



# An updated and extended version of the Melastomataceae probe set for target capture

Léo-Paul M. J. Dagallier  | Fabián A. Michelangeli 

Institute of Systematic Botany, New York Botanical Garden, Bronx, New York, USA

## Correspondence

Léo-Paul M. J. Dagallier, Institute of Systematic Botany, New York Botanical Garden, 2900 Southern Boulevard, Bronx, New York 10458, USA.  
Email: [ldagallier@nybg.org](mailto:ldagallier@nybg.org)

## Abstract

**Premise:** A probe set was previously designed to target 384 nuclear loci in the Melastomataceae family; however, when trying to use it, we encountered several practical and conceptual problems, such as the presence of sequences in reverse complement, intronic regions with stop codons, and other issues. This raised concerns regarding the use of this probe set for sequence recovery in Melastomataceae.

**Methods:** In order to correct these issues, we cleaned the Melastomataceae probe set, extended it with additional sequences, and compared its performance with the original version.

**Results:** The final probe set targets 396 putative nuclear loci represented by 6009 template sequences. The probe set has been made available, along with details on the cleaning process, for reproducibility. We show that the new probe set performs better than the original version in terms of sequence recovery.

**Discussion:** This updated, extended, and cleaned probe set will improve the availability of phylogenomic resources across the Melastomataceae family. It is fully compatible with sequence recovery and extraction pipelines. The cleaning process can also be applied to any plant-targeting probe set that would need to be cleaned or updated if new genomic resources for the targeted taxa become available.

## KEYWORDS

Hyb-Seq, Melastomataceae, phylogenomics, phylogeny, target capture, transcriptome

The study of evolution is greatly enhanced by the use of various types of molecular analyses, from phylogenetic reconstructions to genetic investigations such as estimations of ploidy and hybridization. The development of probe sets for the targeted capture of nuclear genes means these analyses can now incorporate data from hundreds or even thousands of genes. In plants, “universal” probe sets have been developed for gene capture in flagellate land plants (Breinholt et al., 2021) and angiosperms (Johnson et al., 2019). Family-specific probe sets have been successfully developed for many different plant families, such as Annonaceae (Couvreur et al., 2019), Fabaceae (Koenen et al., 2020), Ochnaceae (Shah et al., 2021), Bromeliaceae (Yardeni et al., 2022), Bignoniaceae (Fonseca et al., 2023), and others.

With the advent of plant phylogenomics, a family-specific probe set was also developed for the Melastomataceae (Jantzen

et al., 2020). This probe set was designed to target 384 putative single-copy nuclear genes. The 384 loci are each represented by one to four template sequences representing the variation of nucleotide sequences at the locus. In total, the probe set is composed of 689 template sequences. The 384 loci targeted with this probe set were derived from the Angiosperms353 probe set (Johnson et al., 2019), published transcriptomes (Leebens-Mack et al., 2019), a previously developed low-copy nuclear gene set (Reginato and Michelangeli, 2016), and additional multi-copy functional genes (Jantzen et al., 2020). The probe set was initially designed with a focus on *Memecylon* L. and *Tibouchina* Aubl., two distantly related Melastomataceae genera from the subfamilies Olisbeoideae and Melastomatoideae, respectively, for which flanking intronic regions were assembled to the targeted exons in order to maximize capture within these genera (Jantzen et al., 2020).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Authors. *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

The probe set has been used with apparent success when focusing on the genera *Memecylon* and *Tibouchina* (Amarasinghe et al., 2021a, 2021b; Jantzen et al., 2022); however, when trying to use the probe set on other Melastomataceae taxa, we faced several practical problems and conceptual questions at the time of bioinformatic assembly. First, in the probe set, the different template sequences of the same locus are supposed to have a relatively high similarity and align to each other, but for some loci, we found that some template sequences align better to the reverse complement of the other template sequences of the same locus. These template sequences in reverse complement need to be corrected. Second, the template sequences of some loci have high sequence similarity to the template sequences of other loci. This is the case for eight pairs of loci initially derived from different sources. In such cases, the two loci with highly similar template sequences must be considered the same locus. Finally, unlike other published probe sets (e.g., Couvreur et al., 2019; Johnson et al., 2019) that were designed for exon capture, the Melastomataceae probe set was designed to capture sequences that are not necessarily exons. Indeed, for many template sequences in the Melastomataceae probe set, flanking intronic regions from samples of *Memecylon* and *Tibouchina* were assembled to the exon sequences. This seems to not be a problem when focusing on these two genera, as previous successful results were obtained (Amarasinghe et al., 2021a, 2021b; Jantzen et al., 2022); however, the presence of these genera-specific intronic regions is problematic when focusing on other infrafamilial levels (such as tribes or other genera) or the family level. Indeed, as the template sequences are composed of both intronic and exonic regions, the limits between exons and introns are lost, blurring any downstream exon–intron delineation in the retrieved sequences. Moreover, intronic regions often include stop codons. Programs used to delineate exons and introns rely on the sequence translation to amino acids and cannot deal with stop codons in the middle of the sequences (e.g., Exonerate [Slater and Birney, 2005] as used in HybPiper [Johnson et al., 2016], or Scipio [Keller et al., 2008] as used in Captus [Ortiz, 2022]). Such programs would thus not be able to run when stop codons are found in the template sequences composed of both intronic and exonic regions; for example, when trying to use the probe set with HybPiper version 2.0.1 (Johnson et al., 2016), it returned warnings, flagging many sequences for having unexpected stop codons and for not being multiples of three. In order for the exon–intron delineation to be possible using the available recovery pipelines (e.g., HybPiper, Captus), the template sequences must have the typical exon features, that is, being a multiple of three nucleotides and not including any stop codons.

These conceptual and practical problems raise important concerns regarding the use of this probe set for sequence recovery in any taxa outside of *Memecylon* and *Tibouchina*. To obtain a probe set suitable for Melastomataceae-wide exon capture, a completely new probe set should be redesigned

from scratch; however, this approach would be time-consuming and would prevent the use of data from the taxa already sequenced with the original probe set. We thus adopted the alternative approach of updating the original probe set. We aligned the original probe set to a custom set of reference sequences we built for this purpose and used the alignments to clean the probe sequences. We also extended the cleaned probe set by incorporating extra template sequences from the custom reference set into the final probe set. Note that the purpose of this new, clean, and updated probe set is only the recovery of sequences in silico (i.e., bioinformatically). A new probe set from this work was not physically constructed, so the original probe set is still needed to physically target and enrich the DNA.

To measure the improvements provided by the new probe set, we used publicly available sequence data for the Melastomataceae (Amarasinghe et al., 2021a; Jantzen et al., 2020, 2022) and HybPiper (Johnson et al., 2016) to compare the sequence recovery between the old and new probe sets. The expected improvements are two-fold. First, we expect a better sequence recovery from the new probe set alone, provided by the cleaned sequences and by the addition of extra template sequences for many loci. Indeed, the addition of extra template sequences has been shown to increase sequence recovery (McLay et al., 2021). Second, we expect an improvement from the option to use the probe set in an amino-acids format (translated nucleotides), provided by the reverse complementation of sequences and removal of *Tibouchina*- and *Memecylon*-specific intronic regions. Indeed, for sequence recovery and assembly, the use of the probe set in the amino-acids format is believed to be more efficient (see, for example, <https://github.com/mossmatters/HybPiper/wiki/Troubleshooting,-common-issues,-and-recommendations#20-read-mapping> [accessed 27 November 2023]). In addition, this will allow the recovery of both the targeted exons and associated (partial) introns across the Melastomataceae.

## METHODS

### Term definitions and usage

A “probe set” is a set of sequences used to target a set of loci. Throughout this paper, the term “probe set” is used to refer to a set of sequences in its totality. In a probe set, a locus can be represented by one to several “template sequences,” which are a particular version of the nucleotide sequence for each locus.

### Cleaning process

In summary, we first built a custom set of reference sequences derived from the Melastomataceae transcriptome data, which by definition only represented the exonic regions. We aligned each of the original template sequences with the reference set to identify the matching regions

(hits), then extracted the hits, reversed and complemented them if necessary, and assigned them to a locus name to obtain a final set of cleaned template sequences. Multiple hits from the same template sequence assigned to the same reference locus were concatenated into a single cleaned template sequence. To extend the probe set, we also appended sequences from the reference set to the final probe set. Finally, the final probe set went through a fine-tuned cleaning process (e.g., sequence-by-sequence stop codon removal) and we carried out the final checks (Figures 1 and 2).

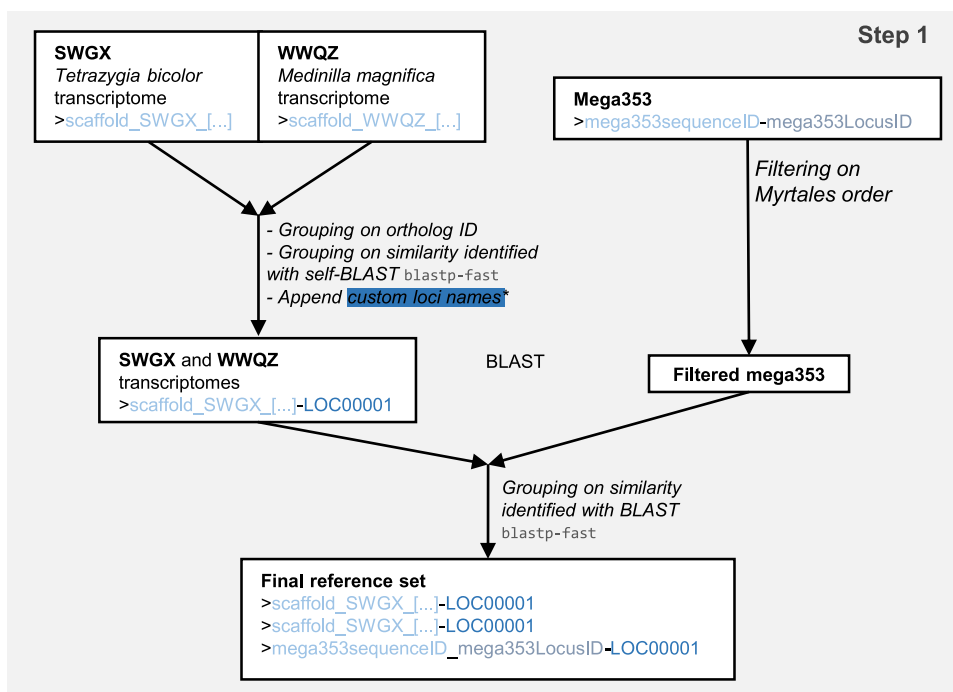
The main steps of the cleaning process are presented below. A more detailed version containing some additional and technical steps with fully reproducible corresponding code is available from [https://github.com/LPDagallier/Clean\\_Melasto\\_probe\\_set](https://github.com/LPDagallier/Clean_Melasto_probe_set) (see Data Availability Statement).

### Step 1: Building a custom set of reference sequences (Figure 1)

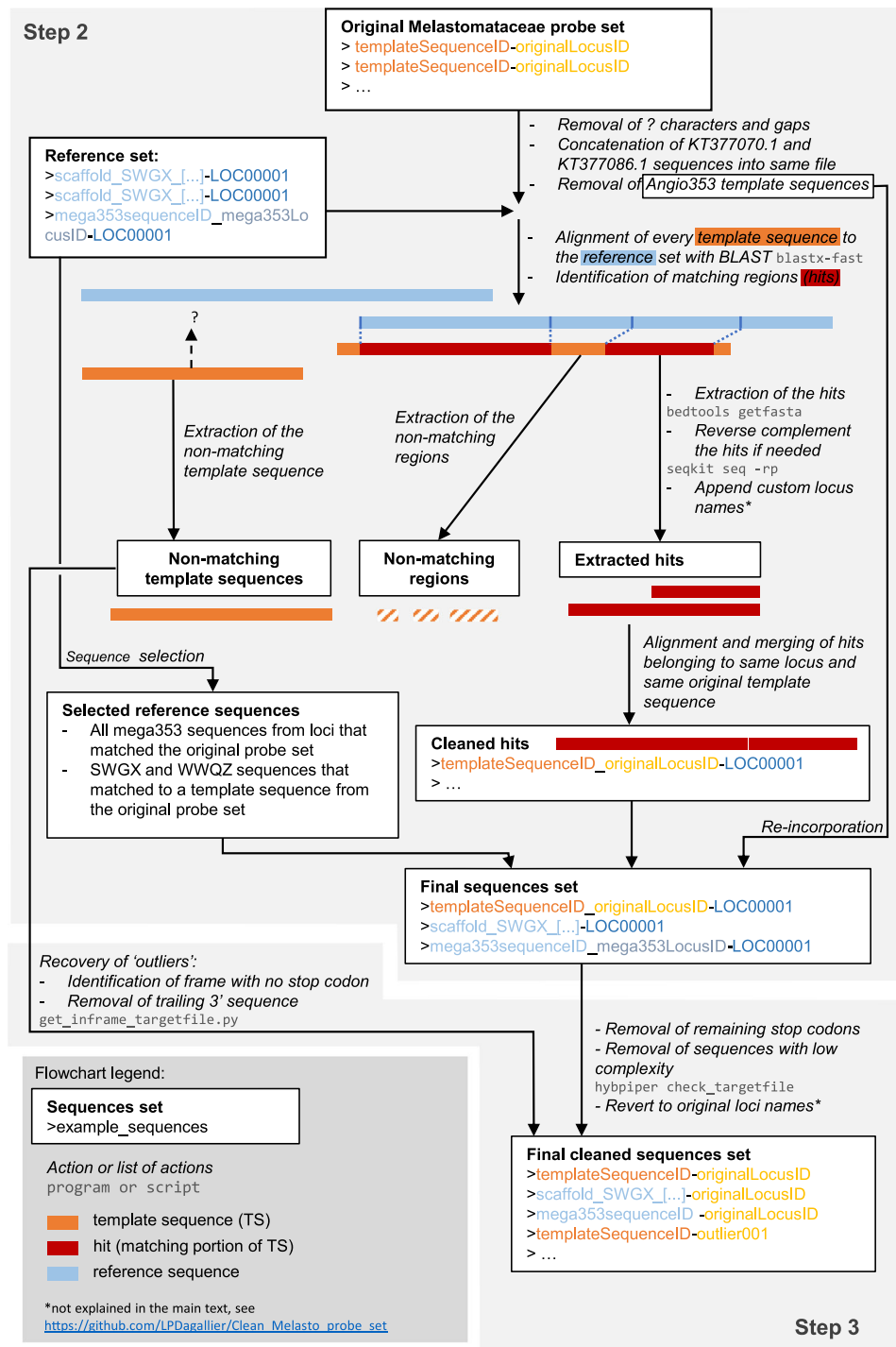
The original probe set (Jantzen et al., 2020) was mainly built upon sequences from Melastomataceae transcriptomes and from sequences from the Angiosperms353 probe set. We thus also used these sequences to build our custom reference set. Specifically, we used the transcriptome sequences of *Tetrazygia bicolor* Cogn. and *Medinilla magnifica* Lindl. (Melastomataceae) from the 1KP project (Leebens-Mack et al., 2019), and the “mega353” probe set

(McLay et al., 2021). The mega353 probe set is a version of Angiosperms353 (Johnson et al., 2019) extended with many additional template sequences from transcriptomic data. Because some sequences from our custom reference set will be included in the new probe set (see the extension step at the end of Step 2 below), and to take advantage of the improved recovery it provides (McLay et al., 2021), we used the mega353 probe set instead of the original Angiosperms-353. Here, we filtered the mega353 probe set to include only the template sequences from the order Myrtales (to which Melastomataceae belong).

In the reference set, the sequences with high similarity were grouped under a single “locus” name. To do so, the transcriptome sequences were first grouped according to their ortholog group, as defined in the 1KP data set. We then used the Basic Local Alignment Search Tool (BLAST) version 2.10.1 (Altschul et al., 1990), running an all-by-all BLAST with the algorithm blastp-fast (Shiryev et al., 2007). We assigned the same locus name to sequences with matches with an *E*-value below 1e-06, and with either (i) a percentage of identity greater than 60%, an alignment length greater than 50 amino acids, a bitscore greater than 100, and a query coverage greater than 50% (meaning at least 50% of the query covers the subject sequence); or (ii) a percentage of identity greater than 80%, an alignment length greater than 50 amino acids, and a bitscore greater than 50. We then matched the mega353 sequences with the transcriptome sequences based on sequence similarity and assigned the same locus name to each group. For that, we



**FIGURE 1** Flowchart representing the construction of a custom set of reference sequences (Step 1 of the cleaning process). White boxes represent intermediate sequence sets, with examples of sequences identified with “>”. Actions are in italics. Programs are in console font. Dark blue represents the custom loci names, while light blue represents the template sequence names.



**FIGURE 2** Flowchart representing the cleaning of the template sequences and the extension of the probe set (Step 2 of the cleaning process), as well as the fine-tuned cleaning and final checks (Step 3 of the cleaning process). White boxes represent intermediate sequences sets, with examples of sequences identified with “>”. Actions are in italics. Programs and scripts are in console font. Different text colors identify different sequence sources: yellow represents the locus name in the original probe set, orange represents the template sequence name in the original probe set, dark blue represents the custom loci names, and light blue represents the template sequence names from the reference set.

used BLAST (blastp-fast) to group under the same locus name the sequences with matches with an *E*-value below  $1e-06$  and with a percentage identity greater than 60% to the matching subject, an alignment length greater than 50 amino acids, a bitscore greater than 100, a frame greater

than 0 (i.e., sequences not in reverse complement), and a query coverage greater than 50% (meaning at least 50% of the query covers the subject sequence). After grouping, the reference set was composed of 50,677 sequences grouped into 11,811 “loci.”

## Step 2: Cleaning the template sequences and extending the set (Figure 2)

The original Melastomataceae template sequences were retrieved from their original publication (Jantzen et al., 2021). Some sequences included question mark characters (?) that were removed using the sed Unix tool as they have no meaning in the International Union of Pure and Applied Chemistry accepted nomenclature (<https://iupac.qmul.ac.uk/misc/naabb.html>). Gaps were also removed.

The template sequences of the loci KT377070.1 and KT377086.1 were both split into two different FASTA files, supposedly representing two versions of these highly divergent loci (Johanna Jantzen, Université de Montréal, personal communication). Instead, we found that the alternative version is not a highly divergent version of the locus, but rather a reverse-complement sequence. As the reverse-complement sequences are addressed later in the cleaning process, we merged the FASTA files in order to have only one file per locus.

In the original probe set, the Angiosperms353 loci are represented by at least one template sequence obtained directly from the Angiosperms353 probe set (Johnson et al., 2019), with some also including additional template sequences assembled from the genome skimming data (Jantzen et al., 2020). We removed the template sequences that came directly from the Angiosperms353 probe set from the original probe set prior to the alignment to the reference set, because the reference set itself also contains these Angiosperms353 template sequences. These removed sequences were later reincorporated during the extension step (see below).

Each template sequence from the original probe set was then aligned to the reference set using BLAST (nucleotide to protein algorithm; blastx-fast) (Shiryev et al., 2007). When a template sequence aligned to a reference sequence, the portions of sequence that matched to a reference (the hits) were then extracted into separate files using getfasta from BEDTools version 2.30.0 (Quinlan and Hall, 2010), and the name of the reference locus was appended to the hits. In cases where a hit aligned with a negative frame, it was reverse-complemented right after extraction using SeqKit version 2.0.0 (seqkit seq -rp) (Shen et al., 2016). The non-matching regions were also identified and separated into different files, together with template sequences with no match in the reference set.

In some cases, multiple hits were extracted from the same template sequence and were assigned to the same reference locus. In such cases, the hits were merged into a single sequence. Because different hits may overlap over their reference sequence, the extracted hits cannot just be concatenated end-to-end. We thus realigned the extracted hits to their reference sequence and drew consensus sequences of the aligned hits using a custom script in R version 4.1.3 (R Core Team, 2022) and the DECIPHER and Biostrings packages (Wright, 2015, 2016; Pagès et al., 2022).

To extend the probe set, we appended several reference sequences to the set of cleaned hits. We selected reference sequences from the loci that previously aligned to the template sequences. Among them, we appended all the Angiosperms353

reference sequences, but as some loci have a very high number of transcriptome reference sequences, we appended only the transcriptome reference sequences that previously aligned to a template sequence. We also reincorporated the previously removed Angiosperms353 template sequences and loci.

## Step 3: Fine-tune cleaning and final checks (Figure 2)

We found stop codons in several template sequences of the final probe set. To remove these stop codons, we examined the sequences in AliView version 1.28 (Larsson, 2014). We aligned the template sequences with stop codons and all the other template sequences from the same locus. The alignments were carried out in AliView using Muscle version 3.8.425 (Edgar, 2004) with a high open gap penalty (“-gapopen -10000”). We then examined the alignment at the stop codon position and replaced with ‘N’ any of the nucleotide(s) from the stop codon that were not consistent with the alignment at this position; for example, a TAA stop codon occurring at a position where the consensus sequence of the alignment is GAA would have been changed to NAA. In some cases, stop codons were found in regions with many ambiguous nucleotides, or in regions that poorly aligned with the other template sequences around the head or tail of the sequences. As regions with many ambiguous nucleotides could be the result of poor sequencing quality, and as regions that poorly align with the other template sequences are possibly the remains of introns, we removed the entire region in cases where stop codons were found.

To recover as much as possible from the original probe set, we recycled the non-matching template sequences (i.e., sequences with no reference found in the reference set). We used a script developed by Chris Jackson ([https://github.com/mossmatters/HybPiper/files/8933179/get\\_inframe\\_targetfile.py.gz](https://github.com/mossmatters/HybPiper/files/8933179/get_inframe_targetfile.py.gz) [accessed 14 November 2022]) that identified the first forward frame that has no stop codon and recovered the in-frame sequence while removing trailing 3’ nucleotides. The retrieved in-frame sequences were renamed with the locus name “outlier###” (with # being a digit). See the Discussion for cautions concerning these outlier loci.

We checked the final template sequences for low-complexity sequences using “hybpiper check\_targetfile” with the default parameters in HybPiper version 2.0.1 (Johnson et al., 2016). For a given locus, all the template sequences flagged as having low complexity were removed, as long as the locus was represented by other template sequences. If all the template sequences of a locus were flagged as having low complexity, the template sequences were not removed.

## Comparing the sequence recovery between the old and new probe sets

To compare the locus sequence recovery between the original and the new probe sets, we used publicly available

data for *Tibouchina* and *Memecylon* (Amarasinghe et al., 2021a; Jantzen et al., 2020, 2022). This data set is composed of 240 samples (144 *Tibouchina* and 96 *Memecylon*) that were sequenced using the old probe set to physically target the DNA (Jantzen et al., 2020), and the comparison here relates only to the in silico sequence recovery. All the sequences from the BioProjects PRJNA573947 and PRJNA576018 (see accession list in Appendix S1) were downloaded from the Sequence Read Archive (National Center for Biotechnology Information [NCBI]; <https://www.ncbi.nlm.nih.gov/sra>) using the SRA-Toolkit's 'prefetch' function. For each accession, the archive was extracted into three FASTQ files (two files with paired reads and one file with unpaired reads) using the 'fastq-dump --split-3' function.

The old probe set was retrieved from the original publication (Jantzen et al., 2020). We merged all the template sequences into a single FASTA file; removed the gaps; removed all the characters that were not A, T, G, C, or N; and formatted the sequences headers to follow the standards of most pipelines (">sequenceID-locusID"). These steps were necessary to ensure the probe set was compatible with HybPiper, otherwise HybPiper would return errors without running. As mentioned earlier (see beginning of Step 2 above), some loci are present under two different names. We thus changed the names KT377070, KT377086, KT377102, and KT377110 to KT377070.1, KT377086.1, KT377102.1, and KT377110.1, respectively, to maintain 384 consistent locus names.

We used HybPiper version 2.1.1 (Johnson et al., 2016) to run three different assemblies of the targeted loci: one with the old probe set in nucleotide format, a second with the new probe set in nucleotide format, and a third with the new probe set in amino-acids format (i.e., translated nucleotides). HybPiper was called with the Burrows–Wheeler Aligner (Li and Durbin, 2009) ('-bwa' option) for the runs with probe sets in the nucleotide format and Diamond (Buchfink et al., 2021) ('-diamond' option) for the run with the probe set in the amino-acids format. HybPiper was run on both paired and unpaired reads, with no stitching of the contigs (option '--no\_stitched\_contigs') and with eight CPUs and 64 GB of RAM allocated. We used HybPiper to compute the assembly's statistics, and used R version 4.2.1 (R Core Team, 2022) and the packages ggplot2 version 3.4.2, ggh4x version 0.2.4, and ggdist version 3.3.0 (Kay and Wiernik, 2023; van den Brand, 2023; Wickham et al., 2023) to visualize and summarize the results.

## RESULTS

The final cleaned probe set is composed of 396 loci (including 14 "outliers," i.e., not matching the references) represented by 6009 template sequences. Every template sequence is a multiple of three nucleotides, free of stop codons in the first-forward frame, and free of intronic regions. The final cleaned probe set is publicly available in the nucleotide format ([https://github.com/LPDagallier/Clean\\_Melasto\\_probe\\_set/blob/master/CLEAN\\_PROBE\\_SET/PROBE\\_SET\\_CLEAN.FNA](https://github.com/LPDagallier/Clean_Melasto_probe_set/blob/master/CLEAN_PROBE_SET/PROBE_SET_CLEAN.FNA)) and in the

translated amino-acid format ([https://github.com/LPDagallier/Clean\\_Melasto\\_probe\\_set/blob/master/CLEAN\\_PROBE\\_SET/PROBE\\_SET\\_CLEAN\\_prot.FAA](https://github.com/LPDagallier/Clean_Melasto_probe_set/blob/master/CLEAN_PROBE_SET/PROBE_SET_CLEAN_prot.FAA)) (see Data Availability Statement).

To infer the efficiency of recovery, we focused on several statistics averaged over the 240 *Tibouchina* and *Memecylon* samples: the percentage of reads on target (i.e., the percentage of reads for each sample that map to the loci in the probe set), the number of loci that have reads mapped to the reference, the number of loci that have an assembled sequence, and the number of loci that have an assembled sequence  $\geq 75\%$  of the reference length. Compared with the old probe set, the new probe set in the nucleotide format had a slightly lower percentage of reads on target (78.94% vs. 83.06%) and number of loci with an assembled sequence  $>75\%$  of the reference length (62.68 vs. 70.91), but a higher number of loci with mapped reads (341.69 vs. 272.21) and loci with assembled sequences (264.74 vs. 221.21) (Table 1, Figure 3). In comparison with the old and new probe sets in the nucleotide format, the new probe set in the amino-acids format showed a lower percentage of reads on target (40.61% vs.  $>78\%$ ) (Table 1). The number of loci with mapped reads and the number of loci with assembled sequences were higher than with the old probe set (328.21 vs. 272.21 and 276.65 vs. 221.11, respectively), but similar to the new probe set in the nucleotide format (Table 1, Figure 3A, B). The number of loci recovered with 75% of the target sequence was higher for the new probe set in the amino-acids format than for the old and new probe sets in the nucleotide format (Table 1, Figure 3C).

## DISCUSSION

During the cleaning process, we corrected previously unexplained mistakes, such as reverse-complement sequences and highly similar sequences separated into different loci. The cleaned probe set is composed of 396 loci (including 14 outliers) vs. 384 loci in the original probe set (Jantzen et al., 2020). This cleaned probe set enables the delineation between the exon–intron boundaries in Melastomataceae taxa outside of *Memecylon* and *Tibouchina*. Contrary to the original probe set, it is fully compatible with commonly used programs for exon–intron delineation, such as Exonerate (Slater and Birney, 2005) and Scipio (Keller et al., 2008), and more broadly with commonly used sequence-recovery pipelines such as HybPiper (Johnson et al., 2016) and Captus (Ortiz, 2022). Finally, we extended the probe set with extra template sequences from our custom reference set. In total, the new probe set is composed of 6009 template sequences, in comparison with the 689 template sequences in the original (Jantzen et al., 2020). The origin of the 14 outlier loci is unclear, and it is possible that they are of intronic nature (not removed because they do not contain any stop codons). They may also be partial exons from loci present in the probe set, but we could not recover them as such because they are too specific to *Tibouchina* or *Memecylon* to match

**TABLE 1** Summary of recovery statistics for the three HybPiper assembly runs. Values are summarized across the 240 samples assembled in each run.

Probe set	Mapping	Mean % of reads on target ( $\pm$ SD)	Mean no. of loci with mapped reads ( $\pm$ SD)	Mean no. of loci with sequence ( $\pm$ SD)	Mean no. of loci at 75% of target sequence ( $\pm$ SD)
Old (nt)	BWA	83.06% ( $\pm$ 8.57%)	272.21 ( $\pm$ 15)	221.11 ( $\pm$ 22.26)	70.91 ( $\pm$ 23.61)
New (nt)	BWA	78.94% ( $\pm$ 9.9%)	341.69 ( $\pm$ 15.75)	264.74 ( $\pm$ 35.71)	62.68 ( $\pm$ 25.23)
New (aa)	Diamond	40.61% ( $\pm$ 4.79%)	328.21 ( $\pm$ 16.18)	276.65 ( $\pm$ 36.75)	107.71 ( $\pm$ 45.62)

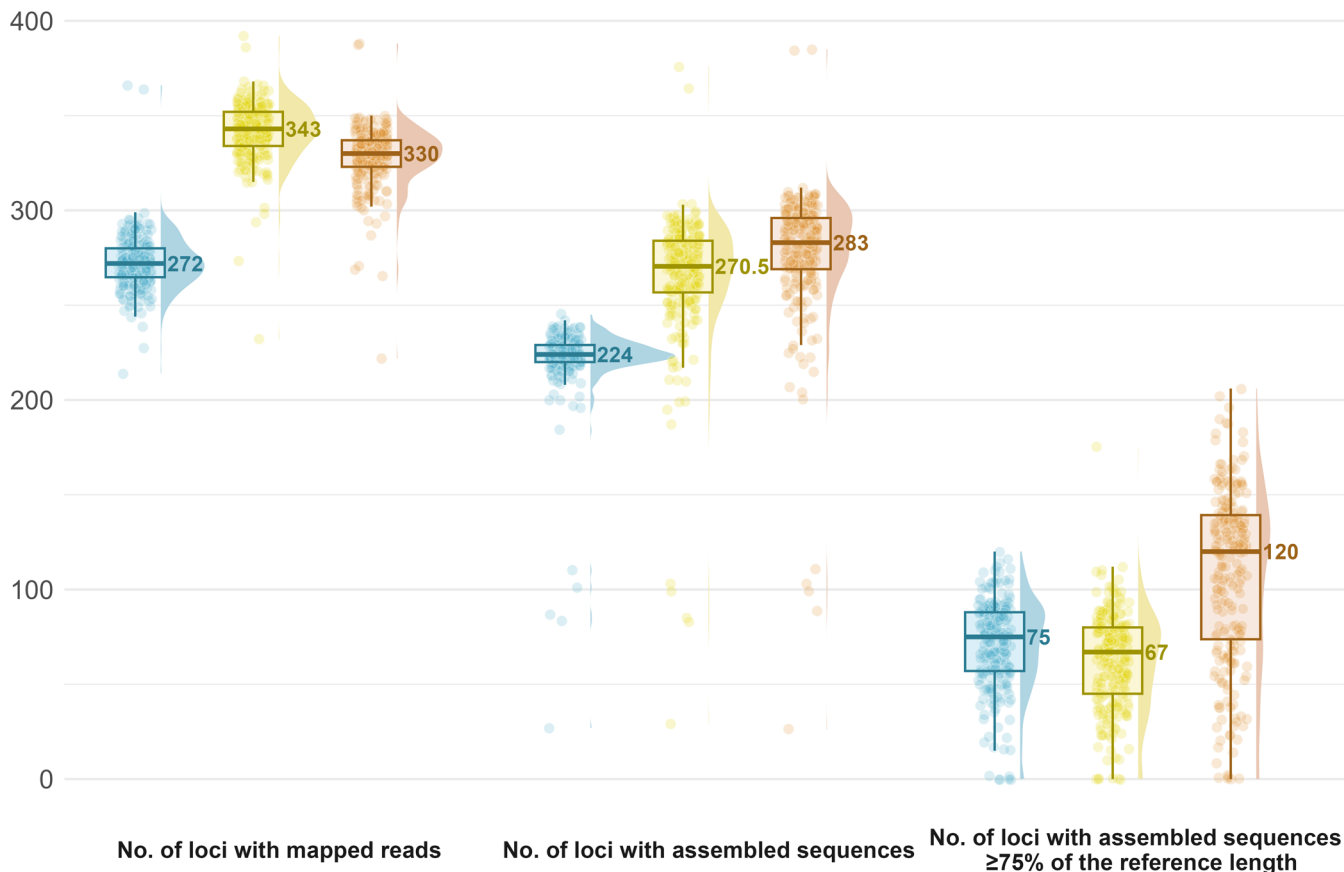
Note: aa = amino acid; BWA = Burrows-Wheeler aligner; nt = nucleotide; SD = standard deviation.

any sequence in the reference set. We included these outlier loci in the new probe set in an effort to recycle as much as possible of the old probe set, but we recommend using them with extreme caution.

Overall, the comparison between the old and new probe sets shows that the latter performs better in terms of sequence recovery. Admittedly, the average percentage of reads that map to a template sequence is higher for the old probe set (Table 1), but this result is expected because it is the old probe set that physically targeted the DNA. Moreover, the number of loci with mapped reads and with assembled sequences (i.e., with sequences that will be used for phylogenetic reconstruction) is higher for the new probe set (Table 1, Figure 3A, B). The better performance of the new probe set can be attributed to the addition of extra template sequences from our custom reference set. As previously shown with the Angiosperms353 probe set, the introduction of greater sequence variation into a probe set improves the sequence recovery and assembly (McLay et al., 2021). The cleaning steps (e.g., removal of introns, reverse complementation) may also play a role in the higher performance of the new probe set, even if their relative contribution is hard to appraise. Regardless, this cleanup was necessary to produce a probe set in the amino-acids format. Importantly, sequence recovery using the new probe set in the amino-acids format was more efficient in terms of the number of assembled loci and in assembling longer sequences than with the new probe set in the nucleotide format (Figure 3). This confirms the unpublished HybPiper benchmark suggesting that mapping the reads with an amino-acid reference results in more and longer assembled sequences (<https://github.com/mossmatters/HybPiper/wiki/Troubleshooting,-common-issues,-and-recommendations#20-read-mapping> [accessed 27 November 2023]). We thus advise future researchers to use the new probe set in the amino-acid format for an even better recovery efficiency.

In conclusion, this extended and cleaned new probe set provides a better and broader sequence recovery across the Melastomataceae. It will enhance the use of phylogenomic resources in this diverse plant family, which will in turn help to resolve questions regarding Melastomataceae evolution, biogeography, and systematics. We recommend using this new probe set for bioinformatic recovery only. To physically target nuclear sequences in the Melastomataceae, users may want to use the old probe set (Jantzen et al., 2020, but bear in mind concerns raised here) or the Angiosperms353 probe set (Johnson et al., 2019), which was previously shown to have good targeting efficiency in this family (Maurin et al., 2021).

Further details about the cleaning process, comparison analysis, and associated code are available from [https://github.com/LPDagallier/Clean\\_Melasto\\_probe\\_set](https://github.com/LPDagallier/Clean_Melasto_probe_set), making this work fully reproducible. This also allows the presented process to be expanded to any other probe set already in use that requires cleaning and/or updating with publicly available transcriptomes.



**FIGURE 3** Summary of recovery statistics computed with HybPiper for the assemblies with the old probe set (blue), the new probe set in the nucleotide format (yellow), and the new probe set in the amino-acids format (orange). Burrows–Wheeler aligner was used to map the reads with nucleotide probe sets, and Diamond was used for the amino-acids probe set. The boxes of the boxplots delineate the first and third quartiles; the upper and lower whiskers extend to the largest and lowest values, respectively, no further than 1.5 times the distance between the first and third quartiles. The numbers to the right of the boxplots are the median values.

### AUTHOR CONTRIBUTIONS

F.A.M. coordinated the project. L.P.M.J.D. designed the cleaning process, wrote the code, carried out the cleaning, and wrote the manuscript. F.A.M. revised the manuscript. Both authors approved the final version of the manuscript.

### ACKNOWLEDGMENTS

The authors thank Johanna Jantzen (Université de Montréal) for discussions about the original probe set, Luo Chen (University of Munich) for the suggested improvements of the earlier version of the cleaned probe set, and Luis Fonseca (Ghent University) and two anonymous reviewers for their constructive comments during the review process. This research was funded by the National Science Foundation (DEB 2001357, DEB 2002270).

### OPEN RESEARCH BADGES



This article has earned an Open Materials badge for making publicly available the components of the research methodology

needed to reproduce the reported procedure and analysis. All materials are available at [https://github.com/LPDagallier/Clean\\_Melasto\\_probe\\_set](https://github.com/LPDagallier/Clean_Melasto_probe_set).

### DATA AVAILABILITY STATEMENT

The step-by-step details about the cleaning process, comparison analysis, and associated code are available from [https://github.com/LPDagallier/Clean\\_Melasto\\_probe\\_set](https://github.com/LPDagallier/Clean_Melasto_probe_set).

The final cleaned probe set is publicly available in the nucleotide format from [https://github.com/LPDagallier/Clean\\_Melasto\\_probe\\_set/blob/master/CLEAN\\_PROBE\\_SET/PROBE\\_SET\\_CLEAN.FNA](https://github.com/LPDagallier/Clean_Melasto_probe_set/blob/master/CLEAN_PROBE_SET/PROBE_SET_CLEAN.FNA) and in the translated amino-acid format from [https://github.com/LPDagallier/Clean\\_Melasto\\_probe\\_set/blob/master/CLEAN\\_PROBE\\_SET/PROBE\\_SET\\_CLEAN\\_prot.FAA](https://github.com/LPDagallier/Clean_Melasto_probe_set/blob/master/CLEAN_PROBE_SET/PROBE_SET_CLEAN_prot.FAA).

### ORCID

Léo-Paul M. J. Dagallier <http://orcid.org/0000-0002-3270-1544>

Fabián A. Michelangeli <http://orcid.org/0000-0001-7348-143X>



## REFERENCES

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.
- Amarasinghe, P., S. Joshi, N. Page, L. S. Wijedasa, M. Merello, H. Kathriarachchi, R. D. Stone, et al. 2021a. Evolution and biogeography of *Memecylon*. *American Journal of Botany* 108: 628–646.
- Amarasinghe, P., P. Pham, R. D. Stone, and N. Cellinese. 2021b. Discordance in a South African *Memecylon* clade (Melastomataceae): Evidence for reticulate evolution. *International Journal of Plant Sciences* 182: 682–694.
- Breinholt, J. W., S. B. Carey, G. P. Tiley, E. C. Davis, L. Endara, S. F. McDaniel, L. G. Neves, et al. 2021. A target enrichment probe set for resolving the flagellate land plant tree of life. *Applications in Plant Sciences* 9: e11406.
- Buchfink, B., K. Reuter, and H.-G. Drost. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* 18: 366–368.
- 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.
- Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.
- Fonseca, L. H. M., M. M. Carlsen, P. V. A. Fine, and L. G. Lohmann. 2023. A nuclear target sequence capture probe set for phylogeny reconstruction of the charismatic plant family Bignoniaceae. *Frontiers in Genetics* 13: 1085692.
- Jantzen, J. R., P. Amarasinghe, R. A. Folk, M. Reginato, F. A. Michelangeli, D. E. Soltis, N. Cellinese, and P. S. Soltis. 2020. A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in Melastomataceae. *Applications in Plant Sciences* 8: e11345.
- Jantzen, J. R., P. Amarasinghe, R. A. Folk, M. Reginato, F. A. Michelangeli, D. E. Soltis, N. Cellinese, and P. S. Soltis. 2021. Data from: A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in Melastomataceae. Website: <https://doi.org/10.5061/dryad.8931zcrm2> [accessed 1 September 2022].
- Jantzen, J. R., P. J. F. Guimarães, L. C. Pederneiras, A. L. F. Oliveira, D. E. Soltis, and P. S. Soltis. 2022. Phylogenomic analysis of *Tibouchina* s.s. (Melastomataceae) highlights the evolutionary complexity of Neotropical savannas. *Botanical Journal of the Linnean Society* 199: 372–411.
- Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: 1600016.
- Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eisehardt, 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.
- Kay, M., and B. M. Wiernik. 2023. ggdist: Visualizations of distributions and uncertainty. R package 3.3.1. Available at Zenodo repository: <https://doi.org/10.5281/zenodo.10236301> [posted 30 November 2023; accessed 22 December 2023].
- Keller, O., F. Odrionitz, M. Stanke, M. Kollmar, and S. Waack. 2008. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* 9: 278.
- Koenen, E. J. M., C. Kidner, É. R. de Souza, M. F. Simon, J. R. Iganci, J. A. Nicholls, G. K. Brown, et al. 2020. Hybrid capture of 964 nuclear genes resolves evolutionary relationships in the mimosoid legumes and reveals the polytomous origins of a large pantropical radiation. *American Journal of Botany* 107: 1710–1735.
- Larsson, A. 2014. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30: 3276–3278.
- Leebens-Mack, J. H., M. S. Barker, E. J. Carpenter, M. K. Deyholos, M. A. Gitzendanner, S. W. Graham, I. Grosse, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.
- Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25: 1754–1760.
- Maurin, O., A. Anest, S. Bellot, E. Biffin, G. Brewer, T. Charles-Dominique, R. S. Cowan, et al. 2021. A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set. *American Journal of Botany* 108: 1087–1111.
- McLay, T. G. B., J. L. Birch, B. F. Gunn, W. Ning, J. A. Tate, L. Nauheimer, E. M. Joyce, et al. 2021. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences* 9: e11420.
- Ortiz, E. M. 2022. Captus, Assembly of phylogenomic datasets from high-throughput sequencing data. Website: <https://github.com/edgardomortiz/Captus> [accessed 23 November 2023].
- Pages, H., P. Aboyoun, R. Gentleman, and S. DebRoy. 2022. Biostrings: Efficient manipulation of biological strings. Website: <https://bioconductor.org/packages/Biostrings> [accessed 23 November 2023].
- Quinlan, A. R., and I. M. Hall. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.
- R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website: <http://www.R-project.org/> [accessed 23 November 2023].
- Reginato, M., and F. A. Michelangeli. 2016. Primers for low-copy nuclear genes in the Melastomataceae. *Applications in Plant Sciences* 4: 1500092.
- Shah, T., J. V. Schneider, G. Zizka, O. Maurin, W. Baker, F. Forest, G. E. Brewer, et al. 2021. Joining forces in Ochnaceae phylogenomics: A tale of two targeted sequencing probe kits. *American Journal of Botany* 108: 1201–1216.
- Shen, W., S. Le, Y. Li, and F. Hu. 2016. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* 11: e0163962.
- Shiryev, S. A., J. S. Papadopoulos, A. A. Schäffer, and R. Agarwala. 2007. Improved BLAST searches using longer words for protein seeding. *Bioinformatics* 23: 2949–2951.
- Slater, G. S. C., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31.
- van den Brand, T. 2023. ggh4x: Hacks for 'ggplot2'. R package 0.2.7. Website: <https://teunbrand.github.io/ggh4x/> [accessed 22 December 2023].
- Wickham, H., W. Chang, L. Henry, T. L. Pedersen, K. Takahashi, C. Wilke, K. Woo, et al. 2023. ggplot2: create elegant data visualisations using the grammar of graphics. R package 3.4.3. Website: <https://ggplot2.tidyverse.org/> [accessed 21 December 2023].
- Wright, E. S. 2015. DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* 16: 322.
- Wright, E. S. 2016. Using DECIPHER v2.0 to analyze big biological sequence data in R. *The R Journal* 8: 352–359.
- Yardeni, G., J. Viruel, M. Paris, J. Hess, C. Groot Crego, M. de La Harpe, N. Rivera, et al. 2022. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources* 22: 927–945.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**Appendix S1.** List of accessions for the 240 samples used in the sequence recovery comparison.

**How to cite this article:** Dagallier, L.-P. M. J., and F. A. Michelangeli. 2024. An updated and extended version of the Melastomataceae probe set for target capture. *Applications in Plant Sciences* 12(1): e11564. <https://doi.org/10.1002/aps3.11564>