

Communication

# Effectiveness of Two Universal Angiosperm Probe Sets Tested In Silico for Caryophyllids Taxa with Emphasis on Cacti Species

Delil A. Chincoya <sup>1,2,\*</sup> and Sofía Solórzano <sup>1,\*</sup>

<sup>1</sup> Laboratorio de Ecología Molecular y Evolución, UBIPRO, FES Iztacala, Universidad Nacional Autónoma de México, Avenida de los Barrios 1, Los Reyes Iztacala, Tlalnepantla de Baz, Estado de México 54090, Mexico

<sup>2</sup> Posgrado en Ciencias Biológicas, UNAM, FES Iztacala, Circuito de Posgrados, Ciudad Universitaria, Coyoacán, Ciudad de México 04510, Mexico

\* Correspondence: dela@comunidad.unam.mx (D.A.C.); solorzanols@unam.mx (S.S.)

**Abstract:** In angiosperms, huge advances in massive DNA sequencing technologies have impacted phylogenetic studies. Probe sets have been developed with the purpose of recovering hundreds of orthologous loci of targeted DNA sequences (TDS) across different plant lineages. We tested in silico the effectiveness of two universal probe sets in the whole available genomes of Caryophyllids, emphasizing phylogenetic issues in cacti species. A total of 870 TDS (517 TDS from Angiosperm v.1 and 353 from Angiosperms353) were individually tested in nine cacti species and *Amaranthus hypochondriacus* (external group) with  $\geq 17$  Gbp of available DNA data. The effectiveness was measured by the total number of orthologous loci recovered and their length, the percentage of loci discarded by paralogy, and the proportion of informative sites (PIS) in the alignments. The results showed that, on average, Angiosperms353 was better than Angiosperm v.1 for cacti species, since the former obtained an average of 275.6 loci that represent 123,687 bp, 2.48% of paralogous loci, and 4.32% of PIS in alignments, whereas the latter recovered 148.4 loci (37,683 bp), 10.38% of paralogous loci, and 3.49% of PIS. We recommend the use of predesigned universal probe sets for Caryophyllids, since these recover a high number of orthologous loci that resolve phylogenetic relationships.



**Citation:** Chincoya, D.A.; Solórzano, S. Effectiveness of Two Universal Angiosperm Probe Sets Tested In Silico for Caryophyllids Taxa with Emphasis on Cacti Species. *Genes* **2022**, *13*, 570. <https://doi.org/10.3390/genes13040570>

Academic Editor: Pablo C. Guerrero

Received: 23 February 2022

Accepted: 21 March 2022

Published: 24 March 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** cactaceae; caryophyllids; phylogenomics; target enrichment; universal probe sets

## 1. Introduction

High-throughput DNA sequencing technologies (HTDNAs) significantly improve denser molecular sampling of DNA than Sanger-based capillarity technology. This increment of whole genomes published either as fully assembled and annotated genomes, or as raw data with poor or null informatics processing, has positively impacted advances in examining fundamental questions in flowering land plants. Fortunately, most of these massive DNA data are being deposited in free-access digital reservoirs, allowing the world's scientific community to use them freely. In the particular case of angiosperms, most problems studied with HTDNAs are those of a phylogenetic nature. Presently, the term phylogenomics is used when a high number of loci (i.e., dozens and even hundreds) from DNA/RNA are sampled in a single or in various genomes (i.e., chloroplast, mitochondria, and nucleus) contained in plants. In land-flowering plants, phylogenomic analysis based on chloroplast genomes [1,2] and transcriptomes [3] has allowed scientists to resolve the relationships of early divergent lineages of angiosperms. The different genome-scale analyses converged in a topology of the species tree in which three orders—Amborellales (Amborellaceae, *Amborella trichopoda*), Nymphaeales (Cabombaceae, Hydatellaceae, and Nymphaeaceae), and Austrobaileyales (Austrobaileyaceae, Schisandraceae and Trimeneaceae)—were, successively, the sister lineages to the whole-monophyletic speciose group of angiosperms [1–3]. Recently, the old problem of determining the origin of angiosperms was reexamined in the broader phylogenetic study of Li et al. [4]. This study integrated previously published data obtained using HTDNAs, as well as other

taxa de novo sequenced by the authors, in order to obtain 80 loci of the genome of the chloroplasts (>82,286 bp) of 2351 ingroup flowering land species. This study concluded that angiosperms emerged in the early Jurassic, which conflicts with the pollen fossil records data that indicate that angiosperms arose in the Cretaceous period. The results of this study revamped an old discussion with those fossil records defenders [5], and it is a fact that massive sequencing data will alter our views of the evolution of the biodiversity of our planet; they could even break scientific paradigms.

In addition, HTDNAs are also fertile tools for those scientific questions focused on smaller scales, either focused on the evolutionary divergence processes that occurred in particular flowering groups or in the biota evolution of local areas. Moreover, phylogenomic sampling may drastically increase the phylogenetic resolution yet at lower taxonomic hierarchical levels and in those plant groups that recently diverged [6–8]. The benefits of having a fully resolved phylogenetic tree species impact many areas: the inherent value is that phylogenetic relationships are clarified—often, in a phylogenetic tree, the diverging process may be inferred, which eventually may help to understand patterns of richness around the planet at different taxonomic hierarchical levels (i.e., orders, families, and genera). Furthermore, a fully resolved phylogeny may support a solution to those pendant taxonomic problems and the species boundaries can be clearly delimited.

Today, it is a fact that HTDNAs are a strategy that may provide denser molecular sampling, and for many research groups around the world, these technologies are a real available methodological alternative. Currently, to access this denser molecular sampling, there are different techniques that include: transcriptomic sequencing, low coverage sequencing (i.e., genome skimming), restriction-site-associated DNA sequencing (e.g., RAD-Seq), and target enrichment (studies cited in McKain et al. [9] and Yu et al. [10]). For the target-enrichment approach, one way is for the scientific team to start from zero in order to design the specific targeted probes (i.e., probe sets, probe kits, enrichment probes, and enrichment kits) for the studied plant taxa [7,11,12]; however, to begin, they will need to generate the primary data, which must be represented by at least a single whole-genome sequence or transcriptome from a taxon. If such primary data were already sequenced with free access, the scientific team would save money and time. If not, researchers would need to generate this primary set of DNA data and, consequently, would require more resources to experimentally produce the primary DNA/RNA data. In this scenario, the scientific team would have to design, essay, and fully complete experimental tests for a set of taxa; however, they would be dealing with complicated bioinformatics analyses. An alternative strategy is to use some of the available predesigned universal targeted probe sets and experimentally test the effectiveness of these on the studied taxa. These probe sets are presumed to be of universal efficiency and are expected to find dozens or hundreds of nuclear orthologous loci across different angiosperm taxa; however, these results are unknown until the experimental and bioinformatics processes are completed. Lastly, we propose here an alternative strategy that consists of testing *in silico* the available universal angiosperm probe sets, without preliminary experimental costs. This *in silico* test requires the availability of primary massive DNA sequences. It is expected that in this quantity of DNA, the nuclear genomes of the angiosperms of interest are represented, and thus it is expected that a high number of variable and orthologous loci will be recovered. In this *in silico* test, the research team must uniquely establish a detailed and appropriate bioinformatic protocol that can be implemented with free-access software, as in this study.

The objective of the present study was to compare the effectiveness *in silico* of two probe sets that promise universal results for angiosperms. The probe set Angiosperm v.1 was designed from the genomes of 25 species and contains 56,862 probes, expected to recover up to 517 individual exons with a maximum total length of 150 kbp [13]. In contrast, the probe set Angiosperms353 was designed from the transcriptomes of over 600 species and contains 75,151 probes, which are expected to recover 353 coding sequences (CDS) composed of one or more exons with a total length of 260 kbp [14]. These two universal probe sets have been applied in specific studies in various lineages of angiosperms for phy-

logenetic purposes. For example, the Angiosperm v.1 probe set was successfully used in the species of the genera *Aristolochia* [15] and *Acer* [16]. Recently, the probe set Angiosperms353 resolved phylogenetic issues in the genera *Nepenthes* [17], in the species of the genus *Cyperus* [18], and in the species of the ten families that comprise the order Cornales [19]. The results of these studies showed that these two probe sets were appropriate alternatives for obtaining loci that were used to construct resolved phylogenetic trees.

Accordingly, in this study, we tested *in silico* these two probe sets (Angiosperm v.1 and Angiosperms353) in ten Caryophyllids species: nine species of cacti (Cactaceae) and the species *Amaranthus hypochondriacus* (Amaranthaceae), used in the phylogenetic analysis as an external group. We focused on cacti species because phylogenetic studies carried out with these taxa repeatedly result in unresolved phylogenetic trees [20–22]. This lack of phylogenetic resolution may be due to the fact that most of these studies used poor molecular sampling (1–12 loci), or may be due to other factors, such as the recent postulation of the origin of Cactaceae (30 to 35 Ma) based on the molecular-clock model [23,24]. Moreover, it has been hypothesized that a global aridification that occurred nearly ~12 million years ago (the late Miocene epoch) promoted the main evolutionary divergence in the Cactaceae family [23]. Consequently, the relatively recent origin of the Cactaceae family and the rapid and recent divergence of cacti species may cause a weak and shallow molecular separation, which has been reflected in those unresolved phylogenetic tree species. In addition, as HTDNAs have shown a significant increase in the number of loci-resolved phylogenetic tree species even in the taxa that have recently diverged [8], we are expecting that these two probe sets have a similar effectiveness in recovering nuclear orthologous loci that serve the cacti species analyzed in the phylogenetic studies.

## 2. Materials and Methods

### 2.1. Genomes of the Species Analyzed

At the beginning of our study, we carried out preliminary tests with the genomes <12 Gbp; however, the bioinformatic protocol from the pipeline used here did not work. For this, we e-searched all cacti species represented by >15 Gb of massive DNA sequencing in the NCBI digital database. A total of eight genomes of different cacti species were downloaded from the SRA repository. These DNA data ranged from 17 Gb (*Pereskia humboldtii*) to 57 Gb (*Selenicereus undatus*) (Table S1). In addition, the complete genome of the small globose cactus *Mammillaria huitzilopochtli* was *de novo* sequenced (raw data are available in S.S.) based on the PE genomic library sequenced in NovaSeq of Illumina. These nine cacti species are grouped into three cacti subfamilies and five tribes. The subfamily Cactoideae was represented by the tribes Cacteae (*M. huitzilopochtli*), Cereeae (*Cereus fernambucensis*), Hylocereeae (*Selenicereus undatus*), and Pachycereeae (*Carnegiea gigantea*, *Lophocereus schottii*, *Pachycereus pringlei*, and *Stenocereus thurberi*). The subfamily Opuntioideae was represented by the tribe Opuntieae (*Opuntia sulphurea*), and the subfamily Pereskioideae by the species *P. humboldtii*. In addition, *Amaranthus hypochondriacus* (Amaranthaceae, Caryophyllales; accession number in genbank, SRR2106212) was used as an external group for phylogenetic analyses of the studied species.

### 2.2. Bioinformatic Process to Test *In Silico* the Two Probe Sets

The massive amount of available raw data of each of the ten species (i.e., ten genomes) were firstly filtered by reads using Trim Galore version 0.6.7 [25] for subsequent bioinformatic analyses. For this, the total of the reads of each genome was trimmed, and the adapters were removed; the only reads used in the analysis were those that had a PHRED quality score  $\geq 15$  and a length  $\geq 80$  bp. The reads per genome that passed this filtering were *in silico* tested with the original predesigned targeted DNA sequences (TDS) from the two probe sets that were compared. Since the probe set Angiosperm v.1 includes 517 TDS, and Angiosperms353 includes 353 TDS, a total of 870 TDS were tested individually per genome following the pipeline available in HybPiper version 1.3.1 [26]. Accordingly, each of the 870 TDS were individually mapped and assembled with the main script (read\_first.py)

of this pipeline. The target files used in HybPiper correspond to the sequences from which the probe sets were designed and were downloaded from [https://github.com/mossmatters/Angiosperms353/blob/master/Angiosperms353\\_targetSequences.fasta](https://github.com/mossmatters/Angiosperms353/blob/master/Angiosperms353_targetSequences.fasta) (Angiosperms353) (accessed on 15 March 2022) and <https://www.biorxiv.org/content/10.1101/086298v2.supplementary-material> (Angiosperm v.1) (accessed on 15 March 2022). Since the sequences of Angiosperm v.1 were 517 alignments in Phylip format, they were converted to fasta format, the gaps were removed, and the sequences were merged. Two scripts in HybPiper (scripts `get_seq_lengths.py` and `hybpiper_stats.py`) were used to obtain the length of the coding sequences (i.e., exons) and the statistical values that estimate the recovery efficiency of each of the two probe sets per genome; lastly, the R script (`gene_recovery_heatmap.R`) was used to plot the heatmaps and to visualize the efficiency of the recovering loci per genome. Finally, the script `retrieve_sequences.py` from the pipeline in HybPiper was used to generate the multifasta files to recover the total number of orthologous nuclear loci per genome. To avoid a possible bias in paralogous identification, we extracted the largest exon from each gene of the Angiosperms353 dataset, and we generated the target file (Supplementary File S1) that contains the individual exons of 3829 of the 4781 CDS (target instances) contained in the Angiosperms353 dataset [14]. This target file was used to run an independent HybPiper pipeline, and the paralogous genes were added to the paralogs identified in the analysis of the complete CDS from Angiosperms353. All the sequences identified as paralogous loci in Angiosperm v.1 and Angiosperms353 (CDS + exon) were discarded for the phylogenetic analysis. The relative effectiveness of each of the two probe sets was estimated per species based on the following parameters: (1) the total number of TDS recovered (i.e., the number of loci) by each probe set; (2) the length (bp) of each locus was summed to have the total length (bp) of all the recovered loci; (3) the total number of paralogous loci was discarded for phylogenetic analysis; and (4) the proportion of parsimoniously informative sites (PIS) was obtained from the alignments of the DNA sequences of the orthologous loci.

### 2.3. Phylogenomics Analysis

The phylogenetic trees were based on the total nuclear orthologous loci recovered by Angiosperms353 (240 loci) and Angiosperm v.1 (71). The alignment of the species of each locus was carried out with MAFFT version v7.310 [27]. We observed a large disparity in the distribution of the orthologous loci among the ten studied species, and in order to diminish the missing data for phylogenetic analysis, we chose those loci that were recorded in the alignments of at least seven species. In addition, for each orthologous locus, its respective proportion of PIS was estimated with the program AMAS [28]. The application ModelFinder [29], implemented in IQ-TREE2 version 2.1.4- $\beta$  [30], was used to select the partitioning scheme and to estimate the substitution models for each partition and for each locus. The DNA sequences of the orthologous loci obtained with each probe set were concatenated to construct the phylogenetic species-trees with IQ-TREE2 that ran 1000 replicates of an ultrafast bootstrap (UFBoot). In order to obtain the gene concordance factors (gCF) for each node of each of the two phylogenetic trees, we constructed an individual gene tree for each locus with IQ-TREE2.

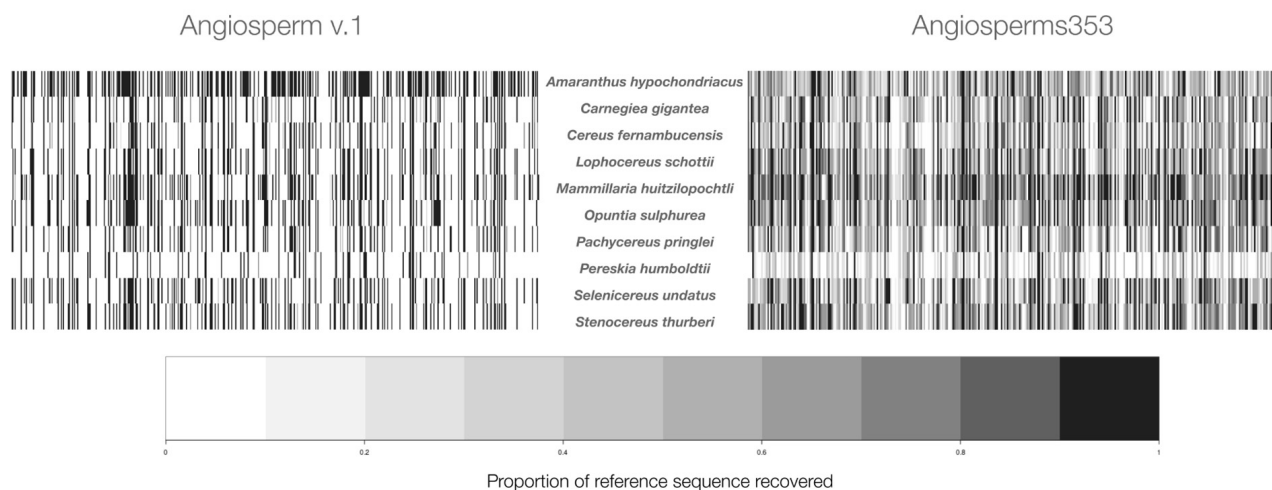
## 3. Results

### 3.1. Effectiveness of the Two Probe Sets

The results showed that, among the ten studied species, the two probe sets tested had clear differences in the global efficiency of the TDS captured (Figure 1, Table 1). For the nine cacti species, Angiosperms353 showed a higher efficiency than Angiosperm v.1, which is supported by the total number of loci recovered (Figure 1), the length of the DNA sequences recovered by the locus, and the number of paralogous loci that were discarded (Table 1). These differences in global efficiency are visualized by the abundant dark lines seen in the heatmap for each species (Figure 1). Each vertical line indicates that an individual TDS was found in the tested genome, and its darkness indicates the proportion of the length of



this sequence that was captured (Figure 1). For example, for *A. hypochondriacus*, the probe set Angiosperms353 showed a higher efficiency in capturing loci (316; Table 1); however, the length of the 216 loci captured with Angiosperm v.1 was larger (68,205 bp; Table 1); therefore, the heatmap is darker for this species (Figure 1).



**Figure 1.** Heatmap per species is presented by raw data. To the left of the species names are the results obtained with Angiosperm v.1, and to the right, the results obtained with Angiosperms353. The gaps (white color) are caused by the absence of vertical lines, which indicates that the targeted DNA sequences did not find those loci in the genome tested.

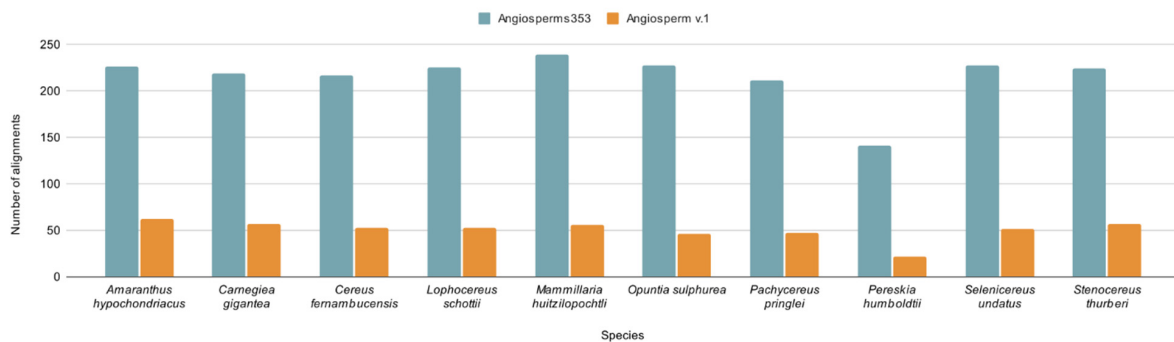
**Table 1.** Comparison of the four parameters used to estimate the effectiveness between the two probe sets in silico tested. The results outside parentheses correspond to results obtained with Angiosperm v.1, and those inside parentheses with Angiosperms353. Total number of loci recovered with each probe set; number of loci that found a proportion  $\geq 50\%$  of the length of the targeted DNA sequence; the total length (bp) of sequences identified as exons and the quotient calculated; total length of exons from Angiosperms353/length of exons from Angiosperm v.1; and the absolute number of paralogous loci identified per species.

Species	Number of Loci Recovered	Number of Loci with $>50\%$ Sequence Length	Total Length of Exons (bp): Quotient	Number of Paralogous Loci
<i>A. hypochondriacus</i>	264 (316)	256 (167)	68,205 (65,658):0.96	24 (8)
<i>C. gigantea</i>	129 (270)	123 (165)	32,133 (124,773):3.88	12 (8)
<i>C. fernambucensis</i>	132 (269)	125 (145)	30,822 (116,277):3.77	13 (1)
<i>L. schottii</i>	137 (287)	135 (194)	35,127 (140,259):3.99	18 (6)
<i>M. huitzilopochtli</i>	177 (319)	168 (245)	43,875 (174,285):3.97	36 (21)
<i>O. sulphurea</i>	155 (292)	149 (204)	40,611 (150,720):3.71	18 (8)
<i>P. pringlei</i>	123 (253)	121 (158)	31,683 (122,844):3.88	6 (4)
<i>P. humboldtii</i>	68 (173)	67 (69)	18,885 (65,658):3.48	2 (2)
<i>S. undatus</i>	154 (287)	149 (170)	37,866 (133,071):3.51	16 (6)
<i>S. thurberi</i>	145 (290)	143 (204)	37,656 (143,322):3.81	16 (8)

The respective heatmaps of the cacti species show abundant gaps with Angiosperm v.1, because a large number of its 517 targeted DNA sequences did not match (Figure 1) in these genomes. This poorer capture of Angiosperm v.1 was documented in the four parameters used to compare its effectiveness to Angiosperms353. The results showed that, in the ten studied species, the probe set Angiosperms353 captured the highest number of nuclear orthologous loci (Table 1). In the ten studied species, with Angiosperms353, the average number of the orthologous loci was  $275.6 \pm 41.3$  SD, and with Angiosperm v.1, it was  $148.4 \pm 49.6$  SD. Moreover, in the nine cacti species studied, the length of the exons with Angiosperms353 was approximately four times larger than with Angiosperm v.1 (Table 1). In the ten studied species, the number of orthologous loci identified with Angiosperms353

was higher than with Angiosperm v.1 (Table 1). Additionally, in the ten studied species, Angiosperm v.1 recovered a total of 70 paralogous loci, and the percentage of these loci varied from 2.9% (*P. humboldtii*) to 20.3% (*M. huitzilopochtli*). In contrast, paralogous loci in Angiosperms353 varied from 0.4% (*C. fernambucensis*) to 6.7% (*M. huitzilopochtli*).

The orthologous loci identified in 70% of the analyzed species totaled 240 alignments with Angiosperms353, and only 71 alignments were obtained with Angiosperm v.1. The lowest number of alignments was obtained for the species *P. humboldtii*: 141 and 22 alignments obtained with Angiosperms353 and Angiosperm v.1, respectively. In contrast, the highest number of alignments was obtained for *A. hypochondriacus* (62) and *M. huitzilopochtli* (239) with Angiosperm v.1 and Angiosperms353, respectively (Figure 2).

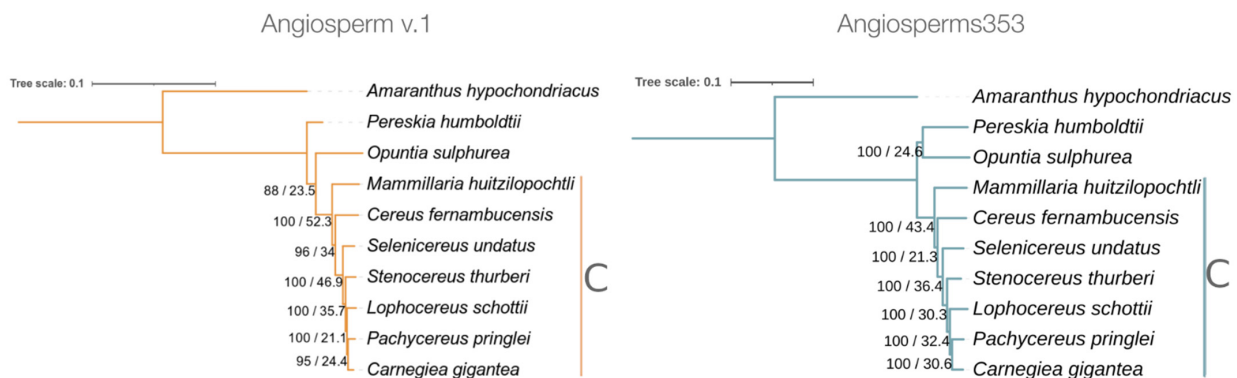


**Figure 2.** Total number of alignments obtained with Angiosperms353 and Angiosperm v.1 per species.

With respect to the number of parsimoniously informative sites (PIS), Angiosperms353 recovered 4.32% of PIS in the sequences of the total number of orthologous loci aligned, whereas Angiosperm v.1 recovered 3.49% of PIS in such alignments.

### 3.2. Phylogenomics Analysis

The phylogenetic trees constructed with the DNA datasets provided by each of the two probe sets showed a fully resolved topology (Figure 3). Interestingly, in these two trees, the seven species of the subfamily Cactoideae were clearly grouped in a single monophyletic node. However, these trees showed subtle differences in the topology of the position occupied by *O. sulphurea*, as well as in the statistical values of the UFBoot and in the gCF. In the tree constructed from the data of Angiosperms353, the nodes were 100% supported by UFBoot values, but in the tree of Angiosperm v.1, these values varied from 88 to 100%. However, these two trees showed low values of gene concordance (<53%) for all the nodes.



**Figure 3.** Phylogenetic trees obtained with ML for the nine studied cacti species; the prince's feather (*A. hypochondriacus*) was the outgroup. The tree constructed with data from Angiosperm v.1 was based on 71 loci (20,082 bp), and that constructed with data obtained from Angiosperms353 was based on 262 loci (186,973 bp). The numbers to the side of the branches correspond to UFBoot (above slash) and gCF (below slash) values. The species grouped in the subfamily Cactoideae are indicated by the letter C.

#### 4. Discussion

Our results showed that the two probe sets tested here were useful for Caryophyllids. However, Angiosperms353 had notably better results than Angiosperm v.1 for the nine cacti species, irrespective of the tribe or the subfamily in which these species are classified. It is evident that this dissimilar effectiveness was not related to the larger number of targeted DNA sequences (517) included in Angiosperms v.1 vs. 353 of Angiosperms353. We consider that these superior results obtained with Angiosperms353 for the cacti species could be explained by factors involved in the design process of the targeted DNA sequences for these probe sets: (1) the relatively wider taxonomic sampling employed to design Angiosperms353 (over 600 taxa) vs. Angiosperm v.1. (25 taxa), which could enhance the probability of sampling a larger number of targeted loci dispersed across different lineages, and that (2) Angiosperms353 was designed to recover up to 260 kbp of exons, whereas Angiosperm v.1 was designed to recover only 150 kbp.

Although these two probe sets were designed to recover only single copy genes, they failed because paralog sequences were recovered, but a substantially lower percentage was discarded with Angiosperms353. In addition, the species *P. humboldtii* showed the lowest global effectiveness with the two probe sets. We consider that these results may be caused by the relatively small quantity of the massive DNA dataset (17 Gb) available for this species, and it is probable that the nuclear genome was not completely sampled. Thus, it is important to consider the size of the nuclear genome in order to choose the appropriate platform for massive sequencing of the interested species.

With respect to *A. hypochondriacus*, the better results were obtained with Angiosperm v.1, except for the parameter of the number of paralogous loci. We consider that this was caused by the fact that, during the design of the targeted sequences of Angiosperm v.1, the authors sampled *β vulgaris* (Chenopodiaceae), a species that is phylogenetically close to the Amaranthaceae family, according to previous studies [31,32]. Consequently, we believe that the taxonomic identity of the taxa sampled during the design of the targeted sequences may impact the success of posterior studies of the predesigned probe sets.

Although, in this study, the primary objective was not to resolve the phylogenetic relationships of cacti species, the phylogenetic trees obtained allowed us to make some considerations. The two probe sets provided data used to obtain trees with similar topologies accompanied by low values of gCF. Consequently, these values of gCF indicated a high gene-tree discordance, and we believe that this was not caused by the number of loci used, since the number of loci obtained with the Angiosperms353 dataset (240) was higher than that obtained with Angiosperm v.1 (71). It is probable that these gCF values are related to the poor taxonomic sampling included in our study; however, previous phylogenomic studies with a broader taxonomic sampling also obtained high levels of gene-tree discordance [32,33]. Among the two trees, there was a punctual difference in the position occupied by *O. sulphurea* with respect to the species grouped in the Cactoideae subfamily, similar to the tree obtained with Angiosperm v.1 in previous phylogenetic studies [32–34]. However, based on the global parameters used to measure effectiveness, we consider that Angiosperms353 provided superior results for the cacti phylogenetic studies to Angiosperm v.1.

#### 5. Conclusions

We concluded that the universal probe sets tested here are a confident strategy for studies that require a high number of single copy orthologous loci. We recommend that researchers compare previous in silico analyses of the selected probe sets in order to estimate their effectiveness. We should mention that, in this study, we focused on exons; thus, the molecular variation of introns was not explored. For introns, we would expect larger variation levels. It is likely that the molecular variation contained in introns may separate the more recent evolutionary processes that occurred at intraspecific levels. Lastly, our results confirmed that non-specific probe sets should not necessarily diminish the molecular sampling across different angiosperm groups. In fact, in a previous study

carried out with the taxa of the genus *Cyperus* (Cyperaceae), the Angiosperms353 probe set provided a similar proportion of PIS compared to that obtained with a family-specific probe set, producing a similar resolution power at infrageneric taxonomic levels [18].

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/article/10.3390/genes13040570/s1>, Table S1: List of the studied cacti species and the subfamily and tribes in which they are classified. Supplementary File S1: Target file with 3829 individual largest exons that were recovered from the total of the 4781 CDS (target instances) contained in Angiosperms353 dataset.

**Author Contributions:** Conceptualization, D.A.C.; methodology, D.A.C.; formal analysis, D.A.C.; writing—original draft preparation, D.A.C. and S.S.; writing—review and editing, D.A.C. and S.S.; supervision, S.S.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by UNAM-PAPIIT IN228619.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** SEMARNAT (SGPA/DGVS/06880/16) supplied collecting permission to sampling *Mammillaria huiquizilpochtli* specimen. D.A.C. (310003715) is a student at Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México; she is granted by Consejo Nacional de Ciencia y Tecnología CONACyT (855061). We thank comments of two anonymous reviewers that improved the quality of this article. Macrogen Inc., Seoul South Korea provided the sequencing service of the whole genome of *M. huiquizilpochtli*.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Ruhfel, B.; Gitzendanner, M.; Soltis, P.S.; E Soltis, D.; Burleigh, J.G. From algae to angiosperms—inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* **2014**, *14*, 23. [[CrossRef](#)] [[PubMed](#)]
2. Gitzendanner, M.A.; Soltis, P.S.; Wong, G.K.-S.; Ruhfel, B.R.; Soltis, D.E. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *Am. J. Bot.* **2018**, *105*, 291–301. [[CrossRef](#)] [[PubMed](#)]
3. Wickett, N.J.; Mirarab, S.; Nguyen, N.; Warnow, T.; Carpenter, E.; Matasci, N.; Ayyampalayam, S.; Barker, M.S.; Burleigh, J.G.; Gitzendanner, M.A.; et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E4859–E4868. [[CrossRef](#)]
4. Li, H.T.; Yi, T.S.; Gao, L.M.; Ma, P.F.; Zhang, T.; Yang, J.B.; Gitzendanner, M.A.; Fritsch, P.W.; Cai, J.; Luo, Y.; et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* **2019**, *5*, 461–470. [[CrossRef](#)]
5. Coiro, M.; Doyle, J.A.; Hilton, J. How deep is the conflict between molecular and fossil evidence on the age of angiosperms? *New Phytol.* **2019**, *223*, 83–99. [[CrossRef](#)] [[PubMed](#)]
6. Parks, M.; Cronn, R.; Liston, A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* **2009**, *7*, 84. [[CrossRef](#)]
7. Nicholls, J.A.; Pennington, R.T.; Koenen, E.J.M.; Hughes, C.E.; Hearn, J.; Bunnefeld, L.; Dexter, K.G.; Stone, G.N.; Kidner, C.A. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Front. Plant Sci.* **2015**, *6*, 710. [[CrossRef](#)]
8. Bagley, J.C.; Uribe-Convers, S.; Carlsen, M.M.; Muchhala, N. Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical *Burmeistera* bellflowers as a case study. *Mol. Phylogenet. Evol.* **2020**, *152*, 106769. [[CrossRef](#)]
9. McKain, M.R.; Johnson, M.G.; Uribe-Convers, S.; Eaton, D.; Yang, Y. Practical considerations for plant phylogenomics. *Appl. Plant Sci.* **2018**, *6*, e1038. [[CrossRef](#)]
10. Yu, X.; Yang, D.; Guo, G.; Gao, L. Plant phylogenomics based on genome-partitioning strategies: Progress and prospects. *Plant Divers.* **2018**, *40*, 158–164. [[CrossRef](#)]
11. Mandel, J.R.; Dikow, R.B.; Funk, V.A.; Masalia, R.R.; Staton, S.E.; Kozik, A.; Michelmore, R.W.; Rieseberg, L.H.; Burke, J.M. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Appl. Plant Sci.* **2014**, *2*, 1300085. [[CrossRef](#)] [[PubMed](#)]



12. Stephens, J.D.; Rogers, W.L.; Heyduk, K.; Cruse-Sanders, J.M.; Determann, R.O.; Glenn, T.C.; Malmberg, R.L. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Mol. Phylogenet. Evol.* **2015**, *85*, 76–87. [[CrossRef](#)] [[PubMed](#)]
13. Buddenhagen, C.; Lemmon, A.R.; Lemmon, E.M.; Bruhl, J.; Cappa, J.; Clement, W.L.; Donoghue, M.; Edwards, E.J.; Hipp, A.L.; Kortyna, M.; et al. Anchored phylogenomics of angiosperms I: Assessing the robustness of phylogenetic estimates. *BioRxiv* **2016**, *2016*, 086298. [[CrossRef](#)]
14. Johnson, M.G.; Pokorny, L.; Dodsworth, S.; Botigué, L.R.; Cowan, R.S.; Devault, A.; Eiserhardt, W.L.; Epitawalage, N.; Forest, F.; Kim, J.T.; et al. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Syst. Biol.* **2019**, *68*, 594–606. [[CrossRef](#)] [[PubMed](#)]
15. Wanke, S.; Mendoza, C.G.; Müller, S.; Paizanni Guillén, A.; Neinhuis, C.; Lemmon, A.R.; Lemmon, E.M.; Samain, M.-S. Recalcitrant deep and shallow nodes in *Aristolochia* (Aristolochiaceae) illuminated using anchored hybrid enrichment. *Mol. Phylogenet. Evol.* **2017**, *117*, 111–123. [[CrossRef](#)]
16. Li, J.; Stukel, M.; Bussies, P.; Skinner, K.; Lemmon, A.R.; Lemmon, E.M.; Brown, K.; Bekmetjev, A.; Swenson, N.G. Maple phylogeny and biogeography inferred from phylogenomic data. *J. Syst. Evol.* **2019**, *57*, 594–606. [[CrossRef](#)]
17. Murphy, B.; Forest, F.; Barraclough, T.; Rosindell, J.; Bellot, S.; Cowan, R.; Golos, M.; Jebb, M.; Cheek, M. A phylogenomic analysis of *Nepenthes* (Nepenthaceae). *Mol. Phylogenet. Evol.* **2020**, *144*, 106668. [[CrossRef](#)]
18. Larridon, I.; Villaverde, T.; Zuntini, A.R.; Pokorny, L.; Brewer, G.E.; Epitawalage, N.; Fairlie, I.; Hahn, M.; Kim, J.; Maguilla, E. Tackling rapid radiations with targeted sequencing. *Front. Plant Sci.* **2020**, *10*, 1655. [[CrossRef](#)]
19. Thomas, S.K.; Liu, X.; Du, Z.-Y.; Dong, Y.; Cummings, A.; Pokorny, L.; Xiang, Q.-Y.; Leebens-Mack, J.H. Comprehending Cornales: Phylogenetic reconstruction of the order using the Angiosperms353 probe set. *Am. J. Bot.* **2021**, *108*, 1112–1121. [[CrossRef](#)]
20. Butterworth, C.A.; Wallace, R.S. Phylogenetic studies of *Mammillaria* (Cactaceae)—Insights from chloroplast sequence variation and hypothesis testing using the parametric bootstrap. *Am. J. Bot.* **2004**, *91*, 1086–1098. [[CrossRef](#)]
21. Griffith, M.P.; Porter, J.M. Phylogeny of Opuntioideae (Cactaceae). *Int. J. Plant Sci.* **2009**, *170*, 107–116. [[CrossRef](#)]
22. Bárcenas, R.T.; Yesson, C.; Hawkins, J.A. Molecular systematics of the Cactaceae. *Cladistics* **2011**, *27*, 470–489. [[CrossRef](#)] [[PubMed](#)]
23. Arakaki, M.; Christin, P.-A.; Nyffeler, R.; Lendel, A.; Egli, U.; Ogburn, R.M.; Spriggs, E.; Moore, M.J.; Edwards, E.J. Contemporaneous and recent radiations of the world’s major succulent plant lineages. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 8379–8384. [[CrossRef](#)] [[PubMed](#)]
24. Magallon, S.; Gomez-Acevedo, S.; Sanchez-Reyes, L.L.; Hernandez-Hernandez, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **2015**, *207*, 437–453. [[CrossRef](#)] [[PubMed](#)]
25. Babraham Bioinformatics. Trim Galore. Available online: [http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) (accessed on 21 December 2021).
26. Johnson, M.G.; Gardner, E.M.; Liu, Y.; Medina, R.; Goffinet, B.; Shaw, A.J.; Zerega, N.J.C.; Wickett, N.J. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. *Appl. Plant Sci.* **2016**, *4*, 1600016. [[CrossRef](#)] [[PubMed](#)]
27. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066. [[CrossRef](#)]
28. Borowiec, M.L. AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ* **2016**, *4*, e1660. [[CrossRef](#)]
29. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.; Von Haeseler, A.; Jermini, L.S. Model Finder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, 587–589. [[CrossRef](#)]
30. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; Von Haeseler, A.; Lanfear, R. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **2020**, *7*, 1530–1534. [[CrossRef](#)]
31. Cuénoud, P.; Savolainen, V.; Chatrou, L.W.; Powell, M.; Grayer, R.J.; Chase, M.W. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. *Am. J. Bot.* **2002**, *89*, 132–144. [[CrossRef](#)]
32. Walker, J.F.; Yang, Y.; Feng, T.; Timoneda, A.; Mikenas, J.; Hutchison, V.; Edwards, C.; Wang, N.; Ahluwalia, S.; Olivieri, J.; et al. From cacti to carnivores: Improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of Caryophyllales. *Am. J. Bot.* **2018**, *105*, 446–462. [[CrossRef](#)] [[PubMed](#)]
33. Wang, N.; Yang, Y.; Moore, M.J.; Brockington, S.F.; Walker, J.F.; Brown, J.W.; Liang, B.; Feng, T.; Edwards, C.; Mikenas, J.; et al. Evolution of Portulacineae marked by gene tree conflict and gene family expansion associated with adaptation to harsh environments. *Mol. Biol. Evol.* **2019**, *36*, 112–126. [[CrossRef](#)]
34. Edwards, E.J.; Nyffeler, R.; Donoghue, M.J. Basal Cactus Phylogeny: Implications of *Pereskia* (Cactaceae) Paraphyly for the Transition to the Cactus Life Form. *Am. J. Bot.* **2005**, *92*, 1177–1188. [[CrossRef](#)] [[PubMed](#)]