OzBaits development V1.0

A hybrid capture RNA bait set for resolving genetic and evolutionary relationships in angiosperms from deep phylogeny to intraspecific lineage hybridization

Michelle Waycott^{1,2}, Kor-jent van Dijk^{1,2}, and Ed Biffin²

¹ School of Biological Sciences, Department of Ecology and Evolutionary Biology and The Environment Institute, University of Adelaide, Adelaide, South Australia, Australia.

² State Herbarium of South Australia, Botanic Gardens and State Herbarium, Department for Water and Environment, Adelaide, South Australia, Australia.

Abstract

Novel multi-gene targeted capture probes have been developed with the objective of obtaining multi-locus high quality sequence reads across any angiosperm lineage. Using existing genomic and transcriptomic data, two independent single assay probe/bait sets have been developed, the first targeting conserved exons from 20 low copy nuclear genes (OzBaits_NR V1.0) and the second, 19 plastid gene regions (OzBaits_CP V1.0). These ‘universal’ bait sets can efficiently generate DNA sequence data that are suitable for systematics and evolutionary studies of flowering plants. The bait sets can be ordered as Daicel-Arbor Sciences custom myBaits. We demonstrate the utility of the bait set in consistently recovering the targeted genomic regions across an evolutionarily broad range of angiosperm taxa.

Introduction

Target enrichment approaches (eg. Cronn et al., 2012; Lemmon et al., 2012; Weitemier et al., 2014) have developed as a method of choice for systematics and evolutionary studies. Target enrichment probe sets can be designed to efficiently target a large number of phylogenetically informative loci and can be readily scaled up to include the dense sampling that is typical of molecular evolutionary studies. In addition, the method appears to work well with degraded samples, and can be used to obtain data from museum specimens (e.g. Hart et al., 2016) and from environmental samples (e.g. Lentz et al., 2021). While probe sets can be designed to target a narrow set of lineages (e.g. angiosperm families) ‘universal’ probe sets are attractive because of reduced cost and effort as well as the potential to generate broadly comparable data across evolutionarily distant lineages (e.g. Johnson et al., 2018; Breinholt et al., 2021).

In this methods manuscript we present two myBaits target enrichment probe sets that target, 1; conserved exons from 20 low copy nuclear genes and 2; 19 chloroplast gene regions. Leveraging existing genomic and transcriptomic data, these bait sets have been developed with the objective to obtain high quality sequence reads across all angiosperms. These new probe sets can be used to generate DNA sequence data that is suitable for a wide range of systematics and evolutionary analyses for any lineage of flowering plants. The bait sets were tested on a small selection of samples that cover a broad phylogeographic range.

Methods

Bait Design and Synthesis OzBaits V1.0

Nuclear genes

Target genes were selected from Duarte et al. (2010) who report a set of putatively orthologous low copy nuclear genes shared in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* (APVO SSC genes sensu Duarte et al., 2010). For each selected gene (Table 1), we downloaded all coding sequences (CDS) for the corresponding gene family available on Phytozome (v. 12; <https://phytozome.jgi.doe.gov>). We used the CDS for *Arabidopsis thaliana* to retrieve putatively homologous transcript sequences

Table 1: Nuclear low copy genes used for OzBaits_NR V1.0, Design ID **D10121VasPI**

Transcript name (<i>Arabidopsis thaliana</i> TAIR10)	Definition	Target coordinates (<i>Arabidopsis thaliana</i> TAIR10 genomic sequence)
AT1G03750	CHROMATIN MODIFICATION-RELATED PROTEIN EAF1	exon 2: 715-1881
AT1G08460	HISTONE DEACTYLASE 8	exon 2: 471-1057
AT1G43580	SPHINGOMYELIN SYNTHETASE	exon 2: 910-1307
AT1G77210	ALCOHOL DEHYDROGENASE I	exon 4: 655-1140
AT3G02300	REGULATOR OF CHROMOSOME CONDENSATION FAMILY PROTEIN	exon 6: 1885-2316; exon 7: 2431-2622
AT3G21540	WD REPEAT CONTAINING PROTEIN 3	exon 3: 766-1884
AT3G51050	FG-GAP REPEAT-CONTAINING PROTEIN	exon 12: 2906-3208; exon 13: 3303-3776
AT4G01040	CHITINASE DOMAIN-CONTAINING PROTEIN 1	exon 7: 1335-1689 ; exon 8: 1783-2039
AT4G36390	METHYLTHIOTRANSFERASE	exon 3: 1886-2386
AT4G36650	PLANT-SPECIFIC TFIIIB-RELATED PROTEIN	exon1: 75-656; exon 3: 1549-2319
AT4G37030	HYPOTHETICAL ANGIOSPERM GENE	exon 5: 1142-1430; exon 8: 1968-2416
AT4G38890	FMN-LINKED OXIDOREDUCTASES SUPERFAMILY PROTEIN	exon 3: 800-1027
AT5G04520	3-OXOACY-[ACYL-CARRIER-PROTEIN] SYNTHASE-LIKE PROTEIN	exon 2: 395-625; exon 3: 702-1052
AT5G13390	NO EXINE FORMATION 1	exon 1: 63-1283; exon 2: 1438-2009
AT5G24760	ALCOHOL DEHYDROGENASE-LIKE 6	exon 4: 1085-1407
AT5G49570	PEPTIDE-N4-(N-ACETYL-BETA-GLUCOSAMINYL) ASPARAGINE AMIDASE	exon 13: 2591-3355
AT5G58760	DNA DAMAGE-BINDING PROTEIN 2	exon 9: 2102-2350; exon 10: 2433-2629
AT5G64150	RNA METHYLTRANSFERASE FAMILY PROTEIN	exon 1: 488-1240; exon 2: 1484-1729
AT5G65860	ANKYRIN REPEAT FAMILY PROTEIN	exon 1: 299-1339

from The 1000 Plants Project (1KP; <https://www.onekp.com>) using the China National Genebank (<https://db.cngb.org>) BLAST portal and the following settings: Discontiguous Mega-Blast, expect value=10, maximum target sequences=1000, selected organisms=Magnoliophyta (taxid:3398). The sequences retrieved from the BLAST search were downloaded and combined with the Phytozome data for each gene and made into a BLAST database in Geneious (Kearse et al., 2012; <https://www.geneious.com>). We queried each BLAST database using the *A. thaliana* gene family member with exon annotations manually added, and the following settings: Discontiguous Mega-Blast, expect value=10, maximum target sequences=1200, results=Hit Table, retrieve=Matching Region with Annotation. We then extracted all sequences matching one or more exon annotations in *A. thaliana* with the caveat that the exon was > 180 bp in length to allow for bait tiling. The

extracted sequences were clustered using CD-HIT-EST (Li and Godzik, 2006; http://weizhong-lab.ucsd.edu/cdhit_suite; Huang et al., 2010) with a sequence identity cut-off fraction of 0.95 and a length similarity fraction of 0.2, and one representative sequence (the longest) per cluster was selected. A total of 9,900 representative sequences were used for bait design with 120-nucleotide baits and ~2X flexible tiling density for a total of 47,001 baits (details of design can be found in Supplemental 1). Bait design (myBaits) and synthesis was performed by Daicel-Arbor Biosciences (formerly Mycroarray; Ann Arbor, Michigan, USA) and has the Design ID **D10121VasPI**.

Chloroplast genes

Although hybrid capture approaches will frequently recover complete or near complete plastid genomes from off-target sequencing reads, the efficiency of recovery is reduced in degraded

Table 2: Chloroplast genes used for OzBaits_CP V1.0, Design ID **D10122Plstd**

Plastid gene name	Definition	Target coordinates of <i>Zostera marina</i> genome (Genbank: MF370229)
psbA	PHOTOSYSTEM II PROTEIN D1	18-1,370
matK	MATURASE K	1,518-3,020
psbK	PHOTOSYSTEM II PROTEIN K	6,526-7,282
atpF	ATP SYNTHASE CF0 SUBUNIT I	10,797-11,827
atpH	ATP SYNTHASE CF0 SUBUNIT III	13,061-13,494
atpI	ATP SYNTHASE CF0 SUBUNIT IV	14,275-15,242
rpoB-rpoC1	RNA POLYMERASE BETA SUBUNIT/RNA POLYMERASE BETA	22,448-23,563
psbD	PHOTOSYSTEM II PROTEIN D2	32,336-33,253
psbZ	PHOTOSYSTEM II PROTEIN Z	35,075-35,742
ndhC-ndhK	NADH-PLASTOQUINONE OXIDOREDUCTASE SUBUNIT 3/NADH-PLASTOQUINONE OXIDOREDUCTASE SUBUNIT K	49,324-50,073
rbcl	RIBULOSE-1,5-BISPHOSPHATE CARBOXYLASE/OXYGENASE LARGE SUBUNIT	51,777-52,659
atpB	ATP SYNTHASE CF1 BETA SUBUNIT	53,256-54,042
accD	ACETYL-COA CARBOXYLASE CARBOXYLTRANSFERASE BETA SUBUNIT	57,899-58,952
petA	CYTOCHROME F	62,109-62,826
psbE-psbF	PHOTOSYSTEM II CYTOCHROME B559 ALPHA SUBUNIT/PHOTOSYSTEM II CYTOCHROME B559 BETA SUBUNIT	63,549-64,591
psbN-psbH	PHOTOSYSTEM II PROTEIN N/PHOTOSYSTEM II PHOSPHOPROTEIN	73,725-74,478
petD	CYTOCHROME B6/F COMPLEX SUBUNIT IV	76,795-77,139
rpl14-rpl16	RIBOSOMAL PROTEIN L14/RIBOSOMAL PROTEIN L16	80,190-81,006
ndhF	NADH-PLASTOQUINONE OXIDOREDUCTASE SUBUNIT 5	118,160-118,825

material such as herbarium specimens (e.g. Bakker et al., 2016). To improve the consistency of recovery of high-coverage plastid data, we designed a set of probes to target a range of chloroplast regions in angiosperms using the NCBI RefSeq plastid sequences (<https://www.ncbi.nlm.nih.gov/genome/organelle/>) for probe design. For each target gene, the corresponding CDS were extracted from the RefSeq genome data and clustered as above, using CD-HIT-EST. A total of 2978 sequences, targeting 19 plastid regions (Table 2) were used for bait design with 120-nucleotide baits and ~2X flexible tiling density for a total of 14,922 baits (details of design can be found in Supplemental 2). Bait design (myBaits) and synthesis was performed by Arbor Biosciences (formerly Mycarray; Ann Arbor, Michigan, USA) and has the Design ID **D10122Plstd**.

Library Preparation and Capture

DNA for targeted capture needs to be of the best quality possible, but this approach has also been successfully tested on degraded DNA (herbarium material from the late 19th century) although sequencing success varies. Most of the libraries prepared in our laboratory used DNA that was extracted with the Qiagen Plant Mini kit. DNA is then normalized to 2ng/uL before proceeding to library preparation, DNAs with lower concentration were used neat, but sequencing success generally decreased. A detailed description of laboratory procedures can be found in Supplemental 3. The described method follows the NEBNext® Ultra™ II DNA Library Prep Kit (v4.0) with Sample Purification Beads Kit (New England Biolabs, Ipswich, Massachusetts, USA). But other kits can be used too, adjustments need to be made to the protocol though. To maintain consistency in the library preparation all liquid handling was done on an Eppendorf epMotion® 5075t - Liquid Handling Workstation, but steps can also be performed manually, both approaches are described in Supplemental 3. To keep cost down and to be able to multiplex in high sample numbers custom stubby Y-adapters were developed and used in the library perps described in Supplemental 3. A detailed description of development with necessary sequence files can be found in Supplemental 4 - 7.

After library preparation, samples are pooled to reduce costs of Hybrid Capture, in Supplemental 3 pools are made with 16 libraries (step 44-47), but if samples are of low quality pools in sets of 8 will generally increase success, or, if you are dealing with ancient DNA you might consider doing captures on single libraries and follow ancient DNA workflows.

Hybrid Capture was performed following the manual provided by myBaits. Our captures were performed at 65°C, but temperatures can be changed depending on the capture efficiency needed. During the hybridisation step Chill-out™ red liquid wax was added to avoid evaporation. After capture libraries were amplified with indexed primers (Supplemental 6), cleaned up with Ampure and pooled to equimolar concentrations (Nuclear and Chloroplast separately). Chloroplast DNA captures are generally orders of magnitude stronger in molarity. This is probably due to the higher DNA extraction yield (number of genome copies recovered) and fewer loci that are used for capture. Only 10% (molarity) of the final library pool was allocated to the Chloroplast library. As a final quality step libraries are size selected according to insert sizes and sequencing platform used to avoid dimers and over-long inserts. If DNA is expected to be degraded size selection should target smaller inserts but avoid primer dimers. In our laboratory we generally sequence on an Illumina NovaSeq S1 flow cell pooling 3-5 plates worth of samples. To test the bait sets OzBaits_NR V1.0 and OzBaits_CP V1.0 a small number of samples covering a broad range of angiosperms was tested (Table 3, Supplemental Table S8). Plant material was collected from either dry herbarium specimens or from the Adelaide Botanic Gardens living collection. Approximately 20-40 mg of dry material was transferred to a 2mL screw cap tube and pulverized in an Omni Bead Ruptor 24. Total genomic DNA was extracted using the DNeasy™ Plant Kit (Qiagen) according to the manufacturer's instructions. Library prep was performed according to Supplemental 3 and targeted capture at a hybridization temperature of 65°C.

Data processing and analyses

High-throughput 150 bp paired-end reads were processed using CLC Genomics Workbench (v. 7; <https://www.qiagenbioinformatics.com/>). De-multiplexing, quality trimming (using a Phred score threshold of 20) and de-novo assembly (length fraction=0.7, similarity fraction=0.9, minimum contig length=300) was performed to obtain contigs at each locus (low coverage cut-off = 10, insert Ns for regions of low coverage, conflict resolution = vote). To identify the 'on target' nuclear contigs from each assembly we first made a BLAST database comprising the full set of reference genes from the Phytozome dataset, and we queried against these using the contigs from each *de-novo* assembly. Contigs corresponding to target loci were then retrieved by creating a BLAST

Table 3: Samples used to test OzBaits_NR and OzBaits_CP V1.0. Detailed information can be found in supplemental Table S8. BioSamples are part of NCBI Bioproject PRJNA680375.

OzBaits#	Biosample	Species	Family	Location
OzB_001	SAMN21154828	<i>Allocasia brisbanensis</i>	ARACEAE	South Australia, Adelaide Botanic Gardens ex Queensland, Tolga
OzB_002	SAMN21154829	<i>Alpinia caerulea</i>	ZINGERBERACEAE	South Australia, Adelaide Botanic Gardens ex Queensland
OzB_026	SAMN21155302	<i>Arthropodium milleflorum</i>	ASPARAGACEAE	Queensland, Palmgrove National Park
OzB_004	SAMN21154831	<i>Banksia sp.</i>	PROTEACEAE	Western Australia, Jingaring
OzB_005	SAMN21154832	<i>Calytrix tetragona</i>	MYRTACEAE	South Australia, Wattle Range
OzB_006	SAMN21154833	<i>Ceratophyllum demersum</i>	CERATOPHYLLACEAE	South Australia, Modbury North, cultivated plants ex Northern Territory
OzB_007	SAMN21154834	<i>Comesperma volubile</i>	POLYGALACEAE	South Australia, Ward Belt
OzB_008	SAMN21154835	<i>Crinum luteolum</i>	AMARYLLIDACEAE	South Australia, Oodnadatta, 20 km N of Willam Creek
OzB_009	SAMN21154836	<i>Cryptocarya mackinoniana</i>	LAURACEAE	South Australia, Adelaide Botanic Gardens ex Queensland, Cairns
OzB_010	SAMN21154837	<i>Drosera binata</i>	DROSERACEAE	New South Wales, Byron Bay, Arakwal National Park
OzB_011	SAMN21154838	<i>Halgania adromedifolia</i>	BORAGINACEAE	South Australia, Eyre Peninsula
OzB_012	SAMN21154839	<i>Halophila johnsonii</i>	HYDROCHARITACEAE	USA: Florida, Indian River Lagoon
OzB_013	SAMN21154840	<i>Magnolia grandiflora</i>	MAGNOLACEAE	South Australia, Adelaide Botanic Gardens Living Collection
OzB_014	SAMN21154841	<i>Myriophyllum aquaticum</i>	HALORAGACEAE	South Australia, Nairne
OzB_015	SAMN21154842	<i>Parapholis incurva</i>	POACEAE	South Australia, Adelaide, St Kilda
OzB_016	SAMN21154843	<i>Peumus boldus</i>	MONIMIACEAE	South Australia, Adelaide Botanic Gardens ex Victoria, Royal Melbourne Botanic Garden
OzB_017	SAMN21154844	<i>Piper novae-hollandiae</i>	PIPERACEAE	South Australia, Adelaide Botanic Gardens ex Queensland, Tolga
OzB_018	SAMN21154845	<i>Prostanthera behriana</i>	LAMIACEAE	South Australia, Northern Lofty
OzB_019	SAMN21154846	<i>Pultenaea pedunculata</i>	FABACEAE	South Australia, Wanilla Conservation Park
OzB_020	SAMN21154847	<i>Rapanea subsessilis</i>	PRIMULACEAE	South Australia, Adelaide Botanic Gardens ex Queensland, Walkamin Yuruga Nursery
OzB_021	SAMN21154848	<i>Romneya coulteri</i>	PAPAVERACEAE	South Australia, Adelaide Botanic Gardens
OzB_022	SAMN21154849	<i>Scaevola albida</i>	GOODENIACEAE	South Australia, Northern Lofty
OzB_023	SAMN21154850	<i>Smilax aspera</i>	SMILACACEAE	South Australia, Adelaide Botanic Gardens ex FRANCE, Corsica, Bonifacio
OzB_024	SAMN21154851	<i>Zostera muelleri</i>	ZOSTERACEAE	Queensland, Moreton Bay, Wanga Wallan Banks
OzB_025	SAMN21154852	<i>Zygophyllum apiculatum</i>	ZYGOPHYLLACEAE	South Australia, Mount Friday, Hiltaba Station

database for each *de-novo* assembly and running a query against the best matching Phytozome reference sequences identified in the first step, using an E-value $\leq 1E-100$ and maximum target sequences set to 20. For each sample, including a diverse set of angiosperm taxa (Table 3, Supplemental Table S8), we assessed per locus recovery, gene copy number (assuming that overlapping and non-identical contigs for a given gene represent paralogues) and the total length of sequence recovered per sample (Table 4).

In order to recover the targeted plastid regions, we used the chloroplast genome sequence of *Arabidopsis lyrata* (KX886355) as a BLAST reference to query the sample specific *de novo* assemblies (as above) using an E-value $\leq 1E-20$ and maximum target sequences set to 20.

Results and Discussion

Recovery of target loci across a diverse set of angiosperm samples was high for both the nuclear (Table 4) and chloroplast target regions (Table 5). For the targeted nuclear gene regions, there appears to be little evidence of a taxonomic effect on gene recovery, suggesting that these should be effectively recovered for any angiosperm. Plastid gene regions are rarely targeted given that

[illegible]

these are commonly recovered from off-target reads (e.g. Weitemier et al., 2014). However, a plastid specific probe set enables consistent recovery of the target regions from a wide range of samples such as environmental and ancient DNA (e.g. Schulte et al., 2021) comprising degraded samples with typically low recovery of genomic DNA. Both chloroplast and nuclear gene regions have been successfully employed in evolutionary analyses at a range of taxonomic levels (Nge et al., 2021;Waycott et al., 2021).

References

- Bakker, F.T., Lei, D., Yu, J., Mohammadin, S., Wei, Z., Van De Kerke, S., Gravendeel, B., Nieuwenhuis, M., Staats, M., Alquezar-Planas, D.E., and Holmer, R. (2016). Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society* 117, 33-43.
- Breinholt, J.W., Carey, S.B., Tiley, G.P., Davis, E.C., Endara, L., Mcdaniel, S.F., Neves, L.G., Sessa, E.B., Von Konrat, M., Chantanaorrapint, S., Fawcett, S., Ickert-Bond, S.M., Labiak, P.H., Larraín, J., Lehnert, M., Lewis, L.R., Nagalingum, N.S., Patel, N., Rensing, S.A., Testo, W., Vasco, A., Villarreal, J.C., Williams, E.W., and Burleigh, J.G. (2021). A target enrichment probe set for resolving the flagellate land plant tree of life. *Applications in Plant Sciences* 9, e11406.
- Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V., and Udall, J. (2012). Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* 99, 291-311.
- Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, P.K., Leebens-Mack, J., and Claude, W.D. (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.
- Hart, M.L., Forrest, L.L., Nicholls, J.A., and Kidner, C.A. (2016). Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65, 1081-1092.
- Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682.
- Johnson, M., Pokorny, L., Dodsworth, S., Botigue, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T., Leebens-Mack, J.H., Leitch, I.J., Maurin, O., Soltis, D.E., Soltis, P.S., Wong, G.K.-S., Baker, W.J., and Wickett, N. (2018). A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-medoids Clustering. *bioRxiv*.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647-1649.

- Lemmon, A.R., Emme, S.A., and Lemmon, E.M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61, 727–744.
- Lentz, D.L., Hamilton, T.L., Dunning, N.P., Tepe, E.J., Scarborough, V.L., Meyers, S.A., Grazioso, L., and Weiss, A.A. (2021). Environmental DNA reveals arboreal cityscapes at the Ancient Maya Center of Tikal. *Scientific Reports* 11, 12725.
- Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659.
- Nge, F.J., Biffin, E., Thiele, K.R., and Waycott, M. (2021). Reticulate Evolution, Ancient Chloroplast Haplotypes, and Rapid Radiation of the Australian Plant Genus *Adenanthos* (Proteaceae). *Frontiers in Ecology and Evolution* 8.
- Schulte, L., Bernhardt, N., Stoof-Leichsenring, K., Zimmermann, H.H., Pestryakova, L.A., Epp, L.S., and Herzsuh, U. (2021). Hybridization capture of larch (*Larix* Mill.) chloroplast genomes from sedimentary ancient DNA reveals past changes of Siberian forest. *Molecular Ecology Resources* 21, 801-815.
- Waycott, M., Van Dijk, K.-J., Calladine, A., Bricker, E., and Biffin, E. (2021). Genomics-based phylogenetic and population genetic analysis of global samples confirms *Halophila johnsonii* Eiseman as *Halophila ovalis* (R.Br.) Hook.f. *Frontiers in Marine Science*.
- Weitemier, K., Straub, S.C.K., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A., and Liston, A. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in plant sciences* 2, apps.1400042.