SYSTEMATICS OF *CHLOROPYRON* (OROBANCHACEAE): IMPLICATIONS OF MISSING DATA ON QUARTET-BASED SPECIES TREE METHODS

A Thesis

Presented in Partial Fulfilment of the Requirements for the

Degree of Master of Science

with a

Major in Biological Sciences

in the

College of Graduate Studies

University of Idaho

by

Ian Spencer Gilman

Major Professor: David C. Tank, Ph.D.

Committee Members: Paul Hohenlohe, Ph.D.; Jack Sullivan, Ph.D.

Department Administrator: James Nagler, Ph.D.

August 2017

Authorization to Submit Thesis

This thesis of Ian Spencer Gilman, submitted for the degree of Master of Science with a major in Biological Sciences and titled "Systematics of *Chloropyron* (Orobanchaceae): Implications of Missing Data on Quartet-based Species Tree Methods," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____Date_____
David C. Tank, Ph.D.

Committee Members: _____Date_____
Paul Hohenlohe, Ph.D.

_____Date_____
Jack Sullivan, Ph.D.

Department Administrator: _____Date_____
James Nagler, Ph.D.

Abstract

Sequence data exists for only 1/5 of plant species, therefore we risk of losing many branches of the tree of life before they are placed into an evolutionary context. This necessitates phylogeny estimation of understudied, rare, and threatened taxa, forcing researchers to utilize historical collections. Reduced representation sequencing approaches allow rapid generation of tens of thousands of loci and are increasingly being used in phylogenetic studies. However, these methods are primarily employed using specimen with low levels of nuclear DNA degradation. We resolve intraspecific relationships in *Chloropyron*, a genus of rare flowering plants, using historical collections, due to high rates of missing data. We characterize the behavior of two commonly used quartet-based species tree methods when rates of missing data are high to assess accuracy of species tree estimation using reduced representation libraries from historical collections. Finally, we elucidate sampling, sequencing, and species tree estimation schemes to better utilize historical samples for phylogenetics.

Acknowledgements

I would like to thank my major professor, Dr. David Tank, for mentoring me as a botanist, researcher, and teacher. Your encouragement and engagement have focused my enthusiasm for systematics and computational biology, while helping me navigate the academic world. I hope to emulate your dedication to your students, family, and research, and never stop riding my skateboard to work.

I would also like to thank my committee members: Dr. Paul Hohenlohe, for building my foundation in molecular evolution during my first few months in Moscow; Dr. Eric Roalson, for your interest in my work, and botanical and methodological insights; and Dr. Jack Sullivan for everything systematics-related, and, above all, teaching me to think critically, believe in my ideas, and occasionally drive to the basket. To Dr. James Foster, thank you for all of the philosophical and Pythonic conservations; they have inspired many ideas. To Dr. Chris Martine, I cannot thank you enough for encouraging me to first take a class with Dave and converting me from a physicist to a botanist. My years at Bucknell, especially as a part of the Martine Machine, will be with me forever.

The students of the Tank and Harmon labs have bolstered my confidence as scientist, and were always happy to take time to give me feedback on writing, identify a plant, or talk about managing life as a graduate student. From my first summer in McCall, Dr. Maribeth Latvis and PhD candidate Sarah Jacobs have been there for me at every step. Sebastian Mortimer, Daniel Caetano, and Dr. Diego Morales-Briones have made the early mornings and late nights in the office some of the most fun and intellectually stimulating. To Austin Anderson, Bryce Blankenship, Grahm Johnson, Johnathon Kaiser, Mason Linscott, Hannah Marx, Austin Patton, Megan Ruffley, Yannik Roell, Amanda Stahlke, Bob Week, and the rest of the students, masters, PhDs, and post-docs, I could not have done it without you. A big thank you to Nic Diaz and Dr. Tommy Stoughton for help in the field.

Finally, thank you to the University of Idaho Department of Biological Sciences, College

of Graduate Studies, Graduate and Professional Student Association, Stillinger Herbarium, Botanical Society of America, American Society of Plant Taxonomists, and National Science Foundation for funding that has made my career as a scientist possible.

Dedication

To my family

For their love, support,

and perpetual encouragement

to explore the world around me

Table of Contents

List of Tables

List of Figures

CHAPTER 1: SPECIES TREE ESTIMATION OF HISTORICAL SPECIMEN FROM DDRADSEQ DATA CONFIRMS MONOPHYLY OF HIGHLY DISJUNCT SPECIES IN *CHLOROPYRON* (OROBANCHACEAE)

1.1 Abstract

Sequence data exists for only about 1/5 of plant species, therefore we are at risk of losing many branches of the tree of life even before they are placed into an evolutionary context. This necessitates methods for phylogeny estimation of understudied, rare, and threatened taxa, which often forces researchers to utilize historical collections. The restriction site-associated DNA sequencing (RADseq) family of reduced representation sequence generation has provided a flexible and efficient method for the rapid generation of hundreds to tens of thousands of loci, and has recently seen adoption for phylogeny estimation. However, these methods have been primarily utilized with freshly collected or well preserved tissue. Here we sample all taxa of a rare genus of flowering plants, *Chloropyron*, from herbarium sheets dating back 25 years and use double digest restriction-site associated DNA sequencing (ddRADseq) to resolve intraspecific relationships. We find all species in *Chloropyron* to be monophyletic, with the inland taxon *C. maritimum* ssp. *canescens* sister to the rest of the coastal *C. maritimum* (ssp. *maritimum* + ssp. *palustre*), and the two distinct subspecies of *C. molle* to be to be monophyletic with strong support. In addition, we demonstrate the utility of reduced representation libraries to address phylogenomic problems in a group of rare species and address pitfalls of accurately inferring relationships when the amount of missing data is large, as is often the case when using historical specimen and rare taxa.

1.2 Introduction

Recent estimates indicate that, of the 500,000+ estimated species of plants (Soltis et al., 2010), about 1/3 are at risk of extinction, a rate 1,000–100,00 times higher than the background extinction rate (Pimm and Joppa, 2015). Currently, sequence data exists for about 116000 species of plants (iPlant Tree of Life, pods.iplantcollaborative.org), and so we are at risk of losing many branches of the plant tree of life even before they are placed into an evolutionary context. This necessitates methods for phylogeny estimation of understudied taxa with no reference genome, poor morphological and/or ecological data, and a lack high quality specimen DNA (particularly of concern for rare, endangered, or extirpated taxa). The restriction site-associated DNA sequencing (RADseq, Baird et al., 2008) family of reduced representation sequence generation has provided a flexible, cost effective, and time efficient approach for the rapid generation of hundreds to tens of thousands of loci (Andrews et al., 2016), which has recently seen adoption for phylogeny estimation (e.g., Cariou et al., 2013; Hipp et al., 2014). However, these methods have been primarily utilized with freshly collected or well preserved tissue. *Chloropyron* (Orobanchaceae) is a rare and ecologically important clade of flowering plants with many historical populations extirpated. This clade has proved to be challenging from a systematics perspective, and would benefit from modern library generation and phylogenetic methods.

The genus *Chloropyron* comprises four species (five subspecies, totaling seven taxa) of annual, hemiparasitic, halophytic herbs native to saline and saline-alkali flats and marshes of western North America (Chuang and Heckard, 1973, 1975a,b, 1986; Tank and Olmstead, 2008; Tank et al., 2009). All taxa are listed from threatened to critically imperiled at the state level in part or all of their range, and *C. maritimum* ssp. *maritimum* is federally endangered. Originally described by Behr (1855) as a distinct genus, *Chloropyron* had been treated primarily as a morphologically and ecologically distinct section (Gray, 1867; Ferris, 1918) or subgenus (e.g., Chuang and Heckard, 1973, 1986) of *Cordylanthus* Nutt. ex

Benth. (Orobanchaceae) until the first molecular phylogeny was erected, which restored *Chloropyron*, after *Cordylanthus* was shown to be paraphyletic (Tank and Olmstead, 2008). Tank and Olmstead (2008) note "the disintegration of *Cordylanthus*, as traditionally recognized, was one of the most surprising results of the molecular phylogenetic analyses." This, most recent, treatment was based on 2 nuclear-ribosomal gene regions (internal and external transcribed spacers; *ITS+ETS*) and 2 chloroplast loci (*rps16* and *trnL/F*). While a monophyletic *Chloropyron* was recovered with high support, interspecific relationships were not fully resolved, not all taxa were sampled, and sampled taxa were represented by a single accession. Furthermore, the nuclear and chloroplast phylogeny estimates showed cyto-nuclear discordance within *Chloropyron*. Plastid data supported the sister relationships *C. maritimum* and *C. tecopense*, and *C. molle* and *C. palmatum*, while nuclear data only supported the sister relationship of *C. maritimum* and *C. molle*.

Both inter- and intraspecific relationships are nontrivial within *Chloropyron* due to complex distributions and independent chromosome number changes (Chuang and Heckard, 1973; Tank et al., 2009). Most subspecies occur allopatrically, with highly disjunct populations, while three of the four species are sympatric or parapatric throughout central California (Figure 1.1). Although four functional stamens (as opposed to two throughout the rest of *Chloropyron*) and a gametic chromosome number of $n = 15$ unite all subspecies of *C. maritimum*, the evidence for a monophyletic *C. maritimum* is challenged by patterns of biogeography and ecology. *Chloropyron maritimum* ssp. *canescens* and *C. tecopense* are the only two taxa distributed east of the Sierra Nevada mountains, and *C. maritimum* ssp. *maritimum* and ssp. *palustre* are the only two strictly coastal taxa. All treatments prior to Chuang and Heckard (Ferris, 1918; Pennell, 1951; Mason, 1957; Munz, 1959), with the exception of Jepson (1925), treat *C. maritimum* as two species (*C. maritimum* = *C. maritimum* ssp. *maritimum* + ssp. *palustre*, the coastal subspecies, and *C. canescens* = *C. maritimum* ssp. *canescens*, the inland subspecies). Therefore, Tank and Olmstead's (2008) representation of *C. maritimum* by a single *C. maritimum* ssp. *canescens* specimen may not

be warranted.

Here, we sample all taxa of *Chloropyron*, with multiple accessions representative of their respective ranges, from historical records dating back 25 years and use double digest restriction-site associated DNA sequencing (ddRADseq, Peterson et al., 2012) to resolve intraspecific relationships. We find all species in *Chloropyron* to be monophyletic, with the inland taxon *C. maritimum* ssp. *canescens* sister to the rest of the coastal *C. maritimum* (ssp. *maritimum* + ssp. *palustre*), and the two distinct subspecies of *C. molle* to be to be monophyletic with strong support. We demonstrate the utility of reduced representation libraries to address phylogenomic problems in a group of rare species and address pitfalls of accurately inferring relationships when the amount of missing data is large, as is often the case when using historical specimen and rare taxa.

1.3 Materials and Methods

1.3.1 Sampling, library preparation, and sequencing

All accessions were sampled from herbarium vouchers dating back to 1983 (Table 1.1), and represent the ranges of taxa except *C. maritimum* ssp. *maritimum* (two individuals from one population) and *C. molle* ssp. *hispidum* (one individual). Genomic DNA was extracted using a modified CTAB protocol (Doyle and Doyle, 1987) for degraded DNA. These modifications included the addition of 2μl per sample proteinase-K to the CTAB solution, 1 hr hot (65°C) incubation with vigorous shaking (175 RPM) followed by 23 hr warm (50°C) incubation with moderate shaking (90 RPM) in CTAB solution, and a 24 hr cold (4°C) incubation in 2-propanol during the alcohol precipitation stage. Following a magnetic bead cleaning procedure to remove short fragments, DNA extractions were visualized on a 2% agarose gel, and quantified using a Qubit 2.0 Fluorometer (ThermoFisher, Carlsbad, CA).

Double digest restriction site-associated libraries were constructed using restriction enzymes *EcoRI* (a 4-base cutter: 5'–G|AATTC–3'; 3'–CTTAA|G–5') and *SbfI* (an 8-base

cutter: 5'–CCTGCA|GG–3'; 3'–GG|ACGTCC–5'). Because no close reference genome is currently available, genome size was coarsely estimated using 2C values of the closest relatives (*Castilleja miniata*, *C. rhexifolia*, and *C. sulphurea*, Hersch-Green and Cronn, 2009), after controlling for ploidy. Libraries were barcoded using a 6bp sequence (minimum distance of 2 between barcodes to reduce demultiplexing error), size selected at 650±50 bp using a PippinPrep (Sage Science, Inc., Beverly, MA), and multiplexed on an Illumina MiSeq at the University of Idaho's IBEST Genomic Resources Core Facility (Moscow, ID) with an expected 30x coverage of 300bp paired-end reads, including adaptor and barcode.

1.3.2 Locus identification

Raw sequence data was analyzed using the software *PyRAD* (Eaton, 2014) and PEAR (Zhang et al., 2014) following the protocol outlined to utilize paired-end ddRADseq data by merging paired-end reads (*PyRAD* manual v.3.0.4). Unless otherwise noted, the following procedures were conducted in *PyRAD*. Briefly, sequences were first demultiplexed by barcode. Next, sequences were input into PEAR, which merged paired-end reads if they overlapped by 10bp or more. Only those reads that were merged (assembled) were retained and input into *PyRAD* for subsequent steps. Assembled sequences were then quality filtered and barcodes, adaptors, and cut sites were removed. Trimmed sequences were then clustered by individual into "stacks" using VSEARCH (Rognes et al., 2016) and aligned via MUSCLE (Edgar, 2004) in *PyRAD*. Sequence error rate and mean heterozygosity were jointly estimated across all stacks in each individual. Error rates were analyzed by eye in FastQC v.0.11.5 (Babraham Bionformatics, Cambridge, United Kingdom). Consensus base calling and paralog filtering were performed before clustering loci across individuals. Finally, alignments for all loci across all samples were generated.

1.3.3 Phylogenetics

Sequence data generated by the RADseq family of methods pose a number of problems for traditional concatenation based maximum likelihood and Bayesian approaches because of coalescent stochasticity, the incongruence of evolutionary histories among loci (Roch and Steel, 2014; Chou et al., 2015), and we could not use commonly employed full likelihood or Bayesian programs such as Garli (Zwickl, 2006), RAxML (Stamatakis, 2006), or BEAST (Drummond et al., 2012) due to computational intractability.Therefore, we employed two classes of species tree methods based on the multi-species coalescent (MSC) to estimate phylogeny. SVDquartets (Chifman and Kubatko, 2014) utilizes site pattern probability distributions at sites with unlinked SNPs to assemble quartets of taxa into a species tree. This method assumes each site has its own genealogy drawn from the MSC and uses all data (all unlinked SNPs) directly. SVDquartets was called in PAUP* v.4.0a.152 (Swofford, 2002) with all possible quartets evaluated and 100 bootstrap replicates. In contrast, ASTRAL-II (Mirarab et al., 2014b; Mirarab and Warnow, 2015) is a summary based method that constructs all possible quartets from a set of unrooted input gene trees. The topology that satisfies the most quartets induced by the input gene trees is selected as the optimal species tree estimate.

To estimate gene trees, models of sequence evolution were evaluated for each locus recovered in *PyRAD* using 'automodel' in PAUP*, and when alternative models were selected by different goodness of fit criteria, the model with fewer parameters was chosen for downstream computational tractability. Gene trees were then constructed using GARLI (Zwickl, 2006). Although ASTRAL-II is consistent under the MSC, the input gene trees are subject to estimation error, which is exacerbated by the short length of RADseq loci and elevated rates of missing data present in reduced representation libraries of historical samples (see Chapter 2).

If phylogenetic signal is low in any one locus, the resulting gene tree may be poorly

estimated. Statistical binning, a graph-theoretical approach that evaluates whether two loci share the same gene tree, can greatly increase the accuracy of species tree reconstruction by leveraging the increased phylogenetic signal in concatenated supergenes (Bayzid and Warnow, 2013; Mirarab et al., 2014a). Concatenation of loci into supergenes was done following the methods outlined in (Mirarab et al., 2014a) and supergenes were used as input into ASTRAL-II. This dataset will hereafter be referred to as the 'supergene' dataset. Both ASTRAL-II datasets were called using the 'multiind' version of ASTRAL-II (v4.10.11), which is tailored to datasets with multiple individuals per taxon, with 100 bootstrap replicates. Finally, all analyses across all datasets were rooted using two species of *Cordylanthus* (Orobanchaceae), *C. capitatus* and *C. eremicus* ssp. *eremicus*.

1.4 Results

An average of 4456 ± 3906 loci were recovered per sample with an average coverage of 16.07 ± 15.45X. The number of loci per sample after quality filtering and removing paralogs and invariant loci was reduced to 96 ± 66. The final dataset comprised 592 loci (average size 422 ± 127 bp) with 1945 parsimony informative sites and 490 unlinked SNPs; this dataset will be referred to the 'all loci' dataset unless the context is clear. The number of loci recovered per taxon was significantly correlated with the number of accessions sampled per taxon (Figure 1.2). Both AICc and BIC selected the same model for 260 (43.9%) loci and 452 (76.4%) loci best fit the simplest (JC) or second simplest (F81) model. The resulting 592 gene trees from GARLI were binned with a threshold of 50% nodal support into 9 supergenes; 7 of 66 loci and 2 of 65 (average size 28311 ± 1351 bp).

All three phylogeny estimates yielded different topologies with inconsistent nodal support (Figure 1.3, 1.4, 1.5). All species containing subspecies were recovered as monophyletic with high support in both the 'all loci' and 'supergene' ASTRAL-II phylogenies. Support was slightly lower in the 'supergene' dataset, but the same relationships among the three subspecies of *C. maritimum* (inland *C. maritimum* ssp. *canescens* sister to coastal *C. mar-*

itimum ssp. *maritimum* + *C. maritimum* ssp. *palustre*) were found. However, interspecific relationships were not congruent between datasets and showed very low support in both analyses. The 'all loci' dataset found *C. molle* sister to sympatric/parapatric *C. palmatum*, and the Mojave Desert endemic *C. tecopense* sister to the rest of *Chloropyron*. The 'supergene' dataset recovered *C. palmatum* + *C. tecopense* sister to *C. maritimum* + *C. molle*. All interspecific nodal support values were less than 50, hence these topologies were not in conflict, but the *Chloropyron* backbone had no resolution.

While SVDquartets also recovered *C. maritimum* as monophyletic (with the same intraspecific relationships highly supported), no other species were found to be monophyletic and support values were lower than either ASTRAL-II topology. *C. molle* ssp. *hispidum* was found sister to *C. maritimum* and *C. molle* ssp. *molle* sister to the rest of *Chloropyron*.

The large contrast in support values for inter- and intraspecific relationships suggested that the majority of informative quartets was much higher within species, rather than between them. To test if these patterns of nodal support were being biased by the lack of informative quartets linking multiple species, we calculated the number of loci shared between at least i) N taxa, ii) N in-group taxa, iii) N species, and iv) N in-group species (Figure 1.6), as well as the number of informative quartets in those subsets (Table 1.2). These data show that the majority of loci (434) were shared between at least 2 different in-group species and only 83 loci (14%) were exclusive to a single taxon. However, very few (23) loci were shared by all in-group species, and only a single locus was shared between all taxa. These subsets have only 296 and 9 informative quartets (as diagnosed by SVDquartets), respectively. We repeated analyses in both ASTRAL-II and SVDquartets on these subsets of loci but found no significant changes in topology or nodal support until the number of informative quartets dropped below 5%, at which point very little resolution was possible anywhere in the tree (Figures 1.7, 1.8, 1.9, 1.10).

Among the 1492 informative quartets found in SVDquartets in the 592 loci concatenated dataset, only 73 unique quartet topologies were observed. While a small number of unique

quartets at high frequencies may indicate that loci were largely congruent, the distribution of quartets among taxa was skewed (Figures 1.11, 1.12). For some taxa, such as *C. maritimum* ssp. *maritimum*, a small number of quartets were supported by hundreds of sites distributed throughout all loci, whereas *C. molle* ssp. *molle* was recovered in just 10 unique quartets, each of which was supported by 10 or fewer sites throughout all loci (Figure 1.11). Furthermore, *C. molle* ssp. *hispidum* and ssp. *molle* were only supported in 3 unique, and 7 total, quartets, although neither the number of unique or total quartets were significantly correlated with the number of accessions per taxon (Supplementary Figures 1.13, 1.14).

ASTRAL-II first induces all possible quartets from all gene trees, but then expands the set of search quartets using heuristic strategies, and so many more quartets were evaluated, although most of these did not inform the species tree. The average number of total quartets induced by all gene trees in 100 bootstrap replicated was 47370 ± 2702. Each replicate species tree induced 2267 ± 295 quartets that were congruent with those induced by all gene trees ($41.7 \pm 1.3\%$ of all species tree induced quartets, $4.8 \pm 0.7\%$ of all gene tree induced quartets) in the 'all loci' dataset. At least one accession of *C. molle* ssp. *molle* was present in 88 loci and ssp. *hispidum* was present in 60 loci, which would induce hundreds to thousands of quartets containing at least one of these taxa from all gene trees. This may have contributed to more quartets informing species tree construction in ASTRAL-II, therefore generating higher support for the monophyly of *C. molle*. This hypothesis would also support the low nodal support values estimated by ASTRAL-II using a small number of supergenes. Binning loci into supergenes may have increased phylogenetic signal at the expense of reducing the number of quartets induced by supergene trees.

1.5 Discussion

The boundaries of the applications of reduced representation sequencing are continuing to expand as genome-scale data becomes easier and cheaper to generate. With this, a body of summary statistics, as well as a better understanding of model behavior in phylogenetic

analyses, are necessary to assess the utility of these data and accuracy of downstream phylogenetic inquiry, such as species tree estimation (Chapter 2). Due to the nature of species tree inference via quartet methods, all data need not completely overlap for hundreds of loci to inform any bipartition and accurately determine relationships between taxa. However, the patterns of missing data that influence the amount of overlap between loci, and the quartets they induce, can have dramatically different effects on the accuracy of species tree methods (reviewed in Eaton et al., 2016). If patterns of missing data are hierarchical, such as those that would result from mutation-disruption or mutation-generation (Rubin et al., 2012), the redundancy in loci/quartets may be significantly decreased, along with the accuracy of species tree estimation.

While the underlying cause for the disparities in number of loci and informative quartets in our dataset is unknown, it may be the result of uneven sampling of taxa. *Chloropyron maritimum* ssp. *canescens* is the most widely distributed member of *Chloropyron* and, despite the extirpation of many historical populations, is not considered strongly threatened. To cover the potential genetic variation in this taxon's distribution, sampling was biased towards *C. maritimum*, producing a hierarchical pattern of present data. In contrast, *C. molle* is known from only a handful of contemporary populations around the San Francisco Bay area, extending slightly into central California along saline estuaries and waterfowl preserves. By limiting sampling to the previous quarter century (to avoid significant degradation of genomic DNA), the number of accessions sampled for federally and state listed taxa were significantly decreased. Although we cannot reject the congruence of these topologies due to extremely low bootstrap support for interspecific relationships, it appears that uneven sampling has yielded data with little power to resolve these relationships because of hierarchical patterns of missing data.

The intraspecific relationships resolved with high confidence, that is *C. maritimum* ssp. *canescens* sister to *C. maritimum* ssp. *maritimum* + ssp. *palustre* confirms previous hypotheses about this highly disjunct species (Chuang and Heckard, 1973). The molecular work

presented here bolsters morphological and cytological evidence gathered over the previous half century. *Chloropyron maritimum* are distinguished morphologically by four functional stamens and entire to slightly-bifid floral bracts. All other members of *Chloropyron* have two functional stamens and floral bracts with one to five deeply cleft lateral lobes. In addition, all subspecies of *C. maritimum* share a chromosome number of 15, which varies throughout the rest of *Chloropyron* (*C. maritimum*, $n = 15$; *C. molle*, $n = 14$; *C. palmatum*, $n = 21$; *C. tecopense*, $n = 14$). These synapomorphies stand in stark contrast to the divergent ecology within *C. maritimum*. Both coastal subspecies inhabit alluvial soils along saline inlets with a diverse community of halophytes including *Limonium* (Plumbaginaceae), *Frankenia* (Frankeniaceae), *Salicornia* (Amaranthaceae), *Cuscuta* (Convolvulaceae), and *Distichlis* (Poaceae). The coastal *C. maritimum* have been found parasitizing multiple of these community members (ISG personal observation) via root haustoria, and host-specificity is rare throughout Orobanchaceae. *C. maritimum* ssp. *canescens* has only been found to parasitize *Distichlis spicata* (Poaceae), and occur exclusively with *D. spicata* in dry, alkali flats throughout the Great Basin, although *Atriplex* (Chenopodiaceae), *Artemisia* (Asteraceae), *Ericameria* (Asteraceae), and other grass species are sometimes present. The disparate ecologies of the coastal and inland taxa do not follow major morphological splits between subspecies. Chaung and Heckard (1973) note that there is no morphological feature that clearly delineates *C. maritimum* ssp. *maritimum* and ssp. *canescens*, but ssp. *palustre* is easily demarcated by its deep pink-purple flowers (ssp. *canescens* and ssp. *maritimum* have white flowers with yellow apices).

The relationships among subspecies of *C. molle* are less clear due to incongruence of species trees estimated by ASTRAL-II and SVDquartets. The number of quartets constructed in SVDquartets, and therefore the power to estimate the species tree, is low. ASTRAL-II does not output the quartets used in species tree construction, and so, although many more quartets informed species tree construction, their distribution among taxa is unknown. We show with simulation studies in Chapter 2 that, at high levels of missing data,

ASTRAL-II is significantly more accurate than SVDquartets. In addition, while binning loci into supergenes can increase phylogenetic signal (Bayzid and Warnow 2013, Mirarab *et al.* 2014), it may reduce the number of quartets induced by supergene trees. The monophyly of *C. molle* in both ASTRAL-II topologies tentatively bolsters previous hypotheses based on morphology, ecology, and cytology.

Chloropyron molle is differentiated from the rest of Chloropyron not by a single feature, such as four functional stamens in *C. maritimum*, but by a suite of traits. *C. molle* has light pink flowers with yellow apices, and ranges from ecologies matching that of the coastal *C. maritimum* subspecies to drier, inland habitats with *Allenrolfea* (Amaranthaceae) and *D. spicata* in central California. The range of *C. molle* overlaps slightly with *C. maritimum* and *C. palmatum*, but there is no evidence of hybridization. *C. molle* is mainly distinguished from *C. maritimum* by the length and density of hair throughout the plants. Subspecies of *C. molle* are primarily differentiated along ecological and geographical lines: ssp. *molle* is endemic to alluvial salt marsh habitats similar to those of the coastal *C. maritimum* subspecies, whereas ssp. *hispidum* occurs in dry alkali flats.

Finally, the inconsistency of the placement of *C. palmatum* and *C. tecopense* within *Chloropyron* adds to a set of complexities of ecologies, morphologies, and cytotypes. *C. palmatum* is morphologically and ecologically very similar to *C. molle*, which have been found within meters of one another (ISG personal observation), but have the largest gap in gametic chromosome number. *Chloropyron tecopense* is the most morphologically distinct taxon due to its small, linear leaves and the overall oppression of its pubescent leaves, bracts, and branches. This taxon is endemic to the alkali flats of Death Valley and the Mojave Desert, similar in habitat to *C. maritimum* spp. *canescens*, but drier, hotter, and with less vegetation. *Chloropyron tecopense* and *C. molle* also share a gametic chromosome number of $n = 14$. The interspecific relationships among these taxa remain blurred by conflicting lines of evidence, but genetic data may still prove to be a useful tool with increased, and equal, sequencing effort. Eaton *et al.* (2016) showed a 10-fold increase in shared loci when doubling

sequence effort. This may or may not be possible for historical collections with limited tissue for DNA extraction, and equal sequencing effort will come at the cost of resampling many of the same populations for narrowly restricted taxa. Future work including specimen collected in the field will hopefully lead to a more robust dataset with more loci spanning more taxa that will allow the creation of a highly supported phylogeny of this small group of extremophiles.

1.6 Conclusions

It is clear that species boundaries in *Chloropyron* cannot be delimited by any one form of evidence gathered thus far. There have been repeated transitions, both within and between species, between coastal salt marshes and the dry, alkali flats of central California, the Mojave Desert, and the Great Basin. There have also been independent gametic chromosome number changes and few individual morphological characters that delineate taxa. When the quality of sample DNA is unknown prior to extraction, such as from tissue on herbarium sheets, even sampling of fewer individuals with deeper sequencing coverage may produce better results for species tree estimation. Finally, as the number of loci and amount of missing data increase, parsing the degree to which data overlap and the informativeness of those data, are paramount to assessing the power of species tree methods.

Table 1.1: Accession numbers for all specimen used in study and herbarium sampled from.

Taxon	Accession No.	Year Collected	Herbarium
Chloropyron maritimum ssp. *canescens*	School craft 2112	1990	UC
Chloropyron maritimum ssp. *canescens*	Wilson 6288	1993	UC
Chloropyron maritimum ssp. *canescens*	Tiehm 12138	1995	UNH
Chloropyron maritimum ssp. *canescens*	Pins 12491	1997	UC
Chloropyron maritimum ssp. *canescens*	Tiehm 12253	1997	RSA
Chloropyron maritimum ssp. *canescens*	Tiehm 12643	1998	UC
Chloropyron maritimum ssp. *canescens*	Tiehm 13336a	2000	RSA
Chloropyron maritimum ssp. *canescens*	Tiehm 13336b	2000	ID
Chloropyron maritimum ssp. *canescens*	Tiehm 14063a	2002	UC
Chloropyron maritimum ssp. *canescens*	Tiehm 14063b	2002	RSA
Chloropyron maritimum ssp. *canescens*	Riefner 04-468	2004	RSA
Chloropyron maritimum ssp. *canescens*	La Doux 115	2005	RSA
Chloropyron maritimum ssp. *canescens*	Andre 10285	2006	RSA
Chloropyron maritimum ssp. *canescens*	Andre 20334	2011	RSA
Chloropyron maritimum ssp. *canescens*	Fraga 4143	2012	RSA
Chloropyron maritimum ssp. *maritimum*	Fraga 4143a	2012	RSA
Chloropyron maritimum ssp. *maritimum*	Fraga 4143b	2012	RSA
Chloropyron maritimum ssp. *palustre*	Chuang 7808	1990	JEPS
Chloropyron maritimum ssp. *palustre*	Wetherwax 2462	1993	JEPS
Chloropyron maritimum ssp. *palustre*	Ruygt & Collins	1993	JEPS
Chloropyron maritimum ssp. *palustre*	Wetherwax 2460	1993	JEPS
Chloropyron maritimum ssp. *palustre*	Slackly 57	2013	JEPS
Chloropyron molle ssp. *hispidum*	Heckard 6740	1990	JEPS
Chloropyron molle ssp. *molle*	Ruygt 1	1993	JEPS
Chloropyron molle ssp. *molle*	Ruygt 2	1994	JEPS
Chloropyron molle ssp. *molle*	Ruygt 3	1994	JEPS
Chloropyron palmatum	Heckard 6145	1983	JEPS
Chloropyron palmatum	Cypher 2004-018	2004	RSA
Chloropyron palmatum	Cypher 2005-022	2005	RSA
Chloropyron palmatum	Cypher 2005-023	2005	RSA
Chloropyron tecopense	Andre 9826	2007	RSA
Chloropyron tecopense	Fraga 3892	2011	RSA
Chloropyron tecopense	Fraga 3870	2011	RSA
Chloropyron tecopense	Andre 9701	2008	RSA
Chloropyron tecopense	Fraga 3769	2010	RSA
Cordylanthus capitatus	Ertter 20380	2010	UC
Cordylanthus eremicus ssp. *eremicus*	Fraga 933	2003	RSA

Table 1.2: Number of loci, characters, and informative quartets in data subsets input into SVDquartets

	Loci	Characters	Informative quartets
Complete dataset	592	250594	1492 (7.98%)
In-group species			
2+	434	183588	1492 (7.98%)
3+	199	84650	1441 (7.71%)
4	23	8142	296 (1.58%)
All species			
2+	460	195347	1492 (7.98%)
3+	273	181803	1486 (7.95%)
4+	99	43463	1128 (6.03%)
5+	24	10072	470 (2.51%)
6	4	1616	20 (0.11%)
In-group taxa			
2+	509	213909	1492 (7.98%)
3+	364	149311	1492 (7.98%)
4+	153	59739	1474 (7.88%)
5+	39	18529	914 (4.89%)
6+	17	6396	274 (1.47%)
7	3	1310	13 (0.07%)
All taxa			
2+	519	218671	1492 (7.98%)
3+	405	168557	1492 (7.98%)
4+	231	94927	1492 (7.98%)
5+	82	33569	1397 (7.47%)
6+	40	16279	959 (5.13%)
7+	18	7623	196 (1.05%)
8+	3	1313	9 (0.05%)
9	1	458	9 (0.05%)

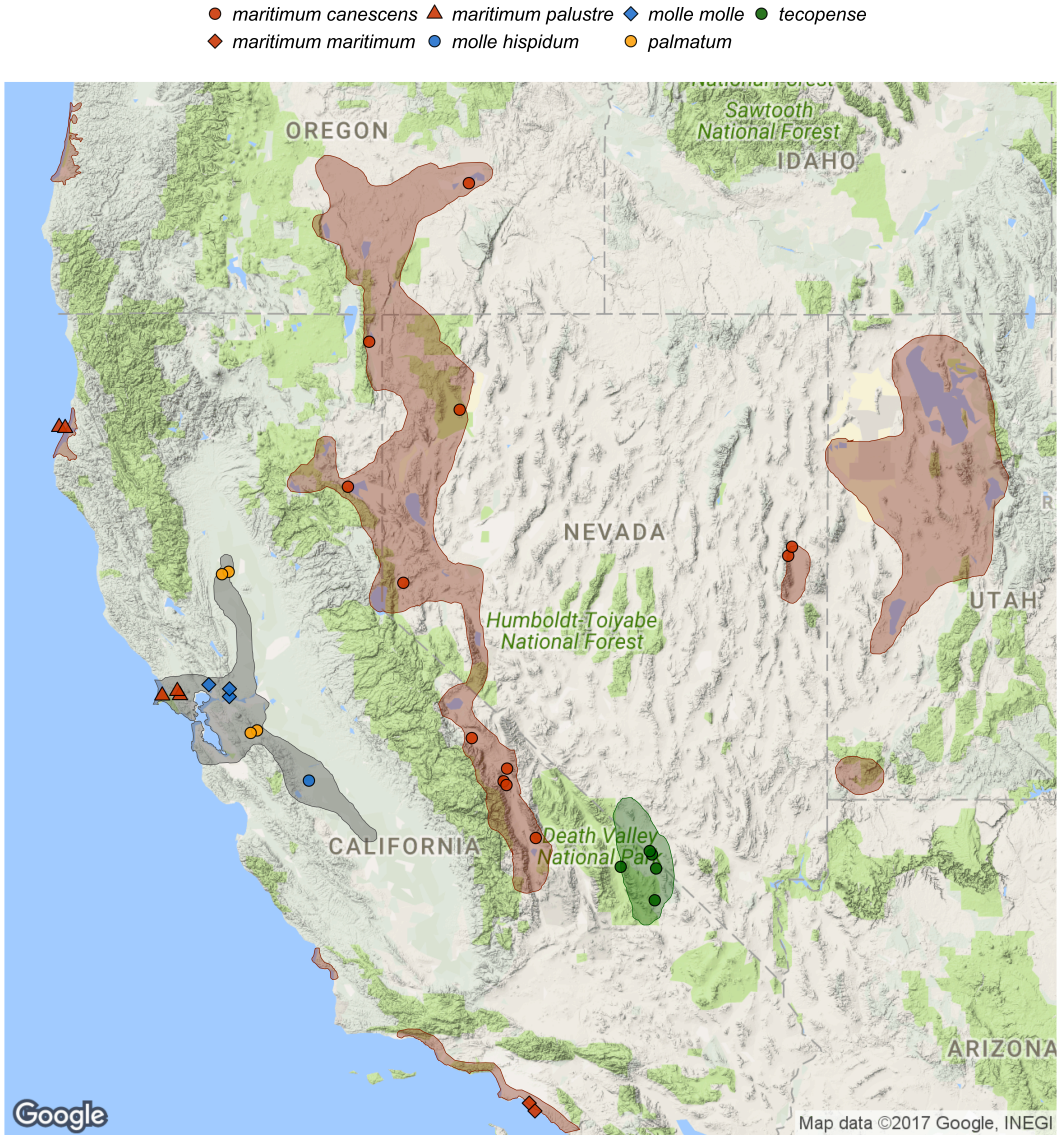

Figure 1.1: Distribution and sampling of *Chloropyron*. Color of distribution corresponds to respective species points except where taxa occur sym- or parapatrically (grey).

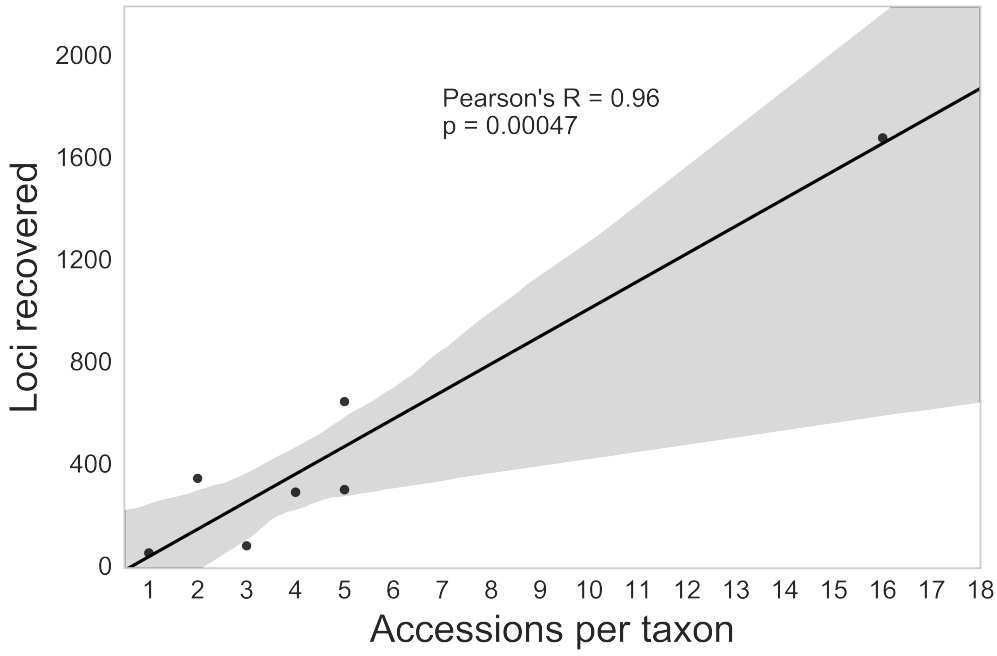

Figure 1.2: Loci recovered per taxon as a function of number of accessions sampled per taxon. Grey shading indicates 95% confidence interval of linear regression.

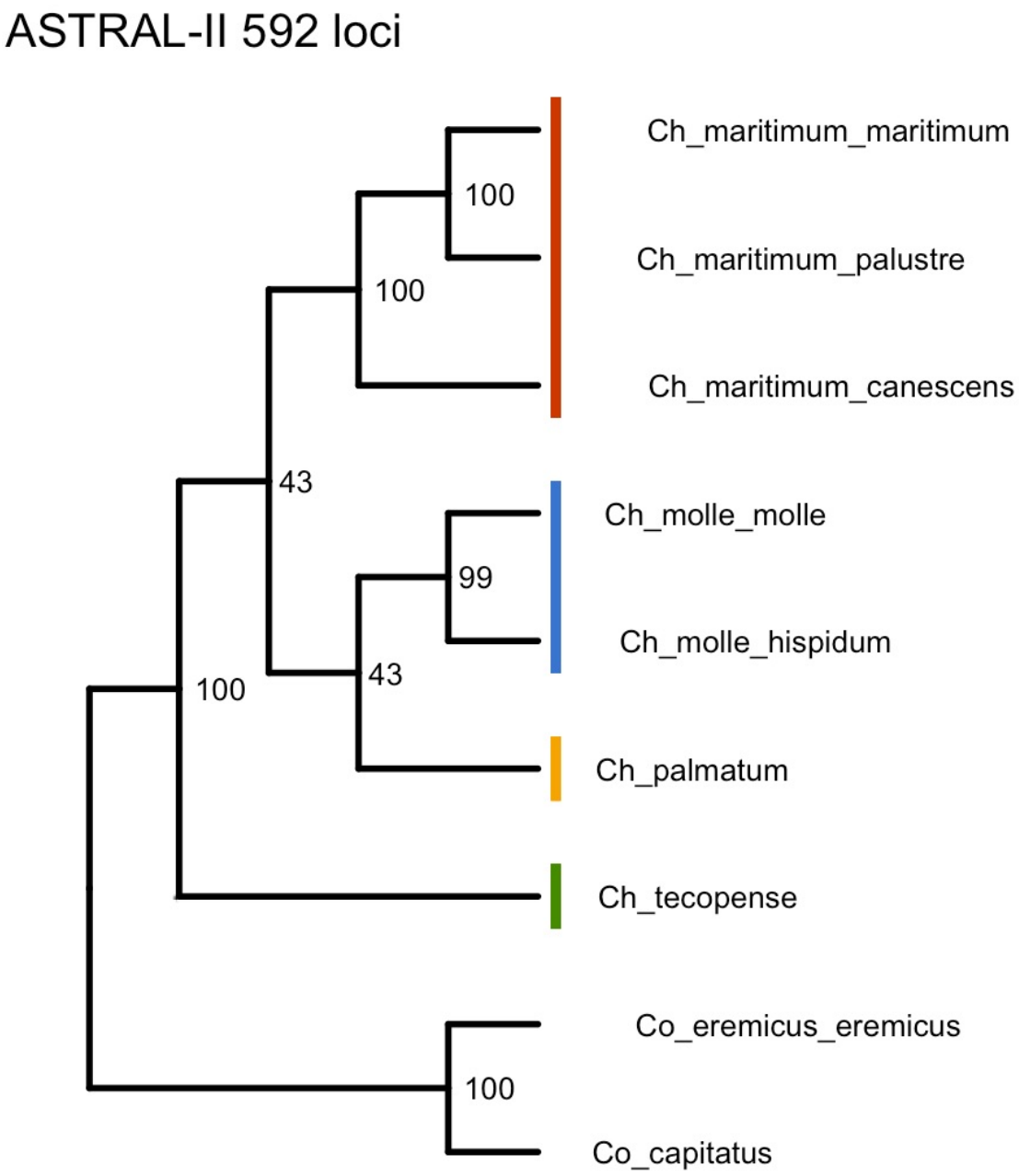

Figure 1.3: Species tree of *Chloropyron* estimated using 592 loci in ASTRAL-II. Node annotations show bootstrap support.

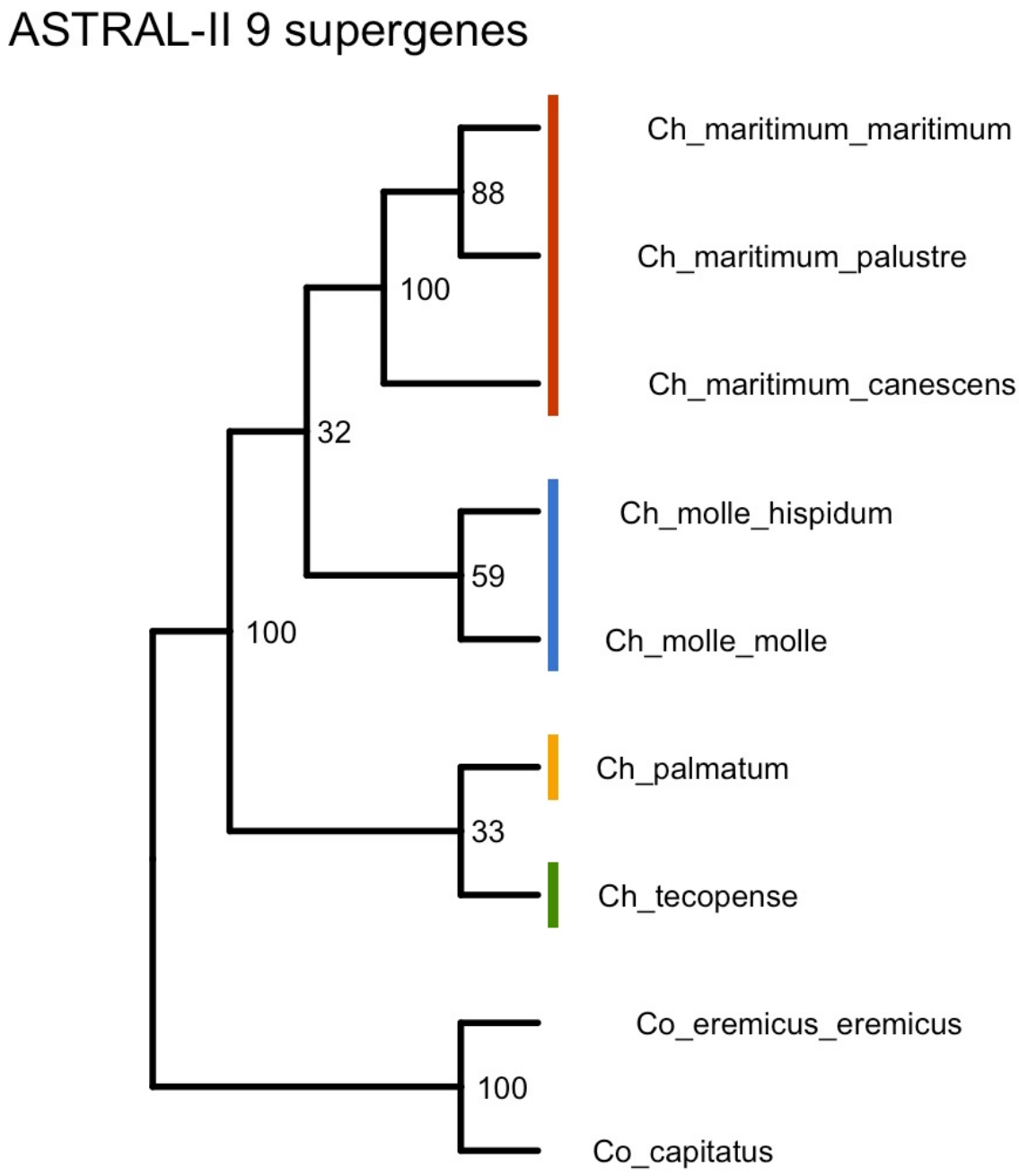

Figure 1.4: Species tree of *Chloropyron* estimated using 9 supergenes in ASTRAL-II. Node annotations show bootstrap support.

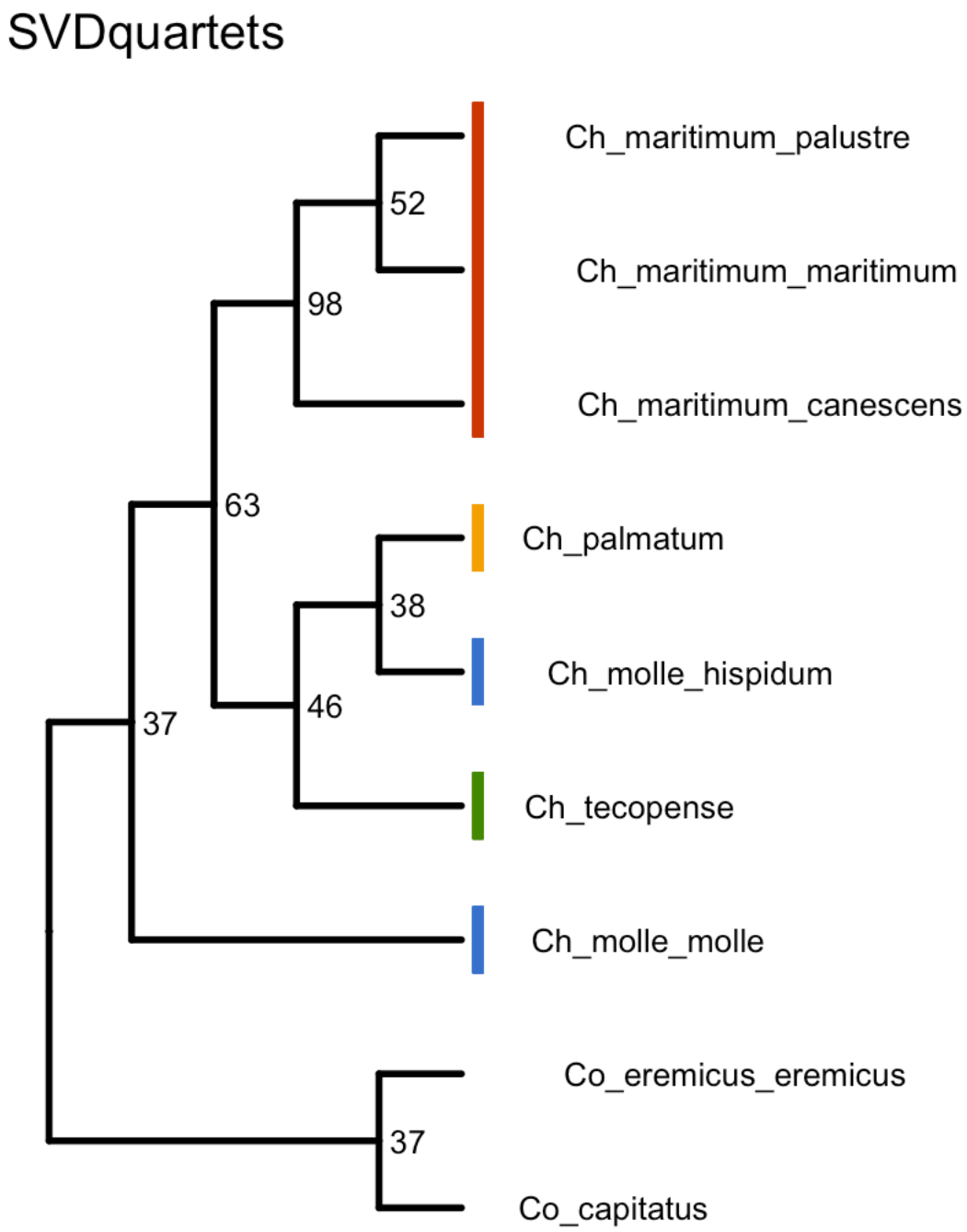

Figure 1.5: Species tree of *Chloropyron* estimated using 1492 quartets in SVDquartets. Node annotations show bootstrap support.

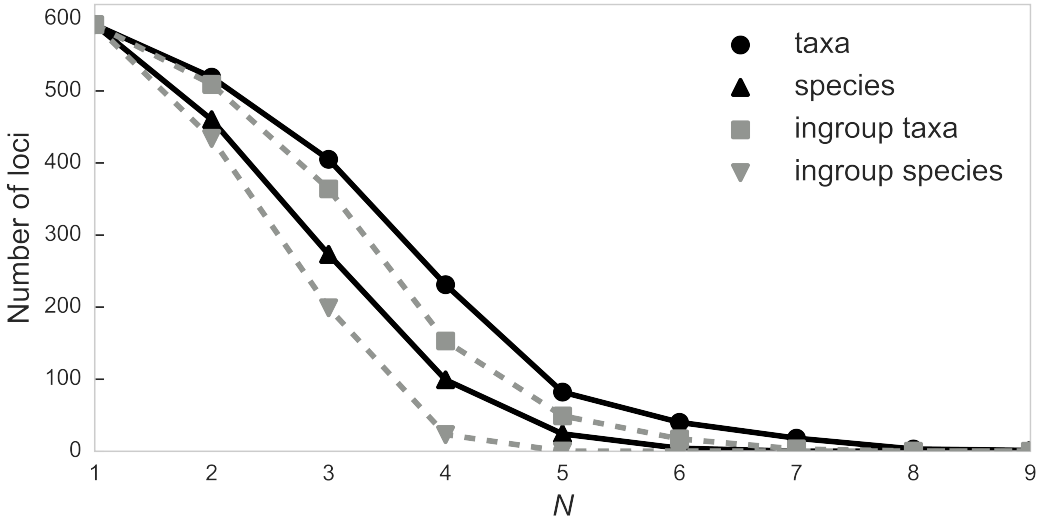

Figure 1.6: The number of loci shared among $N+$ taxa (solid, black circles), species (solid, black triangles), in-group taxa (dashed, grey squares), and in-group species (dashed, grey triangles).

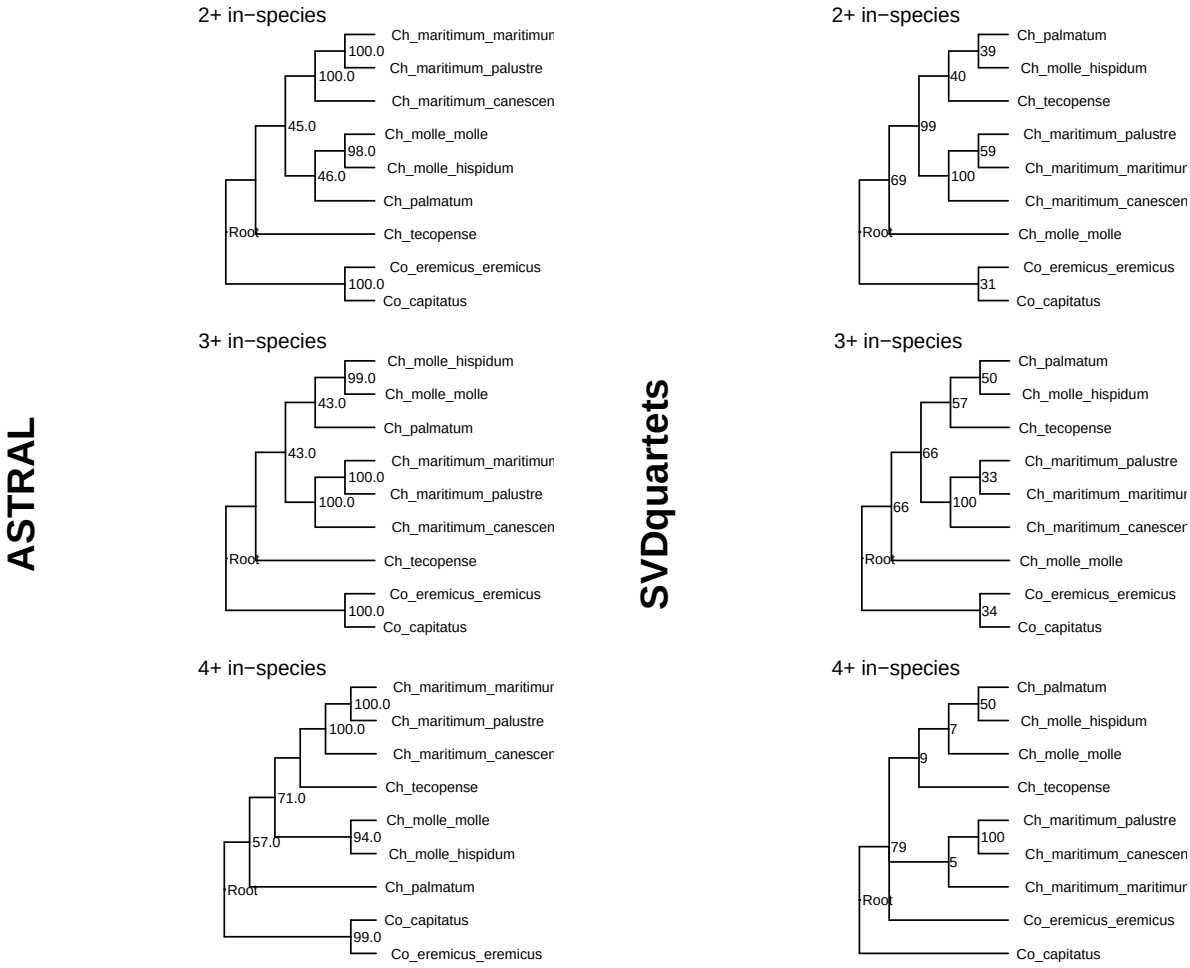

Figure 1.7: Comparison of *Chloropyron* species tree estimates from loci containing $N+$ in-group species in ASTRAL-II (left) and SVDquartets (right). Trees have been rooted with *Cordylanthus capitatus* when the outgroup *C. capitatus* + *C. eremicus* ssp. *eremicus* was not recovered as monophyletic.

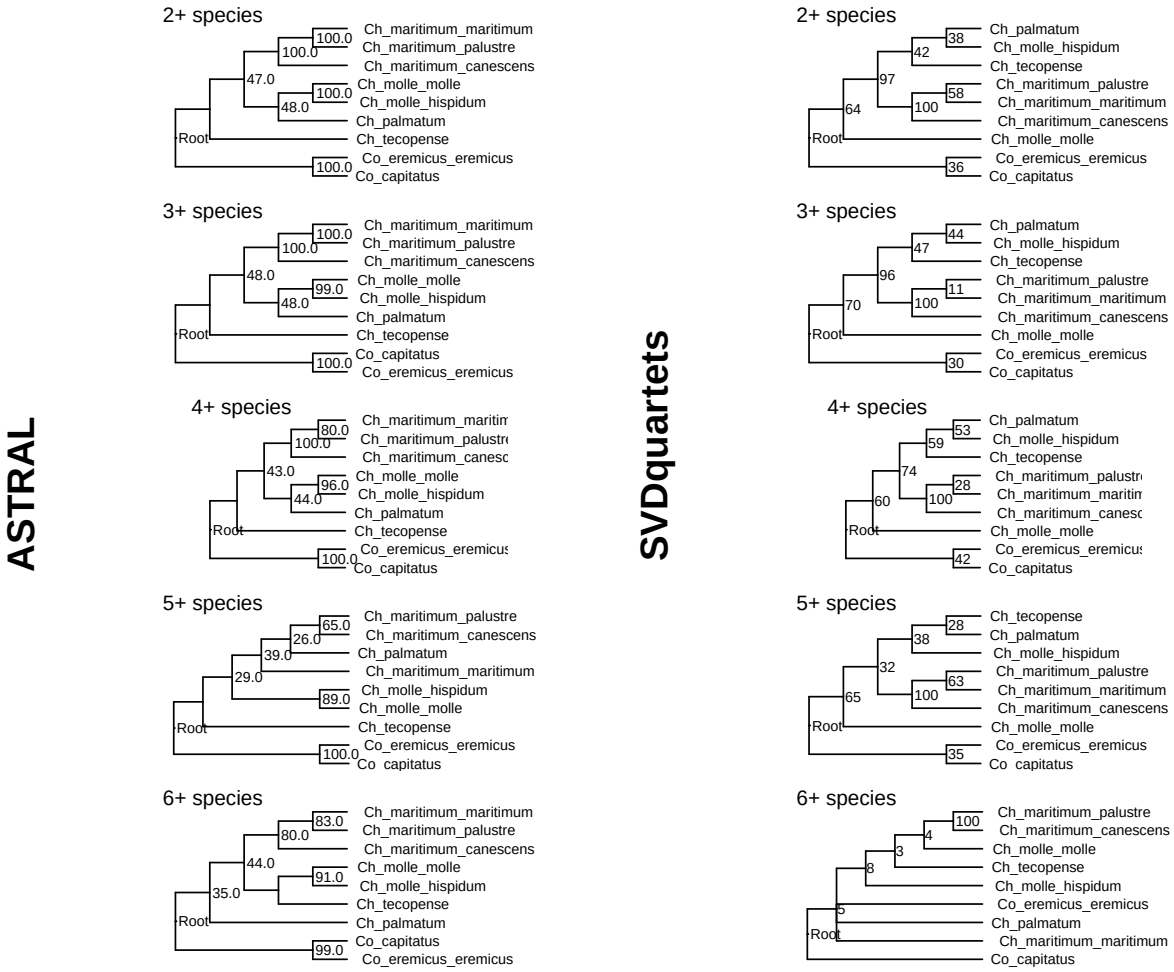

Figure 1.8: Comparison of *Chloropyron* species tree estimates from loci containing $N+$ total species in ASTRAL-II (left) and SVDquartets (right). Trees have been rooted with *Cordylanthus capitatus* when the outgroup *C. capitatus* + *C. eremicus* ssp. *eremicus* was not recovered as monophyletic.

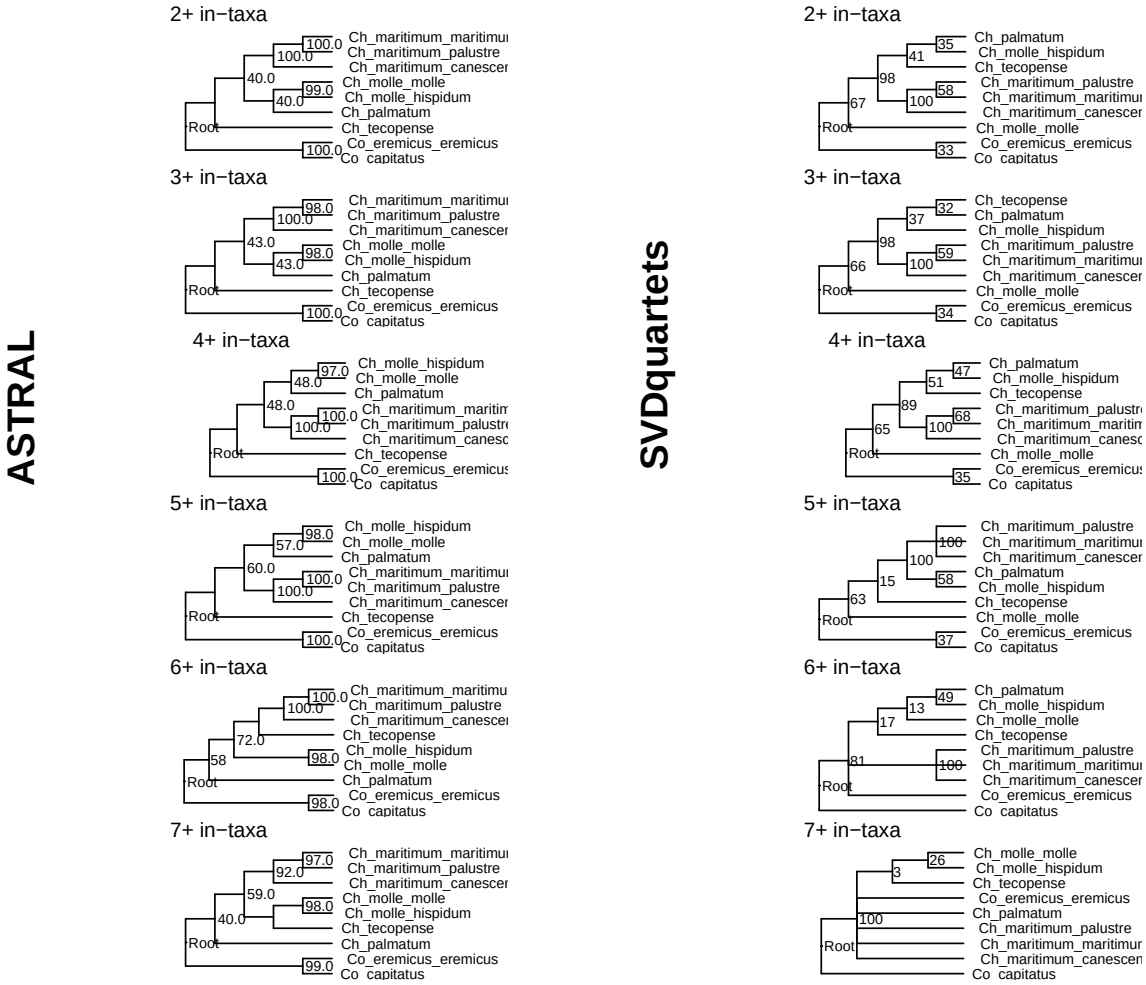

Figure 1.9: Comparison of *Chloropyron* species tree estimates from loci containing $N+$ in-group taxa in ASTRAL-II (left) and SVDquartets (right). Trees have been rooted with *Cordylanthus capitatus* when the outgroup *C. capitatus* + *C. eremicus* ssp. *eremicus* was not recovered as monophyletic.

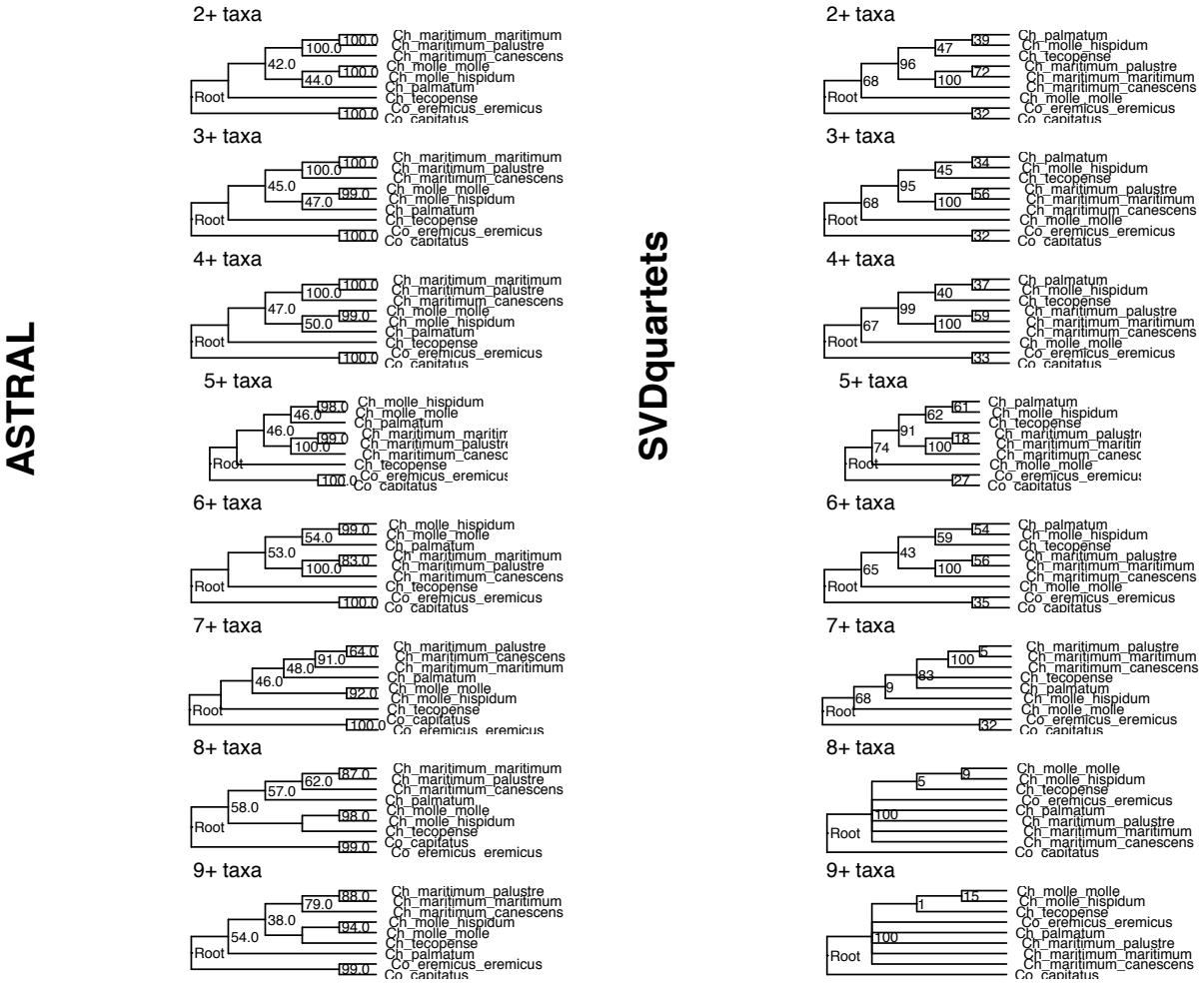

Figure 1.10: Comparison of *Chloropyron* species tree estimates from loci containing $N+$ total taxa in ASTRAL-II (left) and SVDquartets (right). Trees have been rooted with *Cordylanthus capitatus* when the outgroup *C. capitatus* + *C. eremicus* ssp. *eremicus* was not recovered as monophyletic.

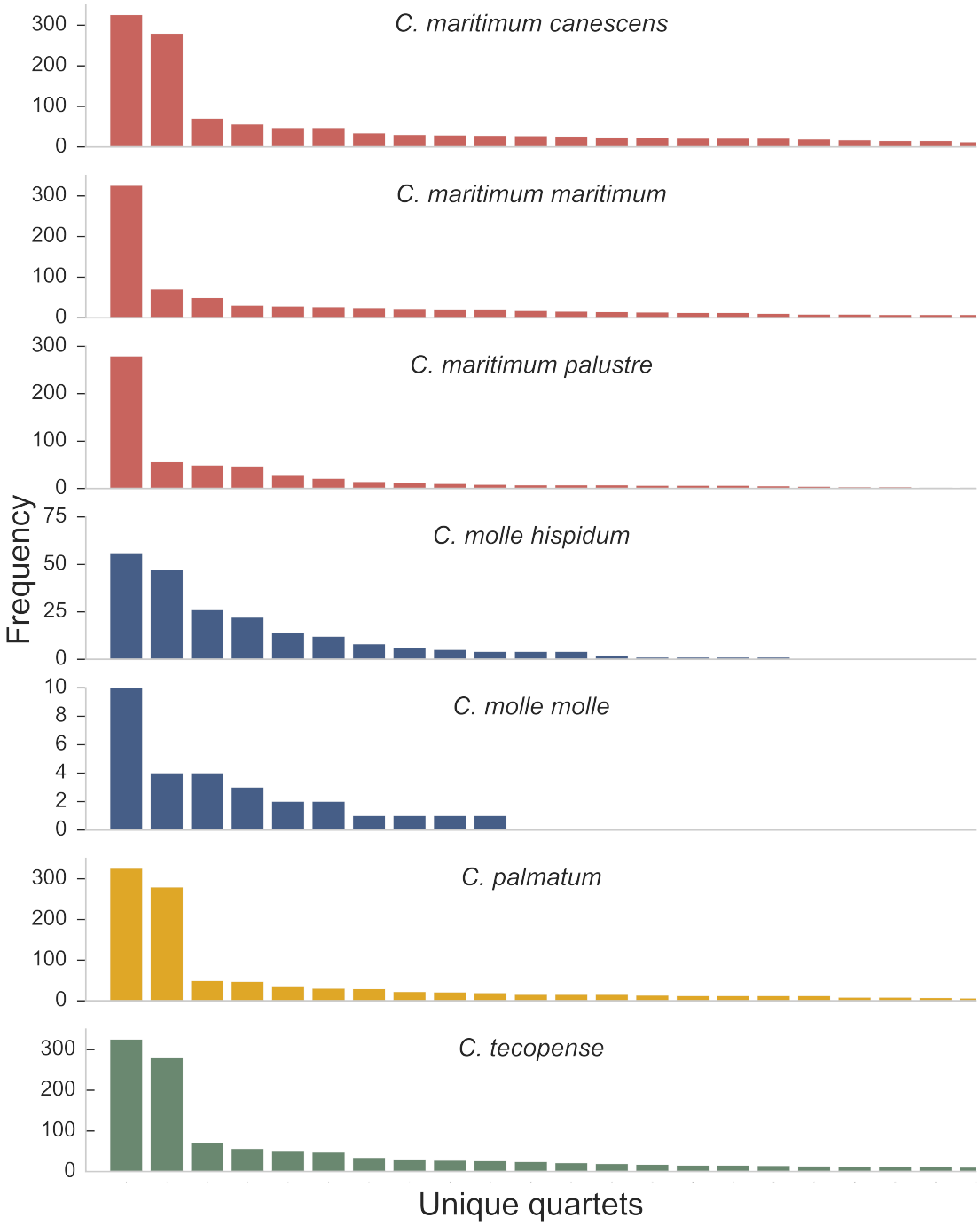

Figure 1.11: Histograms of the 20 most common quartets found in SVDquartets in all in-group taxa. Note scale change in panels *C. molle* ssp. *hispidum* and ssp. *molle*. Bar color corresponds to species distributions in Figure 1.1.

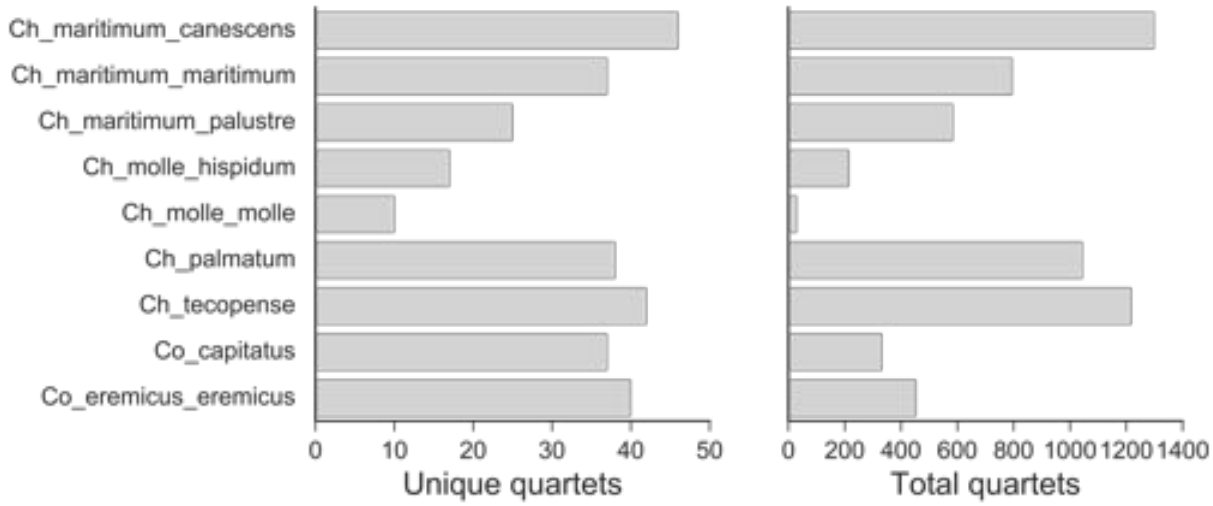

Figure 1.12: Number of unique (left) and total (right) quartets containing each taxon recovered in SVDquartets. A total of 73 unique and 1492 total quartets were recovered.

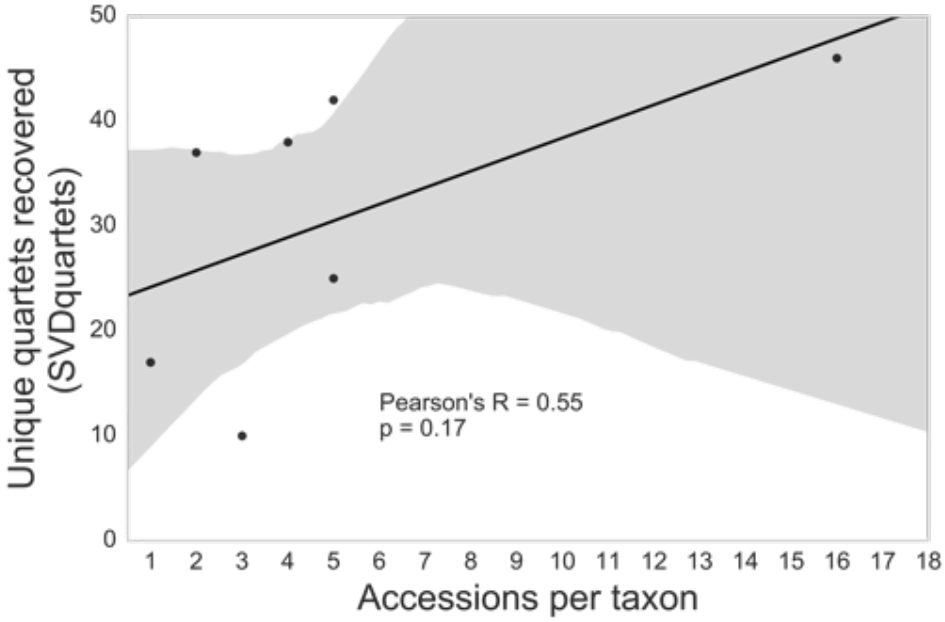

Figure 1.13: Number of unique quartets recovered by SVDquartets per taxon. Grey shading indicates 95% confidence interval of linear regression.

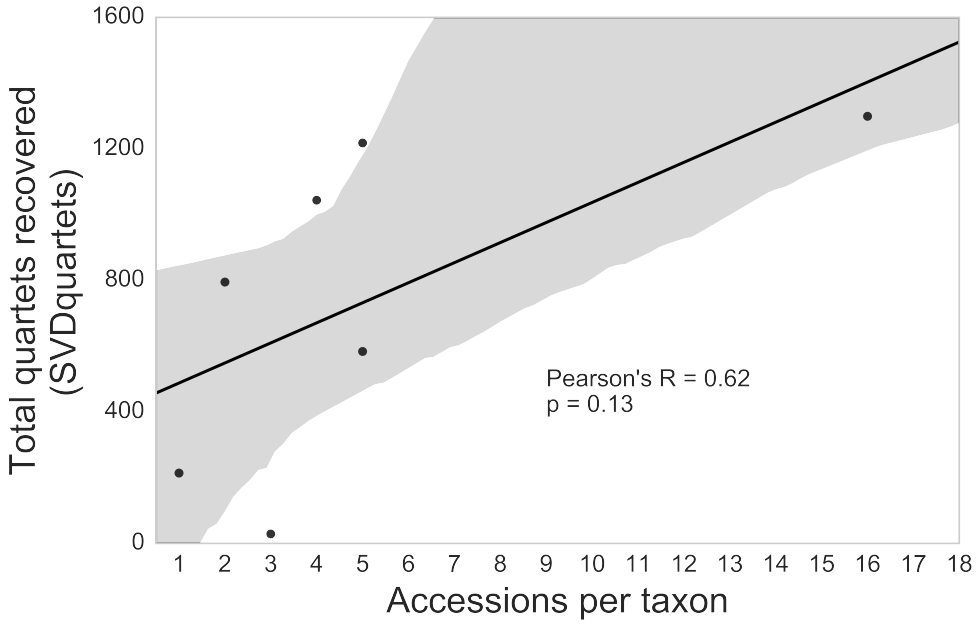

Figure 1.14: Number of total quartets recovered by SVDquartets per taxon. Grey shading indicates 95% confidence interval of linear regression.

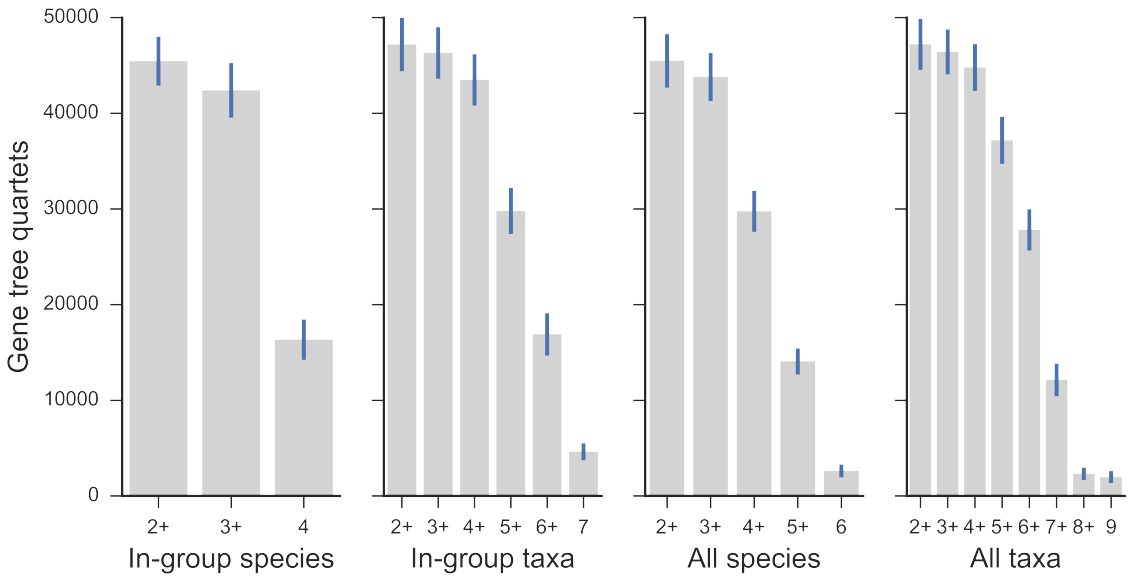

Figure 1.15: Mean number of quartets induced by subsets of input gene trees in ASTRAL-II. Blue bars indicate one standard deviation.

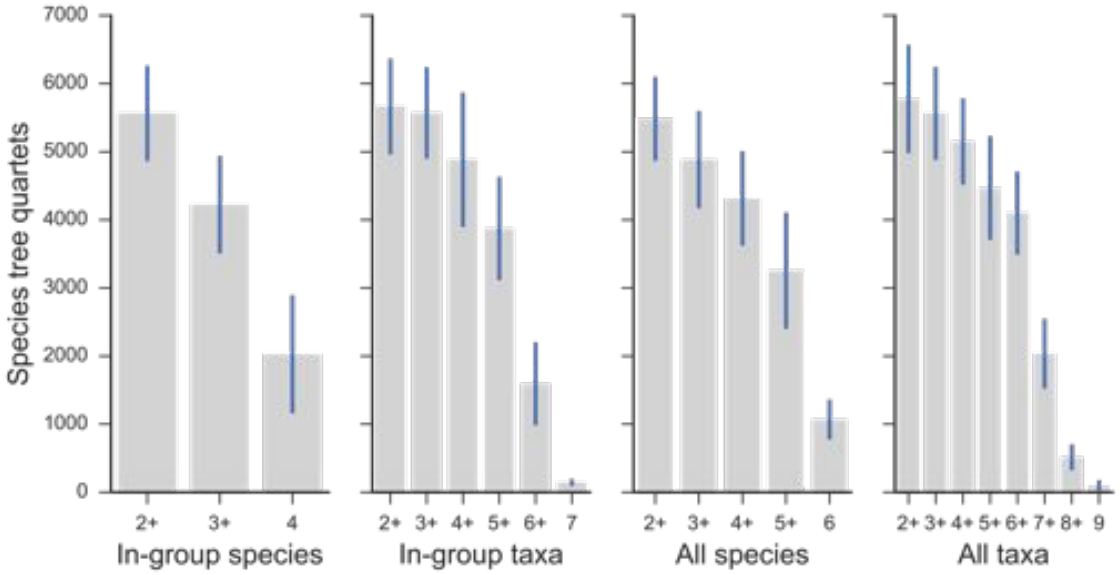

Figure 1.16: Mean number of quartets induced by species trees constructed from subsets of loci in ASTRAL-II. Blue bars indicate one standard deviation.

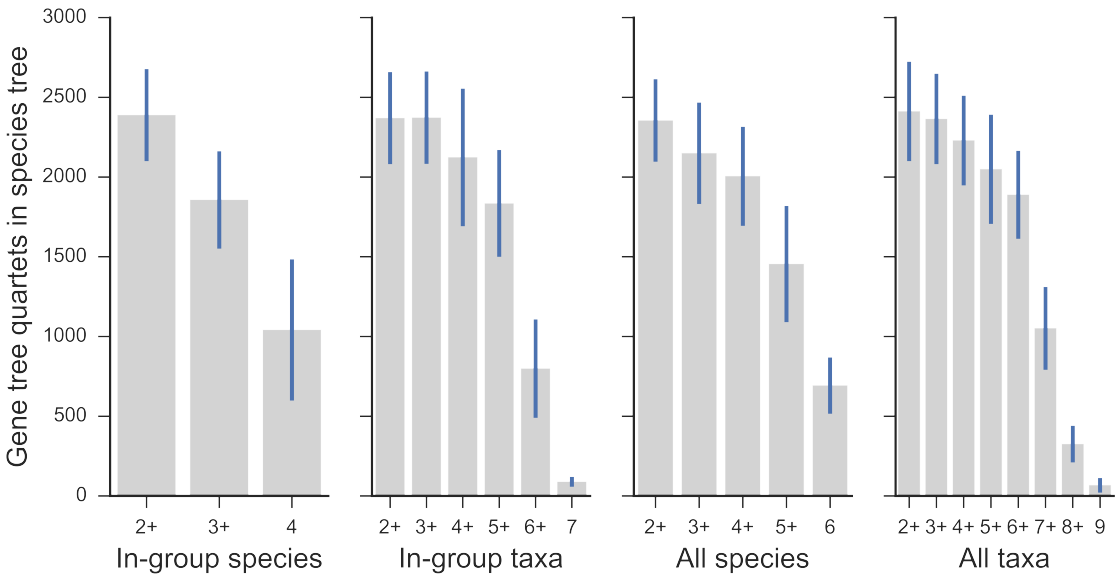

Figure 1.17: Mean number of quartets induced by species trees constructed from subsets of loci in ASTRAL-II and present in input subset of gene trees. Blue bars indicate one standard deviation.

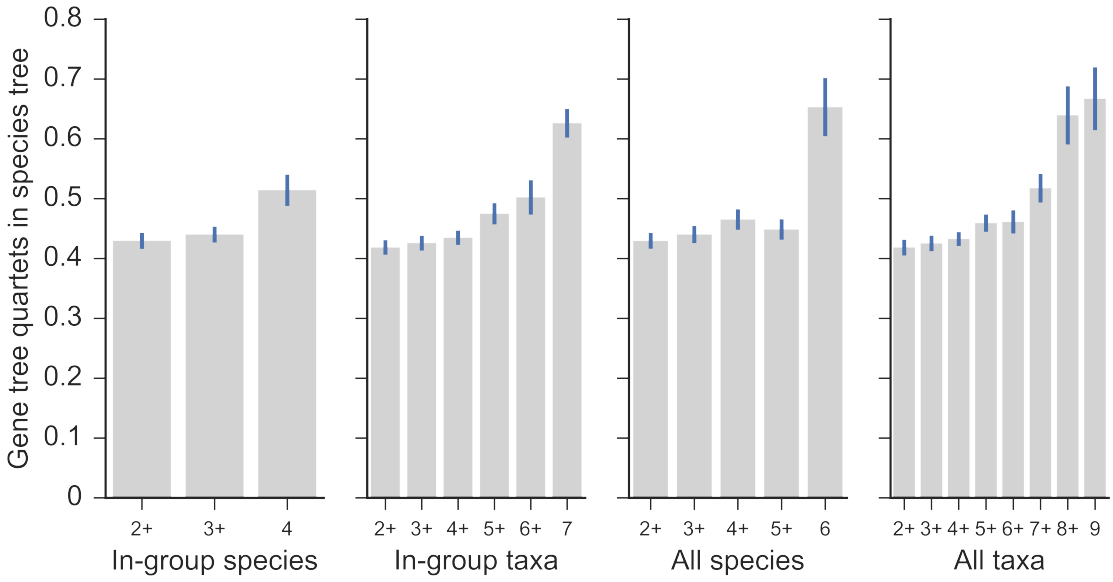

Figure 1.18: Mean fraction of quartets induced by species trees constructed from subsets of loci in ASTRAL-II and present in input subset of gene trees. Blue bars indicate one standard deviation.

CHAPTER 2: ROBUSTNESS OF QUARTET-BASED SPECIES TREE ESTIMATION TO MISSING DATA

2.1 Abstract

As loci from many, often unknown regions of the genome are incorporated into a single dataset for phylogenetic inference, gene trees may show conflicting histories due to a number of evolutionary phenomena including duplication, introgression, and incomplete lineage sorting. Over the past decade, a plethora of multi-species coalescent (MSC) based approaches have sought to ameliorate these issues, but analytical solutions to the resulting likelihood functions are intractable for more than a few taxa. This has spurred the creation of methods that are provably consistent under the MSC; that is, they will converge on the true species tree as the number of loci and sites increases. Here, we compare the accuracy of two major quartet-based species tree methods in the presence of missing data: SVDquartets, a single-site approach that directly utilizes sequence data on a site-by-site basis, and ASTRAL-II, a gene tree-based, summary method. We find higher accuracy employing SVDquartets when missing data is low, but higher accuracy of ASTRAL-II, when missing data is high. Furthermore, we find library and locus size play minor roles once a threshold locus length and number of loci are reached, regardless of method of species tree estimation.

2.2 Introduction

Sequence data generation has sharply increased through technological advancements allowing for longer reads, massive garnering of loci from online databases such as GenBank, and rapid generation of hundreds to tens-of-thousands of loci from reduced representation libraries, such as restriction site-associated DNA sequencing (RADseq, Baird et al., 2008). As loci from many, often unknown, regions of the genome are incorporated into a single dataset, gene

trees may show conflicting histories due to paralogy, introgression, and incomplete lineage sorting, among other causes (Degnan and Salter, 2005; Degnan and Rosenberg, 2006; Xu and Yang, 2016). These problems can make phylogenetic analyses unreliable via traditional concatenation based maximum likelihood or Bayesian methods (Degnan and Rosenberg, 2006; Edwards et al., 2007; Kubatko and Degnan, 2007; Roch and Steel, 2014; Chou et al., 2015).

To match the rising volume of data, data generation techniques, and demand to account for a suite of evolutionary phenomena, species tree estimation methods are being developed at a similar rate. In particular, phylogeny construction using models based on the multi-species coalescent (MSC), an extension of the Kingman coalescent (Kingman, 1982a,b; Hudson, 1983; Tajima, 1983), have sought to ameliorate one of these problems through incorporation of coalescent stochasticity among disparate regions of the genome (reviewed in Xu and Yang, 2016). While a complete likelihood function can be written for the MSC, it is computationally intractable, even for a handful of taxa. A suite of MSC methods, developed primarily over the past decade, utilize full likelihood in a Bayesian framework, summary statistics (e.g. gene tree topologies, branch lengths), or single-site (also referred to as site pattern or invariance methods) methods based on Lake's (1987) theory of phylogenetic invariants (reviewed in Xu and Yang, 2016).

While these methods have elucidated many difficult-to-resolve areas of the tree of life (e.g. early branches in birds, Jarvis *et al.* 2014; rampantly hybridizing American Oaks, Hipp *et al.* 2014), best practices and limiting scenarios with these techniques are not well documented. A standard practice is to report results from multiple methods, which at best have high support for the same topology across techniques, but at worst support multiple, discordant hypotheses. In the latter scenario, inference in hindered by the lack of simulation studies that predict model behavior under a variety of conditions (but see effects of ILS, Chou *et al.* 2015; gene tree discordance, Tian and Kubatko, 2017). To our knowledge, among the few studies that have investigated these models' behaviors, only one (Chou et al.,

2015) compared single-site and summary methods.

SVDquartets (Chifman and Kubatko, 2014) is a single-site method that applies algebraic statistics to calculate the singular value decomposition, or 'SVD', score for all quartets of taxa at sites with unlinked SNPs under the MSC. These quartets are agglomerated into a species tree using the Quartet FM algorithm (Reaz et al., 2014) in PAUP* (Swofford, 2002). ASTRAL-II (Mirarab et al., 2014b; Mirarab and Warnow, 2015) is a summary method that constructs a set of quartets from all input, unrooted gene trees before expanding that set via UPGMA based heuristic searches of those gene tree-induced quartets. The species tree that induces the largest number of congruent quartets from this expanded set, is returned as the best species tree estimate.

One major improvement in ASTRAL-II is the heuristic strategy that increases the search space in the presence of polytomies in input gene trees (Mirarab and Warnow, 2015). Briefly, multiple greedy consensus trees are computed for all gene trees via UPGMA on a similarity matrix of all taxa, with ties broken at random. The similarity of a pair of taxa is given by the number of quartets induced by all gene trees in which the pair appears on the same side of the quartet. The greedy consensus trees are then used to randomly resolve polytomies, and the resulting quartets are added to the search set, which increases the probability that ASTRAL-II will be consistent under the MSC. Although these improvements expand the conditions under which ASTRAL-II will accurately estimate the true species tree, and decreases susceptibility to gene tree estimation error, the relative performance of summary methods to single-site methods across varying levels of missing data, is unknown. As ASTRAL-II and SVDquartets are applied more frequently for species tree estimation from reduced representation libraries, which typically have high levels of missing data (see analysis of 10 RAD datasets, Eaton et al., 2016), the accuracy of these methods with respect to the data is paramount when evaluating phylogenetic hypotheses.

Here we investigate the effects of missing data in species tree estimation via single-site and summary MSC methods in two commonly employed, quartet-based species tree estimation

programs: SVDquartets and ASTRAL-II. We find higher accuracy employing SVDquartets when missing data is low, but higher accuracy with ASTRAL-II when missing data is high. Furthermore, we find library and locus size play minor roles once a threshold locus length and number of loci are reached, regardless of method of species tree estimation.

2.3 Methods

2.3.1 Datasets

To attempt to span the variation of empirical datasets, we altered simulation parameters at both the population and phylogenetic levels, as well as in the *in silico* library generation. We simulated double digest restriction site-associated (ddRAD, Peterson et al., 2012) libraries for a 12-taxon tree under the multi-species coalescent with a Jukes-Cantor model of sequence evolution using *simRRLs* (Eaton et al., 2016), which is built on the Python EggLib module (De Mita and Siol, 2012). All steps from sequence simulation through calculation of the Robinson-Foulds distances (Robinson and Foulds, 1981) were executed in a set of custom Python modules (available on ISG GitHub, github.com/isgilman).

2.3.2 Tree balance

We focused on completely balanced topologies, as they have the highest rates of hierarchical redundancy (*sensu* Eaton, 2016). This refers to the amount of information lost in any one split, due to its hierarchical placement in the tree, when data are missing from the tips. Nodes deeper in the tree can be resolved from more combinations of tips, and therefore have higher hierarchical redundancy (HR). In a completely unbalanced topology, all nodes are adjacent to one tip (the most recent split adjacent to two), and thus HR is generally low. Nodes with high HR are expected be the most efficient in terms of loci/sites needed to accurately estimate the true species tree in a quartet-based framework.

2.3.3 Branch length

The lengths of internal branches in the tree directly affect the rate of deep coalescent events. Longer internal branches allow for more coalescent events between speciation events, and therefore decrease the rate of deep coalescence and incomplete lineage sorting that reduce the efficiency of resolving the true species tree topology (Maddison, 1997). We simulated under three different regimes of internal branch lengths: short (0.5 coalescent units), medium (1.0 coalescent units), and long (2.0 coalescent units).

2.3.4 Effective population size

Effective population size (N_e) also directly affects the rate of deep coalescence. When N_e is high, more coalescent events are required between speciation events to avoid deep coalescence. Thus, incomplete lineage sorting is minimized when effective population sizes are small, increasing the efficiency of species tree estimation. We simulated under three values of N_e that were equal among taxa and constant through time: small ($1e4$), medium ($1e5$), and large ($1e6$).

2.3.5 Library and locus size

We simulated double digest restriction site-associated libraries (ddRAD, Peterson et al., 2012) to assess the accuracy of SVDquartets and ASTRAL-II with our empirical data (Chapter 1). To capture the size range of typical ddRAD libraries we adjusted both the average locus length and the number of loci per library. To span short, single-end to long, paired-end reads we simulated loci of sizes 100, 200, 300, 400, 500, and 600bp. Library size (in number of loci) also varied: small (500 loci), medium (1000 loci), and large (10000 loci).

2.3.6 Missing data

PyRAD (Eaton, 2014) was used to generate the final alignments from raw sequences output by *simRRLs*. *PyRAD* was run with default settings: minimum depth of six reads for within sample clustering of stacks, maximum of 4 low quality sites per read, 88% clustering threshold (within and between samples), minimum of four samples per final locus, and maximum of three samples with a shared heterozygous site in a final locus. For each dataset, a proportion of randomly selected sites were removed before gene and species tree estimation. The proportion ranged from 0% (no data removed) to 80%. Quartet inference at any locus or site relies on data present for at least four taxa; therefore rates of missing data over 75%, on a 12 tip tree will, on average, not be possible.

2.3.7 Gene and species tree estimation

Gene tree estimation

Due to the computational intractability of estimating millions of gene trees, many with large amounts of missing data, in a maximum likelihood or Bayesian framework, gene trees were estimated for each locus via neighbor-joining under a Jukes-Cantor model, with ties broken at random, in PAUP* v.4.0a.152 (Swofford, 2002).

Species tree estimation

Species trees were estimated using SVDquartets (Chifman and Kubatko, 2014) and ASTRAL-II v4.10.12 (Mirarab et al., 2014b; Mirarab and Warnow, 2015). ASTRAL-II was run with the command: `java -jar astral.4.10.12.jar -i [input gene trees] -o [output tree] -t 0`. The final argument of the ASTRAL-II command, '`-t 0`', suppresses branch annotations, which were unnecessary as we were concerned with topology alone. SVDquartets was run in PAUP* with ambiguities set to 'Missing' and all possible quartets evaluated. The accuracy of species trees were evaluated using the average normalized Robinson-Foulds (RF)

distance (Robinson and Foulds, 1981) between the estimated species tree and the true tree simulated under for 10 simulation replicates. The maximum RF distance between two trees with congruent sets of taxa is $2(n-3)$, where n is the number of tips. The normalized RF distance is the measured RF distance divided by $2(n-3) = 18$ for a 12-taxon tree.

2.4 Results

Across all simulation conditions, SVDquartets was more accurate than ASTRAL-II when missing data was below 40%, but less accurate otherwise (Figure 2.1). Even when missing data was absent or low, ASTRAL-II averaged about one bipartition incongruent with the true species tree, which can only result from a polytomy in the estimated species tree that is not incongruent with the true species tree. RF distances in species trees estimated with ASTRAL-II increased less rapidly with the rate of missing data than those estimated with SVDquartets, and ASTRAL-II was much more accurate when missing data was above 50%.

2.4.1 Locus size

The effects of locus size were greater in magnitude, and more frequently significant, in species trees estimated with ASTRAL-II (Figure 2.2A-B, 2.3-2.4). Libraries simulated with small locus sizes (100-200bp) resulted in less accurate species trees for almost all rates of missing data (Figure 2.2A, 2.3). The decrease in accuracy varied from less than 10% to greater than 40% of the maximum RF distance (18), and the gap in accuracy tended to increase with the amount of missing data. The disparity in accuracy between libraries with different sized loci did not, in general, increased with the disparity in locus size. Furthermore, when libraries were simulated with loci of 300bp or greater, there were no significant differences in the accuracy of species tree estimation. The latter two results suggest that there may be a threshold of locus size, below which the accuracy of gene tree estimation and species tree estimation in ASTRAL-II is significantly reduced.

In contrast, the accuracy of species tree estimation in SVDquartets was not reduced, in general, for libraries with shorter loci (Figure 2.2B, 2.4). Significant reductions in accuracy were only observed for the shortest loci (100bp), when rates of missing data were high. The decreases in accuracy were also lower in magnitude, seldom larger than 10% of the maximum RF distance, than in ASTRAL-II. The increase in accuracy of SVDquartets with short-locus libraries was largest when rates of missing data were low (Figure 2.5). When locus size was 300bp, or greater, and rates of missing data were low, the accuracy of SVDquartets was only slightly higher then, or comparable to, ASTRAL-II, but as missing data increased over 40%, SVDquartets tended to be less accurate than ASTRAL-II. The inaccuracy of species tree estimation in ASTRAL-II when loci were less than 300bp significantly inflated RF distances across subsequent analyses. To better broadly compare the behavior of ASTRAL-II and SVDquartets we removed loci less than 300bp from our analyses below. The results of the full data are available in the Supplementary Material.

2.4.2 Library size

When library size (in number of loci) was doubled from 500 to 1000, there was no increase in the accuracy of SVDquartets (Figure 2.2D, 2.6D). However, there were significant increases when library size was enlarged by an order of magnitude or more (Figure 2.2D, 2.6E-F). The effect of library size became larger as the rate of missing data increased past 20%. The magnitude of this effect was relatively small; RF distances of species trees resulting from smaller libraries were, on average 2.5-15% higher than those estimated using the largest library size.

Technical problems prevented the completion of all ASTRAL-II simulations with 10,000 loci. Therefore all results below directly compare ASTRAL-II and SVDquartets simulations with small and medium sized libraries unless explicitly stated.

2.4.3 Internal branch length

Elongation of internal branches resulted in significantly lower RF distances in both ASTRAL-II and SVDquartets (Figure 2.2E-F, 2.7). For species trees estimated in ASTRAL-II, there was a significant decrease in RF distance when internal branch lengths were lengthened from 0.5 to 1.0 coalescent units (CU), but not from 1.0 to 2.0 CU (Figure 2.2E, 2.7A-C). The magnitude of the decrease in RF distance was similar when comparing short (0.5 CU) and medium (1.0 CU), or long (2.0 CU), internal branch lengths: between 5% and 10% maximum RF distance, slightly increasing with the rate of missing data. Significant increases in RF distance in species trees estimated in SVDquartets due to decreased internal branch lengths resulted only when comparing short and medium, or long, internal branch lengths when missing data was moderate to high (Figure 2.2F, 2.7D-F). The magnitude of the increase was larger when the disparity in branch lengths was increased from 0.5 to 1.5 CU, and also increased with the rate of missing data. The accuracy of SVDquartets was greater than, or equal to, that of ASTRAL-II when rates of missing data were less than 40%, when branch lengths were short (Figure 2.2E-F, 2.7G-I). ASTRAL-II was slightly more accurate at high rates of missing data, across all branch lengths.

2.4.4 Effective population size

Varying the effective population size had the largest magnitude effect on the performance of both species tree methods. With ASTRAL-II, small ($1e4$) effective population sizes (N_e) were roughly 20-60% and 15-75% less accurate than medium ($1e5$) and large ($1e6$) N_e, respectively (Figure 2.2G-H, 2.8A-B). The magnitude of the difference in accuracy tended to increase with the rate of missing data, and a significant difference in accuracy was only present when comparing medium and large N_e at rates of missing data of 70-80% (Figure 2.2G, 2.8C). The behavior of SVDquartets was similar (Figure 2.2H, 2.8D-F), but when the rate of missing data was low ($\leq 20\%$) there was very little or no difference in the accuracy

between simulations with different N_e.

There was no, or nearly no, difference in accuracy of ASTRAL-II and SVDquartets when effective population sizes were medium or high and missing data rates were low or moderate (Figure 2.8H-I). Only with small N_e did SVDquartets outperform ASTRAL-II. When the rate of missing data was less than 50%, SVDquartets was up to 17% more accurate, but ASTRAL-II was more accurate when rates were very high (60+%).

2.5 Discussion

2.5.1 Missing data

A growing number of studies have shown that reduced representation sequencing, specifically techniques in the RADseq family, are information-rich enough for even deep-scale phylogenetic inference (reviewed in Eaton et al., 2016). Previous work has primarily focused on the ability to generate phylogenetically-informative data via these techniques, but not the species tree methods employed to erect phylogenetic hypotheses with these data. We show that accurate species trees can be estimated with moderate or large rates of randomly distributed missing data, using both summary and single-site quartet-based methods.

Species trees estimated using SVDquartets, a single-site method, tended to be more accurate when missing data was less than 50%, but less accurate than ASTRAL-II, a summary method, when missing data was high, which is common for RADseq datasets. Furthermore, when missing data was low, SVDquartets was able to recover the true tree over a wider array of simulation conditions. With sufficient data, both ASTRAL-II and SVDquartets are statistically consistent under the multi-species coalescent, but ASTRAL-II relies on the input of well estimated gene trees. While we believe that our use of neighbor-joining gene tree estimation under a Jukes-Cantor model was justified due to the size and scope of our study, best practices would entail assessing models of sequence evolution on a locus-by-locus basis before gene tree construction in a maximum likelihood or Bayesian framework. Hence

the baseline level of incongruence between estimated and true species tree in ASTRAL-II may be nonzero due to the extra step of gene tree estimation in the ASTRAL-II pipeline.

The effects of gene tree estimation may also be responsible for the higher accuracy of ASTRAL-II when missing data is high. When rates of missing data were high, many more gene trees had polytomies, which ASTRAL-II uses as part of its heuristic search strategy to increase its species tree search space (Mirarab and Warnow, 2015), whereas SVDquartets gains no information from sites that do not generate quartets. In Chapter 1 we showed that high rates of missing data led to fewer informative quartets in analyses performed by SVDquartets than in those performed in ASTRAL-II. While estimation was rarely completely accurate in ASTRAL-II at high levels of missing data, ASTRAL-II's heuristic search provided more information leading to higher relative accuracy.

2.5.2 Library generation

As the rate of missing data and underlying species tree topology are (generally) not amendable, library generation is one of the few processes in which researchers have direct control over the potential informativeness of their data. We found that loci smaller than 300bp, in general, resulted in significantly less accurate species tree estimates in ASTRAL-II. This is most likely due to gene tree estimation error on small loci. The number of informative sites in a locus is directly related to its length, and so many of our smallest loci had very few (or no) informative sites for gene tree estimation. Because SVDquartets takes a concatenated dataset as input, the various locus sizes were essentially linear scaling factors for the size of the total data matrix, which did not significantly improve accuracy.

A twofold increase in the library size (in number of loci) did not significantly improve the accuracy of either species tree method, although a further increase, of an order of magnitude, did. Despite the fact that doubling the number of loci produced a linear increase in the number of sites used for inference in SVDquartets, it was not trivial that there would be no significant effect. Doubling the length of a locus will double the sites used in analyses, but

the sources of that information will remain relatively unchanged. Doubling the number of loci will also double the information in the data, but sample twice as many regions of the genome. There is potentially more disparate information because those sites sampled may or may not have the same evolutionary histories.

2.5.3 Tree shape

In a multi-species coalescent framework, tree shape consists of three parts: balance, length, and width. Balance refers to the evenness in the number of daughter nodes at any node rank in the tree, and here, by length we mean internal branch length in coalescent units. Width refers to the analogy of lineages tracing gene trees in wider, tube-like species trees (see Maddison, 1997, Figure 4). Wide branches contain many lineages, while narrow branches contain few, and we use effective population size as a proxy for branch width. The effects of internal branch length and width on the efficiency and accuracy of species tree estimation from hundreds or thousands of loci has been a subject of recent attention (Chou et al., 2015; Collins and Hrbek, 2015; Eaton et al., 2016; Shekhar et al., 2017), although reconciliation among gene trees and between gene and species trees in the presence of incomplete lineage sorting (ILS) has been a focal point of phylogenetics for decades (e.g., Maddison, 1997; Degnan and Salter, 2005; Knowles and Carstens, 2007; Edwards et al., 2007; Reid et al., 2011). We expect that species tree estimation will be less efficient, and potentially less accurate, when ILS is high. The particular source of discordance modeled in our study resulted from coalescent stochasticity alone, as this is the only source of incongruence that most multi-species coalescent based approaches account for when agglomerating discordant information. Incomplete lineage sorting resulting from deep coalescent events is expected to be highest when i) internal branch lengths are short, therefore decreasing the time for allele sorting, and ii) internal branches are wide, increasing the number of alleles that must be sorted.

Chou *et al.* (2015) showed that ASTRAL-II tended to outperform SVDquartets (among

other species tree methods) when ILS was high, but that SVDquartets had generally high accuracy across simulation parameters. We found a general decrease in performance of both methods when internal branch lengths were short, as well as when effective population sizes were low. Surprisingly, the primary driver of accuracy across our simulations was effective population size, which dramatically reduced performance of ASTRAL-II and SVDquartets when low. When considering only medium and large N_e, both methods had perfect, or near perfect, accuracy when missing data was below 70%, regardless of internal branch lengths. This counterintuitive behavior may have been an artifact of a lack of polymorphism resulting from low N_e and low mutation rate. Shekhar *et al.* (2017) recently found similarly counterintuitive behavior of ASTRAL-II when very long branch lengths were present in a completely balanced topology that they had expected to be the most efficient topology to recover in terms of number of gene trees. This result is analogous to long branch attraction in sequence-based species tree methods and is caused by one branch with essentially no gene tree discordance above branches with relatively high discordance. The lack of discordance along the anomalous branch will cause the three possible induced quartet topologies to occur with near equal frequencies, and thus have little power to infer the correct relationship. These results may underly previous findings that concatenation outperforms multi-species coalescent based approaches when ILS is very low (Chou et al., 2015; Tonini et al., 2015). Taken together, these results show that future work identifying and resolving anomalous tree shapes with quartet-based species tree methods is needed.

2.6 Conclusion

We explored the accuracy of species tree estimation using two commonly employed quartet-based, multi-species coalescent methods in the presence of missing data and found higher accuracy using SVDquartets, a single-site method, when missing data was low, but higher accuracy with ASTRAL-II, a summary method, when missing data was high. We believe that the heuristic expansion of the search space when polytomies are frequent in gene trees

improved accuracy when rates of missing data were high, and advise careful analysis of levels of missing data when assessing the performance of quartet-based species tree programs. In general, we find library and locus size play minor roles once a threshold locus length and number of loci are reached, regardless of method of species tree estimation. Future work to characterize the performance of these techniques with anomalous tree shapes, especially with adjacent long and short branches, are also needed.

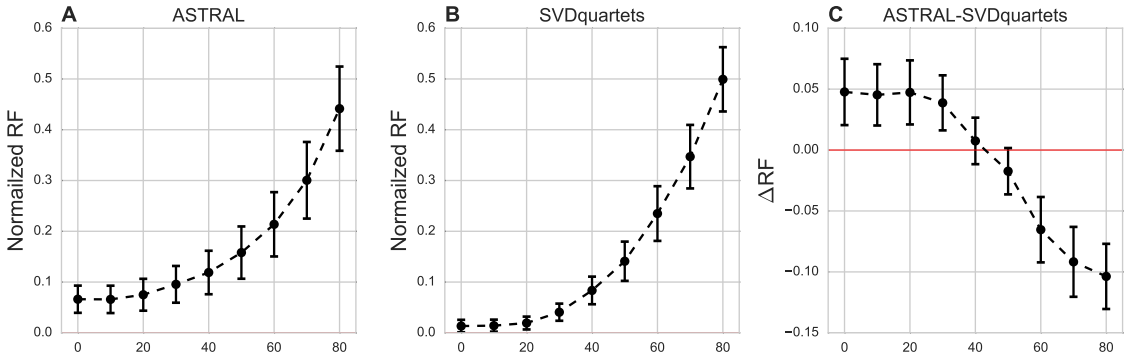

Figure 2.1: Comparisons of normalized RF distances across all simulations using ASTRAL-II (A), SVDquartets (B), and the difference between them (C). Points above, and below, 0 (C, red lines) indicate higher RF distances, and therefore lower accuracy, for species trees estimated with ASTRAL-II, and SVDquartets, respectively. Vertical bars represent 95% confidence interval.

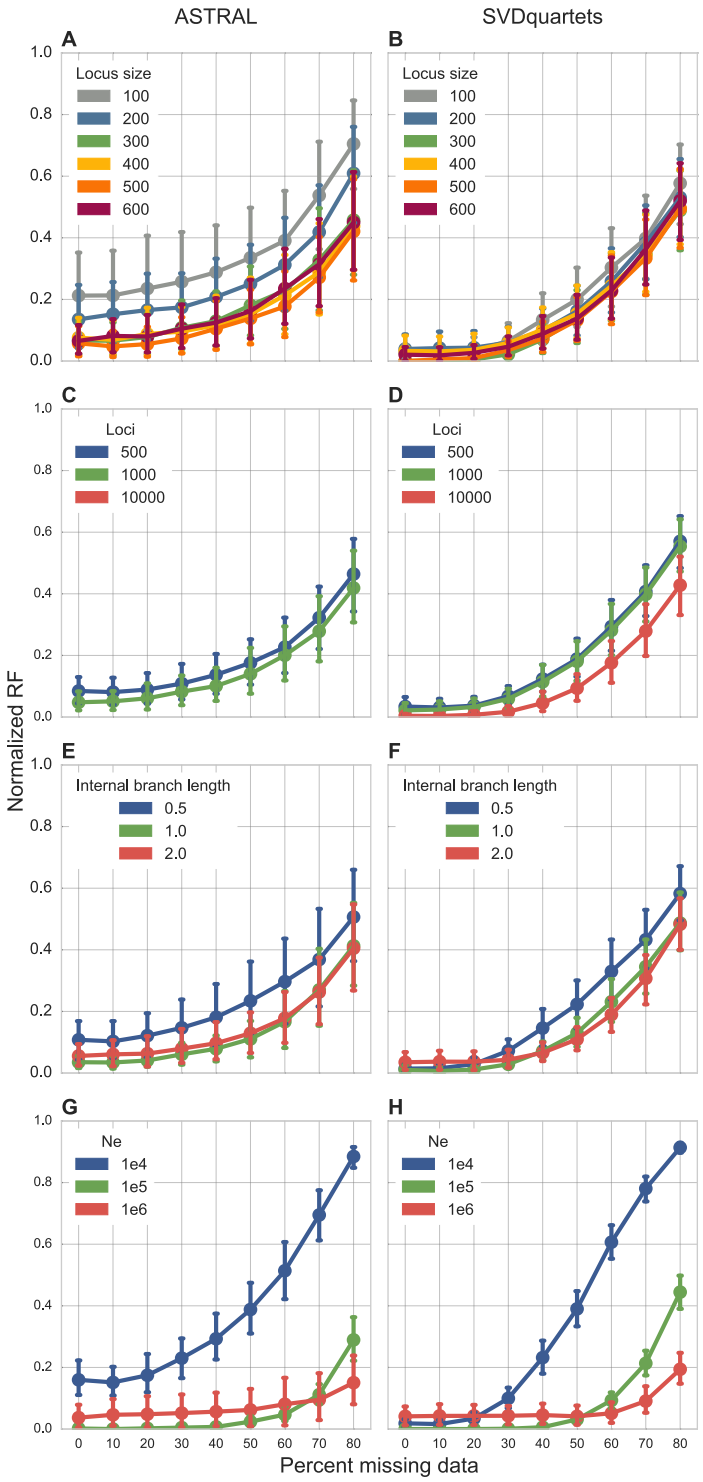

Figure 2.2: Normalized RF distances of estimated species tree using ASTRAL-II (left) and SVDquartets (right) under various locus sizes (A-B), number of loci (C-D), internal branch lengths (E-F), and effective population sizes (G-H). Vertical bars represent 95% confidence intervals.

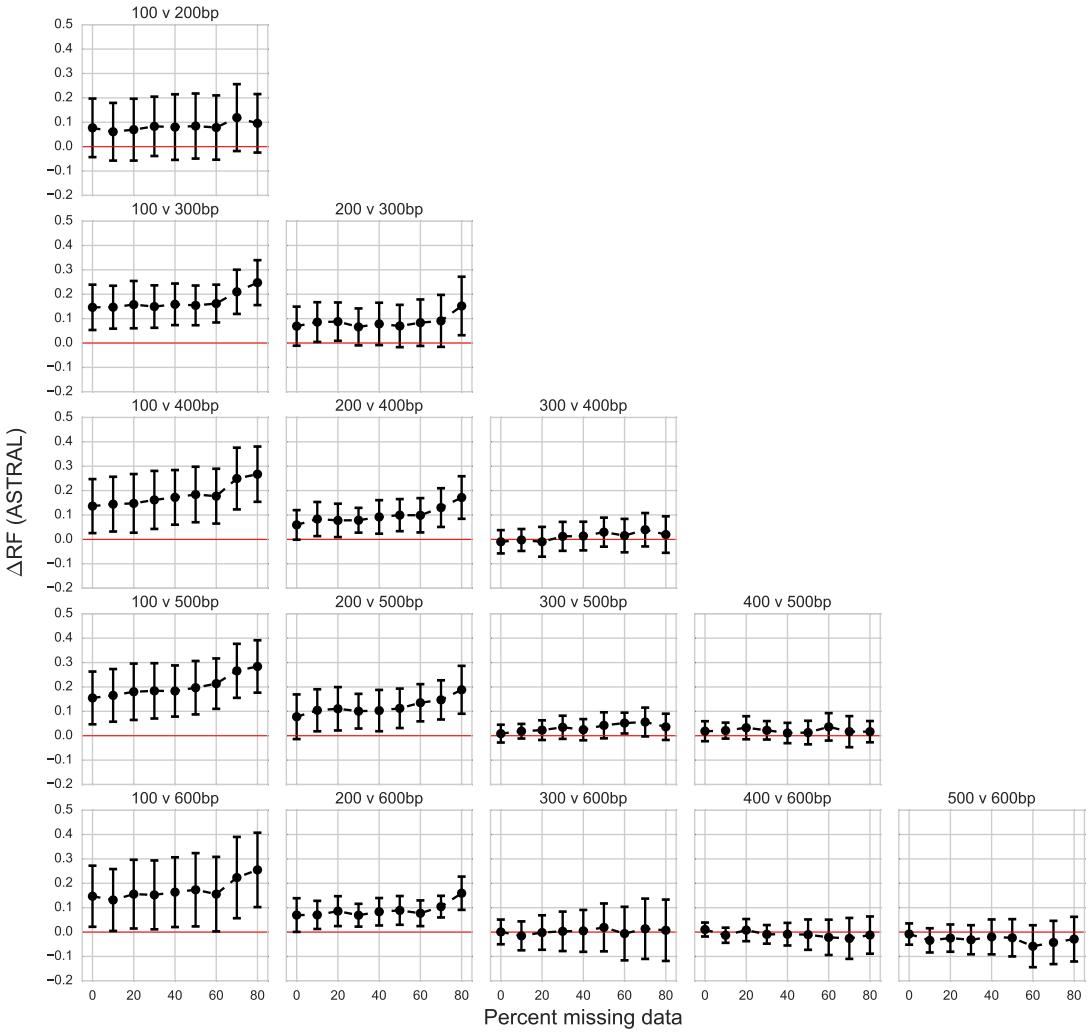

Figure 2.3: Comparisons of normalized RF distances for species trees estimated in ASTRAL-II between libraries with different locus sizes. Points significantly above, and below, 0 (red lines) indicate lower accuracy of shorter, and longer loci, respectively. Vertical bars represent 95% confidence interval.

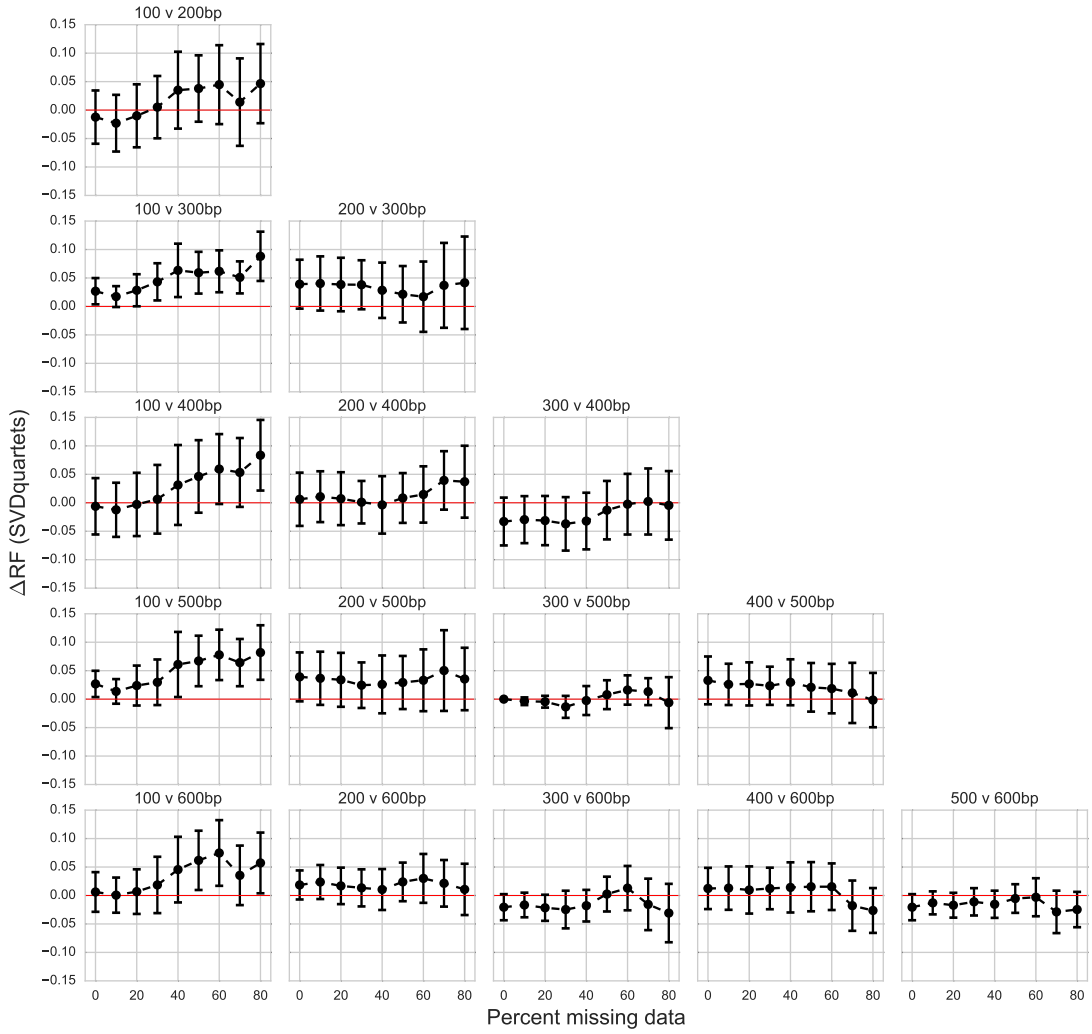

Figure 2.4: Comparisons of normalized RF distances for species trees estimated in SVDquartets between libraries with different locus sizes. Points significantly above, and below, 0 (red lines) indicate lower accuracy of shorter, and longer loci, respectively. Vertical bars represent 95% confidence interval.

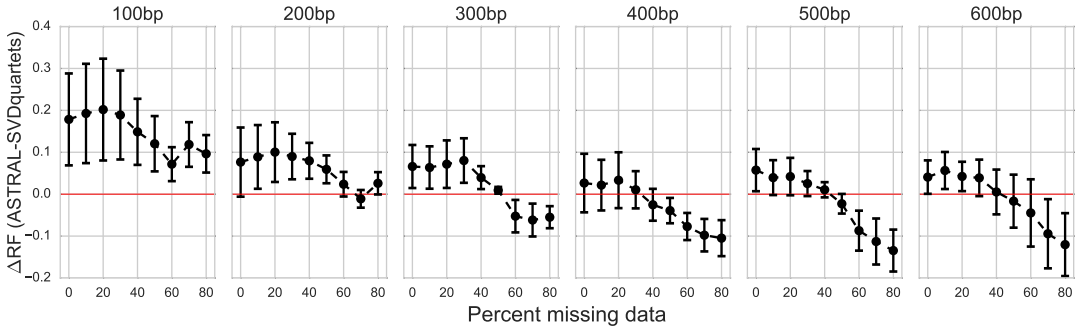

Figure 2.5: Comparisons of normalized RF distances between species trees estimated in ASTRAL-II and SVDquartets using libraries with various locus sizes. Points significantly above, and below, 0 (red lines) indicate higher RF distances, and therefore lower accuracy, for species trees estimated with ASTRAL-II, and SVDquartets, respectively. Vertical bars represent 95% confidence interval.

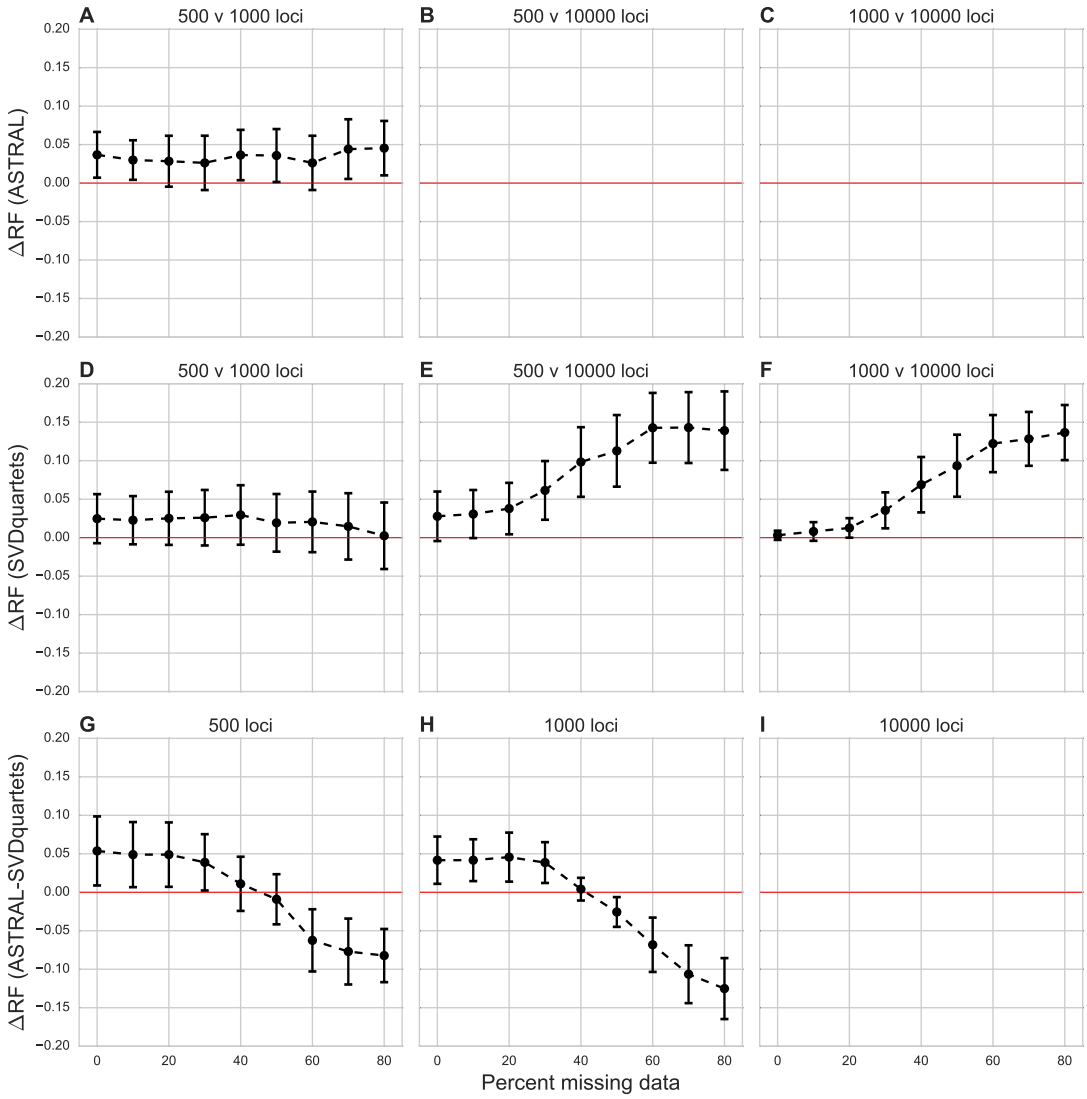

Figure 2.6: Comparisons of normalized RF distances between different sized libraries using ASTRAL-II (A-C) and SVDquartets (D-F). Points significantly above, and below, 0 (red lines) indicate lower accuracy of smaller, and larger libraries, respectively. In comparisons of RF distances between ASTRAL-II and SVDquartets with small, medium, and large libraries (G-I), points significantly above, and below, 0 (red lines) indicate higher RF distances, and therefore lower accuracy, for species trees estimated with ASTRAL-II, and SVDquartets, respectively. Vertical bars represent 95% confidence interval.

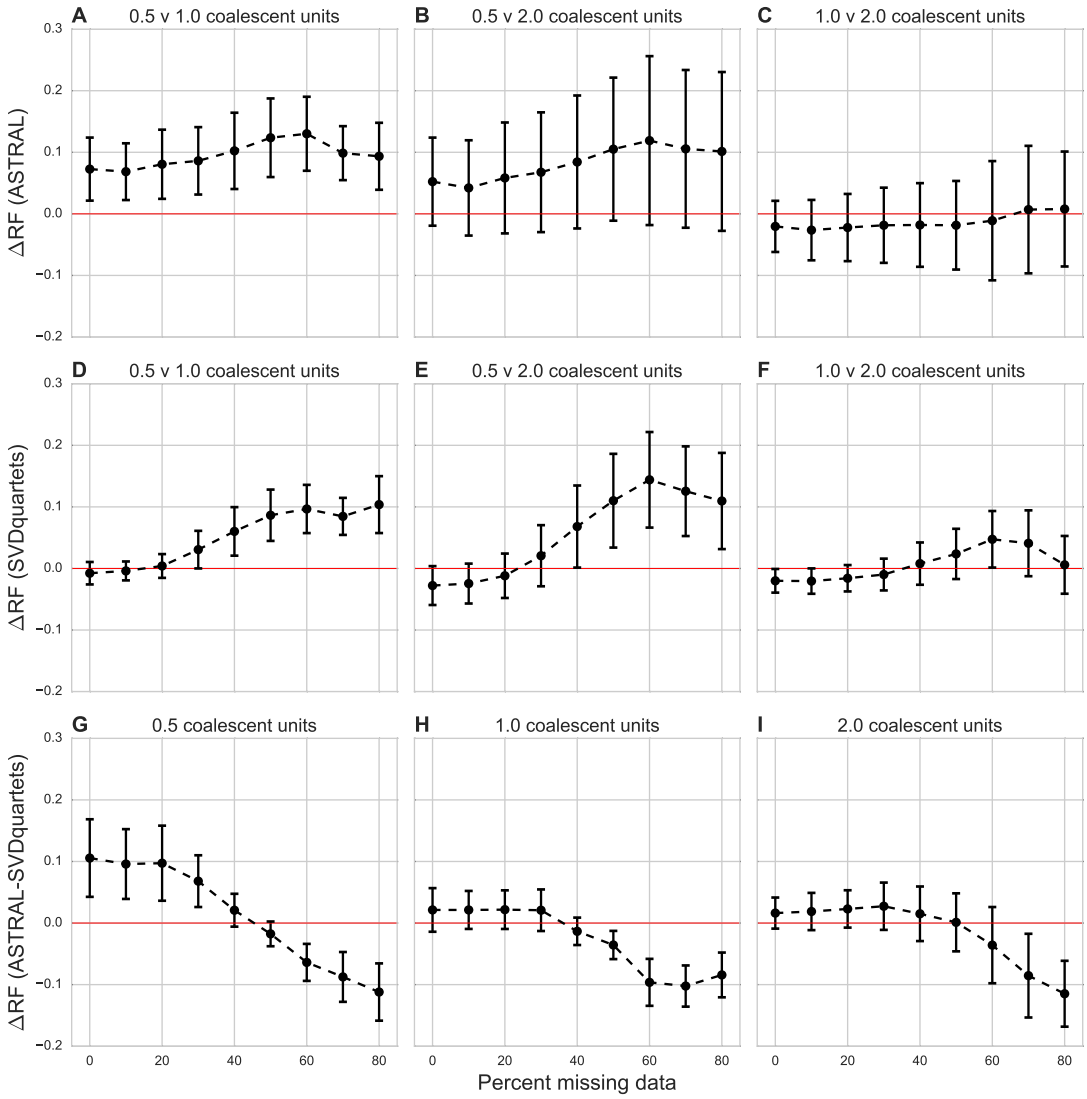

Figure 2.7: Comparisons of normalized RF distances between topologies simulated with varying internal branch lengths using ASTRAL-II (A-C) and SVDquartets (D-F). Points above, and below, 0 (red lines) indicate lower accuracy of shorter, and longer, internal branch lengths, respectively. In comparisons of RF distances between ASTRAL-II and SVDquartets with short, medium, and long internal branches (G-I), points above, and below, 0 (red lines) indicate higher RF distances, and therefore lower accuracy, for species trees estimated with ASTRAL-II, and SVDquartets, respectively. Vertical bars represent 95% confidence interval.

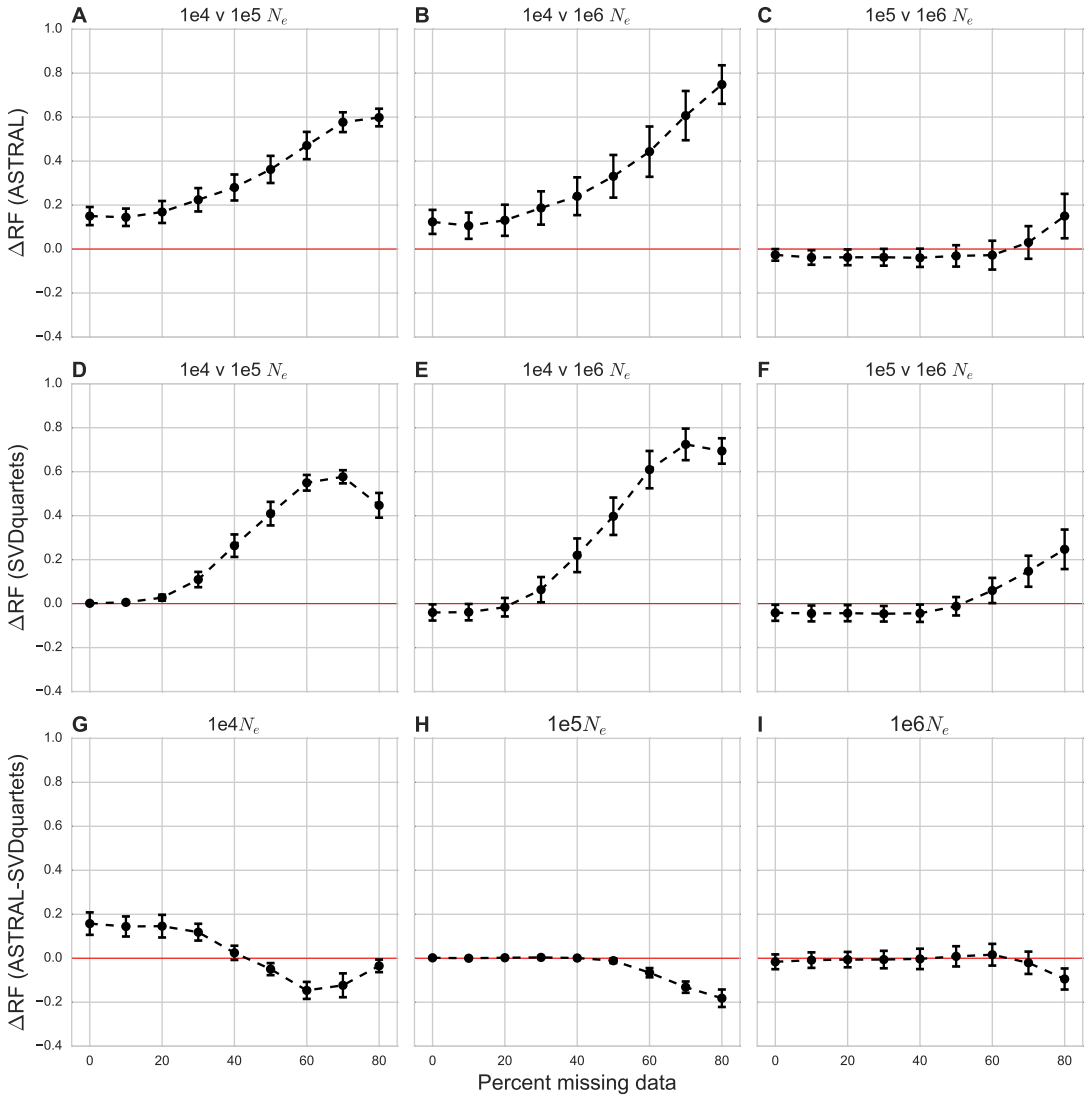

Figure 2.8: Comparisons of normalized RF distances between topologies simulated with varying effective population sizes (N_e) using ASTRAL-II (A-C) and SVDquartets (D-F). Points above, and below, 0 (red lines) indicate lower accuracy of smaller, and larger N_e, respectively. In comparisons of RF distances between ASTRAL-II and SVDquartets with small, medium, and large N_e (G-I), points above, and below, 0 indicate higher RF distances, and therefore lower accuracy, for species trees estimated with ASTRAL-II, and SVDquartets, respectively. Vertical bars represent 95% confidence interval.

Bibliography

Andrews, K. R., J. M. Good, M. R. Miller, and G. Luikart. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics .

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 3:e3376–17.

Bayzid, M. S. and T. Warnow. 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29:2277–2284.

Behr, H. H. 1855. A new genus and species, *Chloropyron palustre*. Proceedings of the California Academy of Sciences 1:62.

Cariou, M., L. Duret, and S. Charlat. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecology and Evolution 3:846–852.

Chifman, J. and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.

Chou, J., A. Gupta, S. Yaduvanshi, R. Davidson, M. Nute, S. Mirarab, and T. Warnow. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics 16:S2.

Chuang, T. I. and L. R. Heckard. 1973. Taxonomy of Cordylanthus subgenus Hemistegia (Scrophulariaceae). Brittonia 25:135–158.

Chuang, T. I. and L. R. Heckard. 1975a. Re-evaluation of bract morphology in taxonomy of Cordylanthus (Scrophulariaceae). Madroño .

Chuang, T. I. and L. R. Heckard. 1975b. Taxonomic status of Cordylanthus (subg. Dicranostegia) orcuttianus (Scrophulariaceae). Madroño .

Chuang, T. I. and L. R. Heckard. 1986. Systematics and Evolution of Cordylanthus (Scrophulariaceae-Pedicularieae)(Including the Taxonomy of Subgenus Cordylanthus). Systematic Botany Monographs 10:1.

Collins, R. A. and T. Hrbek. 2015. An in silico comparison of reduced-representation and sequence-capture protocols for phylogenomics. bioRxiv Pages 1–51.

De Mita, S. and M. Siol. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC genetics 13:27.

Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. PLoS Genetics 2:e68–7.

Degnan, J. H. and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Doyle, J. J. and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19:11–15.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29:1969–1973.

Eaton, D. A. R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30:1844–1849.

Eaton, D. A. R., E. L. Spriggs, B. Park, and M. J. Donoghue. 2016. Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. Systematic Biology Page syw092.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32:1792–1797.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences 104:5936–5941.

Ferris, R. S. 1918. Taxonomy and distribution of *Adenostegia*. Bulletin of the Torrey Botanical Club 45:399–423.

Gray, A. 1867. Characters of new plants of California and elsewhere, principally of those collected by H. N. Bolander in the state geological survey. Proceedings of the American Academy of Arts 7:327–402.

Hersch-Green, E. I. and R. Cronn. 2009. Tangled trios?: Characterizing a hybrid zone in Castilleja (Orobanchaceae). American Journal of Botany 96:1519–1531.

Hipp, A. L., D. A. R. Eaton, J. Cavender-Bares, E. Fitzek, R. Nipper, and P. S. Manos. 2014. A Framework Phylogeny of the American Oak Clade Based on Sequenced RAD Data. PLoS ONE 9:e93975–12.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution Pages 203–217.

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, M. S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldon, S. Capella-Gutierrez, J. Huerta-Cepas, B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E. Green, M. W. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry, M. F. Bertelsen, F. H. Sheldon, R. T. Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M. P. Cruz Schneider, F. Prosdocimi, J. A. Samaniego, A. M. Vargas Velazquez, A. Alfaro-Nunez, P. F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou, P. Perelman, A. C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang,

J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldsa, L. ORLANDO, F. K. Barker, K. A. Jonsson, W. Johnson, K.-P. Koepfli, S. O'Brien, D. Haussler, O. A. Ryder, C. Rahbek, E. WILLERSLEV, G. R. Graves, T. C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alstrom, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow, W. Jun, M. T. P. GILBERT, and G. Zhang. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Jepson, H. L. 1925. *A manual of the flowering plants of California*. University of California Press, Berkeley and Los Angeles.

Kingman, J. 1982a. The coalescent. Stochastic processes and their applications 13:235–248.

Kingman, J. F. C. 1982b. On the genealogy of large populations. Journal of Applied Probability 19:27–43.

Knowles, L. L. and B. Carstens. 2007. Delimiting Species without Monophyletic Gene Trees. Systematic Biology 56:887–895.

Kubatko, L. S. and J. H. Degnan. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. Systematic Biology 56:17–24.

Lake, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Molecular Biology and Evolution Pages 167–191.

Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology 46:523–536.

Mason, H. L. 1957. A flora of the marshes of California. University of California Press, Berkeley and Los Angeles.

Mirarab, S., M. S. Bayzid, B. Boussau, and T. Warnow. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346:1250463–1250463.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014b. ASTRAL: genome-scale coalescent-based species tree estimation . Bioinformatics 30:I541–I548.

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

Munz, P. A. 1959. A California Flora. University of California Press, Berkeley and Los Angeles.

Pennell, F. W. 1951. Scrophulariaceae. *In* L. Abrams [ed.], Illustrated flora of the Pacific states. Stanford University Press, Stanford.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. 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:e37135–11.

Pimm, S. L. and L. N. Joppa. 2015. How Many Plant Species are There, Where are They, and at What Rate are They Going Extinct? Annals of the Missouri Botanical Garden 100:170–176.

Reaz, R., M. S. Bayzid, and M. S. Rahman. 2014. Accurate Phylogenetic Tree Reconstruction from Quartets: A Heuristic Approach. PLoS ONE 9:e104008–13.

Reid, N., J. R. Demboski, and J. Sullivan. 2011. Phylogeny Estimation of the Radiation of Western North American Chipmunks (Tamias) in the Face of Introgression Using Reproductive Protein Genes. Systematic Biology 61:44–62.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences Pages 131–147.

Roch, S. and M. Steel. 2014. Likelihood-based tree reconstruction on a concatenation of

aligned sequence data sets can be statistically inconsistent. Theoretical population biology 100C:56–62.

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

Rubin, B. E. R., R. H. Ree, and C. S. Moreau. 2012. Inferring Phylogenies from RAD Sequence Data. PLoS ONE 7:e33394–13.

Shekhar, S., S. Roch, and S. Mirarab. 2017. Species tree estimation using ASTRAL: how many genes are enough? arXiv.org .

Soltis, D. E., M. J. Moore, J. G. Burleigh, C. D. Bell, and P. S. Soltis. 2010. Assembling the Angiosperm Tree of Life: Progress and Future Prospects. Annals of the Missouri Botanical Garden 97:514–526.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Swofford, D. 2002. PAUP 4.0 b10: phylogenetic analysis using parsimony. Sunderland (MA): Sinauer Associates. .

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Tank, D. C., J. M. Egger, and R. G. Olmstead. 2009. Phylogenetic classification of subtribe Castillejinae (Orobanchaceae). Systematic botany .

Tank, D. C. and R. G. Olmstead. 2008. From annuals to perennials: phylogeny of subtribe Castillejinae (Orobanchaceae). American Journal of Botany 95:608–625.

Tian, Y. and L. S. Kubatko. 2017. Expected pairwise congruence among gene trees under the coalescent model. Molecular Phylogenetics and Evolution 106:144–150.

Tonini, J. a., A. Moore, D. Stern, M. Shcheglovitova, and G. Ortì. 2015. Concatenation and Species Tree Methods Exhibit Statistically Indistinguishable Accuracy under a Range of Simulated Conditions. PLoS Currents 7.

Xu, B. and Z. Yang. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. Genetics 204:1353–1368.

Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620.

Zwickl, D. J. 2006. Genertic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Unpublished D. Phil. Thesis, University of Texas at Austin .