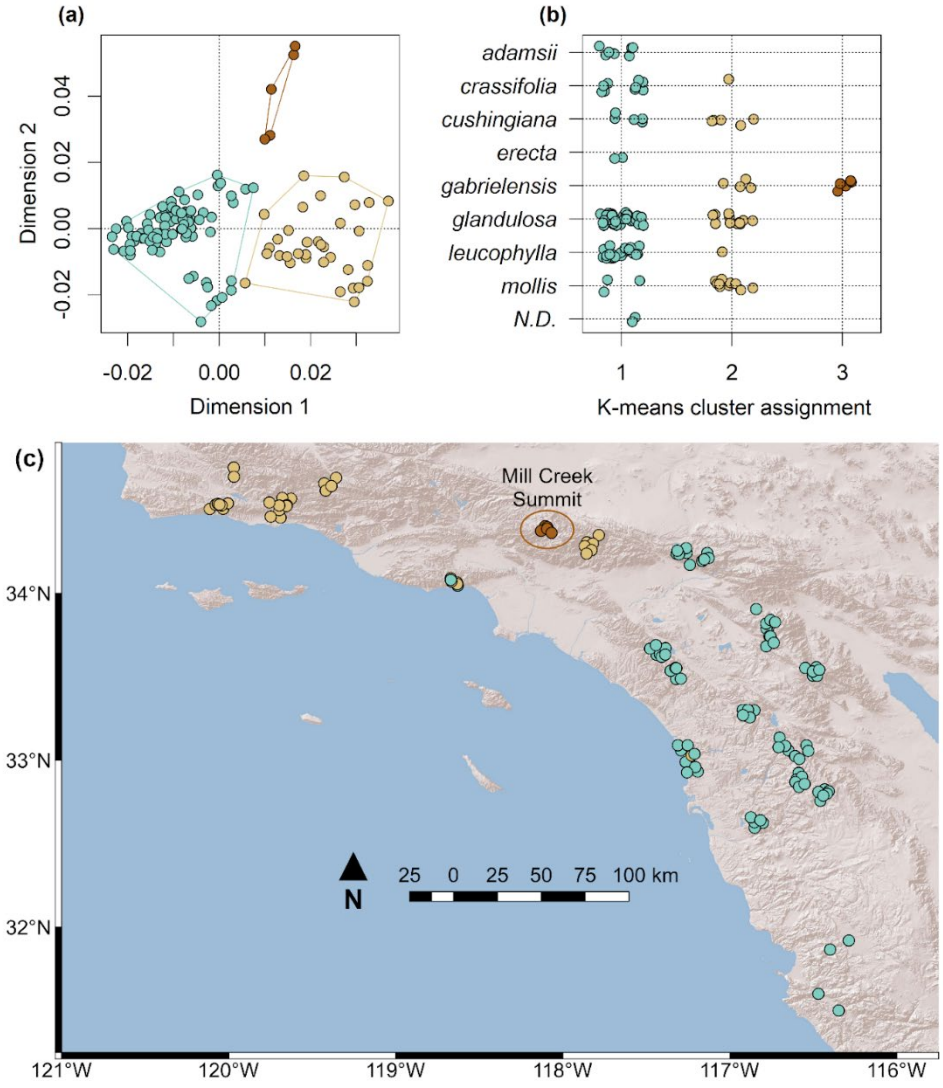
Appendix S4.10 Statistical evaluations for the best supported number of clusters in the k-means clustering analyses of MDS dimensions. The top, middle, and bottom rows show evaluations for analyses based on MDS of the 4N, 2N, and 2N-biallelic SNP datasets, respectively. The left, middle, and right columns show evaluations using the within cluster sum-of-squares method, silhouette, and gap statistic methods. Plots of gap statistic results (in the right column) show mean values for 100 bootstrap replicates as points, and error bars show plus or minus one standard error from the mean. Vertical dashed lines indicate the inferred best supported number of clusters. No vertical lines are provided for the within cluster sum-of-squares method, as this method relies on visual interpretation of the slope of the curve, rather than finding extrema.



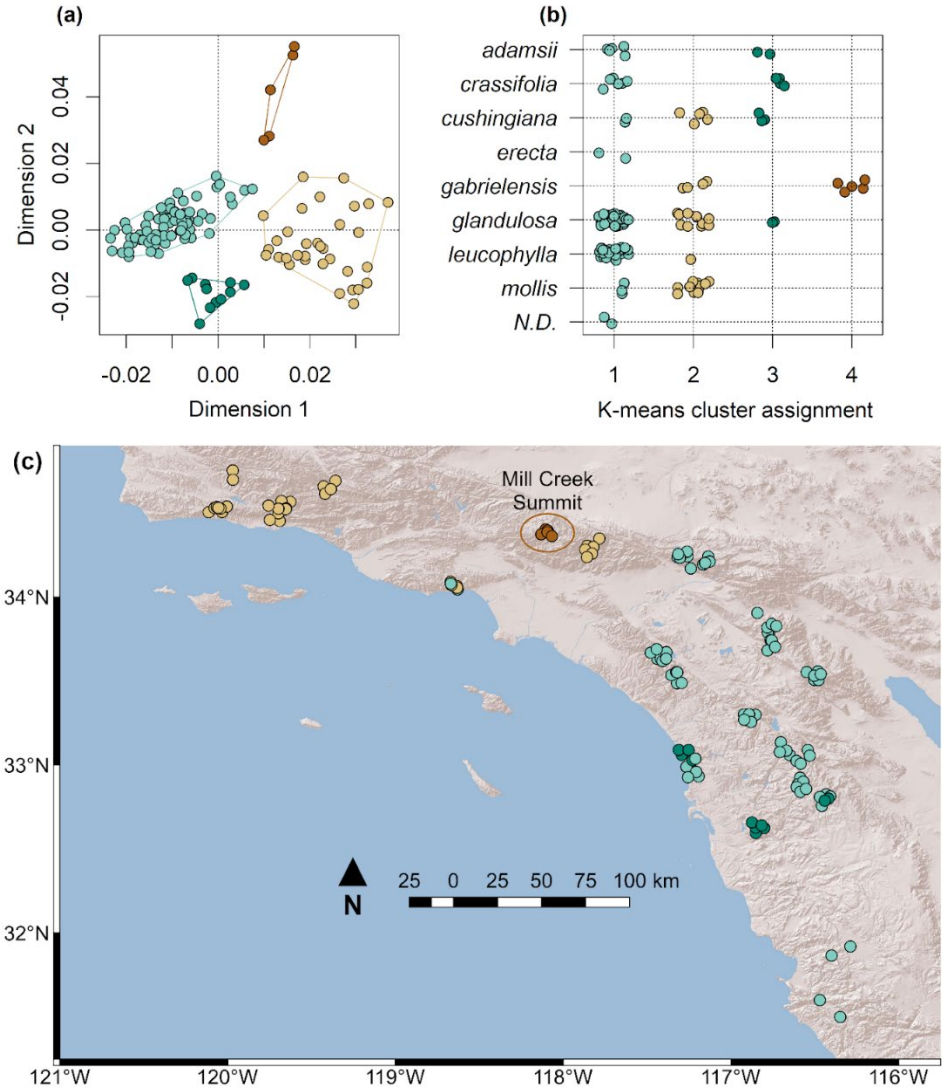
Appendix S4.11 Results of k-means clustering, for *k* = 2, on the MDS of the 4N SNP data set. a. Two-dimensional MDS. Polygons mark the boundaries of the k-means clusters. b. Plot of cluster assignment by subspecies identity. Points are individual samples, plotted with random jitter at each subspecies-cluster intersection. c. Map of collection localities for each sample, colored to indicate cluster assignment. Because some samples would overlap on the map, a random jitter value between -0.15 and 0.15 degrees longitude and latitude was applied to each point.



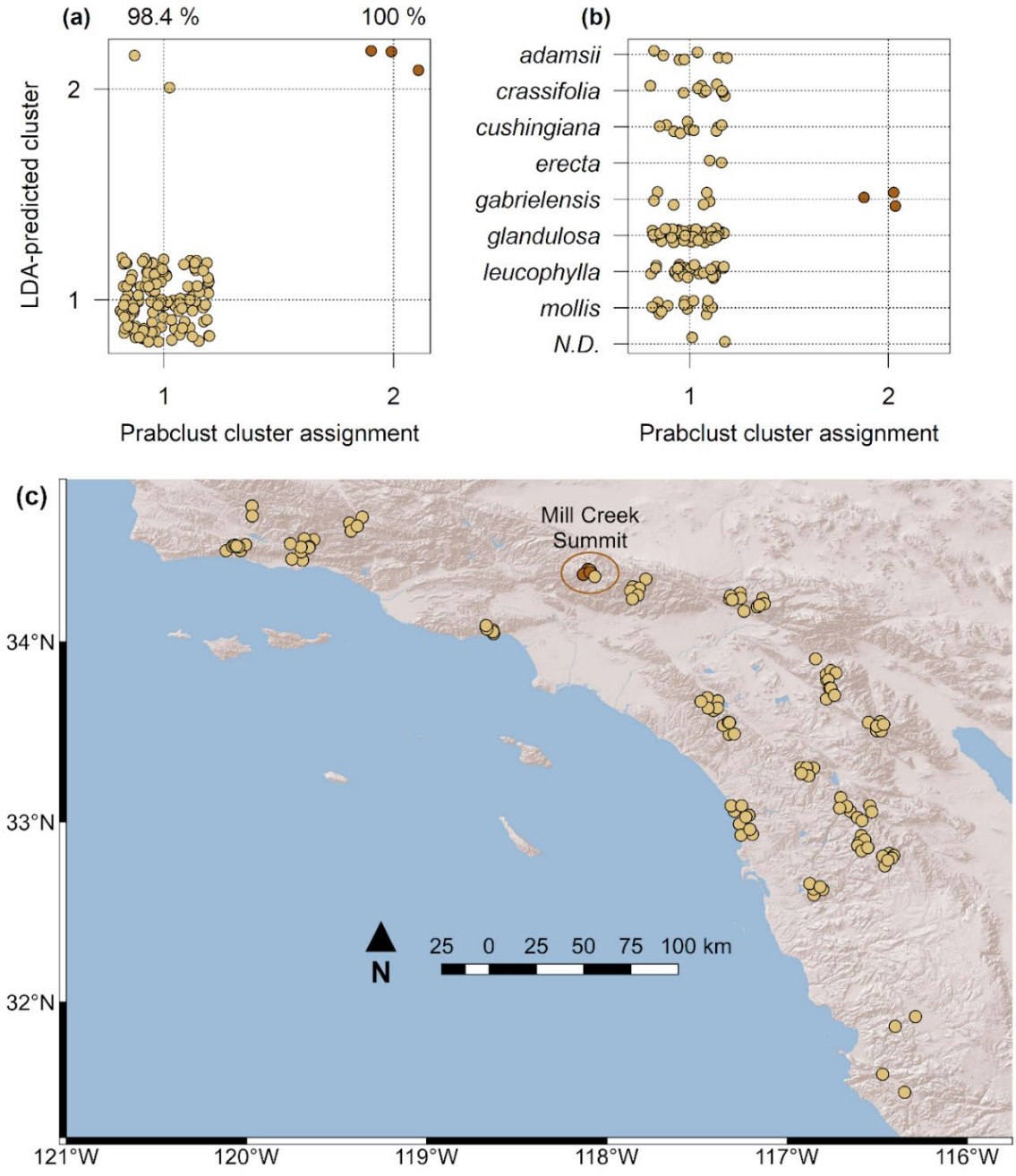
Appendix S4.12 Results of k-means clustering, for *k* = 3, on the MDS of the 4N SNP data set. a. Two-dimensional MDS. Polygons mark the boundaries of the k-means clusters. b. Plot of cluster assignment by subspecies identity. Points are individual samples, plotted with random jitter at each subspecies-cluster intersection. c. Map of collection localities for each sample, colored to indicate cluster assignment. Because some samples would overlap on the map, a random jitter value between -0.15 and 0.15 degrees longitude and latitude was applied to each point.



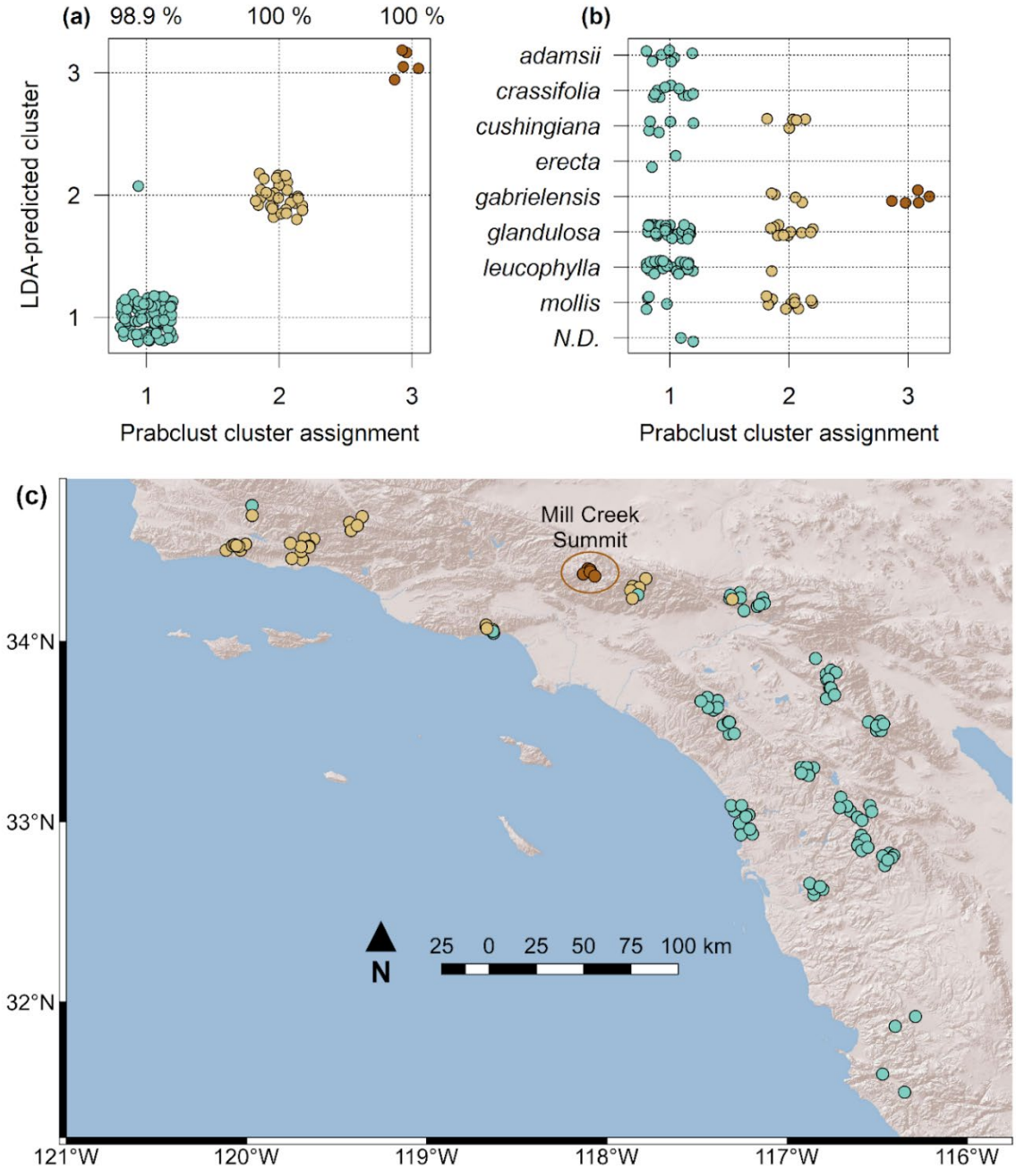
Appendix S4.13 Results of k-means clustering, for *k* = 4, on the MDS of the 4N SNP data set. a. Two-dimensional MDS. Polygons mark the boundaries of the k-means clusters. b. Plot of cluster assignment by subspecies identity. Points are individual samples, plotted with random jitter at each subspecies-cluster intersection. c. Map of collection localities for each sample, colored to indicate cluster assignment. Because some samples would overlap on the map, a random jitter value between -0.15 and 0.15 degrees longitude and latitude was applied to each point.



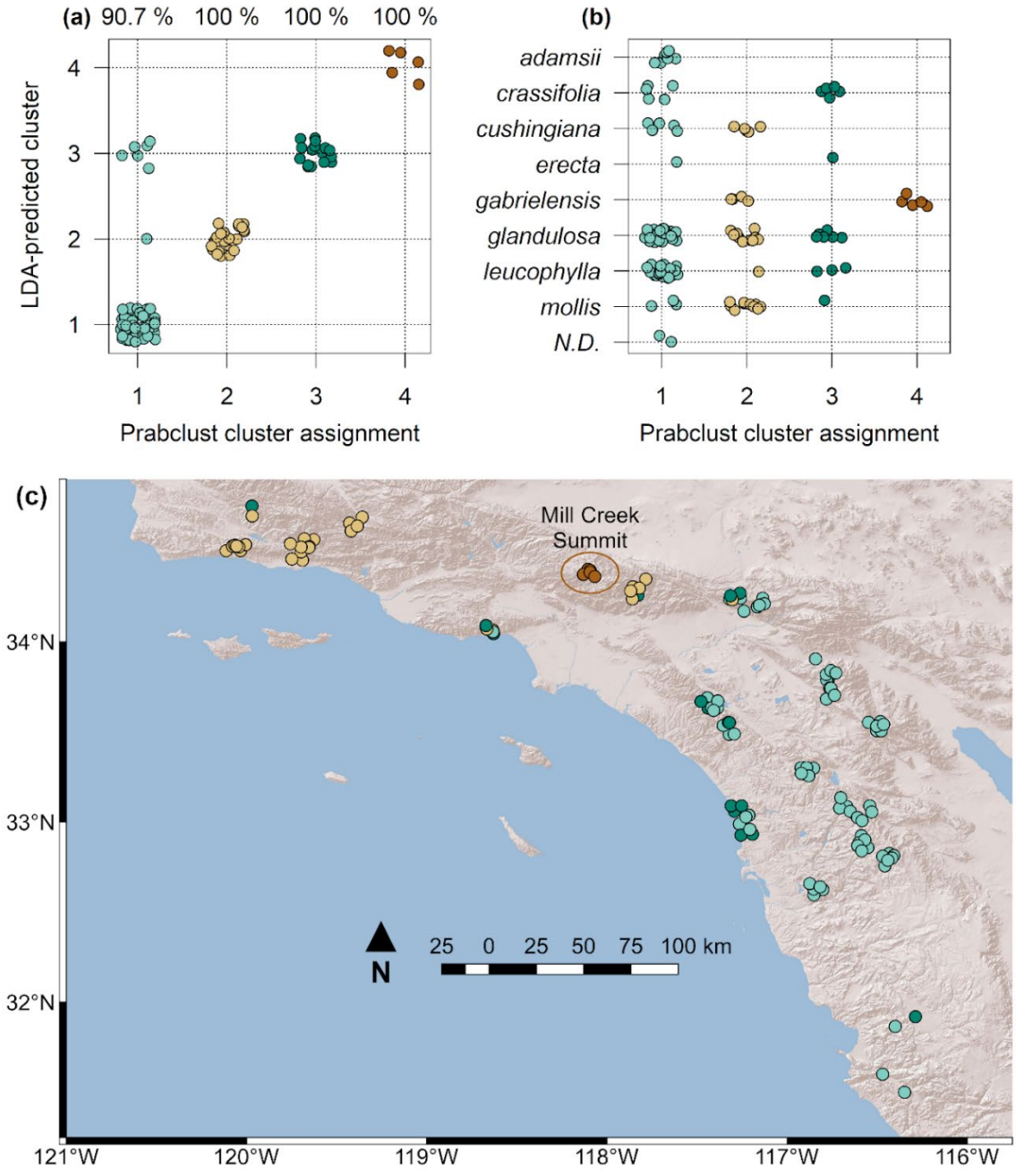
Appendix S4.14 Results of Gaussian clustering with prabclust, for *k* = 2, on the NMDS of the 4N SNP data set. a. Assessment of cluster distinction by linear discriminant analysis (LDA) with leave-one-out cross-validation. X-axis shows cluster assignment and y-axis shows the LDA-predicted cluster assignment. Percentage values plotted at top are the percentage of samples from each cluster with correct LDA-predicted cluster assignment. b. Plot of cluster assignment by subspecies identity. Points are individual samples, plotted with random jitter at each subspecies-cluster intersection. c. Map of collection localities for each sample, colored to indicate cluster assignment. Because some samples would overlap on the map, a random jitter value between -0.15 and 0.15 degrees longitude and latitude was applied to each point.


(a)
98.4 %
100 %
LDA-predicted cluster
2
1
1
2
Prabclust cluster assignment
(b)
adamsii
crassifolia
cushingiana
erecta
gabrielensis
glandulosa
leucophylla
mollis
N.D.
1
2
Prabclust cluster assignment
(c)
Mill Creek
Summit
34°N
33°N
32°N
N
25 0 25 50 75 100 km
121°W
120°W
119°W
118°W
117°W
116°W

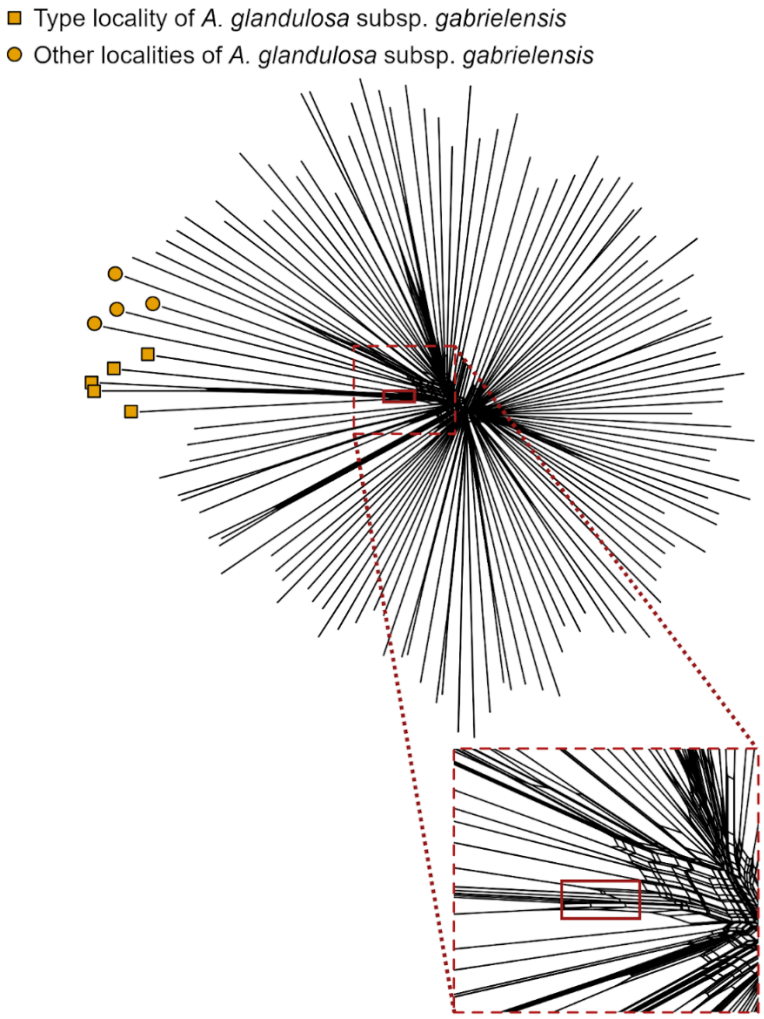
Appendix S4.15 Results of Gaussian clustering with prabclust, for *k* = 3, on the NMDS of the 4N SNP data set. a. Assessment of cluster distinction by linear discriminant analysis (LDA) with leave-one-out cross-validation. X-axis shows cluster assignment and y-axis shows the LDA-predicted cluster assignment. Percentage values plotted at top are the percentage of samples from each cluster with correct LDA-predicted cluster assignment. b. Plot of cluster assignment by subspecies identity. Points are individual samples, plotted with random jitter at each subspecies-cluster intersection. c. Map of collection localities for each sample, colored to indicate cluster assignment. Because some samples would overlap on the map, a random jitter value between -0.15 and 0.15 degrees longitude and latitude was applied to each point.


(a)
98.9 %
100 %
100 %
LDA-predicted cluster
Prabclust cluster assignment
(b)
adamsii
crassifolia
cushingiana
erecta
gabrielensis
glandulosa
leucophylla
mollis
N.D.
Prabclust cluster assignment
(c)
Mill Creek
Summit
34°N
33°N
32°N
N
25 0 25 50 75 100 km
121°W
120°W
119°W
118°W
117°W
116°W

Appendix S4.16 Results of Gaussian clustering with prabclust, for *k* = 4, on the NMDS of the 4N SNP data set. a. Assessment of cluster distinction by linear discriminant analysis (LDA) with leave-one-out cross-validation. X-axis shows cluster assignment and y-axis shows the LDA-predicted cluster assignment. Percentage values plotted at top are the percentage of samples from each cluster with correct LDA-predicted cluster assignment. b. Plot of cluster assignment by subspecies identity. Points are individual samples, plotted with random jitter at each subspecies-cluster intersection. c. Map of collection localities for each sample, colored to indicate cluster assignment. Because some samples would overlap on the map, a random jitter value between -0.15 and 0.15 degrees longitude and latitude was applied to each point.


(a) 90.7 % 100 % 100 % 100 %
LDA-predicted cluster
Prabclust cluster assignment
(b)
adamsii
crassifolia
cushingiana
erecta
gabrielensis
glandulosa
leucophylla
mollis
N.D.
Prabclust cluster assignment
(c)
Mill Creek
Summit
34°N
33°N
32°N
N
25 0 25 50 75 100 km
121°W
120°W
119°W
118°W
117°W
116°W

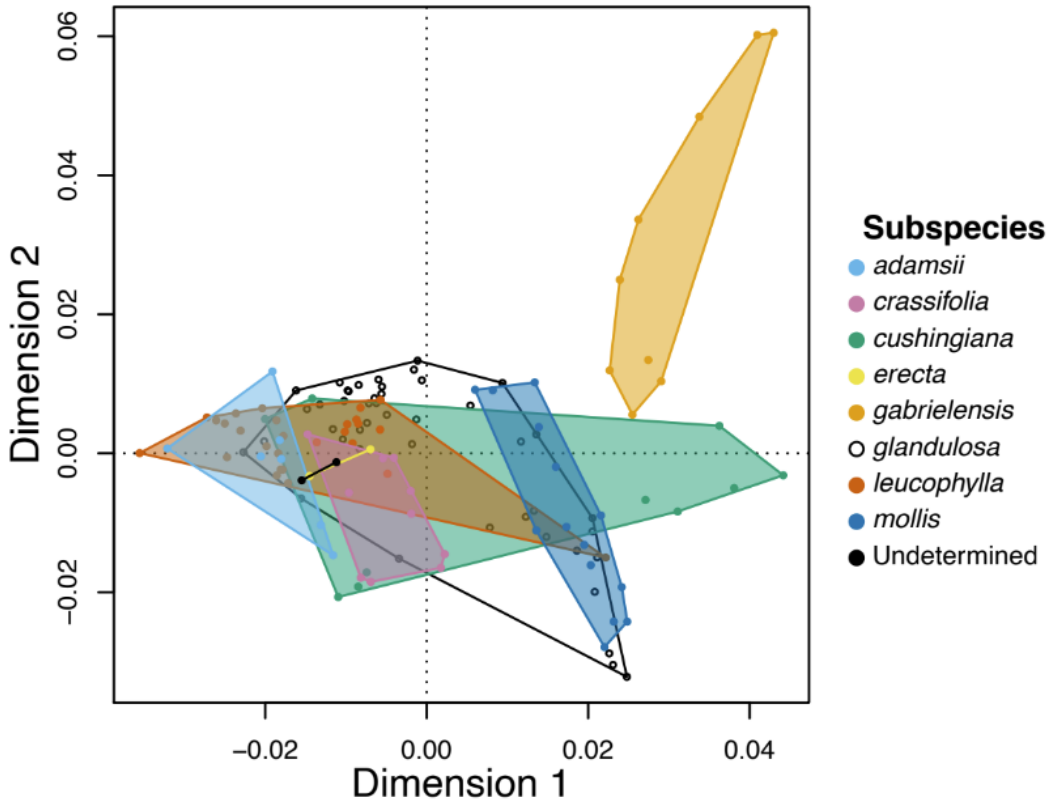
Appendix S4.17 NeighborNetwork of the 4N SNP data set, with tips representing samples of *A. glandulosa* subsp. *gabrielensis* marked with orange points. Square and round points represent samples from the type locality and from other localities, respectively. Unmarked tips represent samples not identified as *A. glandulosa* subsp. *gabrielensis*. The inset at bottom is a zoomed-in view of the shared edge length (segments in the red rectangle) unique to the five samples of *A. glandulosa* subsp. *gabrielensis* from the type locality.



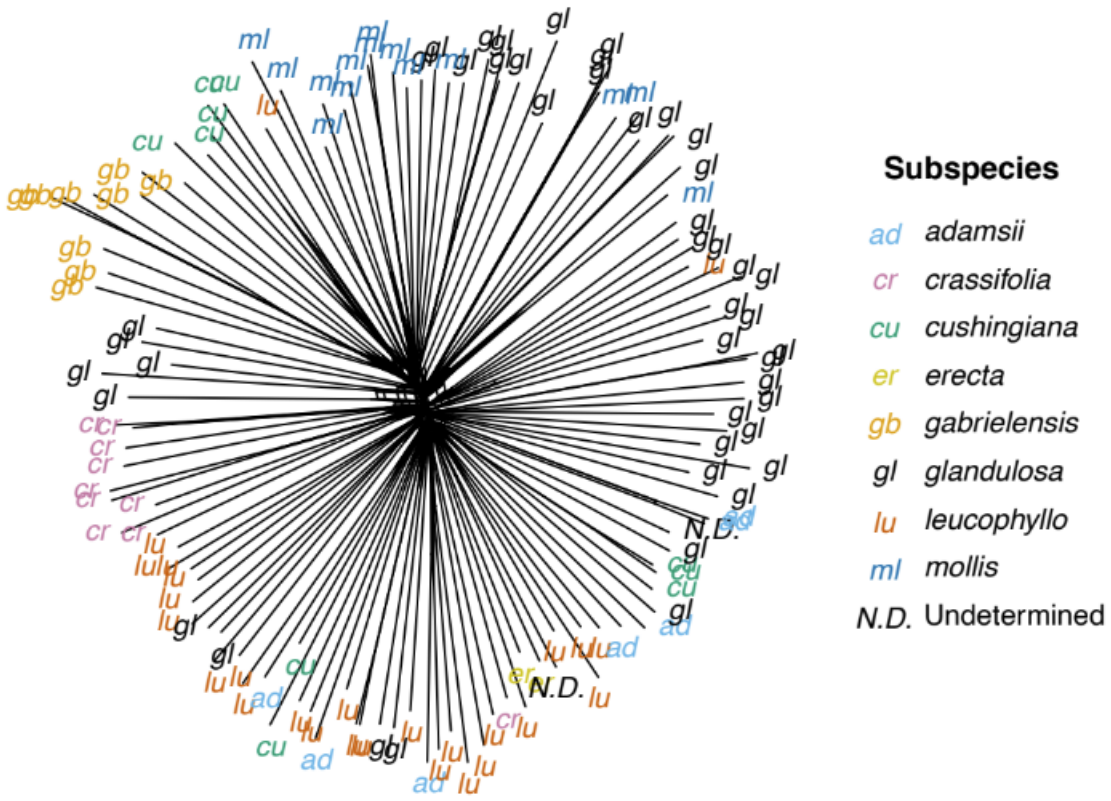
Appendix S4.18 Statistical values used to determine the optimal number of clusters (*k*) in STRUCTURE analyses, using the Delta-k method.

K	*Reps*	*Mean LnP(K)*	*Stdev LnP(K)*	*Ln'(K)*	*|Ln''(K)|*	*Delta K*
1	15	-29194.7733	0.128	NA	NA	NA
2	15	-28132.6333	0.9053	1062.140000	357.120000	394.487626
3	15	-27427.6133	1.4525	705.020000	263.113333	181.142865
4	15	-26985.7067	10.4442	441.906667	17.873333	1.711311
5	15	-26561.6733	22.9985	424.033333	59.486667	2.586543
6	15	-26197.1267	14.5175	364.546667	42.393333	2.920144
7	15	-25874.9733	24.9218	322.153333	31.860000	1.278400
8	15	-25584.68	23.2985	290.293333	1.686667	0.072394
9	15	-25292.7	40.8335	291.980000	NA	NA

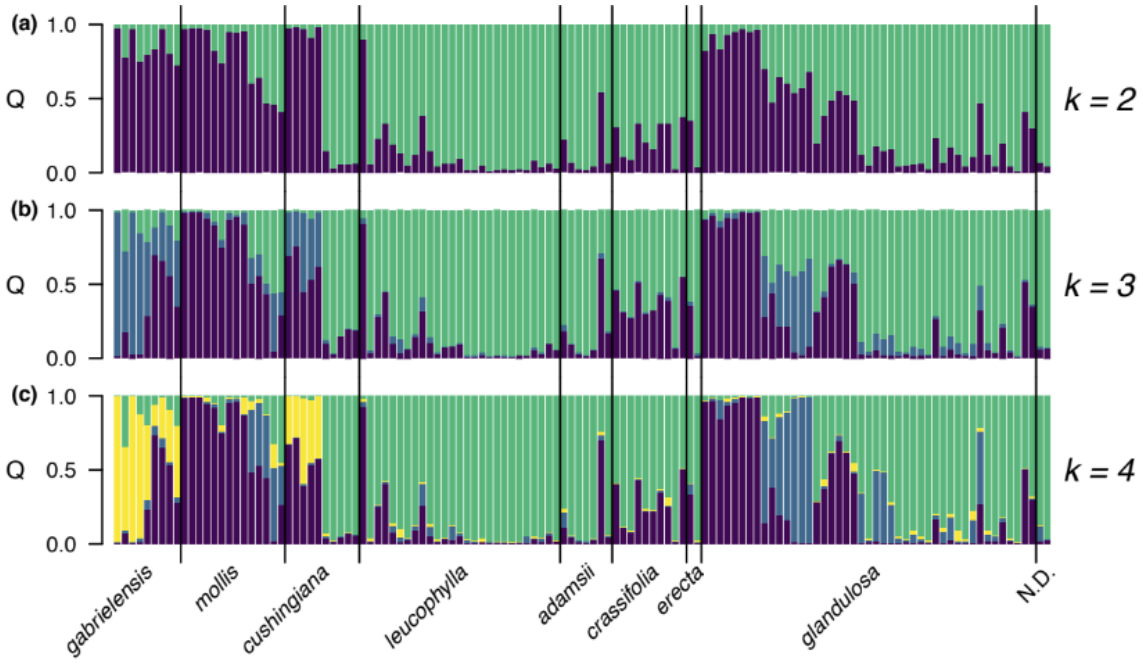
Appendix S4.19 MDS analysis of 2N-biallelic data set. Two dimensional representation of genetic distance among samples, calculated using multidimensional scaling (MDS). Points and polygons are colored by subspecies identification. Polygons are minimum areas that enclose all samples of each subspecies.



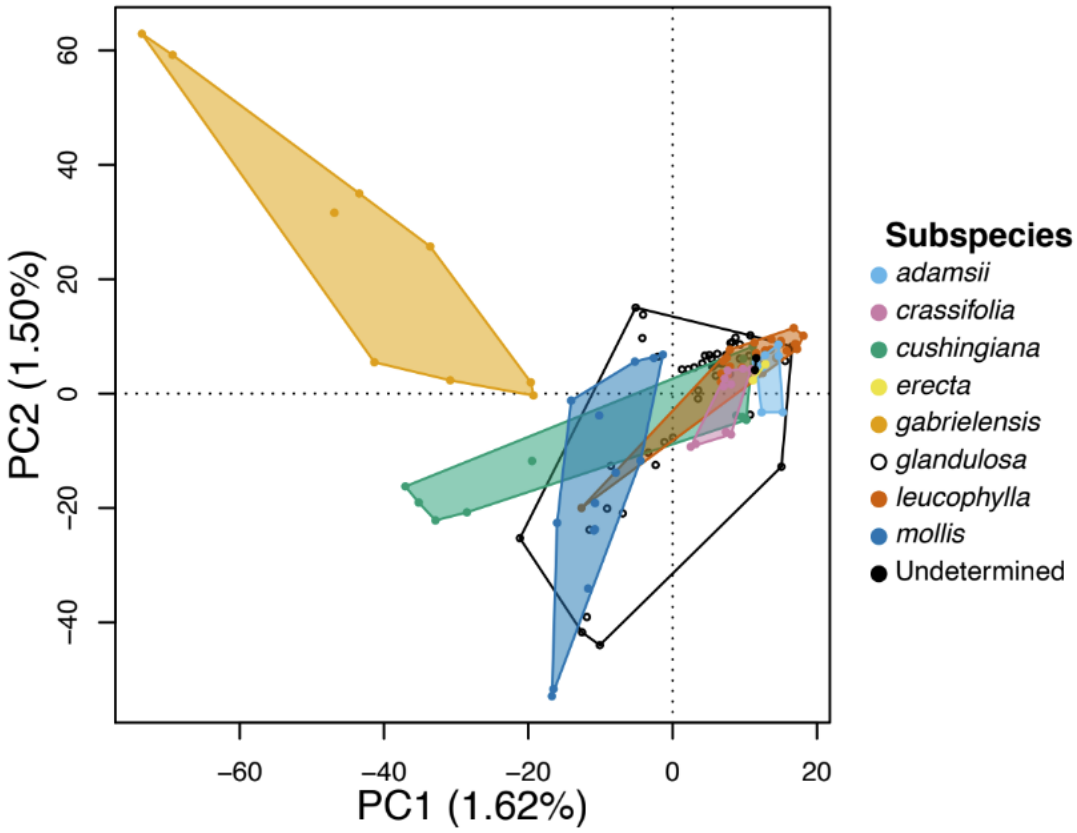
Appendix S4.20 NeighborNetwork for 2N-biallelic data set. Tips are labelled and colored by subspecies identification.



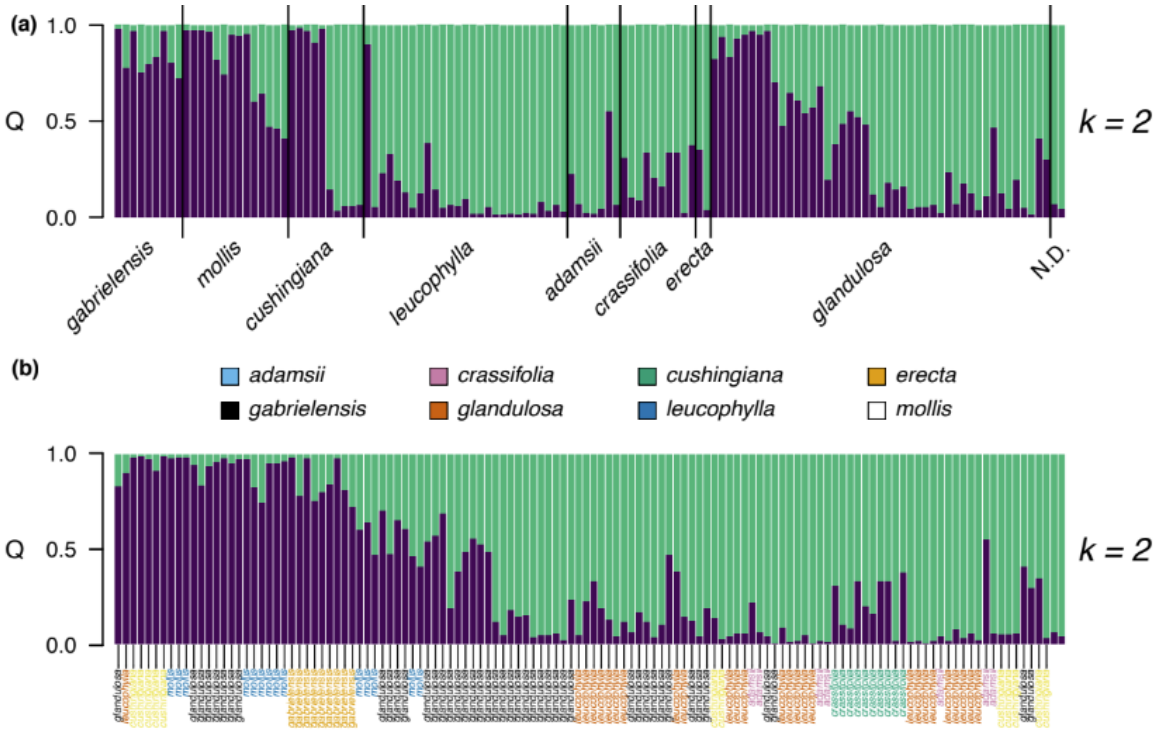
Appendix S4.21 STRUCTURE results for k = 2, 3 and 4 for the 2N-biallelic data set. Vertical bars represent individuals. Colors within bars represent ancestral clusters of differing genotypes. The proportion of each color in each bar represents the probability of assignment (Q) to each cluster. Individuals are sorted along the x-axis by subspecies, then by latitude of collection. Two samples undetermined to subspecies (farthest right in graph) are not labeled.



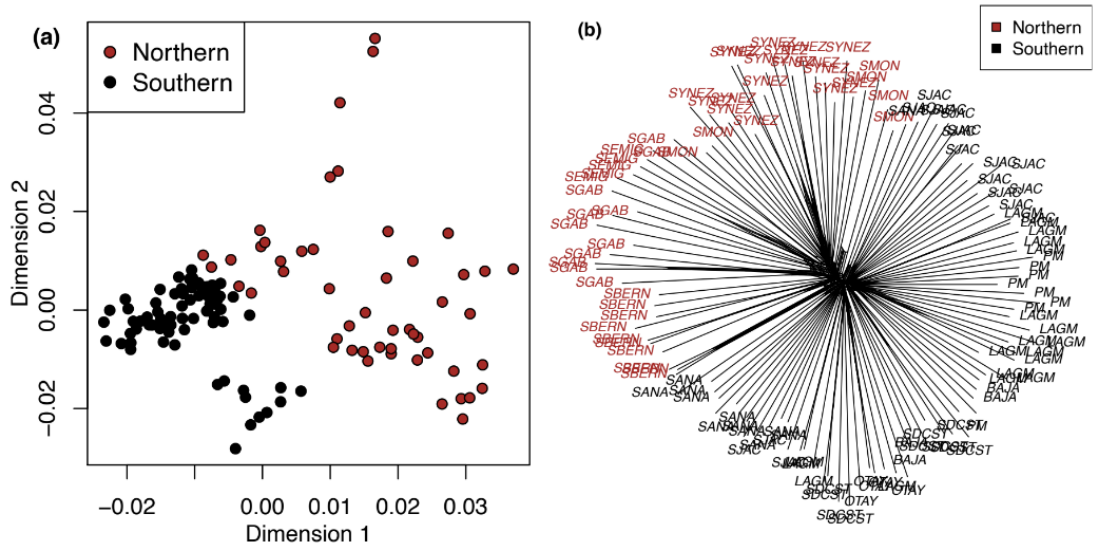
Appendix S4.22 PCA for the 2N-biallelic data set. Points represent individual samples. Polygons are the minimum areas that enclose all samples of a given subspecies. Points and polygons are colored by subspecies identification.



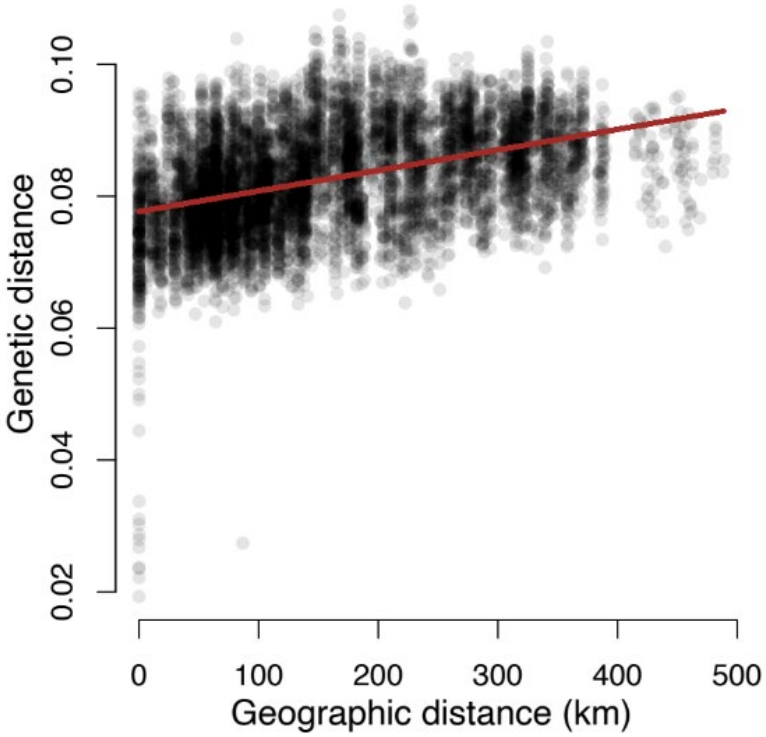
Appendix S4.23 Comparison of 4N STRUCTURE results sorted by subspecies then latitude (a), with results sorted by latitude then subspecies (b). Vertical bars represent individuals. Colors within bars represent ancestral clusters of differing genotypes. The proportion of each color in each bar represents the probability of assignment (Q) to each cluster. Two samples in (b) undetermined to subspecies (farthest right in graph) are not labeled.



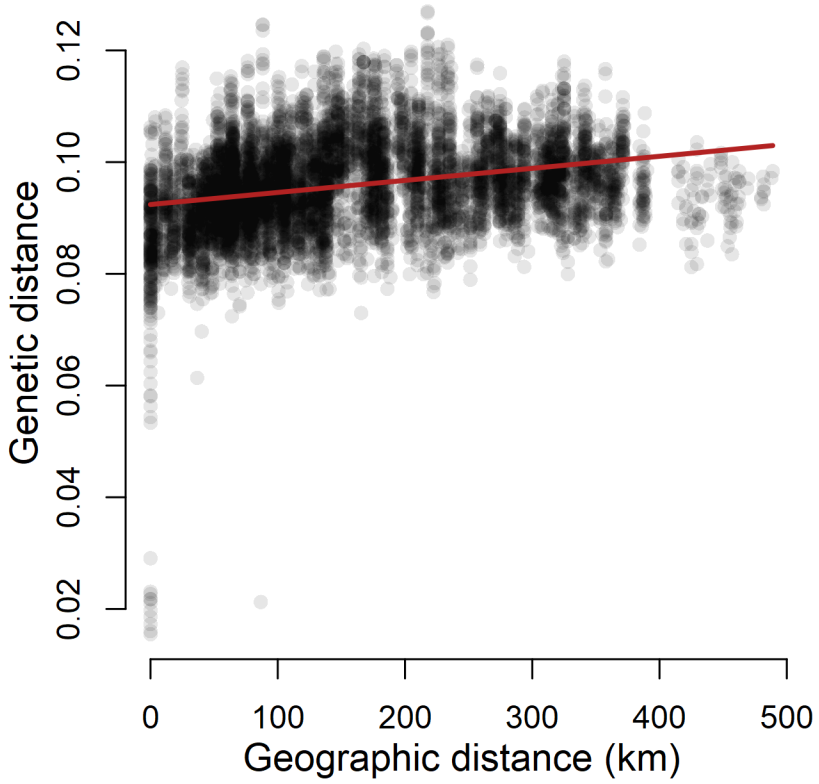
Appendix S4.24 Geographically labelled MDS (a) and NeighborNetwork (b) analyses of the 4N data set. (a) Two dimensional representation of genetic distance among samples in the 4N data set. Points are colored by region of collection: red points were collected in the northern part of the collection area (Transverse Ranges); black points were collected south of Transverse Ranges. (b) Black tips are samples collected south of the Transverse Ranges and red tips are those collected within the Transverse Ranges. Shorthand labels at tips indicate mountain range or area of collection: BAJA = NW Baja California; LAGM = Laguna Mountains; OTAY = Otay Mountain; PM = Palomar Mountain; SANA = Santa Ana Mountains; SBERN = San Bernardino Mountains; SDCST = San Diego Coast; SEMIG = San Emigdio Mountains; SGAB = San Gabriel Mountains; SJAC = San Jacinto Mountains; SMON = Santa Monica Mountains; SYNEZ = Santa Ynez Mountains.



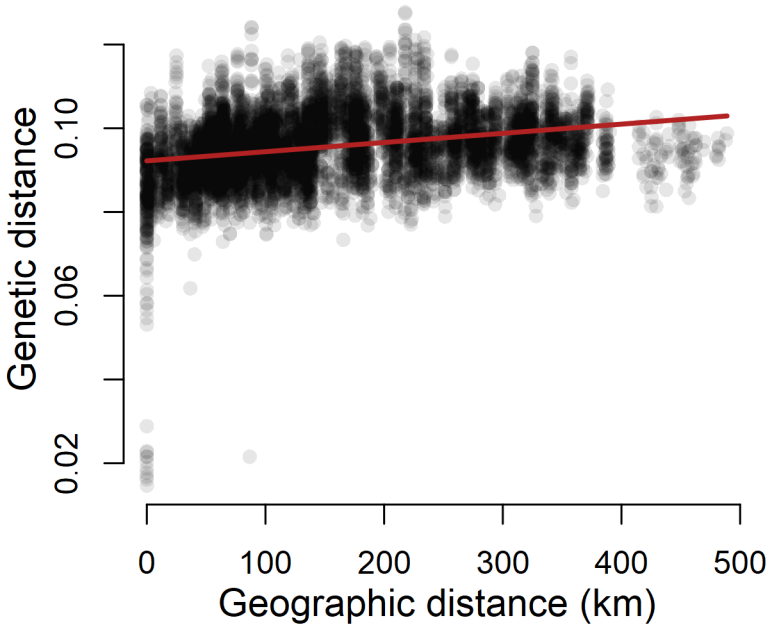
Appendix S4.25 Uncorrected *p* genetic distance plotted as a function of geographic distance (in kilometers) between every sample pair in the 4N data set. Points represent sample pairs. Points are plotted with 5% opacity to aid visual interpretation, as point density is high.



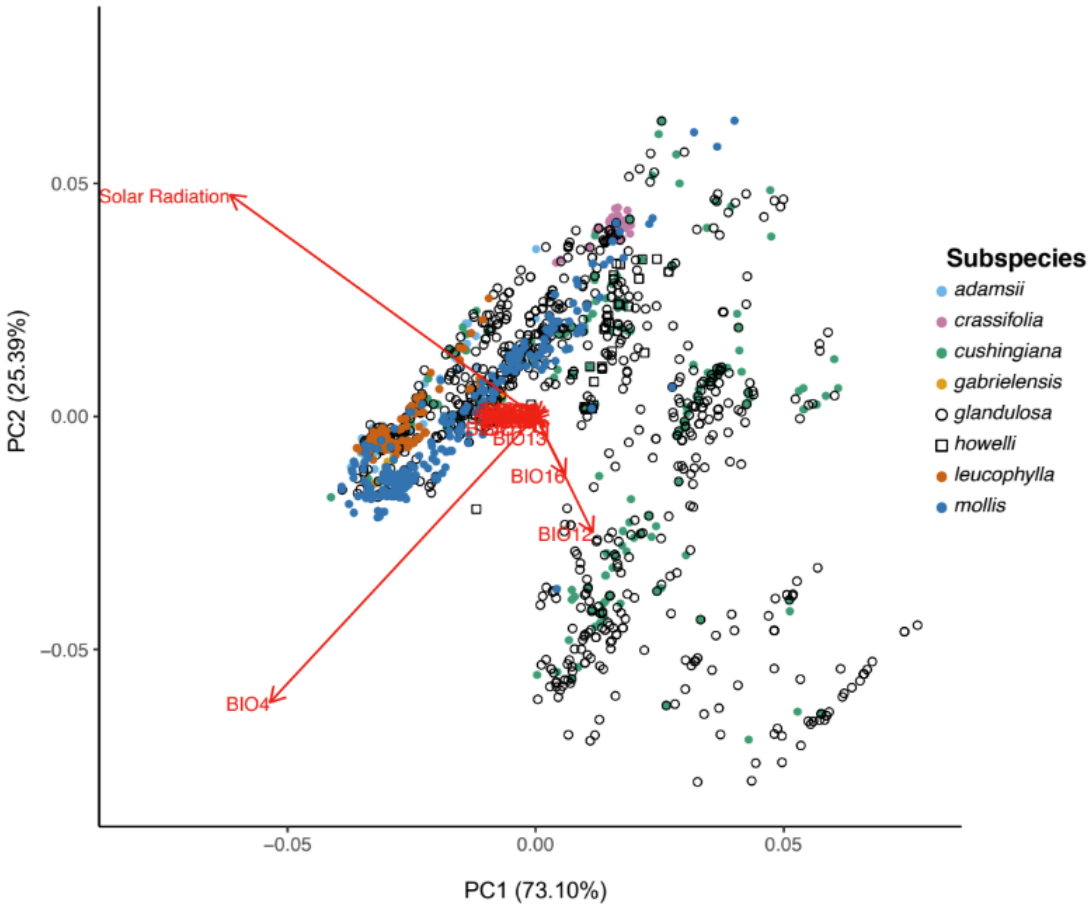
Appendix S4.26 Uncorrected *p* genetic distance plotted as a function of geographic distance (in kilometers) between every sample pair in the 2N data set. Points represent sample pairs. Points are plotted with 5% opacity to aid visual interpretation, as point density is high.



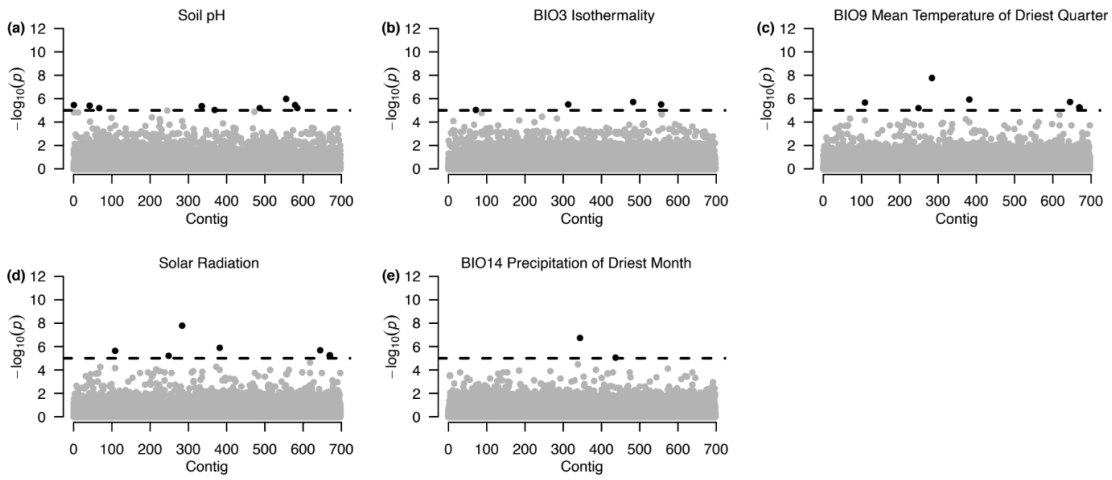
Appendix S4.27 Uncorrected *p* genetic distance plotted as a function of geographic distance (in kilometers) between every sample pair in the 2N-biallelic data set. Points represent sample pairs. Points are plotted with 5% opacity to aid visual interpretation, as point density is high.



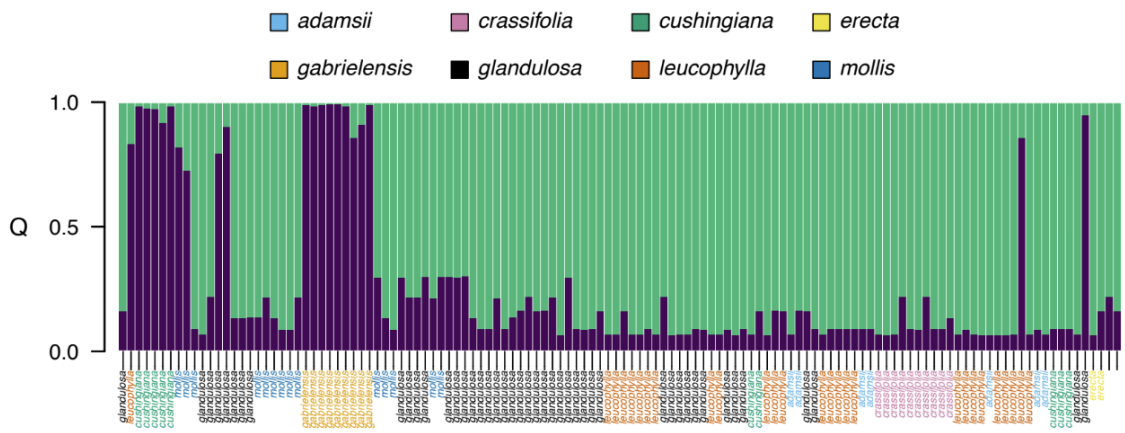
Appendix S4.28 PCA of the environmental data (19 Bioclimatic variables, soil pH and solar radiation). Points represent individual samples and are colored by subspecies identification. Arrows represent environmental variables, which are presented as the vectors. Four environmental variables contribute most to the separation of samples: solar radiation, BIO 4 Temperature Seasonality (standard deviation * 100), BIO 12 Annual Precipitation and BIO16 Precipitation of Wettest Quarter. The remaining 17 variables do not contribute significantly to variation across samples and appear superimposed as a red rectangle at the origin.



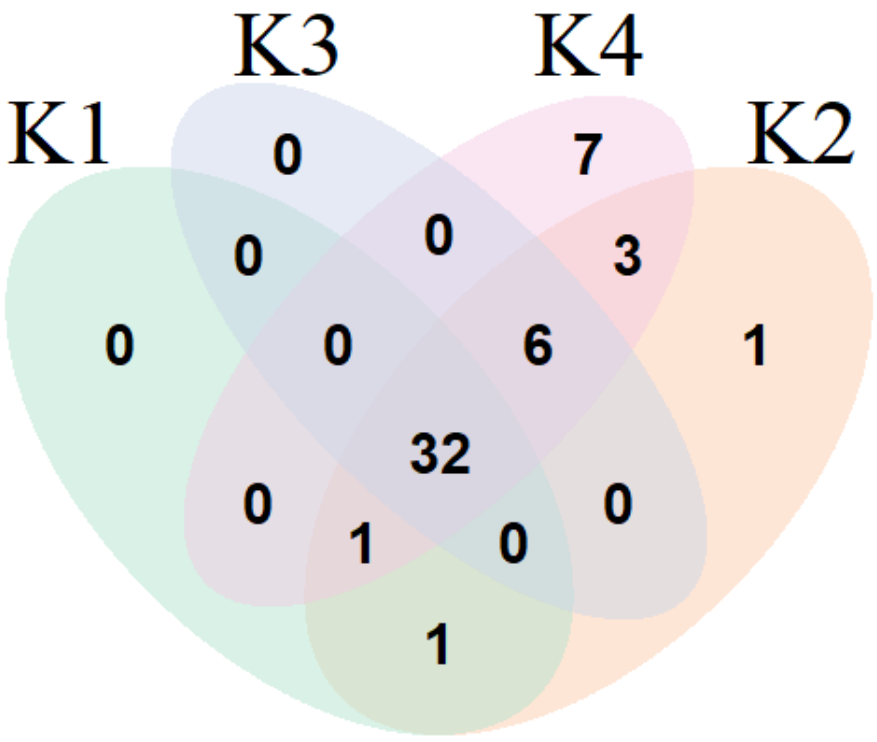
Appendix S4.29 SNPs associated with five environmental variables in latent factor mixed models (LFMM) in which the number of latent factors is set as two. Points represent SNPs and the dashed line represents the threshold for statistical significance at p = 1e-5 for the correlation of SNP and environmental variable. Solid points are significantly associated with that environmental variable.



Appendix S4.30 STRUCTURE results for *k* = 2, for the environment-associated SNP data set. Vertical bars represent individuals. Colors represent ancestral clusters of differing genotypes. The proportion of each color in each bar represents the probability of assignment (Q) to each ancestral cluster. Individuals are sorted along the x-axis by descending latitude of collection, and then by subspecies identification.

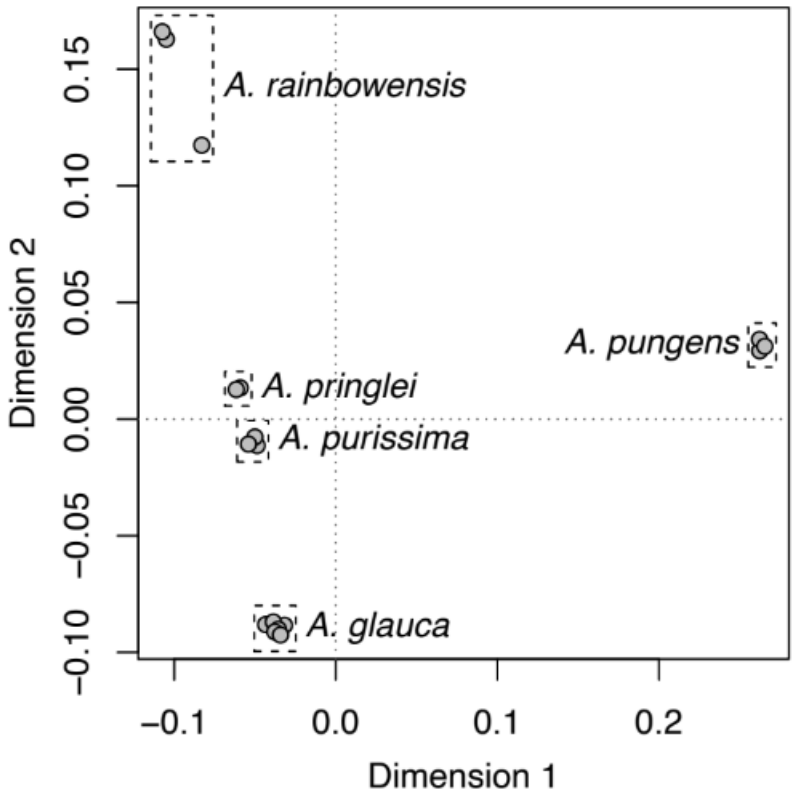


Appendix S4.31 Venn Diagram showing the overlap among contigs identified using K=1, 2, 3, and 4 in LFMM analysis.

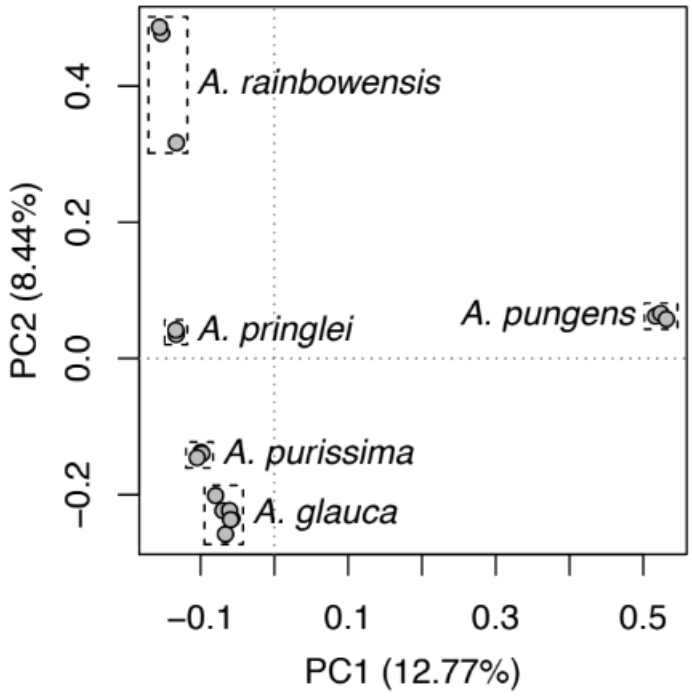


Appendix S4.32 Genes containing the environment-associated SNPs and their predicted functions based on BLAST search results.

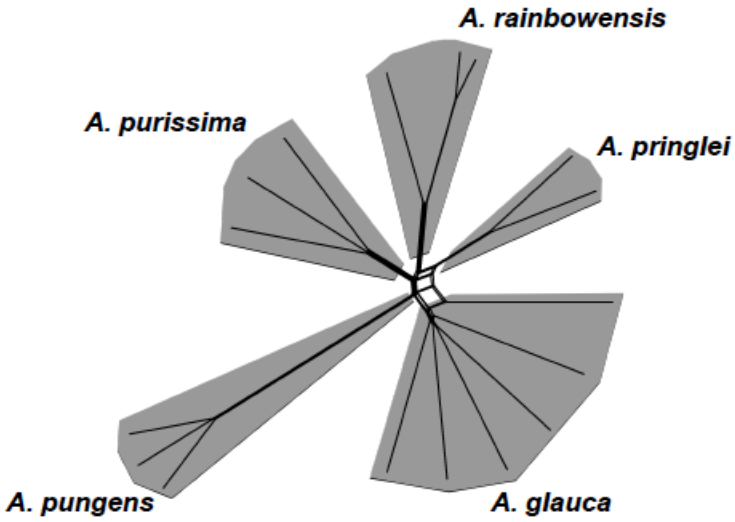
Appendix S4.33 MDS for the multispecies data set. Two dimensional representation of genetic distance among samples. Dashed boxes delimit samples from each species.



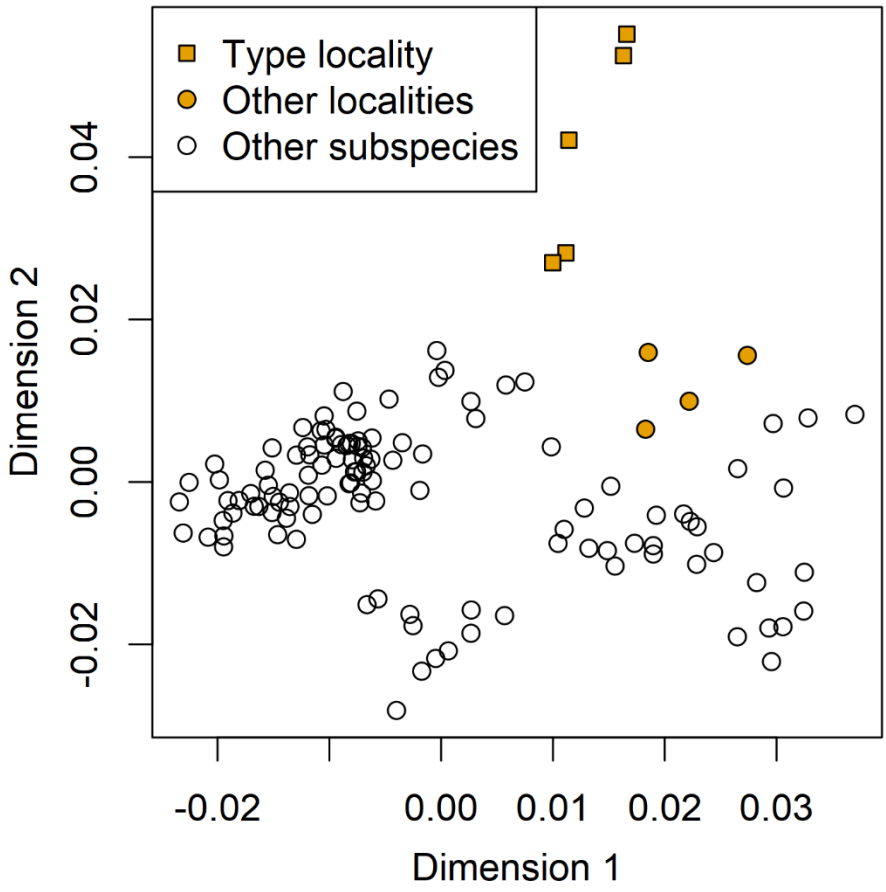
Appendix S4.34 PCA for the multispecies data set. Dashed boxes delimit samples from each species.



Appendix S4.35 NeighborNetwork for the multispecies data set. Shaded areas were drawn in to highlight the clustering of tips belonging to each species in the analysis.



Appendix S4.36 Two-dimensional MDS of the 4N SNP data set, with points representing samples of *A. glandulosa* subsp. *gabrielensis* marked with orange points. Square and round orange points represent samples from the type locality and from other localities, respectively. Hollow points represent samples not identified as *A. glandulosa* subsp. *gabrielensis*.



5 Conclusion

The diversity and essential ecological roles of manzanitas have long fascinated scientists in diverse areas, including taxonomy, ecology, evolution, and conservation. In the second chapter, I annotated the first genome assembly of manzanitas and compared it with other Ericales assemblies to gain insight into manzanita adaptation and evolution. Multiple evaluation statistics, visualization of genomic architecture across 13 pseudo chromosomes, and syntenic relationships between the manzanita genome and other chromosome-level assemblies of Ericaceae species support the conclusion that we have obtained a well-annotated chromosome-level assembly for the Big Berry Manzanita (*A. glauca*). Synteny analyses support multiple independent, family-specific whole-genome duplication (WGD) events for the Ericales, the WGD history of which has the subject of debate (Larson et al. 2020; Wang et al. 2021; One Thousand Plant Transcriptomes 2019). Comparative analyses between manzanita genome and other genomes of Ericales support my hypothesis and highlight the potential importance of terpenoids in the fire- and drought-adaptation of manzanitas in the California Floristic Province. This annotation will serve as an invaluable reference resource for ecological, evolutionary, and conservation studies of manzanitas, and facilitate the identification and classification of genetic variants across different populations, subspecies and/or species.

In the third chapter, I presented the first quantitative analysis of habitat diversification of manzanita species. In contrast to my hypothesis, my results did not show that any narrowly-distributed manzanita species are ecologically distinct when comparing all available species and weighting all available environmental variables equally. However, when I restricted the analyses to species from the same geographic

region and altered the machine learning algorithm to eliminate the effect of some invariant environmental factors, some species appeared ecologically distinct from other species in the same region. We found that the only species on Santa Catalina island, *A. catalinae*, appears to live in a different habitat, which differs in the temperature, precipitation, solar radiation, soil carbon and nitrogen stock, from other species from the Southern California-Baja California region. This finding highlights the importance of conserving the unique habitat of Santa Catalina Island to the preservation of *A. catalinae* and other flora endemic to the island. In addition, this study revealed 11 manzanita species that should potentially be considered critically endangered based on their small geographical range, but that have not been given any conservation status by state and federal agencies.

In the fourth chapter, we applied reduced-representation genomic sequencing technology to test the genetic distinctiveness of Eastwood Manzanita (*A. glandulosa*) subspecies, especially two that are state or federally listed. I found that most Eastwood manzanita subspecies are not differentiated by genetic data, except for one of the two rare subspecies, *A. glandulosa* subsp. *gabrielensis*. This finding highlights the importance of genotype preservation in the conservation of this subspecies. We did not find similar genetic distinctiveness for *A. glandulosa* ssp. *crassifolia*, thus our results do not support targeted conservation efforts for this federally endangered subspecies. Nevertheless, more data are needed to reach conclusions about the conservation status of this subspecies. In addition, our study suggests that genetic differentiation among manzanita taxa does not correspond to current species/subspecies delineation, but rather forms a gradient that follows a northwest-southeast geographic pattern. This study suggests that similar genetic studies in other *Arctostaphylos* species with currently

recognized subspecies, as well as other rare or endangered species and subspecies, will clarify our understanding of differentiation within this diverse woody genus.

5.1 References

Larson, Drew A., Joseph F. Walker, Oscar M. Vargas, and Stephen A. Smith. 2020. 'A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales', *Am. J. Bot.*, 107: 773-89.

One Thousand Plant Transcriptomes, Initiative. 2019. 'One thousand plant transcriptomes and the phylogenomics of green plants', *nature*, 574: 679-85.

Wang, Ya, Fei Chen, Yuanchun Ma, Taikui Zhang, Pengchuan Sun, Meifang Lan, Fang Li, and Wanping Fang. 2021. 'An ancient whole-genome duplication event and its contribution to flavor compounds in the tea plant (Camellia sinensis)', *Hortic Res*, 8: 176.