A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering

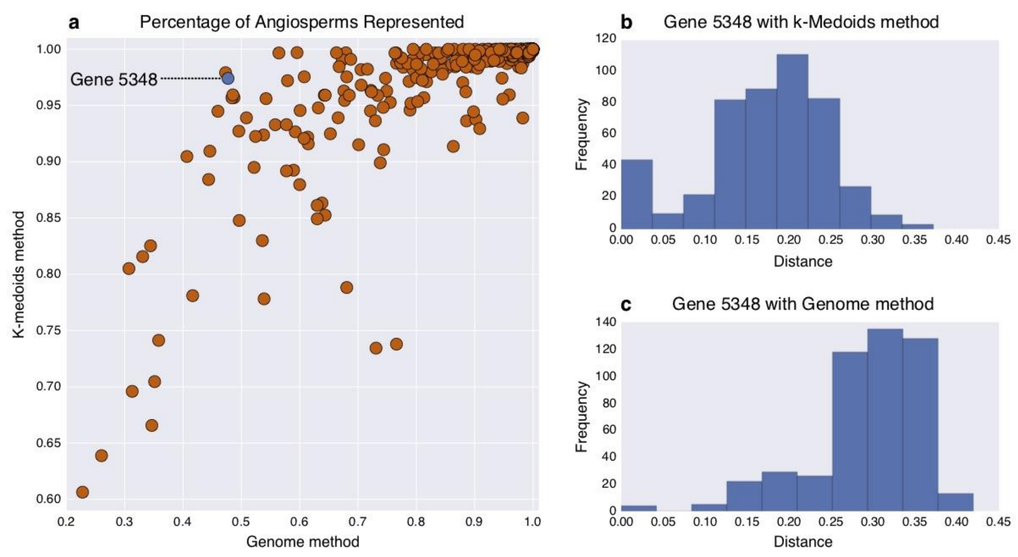
Matthew G. Johnson[1], Lisa Pokorny[2], Steven Dodsworth[3], Alison Devault[4], Laura R. Botigué[5], Robyn S. Cowan[2], Wolf L. Eiserhardt[6], Niroshini Epitawalage[2], Felix Forest[2], Paul Kersey[2], Jan T. Kim[2], James H. Leebens-Mack[7], Ilia J. Leitch[2], Olivier Maurin[2], Douglas E. Soltis[8], Pamela S. Soltis[8], Gane K.-S. Wong[9], William J. Baker[2], Norman J. Wickett[10]

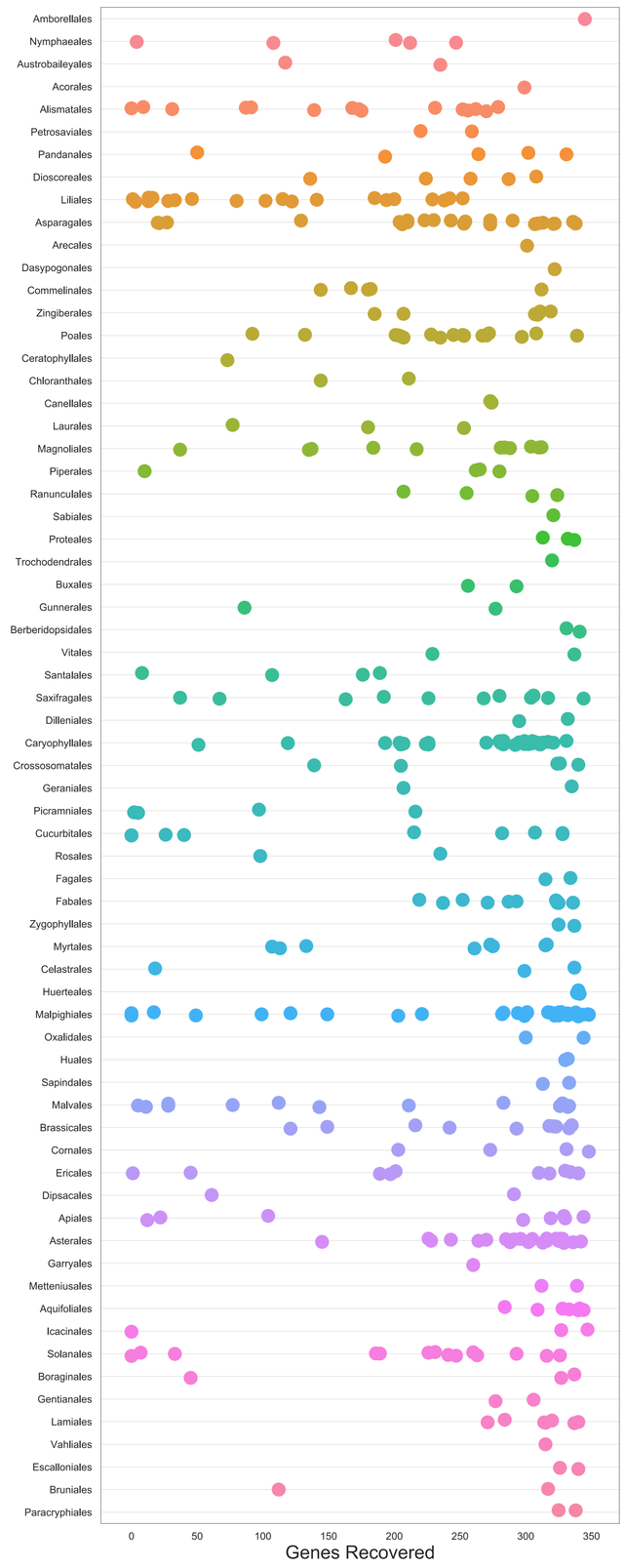
1 Texas Tech University | 2 Royal Botanic Gardens, Kew | 3 University of Bedfordshire | 4 Arbor Biosciences | 5 Centre for Research in Agriculture Genomics
6 Aarhus University | 7 University of Georgia | 8 University of Florida | 9 University of Alberta | 10 Chicago Botanic Garden

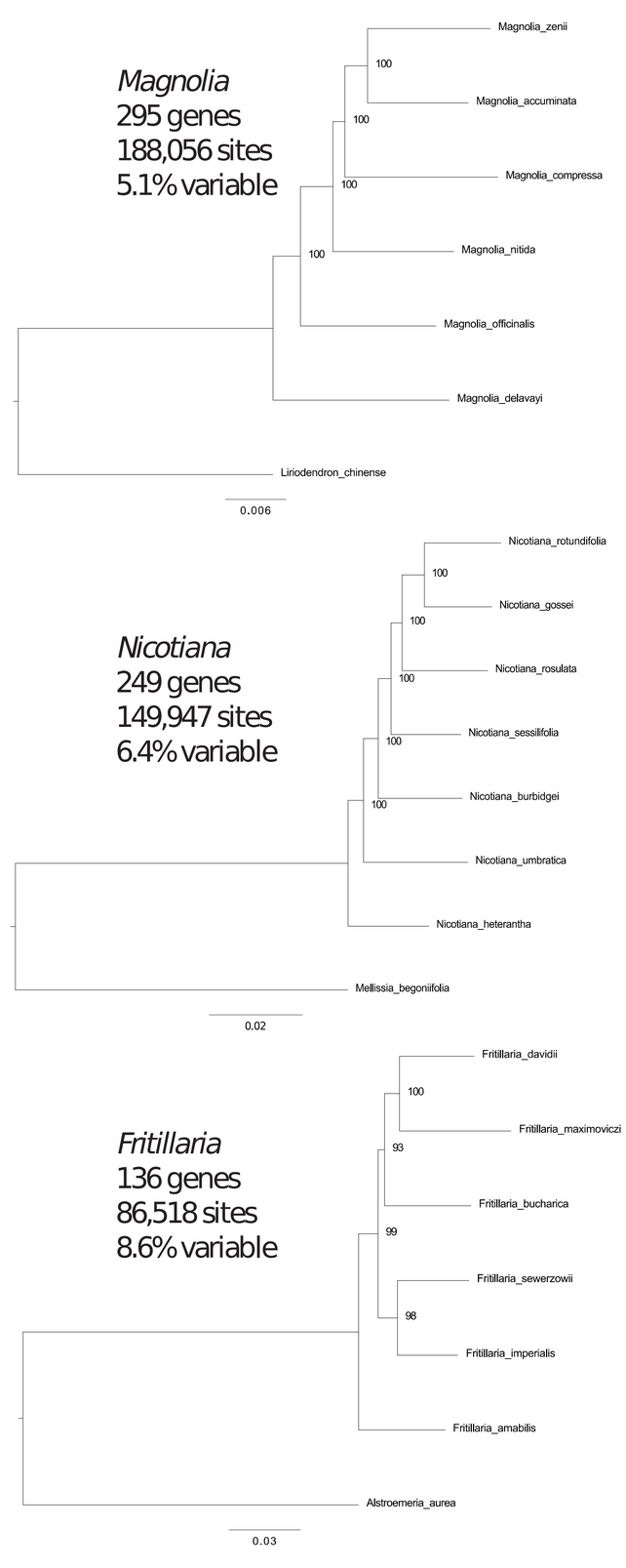
Sequencing of target enriched libraries is an efficient and cost-effective method for obtaining DNA sequence data from hundreds of nuclear loci for phylogenetic reconstruction. Much of the cost associated with developing targeted sequencing approaches is preliminary data needed for identifying orthologous loci for probe design. In plants, identifying orthologous loci has proven difficult due to a large number of whole genome duplication events, especially in the angiosperms. **We used alignments of over 600 angiosperms for 353 putatively single-copy protein coding genes to design targeted sequencing probes that would be useful for phylogenetics in any flowering plant.** To select between 5 and 15 sequences for each locus to use for probe design, we employed a k-medoids clustering approach which more efficiently represented angiosperm sequence diversity compared to using only published genomes. To test our probe sequences, we captured sequences from 42 species representing nearly all orders of angiosperms. We recovered exon sequence for 100 or more loci in all species, which was not affected by similarity to the taxa selected as medoids in probe design, suggesting that the probe design is effective in any group of flowering plants and would be useful for studies in phylogenetics of lineages at any taxonomic level. Additional enrichments have now been successful in 269 families representing 67 orders.

Probe design

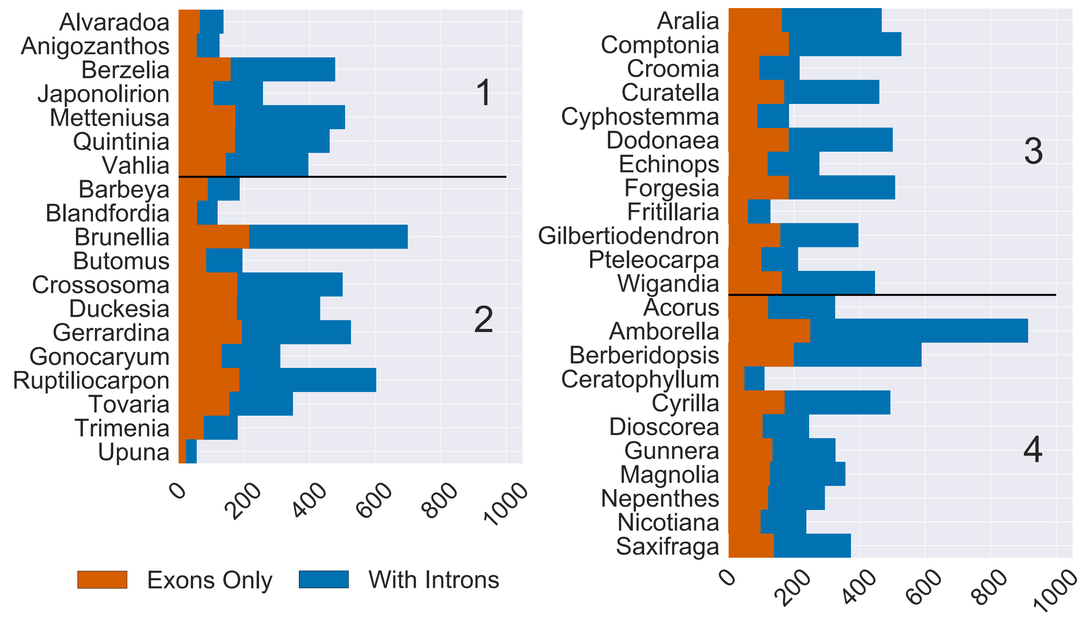
Initial sequences used for probe design represented >600 angiosperm species from 410 protein-coding loci from the OneKP initiative (onekp.org). We employed a k-medoids clustering algorithm (Bauckhage 2015) to partition sequences into groups, centered around a set number of sequences (the medoids). We identified 353 genes (targets) for which 95% of angiosperm sequences could be represented by 15 or fewer target instances. A sequence was considered represented if it was within 30% sequence divergence of one of the target instances. The number of required medoid sequences ranged from six (in 64 genes) to 15 (in 125 genes) with an average of 11.1 medoid sequences (median 13 sequences). The k-medoid method of selecting target sequences outperformed selection from genome sequences alone (see figure below) by decreasing the average distance between OneKP transcript sequences and representative sequences used for probe design. Using the mean length of target instances from each gene, the total length of coding sequence targeted was 260,802 bp. 75,151 3x-tiled 120-mer RNA probe sequences were designed from all selected target instances for each gene and synthesized by Arbor Biosciences.







Coding and flanking region recovery across angiosperm genera



Total Recovered Per-Locus Sequence Length (at 8x depth in Kbp) for Exon and Flanking Regions. We tested the probes using 42 "input" taxa that were not included in the OneKP transcriptome data, as well as one that was (*Amborella trichopoda*). For each of the 353 loci in the panel, the total amount of sequence recovered when mapped to exons alone (in orange) or exons + flanking non-coding/intron regions (in blue) is shown. Total targeted length of the panel (exons only) was 261Kb; median sequence recovered was 137Kb total for exons and 217Kb for flanking regions. Modified from Figure 4, Johnson, Pokorny, Dodsworth et al. (2018), Syst Biol.

Angiosperm353 probes are useful for phylogenetic analysis at both deep and shallow scales. Gene recovery is high in 407 species sequenced across 67 angiosperm orders (left) with most variation due to input DNA quality, not taxonomic bias. Sequence variation in coding regions is sufficient for within-genus phylogenetics (right), and will be further enhanced by incorporating noncoding sequences. Ongoing work will further optimize the procedure to make Angiosperms353 an effective tool across flowering plants.

Our recent publication:
Johnson, Pokorny, Dodsworth et al. (2018), Systematic Biology
doi.org/10.1093/sysbio/syy086

We thank BC, WAG, and Arboretum Wespelaar for providing tissue for DNA extraction. This work was supported by funding from the College of Arts and Sciences at TTU to MGJ, NSF grants (DEB-1239992 and DEB-1342873) to NJW, and grants from the Calleva Foundation, the Garfield Weston Foundation, and the Sackler Trust to the Royal Botanic Gardens, Kew. Probe sequences are publicly available under a CC-BY-SA license at github.com/mossmatters/Angiosperms353.