bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Resolving Rapid Radiations Within Angiosperm Families Using Anchored Phylogenomics

Étienne Léveillé-Bourret[1,*], Julian R. Starr[1], Bruce A. Ford[2], Emily Moriarty Lemmon[3], Alan R. Lemmon[4]

[1]*Department of Biology, University of Ottawa, K1N 6N5, Ottawa, Ontario, Canada.*

[2]*Department of Biological Sciences, University of Manitoba, R3T 2N2, Winnipeg, Manitoba, Canada.*

[3]*Department of Biological Science, Florida State University, Tallahassee, Florida 32306, United States.*

[4]*Department of Scientific Computing, Florida State University, Tallahassee, Florida 32306, United States.*

[*]*Corresponding author. E-mail:* eleve060@uottawa.ca

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract. – Despite the promise that molecular data would provide a seemingly unlimited source of independent characters, many plant phylogenetic studies are based on only two regions, the plastid genome and nuclear ribosomal DNA (nrDNA). Their popularity can be explained by high copy numbers and universal PCR primers that make their sequences easily amplified and converted into parallel datasets. Unfortunately, their utility is limited by linked loci and limited characters resulting in low confidence in the accuracy of phylogenetic estimates, especially when rapid radiations occur. In another contribution on anchored phylogenomics in angiosperms, we presented flowering plant-specific anchored enrichment probes for hundreds of conserved nuclear genes and demonstrated their use at the level of all angiosperms. In this contribution, we focus on a common problem in phylogenetic reconstructions below the family level: weak or unresolved backbone due to rapid radiations (≤ 10 million years) followed by long divergence, using the Cariceae-Dulichieae-Scirpeae clade (CDS, Cyperaceae) as a test case. By comparing our nuclear matrix of 461 genes to a typical Sanger-sequence dataset consisting of a few plastid genes (matK, ndhF) and an nrDNA marker (ETS), we demonstrate that our nuclear data is fully compatible with the Sanger dataset and resolves short backbone internodes with high support in both concatenated and coalescence-based analyses. In addition, we show that nuclear gene tree incongruence is inversely proportional to phylogenetic information content, indicating that incongruence is mostly due to gene tree estimation error. This suggests that large numbers of conserved nuclear loci could produce more accurate trees than sampling rapidly evolving regions prone to saturation and long-branch attraction. The robust phylogenetic estimates obtained here, and high congruence with previous morphological and molecular analyses, are strong evidence for a complete tribal revision of CDS. The anchored hybrid enrichment probes used in this study should be similarly effective in other flowering plant groups. [*Carex*, coalescent based species

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tree, flowering plants, low-copy nuclear genes, low-level phylogenetics, universal hybrid
40 enrichment probes]

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

One of the strongest arguments for the use of molecular data in phylogenetic reconstruction was that it provided a seemingly unlimited source of independent characters (Hillis 1987; Hillis and Wiens 2000; Scotland et al. 2003). Although sound in theory, most plant phylogenetic studies are still limited to just two regions, the plastid genome and the nuclear ribosomal DNA (nrDNA; Hughes et al. 2006) region. These regions are widely used because they are easily amplified due to high copy numbers (Álvarez and Wendel 2003) and the availability of universal PCR primers for many of their loci (e.g. White et al. 1990; Taberlet et al. 1991; Baldwin 1992). These regions are also particularly attractive for phylogenetic research because they consist of coding and non-coding loci which evolve at different rates (White et al. 1990; Wicke and Schneeweiss 2015), and they can be used in combination to study processes such as hybridization (Rieseberg et al. 1990; Feliner and Rosselló 2007).

Accessing other sources of molecular characters in plants has not been easy. Although the mitochondrial genome should be a prime source of characters due to high copy numbers, low sequence variation, major structural rearrangements and frequent lateral gene transfer has rendered its sequences impractical for most applications (Palmer and Herbon 1988; Knoop 2004; Bergthorsson et al. 2003; Richardson and Palmer 2006). The plant nuclear genome is equally problematic. Its vast size (63.4–700,000 Mbp; Greilhuber et al. 2006; Pellicer et al. 2010), independent genealogical histories and biparental inheritance are favourable characteristics; however, its generally higher evolutionary rates, low copy numbers and the scarcity of complete model genomes means that designing broadly applicable PCR primers in non-model organisms is rarely successful. This is true even for the most interesting and well-known plant nuclear loci

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Hughes et al. 2006) such as the *Waxy* (granule-bound starch synthase) or *LEAFY* genes, and other challenges, such as gene duplication, often necessitate extensive rounds of cloning to differentiate paralogs (e.g., Mason-Gamer et al. 1998; Hoot and Taylor 2001). Even when primers are successfully designed, they often cannot amplify low copy loci from degraded tissue samples, such as herbarium specimens, which means that their usefulness for studies on species-diverse or geographically widespread groups is limited. Consequently, most plant molecular phylogenetic studies consist of only a handful of linked loci from the plastid genome (e.g., *matK*, *ndhF*, *trnL-F*) and nrDNA cistron (ITS region, ETS; Hughes et al. 2006) and are rarely comprised of combined analyses of more than five sequenced regions from these two character sources.

Next-generation sequencing (NGS) technologies are directly and indirectly facilitating the exploration of new sources of molecular characters in non-model plants. Although full genome sequencing remains beyond the reach of most systematists, NGS promotes the development of genomic resources for new model organisms, which in turn provides data useful for the design of new Sanger-based markers such as low copy nuclear genes (Blischak et al. 2014; Chamala et al. 2015) or microsatellites (Gardner et al. 2011). In addition, the development of efficient multiplexing and enrichment methods are making NGS increasingly accessible as a method to directly gather data for larger species samples in non-model organisms (Cronn et al. 2012; Lemmon and Lemmon 2013). Low-coverage shotgun sequencing (genome skimming) and organellar genome enrichment permit rapid and efficient sequencing of large phylogenomic matrices from the high-copy regions of genomes (organelles and nrDNA; Straub et al. 2012). However, these approaches are limited by the finite size and generally linked nature of the

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

targeted regions, as they were prior to the invention of NGS. Moreover, they have been unable to completely resolve several important plant radiations (Xi et al. 2012; Barrett et al. 2013, 2014; Ma et al. 2014; Straub et al. 2014). While cost-efficient alternatives, including RADseq (Baird et al. 2008) and transcriptome sequencing (e.g., Wen et al. 2013), can provide data from thousands of unlinked nuclear loci, they both have limitations for phylogenetic analysis. Indeed, RADseq datasets are characterized by short loci of uncertain homology and high amounts of missing data (Rubin et al. 2012; Huang and Knowles 2016), and although it has been used in phylogenetic studies of radiations at least as old as 60 Ma (Gonen et al. 2015; Eaton et al. 2016), the existence of many different protocols and the anonymous nature of RADseq loci (lacking a reference genome) does not facilitate data sharing and reuse across study groups (Harvey et al. 2016). On the other hand, transcriptome sequencing has the potential to be useful at any taxonomic level, but important drawbacks include the complexity of working with RNA (Johnson et al. 2012), especially when living material is not available, and the computational burden of gene assembly and orthology inference in plant genomes where gene families, paralogs, and splice variants are common (Cronn et al. 2012).

A more flexible and promising approach is hybrid enrichment, a method that reduces the bioinformatics and laboratory complexity of transcriptome sequencing by using probes designed from existing genomic or transcriptomic sequences to enrich a fixed set of molecular targets (Lemmon and Lemmon 2013). Low-copy nuclear gene enrichment probes have already been designed to work across vertebrates (Faircloth et al. 2012; Lemmon et al. 2012), and they have been used in the phylogenetic analysis of birds (Prum et al. 2015), snakes (Pyron et al. 2014;

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ruane et al. 2015), lizards (Leaché et al. 2014; Brandley et al. 2015; Pyron et al. 2016), frogs (Peloso et al. 2014) and fishes (Eytan et al. 2015) amongst others. In plants, hybrid enrichment probes have been designed for several genera (e.g., de Sousa et al. 2014; Weitemier et al. 2014; Nicholls et al. 2015; Schmickl et al. 2015; Stephens et al. 2015; Heyduk et al. 2016; Johnson et al. 2016), a subfamily of palms (Arecoideae, Arecaceae; Comer et al. 2016), a subfamily of grasses (Chloridoideae, Poaceae; Fisher et al. 2016), and for the sunflower family (Asteraceae; Mandel et al. 2015). Although this taxon-specific approach in plants has been successful, it requires new probes to be designed for every group, and the data generated from such studies has limited potential to be reused because the targeted regions are group-specific. On the contrary, if conserved targets are selected, hybrid enrichment probes can be designed to work on broad taxonomic scales. This method, known as “anchored phylogenomics” (Lemmon and Lemmon 2012), has the potential to provide parallel datasets for a fixed set of loci across large taxonomic groups, like flowering plants. In other words, anchored phylogenomics has the potential to become the modern NGS equivalent of the “universal” PCR primer papers that resulted in an explosion of phylogenetic studies in non-model organisms during the past decades (White et al. 1990; Taberlet et al. 1991; Baldwin, 1992).

The success of hybrid enrichment in several isolated plant groups has motivated us to design a new set of flowering plant-specific probes that can enrich nuclear genes across all flowering plants. In a previous contribution, we identified 517 target loci using 25 angiosperm genomes and we demonstrated their universality and broad utility in flowering plants (Buddenhagen et al., in prep.). This new resource has the potential to greatly simplify and

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

accelerate plant phylogenomic research by reducing the burden of marker choice and probe design, and by promoting the accumulation of parallel data from a standard set of nuclear genes shared by all plant families. Moreover, it would be especially useful if it was able to resolve relationships at both higher and lower taxonomic levels. In fact, the angiosperm probe kit is already being widely adopted through numerous ongoing collaborations to collect data from nearly 90 angiosperm families for more than 2000 samples (to date) at the Center for Anchored Phylogenomics (e.g. Mitchell et al. 2017; www.anchoredphylogeny.com). Building upon Buddenhagen et al. (in prep.) where we examine the relationships of major angiosperm lineages, this contribution demonstrates the utility of the method to resolve difficult branches in a rapid radiation of tribes and genera (Cyperaceae, Cariceae-Dulichieae-Scirpeae).

The radiation of *Carex*, the most diverse flowering plant genus of the northern hemisphere (Starr et al. 2015), and its relatives within the Cariceae-Dulichieae-Scirpeae clade (hereafter CDS; Cyperaceae) provides an ideal case to test the utility of universal nuclear gene enrichment probes at medium and shallow phylogenetic depth in plants. With 16 genera and over 2,200 species, this cosmopolitan clade is of considerable evolutionary interest due to its habitat variety (deserts to rain forests), biogeographic patterns (e.g., bipolar, Gondwanan, Amphiatlantic; Croizat 1952), and unique cytology (n = 6 to 56) promoted by agmatoploid chromosomal fusion and fragmentation (Hipp et al. 2009). Previous phylogenies of CDS have identified seven major lineages using the traditional combination of plastid and nrDNA markers (Muasya et al. 2009; Léveillé-Bourret et al. 2014; Léveillé-Bourret et al. 2015). However, like many plant groups, the backbone of the tree remains unresolved possibly due to a relatively old crown age (>40 Ma) and

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

an early radiation that occurred over just 10 million years (Escudero et al. 2013; Spalink et al. 2016). As a result, the most rapidly-evolving plastid genes contain few, if any, characters to support the backbone topology, whereas non-coding plastid and nrDNA regions have diverged so much they cannot be confidently aligned across the whole group. These factors suggest that large numbers of nuclear genes are needed to resolve the backbone phylogeny of CDS, and universal anchored phylogenomics probes could provide the quick and efficient means to obtain them.

The aims of this study were twofold to: 1) test the utility of anchored phylogenomics in closely related genera of flowering plants showing evidence of rapid diversification; and 2) to resolve long-standing taxonomic problems in CDS by estimating a robust phylogeny of the major lineages of the clade. Using the first set of universal probes available for nuclear gene enrichment in flowering plants, we collected data from hundreds of loci in 34 species not included in the initial probe design and representing the full phylogenetic diversity of the CDS clade. To determine whether the results of such an analysis would be compatible with prior analyses, we compared our genomic tree to a phylogeny estimated from a typical plastid plus nrDNA Sanger-derived dataset that is still commonly generated by many researchers today (i.e., a plastid and nrDNA analysis). We discuss the value of anchored phylogenomics for resolving rapid radiations in flowering plants, and the implications of our results on the taxonomy and evolution of the Cariceae-Dulichieae-Scirpeae clade.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MATERIALS AND METHODS

Taxon Sampling and DNA extraction

A total of 32 ingroup taxa were included to represent the all major clades of the CDS clade as based on a previous phylogenetic study with extensive taxonomic sampling of the clade (Léveillé-Bourret et al. 2014). This includes *Dulichium arundinaceum* (Dulichieae; comprising ca. 7 spp.), *Khaosokia caricoides* (incertae sedis), both *Calliscirpus* species (Calliscirpus Clade; 2 spp.), *Amphiscirpus nevadensis* (Zameioscirpus Clade; comprising ca. 8 spp.), 5 *Scirpus* and 3 *Eriophorum* species (Scirpus Clade; comprising ca. 48 spp.), 2 *Trichophorum* species (Trichophorum Clade; comprising ca. 18 spp.), and 18 *Carex* species (Cariceae; comprising ca. 2,150 spp.) representing all major Cariceae lineages identified by Starr et al. (2015). Two outgroup taxa (*Eleocharis obtusa* (Willd.) Schult.; Eleocharideae and *Erioscirpus comosus* (Wall.) Palla; Cypereae) were selected from the CDS sister group, the Abildgaardieae-Eleocharideae-Cypereae-Fuireneae clade (Muasya et al. 2009). Two accessions of *Scirpus atrovirens* were included to test the repeatability of the hybrid-enrichment methodology. Leaves collected fresh in the field and dried immediately in silica gel were used in whole genomic DNA extractions using the silica-column based protocol of Alexander et al. (2007) as modified by Starr et al. (2009). However, increased quantities of leaf tissue (80–100 mg instead of 20 mg) and reagents were used to account for the greater mass of DNA required for NGS protocols. We aimed for 1-3 μg of DNA for hybrid enrichment, although some samples with as little as 0.15 μg of DNA worked very well with our methodology. Voucher information is available in Appendix 1.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hybrid Enrichment Data Collection

Data were collected following the general methodology of Lemmon et al. (2012) through the Center for Anchored Phylogenomics at Florida State University (http://anchoredphylogeny.com/). After extraction, genomic DNA was sonicated to a fragment size of ~300-800 bp using a Covaris E220 Focused-ultrasonicator with Covaris microTUBES. Subsequently, library preparation and indexing were performed on a Beckman-Coulter Biomek FXp liquid-handling robot following a protocol modified from Meyer and Kircher (2010). Briefly, sonication is followed by blunt-end repair using T4 DNA polymerase, two different adapters are ligated to both ends of the DNA molecules using T4 DNA ligase, and indexes and full length adapter sequences are added by amplification with 5'-tailed primers. An important modification of this protocol is the addition of a size-selection step after blunt-end repair using SPRI select beads (Beckman-Coulter Inc.; 0.9× ratio of bead to sample volume). Indexed samples were then pooled at equal molarities (typically 16-18 samples per pool), and then each pool was enriched using the Angiosperm v.1 kit (Agilent Technologies Custom SureSelect XT kit), which contained probes for 517 flowering plant exons (average: 287 bp, median: 225 bp) as described by Buddenhagen et al. (in prep.). Briefly, the probes were designed by selecting genes that are putatively single copy in *Arabidopsis*, poplar (*Populus*), grape (*Vitis*) and rice (*Oryza*), filtering out exons below the minimum size necessary for enrichment, and then narrowing down on the exons that had ≥55% similarity between *Arabidopsis* and rice. Using these two taxa as reference, orthologous regions from 33 complete flowering plant genomes were identified, and the 517 exons that had an average copy number ≤1.2 per genome were selected for probe design. More

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

details, including the probe sequences, can be found in Buddenhagen et al. (in prep.).After enrichment, 3–4 enrichment reactions were pooled in equal quantities for each sequencing lane and sequenced on paired-end 150-bp Illumina HiSeq 2500 lanes at the Translational Science Laboratory in the College of Medicine at Florida State University.

Assembly

Reads passing quality filtering were checked for overlap and merged following Rokyta et al. (2012). Reads that could not be merged were treated as unpaired during assembly. Reads were assembled on the flowering plant anchored enrichment references following Buddenhagen et al. (in prep.) and contigs were extended into flanking regions using a *de novo* assembler. Briefly, preliminary matches between each read and the reference sequences were called if 17 bases matched a library of spaced 20-mers derived from the references. Reads were then considered mapped if 55 matches were found over 100 consecutive bases in the reference sequences (all possible gap-free alignments between the read and the reference were considered). The approximate alignment position of mapped reads were estimated using the position of the spaced 20-mer, and all 60-mers existing in the read were stored in a hash table used by the *de novo* assembler. Then, the *de novo* assembler maps additional reads by identifying exact matches between a read and one of the 60-mers in the hash table. Simultaneously using the two levels of assembly described above, the reference sequences were traversed repeatedly until a pass produced no additional mapped reads, enabling extension of assemblies into variable flanking regions. Contigs were estimated from 60-mer clusters. For each locus, a list of all 60-mers found in the mapped reads was compiled, and the 60-mers were clustered if found together in at least

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

two reads. Each cluster of 60-mers was then used to separate the reads into contigs. Relative alignment positions of reads within each contig were then refined in order to increase the agreement across the reads. Up to one gap was also inserted per read if needed to improve the alignment. In the absence of contamination, low coverage or gene duplication, each locus should produce one assembly cluster. Consensus bases were called from assembly clusters as ambiguous base calls (IUPAC ambiguity codes) only when polymorphisms could not be explained as sequencing error (assuming a 0.1 probability of error and alpha equal to 0.05, Buddenhagen et al., in prep.). Called bases were soft-masked (made lowercase) for sites with coverage lower than 5. Assembly contigs derived from less than 10 reads were removed in order to reduce the effects of cross contamination and rare sequencing errors in index reads.

Orthology, Filtering and Alignment

Orthology was determined for genes with multiple copies following Prum et al. (2015). Briefly, for each locus, the distance between each pair of contig sequences was computed as the proportion of shared 20-mers. The list of 20-mers was constructed from both consecutive bases and spaced bases (every third base). Contig sequences were then clustered with neighbor-joining (NJ) using this alignment-free distance measure, allowing at most one sequence per species in each NJ cluster. This results in multiple clusters, each containing at most one sequence per species. Each cluster is then treated as a probable paralog. Gene copies were efficiently sorted using their variable flanking regions recovered during extension assembly. Clusters containing fewer than 50% of the species were removed from downstream processing. Finally, alignments of the remaining orthologous sequence clusters were performed with MAFFT v. 7.023b (Katoh

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2013), with the --genafpair and --maxiterate 1000 flags utilized, and alignments were trimmed/masked using the steps from Prum et al. (2015) and Buddenhagen et al. (in prep.). Briefly, a sliding window of 20 bp was used to mask regions where <10 sites had the most common character present in at least 40% of the sequences. Sites with fewer than 12 unmasked bases were also removed from the alignments. Because of the relatively deep phylogenetic timescale of this study, many variable sites in the regions flanking the conserved exonic core of the probes were masked because they were too variable to align across all taxa.

After the initial automatic alignment in MAFFT, there remained several obviously misaligned regions that were were not removed by the previous filtering step. We initially tried using Gblocks 0.91b (Castresana 2000) to exclude poorly aligned or highly divergent regions, but no parameter combinations could remove some clearly misaligned regions, whereas apparently well-aligned and informative regions were often excluded. This was probably due to the fact that most ambiguous stretches consisted of a few completely misaligned sequences within well-conserved blocks, a situation which is known to confound Gblocks (Castresana 2000). In consequence, all nuclear gene alignments were visually examined and sites containing misaligned bases, diagnosed by long (>3 bp) stretches of disagreements to the consensus sequence in one or a few taxa, were excluded. All the separate alignments were combined in a single concatenated alignment for concatenated analyses or kept separate for analyses based on gene trees.

Phylogenetic Analyses

Parsimony analyses.—Heuristic maximum parsimony (MP) searches on the concatenated

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

alignment were performed in PAUP* v4.0 (Swofford 2003) using 1,000 random addition sequence (RAS) replicates, tree-bisection-reconnection branch swapping, holding 5 trees at each
270 step and with the STEEPEST option ON. To prevent undersampling-within-replicate and frequency-within-replicate artefacts, support was assessed with 1000 jackknife 50% replicates using 10 RAS replicates, saving a maximum of 10 trees per RAS, and using the strict-consensus jackknife (GRPFREQ=NO) following the recommendations of Simmons and Freudenstein (2011). Single-locus parsimony analyses and partitioned, hidden and total Bremer support values
275 were calculated in PAUP with the help of ASAP, a perl script provided by Sarkar et al. (2008). Concatenated MP analyses excluding outgroups and/or Cariceae were made to determine whether they could be causing long branch attraction problems affecting ingroup topology.

Concatenated maximum likelihood analyses.—Concatenated maximum likelihood (ML) analyses were performed in RAxML 8.1.11 (Stamatakis 2014) on the CIPRES Science Gateway
280 v3.3 (Miller et al. 2010). The partitioning scheme was selected among all locus subsets with PartitionFinder v1.1.1 (Lanfear et al. 2012) using the relaxed hierarchical clustering algorithm (Lanfear et al. 2014) based only on subset similarity (--weights 1,0,0,0), with --rclust-percent settings of 1%, 2%, 4% and 10% and using the Bayesian information criterion (BIC), with GTR+G as the only allowed model. The best scoring scheme (BIC = 3,277,453.37809) was found
285 at a --rclust-percent setting of 4% and comprised 19 partitions with 2–127 loci and 395–31,991 distinct alignment patterns. RAxML searches were made with 100 randomized maximum parsimony starting trees and the new rapid hill-climbing algorithm (Stamatakis et al. 2007). Branch support was assessed with 100 (standard) bootstrap replicates (Felsenstein, 1985). Single-

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

locus ML searches were done using the rapid hill-climbing algorithm and 200 rapid bootstrap
290 replicates (option -f a) using a python script to input the parameters to RAxML 8.1.21. Internode and tree certainty values were calculated in RAxML 8.2.4 (Salichos et al. 2014).

Species tree analyses.—Our phylogenetic problem is characterized by very short backbone branches susceptible to gene tree incongruence (Degnan and Rosenberg 2006). However, fully parametric coalescent-based species tree estimation such as *BEAST is too computationally
295 demanding to be used with hundreds of loci and 34 taxa (Ogilvie et al. 2016). We therefore estimated the species tree with ASTRAL-II v4.10.12 (hereafter ASTRAL), a "summary" species tree method that has been shown to be more accurate and less sensitive to gene tree estimation error than alternatives (e.g. MP-EST) in simulation studies (Mirarab et al. 2014; Chou et al. 2015; Mirarab and Warnow 2015). Trees from the single-locus RAxML searches were used as input in
300 ASTRAL, and branch support was assessed with local posterior probabilities (Sayyari and Mirarab 2016) and with 100 gene and site bootstrap replicates (option -g -r). Internode certainty of the branches of this tree were calculated in RAxML 8.2.4 based on all bipartitions of each quartet, and using the lossless adjustment scheme to correct for incomplete gene trees (Salichos et al. 2014, Kobert et al. 2015).

305 *Substitution saturation.*— Site-specific substitution rates were estimated in IQ-TREE 1.5.0 (Nguyen et al. 2015) by re-optimizing a single GTR+G (16 rate categories) model to the whole concatenated alignment using the ML topology found by RAxML, and using the "-wsr" option. Using these site-specific rates, the ⅓ fastest-evolving sites were extracted and tested for

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

substitution saturation. These sites should be dominated by 3rd codon positions and non-coding bases. Substitution saturation was assessed by plotting raw number of transversions and transitions against GTR distances and noting whether a plateau is attained. Additionally, the statistical test of substitution saturation of Xia et al. (2003) was made in DAMBE (Xia and Lemey 2009; Xia 2015) with 1,000 jackknife replicates on subsets of 4, 8, 16 and 32 taxa.

Sources of incongruence.— Multiple factors such as lateral gene transfer, gene duplication, hybridization, incomplete lineage sorting and gene tree estimation error cause incongruence between gene trees and their associated species tree. Different types of exploratory data analyses were therefore done to pinpoint the sources of incongruence in our dataset, and especially to determine whether the observed incongruence between gene trees is caused by biological factors (hard incongruence) or is simply due to gene tree estimation error (soft incongruence). Robinson-Foulds distances between ML-estimated gene trees and ASTRAL-estimated species trees were calculated with the ape package (Paradis et al. 2004) in R (R Core Team 2016). We made a linear regression of gene tree to species tree distances against the average gene tree bootstrap support. In addition, bootstrap support of gene tree branches present or absent in the species tree were compared with the help of histograms. Several reduced concatenated MP analyses were made including only loci conflicting with selected backbone branches (Fig. 1) to determine whether combined analyses would give similarly conflicting results, which would be expected in the case of hard incongruence, or whether combined analysis would negate conflict, which should happen if incongruence was simply due to gene tree estimation error. Scatter plots, histograms and linear regression coefficients were drawn and calculated with R.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Reduced analyses.—The influence of the number of loci and analysis method on the reconstructed phylogeny was assessed with reduced analyses. In each analysis, a number of loci were randomly selected (without replacement) and analyzed either by concatenation in RAxML or with ASTRAL. Then, Robinson-Foulds distance between the resulting tree and our best estimate of the species tree (ASTRAL) was calculated. This procedure was repeated 200 times for 5, 10, 20, 50, 100, 200 and 400 loci in ASTRAL, and 50 times for the same number of loci in RAxML. To determine if the information content of the selected loci has an effect on phylogenetic analyses, we ranked loci based on number of informative characters and repeated the ASTRAL reduced analyses with the 33% highest ranking loci and the 33% lowest ranking loci, making 200 replicate analyses with 5, 10, 20, 50 and 100 loci. Boxplots were drawn with R (R Core Team 2016).

Comparative Sanger Matrix

To compare phylogenetic results obtained with hybrid enrichment, data from the plastid genes *matK* and *ndhF* (Gimour et al. 2013), as well as the nrDNA region ETS-1f, were obtained from Genbank for the same species as those used in phylogenomics analysis (but replacing the outgroup *Eleocharis obtusa* with *E. acicularis* and including only one *Scirpus atrovirens* accession). Five sequences were newly obtained by PCR and Sanger-sequencing following the procotols in Léveillé-Bourret et al. (2015). Genbank accession numbers and voucher information are available in Appendix 2.

The sequences were concatenated by species, aligned with the MAFFT v7.017b (Katoh and

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Standley 2013) plugin in Geneious 8.1.7 (http://www.geneious.com; Kearse et al. 2012), and the resulting alignments were corrected by hand. Concatenated maximum likelihood (ML) analyses were done in RAxML 8.2.4 (Stamatakis 2014). The partitioning scheme was selected among all codon and locus subsets with PartitionFinder v1.1.1 (Lanfear et al. 2012) using an exhaustive search (Lanfear et al. 2014) and using the Bayesian information criterion (BIC), with GTR+G as the only allowed model. The best scoring scheme (BIC = 31,271.0585328) comprised three partitions: codons 1 and 2 of *matK* and *ndhF*, codon 3 of *matK* and *ndhF*, and ETS-1f. RAxML searches were done using the rapid hill-climbing algorithm and support estimated with 500 rapid bootstrap replicates (option -f a). Support in parsimony was assessed with 500 bootstrap replicates in PAUP* v4.0, using 10 RAS replicates, saving a maximum of 20 trees per RAS, and using the strict-consensus bootstrap (GRPFREQ=NO).

RESULTS

Sequence Characteristics

A total of 462 loci were recovered from the Illumina reads, including 456 targets in single-copy, three targets duplicated into two paralogs each, and only 59 targets not recovered (11.4% of the 517). A single locus possessed no informative variation and was excluded. The average base coverage of loci before trimming was 0–5576 (mean = 234) across terminals, and maximum number of distinct copies per locus (before orthology filtering) ranged from 1 to 4 (mean = 1.4). After alignment and removal of misaligned bases, loci were 219–1,875 bp long (mean = 649), with 1–80% of their length consisting of flanking sequence (mean = 56%), and each with 26–371 parsimony informative characters (mean = 120). Two hundred seventy-nine loci were missing

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

some terminals, but 99% of all loci had at least 29 (83%) terminals. Each terminal had data for 425–462 loci, averaging 457 loci per terminal (98.9% of all loci). The combined dataset was 299,241 bp long after trimming and exclusion of 6,649 misaligned sites. This included 55,417 (18.5%) parsimony informative characters, 4.7% missing, 0.4% ambiguous bases and a GC-content of 40%. Sequence statistics for each locus are found in online Appendix 1 (available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.55h30). The two *Scirpus atrovirens* accessions had 99.3% identical sites and almost identical coverage, with only 5,865 sites (ca. 1% of total aligned length) present in one accession but absent in the other. No evidence of strong substitution saturation was found in the ⅓ fastest-evolving sites. A non-linear trend was visible in the GTR vs transversions and the GTR vs transition plots (online Appendix 2), but the Iss values (0.364–0.424) were significantly smaller ($p < 0.001$) than critical thresholds (0.603–0.860) for all subset sizes and for symmetrical and asymmetrical topologies, which suggest little saturation (Xia et al. 2003).

Phylogenetic Results

Concatenated parsimony searches on the phylogenomic matrix found a single shortest tree of 219,733 steps (consistency index = 0.71, retention index = 0.74; online Appendix 3). The best tree found by concatenated ML searches had a log-likelihood of -1,636,930.180799 as calculated by RAxML (online Appendix 4). The best MP and ML trees were almost identical to the ASTRAL species tree (Fig. 1), except for an unsupported sister-relationship between *Carex canescens* (representing the Vignea Clade) and the Core Carex Clade in MP, and a highly supported (100% BS) sister-relationship between *Scirpus cyperinus* and *S. atrovirens* in MP and

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ML. Relationships between major lineages of the CDS clade, the focus of this study, were identical in all analyses, and their relative position remained stable when outgroups and/or Cariceae were excluded from MP analyses. The MP and ML trees obtained with the comparative Sanger matrix (*matK* + *ndhF* + ETS-1f) were completely congruent with the phylogenomics results, except that most backbone branches had low support in the the Sanger matrix, but very high support in the phylogenomic matrix (Fig. 1).

Phylogenetic analyses position Dulichieae and *Khaosokia* as successive sisters to a highly supported Cariceae + Scirpeae clade. Within this clade, Scirpeae forms four major lineages in three monophyletic groups: a Calliscirpus Clade (*Calliscirpus*) sister to everything else, *Amphiscirpus* (representing the Zameioscirpus Clade) sister to a Scirpus Clade (*Scirpus* + *Eriophorum*), and a Trichophorum Clade (*Trichophorum*) sister to Cariceae. These backbone relationships are highly supported by all analyses except for the position of *Calliscirpus*, which is highly supported in MP and ML, but is supported in only 68% of ASTRAL bootstrap replicates, with the most frequent conflicting BS replicate trees (31%) putting *Calliscirpus* sister to the clade comprising the Zameioscirpus Clade + Scirpus Clade.

Incongruence Between Gene Trees

Incongruence between ML estimated gene trees was high in the short backbone branches of the phylogeny. This is reflected by small internode certainty values and relatively high numbers of loci with negative partitioned Bremer support (supporting conflicting clades in parsimony) for these short branches identified by letters A to E in Figure 1 (Table 1). However, support and

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

resolution of backbone branches was low in most estimated gene trees, and there was a clear negative relationship between average ML bootstrap (across all branches) of a gene tree and its distance to the ASTRAL species tree (Fig. 2, $R^2 = 0.24$, slope = -0.0063, slope p-value < 10^{-15}). In addition, the average bootstrap support for branches present in the ML gene trees, but absent in the species tree, was considerably lower than support for branches present in both (Fig. 3). Combined analyses identified extensive emergent support for the backbone branches of the estimated species tree even in loci that are apparently conflicting in single-locus analyses, as shown by the high proportion of hidden Bremer support for backbone branches (Table 1). Likewise, concatenated MP analyses of all loci conflicting with selected backbone branches (identified with letters in Fig. 1) always gave highly supported trees completely congruent with the backbone of our estimated species tree, consistent with soft incongruence due to gene tree estimation error rather than hard incongruence due to biological factors.

Reduced Analyses

The 33% highest and 33% lowest ranking loci in terms of number of potentially informative characters had an average of 161.2 (sd = 34.4) and 81.6 (sd = 16.7) informative characters, respectively. Results obtained in reduced analyses by using all loci in ASTRAL, RAxML, or with only the highest or lowest ranking loci in ASTRAL, were all broadly similar. Distance between trees estimated in reduced analyses and the best species tree diminished with increasing number of loci per jackknife replicate, with the majority of replicates having less than 10% conflicting bipartitions with 100 loci or more (Fig. 4). With ASTRAL and 200 loci or more, all replicates had a backbone identical to the best species tree, whereas the position of

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Calliscirpus was inconsistent in a minority of replicates with 100 loci. Results were similar when using the highest and lowest ranking loci in ASTRAL. With RAxML, 100 loci were sufficient to get a backbone identical to the species tree in all replicates.

DISCUSSION

Targeted NGS of Conserved Nuclear Genes in Phylogenetic Inference

Using the first set of universal probes available for nuclear gene enrichment in flowering plants, we were able to collect data from hundreds of loci in 34 taxa representing a typical flowering plant radiation encompassing >40 million years of evolution (Escudero et al. 2013; Spalink et al. 2016). Despite short backbone internodes connected to long branches, typical of ancient rapid radiations (Whitfield and Lockhart 2007), the inferred backbone relationships were well supported in both concatenation and coalescence-based analyses. These results illustrate the great promise of anchored phylogenomics for the resolution of rapid ancient radiations of non-model organisms.

Important amounts of incongruence between estimated gene trees was found in our dataset. The observation that gene tree incongruence was inversely proportional to the amount of phylogenetic information content (as measured by average gene tree ML bootstrap) indicates that at least part of the incongruence must be due to gene tree estimation error. This is corroborated by the lower bootstrap support of gene tree branches absent in the species tree, and the high amount of hidden Bremer support in the shortest branches of the backbone. This also explains why ASTRAL analyses necessitate more loci (ca. 200) than ML analyses (ca. 100) to get consistent

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

results on the backbone relationships of CDS: the estimated gene trees that ASTRAL takes as input are highly affected by estimation error, whereas concatenation presumably amplifies the phylogenetic signal common to all loci, thus reducing the relative influence of noise on the results (Townsend et al. 2012; Bayzid et al. 2015; Warnow 2015; Meiklejohn et al. 2016). The same effect is seen in several simulation studies that have shown higher efficiency of concatenation relative to summary coalescence methods when incomplete lineage sorting is low (e.g. Bayzid and Warnow 2013; Chou et al. 2015; Mirarab et al. 2016). Because the probes used for enrichment were designed to be universal for flowering plants, and since many sites in the variable flanking regions were filtered out because of the phylogenetic depth of the study, the anchored loci tend to be slow-evolving in this study. This resulted in modest numbers of informative characters per locus and low levels of support for individual gene phylogenies. However, it should be noted that the faster evolving flanking regions of the probes could be retained in studies focusing on shallower divergences where additional sequence variation is needed.

It has been argued that small numbers of highly-informative loci are preferable to larger numbers of more slowly-evolving loci when attempting to resolve phylogenies with short branches (Salichos and Rokas 2013). However, the matter is certainly more complex, because more variable loci are often noisier due to multiple substitutions (Townsend et al. 2012) and they have a higher susceptibility to long-branch attraction (Felsenstein 1978; Bergsten 2005). In the case of rapid ancient divergences, difficulties arise because of multiple factors: short backbone branches offer poor phylogenetic signal and increase the probability of deep coalescences,

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

whereas long terminal branches are susceptible to problems of substitution saturation and long-
475 branch attraction (Whitfield and Lockhart 2007). This creates an apparent tradeoff, since fast-evolving loci have a higher probability of containing variation informative for short backbone branches, but are also more susceptible to substitution saturation and long-branch attraction. Indeed, selection of loci should not aim for enormous amounts of variation dominated by noise, but rather for sufficient variation with a high signal/noise ratio and good taxonomic coverage
480 (Philippe et al. 2011; Betancur-R. et al. 2014; Hedtke et al. 2006). For this reason, slowly-evolving, homoplasy-free markers have been suggested to be optimal for the resolution of ancient rapid radiations (Whitfield and Lockhart 2007). Empirical results and simulation studies also indicate that unresolved gene trees are less problematic in a species tree framework than gene trees potentially biased by substitution saturation or long-branch attraction (Chiari et al. 2012; Xi
485 et al. 2015). All these considerations seem to indicate that resolving rapid radiations would profit more from large numbers of conserved loci than from similar numbers of more variable loci. This idea is clearly supported by our success in resolving the polytomy at the base of the CDS clade with hundreds of conserved nuclear genes, despite low levels of support for individual gene trees. Our reduced analyses likewise indicate that using loci with higher or lower numbers of
490 informative characters has almost no effect on our results, whereas the number of loci had a very significant effect on precision of species tree and concatenation analyses. This suggests that future phylogenomic studies based on conserved nuclear loci could profit more from large numbers of loci than from sampling more characters per locus, despite the likelihood that longer loci could decrease gene tree estimation error. However, Meiklejohn et al. (2016) found that

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

species tree estimation methods gave inconsistent results when using gene trees with very low signal (<25 informative characters), whereas more informative loci gave more consistent results. The lack of relationship we found between gene tree information content and species tree accuracy could be explained by the fact that all our loci contained at least 26 informative characters, which suggests that there could exist a threshold below which methods based on estimated gene trees might loose their accuracy. This subject should be explored further to determine whether such a threshold exists.

Our test case suggests that the efficiency of anchored phylogenomics to enrich specific and universal flowering plant loci is highly promising, and the method could thus simplify next-generation sequencing data collection and sharing across diverse flowering plant groups. Of the 517 anchored phylogenomics loci targeted in this study, approximately 90% produced useable data. In addition, different accessions of the same species provided almost identical sequence and coverage, suggesting that the methodology is reproducible and will provide data that can be re-used and combined in future phylogenetic studies. Other next-generation sequencing approaches in plants have up to now focused on target-enrichment of lineage-specific nuclear loci (de Sousa et al. 2014; Nicholls et al. 2015; Stephens et al. 2015; Heyduk et al. 2016) or on the anonymous and very short markers provided by RADseq (e.g. Eaton and Ree 2013; Escudero et al. 2014; Hipp et al. 2014; Gonzalez 2014; Massatti et al. 2016). Lineage-specific target-enrichment approaches have the advantage of being tailored for the group of interest, and are thus expected to perform better on average. However, this is counter-balanced by the additional cost and time needed to design new probes for every taxonomic group, and the limitations that lineage-specific

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

markers impose on data sharing and reuse across taxonomic groups. RADseq, on the other hand, enables rapid and cost-efficient production of tens to hundreds of thousands of loci in large numbers of individuals without the need for genomic references. The short length of RADseq loci (mostly limited by read-length) makes determination of homology difficult, for instance creating
520 a tradeoff between the number of putative loci retained for analysis and the proportion of loci which are truely orthologous (Rubin et al. 2012; Harvey et al. 2016). High levels of missing data due to uneven coverage, mutation-induced locus-dropout or other causes (ca. 30-80% in published analyses; Mastretta-Yanes et al. 2015; Eaton et al. 2016) creates a similar tradeoff where excluding loci with missing data also signicantly reduces the total number of informative
525 characters in the dataset. Despite this, several studies have now demonstrated that radiations at least as old as 60 Ma can be successfully resolved using lax similarity cutoffs during assembly and inclusion of all loci with at least 4 terminals, which suggests that paralogy and missing data may not be problematic for RADseq in most applications (Gonen et al. 2015; Eaton et al. 2016; Huang and Knowles 2016). One advantage of RADseq compared to Anchored Phylogenomics is
530 the ability to tailor the number of loci to the phylogenetic question and ressources available, while there is a hard limit to the number of loci in hybridization-based approaches (517 loci in our case). On the other hand, data sharing and reuse remains an issue with RADseq because of the use of different enzymes and library preparation methods, the difficulty of assessing orthology at deeper evolutionary timescales and the anonymity of RADseq loci when lacking
535 reference genomes (Ree and Hipp 2015; Harvey et al. 2016). Compared to both lineage-specific approaches and RADseq, the universality and easy comparability of anchored phylogenomics

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

results in significant savings in cost and time whilst simplifying data analysis and sharing.

Phylogenetic and Taxonomic Implications

After nearly two decades of molecular work on the higher-level phylogeny of the sedge family (Cyperaceae; starting with Muasya et al. 1998), one of the most enduring problems has been the placement of tribe Cariceae and the identification of its sister group. This is true despite considerable interest in the evolution, biogeography and ecology of the tribe (more than 140 articles per year since 2010 according to Web of Science), which is derived in part from its exceptional diversity (ca. 2,000 species), global distribution, and peculiar cytology (agmatoploidy, n = 6 to 56, holocentric chromosomes). Lack of knowledge of Cariceae's sister group has important implications, since the accuracy of any morphological, ecological or geographical character reconstruction is affected by outgroup choice, outgroup relationships, and its effects on ingroup topology (Lyons-Weiler et al. 1998; Wheeler 1990; Graham et al. 2002; Wilberg 2015).

Continued work towards inclusion of more informative molecular regions or increasing taxonomic sampling has resulted in good support for seven major lineages within a clade consisting of tribes Cariceae, Dulichieae and Scirpeae (CDS), but relationships between these major lineages and the identity of Cariceae's sister group has remained elusive (Léveillé-Bourret et al. 2014). Previous studies have placed Cariceae sister to a monophyletic Scirpeae (Muasya et al. 2009) or nested within a paraphyletic Scirpeae and sister to either a Trichophorum Clade (Léveillé-Bourret et al. 2014), the genus *Calliscirpus* (Gilmour et al. 2013) or a clade consisting

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of the Scirpus Clade + Zameioscirpus Clade (Jung and Choi 2012). The only consistency has been the poor support for all backbone relationships, a consequence of a rapid radiation (ca. 10 My) followed by long divergence (30–40 My) between major CDS lineages (Escudero et al. 2013; Spalink et al. 2016). Our highly supported results, based on data from hundreds of nuclear genes encompassing hundreds of thousands of base pairs, identify the Trichophorum Clade as sister to Cariceae and are in complete agreement with the results of the most inclusive plastid phylogeny of CDS (Léveillé-Bourret et al. 2014). Such a high congruence between phylogenetic estimates based on the nuclear and plastid genomes gives us confidence in the robustness of the results and confirms the usefulness of targeted-enrichment of conserved nuclear genes for phylogenetic analysis at the tribal level and above in sedges (Cyperaceae). Moreover, since the enrichment probes we used are universal and are flanked by regions of variable evolutionary rates, they could be equally effective for low-level phylogenetic investigation of flowering plants in general.

Our analyses strongly support the paraphyly of tribe Scirpeae, a long-expected result given the likely plesiomorphic nature of its defining characteristics (Goetghebeur 1998). Moreover, the isolated phylogenetic position of *Khaosokia* definitely excludes it from any currently recognized Cyperaceae tribe. These phylogenetic results are congruent with previously identified morphological and embryological variation (Léveillé-Bourret et al. 2014), but a lack of support in previous phylogenetic analyses has prevented taxonomic changes from being made. The robust phylogenetic estimates obtained in this study now provide a solid foundation for a complete revision of the tribal taxonomy of the CDS clade. It is clear that preservation of the highly

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

distinctive Cariceae within a natural and inclusive tribal classification will necessitate the naming of at least three new tribes. Such strongly supported results would probably never have been achieved without genome-scale phylogenetic analyses, which clearly demonstrates the importance of new data acquisition and analysis methodologies in the progress of systematics and taxonomy.

ACKNOWLEDGMENTS

The authors thank the curators of the following herbaria for loaning material and permission to sample specimens for DNA analysis: Jennifer Doubt (CAN), Dr. Paul Catling (DAO), Dr. Paul E. Berry (MICH) and Dr. Luc Brouillet (MT). We also thank Marie-Ève Garon-Labrecque (DAO), Kellina Higgins (MT), Jacques Cayouette (DAO) and Alexandre Bergeron (MT) for having collected or helped with the collection of canadian material. Nguyễn Thị Kim Thanh (HNU), Vũ Anh Tài (HNU) and Jack Regalado (MO) provided field assistance in Vietnam. Michelle Kortyna, Alyssa Bigelow and Sean Holland help with the lab work. Kirby Birch helped with the bioinformatics. Erika Edwards, Mark Fishbein and five anonymous reviewers made helpful comments on the manuscript. This research was conducted as part of the requirements for a Ph.D. at the University of Ottawa (UofO), with support from an Alexander-Graham-Bell Research Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) and an Excellence Scholarship from the UofO to ÉLB. This research was also funded by a NSERC Discovery Grant to JRS for laboratory and field work, and National Geographic Society Research and Exploration Grants (#9035-11, #9441-14) to BAF (Principal Investigator) for field-work in Vietnam. Additional funding was provided by the National Science

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Foundation through awards to ARL and EML, (NSF IIP-1313554), and to EML (NSF DEB-
600 1120516).

LITERATURE

Álvarez I., Wendel J.F. 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29: 417–434.

605 Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U., Cresko W.A., Johnson E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3: e3376.

Baldwin B.G. 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the compositae. Mol. Phylogenet. Evol. 1: 3–16.

610 Barrett C.F., Davis J.I., Leebens-Mack J., Conran J.G., Stevenson D.W.. 2013. Plastid genomes and deep relationships among the commelinid monocot angiosperms. Cladistics 29: 65–87.

Barrett C.F., Specht C.D., Leebens-Mack J., Stevenson D.W., Zomlefer W.B., Davis J.I. 2014. Resolving ancient radiations: can complete plastid gene sets elucidate deep relationships among the tropical gingers (Zingiberales)? Ann. Bot. 113: 119–133.

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

Bayzid M.S., Mirarab S., Boussau B., Warnow T. 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses (N Cellinese, Ed.). PLoS ONE 10: e0129183.

620 Bergsten J. 2005. A review of long-branch attraction. Cladistics 21: 163–193.

Bergthorsson U., Adams K.L., Thomason B., Palmer J.D. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424: 197–201.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Betancur-R. R., Naylor G.J.P., Ortí G. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst. Biol. 63: 257–262.

Blischak P.D., Wenzel A.J., Wolfe A.D. 2014. Gene prediction and annotation in *Penstemon* (Plantaginaceae): a workflow for marker development from extremely low-coverage genome sequencing. Appl. Plant Sci. 2: 1400044.

Brandley M.C., Bragg J.G., Singhal S., Chapple D.G., Jennings C.K., Lemmon A.R., Lemmon E.M., Thompson M.B., Moritz C. 2015. Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian *Eugongylus* group scincid lizards. BMC Evol. Biol. 15: 14.

Buddenhagen C., Lemmon A.R., Lemmon E.M., Bruhl J., Cappa J., Clement W.L., Donoghue M.J., Edwards E.J., Hipp A., Kortyna M., Mitchell N., Prychid C., Simmons M.P., Soltis P., Wanke S., Mast A. In prep.. Anchored Phylogenomics of Angiosperms I: Assessing the Robustness of Phylogenetic Estimates. bioRxiv doi: 10.1101/086298

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17: 540–552.

Chamala S., García N., Godden G.T., Krishnakumar V., Jordon-Thaden I.E., Smet R.D., Barbazuk W.B., Soltis D.E., Soltis P.S. 2015. MarkerMiner 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. Appl. Plant Sci. 3: 1400115.

Chiari Y., Cahais V., Galtier N., Delsuc F. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10: 65.

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

Comer J.R., Zomlefer W.B., Barrett C.F., Stevenson D.W., Heyduk K., Leebens-Mack J.H. 2016. Nuclear phylogenomics of the palm subfamily Arecoideae (Arecaceae). Mol. Phyogenet. Evol. 97: 32–42.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

650 Croizat L. 1952. Manual of phytogeography. The Hague (Netherlands): W. Junk. 587 p.

Cronn R., Knaus B.J., Liston A., Maughan P.J., Parks M., Syring J.V., Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99: 291–311.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2: e68.

655 Eaton D.A.R., Ree R.H. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol. 62: 689–706.

Escudero M., Hipp A. 2013. Shifts in diversification rates and clade ages explain species richness in higher-level sedge taxa (Cyperaceae). Am. J. Bot. 100: 2403–2411.

Escudero M., Eaton D.A.R., Hahn M., Hipp A.L. 2014. Genotyping-by-sequencing as a tool to
660 infer phylogeny and ancestral hybridization: a case study in *Carex* (Cyperaceae). Mol. Phylogenet. Evol. 79: 359–367.

Eytan R.I., Evans B.R., Dornburg A., Lemmon A.R., Lemmon E.M., Wainwright P.C., Near T.J. 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. BMC Evol. Biol. 15: 20.

665 de Sousa F., Bertrand Y.J.K., Nylinder S., Oxelman B., Eriksson J.S., Pfeil B.E. 2014. Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. PLoS ONE 9: e109704.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary
670 timescales. Syst. Biol. 61: 717–726.

Feliner G.N., Rosselló J.A. 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Mol. Phylogenet. Evol. 44: 911–919.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively
675 misleading. Syst. Biol. 27: 401–410.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783–791.

Fisher A.E., Hasenstab K.M., Bell H.L., Blaine E., Ingram A.L., Columbus J.T. 2016. Evolutionary history of chloridoid grasses estimated from 122 nuclear loci. Mol. Phylogenet. Evol. 105:1–14.

Gardner M.G., Fitch A.J., Bertozzi T., Lowe A.J. 2011. Rise of the machines - recommendations for ecologists when using next generation sequencing for microsatellite development. Mol. Ecol. Resour. 11: 1093–1101.

Gilmour C.N., Starr J.R., Naczi R.F.C. 2013. Calliscirpus, a new genus for two narrow endemics of the California Floristic Province, *C. criniger* and *C. brachythrix* sp. nov. (Cyperaceae). Kew Bull. 68: 84–105.

Goetghebeur P. 1998. Cyperaceae. The families and genera of vascular plants. In: Kubitzki, K. (Ed.) Flowering plants, Monocotyledons□: Alismatanae and Commelinanae (except Gramineae). New York: Springer. p. 141–190.

Gonen S., Bishop S.C., Houston R.D. 2015. Exploring the utility of cross-laboratory RAD-sequencing datasets for phylogenetic analysis. BMC Res. Notes 8: 299.

Gonzalez L.A. 2014. Phylogenetics and mating system evolution in the Southern South American *Valeriana* (Valerianaceae). M.Sc. Thesis, University of New Orleans. 46 p.

Graham S.W., Olmstead R.G., Barrett S.C.H. 2002. Rooting phylogenetic trees with distant outgroups: a case study from the Commelinoid monocots. Mol. Biol. Evol. 19: 1769–1781.

Greilhuber J, Borsch T, Müller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest angiosperm genomes found in lentibulariaceae, with chromosomes of bacterial size. Plant Biol. 8: 770–777.

Hedtke S.M., Townsend T.M., Hillis D.M. 2006. Resolution of phylogenetic conflict in large data sets by increased sampling. Syst. Biol. 55: 522–529.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Heyduk K., Trapnell D.W., Barrett C.F., Leebens-Mack J. 2016. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. Biol. J. Linn. Soc. 117: 106–120.

Hillis D.M. 1987. Molecular versus morphological approaches to systematics. Ann. Rev. Ecol. Syst. 18: 23–42.

Hillis D.M., Wiens J.J. 2000. Molecules versus morphology in systematics: conflicts, artifacts, and misconceptions. Phylogenetic analysis of morphological data. Washington (United-States): Smithsonian Institution Press, 1–19.

Hipp A.L., Rothrock P., Roalson E. 2009. The evolution of chromosome arrangements in *Carex* (Cyperaceae). Bot. Rev. 75: 96–109.

Hipp A.L., Eaton D.A.R., Cavender-Bares J., Fitzek E., Nipper R., Manos P.S. 2014. A framework phylogeny of the american oak clade based on sequenced RAD data (S Joly, Ed.). PLoS ONE 9: e93975.

Hoot S.B., Taylor W.C. 2001. The utility of nuclear ITS, a LEAFY homolog intron, and chloroplast atpB-rbcL spacer region data in phylogenetic analyses and species delimitation in *Isoëtes*. Am. Fern J. 91: 166–177.

Huang H., Knowles L.L. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst. Biol. 65: 357–365.

Hughes C.E., Eastwood R.J., Donovan Bailey C. 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. Philos. Trans. R. Soc. B: Biol. Sci. 361: 211–225.

Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett N.J. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. in Plant Sci. 4:1600016.

Johnson M.T.J., Carpenter E.J., Tian Z., Bruskiewich R., Burris J.N., Carrigan C.T., Chase M.W., Clarke N.D., Covshoff S., dePamphilis C.W., Edger P.P., Goh F., Graham S., Greiner S., Hibberd J.M., Jordon-Thaden I., Kutchan T.M., Leebens-Mack J., Melkonian M., Miles N.,

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Myburg H., Patterson J., Pires J.C., Ralph P., Rolf M., Sage R.F., Soltis D., Soltis P., Stevenson D., Stewart C.N., Surek B., Thomsen C.J.M., Villarreal J.C., Wu X., Zhang Y., Deyholos M.K., Wong G.K.-S. 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. PLoS ONE 7: e50226.

Jung J., Choi H.-K. 2012. Recognition of two major clades and early diverged groups within the subfamily Cyperoideae (Cyperaceae) including Korean sedges. J. Plant Res. 126: 1–15.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30: 772–780.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649.

Knoop V. 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr. Genet. 46: 123–139.

Kobert K., Salichos L., Rokas A., Stamatakis A. 2015. Computing the internode certainty and related measures from partial gene trees. bioRxiv.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29: 1695–1701.

Lanfear R., Calcott B., Kainer D., Mayer C., Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14: 82.

Leaché A.D., Wagner P., Linkem C.W., Böhme W., Papenfuss T.J., Chong R.A., Lavin B.R., Bauer A.M., Nielsen S.V., Greenbaum E., Rödel M.-O., Schmitz A., LeBreton M., Ineich I., Chirio L., Ofori-Boateng C., Eniang E.A., Baha El Din S., Lemmon A.R., Burbrink F.T. 2014. A hybrid phylogenetic–phylogenomic approach for species tree estimation in African *Agama* lizards with applications to biogeography, character evolution, and diversification. Mol. Phylogenet. Evol. 79: 215–230.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

755 Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61: 727–744.

Lemmon E.M., Lemmon A.R. 2013. High-throughput genomic data in systematics and phylogenetics. Ann. Rev. Ecol. Evol. Syst. 44: 99–121.

Léveillé-Bourret É., Gilmour C.N., Starr J.R., Naczi R.F.C., Spalink D., Systma K.J. 2014.
760 Searching for the sister to sedges (*Carex*): resolving relationships in the Cariceae-Dulichieae-Scirpeae clade (Cyperaceae). Bot. J. Linn. Soc. 176: 1–21.

Léveillé-Bourret É., Donadío S., Gilmour C.N., Starr J.R. 2015. *Rhodoscirpus* (Cyperaceae: Scirpeae), a new South American sedge genus supported by molecular, morphological, anatomical and embryological data. Taxon 64: 931–944.

765 Lyons-Weiler J., Hoelzer G.A, Tausch R.J. 1998. Optimal outgroup analysis. Biol. J. Linn. Soc. 64: 493–511.

Ma P.-F., Zhang Y.-X., Zeng C.-X., Guo Z.-H., Li D.-Z. 2014. Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). Syst. Biol. 63: 933–950.

770 Mandel J.R., Dikow R.B., Funk V.A., Masalia R.R., Staton S.E., Kozik A., Michelmore R.W., Rieseberg L.H., Burke J.M. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. App. Plant Sci. 2: 1300085.

Mason-Gamer R.J., Weil C.F., Kellogg E.A. 1998. Granule-bound starch synthase: structure,
775 function, and phylogenetic utility. Mol. Biol. Evol. 15: 1658–1673.

Massatti R., Reznicek A.A., Knowles L.L. 2016. Utilizing RADseq data for phylogenetic analysis of challenging taxonomic groups: a case study in *Carex* sect. *Racemosae*. Am. J. Bot. 103: 337–347.

Mastretta-Yanes A., Arrigo N., Alvarez N., Jorgensen T.H., Piñero D., Emerson B.C. 2015.
780 Restriction site-associated DNA sequencing, genotyping error estimation and de novo

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

assembly optimization for population genetic inference. Molecular Ecology Resources 15: 28–41.

Meiklejohn K.A., Faircloth B.C., Glenn T.C., Kimball R.T., Braun E.L. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. Syst. Biol.. 65:612–627.

Meyer M., Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010: pdb.prot5448-prot5448.

Miller M.A., Pfeiffer W., Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE). New Orleans, 1–8.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30: i541–i548.

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

Mirarab S., Bayzid M.S., Warnow T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65: 366–380.

Mitchell N., Lewis P.O., Lemmon E.M., Lemmon A.R., Holsinger K.E. 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. Am. J. Bot.. Online first.

Muasya A.M., Simpson D.A., Chase M.W., Chulham A. 1998. An assessment of suprageneric phylogeny in Cyperaceae using rbcL DNA sequences. Plant Systematics and Evolution 211: 257–271.

Muasya A.M., Simpson D.A., Verboom G.A., Goetghebeur P., Naczi R.F.C., Chase M.W., Smets E. 2009. Phylogeny of Cyperaceae based on DNA sequence data: current progress and future prospects. Bot. Rev. 75: 2–21.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32: 268–274.

Nicholls J.A., Pennington R.T., Koenen E.J.M., Hughes C.E., Hearn J., Bunnefeld L., Dexter K.G., Stone G.N., Kidner C.A. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). Frontiers in Plant Science 6: 1–20.

Ogilvie H.A., Heled J., Xie D., Drummond A.J. 2016. Computational performance and statistical accuracy of *BEAST and comparisons with other methods. Syst. Biol. 65: 381–396.

Palmer J.D., Herbon L.A. 1988. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. Journal of Molecular Evolution 28: 87–97.

Paradis E., Claude J., Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.

Pellicer J., Fay M.F., Leitch I.J. 2010. The largest eukaryotic genome of them all? Bot. J. Linn. Soc. 164: 10–15.

Peloso P.L.V., Frost D.R., Richards S.J., Rodrigues M.T., Donnellan S., Matsui M., Raxworthy C.J., Biju S.D., Lemmon E.M., Lemmon A.R., Wheeler W.C. 2015. The impact of anchored phylogenomics and taxon sampling on phylogenetic inference in narrow-mouthed frogs (Anura, Microhylidae). Cladistics Early View: 28.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough (D Penny, Ed.). PLoS Biology 9: e1000602.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature.

Pyron R.A., Hendry C.R., Chou V.M., Lemmon E.M., Lemmon A.R., Burbrink F.T. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). Mol. Phylogenet. Evol. 81: 221–231.

Pyron R.A., Hsieh F.W., Lemmon A.R., Lemmon E.M., Hendry C.R. 2016. Integrating phylogenomic and morphological data to assess candidate species-delimitation models in brown and red-bellied snakes (*Storeria*). Zoological Journal of the Linnean Society Early View.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Ree R.H., Hipp A.L. 2015. Inferring phylogenetic history from restriction site associated DNA (RADseq). In: Hörandl E. Appelhans M.S, ed. Regnum Vegetabile. Next-generation sequencing in plant systematics. Königstein (Germany): Koeltz Scientific Books, 181–204.

Richardson A.O., Palmer J.D. 2006. Horizontal gene transfer in plants. Journal of Experimental Botany 58: 1–9.

Rieseberg L.H., Carter R., Zona S. 1990. Molecular tests of the hypothesized hybrid origin of two diploid *Helianthus* species (Asteraceae). Evolution 44: 1498.

Rokyta D.R., Lemmon A.R., Margres M.J., Aronow K. 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). BMC Genomics 13: 312.

Ruane S., Raxworthy C.J., Lemmon A.R., Lemmon E.M., Burbrink F.T. 2015. Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: an empirical study on Malagasy pseudoxyrhophiine snakes. BMC Evol. Biol. 15: 14.

Rubin B.E.R., Ree R.H., Moreau C.S. 2012. Inferring phylogenies from RAD sequence data (S-O Kolokotronis, Ed.). PLoS ONE 7: e33394.

Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31: 1261–1271.

Sarkar I.N., Egan M.G., Coruzzi G., Lee E.K., DeSalle R. 2008. Automated simultaneous analysis phylogenetics (ASAP): an enabling tool for phylogenomics. BMC Bioinformatics 9: 103.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Schmickl R., Liston A., Zeisek V., Oberlander K., Weitemier K., Straub S.C.K., Cronn R.C., Dreyer L.L., Suda J. 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). Mol. Ecol. Resour. 16:1124–1135.

Scotland R.W., Olmstead R.G., Bennett J.R. 2003. Phylogeny reconstruction: the role of morphology. Syst. Biol. 52: 539–548.

Simmons M.P., Freudenstein J.V. 2011. Spurious 99% bootstrap and jackknife support for unsupported clades. Mol. Phylogenet. Evol. 61: 177–191.

Spalink D., Drew B.T., Pace M.C., Zaborsky J.G., Starr J.R., Cameron K.M., Givnish T.J., Sytsma K.J. 2016. Biogeography of the cosmopolitan sedges (Cyperaceae) and the area-richness correlation in plants. Journal of Biogeography Online Advance Access.

Stamatakis A., Blagojevic F., Nikolopoulos D.S., Antonopoulos C.D. 2007. Exploring new search algorithms and hardware for phylogenetics; RAxML meets the IBM cell. The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology 48: 271–286.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

Starr J.R., Naczi R.F.C., Chouinard B.N. 2009. Plant DNA barcodes and species resolution in sedges (Carex, Cyperaceae). Molecular Ecology Resources 9: 151–163.

Starr J.R., Janzen F.H., Ford B.A. 2015. Three new, early diverging *Carex* (Cariceae, Cyperaceae) lineages from East and Southeast Asia with important evolutionary and biogeographic implications. Mol. Phylogenet. Evol. 88: 105–120.

Stephens J.D., Rogers W.L., Heyduk K., Cruse-Sanders J.M., Determann R.O., Glenn T.C., Malmberg R.L. 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. Mol. Phylogenet. Evol. 85: 76–87.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Straub S.C.K., Parks M., Weitemier K., Fishbein M., Cronn R.C., Liston A. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. 99: 349–364.

Straub S.C.K., Moore M.J., Soltis P.S., Soltis D.E., Liston A., Livshultz T. 2014. Phylogenetic signal detection from an ancient rapid radiation: effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. Mol. Phylogenet. Evol. 80: 169–185.

Swofford D.L. 2002. PAUP: Phylogenetic analysis using parsimony *and other methods: Version 4.0 beta. Sunderland, Massachussetts: Sinauer Associates.

Taberlet P., Gielly L., Pautou G., Bouvet J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Molecular Biology 17: 1105–1109.

Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61: 835–849.

Warnow T. 2015. Concatenation analyses in the presence of incomplete lineage sorting. PLoS Curr. 7.

Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2:1400042.

Wen J., Xiong Z., Nie Z.-L., Mao L., Zhu Y., Kan X.-Z., Ickert-Bond S.M., Gerrath J., Zimmer E.A., Fang X.-D. 2013. Transcriptome sequences resolve deep relationships of the grape family. PLoS ONE 8: e74394.

Wheeler W.C. 1990. Nucleic acid sequence phylogeny and random outgroups. Cladistics 6: 363–367.

White T.J., Bruns T., Lee S., Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis M.A., Gelfand D.H., Sninsky J.J., White T.J. (Eds.) PCR protocols: a guide to methods and applications. New York: Academic Press. p. 315–322.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Whitfield J.B., Lockhart P.J. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22: 258–265.

Wicke S., Schneeweiss G.M. 2015. Next generation organellar genomics. In: Hörandl E, Appelhans M.S. e(Ed.) Regnum Vegetabile. Next-generation sequencing in plant systematics. Königstein (Germany): Koeltz Scientific Books. p. 9–50.

Wilberg E.W. 2015. What's in an outgroup? The impact of outgroup choice on the phylogenetic position of *Thalattosuchia* (Crocodylomorpha) and the origin of Crocodyliformes. Syst. Biol. 64: 621–637.

Xi Z., Ruhfel B.R., Schaefer H., Amorim A.M., Sugumaran M., Wurdack K.J., Endress P.K., Matthews M.L., Stevens P.F., Mathews S., Davis C.C. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc. Natl. Acad. Sci. 109: 17519–17524.

Xi Z., Liu L., Davis C.C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92: 63–71.

Xia X., Xie Z., Salemi M., Chen L., Wang Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26: 1–7.

Xia X., Lemey P. 2009. Assessing substitution saturation with DAMBE. In: Salemi M., Vandamme A.-M., Lemey P. (Eds.) The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge: Cambridge University Press. p. 611–626.

Xia X. 2013. DAMBE5: A Comprehensive Software Package for Data Analysis in Mol. Biol. Evol. 30: 1720–1728.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TABLE 1. Comparison of branch support measures for selected backbone branches (A-E; see Fig. 1). Branch length (expected changes per site in ML), internode certainty (ICA) based on estimated ML gene trees, ASTRAL bootstrap (BS), number of unambiguous synapomorphies, number of loci with positive and negative partitionned Bremer support, total Bremer support and proportion of hidden Bremer support.

Branch	Branch length	ICA	BS	Unambiguous synapomorphies	Loci positive	Loci negative	Total Bremer	Proportion hidden Bremer
A	0.0047	0.180	100	1404	296	86	790	(43%)
B	0.0023	0.066	99	646	209	90	327	(66%)
C	0.0014	0.006	68	304	195	203	95	(43%)
D	0.0058	0.380	100	1082	299	48	921	(57%)
E	0.0017	0.024	100	364	195	202	137	(42%)

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FIGURE 1. Best phylogenetic hypotheses for the CDS Clade, with the species tree estimated by ASTRAL using 461 NGS nuclear loci on the left, and the ML tree estimated using *matK* + *ndhF* + ETS-1f on the right (arrows indicate topological differences with the ASTRAL species tree). The smaller trees on either side represent relative ML branch lengths for each dataset. Branches without values have 100% support for all measures used. When at least one measure was <100%, the support values are reported above branches as follows: ASTRAL tree (bold italics: local posterior probabily, normal text: ASTRAL multilocus bootstrap/ML bootstrap/MP jackknife) and ML tree (ML bootstrap/MP bootstrap). An asterisk (*) indicates 100% bootstrap support, and a dash (-) indicates less than 50% bootstrap support. Branch width is a function of support in parsimony. Letters under branches refer to clades in Table 1. Legend; Out: outgroups, DUL: Dulichieae, Kha: *Khaosokia*, Cal: Calliscirpus Clade, Zam: Zameioscirpus Clade, Sci: Scirpus Clade, Tri, Trichophorum Clade, CAR: Cariceae, Eri: Eriophorum Clade, Min: Minor Carex Alliance, Vig: *Carex* subg. *Vignea*, Uni: Unispicate Carex Clade, Car: Core Carex Clade. Scirpeae = Cal + Zam + Sci + Tri.

FIGURE 2. Negative relationship between average ML bootstrap of estimated gene trees and Robinson-Foulds distance between that gene tree and the ASTRAL species tree. Regression line estimated by standard linear regression.

FIGURE 3. Relative frequency of bipartitions of estimated gene trees as a function of their ML bootstrap support. Bipartitions compatible with the ASTRAL species tree in blue, incompatible in yellow, showing that most gene tree bipartitions are compatible with the species tree beyond about 80% BS.

FIGURE 4. Results of reduced analyses, showing the distribution of Robinson-Foulds distance between reduced analyses tree and the best ASTRAL species tree, as a function of analytic method and number of loci. Note that since all trees were bifurcating, the Robinson-Foulds distance is equivalent to the proportion of conflicting nodes.

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

965 APPENDIX 1. Samples used for anchored phylogenomics.

Species	DNA number	NGS number	Collectors	Collection number	Herb.	Origin
Amphiscirpus nevadensis (S.Watson) Oteng-Yeb.	STA2640	I8317	Starr, Julian R.	1301-08	CAN	Canada, British Columbia
Calliscirpus brachythrix C.N.Gilmour	STA2625	I8333	Starr, Julian R.	07-037	CAN	United-States, California
Calliscirpus criniger (A.Gray) C.N.Gilmour et al.	STA2629	I8311	Starr, Julian R.	10S-055 (P55-10)	CAN	United-States, California
Carex atrivaginata Nelmes	STA2647	I8330	Ford & al.	1241A	WIN	Vietnam, Lao Cai
Carex bavicola Raymond	STA2644	I8327	Ford & al.	1220	WIN	Vietnam, Hanoi
Carex camptoglochin V.I.Krecz.	STA2656	I8334	Starr & al.	10-001	CAN	Argentina, Tierra del Fuego
Carex canescens L.	STA2654	I8332	Starr	10S-005 (P5-19)	CAN	United States, New Mexico
Carex capitata Sol.	STA2653	I8323	Starr & Villaverde	10-023 (P18-27)	CAN	Argentina, Santa Cruz
Carex dimorpholepis Steud.	STA2645	I8328	Ford & al.	1240A	WIN	Vietnam, Lao Cai
Carex filicina Nees	STA2651	I8321	Ford & al.	1229	WIN	Vietnam, Lao Cai
Carex hypolytroides Ridl.	STA2442	I8306	Ford & al.	1255A	WIN	Vietnam, Lao Cai
Carex kucyniakii Raymond	STA2639	I8316	Ford & al.	1258A	WIN	Vietnam, Lao Cai
Carex microglochin Wahlenb.	STA2658	I8336	Starr	10S-035 (P35-1 & P35-4)	CAN	United States, Colorado
Carex myosuroides Vill.	STA2630	I8312	Starr	10S-012 (P12-9)	CAN	United States, New Mexico
Carex myosurus Nees	STA2636	I8325	Ford & al.	1224	WIN	Vietnam, Lao Cai
Carex nardina (Hornem.) Fr.	STA2657	I8335	Starr	10S-051 (P51-20)	CAN	United States, Utah
Carex phleoides Cav.	STA2652	I8322	Starr & Villaverde Hidalgo	10-026	CAN	Argentina, Neuquén
Carex plantaginea Lam.	STA0178	I8304	Bakowski	96-174	WIN	Canada, Ontario
Carex pulicaris L.	STA0105	I8303	Starr & Scott	98001	FHO	England, North Yorkshire
Carex siderosticta Hance	STA2681	I8337	Léveillé-Bourret	545	CAN	Garden

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Carex speciosa Kunth	STA2646	I8329	Ford & al.	1236A	WIN	Vietnam, Lao Cai
Khaosokia caricoides D.A.Simpson et al.	STA0387B	I8305	Middleton & al.	4071	MICH	Thailand, Surat Thani
Dulichium arundinaceum Pers. **var.** ***arundinaceum***	STA2682	I8338	Starr	16-001	OTT	Canada, Québec
Eleocharis obtusa (Willd.) Schult.	STA2632	I8314	Bergeron	12-272	MT	Canada, Québec
Erioscirpus comosus (Wall.) Palla	STA2648	I8331	Ford & al.	1269C	WIN	Vietnam, Ha Giang
Eriophorum angustifolium Honck.	STA2631	I8313	Starr	10S-011 (P11-2V)	CAN	United States, New Mexico
Eriophorum vaginatum **subsp.** ***spissum*** (Fern.) Hultén	STA2634	I8324	Léveillé-Bourret	632	DAO	Canada, Ontario
Eriophorum virginicum L.	STA2608	I8308	Léveillé-Bourret	633	DAO	Canada, Ontario
Scirpus atrovirens Willd.	STA2567	I8307	Léveillé-Bourret	610	DAO	Canada, Québec
Scirpus atrovirens Willd.	STA2638	I8326	Léveillé-Bourret	609	DAO	Canada, Québec
Scirpus cyperinus (L.) Kunth	STA2609	I8309	Léveillé-Bourret	634	DAO	Canada, Ontario
Scirpus pendulus Muhl.	STA2643	I8319	Léveillé-Bourret	611	DAO	Canada, Québec
Scirpus rosthornii Diels	STA2650	I8320	Ford & al.	1223A	WIN	Vietnam, Lao Cai
Trichophorum alpinum (L.) Pers.	STA2633	I8315	Garon-Labrecque	129	MT	Canada, Northwest Territories
Trichophorum cespitosum (L.) Schur	STA2628	I8310	Garon-Labrecque	130	MT	Canada, Northwest Territories

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

APPENDIX 2. Samples and Genbank accession numbers used in the comparative Sanger-based analyses.

Species	DNA number	Collectors	Collection number	Herb.	Origin	*matK*	*ndhF*	ETS-1f
Amphiscirpus nevadensis (S.Watson) Oteng-Yeb.	STA2141	Hudson	5177	CAN	Canada, Saskatchewan	JX065075	JX074631	KP705256
Calliscirpus brachythrix C.N.Gilmour	STA2625	Ahart & Oswald	5099	CHS	United States, California	JX065078	JX074634	JX065112
Calliscirpus criniger (A.Gray) C.N.Gilmour et al.	STA2629	Chambers	2973	DAO	United States	JX074655	KJ513488	JX065099
Carex atrivaginata Nelmes	STA2419	Ford & al.	1230A	WIN	Vietnam, Lao Cai	newly submitted	newly submitted	newly submitted
Carex bavicola Raymond	STA2389	Ford & al.	1215B	WIN	Vietnam, Hanoi	KP273672	KP273726	KP273600
Carex camptoglochin V.I.Krecz.	STA0027	Molau & al.	2329	GB	Ecuador, Chimborazo	KJ513584	KJ513492	AY244520
Carex canescens L.	-	Kaantonen	156/94	H	Finland	KP980061	-	-
Carex canescens L.	-	Bond	s.n.	MTMG	Canada, Québec	-	-	AY757384
Carex capitata Sol.	STA2653	Starr & Thibeault	6016	CAN	United States	KJ513585	KJ513493	-
Carex capitata Sol.	STA1411	Ford	02379	WIN	Canada	-	-	DQ115119
Carex dimorpholepis Steud.	STA2645	[not available]	MAK accession no. 99052601	MAK	Japan, Nara Pref.	AB079435	AB079422	-
Carex filicina Nees	STA2459	Ford & al.	1247A	WIN	Vietnam, Lao Cai	KP273682	KP273736	KP273608
Carex hypolytroides Ridl.	STA2442	Ford & al.	1255A	WIN	Vietnam, Lao Cai	KP273688	KP273742	KP273610
Carex kucyniakii Raymond	STA2351	Ford & al.	1261A	WIN	Vietnam, Lao Cai	KP273693	KP273747	KP273615

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Carex microglochin Wahlenb.	STA2658	Starr & al.	10-008 (P5-9)	CAN	Argentina, Tierra del Fuego	KP273698	KP273752	-
Carex microglochin Wahlenb.	STA0106	Starr & Scott	98017	FHO	Scotland, County of Perth	-	-	AY244518
Carex myosuroides Vill.	B1518	Jones	146	UBC	Canada, British Columbia	KJ513622	KJ513529	-
Carex myosuroides Vill.	STA0101	Playford & al.	9084	FHO	France, Hautes-Alpes/Savoie	-	-	AH012966
Carex myosurus Nees	STA2456	Ford & al.	1246A	WIN	Vietnam, Lao Cai	KP273700	KP273754	KP273620
Carex nardina (Hornem.) Fr.	-	Aiken & al.	86-091	CAN	Canada, Nunavut	FJ548120	-	-
Carex nardina (Hornem.) Fr.	STA0839	Ford & al.	02230	WIN	Canada, Manitoba	-	-	DQ115221
Carex phleoides Cav.	-	Danton	G-(1377)-1142	[not available]	[not available]	-	AM999972	-
Carex phleoides Cav.	STA0034	Vann	s.n.	FHO	Chile, Chiloé	-	-	AH010381
Carex plantaginea Lam.	STA0178	Waterway	2000.002	MTMG	Canada, Québec	-	-	AY757674
Carex pulicaris L.	STA0105	Starr & Scott	98001	FHO	England, North Yorkshire	KJ513590	KJ513576	AY242019
Carex siderosticta Hance	STA0733	Léveillé-Bourret	545	CAN	Garden	KJ513592	KJ513499	-
Carex siderosticta Hance	-	Waterway	2004.268	MTMG	MTMG	-	-	DQ998892
Carex speciosa Kunth	STA2417	Ford & al.	1236A	WIN	Vietnam, Lao Cai	KP273706	KP273760	KP273625
Khaosokia caricoides D.A.Simpson et al.	STA0387 B	Middleton & al.	4071	MICH	Thailand, Surat Thani	JX065087	JX074643	-
Dulichium arundinaceum	STA0154	Ford & Punter	94233	FHO	Canada, Manitoba	JX065083	JX074639	-

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Pers. **var.** ***arundinaceum***								
Dulichium arundinaceum Pers. **var.** ***arundinaceum***	STA2469	Bergeron & al.	81113	CAN	Canada, Québec	-	-	KP705281
Eleocharis acicularis (Willd.) Schult.	d484	Fields	2583	WIS	United States, Wisconsin	KJ513595	KJ513502	-
Erioscirpus comosus (Wall.) Palla	STA2092	Hing & al.	22413	A	China, Yunnan	KJ513619	KJ513526	-
Erioscirpus comosus (Wall.) Palla	-	Ikeda & al.	4032	TI	Nepal	-	-	KM462231
Eriophorum angustifolium Honck.	STA1777	Scoggan	10947	CAN	Canada	KJ513597	KJ513504	
Eriophorum angustifolium Honck.	STA2547	Keleher	755	CAN	Canada	-	-	KP705276
Eriophorum vaginatum **subsp.** ***spissum*** (Fern.) Hultén	STA1804	Porsild	12	CAN	Canada	KJ513614	KJ513521	-
Eriophorum vaginatum **subsp.** ***spissum*** (Fern.) Hultén	STA2607	Léveillé-Bourret	632	DAO	Canada	-	-	KP705278
Eriophorum virginicum L.	STA1807	Dickson & Brunton	3214	CAN	Canada	KJ513617	KJ513524	-
Eriophorum virginicum L.	STA2608	Léveillé-Bourret	633	DAO	Canada	-	-	KP705269
Scirpus atrovirens Willd.	dS043	Spalink	186	WIS	United States, Ohio	KJ513630	KJ513537	-
Scirpus cyperinus (L.) Kunth	STA1777	Lindsey	1025	CAN	Canada, Ontario	JX065092	JX074648	-
Scirpus pendulus Muhl.	STA1878	Cruise	1388	CAN	Canada	KJ513649	KJ513556	-
Scirpus pendulus Muhl.	STA2569	Léveillé-Bourret	611	DAO	Canada	-	-	KP705270

bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Scirpus rosthornii Diels	STA2437	Ford & al.	1260A	WIN	Vietnam, Lao Cai	newly submitted	-	newly submitted
Trichophorum alpinum (L.) Pers.	STA1815	Spetzman	4941	CAN	United States, Alaska	JX065093	JX074649	KP705266
Trichophorum cespitosum (L.) Schur	STA1819	Aiken & Iles	02-048	CAN	Canada, Nunavut	KJ513657	KJ513564	KP705267

970

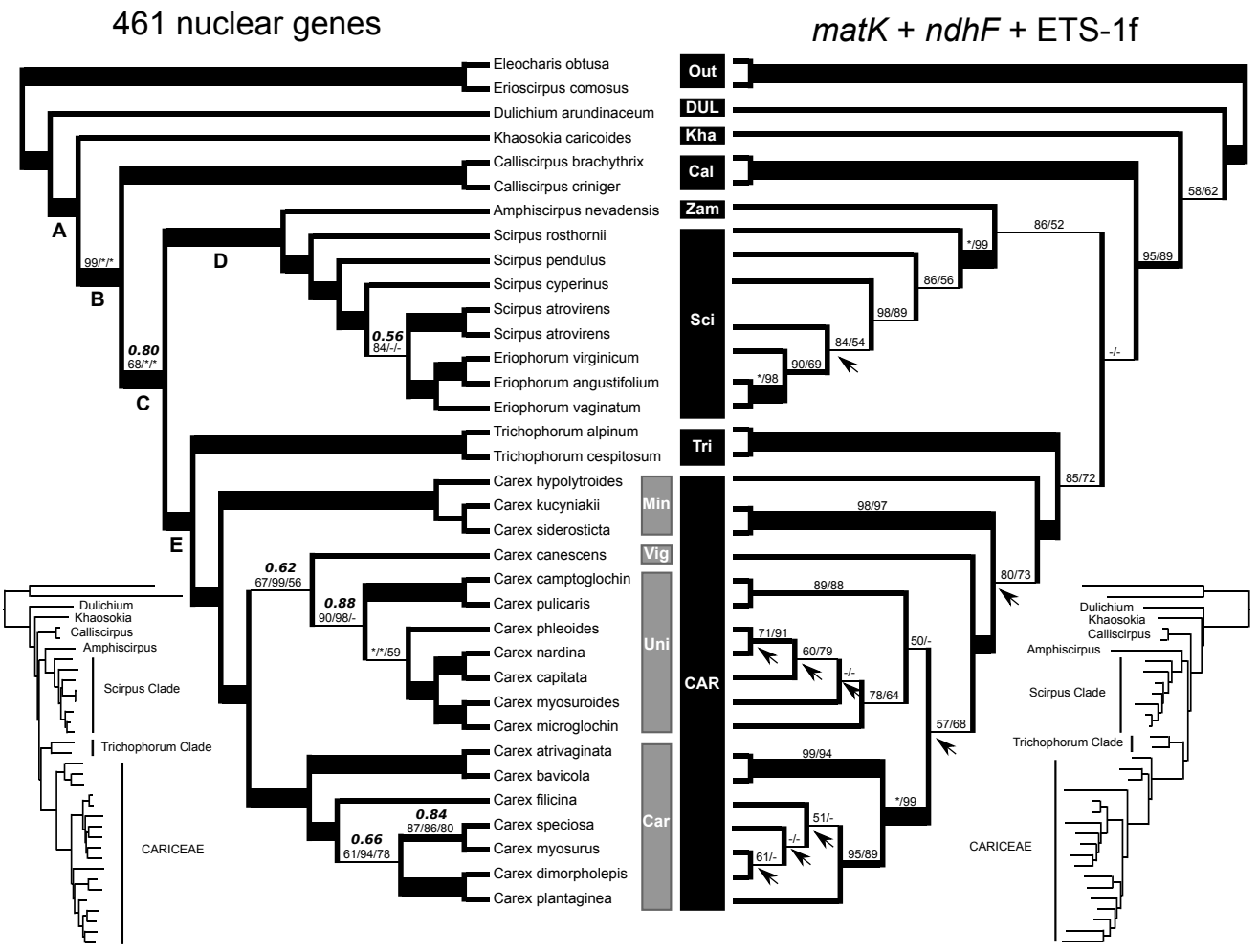
bioRxiv preprint doi: https://doi.org/10.1101/110296; this version posted February 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

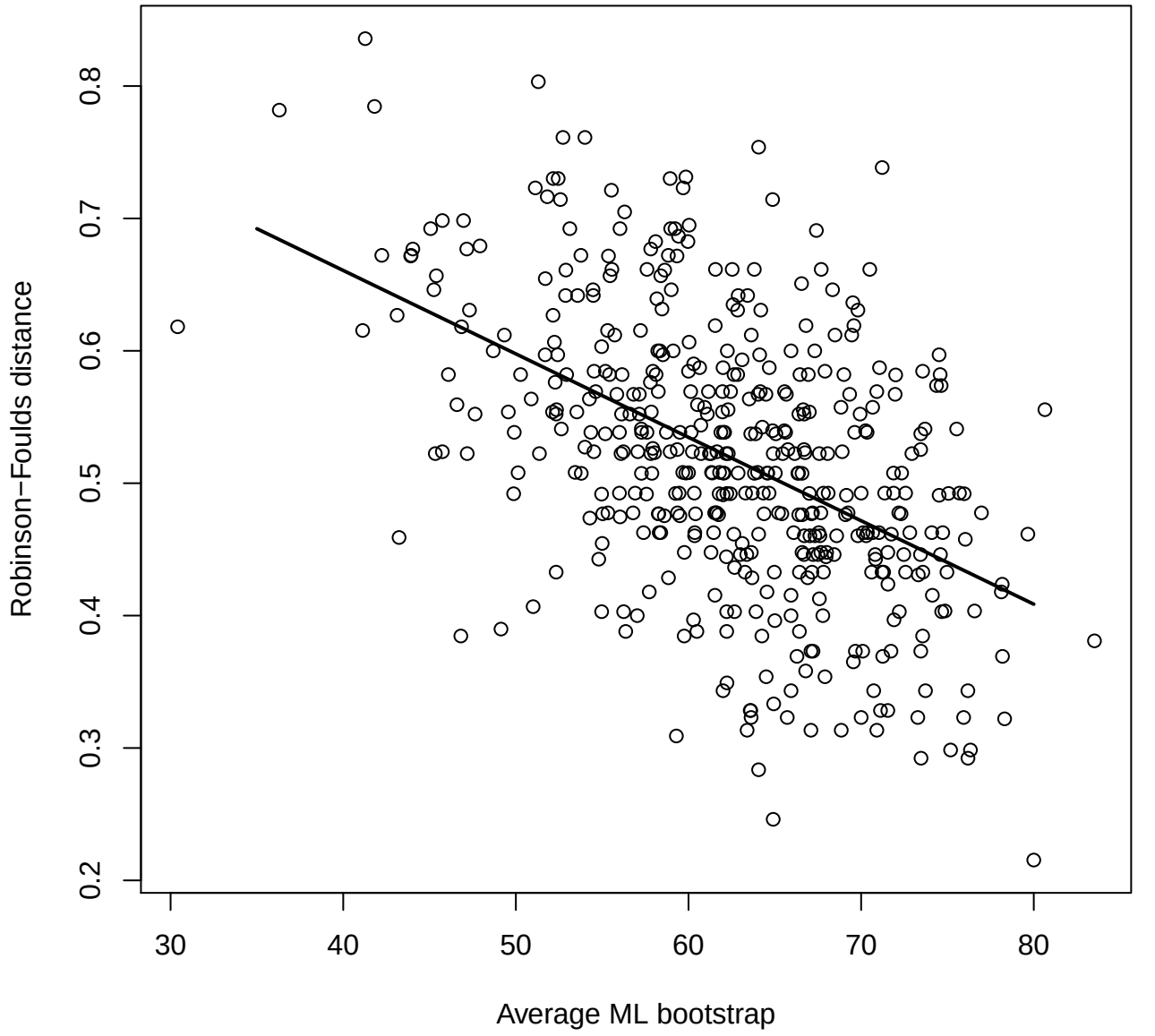
ONLINE APPENDIX 1. Characteristics of loci used in anchored phylogenomics analyses after masking and trimming of alignments. AVE: average value for all loci, MIN: minimum value for all loci, MAX: maximum value for all loci, Locus#: locus identification number, New#: new number assigned to locus for this analysis (necessary because of paralogs), cov: average coverage across all species, ntax: number of taxa possessing the locus, flank%: percentage of total locus length estimated to be in the "flanking region", aveC: average number of copies of locus across species, before orthology filtering, maxC: maximum number of distinct copies of locus, before orthology filtering, len: alignment length, var: number of variable characters, pi: number of parsimony informative characters, GC%: percentage of GC across all non-gapped bases, amb%: percentage of ambiguous bases, miss%: percentage of missing data.

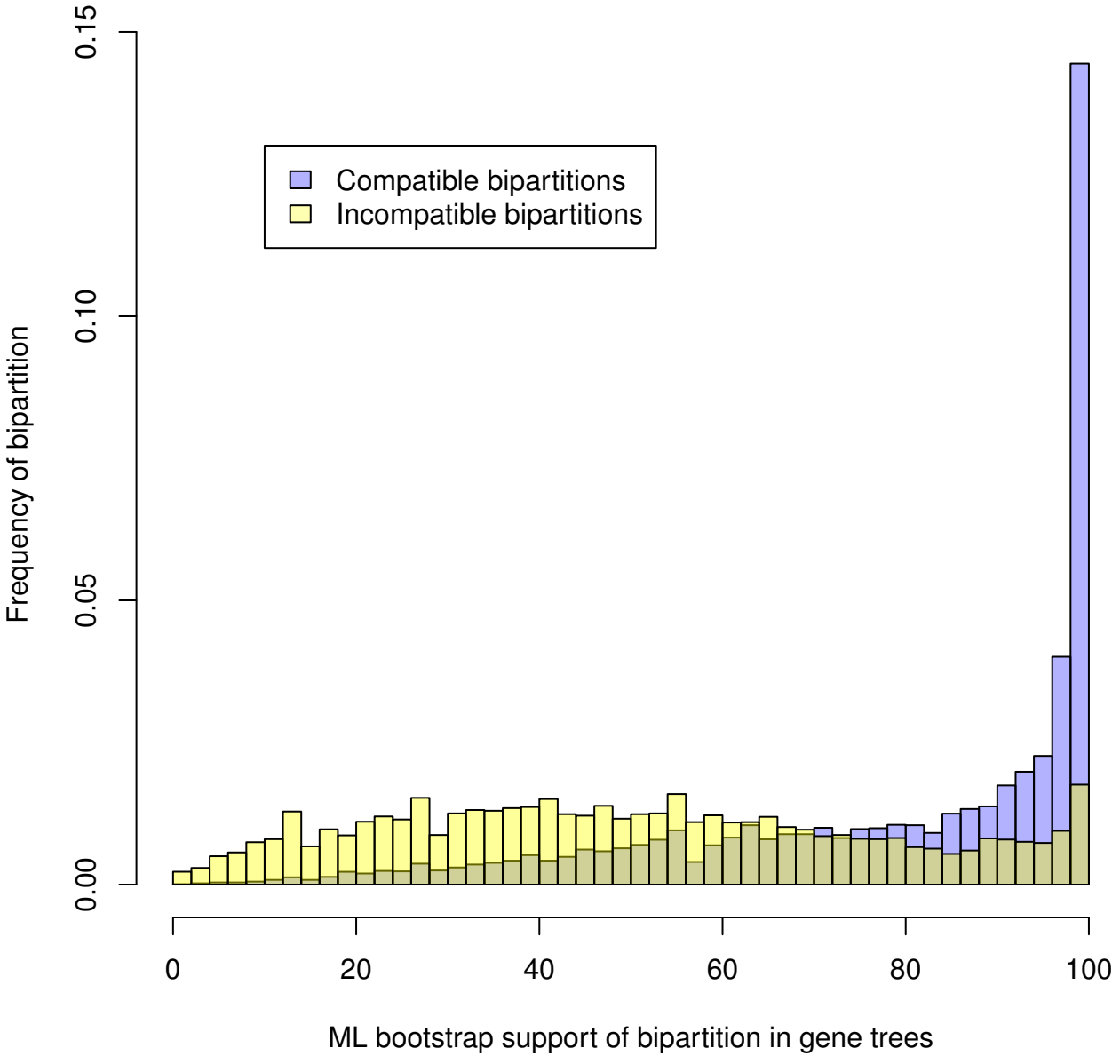
ONLINE APPENDIX 2. Saturation plot as calculated in DAMBE, showing GTR distance estimated for the ⅓ fastest-evolving sites in relation to the number of transitions and transversions at these sites. Regression lines (forced through the origin) and slope coefficients are shown. Dashed line indicates slope of 1.

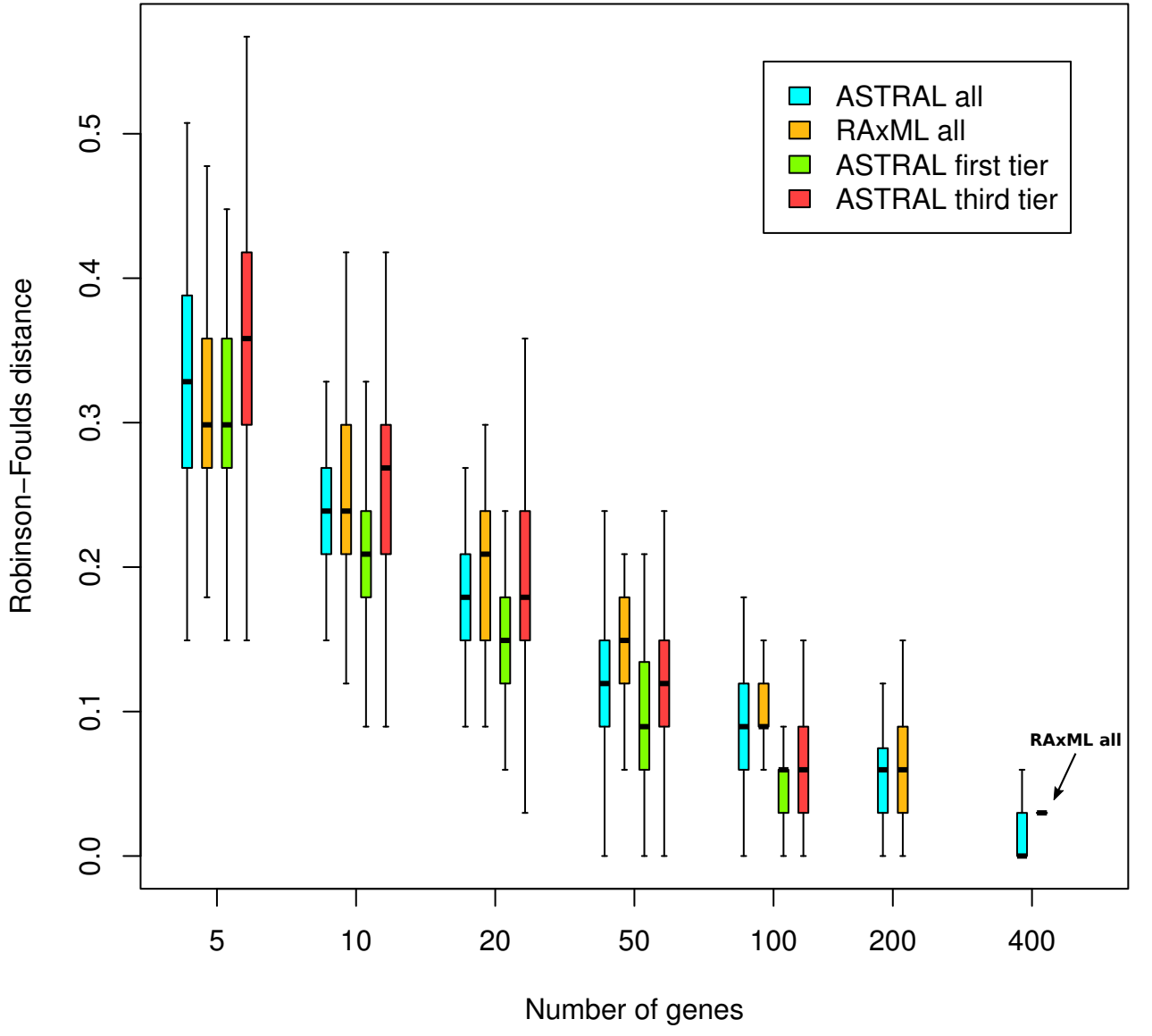
ONLINE APPENDIX 3. Single shortest tree found in PAUP* maximum parsimony searches. Support as MP jackknife, and ACCTRAN branch lengths.

ONLINE APPENDIX 4. Maximum likelihood tree in RAxML searches. Support as ML bootstrap.


461 nuclear genes
matK + ndhF + ETS-1f
Eleocharis obtusa
Erioscirpus comosus
Dulichium arundinaceum
Khaosokia caricoides
Calliscirpus brachythrix
Calliscirpus criniger
Amphiscirpus nevadensis
Scirpus rosthornii
Scirpus pendulus
Scirpus cyperinus
Scirpus atrovirens
Scirpus atrovirens
Eriophorum virginicum
Eriophorum angustifolium
Eriophorum vaginatum
Trichophorum alpinum
Trichophorum cespitosum
Carex hypolytroides
Carex kucyniakii
Carex siderosticta
Carex canescens
Carex camptoglochin
Carex pulicaris
Carex phleoides
Carex nardina
Carex capitata
Carex myosuroides
Carex microglochin
Carex atrivaginata
Carex bavicola
Carex filicina
Carex speciosa
Carex myosurus
Carex dimorpholepis
Carex plantaginea
A
B
C
D
E
99/*/*
0.80
68/*/*
0.56
84/-/-
0.62
67/99/56
0.88
90/98/-
//59
0.84
87/86/80
0.66
61/94/78
Out
DUL
Kha
Cal
Zam
Sci
Tri
CAR
Min
Vig
Uni
Car
58/62
86/52
*/99
95/89
86/56
98/89
84/54
90/69
*/98
-/-
85/72
98/97
80/73
89/88
71/91
60/79
-/-
78/64
50/-
57/68
99/94
*/99
51/-
-/-
61/-
95/89
Dulichium
Khaosokia
Calliscirpus
Amphiscirpus
Scirpus Clade
Trichophorum Clade
CARICEAE
Dulichium
Khaosokia
Calliscirpus
Amphiscirpus
Scirpus Clade
Trichophorum Clade
CARICEAE


Robinson−Foulds distance
0.2
0.3
0.4
0.5
0.6
0.7
0.8
30
40
50
60
70
80
Average ML bootstrap


Compatible bipartitions
Incompatible bipartitions
Frequency of bipartition
0.00
0.05
0.10
0.15
0
20
40
60
80
100
ML bootstrap support of bipartition in gene trees


ASTRAL all
RAxML all
ASTRAL first tier
ASTRAL third tier
RAxML all
Robinson−Foulds distance
0.0
0.1
0.2
0.3
0.4
0.5
5
10
20
50
100
200
400
Number of genes