Title
A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae.

Permalink
https://escholarship.org/uc/item/11c0g490

Journal
Applications in plant sciences, 2(2)

ISSN
2168-0450

Authors
Mandel, Jennifer R
Dikow, Rebecca B
Funk, Vicki A
et al.

Publication Date
2014-02-06

DOI
10.3732/apps.1300085

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

JENNIFER R. MANDEL2,9, REBECCA B. DIKOW3, VICKI A. FUNK4, RISHI R. MASALIA5, S. EVAN STATON6, ALEX KOZIK7, RICHARD W. MICHELMORE7, LOREN H. RIESEBERG8, AND JOHN M. BURKE5

2Department of Biological Sciences, University of Memphis, Memphis, Tennessee 38152 USA; 3Center for Conservation and Evolutionary Genetics, National Zoological Park and Division of Mammals, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560 USA; 4Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560 USA; 5Department of Plant Biology, Miller Plant Sciences, University of Georgia, Athens, Georgia 30602 USA; 6Department of Genetics, Davison Life Sciences Building, University of Georgia, Athens, Georgia 30602 USA; 7The Genome Center, University of California, Davis, California 95616 USA; and 8Department 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 the 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.

doi:10.3732/apps.1300085
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; Cichorioidae) 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).

METHODS AND RESULTS

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; Cichorioidae) 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).

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; Cichorioidae) 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).
| Tribe            | Genus     | Species           | Authority | Enr. | No. of loci b | Location                                                                 | Collection date | Collector(s) | Collection no. |
|------------------|-----------|-------------------|-----------|------|---------------|---------------------------------------------------------------------------|----------------|--------------|----------------|
| Calyceraceae     | Nastanthus| patagonicus       | Spreng.   | 17   | 306           | Argentina: Santa Cruz, Rio Chico                                          | 14-Dec-2009     | Bonifacino & Funk | 4016*          |
| Barnadesieae     | Fulcaldea | stuessyi          | Roque & V. A. Funk | 4    | 174           | Brazil: Bahia                                                            | 09-Aug-2010     | Abreu, I. S.   | 123*           |
| Mutisieae        | Gerbera   | hybrida           | n/a       | 19   | 279           | Greenhouse grown cutting, Terra Nigra, USA                                | 22-Oct-2013     | Mandel, J. R.  | 105*           |
| Cardueae         | Carthamus | tinctorius        | L.        | 21   | 396           | Voucher n/a, USDA, PI 592391                                             | n/a             | n/a          | n/a            |
| Cichorieae       | Taraxacum | kok-saghyz        | L. E. Rodin | 25   | 407           | Greenhouse grown seed, USDA, W6 35156                                    | 27-Aug-13       | Mandel, J. R.  | 102*           |
| Vernonieae       | Centrapalus| pauciflorus      | (Willd.) H. Rob. | 28   | 408           | Greenhouse grown seed, USDA, PI 312852                                   | 22-Oct-2013     | Mandel, J. R.  | 104*           |
| Senecioneae      | Senecio   | vulgaris          | L.        | 28   | 316           | Washington, D.C.: NMNH                                                   | 07-Nov-2011     | Funk, V. A.   | 12774*         |
| Gnaphalieae      | Pseudognaphalium| obtusifolium | (L.) Hilliard & B. L. Butt   | 15   | 208           | USA: Fairfax Co., Falls Church, Virginia                                 | 12-Sep-2011     | Funk, V. A.   | 12773*         |
| Eupatorieae      | Conoclinium| coelestinum      | (L.) DC.  | 30   | 416           | USA: Fairfax Co., Falls Church, Virginia                                 | 11-Sep-2001     | Funk, V. A.   | 12769*         |
| Heliantheae      | Phoebanthus| tenufolius        | S. F. Blake | 3'   | 372           | Greenhouse grown seed collected, USA: Liberty Co., Florida               | 10-Sep-2010     | Mason, C. M.  | 101*           |
| Heliantheae      | Helianthus | porteri           | (A. Gray) Pruski | 46   | 325           | Greenhouse grown seed collected, USA: DoKab Co., Georgia                | 22-Oct-2013     | Mandel, J. R.  | 103*           |
| Heliantheae      | Helianthus | verticillatus    | Small     | 33   | 379           | Greenhouse grown seed collected, USA: Madison Co., Tennessee            | 01-Sep-2004     | Mandel, J. R.  | 101*           |
| Heliantheae      | Helianthus | niveus subsp. tephrodes | (A. Gray) Heiser | 62   | 395           | Voucher n/a, USDA, PI 613758                                             | n/a             | n/a          | n/a            |
| Heliantheae      | Helianthus | argophyllus      | Torr. & A. Gray | 34   | 339           | Voucher n/a, USDA, PI 435623                                             | n/a             | n/a          | n/a            |
| Heliantheae      | Helianthus | annuus           | L.        | 38   | 385           | Voucher n/a, USDA, PI 603989                                             | n/a             | n/a          | n/a            |

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

* 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 biotinylated 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 C. The reads (Table 1). The biotinylated 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 so that probes were positioned so as not to span putative splice sites. Putative intron positions for these loci were determined following the methods of Chapman et al. (2007); briefly, the Arabidopsis genome 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 putative exon lengths). The sequences for the baits and source ESTs can be located in Appendix S1. To get a first look at how many COS loci were recovered following the enrichment process, we concatenated all loci contigs to the COS loci using the same parameters as before and noted the number of COS loci obtained as a best hit for each species (see Appendix S1). The most basal taxon, F. stuessyi Roque & V. A. Funk, which had the lowest number of BLAST filtered reads, also had the lowest number of Velvet contigs and COS BLAST hits—even lower than the outgroup. At the present time, we do not know if this is related to DNA/library quality or whether this taxon is extremely divergent in the family. Future studies will rely more on taxon sampling from this area of the Compositae and probes designed from additional basal taxa. At F. stuessyi, the number of COS BLAST loci generally followed a trend of recovering more loci the closer the query was in relatedness to the species used for probe design, though overall a good number of loci were recovered for the majority of these species.

Following assembly of the reads, contigs from each of the 15 species were analyzed in the PHYLUCe pipeline (version 0.1.0; Faircloth et al., 2012) specifically using the programs: match_contigs_to_probes.py, get_match_counts.py, get_fastas_from_match_counts.py, and seqcap_align_2.py. Briefly, the program uses LASTZ (version 1.02.00; Harris, 2007) to align the baits/probes to the assembled contigs from each taxon to determine which contigs match the COS loci. PHYLUCe then associates the contigs across species that are putative orthologs, and is quite conservative in that it rejects putative orthologs with more than one match assuming possible paralogy (i.e., ensuring that only one contig matches probes from one COS locus and that only probes from one COS locus match one contig). In addition to the 15 species for which sequence capture was performed, we also included lettuce sequence derived from the publicly available whole genome assembly (version 4; https://genome.usDA.genes). This was done by creating a BLAST database from the lettuce genome scaffolds and performing BLAST searches of each captured COS (from each species) to the lettuce genome. Nucleotide alignments were performed in MAFFT (version 7.029b; Katoh et al., 2002; as implemented in PHYLUCe) for the 763 COS loci with sufficient coverage in a minimum of three species. The number of COS loci analyzed for each species is listed in Table 1. A comparison of the COS loci recovered via BLAST for each species (Appendix S1) and the COS loci retained following PHYLUCe revealed that, in general, the hybridization and sequencing captured a large portion of the 1061 targeted loci for each taxon, and the stage where the majority of loci were lost (i.e., not recovered for phylogenetic analyses out of the 1061) occurred at the orthology assignment step in PHYLUCe. Notably, the one species that is a polyploid, Senecio vulgaris L., did not suffer from fewer loci recovered, despite the more difficult process of recovering potential paralogs in PHYLUCe as we as sess the utility of the probe set within closely related species, and to ascertain whether PHYLUCe would homologize additional data, we also ran the PHYLUCe pipeline using only the taxa from the Heliantheae tribe, and this resulted in 415 COS aligned with MAFFT. These 415 loci had a mean length of 404 bp (range: 200–1101 bp), while the original loci across all taxa had a mean length of 353 bp (range: 27–1545 bp). Alignments have been deposited in the Dryad Data Repository (http://doi.org/10.5061/dryad.gr93t; Mandel et al., 2014). Phylogenetic analyses of concatenated data sets were completed for all 763 COS loci (269,585 bp; 59% missing data), all COS loci for which 10/16 (lettuce included) species were represented (186 loci; 49,918 bp; 37% missing data), and all COS loci for which 8/16 species were represented (347 loci; 96,649 bp; 45% missing data) and all COS loci (269,585 bp; 59% missing data), all COS loci for which 10/16 (lettuce included) species were represented (186 loci; 49,918 bp; 37% missing data), and all COS loci for which 8/16 species were represented (347 loci; 96,649 bp; 45% missing data). The 763-genome alignment contained a total of 28,324 parsimony informative characters. Comparison with Fig. 1A shows that the relationships found in the 763-genome 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-loci tree (not shown) was congruent with the 763-genome tree, and the 347-locus tree (also not shown) differed

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

Fig. 2. COS workflow. Schematic of the laboratory method, bioinformatics, and phylogenetic analyses for this project.
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

BALDWIN, B. 2009. Heliantheae alliance. In V. A. Funk, A. Susanna, T. F. Stuessy, and R. J. Bayer [eds.], Systematics, evolution, and biogeography of Compositae, 689–711. International Association for Plant Taxonomy, Vienna, Austria.

CHAPMAN, M. A., J.-C. CHANG, D. WEISMAN, R. V. KESSELI, AND J. M. HUPALO, AND A. D. KERN. 2013. Conservation and functional element discovery in 20 angiosperm plant genomes. Molecular Biology and Evolution 30: 1729–1744.

KATOH, K., K. MISAWA, K. KUMA, AND T. MIYATA. 2002. MAFFT: A novel method for multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3599–3606.

KRITAS, S. S., E. WUEST, D. HUPALO, A. D. KERN, T. WECKER, AND U. GRENNERKLAUS. 2012. Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. Genome Research 22: 2455–2466.

MANDEL, J. R., R. B. DKROW, V. A. FUNK, R. R. MASALIA, S. E. STATON, A. KOZIK, R. W. MICHELMORE, ET AL. 2014. Data from: A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. Dryad Digital Repository. http://doi.org/10.5061/dryad.gr93t.

MCKAY, M. R., N. WICKETT, Y. ZHANG, S. AYYAMPALAYAM, W. R. McCOMBIE, M. W. CHASE, J. C. PRIES, ET AL. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavaceae (Asparagaceae). American Journal of Botany 99: 397–406.

NEVES, L. G., J. M. DAVIS, W. B. BARBAZUK, AND M. KIRST. 2013. Whole-exome targeted sequencing of the uncharacterized pine genome. Plant Journal 75: 146–156.

ORTZ, S., J. M. BONFACINO, J. V. CRBECI, V. A. FUNK, H. HANSEN, D. J. N. HIND, L. KATINAS, ET AL. 2009. The basal grade of Compositae: Mutisieae (sensu Cabrera). In V. A. Funk, A. Susanna, T. F. Stuessy, and R. J. Bayer [eds.], Systematics, evolution, and biogeography of Compositae, 193–213. International Association for Plant Taxonomy, Vienna, Austria.

PARKS, M., R. CRONN, AND A. Liston. 2012. Separating the wheat from the chaff: Mitigating the effects of noise in a plastome phylogenomic data set from Pinus L. (Pinaceae). BMC Evolutionary Biology 12: 100.

RENEKER, J., E. LYONS, G. C. CONANT, J. C. PRIES, M. FREELING, C.-R. SHYU, AND D. KORKIN. 2012. Long identical multispecies elements in plant and animal genomes. Proceedings of the National Academy of Sciences, USA 109: E1183–E1191.

SALMON, A., J. A. UDALL, J. A. JEDDELOH, AND J. WENDEL. 2012. Targeted capture of homologous coding and noncoding sequence in polyploid cotton. G3: Genes|Genomes|Genetics 2: 921–930.

SCHILLING, E. E., AND J. L. PANERO. 2011. A revised classification of subtribe Helianthinae (Asteraceae: Heliantheae). II. Derived lineages. Botanical Journal of the Linnean Society 167: 311–331.

SCHMIDER, R., AND R. EDWARDS. 2011. Quality control and preprocessing of metagenomic databases. Bioinformatics (Oxford, England) 27: 863–864.

STULL, G. W., M. J. MOORE, V. S. MANDALA, N. A. DOUGLAS, H.-R. KATES, X. QI, S. F. BROCKINGTON, ET AL. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. Applications in Plant Sciences 1: 1200497.

TENNENSEN, J. A., R. GOVINDARAJU, AND A. LISTON. 2013. Targeted sequence capture provides insight into genome structure and genetics of male sterility in a gynodioecious diploid strawberry, Fragaria vesca ssp. bracteata (Rosaceae). G3: Genes|Genomes|Genetics 3: 1341–1351.

TIMME, R. E., B. B. SIMPSON, AND C. R. LINDER. 2007. High-resolution phylogeny for Helianthus (Asteraceae) using the 18S-25S ribosomal DNA external transcribed spacer. American Journal of Botany 94: 1837–1852.

ZEHRING, D. R., AND E. BRNEY. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821–829.

ZHOU, L., AND J. A. HOLLIDAY. 2012. Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. BMC Genomics 13: 703.

ZWICKL, 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.