A TARGET ENRICHMENT METHOD FOR GATHERING PHYLOGENETIC INFORMATION FROM HUNDREDS OF LOCI: AN EXAMPLE FROM THE COMPOSITAE

JENNIFER R. MANDEL, REBECCA B. DIKOW, VICKI A. FUNK, RISHI R. MASALIA, S. EVAN STATON, ALEX KOZIK, RICHARD W. MICHELMORE, LOREN H. RIESEBERG, AND JOHN M. BURKE

Department of Biological Sciences, University of Memphis, Memphis, Tennessee 38152 USA; Center for Conservation and Evolutionary Genetics, National Zoological Park and Division of Mammals, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560 USA; Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560 USA; Department of Plant Biology, Miller Plant Sciences, University of Georgia, Athens, Georgia 30602 USA; Department of Genetics, Davison Life Sciences Building, University of Georgia, Athens, Georgia 30602 USA; The Genome Center, University of California, Davis, California 95616 USA; and Department of Botany, University of British Columbia, Vancouver, British Columbia V6T 1Z4 Canada

- Premise of the study: The Compositae (Asteraceae) are a large and diverse family of plants, and the most comprehensive phylogeny to date is a meta-tree based on 10 chloroplast loci that has several major unresolved nodes. We describe the development of an approach that enables the rapid sequencing of large numbers of orthologous nuclear loci to facilitate efficient phylogenomic analyses.

- Methods and Results: We designed a set of sequence capture probes that target conserved orthologous sequences in the Compositae. We also developed a bioinformatic and phylogenetic workflow for processing and analyzing the resulting data. Application of our approach to 15 species from across the Compositae resulted in the production of phylogenetically informative sequence data from 763 loci and the successful reconstruction of known phylogenetic relationships across the family.

- Conclusions: These methods should be of great use to members of the broader Compositae community, and the general approach should also be of use to researchers studying other families.

Key words: base tree; conserved; exon capture; next-generation sequencing; orthologs; phylogenomics.

Ten to twelve percent of all flowering plants (25,000–33,000 species) belong to the Compositae family (Asteraceae). They occur throughout the world, but are most abundant in open areas with seasonal climates such as Mediterranean climates, deserts, prairies and steppes, and mountains. Some family members are widespread and a few are aggressive weeds; most, however, have restricted ranges, and many are in danger of extinction. The Compositae are monophyletic based on morphology as well as molecular genetic data. The most comprehensive phylogeny to date is a meta-tree (Funk and Specht, 2007) that was constructed using a base tree (i.e., the tree used to build the larger phylogeny/meta-tree) of 10 chloroplast loci (Funk et al., 2009; based on Panero and Funk, 2008; Funk and Chan, 2009; Pelser and Watson, 2009; Baldwin, 2009); this meta-tree included ~900 of the 1700 genera found in the family. Several areas of the tree remain poorly resolved (Fig. 1A). These areas are important because the unresolved taxa vary in key morphological traits; thus, well-supported hypotheses of character evolution cannot be developed (e.g., Ortiz et al., 2009).

The combination of next-generation sequencing and large-scale phylogenomics is a promising avenue for efficiently assaying hundreds of loci across multiple taxa to resolve species relationships. One potential approach to identify and sequence loci for phylogenetic analysis is transcriptome sequencing (e.g., RNA-seq; e.g., McKain et al., 2012); however, in many cases where obtaining fresh RNA is difficult or not feasible, a method based on genomic DNA would be preferable. This would also enable the use of museum specimens. Recent work in vertebrates has used DNA sequence capture of mostly noncoding nuclear regions flanking and including so-called “ultraconserved elements” (UCEs; Faircloth et al., 2012). The use of UCEs for phylogenomics is promising, but their detection in plant genomes may be more limited than in vertebrates.

1 Manuscript received 12 November 2013; revision accepted 13 January 2014.

The authors thank Cristin Walters at the University of Georgia Herbarium for assistance in preparing voucher specimens and John Bowers, Suvitha Nambeesan, and Katrien Devos for helpful discussion of the project. We also thank Aaron Liston and another anonymous reviewer for comments on a previous version of this manuscript. The authors thank the Undersecretary for Science, Smithsonian Institution, for the Next Generation Sequencing Small Grant to V.A.F. Funding was also provided by Genome BC, Genome Canada, and the National Science Foundation Plant Genome Research Program (DBI-0820451) to J.M.B., R.W.M., and L.H.R.

9 Author for correspondence: jmandel@memphis.edu
METHODS AND RESULTS

Taxon selection—Fifteen taxa were selected for this study to serve two purposes. First, 10 species were selected to span the entire family and its sister group, the Calyceraceae (species list in Table 1). These taxa allowed us to investigate the utility of our sequence capture probes across the family. Second, we added four additional representatives of the genus *Helianthus* (*H. annuus* L. was one of the original 10) and a taxon from its sister genus, *Phoebanthus*, so that we could also investigate the ability of our COS loci to resolve relationships among more closely related species.

Identification of conserved orthologous sequences—A set of ~1300 conserved genes including approximately 300 single- or low-copy genes for the Compositae was previously identified via BLAST (version 2.2.6) searches of *H. annuus* (sunflower; Asteroidae) and *Lactuca sativa* L. (lettuce; Cichorioideae) ESTs against a set of *Arabidopsis* single-copy genes (the spliced gene models only; see putative intron position determination below) (Kozik et al., unpublished; see http://www.cgpdb.ucdavis.edu/COS_Arabidopsis/ for a description of the pipeline and sequence files). To broaden the representation of Compositae sequences in our analysis, we subsequently used ca. 19,000 *Carthamus tinctorius* L. (safflower; Carduoideae) unigenes derived from ca. 41,000 ESTs (data available at http://www.cgpdb.ucdavis.edu/asteraceae_assembly/) in a BLAST (version 2.2.26) search against the set of ~1300 genes (hereafter simply referred to as the conserved ortholog set loci, or COS loci). The best safflower hits with an E-value ≤E-40 and spanning ≥150 bp were added to the COS alignments using MUSCLE (version 3.8; Edgar, 2004). We were able to generate safflower alignments to 624 out of the ~1300 COS loci. These sequences and the alignments are deposited in the Dryad Digital Repository: http://doi.org/10.5061/dryad.gr93t (Mandel et al., 2014).
<table>
<thead>
<tr>
<th>Tribe</th>
<th>Genus</th>
<th>Species</th>
<th>Authority</th>
<th>Enr.</th>
<th>No. of loci</th>
<th>Location</th>
<th>Collection date</th>
<th>Collector(s)</th>
<th>Collection no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calyceraceae</td>
<td>Nastanthus</td>
<td>patagonicus</td>
<td>Spreg.</td>
<td>17</td>
<td>306</td>
<td>Argentina: Santa Cruz, Rio Chico</td>
<td>14-Dec-2009</td>
<td>Bonifacino &amp; Funk</td>
<td>4016*</td>
</tr>
<tr>
<td>Barnadesieae</td>
<td>Fulcaldea</td>
<td>suessyi</td>
<td>Roque &amp; V. A. Funk</td>
<td>4</td>
<td>174</td>
<td>Brazil: Bahia</td>
<td>09-Aug-2010</td>
<td>Abreu, I. S.</td>
<td>123*</td>
</tr>
<tr>
<td>Mutisieae</td>
<td>Gerbera</td>
<td>hybrid</td>
<td>n/a</td>
<td>19</td>
<td>279</td>
<td>Greenhouse grown cutting, Terra Nigra, USA</td>
<td>22-Oct-2013</td>
<td>Mandel, J. R.</td>
<td>105*</td>
</tr>
<tr>
<td>Cardueae</td>
<td>Carthamus</td>
<td>tinctorius</td>
<td>L.</td>
<td>21</td>
<td>396</td>
<td>Voucher n/a, USDA, PI 592391</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Vernonieae</td>
<td>Centrapalus</td>
<td>paeflorus</td>
<td>(Willd.) H. Rob.</td>
<td>28</td>
<td>408</td>
<td>Greenhouse grown seed, USDA, PI 312852</td>
<td>22-Oct-2013</td>
<td>Mandel, J. R.</td>
<td>104*</td>
</tr>
<tr>
<td>Gnaphalieae</td>
<td>Pseudognaphalium</td>
<td>obtusifolium</td>
<td>(L.) Hilliard &amp; B. L. Burtt</td>
<td>15</td>
<td>208</td>
<td>USA: Fairfax Co., Falls Church, Virginia</td>
<td>12-Sep-2011</td>
<td>Funk, V. A.</td>
<td>12773*</td>
</tr>
<tr>
<td>Eupatorieae</td>
<td>Conoclinium</td>
<td>coelestinum</td>
<td>(L.) DC.</td>
<td>30</td>
<td>416</td>
<td>USA: Fairfax Co., Falls Church, Virginia</td>
<td>11-Sep-2001</td>
<td>Funk, V. A.</td>
<td>12769*</td>
</tr>
<tr>
<td>Heliantheae</td>
<td>Phoebanthus</td>
<td>tenufolius</td>
<td>S. F. Blake</td>
<td>3</td>
<td>372</td>
<td>Greenhouse grown seed collected USA: Liberty Co., Florida</td>
<td>10-Sep-2010</td>
<td>Mason, C. M.</td>
<td>101*</td>
</tr>
<tr>
<td>Heliantheae</td>
<td>Helianthus</td>
<td>porteri</td>
<td>(A. Gray) Pruski</td>
<td>46</td>
<td>325</td>
<td>Greenhouse grown seed collected USA: DeKalb Co., Georgia*</td>
<td>22-Oct-2013</td>
<td>Mandel, J. R.</td>
<td>103*</td>
</tr>
<tr>
<td>Heliantheae</td>
<td>Helianthus</td>
<td>verticillatus</td>
<td>Small</td>
<td>33</td>
<td>379</td>
<td>Greenhouse grown seed collected USA: Madison Co., Tennessee</td>
<td>01-Sep-2004</td>
<td>Mandel, J. R.</td>
<td>101*</td>
</tr>
<tr>
<td>Heliantheae</td>
<td>Helianthus</td>
<td>niveus subsp. tephrodes</td>
<td>(A. Gray) Heiser</td>
<td>62</td>
<td>395</td>
<td>Voucher n/a, USDA, PI 61 3758</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Heliantheae</td>
<td>Helianthus</td>
<td>argophyllus</td>
<td>Torr. &amp; A. Gray</td>
<td>34</td>
<td>339</td>
<td>Voucher n/a, USDA, PI 435623</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Heliantheae</td>
<td>Helianthus</td>
<td>annuus</td>
<td>L.</td>
<td>38</td>
<td>385</td>
<td>Voucher n/a, USDA, PI 60 3989</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Note: Enr. = fold enrichment; n/a = not available; NMNH = National Museum of Natural History; USDA = U.S. Department of Agriculture.

*a* Five hundred ninety-one COS loci were extracted from the lettuce genome. DNA from some taxa studied had been extracted for other projects and the plants were not vouchered; they are listed as n/a.

*b* Number of COS loci identified by PHYLUCE.

* Deposited at the United States National Herbarium, Smithsonian Institution (US), Washington, D.C., USA.

† Deposited at the University of Georgia Herbarium (GA), Athens, Georgia, USA.

* Voucher specimen is an individual from the same population.

^ The *Phoebanthus* WGS sample resulted in far fewer reads than other taxa (~2% of the average for other WGS samples) and could bias the enrichment calculation downward.
Probe design, sequence capture, and sequencing—Custom bioinylated RNA bait/probe libraries were designed using the MYbaits target enrichment system (MYcroarray, Ann Arbor, Michigan, USA) to enrich genomic DNA libraries from 15 species across the family. In total, 15,642 COS-1 bp probes were designed (species list in Table 1). The bioinylated 120-mer baits were tiled across each locus with a 60-base overlap between baits. The putative intron positions of each locus were taken into account during the design process such that probes were positioned so as to span putative splice sites. Putative intron positions for these loci were determined following the methods of Chapman et al. (2007); briefly, the Arabidopsis region from each alignment was used in a BLAST search against the full Arabidopsis genome database (available from http://www.arabidopsis.org/). The BLAST output was mapped onto the global alignments (i.e., the multispecies alignments) via custom scripts (available from Chapman et al., 2007), thus allowing the identification of putative intron positions for these loci. For each COS locus, we designed three probes when possible, i.e., matching the lettuce, sunflower, and safflower sequences. Ultimately, the probe set included 9678 baits targeting 1061 orthologous genes (note that some genes were not able to be covered by baits during the MYcroarray design process due to short putative exon lengths). The sequences for the baits and source ESTs can be found in Appendix S1.

Genomic DNA of each species was extracted using the DNeasy Plant Mini Kit (QIAGEN, Valencia, California, USA), and a barcoded sequencing library was constructed using the TruSeq DNA Sample Preparation kit (Illumina, San Diego, California, USA). Sequence capture was performed following the manufacturer’s protocol (MYcroarray) with two additional steps to improve the hybridization efficiency of the baits to the DNA: (1) the amount of “Hyb #1” used in Step III of the manufacturer’s protocol was reduced from 20 µL to 15 µL per sample, and (2) the magnetic beads were incubated for 30 min at room temperature with 5 µg of denatured salmon sperm DNA prior to transferring the hybridization/capture solution to the beads. We also produced libraries with unique TruSeq barcodes of noncaptured DNA (i.e., an unenriched library) for each species for whole genome shotgun (WGS) sequencing. The enriched samples (hereafter referred to as COS-DNA) and WGS-DNA libraries were quantified using a Qubit 2.0 Fluorometer (Life Technologies, Grand Island, New York, USA), pooled, and sequenced on two lanes of an Illumina HiSeq 2000 sequencer (paired end, 100 base reads). The number of reads for each taxon is located in Appendix S1. Overall, we obtained a high number of reads for each sample (COS-DNA and WGS-DNA). The only exception was for the Phoebe-thus WGS-DNA sample, which had poor sequencing results. Fold-enrichment was calculated for each species by using BLAST similarity (blastn, “e-1e-5”) of the COS-DNA reads and WGS-DNA reads (both quality trimmed; see below for details) to the COS loci and taking the percentage of reads that had a significant BLAST hit out of the total number of reads for each species (COS-DNA vs. WGS-DNA). Fold enrichment is indicated for each species in Table 1. There was no overwhelming trend with respect to the fold enrichment for each species and the phylogenetic distance from the species used in bait design. However, species in the Heliantheae tended to have higher enrichment when compared to the other taxa, and the outgroup and most basal taxon had lower values of enrichment.

Bioinformatic and phylogenetic analyses—All custom scripts, FASTA files, and links to publicly available code have been placed on GitHub (https://github.com/Smithsonian/Compositae-COS-workflow). The workflow is also illustrated in Fig. 2. Note that we also included two optional Perl wrapper scripts on GitHub that can be used to streamline several of the steps performed here; see the README files on the GitHub repository for details. The bioinformatic and phylogenetic workflow was as follows: for each sample, the COS-DNA paired-end reads (the following steps were performed on the two pairs from each species separately: R1, R2) were quality trimmed and reformatted from FASTQ to FASTA using PRINSEQ (version 0.18.2; Schmieder and Edwards, 2011) with the parameters: -min_len 40 –minqap -min_qual_mean 15 –lcr_method entropy –lcr_threshold 60 -trim_nst_right 10 –max_m_p 20 –r remove low-quality and short sequences. While we enriched for the targeted loci by performing the sequence capture, the abundance of remaining off-target reads, combined with the sheer amount of data, made de novo assemblies (see below) computationally demanding. Therefore, we performed a prefiltering step to screen for reads that contained DNA of targeted loci. To do this, the cleaned reads in FASTA format were subjected to BLAST for a similarity search to the EST sequences used for probe design (COS_suff_lett_saff_allfasta), and a custom Perl script (Top_Hits.pl) was used to take only the best read hit from the BLAST output and create a new file with only those read sequences from the original FASTA reads files. The BLAST-filtered Illumina reads have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) as BioProject PRJNA236448. The R1 and R2 reads were paired with PairTim (https://github.com/sestaton/PairTim) and then shuffled (using shuffleSequences fasta.pl) from the Velvet de novo sequence assembler to create replicates (100 search replicates; 1000 bootstrap replicates). The signifi cant fraction of missing data is indicative of the stringency of PHYLUCE. In future work, we are planning to implement strategies that will allow us to explore the effects of different methods of orthology detection. The 763-locus tree, along with bootstrap information, is provided in Fig. 1B. This 763-locus alignment contained a total of 28,324 parsimony informative characters. Comparison with Fig. 1A shows that the relationships found in the 763-locus tree mirror what is expected based on the base tree (Funk et al., 2009; Fig. 1A), and bootstrap support was greater than 75% on almost all nodes. The 186-locus tree (not shown) was congruent with the 763-locus tree, and the 347-locus tree (also not shown) differed
only in the placement of species within Heliantheae, suggesting there may be a tradeoff between taxon representation and total data/number of genes when compared to the 186- and 763-locus trees. The Heliantheae-only tree (415 loci; 167,650 bp) also reconstructed the expected relationships (and the same as the 763-locus tree) based on current phylogenetic information for the taxa studied here with high bootstrap support (Timme et al., 2007; Schilling and Panero, 2011). Bootstrap values for this data set are displayed on Fig. 1B with arrows. As noted previously, restricting the PHYLUCE pipeline to only taxa within the Heliantheae often produced longer alignments for the 415 loci when compared to these same loci in the whole data set alignments (see MAFFT alignments in the Dryad Data Repository [http://doi.org/10.5061/dryad.gr93t; Mandel et al., 2014]). This may, in part, be due to difficulty in aligning across introns, or other highly divergent regions of a locus, when including taxa from across the entire family. These findings suggest that orthology detection is dependent on taxon sampling and the development of new strategies that take into account the hierarchical nature of homology, and how to include as many data as possible in the initial orthology assessment, will be explored in the future. A broader study of the phylogenetic relationships within the Compositae using this method and their relationship with chloroplast loci is underway (Mandel et al., in prep). Data matrices and phylogenetic trees have been deposited in the Dryad Data Repository (http://doi.org/10.5061/dryad.gr93t; Mandel et al., 2014).
CONCLUSIONS

Targeted sequence capture of COS loci facilitated phylogenetic analyses based on a large number of genes across the Compositae. To date, phylogenetic reconstruction in the Compositae has mostly relied on a small number of chloroplast DNA loci, and many relationships among family members remain poorly understood. The motivation for choosing the taxa sequenced here was that the relationships among them are well-established (based on both morphological and molecular data); thus, these known relationships provide a benchmark against which the COS-DNA phylogenies can be compared. We were able to generate usable sequence information for a total of 763 loci and to recover a phylogeny consistent with known relationships within the family with high bootstrap support at most of the nodes within the tree. Moreover, our workflow also proved successful in reconstructing relationships within the Heliantheae tribe, demonstrating that this method is useful for both broader- and finer-scaled phylogenetic reconstruction. We are continuing to add taxa to this base tree using the methods described herein with an ultimate goal of generating a phylogeny that includes at least 20 species selected to represent key nodes within the family. These methods should be of great use to members of the broader Compositae community, and the general approach outlined herein should also be of use to researchers studying other families.

LITERATURE CITED


ZWEICKEL, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, University of Texas at Austin, Austin, Texas, USA.

http://www.bioone.org/loi/apps