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 version 1.4 (Faircloth et al., 2012; Faircloth, 2014), 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 version 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 data set. 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. GitHub repository: https://github.com/faircloth-lab/phyluce [accessed 15 July 2014].

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(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.