Applications in Plant Sciences

Botanical Society of America
leading scientists and educators since 1893

PROTOCOL NOTE

INVITED SPECIAL ARTICLE
For the Special Issue: Exploring Angiosperms353: a Universal Toolkit for Flowering Plant Phylogenomics

The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions

Kasper P. Hendriks[1,2,12], Terezie Mandáková[3], Nikolai M. Hay[4], Elfy Ly[1], Alex Hooft van Huysduynen[1], Rubin Tamrakar[5], Shawn K. Thomas[6], Oscar Toro-Núñez[7], J. Chris Pires[6], Lachezar A. Nikolov[8], Marcus A. Koch[9], Michael D. Windham[4], Martin A. Lysak[3], Félix Forest[10], Klaus Mummenhoff[2], William J. Baker[10], Frederic Lens[1,11], and C. Donovan Bailey[5]

Manuscript received 15 November 2020; revision accepted 16 March 2021.

[1] Naturalis Biodiversity Center, Research Group Functional Traits, P.O. Box 9517, 2300RA Leiden, The Netherlands

[2] Department of Biology/Botany, Osnabrück University, Osnabrück 49076, Germany

[3] CEITEC, Masaryk University, Brno CZ-625 00, Czech Republic

[4] Department of Biology, Duke University, Durham, North Carolina 27708, USA

[5] Department of Biology, New Mexico State University, Las Cruces, New Mexico 88001, USA

[6] Division of Biological Sciences, University of Missouri, Columbia, Missouri 65211, USA

[7] Departamento de Botánica, Universidad de Concepción, Concepción, Chile

[8] Department of Molecular, Cell and Developmental Biology, University of California Los Angeles, Los Angeles, California 90095, USA

[9] Centre for Organismal Studies COS, Biodiversity and Plant Systematics, Heidelberg University, 69120 Heidelberg, Germany

[10] Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, United Kingdom

[11] Institute of Biology Leiden, Plant Sciences, Leiden University, Sylviusweg 72, 2333 BE, Leiden, The Netherlands

[12] Author for correspondence: kasper.hendriks@naturalis.nl

Citation: Hendriks, K. P., T. Mandáková, N. M. Hay, E. Ly, A. Hooft van Huysduynen, R. Tamrakar, S. K. Thomas, et al. 2021. The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. *Applications in Plant Sciences* 9(7): e11438.

doi:10.1002/aps3.11438

PREMISE: Researchers adopting target-enrichment approaches often struggle with the decision of whether to use universal or lineage-specific probe sets. To circumvent this quandary, we investigate the efficacy of a simultaneous enrichment by combining universal probes and lineage-specific probes in a single hybridization reaction, to benefit from the qualities of both probe sets with little added cost or effort.

METHODS AND RESULTS: Using 26 Brassicaceae libraries and standard enrichment protocols, we compare results from three independent data sets. A large average fraction of reads mapping to the Angiosperms353 (24–31%) and Brassicaceae (35–59%) targets resulted in a sizable reconstruction of loci for each target set ($\bar{x} \geq 70\%$).

CONCLUSIONS: High levels of enrichment and locus reconstruction for the two target sets demonstrate that the sampling of genomic regions can be easily extended through the combination of probe sets in single enrichment reactions. We hope that these findings will facilitate the production of expanded data sets that answer individual research questions and simultaneously allow wider applications by the research community as a whole.

KEY WORDS Brassicaceae; combining probes; enrichment; Hyb-Seq; phylogenomics; phylogeny; population biology; target enrichment.

Target capture approaches to DNA analyses (e.g., Mandel et al., 2014; Weitemier et al., 2014) are emerging as one of the most important tools in evolutionary biology, especially phylogenomics. Researchers adopting these methods are clear on the importance and utility of the data generated (e.g., Johnson et al., 2019), but often face a difficult decision during the early stages of project design. They must typically choose between the use of a universal probe set (e.g., Buddenhagen et al., 2016; Johnson et al., 2019) developed to work across larger taxonomic scales (e.g., the angiosperms), or a narrower lineage-specific probe set designed for the group of interest (e.g., Mandel et al., 2014; Weitemier et al., 2014; Vatanparast et al., 2018; Gardiner et al., 2019; Koenen et al., 2020). When considering target enrichment options, the core exons of universal probe sets are perhaps viewed as best suited for higher-level phylogenetic problems, where their conserved nature tends to have greatest utility (but see Mitchell et al., 2017; Wanke et al., 2017). Such probe sets, which have now been applied across nearly all angiosperm families (e.g., Baker et al., 2017; Dodsworth et al., 2019), produce data that can be easily integrated with studies from other labs focused on alternative samples or even different lineages including outgroup species (e.g., Buddenhagen et al., 2016; Johnson et al., 2019). The potential utility of these markers and their associated flanking regions are also being explored for the elucidation of species complexes (e.g., Larridon et al., 2020) and population-level studies (e.g., Slimp et al., 2020). By contrast, well-designed lineage-specific probes, incorporating local information on single-copy genes and greater fidelity between probe and target, can successfully select and recover a larger portion of orthologous gene space (e.g., Soto Gomez et al., 2019). They may also maximize the phylogenetic signal per region sequenced (e.g., Folk et al., 2015), generating data even more amenable to solving problems with both recalcitrant nodes in phylogenetic trees and questions in population biology. However, lineage-specific data tend not to be readily combinable with data generated using other probe sets.

The choice between universal and lineage-specific probe sets can be further complicated when previously generated lineage-specific data are available for some samples, resulting in a hesitancy to engage a universal set because of the inability to integrate existing data. The tradeoffs associated with these choices can have long-term consequences, both for the source study and for the downstream utility of the data generated. In an ideal world, researchers would interrogate the same set of comprehensive loci, with targets able to address evolutionary questions ranging from the divergence of major clades to population-level studies, or even "next generation barcoding" (Johnson et al., 2019). However, the molecular evolution of plant genomes largely dictates that no one set of sampled loci is likely to fit this ideal range of desired qualities for all scales and levels of investigation; thus, researchers continue to struggle with the decision associated with adopting universal probes or designing and applying a lineage-specific set, leading to suggestions that both classes of probe sets might be engaged in some projects (e.g., Couvreur et al., 2019).

As part of a collaboration between the Plant and Fungal Trees of Life project (PAFTOL; https://www.kew.org/science/our-science/projects/plant-and-fungal-trees-of-life) (Baker et al., 2021) and a group of Brassicaceae systematists, we faced this issue when selecting probes for target enrichment–based phylogenomic studies of the Brassicaceae. A confluence of several previously independent research projects has led us to envision performing target capture sequencing for all 4000 species in the family. In this context, a case can be made to favor the use of the universal Angiosperms353 probe set (Johnson et al., 2019), with obvious emphasis on the long-term added value of sequencing loci that could be combined with data from similar ongoing studies across the angiosperms. However, it could also be argued that a recently published Brassicaceae-specific probe set (Nikolov et al., 2019), targeting more variable loci and four-fold greater base pair representation, is better suited to resolving the fine details of the family's phylogenetic relationships. With the availability of both the Angiosperms353 and Brassicaceae probe sets, and the amount of existing data generated using the latter, our path forward was not entirely clear. We all agreed that one of the least desirable options was embarking on separate, partially overlapping projects applying different probe sets.

Ultimately, we settled on a pilot study to investigate the feasibility of applying both probe sets by combining them in a single hybridization reaction and sequencing captured targets simultaneously. Ideally, this would facilitate the capture of universal and lineage-specific loci with minimal extra effort and only a small additional cost per sample associated with the purchase of two probe sets. Here, we test the efficacy of combining two probe sets that share just 30 loci, the Angiosperms353 probes (353 loci, 260 kbp total length) and the Brassicaceae-specific set (1827 exons ["Nikolov1827"] derived from 764 loci, 940 kbp total length), using three different sets of Brassicaceae gDNA samples and enriched libraries generated in two independent labs. Because neither lab had prior experience with these approaches, the study offers both an assessment of combining probe sets and the feasibility of doing so in a variety of labs with limited experience in the generation of target capture data.

METHODS AND RESULTS

DNA extraction and library preparation

The DNA samples (Appendix 1) used as part of our broader study were obtained from a combination of new extractions using a QIAGEN DNeasy PowerPlant Pro Kit (with subsequent purification of greenish extracts using the DNeasy PowerClean CleanUp Kit; QIAGEN, Hilden, Germany) and existing extractions from a prior project generated using the extraction protocol of Alexander et al. (2006). These extractions were used to develop three example target-enrichment Brassicaceae data sets (Table 1) from two independent labs, the Bailey lab (New Mexico State University, Las Cruces, New Mexico, USA) and Naturalis Biodiversity Center (Leiden, The Netherlands; principal investigator: Frederic Lens). Example enrichment sets 1 (six libraries) and 2 (10 libraries) were generated in the Bailey lab, while set 3 (10 libraries) came from Naturalis. The Bailey lab samples were all representatives of the tribe Boechereae, while the Naturalis samples (obtained from collections at the University of Osnabrück, Osnabrück, Germany) represent a broader sampling across the Brassicaceae.

Initially, the Bailey lab generated libraries from six silica gel–dried DNA extractions (set 1) of Boechereae species (Table 1). This set derived from fresh silica gel–dried leaves and included four taxa, with three technical replicates of one taxon (PJA370) to investigate reproducibility. Later, the Bailey lab generated results from hybridization reactions including 23–26 herbarium sample–derived libraries per enrichment. Ten samples, with between 1.5 million and 4 million recovered reads, were randomly selected for evaluation and presented in set 2 (Table 1). Similarly, Naturalis generated larger

TABLE 1. Samples included in each set of example enrichments. Sample sets 1 and 2 were generated by the Bailey lab (New Mexico State University), while set 3 came from the Naturalis Biodiversity Center.

Sample set	Species	DNA extraction label[a]	NCBI SRA ID
1	*Boechera sanluisensis* P. J. Alexander	PJA296A	SAMN17836232
	Cusickiella douglasii (A. Gray) Rollins	PJA370A	SAMN17836233
	Cusickiella douglasii	PJA370B	SAMN17836234
	Cusickiella douglasii	PJA370C	SAMN17836235
	Halimolobos jaegeri (Munz) Rollins	PJA244	SAMN17836236
	Sandbergia whitedii Greene	PJA248	SAMN17836237
2	*Boechera paupercula* (Greene) Windham & Al-Shehbaz	JB242	SAMN17836238
	Boechera pendulina (Greene) W. A. Weber	JB152	SAMN17836239
	Boechera pendulina	w4485	SAMN17836246
	Boechera platysperma (A. Gray) Al-Shehbaz	FW443	SAMN17836245
	Boechera rectissima (Greene) Al-Shehbaz	JB274	SAMN17836240
	Boechera retrofracta (Graham) Á. Löve & D. Löve	FW562	SAMN17836241
	Boechera schistacea (Rollins) Dorn	LA474	SAMN17836242
	Boechera shevockii Windham & Al-Shehbaz	FW757	SAMN17836243
	Boechera suffrutescens (S. Watson) Dorn	JB967	SAMN17836244
	Yosemitea repanda (S. Watson) P. J. Alexander & Windham	JB171	SAMN17836247
3	*Diptychocarpus strictus* Trautv.	S0673	SAMN17103305
	Draba nuda (Bél.) Al-Shehbaz & M. Koch	S0658	SAMN17103302
	Heliophila diffusa DC.	S0807	SAMN17103309
	Heliophila elata Sond.	S0797	SAMN17103308
	Heliophila linearis DC.	S0816	SAMN17103310
	Heliophila suavissima Burch. ex DC.	S0775	SAMN17103306
	Morettia canescens Boiss.	S0791	SAMN17103307
	Notoceras bicorne Amo	S0642	SAMN17103301
	Rorippa sylvestris (L.) Besser	S0672	SAMN17103304
	Rytidocarpus moricandioides Coss.	S0668	SAMN17103303

Note: NCBI SRA ID = National Center for Biotechnology Information Sequence Read Archive identification number.
[a]Abbreviations that link vials of gDNA to specific DNA samples and genomic libraries.

data sets with 15 or 16 herbarium-derived libraries per hybridization, with 10 samples randomly selected for set 3 (Table 1).

In the Bailey lab, the genomic libraries were generated using the NEBNext Ultra II FS kit (New England Biolabs, Ipswich, Massachusetts, USA). All library steps followed the production manual (E7805L kit, version 5.0), with a fragmentation time of 5–10 min and six (set 1) or seven (set 2) cycles of PCR amplification. New England Biolabs single- and dual-index adapters were applied to sets 1 and 2, respectively. Libraries generated at Naturalis (set 3) used the same library kit and protocol, but with a 1-min fragmentation using sonication in an M220 Focused-ultrasonicator (Covaris, Woburn, Massachusetts, USA), indexing with IDT 10 primers (Integrated DNA Technologies, Coralville, Iowa, USA), and nine cycles of PCR amplification.

Target enrichment and sequencing

We employed the Brassicaceae-specific bait set developed by Nikolov et al. (2019), along with Angiosperms353 (Johnson et al., 2019), both of which are available as Arbor Biosciences "myBaits" kits (Arbor Biosciences, Ann Arbor, Michigan, USA; https://arborbiosci.com/genomics/targeted-sequencing/mybaits/). These kits have just 30 loci in common. Staff at Arbor Biosciences (Brian Brunelle, personal communication) noted that combined bait-set approaches had been successfully applied and that the logical starting point for exploring a mixture of baits was to maintain the relative representation of each set in the hybridization reaction. The Angiosperms353 and Nikolov1827 kits include 80,000 and 40,000 probes, respectively. To maintain twice as many Angiosperms353 probes, the standard 5.5 μL of a single bait set used in the myBaits hybridization protocol ("Hybridization Capture for Targeted NGS" protocol, version 4.01 [April 2018]) was replaced with a 2 : 1 (v/v) mixture of Angiosperms353 : Nikolov1827 baits. All other hybridization steps followed the myBaits protocol with the 0.2-mL plate format and four washing steps.

For the Bailey lab enrichments, sets 1 and 2 targeted the equal inclusion of libraries based on mass (Qubit dsDNA HS Assay Kit; Thermo Fisher Scientific, Waltham, Massachusetts, USA), with 100 ng and 20 ng DNA per library, respectively. For set 2, the libraries were combined based on similar size distributions (400–450 bp, 450–500 bp, 500–550 bp, or >600 bp), as determined using a 0.7% agarose gel. The post-hybridization libraries were subjected to 19 cycles of PCR with the KAPA HiFi amplification kit (Roche Sequencing, Pleasanton, California, USA) and IDT xGen amplification primers. The final post-amplification cleanups were performed using ABM beads (Applied Biological Materials, Richmond, British Columbia, Canada). Quality control checks, the combining of enriched pools (set 2 only), and sequencing were performed by Novogene (Beijing, China). Set 1 was sequenced using an Illumina 150-bp paired-end (PE) MiSeq Micro (Illumina, San Diego, California, USA; targeting approximately 2 million reads/sample), while set 2 ran with 96 multiplexed samples on a lane of an Illumina HiSeq4000 (150 bp PE, targeting approximately 3 million reads/sample). A protocol for the hybridization reactions is provided in Appendix 2.

The Naturalis-derived enrichments (set 3) included 15.6 ng (in hybridization reactions with a total of 250 ng) or 33.3 ng (reactions with 500 ng) of each library in the target mixture. The DNA concentrations from libraries included in this study ranged between

1.0 and 25.9 ng/μL. Libraries were pooled into reactions based on the similarity of the fragment length distributions, as measured on a Fragment Analyzer with an HS Small Fragment DNF-477 kit (Agilent Technologies, Santa Clara, California, USA). The post-hybridization library was subjected to 20 cycles (plus five additional cycles for library S0775) of PCR with a KAPA HiFi HotStart Library Amp Kit (Roche Sequencing) and the general amplification primers (matching IDT i7 and i5 index primers), followed by a bead cleanup (Macherey-Nagel, Düren, Germany). The amplified libraries were sequenced as 150 bp PEs using an Illumina NovaSeq 6000 at BaseClear (Leiden, The Netherlands), with a targeted sequence coverage of 325×. All raw data were uploaded to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA; BioProjects PRJNA678873 and PRJNA700668).

Data analysis

The raw reads were downloaded onto a Supermicro H8QG6 server with 64 AMD 6272 processors and 512 GB of RAM. Their analyses primarily employed SuperDeduper (version 1.3.0, https://github.com/s4hts/HTStream) for tests of PCR duplicate removal, Trimmomatic (version 0.39; Bolger et al., 2014) for adapter removal and quality trimming (with the arguments ILLUMINACLIP:../TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50), and HybPiper (version 1.3.1; installed from https://github.com/mossmatters/HybPiper.git) for locus mapping and reconstruction (applying the script "reads_first.py") and the generation of comparative statistics (applying scripts "get_seq_lengths.py" and "hybpiper_stats.py"). HybPiper is a wrapper that utilizes a variety of publicly available tools. Our analyses utilized elements that applied BWA version 0.7.12-r1039 (Li and Durbin, 2010) for mapping reads to the target sets, Biopython (Cock et al., 2009) for handling reads, SAMtools version 1.9 (Li et al., 2009) for sorting reads, SPAdes version 3.13.0 (Bankevich et al., 2012) for the de novo assembly of loci, and GNU Parallel (Tange, 2011) for multithreading on the server. The target locus files were the Angiosperms353 set (https://github.com/mossmatters/Angiosperms353/blob/master/Angiosperms353_targetSequences.fasta) and the Nikolov et al. (2019) set obtained directly from the author (L. A. Nikolov, personal communication). Scripts for the applied informatics are available from GitHub (https://github.com/cdb3ny/combined_enrichment_probes). In short, "reads_first.py" generated the de novo reconstruction of each locus while "get_seq_lengths.py" provided the sequence lengths for the downstream statistical summaries that were generated through "hybpiper_stats.py". Default parameters were applied in all cases. The reported "percent enrichments" represent the number of reads from a sample mapping to the target sequences relative to the total number of reads for that sample ([no. of mapped reads] / [no. of total reads] × 100). Given that the target sequences represent less than 1% of the total genome size for these taxa, this simple measure denotes the relative enrichment in the raw reads while providing a fairly accurate (± <1%) representation of the target enrichment component in relation to the general genome representation in the recovered reads.

Results

Three pipelines were applied to each of the sequenced sets of enriched libraries and target locus sets. These included running all raw paired data through: (1) SuperDeduper, Trimmomatic, and recovered PE data only through HybPiper; (2) Trimmomatic and recovered PE data only through HybPiper; or (3) Trimmomatic and all recovered reads (PE and single-end [SE]) through HybPiper. A summary of key results is presented in Table 2. We also report the percentage of cleaned reads mapping to the target set and the percentage of loci recovered with at least 75% sequence length as a primary measure of sequence enrichment and locus recovery for the samples within each data set (Appendix 3).

Whenever PCR deduplication was applied as the first step in the pipeline, we observed a considerable loss of reads recovered and subsequently available for mapping to loci (Appendix 3). This was especially pronounced for samples with low levels of recovered raw reads (e.g., <1 million), highlighting problems with including a PCR deduplication step. This issue was noted by the author of HybPiper, resulting in his not recommending the use of deduplication when applying the pipeline (M. G. Johnson, Texas Tech University, personal communication). The PCR deduplication–derived results are not discussed or presented further.

The two remaining implementations, both excluding deduplication, produced similar results. Unsurprisingly, the use of all reads (PE and SE) recovered a few additional loci (Appendix 3). The utility of adding SE data to the PE data was particularly pronounced with the MiSeq results, which are known to generate lower-quality reverse reads under some circumstances (M. G. Johnson, personal communication). Thus, the MiSeq data retained more SE forward read–only sequences than SE reverse reads after quality trimming.

TABLE 2. Summary of the enrichment and locus reconstruction results for assemblies based on all (paired-end and single-end) trimmed reads without PCR deduplication. A locus was considered "recovered" from a sample when at least 75% of its read length was reconstructed.

	Data set		
Statistic	**Set 1**	**Set 2**	**Set 3**
No. of samples included	6	10	10
Samples in the hybridization reaction	6	23–26	15–16
Raw reads per sample (mean (range))	2.1 M[a] (100,000–6.25 M)	3 M (1.9 M–3.9 M)	1.6 M[b] (512,000–4.3 M)
Mean % of Angiosperms353 enrichment	31	24.80	24.50
Mean % of Angiosperms353 targets recovered	70	88	84
Mean Angiosperms353 theoretical read coverage	338	343	219
Mean % of Nikolov1827 enrichment	59	43	35
Mean % of Nikolov1827 targets recovered	75	94	79
Mean Nikolov1827 theoretical read coverage	180	167	88

Note: M = million.
[a]Two of the six samples had fewer than 500,000 reads.
[b]Two of 10 samples had fewer than 1 M reads.

Even so, the difference in the percentage of loci recovered was minimal (Appendix 3).

The most important take home message from either the PE-only or PE+SE results is the high degree of sequence enrichment achieved for both groups of target loci (Table 2, Appendix 3). From this point, we use the results from the PE+SE analyses (Table 2) to discuss the potential for mixing probe sets in one hybridization reaction. Considering each of the three example data sets, the average percent of cleaned reads mapping to the Angiosperms353 and Nikolov1827 targets were 24–31% and 35–59%, respectively. For some samples, 90% of cleaned reads mapped to the target sequences. These high levels of enrichment were most pronounced in set 1, which included just six libraries. A modest decrease in enrichment efficiency was observed for sets 2 and 3, which each included at least 15 samples per hybridization reaction (Table 2, Appendix 3).

The Angiosperms353 and Nikolov1827 bait sets correspond to 260 kbp and 940 kbp of exon-derived data, respectively; thus, an increased fraction of reads mapping to the Nikolov1827 targets (Table 2) is important for reconstructing a larger portion of genome space. Using the genomic portion represented by each probe set, the total number of reads mapped per sample, and an estimated 145 bp length for the average retained cleaned reads, we calculated an average theoretical coverage across loci ([no. of reads × 145 bp] / [bp of genomic space of each target file per sample]) (Table 2, Appendix 3). The theoretical coverage of Angiosperms353 loci was 1.8–2.5 times greater than that for the Nikolov1827 probe set (Table 2); nonetheless, the percentage recovery of loci was similar (differing by less than 5% within data sets).

Hale et al. (2020) suggested that between 300,000 and 1 million reads represented a reasonable target for the 300 bp PE data generated by a MiSeq run for the high recovery of Angiosperms353 loci. Our data are 150 bp PE, making a corresponding estimate for our data of 600,000 to 2 million reads per sample, which fits well with the generally high recovery of loci from both probe sets (Appendix 3). Our results from the simultaneous hybridization of two different probe sets were supportive of the 2 : 1 Angiosperms353 : Nikolov1827 bait ratio, without requiring a greater sequencing depth than one might have applied for a single bait set. We feel that the simultaneous enrichment, using two different groups of probes, is strikingly balanced considering the mixture of up to 26 libraries in the enrichments and the fact that post-enrichment libraries were subjected to ≥19 cycles of PCR.

We consider the results presented here to be a promising outcome, one that is currently guiding the generation of new data for Brassicaceae. Thus far, the larger-scale preliminary results from those data (Bailey et al. and Hendriks et al., unpublished data) are similar to those presented here. Nonetheless, when choosing bait by taxon combinations with lower hybridization efficiency, adjustments may be needed in both the bait ratio and the depth of sequencing required for the recovery of a high percentage of loci from each target set.

CONCLUSIONS

The high levels of enrichment and locus reconstruction for two different sets of loci, obtained through one enrichment step, demonstrate that target-enrichment projects can be easily expanded to include a greater portion of genome space. Prior studies suggest that hybridization efficiency can range from around 15% to 80% (Hale et al., 2020). The high degree of hybridization efficiency observed here, ranging up to 90% of cleaned reads mapping to one target file or the other, are likely the outcome of the high sequence similarity between our Boechereae samples and other Brassicaceae samples and between the orthologs used in the design of both sets of probes, which drew heavily on the *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae) genome. In the case of Angiosperms353, for which 15 or fewer target instances were selected from across the angiosperms for each of the target loci using *k*-medoids clustering, a further three instances were added from the *A. thaliana*, *Oryza sativa* L., and *Amborella trichopoda* Baill. genomes, rendering the probe set especially effective in their respective families. This ensures a fair comparison of probe performance (in terms of reads on target) as presented here. When implementing a similar approach using probe mixes whose design lacked a closely matching genome for the study group, lower enrichment efficiencies are likely. It will be prudent to invest in similar preliminary studies early in the project. If an imbalance in recovered loci is detected, adjustments in the ratio of baits can easily be made.

This study illustrates the potential ease with which new target capture data can be simultaneously generated for multiple probe sets, with relatively little extra cost or work per sample. Our robust results suggest that researchers interested in combining multiple probe sets (e.g., a universal plus lineage-specific, multiple universal, or even multiple lineage-specific sets) can achieve this in one step. The successful simultaneous application of bait sets will hopefully be adopted in other projects to maximize the generation of useful data for wide-ranging investigations in evolutionary biology. As the availability of bait sets increases and the cost of sequencing continues to decline, there is no obvious reason to limit the combination of probes to just two sets. It should be possible to mix multiple bait sets (e.g., universal, lineage-specific, or gene family [e.g., nodulation or others]), perhaps even including baits that target different taxa in shared tissues (e.g., endosymbionts and parasites). It is hoped that these practical findings will relieve researchers of some difficult decision-making, ultimately leading to the generation of a broader spectrum of loci serving the interests of our research communities in terms of generating data with wider downstream applications.

ACKNOWLEDGMENTS

The authors are grateful to Matt Johnson (Texas Tech University) for advice on target enrichment and elements of the informatics. We thank Brian Brunelle (Arbor Biosciences) for advice on mixing probes for enrichments and Barbara Neuffer (University of Osnabrück) for help with sampling from the University of Osnabrück herbarium. This work was supported by the Czech Science Foundation (project no. 21-06839S); a grant from the Deutsche Forschungsgemeinschaft (MU1137/17-1; Reaching for the sky: exploring massive convergent evolution towards woodiness in Brassicaceae); and the Czech Ministry of Education, Youth, and Sports within the program INTER-EXCELLENCE (project no. LTAUSA17002).

AUTHOR CONTRIBUTIONS

All authors contributed to the design and writing and/or revision of the manuscript. K.P.H., T.M., and A.H.H. isolated the gDNA. C.D.B.

and K.P.H. generated the libraries and enrichment data. C.D.B., K.P.H., N.M.H., and E.L. conducted analyses related to the project. C.D.B., W.J.B., F.L., and K.P.H. wrote the primary body of the manuscript. All authors agreed with the final version of the manuscript and its submission for publication.

DATA AVAILABILITY

All raw data generated as part of the project are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (BioProjects PRJNA678873 and PRJNA700668).

LITERATURE CITED

Alexander, P. J., G. Rajanikanth, C. Bacon, and C. D. Bailey. 2006. Rapid inexpensive recovery of high quality plant DNA using a reciprocating saw and silica-based columns. *Molecular Ecology Notes* 7: 5–9.

Baker, W., A. Barker, L. Botigué, S. Dodsworth, W. Eiserhardt, E. Gaya, J. Kim, et al. 2017. PAFTOL First Annual Report. Available from: https://www.kew.org/sites/default/files/2019-07/PAFTOL%201st%20annual%20report.pdf [accessed 21 April 2021].

Baker, W. J., P. Bailey, V. Barber, A. Barker, S. Bellot, D. Bishop, L. R. Botigué, et al. 2021. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology*: syab035. https://doi.org/10.1093/sysbio/syab035

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Buddenhagen, C., A. R. Lemmon, E. M. Lemmon, J. Bruhl, J. Cappa, W. L. Clement, M. J. Donoghue, et al. 2016. Anchored phylogenomics of angiosperms I: Assessing the robustness of phylogenetic estimates. bioRxiv 086298 [Preprint]. Posted 28 November 2016 [accessed 21 April 2021]. Available from: https://doi.org/10.1101/086298.

Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, et al. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422–1423.

Couvreur, T. L. P., A. J. Helmstetter, E. J. M. Koenen, K. Bethune, R. D. Brandão, S. A. Little, H. Sauquet, and R. H. J. Erkens. 2019. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. *Frontiers in Plant Science* 9: 1941.

Dodsworth, S., L. Pokorny, M. G. Johnson, J. T. Kim, O. Maurin, N. J. Wickett, F. Forest, and W. J. Baker. 2019. Hyb-seq for flowering plant systematics. *Trends in Plant Science* 24: 887–891.

Folk, R. A., J. R. Mandel, and J. V. Freudenstein. 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: A phylogenomic example from *Heuchera* (Saxifragaceae). *Applications in Plant Sciences* 3: 1500039.

Gardiner, L.-J., T. Brabbs, A. Akhunov, K. Jordan, H. Budak, T. Richmond, S. Singh, et al. 2019. Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. *GigaScience* 8: giz018.

Hale, H., E. M. Gardner, J. Viruel, L. Pokorny, and M. G. Johnson. 2020. Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Applications in Plant Sciences* 8: e11337.

Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68: 594–606.

Koenen, E. J. M., D. I. Ojeda, R. Steeves, J. Migliore, F. T. Bakker, J. J. Wieringa, C. Kidner, et al. 2020. Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytologist* 225: 1355–1369.

Larridon, I., T. Villaverde, A. R. Zuntini, L. Pokorny, G. E. Brewer, N. Epitawalage, I. Fairlie, et al. 2020. Tackling rapid radiations with targeted sequencing. *Frontiers in Plant Science* 10: 1655.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589–595.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 27: 2078–2079.

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.

Mitchell, N., P. O. Lewis, E. M. Lemmon, A. R. Lemmon, and K. E. Holsinger. 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *American Journal of Botany* 104: 102–115.

Nikolov, L. A., P. Shushkov, B. Nevado, X. Gan, I. A. Al-Shehbaz, D. Filatov, C. D. Bailey, and M. Tsiantis. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist* 222: 1638–1651.

Slimp, M., L. D. Williams, H. Hale, and M. G. Johnson. 2020. On the potential of Angiosperms353 for population genomics. *Applications in Plant Sciences* 9(7): e11419.

Soto Gomez, M., L. Pokorny, M. B. Kantar, F. Forest, I. J. Leitch, B. Gravendeel, P. Wilkin, et al. 2019. A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). *Applications in Plant Sciences* 7: e11254.

Tange, O. 2011. GNU Parallel: The command-line power tool. *USENIX Magazine* 36: 42–47.

Thiers, B. 2021. (continuously updated). Index Herbariorum. Website http://sweetgum.nybg.org/science/ih/ [accessed 21 May 2021].

Vatanparast, M., A. Powell, J. J. Doyle, and A. N. Egan. 2018. Targeting legume loci: A comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Applications in Plant Sciences* 6: e1036.

Wanke, S., C. Granados Mendoza, S. Müller, A. Paizanni Guillén, C. Neinhuis, A. R. Lemmon, E. M. Lemmon, and M.-S. Samain. 2017. Recalcitrant deep and shallow nodes in *Aristolochia* (Aristolochiaceae) illuminated using anchored hybrid enrichment. *Molecular Phylogenetics and Evolution* 117: 111–123.

Weitemier, K., S. C. K. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: 1400042.

APPENDIX 1. Sample voucher information.

Species	Voucher specimen, collection no., herbarium[a]	Collection locality	Geographic coordinates
Boechera sanluisensis	P.J. Alexander 599B, NMC	Carson National Forest, ±1.75 miles WNW of Tres Piedras, ±0.15 miles N of US Highway 64, Rio Ariba County, New Mexico, USA	36.6544, –105.9974
Halimolobos jaegeri	Erik Schranz, 1074, personal collection	USA	NA
Sandbergia whitedii	Erik Schranz, 1080, personal collection	USA	NA
Boechera paupercula	Alexander 1107, DUKE	Tulare County, California, USA	36.4003, –118.5727
Boechera platysperma s.l.	Howden 12, UC	Alpine County, California, USA	38.4704, –119.9967
Boechera pendulina	Windham et al. 3709a, DUKE	Clark County, Nevada, USA	36.2609, –115.6086
Boechera rectissima	Alexander 1026, DUKE	Fresno County, California, USA	37.0542, –119.1551
Boechera retrofracta	Soper 5470, CAN	Bruce County, Ontario, Canada	44.9323, –81.1343
Boechera schistacea	Windham & Allphin 4307, DUKE	Uinta County, Wyoming, USA	41.0756, –110.3806
Boechera shevockii	Shevock 10098, GH	Tulare County, California, USA	36.0210, –118.4167
Boechera suffrutescens	Cusick s.n., ORE	Baker County, Oregon, USA	44.9718, –116.862
Boechera platysperma	Howden 12, UC	Alpine County, California, USA	38.4704, –119.9967
Boechera pendulina	Windham et al. 4435, DUKE	Fremont County, Wyoming, USA	42.4302, –109.0342
Cusickiella douglasii	M.D. Windham & L. Allphin 3362, NMC	Box Elder County, Utah, USA	41.7675, –113.9419
Yosemitea repanda	Alexander et al. 845f, DUKE	Inyo County, California, USA	37.209, –118.6124
Diptychocarpus strictus	TUH35369, TUH	60 km away from Delijan from Esfahan, Esfahan Province, Iran	33.017, –51.567
Draba nuda	Solomon et al. 21443, Gomez-Campo Collection	Tajikistan	NA
Heliophila diffusa	NGS311, NBG	Clanwilliam, Cederberg, Western Cape, South Africa. Road to Pakhuis Pass, at Leipoldt's Grave	32.135, –18.989
Heliophila elata	Mummenhoff & Ramdhani 65, personal collection	South Africa. Along road 364 from Butterkloof Pass to Clanwilliam, 200 m W of Elizabethsfontein junction	NA
Heliophila linearis	Linder P14, personal collection	Geelbek Lagoon, Darling District, South Africa	NA
Heliophila suavissima	Clark et al. 135, GRA	Farm Puttersvlei 190, Karoo National Park (Beaufort West), Western Cape, South Africa	32.264, –22.499
Morettia canescens	Staudinger, 13669, OSBU	Jbel Sarho, Zagora, Morocco	NA
Notoceras bicorne	Neuffer, 19678, OSBU	Fermes, Lanzarote, Canary Islands, Spain	28.883, –13.750
Rorippa sylvestris	Neuffer, Hurka, Friesen, 18646, OSBU	Bezirk Smolenskoje, Altaijski Kraij, Siberia, Russia. About 35 km south of Bijsk and 10 km south of Smolenskoje along the Pestschanaja river	37.478, –71.603
Rytidocarpus moricandioides	GCC0708-67, Gomez-Campo Collection	Botanical Garden Paris, France	NA

Note: NA = not available.
[a] Herbarium abbreviations are per Thiers et al. (2021).

APPENDIX 2. Full wet-lab protocol for the hybridization reactions used in this study. The following procedure is a slight modification from the Arbor Biosciences "Hybridization Capture for Targeted NGS" protocol, version 4.01 (April 2018; available from: https://arborbiosci.com/wp-content/uploads/2018/04/myBaits-Manual-v4.pdf). Arbor Biosciences has granted permission for the reprint of their elements herein. The significantly modified or added elements are highlighted in italics.

Part 1: Hybridization

A. Materials required (*when removing reagents from freezer/refrigerator, only remove what is needed for your reactions*). All reagent names refer to materials provided in the Arbor Biosciences myBaits kits.

- Hyb reagents (Boxes 1 [4°C] and 2 [–20°C])
- Block reagents (Box 2)
- Baits (Box 3 [–80°C; aliquot to 12 µL]) – *keep on ice*
- Sequencing libraries to be enriched, in a final volume of 7 µL per reaction
- 1.7-mL nuclease-free low-bind tubes (×2)
- 0.2-mL low-bind tubes with individual caps (×2 per reaction)
- Pipettors and tips (20-µL multichannel pipette)
- SpeedVac
- *Stoichiometrically combined libraries of similar size. Each combined set of libraries (ca. 24 libraries per Hyb-Seq reaction) will be run through as one hybridization reaction. They should contain 100–500 ng total DNA. Small libraries (<300 bp including the 140 bp of adapters) and larger libraries (ca. 350–700 bp including adapters) should be pooled and used in the separate hybridization reactions. Once combined, use the SpeedVac to concentrate the set down to a total volume of 7 µL.*

B. Hybridization mix setup

1. Thaw the Hyb reagents (Boxes 1 and 2), vortex to homogenize, and centrifuge briefly. (Note: If Hyb N and/or Hyb S have visible precipitate after thawing, heat them to 60°C and vortex until the precipitate dissolves.)
2. *For the baits, combine different probe sets based on the number of probes per set. In our case, it was a 2 : 1 mixture of Angiosperms353 (80,000 probes) to Nikolov1827 (40,000 probes).*

3. Assemble the Hybridization Mix in a 0.2-mL low-bind tube for fewer than eight reactions or a 1.5-mL tube for larger numbers of reactions.

Component	Amount per reaction (µL)	Amount for four reactions (µL)
Hyb N	9.25	37
Hyb D	3.5	14
Hyb S	0.5	2
Hyb R	1.25	5
Baits	5.5	22
TOTAL	20	80

Note: The introduction of Hyb S will cause cloudiness; the mixture will clarify after step 3.

4. Incubate the Hybridization Mix at 60°C for 10 min in the heat block and heated lid, vortexing occasionally to collect condensed evaporate from the tube lid. Remove the mix from the heat block, briefly spin down, and allow to sit at room temperature for 5 min.
5. For each capture reaction, aliquot 18.5 µL of Hybridization Mix into a 0.2-mL tube. These are hereafter referred to as HYB tubes.

C. Blockers Mix setup

1. Assemble the Blockers Mix in a 0.2-mL no-bind tube and mix by pipetting.

Component	Amount per reaction (µL)	Amount for four reactions (µL)
Block A	0.5	2
Block C	2.5	10
Block O	2.5	10
TOTAL	5.5	22

2. For each capture reaction, aliquot 5 µL of Blockers Mix into a 0.2-mL low-bind tube.
3. Add 7 µL of library (100–500 ng recommended) to each Blockers Mix aliquot and mix by pipetting. The resulting mix will be referred as LIB reactions.

D. Reaction assembly

Thermal program for thermal cycler (using heated lid)

Step	Temperature	Time
1	95°C	5 min
2	Hybridization temperature (65°C)	5 min
3	Hybridization temperature (65°C)	∞

1. Put the LIBs in the thermal cycler, close the lid, and start the thermal program.
2. Once the cycler reaches the hybridization temperature during step 2, pause the program, put the HYBs in the thermal cycler, close the lid, and resume the program.
3. After step 2 of the program is complete, leaving all tubes in the thermal cycler, pipette 18 µL of each HYB into each LIB using a multichannel pipette. Gently homogenize by pipetting up and down five times.
4. Dispose of the HYB tubes. Briefly spin down the LIBs, return to the thermal cycler, close the lid, and allow the reactions to incubate at the hybridization temperature (using heated lid) for your chosen time. For this study, we used 24 h.

Part 2: Bind and wash (cleanup)

A. Begin assembly of materials at least 90 min before the end of the hybridization reaction.

B. Materials required
Note: Bring the solutions to room temperature prior to use. Warm gently to dissolve precipitate if necessary.

- Hyb S
- Binding Buffer
- Wash Buffer
- Arbor Beads (Streptavidin bound)
- Nuclease-free sterile water (up to 900 µL per cleanup)
- 10 mM Tris-Cl, 0.05% TWEEN-20 solution (pH 8.0–8.5)
- Magnetic particle concentrator(s) (MPC) for 0.2-mL PCR strips/plates
- Incubator and water bath set at 65°C
- 50-mL nuclease-free tube

C. Wash Buffer X preparation
1. Thaw and thoroughly homogenize the Wash Buffer and Hyb S prior to aliquoting in order to dissolve any visible precipitate; warm slightly if necessary.
2. For each enrichment reaction, combine the following in a 1.5-mL nuclease-free sterile tube, vortex, and label as "Wash Buffer X."

Reagents	Amount per reaction (µL)	Amount for four reactions (µL)
Hyb S	9	36
NF water	900	3600
Wash buffer	227	908

D. Bead preparation

Note: Prepare beads immediately prior to use.

1. For each capture reaction, aliquot 30 µL of beads into a 1.7-mL low-bind tube.
2. Pellet the beads in the MPC until the suspension is clear (1–2 min). Leaving the tubes on the magnet, remove and discard the supernatant without disturbing the beads.
3. Add 200 µL Binding Buffer to each bead aliquot. Vortex to resuspend the beads and centrifuge briefly. Pellet in the MPC, remove, and discard the supernatant without disturbing the beads.
4. Repeat Step 3 twice more for a total of three washes.
5. Resuspend each bead aliquot in 70 µL Binding Buffer.
6. Transfer the bead aliquot to 0.2-mL plate tubes for 96-well processing with MPC-style magnets. Other options are available here (see the original myBaits protocol).

E. Binding beads and hybrids

1. Heat the bead aliquots (sealed in their 0.2-mL well) to the hybridization temperature (65°C) for at least 2 min in thermal cycler.

2. Transfer each capture reaction to the heated bead aliquots and mix by pipetting. Seal the tops to the tubes (strip cap lids work well).
3. Incubate the libraries+beads on the thermal cycler for 5 min. Agitate at the 2.5-min mark by pipetting (briefly centrifuging to collect if necessary).
4. After 5 min, pellet the beads with the MPC until the solution is clear. Remove and discard the supernatant without disturbing the beads. Immediately move to the next step.

F. Bead washing

1. Remove samples from the MPC and add 180 µL warmed Wash Buffer X to the beads, mixing by pipetting. If necessary, briefly centrifuge to collect.
2. Incubate for 5 min at the hybridization temperature in the heat block or thermal cycler. Agitate at the 2.5-min mark via pipetting (briefly centrifuge if necessary).
3. Pellet the beads with the MPC and discard the supernatant without disturbing the bead portions.
4. Repeat steps 1–3 three times for the 0.2-mL format (four washes total). After the last wash and pelleting, remove as much fluid as possible without touching the bead pellet.

Part 3: Library resuspension and amplification

A. Materials required

- 10 mM Tris-Cl, 0.05% TWEEN-20 solution (pH 8.0–8.5)
- Reagents for library amplification using universal primers
- PCR purification system, solid-phase reversible immobilization (SPRI) beads

B. Enriched library resuspension

1. Add 30 µL of 10 mM Tris-Cl, 0.05% TWEEN-20 solution (pH 8.0–8.5) to the washed beads and thoroughly resuspend by pipetting.

Note: Beads can be frozen at –20°C if you are not moving on to amplification immediately.

C. Library amplification

1. For each sample, assemble the following PCR master mix:

Component	Final concentration	Amount per reaction (µL)
Nuclease-free water	—	8.75
2× KAPA HiFi HotStart ReadyMix	1×	25
IDT xGEN amp primers (20 µM)	500 nM	1.25
Enriched library (pellet the beads before pulling off the 15-µL aliquot)	—	15*
Total		50

*The remaining bead-bound library can be stored at –20°C for several months.

2. Amplify the reactions using the program below. *Note: the number of cycles needed can be highly variable and can be influenced by the sequencing provider's requirements and the sequencing platform. For our Illumina HiSeq4000 runs performed by Novogene, we used 14 initial cycles of PCR, then paused the PCR program at 4°C to quickly run a Qubit dsDNA HS estimate concentration, then ran additional cycles to reach our desired concentration (we targeted around 4–10 ng/µL, which gave us >2 µM libraries after SPRI cleanup).*

Use the calculated temperature setting:

Step	Temperature	Time
1	98°C	2 min
2	98°C	20 s
3	60°C	30 s
4	72°C	Length-dependent[a]
5	Return to step 2 for appropriate number of cycles[b]	
5	72°C	5 min
6	8°C	∞

[a]Extension time can be library-size dependent (when in doubt, a slightly longer time is acceptable). A mean length <500 bp requires 30 s, a mean of 500–700 bp requires 45 s, while a mean length >700 bp requires 1 min.
[b]The number of cycles needs to be empirically determined. For this study, we used 17 cycles total.

3. Purify the reaction using your preferred PCR cleanup (e.g., SPRI beads or Column cleanup). *In our hands both worked, but the SPRI cleanup recovered a higher amount of the DNA. The enriched libraries were then ready for sequencing.*

   a. SPRI bead purification using ABM magnetic beads (performed in 96-well format).
      - Add 90 µL of room-temperature and resuspended ABM SPRI beads to the 50-µL PCR reaction (1.8 SPRI : 1 PCR v/v).
      - Pipette up and down 10 times to mix and incubate at room temp for 5 min.
      - Place on the MPC until the beads have cleared from the solution (2–5 min typically).
      - Carefully remove and discard supernatant without taking up any beads. In this step, it may be hard not to accidentally pick up beads, so you can leave a bit of liquid behind if needed.
      - Keeping the tubes on the MPC, add 200 µL of freshly made 80% ethanol, incubate for 30 s, and remove and discard the supernatant. The beads are not as easily disturbed now and you can remove all liquid.
      - Repeat one more wash with 200 µL 80% ethanol.
      - Air dry beads for 1 min.
      - Remove the plate from the MPC and elute the DNA from beads with 30 µL of 0.1× TE (1× TE [10 mM Tris, 1 mM EDTA, pH 8] diluted 1 : 10). If the concentration is a concern, you could recover the DNA in a lesser volume of 0.1× TE.
      - Pellet the beads with MPC and transfer the newly suspended DNA into a clean tube.
      - Store at –20°C or –80°C.

APPENDIX 3. Results for the analyses of all three example data sets analyzed for both the Angiosperms353 (Angio353) and Nikolov1827 targets.

Set	Sample	Source and analysis pipeline	Target	No. of raw reads	No. of trimmed reads	No. of reads mapped	Fraction mapped to target	Loci with at least 75% of the target sequence length recovered	Theoretical coverage	Percentage of loci recovered with 75%
1	PJA244_S6	Bailey, SDD+T+PE	Angio353	109,024	23,768	7415	0.31	5	4.14	1.42
	PJA248_S5	Bailey, SDD+T+PE	Angio353	207,784	47,499	14,948	0.32	27	8.34	7.65
	PJA296A_S4	Bailey, SDD+T+PE	Angio353	6,255,118	1,944,653	570,887	0.29	323	318.38	91.50
	PJA370-A_S1	Bailey, SDD+T+PE	Angio353	1,946,754	664,426	194,666	0.29	296	108.56	83.85
	PJA370-B_S2	Bailey, SDD+T+PE	Angio353	2,031,530	623,004	185,564	0.30	291	103.49	82.44
	PJA370-C_S3	Bailey, SDD+T+PE	Angio353	2,315,704	684,511	205,267	0.30	294	114.48	83.29
	Averages			**2,144,319.00**	**664,643.50**	**196,457.83**	**0.30**	**206.00**	**109.56**	**58.36**
	PJA244_S6	Bailey, T+PE	Angio353	109,024	64,581	20,671	0.32	53	11.53	15.01
	PJA248_S5	Bailey, T+PE	Angio353	207,784	129,525	41,656	0.32	122	23.23	34.56
	PJA296A_S4	Bailey, T+PE	Angio353	6,255,118	5,304,171	1,661,275	0.31	327	926.48	92.63
	PJA370-A_S1	Bailey, T+PE	Angio353	1,946,754	1,713,693	514,130	0.30	317	286.73	89.80
	PJA370-B_S2	Bailey, T+PE	Angio353	2,031,530	1,657,000	504,347	0.30	316	281.27	89.52
	PJA370-C_S3	Bailey, T+PE	Angio353	2,315,704	1,845,307	564,851	0.31	318	315.01	90.08
	Averages			**2,144,319.00**	**1,785,712.83**	**551,155.00**	**0.31**	**242.17**	**307.37**	**68.60**
	PJA244_S6	Bailey, T+PE+SE	Angio353	109,024	86,534	26,734	0.31	60	14.91	17.00
	PJA248_S5	Bailey, T+PE+SE	Angio353	207,784	168,751	52,743	0.31	137	29.41	38.81
	PJA296A_S4	Bailey, T+PE+SE	Angio353	6,255,118	5,783,880	1,811,539	0.31	327	1010.28	92.63
	PJA370-A_S1	Bailey, T+PE+SE	Angio353	1,946,754	1,839,458	554,773	0.30	320	309.39	90.65
	PJA370-B_S2	Bailey, T+PE+SE	Angio353	2,031,530	1,850,926	563,495	0.30	319	314.26	90.37
	PJA370-C_S3	Bailey, T+PE+SE	Angio353	2,315,704	2,086,092	636,019	0.31	319	354.70	90.37
	Averages			**2,144,319.00**	**1,969,273.50**	**607,550.50**	**0.31**	**247.00**	**338.83**	**69.97**
	PJA244_S6	Bailey, SDD+T+PE	Nikolov1827	109,024	23,734	14,819	0.62	18	2.29	0.99
	PJA248_S5	Bailey, SDD+T+PE	Nikolov1827	207,784	47,349	28,081	0.59	122	4.33	6.68
	PJA296A_S4	Bailey, SDD+T+PE	Nikolov1827	6,255,118	1,942,309	1,123,814	0.58	1782	173.35	97.54
	PJA370-A_S1	Bailey, SDD+T+PE	Nikolov1827	1,946,754	661,998	380,937	0.58	1500	58.76	82.10
	PJA370-B_S2	Bailey, SDD+T+PE	Nikolov1827	2,031,530	621,380	362,952	0.58	1509	55.99	82.59
	PJA370-C_S3	Bailey, SDD+T+PE	Nikolov1827	2,315,704	683,370	403,143	0.59	1568	62.19	85.82
	Averages			**2,144,319.00**	**663,356.67**	**385,624.33**	**0.59**	**1083.17**	**59.48**	**59.29**
	PJA244_S6	Bailey, T+PE	Nikolov1827	109,024	64,490	40,887	0.63	318	6.31	17.41
	PJA248_S5	Bailey, T+PE	Nikolov1827	207,784	129,109	77,504	0.60	636	11.96	34.81
	PJA296A_S4	Bailey, T+PE	Nikolov1827	6,255,118	5,297,552	3,250,608	0.61	1813	501.42	99.23
	PJA370-A_S1	Bailey, T+PE	Nikolov1827	1,946,754	1,706,716	998,214	0.59	1754	153.98	96.00
	PJA370-B_S2	Bailey, T+PE	Nikolov1827	2,031,530	1,651,985	978,439	0.59	1741	150.93	95.29
	PJA370-C_S3	Bailey, T+PE	Nikolov1827	2,315,704	1,841,611	1,100,963	0.60	1758	169.83	96.22
	Averages			**2,144,319.00**	**1,781,910.50**	**1,074,435.83**	**0.60**	**1336.67**	**165.74**	**73.16**
	PJA244_S6	Bailey, T+PE+SE	Nikolov1827	109,024	86,310	52,477	0.61	407	8.09	22.28
	PJA248_S5	Bailey, T+PE+SE	Nikolov1827	207,784	167,817	97,136	0.58	758	14.98	41.49
	PJA296A_S4	Bailey, T+PE+SE	Nikolov1827	6,255,118	5,774,149	3,528,033	0.61	1812	544.22	99.18
	PJA370-A_S1	Bailey, T+PE+SE	Nikolov1827	1,946,754	1,826,576	1,059,310	0.58	1758	163.40	96.22
	PJA370-B_S2	Bailey, T+PE+SE	Nikolov1827	2,031,530	1,840,718	1,075,680	0.58	1742	165.93	95.35
	PJA370-C_S3	Bailey, T+PE+SE	Nikolov1827	2,315,704	2,077,958	1,225,445	0.59	1759	189.03	96.28
	Averages			**2,144,319.00**	**1,962,254.67**	**1,173,013.50**	**0.59**	**1372.67**	**180.94**	**75.13**
2	FW443	Bailey, SDD+T+PE	Angio353	1,973,768	173,176	51,923	0.3	106	28.96	30.03
	FW562	Bailey, SDD+T+PE	Angio353	3,873,080	1,185,052	117,724	0.099	276	65.65	78.19
	FW757	Bailey, SDD+T+PE	Angio353	1,916,118	284,850	73,289	0.257	217	40.87	61.47
	JB152	Bailey, SDD+T+PE	Angio353	2,225,116	687,086	119,150	0.173	289	66.45	81.87
	JB171	Bailey, SDD+T+PE	Angio353	2,786,144	795,811	125,353	0.158	271	69.91	76.77
	JB242	Bailey, SDD+T+PE	Angio353	2,875,092	953,137	205,360	0.215	312	114.53	88.39
	JB274	Bailey, SDD+T+PE	Angio353	3,896,534	1,459,263	195,111	0.134	306	108.81	86.69
	JB967	Bailey, SDD+T+PE	Angio353	3,486,402	327,501	94,990	0.29	258	52.98	73.09
	LA474	Bailey, SDD+T+PE	Angio353	3,933,178	517,705	138,434	0.267	303	77.20	85.84
	W4485	Bailey, SDD+T+PE	Angio353	3,744,160	376,946	110,826	0.294	284	61.81	80.45
	Averages			**3,070,959.2**	**676,052.7**	**123,216**	**0.2187**	**262.2**	**68.72**	**74.28**
	FW443	Bailey, T+PE	Angio353	1,973,768	999,416	309,278	0.309	229	172.48	64.87
	FW562	Bailey, T+PE	Angio353	3,873,080	3,113,683	549,261	0.176	322	306.32	91.22
	FW757	Bailey, T+PE	Angio353	1,916,118	1,194,875	335,728	0.281	300	187.23	84.99
	JB152	Bailey, T+PE	Angio353	2,225,116	1,690,398	408,754	0.242	322	227.96	91.22
	JB171	Bailey, T+PE	Angio353	2,786,144	1,797,747	382,992	0.213	315	213.59	89.24
	JB242	Bailey, T+PE	Angio353	2,875,092	2,367,833	535,158	0.226	329	298.45	93.20
	JB274	Bailey, T+PE	Angio353	3,896,534	3,072,867	523,538	0.17	314	291.97	88.95
	JB967	Bailey, T+PE	Angio353	3,486,402	1,669,004	530,602	0.318	314	295.91	88.95
	LA474	Bailey, T+PE	Angio353	3,933,178	2,715,849	812,459	0.299	322	453.10	91.22
	W4485	Bailey, T+PE	Angio353	3,744,160	2,031,630	642,825	0.316	318	358.50	90.08

(Continues)

APPENDIX 3. (Continued)

Set	Sample	Source and analysis pipeline	Target	No. of raw reads	No. of trimmed reads	No. of reads mapped	Fraction mapped to target	Loci with at least 75% of the target sequence length recovered	Theoretical coverage	Percentage of loci recovered with 75%
	Averages			**3,070,959.2**	**2,065,330.2**	**503,059.5**	**0.255**	**308.5**	**280.55**	**87.39**
	FW443	Bailey, T+PE+SE	Angio353	1,973,768	1,451,430	414,499	0.286	244	231.16	69.12
	FW562	Bailey, T+PE+SE	Angio353	3,873,080	3,476,610	601,370	0.173	322	335.38	91.22
	FW757	Bailey, T+PE+SE	Angio353	1,916,118	1,524,794	417,499	0.274	300	232.84	84.99
	JB152	Bailey, T+PE+SE	Angio353	2,225,116	1,938,915	461,653	0.238	324	257.46	91.78
	JB171	Bailey, T+PE+SE	Angio353	2,786,144	2,221,872	462,038	0.208	321	257.68	90.93
	JB242	Bailey, T+PE+SE	Angio353	2,875,092	2,622,550	582,996	0.222	325	325.13	92.07
	JB274	Bailey, T+PE+SE	Angio353	3,896,534	3,442,445	574,956	0.167	317	320.65	89.80
	JB967	Bailey, T+PE+SE	Angio353	3,486,402	2,525,497	773,699	0.306	314	431.49	88.95
	LA474	Bailey, T+PE+SE	Angio353	3,933,178	3,299,070	982,381	0.298	325	547.87	92.07
	W4485	Bailey, T+PE+SE	Angio353	3,744,160	2,843,615	887,927	0.312	321	495.19	90.93
	Averages			**3,070,959.20**	**2,534,679.80**	**615,901.80**	**0.25**	**311.30**	**343.48**	**88.19**
	FW443	Bailey, SDD+T+PE	Nikolov1827	3,744,160	171,161	82,023	0.479	515	12.65	28.19
	FW562	Bailey, SDD+T+PE	Nikolov1827	3,486,402	1,185,811	229,965	0.194	1244	35.47	68.09
	FW757	Bailey, SDD+T+PE	Nikolov1827	2,225,116	284,213	132,267	0.465	921	20.40	50.41
	JB152	Bailey, SDD+T+PE	Nikolov1827	1,973,768	686,733	205,985	0.3	1255	31.77	68.69
	JB171	Bailey, SDD+T+PE	Nikolov1827	2,786,144	795,638	219,696	0.276	1243	33.89	68.04
	JB242	Bailey, SDD+T+PE	Nikolov1827	3,933,178	955,986	363,096	0.38	1472	56.01	80.57
	JB274	Bailey, SDD+T+PE	Nikolov1827	2,875,092	1,467,415	344,830	0.235	1493	53.19	81.72
	JB967	Bailey, SDD+T+PE	Nikolov1827	3,873,080	326,346	155,101	0.475	1076	23.93	58.89
	LA474	Bailey, SDD+T+PE	Nikolov1827	3,896,534	517,028	231,371	0.448	1376	35.69	75.31
	W4485	Bailey, SDD+T+PE	Nikolov1827	1,916,118	375,095	185,612	0.495	1255	28.63	68.69
	Averages			**3,070,959.20**	**676,542.60**	**214,994.60**	**0.37**	**1185.00**	**33.16**	**64.86**
	FW443	Bailey, T+PE	Nikolov1827	3,744,160	987,931	484,589	0.491	1371	74.75	75.04
	FW562	Bailey, T+PE	Nikolov1827	3,486,402	3,118,779	1,075,245	0.345	1762	165.86	96.44
	FW757	Bailey, T+PE	Nikolov1827	2,225,116	1,190,913	605,666	0.509	1611	93.43	88.18
	JB152	Bailey, T+PE	Nikolov1827	1,973,768	1,688,260	698,311	0.414	1736	107.72	95.02
	JB171	Bailey, T+PE	Nikolov1827	2,786,144	1,795,223	664,251	0.37	1687	102.46	92.34
	JB242	Bailey, T+PE	Nikolov1827	3,933,178	2,374,936	936,372	0.394	1768	144.44	96.77
	JB274	Bailey, T+PE	Nikolov1827	2,875,092	3,098,713	914,547	0.295	1736	141.07	95.02
	JB967	Bailey, T+PE	Nikolov1827	3,873,080	1,661,855	843,329	0.507	1655	130.09	90.59
	LA474	Bailey, T+PE	Nikolov1827	3,896,534	2,710,952	1,326,774	0.489	1766	204.66	96.66
	W4485	Bailey, T+PE	Nikolov1827	1,916,118	2,018,357	1,056,289	0.523	1732	162.94	94.80
	Averages			**3,070,959.2**	**2,064,591.9**	**860,537.3**	**0.4337**	**1682.4**	**132.74**	**92.09**
	FW443	Bailey, T+PE+SE	Nikolov1827	1,973,768	1,439,504	663,559	0.461	1446	102.36	79.15
	FW562	Bailey, T+PE+SE	Nikolov1827	3,873,080	3,482,342	1,183,642	0.34	1768	182.58	96.77
	JB152	Bailey, T+PE+SE	Nikolov1827	2,225,116	1,936,510	791,047	0.408	1751	122.02	95.84
	JB171	Bailey, T+PE+SE	Nikolov1827	2,786,144	2,218,044	800,040	0.361	1715	123.41	93.87
	JB242	Bailey, T+PE+SE	Nikolov1827	2,875,092	2,629,961	1,026,062	0.39	1770	158.28	96.88
	JB274	Bailey, T+PE+SE	Nikolov1827	3,896,534	3,470,509	1,007,944	0.29	1749	155.48	95.73
	JB967	Bailey, T+PE+SE	Nikolov1827	3,486,402	2,518,124	1,255,748	0.499	1715	193.71	93.87
	LA474	Bailey, T+PE+SE	Nikolov1827	3,933,178	3,294,200	1,623,201	0.493	1777	250.39	97.26
	W4485	Bailey, T+PE+SE	Nikolov1827	3,744,160	2,820,469	1,442,458	0.511	1756	222.51	96.11
	Averages			**3,070,959.20**	**2,535,691.99**	**1,064,593.39**	**0.43**	**1714.67**	**164.22**	**93.85**
3	S0642	Naturalis, SDD+T+PE	Angio353	2,241,558	1,678,382	201,054	0.12	225	112.13	63.74
	S0658	Naturalis, SDD+T+PE	Angio353	2,341,630	1,283,758	298,272	0.232	284	166.34	80.45
	S0668	Naturalis, SDD+T+PE	Angio353	4,323,224	3,010,397	553,899	0.184	270	308.91	76.49
	S0672	Naturalis, SDD+T+PE	Angio353	1,005,866	715,945	181,459	0.253	267	101.20	75.64
	S0673	Naturalis, SDD+T+PE	Angio353	512,280	375,309	87,379	0.233	222	48.73	62.89
	S0775	Naturalis, SDD+T+PE	Angio353	1,855,986	89,089	18,077	0.203	20	10.08	5.67
	S0791	Naturalis, SDD+T+PE	Angio353	1,403,254	554,884	72,182	0.13	184	40.26	52.12
	S0797	Naturalis, SDD+T+PE	Angio353	1,266,122	497,669	78,779	0.158	189	43.93	53.54
	S0807	Naturalis, SDD+T+PE	Angio353	1,060,492	784,329	169,397	0.216	184	94.47	52.12
	S0816	Naturalis, SDD+T+PE	Angio353	648,334	469,121	103,943	0.222	202	57.97	57.22
	Averages			**1,665,874.6**	**945,888.3**	**176,444.1**	**0.1951**	**204.7**	**98.40**	**57.99**
	S0642	Naturalis, T+PE	Angio353	2,241,558	2,196,887	366,430	0.167	302	204.36	85.55
	S0658	Naturalis, T+PE	Angio353	2,341,630	2,308,263	630,132	0.273	333	351.42	94.33
	S0668	Naturalis, T+PE	Angio353	4,323,224	4,282,247	997,557	0.233	329	556.33	93.20
	S0672	Naturalis, T+PE	Angio353	1,005,866	996,920	295,027	0.296	324	164.53	91.78
	S0673	Naturalis, T+PE	Angio353	512,280	504,950	143,406	0.284	307	79.98	86.97
	S0775	Naturalis, T+PE	Angio353	1,855,986	1,832,692	513,304	0.28	201	286.27	56.94
	S0791	Naturalis, T+PE	Angio353	1,403,254	1,381,003	239,206	0.173	300	133.40	84.99
	S0797	Naturalis, T+PE	Angio353	1,266,122	1,236,899	266,014	0.215	293	148.35	83.00
	S0807	Naturalis, T+PE	Angio353	1,060,492	1,047,471	273,689	0.261	276	152.63	78.19

(Continues)

APPENDIX 3. (Continued)

Set	Sample	Source and analysis pipeline	Target	No. of raw reads	No. of trimmed reads	No. of reads mapped	Fraction mapped to target	Loci with at least 75% of the target sequence length recovered	Theoretical coverage	Percentage of loci recovered with 75%
	S0816	Naturalis, T+PE	Angio353	648,334	637,945	170,071	0.267	291	94.85	82.44
	Averages			**1,665,874.6**	**1,642,527.7**	**389,483.6**	**0.2449**	**295.6**	**217.21**	**83.74**
	S0642	Naturalis, T+PE+SE	Angio353	2,241,558	2,219,047	367,768	0.166	302	205.10	85.55
	S0658	Naturalis, T+PE+SE	Angio353	2,341,630	2,325,982	633,575	0.272	333	353.34	94.33
	S0668	Naturalis, T+PE+SE	Angio353	4,323,224	4,330,678	1,020,020	0.236	329	568.86	93.20
	S0672	Naturalis, T+PE+SE	Angio353	1,005,866	1,002,483	296,353	0.296	324	165.27	91.78
	S0673	Naturalis, T+PE+SE	Angio353	512,280	508,547	144,078	0.283	307	80.35	86.97
	S0775	Naturalis, T+PE+SE	Angio353	1,855,986	1,846,386	516,246	0.28	201	287.91	56.94
	S0791	Naturalis, T+PE+SE	Angio353	1,403,254	1,391,718	240,200	0.173	301	133.96	85.27
	S0797	Naturalis, T+PE+SE	Angio353	1,266,122	1,251,466	267,543	0.214	292	149.21	82.72
	S0807	Naturalis, T+PE+SE	Angio353	1,060,492	1,054,416	274,941	0.261	276	153.33	78.19
	S0816	Naturalis, T+PE+SE	Angio353	648,334	643,131	170,942	0.266	291	95.33	82.44
	Averages			**1,665,874.60**	**1,657,385.40**	**393,166.60**	**0.24**	**295.60**	**219.27**	**83.74**
	S0642	Naturalis, SDD+T+PE	Nikolov1827	2,241,558	1,676,774	340,122	0.203	1392	189.68	76.19
	S0658	Naturalis, SDD+T+PE	Nikolov1827	2,341,630	1,282,734	443,602	0.346	1572	247.39	86.04
	S0668	Naturalis, SDD+T+PE	Nikolov1827	4,323,224	2,990,730	1,024,972	0.343	1664	571.62	91.08
	S0672	Naturalis, SDD+T+PE	Nikolov1827	1,005,866	714,302	289,264	0.405	1547	161.32	84.67
	S0673	Naturalis, SDD+T+PE	Nikolov1827	512,280	375,424	143,552	0.382	1214	80.06	66.45
	S0775	Naturalis, SDD+T+PE	Nikolov1827	1,855,986	88,918	26,420	0.297	92	14.73	5.04
	S0791	Naturalis, SDD+T+PE	Nikolov1827	1,403,254	554,750	114,676	0.207	965	63.95	52.82
	S0797	Naturalis, SDD+T+PE	Nikolov1827	1,266,122	497,342	125,828	0.253	945	70.17	51.72
	S0807	Naturalis, SDD+T+PE	Nikolov1827	1,060,492	783,710	289,910	0.37	1324	161.68	72.47
	S0816	Naturalis, SDD+T+PE	Nikolov1827	648,334	468,910	167,812	0.358	1143	93.59	62.56
	Averages			**1,665,874.60**	**943,359.40**	**296,615.80**	**0.32**	**1185.80**	**165.42**	**64.90**
	S0642	Naturalis, T+PE	Nikolov1827	2,241,558	2,194,824	528,889	0.241	1524	81.58	83.42
	S0658	Naturalis, T+PE	Nikolov1827	2,341,630	2,306,466	864,500	0.375	1701	133.35	93.10
	S0668	Naturalis, T+PE	Nikolov1827	4,323,224	4,249,900	1,651,136	0.389	1758	254.70	96.22
	S0672	Naturalis, T+PE	Nikolov1827	1,005,866	994,699	423,727	0.426	1625	65.36	88.94
	S0673	Naturalis, T+PE	Nikolov1827	512,280	505,046	209,018	0.414	1341	32.24	73.40
	S0775	Naturalis, T+PE	Nikolov1827	1,855,986	1,829,251	640,249	0.35	841	98.76	46.03
	S0791	Naturalis, T+PE	Nikolov1827	1,403,254	1,380,874	324,424	0.235	1481	50.04	81.06
	S0797	Naturalis, T+PE	Nikolov1827	1,266,122	1,236,252	363,752	0.294	1455	56.11	79.64
	S0807	Naturalis, T+PE	Nikolov1827	1,060,492	1,046,600	420,874	0.402	1437	64.92	78.65
	S0816	Naturalis, T+PE	Nikolov1827	648,334	637,645	250,449	0.393	1313	38.63	71.87
	Averages			**1,665,874.60**	**1,638,155.70**	**567,701.80**	**0.35**	**1447.60**	**87.57**	**79.23**
	S0642	Naturalis, T+PE+SE	Nikolov1827	2,241,558	2,216,884	530,662	0.239	1527	81.86	83.58
	S0658	Naturalis, T+PE+SE	Nikolov1827	2,341,630	2,324,136	869,165	0.374	1701	134.07	93.10
	S0668	Naturalis, T+PE+SE	Nikolov1827	4,323,224	4,285,603	1,655,919	0.386	1760	255.43	96.33
	S0672	Naturalis, T+PE+SE	Nikolov1827	1,005,866	1,000,101	425,426	0.425	1625	65.62	88.94
	S0673	Naturalis, T+PE+SE	Nikolov1827	512,280	508,642	210,146	0.413	1342	32.42	73.45
	S0775	Naturalis, T+PE+SE	Nikolov1827	1,855,986	1,842,191	642,761	0.349	849	99.15	46.47
	S0791	Naturalis, T+PE+SE	Nikolov1827	1,403,254	1,391,540	325,898	0.234	1481	50.27	81.06
	S0797	Naturalis, T+PE+SE	Nikolov1827	1,266,122	1,250,748	365,776	0.292	1454	56.42	79.58
	S0807	Naturalis, T+PE+SE	Nikolov1827	1,060,492	1,053,332	422,400	0.401	1441	65.16	78.87
	S0816	Naturalis, T+PE+SE	Nikolov1827	648,334	642,818	251,766	0.392	1315	38.84	71.98
	Averages			**1,665,874.60**	**1,651,599.50**	**569,991.90**	**0.35**	**1449.50**	**87.92**	**79.34**

Note: PE = recovered paired-end-only data; SDD = SuperDeduper; SE = single end; T = Trimmomatic.