A peer-reviewed version of this preprint was published in PeerJ on 25 July 2017.

View the peer-reviewed version (peerj.com/articles/3569), which is the preferred citable publication unless you specifically need to cite this preprint.

Kadlec M, Bellstedt DU, Le Maitre NC, Pirie MD. 2017. Targeted NGS for species level phylogenomics: “made to measure” or “one size fits all”? PeerJ 5:e3569 https://doi.org/10.7717/peerj.3569

PeerJ Preprints NOT PEER-REVIEWED

Targeted NGS for species level phylogenomics: "made to measure" or "one size fits all"?

Malvina Kadlec[1,3], Dirk U. Bellstedt[2], Nicholas C. Le Maitre[2], and Michael D. Pirie[1,2]

[1]Institut für Organismische und Molekulare Evolutionsbiologie, Johannes Gutenberg-Universität, Anselm-Franz-von-Bentzelweg 9a, 55099 Mainz, Germany

[2]Department of Biochemistry, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa

[3]Author for correspondence: mkadlec@uni-mainz.de

Abstract

Targeted high-throughput sequencing using hybrid-enrichment offers a promising source of data for inferring multiple, meaningfully resolved, independent gene trees suitable to address challenging phylogenetic problems in species complexes and rapid radiations. The targets in question can either be adopted directly from more or less universal tools, or custom made for particular clades at considerably greater effort. We applied custom made scripts to select sets of homologous sequence markers from transcriptome and WGS data for use in the flowering plant genus *Erica* (Ericaceae). We compared the resulting targets to those that would be selected both using different available tools (Hyb-Seq; MarkerMiner), and when optimising for broader clades of more distantly related taxa (Ericales; eudicots). Approaches comparing more divergent genomes (including MarkerMiner, irrespective of input data) delivered fewer and shorter potential markers than those targeted for *Erica*. The latter may nevertheless be effective for sequence capture across the wider family Ericaceae. We tested the targets delivered by our scripts by obtaining an empirical dataset. The resulting sequence variation was lower than that of standard nuclear ribosomal markers (that in *Erica* fail to deliver a well resolved gene tree), confirming the importance of maximising the lengths of individual markers. We conclude that rather than searching for "one size fits all" universal markers, we should improve and make more accessible the tools necessary for developing "made to measure" ones.

Keywords: Ericaceae; hybridization enrichment; marker development, next-generation sequencing; phylogeny; targeted sequence capture; target enrichment; transcriptome

Introduction

DNA sequence data is the cornerstone of comparative and evolutionary research, invaluable for inference of population-level processes and species delimitation through to higher level relationships. Sanger sequencing (Sanger, Nicklen & Coulson, 1977) and Polymerase Chain Reaction (PCR) amplification (Saiki et al., 1985) have been standard tools for decades, aided by the development of protocols that can be applied across closely and distantly related organisms. In plants, universal primers such as for plastid (Taberlet et al., 1991), nuclear ribosomal (White et al., 1990) and even single or low copy nuclear (Blattner, 2016) sequences have been widely applied to infer evolutionary histories. Many empirical studies are still limited to these few independent markers, the phylogenetic signal of which may not reflect the true sequence of speciation events (Kingman, 1982; White et al., 1990). Additionally, the resulting gene trees are often poorly resolved, particularly when divergence of lineages was rapid. When it is not possible to generate a robust and unambiguous phylogenetic hypothesis using standard universal markers, protocols for alternative low copy genes are highly desirable (Sang, 2002; Hughes, Eastwood & Bailey, 2006).

With the development of next generation sequencing (NGS) techniques, we now have potential access to numerous nuclear markers allowing us to address evolutionary questions without being constrained by the generation of sequence datasets per se. In principle, the whole genome is at our disposal, but whole genome sequencing (WGS) is currently relatively expensive, time-consuming and computationally difficult, especially for non-model organisms and eukaryote genomes in general (Jones & Good, 2016). These disadvantages will doubtless reduce in the near future, but nevertheless much of the data that might be obtained through WGS is irrelevant for particular purposes. In the case of phylogenetic problems, repetitive elements and multiple copy genes are not useful; neither are sequences that are highly constrained and hence insufficiently variable, nor indeed those that are too variable and impossible to align; nor those subject to strong selection pressure. We need strategies to identify and target sequencing of markers appropriate for phylogenomic analysis in different clades and at different taxonomic levels, and are currently faced with an array of options.

Different methods, referred to in general as “genome-partitioning approaches”, or “reduced-representation genome sequencing”, have been developed that are cheaper, faster and computationally less demanding than WGS, and as such are currently more feasible for analyses of numerous samples for particular purposes (Mamanova et al., 2010). These include restriction-site-associated DNA sequencing (RAD-seq; Miller et al., 2007), and

similar Genotyping by sequencing (GBS) approaches (Elshire et al., 2011), and whole-transcriptome shotgun sequencing (RNA-seq; Wang, Gerstein & Snyder, 2009). These methods can be applied to non-model species (Johnson et al., 2012) but do not necessarily deliver the most informative data for phylogenetic inference. RAD-seq/GBS sequences are short, generally used for obtaining (independent) single nucleotide polymorphisms (SNPs) from across the genome, suitable for population genetic analyses. RNA-seq transcriptome data cannot be obtained from dried material (such as herbarium specimens), restricting its application, and the sequences that are obtained are functionally conserved and therefore may be more suitable for analysing more ancient divergences, such as the origins of land plants (Wickett et al., 2014). Neither approach is ideal for inferring meaningfully resolved independent gene trees of closely related species as they will inevitably present limited numbers of linked, informative characters.

Alternative approaches can be used to target more variable, longer contiguous sequences involving selective enrichment of specific subsets of the genome before using NGS through PCR based, or sequence capture techniques. PCR based enrichment, or multiplex and microfluidic amplification of PCR products, is the simultaneous amplification of multiple targets (e.g. 48, as used in Uribe-Convers, Settles & Tank, 2016; to potentially hundreds or low thousands per reaction). Although this method dispenses with the need for time-consuming library preparation, it requires prior knowledge of sequences for the design of primers; such primers must be restricted to within regions that are known to be conserved across the study group.

Current targeted sequence capture methods involve hybridization in solution between genomic DNA fragments and biotinylated RNA “baits” (also referred to as “probes” or the “Capture Library’’) between 70 and 120 bp long. Hybridization capture can be used with non-model organisms (as is the case for RAD-seq/GBS and RNA-seq), and shows promising results with fragmented DNA (such as might be retrieved from museum specimens) (Lemmon & Lemmon, 2013; Zimmer & Wen, 2015; Hart et al., 2016; Budenhagen et al., 2016). Moreover, even without baits specifically designed using organelle genomes, plastid and mitochondrial sequences can also be retrieved during the hybrid-enrichment process (Tsangaras et al., 2014). Use of targeted sequence capture for phylogenetic inference is on the increase but still somewhat in its infancy, with a range of different more or less customised laboratory and bioinformatic protocols being applied to different organismal groups and in different laboratories. The protocols follow two general approaches: One is to design baits for

PeerJ Preprints NOT PEER-REVIEWED

use in specific organismal groups (e.g. Compositae, Mandel et al., 2014; cichlid fish, Ilves & Lopez-Fernandez, 2014; and Apocynaceae, Weitemier et al., 2014□). To this end, conserved orthologous sequences of genes of the species of interest are identified e.g. using a BLASTn or BLASTx search (or equivalent) with transcriptome data, expressed sequences tags (ESTs) and/or WGS. Alternatively, and with considerably less effort, pre-designed sets of more universal baits are used (Faircloth et al., 2012; Lemmon, Emme & Lemmon, 2012). Of the latter, "Ultra Conserved Elements" (UCE) (Faircloth et al., 2012) and "Anchored Hybrid Enrichment" (AHE) (Lemmon, Emme & Lemmon, 2012) approaches have been applied in phylogenetic analyses of animal (e.g. snakes, Pyron et al., 2014; lizards, Leaché et al., 2014; frogs, Peloso et al., 2016; and spiders, Hamilton et al., 2016) and plant (*Medicago,* De Sousa et al., 2014; *Sarracenia,* Stephens et al., 2015; palms, Comer et al., 2016; Heyduk et al., 2016; *Heuchera,* Folk, Mandel & Freudenstein, 2015; *Inga,* Nicholls et al., 2015; and *Protea,* Mitchell et al., 2017) clades.

Universal protocols are an attractive prospect, in terms of reduced cost and effort, and because they might generate broadly comparable data suitable for wider analyses (or even DNA barcoding; Blattner, 2016). However, the resulting sequence markers may not be optimal for all purposes. For phylogenetic inference, low-copy markers are required to avoid paralogy issues, and for successful hybridisation capture similarity of baits to target sequences must fall within c. 75-100% (Lemmon & Lemmon, 2013). This places a restriction on more universal markers that will necessarily exclude potentially useful low copy, high variability markers where these are subject to duplications or too variable in particular lineages.

The selection of appropriate sequence markers may therefore be crucial in determining the success of this kind of analysis, especially for non-model species. Transcriptome data for increasing numbers of non-model organisms are available (Matasci et al., 2014) and can already be used for marker selection in many plant clades. Bioinformatics tools are available that can assist in the selection of markers and design of baits, taking transcriptome and/or whole genome sequences of relevant taxa as input. These include MarkerMiner (Chamala et al., 2015), Hyb-Seq (Weitemier et al., 2014; Schmickl et al., 2016) and BaitsFisher (Mayer et al., 2016). The question for researchers embarking on phylogenomic analyses is whether it is worth the additional cost and effort involved in designing custom baits, and how to select sequence markers in order to get the most information out of a given investment of time and funds.

Our ongoing research addresses the challenge of resolving potentially complex phylogenetic

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

relationships between closely related populations and species of a non-model flowering plant group, the genus *Erica* (Ericaceae; one of 22 families of the asterid order Ericales; Stevens, 2001). The c. 700 South African species of *Erica* represent the most species rich 'Cape clade' in the spectacularly diverse Cape Floristic Region (Linder, 2003; Pirie et al., 2016). Analyses of the *Erica* clade as a whole offer a rich source of data in terms of numbers of evolutionary events, and our ability to infer such events accurately is arguably greatest in the most recently diverged species and populations. In such clades, the historical signal for shifts in key characteristics and geographic ranges are in general less likely to have been overwritten by subsequent shifts and (local) extinction. However, phylogenetic inference in rapid species radiations, such as that of Cape *Erica* (Pirie et al., 2016), Andean *Lupinus* (Hughes & Eastwood, 2006) or Lake Malawi cichlid fish (Santos & Salzburger, 2012) presents particular challenges. These include low sequence divergence confounded by the impact of both reticulation and coalescence on population-level processes. To infer a meaningful species tree under such circumstances, we need data suitable to infer multiple, maximally informative, independent gene trees.

The aims of this paper are to compare the performance of custom versus more universal approaches to marker selection for groups of closely related species/populations. Applying new scripts and a number of similar existing methods for marker selection, we compare predicted sequence lengths and variability of the resulting markers as estimates of their potential for delivering multiple independent and informative gene trees. We further compare different options implemented in our scripts for optimising e.g. intron numbers/lengths for a given number of baits. In so doing, we generate a tool for low-level phylogenetic inference in *Erica*, we test it experimentally by generating empirical data, and we assess its potential application across a wider group, e.g. the family Ericaceae.

Materials & Methods

Our first aim was to identify homologous, single-copy sequence markers for which we could design baits (probes) with similarity of ≥75% (as hybridization between target and probe tolerates a maximum of 25% divergence) that would be predicted to deliver the greatest numbers of informative characters. Baits currently represent a relatively large proportion of the total cost of the protocol (which is expensive on a per sample basis compared to e.g. PCR enrichment). We therefore restricted the total length of hybridisation baits to 692,400 bp

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

(5770 individual 120 bp baits), representing a total "capture footprint" (i.e. cumulative sequence length) of 173,100 bp given probe overlap representing 4x coverage. With our lab protocol (see below) this permits dilution of the baits to capture five samples per unit of baits instead of just one. We developed custom-made Python 2.7.6 scripts to identify the wider pool of all potential target sequences from transcriptome and WGS data (both of which were available from published sources; details below), as well as applying already available scripts/software for comparison. We subsequently implemented in further scripts different options for prioritising target variability, length and/or intron numbers and lengths to select optimal sequence markers from these pools of potential targets. We then compared the lengths and numbers of the sequences in the different resulting potential and optimal marker sets.

Identifying potential target sequences

Our custom-made script (AllMarkers.py; summarised in Fig. 1 available at Github: https://github.com/MaKadlec/Select-Markers/tree/AllMarkers) requires at least two transcriptomes, ideally of taxa closely related to the focal group. Where WGS/genome skimming data of one or more such taxa is available, it can be used too, as in Folk, Mandel & Freudenstein (2015). AllMarkers.py implements the following steps: First, two or more transcriptomes are compared to identify homologues, retaining those found in at least two transcriptomes. These are hence likely to also be found in related genomes. We have successfully used up to eight transcriptomes; on eight cores of a fast desktop PC the analyses ran for up to two days. Particularly when larger numbers of larger transcriptomes are compared, an additional filter can be applied prior to this step to remove shorter sequences (e.g. those <1,000 bp) and thereby improve speed. Next, multiple copy sequences are identified, for which homology assessment might be problematic. When WGS data is available, this is achieved using BLASTn of transcriptome against WGS. When no WGS data is available it is by comparison to the classification of proteins as single/mostly single copy across angiosperms by De Smet et al. (2013), using BLASTx following the approach used in MarkerMiner (Chamala et al., 2015). Multiple-copy sequences are then excluded. Finally, a filter for similarity ≥75% is applied. This series of steps is comparable to but differs from those implemented in Hyb-Seq (Weitemier et al., 2014) and in MarkerMiner (Chamala et al., 2015) (Fig. 1), which we also applied here.

The Hyb-Seq pipeline uses transcriptome and WGS sequences of closely related species to

PeerJ Preprints NOT PEER-REVIEWED

select marker sequences. This pipeline employs BLAT (BLAST-like Alignment Tool), rather than BLAST as in AllMarkers.py, to identify single-copy sequences with identity > 99%. After isoform identification, sequences with exons <120 bp and those of total length <960 bp are removed. This represents a further filtering of potential targets that is comparable in part to the next steps in our own scripts, as described below. Then orthologous sequences are identified using a transcriptome of a closest related species or of one of four angiosperms (*Arabidopsis thaliana, Oryza sativa, Populus trichocarpa* and *Vitis vinifera*), as opposed to by comparison to two or more transcriptomes in AllMarkers.py.

For MarkerMiner, WGS data is neither required (as in HybSeq) nor used if available (as in AllMarkers.py). This pipeline involves selecting sequences by size in input transcriptomes (we set length parameter to >1000 bp) then using reciprocal BLAST between transcriptomes and a reference proteome to select sequences above 70% similarity. The proteome most closely related to *Erica* implemented in MarkerMiner in August 2016 was that of *Vitis vinifera* (Vitaceae; Vitales; core eudicots; Stevens, 2001). This minimum similarity threshold does not directly reflect that required for successful probe hybridisation, and particularly given comparison to a relatively distantly related proteome (as opposed to more closely related transcriptomes with AllMarkers.py and HybSeq) can be expected to be conservative. In the final step, MarkerMiner retains putative single copy ortholog pairs following De Smet et al. (2013), as also implemented in AllMarkers.py when no WGS is available.

Selection of optimal target sequences from pools of potential targets

The above steps result in potentially large pools of potentially highly suboptimal targets, in particular shorter and/or invariable sequences that, given rapid lineage divergence, may not deliver enough informative characters to discern meaningfully resolved independent gene trees. In order to select optimal markers from these pools given a limited number of baits we designed a further script (available at Github: https://github.com/MaKadlec/Select-Markers/tree/BestMarkers.py). Depending on the phylogenetic problem to hand (e.g. recent, species level divergence versus older radiations) and available information (e.g. about sequence variability in the focal clade; positions and lengths of potentially more variable introns), various options are possible. In our case, from WGS and transcriptome data we know where introns are likely to be found, but in the absence of sequences from multiple accessions of our ingroup, the only indication of sequence variability comes from comparison

of coding regions of relatively distantly related taxa, i.e. single species of *Rhododendron*, *Vaccinium* and *Erica*. We therefore assessed two options: 1) simply selecting the longest sequences. 2) Selecting the longest sequences, but taking into account the (likely) additional length of introns. Using WGS data, we assessed the number and length of introns. For the purpose of ranking potential markers, we decided to use mean intron length in order to avoid favouring the selection of sequences with large introns that a) might not be efficiently captured/sequenced; or b) might not be so large in the focal clade. Finally, the longest sequences were selected that could be captured with our maximum number of baits. Coding regions <120 bp long are shorter than the baits and are likely to be ineffectively captured. For this reason, in the Hyb-Seq approach (Weitemier et al., 2014) all sequences including exons <120 bp are excluded; however, this is at the expense of excluding otherwise optimal markers that may include individual exons of <120 bp. We therefore opted to retain sequences including one or more coding regions ≥120 bp, whilst excluding individual exons <120 bp as potential targets for baits.

In silico *comparison with empirical data*

Our custom scripts (AllMarkers.py and BestMarkers.py), the Hyb-Seq and MarkerMiner pipelines were each applied to transcriptomes and (except for MarkerMiner) WGS of representatives of the Ericaceae subfamily Ericoideae. Transcriptome data was of *Rhododendron scopulorum* (18,307 gene sequences; 1KP project; Matasci et al., 2014) and (diploid) cranberry *Vaccinium macrocarpon* (48, 270 sequences, PRJNA260125 NCBI). WGS was of *V. macrocarpon* (PRJNA246586) and *Erica plukenetii* (Le Maitre & Bellstedt, unpublished data). We compared the (potential) length and identity of the resulting targets.

We then compared these “made to measure” (*Erica*/Ericoideae-specific) targets with those that might be selected using a more “one size fits all” (universal) approach to probe design. For this purpose, we used transcriptomes from increasingly distantly related plants as available on NCBI. First we included different families of the wider order Ericales: Actinidiaceae (*Actinidia chinensis*; 10,000 sequences; PRJNA277383), Primulaceae (*Aegiceras corniculatum*; 49,412 sequences; PRJNA269022), Theaceae (*Camellia reticulata* ; 139,145 sequences; PRJNA297756), Ebenaceae (*Diospyros lotus*; 413, 775 sequences; PRJNA261339), and Ericaceae (*R. scorpulum and V. macrocarpon*, as above). Then we expanded to different orders of eudicots: Ranunculales (*Anemone flaccida*; 46,945 sequences;

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PRJNA277332), Asterales (*Dahlia pinnata*; 35,638 sequences; PRJNA189243), Proteales (*Gevuina avellana*; 185,089 sequences; PRJNA299715), Caryophyllales (*Mesembryanthemum crystallinum*; 24,204 sequences; PRJNA217685), Solanales (*Solanum chacoense*; 42,873 sequences; PRJNA299204), Fabales (*Vigna radiata*; 78,617 sequences; PRJNA266360), Vitales (*Vitis vinifera*; 52,310 sequences; PRJNA239278) and Ericales (*R. scorpulum*, as above). Because in this wider context it is no longer appropriate to identify single copy markers on the basis of Ericoideae data alone, we instead used the option to compare to the angiosperm-wide database (De Smet et al., 2013) following an approach similar to MarkerMiner (Chamala et al., 2015). We compared the resulting targets to those of the *Erica*-specific approach, as above.

Generation of a novel empirical dataset

In order to confirm that our scripts can be used to obtain datasets of single-copy markers, we applied them to our empirical study on Cape *Erica*. We used the 132 sequences resulting from our custom scripts, taking into account the potential intron lengths (see results and discussion).

In addition to these targets, we added two additional markers that were not otherwise selected as optimal, for the purpose of comparison with other datasets. These were rpb2 (as used in phylogenetic reconstruction in *Rhododendron;* Goetsch, Eckert & Hall, 2005) and topoisomerase B (as proposed for use across flowering plants; Blattner, 2016).

Laboratory methods: Plant material was collected in the field under permit (Cape Nature: 0028-AAA008-00134; South Africa National Parks: CRC-2009/007-2014) or obtained from cultivation. DNA was extracted from one sample of *Rhododendron camtschaticum*, supplied by Dirk Albach and Bernhard von Hagen from collections of the Botanic Garden, Carl von Ossietzky Universität, Oldenburg, Germany; and 12 of *Erica* (Table 1) using Qiagen DNAeasy kits (Qiagen, Hilden, Germany). DNA extraction in *Erica* is generally challenging (Bellstedt et al., 2010) and the quantity and quality of DNA obtained differed considerably between species. To reach the correct amount of DNA required for library preparation, multiple DNA extractions from the same sample were combined.

For library preparation and hybridisation enrichment, we used the Agilent SureSelectXT

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

protocol (G7530-90000), incorporating sample-specific indexes for pooled sequencing, with a 1kb-499kb SureSelectXT *Custom* capture library designed using the SureDesign Custom Design Tool for NGS Target Enrichment, specifying 4x coverage and probe length 120 bp. For the library preparation, amount of gDNA used was between 1 and 3 µg, and during the hybridisation and capture step, we used a diluted capture library (1 part Agilent baits solution to 4 of ddH_2O). Sequencing was performed with Illumina NextSeq500 (StarSeq, Mainz, Germany) to generate 25 million paired-end reads of length 150 bp.

Bioinformatic analysis: As the total footprint of the capture library (the cumulative sequence lengths of all the selected markers) was small, *de novo* assembly was possible. We chose to use MIRA (version 4.0) (Chevreux, Wetter & Suhai, 1999), in part because MIRA can be used to perform both *de novo* assembly and mapping. The two options were used with default parameters for Illumina (overlap value=80 for *de novo* and 160 for mapping assembly; quality level=accurate). Reads were assembled into contiguous sequences (contigs). We then compared using BLASTn against the sequence targets (complete sequences and coding region sequences) as well as against nuclear ribosomal (nrDNA), plastid, and mitochondrial data. Contigs for which overlap with targets was under 100 bp and similarity to target sequences was less than 75% were removed. Using the L-INS-i (iterative refinement method incorporating local pairwise alignment information) method of MAFFT (Katoh et al., 2002), we aligned contigs with each other and with the sequence targets (complete sequences and coding region sequences). Contigs were checked with Gap5 (Bonfield & Whitwham, 2010) and by comparison to the alignments to identify and confirm remaining separate overlapping contigs without sequence differences. We used custom made scripts to merge and remove redundant contigs, combining only those with identical overlapping sequences (minimum overlap of 30 bp) or which differed by a single base only (in which case this position was coded with IUPAC ambiguity codes). Contigs differing by more than one base, or which did not overlap, were not combined. This should avoid combining non-continuous contigs representing different copies or alleles, at the cost of tending to overestimate the numbers of such copies where overlap of contigs is incomplete. We then attempted to add to the alignments any <100 bp sequences or sequences under 75% similarity that matched the target according to BLASTn, combining (or not) contigs using the same principles as above.

We excluded alignment positions representing indels or missing data in one or more samples and then calculated the percentage of variable sites per marker, including combined

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

mitochondrial and plastid sequences and individual nrDNA sequences representing Internal and External Transcribed Spacer regions (ITS and ETS) as obtained using Sanger sequencing in previous work (Pirie, Oliver & Bellstedt, 2011;Pirie et al., 2017). Gene trees were inferred using RAxML (Stamatakis, 2014) and used as a rough test for potential paralogy, under the assumption that the ingroup (comprising all samples except *Rhododendron* and the more closely related outgroups *Erica abietina and Erica plukenetii*) is monophyletic. We summarised 70% bootstrap consensus trees using DendroPy (Sukumaran & Holder, 2010) with SumTrees (https://github.com/jeetsukumaran/DendroPy).

Results

Similarity, length and overlap of selected markers: "made to measure" versus "one size fits all"

The lengths of sequences selected using the different scripts are presented in Fig. 2. Summary comparisons by method are presented in Table 2 (sequence numbers, lengths and similarity). In general, the additional filter that includes mean intron length resulted in an increased number of shorter targets that might nevertheless deliver greater final sequence lengths, if average lengths of flanking introns are effectively captured (Fig. 2).

Made to measure: We identified 4649 potential markers using our custom script AllMarkers.py. Applying script BestMarkers.py to this pool to optimise for length, two different subsets of optimal markers were obtained: 132 with median length (of coding region) of 2,187 bp when taking intron lengths into account; 79 of median length 2,631 bp when not. Sequence identity was similar (Table 2).

With the Hyb-Seq pipeline, 782 sequences were obtained, which after applying BestMarkers.py, was reduced to 55 of median length 2,157 bp when taking introns into account and 66 of median length 2,184 bp when not. Sequence identity was similar, and similar to that resulting from AllMarkers.py (Table 2).

With MarkerMiner, target sequences are delivered separately for each transcriptome provided. We selected a total pool of 544 potential target sequences, of which 389 are represented in the *R. scopulorum* data and 222 in *V. macrocarpon*. By comparison using our own scripts (available on request) we identified just 67 that were common to both (whereby it should be noted that AllMarkers.py by default retains only those found in at least two transcriptomes). Of the 544 sequences, 519 are indicated by MarkerMiner as mostly single copy and 25 as

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

strictly single copy in angiosperms. After applying BestMarkers.py we retained 254 sequence targets when taking introns into account and 207 sequences when not. Use of MarkerMiner resulted in the selection of greater numbers of shorter and slightly more conserved markers compared to both AllMarkers.py and HybSeq (Table 2, Figs. 2-3).

One size fits all: Applying AllMarkers.py/BestMarkers.py to transcriptomes of Ericales resulted in a pool of 2,354 potential markers and final datasets of 409 sequences when taking introns into account and 171 when not. With the Eudicot transcriptomes, the total pool included 461 potential markers and final datasets 249 (when taking introns into account) and 130 sequences (when not) (Table 2). In the latter, there is a slight increase in similarity (≥85%, similar to MarkerMiner; Fig. 3), and in both, sequences are shorter (Table 2, Fig. 2).

The numbers of markers in common given the different methods for selecting them, before and after applying BestMarkers.py are presented in Fig. 4. Fig. 4a illustrates both the low overlap and large differences in numbers between the complete pools of potential markers identified using the different methods/input data. Expanding in taxonomic scope from *Erica* (identifying single-copy genes on the basis of WGS data) to Ericales and to eudicots (adopting single copy markers from the database of De Smet et al. (2013) resulted in a decrease in numbers of potential markers, and the use of MarkerMiner a further decrease. Fig. 4b illustrates the differences in the optimal markers selected using BestMarkers.py on these pools. There is limited overlap and considerable differences in both target numbers and lengths: overall, AllMarkers.py/BestMarkers.py and HybSeq delivered the longest sequences, whereby the former delivered more markers for the same number of baits. Both the Ericales and eudicot analyses and MarkerMiner delivered greater numbers of shorter sequences.

Empirical data

We performed selective enrichment of 134 markers (132 selected using AllMarkers.py/BestMarkers.py, plus the two ‘universal’ markers added for the purposes of comparison). Exon sequences used for probe design are presented in supplementary data 1 and sequence alignments in supplementary data 2. Raw sequence reads are deposited on NCBI (PRJNA388814). With the exception of a single marker, capture was equally effective in the single *Rhododendron* sample and thirteen *Erica* samples. One marker was captured only in *Rhododendron*, and two others was not captured at all. All of the remaining 129 markers plus rpb2 and topoisomerase B were recovered, at least in part, from all thirteen

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

samples analysed (supplementary data 3). Of these, 6 were single copy without allelic variation; 83 included sequence polymorphisms corresponding to two distinguishable putative alleles in one or more (but not all) individual samples. A further 40 included sequence polymorphisms in all samples, which exhibited two or more copies. Of the latter 40, 28 represented paralogs that were easily distinguished on the basis of high sequence divergence in one or more coding region(s) and could thus be segregated into separate matrices of homologous sequences. The remainder (12) included multiple contigs that could not obviously be combined into single homologous sequences or pairs of alleles. Inspection of individual gene trees (supplementary data 4) failed to reject the monophyly of the ingroup in all but five cases.

Comparison of sequence length/variability was limited by uneven sequencing coverage, but we could confirm the capture of complete intron sequences of up to c. 1000 bp and partial introns/flanking non-coding regions of up to c. 500 bp. In addition, large stretches of homologous high copy nuclear ribosomal and mitochondrial sequences were captured for all samples, as well as more fragmented plastid sequences.

Despite incomplete sequencing coverage, the average alignment length of single copy nuclear sequences was 1810 bp, with a range between 823 and 5574 bp. With all gaps and missing data excluded (resulting in alignments of between 327 and 4716 bp), the single copy nuclear sequences in the ingroup presented between 5 and 412 variable positions each, representing a range of 2.6 – 26.1 % variability. Variability of rpb2 was 3.4%; topoisomerase B: 7.5%; ETS: 22.1%; ITS: 17.9%; mitochondrial: 6.3%; and plastid sequences: 0.54%. A plot of original predicted length of markers (instead of real length since in most cases complete sequences were not obtained) against variability is presented in Fig. 5. There was no obvious relationship between sequence length and variability. A further plot of observed sequence variability against variability of the corresponding transcriptome data (*Rhododendron* compared to *Empetrum*) is presented in Appendix 1; there was also no obvious relationship. Gene trees inferred under ML are documented in Supplementary Material 3 (with further details in Supplementary Material 4), with eight based on selected markers (ITS, mitochondrion, and six single copy nuclear markers that delivered the greatest numbers of clades supported by ≥70% BS) illustrated in Fig. 6.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

Discussion

Comparing closely versus distantly related genomes for marker selection

It seems intuitively obvious that optimal markers for a given phylogenetic problem will be those informed by comparison to transcriptomes/WGS of the most closely related representative taxa. With such data, lineage specific gene duplications can be identified and the number of potential targets of appropriate variability maximised. However, the genomic data available for a given focal group (such as transcriptome data from the 1KP project; Matasci set al., 2014) may represent taxa more or less distantly related to it, and particular researchers may or may not wish to go to the trouble of designing and applying custom protocols. Indeed, if an off-the-shelf tool will provide appropriate data, it would be a great deal simpler just to use it. Hence, before embarking on expensive and time-consuming lab procedures, we need to know to what degree targets designed for one group might be applied to more distantly related ones (e.g. in this case the utility of *Erica* baits across Ericaceae, or Ericales); and conversely, how suboptimal baits designed for universal application (e.g. across angiosperms) are likely to be for a given subclade.

Using our own custom scripts, we compared the pools of markers that might be selected on the basis of comparison of relatively closely related genomes with those on the basis of more distantly related ones (i.e. within the subfamily Ericoideae as opposed to within the order Ericales or across eudicots). Our results showed that both the pools and the best marker sets from those pools differed considerably, and that the sequences of the latter were considerably shorter (Table 2, Figs. 2 and 3). On the other hand, sequence variability within Ericales (minimum sequence identity between Ericaceae and Actinidiaceae: 73%) suggests that baits designed for *Erica* are also potentially suited for use at least across Ericaceae, including in *Rhododendron* and *Vaccinium* (both species-rich genera for which such tools might be particularly useful (Kron, Powell & Luteyn, 2002; Goetsch, Eckert & Hall, 2005). In general, our results confirm both the greater potential of custom baits developed for specific clades; and show that once obtained, such tools are nevertheless likely to apply across a fairly broad range of related taxa.

The impact of method for marker selection

Having decided to design custom baits, the next question that we might ask is which method to use for probe selection/design. Our results suggest that this is also likely to have a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

significant impact on the resulting datasets. We compared three approaches to marker selection: our own custom scripts; those presented in the Hyb-Seq approach (Weitemier et al., 2014) and MarkerMiner (Chamala et al., 2015).

Of these three, MarkerMiner is arguably the most user-friendly, which is important given that its user base ought ideally to include biologists without extensive bioinformatics skills. However, in our comparisons it delivered the shortest sequence lengths (Table 2). The reasons for this are two-fold. First (and perhaps most importantly), because the transcriptomes used, irrespective of their similarity one to another, are compared to what is likely to be a rather distantly related proteome. Second, because the approach for identifying single or low-copy markers involves comparison to a general database (in this case for flowering plants), rather than a case-by-case assessment. Hence, in the current implementation of MarkerMiner it is to be expected that the most variable sequences will be excluded. So will some that are single copy in the focal group (or with easily discerned paralogs, as was the case here and also at lower taxonomic levels in Budenhagen et al. 2016), but not in other clades; and some that are multiple-copy may in fact be included. This is reflected in our results by the low number of potential target sequences recovered in total; in the low proportion of those that were recovered also being recovered using our own custom scripts and Hyb-Seq; and in the lower sequence length: the removal of more variable sequences arbitrarily results in the removal of longer ones too (Table 2). This phenomenon is apparently also reflected in the even shorter sequences reported by Budenhagen et al. (2016), using universal angiosperm probes (average 764 bp, derived from targets averaging 343 bp).

The Hyb-Seq approach is more similar to our own, but nevertheless results in a different dataset of selected sequences. The main differences lie in the search tool and filters. Our script uses BLAST, whereas Hyb-Seq uses BLAT. BLAT is faster than BLAST, but needs an exact or nearly-exact match to return a hit. Significantly, the exclusion in HybSeq of all sequences including any exons <120 bp is at the loss of markers including variable introns; in our approach the problem of short exon/probe mismatch is avoided simply by ignoring such exons during probe design. The net result is that while both approaches deliver long target sequences, ours can deliver those including more introns (which can therefore be captured using fewer baits).

Selecting optimal markers from within a pool of potential candidates

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

Our approach includes not just a means to select potentially appropriate markers (AllMarkers.py; as is the case with the other approaches compared) but also a second step (BestMarkers.py) that selects putatively optimal markers from amongst that pool. Obviously, it is possible to capture and sequence the entire pool (following Ilves & Lopez-Fernandez, 2014; Mandel et al., 2014; Weitemier et al., 2014). However, by targeting a smaller number of the most appropriate markers, more samples can be analysed less expensively. A given bait solution can be used for a greater number of samples (because it includes fewer different baits, each at higher concentration), whilst sequencing effort can be reduced by eliminating a potentially large number of less informative (or perhaps even entirely uninformative) markers.

AllMarkers.py identifies and reports the positions of introns from comparison of WGS to transcriptome data. Subsequently optimising for intron numbers/length, as implemented in BestMarkers.py, would seem appropriate for the purpose of identifying regions that are likely to be both longer and more variable (Folk, Mandel & Freudenstein, 2015). Hybrid capture can result in sequencing of potentially long stretches of flanking regions (Tsangaras et al., 2014) without requiring matching baits, and introns should be less constrained, possibly with informative length variation too. Hence, taking into account the additional length of introns in marker selection can result in greater numbers of longer (and likely more variable) obtained sequences. Our empirical results support this approach: sequences showed intron capture of up to 1,000 bp, including regions in which multiple introns are interspersed with short (<120 bp) exons for which no probes were used. Intron sequences from WGS data can nevertheless be included in the output of AllMarkers.py and used to design probes. This may be effective at low taxonomic levels when WGS appropriate to assess sequence similarity within the focal group is available. Alternatively, if the problem to be addressed represents older divergences (e.g. phylogenetic uncertainty within Ericaceae; Freudenstein, Broe & Feldenkris, 2016) for which length variation in introns would be unhelpful, BestMarkers.py can be used to optimise the length of exons alone.

An alternative to optimising for sequence length (with or without taking introns into account) would be to optimise for variability (or combined length and variability). We included this option in BestMarkers.py, but in the absence of data with which to compare within our ingroup, decided *a priori* that we would be more likely to optimise total per sequence variation by selecting on the basis of length alone. This decision was supported by the empirical results: as might be expected, there was no obvious relationship between sequence length and variability (Fig. 5) and the numbers of informative characters provided by a given

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

target could not be predicted from the similarity of the *Vaccinium* and *Rhododendron* transcriptomes (Appendix 1).

The variability of the data we obtained can be compared to that of nrDNA, plastid and mitochondrial sequences (and which were also obtained here without the need for matching baits due to their high copy number) and to two generally single copy nuclear genes, topoisomerase B and rpb2 (Fig. 5). Consistent with the results presented by Nichols et al. (2015), the variability of the nrDNA spacer regions (ITS and ETS) that are frequently used in empirical studies of plants is at the upper end of that observed in the sequences we obtained (of which topoisomerase B and rpb2 were fairly typical); plastid (and mitochondrial) sequences at the lower end. Given the comparably modest variability of most alternative nuclear markers, this suggests that even in cases where ITS/ETS present sufficient information to infer a well resolved nrDNA gene tree (not the case in Cape *Erica*, Pirie et al., 2011; Fig. 6), considerably longer sequences will be needed to infer comparably resolved independent gene trees. Difficult phylogenetic problems arise when gene trees can be expected to differ, but those inferred from standard markers are not sufficiently resolved to actually reveal it. Low information content of individual markers limits accuracy of species tree inference methods (Lanier, Huang & Knowles, 2014), and when relationships are contentious, resolution can be influenced disproportionately by small numbers of individual markers or sites (Shen, Hittinger & Rokas, 2017). These are the cases for which targeted capture approaches offer the greatest potential. We need to target markers that might deliver a forest of trees, rather than just more bushes, and not all targeted enrichment strategies are optimised to deliver this kind of data.

Conclusions

When sequence variation is appropriate and gene trees are consistent, standard Sanger sequencing of a small number of markers may be all that is required to infer robust and meaningful phylogenetic trees. For species complexes and rapid radiations (either ancient or recent) where this is not the case, the usefulness of sequence datasets will inevitably be limited by the resolution of individual gene trees. Our results suggest that under these circumstances, where the need for NGS and targeted sequence capture, such as hybrid enrichment, is greatest, “made to measure” markers identified using both transcriptome and WGS data of related taxa will deliver results that are superior to those that might be obtained

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

using a more universal "one size fits all" approach. Once available, such markers may nevertheless be useful across a fairly wide range of related taxa: e.g. those presented here, targeted for use in *Erica*, fall within the range of sequence variation that would in principle be applicable across the family Ericaceae. Transcriptome data for many flowering plant groups are now available; these would ideally (but not necessarily) be complemented with WGS or genome skimming data of one or more focal taxa for use in marker selection. With such data to hand, biologists are still reliant on bioinformatics skills or user-friendly tools (such as MarkerMiner). In either case, the full potential of the techniques will only be harnessed if comparisons to distantly related genomes and generalisations of single/low copy genes across wide taxonomic groups are avoided. We would conclude that rather than searching for "one size fits all" universal markers, we should be improving and making more accessible the tools necessary for developing our own "made to measure" ones.

References

Bellstedt DU., Pirie MD., Visser JC., de Villiers MJ., Gehrke B. 2010. A rapid and inexpensive method for the direct PCR amplification of DNA from plants. *American Journal of Botany*. DOI: 10.3732/ajb.1000181.

Blattner FR. 2016. TOPO6: a nuclear single-copy gene for plant phylogenetic inference. *Plant Systematics and Evolution*. DOI: 10.1007/s00606-015-1259-1.

Bonfield JK., Whitwham A. 2010. Gap5-editing the billion fragment sequence assembly. *Bioinformatics*. DOI: 10.1093/bioinformatics/btq268.

Budenhagen C., Lemmon AR., Lemmon EM., Bruhl J., Cappa J., Clement WL., Donoghue M., Edwards EJ., Hipp AL., Kortyna M., Mitchell N., Moore A., Prychid CJ., Segovia-Salcedo MC., Simmons MP., Soltis PS., Wanke S., Mast A. 2016. Anchored Phylogenomics of Angiosperms I: Assessing the Robustness of Phylogenetic Estimates. *bioRxiv*:86298. DOI: 10.1101/086298.

Chamala S., García N., Godden GT., Krishnakumar V., Jordon-Thaden IE., De Smet R., Barbazuk WB., Soltis DE., Soltis PS. 2015. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in plant sciences* 3:1400115. DOI: 10.3732/apps.1400115.

Chevreux B., Wetter T., Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) '99, GCB, Hannover, Germany.*:45–56. DOI: 10.1.1.23/7465.

Comer JR., Zomlefer WB., Barrett CF., Stevenson DW., Heyduk K., Leebens-Mack JH. 2016. Nuclear phylogenomics of the palm subfamily Arecoideae (Arecaceae). *Molecular Phylogenetics and Evolution* 97:32–42. DOI: 10.1016/j.ympev.2015.12.015.

Elshire RJ., Glaubitz JC., Sun Q., Poland JA., Kawamoto K., Buckler ES., Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6.

Faircloth BC., McCormack JE., Crawford NG., Harvey MG., Brumfield RT., Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61:717–726.

Folk RA., Mandel JR., Freudenstein J V. 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.

Freudenstein J V., Broe MB., Feldenkris ER. 2016. Phylogenetic relationships at the base of Ericaceae : Implications for vegetative and mycorrhizal evolution. *Taxon* 65:1–11. DOI: 10.12705/654.7.

Goetsch L., Eckert AJ., Hall BD. 2005. The Molecular Systematics of <I>Rhododendron</I> (Ericaceae): A Phylogeny Based Upon <I>RPB2</I> Gene Sequences. *Systematic Botany* 30:616–626. DOI: 10.1600/0363644054782170.

Hamilton CA., Lemmon AR., Lemmon EM., Bond JE. 2016. Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. *BMC Evolutionary Biology* 16:212. DOI: 10.1186/s12862-016-0769-y.

Hart M., Forrest LL., Nicholls J., Kidner C. 2016. Robust reconstruction of hundreds of nuclear loci from herbarium specimens. *Taxon* in review:1081–1092. DOI: http://dx.doi.org/10.12705/655.9.

Heyduk K., Trapnell DW., Barrett CF., Leebens-Mack J. 2016. Phylogenomic analyses of species relationships in the genus Sabal (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society* 117:106–120.

Hughes C., Eastwood R. 2006. Island radiation on a continental scale: exceptional rates of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

plant diversification after uplift of the Andes. *Proceedings of the National Academy of Sciences of the United States of America* 103:10334–10339. DOI: 10.1073/pnas.0601928103.

Hughes CE., Eastwood RJ., Bailey CD. 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 361:211–225. DOI: 10.1098/rstb.2005.1735.

Ilves KL., Lopez-Fernandez H. 2014. A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. *Molecular Ecology Resources* 14:802–811.

Johnson MTJ., Carpenter EJ., Tian Z., Bruskiewich R., Burris JN., Carrigan CT., Chase MW., Clarke ND., Covshoff S., Depamphilis CW., Edger PP., Goh F., Graham S., Greiner S., Hibberd JM., Jordon-Thaden I., Kutchan TM., Leebens-Mack J., Melkonian M., Miles N., Myburg H., Patterson J., Pires JC., Ralph P., Rolf M., Sage RF., Soltis D., Soltis P., Stevenson D., Stewart CN., Surek B., Thomsen CJM., Villarreal JC., Wu X., Zhang Y., Deyholos MK., Wong GK-S. 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PloS one* 7:e50226. DOI: 10.1371/journal.pone.0050226.

Jones MR., Good JM. 2016. Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*. DOI: 10.1111/mec.13304.

Katoh K., Misawa K., Kuma K., Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30:3059–3066. DOI: 10.1093/nar/gkf436.

Kingman J. 1982. On the Genealogy of Large Populations. *Journal of applied probability* 19:27–43. DOI: 10.2307/3213548.

Kron KA., Powell EA., Luteyn JL. 2002. Phylogenetic relationships within the blueberry tribe (Vaccinieae, Ericaceae) based on sequence data from matK and nuclear ribosomal ITS regions, with comments on the placement of Satyria. *American Journal of Botany* 89:327–336. DOI: 10.3732/ajb.89.2.327.

Lanier HC., Huang H., Knowles LL. 2014. Molecular Phylogenetics and Evolution How low can you go ? The effects of mutation rate on the accuracy of species-tree estimation. *Molecular Phylogenetics and Evolution* 70:112–119. DOI:

10.1016/j.ympev.2013.09.006.

Leaché AD., Wagner P., Linkem CW., Böhme W., Papenfuss TJ., Chong RA., Lavin BR., Bauer AM., Nielsen S V., Greenbaum E., Rödel MO., Schmitz A., LeBreton M., Ineich I., Chirio L., Ofori-Boateng C., Eniang EA., Baha El Din S., Lemmon AR., Burbrink FT. 2014. A hybrid phylogenetic-phylogenomic approach for species tree estimation in african agama lizards with applications to biogeography, character evolution, and diversification. *Molecular Phylogenetics and Evolution* 79:215–230. DOI: 10.1016/j.ympev.2014.06.013.

Lemmon AR., Emme SA., Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61:727–744.

Lemmon EM., Lemmon AR. 2013. High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 44:99–121.

Linder HP. 2003. The radiation of the Cape flora, southern Africa. *Biological reviews of the Cambridge Philosophical Society* 78:597–638. DOI: Doi 10.1017/S1464793103006171.

Mamanova L., Coffey AJ., Scott CE., Kozarewa I., Turner EH., Kumar A., Howard E., Shendure J., Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nature methods* 7:111–8. DOI: 10.1038/nmeth.1419.

Mandel JR., Dikow RB., Funk V a., Masalia RR., Staton SE., Kozik A., Michelmore RW., Rieseberg LH., Burke JM. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences* 2:1300085.

Matasci N., Hung L-H., Yan Z., Carpenter EJ., Wickett NJ., Mirarab S., Nguyen N., Warnow T., Ayyampalayam S., Barker M., Burleigh JG., Gitzendanner MA., Wafula E., Der JP., dePamphilis CW., Roure B., Philippe H., Ruhfel BR., Miles NW., Graham SW., Mathews S., Surek B., Melkonian M., Soltis DE., Soltis PS., Rothfels C., Pokorny L., Shaw JA., DeGironimo L., Stevenson DW., Villarreal JC., Chen T., Kutchan TM., Rolf M., Baucom RS., Deyholos MK., Samudrala R., Tian Z., Wu X., Sun X., Zhang Y., Wang J., Leebens-Mack J., Wong GK-S. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* 3:17.

Mayer C., Sann M., Donath A., Meixner M., Podsiadlowski L., Peters RS., Petersen M., Meusemann K., Liere K., Wägele J-W., Misof B., Bleidorn C., Ohl M., Niehuis O. 2016.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

BaitFisher: A software package for multi-species target DNA enrichment probe design. *Molecular Biology and Evolution*:1–27. DOI: 10.1093/molbev/msw056.

Miller MR., Dunham JP., Amores A., Cresko WA., Johnson EA. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17:240–248.

Mitchell N., Lewis PO., Lemmon EM., Lemmon AR., Holsinger KE. 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *American Journal of Botany* 104:102–115. DOI: 10.3732/ajb.1600227.

Nicholls JA., Pennington RT., Koenen EJM., Hughes CE., Hearn J., Bunnefeld L., Dexter KG., Stone GN., Kidner CA. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae: Mimosoideae). *Frontiers in plant science* 6:710.

Peloso PL V., Frost DR., Richards SJ., Rodrigues MT., Donnellan S., Matsui M., Raxworthy CJ., Biju SD., Lemmon EM., Lemmon AR., Wheeler WC. 2016. The impact of anchored phylogenomics and taxon sampling on phylogenetic inference in narrow-mouthed frogs (Anura, Microhylidae). *Cladistics* 32:113–140.

Pirie MD., Oliver EGH., Bellstedt DU. 2011. A densely sampled ITS phylogeny of the Cape flagship genus Erica L. suggests numerous shifts in floral macro-morphology. *Molecular Phylogenetics and Evolution* 61:593–601.

Pirie MD., Oliver EGH., Gehrke B., Heringer L., Mugrabi de Kuppler A., Le Maitre NC., Bellstedt DU. 2017. Underestimated regional species diversity in the Cape Floristic Region revealed by phylogenetic analysis of the Erica abietina/E. viscaria-clade (Ericaceae). *Botanical Journal of the Linnean Society* in press. DOI: 10.1093/botlinnean/box021.

Pirie MD., Oliver EGH., Mugrabi de Kuppler A., Gehrke B., Le Maitre N., Kandziora M., Bellstedt DU. 2016. The biodiversity hotspot as evolutionary hot-bed: spectacular radiation of Erica in the Cape Floristic Region. *BMC Evolutionary Biology* 16:190. DOI: 10.1186/s12862-016-0764-3.

Pyron RA., Hendry CR., Chou VM., Lemmon EM., Lemmon AR., Burbrink FT. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). *Molecular Phylogenetics and Evolution* 81:221–231. DOI: 10.1016/j.ympev.2014.08.023.

Saiki RK., Scharf S., Faloona F., Mullis KB., Horn GT., Erlich HA., Arnheim N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230:1350–1354.

Sang T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical reviews in biochemistry and molecular biology* 37:121–147. DOI: 10.1080/10409230290771474.

Sanger F., Nicklen S., Coulson a R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74:5463–7. DOI: 10.1073/pnas.74.12.5463.

Santos ME., Salzburger W. 2012. How Cichlids Diversify. *Science* 338:619–621. DOI: 10.1126/science.1224818.

Schmickl R., Liston A., Zeisek V., Oberlander K., Weitemier K., Straub SCK., Cronn RC., Dreyer LL., Suda J. 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African Oxalis (Oxalidaceae). *Molecular Ecology Resources* 16:1124–1135.

Shen X-X., Hittinger CT., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*. DOI: 10.1038/s41559-017-0126.

De Smet R., Adams KL., Vandepoele K., Van Montagu MCE., Maere S., Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* 110:2898–903. DOI: 10.1073/pnas.1300127110.

De Sousa F., Bertrand YJK., Nylinder S., Oxelman B., Eriksson JS., Pfeil BE. 2014. Phylogenetic properties of 50 nuclear loci in Medicago (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. *PLoS ONE* 9.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Stephens JD., Rogers WL., Heyduk K., Cruse-Sanders JM., Determann RO., Glenn TC.,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

Malmberg RL. 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus Sarracenia using target enrichment. *Molecular Phylogenetics and Evolution* 85:76–87.

Stevens PF. 2001.Angiosperm Phylogeny Website. Version 12, July 2012

Sukumaran J., Holder MT. 2010. DendroPy: A Python library for phylogenetic computing. *Bioinformatics*. DOI: 10.1093/bioinformatics/btq228.

Taberlet P., Gielly L., Pautou G., Bouvet J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* 17:1105–1109.

Tsangaras K., Wales N., Sicheritz-Pontén T., Rasmussen S., Michaux J., Ishida Y., Morand S., Kampmann ML., Gilbert MTP., Greenwood AD. 2014. Hybridization capture using short PCR products enriches small genomes by capturing flanking sequences (CapFlank). *PLoS ONE*. DOI: 10.1371/journal.pone.0109101.

Uribe-Convers S., Settles ML., Tank DC. 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: Resolving the diversity within the south American species of Bartsia L. (Orobanchaceae). *PLoS ONE* 11.

Wang Z., Gerstein M., Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* 10:57–63.

Weitemier K., Straub SCK., Cronn RC., Fishbein M., Schmickl R., McDonnell A., Liston A. 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. *Applications in Plant Sciences* 2:1400042. DOI: 10.3732/apps.1400042.

White TJ., Bruns S., Lee S., Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: *PCR Protocols: A Guide to Methods and Applications*. 315–322. DOI: citeulike-article-id:671166.

Wickett NJ., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker MS., Burleigh JG., Gitzendanner MA., Ruhfel BR., Wafula E., Der JP., Graham SW., Mathews S., Melkonian M., Soltis DE., Soltis PS., Miles NW., Rothfels CJ., Pokorny L., Shaw AJ., DeGironimo L., Stevenson DW., Surek B., Villarreal JC., Roure B., Philippe H., dePamphilis CW., Chen T., Deyholos MK., Baucom RS., Kutchan TM., Augustin MM., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong GK-S., Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

the United States of America 111:E4859-68. DOI: 10.1073/pnas.1323926111.

Zimmer EA., Wen J. 2015. Using nuclear gene data for plant phylogenetics: Progress and prospects II. Next-gen approaches. *Journal of Systematics and Evolution* 53:371–379.

Acknowledgements

The authors thank Cape Nature and South Africa National Parks for assistance with permits (Cape Nature: 0028-AAA008-00134; South Africa National Parks: CRC-2009/007-2014); Mark Chase and the 1,000 Plants (1KP) project for access to *Rhododendron* transcriptome data; Kai Hauschulz (Agilent), Abigail Moore (University of Oklahoma), and Frank Blattner, Nadine Bernhardt and Katja Herrmann (IPK Gatersleben) for help and advice with lab protocols; and reviewers Daniel Campo and Ryan Folk and handling editor Keith Crandall for constructive criticism.

Table 1: Samples used for DNA extraction and their collection localities. Vouchers were lodged at herbarium NBG (MP: Pirie).

Voucher	Sample #	Species	Locality (unless specified, within the Western Cape, South Africa)
MP1320	*78*	*E. abietina* L. ssp. *aurantiaca*	Du Toit's Pass
MP1330	*74*	*E. coccinea* L.	RZE, Greyton
MP1336	*81*	*E. coccinea* L.	Groot Hagelkraal
MP1318	*72*	*E. imbricata* L.	Flouhoogte
MP1319	*73*	*E. imbricata* L.	Stellenbosch
MP1334	*74*	*E. imbricata* L.	Groot Hagelkraal
MP1311	*69*	*E. imbricata* L.	Boskloof
MP1312	*80*	*E. lasciva* Salisb.	Boskloof
MP1325	*83*	*E. lasciva* Salisb.	Albertinia
MP1309	*71*	*E. penicilliformis* Salisb.	Boskloof
MP1339	*75*	*E. placentiflora* Salisb.	Cape Hangklip
MP1333	*82*	*E. plukenetii* L.	Groot Hagelkraal
	68	*R. camtschaticum* Pall.	Oldenburg Botanical Garden, Germany (cultivated)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

783 Table 2: Range, median and average length of selected markers in *Rhododendron*, with and
784 without taking introns into account, and similarities to homologues in *Vaccinium.*

		Length of CR (bp)		Similarity (%)		Predicted length (bp)	
		Range	Mean Median sd	Range	Mean Median sd	Range	Mean Median sd
Erica	AllMarkers.py (without intron length) - 79 seq	2316-4815	2834 2631 535	82-96	90 90 3,1	2412-7425	3541 3342 998
	AllMarkers.py (with intron length) - 132 seq	1053-4815	2287 2187 736	81-97	91 92 3,5	2847-7425	3579 3339 773
	HybSeq (without intron length) - 66 seq	1170-4146	2350 2184 549	77-95	89 91 5	1839-9326	3285 3013 1181
	HybSeq (with intron length) - 55 seq	993-4146	2226 2157 719	77-95	89 91 5	2943-9326	3835 3614 1032
	MarkerMiner (without intron length) - 207 seq	1293-4146	1726 1596 419	85-97	93 94 2	1338-5849	2411 2307 649
	MarkerMiner (with intron length) - 254 seq	1011-4146	1600 1518 454	85-97	93 94 2	1665-5849	2329 2210 611
Ericales	AllMarkers.py (without intron length) - 171 seq	1002-4146	1400 1266 460	82-97	93 93 2,6	1014-8546	2389 2121 1153
	AllMarkers.py (with intron length) - 408 seq	342-4146	1014 924 458	82-97	93 93 2,3	1003-8546	1830 1623 928
Eudicots	AllMarkers.py (without introns length) - 130 seq	1002-4146	1427 1283 487	85-97	93 93 2,4	1014-7657	2379 2093 1089
	AllMarkers.py (with introns length) - 249 seq	369-4146	1112 1017 494	85-97	93 94 2,2	1002-7647	1895 1689 960

785

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

Figures:

Figure 1: Flowchart(s) illustrating the methods used for marker selection.

Figure 2: Summary of a) exon lengths and b) predicted exon plus intron lengths of markers selected using AllMarkers.py (shades of green), Hyb-Seq (blue) and MarkerMiner (purple) followed by BestMarkers.py. Each pair of plots represents the markers selected when optimising for exon lengths (left) and predicted exon plus intron lengths (right). From left to right, the first three pairs represent markers targeted for Erica/Ericoideae (comparing by method); the final two for Ericales and eudicots respectively (using AllMarkers.py only).

Figure 3: Length versus variability of potential sequence markers (grey dots) and those selected using BestMarkers.py from the pools generated by the different methods (coloured symbols).

Figure 4: Venn diagrams produced using http://bioinformatics.psb.ugent.be/webtools/Venn/ comparing overlap in markers selected given the different methods, superimposed with their numbers. a) The complete pools of potential markers; b) the subsets of markers selected using BestMarkers.py, optimising for total predicted length (exons and introns).

Figure 5: Sequence variability observed in the empirical data plotted against predicted sequence length. “Universal” markers rpb2 and topoisomerase B are indicated and plastid, mitochondrial and nrDNA are included with indication of sequence lengths derived from the literature.

Figure 6: Selected 70% bootstrap support (BS) consensus gene trees inferred under maximum likelihood with RAxML, summarised with DendroPy/SumTrees and presented using Dendroscope 3.5.7 (http://dendroscope.org/). The six nuclear markers that delivered the greatest numbers of nodes supported by ≥70% BS are presented along with those based on ITS and mitochondrial sequences. Terminals correspond to collection codes and species names (Table 1). Some taxa are represented twice in some trees due to the presence of alleles, including two distinct copies of ITS in *E. abietina* ssp. *aurantiaca* (confirming previous work using cloning; Pirie et al., in press). Node labels represent bootstrap support.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

Appendices:

Appendix 1: Plot of sequence similarity (transcriptome data; *Rhododendron* and *Vaccinium*) against sequence similarity (empirical dataset generated here; *Rhododendron* and *Erica spp*.) for individual markers.

Supplementary data:

Supplementary data 1: Exon sequences corresponding to the 134 markers selected for the empirical study and the complete pools of markers selected using each of the methods compared (fasta format).

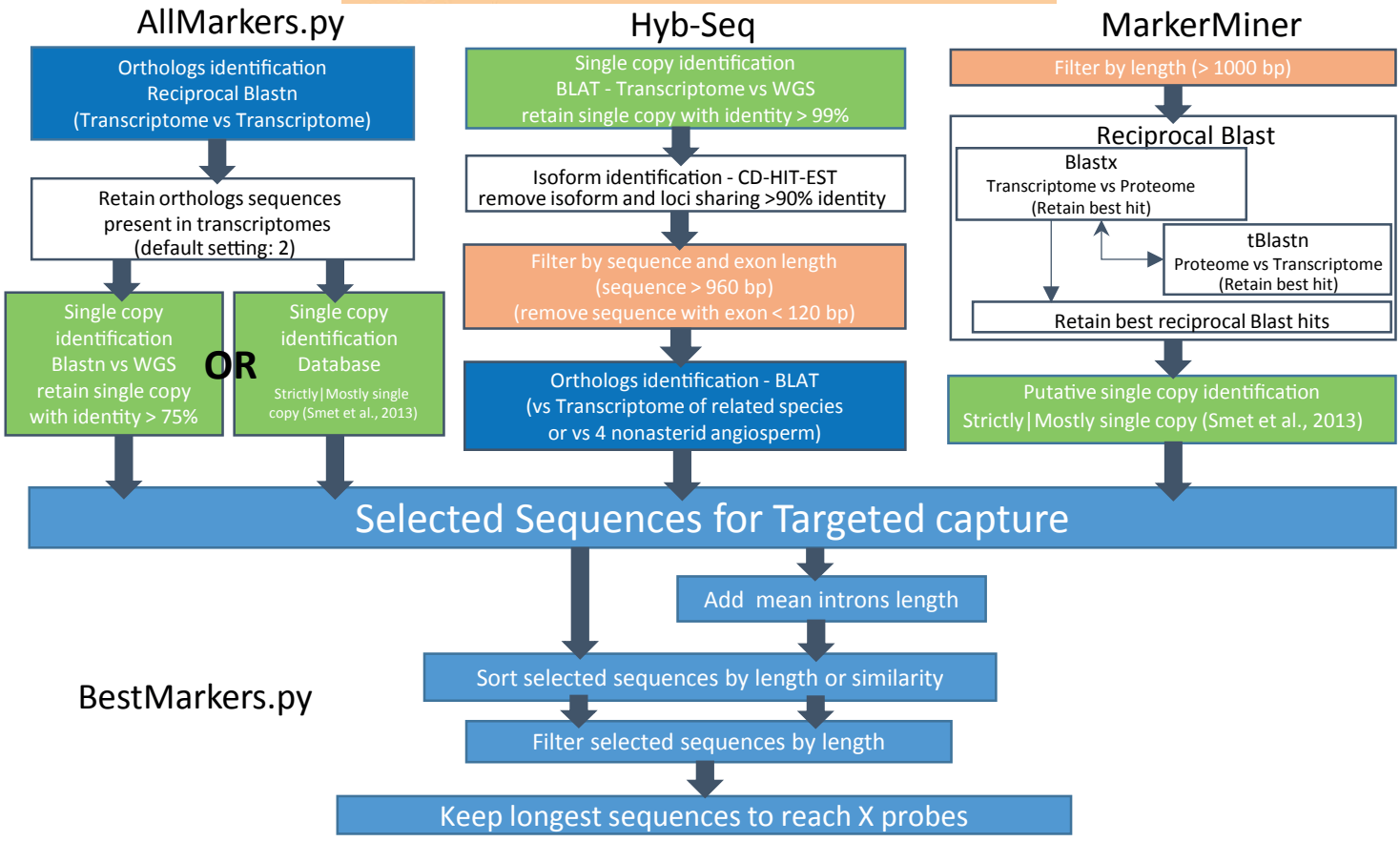
Supplementary data 2: Sequence alignments

Supplementary data 3: Table documenting markers as represented in Supplementary data 1-2 and 4.

Supplementary data 4: Gene trees inferred under RAxML (excluding multiple copy markers for which paralogues could not be distinguished).

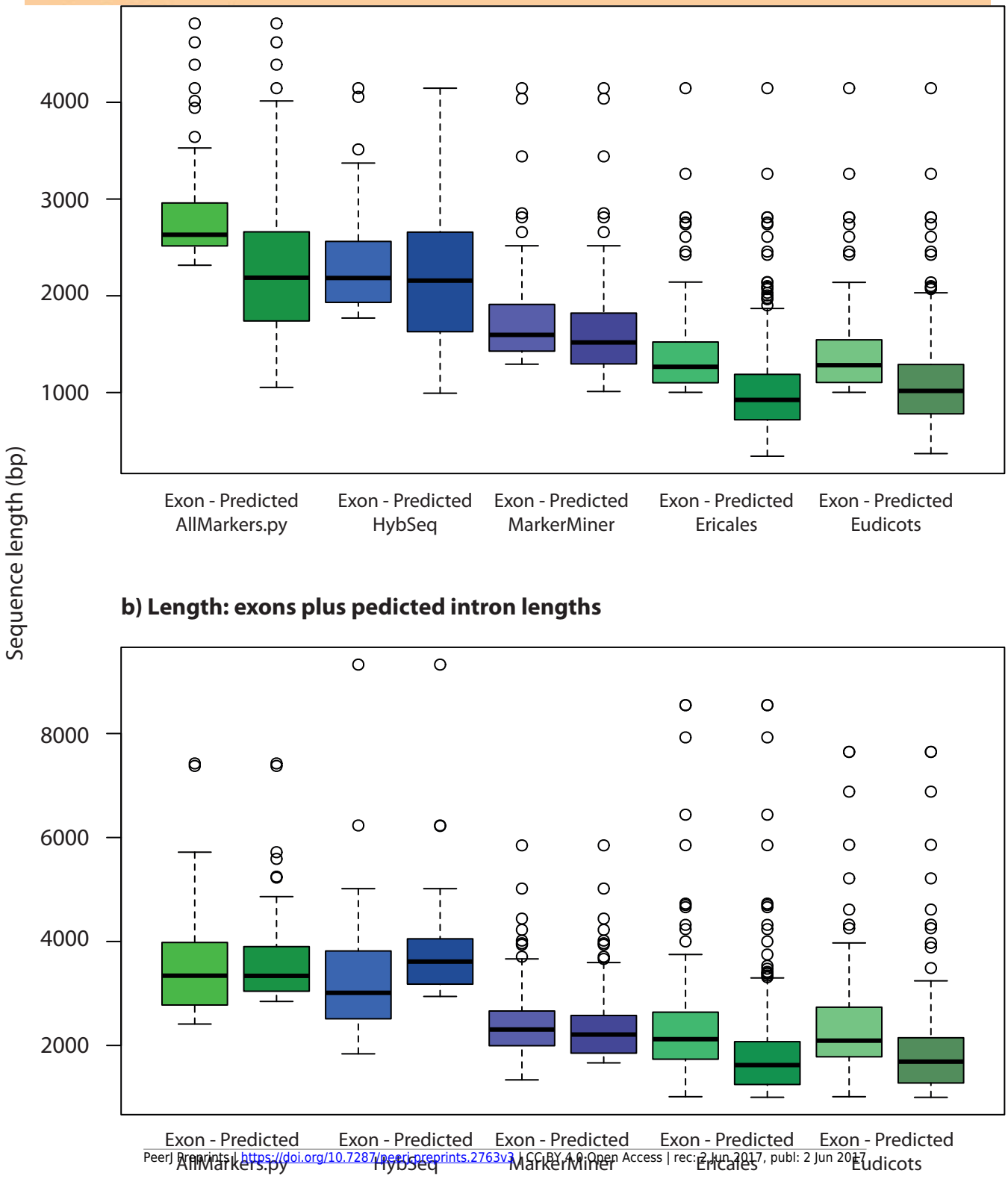
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED



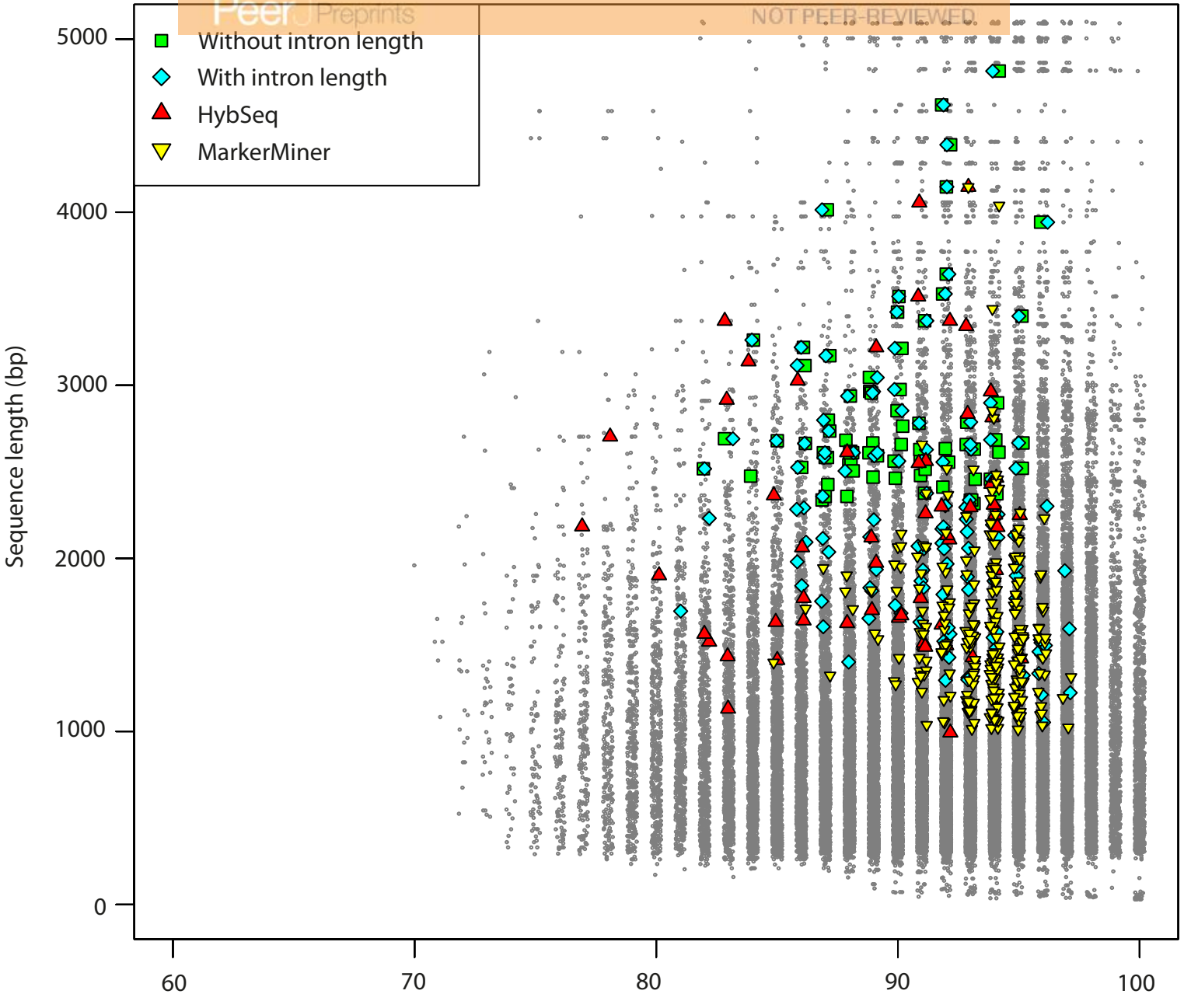
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED



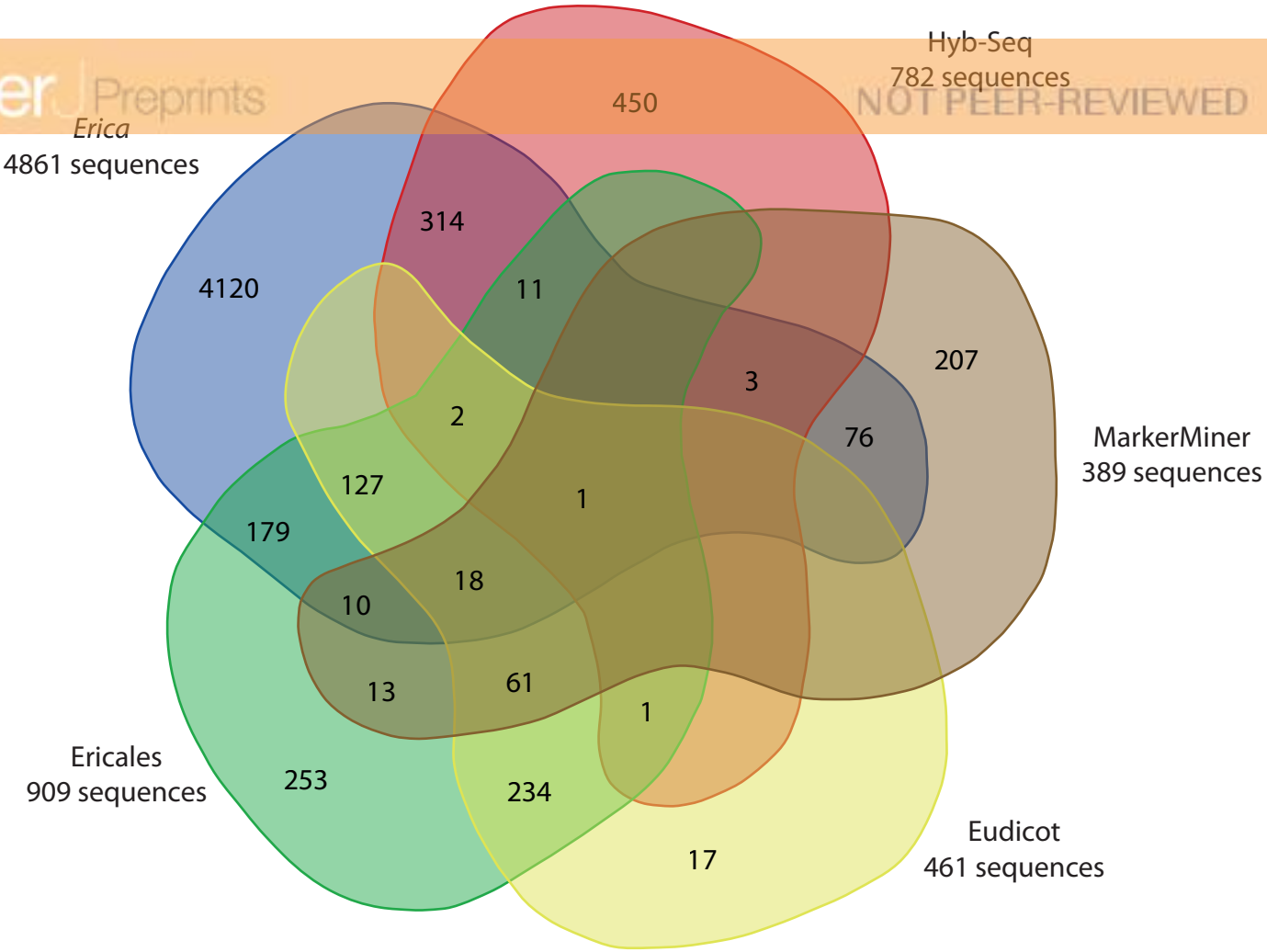
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints NOT PEER-REVIEWED

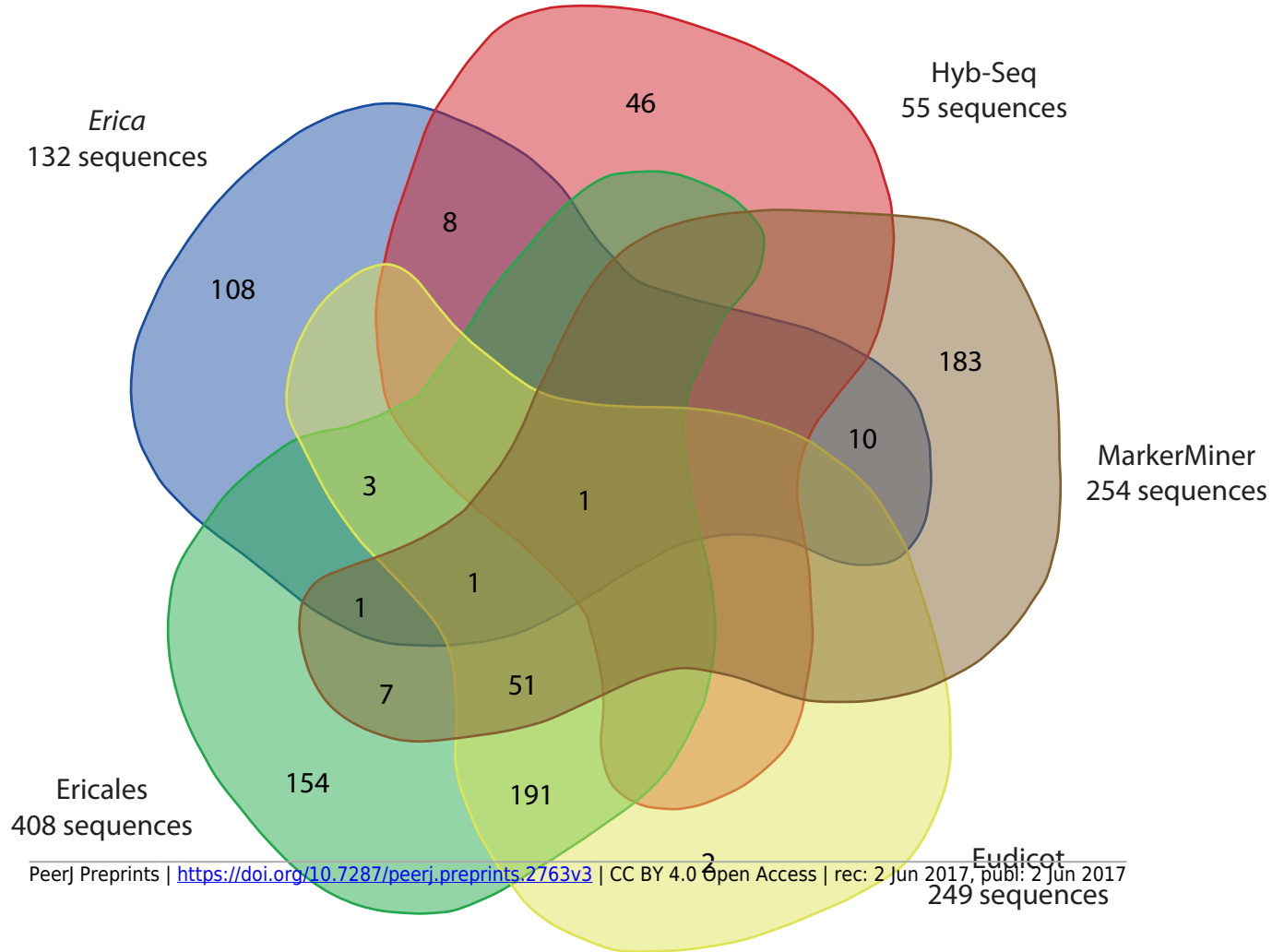


PeerJ Preprints NOT PEER-REVIEWED

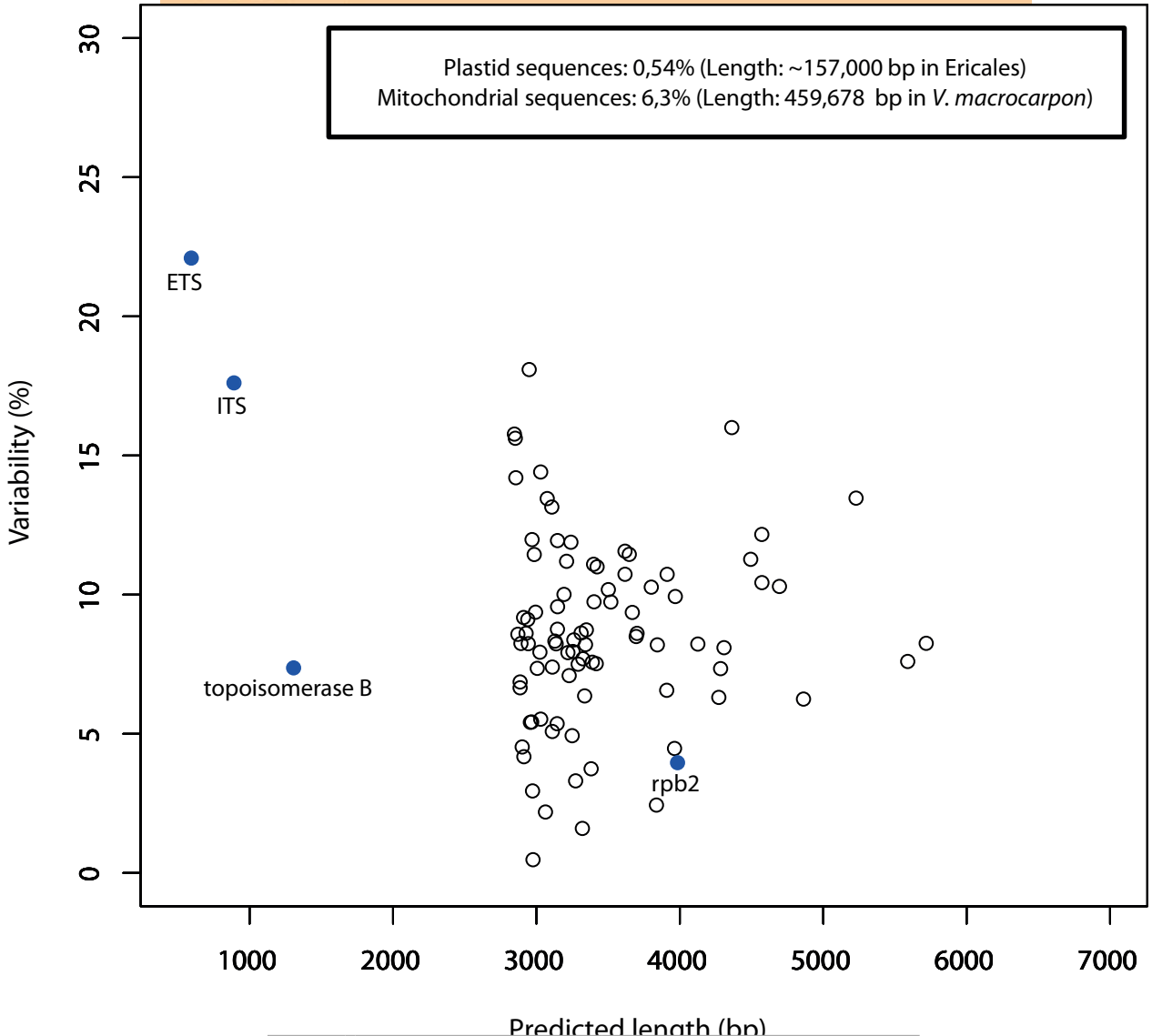
a)



b)



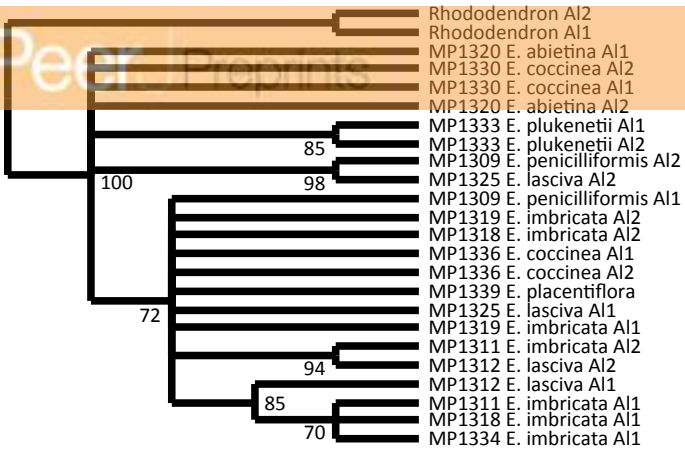
PeerJ Preprints NOT PEER-REVIEWED



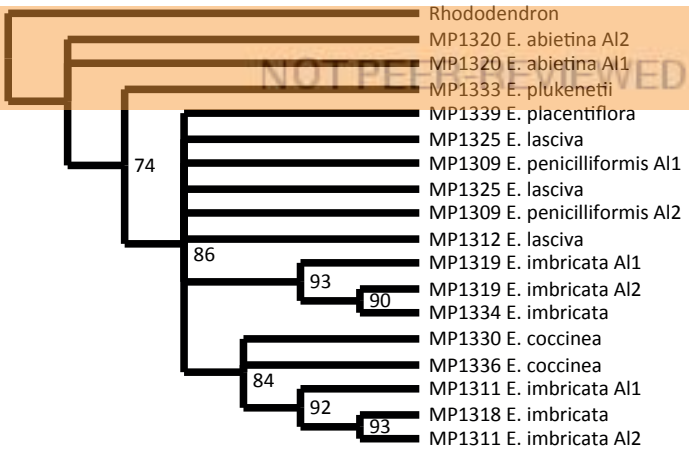
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

PeerJ Preprints
NOT PEER-REVIEWED

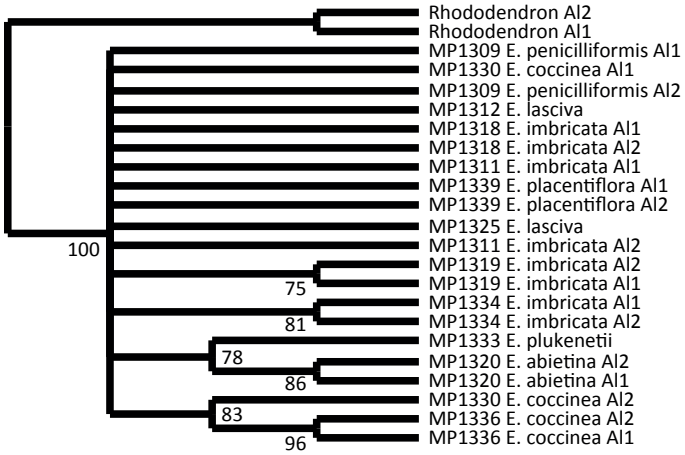
4430



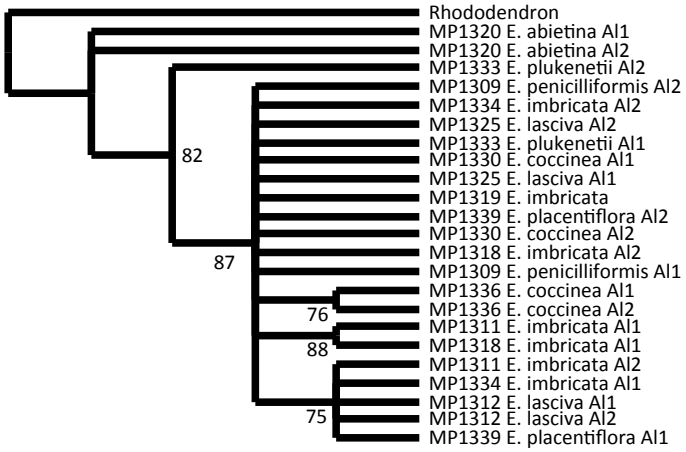
12303



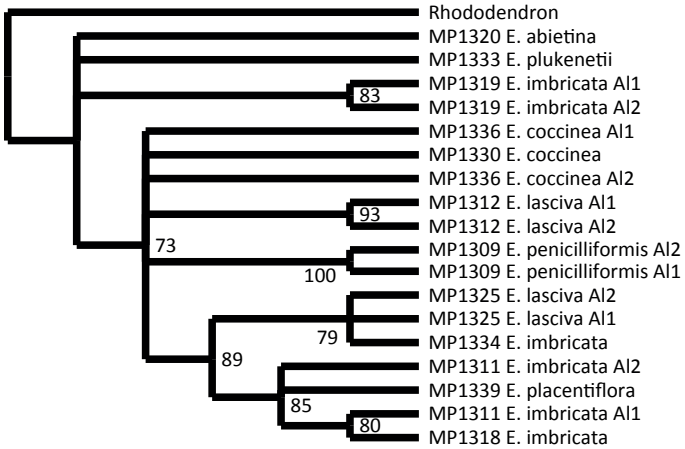
14220



17845

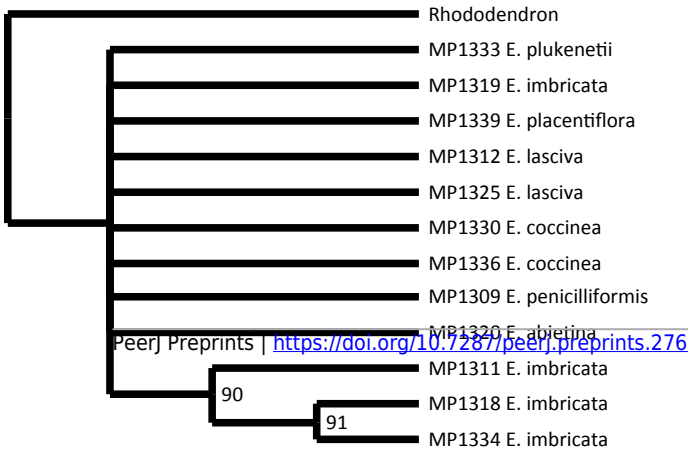


20893

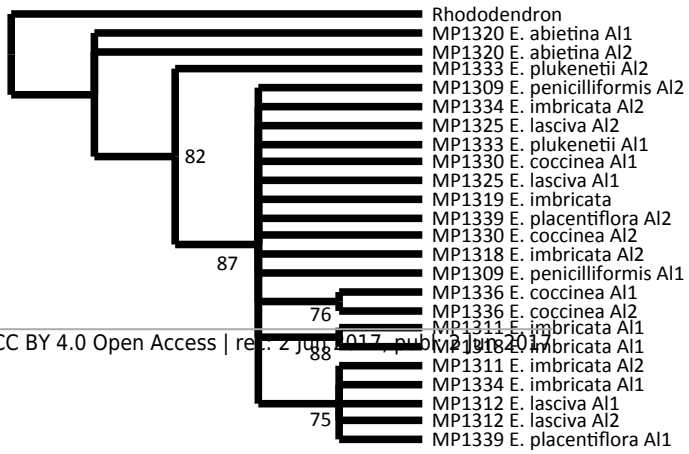


22868

Mitchondrial



ITS



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2763v3 | CC BY 4.0 Open Access | rec: 2 Jun 2017, pub: 21 Jun 2017