Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L.\(^1\)

Nora Mitchell\(^{2,5}\), Paul O. Lewis\(^2\), Emily Moriarty Lemmon\(^3\), Alan R. Lemmon\(^4\), and Kent E. Holsinger\(^2\)

**PREMISE OF THE STUDY:** Estimating phylogenetic relationships in relatively recent evolutionary radiations is challenging, especially if short branches associated with recent divergence result in multiple gene tree histories. We combine anchored enrichment next-generation sequencing with species tree analyses to produce a robust estimate of phylogenetic relationships in the genus *Protea* (Proteaceae), an iconic radiation in South Africa.

**METHODS:** We sampled multiple individuals within 59 out of 112 species of *Protea* and 6 outgroup species for a total of 163 individuals, and obtained sequences for 498 low-copy, orthologous nuclear loci using anchored phylogenomics. We compare several approaches for building species trees, and explore gene tree–species tree discrepancies to determine whether poor phylogenetic resolution reflects a lack of informative sites, incomplete lineage sorting, or hybridization.

**KEY RESULTS:** Phylogenetic estimates from species tree approaches are similar to one another and recover previously well-supported clades within *Protea*, in addition to providing well-supported phylogenetic hypotheses for previously poorly resolved intrageneric relationships. Individual gene trees are markedly different from one another and from species trees. Nonetheless, analyses indicate that differences among gene trees occur primarily concerning clades supported by short branches.

**CONCLUSIONS:** Species tree methods using hundreds of nuclear loci provided strong support for many previously unresolved relationships in the radiation of the genus *Protea*. In cases where support for particular relationships remains low, these appear to arise from few informative sites and lack of information rather than strongly supported disagreement among gene trees.

**KEY WORDS** anchored phylogenomics; coalescence; phylogenetics; Proteaceae; radiation

Evolutionary radiations, which are typically associated with rapid bursts of diversification into many species and morphological forms, provide ideal systems for studying evolution. They can be found at both deep and shallow taxonomic levels, from the origin and explosion in diversity of all flowering plants (Crepet, 2000) to individual families (e.g., Restionaceae; Linder, Eldenas, and Briggs, 2003), groups within families (e.g., Hawaiian silversword alliance; Baldwin and Sanderson, 1998), genera (e.g., *Pelargonium*; Bakker et al., 2004), and even subclades (e.g., the white proteas; Prunier and Holsinger, 2010). Robust estimates of relationships among taxa are a prerequisite for studying the morphological, ecological, and often cytological diversity in radiating groups and understanding trait evolution. However, the rapid evolution that makes these systems so interesting also makes it difficult to build well-resolved phylogenies (Knowles and Chan, 2008). Rapid radiation leads to many short branches with few nucleotide differences reflecting shared ancestry. Gene trees may not reflect the same history as the species tree because of incomplete lineage sorting (ILS), where alleles coalesce prior to the splitting of species, gene duplication, or loss, or because of hybridization (Maddison, 1997). Empirical data sets in many systems have found evidence for discordance among nuclear loci (reviewed in Degnan and Rosenberg, 2009), and in particular, several studies have highlighted the high frequency of gene tree discordance in more recent radiations (Knowles, 2009;
show substantial diversity in leaf shape and size. Species differ in many functional traits, and several are correlated with important environmental variables such as seasonality and mean annual precipitation and temperature (Mitchell et al., 2015). Common-garden experiments have demonstrated both inter- and intraspecific adaptive differences in physiological and functional traits (Carlson et al., 2011; Prunier et al., 2012; Carlson et al., 2015).

Previous phylogenetic analyses of this genus used only a few molecular markers: the nuclear ribosomal DNA region ITS, a set of plastid noncoding regions, the nuclear gene ncpGS, and 138 AFLP loci (Valente et al., 2010; Schnitzler et al., 2011; the latter species tree approach used the same data without AFLPs and included additional taxa). Although these analyses identified a few well-supported clades, many relationships were poorly resolved. In addition, some groups that have long been recognized based on morphological characters (Rourke, 1982; Rebelo, 2001) are not supported in the published phylogenies. For example, P. laurifolia and P. nerifolia are morphologically very similar and replace one another geographically, yet molecular phylogenies suggest they are distant relatives (Valente et al., 2010; Schnitzler et al., 2011). Hybridization is also known to occur in *Protea* (Prunier and Holsinger, 2010), possibly contributing to regions of the tree with low support (Valente et al., 2010). To build a more-strongly supported phylogeny for *Protea* as a basis for future analyses of trait evolution, we collected samples from multiple populations throughout South Africa and used targeted sequencing techniques to greatly increase the number of DNA sequence markers available for phylogenetic inference.

We aim to (1) resolve relationships within the rapid, recent radiation of *Protea*; (2) compare widely used concatenation and species-tree approaches; and (3) explore the causes of differences in gene tree and species tree topologies.

**MATERIALS AND METHODS**

**Taxon sampling**—We collected DNA samples from fresh leaf tissue in the field from 2011–2014 using locations from the Protea Atlas database (http://www.proteaatlas.org.za/), live accessions at Kirstenbosch Botanical Gardens, and greenhouse-grown individuals derived from wild-collected seed (Prunier et al., 2012 Fig. 1; see Appendix S1 for a full list of species and voucher information in the Supplemental Data with this article). In all, our initial data set includes samples from 163 individuals collected from 65 species, including 6 outgroup species (*Serruria* and *Faurea*) and 59 *Protea* species (Appendix 1); DNA was collected into a concentrated CTAB solution (Doyle and Doyle, 1987).

These samples represent a substantial fraction of species in *Protea*: over half of the total species, and approximately 70% of the species found in South Africa (Rourke, 1982), but our own sampling is insufficient to build the most complete phylogeny possible with available DNA data. To supplement our samples, we used sequence data from Valente et al. (2010) and Schnitzler et al. (2011) as downloaded from TreeBase.org (S11312). The Schnitzler data set includes 32 additional species of *Protea* and 7 outgroup taxa with sequence information from 4199 additional bases in plastid noncoding regions, as well as the nuclear genes ITS and ncpGS. This AUGMENTED data set thus uses the sequence data from Schnitzler et al. (2011), but builds on trees constructed using our samples. We used this data set to construct a phylogenetic estimate for *Protea* that includes 91 of the 112 extant taxa plus 13 outgroup taxa (see Results section).
identifi ed as important in selenium tolerance as in Buddenhagen et al. (2016) using the general methods of Lemmon et al. (2012). Loci complications due to whole-genome duplication. Sequences for an-

numbers are not identical to Buddenhagen et al. (2016), but (2016) using the general methods of Doyle and Doyle, 1987). Further molecular work was conducted at the Center for Anchored Phylogenomics (www.anchoredphylogeny.com) at Florida State University. The lack of polyploidy in Protea (Oberlander et al., 2016) facilitates the assembly of anchored phylogenomics data and avoids the possibility of complications due to whole-genome duplication. Sequences for anchored nuclear loci were obtained from 498 low-copy orthologous regions identifi ed and designed across angiosperms, 482 based on the orthologs of Duarte et al. (2010), and an additional 16 genes identifi ed as important in selenium tolerance as in Buddenhagen et al. (2016) using the general methods of Lemmon et al. (2012). Loci numbers are not identical to Buddenhagen et al. (2016), but Arabidopsis gene identifi ers match (Appendix S2). In short, extracted DNA was sonicated via a Covaris E220 Focused-ultrasonicator (Woburn, MA) to obtain 300–800 bp fragments. Libraries were prepared and indexed on a liquid-handling robot (Beckman-Coulter Biomek FXp, Brea, California) using the protocol of Meyer and Kircher (2010). One modifi cation of the protocol included a size-selection step, removing fragments <200 bp in length, after blunt-end repair using Solid Phase Reversible Immobilization (SPRI) select beads (Beckman-Coulter). After indexing, samples were pooled in equal quantities (16–18 samples per pool), and each pool was enriched using an Agilent Custom SureSelect kit (Agilent Technologies, Lexington, Massachusetts). Enrichment pools were run in equal quantities for sequencing on replicate PE150 Illumina HiSeq2500 lanes (typically three pools, which included ~48 samples per lane). Sequencing was performed in the Translational Science Laboratory in the College of Medicine at Florida State University.

**Bioinformatics processing**—Reads were processed following Prum et al. (2015), Ruane et al. (2015), and Pyron et al. (2016). For paired-read merging, the probability of overlapping to a given degree by chance was calculated, and read pairs with signifi cant matches were merged (see Rokyta et al., 2012). Base-specific quality scores were used to reconcile differences and were combined to produce quality scores for the merged reads. For the target regions, divergent reference assembly was used to map reads to the probe region sequences for Arabidopsis thaliana, Aquilegia coerulea, and Nelumbo nucifera, and de novo assembly was then used to extend these to flanking regions (see Prum et al., 2015 for details). A coverage filter removed low-coverage contigs (<20 reads) to remove reads from potential cross-contamination. To assess putative orthology among consensus sequences at each locus, pairwise distances between two sequences were computed using the percent of 20-mers found in common between each pair of sequences. A Neighbor-Joining algorithm was then used to cluster sequences based on these pairwise distance measures (see Prum et al., 2015 for details). Two alleles were phased per consensus sequence following Pyron et al. (2016), using a Bayesian approach that estimates the posterior distribution of phasing solutions from assembled reads.

Sequences in each orthologous set were aligned using MAFFT v7.023b (Katoh and Standley, 2013). Alignments were trimmed by identifying “good sites” (sites where the most common state was present in >50% of the sequences), masking 20 bp regions that contained <14 good sites, and removing sites with <240 unmasked bases. After the pipeline of filtering, orthology, trimming, and masking, the sequence data consisted of both target regions and variable flanks for 498 target loci.

**Phylogenetic analysis**—Our fi rst goal was to build species-level phylogenies for Protea using two concatenation and two species tree methods, as well as gene trees for individual loci (Table 1). Figures for trees were created using TreeGraph2 (Stöver and Müller, 2010).

We used four different sets of data derived from the raw sequence data in our analyses (Table 2). The complete set (hereafter referred to as the COMPLETE data set) includes 163 individuals (from 59 species of Protea and 6 outgroup species) and sequences from both alleles for up to 498 loci. Not all loci were captured for all taxa; in this data set, nucleotides at these loci were coded as missing values (Appendix S1). Analysis of this data set allows us to assess monophyly of most species-level taxa in our data set, and can also be used in multi-individual modes in ASTRAL-II and SVDquartets. To reduce the computational burden for other analyses and to build species-level...
TABLE 1. Summary of tree-building methods. “Gene Trees” indicates that all samples were used as terminals, “Full Trees” means that species were used as terminals.

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Output</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene tree building</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAxML</td>
<td>Sequence</td>
<td>Gene Trees</td>
<td>Concatenated</td>
</tr>
<tr>
<td>Species tree building</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAxML</td>
<td>Sequence</td>
<td>Full Tree</td>
<td>Concatenated</td>
</tr>
<tr>
<td>MrBayes</td>
<td>Sequence</td>
<td>Full Tree</td>
<td>Concatenated</td>
</tr>
<tr>
<td>SVDOquartets</td>
<td>Sequence</td>
<td>Full Tree</td>
<td>Species Tree</td>
</tr>
<tr>
<td>ASTRAL-II</td>
<td>Gene Trees</td>
<td>Full Tree</td>
<td></td>
</tr>
</tbody>
</table>

We also saved the best maximum likelihood gene trees from the ONEPER data set and analyses.

Six individuals (one Protopsidae, one P. nubigena, one P. reconspita, and three Serruria samples) failed to recover many loci, and likewise, over a quarter of the loci were not recovered for many individuals (Appendix S1). We used RAxML for making tree comparisons to be consistent with the sequence data used. Taking species-level consensus sequences may mask information, especially if species are not monophyletic. To check this, we also created a ONEPER data set that contains one arbitrarily selected sequence per species, rather than a consensus across all sequences for each species (see Appendix S3 for details on the ONEPER data set and analyses).

Individual gene trees for all 498 loci in the COMPLETE and ONEPER data sets were obtained in RAxML v8.3.17 (Stamatakis, 2014) using a GTR_GAMMA model and 100 bootstrap replicates. For each locus, we saved 100 bootstrap replicates for each of the gene trees and used them in the subsequent ASTRAL-II analysis. We also saved the best maximum likelihood gene trees from the COMPLETE analysis and used these to compute distances and internode certainty (IC) values on our species trees in RAxML (Salichos et al., 2014). Internode certainty takes into account the frequency of the most common taxon bipartition in comparison to the most observed conflicting bipartition.

Analyses of concatenated data were conducted in RAxML v8.3.17 (Stamatakis, 2014), also using a GTRGAMMA model and 100 bootstrap replicates. Analyses were conducted on the CONSENSUS and ONEPER data sets (all 274,405 bp from the 498 loci as one sequence per sample, 65 species total, with separate partitions for each locus) to obtain a species tree estimated using concatenation. We did not partition by codon position because identifying coding regions and codon position is difficult for this type of data set. To check for species monophyly, we ran RAxML over the COMPLETE data set with the same settings. We also conducted a concatenated, unpartitioned Bayesian analysis using the CONSENSUS and ONEPER sequences in MrBayes v3.2.1 (Ronquist and Huelsenbeck, 2003) using the GTR +I +G model with four chains for 5 million generations, thinned to save one sample every 1000 generations. Parameters were visually checked in Tracer to confirm convergence, and a consensus tree (plus other compatible groupings) was computed in PAUP* after a 500 tree burn-in and used as a species tree for further analyses. We did not use MrBayes to check for species monophyly because of the size of the data set.

ASTRAL-II (Mirarab et al., 2014, Mirarab and Warnow, 2015) estimates a species tree from input gene trees, and has been shown to be statistically consistent under the multispecies coalescent model. ASTRAL-II finds the species tree that maximizes the number of embedded quartet trees in the given gene trees; it works efficiently by limiting the number of bipartitions explored to those included in the supplied gene trees. Additionally, ASTRAL-II is capable of taking information from bootstrap replicates of these gene trees, as well as including multiple individuals per species. We employed the bootstrapping method in ASTRAL-II v4.7.9 to estimate a species tree using this coalescent-based approach, as well as the multi-individual feature using the /multiind branch of the ASTRAL-II GitHub repository (https://github.com/smirarab/ ASTRAL/tree/multiind). This option allows for a species-level estimation rather than building a tree with multiple accessions per species. Best trees and bootstrap replicates were estimated in RAxML separately for each locus in the COMPLETE data set. These best trees, bootstrap files, and a species-to-allele file were provided for each locus and run for 100 bootstrap replicates. We also ran ASTRAL-II using the best trees and bootstrap files from the ONEPER and CONSENSUS data sets to obtain species trees and the COMPLETE best trees and bootstrap files to check for species monophyly.

SVDOquartets (Chifman and Kubatko, 2014) is a recent quartet-based species tree method that is robust to ILS given data that is reasonably clock-like. This method treats each single nucleotide polymorphism (SNP) as an independent sample from a species tree with a coalescent history within species. It produces a species tree estimate, rather than estimates of individual gene trees. We performed the SVDOquartets analysis on the COMPLETE data set in the test version of PAUP* 4.0a146 (Swofford, 2003) using the QFM matrix agglomeration method (Reaz et al., 2014). We used the multispecies coalescent approach with the species-membership partition, searching 1 million quartets, and did a bootstrap analysis of 100 replicates. For the CONSENSUS and ONEPER data sets we used the same settings, but did not include the species-membership partition scheme. To check for species monophyly, we searched a reduced set of 10,000 quartets because of the increased number of tips in the tree. Bootstrap trees for each data set were saved and an SVDOquartets consensus tree was computed in PAUP*.

TABLE 2. Summary of data sets used. Note that not all loci are represented by all species in the data set except in the case of the REDUCED set. Tips indicate the terminals in the tree: sequences were used either associated with alleles, individuals, or species (see brief description).

<table>
<thead>
<tr>
<th>Data set</th>
<th># Loci</th>
<th># Tips</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPLETE</td>
<td>498</td>
<td>163</td>
<td>All alleles and individuals</td>
</tr>
<tr>
<td>CONSENSUS</td>
<td>498</td>
<td>65</td>
<td>Species-level consensus</td>
</tr>
<tr>
<td>ONEPER</td>
<td>498</td>
<td>65</td>
<td>One individual selected per species</td>
</tr>
<tr>
<td>REDUCED</td>
<td>354</td>
<td>60</td>
<td>All loci found in all CONSENSUS species</td>
</tr>
<tr>
<td>AUGMENTED</td>
<td>3</td>
<td>99</td>
<td>Backbone from this study, sequences from</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Schnitzler et al. (2011)</td>
</tr>
</tbody>
</table>
Species topologies as constraint trees—SVDquartets and ASTRAL-II produce species-level phylogenies that take into account multiple individuals per species. To estimate branch lengths, the 80% majority rule topologies from each of the ASTRAL-II and SVDquartets trees were input as constraint trees and run in RAxML with the CONSENSUS sequence data. The topology from the best RAxML concatenated tree has maximum-likelihood lengths associated with each branch.

We used the species trees as backbones on which to place the 39 additional species included in Schnitzler et al. (2011), the AUGMENTED data set. We did this by removing species for which Schnitzler et al. (2011) had no data and using the remainder to construct 80% majority-rule consensus trees from the bootstrap replicates for the SVDquartets, RAxML, and ASTRAL-II trees. For the MrBayes species tree, we calculated the 80% SVDquartets majority-rule tree after a 500-tree burn-in in the second run of our analysis, which converged more quickly. We then used these backbones as constraint trees in RAxML for analyses of the concatenated sequence data (4199bp) from Schnitzler et al. (2011) under the GTRGAMMA model. This method may be problematic where our data and the sequences from Schnitzler et al. (2011) suggest different topologies, but it can provide a rough estimate of placement of additional species for which anchored phylogenomics data are not available.

Relevant alignment and tree files are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.vj32s. Raw sequence reads are deposited in the NCBI SRA BioProject ID PRJNA354967, SRA study SRP093931.

Hypothesis testing: poor support due to few changes or ILS—We used the REDUCED data set to estimate distances between gene and species trees. Each of the 354 individual gene alignments was run in RAxML (using the above settings) to obtain the best and bootstrapped gene trees. We also generated 354 random topologies in PAUP*, using the proportional-to-distinguishable model, to compare the distribution of gene tree distances to those of randomly generated trees, and we simulated 354 gene trees from the ASTRAL-II species tree in the program COAL, which is used for computing gene tree distributions (Degnan and Salter, 2005), using branch lengths of one (branch lengths are equal to the number of generations / (2 * the effective population size)). Branch lengths of 0.5 and 0.2 did not significantly affect simulated tree topologies.

To determine which trees were most similar and which may be most reliable, we compared all tree topologies using an adjusted Robinson-Foulds (RF) distance. We calculated raw distances (RF) using RF.dist() from phangorn (Schliep, 2011) in R v3.1.3 (R Core Team, 2015) and adjusted RF distances as RFadj = RF/(2n−6) where n is the number of nodes on the tree (Steel and Penny, 1993). The RFadj values can range from zero (topologically identical) to one (completely dissimilar). We calculated several sets of distances: (1) among individual RAxML gene trees, (2) among random trees, (3) among trees simulated from the coalescent, (4) between individual gene trees and species trees produced via concatenation or species tree building methods, and (5) among the species trees. We compared the distributions of RFadj between different sets of trees by performing two-tailed T-tests in R. We adjusted the sample size in our T-tests to be more conservative, using only the number of trees compared, not the number of pairwise comparisons, which are nonindependent.

Low levels of support at any particular branch of a species tree could reflect either short branches or a balance between strongly supported but conflicting gene histories (i.e., ILS). To distinguish between these possibilities in clades of particular interest, we took the existing, fully bifurcating ASTRAL-II tree topology and constructed trees with alternative resolutions to match the different species trees from the COMPLETE analysis (“A = ASTRAL-II”, “B = SVDquartets”, “C = RAxML/MrBayes”) at only branches of interest. We then measured distances from both the RAxML “best” gene trees and the bootstrap replicates for each of those gene trees to the “A”, “B”, or “C” species tree topologies. We call these ad hoc tests “alternative placement tests”. If poor support is the result of a small number of changes, we expect that the distances from gene trees to each species tree will be sampled from a single underlying distribution. If poor support is the result of balance between strongly supported conflicting topologies, one set of genes will have a shorter distance to one topology, and another set will have a shorter distance to an alternate topology. We are particularly interested in the relative placement of P. repens, because it is the most widespread South African endemic in Protea, and much recent work has focused on intraspecific variation and local adaptation in this species at the morphological, physiologically, genomic, and transcriptomic levels (Akman et al., 2015; Carlson et al., 2015; Prunier et al., personal communication). In our species trees, P. repens is sometimes sister to P. rupicola, and other times has a more complicated relationship, leading us to focus on the placement of these two species as a case study.

Hypothesis testing: poor support due to hybridization—If hybridization has caused a lack of confidence in relationships in the phylogeny of Protea, we would expect to see evidence of reticulation in areas of the tree associated with low bootstrap support values. We built a phylogenetic network to visually identify regions of the Protea phylogeny possibly associated with hybridization, which could potentially generate observed conflict among gene trees, in Splitstree4 (v.4.13.1) (Huson and Bryant, 2006). We used the COMPLETE sequences for this analysis, and we excluded outgroup species to emphasize ingroup relationships. We used the JC69 model to estimate species distance with the “NeighborNet” distance transformation. We ran Splitstree4 on the ONEPER and CONSENSUS data sets using the same settings.

RESULTS

Target enrichment—We captured up to 498 loci across the 163 specimens, with both alleles assessed for each of our samples in the COMPLETE data set. The concatenated sequence contains 274,405 bp with an average locus length of 551 bp. The COMPLETE data set contained 67,677 parsimony-informative (PI) sites and an average of 139 PI sites per locus with 7.5% of characters coded as gaps/missing. When we create species-level CONSENSUS sequences, these numbers drop to 31,422 PI sites total and 66 PI sites per locus with 4.85% of data missing or coded as gaps. The ONEPER analysis had a total of 35,712 PI sites, with an average of 72 per locus and 7.5% of data coded as missing/gaps. The REDUCED data set had only 14,612 PI sites with an average of 41 per locus and only 1.3% missing data. See Appendix S2 for additional locus information.

Species monophyly—Phylogenetic tree topologies for the COMPLETE data set revealed monophyly for most species sampled, with a few exceptions (Appendix S4). The trees generated using the three
methods were fairly dissimilar as measured by the adjusted Robinson-Foulds distance: (RAxML-SVDquartets = 0.484, RAxML-ASTRAL-II = 0.413, SVDquartets-ASTRAL-II = 0.587). Notably, the two methods that incorporate the multispecies coalescent (SVDquartets and ASTRAL-II) were the most dissimilar when examining all individuals and all alleles per species. The main instances of consistent nonmonophyletic species were within the white protea clade, for which species and subspecies were highly mixed. Other regions included grades or very close placement rather than true monophyly (e.g., P. piscina), or divergent individuals with anomalous placement (e.g., P. scolopendrifolia 246, P. cordata 42B, P. recondita 58A, P. burchellii 1476).

**Species-level phylogeny estimation**—Tree topologies derived from concatenated vs. species tree strategies were fairly similar for the CONSENSUS data set, in terms of adjusted Robinson-Foulds distance with an average of 0.276 across the six pairwise comparisons. The two concatenated trees were topologically identical (RAxML-MrBayes = 0), while differences between the two species tree-build trees were greater (SVDquartets-ASTRAL-II = 0.290) (Fig. 2, Appendix S5). Comparisons across methodologies were more dissimilar (RAxML-SVDquartets = 0.381, RAxML-ASTRAL-II = 0.302). Results were similar if we used the COMPLETE data set (for SVDquartets-ASTRAL-II = 0.129) instead of CONSENSUS. Results for the ONEPER data set were qualitatively similar; see Appendix S3 for trees built using the ONEPER data set and comparisons across data sets. In spite of these differences, all four approaches produced trees that were much more similar to one another than any of the gene trees were to each other. They were also more similar to one another than any of the gene trees were to any of the species trees. We examined relationships further in all four species trees, but displayed the ASTRAL-II tree as a representative (Fig. 3). Although we cannot say that one tree is more accurate than another, ASTRAL-II uses bootstrapped gene trees, and topological differences among the species trees are relatively minor. The remaining trees are found in Appendix S5.

The species tree topologies in the COMPLETE data set all had three strongly supported clades, which are largely consistent with the previously published trees of Valente et al. (2010) and Schnitzler et al. (2011). These include the snow proteas plus *P. cynaroides*, which are sister to all remaining species within *Protea*, a large clade containing the non-Cape clade, rose, shale, penduline, and western ground proteas, and another clade largely containing the white, rodent, spoon-bract, bearded, dwarf-tuffed, and eastern ground proteas. Morphologically defined clades within these large groupings in *Protea* do not consistently reflect evolutionary relationships among species either in previous studies or in our species trees (see Appendix S1 for classifications from Rebele, 2001). There is also a lack of consistency and confidence in the placement of two species in particular: *P. repens* and *P. rupicola*, which we investigate further.

The average bootstrap value across branches was 93% in the ASTRAL-II analysis and 90% for the SVDquartets analysis using the COMPLETE data sets. Values were lower using the CONSENSUS data sets (92% for ASTRAL-II and 86% for SVDquartets, 87% for RAxML). As is commonly seen, posterior support for branches in the CONSENSUS Bayesian analysis was higher than bootstrap support (an average posterior probability of 0.98 in the MrBayes analysis; Appendix S5). The COMPLETE ASTRAL-II tree had only 7 branches with less than 80% bootstrap support and a total of 12 with less than 95%; SVDquartets had 12 branches with less than 80% and 18 with less than 95%; CONSENSUS RAxML had 14 branches with less than 80% and 21 with less than 95%. MrBayes had 3 branches with less than 0.95 posterior probability, and 10 that were under 1.00.

We incorporated sequence data from Schnitzler et al. (2011) with our anchored phylogenomics set (AUGMENTED data set) using 80% majority rule consensus species trees as constraints. Our resulting species trees were quite different from the maximum clade credibility tree in Schnitzler et al. (2011; see their Figure S3), with an average RFadj value of 0.587 between it and the four constraint species trees (Appendix S6). The average RFadj value among our four species trees was 0.365, although trees built using the same method were not more similar as they were in analyses of the COMPLETE or CONSENSUS data sets.

Branch lengths across the phylogeny tended to be very short. Omitting branches associated with outgroup taxa, the average branch lengths for all four trees were consistent at 2.84 × 10⁻³ substitutions per site for internal branches. The minimum values ranged from 2.23 × 10⁻³ for the SVDquartets tree to 1.04 × 10⁻² for the RAxML/MrBayes tree. The maximum branch lengths ranged from 0.0941 for the RAxML/MrBayes tree to 0.0947 in the SVDquartets tree (Appendix S7).

**Hypothesis testing: poor support due to few changes or ILS?**—Branches may have low support either because there is little information across all loci to resolve relationships (due to sites evolving too slowly or too rapidly, resulting in saturation) or because the different sets of genes have histories that are incompatible with species trees that either ignore gene tree topological variation, or account for it with ILS alone. The 50% SVDquartets consensus tree from the 354 REDUCED set best gene trees resulted in a topology with only two ingroup branches resolved, indicating either that individual gene trees lack sufficient information to support strongly resolved relationships, or that at least some genes suggest strongly supported histories that conflict with those supported by other genes.

In general, branches that conflict between species trees are not well supported in any of the trees. For instance, in the ASTRAL-II analysis, *P. repens* is placed as the sister taxon to the entire genus

![FIGURE 2](image-url)  
**FIGURE 2** Adjusted Robinson-Foulds distances among the three CONSENSUS species trees (first column, shown as horizontal bars) and between each species tree and each gene tree.
FIGURE 3  Species tree generated using ASTRAL-II. Branches with 100% bootstrap support are indicated with thick black lines; branches with less than 100% bootstrap support are orange and have support values written; branches with less than 50% bootstrap support have been collapsed. Outgroups have been removed to show details within Protea; for whole trees, see newick files in the supporting online material. Branch lengths correspond to the mean number of substitutions per site. Representative species are shown to demonstrate floral diversity. From top to bottom: Protea aurea subsp. aurea, P. punctata, P. montana, P. cordata, P. laurifolia, P. magnifica, P. longifolia, P. repens, P. susannae, P. caffra, P. gaguedi, P. lanceolata, P. sulphurea, P. nitida, P. repens, P. cynaroides. Photos credit: N. Mitchell, J. E. Carlson, and C. S. Adams.
except for the snow proteas, and *P. rupicola* is then sister to a clade nested within the remainder of the group. However, these groupings are not well supported, with bootstrap values of 49% and 42% (Fig. 3, Fig. 4A). In the SVDquartets analysis, *P. rupicola* and *P. repens* are sister species that are collectively sister to a larger clade containing all of *Protea* except for the snow proteas (Appendix S5, Fig. 4B). However, the bootstrap support underlying the sister pairing of these species is only 64%. In contrast, in the RAxML topology, these taxa form a grade within one of the major clades, with *P. rupicola* sister to a grade with *P. repens*, and *P. repens* sister to another large clade, supported with low bootstrap values of 78% and 57%. The topology from the MrBayes analysis is the same as that from RAxML, with fairly high mean posterior probabilities of 1.0 and 0.97 (Appendix S5, Fig. 4). These topologies differ from the published topologies found in Valente et al. (2010) and Schnitzler et al. (2011).

To understand the discordant placement of *P. repens* and *P. rupicola*, we performed ad hoc “alternative placement tests". Three pairwise comparisons among alternative topologies found in our trees indicate that most genes provide no information regarding the placement of these taxa. For these analyses, we use the ASTRAL-II tree as a base and then manipulated the placement of only *P. repens* and *P. rupicola* on that background. Selection of the base tree is not of extreme importance, because we are looking at differences related only to the placement of the two taxa of interest, which is manipulated by the user. Use of other species trees as backbones does not change the outcome. A gene tree having an RFadj distance closer to one topology over the other implies that that gene tree is more similar to that particular topology. In this way we can see whether genes differ strongly regarding this particular placement while controlling the rest of the tree. Here, the “A" topology reflects the topology found in the ASTRAL-II tree, “B" is the topology found in the SVDquartets tree, and “C" is the same base topology with the RAxML/MrBayes placement (Fig. 4A-C). Comparing “A" and “B", we found that 22 best gene trees were closer to “A" and 25 were closer to “B", with 307 not differing (distance was the same to either topology). Similarly, when
comparing “A” and “C,” 31 genes were closer to “A,” and only 7 were closer to “C,” and 316 did not differ. Finally, when comparing “B” and “C,” 30 were closer to “B,” 7 were closer to “C,” and 317 did not differ. Moreover, 220 out of 354 gene trees had at least one bootstrap replicate closer to one topology or the other. Within a locus, there are bootstrap replicates closer to either topology in pairwise comparisons (Fig. 4D-F).

Internode certainty (IC, which ranges from negative one to one) values across the REDUCED 60-taxon ASTRAL-II topology were on average very low (0.076) despite overall high bootstrap support, suggesting that conflicting clades in the gene trees have as much support as the focal clade in the species tree. In fact, almost half (25/56) had zero or negative IC values, indicating that a different resolution was favored by gene tree topologies other than those found in the species tree. Even excluding negative or zero values, the average is still low (0.266). Consistent with gene trees lacking resolution, we found a positive relationship between log-transformed branch length and internode certainty (Pearson’s correlation = 0.587, t = 54, P < 0.001; Appendix S7). Relationships between branch lengths and IC for the other species trees had similar patterns: SVD had 25 zero or negative values (average IC of 0.059, 0.258 with positives only), while the identical MrBayes and RAxML trees had 26 zero or negative IC scores (average of 0.058, 0.261 positives only).

Hypothesis testing: poor support due to gene tree discrepancy—

Each of the 354 individual gene trees generated in RAxML using species-level CONSENSUS data had extremely low support at nearly every branch. On average, gene trees had only 15 branches with 50% or greater support, only 5 branches with 80% or greater support, and only 2 branches with 95% or greater support (out of a possible 57 nodes in REDUCED unrooted gene trees, average support across all: 29.6%), with none of the 354 gene trees having more than 48.7% average support, and a lowest average support value of 6.5%. If we look at this in a Bayesian context, on average the majority-rule consensus gene trees from the MrBayes analysis had 14 branches with a posterior probability of 50% or higher, 10 branches with an average posterior probability of 80% or higher, and 7 branches 95% or over. The average posterior probability was only 28.3% within any individual gene tree. Bayesian analyses are not dependent on containing a reasonable chance that sites change along a branch, and thus, single sites can give strong support. It therefore seems likely that individual sites within a locus may lend some support to relationships, although there may be few sites per locus with substantial information for phylogenetic inference.

Adjusted Robinson-Foulds distances (RFadj) among the gene trees were very high, with an average of 0.912 and range of 0.667 to 1.00 across the 62,148 comparisons. In spite of these very large distances, they were more similar to each other than randomly simulated gene trees (average of 0.998, range 0.930 to 1.00), (t = 43, df = 375, P < 0.001). Gene trees simulated under the coalescent process in COAL were more similar to each other (average of 0.614, range 0.253 to 0.895) than the observed best trees (t = 67, df = 517, P < 0.001) or randomly generated trees (t = 96, df = 358, P < 0.001), indicating that ILS alone cannot explain the discrepancies among gene trees (Appendix S8).

Gene tree-to-species tree RFadj values were fairly high for the four species trees, with averages of 0.820, 0.839, and 0.828 for ASTRAL-II, SVDquartets, and RAxML/MrBayes, respectively. Gene tree-to-species tree differences were significantly different for

SVDquartets-gene and ASTRAL-gene distances (t = 4.13, df = 702, P < 0.001) and SVDquartets-gene and RAxML/MrBayes-gene (t = 2.39, df = 705, P < 0.05), though not significantly different for ASTRAL-gene and RAxML/MrBayes-gene (t = 1.74, df = 705, P = 0.08) (Fig. 2). Note that RAxML and MrBayes generated identical topologies, so distances are identical between these trees and other trees. This pattern of relatively high gene tree distance from the species trees may be due to essentially random resolution of the mostly unresolved best gene trees (given the average bootstrap support of approximately 30% for bipartitions in individual gene trees).

**Hypothesis testing: poor support due to hybridization**—Using the COMPLETE data set, the SplitsTree analysis (Fig. 5), identified a handful of species possibly involved in reticulation (P. glabra, P. nitida, P. acaculao, and P. ruipola) as well as some divergent individuals or sequences. Three of these (P. scolopendriifolia 246, P. cordata 42B, and P. recondita 58A) also had nonmonophyletic placement in the COMPLETE phylogenies, while P. venusta 148 was contained within the white proteas, but its two sequences were separated. Apart from these examples, the network has a distinctively tree-like topology. SplitsTree analyses for the ONEPER and COMPLETE data set showed similar patterns (Appendix S9).

**DISCUSSION**

**Phylogenetic support**—Analyses of large, multilocus data sets have improved support and enhanced resolution in many radiations and allowed for robust insights into lineage-specific hypotheses related to biogeography, trait evolution, and timing of events (Leaché et al., 2014; Tonnabel et al., 2014a; Shen et al., 2015). Similarly, the Protea phylogeny presented here represents a significant improvement over trees estimated from a handful of molecular markers and AFLP loci. It is difficult to compare these data sets, given nonoverlapping taxa, but for instance, in the MrBayes species tree, 59 branches (95% of the total 62) were supported with posterior probabilities over 0.95. In contrast, only 25 of 88 branches in the Schnitzler et al. (2011) analysis received posterior probabilities over 0.95 (28%) and 29 of 86 (34%) branches in Valente et al. (2010). These results are similar to those from the related genus Leucadendron (Proteaceae) in which adding nuclear markers led to significantly improved resolution over that achieved using ITS alone (Tonnabel et al., 2014b).

**Consistency of species tree methods**—There is a large body of literature dedicated to comparing and contrasting different concatenation and species tree methods with both simulated and empirical data sets (Edwards, Liu, and Pearl, 2007; Kubatko and Degnan, 2007; Gatesy and Springer, 2014; Xi et al., 2014; Tonini et al., 2015). Broadly speaking, the methods can be divided into two groups: (1) those that implicitly assume that all loci reflect the same genealogy and analyze concatenated sequences, and (2) those that allow different loci to have different genealogies and account for ILS when estimating a species tree consistent with different gene genealogies. Although there is considerable disagreement about the virtues of each approach, agreement between species tree methods and concatenation approaches suggest low levels of ILS, hybridization, or other forms of gene tree discordance. To the extent that the approaches disagree, the areas of disagreement identify clades in the tree that warrant further investigation.
In *Protea*, concatenation and species tree methods produced similar species trees. Nonetheless, topologies generated using the two different species tree methods were more similar to each other than they were to the two concatenation-based trees. Given the rapid diversification in this genus, we expect moderate-to-high levels of incomplete lineage sorting, but we are unable to definitively declare that one method or one program works better than the others, using this data set. It could be the case that concatenation methods are able to accommodate moderate levels of ILS, and that species tree methods suffer from inaccurate gene tree inference (in the case of ASTRAL-II), which we do observe in this data set. However, the greater similarity of ASTRAL-II and SVDquartets trees to one another than to the RAxML/MrBayes tree suggests that the species tree approaches resolve possible cases of ILS in similar ways. Admittedly, this could be due to the use of the CONSENSUS data set for the concatenation analyses, and the COMPLETE data set for the
species tree methods. However, part of the appeal of programs such as SVDquartets and ASTRAL-II is the ability to account for the sampling of multiple individuals per taxon. In parts of the tree where concatenation approaches differ from the ASTRAL-II or SVDquartets-based trees, ILS or other phenomena may be invoked.

**Conflicting or poorly supported clades**—Within phylogenies, some branches are likely to have stronger statistical support than others. Branches that have low support may reflect short branches with few shared changes on them, incomplete lineage sorting, or hybridization. In *Protea*, branch lengths in the species trees are positively related to internode certainty, suggesting a possible lack of shared changes. However, short branches are also associated with the presence of ILS, or branches could be artificially shortened because of admixture. This finding is consistent with low support for many branches in other phylogenies associated with inferred rapid radiations, e.g., the caenophidian snakes (Pyron et al., 2014) and the diploid *Helianthus* (Stephens et al., 2015b). In addition, when we examine individual cases of species tree discrepancies using ad hoc alternative placement tests, like the placement of *P. rupicola* and *P. repens*, we find no evidence for the strongly conflicting gene trees. Although there is limited asymmetry in support for alternate topologies, the overwhelming majority of gene trees do not differ when it comes to this particular case, and there is evidence for conflicting support within individual gene trees.

**Species tree: more than the sum of its parts?**—When the amount of phylogenetic information contained in any one locus is quite small, the best tree at that locus is not expected to be a good estimate of species relationships. For example, many IC values were negative, meaning that the most common bipartition in the bootstrap sample was not included in the best tree. These results suggest that taking the information from best gene trees alone may not produce reliable estimates of species trees. Additionally, gene trees were topologically very different from each other and from species trees. Nonetheless, both concatenation and species tree methods produce well-resolved trees that are largely congruent. Taken together, these results suggest that while the signal at any one locus is relatively low, the signal is correlated across loci leading to a relatively strong phylogenetic signal when information from many loci is combined.

**Species reciprocal monophyly**—Our analyses included samples from multiple individuals for most species, and two alleles per individual, allowing us to test for reciprocal monophyly among species (Appendix S4). Overall, species formed clades, except within the white protea. This smaller radiation within the larger radiation of *Protea* is apparently quite recent, the lack of time for divergences to be reflected in short branches within the species-level phylogeny and highly intermixed groupings in the allele-level phylogeny. There is little evidence for reticulation in this group. Instead, the SplitsTree analysis suggests a star-like radiation (Fig. 5). Nonmonophyly in the white protea could affect estimates of species-level relationships in our CONSENSUS trees and could contribute to poor resolution in these analyses. Outside the white protea, a few individuals had anomalous placement, perhaps associated with high amounts of missing data (for instance, *P. recondita* 58A had only 19 loci recovered). The extent to which high amounts of missing data affect tree topologies in phylogenomics studies remains unclear, but extremes do appear to affect placement of individuals.

**Hybridization**—In addition to the possibility of ILS contributing to discordance between gene and species trees, there is both genetic (Prunier and Holsinger, 2010) and anecdotal (A. G. Rebelo, pers. communication) evidence of hybridization in wild populations, and breeders commonly hybridize species in cultivation for the cut flower trade (Coetzee and Littlejohn, 2007). Much of the evidence for hybridization comes from observations in the white protea sub-clade and the bearded sugarbushes as defined by Rebelo (2001) (e.g., *P. magnifica*, *P. longifolia*, *P. laurifolia*, *P. lepidocarpodendron*, *P. burchellii*, etc.; Appendix S1; Coetzee and Littlejohn, 2007). The phylogenetic network from SplitsTree does not provide a formal test for hybridization, but suggests that hybridization has not played an important role in the diversification of *Protea*. We expected to find evidence for hybridization between *P. punctata* and *P. venusta* of the white proteas, because population genetic analyses have previously detected evidence of introgression between these species (Prunier and Holsinger, 2010), yet the SplitsTree analysis does not detect evidence for this hybridization. Notably, there is also a lack of evidence for hybridization in the bearded sugarbushes. Areas that do seem more network-like are associated with certain individuals with divergent sequences. Apparent reticulation involving *P. recondita* 58A may be associated with large amounts of missing data. The apparent reticulation involving *P. nitida*, *P. glabra*, *P. acaulos*, and *P. rupicola* is surprising given that these species are morphologically very different. If hybridization is occurring, it may be responsible for discrepancies in the placement of *P. rupicola* in species trees estimates derived from different methods.

In collecting samples for this analysis, we avoided sampling from individuals of questionable origin in an attempt to avoid individuals of recent admixture. Thus, these results cannot be used to infer the frequency of hybridization among extant populations of species in *Protea*. Additional population genetic work and more formal tests are necessary to verify the existence of recent interspecific gene flow.

**Major clades of Protea**—Our analyses led to a highly resolved phylogeny for *Protea*, although the traditional morphological groups defined by Rebelo (2001) are still not easily defined. Our results are, however, consistent with the strongly supported clades found in Valente et al. (2010) and Schnitzler et al. (2011). For example, the snow proteas are still very well supported as having the earliest split between the non-Cape and the non-Cape sub-clade with low support (Valente et al., 2010), but it is now consistently found in a more nested position in a group outside the non-Cape clade. The non-Cape clade is also highly supported, although *P. lanceolata* is consistently sister to this group. Previously, *P. sulphurea* was found to be sister to the non-Cape clade with low support (Valente et al., 2010), but it is now consistently found in a more nested position in a group outside the non-Cape species. We recover the rodent proteas as monophyletic, with high resolution within the group, despite evidence for hybridization. Many of the morphologically classified spoonbract and bearded proteas (Rebelo, 2001) form a monophyletic group as previously reported, yet several species from each group are found in different or very different parts of the tree (e.g., *P. coronata*, *P. grandiceps*, *P. nitida*, *P. glabra*) suggesting a possible role...
for strong convergence in form. Additionally, we find strong evidence in all species tree topologies that the morphologically similar, but geographically distinct species, *P. laurifolia* and *P. nerifolia* are sister taxa, in contrast with previous findings (Valente et al., 2010; Schnitzler et al., 2011). The placement of *P. grandiceps* as sister to the white proteas is still surprising given morphology, although this relationship is not well-supported and does not differ greatly from the topologies published in previous trees (Valente et al., 2010; Schnitzler et al., 2011). The placement of *P. witzenbergiana* with *P. recondita* is also surprising, given that *P. witzenbergiana* is morphologically more similar to *P. nana*, hinting at another instance of convergence.

**Incorporating sequence data from other sources**—Our study included field-collected samples across South Africa, but focused on the highly diverse Cape Floristic Region in South Africa. Our samples included over half of all known *Protea* species, but it did not include some rare species, and it did not include any species outside of South Africa. To build the most complete species-level phylogeny possible, we included sequences from a different set of loci and the 39 additional taxa included in Schnitzler et al. (2011). We constrained phylogenetic estimates using these data to the 80% majority-rule consensus tree for the AUGMENTED analyses. Not unexpectedly, many branches had poor support, resulting in several “combs” where we only had sequence data from Schnitzler et al. (2011) and no anchored phylogenomics data. The trees built using different methods were, on average, more dissimilar than our nonaugmented trees, and even more different from the tree published by Schnitzler et al. (2011). These differences appear to be mostly within major clades and the placement of some clades relative to each other, likely associated with poor support and what “random” resolution. These additional species do change some of our sister species groupings; for instance *P. repens* is sister to *P. aristata* in these trees, but their relative placement is still uncertain.

Although this method incorporates additional species, it is important to note that most of these relationships are still very uncertain. This is likely due to a lack of information in the Schnitzler et al. (2011) data set, but could be due to disagreement between our consensus tree and the additional data. We also do not trust branch lengths for this analysis, and therefore have not included them. These trees have limited utility, but can give a general sense of where additional species might fit. Additional statistical phylogenetic work is necessary to truly combine information from data sets with nonoverlapping sequences in a way that does not include massive amounts of missing data.

**CONCLUSIONS**

Using a broadly expanded phylogenomic data set, we were able to build well-resolved species-level phylogenies for the rapid radiation of *Protea*. The use of multiple approaches to tree-building allows us to identify potential areas of interest across the topology for investigating the influence of phenomena such as ILS or hybridization in the history of this group. The phylogenies generated here will allow for increased confidence in analyses of evolutionary questions in *Protea*, providing a basis for asking how diversity has been generated in this morphologically diverse, speciose, iconic plant lineage.

**ACKNOWLEDGEMENTS**

The authors thank J. E. Carlson and C. S. Adams for extensive field support and data collection, A. G. Rebelo and L. Nurrish for field and garden sampling support, and M. Kortyna and A. Bigelow for laboratory assistance. We thank Cape Nature, the Eastern Cape Province Department of Economic Development and Environmental Affairs, and the Eastern Cape Parks Board for collecting permits. Large-scale analyses were performed at the Computational Biology Core Facility of the University of Connecticut. Additional thanks to J. Lary for bioinformatics cluster assistance, and C. S. Jones, C. D. Schlichting, E. L. Jockusch, and two anonymous reviewers for feedback on this manuscript. K. E. Holsinger and N. Mitchell were supported by NSF DEB-1046328 and NSF DEB-1242294, P. O. Lewis was supported by NSF DEB-1354146, E. Moriarty Lemmon was supported by NSF DEB-1120516, and both E. Moriarty Lemmon and A. R. Lemmon were supported by NSF IIP-1313554.

**LITERATURE CITED**


APPENDIX 1.

Voucher information for specimens used in this study, accession number at CONN herbarium. Protea acaulos 266402, 227586; P. acuminate 227590; P. amplicaulis 256139, 266398, 261485, P. aurea-aurea 1411487, 141483; P. aurea-potbergensis 141481, 141482; P. burchellii 266405; P. caffra 248100, 245131; P. canaliculara 227575; P. compacta 227564; P. comptonii 245134, 245135; P. cordata 227595, 230152; P. coronata 227581; P. cryophila 228689, 248060; P. cyanaroides 227574; P. decurrens 230364; P. denticula 230371, P. dracomontana 248046; P. eximia 266413, 266405; P. gagueyi 248057, 248040; P. globa 255946, P. granuliceps 230363; P. intonsa 256140; P. lacticolor 141479, 141623, 141486; P. lactea 248061, P. lanceolata 266416, 230374; P. lauifolia 256127, 227477, 255937; P. lepidocarpos 230361, 255944, 230361; P. longifolia 227492, 266419; P. longifolia 266415, 227469, 227567; P. magnifica 227599; P. montana 256141; P. mundii 266410, 141617; P. mundii-east 141619, 141618; P. nau 256129; P. nenifolia 230153, 266411, 230306; P. nitida 266408, P. nigroga 248038, P. odorata 256132; P. parvula 248054; P. piscina 227578, 255939; P. pruinosa 256133; P. punctata 141608, 141615, 141621, 141613; P. recurvata 266397, 227598; P. repens 256128, 266412, 266400, 255942, 266403, 266414, 26407; P. roupelliae 245139, 248055; P. rubropilosa 248041; P. rupicola 256137; P. scabra 227571; P. scolopendrifolia 256130, 227583; P. speciosa 227572; P. subvestita 141609, 141504; P. sulphurea 266399; P. susannae 230372; P. tenax 230302, P. venusta 256134, 141500, 141606, 227596, 227695; P. weilwitschi 245138; P. witzenbergiana 227598, Faurea rochetiana; F. saligna; Serruria adscendens; S. furcellata; S. phylicoides 266418; S. trilophia.


