Anchored Phylogenomics of Angiosperms I:

Assessing the Robustness of Phylogenetic Estimates

Chris Buddenhagen*[1,16], Alan R. Lemmon*[2,17], Emily Moriarty Lemmon[1,18], Jeremy Bruhl[3,19], Jennifer Cappa[4,20], Wendy L. Clement[5,21], Michael J. Donoghue[6,22], Erika J. Edwards[7,23], Andrew L. Hipp[8,24], Michelle Kortyna[1,25], Nora Mitchell[9,26], Abigail Moore[10,27], Christina J. Prychid[3,11,28], Maria C. Segovia-Salcedo[12,29], Mark P. Simmons[13,30], Pamela S. Soltis[14,31], Stefan Wanke[15,32], and Austin Mast[1,33]

*Co-first authors

[1] *Department of Biological Science, Florida State University, 319 Stadium Dr., P.O. Box 3064295, Tallahassee, FL, 32306-4295*;
[2] *Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL, 32306-4102*;
[3] *School of Environmental & Rural Science, University of New England, Armidale, NSW, Australia, 2351*;
[4] *Department of Biology, Colorado State University, Fort Collins, CO 80523-1878*;
[5] *The College of New Jersey, 200 Pennington rd, Ewing, NJ 08628*
[6] *Department of Ecology and Evolutionary Biology, Yale University, PO Box 208105, New Haven, CT 06520*
[7] *Ecology and Evolutionary Biology, Brown University, Providence, RI 02912*
[8] *The Morton Arboretum, 4100 Illinois Rte 53, Lisle, IL 60532*
[9] *Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Rd, U-3043, Storrs, CT 06269-3043*
[10] *Oklahoma Biological Survey and Department of Microbial and Plant Biology, University of Oklahoma, 770 Van Vleet Oval, Norman, OK 73019-6131*
[11] *Comparative Plant and Fungal Biology, Royal Botanic Gardens Kew, Richmond, Surrey, United Kingdom, TW9 3DS*
[12] *Departamento de Ciencias de la vida y Agricultura, Universidad de las Fuerzas Armadas ESPE. Av. General Rumiñahui s/n. Sangolquí, Ecuador. PO-BOX 171-5-231B*
[13] *Department of Biology, Colorado State University, Fort Collins, CO 80523-1878*
[14] *Florida Museum of Natural History, University of Florida, Dickinson Hall, Gainesville, FL 32611*
[15] *Institut für Botanik, Technische Universität Dresden, D-01062 Dresden, Germany*

[16] cbuddenhagen@gmail.com, [17] alemmon@fsu.edu,
[18] chorusfrog@bio.fsu.edu, [19] jbruhl@une.edu.au, [20] jjlarson@rams.colostate.edu, [21] clementw@tcnj.edu
[22] michael.donoghue@yale.edu, [23] erika_edwards@brown.edu, [24] ahipp@mortonarb.org, [25] mk09g@my.fsu.edu, [26] nora.mitchell@uconn.edu, [27]abigail.j.morre@ou.edu, [28] C.Prychid@kew.org, [29]mcsegovia@espe.edu.ec, [30] Mark.Simmons@colostate.edu, [31] psoltis@flmnh.ufl.edu, [32] stefan.wanke@tu-dresden.de
[33] amast@bio.fsu.edu

Corresponding Author:
Alan R. Lemmon
Department of Scientific Computing
Florida State University
Dirac Science Library
Tallahassee, FL, 32306-4102
Email: alemmon@fsu.edu
Phone: (850)-445-8946

ABSTRACT

An important goal of the angiosperm systematics community has been to develop a shared approach to molecular data collection, such that phylogenomic data sets from different focal clades can be combined for meta-studies across the entire group. Although significant progress has been made through efforts such as DNA barcoding, transcriptome sequencing, and whole-plastid sequencing, the community current lacks a cost efficient methodology for collecting nuclear phylogenomic data across all angiosperms. Here, we leverage genomic resources from 43 angiosperm species to develop enrichment probes useful for collecting ~500 loci from non-model taxa across the diversity of angiosperms. By taking an anchored phylogenomics approach, in which probes are designed to represent sequence diversity across the group, we are able to efficiently target loci with sufficient phylogenetic signal to resolve deep, intermediate, and shallow angiosperm relationships. After demonstrating the utility of this resource, we present a method that generates a heat map for each node on a phylogeny that reveals the sensitivity of support for the node across analysis conditions, as well as different locus, site, and taxon schemes. Focusing on the effect of locus and site sampling, we use this approach to statistically evaluate relative support for the alternative relationships among eudicots, monocots, and magnoliids. Although the results from supermatrix and coalescent analyses are largely consistent across the tree, we find support for this deep relationship to be more sensitive to the particular choice of sites and loci when a supermatrix approach as employed. Averaged across analysis approaches and data subsampling schemes, our data support a eudicot-monocot sister relationship, which is supported by a number of recent angiosperm studies.

KEYWORDS: hybrid enrichment, probe set, phylogenetic informativeness, nuclear loci, eudicots, monocots, magnoliids

78 A longstanding aim of angiosperm systematists has been to develop a unified approach to
79 data collection across the entire clade. Previous efforts have demonstrated the importance of
80 community coordination (e.g., *rbcL* sequencing; Chase et al. 1993). As the systematics
81 community moves into the genomics era, however, new technologies are enabling researchers to
82 transition from targeting a small number of primarily organellar loci, to targeting hundreds or
83 thousands of loci that are both broadly distributed across the genome and broadly homologous
84 across deep clades (Mamanova et al. 2010; Griffin et al. 2011; Crawford et al. 2012; Cronn et al.
85 2012; Lemmon and Lemmon 2013; Davis et al. 2014; Wickett et al. 2014; Prum et al. 2015;
86 Wicke and Schneeweiss 2015). Although progress has been made in plants recently using other
87 approaches such as whole genome sequencing, whole-plastid sequencing (Ruhfel et al. 2014),
88 and transcriptome sequencing (Burleigh et al. 2011; Lee et al. 2011; Wickett et al. 2014; Zeng et
89 al. 2014), perhaps the most promising methodology to date is hybrid enrichment, which is
90 rapidly resolving many previously intractable branches across the Tree of Life (Stull et al. 2013;
91 Pyron et al. 2014; Brandley et al. 2015; Crawford et al. 2015; Eytan et al. 2015; Prum et al. 2015;
92 Bryson et al. 2016; Meiklejohn et al. 2016; Tucker et al. 2016; Young et al. 2016). This method
93 has the advantages of cost-effectiveness, high-throughput data production, and success with
94 historical specimens, such as herbarium samples (Bi et al. 2013; Stull et al. 2013; Bakker 2015;
95 Beck and Semple 2015). Hybrid enrichment is a method wherein short oligonucleotides (termed
96 "probes" or "baits") are designed to complement target loci across the genomes of a clade (e.g.,
97 vertebrates, angiosperms, etc.). The researcher can choose the number, length, and type of loci to
98 be targeted. In the laboratory, probes hybridize to genomic DNA libraries, facilitating
99 enrichment of target loci relative to the remainder of the genome. The enriched DNA libraries
100 are then sequenced on a high-throughput sequencing platform. Because of the flexible and

customizable nature of the method, hybrid enrichment has the potential to unify data collection for many different types of research questions across angiosperms.

Some versions of hybrid enrichment are generally less appropriate than others for phylogenetic studies in plants. For example, the "UCE" method, as originally defined (Faircloth et al. 2012), utilizes thousands of relatively-short ultraconserved elements scattered across the genome as target regions. The use of UCEs in plants may be limited because appear to be scarce in plants compared to other systems (6 non-repetitive, invariant UCEs in plants vs. 1120 in animals; Reneker et al. 2012). An alternative method, Anchored Hybrid Enrichment (AHE; Lemmon et al. 2012; Lemmon and Lemmon 2013), ensures high enrichment efficiency by leveraging genomic and transcriptomic resources to develop probe sets that reflect the natural sequence variation occurring across a clade. By incorporating probes reflecting this phylogenetic diversity, regions with moderate levels of sequence variation can be efficiently enriched. Following recommendations from simulation studies (Leaché and Rannala 2011), this method incorporates a moderate number (~500-1000) of moderately-conserved target loci. Thus, the phylogenetic information content of the data sets generated is high, and the reasonable sequencing effort required makes it affordable for researchers to collect data from hundreds to thousands of species across evolutionary timescales. This method has been applied to diverse taxa across the Tree of Life, leading to complete or nearly complete resolution of many clades of different taxonomic depth, including birds (Prum et al. 2015; Morales et al. 2016), snakes (Pyron et al. 2014; Ruane et al. 2015), flies (Young et al. 2016), lizards (Brandley et al. 2015; Tucker et al. 2016), teleost fish (Eytan et al. 2015), frogs (Peloso et al. 2016). Given the success of AHE in many other systems, and the success to date of hybrid enrichment efforts targeting individual angiosperm clades (Table 1; e.g., Asteraceae, Mandel et al. 2014), the time seems ripe to extend

124 AHE to angiosperm phylogenetics though development and implementation of a general
125 angiosperm-wide enrichment resource.

126 Despite tremendous efforts to reconstruct the phylogeny of angiosperms, several deep
127 nodes continue to be debated (Soltis et al. 2011; Burleigh et al. 2011; Wickett et al. 2014;
128 Magallón et al. 2013). One reason for this difficulty is the large diversity in angiosperms—
129 estimates suggest approximately 400,000 extant species (~300,000 described; Pimm and Joppa
130 2015). A second reason is the limitation of many previous studies to a small number of plastid
131 and nuclear ribosomal markers available at the time, which do not always contain sufficient
132 information for resolving challenging nodes (e.g., Zimmer and Wen 2012), due to processes such
133 as incomplete lineage sorting and hybridization. A third reason is that many previous studies
134 have only presented results from a narrow set of data subsampling and methodological
135 conditions (e.g., Wickett et al. 2014). An especially intractable node that may be influenced by
136 these problems is the relationship among eudicots, monocots, and magnoliids at the base of the
137 angiosperms. Previous work has suggested all possible alternative resolutions of these three
138 clades depending upon type of data or taxa included and analysis method (eudicots-monocots
139 sister: Soltis et al. 2011; eudicots-magnoliids sister: Burleigh et al. 2011, Lee et al. 2011;
140 magnoliids-monocots sister: Wickett et al. 2014). Elucidating the true history of this group and
141 other difficult clades will require a more comprehensive approach that accounts for effects of
142 both taxa and data inclusion.

143 Systematists are increasingly becoming aware of the fact that although sufficient genomic
144 and taxonomic samplings are necessary for accurate phylogeny estimation, proper data analysis
145 is also critical. Accurate reconstruction requires careful navigation of a phylogenetic workflow
146 that is becoming increasingly complex and sophisticated (reviewed by Lemmon and Lemmon

147 2013). For example, evolutionary processes must be modeled adequately or estimates of
148 topology and branch lengths may be biased (Posada and Crandall 1998; Lemmon and Moriarty
149 2004; Brown and Lemmon 2007; Lanfear et al. 2012; Hoff et al. 2016). Second, proper
150 orthology assessment is essential to avoid including paralogous sequences (Chen et al. 2007;
151 Lemmon and Lemmon 2013). Third, construction of complete or nearly-complete data matrices
152 can be important for minimizing the potential effects of missing data (Lemmon et al. 2009;
153 Roure et al. 2013; Hosner et al. 2015). Finally, a particularly difficult aspect of phylogenomic
154 analysis is how to accommodate multilocus data with different underlying histories. The
155 traditional approach has been to concatenate the loci into a single supermatrix in the hope that
156 the signal of the true underlying species history overwhelms any discrepancies among the
157 individual histories of the loci sampled (Kluge 1989; de Queiroz and Gatesy 2007). More
158 recently, however, theory and simulation studies have shown that explicitly accounting for
159 incongruent gene histories (e.g., using a coalescent model or networks) can be essential for
160 accurate estimation of species history (Edwards 2009; Morrison 2011; Mirarab and Warnow
161 2015; Edwards et al. 2016). Incongruence among loci has particular bearing on angiosperm
162 systematics because relationships among some of the key groups change depending upon
163 whether a supermatrix or coalescent (species-tree) approach is employed (e.g., Wickett et al.
164 2014; Edwards et al. 2016; Sun et al. 2016), although it is possible that increased taxon sampling
165 could mitigate these effects. What is currently lacking is a framework for systematically
166 comparing results from different analysis conditions, while also considering the sensitivity of
167 these results to particular combinations of taxa and loci within the data set.

168 Here, we develop and test an angiosperm-wide hybrid enrichment resource for general use
169 by the angiosperm systematics community in order to facilitate unification of data collection and

accelerate resolution within this clade. After identifying 499 target loci with appropriate characteristics derived from Duarte et al. (2010), we develop a robust AHE probe kit that utilizes genomic resources from 25 genomes distributed broadly across angiosperms. We demonstrate the value of this resource for use at deep, intermediate, and shallow scales by carrying out hybrid enrichment across six orders and 10 plant families, and we incorporate previously published data to estimate a phylogenetic tree of 30 of the 79 angiosperm orders (APG IV 2016). Note that our sampling is designed to provide sufficient taxonomic coverage to evaluate the effectiveness of the enrichment probes, not to produce a comprehensive phylogeny of angiosperms. Independent of the phylogenetic data collection, we illustrate the potential for collecting data from functional loci by including probes for 18 selenium-tolerance loci (Freeman et al. 2010). Finally, we present results from a sensitivity analysis that will allow for a more systematic distinction between robust relationships that do not require further study versus tenuous relationships that require a more thorough investigation. Our goals are to: (1) test the efficiency of the enrichment resource and its effectiveness in resolving phylogenies and (2) provide a novel framework for evaluating the robustness of phylogenetic estimates and testing alternative topological hypotheses. The hybrid enrichment resource we present here is broad both genomically (loci are selected from throughout the genome; Suppl. Fig. 1) and taxonomically (taxa represented are distributed across the clade) in order to provide a useful tool for widespread application in angiosperm systematics studies. A practical demonstration of the application of this methodology to resolving difficult nodes below the family level is provided in another contribution (Léveillé-Bourret et al., in review).

192 **METHODS**

193 The primary objective of this work is to develop an enrichment kit capable of efficiently
194 collecting a large number of informative nuclear loci across all of angiosperms. A summary of
195 previous efforts to develop enrichment tools for plants is given in Table 1. Although a handful of
196 previous researchers have developed enrichment kits for broad taxonomic groups (e.g., eudicots),
197 their aim has been plastid genomes (Stull et al. 2013). Other researchers have targeted nuclear
198 loci, but only for specific clades (e.g., *Asclepias*, Weitemier et al. 2014). Moreover, studies
199 targeting nuclear loci have incidentally recovered plastid genomes as a byproduct of nuclear
200 enrichment (e.g., Stephens et al. 2015b; Schmickl et al. 2016), suggesting that specifically
201 targeting plastid genomes may be unnecessary. Our effort builds upon the effort of previous
202 studies by developing a general angiosperm-wide enrichment resource that targets loci of
203 moderate conservation distributed across the nuclear genome; we achieve this goal by including
204 probes from divergent lineages representing the phylogenetic breadth of angiosperms.

205 ***Locus Selection and Probe Design***

206 Based on previous experience in vertebrates (Lemmon et al. 2012), including birds (Prum
207 et al. 2015; Morales et al. 2016), snakes (Pyron et al. 2014; Ruane et al. 2015), lizards (Brandley
208 et al. 2015; Tucker et al. 2016), fish (Eytan et al. 2015), frogs (Peloso et al. 2016) and
209 invertebrates, including flies (Young et al. 2016), butterflies/moths (Breinholt et al. *in press*),
210 spiders (Hamilton et al. 2016), and other taxa, development of an efficient enrichment kit that is
211 useful across broad taxonomic scales requires careful selection of loci that are of sufficient
212 length for probe design, relatively free of indels and introns, and of moderate conservation across
213 the target group. In order to meet these requirements, we began with the 959 genes identified as
214 single-copy orthologs among *Populus*, *Vitis*, *Oryza,* and *Arabidopsis* by Duarte et al. (2010).

215 Since many of these genes contain introns, we divided these genes into exons using *Arabidopsis*
216 annotations and retained only those that were at least 150 bp, which is long enough to contain at
217 least two probes at 1.25 tiling density (i.e. 90 bp overlap; Fig. 1). This process yielded 3050
218 exons. In order to ensure we were not targeting loci that were too divergent for efficient
219 enrichment, we removed exons with <55% sequence similarity between *Arabidopsis* and *Oryza*.
220 The remaining 1721 exons served as our preliminary target loci.

221 In order to ensure that we developed enrichment probes that both represent the breadth of
222 angiosperm diversity and also ensure that the loci captured are low in copy number across
223 angiosperms, we leveraged a large set of published and unpublished genomic and transcriptomic
224 resources. Specifically, we utilized complete genomes from 33 species across angiosperms
225 (except magnoliids) together with newly obtained genomic data from nine non-model members
226 of Poales (Table 2; Suppl. Table 1). For each of these species, we identified homologous regions
227 following Prum et al. (2015), using *Arabidopsis* and *Oryza* as references (>55% similarity to one
228 of the references was required). We identified the final 499 target loci as those present in >85%
229 of the 41 references species and existing in ≤1.2 copies on average. These loci are distributed
230 across the *Arabidopsis* genome (Suppl. Fig. 1). After selecting homologous sequences with the
231 greatest sequence similarity to *Arabidopsis* and *Oryza* orthologs, we aligned sequences and
232 trimmed the resulting alignments to the exon ends using the *Arabidopsis* and *Oryza* sequences as
233 a guide. Prior to probe design, we subsampled the alignments to include only 25 species that
234 more evenly represented angiosperm diversity. The 25 references include *Amborella*, 6 monocot
235 families, 11 rosid families, 4 asterid families, and 4 non-rosid and non-asterid eudicot families
236 (Table 2). Inclusion of this large number of references was necessary because enrichment

237 success is inversely related to evolutionary distance of references and target species (Lemmon et
238 al. 2012; Bi et al. 2013; Sass et al. 2016).

239 In order to test the potential for simultaneously targeting loci of known function in
240 addition to anchor loci, we incorporated probes targeting 18 selenium-tolerance genes. The 18
241 genes are candidates for the selenium-tolerance phenotype obtained from a comparative
242 transcriptome analysis between selenium hyperaccumulator *S. pinnata* and selenium non-
243 hyperaccumulator *S. elata* (unpublished data) and a macroarray study between *S. pinnata* and *S.*
244 *albescens* (Freeman et al. 2010). Selenium is atomically similar to sulfur and can be assimilated
245 into selenocysteine via the sulfur assimilation pathway (Terry et al. 2000). These genes include
246 sulfur transporters and key enzymes in sulfur assimilation. Alignments for 18 of these genes
247 were obtained from TAIR (Berardini et al. 2015). Following the methods described above, we
248 scanned genomic resources for the 25 references species for selenium homologs using
249 *Arabidopsis* as a reference. Alignments of the homologs were then constructed using MAFFT
250 (v7.023b; Katoh and Standley 2013) and trimmed as described above.

251 After masking high-copy and repetitive regions, we tiled 120 bp probes across each of the
252 sequences at 2.8x coverage (neighboring probes overlap by approximately 48 bp). Probes were
253 synthesized by Agilent Technologies (Santa Clara, CA) to develop the kit hereafter referred to as
254 the Angiosperm v.1 design. The alignments and probe design are available as Supplemental
255 Material (Dryad accession XXXX).

256 ***Taxon Sampling***

257 Data were collected from a total of 104 angiosperm species representing 30 orders and 48
258 families (Suppl. Table 2). For 53 of these species (representing six orders, 10 families, and 21
259 genera), the Angiosperm v.1 kit was used to enrich genomic DNA for the target regions prior to

260 sequencing. Sequences for the remaining 51 species were obtained from low-coverage genomic
261 reads (not enriched), assembled transcriptomes, and assembled genomes (Suppl. Table 2).

262 ***Hybrid Enrichment Data Collection***

263 In order to test the efficiency and utility of the AHE resource described above, we
264 enriched 53 samples from across the phylogenetic breadth of angiosperms following the general
265 methods of Lemmon et al. (2012), with adaptations made for plant samples. Genomic DNA was
266 extracted using the DNEasy Plant Mini Kit (Qiagen, Valencia, California, USA). The protocol
267 was modified following suggestions (Costa and Roberts 2014) to compensate for frequent low
268 yields obtained with the standard protocol. After extraction, genomic DNA was sonicated to a
269 fragment size of ~300-800 bp using a Covaris E220 Focused-ultrasonicator with Covaris
270 microTUBES. Subsequently, library preparation and indexing were performed on a Beckman-
271 Coulter Biomek FXp liquid-handling robot following a protocol modified from Meyer and
272 Kircher (2010). A size-selection step was also applied after blunt-end repair using SPRI select
273 beads (Beckman-Coulter Inc.; 0.9x ratio of bead to sample volume). Indexed samples were then
274 pooled at approximately equal quantities (typically 16-18 samples per pool), and then each pool
275 was enriched using the Angiosperm v.1 kit (Agilent Technologies Custom SureSelect XT kit
276 ELID 623181). After enrichment, 3–4 enrichment reactions were pooled in equal quantities for
277 each sequencing lane and sequenced on PE150 Illumina HiSeq 2500 lanes at the Translational
278 Science Laboratory in the College of Medicine at Florida State University.

279 ***Paired-Read Merging***

280 In order to increase read accuracy and length, paired reads were merged prior to assembly
281 following Rokyta et al. (2012). In short, for each degree of overlap each read pair was evaluated
282 with respect to the probability of obtaining the observed number of matches by chance. The

283 overlap with the lowest probability was chosen if the p-value was less than 10^{-10}. This low p-
284 value avoids chance matches in repetitive regions. Read pairs with a p-value below the threshold
285 were merged and quality scores were recomputed for overlapping bases (see Rokyta et al. [2012]
286 for details). Read pairs failing to merge were utilized but left unmerged during the assembly.

287 ***Read Assembly***

288 Divergent reference assembly was used to map reads to the probe regions and extend the
289 assembly into the flanking regions (Fig. 2; Prum et al. 2015). More specifically, a subset of the
290 taxa used during probe design were chosen as references for the assembly: *Arabidopsis thaliana*,
291 *Billbergia nutans*, and *Carex lurida*. Matches were called if 17 bases matched a library of spaced
292 20-mers derived from the conserved reference regions (i.e., those used for probe design).
293 Preliminary reads were then considered mapped if 55 matches were found over 100 consecutive
294 bases in the reference sequences (all possible gap-free alignments between the read and the
295 reference were considered). The approximate alignment position of mapped reads were estimated
296 using the position of the spaced 20-mer, and all 60-mers existing in the read were stored in a
297 hash table used by the de-novo assembler. Simultaneously using the two levels of assembly
298 described above, the three read files were traversed repeatedly until a pass through the reads
299 produced no additional mapped reads. For each locus, a list of all 60-mers found in the mapped
300 reads was compiled and 60-mers were clustered if found together in at least two reads. Contigs
301 were estimated from 60-mer clusters. In the absence of contamination, low coverage, and gene
302 duplication each locus should produce one assembly cluster. Consensus bases were called from
303 assembly clusters as unambiguous base calls if polymorphisms could be explained as sequencing
304 error (assuming a binomial probability model with the probability of error equal to 0.1 and alpha
305 equal to 0.05). Otherwise ambiguous bases were called (e.g., 'R' was used if 'A' and 'G' were

306 observed). Called bases were soft-masked (made lowercase) for sites with coverage lower than
307 five. Assembly contigs derived from less than 10 reads were removed in order to reduce the
308 effects of cross contamination and rare sequencing errors in index reads. Scripts used in the
309 assembly process are available on Dryad (accession XXXX).

310 ***Orthology Estimation***

311 After grouping homologous sequences obtained by enrichment and whole genomes,
312 putative orthologs were identified for each locus following Prum et al. (2015). For each locus,
313 pairwise distances among homologs were computed using an alignment-free approach based on
314 percent overlap of continuous and spaced 20-mers. Using the distance matrix, sequences were
315 clustered using the neighbor-joining algorithm, but allowing at most one sequence per species to
316 be present in a given cluster. Note that flanks recovered through extension assembly contain
317 more variable regions and allow gene copies to be sorted efficiently. Gene duplication before the
318 ancestor of the clade results in two distinct clusters that are easily separated. Duplication within
319 the clade typically results in two clusters, one containing all of the taxa and a second containing a
320 subset of the taxa (missing data). Gene loss also results in missing data. In order to reduce the
321 effects of missing data, clusters containing fewer than 50% of the species in the taxon set were
322 removed from downstream processing.

323 ***Alignment and Trimming***

324 Putatively orthologous sequences were processed using a combination of automated and
325 manual steps in order to generate high quality alignments in a reasonable amount of time.
326 Sequences in each orthologous cluster were aligned using MAFFT v7.023b (Katoh and Standley
327 2013), with --genafpair and --maxiterate 1000 flags utilized. The alignment for each locus was
328 then trimmed/masked using the steps from Prum et al. (2015). First, each alignment site was

329 identified as "conserved" if the most commonly observed character state was present in > 40% of
330 the sequences. Second, using a 20 bp window, each sequence was scanned for regions that did
331 not contain at least 10 characters matching to the common state at the corresponding conserved
332 site. Characters from regions not meeting this requirement were masked. Third, sites with fewer
333 than 12 unmasked bases were removed from the alignment. A visual inspection of each masked
334 alignment was carried out in Geneious version 7 (www.geneious.com; Kearse et al. 2012).
335 Regions of sequences identified as obviously misaligned or non-homologous were removed.

336 ***Phylogeny Estimation***

337 Adequate modeling of sequence evolution is an important prerequisite for accurate
338 phylogenetic estimation (Hillis et al. 1994; Yang 1994; Posada and Crandall 1998; Lemmon and
339 Moriarty 2004; Ripplinger and Sullivan 2008; Lanfear et al. 2012; Hoff et al. 2016). Site
340 partitioning is one aspect of modeling that has recently gained appreciation (Brown and Lemmon
341 2007; Lanfear et al. 2012; Darriba and Posada 2015; Frandsen et al. 2015). When the loci being
342 modeled are from protein coding regions, some authors choose to partition by codon position and
343 remove third codon positions in order to reduce the effects of substitutional saturation (e.g.,
344 Breinholt and Kawahara 2013; Wickett et al. 2014; Frandsen et al. 2015). Although the anchored
345 loci we target here were derived from protein-coding exonic regions, partitioning by coding
346 position would not be sufficient because captured flanks could contain intronic and intergenic
347 regions that are also likely to vary both by model and rate of evolution. Moreover, the degree of
348 saturation and its effects on downstream analyses depend heavily on the phylogenetic breadth
349 and depth of sampling. Consequently, we used an agnostic approach to site partitioning that does
350 not rely on a prior knowledge of possible site partitions. More specifically, we used
351 PartitionFinder v2.0.0-pre14 (Frandsen et al. 2015), which utilizes a k-means approach to

352 partition sites into clusters based on similarity with respect to model of sequence evolution
353 (including rate). This approach produces partitioning schemes that fit alignments better than
354 schemes determined by methods requiring definition of *a priori* partitions (e.g., by codon
355 positions and locus boundaries; Frandsen et al. 2015). An added benefit to using this approach is
356 that it produces estimates of site-specific rates of evolution.

357 In order to allow for estimation of trees under both coalescent and supermatrix
358 approaches, RAxML v8.1.21 (Stamatakis 2014) was used to estimate gene trees from individual
359 locus alignments and also the concatenated supermatrix (GTRGAMMA model, partitioned as
360 described above, with -f -a flags). In each case, *Amborella* was specified as the outgroup, and
361 100 rapid bootstrap replicates were collected in order to measure support for the maximum
362 likelihood estimate. All other parameters remained at default settings. Species trees were
363 estimated in ASTRAL-II v.4.9.7 (Mirarab and Warnow 2015) using bootstrap replicates from the
364 RAxML-estimated gene trees.

365 ***Evaluating the Size and Variability of Flanking Regions for Family- and Genus-Level Clades***

366 One feature of AHE is that data collected by the approach can be useful at both deep,
367 intermediate, and shallow timescales. Recall that in order to ensure efficient enrichment across
368 deep time scales, regions targeted by the enrichment kit described above consist of moderately
369 conserved exons. Consequently, flanking regions are expected to contain intronic and intergenic
370 sites that may only be accurately aligned at shallower taxonomic scales. We expect, therefore,
371 that the majority of flanking bases would be filtered out when the angiosperm-wide alignment is
372 trimmed. In order to study the effect of phylogenetic depth on the size of usable flanks, we also
373 performed the post-assembly analysis steps on subsamples of the taxa. More specifically, we
374 performed orthology assessment and constructed trimmed alignments specific to the following

375 clades: Cyperaceae (21 samples), Caryophyllales (5 samples), Dipsacales (6 samples), *Magnolia*
376 (7 samples), *Protea* (4 samples), and *Stanleya* (5 samples). The size and variation of flanking
377 regions for these alignments was compared to corresponding sections of the angiosperm-wide
378 alignment.

379 ***Assessing Robustness of Phylogenetic Estimates***

380 Phylogenetic estimates can be highly sensitive to not only the model of evolution
381 assumed, but also the choice of data being analyzed (Ripplinger and Sullivan 2008; Wickett et al.
382 2014; Hosner et al. 2015; Simmons et al. 2016). Inclusion of saturated sites and/or outlier loci,
383 for example, can have particularly profound effects on phylogenetic estimates (Breinholt and
384 Kawahara 2013; Simmons et al. 2016). These effects can result in conflict among studies,
385 especially for phylogenetic relationships that are particularly difficult (e.g., the relationships
386 among eudicots, monocots, and magnoliids). Given this difficulty, it is surprising that more
387 authors do not present results from a more exhaustive exploration of the effects of their data
388 selection on their estimates of phylogeny.

389 Here, we present an approach to evaluate the robustness of phylogenetic estimates to
390 differential character and taxon sampling (Fig. 3; all scripts available in Dryad). Although we
391 focus on the effects of variation in sampled loci included, other aspects of the data set and
392 analysis could also be incorporated (e.g., taxon sampling). The approach begins by
393 independently varying two or more aspects of the data set or analysis. In this case we varied both
394 the sites included at each locus and also the loci included. We ordered the sites based on
395 evolutionary rate and the loci based on similarity to other loci estimated via gene tree distances;
396 then we analyzed the data with successively less variable and less distant loci included (see
397 details below). Data sets generated under each of these conditions were then used to estimate the

398 phylogeny and support values inferred using both supermatrix and coalescent approaches. The
399 novel aspect of this approach is the representation of the resulting support values for particular
400 nodes as heat maps in the parameter space defined by the aspects of the data set or analysis that
401 we varied.

402 High evolutionary rate estimates can result from alignment error, substitutional
403 saturation, and convergent evolution of base composition. Removal of the sites with highest
404 estimated rates can improve the support for the topology by removing noise from the data set
405 (Castresana 2000; Talavera and Castresana 2007; Xia and Lemey 2009). Removing too many
406 sites, however, will eventually result in loss of signal and subsequently in poorly supported
407 topologies. In some cases, conversely, the most rapidly evolving sites contain strong
408 phylogenetic signal, and so removing these sites may actually reduce branch support values.
409 Given the difficulty in knowing *a priori* which rate threshold optimizes this tradeoff for a
410 particular data set, a useful approach is to test several thresholds along a continuum. We utilized
411 site-specific rates estimated by PartitionFinder v2.0.0-pre14 from a concatenated matrix
412 containing all loci using the kmeans option (default settings otherwise; Frandsen et al. 2015).
413 Fourteen site inclusion thresholds along the observed distribution were then chosen based on the
414 following percentiles: 100%, 99%, 97%, 95%, 92%, 89%, 86%, 82%, 76%, 70%, 63%, 54%,
415 43%, and 31%. These thresholds progressively remove sites such that our test would be more
416 sensitive to high levels of site inclusion. Based on preliminary tests with this data set, we
417 expected phylogenetic support to deteriorate rapidly when less than 63% of sites were included.
418 New locus-specific alignments were then constructed based on each of these site inclusion levels,
419 and gene trees were estimated for each of these data sets using RAxML as described above (14
420 estimated gene trees per locus; Fig. 4a).

421 Inclusion of loci with insufficient signal can result in imprecise or inaccurate estimates of
422 species trees (Hosner et al. 2015; Simmons et al. 2016). Moreover, inclusion of loci with
423 estimated gene-tree-histories that are substantially different than the underlying species history
424 can also result in imprecise or inaccurate estimates of species trees (Meredith et al. 2011;
425 Townsend et al. 2011; Simmons and Gatesy 2015; Simmons et al. 2016; Springer and Gatesy
426 2016). Therefore, removal of outlier gene trees may result in more accurate and precise species
427 trees. As with inclusion of sites, however, inclusion of too many loci will eventually result in
428 reduced accuracy, precision, and support. In order to identify the level of locus inclusion that
429 maximizes accuracy and precision, we tested several thresholds along a continuum. For each of
430 the 14 sets of gene trees described above, pairwise tree distances were estimated using treeCmp
431 (v1.0-b29; Bogdanowicz et al. 2012), via the triple metric for rooted trees (tt option; Critchlow et
432 al. 1996). The resulting matrix of pairwise tree distances was visualized using multidimensional
433 scaling using the R package plotrix version 3.6-2 (Fig. 4b; Fig. 3; from Lemon [2006]).
434 Euclidian distance from each tree to the center of the trees was then used to rank the loci, with
435 the greatest distance (i.e., greatest outlier) having the highest rank. Fourteen locus-inclusion
436 thresholds were then applied: 384, 381, 374, 363, 348, 329, 306, 279, 248, 213, 174, 131, 84, and
437 33 loci. These thresholds were chosen to cover a broad range of locus inclusion levels, but still
438 allow the effects of removing small numbers of outliers to be studied. These 14 locus-inclusion
439 thresholds were applied to each of 14 site-inclusion thresholds described above (Fig. 4d).
440 Phylogenies were then inferred using each of the 196 combinations in RAxML and ASTRAL-II,
441 using corresponding concatenated matrices and gene trees sets, respectively.

442 The primary goal of the extensive analyses described above is to produce a mechanism
443 for visualizing the sensitivity of phylogenetic estimates to variation in site and locus inclusion.

444 To this aim, nodal support for trees at each filter level were estimated using the R package ape
445 (Paradis et al. 2004) with the "prop.clades" command, which counts the number of times the
446 bipartitions present in a reference tree are present in a list of trees. After selecting a reference tree
447 (the best tree from the full unfiltered data set), the occurrence of a clade in the bootstrap trees
448 was assessed for each of the 196 ASTRAL-II and 196 RAxML concatenated analyses. This step
449 produced support values for each combination of thresholds at each node in the reference tree.
450 Heat maps were then constructed for each internal node in the reference tree (Fig. 4c). Note that
451 alternate reference trees can be used to generate heat maps representing support for alternative
452 hypotheses.

453 ***Testing Alternative Hypotheses***

454 The large number of estimates produced by the above procedure provides an opportunity
455 to test alternative phylogenetic hypotheses and alternative phylogeny estimation approaches
456 across a broad range of conditions. To demonstrate the potential for testing alternative
457 phylogenetic hypotheses, we focused on the relationship between eudicots, monocots, and
458 magnoliids that has been debated in the literature (Lee et al. 2011; Soltis et al. 2011; Burleigh et
459 al. 2011; Wickett et al. 2014). Specifically, support values represented by heat maps
460 corresponding to support for the three alternative relationships among these clades (magnoliids-
461 monocots sister, eudicots-monocots sister, and magnoliids-eudicots sister) were compared using
462 randomization tests as follows. Average support across all analysis conditions (a heat map) was
463 computed for each of the three relationships. A test statistic was then computed as the difference
464 between the mean support (across analysis conditions) for two alternative relationships. To
465 generate a null distribution with which to compare the test statistic, support values found in heat
466 map matrices were then randomized the relevant pair of matrices, while keeping the analysis

467 condition constant. In other words, the two support values (one for each relationship) for a given
468 analysis condition were randomized with respect to the matrix in which they were placed. After
469 randomizing, the difference between the means of the two randomized matrices was computed as
470 a value in the null distribution. This process was repeated 10,000 times to generate the null
471 distribution. To demonstrate the potential for testing support given by alternative analysis
472 methods, we used the randomization test described above, but instead used matrices derived
473 from the supermatrix (RAxML concatenated) and species tree (ASTRAL-II) analyses.

474 We also tested the hypothesis that support produced by the species tree approach is more
475 robust to variation in data sets than the support produced under the supermatrix approach (Song
476 et al. 2012; Edwards et al. 2016). In particular, we compared the overall sensitivity as the
477 variance in support values across analysis conditions, averaging the variance across nodes (a
478 value of zero indicates high robustness). We then computed the test statistic as the difference
479 between the sensitivity measure for the coalescent and supermatrix (with supermatrix -
480 coalescent > 0 indicating support that the coalescent approach is more sensitive to changes in
481 analysis conditions). Ten thousand points from the null distribution were computed by shuffling
482 corresponding support values between coalescent- and supertree-derived matrices, before
483 recomputing the test statistic.

484 In addition to the randomization tests described above, we performed more traditional
485 hypothesis tests using point estimates of phylogeny from the analyses of the full data set. The
486 best-constrained trees for each topology were estimated in RAxML using the -g option under a
487 GTRGAMMA model. The per-site log likelihoods for the three best constrained trees were
488 written to file and loaded in the CONSEL (v0.1i) program (Shimodaira and Hasegawa 2001).
489 This program ranks alternative hypotheses in order of the likelihood via p-value calculated from

490 the multi-scale bootstrap using several tests including the widely used Shimodaira–Hasegawa
491 (SH) and the approximately unbiased (AU) test (Shimodaira and Hasegawa 2001; Shimodaira
492 2002).

493

494 **RESULTS**

495 ***Locus Selection and Probe Design***

496 The procedure followed for locus selection (Fig. 1) yielded a large number of target loci,
497 with minimal missing data and optimal levels of sequence variation for hybrid-enrichment-based
498 phylogenomics. Target probes were designed for 499 anchor loci and 18 functional loci (517 loci
499 in total). Alignments of sequences for the 25 reference species averaged 343 sites per target locus
500 (range: 124 to 2176 sites) and contained only 10% missing characters. Sequence variation in
501 target regions was moderate: 62% of the sites were variable, 51% of the sites were parsimony
502 informative, pairwise sequence divergence averaged 27%, and maximum pairwise sequence
503 divergence within loci averaged 39%. The probe design contained a total of 56,862 probes (~4.4
504 probes per species per locus; See Supplemental Materials).

505 ***Hybrid Enrichment***

506 The probe kit developed above performed well on the 53 test samples taken from across
507 angiosperms; >90% of target loci were obtained for 90% of the samples (Table 3). Over 91
508 billion nucleotides were sequenced on the two PE150 Illumina lanes. Assemblies resulted in
509 recovery of an average of 470 loci (91%) with contigs longer than 250 bp. The average
510 consensus sequence recovered was 764 bp, more than twice the length of average probe region
511 (343 bp). On average, 14% of reads mapped to target plus flanking regions (i.e., were
512 incorporated in the assemblies). As expected, capture efficiency was lower for taxa more

513 evolutionary divergent from reference sequences (Fig. 5). For example, an average of only 329
514 loci were recovered for members of Piperales, which shares an ancestor with the closest
515 reference 124 million years ago (Magallón et al. 2013). In contrast, for members of Brassicales
516 (sharing a common ancestor with the closest reference 24 million years ago), an average of 512
517 loci were recovered.

518 ***Alignments***

519 For the angiosperm taxon set, a large number of loci were retained through the post-
520 assembly pipeline. More specifically, 384 alignments containing orthologous sequences were
521 obtained (available as Supplemental Material). The final concatenated angiosperm alignment
522 contained 138,616 bases. In general, alignments contained a modest amount of missing data (7%
523 if missing taxa were excluded, 22% if missing taxa were included). The number of missing taxa
524 varied among loci (mean=15, range=2-40). Note for the angiosperm-wide alignments, flanks
525 were typically trimmed off because they contained non-homologous or unalignable bases at this
526 deep scale (see below for intermediate- and shallow-clade alignments). As these alignments
527 contain primarily probe regions, it is not surprising that final, trimmed alignments were only
528 slightly longer than the alignments used to design probes (360 vs. 343, respectively). Also note
529 that none of the 18 functional loci were retained for the angiosperm-wide taxon set.

530 For the intermediate and shallow-clade taxon sets, an even larger number of loci was
531 retained through the post-assembly pipeline (Table 3). An average of 410 alignments containing
532 orthologous sequences were obtained (range 189 to 540). Because they contained flanking
533 regions, alignments for these genus-level taxon sets were substantially longer (68% longer on
534 average) than those of the angiosperm-level taxon set (Table 4). Concatenated alignments
535 averaged 233,993 bases (107,918 to 365,941), and locus-specific alignments averaged 577 bases

536 (375 to 670). Remarkably, all taxa were present for each of the final locus sets, indicating that
537 missing data/taxa in the angiosperm-wide alignments resulted from differences in locus recovery
538 across families. One indication of this cause is the consistent absence of some target loci in
539 Piperales and Proteales, in which only 189 and 159 orthologous loci were retained, respectively.
540 As a result of the increased length and variability of the loci, support for the intermediate and
541 shallow-scale clades improved (Suppl. Figs. 2-6). Although the estimated topologies remained
542 unchanged, half of the nodes that did not receive full support (100% bootstrap) in the
543 Angiosperm-wide alignment became fully supported by the intermediate and shallow-scale
544 alignments.

545 One reason why a greater number of orthologous loci were retained for the intermediate
546 and shallow taxon sets is that orthology could more easily be established for duplicated genes.
547 For example, most of the functional locus targets contained multiple copies (average 2.16; Table
548 3). At one functional target locus in Proteaceae, in fact, 14 orthologs were identified. Multi-copy
549 loci such as these were removed from the angiosperm-wide taxon set during the
550 orthology/trimming process because copy number for these loci varied wildly across genera (i.e.,
551 multiple independent duplications resulting in these copies occurred within angiosperms).
552 Brassicales, in which many species have high selenium tolerance (likely due to the propensity for
553 sulfur accumulation and assimilation into unique sulfur-containing compounds), produced the
554 greatest number of orthologous alignments (n=574).

555 ***Robustness of Phylogenetic Estimates***

556 Although the majority of the angiosperm relationships were unequivocally supported
557 across the majority of analysis conditions, some relationships were less robust to variation in
558 levels of data filtering and/or method of phylogeny estimation. The robustness of each estimated

559 node is shown as a heat map in Figure 6, which provides not only a comprehensive view of
560 support across many angiosperms orders, but also a clear indication of which relationships are
561 only supported under specific conditions. Many nodes were well supported in all but the most
562 extreme thresholds for site and locus removal (i.e., >60% of sites removed and/or >60% loci
563 removed). Overall, support values were consistent between coalescent and supermatrix
564 approaches as long as more than ~100 loci were included. Nonetheless, support for several nodes
565 varied wildly across analysis conditions, with highest support occurring when a moderate to
566 large numbers of sites or loci were removed.

567 For the relationship of particular interest, that between monocots, eudicots and
568 magnoliids, support for alternative arrangements were highly dependent upon the level of site
569 filtering, the level of locus filtering, and also the approach to phylogeny estimation (Fig. 7).
570 Coalescent-based analyses supported the eudicots-monocots sister hypothesis with moderate to
571 high support for the majority of site- and locus-filtering conditions. Support for the other two
572 arrangements was almost entirely lacking, regardless of the level of site- and locus-filtering.
573 Conversely, supermatrix-based analyses strongly supported the eudicots-magnoliids sister
574 relationship when all data were included, but strongly supported the eudicots-monocots sister
575 relationship when 54–76% of the most rapidly evolving sites were removed.

576 ***Testing Alternative Hypotheses***

577 Statistical tests indicated that the eudicots and monocots sister relationship was the best-
578 supported alternative (compared to eudicots-magnoliids sister and compared to monocots-
579 magnoliids sister) when results were averaged across all site- and locus-filtering levels. This
580 result was obtained regardless of whether a coalescent (test statistic eudicots-monocots vs.
581 eudicots-magnoliids = 25.09; eudicots-monocots vs. monocots-magnoliids = 52.88; $p < 0.0001$,

582 both tests) or supermatrix (test statistic eudicots-monocots vs. eudicots-magnoliids = 42.63
583 eudicots-monocots vs. monocots-magnoliids = 24.83; $p < 0.0001$, both tests) approach was
584 applied. This result may seem somewhat surprising given the inconsistent support generated by
585 the supermatrix and coalescent approaches for this node (Fig. 7). Recall, however, that strong
586 support for the eudicots-magnoliids sister hypothesis was only found in the narrow condition in
587 which >95% of the sites were included in the supermatrix analysis. Support for alternative
588 relationships of these taxa was found to be more sensitive to data filtering when the supermatrix
589 instead of the coalescent approach was employed (test statistic = 38.69, $p < 0.0001$). The
590 approximately unbiased (AU) test (Shimodaira and Hasegawa 2001) rejected the monocots-
591 magnoliids sister hypothesis ($p = 0.015$), but failed to reject the eudicots-magnoliids sister
592 hypothesis ($p = 0.119$).

593 Although estimates based on supermatrix and coalescent approaches were found to be at
594 odds for some of the difficult nodes for some data-filtration levels, estimates based on the two
595 approaches were quite consistent overall as long as ~150 or more loci were included. For
596 example, analysis of the full dataset under the supermatrix and coalescent approaches yielded
597 identical topologies within monocots and magnoliids (some topological differences were evident
598 within eudicots, however). Interestingly, inclusion of more than 150 loci did not appear to
599 improve the consistency of estimates between supermatrix and coalescent approaches.
600

601 **DISCUSSION**

602 The anchored hybrid enrichment resource developed here provides systematists with the
603 means to collect hundreds of phylogenomic loci for resolving species relationships across
604 angiosperms. By leveraging available genomic resources we have developed an enrichment

605 resource containing probes that represent much of the ordinal-level diversity within angiosperms,
606 which facilitates efficient enrichment of up to 499 anchor loci. The targeted anchor loci are
607 sufficiently conserved to allow enrichment in most groups within the clade, yet are still contain
608 enough variation to robust resolution of the large majority of nodes at a broad range of
609 phylogenetic scales. For instance, Léveillé-Bourret et al. (in review) demonstrate the success of
610 the methodology in resolving the most difficult nodes of a rapid Cyperaceae radiation involving
611 thousands of species. Our data indicate that resolution at intermediate and shallow scales is
612 improved by including sequence data from flanking regions around the probe region because the
613 flanks contain a greater degree of variation. Given the broad taxonomic span, this resource has
614 the potential to unify the field of angiosperm systematics by providing a common set of nuclear
615 loci that have a high level of phylogenetic information across evolutionary scales. In fact, the
616 angiosperm probe v.1 kit has already been widely adopted through 21 ongoing collaborations to
617 collect data from 87 angiosperm families for more than 2000 taxa (as of October 2016) at the
618 Center for Anchored Phylogenomics (www.anchoredphylogeny.com).

619 Our results indicate that the anchor loci targeted by the angiosperm enrichment probes
620 contain a sufficient amount of data to resolve relationships at a broad set of taxonomic scales.
621 More than 75% of the nodes in this angiosperm-wide test were strongly supported under a broad
622 range of conditions as long as at least 60% of the sites and 25% of the loci were included. Of the
623 remaining nodes, only a small minority required hundreds of loci to achieve strong support.
624 Léveillé-Bourret et al. (in review) found highly similar results in a tribal-level study, with the
625 backbone topology stabilizing and support peaking with 100–200 loci in both concatenation and
626 coalescent-based analyses. This pattern is not surprising given results from simulations
627 indicating that only a few informative loci are needed to resolve easy nodes and a few hundred

628 informative loci are sufficient to resolve more difficult nodes (Leaché and Rannala 2011).
629 Despite the robustness of support for the majority of nodes, support values for some nodes were
630 quite sensitive to the number of loci included, the number of sites included, and/or the analysis
631 approach employed (supermatrix or coalescent). In most of these cases, the support peaked when
632 an intermediate number of loci were included, suggesting that increasing the number of targeted
633 loci beyond 499 may have a small effect on the accuracy or precision of estimates. Focusing
634 resources on other aspects of study design, such as density of taxon sampling or increasing locus
635 lengths, may prove to be more fruitful (Zwickl and Hillis 2002; Betancur-R. et al. 2014; Peloso
636 et al. 2016; Simmons et al. 2016). For very shallow and/or rapidly-radiating groups, for example,
637 robust estimates may require increased locus lengths in order to ensure well-resolved and well-
638 supported gene trees. One approach to improve gene tree resolution for shallow-scale studies is
639 to extend probe regions into flanks using low-coverage genome data; based on our experience,
640 >15x coverage is typically necessary. We have had good success with this approach in orchids
641 (Gravendeel, et al., *in prep*), lobelias (Givnish et al., *in prep*), lilies (Givnish et al., *in prep*), and
642 goldenrods (Beck et al. *in prep*). The extended loci contain more variable regions and yield more
643 strongly supported gene and species trees. Extending anchor loci is often preferable to targeting
644 anonymous loci because inclusion of the anchor regions allows the data to also be useful for
645 broad-scale meta-studies.

646 Despite our efforts to incorporate genomic resources from across angiosperms, enrichment
647 efficiency could be improved further in underrepresented regions of the tree (e.g., magnoliids
648 and basal eudicots). Fortunately, it is fairly straightforward to improve the resource by
649 incorporating additional probes representing additional taxa as genomic resources become
650 available. For example, as part of the orchid, lobelia, and lily projects described above, we have

651 added probes to improve the Angiosperm v.1 kit. In each of these cases, inclusion of the
652 additional probes substantially increased enrichment efficiency for samples from those groups
653 (Lemmon et al., unpub. data).

654 The heat map approach representing robustness of phylogenetic estimates is useful in
655 several contexts. First and foremost, it is a useful means to separate nodes with tenuous support
656 from those with robust support. As seen with several difficult nodes in the angiosperm tree, high
657 support values occurring under a narrow range of conditions (e.g., no site or locus filtering) has
658 the potential to be misleading, since alternative hypotheses can be supported across a broader
659 range of conditions. Second, the approach provides a framework for testing the robustness and
660 consistency of alternative methods of data analysis (e.g., supermatrix and coalescent). For the
661 monocot-eudicot-magnoliid split, for example, our results support the claim of Edwards et al.
662 (2016) that the supermatrix approach has the tendency to produce confident yet conflicting
663 results in different regions of parameter space, although other interpretation are possible
664 (Simmons and Gatesy, 2016). In contrast, while the coalescent approach tends to require more
665 data to attain high support for a hypothesis, it has a much lower tendency to produce conflicting
666 results under different conditions. Given the potential for obtaining strong but unstable support
667 under some conditions, we recommend utilization of the heat map approach we develop here to
668 ensure more thorough analyses in difficult regions of the tree. Adoption of the heat-map
669 approach may temper the claims of systematists favoring relationships that appear to be strongly
670 supported yet are sensitive to analysis conditions or the particular choice of sites and loci to
671 include in their dataset.

672 We made a substantial effort to target only low-copy loci when developing the AHE
673 resource. Use of low-copy loci in hybrid enrichment-based phylogenomics is especially

674 important because probes enrich not only the target regions, but also regions with moderate to
675 high sequence similarity to the target regions. Paralogous gene regions will be enriched to a
676 degree inversely proportional to their evolutionary divergence from the target region (Lemmon et
677 al. 2012). In order to minimize potential error arising when enriched sequences are assessed for
678 orthology, we estimated copy number in 45 divergent angiosperms species prior to selecting
679 target loci for the Angiosperm v.1 design. Use of a much smaller set of taxa tends to result in an
680 over-fitting problem wherein the taxa happen by chance to each have a single copy for some
681 portion of loci (e.g., due to coincidental gene loss). Unsurprisingly, many of the loci previously
682 identified as low copy based on only four taxa (*Arabidopsis, Populus, Vitis,* and *Oryza*; Duarte et
683 al. 2010) were found to have a moderate to high number of copies in other angiosperm species
684 (Table 3). This finding largely explains why many of the initial 959 loci identified by Duarte et
685 al. (2010) were removed from our locus list as we refined our target set to 499 loci. The potential
686 for over-fitting is especially likely in taxa with a history of whole genome duplications and
687 subsequent gene loss (Glasauer and Neuhauss 2014) . In teleost fishes, for example,
688 identification of a substantial number of single-copy hybrid enrichment targets proved to be
689 impossible despite work by Li et al. (2007), who identified 154 putatively single copy loci. In
690 this case, Stout et al. (*in press*) circumvented this problem by developing a teleost-wide
691 enrichment kit that targets up to six homologous copies per locus and then sorting out the gene
692 copies into orthologs bioinformatically after sequencing. Analysis of enriched loci in teleosts and
693 other taxa with extensive gene duplications/losses may require simultaneous estimation of
694 speciation and duplication history (Cui et al. 2006; De Smet et al. 2013).

695 Many of the published studies utilizing hybrid enrichment in angiosperms have targeted
696 the plastome, either exclusively or in combination with nuclear loci (Table 1). One challenge

697 with simultaneously targeting the plastome and nuclear genome is overcoming the unequal
698 coverage resulting from the large disparity in copy number between the two genomes. Some
699 researchers desiring data from both genomes have mitigated the coverage disparity by
700 performing two separate enrichments using kits designed separately for the two genomes
701 (Heyduk et al. 2016; Sass et al. 2016). While this approach can be successful, it is more
702 expensive and time consuming than AHE, and it also requires some knowledge of approximate
703 plastid copy number. An alternative approach that has only recently been discovered is to take
704 advantage of the fact that enrichment of nuclear targets is not perfectly efficient. Since some
705 fraction of the sequenced library fragments are from off-target regions, it is reasonable to expect
706 that some portion of this by-catch will be derived from the plastid genome. Indeed, one can often
707 assemble whole plastid genomes from the enrichment by-catch (Weitemierx x et al. 2014;
708 Stephens et al. 2015a, 2015b; Schmickl et al. 2016). Although the success of this approach
709 depends on plastome copy number and enrichment efficiency of targeted nuclear loci, the levels
710 of enrichment we obtained across angiosperms (2%-22% on target; Table 3) suggest that the by-
711 catch of AHE in angiosperms is likely to contain sufficient plastome sequence reads to enable
712 reliable reconstruction of whole plastid genomes. Note, however, that this expectation may not
713 be met when target taxa are very closely related to one of the 25 references used to develop the
714 probes, since enrichment efficiency in these taxa is expected to be highest. Although
715 reconstruction of plastid genomes is not a focus of this study, the raw reads generated during
716 enrichment of the 53 taxa used in this study could be mined for plastid genomes. For example,
717 using a reference species from the clade as a guide, assemblies were easily constructed for the
718 plastid genome in Cyperaceae (Buddenhagen, unpubl. data).

719 The diversification of angiosperm lineages continues to be debated despite the
720 application of large genomic data sets and sophisticated phylogenetic models, leading some to
721 wonder if we are closer to resolving Darwin's (1903) "abominable mystery". Our study
722 demonstrates that despite the conflict in recent studies of deep angiosperm relationships (e.g.,
723 Duarte et al. 2010; Lee et al. 2011; Wickett et al. 2014), one hypothesis predominates (eudicots-
724 monocots sister) over the majority of parameter space under both supermatrix and species tree
725 methods of inference. This relationship is supported by the most recent studies focusing on the
726 plastid genome (Moore et al. 2007; Soltis et al. 2011; Ruhfel et al. 2014) and the Angiosperm
727 Phylogeny Group's recent synthesis (APG IV 2016). The predominance of one hypothesis is not
728 obvious at first glance, however, because an alternative hypothesis (i.e., eudicots-magnoliids
729 sister) is strongly supported under a narrow but commonly chosen set of conditions (supermatrix
730 with < 5% of sites removed). The placement of Malpighiales within the Malvids has also been
731 debated (APG IV 2016). Despite support for alternative placement of Malpighiales by the
732 Angiosperm Phylogeny Group, our analysis suggests robust support for Malpighiales within
733 Malvids across a broad range of conditions. Our heat map-based analyses of robustness
734 indicates that the differences between alternative studies of these and other deep angiosperm
735 relationships could be reconciled through more detailed exploration of the sensitivity of each
736 study's findings to the particular conditions under which these studies were conducted. One
737 especially interesting extension to the study presented herein would be to incorporate the degree
738 of taxon sampling into the analysis of robustness, especially with respect to inclusion of the early
739 branching taxa that were admittedly under-sampled in this study.

740 The taxa sampled in this study were chosen to allow the enrichment efficiency of the
741 probe kit and the phylogenomic signal to be assessed at deep, intermediate, and shallow scales.

742 Data from additional taxa were obtained from available genomic resources in order to increase
743 the coverage across the angiosperm clade. The phylogeny estimated using these samples was
744 useful for demonstrating the utility of the heat map approach at identifying robust phylogenetic
745 estimates and for testing alternative hypotheses. Nonetheless, the taxon sampling could certainly
746 be improved especially for basal lineages. Although the sensitivity of the phylogenetic estimates
747 to differential taxon sampling could be assessed using the heat map approach, such an analysis is
748 outside the scope of this paper, and so we leave this effort for future studies.

749

750 **CONCLUDING REMARKS**

751 The anchored enrichment resource we developed provides an efficient way of obtaining a
752 sufficient amount of phylogenomic data for deep, intermediate, and shallow angiosperm studies.
753 Although most angiosperm relationships could be robustly resolved, support for a small minority
754 of relationships varied substantially depending on the particular subset of sites and loci analyzed,
755 suggesting that the sensitivity approach we developed here may helpful for avoiding false
756 confidence in hypotheses that supported under only a narrow set of conditions. Use of this
757 holistic approach may enable systematists to reconcile differences across studies as they work
758 towards a unified understanding the history of angiosperm diversification.

759

760 **ACKNOWLEDGEMENTS**

761 We are grateful to Alyssa Bigelow, Kirby Birch, Ameer Jalal, and Sean Holland at FSU's
762 Center for Anchored Phylogenomics for assistance with molecular data collection and
763 bioinformatics analysis. We are also grateful to Roger Mercer and Yanming Yang at FSU's
764 College of Medicine for sequencing services. This research was partially funded by NSF DDIG

765 (DEB-1311150) to CEB and by a FSU Planning Grant awarded to ARM. EML acknowledges
766 support from NSF DEB-1120516. ARL and EML were supported by NSF IIP-1313554. CP
767 would like to acknowledge the generous support of the Hermon Slade Foundation, Emily
768 Holmes Memorial Scholarship; Hansjörg Eichler Scientific Research Fund; Systematics
769 Research Fund, without which research material used in this publication would not have been
770 collected. NM was supported by NSF DEB-1046328. SW thanks the TU Dresden and especially
771 Christoph Neinhuis for continuous support. AM was supported by NSF DEB-1252901 to EJE.
772 PSS is grateful for support from iDigBio (NSF grant EF-1115210) and a Dimensions of
773 Biodiversity grant (DEB-1442280).

774

775 **LITERATURE CITED**

776 APG IV 2016. An update of the Angiosperm Phylogeny Group classification for the orders and
777 families of flowering plants: APG IV. Bot. J. Linn. Soc. 181:1–20.

778 Bakker F.T. 2015. DNA sequences from plant herbarium tissue. In: Appelhans M.S., Hörandl E.,
779 editors. Next-Generation Sequencing in Plant Systematics. International Association for
780 Plant Taxonomy.

781 Beck J.B., Semple J.C. 2015. Next-generation sampling: Pairing genomics with herbarium
782 specimens provides species-level signal in *Solidago* (Asteraceae). Appl. Plant Sci.
783 3:apps.1500014.

784 Betancur-R. R., Naylor G.J.P., Ortí G. 2014. Conserved Genes, Sampling Error, and
785 Phylogenomic Inference. Syst. Biol. 63:257–262.

786 Bi K., Linderoth T., Vanderpool D., Good J.M., Nielsen R., Moritz C. 2013. Unlocking the vault:
787 next-generation museum population genomics. Mol. Ecol. 22:6018–6032.

788 Bogdanowicz D., Giaro K., Wróbel B. 2012. TreeCmp: Comparison of trees in polynomial time.
789 Evol. Bioinform. 8:475–487.

790 Brandley M.C., Bragg J.G., Singhal S., Chapple D.G., Jennings C.K., Lemmon A.R., Lemmon
791 E.Moriarty, Thompson M.B., Moritz C. 2015. Evaluating the performance of anchored
792 hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian
793 *Eugongylus* group scincid lizards. BMC Evol. Biol. 15:1–14.

794 Breinholt J.W., Kawahara A.Y. 2013. Phylotranscriptomics: Saturated Third Codon Positions
795 Radically Influence the Estimation of Trees Based on Next-Gen Data. Genome Biol. Evol.
796 5:2082–2092.

797 Breinholt, J.W., Lemmon A.R., Lemmon E.M., Xiao L., and Kawahara A.Y. Anchored hybrid
798 enrichment in Lepidoptera: Leveraging genomic data for studies on the megadiverse
799 butterflies and moths. Syst. Biol. *Accepted.*

800 Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of Bayes
801 factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.

802 Bryson Jr R.W., Faircloth B.C., Tsai W.L., McCormack J.E., Klicka J. 2016. Target enrichment
803 of thousands of ultraconserved elements sheds new light on early relationships within New
804 World sparrows (Aves: Passerellidae). The Auk. 133:451–458.

805 Burleigh J.G., Bansal M.S., Eulenstein O., Hartmann S., Wehe A., Vision T.J. 2011. Genome-
806 Scale Phylogenetics: Inferring the Plant Tree of Life from 18,896 Gene Trees. Syst. Biol.
807 60:117–125.

808 Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in
809 Phylogenetic Analysis. Mol. Biol. Evol. 17:540–552.

810 Chase M.W., Soltis D.E., Olmstead R.G., Morgan D., Les D.H., Mishler B.D., Duvall M.R.,
811 Price R.A., Hills H.G., Qiu Y.-L., Kron K.A., Rettig J.H., Conti E., Palmer J.D., Manhart
812 J.R., Sytsma K.J., Michaels H.J., Kress W.J., Karol K.G., Clark W.D., Hedren M., Gaut B.S.,
813 Jansen R.K., Kim K-J., Wimpee C.F., Smith J.F., Furnier G.R., Strauss S.H., Xiang Q-Y,
814 Plunkett G.M., Soltis P.S., Swensen S.M., Williams S.E., Gadek P.A., Quinn C.J., Eguiarte
815 L.E., Golenberg E., Learn Jr. G.H., Graham S.W., Barrett S.C.H., Dayanandan S., Albert
816 V.A. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid
817 gene rbcL. Ann. Mo. Bot. Gard. 80:528–580.

818 Chen F., Mackey A.J., Vermunt J.K., Roos D.S. 2007. Assessing performance of orthology
819 detection strategies applied to eukaryotic genomes. PloS One. 2:e383.

820 Costa C.M., Roberts R.P. 2014. Techniques for improving the quality and quantity of DNA
821 extracted from herbarium specimens. Phytoneuron. 48:1–8.

822 Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012.
823 More than 1000 ultraconserved elements provide evidence that turtles are the sister group of
824 archosaurs. Biol. Lett. 8:783–786.

825 Crawford N.G., Parham J.F., Sellas A.B., Faircloth B.C., Glenn T.C., Papenfuss T.J., Henderson
826 J.B., Hansen M.H., Simison W.B. 2015. A phylogenomic analysis of turtles. Mol.
827 Phylogenet. Evol. 83:250–257.

828 Critchlow D.E., Pearl D.K., Qian C. 1996. The triples distance for rooted bifurcating
829 phylogenetic trees. Syst. Biol. 45:323–334.

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

832 Cui L., Wall P.K., Leebens-Mack J.H., Lindsay B.G., Soltis D.E., Doyle J.J., Soltis P.S., Carlson
833 J.E., Arumuganathan K., Barakat A. 2006. Widespread genome duplications throughout the
834 history of flowering plants. Genome Res. 16:738–749.

835 Darriba D., Posada D. 2015. The impact of partitioning on phylogenomic accuracy. bioRxiv doi:
836 http://dx.doi.org/10.1101/023978.

837 Darwin C., Darwin F., and Seward A.C. 1903. More letters of Charles Darwin: a record of his
838 work in a series of hitherto unpublished letters. D. Appleton and Co., New York.

839 Davis C.C., Xi Z., Mathews S. 2014. Plastid phylogenomics and green plant phylogeny: almost
840 full circle but not quite there. BMC Biol. 12:1–4.

841 De Smet R., Adams K.L., Vandepoele K., Van Montagu M.C.E., Maere S., Van de Peer Y. 2013.
842 Convergent gene loss following gene and genome duplications creates single-copy families
843 in flowering plants. Proc. Natl. Acad. Sci. 110:2898–2903.

844 Duarte J.M., Wall P.K., Edger P.P., Landherr L.L., Ma H., Pires J.C., Leebens-Mack J. 2010.
845 Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza
846 and their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10:61.

847 Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution.
848 63:1–19.

Edwards S.V. 2016. Phylogenomic subsampling: a brief review. Zoologica Scripta 45: 63-74.

Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94:447–462.

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. 15:1.

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 timescales. Syst. Biol. 61:717–726.

Frandsen P.B., Calcott B., Mayer C., Lanfear R. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol. Biol. 15:1–17.

Freeman J.L., Tamaoki M., Stushnoff C., Quinn C.F., Cappa J.J., Devonshire J., Fakra S.C., Marcus M.A., McGrath S.P., Van Hoewyk D., Pilon-Smits E.A.H. 2010. Molecular mechanisms of selenium tolerance and hyperaccumulation in *Stanleya pinnata*. Plant Physiol. 153:1630–1652.

Glasauer S.M., Neuhauss S.C. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol. Genet. Genomics. 289:1045–1060.

868 Griffin P., Robin C., Hoffmann A. 2011. A next-generation sequencing method for overcoming
869 the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. BMC
870 Biol. 9:1–19.

871 Grover C.E., Gallagher J.P., Jareczek J.J., Page J.T., Udall J.A., Gore M.A., Wendel J.F. 2015.
872 Re-evaluating the phylogeny of allopolyploid *Gossypium* L. Molecular Phylogenetics and
873 Evolution. 92:45–52.

874 Hamilton C.A., Lemmon A.R., Lemmon E. Moriarty, Bond J.E. 2016. Expanding anchored
875 hybrid enrichment to resolve both deep and shallow relationships within the spider tree of
876 life. BMC Evolutionary Biology 16:212.

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

880 Hillis D.M., Huelsenbeck J.P., Cunningham C.W. 1994. Application and accuracy of molecular
881 phylogenies. Science 264:671–677.

882 Hoff M., Orf S., Riehm B., Darriba D., Stamatakis A. 2016. Does the choice of nucleotide
883 substitution models matter topologically? BMC Bioinformatics 17:1–13.

884 Hosner P.A., Faircloth B.C., Glenn T.C., Braun E.L., Kimball R.T. 2015. Avoiding missing data
885 biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes).
886 Mol. Biol. Evol. 33:1110–1125.

887 Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho
888 L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C.,

889 Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and
890 angiosperms. Nature. 473:97–100.

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

893 Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A.,
894 Markowitz S., Duran C. 2012. Geneious Basic: an integrated and extendable desktop
895 software platform for the organization and analysis of sequence data. Bioinformatics.
896 28:1647–1649.

897 Kluge A.G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among
898 *Epicrates* (Boidae, Serpentes). Syst. Biol. 38:7–25.

899 Lanfear R., Calcott B., Ho S.Y., Guindon S. 2012. PartitionFinder: combined selection of
900 partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol.
901 29:1695–1701.

902 Leaché A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: a
903 comparison of methods. Syst. Biol. 60:126–137.

904 Lee E.K., Cibrian-Jaramillo A., Kolokotronis S.-O., Katari M.S., Stamatakis A., Ott M., Chiu
905 J.C., Little D.P., Stevenson D.W., McCombie W.R. 2011a. A functional phylogenomic view
906 of the seed plants. PLoS Genet. 7:e1002411.

907 Lemmon A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E. Moriarty. 2009. The effect of
908 ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian
909 inference. Syst. Biol. 58:130–145.

910 Lemmon A.R., Emme S.A., Lemmon E. Moriarty 2012. Anchored Hybrid Enrichment for
911 Massively High-Throughput Phylogenomics. Syst. Biol. 61:727–744.

912 Lemmon A.R., Moriarty E.C. 2004. The importance of proper model assumption in Bayesian
913 phylogenetics. Syst. Biol. 53:265–277.

914 Lemmon E.M., Lemmon A.R. 2013. High-Throughput Genomic Data in Systematics and
915 Phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44:99–121.

916 Lemon J. 2006. Plotrix: a package in the red light district of R. R-News. 6:8–12.

917 Léveillé-Bourret E., Starr J.R., Ford B.A., Lemmon E.M., Lemmon A.R. *In review*. Anchored
918 phylogenomics for angiosperms II: Resolving rapid generic and tribal-level radiations. Syst.
919 Biol.

920 Li C., Ortí G., Zhang G., Lu G. 2007. A practical approach to phylogenomics: the phylogeny of
921 ray-finned fish (Actinopterygii) as a case study. BMC Evol. Biol. 7:44–44.

922 Magallón S., Hilu K.W., Quandt D. 2013. Land plant evolutionary timeline: Gene effects are
923 secondary to fossil constraints in relaxed clock estimation of age and substitution rates. Am.
924 J. Bot. 100: 556–573.

925 Mamanova L., Coffey A.J., Scott C.E., Kozarewa I., Turner E.H., Kumar A., Howard E.,
926 Shendure J., Turner D.J. 2010. Target-enrichment strategies for next-generation sequencing.
927 Nat. Methods. 7:111–118.

928 Mandel J.R., Dikow R.B., Funk V.A., Masalia R.R., Staton S.E., Kozik A., Michelmore R.W.,
929 Rieseberg L.H., Burke J.M. 2014. A Target Enrichment Method for Gathering Phylogenetic

930 Information from Hundreds of Loci: An Example from the Compositae. Appl. Plant Sci.
931 2:1300085.

932 McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C.
933 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental
934 mammal phylogeny when combined with species-tree analysis. Genome Res. 22:746–754.

935 Meiklejohn K.A., Faircloth B.C., Glenn T.C., Kimball R.T., Braun E.L. 2016. Analysis of a
936 Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some
937 Multispecies Coalescent Methods. Syst. Biol. 65:612-627.

938 Meredith R.W., Janecka J.E., Gatesy J., Ryder O.A., Fisher C.A., Teeling E.C., Goodbla A.,
939 Eizirik E., Simao T.L.L., Stadler T., Rabosky D.L., Honeycutt R.L., Flynn J.J., Ingram C.M.,
940 Steiner C., Williams T.L., Robinson T.J., Burk-Herrick A., Westerman M., Ayoub N.A.,
941 Springer M.S., Murphy W.J. 2011. Impacts of the Cretaceous terrestrial revolution and KPg
942 extinction on mammal diversification. Science 334:521–524.

943 Meyer M., Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed
944 target capture and sequencing. Cold Spring Harb. Protoc. 2010; doi:10.1101/pdb.prot5448.

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

947 Moore M.J., Bell C.D., Soltis P.S., Soltis D.E. 2007. Using plastid genome-scale data to resolve
948 enigmatic relationships among basal angiosperms. Proc. Natl. Acad. Sci. 104:19363–19368.

949 Morales H.E., Pavlova A., Sunnucks P., Major R., Amos N., Joseph L., Lemmon A.R., Endler
950 J.A., Delhey K. Neutral and selective drivers of colour evolution in a widespread Australian
951 passerine. J. Biogeogr.; http://hdl.handle.net/11858/00-001M-0000-002B-11FF-D *In review*

952 Morrison D. 2011. Introduction to Phylogenetic Networks. RJR Publications. Uppsala, Sweden.

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

955 Parks M., Cronn R., Liston A. 2012. Separating the wheat from the chaff: mitigating the effects
956 of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). BMC Evol. Biol.
957 12:1–17.

958 Peloso P.L.V., Frost D.R., Richards S.J., Rodrigues M.T., Donnellan S., Matsui M., Raxworthy
959 C.J., Biju S.D., Lemmon E.M., Lemmon A.R., Wheeler W.C. 2016. The impact of anchored
960 phylogenomics and taxon sampling on phylogenetic inference in narrow-mouthed frogs
961 (Anura, Microhylidae). Cladistics 32:113–140.

962 Pimm S.L., Joppa L.N. 2015. How many plant species are there, where are they, and at what rate
963 are they going extinct? Ann. Mo. Bot. Gard. 100:170–176.

964 Posada D., Crandall K.A. 1998. Modeltest: testing the model of DNA substitution.
965 Bioinformatics 14:817–818.

966 Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R.
967 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA
968 sequencing. Nature. 526:569–573.

969 Pyron R.A., Hendry C.R., Chou V.M., Lemmon E.M., Lemmon A.R., Burbrink F.T. 2014.
970 Effectiveness of phylogenomic data and coalescent species-tree methods for resolving
971 difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). Mol.
972 Phylogenet. Evol. 81:221–231.

973 de Queiroz A., Gatesy J. 2007. The supermatrix approach to systematics. Trends Ecol. Evol.
974 22:34–41.

975 Reneker J., Lyons E., Conant G.C., Pires J.C., Freeling M., Shyu C.-R., Korkin D. 2012. Long
976 identical multispecies elements in plant and animal genomes. Proc. Natl. Acad. Sci.
977 109:E1183–E1191.

978 Ripplinger J., Sullivan J. 2008. Does choice in model selection affect maximum likelihood
979 analysis? Syst. Biol. 57:76–85.

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

982 Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from
983 empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

984 Ruane S., Raxworthy C.J., Lemmon A.R., Lemmon E.M., Burbrink F.T. 2015. Comparing
985 species tree estimation with large anchored phylogenomic and small Sanger-sequenced
986 molecular datasets: an empirical study on Malagasy pseudoxyrhophiine snakes. BMC Evol.
987 Biol. 15:1.

988 Ruhfel B.R., Gitzendanner M.A., Soltis P.S., Soltis D.E., Burleigh J.G. 2014. From algae to
989 angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid
990 genomes. BMC Evol. Biol. 14:23.

991 Sass C., Iles W.J.D., Barrett C.F., Smith S.Y., Specht C.D. 2016. Revisiting the Zingiberales:
992 using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a
993 charismatic plant lineage. Peerj. 4:e1584.

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

998 Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree
999 selection. Bioinformatics. 17:1246–1247.

1000 Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol.
1001 51: 492-508.

1002 Simmons, M. P., J. Gatesy. 2015. Coalescence vs. concatenation: sophisticated analyses vs. first
1003 principles applied to rooting the angiosperms. Mol. Phylogenet. Evol. 91: 98-122.

1004 Simmons, M. P., J. Gatesy. 2016. Biases of tree-independent-character-subsampling methods.
1005 Mol. Phylogenet. Evol. 100: 424-443.

1006 Simmons M.P., Sloan D.B., Gatesy J. 2016. The effects of subsampling gene trees on coalescent
1007 methods applied to ancient divergences. Mol. Phylogenet. Evol. 97:76–89.

1008 Song S., Liu L., Edwards S.V., Wu S., 2012. Resolving conflict in eutherian mammal phylogeny
1009 using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA
1010 109:14942–14947.

1011 Springer M. S., J. Gatesy. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1-33.

1012 Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-
1013 Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S. 2011. Angiosperm phylogeny: 17
1014 genes, 640 taxa. Am. J. Bot. 98:704–730.

1015 de Sousa F., Bertand Y.J.K., Nylinder S., Oxelman B., Eriksson J.S., Pfeil B.E. 2014.
1016 Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using
1017 multiplexed sequence capture and next-generation sequencing. PLoS ONE 9:e109704. doi:
1018 10.1371/journal.pone.0109704.

1019 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1020 large phylogenies. Bioinformatics. 30:1312–1313.

1021 Stephens J.D., Rogers W.L., Heyduk K., Cruse-Sanders J.M., Determann R.O., Glenn T.C.,
1022 Malmberg R.L. 2015a. Resolving phylogenetic relationships of the recently radiated
1023 carnivorous plant genus *Sarracenia* using target enrichment. Mol. Phylogenet. Evol. 85:76–
1024 87.

1025 Stephens J.D., Rogers W.L., Mason C.M., Donovan L.A., Malmberg R.L. 2015b. Species tree
1026 estimation of diploid *Helianthus* (Asteraceae) using target enrichment. Am. J. Bot. 102:910–
1027 920.

1028 Stout, C.C., Tan M., Lemmon A.R., Lemmon E.M., Armbruster J.W. Resolving Cypriniformes
1029 Relationships Using An Anchored Enrichment Approach. BMC Evol. Biol. *In Review.*

1030 Stull G.W., Moore M.J., Mandala V.S., Douglas N.A., Kates H.-R., Qi X., Brockington S.F.,
1031 Soltis P.S., Soltis D.E., Gitzendanner M.A. 2013. A targeted enrichment strategy for
1032 massively parallel sequencing of angiosperm plastid genomes. Appl. Plant Sci. 1:1200497.

1033 Sun Y., Moore M.J., Zhang S., Soltis P.S., Soltis D.E., Zhao T., Meng A., Li X., Li J., Wang H.
1034 2016. Phylogenomic and structural analyses of 18 complete plastomes across nearly all
1035 families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene
1036 content evolution. Mol. Phylogenet. Evol. 96:93–101.

1037 Syring J., Cronn R., Tennessen J.A., Jennings T.N., Scelfo-Dalbey C., Wegrzyn J. 2016.
1038 Targeted capture sequencing in whitebark pine reveals range-wide demographic and adaptive
1039 patterns despite challenges of a large, repetitive genome. Frontiers in Plant Science. 7:484.

1040 Talavera G., Castresana J. 2007. Improvement of phylogenies after removing divergent and
1041 ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–577.

1042 Berardini T.Z., Reiser L, Li D., Mezheritsky Y., Muller R., Strait E., Huala E. 2015. The
1043 Arabidopsis Information Resource: Making and mining the "gold standard" annotated
1044 reference plant genome. Genesis. 53:474-485.

1045 Terry N., Zayed A.M., de Souza M.P., Tarun A.S. 2000. Selenium in higher plants. Ann. Rev.
1046 Plant Physiol. Plant Mol. Biol. 51:401–432.

1047 Townsend, T.M., Mulcahy, D.G., Noonan, B.P., Sites, J.W., Kuczynski, C.A., Wiens, J.J.,
1048 Reeder, T.W., 2011. Phylogeny of iguanian lizards inferred from 29 nuclear loci, and a

1049 comparison of concatenated and species-tree approaches for an ancient, rapid radiation. Mol.
1050 Phylogenet. Evol. 61:363–380.

1051 Tucker D.B., Colli G.R., Giugliano L.G., Hedges S.B., Hendry C.R., Lemmon E. Moriarty,
1052 Lemmon A.R., Pyron R.A. 2016. Methodological congruence in phylogenomic analyses with
1053 morphological support for teiid lizards (Sauria: Teiidae). Mol. Phylogenet. Evol. 103:75–84.

1054 Weitemier K., Straub S.C., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A.
1055 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant
1056 phylogenomics. Appl. Plant Sci. 2:1400042.

1057 Wicke S., Schneeweiss G.M. 2015. Next-generation organellar genomics: Potentials and pitfalls
1058 of high-throughput technologies for molecular evolutionary studies and plant systematics. In:
1059 Horandl E., Appelhans M.S., editors. Next Generation Sequencing in Plant Systematics.
1060 Koenigstein: A R G Gantner Verlag K G. Pp. 9–50.

1061 Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S.,
1062 Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham
1063 S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J.,
1064 Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B.,
1065 Philippe H., dePamphilis C.W., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M.,
1066 Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S.,
1067 Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of
1068 land plants. Proc. Natl. Acad. Sci. 111:E4859–E4868.

1069 Xia X., Lemey P. 2009. Assessing substitution saturation with DAMBE. In: Lemey P., Salemi
1070 M., and Vandamme A.-M., editors. The Phylogenetic Handbook: a Practical Approach to

1071 Phylogenetic Analysis and Hypothesis Testing. Cambridge University Press, UK. Pp. 611–
1072 626.

1073 Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable
1074 rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

1075 Young A.D., Lemmon A.R., Skevington J.H., Mengual X., Ståhls G., Reemer M., Jordaens K.,
1076 Kelso S., Lemmon E. Moriarty, Hauser M. 2016. Anchored enrichment dataset for true flies
1077 (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). BMC
1078 Evol. Biol. 16:143.

1079 Zeng L., Zhang Q., Sun R., Kong H., Zhang N., Ma H. 2014. Resolution of deep angiosperm
1080 phylogeny using conserved nuclear genes and estimates of early divergence times. Nat
1081 Commun. 5:4956.

1082 Zimmer E.A., Wen J. 2012. Using nuclear gene data for plant phylogenetics: Progress and
1083 prospects. Mol. Phylogenet. Evol. 65:774–785.

1084 Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error.
1085 Syst. Biol. 51:588–598.

1086

FIGURE CAPTIONS

Figure 1. Pipeline for locus selection and probe design. Candidate loci were derived from Duarte et al. (2010) and filtered to produce anchor loci that each contained one low-copy exon with sufficient degree of conservation across angiosperms. Eighteen functional targets involved in selenium tolerance were added prior to probe tiling. Scripts used in probe design are available in Dryad (accession XXXX).

Figure 2. Standard pipeline for processing anchored phylogenomic data to produce final alignments. Merged read pairs are assembled using a quasi-de novo approach that utilizes divergent references from the kit. After orthologous sets of consensus sequences are identified, raw alignments are constructed using MAFFT and trimmed/masked to remove probable misaligned regions. Alignments are finalized after being manually inspected and trimmed in Geneious. Scripts are available in Dryad (accession XXXX).

Figure 3. Pipeline for generating heat maps representing robustness of phylogenetic estimates to variation in site and locus filtering. Labeled arrows involve steps performed using standard methods, whereas unlabeled arrows involve steps performed using custom scripts available in Dryad (accession XXXX).

Figure 4. Strategy for generating 196 data filtering conditions to assess robustness of phylogenetic estimates. (a) Sites are ranked by rate as estimated using kmeans in PartitionFinder. Invariable sites are plotted to right of the distribution. Fourteen site-inclusion thresholds (gray lines/arrows on histogram) are applied, with the most variable sites being excluded at the lowest thresholds. Gene trees constructed after applying each of these thresholds are compared pairwise using the triplet measure in TreeComp to produce a pairwise tree-distance matrix. (b) Plotting loci in multidimensional space (MDS) using corresponding tree-distances allows for the identification of outliers (furthest from center) and 14 thresholds for inclusion of loci for phylogeny estimation (gray circles). (c) Example of heat-map for one node showing phylogenetic support as color for each of the 14×14=196 combinations/data sets, with white indicating no support for the node and black indicating 100% support for the node. (d) Thresholds

1110 (gray lines/arrows) plotted on the distribution of loci ranked by distance from origin on the MDS plot.
1111 Robust estimates show relatively consistent level of support (similar color) across the heat map in (c).

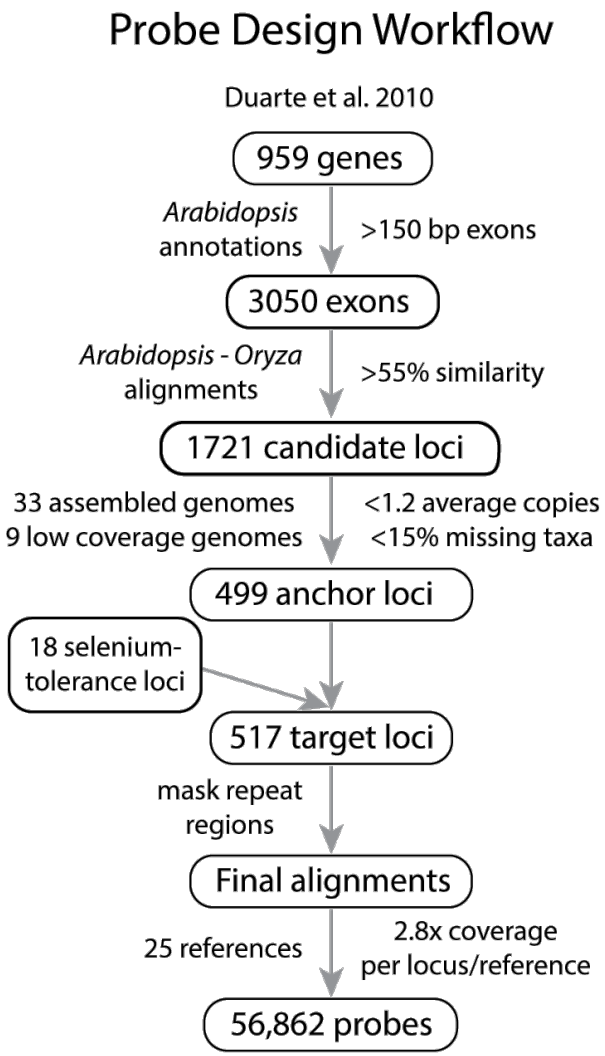
1112 **Figure 5.** The relationship between locus recovery and taxonomic representation in the enrichment kit.
1113 For each enriched sample, the number of loci recovered with consensus sequence greater than 250 bp is
1114 plotted against the divergence time (in millions of years - see Supplemental Methods for details) between
1115 that sample and the nearest reference species used for probe design. The curve represents the best-fit
1116 quadratic polynomial ($y = -0.0083x^2 + 0.3049x + 501.43$; $R^2 = 0.35734$).

1117 **Figure 6**. Estimates of the angiosperm phylogeny under a coalescent (ASTRAL) or supermatrix
1118 (concatenated RAxML) framework with all data included. Heat-maps on each internal node indicate
1119 bootstrap support across 196 data inclusion conditions (example shown in upper right), with lower-left
1120 corner of each heat map indicating support when all data are included (colors described in Fig. 4). Solid
1121 black heat maps indicate high support for a node under a broad range of conditions. Heat maps for nodes
1122 not present in both topologies are outlined in red.

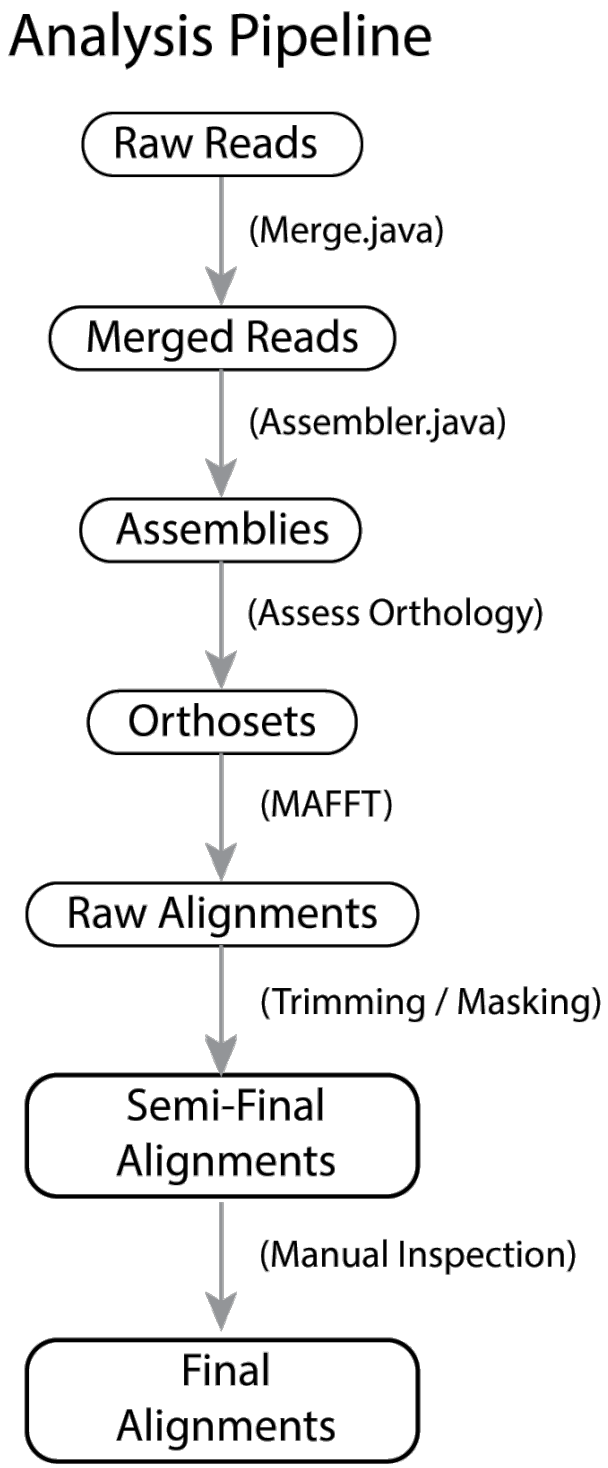
1123 **Figure 7** Using heat maps to compare robustness of support for key nodes across analysis frameworks.
1124 (a) Support for three alternative relationships among eudicots, monocots, and magnoliids under coalescent
1125 and supermatrix frameworks. Under the coalescent framework, a eudicot-monocot sister relationship is
1126 supported across a broad range of conditions (>63% of sites are included). Under the supermatrix
1127 approach, however, support shifts from strong support for a eudicot-magnoliid sister relationship when all
1128 sites are included to strong support for a eudicot-monocot sister relationship when 54%-76% of sites are
1129 included. (b) An additional approach for visualizing support for alternative hypotheses. In this
1130 representation, support for each of the three alternatives is indicated by a different color in the rgb (red-
1131 green-blue) color space, with the opacity of the color indicating the sum of support across the three
1132 alternatives (white indicates no support for any hypothesis). (c) A triangle representing rgb space can also
1133 be used to indicate support for alternative hypotheses. In this representation, each vertex of the triangle

correponds to 100% suport for one of the three alternative hypotheses (colors have same meaning as in [b]). For each one of the data sampling schemes (pixel in [b]) a yellow point is plotted on the triangle, at the location corresponding to the relative support for the three hypotheses (a point in the middle would indicate 33% support for each of the three alternative hypotheses, whereas a point half-way along one of the edges of the triangle would indicate 50% support for each of two alternatives).

Figure 1.


Probe Design Workflow
Duarte et al. 2010
959 genes
Arabidopsis annotations
>150 bp exons
3050 exons
Arabidopsis - Oryza alignments
>55% similarity
1721 candidate loci
33 assembled genomes
9 low coverage genomes
<1.2 average copies
<15% missing taxa
499 anchor loci
18 selenium-tolerance loci
517 target loci
mask repeat regions
Final alignments
25 references
2.8x coverage per locus/reference
56,862 probes

1147 **Figure 2.**



1148

1149

Figure 3.

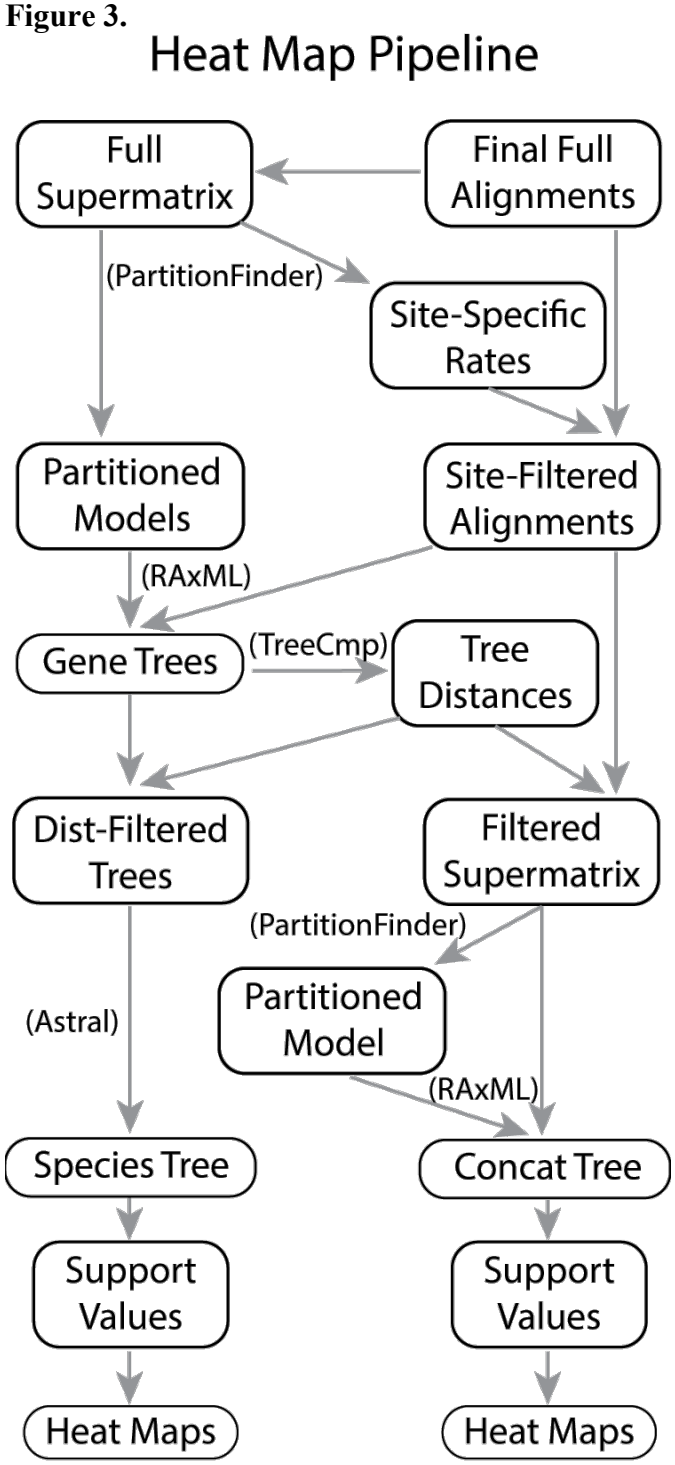

Heat Map Pipeline
Full Supermatrix
Final Full Alignments
(PartitionFinder)
Site-Specific Rates
Partitioned Models
Site-Filtered Alignments
(RAxML)
Gene Trees
(TreeCmp)
Tree Distances
Dist-Filtered Trees
Filtered Supermatrix
(PartitionFinder)
Partitioned Model
(Astral)
(RAxML)
Species Tree
Concat Tree
Support Values
Support Values
Heat Maps
Heat Maps

Figure 4.

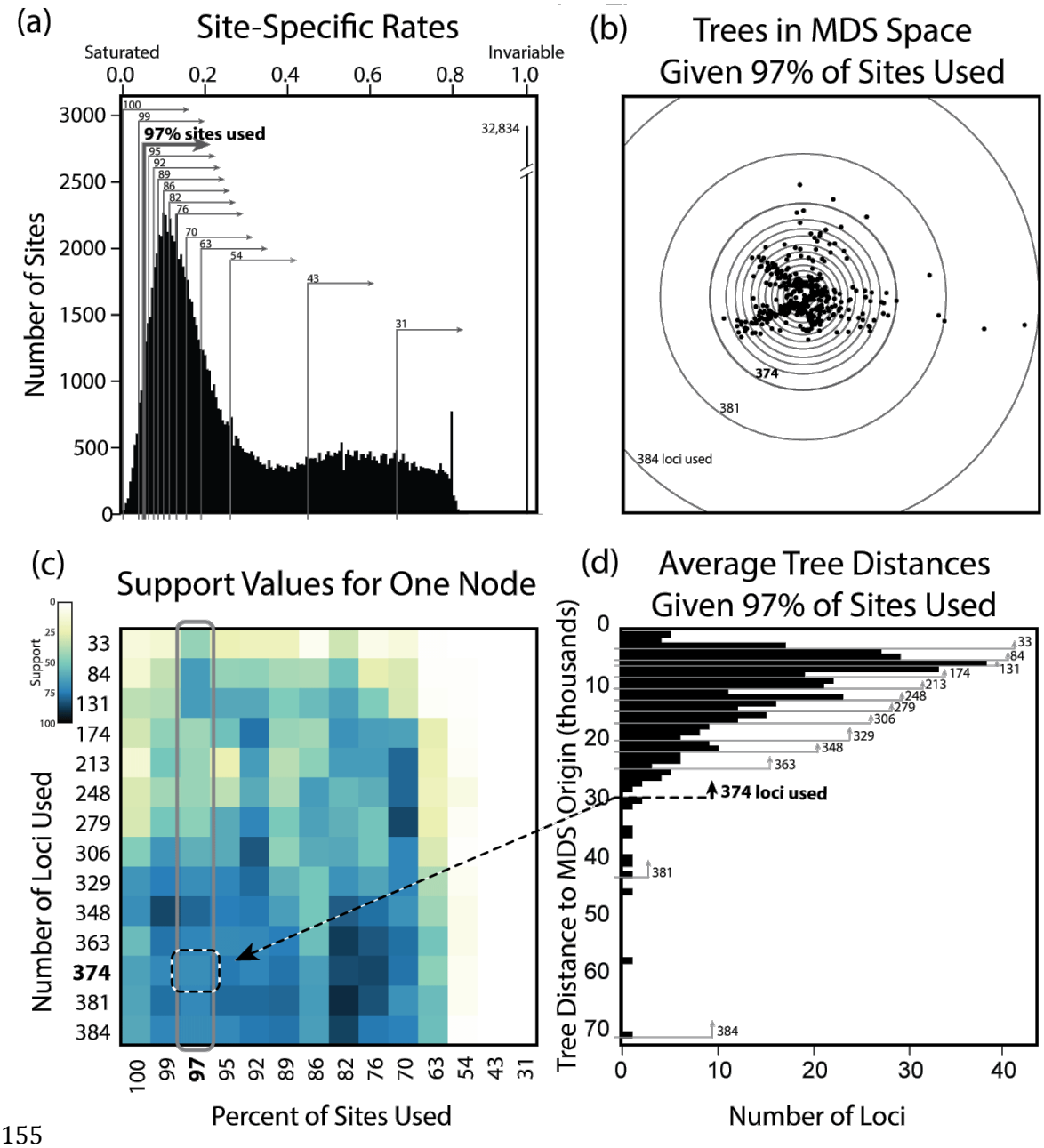

(a)
Site-Specific Rates
Saturated
Invariable
0.0
0.2
0.4
0.6
0.8
1.0
Number of Sites
3000
2500
2000
1500
1000
500
0
97% sites used
32,834
100
99
95
92
89
86
82
76
70
63
54
43
31
(b)
Trees in MDS Space
Given 97% of Sites Used
374
381
384 loci used
(c)
Support Values for One Node
Support
0
25
50
75
100
Number of Loci Used
33
84
131
174
213
248
279
306
329
348
363
374
381
384
Percent of Sites Used
(d)
Average Tree Distances
Given 97% of Sites Used
Tree Distance to MDS Origin (thousands)
374 loci used
Number of Loci

Figure 5.

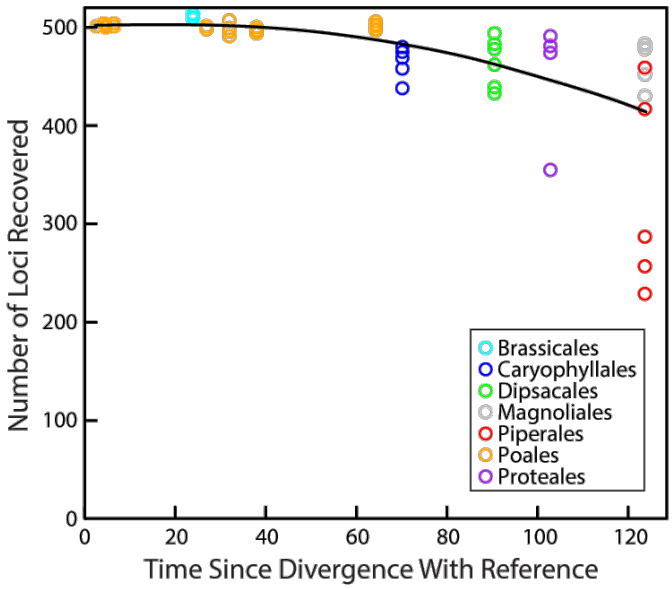

Number of Loci Recovered
Time Since Divergence With Reference
0
100
200
300
400
500
0
20
40
60
80
100
120
Brassicales
Caryophyllales
Dipsacales
Magnoliales
Piperales
Poales
Proteales

Figure 6.

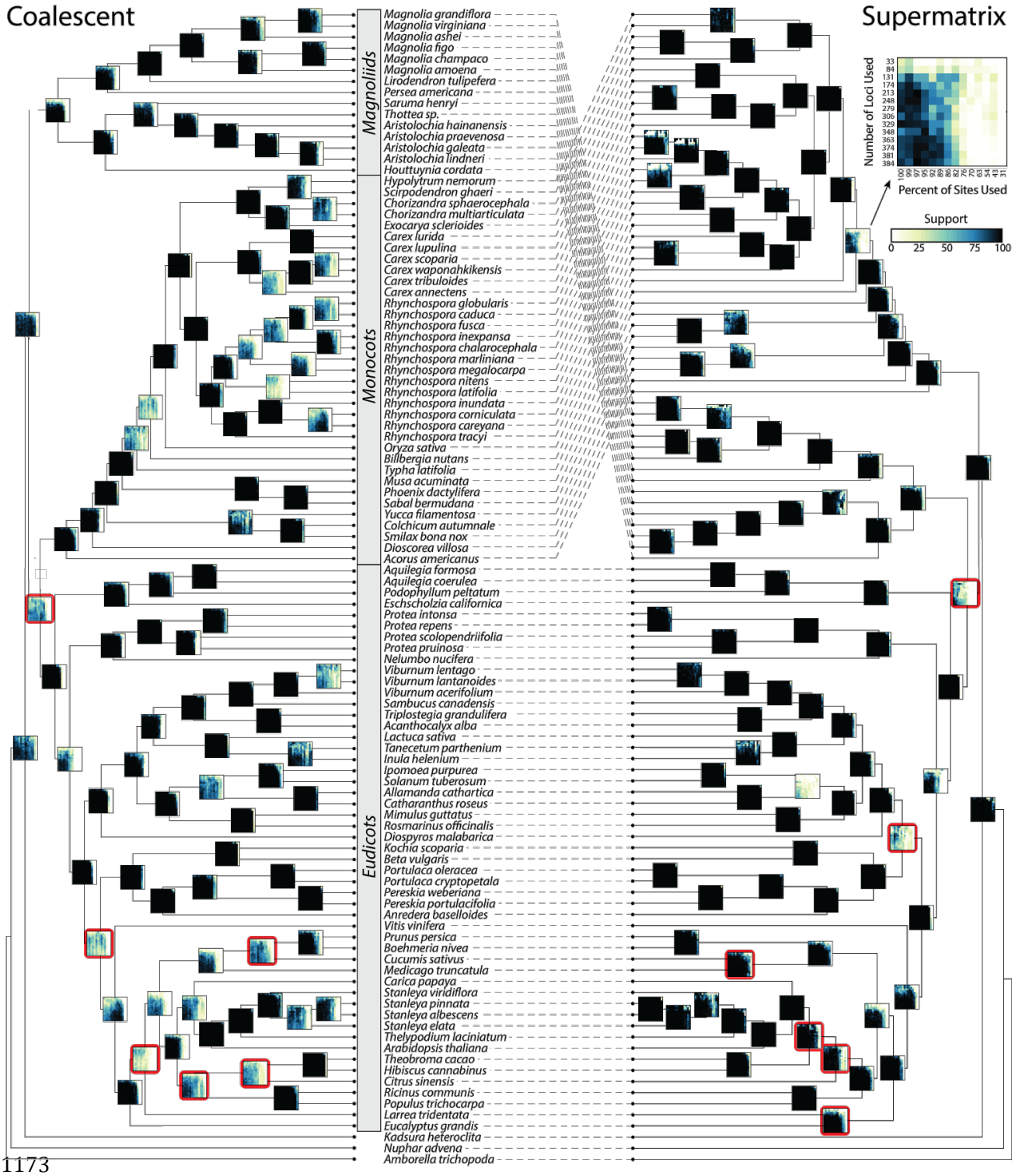
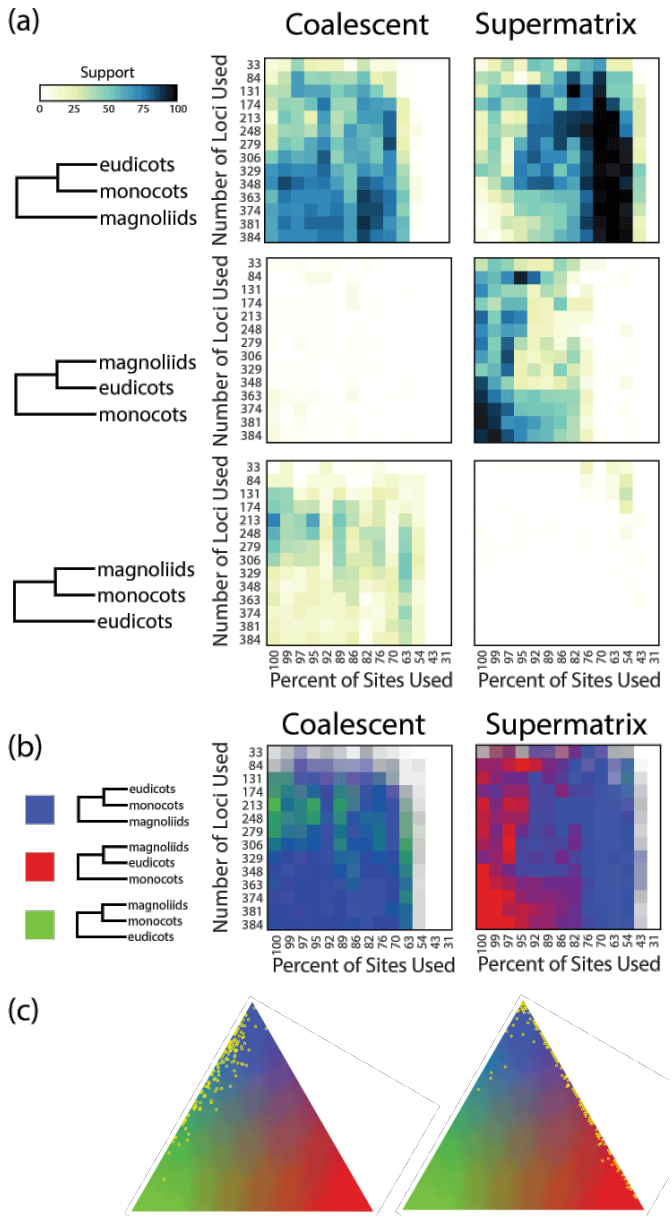

Coalescent
Supermatrix
Magnoliids
Monocots
Eudicots
Magnolia grandiflora
Magnolia virginiana
Magnolia ashei
Magnolia figo
Magnolia champaca
Magnolia amoena
Lirodendron tulipefera
Persea americana
Saruma henryi
Thottea sp.
Aristolochia hainanensis
Aristolochia praevenosa
Aristolochia galeata
Aristolochia lindneri
Houttuynia cordata
Hypolytrum nemorum
Scirpodendron ghaeri
Chorizandra sphaerocephala
Chorizandra multiarticulata
Exocarya sclerioides
Carex lurida
Carex lupulina
Carex scoparia
Carex waponahkikensis
Carex tribuloides
Carex annectens
Rhynchospora globularis
Rhynchospora caduca
Rhynchospora fusca
Rhynchospora inexpansa
Rhynchospora chalarocephala
Rhynchospora marliniana
Rhynchospora megalocarpa
Rhynchospora nitens
Rhynchospora latifolia
Rhynchospora inundata
Rhynchospora corniculata
Rhynchospora careyana
Rhynchospora tracyi
Oryza sativa
Billbergia nutans
Typha latifolia
Musa acuminata
Phoenix dactylifera
Sabal bermudana
Yucca filamentosa
Colchicum autumnale
Smilax bona nox
Dioscorea villosa
Acorus americanus
Aquilegia formosa
Aquilegia coerulea
Podophyllum peltatum
Eschscholzia californica
Protea intonsa
Protea repens
Protea scolopendriifolia
Protea pruinosa
Nelumbo nucifera
Viburnum lentago
Viburnum lantanoides
Viburnum acerifolium
Sambucus canadensis
Triplostegia grandulifera
Acanthocalyx alba
Lactuca sativa
Tanecetum parthenium
Inula helenium
Ipomoea purpurea
Solanum tuberosum
Allamanda cathartica
Catharanthus roseus
Mimulus guttatus
Rosmarinus officinalis
Diospyros malabarica
Kochia scoparia
Beta vulgaris
Portulaca oleracea
Portulaca cryptopetala
Pereskia weberiana
Pereskia portulacifolia
Anredera baselloides
Vitis vinifera
Prunus persica
Boehmeria nivea
Cucumis sativus
Medicago truncatula
Carica papaya
Stanleya viridiflora
Stanleya pinnata
Stanleya albescens
Stanleya elata
Thelypodium laciniatum
Arabidopsis thaliana
Theobroma cacao
Hibiscus cannabinus
Citrus sinensis
Ricinus communis
Populus trichocarpa
Larrea tridentata
Eucalyptus grandis
Kadsura heteroclita
Nuphar advena
Amborella trichopoda
Number of Loci Used
33
84
131
174
213
248
279
306
329
348
363
374
381
384
100
99
97
95
92
89
86
82
76
70
63
54
43
31
Percent of Sites Used
Support
0
25
50
75
100

Figure 7.


(a)
Coalescent
Supermatrix
Support
0 25 50 75 100
eudicots
monocots
magnoliids
Number of Loci Used
Percent of Sites Used
(b)
(c)

1177 **Table 1**. Recent phylogenetic or population genetic studies that used hybrid-enrichment methods, including plastid-oriented studies.

1178 Studies targeting only nuclear regions but that utilized plastid sequences found in bycatch are denoted by an asterisk (*). # Ref. Taxa

1179 refers to the number of reference taxa included in a probe design. # Sp. Copy Assessment indicates how many species the study

1180 included in their assessment of copy number for each locus included in the probe kit developed.

Study	Basal Group	Major Group	Family	Genus	# Ref. Taxa	# Nuclear Loci Targeted	Probes Target Plastids	# Sp. Copy Assessment
Parks et al. (2012)	Gymnosperms	-	Pinaceae	*Pinus*	2	0	yes	2
Stull et al. (2013)	Angiosperms	eudicots	-	-	22	0	yes	0
Mandel et al. (2014)	Angiosperms	eudicots	Asteraceae	-	1	1061	no	4
Weitmier et al. (2014)	Angiosperms	eudicots	Asteraceae	*Asclepias*	1	3385	no*	6
de Sousa et al. (2014)	Angiosperms	eudicots	Leguminosae	*Medicago*	1	319	no	5
Grover et al. (2015)	Angiosperms	eudicots	Malvaceae	*Gossypium*	1	500	no	1
Stephens et al. (2015a)	Angiosperms	eudicots	Sarraceniaceae	*Sarracenia*	2	646	no*	2
Stephens et al. (2015b)	Angiosperms	eudicots	Asteraceae	*Helianthus*	4	598	no*	2
Schmickl et al. (2016)	Angiosperms	eudicots	Oxalidaceae	*Oxalis*	1	1164	no*	3
Heyduk et al. (2016)	Angiosperms	monocots	Arecaceae	*Sabal*	15	837	yes	15
Syring et al. (2016)	Gymnosperms	-	Pinaceae	*Pinus albicaulis*	1	7849	no	1
Sass et al. (2016)	Angiosperms	monocots	Zingiberaceae	-	8	494	yes	8
This Study	**Angiosperms**	**All groups**	**-**	**-**	**25**	**517**	**no***	**43**

1181 **Table 2.** Genomic sources for taxa used in the Angiosperm v.1 probe design. All taxa included in
1182 the design (n=25) are denoted with an asterisk (*); the other taxa were utilized for locus copy
1183 number assessments only (n=18). Data for taxa indicated in bold were obtained from low-
1184 coverage genome sequencing of five vouchered specimens (Index Herbariorum code in
1185 parentheses). All other data were whole genome data downloaded from the sources provided
1186 [accessed June 20, 2013]. Sources of genomic data are indicated by superscripts as follows:
1187 1=www.amborella.org; 2=www.phytozome.net; 3=*Chris Buddenhagen 13041601* (FSU); 4=
1188 *Loran C. Anderson 24918* (FSU); 5=*Loran C. Anderson 23871* (FSU);
1189 6=http://bioinformatics.psb.ugent.be/plaza/; 7=*Rob Naczi 12038* (NY); 8= *Loran C. Anderson*
1190 *26956* (FSU); 9= *Loran C. Anderson 26960* (FSU); 10=*Loran C. Anderson 24932* (FSU);
1191 11=*Loran C. Anderson 23871* (FSU).

Major Group	**Family**	**Species**
Amborellales	Amborellaceae	*Amborella trichopoda*[1]*
eudicot	Ranunculaceae	*Aquilegia coerulea*[2]*
eudicot	Brassicaceae	*Arabidopsis thaliana*[2]*
eudicot	Chenopodiaceae	*Beta vulgaris*[2]*
monocot	Bromeliaceae	***Billbergia nutans*[3]***
monocot	Poaceae	*Brachypodium distachyon*[2]
eudicot	Brassicaceae	*Brassica rapa*[2]
eudicot	Brassicaceae	*Capsella rubella*[2]
monocot	Cyperaceae	*Carex lurida*[4]*
monocot	Cyperaceae	***Carex tenax*[5]**
eudicot	Caricaceae	*Carica papaya*[6]*
eudicot	Rutaceae	*Citrus sinensis*[6]*
eudicot	Cucurbitaceae	*Cucumis sativus*[2]*
eudicot	Myrtaceae	*Eucalyptus grandis*[6]*
eudicot	Rosaceae	*Fragaria vesca*[2]
eudicot	Fabaceae	*Glycine max*[2]
eudicot	Malvaceae	*Gossypium raimondii*[2]
eudicot	Asteraceae	*Lactuca sativa*[2]*
eudicot	Linaceae	*Linum usitatissimum*[2]

eudicot	Rosaceae	*Malus domestica*[2] 1192
eudicot	Euphorbiaceae	*Manihot esculenta*[2]
eudicot	Fabaceae	*Medicago truncatula*[2]*
eudicot	Phrymaceae	*Mimulus guttatus*[2]*
monocot	Musaceae	*Musa acuminata**
eudicot	Nelumbonaceae	*Nelumbo nucifera*[2]*
monocot	Poaceae	*Oryza sativa*[2]*
eudicot	Fabaceae	*Phaseolus vulgaris*[2]
monocot	Arecaceae	*Phoenix dactylifera*[2]*
eudicot	Salicaceae	*Populus trichocarpa*[2]*
eudicot	Rosaceae	*Prunus persica*[2]*
monocot	Cyperaceae	***Rhynchospora chalarocephala***[7]*
eudicot	Euphorbiaceae	*Ricinus communis*[2]*
monocot	Cyperaceae	*Scirpus divaricatus*[8]
monocot	Cyperaceae	*Scleria oligantha*[9]
monocot	Poaceae	*Setaria italica*[2]
eudicot	Solanaceae	*Solanum pimpinellifolium*[2]
eudicot	Solanaceae	*Solanum tuberosum*[2]*
monocot	Poaceae	*Sorghum bicolor*[2]
monocot	Typhaceae	***Sparganium americanum***[10]
eudicot	Brassicaceae	*Thellungiella halophila*[2]
eudicot	Malvaceae	*Theobroma cacao*[2]*
monocot	Typhaceae	***Typha latifolia***[11]*
eudicot	Vitaceae	*Vitis vinifera*[2]*

1193 **Table 3**. Overall success for targeted loci, in terms of regions remaining in the alignment for phylogenetic estimation after orthology
1194 and trimming phases of the bioinformatics pipeline were completed. Targets include the 499 low-copy nuclear loci and the 18
1195 functional mostly high-copy genes associated with selenium tolerance. The total number of orthologs indicates the total number of loci
1196 in the alignment, which may be larger than the number of targets since some targets appear to be multicopy.

	Angiosperms	Brassicales	Caryophyllales	Dipsacales	Magnoliales	Piperales	Poales	Proteales
Anchor Loci								
# Recovered	470	512	464	465	470	330	500	450
% on target	14.1	20.6	10.3	8.1	7.0	1.7	21.9	6.3
Contig length	764	838	742	581	622	631	912	609
# Orthologs	384	540	405	488	483	189	488	159
Avg Copy	1.0	1.1	1.1	1.1	1.0	1.0	1.0	1.7
Max Copy	1.0	4	2	3	2	2	2	4
Functional Loci								
# Recovered	3	15	9	12	10	7	8	11
# Orthologs	0	34	17	26	20	10	13	55
Avg # Copy	n/a	2.3	1.9	2.2	2.0	1.3	1.6	5.0
Max # Copy	n/a	5	4	4	4	2	3	14

1197

1198 **Table 4.** Properties of probe regions and flanks for deep, intermediate, and shallow-scale alignments.

	Probe Region				**Flanks**			
	Total Sites	Var. Sites	Inform. Sites	% Var.	Total Sites	Var. Sites	Inform. Sites	% Var.
Angiosperms	138,616	103,616	94,185	74.8	N/A	N/A	N/A	N/A
Carex	114,623	16,512	3,592	13.9	188,315	43,195	11,307	23.6
Caryophyllales	96,372	23,355	9,171	23.6	66,471	19,490	6,964	29.2
Cyperaceae	120,774	25,082	7,990	19.8	224,529	70,104	20,716	31.6
Dipsacales	80,064	21,501	10,151	26.3	12,429	3,166	1,315	23.4
Magnolia	110,509	6,460	1,181	5.6	87,905	10,450	1,843	12.4
Protea	82,258	11,391	544	13.7	66,269	4,644	536	7.3
Rhynchospora	132,656	29,530	16,615	20.9	164,027	60,759	35,824	36.9
Stanleya	129,500	2,036	149	1.4	178,670	11,933	1,975	6.9

1199