Species Tree Estimation Using ddRADseq Data from Historical Specimens Confirms the Monophyly of Highly Disjunct Species of Chloropyron (Orobanchaceae)

Authors: Ian S. Gilman, and David C. Tank
Source: Systematic Botany, 43(3) : 701-708
Published By: American Society of Plant Taxonomists
URL: https://doi.org/10.1600/036364418X697418

Systematic Botany (2018), 43(3): pp. 701–708

DOI 10.1600/036364418X697418
Date of publication August 10, 2018

Species Tree Estimation using ddRADseq Data from Historical Specimens Confirms the Monophyly of Highly Disjunct Species of *Chloropyron* (Orobanchaceae)

Ian S. Gilman[1,2,3,4,5] and David C. Tank[1,2,3]

[1]Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA
[2]Stillinger Herbarium, University of Idaho, Moscow, ID 83844, USA
[3]Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, USA
[4]Current address: Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA
[5]Author for correspondence (ian.gilman@yale.edu)

Communicating Editor: Sasa Stefanovic

Abstract—Sequence data exist 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 a molecular 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 genus of rare flowering plants, *Chloropyron* (Orobanchaceae), from herbarium sheets dating up to 25 yr 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 each other's closest relative 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 specimens and rare taxa.

Keywords—Missing data, phylogenomics, restriction site-associated DNA sequencing, halophyte, hemiparasite.

Recent estimates indicate that of the 500,000 + species of plants (Soltis et al. 2010), about 1/3 are at risk of extinction, a rate 1000–10,000 times higher than the background extinction rate (Pimm and Joppa 2015). Currently, sequence data exists for about 138,000 species of plants (NCBI Taxonomy Browser, accessed November 2017), and so we are at risk of losing many branches of the plant tree of life before they are placed into an evolutionary context using molecular data. This necessitates methods for phylogeny estimation of understudied taxa with no reference genome, poor morphological and/or ecological data, and a lack of 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 (reviewed in Andrews et al. 2016). Hence, RADseq has recently seen adoption for phylogeny estimation (e.g. Cariou et al. 2013; Hipp et al. 2014; Eaton et al. 2017). However, reduced representation methods have been primarily utilized with freshly collected or well preserved tissue (but see Beck and Semple 2015, for an example of genotype-by-sequencing in *Solidago* (Asteraceae)).

The genus *Chloropyron* Behr (Orobanchaceae) is a rare and ecologically important clade of flowering plants with many historical populations that have been extirpated. This clade has proven to be challenging from a systematics perspective (Chuang and Heckard 1973; Tank and Olmstead 2008; Tank et al. 2009), and would benefit from the application of phylogenomic approaches aimed at investigating both inter- and intraspecific relationships. *Chloropyron* comprises four species (five subspecies, totaling seven taxa; Fig. 1) 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 et al. 2009). All taxa are listed from threatened to critically imperiled at the state level in part or all of their ranges, 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 their phylogenetic analyses of subtribe Castillejinae. This most recent treatment was based on two nuclear ribosomal DNA gene regions (internal and external transcribed spacers; ITS + ETS) and two chloroplast DNA regions (the *rps16* intron and the *trnL/F* region). While a monophyletic *Chloropyron* was recovered with high support, interspecific relationships were not fully resolved and there was considerable discordance between nuclear and plastid data; plastid data supported the sister relationships *C. maritimum* + *C. tecopense*, and *C. molle* + *C. palmatum*, while nuclear data only supported the sister relationship of *C. maritimum* + *C. molle*. These inconsistencies may be the results of insufficient taxon sampling, which we address here by including members of all species and subspecies of *Chloropyron*, with multiple accessions for most taxa.

Resolving both inter- and intraspecific relationships is nontrivial within *Chloropyron* due to differing base chromosome numbers, intraspecific variation in morphology and ecology, and complex distributions (Chuang and Heckard 1973; Tank et al. 2009). Most subspecies have allopatric distributions, with highly disjunct populations, but three of the four species are sympatric or parapatric throughout central California (Fig. 2). For example, the subspecies of the most widely distributed species, *Chloropyron maritimum*, are united by four functional stamens (as opposed to two throughout the rest of *Chloropyron*) and a gametic chromosome number of $n = 15$, but monophyly of *C. maritimum* is challenged by biogeographic patterns and ecology. *Chloropyron maritimum* ssp.

Fig. 1. Taxa of *Chloropyron*. A. *Chloropyron maritimum* ssp. *canescens*. B. *C. maritimum* ssp. *maritimum*. C. *C. maritimum* ssp. *palustre*. D. *C. molle* ssp. *hispidum*. E. *C. molle* ssp. *molle*. F. *C. palmatum*. G. *C. tecopense*. Photographs by Nicolas Diaz (A, B, G), Ian S. Gilman (C), and Thomas R. Stoughton (D, E, F).

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 does not accurately represent all putative lineages.

Here, we sample all taxa of *Chloropyron*, including multiple accessions representative of their respective ranges, from historical records dating back up to 25 yr, and use double digest restriction site-associated DNA sequencing (ddRADseq, Peterson et al. 2012) to investigate 1) intraspecific relationships between highly disjunct and ecologically divergent subspecies of *C. maritimum*, 2) interspecific relationships across *Chloropyron*, and 3) the applicability of RADseq (specifically ddRADseq) to historical specimens. We find all species of *Chloropyron* to be monophyletic, with the inland taxon *C. maritimum* ssp. *canescens* sister to the two coastal taxa of *C. maritimum* (ssp. *maritimum* + ssp. *palustre*), and the two distinct subspecies of *C. molle* to be each other's closest relative with strong support. We demonstrate the utility of reduced representation libraries to infer phylogenetic relationships 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 specimens, rare taxa, and reduced representation phylogenomic approaches.

Materials and Methods

Sampling, Library Preparation, and Sequencing—All accessions were sampled from herbarium vouchers dating back to 1983 (Appendix 1) and represent the range of taxa except *C. maritimum* ssp. *maritimum* (two individuals from one population) and *C. molle* ssp. *hispidum* (one individual), which are both rarely collected. Two members of the genus *Cordylanthus* were included to root gene and species trees. Genomic DNA was extracted using a modified CTAB protocol (Doyle and Doyle 1987) for degraded DNA. These modifications included the addition of 2 μl 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 (openwetware.org/wiki/SPRI bead mix), DNA extractions were visualized

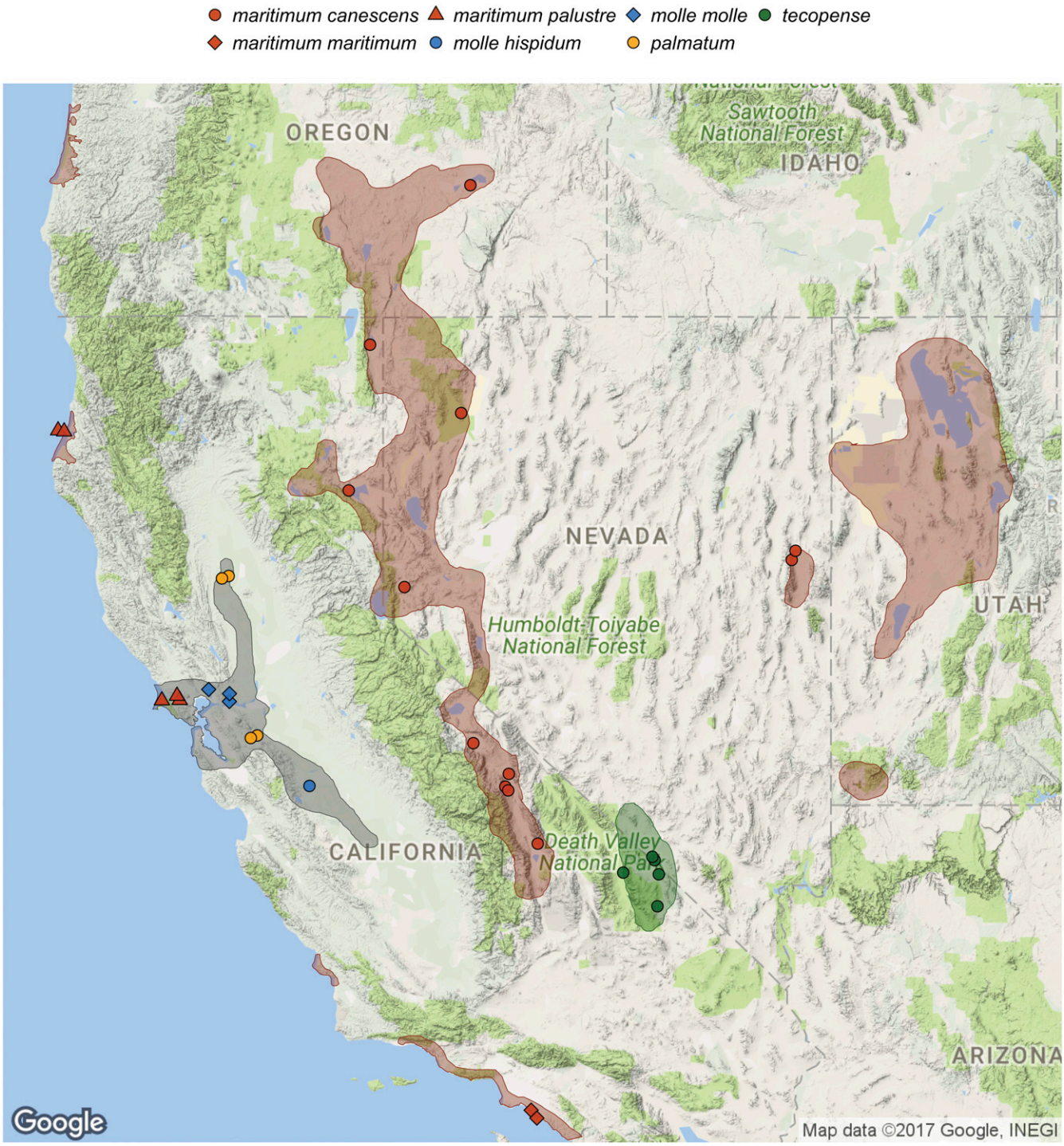

FIG. 2. Distribution and sampling of *Chloropyron*. Points represent the collection localities of specimens used in this study. Color of distribution corresponds to respective species points except where taxa occur sym- or parapatrically (grey).

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 6-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) controlling for ploidy. Libraries were barcoded using a 6 bp sequence (minimum distance of two 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 Institute for Bioinformatics and Evolutionary Studies (IBEST) Genomic Resources Core Facility with an expected 30 × coverage of 300 bp paired-end reads, including adaptor and barcode. Error rates were analyzed by eye in FastQC v. 0.11.5 (Babraham Bionformatics, Cambridge, United Kingdom). All molecular data have been deposited in Dryad (Gilman and Tank 2018).

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 (bases with phred quality scores less than 20 were changed to "N" and reads with more low quality sites were discarded), and barcodes, adaptors, and cut sites were removed. Trimmed sequences were then clustered by individual into "stacks" with a minimum of 85% similarity 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. Finally, alignments for all loci with at least 4 samples across all samples were generated.

Phylogenetics—To estimate phylogenetic relationships, we employed two classes of species tree methods based on the multi-species coalescent (MSC) that are appropriate for sequence data generated by the RADseq family of methods. SVDquartets (Chifman and Kubatko 2014) utilizes site pattern probability distributions at sites with SNPs to assemble quartets of taxa into a species tree. This method assumes each locus is unlinked and has its own genealogy drawn from the MSC. SVDquartets was called in PAUP* v. 4.0a.152 (Swofford 2002) with all possible quartets evaluated, 100 bootstrap replicates, and missing data treated as "missing" as opposed to "distributed." 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, first, 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 (i.e. AICc or BIC), the model with fewer parameters was chosen for downstream computational tractability. Gene trees were then constructed using GARLI v. 2.01.1067 (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 generated from historical samples. 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, has been shown to 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 (v. 4.10.11), which is tailored to datasets with multiple individuals per taxon, and statistical support was evaluated via 100 bootstrap replicates. Finally, all analyses across all datasets were rooted using two species of *Cordylanthus* (Orobanchaceae), *C. capitatus* and *C. eremicus* ssp. *eremicus* (Appendix 1).

Results

An average of 4456 ± 3906 loci were recovered per sample with an average coverage of 16.07 ± 15.45×. The number of loci per sample after quality filtering and removing paralogs and invariant loci was reduced to 96 ± 66. The final dataset comprised 5924 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 highly correlated with the number of accessions sampled per taxon (Fig. 3). 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 nine supergenes: seven of 66 loci and two of 65 (average size 28,311 ± 1351 bp).

All three phylogeny estimates yielded different topologies with inconsistent nodal support (Fig. 4). All species containing subspecies were recovered as monophyletic with high support in both the 'all loci' and 'supergene' ASTRAL-II phylogenies.

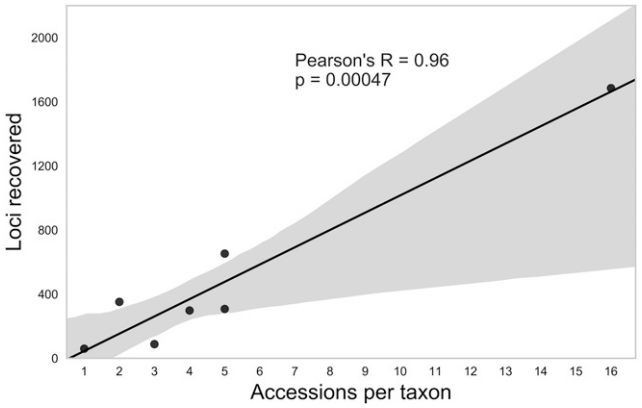

Fig. 3. Loci recovered per taxon by number of accessions sampled per taxon. Grey shading indicates 95% confidence interval.

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. maritimum* 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 necessarily in conflict, but the *Chloropyron* backbone had little resolution.

While SVDquartets also recovered *C. maritimum* as monophyletic (with the same intraspecific relationships highly supported), *Chloropyron molle* was not recovered as monophyletic, and support values were lower than either ASTRAL-II topology. Instead, *C. molle* ssp. *hispidum* was found sister to *C. palmatum* and *C. molle* ssp. *molle* sister to the rest of *Chloropyron*, albeit with low bootstrap support. The large contrast in support values for inter- and intraspecific relationships suggested that the number of informative quartets was much higher within species, rather than between them.

To test if these patterns of nodal support were biased by the lack of informative quartets linking multiple species, we calculated the number of loci shared between *N* or more i) taxa, ii) in-group taxa, iii) species, and iv) in-group species, as well as the number of informative quartets in those subsets as diagnosed by SVDquartets (Fig. 5; Table 1). 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, 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.

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

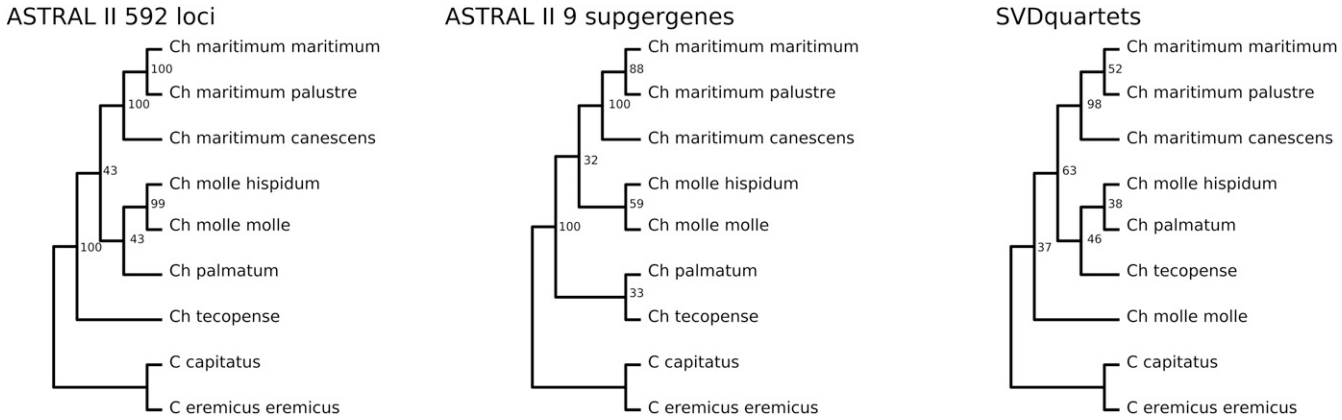

FIG. 4. Phylogeny of *Chloropyron* based on 592 loci in ASTRAL-II (left), 9 supergenes in ASTRAL-II (center), and all loci concatenated in SVDquartets (right). Node labels indicate bootstrap support values.

skewed. Some taxa, such as *C. maritimum* ssp. *maritimum*, were recovered in a small number of unique quartets that were supported by hundreds of sites distributed throughout all loci (suggesting consistent signal throughout the genome). In contrast, *C. molle* ssp. *molle* was recovered in just 10 unique quartets, each of which was supported by 10 or fewer sites throughout all loci (suggesting low concordance, but also reflective of the low number of recovered loci). Furthermore, *C. molle* ssp. *hispidum* and *C. molle* ssp. *molle* were only recovered together in three unique, and seven total, quartets, although neither the number of unique nor total quartets were significantly correlated with the number of accessions per taxon.

In contrast to SVDquartets, ASTRAL-II first induces all possible quartets from all gene trees, and then expands the set of search quartets using heuristic strategies; therefore, many more quartets were evaluated, although many of these did not inform the species tree. The average number of total quartets induced by all gene trees in 100 bootstrap replicates was 47,370 ± 2702. Each species tree resulting from a bootstrap replicate induced 2267 ± 295 quartets that were congruent with those induced by all gene trees (41.7 ± 1.3% of all species tree induced quartets, 4.8 ± 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 *C. molle* 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 possible quartets induced by supergene trees.

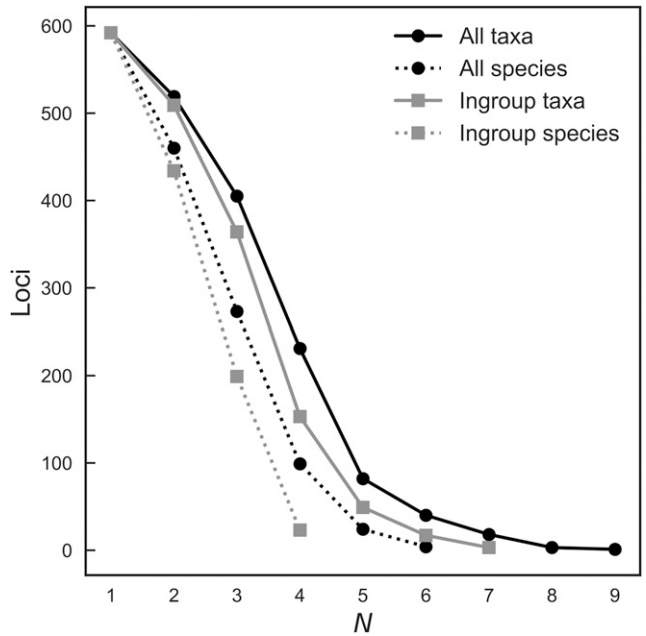

FIG. 5. The number of recovered loci shared among *N* or more taxa (solid, black circles), species (dashed, black circles), in-group taxa (solid, grey squares), and in-group species (dashed, grey squares).

TABLE 1. Number of loci, characters, and informative quartets in data subsets recovered by SVDquartets. Fraction of informative quartets is in reference to the total number of quartets across all sites (18,697).

	Loci	Characters	Informative quartets
Complete dataset	592	250,594	1492 (7.98%)
In-group species			
2+	434	183,588	1492 (7.98%)
3+	199	84,650	1441 (7.71%)
4	23	8142	296 (1.58%)
All species			
2+	460	195,347	1492 (7.87%)
3+	273	181,803	1486 (7.95%)
4+	99	43,463	1128 (6.03%)
5+	24	10,072	470 (2.51%)
6+	4	1616	20 (0.11%)
In-group taxa			
2+	509	213,909	1492 (7.98%)
3+	364	149,311	1492 (7.98%)
4+	153	59,739	1474 (7.88%)
5+	39	18,529	914 (4.89%)
6+	17	6396	274 (1.47%)
7	3	1310	13 (0.07%)
All taxa			
2+	519	218,671	1492 (7.98%)
3+	405	168,557	1492 (7.98%)
4+	231	94,927	1492 (7.98%)
5+	82	33,569	1397 (5.13%)
6+	40	16,279	959 (5.13%)
7+	18	7623	196 (1.05%)
8+	3	1313	9 (0.05%)
9	1	458	9 (0.05%)

Discussion

The boundaries of the applications of reduced representation sequencing are continuing to expand as genome-scale data becomes easier and cheaper to generate per basepair per taxon. 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. DNA degradation has been shown to significantly decrease the utility of RADseq in population genetic studies (Graham et al. 2015) but, to our knowledge, there is no corresponding research that quantifies the decrease in phylogenetic information. 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 among 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. 2017). For example, if patterns of missing data are hierarchical, such as those that would result from mutation-disruption or mutation-generation, the redundancy in loci/quartets may be significantly decreased, along with the accuracy of species tree estimation (Rubin et al. 2012).

Building off of these recent works that examined the effects of missing data in studies that employed reduced representation and/or quartet-based species tree methods, here we report the absolute number of gene tree induced (GTI) and species tree induced (STI) quartets, as well as the overlap between the two sets for ASTRAL-II. Of particular interest is the overlap between GTI and STI quartets. ASTRAL-II's bootstrapping method operates by randomly resampling bootstrap replicate gene trees with replacement for every gene included in the analysis. Therefore, for any one gene tree, if bootstrap replicates are highly congruent, the set of GTI quartets will remain relatively constant. In contrast, if replicates are divergent in topology, replicate gene trees will produce different GTI quartets, thus expanding the total number of GTI quartets. Hence, the variance in GTI quartets over all ASTRAL bootstrap replicates may reflect uncertainty in gene tree estimation. Furthermore, because STI quartets will always be a subset of GTI quartets, the percent of STI quartets in common with GTI quartets will be maximized when every gene tree is exactly congruent with the species tree. As the number of gene trees that differ from the species tree, and the degree to which they differ, increases, the pool of GTI quartets increases, and the fraction of GTI quartets in common with STI quartets decreases. Therefore, we propose four possible scenarios: 1) gene trees well estimated and convergent (with each other and the species tree) implies high overlap between STI and GTI quartets, with low variance in the overlap between bootstrap replicates, 2) gene trees well estimated but divergent implies low overlap between STI and GTI quartets, with low variance in the overlap between replicates, 3) gene trees poorly estimated, yet still convergent, implies moderate to high overlap between STI and GTI quartets with high variance in the overlap between replicates, and 4) gene trees poorly estimated and divergent implies low overlap between STI and GTI quartets with high variance in the overlap between replicates. The degree to which STI-GTI quartet overlap truly reflects these four scenarios is, to our knowledge, unknown and in need of investigation through simulation and reexamination of empirical datasets that have employed quartet-based species tree inference.

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 throughout its range. 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 our resulting species trees 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) and bolsters morphological and cytological evidence gathered over the previous half century. Despite widely disjunct populations, *Chloropyron maritimum* is distinguished morphologically by four functional stamens and entire to minutely-lobed, glabrous-pubescent floral bracts (Figs. 1, 2). 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 base chromosome number of 15, which varies throughout the rest of *Chloropyron* (*C. molle*, $n = 14$; *C. palmatum*, $n = 21$; *C. tecopense*, $n = 14$). Although Chuang and Heckard (1986) hypothesized that extant members of *Chloropyron* arose from an $n = 7$, diploid ancestor, more recent work by Tank (2006) has shown that the broader subtribe of Castillejinae (Orobanchaceae) has a base chromosome number of $n = 14$ ancestrally. This implies two dysploid events leading to *C. maritimum* ($n = 15$) and *C. palmatum* ($n = 21$).

These synapomorphies stand in stark contrast to the divergent ecologies and morphologies 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 (Gilman, pers. obs.) via root haustoria, and host-specificity is rare throughout hemiparasitic Orobanchaceae. However, *Chloropyron maritimum* ssp. *canescens* has only been found to parasitize *Distichlis spicata* (Poaceae), and occurs 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. Chuang 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. Although both SVDquartets and ASTRAL-II are consistent under the MSC, and therefore should converge on the same topology given enough data, the way that they utilize sequence data is very different, which may have consequences when the amount of missing data is high. In particular, our use of SVDquartets treats missing data from sites as missing, not ambiguous, and therefore will not attempt to resolve quartets from taxa with missing data. In contrast, ASTRAL-II does not directly "see" these missing data directly, as missing data only affect the resolution of gene trees, which is an indirect way to "view" the missing data.

The number of quartets constructed in SVDquartets, and therefore the power to estimate the species tree, is low. The version of ASTRAL-II used for this study 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. Furthermore, while binning loci into supergenes can increase phylogenetic signal (Bayzid and Warnow 2013; Mirarab et al. 2014a), it may reduce the number of quartets induced by supergene trees. Ongoing work investigating the behavior of quartet based species tree methods via simulations suggests that, at high levels of missing data, ASTRAL-II is significantly more accurate than SVDquartets (Gilman 2017). Therefore, 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 the four functional stamens in *C. maritimum*, but by a suite of traits. *Chloropyron 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 *Distichlis spicata* in central California. The range of *C. molle* overlaps slightly with *C. maritimum* and *C. palmatum*, but there is no evidence of hybridization. *Chloropyron molle* is distinguished from *C. maritimum* by its lobed floral bracts, the length and density of hair throughout the plants, and base chromosome number. Subspecies of *C. molle* are primarily differentiated along ecological and geographical lines: *C. molle* ssp. *molle* is endemic to alluvial salt marsh habitats similar to those of the coastal *C. maritimum* subspecies, whereas *C. molle* ssp. *hispidum* occurs in dry alkali flats.

Finally, the inconsistent placement of *C. palmatum* and *C. tecopense* within *Chloropyron* adds to a complex set of ecologies, morphologies, and cytotypes. *Chloropyron palmatum* is morphologically and ecologically very similar to *C. molle*, which have been found within meters of one another (Gilman pers. obs.), 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 appression of its pubescent leaves, bracts, and branches (Fig. 1). This taxon is endemic to the alkali flats of Death Valley and the Mojave Desert of California and adjacent Nevada, similar in habitat to the inland distribution of *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, including between sources and types of genetic information. The only topology presented here that is consistent with previous molecular phylogenies is the ASTRAL-II analysis using individual loci. These relationships are consistent with the plastid data of Tank and Olmstead (2008), but that phylogeny only supports *C. maritimum* and *C. tecopense* as sister, while all other relationships form a polytomy. Genomic data may still prove to be a useful tool with increased, and equal, sequencing effort. Although perhaps counterintuitive from a systematics/macroevolutionary perspective, Eaton et al. (2017) showed a 10-fold increase in shared loci when doubling sequence effort (i.e. resequencing). 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. In addition, modifications to the original ddRAD protocol (Peterson et al. 2012) not performed here can reduce biases in sequencing depth. For example, amplifying barcoded samples prior to pooling, instead of amplification post-pooling, can reduce PCR, and downstream sequencing, bias. Future work including DNA isolated from field-collected specimens may lead to a more robust dataset with more loci spanning more taxa, which would allow the creation of a highly supported phylogeny of this small group of extremophiles. However, despite limitations with the current dataset, we have been able to confirm the monophyly of several highly disjunct taxa of *Chloropyron* and set the stage for detailed biogeographic analyses investigating historical connectivity of saline and alkaline habitats of western North America.

Acknowledgments

The authors thank the curators and directors of the following herbaria for providing herbarium specimens and sampling permission: JEPS, RSA, UC, and ID. We also thank Nicolas A. Diaz and Thomas R. Stoughton for field assistance, Jack Sullivan for helpful discussion, members of the Donoghue and Near Labs at Yale University for suggestions to improve ddRADseq techniques for phylogenomics, and two anonymous reviewers for insightful criticism on the manuscript. This research was partially supported by the American Society of Plant Taxonomist's Rogers McVaughn Graduate Student Award, the Botanical Society of America's Graduate Student Research Award to ISG and the National Science Foundation grant DEB-1253463 to DCT, and represents a portion of the M.S. thesis of Ian S. Gilman.

Author Contributions

ISG and DCT conceived of the project; ISG performed experiments and analyzed resulting data; ISG wrote the first draft of the manuscript, and ISG and DCT both contributed to revisions.

Literature Cited

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews. Genetics* 17: 81–92.

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.

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

Beck, J. B. and J. C. Semple. 2015. Next-generation sampling: Pairing genomics with herbarium specimens provides species-level signal in *Solidago* (Asteraceae). *Applications in Plant Sciences* 3: 1500014.

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.

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* 23: 169–173.

Chuang, T. I. and L. R. Heckard. 1975b. Taxonomic status of *Cordylanthus* (subg. Dicranostegia) *orcuttianus* (Scrophulariaceae). *Madroño* 23: 88–95.

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: 1–105.

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.

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. 2017. Misconceptions on missing data in rad-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology* 66: 399–412.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

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

Gilman, I. S. 2017. Systematics of *Chloropyron* (Orobanchaceae): Implications of missing data on quartet-based species tree methods. Master's thesis. Moscow, Idaho: University of Idaho.

Gilman, I. S. and D. C. Tank. 2018. Data from: Species tree estimation using ddRADseq data from historical specimens confirms the monophyly of highly disjunct species of *Chloropyron* (Orobanchaceae). Dryad Digital Repository. doi:10.5061/dryad.6j58m7g.

Graham, C. F., T. C. Glenn, A. G. McArthur, D. R. Boreham, T. Kieran, S. Lance, R. G. Manzon, J. A. Martino, T. Pierson, S. M. Rogers, J. Y. Wilson, and C. M. Somers. 2015. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources* 15: 1304–1315.

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* 15: 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.

Jepson, H. L. 1925. *A Manual of the Flowering Plants of California*. Berkeley and Los Angeles: University of California Press.

Mason, H. L. 1957. *A Flora of the Marshes of California.* Berkeley and Los Angeles: University of California Press.

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.

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.

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.

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

Pennell, F. W. 1951. Scrophulariaceae. In *Illustrated Flora of the Pacific States*, ed. L. Abrams. Palo Alto: Stanford University Press.

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.

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.

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.

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.

Swofford, D. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods), v. 4.0 b10. Sunderland: Sinauer Associates.

Tank, D. C. 2006. *Molecular phylogenetics of* Castilleja *and Castillejinae (Orobanchaceae).* Ph.D. dissertation. Seattle, Washington: University of Washington.

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

Tank, D. C., J. M. Egger, and R. G. Olmstead. 2009. Phylogenetic classification of subtribe Castillejinae (Orobanchaceae). *Systematic Botany* 34: 182–197.

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. *Generic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion*. Ph.D. dissertation. Austin: University of Texas at Austin.

Appendix 1. List of species and vouchers for all specimens used in this study, arranged by taxon and collection year (* denotes accession sampled from UC herbarium while on loan). ID = University of Idaho; JEPS = University of California Jepson Herbarium; UC = University of California, Berkeley; RSA = Rancho Santa Ana Botanic Garden; RENO-V = University of Nevada Reno.

Chloropyron maritimum (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Schoolcraft 2112* (UC), 1990; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Wilson 6288* (UC), 1993; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Tiehm 12138* (RENO-V)*, 1995; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Pinzl 12491* (UC), 1997; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Tiehm 12253* (RSA), 1997; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Tiehm 12643* (UC), 1998; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Tiehm 13336a* (RSA), 2000; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Tiehm 13336b* (ID), 2000; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Tiehm 14063a* (RSA), 2002; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Tiehm 14063b* (RSA), 2002; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Riefner 04-468* (RSA), 2004; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *La Doux 115* (RSA), 2005; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Andre 10285* (RSA), 2006; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *canescens* (A. Gray) Tank & J.M. Egger: *Andre 20334* (RSA), 2011; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *maritimum*: *Fraga 4143a* (RSA), 2012; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *maritimum*: *Fraga 4143b* (RSA), 2012; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *palustre* (Behr) Tank & J.M. Egger: *Chuang 7808* (JEPS), 1990; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *palustre* (Behr) Tank & J.M. Egger: *Ruygt & Collins* (JEPS), 1993; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *palustre* (Behr) Tank & J.M. Egger: *Wetherwax 2460* (JEPS), 1993; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *palustre* (Behr) Tank & J.M. Egger: *Wetherwax 2462* (JEPS), 1993; *Chloropyron maritimum* (Nutt. Ex Benth) A. Heller ssp. *palustre* (Behr) Tank & J.M. Egger: *Slackly 57* (JEPS), 2013; *Chloropyron molle* (A. Gray) A. Heller ssp. *palustre* (Pennell) Tank & J.M. Egger: *Heckard 6740* (JEPS), 1990; *Chloropyron molle* (A. Gray) A. Heller ssp. *molle*: *Ruygt 1* (JEPS), 1993; *Chloropyron molle* (A. Gray) A. Heller ssp. *molle*: *Ruygt 2* (JEPS), 1994; *Chloropyron molle* (A. Gray) A. Heller ssp. *molle*: *Ruygt 3* (JEPS), 1994; *Chloropyron palmatum* (Ferris) Tank & J.M. Egger: *Heckard 6145* (JEPS), 1983; *Chloropyron palmatum* (Ferris) Tank & J.M. Egger: *Cypher 2004-018* (RSA), 2004; *Chloropyron palmatum* (Ferris) Tank & J.M. Egger: *Cypher 2005-022* (RSA), 2005; *Chloropyron tecopense* (Munz & J.C. Roos) Tank & J.M. Egger: *Andre 9826* (RSA), 2007; *Chloropyron tecopense* (Munz & J.C. Roos) Tank & J.M. Egger: *Fraga 3892* (RSA), 2011; *Chloropyron tecopense* (Munz & J.C. Roos) Tank & J.M. Egger: *Fraga 3870* (RSA), 2011; *Chloropyron tecopense* (Munz & J.C. Roos) Tank & J.M. Egger: *Andre 9701* (RSA), 2008; *Chloropyron tecopense* (Munz & J.C. Roos) Tank & J.M. Egger: *Fraga 3769* (RSA), 2010; *Cordylanthus capitatus* Nutt. Ex Benth.: *Ertter 20380* (UC), 2010; *Cordylanthus eremicus* (Corville & C.V. Morton) Munz ssp. *eremicus*: *Fraga 933* (RSA), 2003.