Short Title for Running Head: Weitemier et al. – Hyb-Seq for plant phylogenomics

Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics[1]

Kevin Weitemier[2,7], Shannon C.K. Straub[2,7], Richard C. Cronn[3], Mark Fishbein[4], Roswitha Schmickl[5], Angela McDonnell[4], and Aaron Liston[2,6]

[2] Department of Botany and Plant Pathology, Oregon State University, 2082 Cordley Hall, Corvallis, Oregon 97331 USA

[3] Pacific Northwest Research Station, USDA Forest Service, 3200 SW Jefferson Way, Corvallis, Oregon 97331, USA

[4] Department of Botany, Oklahoma State University, 301 Physical Sciences, Stillwater, Oklahoma 74078, USA

[5] Institute of Botany, Academy of Sciences of the Czech Republic, CZ-25243 Průhonice, Czech Republic

Email addresses: KW: weitemik@science.oregonstate.edu

SCKS: straubs@science.oregonstate.edu

RCC: rcronn@fs.fed.us

MF: mark.fishbein@okstate.edu

RS: roswitha.schmickl@ibot.cas.cz

AM: arein@okstate.edu

AL: listona@science.oregonstate.edu

Number of words: 2830

[1] Manuscript received _____; revision accepted _____.

[6] Author for correspondence: listona@science.oregonstate.edu

[7] These authors contributed equally to this work.

Acknowledgements: We thank M. Dasenko, Z. Foster, K. Hansen, S. Jogdeo, Z. Kamvar, L. Mealy, M. Parks, M. Peterson, C. Sullivan, and L. Worchester for laboratory, sequencing, data analysis, and computational support. J.M. Rouillard with MYcroarray provided valuable technical support. We thank J. Mandel and another anonymous reviewer for comments on a previous version of this manuscript. This work was funded by the US National Science Foundation (DEB 0919583).

ABSTRACT

- *Premise of the study:* Hyb-Seq, the combination of target enrichment and genome skimming allows simultaneous data collection for low-copy nuclear genes and high-copy genomic targets for plant systematics and evolution studies.
- *Methods and results:* Genome and transcriptome assemblies for milkweed (*Asclepias syriaca*) were utilized to design enrichment probes for 3385 exons from 768 genes (>1.6 Mbp) followed by Illumina sequencing of enriched libraries. Hyb-Seq of twelve individuals (ten *Asclepias* species and two related genera) resulted in at least partial assembly of 92.6% of exons and 99.7% of genes and an average assembly length >2 Mbp. Importantly, complete plastomes and nrDNA cistrons were assembled using off-target reads. Phylogenomic analyses demonstrated signal conflict between genomes.
- *Conclusions:* The Hyb-Seq approach enables targeted sequencing of thousands of low-copy nuclear exons and flanking regions, as well as genome skimming of high-copy repeats and organellar genomes, to efficiently produce genome-scale datasets for phylogenomics.

Key words: genome skimming; Hyb-Seq; nuclear loci; phylogenomics; species tree; target enrichment

INTRODUCTION

The importance of incorporating low-copy nuclear genes in phylogenetic reconstruction is well-recognized, but has largely been constrained by technical limitations (Zimmer and Wen, 2013). These data are essential for reconstructing the evolutionary history of plants, including understanding the causes of observed incongruities among gene tress that arise from incomplete lineage sorting and introgressive hybridization. The combination of solution hybridization for target enrichment of specific genomic regions and the high sequencing throughput of current platforms (e.g., Illumina), provides the opportunity to sequence hundreds or thousands of low-copy nuclear loci appropriate for phylogenetic analyses in an efficient and cost-effective manner (Cronn et al., 2012; Lemmon and Lemmon, 2013). Most efforts to date for targeted sequencing of plant genomes for phylogenetics have been directed at the plastome (e.g., Parks et al., 2012; Stull et al., 2013). Recently, conserved orthologous sequences in Asteraceae (Chapman et al., 2007) were obtained via target enrichment for phylogenomics (Mandel et al., 2014).

Methods have been developed to target highly- or ultra-conserved elements (UCEs) in animal genomes (Faircloth et al., 2012; Lemmon and Lemmon, 2013; McCormack et al., 2013). However, UCEs in plants are nonsyntenic, and are hypothesized to have originated via horizontal transfer from organelles or de novo evolution (Reneker et al., 2012). Whatever their origin, their potential for non-orthology among species makes them unsuitable as phylogenetic markers in plants. The frequency of polyploidy throughout angiosperm evolution (Jiao et al., 2011) also impedes obtaining a large set of conserved orthologous single-copy loci transferable across plant lineages, which in combination with the lack of orthologous UCEs, means that design of targeted sequencing strategies for plant nuclear genomes will necessarily be lineage-specific.

Here we present Hyb-Seq, a protocol that combines target enrichment of low-copy nuclear genes and genome skimming (Straub et al., 2012), the use of low coverage shotgun sequencing to assemble high-copy genomic targets. Our protocol improves upon the methods of Mandel et al. (2014) by 1) utilizing the genome and transcriptome of a single species for probe design, which makes our approach

more generally applicable to any plant lineage, 2) obtaining additional data from the procedure through combination with genome skimming, and 3) developing a data analysis pipeline that maximizes the data usable for phylogenomic analyses. Furthermore, we assemble sequences from the flanking regions (the "splash zone") of targeted exons, yielding non-coding sequence from introns or sequence 5′ or 3′ to genes, which are potentially useful for resolving relationships at low taxonomic levels. We demonstrate the feasibility and utility of Hyb-Seq in a recent, rapid evolutionary radiation: *Asclepias* L. (Apocynaceae). Target enrichment probes were designed using the *Asclepias syriaca* L. draft genome and transcriptome sequences in concert to identify nuclear loci of sufficient length (>960 bp) for robust gene tree reconstruction with a high probability of being single copy (>10% divergence from all other loci in the target genome). We also demonstrate the utility of the *Asclepias* data for phylogenomic analysis and explore the utility of the probes for Hyb-Seq in another genus of the same subtribe, *Calotropis* R. Br. (Asclepiadineae), and a more distantly related genus, *Matelea* Aubl. (Gonolobineae). The Hyb-Seq approach presented here efficiently obtains genome-scale data appropriate for phylogenomic analyses in plants, and highlights the utility of extending genomic tools developed from a single individual for use at deeper phylogenetic levels.

METHODS AND RESULTS

***Targeted enrichment probe design*—** An approach was developed for Hyb-Seq probe design (Table 1) in *Asclepias* utilizing a draft assembly of the *A. syriaca* nuclear genome (Weitemier et al., unpublished data), which was assembled using Illumina paired-end data from libraries with insert sizes of 200 and 450 bp and a k-mer size of 79 in ABySS v. 1.3.2 (Simpson et al., 2009) with reads of plastid or mitochondrial origin removed prior to assembly. A transcriptome of *A. syriaca* leaf and bud tissue (Straub et al., unpublished data) was assembled de novo using Trinity version Trinityrnaseq_r20131110 (Grabherr et al., 2011) and refined using transcripts_to_best_scoring_ORFs.pl (included with Trinity). Probe design was based on data from the draft genome, which was combined with transcriptome assembly data in order to target the exons of hundreds of low copy loci. Contigs from the draft nuclear genome were matched

against those sharing 99% sequence identity from the transcriptome using the program BLAT v. 32x1. BLAT accommodates large gaps in matches between target and query sequences, and is suitable for matching the exon-only sequence of transcripts with the intron-containing genomic sequence, allowing the locations of potential intron/exon boundaries to be identified. Additionally, in an effort to prevent loci present in multiple copies within the genome from being targeted, only those transcripts with a single match against the genome were retained. To prevent probes from enriching multiple similar loci, any targets sharing ≥90% sequence similarity were removed using CD-HIT-EST. The remaining transcriptome contigs were filtered to retain only those containing exons ≥120 bp totaling at least 960 bp. The lower cutoff was necessary to provide sufficiently long sequences for probe design (=120 bp), and the upper cutoff was chosen to exclude short loci less likely to include phylogenetically informative sites. Of the loci that passed filtering, all of those matching (70% sequence identity over 30% of its length) a previously characterized putative ortholog from Apocynaceae (ESTs from *Catharanthus roseus* G. Don; Murata et al., 2006), the asterids (COSII; Wu et al., 2006), or four non-asterid angiosperms (Duarte et al., 2010) were retained (1335 exons in 350 loci). Additional loci that passed filtering were added to the set of targeted loci until the total length of the target probes approached the minimum required for oligo synthesis (2050 exons in 418 loci). The final probe set also contained probes intended to generate data for other projects [157 defense-related and floral development genes and 4000 single nucleotide polymorphisms (SNPs)], which were only included here where necessary for calculations of hybridization efficiency and assembly length. Note that care should be taken during the probe design process to avoid targeting organellar sequences together with nuclear sequences, because enrichment of organellar targets will be proportional to their presence in the genomic DNA extractions used to prepare sequencing libraries and may greatly exceed nuclear targets (See Appendix S1).

Illumina library preparation and Hyb-Seq— DNA was extracted from ten species of *Asclepias, Calotropis procera,* and *Matelea cynanchoides* (Appendix 1) using either a modified CTAB protocol (Doyle and Doyle, 1987), DNeasy (Qiagen, Valencia, California, USA), FASTDNA (MP Bio, Santa Ana,

California, USA), or Wizard kits (Promega, Madison, Wisconsin, USA). Most indexed Illumina libraries were prepared as described by Straub et al. (2012). Two exceptions were *A. cryptoceras* (prepared with a NEXTflex DNA barcode; Bioo Scientific, Austin, Texas, USA) and *M. cynanchoides* (TruSeq library preparation kit; Illumina, San Diego, California, USA). Libraries were then pooled in 11- or 12-plexes with approximately equimolar ratios (some samples included in the pools were not included in the present study). Solution hybridization with MyBaits biotinylated RNA baits (Mycroarray, Ann Arbor, Michigan, USA) and enrichment followed Tennessen et al., (2013) with approximately 350 - 480 ng of input DNA per pool and 12 rather than 15 cycles of polymerase chain reaction (PCR) enrichment to decrease the production of PCR duplicates. These target-enriched libraries were then sequenced on an Illumina MiSeq at either Oregon Health Science University (v. 2 chemistry) to obtain 2 x 251 bp reads or Oregon State University (v. 3 chemistry) to obtain 2 x 76 bp reads. Raw Illumina data were submitted to the NCBI Sequence Read Archive (SRP043058).

Data analysis pipelines— Raw data were filtered for adapter sequences either by the sequencing centers, using Trimmomatic v. 0.20 or 0.30, or using custom scripts. Internal sequence barcodes were deconvoluted using bc_sort_pe.pl (Knaus, 2012). Reads were quality filtered using Trimmomatic to remove bases at read ends with qualities lower than Q20, trim the rest of the read when average quality in a 5 bp window was <Q20, and to remove reads shorter than 36 bp following trimming. For *A. cryptoceras*, only read ends were trimmed to Q20. Duplicate reads were removed using the FASTX-Toolkit (Gordon, 2010). For target assembly, a reference-guided approach utilizing a pseudo-reference consisting of targeted exons separated by 200 Ns was implemented in Alignreads v. 2.25. BLAT was used to identify contigs in the final assembly with sequence similarity to targeted exons. A custom script extracted the longest assembled sequence corresponding to each exon and constructed matrices for multiple sequence alignment, while adding Ns to the matrix if an exon was missing for a particular species. Exons were aligned using default settings in MAFFT v. 6.864b. Following alignment, exons of the same gene were joined using catfasta2phyml.pl. Splash-zone sequences were not included in this

analysis. The same read pools were then used for reference-guided assembly of high-copy sequences, the plastome and nuclear ribosomal DNA (nrDNA) cistrons (18S-5.8S-26S), using Alignreads. The references used for each species were generated through analysis of previously-collected genome skim data of the same libraries used for this study (Straub et al., 2012; Straub et al., unpublished data). References from a different *A. cryptoceras* individual were used for that species and reads were retrimmed using the Trimmomatic setting described above. The *M. biflora* plastome (GenBank: KF539850.1) and the *C. procera* nrDNA sequence served as references for *M. cynanchoides*. MAFFT was used for alignment. Appendix S2 provides additional details on bioinformatic analyses.

***Analyses of assembled sequences*—** The total length of assembled sequence, numbers of targeted exons and genes assembled, amount of flanking sequence assembled from the splash zone, percentage of plastome and nrDNA cistron sequence assembled from the off-target reads based on the lengths of the reference sequences, and percent divergence from the *A. syriaca* exon sequences were calculated for each species. The Hyb-Seq data were also analyzed using the phyluce v. 1.4 pipeline (Faircloth et al., 2012; Faircloth, 2014) utilized by Mandel et al. (2014), using both the native de novo assembly option and using the contigs produced by the reference-guided assembly in Alignreads as input data (See Appendix S3 for detailed methods). To demonstrate the utility of the data for phylogenomics, analyses of *Asclepias* and outgroup *Calotropis* were conducted for nuclear genes individually (excluding seven genes with terminals with all missing data), a concatenation of all nuclear genes, and whole plastomes using RAxML v. 7.3.0 with a GTR + Γ model of nucleotide substitution. One hundred and 1000 rapid bootstrap replicates were conducted for nuclear and plastid analyses, respectively. Prior to analysis, the plastome matrix was edited following Straub et al. (2012). RAxML nuclear gene trees were then used for phylogenomic analyses of all targeted loci with complete taxon sampling (n=761) using the MP-EST species tree approach with bootstrap evaluation of clade support implemented through the STRAW webserver (Shaw et al., 2013). Targeted nuclear exon sequences, data matrices, and trees were submitted

to Figshare (http://dx.doi.org/10.6084/m9.figshare.1024614). See Appendices S1 and S2 for detailed discussion of the protocol from probe design to data analysis.

***Hyb-Seq Results*—** We identified 768 putatively single-copy genes (3385 exons, ca. 1.6 Mbp) meeting the criteria of sufficient length and divergence from all other genes in the genome. Of these genes, 139 genes were among asterid COSII sequences and 47 were among genes conserved across four angiosperm genomes; only 15 out of 155 possible overlapping genes were shared by both conserved sets. Enrichment, sequencing, and assembly of the targeted putatively single-copy genes was successful in *Asclepias* and related Apocynaceae with at least partial assembly of an average of 92.6% of exons and 99.7% of genes and a total average assembly length for all genes in the probe set of ca. 2.2 Mbp from 1.7 Mbp of targeted exons (including the defense and floral development genes; Table 2). Lower read numbers (due to unequal library pooling) resulted in reduced target capture and assembly efficiency in *A. eriocarpa* and *A. involucrata* (Table 2), while a combination of lower read number and sequence divergence (average 4.5%) between *Matelea* and the probes is likely responsible for its somewhat lower success (Fig. 1; Table 2). In contrast, target capture in *Calotropis* (average 3.2% divergence from *A. syriaca*) was similar to *Asclepias* (Table 2). Given that the probes should work well up to 10% sequence divergence, this probe set is likely useful for enrichment of the targeted genes across Asclepiadoideae (Fig. 1). Extending the comparison to the more distantly-related *Catharanthus roseus* (Gongora-Castillo et al., 2012), BLAT analysis reveals an average 12% divergence between *A. syriaca* exons and orthologous transcripts (Fig. 1). This result predicts that a smaller, but not insignificant, amount of sequence data could be obtained from the rest of Apocynaceae. Modification of hybridization conditions could further increase success for more divergent species (Li et al., 2013). In addition to the targeted nuclear loci, reference-guided assembly of the off-target reads yielded complete or nearly-complete plastome and nrDNA cistron sequences (Table 2).

The data analysis pipeline presented here resulted in a data set with few missing genes for each species. In contrast, the phyluce pipeline recovered comparatively few loci for phylogenomic analysis

(Table 3). Phyluce was designed for the analysis of UCE data and its adoption for analysis of single-copy genes where multiple exons have been targeted is inappropriate because exons are often assembled on separate contigs and phyluce views multiple contigs matching a targeted locus as an indication of paralogy (see Appendix S3 for further discussion). The use of reference-guided assembly in the pipeline presented here, rather than the de novo approach of phyluce, also results in a greater amount of data recovery for use in phylogenomic analyses (Table 3).

***Phylogenomics*—** Percentage of variable sites within 768 sequence alignments ranged from 1.8% to 12.5%, with a mean of 5.9% (Appendix S4). The concatenated data matrix was 1,604,805 bp, with 104,717 variable sites, 10,210 of which were parsimony informative. Phylogenomic analysis of the maximum likelihood gene trees for the 761 putatively single-copy genes containing information for all 12 taxa resulted in a species tree topology in which most nodes received high bootstrap support, and which differed from the concatenation species tree in the placement of *A. eriocarpa* (Fig. 2, left). This result highlights the importance of utilizing species tree methods and approaches for assessing clade support that take into account discordance among gene trees, since concatenation approaches can result in strongly supported, but misleading inferences of evolutionary relationships (Kubatko and Degnan, 2007; Salichos and Rokas, 2013). Maximum likelihood analysis of plastomes resulted in a resolved and well-supported phylogeny with a topology in conflict with that of the species tree, especially among temperate North American species (Fig. 2, right). Relationships in this clade estimated from non-coding plastid sequences have been shown to be at odds with expectations based on morphology (Fishbein et al., 2011).

CONCLUSIONS

Hyb-Seq, the combined target enrichment and genome skimming approach presented here, efficiently generates copious data from both the low-copy nuclear genome and high-copy elements (e.g., organellar genomes) appropriate for phylogenomic analyses in plants. With a small investment to generate a genome and transcriptome for an exemplar or the utilization of quickly-growing resources

from the many publicly-available genome and transcriptome projects, a probe set can be designed that will target conserved regions that are phylogenetically informative across plant genera or families. Because this approach recovers sequences that are hundreds of base pairs in length from hundreds to thousands of loci, even with modest levels of variation the data are appropriate for addressing questions at low taxonomic levels. Furthermore, sequences flanking the conserved target regions will generally evolve more rapidly, providing additional potentially informative variation.

The Hyb-Seq protocol based on taxon-specific genome and transcriptome data has advantages over alternative approaches, such as transcriptome sequencing or genome reduction via restriction digest. Transcriptome sequencing results in thousands of orthologous nuclear loci, but requires living, flash frozen, or specially preserved tissue for RNA extraction, is subject to large amounts of missing loci across samples, and does not as effectively sample rapidly-evolving non-coding regions. In contrast, target capture and genome skimming can utilize small amounts of relatively degraded DNA, such as extractions from herbarium specimens (Cronn et al., 2012; Straub et al., 2012), and consistently yield intron and 5′ and 3′ untranslated region sequence. Genome reduction methods utilizing restriction digests (e.g., RAD-Seq, genotyping-by-sequencing; Davey et al., 2011) also produce thousands of loci, and have been effective in resolving phylogenetic relationships and patterns of introgression (e.g., Eaton and Ree, 2013). However, the effectiveness of these approaches with poor quality or degraded DNA has not been demonstrated and the anonymous nature of these loci makes it more challenging to determine orthology. Most importantly, the data obtained (SNPs or 30-200 bp sequences) are not appropriate for applying phylogenetic approaches that estimate species trees from a large number of gene trees. Focusing on orthologous targets through Hyb-Seq also reduces the amount of missing data relative to both transcriptome and RAD-Seq studies. Until the sequencing of whole genomes for every species of interest becomes practical and affordable, the protocol presented here is poised to become the standard for quickly and efficiently producing genome-scale datasets to best advance our understanding of the evolutionary history of plants.

LITERATURE CITED

BOLGER, A. M., M. LOHSE, and B. USADEL. In press. Trimmomatic: a flexible trimmer for Illumina Sequence Data. *Bioinformatics*.

CHAPMAN, M. A., J. CHANG, D. WEISMAN, R. V. KESSELI, and J. M. BURKE. 2007. Universal markers for comparative mapping and phylogenetic analysis in the Asteraceae (Compositae). *Theoretical and Applied Genetics* 115: 747-755.

CRONN, R., B. J. KNAUS, A. LISTON, P. J. MAUGHAN, M. PARKS, J. V. SYRING, and J. UDALL. 2012. Targeted Enrichment Strategies for Next-Generation Plant Biology. *American Journal of Botany* 99: 291-311.

DAVEY, J. W., P. A. HOHENLOHE, P. D. ETTER, J. Q. BOONE, J. M. CATCHEN, and M. L. BLAXTER. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12: 499-510.

DOYLE, J.J., and J.L. DOYLE. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19:11-15.

DUARTE, J. M., P. K. WALL, P. P. EDGER, L. L. LANDHERR, H. MA, J. C. PIRES, J. LEEBENS-MACK, et al. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* 10: 61.

EATON, D. A. R., and R. H. REE. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* 62: 689-706.

FAIRCLOTH, B. C. 2014. phyluce: phylogenetic estimation from ultraconserved elements. doi: 10.6079/J9PHYL.

FAIRCLOTH, B. C. , J. E. MCCORMACK, N. G. CRAWFORD, M.G. HARVEY, R.T. BRUMFIELD, T.C. GLENN. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61: 717-726.

FISHBEIN, M., D. CHUBA, C. ELLISON, R. J. MASON-GAMER, and S. P. LYNCH. 2011. Phylogenetic relationships of *Asclepias* (Apocynaceae) inferred from non-coding chloroplast DNA sequences. *Systematic Botany* 36: 1008-1023.

GONGORA-CASTILLO, E., K. L. CHILDS, G. FEDEWA, J. P. HAMILTON, D. K. LISCOMBE, M. MAGALLANES-LUNDBACK, K. K. MANDADI, et al. 2012. Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. *PLoS One* 7: e52506.

GORDON, A. 2010. FASTX-Toolkit.Website http://hannonlab.cshl.edu/fastx_toolkit/ [Accessed 15 May 2014].

GRABHERR, M. G., B. J. HAAS, M. YASSOUR, J. Z. LEVIN, D. A. THOMPSON, I. AMIT, X. ADICONIS, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644-652.

JIAO, Y., N. J. WICKETT, S. AYYAMPALAYAM, A. S. CHANDERBALI, L. LANDHERR, P. E. RALPH, L. P. TOMSHO, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97-100.

KATOH, K., and H. TOH. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* 9: 286-298.

KENT, W. J. 2002. BLAT—the BLAST-Like Alignment Tool. *Genome Research* 12: 656-664.

KNAUS, B. 2012. Short read toolbox. Website http://brianknaus.com/software/srtoolbox/ [Accessed 15 May 2014].

KUBATKO, L., and J. DEGNAN. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56: 17-24.

LEMMON, E. M., and A. R. LEMMON. 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 44: 99-121.

LI, C., M. HOFREITER, N. STRAUBE, S. CORRIGAN, and G. J. NAYLOR. 2013. Capturing protein-coding genes across highly divergent species. *Biotechniques* 54: 321-326.

LI, W., and A. GODZIK. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658-1659.

LIU, L., L. YU, and S. V. EDWARDS. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* 10: 302.

MANDEL, J. R., R. B. DIKOW, V. A. FUNK, R. R. MASALIA, S. E. STATON, A. KOZIK, R. W. MICHELMORE, et al. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences* 2: 1300085.

MCCORMACK, J. E., M. G. HARVEY, B. C. FAIRCLOTH, N. G. CRAWFORD, T. C. GLENN, and R. T. BRUMFIELD. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS One* 8: e54848.

MURATA, J., D. BIENZLE, J. E. BRANDLE, C. W. SENSEN, and V. DE LUCA. 2006. Expressed sequence tags from Madagascar periwinkle (*Catharanthus roseus*). *FEBS Letters* 580: 4501-4507.

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.

NYLANDER, J.A.A. 2011. Catfasta2pyml.pl. Website http://www.abc.se/~nylander/catfasta2phyml/ [Accessed 15 May 2014].

RATAN, A. 2009. Assembly algorithms for next-generation sequence data. Ph.D. Dissertation, The Pennsylvania State University, University Park, Pennsylvania, USA.

RENEKER, J., E. LYONS, G. C. CONANT, J. C. PIRES, 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.

SALICHOS, L., and A. ROKAS. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497: 327-331.

SHAW, T. I., Z. RUAN, T. C. GLENN, and L. LIU. 2013. STRAW: Species TRee Analysis Web server. *Nucleic Acids Research* 41: W238-W241.

SIMPSON, J. T., K. WONG, S. D. JACKMAN, J. E. SCHEIN, S. J. M. JONES, and İ. BIROL. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Research* 19: 1117-1123.

STAMATAKIS, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688-2690.

STRAUB, S. C. K., M. PARKS, K. WEITEMIER, M. FISHBEIN, R. C. CRONN, and A. LISTON. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349-364.

STRAUB, S. C. K., M. FISHBEIN, T. LIVSHULTZ, Z. FOSTER, M. PARKS, K. WEITEMIER, R. C. CRONN, and A. LISTON. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.

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.

TENNESEN, J. A., R. GOVINDARAJULU, A. LISTON, and T.-L. ASHMAN. 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.

WU, F., L. A. MUELLER, D. CROUZILLAT, V. PETIARD, and S. D. TANKSLEY. 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics* 174: 1407-1420.

ZIMMER, E. A., and J. WEN. 2013. Using nuclear gene data for plant phylogenetics: progress and prospects. *Molecular Phylogenetics and Evolution* 66: 539-550.

TABLES

Table 1. Hyb-Seq target enrichment probe design and bioinformatics pipeline. A script combining and detailing the steps of the probe design process, Building_exon_probes.sh, is provided in the supplementary materials.

Steps	Comments	Primary program or custom script
Probe Design		
Match	Find genome and transcriptome sequences with 99% identity	BLAT[a]
Filter	Retain single hits of substantial length	Part of Building_exon_probes.sh[b]
Cluster	Remove isoforms and loci sharing >90% identity	CD-HIT-EST[c], grab_singleton_clusters.py[b]
Filter	Retain loci with long exons summing to desired length	blat_block_analyzer.py[b]
Cluster	Remove exons sharing >90% identity	CD-HIT-EST[c], grab_singleton_clusters.py[b]
Bioinformatics		
Read processing	Adapter trimming, quality filtering	Trimmomatic[d]
Exon assembly	Reconstruct a sequence for each sample, for each exon	YASRA[e], Alignreads[f]

Identify assembled contigs	If contig identity is unknown, identify which targeting exon(s) it corresponds to	BLAT[a]
Sequence alignment I: Collate exons	Cluster orthologous exons across samples	assembled_exons_to_fasta.py[b]
Sequence alignment II: Perform alignment	Align homologous bases within each exon	Mafft[g]
Concatenate Exons	For each locus, concatenate the aligned exons	catfasta2phyml.pl[h]
Gene tree construction	For each locus, estimate the maximum likelihood gene tree	RAxML[i]
Species tree construction	Estimate the species tree from independent gene trees in a coalescent framework	MP-EST[j]

[a]Kent (2002)
[b] New scripts written for this protocol, an example data set, and any future updates are available at https://github.com/listonlab/.
[c]Li and Godzik (2006)
[d]Bolger, Lohse, and Usadel (In press)
[e]Ratan (2009)
[f]Straub et al. (2011)
[g]Katoh and Toh (2008)
[h]Nylander (2011)

[i]Stamatakis (2006)
[j]Liu et al. (2010)

Table 2. Success of Hyb-Seq for targeted sequencing and assembly of nuclear genes combined with genome skimming of high-copy targets in *Asclepias* and related species of Apocynaceae.

Species	Reads[a]	Quality-filtered reads	Unique, on-target, quality filtered reads (%)[b,c]	Assembly length (Mbp)[b]	Splash zone assembly length (Mbp)[b]	Single-copy gene exons assembled[d]	Single-copy genes assembled[d]	% Divergence from single-copy gene probes[e]	% Missing Data in Matrix	% Completion of plastome	% Completion of nrDNA cistron
Asclepias cryptoceras	1174294	1149278	746909 (65.0%)	3.2	1.6	3349	768	0.9	7.4	99.7	100
Asclepias engelmanniana	1943370	1804956	523477 (29.0%)	2.7	1.0	3359	767	0.8	3.6	97.8	98.3
Asclepias eriocarpa	393048	384595	72200 (18.8%)	1.1	0.5	2260	762	0.9	69.0	81.9	94.4
Asclepias flava	1457860	1301608	397798 (30.6%)	2.2	0.8	3313	768	1.5	14.7	98.4	100
Asclepias humistrata	918608	843463	234502 (27.8%)	2.0	0.8	3163	768	1.0	27.1	93.1	97.0
Asclepias involucrata	664820	645580	139407 (21.6%)	1.7	0.7	2978	768	0.9	41.8	90.5	99.4
Asclepias masonii	1097532	971606	270123 (27.8%)	2.1	0.9	3275	768	1.4	30.6	99.1	100
Asclepias nyctaginifolia	2482686	2295691	558822 (24.3%)	2.4	0.8	3369	768	0.9	2.1	96.0	100
Asclepias scheryi	1345732	1295739	384451 (29.7%)	2.4	0.8	3314	768	1.0	4.9	98.7	100
Asclepias tomentosa	1248940	1111909	310020 (27.9%)	2.1	0.8	3208	768	0.9	26.7	95.2	99.7
Calotropis procera	1172456	1135014	380155 (33.5%)	2.6	1.0	3287	768	3.2	5.0	96.0	100
Matelea cynanchoides	418590	388064	208835 (53.8%)	1.7	0.4	2718	757	4.5	n/a	99.4	100
Average	**1190419**	**1110625**	**352225 (32.5%)**	**2.2**	**0.8**	**3133**	**767**	**1.5**	**21.2**	**95.5**	**99.1**

[a] Most samples were sequenced in a single MiSeq run (11-plex 2 x 251 bp v. 2 chemistry) except for *A. cryptoceras* and *M. cynanchoides*, which were each sequenced in different MiSeq runs (12-plex 2 x 251 bp v. 2 chemistry and 15-plex 2 x 76 bp v. 3 chemistry, respectively).

[b] These values were calculated using the entire probe set, including single-copy gene, defense and floral development genes, and SNPs.

[c] These estimates are lower than the true overall efficiency due to quality filtering and the removal of duplicate reads. Except for *A. cryptoceras* and *M. cynanchoides*, the libraries were made with internal barcodes, which apparently contributed to suboptimal base calling and lower quality scores, leading to apparent suboptimal target capture efficiency.

[d] These estimates are based on a minimum 90% sequence identity to the *A. syriaca* probes, and are therefore conservative; especially so for *C. procera* and *M. cynanchoides*, which are expected to have higher sequence divergence.

[e] These estimates are based on a minimum 75% sequence identity to the *A. syriaca* probes.

Table 3. Number of single-copy genes recovered for phylogenomic analysis with different data analysis pipelines

Species	Hyb-Seq	phyluce	phyluce with Alignreads contigs
Asclepias cryptoceras	768	16	145
Asclepias engelmanniana	767	69	201
Asclepias eriocarpa	762	10	23
Asclepias flava	768	28	109
Asclepias humistrata	768	27	62
Asclepias involucrata	768	3	24
Asclepias masonii	768	8	38
Asclepias nyctaginifolia	768	13	198
Asclepias scheryi	768	69	186
Asclepias tomentosa	768	21	54
Calotropis procera	768	84	203
Matelea cynanchoides	757	51	98
Average	**767**	**33**	**112**

APPENDICES

Appendix 1. Voucher information for species of *Asclepias* and related genera used in this study.

Species	Voucher Specimen	Collection locality	GPS Coordinates[a]
Asclepias cryptoceras S. Watson	Weitemier 12-23 [OSC]	Grant Co., Oregon, USA	44.47970, -119.57758
A. engelmanniana Woodson	Lynch 11224 [LSUS]	Barber Co., Kansas, USA	37.3, -98.7
A. eriocarpa Torr.	Lynch 10923 [LSUS]	Lassen Co., California, USA	41.09, -121.30
A. flava (Kuntze) Lillo non N.E. Br.	Zuloaga & Morrone 7069 [OKLA]	Dist. Jujuy, Argentina	-24, -63.35
A. humistrata Walter	Fishbein 5596 [OKLA]	Polk Co., Florida, USA	27.761, -81.465
A. involucrata Englem. ex Torr.	Lynch 12050 [LSUS]	Apache Co., Arizona, USA	36.7 -109.7
A. masonii Woodson	Fishbein 3101 [OKLA]	Mpio. Comondu, Baja California Sur, Mexico	24.63, -112.14
A. nyctaginifolia A. Gray	Fishbein 2445 [ARIZ]	Pima Co., Arizona, USA	31.80, -110.81
A. scheryi Woodson	Fishbein 5137 [OKLA]	Mpio. Cuautitlán, Jalisco, Mexico	19.561, -114.203
A. tomentosa Elliott	Fishbein 5608 [MISSA]	Franklin Co., Florida, USA	29.916, -84.369
Calotropis procera (Aiton) W.T. Aiton	Fishbein 5427 [OKLA]	Cultivated	
Matelea cynanchoides (Engelm. & A. Gray) Woodson	Rein 106 [OKLA]	Angelina Co., Texas, USA	31.07995, -94.27735

[a] GPS coordinates reported to the accuracy recorded or based on coarse geo-referencing based on the collection locality.

FIGURE LEGENDS

Fig. 1. Histogram of exon sequence divergence between the species used for probe design, *Asclepias syriaca*, and four other species: the most divergent species of *Asclepias*, *A. flava*; another member of Asclepiadinae (Asclepiadeae:Asclepiadoideae), *Calotropis procera*; a member of Gonolobinae (Asclepiadeae:Asclepiadoideae), *Matelea cynanchoides*; and a member of a different subfamily *Catharanthus roseus* (Rauvolfioideae). Note that a maximum sequence divergence of 75% was allowed for BLAT and that exons with >10% divergence were less likely to be observed in *Calotropis* and *Matelea* because they were less likely to be enriched by the probes, while the *Catharanthus* data were from transcriptome sequences of multiple tissues and not subject to target enrichment bias.

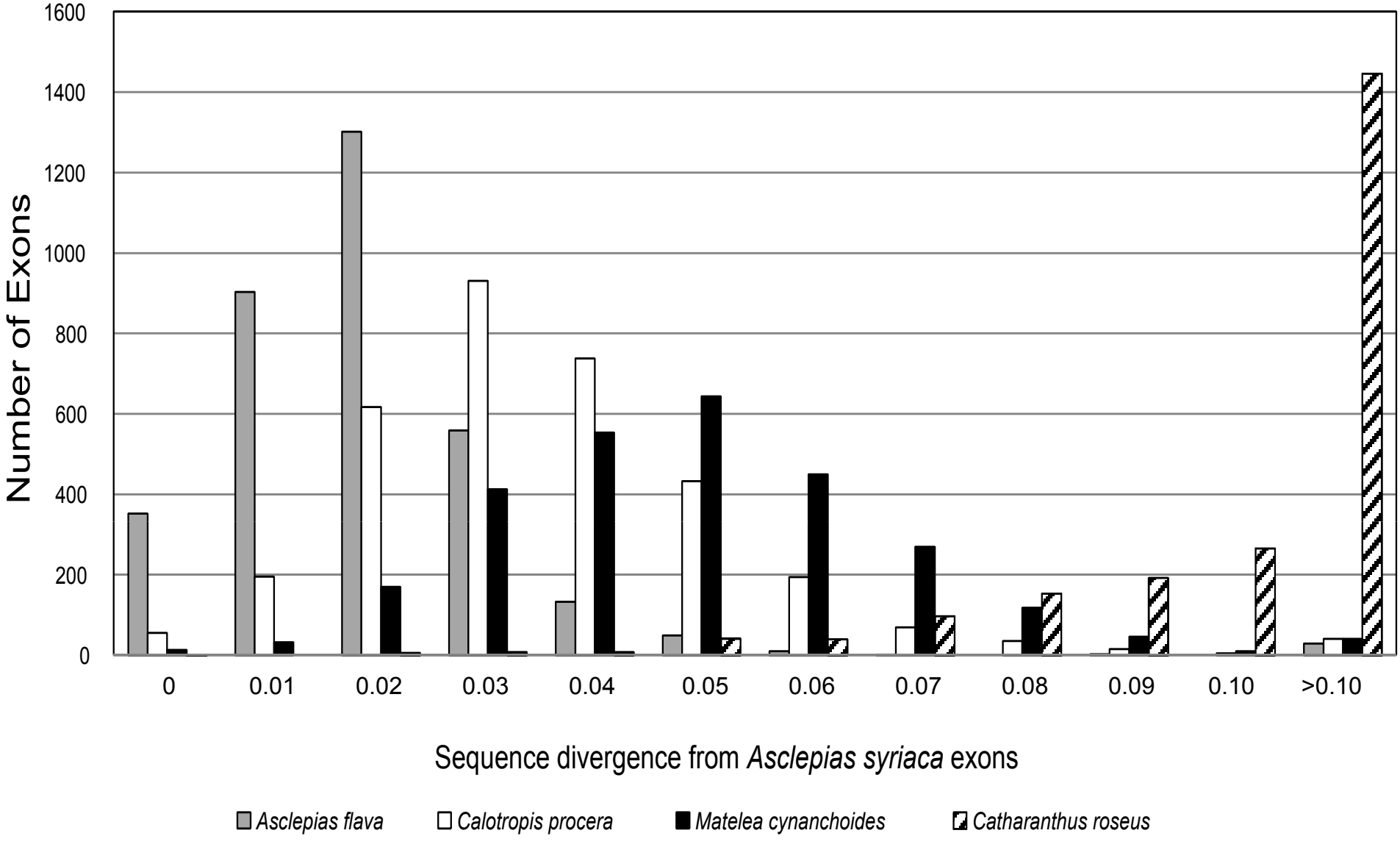
Fig. 2. Comparison of the species tree of *Asclepias* based on 761 putatively single-copy loci and the whole plastome phylogeny. The MP-EST tree is shown at left and the difference between this topology and that recovered through an analysis of the concatenated nuclear gene data set is indicated by the red arrow. Solid lines connect each species to its placement in the plastome phylogeny (right). Values near the branches are bootstrap support values (*=100%). Colors reflect the plastid clades of Fishbein et al. (2011): temperate North America (green), unplaced (orange), highland Mexico (purple), series Incarnatae sensu Fishbein (pink), Sonoran Desert (blue), and outgroup (black).

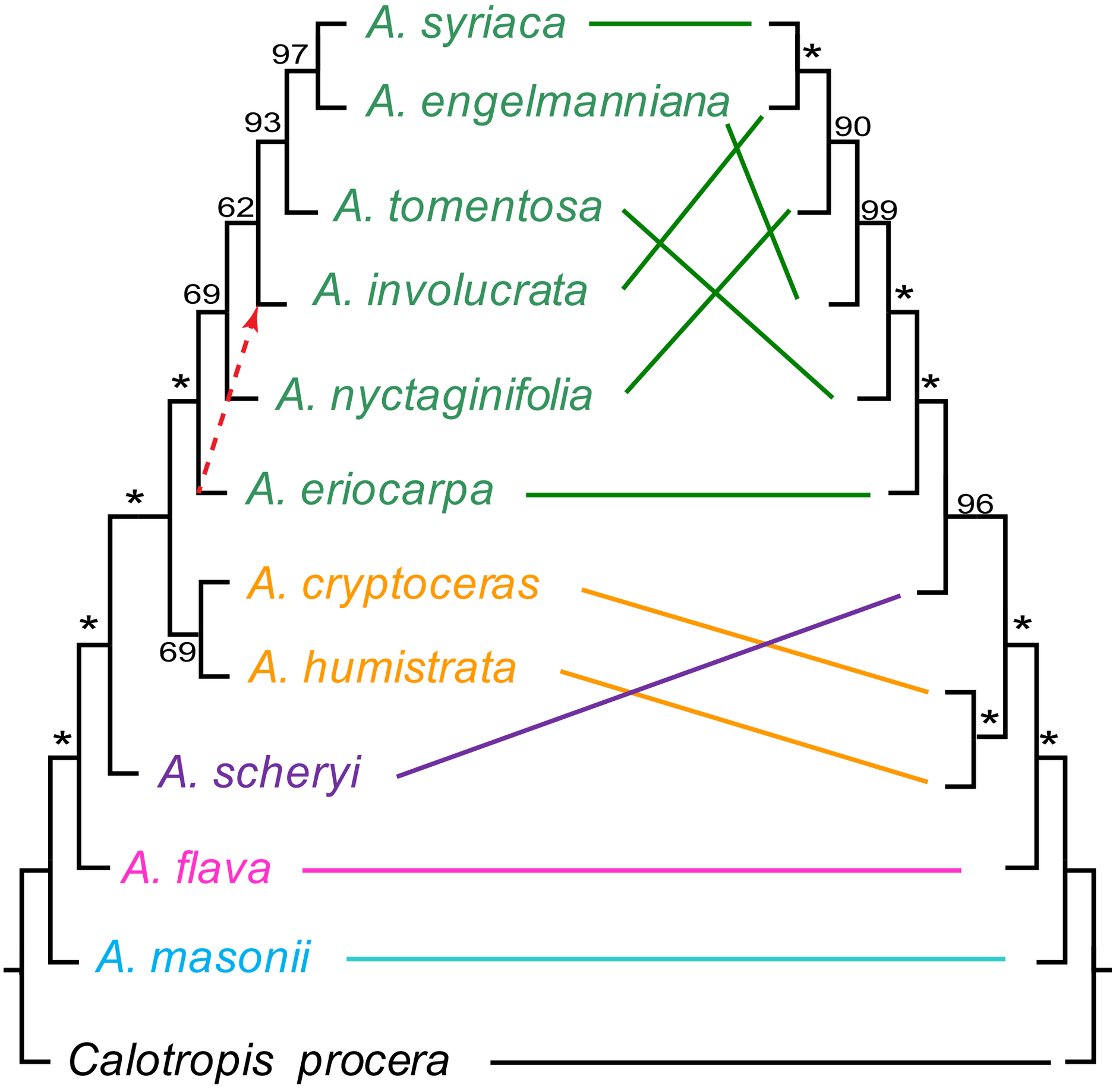
Appendix S1. Detailed target enrichment probe design protocol.

Appendix S2. Detailed bioinformatics pipeline protocol.

Appendix S3. Comparison of post-sequencing analyses.

Appendix S4. Percentage of variable sites within sequence alignments of 12 taxa at 768 loci.


1600
1400
1200
1000
800
600
400
200
0
Number of Exons
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
>0.10
Sequence divergence from *Asclepias syriaca* exons
Asclepias flava
Calotropis procera
Matelea cynanchoides
Catharanthus roseus


A. syriaca
A. engelmanniana
A. tomentosa
A. involucrata
A. nyctaginifolia
A. eriocarpa
A. cryptoceras
A. humistrata
A. scheryi
A. flava
A. masonii
Calotropis procera
97
93
62
69
69
90
99
96
*

```
#/bin/tcsh

#                       Appendix S1
#                   Building_exon_probes.sh
#       Workflow and script for development of Hyb-Seq target probes


################################################################################
#
#                 *** PLEASE CONTINUE READING ***
#
#THIS FILE IS BOTH A DETAILED DESCRIPTION OF THE HYB-SEQ PROBE DESIGN PROCESS,
#AND A PROGRAM THAT CAN BE RUN IN A LINUX ENVIRONMENT.
#
#COMMENTS ARE CONTAINED ON LINES BEGINNING WITH A # SYMBOL.
#
################################################################################


#Disclaimer:
#Although these commands should function with just the input of two fasta
#files, genome.fasta and transcriptome.fasta, their proper execution is not
#guaranteed. Given the idiosyncrasies in file formats and operating
#environments, these commands are meant more as a starting point, to be
#modified as needed by the user. For example, the presence of spaces or tabs in
#the ID line of the fasta files may cause problems downstream.

#Copyright (c) 2014
#K. Weitemier, S.C.K. Straub, R. Cronn, M. Fishbein, R. Schmickl, A. McDonnell,
#and A. Liston.
#This script is free software: you can redistribute it and/or modify it under
#the terms of the GNU General Public License as published by the Free Software
#Foundation, either version 3 of the License, or (at your option) any later
#version. A copy of this license is available at <http://www.gnu.org/licenses/>.
#Great effort has been taken to make this software perform its said
#task, however, this software comes with ABSOLUTELY NO WARRANTY,
#not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

#TO RUN THIS SCRIPT (in a Linux environment):
#Copy this script into a new directory along with:
#1) A fasta file of genomic contigs, named "genome.fasta"
#2) A fasta file of transcriptome contigs, named "transcriptome.fasta"
#3) the file provided with this script named "grab_singleton_clusters.py" and
#4) the file provided with this script named "blat_block_analyzer.py"
#You may need to make this script and the other two programs executable, or able
#to be recognized as programs. To do this, run the following command:
#
#  chmod +x Building_exon_probes.sh grab_singleton_clusters.py blat_block_analyzer.py
#
#Finally, to run the script, type the following command:
#
#  ./Building_exon_probes.sh
#
```

#Additional notes: This script calls two third-party programs, BLAT and
#CD-HIT-EST. These need to be installed on your system so that they can be
#called simply from the commands "blat" and "cd-hit-est", respectively. It is
#also necessary to have Python 2.x installed to run grab_singleton_clusters.py
#and blat_block_analyzer.py.

###
#Match genome and transcriptome sequences
###
#Here we use BLAT to identify transcripts that have high identity to genomic
#contigs. An important consideration when preparing the input genome and
#transcriptome assemblies is to first remove those contigs that match the
#chloroplast or mitochondrial genomes, so that they are not targeted in the same
#pool as nuclear loci. The presence of multiple copies of these genomes in
#prepared libraries, relative to the nuclear genome, will skew the ratio of
#captured fragments by an equal proportion (e.g., if 100 copies of a targeted
#chloroplast locus are present relative to a targeted nuclear locus, the pool of
#enriched fragments, and sequenced reads, will over-represent the chloroplast
#locus by about 100:1). High copy loci such as the chloroplast will represent a
#large fraction of off-target sequenced reads and can likely be assembled
#without enrichment. If enrichment of organellar or other high copy regions is
#desirable (as might be required for very high multiplexing), target enrichment
#should be done in a separate reaction using a separate set of probes. After
#enrichment, the individual pools representing different targeted fractions
#(e.g., nuclear and chloroplast targets) can be re-mixed at desired ratios for
#multiplex sequencing.

#As a default in this step the matching identity is set very high (99%) because
#the transcripts and genomic contigs are expected to come from the same
#individual. If this is not the case, particularly if the transcriptome and
#genome originate from different taxa, you should reduce the stringency by
#modifying the "-minIdentity" flag.

#The fasta files should be formatted so that each sequence takes up exactly two
#lines: the ID line and the sequence line.

#This step may take a very long time.

```
echo """Comparing genome and transcriptome. This may take a very long time."""

date

blat genome.fasta transcriptome.fasta -tileSize=8 -minIdentity=99 -noHead -out=pslx
genome_v_transcriptome.pslx
```

###
#Find and extract transcriptome sequences with only one match against the genome
###
#This step removes transcripts that have matches against more than one genomic
#contig, based on the assumption that multiple hits may indicate loci with
#several copies in the genome. However, this step may exclude loci that are
#truly single copy in cases where a locus is split across more than one genomic
#contig in a fragmented genome assembly.

```
echo """Finding and extracting matches with a single hit."""

date

cut -f10 genome_v_transcriptome.pslx | sort | uniq -c | grep '      1 ' | sed -e
's/      1 /\\\</' -e 's/$/\\\>/' > single_hits_vs_genome.txt

grep -f single_hits_vs_genome.txt genome_v_transcriptome.pslx >
single_hit_genome_v_transcriptome.pslx



################################################################################
#Keep only matches hitting at least 95% of the transcript.
################################################################################


echo """Finding and extracting transcripts with single, substantial hits."""

date

awk '{if (($13-$12) >= ($11 * 0.95)) print $0}'
single_hit_genome_v_transcriptome.pslx >
whole_gene_single_hit_genome_v_transcriptome.pslx


#Extract these transcripts from the transcriptome assembly

cut -f10 whole_gene_single_hit_genome_v_transcriptome.pslx >
single_whole_hits_vs_genome.txt

sed -i'' -e 's/^/^>/' -e 's/$/\\\>/' single_whole_hits_vs_genome.txt

grep -A1 -f single_whole_hits_vs_genome.txt transcriptome.fasta | sed '/^--$/d' >
single_transcript_hits.fasta


################################################################################
#Cluster any nested transcripts
################################################################################
#In some cases, one transcript may exactly match, but be nested within, another
#transcript. This program finds such transcripts and retains only the largest
#sequence.

echo """Clustering transcripts with 100% identity."""

date

cd-hit-est -i single_transcript_hits.fasta -o
single_transcript_hits_cluster_100.fasta -d 0 -c 1.0 -p 1 >
cluster_100_single_transcript_hits_log.txt


################################################################################
#Remove transcripts with 90% or greater similarity
################################################################################
#This step removes transcripts that share high identity. This should help reduce
#the number of targeted loci with multiple copies within the genome, and reduce
#the chance of capturing similar, but off-target, loci.

echo """Clustering and removing transcripts with 90% identity."""
```

```
date

cd-hit-est -i single_transcript_hits_cluster_100.fasta -o
single_transcript_hits_cluster_90.fasta -d 0 -c 0.9 -p 1 -g 1 >
cluster_90_single_transcript_hits_log.txt

python grab_singleton_clusters.py -i single_transcript_hits_cluster_90.fasta.clstr
-o unique_single_transcript_hits_cluster_90.fasta.clstr

grep -v '>Cluster' unique_single_transcript_hits_cluster_90.fasta.clstr | cut -d' '
-f2 | sed -e 's/\.\.\./\\\>/' -e 's/^/^/' > unique_single_transcript_hits

grep -A1 -f unique_single_transcript_hits single_transcript_hits_cluster_100.fasta
| sed '/^--$/d' > unique_single_transcript_hits.fasta

sed -i'' -e 's/\^>/\\\</' unique_single_transcript_hits

grep -f unique_single_transcript_hits
whole_gene_single_hit_genome_v_transcriptome.pslx > unique_single_hits.pslx


################################################################################
#Find loci and exons that meet length requirements
################################################################################
#This step finds every remaining locus and all the remaining exons that make up
#that locus. If an exon is above a specified length, the program adds it to the
#total length for that locus. It outputs the sequences of all the large exons
#for each locus that exceeds a specified minimum length. The default minimum
#length for exons is 120 bp, which matches the length of the probe oligos used
#in the target capture reaction. It can be modified by altering the "-s" flag.
#The default minimum length for loci is 960 bp. This was chosen because it is a
#multiple of 120 and was considered a large enough length for reliable gene tree
#estimation. It can be modified by altering the "-l" flag.

echo """Finding loci and exons that meet length requirements."""

date

python blat_block_analyzer.py -i unique_single_hits.pslx -o
large_enough_unique_single_hits.fasta -l 960 -s 120


################################################################################
#Remove exons with 90% or greater similarity
################################################################################
#Finally, any remaining individual exons that share >=90% identity are removed.
#This is to prevent any problems caused by the cross-enrichment of similar
#exons across different loci.

echo """Removing individual exons with high identity."

date

cd-hit-est -i large_enough_unique_single_hits.fasta -o
large_enough_unique_single_hits_cluster90.fasta -d 0 -c 0.9 -p 1 -g 1 >
cluster_90_large_enough_unique_single_hits_log.txt

python grab_singleton_clusters.py -i
large_enough_unique_single_hits_cluster90.fasta.clstr -o
unique_blocks_large_single_hits_cluster90.fasta.clstr

grep -v '>Cluster' unique_blocks_large_single_hits_cluster90.fasta.clstr | cut -d'
```

```
' -f2 | sed -e 's/\.\.\./\\\>/' -e 's/^/^/' > unique_blocks_large_single_hits

grep -A1 -f unique_blocks_large_single_hits large_enough_unique_single_hits.fasta |
sed '/^--$/d' > blocks_for_probe_design.fasta

echo """Process complete."""

date
```

#The final output file, blocks_for_probe_design.fasta, contains sequences of
#each exon for each locus that are thought to be low-copy within the genome and
#meet the minimum length standards. The sequences that are output are ultimately
#derived from the original genome assembly (as opposed to the transcriptome
#assembly, if there are differences between the two).
#The ID line of each sequence is comma delimited and contains the name of the
#locus it originates from (from the transcriptome ID), the name of the genomic
#contig it matches followed by an underscore and a number indicating the exon
#within the locus (exon numbering within a locus is arbitrary), and the length
#of the exon. As an example, the following ID line identifies the second exon
#found within transciptome locus m.33568 and genome contig 5193133, which has a
#length of 186 bp:
#
>m.33568,5193133_2,186
ctca...

#End of File.

Appendix S2
Hyb-Seq workflow from raw reads to species tree

This document describes the generalized process for analyzing raw data from a Hyb-Seq library targeting hundreds or thousands of loci and exons, beginning with raw sequenced reads and ending with estimation of a species tree. This description includes the methods used in this study. However, due to idiosyncrasies in data sets, analysis preferences, bioinformatic expertise, operating environments and software availability, and the rapidly changing nature of trends and methods in sequencing technologies and analyses, we realize that applying these exact methods to other studies may not be feasible. Therefore, instead of providing a strict set of commands or a stand-alone program to perform all of these steps, we describe the motivation and reasoning behind our choices at each step. The reader is encouraged to consider their own needs, preferences, and resources when performing each step for their own analyses.

This document assumes that the enrichment probes will be targeting several hundred genes or loci, each of which may be constructed of multiple exons. It describes a strategy of assembling sequence for each exon, then combining exons into loci, then analyzing loci under a coalescent framework. This matches the framework that was adopted for this study, and we expect it will be a typical strategy for studying samples across several genera. However, the smallest units of contiguous targeted sequence do not necessarily need to be exons, but could be any low-copy region of the genome. In such a case the protocols described here would remain largely unchanged.

Throughout this document, wherever there is a reference to the “probe sequences” it is referring to the sequence of each exon used as a template for the probe design, not to the actual 80-120 bp oligonucleotide probes.

Read processing

Decisions about how raw reads with quality scores are processed prior to analysis can have a dramatic influence on final results. Typical manipulations include, but are not limited to, trimming reads that include adapter sequence, trimming portions of reads where the base quality is below a certain threshold, and removing reads that are exact duplicates of another read. We often find that a stricter filtering scheme results in more complete assemblies, even though a substantial portion of the raw data may be discarded. The removal of exact duplicate reads is intended to mitigate the effect of PCR bias preferentially amplifying some genome fragments over others.

We used the program Trimmomatic (Bolger, Lohse, and Usadel, 2014) to perform adapter trimming and quality filtering. This program can perform several read processing steps simultaneously, and is available in Java .jar format for availability across platforms. Duplicate removal was performed with the fastx_collapser program in the FASTX-Toolkit (available at http://hannonlab.cshl.edu/fastx_toolkit/), a suite of tools for performing several read processing actions.

Exon assembly

In this step the processed reads are assembled into the targeted exons and non-targeted high copy loci for each sample. Many programs are available for assembling Illumina sequence data. We used an iterative reference-guided assembly approach, where the probe sequences were used as a pseudo-reference to guide the assembly of the targeted exons from each sample. We performed this analysis with the program YASRA (Ratan, 2009), as implemented in the Alignreads pipeline (Straub et al, 2011), but other iterative assemblers, such as that included in the proprietary Geneious bioinformatics suite (Biomatters Ltd.; Auckland, New Zealand), would also be suitable. YASRA tolerates divergence from the reference and therefore allows the assembled sequence to have indels and substitutions relative to the reference. This feature also makes it useful for assembly of non-targeted loci by using sequences from related taxa as a reference (e.g., for plastome or nrDNA assembly; Straub et al., 2012). YASRA will continue to assemble sequence beyond the edges of the reference, which is useful in this application for assembly of the "splash zone." Assembly of the "splash zone" could also be accomplished by using a reference that contains introns of the expected size, such as the original genomic contigs from which the probes were designed. We performed the reference-guided assembly using a single YASRA run for each sample (as opposed to one run per exon per sample) by constructing a single reference sequence containing each exon separated by a string of 200 Ns.

De novo assembly can be used as an alternative to a reference-guided approach. This may be useful for locating novel intron-exon boundaries among samples, and may be able to simultaneously assemble both the targeted loci and non-targeted high-copy loci. However, a de novo approach may also be more computationally expensive and some programs may have difficulty with the differences in read coverage between targeted and non-targeted regions.

Identify assembled contigs (i.e. Assign orthology)

Regardless of the method used for exon assembly, the resulting assembled sequence will exist as a set of contigs that correspond to the set of targeted exons. These contigs may not be labeled with which exon they correspond to, as in the case of de novo assembly or in the present case of reference-guided assembly from concatenated exons, and they will contain sequence from the non-targeted "splash zone" beyond the boundaries of the targeted exons. To identify the exon that served as the reference for each contig, we matched the set of contigs against the set of exons using the program BLAT (Kent, 2002). This program allows indels between the database and query sequences, and can output the nucleotide sequence of the matching portion by outputting the results in the .pslx format (i.e.using the option "-out=pslx").

Sequence alignment: Collate exons and perform alignment

This first step in constructing a sequence alignment for each exon simply entails gathering together for each exon the sequence of each sample. We have written a program, assembled_exons_to_fasta.py, that performs this sorting if the previous step of matching exons to contigs was performed using BLAT. The BLAT output needs to be in the .pslx format and needs to have the exon probe sequences input as

the database (or targets), and the contigs to be labeled as the query. The user provides a fasta file of the probe sequences and a file containing a list of the .pslx files to be analyzed. The program outputs a fasta file for each targeted exon. Each fasta file contains the sequence for each sample that had the largest match to the exon. If a sample has no match to that exon, it is still included but given a sequence consisting of Ns equal to the length of the exon. Note that this program was designed for studies of taxa across several genera, so it includes only exon sequence and excludes introns or untranslated regions that may be present in the "splash zone." In applications where the "splash zone" is desired, BLAT could still be used to identify contigs, but other tools would be needed to extract the desired sequence. For each exon file we then performed a standard sequence alignment using the program MAFFT (Katoh and Toh, 2008).

Concatenate exons

At this point a sequence alignment is constructed for each locus. This can be done by simply concatenating the sequences for each exon of that locus. We performed this step using the program catfasta2phyml.pl (available at http://www.abc.se/~nylander/catfasta2phyml/) which simultaneously concatenates the sequences and transforms them into phyml format for downstream analysis. For each locus we concatenated the exons in an arbitrary order. This should have no effect on phylogenetic analyses under the assumption of independently evolving sites. However, some partitioning schemes, such as those that use codon position, may be more accurate if applied at the exon level rather than the locus level, and should be performed prior to exon concatenation. It is important to recall that the concatenated exons are not the equivalent to complete genes or cDNA sequences: in addition to the arbitrary ordering of exons, many exons will be missing either because they were excluded from the original probe set due to their short size or were simply not sufficiently enriched to be assembled in a particular sample.

Tree estimation

A rich literature exists on methods of gene tree estimation, and a rapidly expanding literature is being developed on species tree estimation. We favored a strategy of estimating gene trees, including bootstrap replicates, individually for each locus, and then using those sets of gene trees to estimate a species tree. Gene trees were estimated with RAxML (Stamatakis, 2006), with a species tree being estimated using MP-EST via the STRAW webserver, which incorporates a coalescent framework (Liu et al., 2010; Shaw et al., 2013).

Literature cited

BOLGER, A. M., M. LOHSE, and B. USADEL. In press. Trimmomatic: a flexible trimmer for Illumina Sequence Data. *Bioinformatics*.

KATOH, K., and H. TOH. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* 9: 286-298.

KENT, W. J. 2002. BLAT—the BLAST-Like Alignment Tool. *Genome Research* 12: 656-664.

LIU, L., L. YU, and S. V. EDWARDS. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* 10: 302.

RATAN, A. 2009. Assembly algorithms for next-generation sequence data. Ph.D. Dissertation, The Pennsylvania State University, University Park, Pennsylvania, USA.

SHAW, T. I., Z. RUAN, T. C. GLENN, and L. LIU. 2013. STRAW: Species TRee Analysis Web server. *Nucleic Acids Research* 41: W238-W241.

STAMATAKIS, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688-2690.

STRAUB, S. C. K., M. FISHBEIN, T. LIVSHULTZ, Z. FOSTER, M. PARKS, K. WEITEMIER, R. C. CRONN, et al. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.

STRAUB, S. C. K., M. PARKS, K. WEITEMIER, M. FISHBEIN, R. C. CRONN, and A. LISTON. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349-364.

Appendix S3
Comparison of post-sequencing analyses

Tools designed for the analysis of ultra-conserved elements (UCEs) have been used to analyze sequences resulting from the enrichment of genes containing multiple exons (Mandel et al., 2014). We compared one such analysis pipeline, phyluce v. 1.4 (Faircloth, 2014; Faircloth et al., 2012), with the bioinformatics pipeline presented here (Hyb-Seq pipeline). Major differences between the Hyb-Seq pipeline and phyluce include the method of sequence assembly (reference-guided or de novo, respectively) and the method of removing high-copy loci. The Hyb-Seq pipeline does much of the screening against high-copy loci during the probe design process, and after reference-guided assembly chooses the assembled contigs with the longest match against the targeted exon (if there are multiple hits). The phyluce pipeline filters out high-copy loci by removing those assembled contigs with matches against multiple targeted loci, and by removing targeted loci with matches against multiple assembled contigs. To better understand the effects of these differences we compared results from the Hyb-Seq pipeline against 1) the entire phyluce pipeline, and 2) the phyluce pipeline using the reference-guided assembled contigs used in the Hyb-Seq pipeline.

Assembly of adapter- and quality-trimmed reads was performed with Velvet (de novo assembly; Zerbino and Birney, 2008), as implemented in phyluce. Best k-mer length, k=23, for de novo assembly was estimated with KmerGenie v. 1.6663 (Chikhi and Medvedev, 2014), which searches for the optimal trade-off between largest k-mer length and maximum number of genomic k-mers in the dataset. Contigs from reference-guided and de novo assembly strategies were processed in phyluce with the following parameters: matching of contigs to probe sequences was performed with 90% minimum sequence identity, the “incomplete matrix” option allowed for missing data from taxa and genes, and genes with fewer than three taxa with sequence data were excluded.

Although the most liberal settings to allow for missing data were used, a large number of genes were dropped in both phyluce analyses compared to the Hyb-Seq pipeline (Table 3). This occurred primarily in the orthology assignment step, but also during the alignment step. The dramatic reduction in useable loci between the Hyb-Seq pipeline and phyluce, as it is currently implemented, is due to the targeted loci consisting of genes containing multiple exons. This is inappropriate for the filtering against multiple target/contig matches performed by phyluce, which was designed for UCEs that are expected to be assembled as single contigs. The several exons contained within a targeted gene might very well be assembled on separate contigs, especially under de novo assembly, and subsequently be excluded from the phyluce pipeline when it finds multiple contigs matching a single locus. We conclude that the current implementation of phyluce is inappropriately conservative for analyses of data sets similar to the one collected here and that of Mandel et al. (2014).

Literature cited

FAIRCLOTH, B. C. 2014. PHYLUCE: PHYLOGENETIC ESTIMATION FROM ULTRACONSERVED ELEMENTS. DOI: 10.6079/J9PHYL.

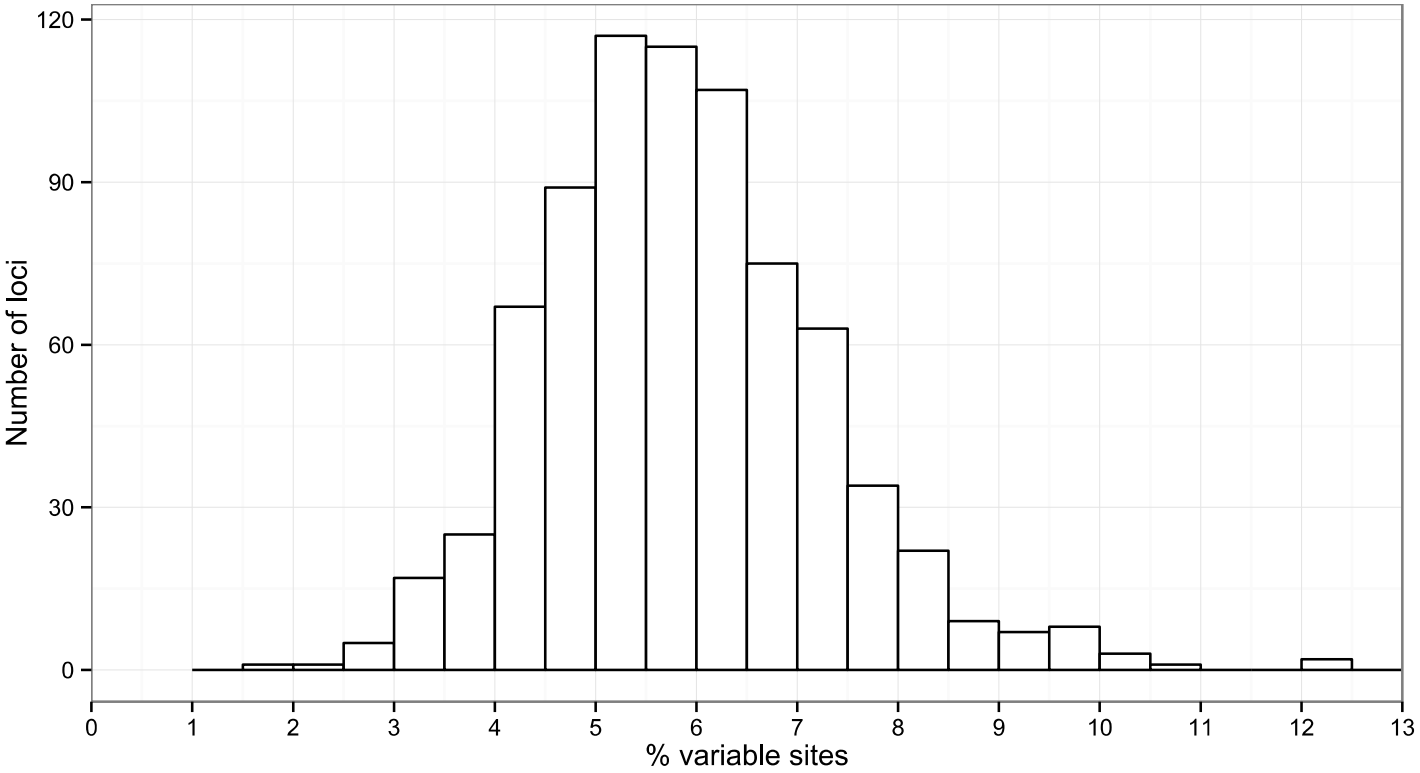
FAIRCLOTH, B. C., J. E. MCCORMACK, N. G. CRAWFORD, M. G. HARVEY, R. T. BRUMFIELD, AND T. C. GLENN. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61: 717- 726.

CHIKHI, R., AND P. MEDVEDEV. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30: 31-37.

MANDEL, J. R., R. B. DIKOW, V. A. FUNK, R. R. MASALIA, S. E. STATON, A. KOZIK, R. W. MICHELMORE, ET AL. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Science* 2:1300085.

ZERBINO, D., AND E. BIRNEY. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821-829.

Appendix S4



Percentage of variable sites within sequence alignments of 12 taxa at 768 loci.