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 et al., 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. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* 30: in press.

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.