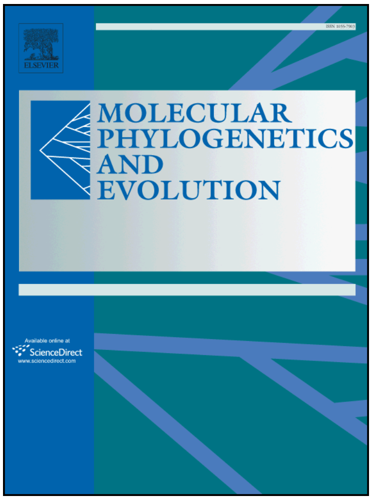
Accepted Manuscript



Editor's Choice Article

A pilot study applying the Plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus *Calosphace*; Lamiaceae)

Itzi Fragoso-Martínez, Gerardo A. Salazar, Martha Martínez-Gordillo, Susana Magallón, Luna Sánchez-Reyes, Emily Moriarty Lemmon, Alan R. Lemmon, Federico Sazatornil, Carolina Granados Mendoza

PII:	S1055-7903(16)30325-6
DOI:	http://dx.doi.org/10.1016/j.ympev.2017.02.006
Reference:	YMPEV 5746
To appear in:	*Molecular Phylogenetics and Evolution*
Received Date:	7 November 2016
Revised Date:	6 February 2017
Accepted Date:	6 February 2017

Please cite this article as: Fragoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Moriarty Lemmon, E., Lemmon, A.R., Sazatornil, F., Granados Mendoza, C., A pilot study applying the Plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus *Calosphace*; Lamiaceae), *Molecular Phylogenetics and Evolution* (2017), doi: http://dx.doi.org/10.1016/j.ympev.2017.02.006

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

A pilot study applying the Plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus *Calosphace*; Lamiaceae)

Itzi Fragoso-Martínez [a,b,*], Gerardo A. Salazar [b], Martha Martínez-Gordillo [c], Susana Magallón [b], Luna Sánchez-Reyes [b], Emily Moriarty Lemmon [d], Alan R. Lemmon [e], Federico Sazatornil [f] & Carolina Granados Mendoza [b,g,*]

[a] Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México; Av. Universidad 3000, 04510, Coyoacán, Mexico City, Mexico.

[b] Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, Apartado Postal 70–367, 04510 Coyoacán, Mexico City, Mexico.

[c] Herbario de la Facultad de Ciencias (FCME), Universidad Nacional Autónoma de México, Apartado Postal 70-399, 04510 Coyoacán, Mexico City, Mexico.

[d] Department of Biological Science, Florida State University, 319 Stadium Drive, P.O. Box 3064295, Tallahassee, FL 32306–4295, USA.

[e] Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4102, USA.

[f] Laboratorio de Ecología Evolutiva – Biología Floral, Instituto Multidisciplinario de Biología Vegetal (IMBIV), Universidad Nacional de Córdoba, CONICET, CC 495, X5000ZAA Córdoba, Argentina.

[g] CONACyT División de Biología Molecular, Instituto Potosino de Investigación Científica y Tecnológica A.C., Camino a la Presa de San José 2055, Lomas 4a. sección, C.P. 78216 San Luis Potosí, San Luis Potosí, Mexico.

*Shared corresponding authors

E-mail addresses: i.fragoso@ciencias.unam.mx (I. Fragoso-Martínez), gasc@ib.unam.mx (G.A. Salazar), mjmg_unam@yahoo.com (M. Martínez-Gordillo), s.magallon@ib.unam.mx (S. Magallón), sanchez.reyes.luna@gmail.com (L. Sánchez-Reyes), chorusfrog@bio.fsu.edu (E.M. Lemmon), alemmon@fsu.edu (A. Lemmon), federicosaza@gmail.com (F. Sazatornil), carolina.granados@ipicyt.edu.mx (C. Granados Mendoza).

Abstract

We conducted a pilot study using Anchored Hybrid Enrichment to resolve relationships among a mostly Neotropical sage lineage that may have undergone a recent evolutionary radiation. Conventional markers (ITS, *trnL-trnF* and *trnH-psbA*) have not been able to resolve the relationships among species nor within portions of the backbone of the lineage. We sampled 12 representative species of subgenus *Calosphace* and included one species of *Salvia*'s s.l. closest relative, *Lepechinia,* as outgroup. Hybrid enrichment and sequencing were successful, yielding 448 alignments of individual loci with an average length of 704 bp. The performance of the phylogenomic data in phylogenetic reconstruction was superior to that of conventional markers, increasing both support and resolution. Because the captured loci vary in the amount of net phylogenetic informativeness at different phylogenetic depths, these data are promising in phylogenetic reconstruction of this group and likely other lineages within Lamiales. However, special attention should be placed on the amount of phylogenetic noise that the data could potentially contain. A prior exploration step using phylogenetic informativeness profiles to detect loci with sites with disproportionate high substitution rates (showing "phantom" spikes) and, if required, the ensuing filtering of the problematic data is recommended. In our dataset, filtering resulted in increased support and resolution for the shallow nodes in maximum likelihood phylogenetic trees resulting from concatenated analyses of all the loci. Additionally, it is expected that an increase in sampling (loci and taxa) will aid in resolving weakly supported, short deep internal branches.

Key words

anchored phylogenomics; Lamiales, maximum likelihood; phylogenetic noise; target enrichment

1. Introduction

1.1 Genomic partitioning strategies and hybrid enrichment methods

Genomic partitioning strategies incorporate selectively large-scale genomic data that can be obtained and analysed at lower costs compared to sequencing whole genomes. The genomic partitioning strategies are methods for enriching sequence libraries for specific regions of the genome (Turner et al., 2009). These methods include, among others, transcriptome sequencing, multiplex PCR, reduced representation methods (RAD-seq, RRL) and hybrid enrichment (Lemmon and Lemmon, 2013). These strategies can recover large amounts of homologous loci from multiple individuals in shorter periods of time and with less effort compared to Sanger sequencing (Blaimer et al., 2015). Data derived from different genomic partitioning strategies have been incorporated into phylogenetic analyses, increasing the amount of informative characters and improving the resolution and support of relationships in some groups (e.g. pines: Parks et al., 2009; birds: Prum et al., 2015; grape family: Wen et al., 2013).

Hybrid Enrichment is a group of genomic partitioning strategies that use probes which hybridize with selected genome regions to capture them before high-throughput sequencing. This combination of methods yields hundreds or thousands of loci from nuclear and/or organellar compartments that are potentially informative at different phylogenetic scales (Faircloth et al., 2012; Jones and Good, 2015; Lemmon et al., 2012). These methods are starting to play an important role in large-scale data generation for phylogenetic and ecological studies. Another appealing property of these methods is their potential to use highly degraded DNA from tissue recovered from museum specimens (Cronn et al., 2012; Faircloth et al., 2015; Mamanova et al., 2010).

All hybrid enrichment methods involve as a first step a phase of bioinformatic probe design, using as references genomic resources of species from the group of interest or related groups, generated *de novo* or available in public databases (Blaimer et al., 2015; Faircloth et al., 2012; Lemmon et al., 2012). The probe design phase allows the researchers to tailor probes to answer biological questions at deep or shallow evolutionary scales (Barrett et al., 2016). Moreover, the flexibility of hybrid enrichment methods allows to use previously available markers, that have been traditionally obtained through Sanger sequencing, into the target enrichment probe design (Grover et al., 2012). Probes, or "baits," are usually RNA oligonucleotides (~60–120 bp long) complementary to conserved regions of the genome such as exons (Lemmon et al., 2012; Portik et al., 2009; Sass et al., 2016), conserved orthologous sequences (COS; Mandel et al., 2014) or ultraconserved elements (UCEs; Faircloth et al., 2012; McCormack et al., 2012; Sun et al., 2014), flanked by highly-variable regions. Capturing both conserved and variable regions makes the sequences recovered potentially informative at a wide range of phylogenetic levels (Faircloth et al., 2012; Lemmon et al., 2012). Probes are designed such that captured DNA fragments overlap, allowing the targeted regions to be assembled into larger contigs, thus increasing the coverage when sequenced using next-generation sequencing (NGS) platforms (McCormack et al., 2013).

The next step in hybrid enrichment methods is library preparation, which involves DNA fragmentation, usually by sonication, and a subsequent ligation of adapters specific for particular high-throughput sequencing platforms. These adapters can be indexed to help identify samples from different individuals when they are pooled for sequencing (Lemmon and Lemmon, 2013; Prum et al., 2015; Young et al., 2016). Libraries are enriched with the designed probes and hybridization can be done either in solution (e.g. Gnirke et al., 2009; Faircloth et al., 2012; Lemmon et al., 2012) or in a solid structure such as microarrays (e.g. Bi et al., 2012). After hybridization, the captured DNA fragments are separated from the non-targeted DNA, usually by means of streptavidin-coated beads that attract biotinylated probes, or by washing away the background DNA from the microarrays (Mamanova et al., 2010). Before sequencing, targeted DNA is subjected to amplification using adapter primers to increase copy number (Prum et al., 2015; Young et al., 2016; Buddenhagen et al., 2016).

In land plants, different hybrid enrichment tools have been used to obtain phylogenomic data, including plastome sequencing through organelle capture in gymnosperms and angiosperms (Cronn et al., 2012; Heyduk et al., 2015; Parks et al., 2012; Stull et al., 2013), nuclear loci using exon capture (Grover et al., 2015; Heyduk et al., 2015; Mandel et al., 2014; Sass et al., 2016; Sousa et al., 2014; Stephens et al., 2015a; Stephens et al., 2015b), Anchored Hybrid Enrichment (Buddenhagen et al., 2016; Mitchell et al., 2017), and a combination of both plastid and nuclear data through Hyb-Seq (Folk et al., 2015; Schmickl et al., 2016; Weitemier et al., 2014).

1.2 The Anchored Hybrid Enrichment method

The Anchored Hybrid Enrichment (AHE) method is a genomic partitioning strategy originally designed for vertebrates, that uses probes to capture hundreds of loci (~500), that are potentially useful in recovering relationships at different phylogenetic depths. The universality of these probes results from designing them using genomic resources from representative species across the lineage of interest (Lemmon et al., 2012). The AHE method has been applied to different groups of animals, improving the resolution and support of their phylogenetic trees (Brandley et al., 2015; Eytan et al., 2015; Prum et al., 2015; Ruane et al., 2015; Hamilton et al., 2016; Stout et al., 2016; Tucker et al., 2016; Young et al., 2016). Recently, the AHE method has been applied successfully to flowering plants by Buddenhagen et al. (2016) and Mitchell et al. (2017), using a probe kit designed from publicly available and newly generated genomic resources for 25 species distributed across the phylogeny of flowering plants. The Angiosperm v. 1 kit (Buddenhagen et al., 2016) targets 517 nuclear loci that are potentially shared as a single copy among various lineages of angiosperms. This capture kit was tested in 53 species belonging to 10 angiosperm families distributed in six different orders, resulting in a high enrichment success (Buddenhagen et al., 2016) that makes it a promising tool for phylogenetic reconstruction in angiosperms (e.g. to resolve the rapid radiation in the genus *Protea*; Mitchell et al. 2017).

1.3 Phylogenetic noise reduction

Phylogenetic noise as defined by Straub et al. (2014) consists of some features of DNA sequences that contradict or obscure the phylogenetic signal of a true gene genealogy. This noise can have an impact in phylogenetic reconstruction, causing some nodes to have low support or providing high support for wrong relationships (Rokas and Carroll, 2006; Townsend et al., 2012; Wenzel and Siddall, 1999). In some cases, phylogenetic noise can be due to substitution saturation or alignment errors, and exclusion of unusually fast-evolving sites has been recommended as a strategy of noise reduction (Buddenhagen et al., 2016; Goremykin et al., 2015, 2010; Granados Mendoza et al., 2013; López-Giráldez and Townsend, 2011; Straub et al., 2014; Zhong et al., 2011).

1.4 *Salvia* subgenus *Calosphace* (Lamiaceae; Lamiales)

With ca. 900 species, *Salvia* L. (sages) is the most diverse genus of Lamiaceae, the mint family (Harley et al., 2004). *Salvia* is not monophyletic, and it contains three main lineages with other five genera embedded (Walker et al., 2004, 2015). Nevertheless, *Salvia* clade II (Walker et al., 2004) constitutes a lineage endemic to the New World that includes the only two monophyletic subgenera of *Salvia*: *Calosphace* (Benth.) Epling and its sister subgenus, *Audibertia* J. B. Walker, B. T. Drew & K. J. Sytsma (Walker et al., 2015). Subgenus *Calosphace* (in the following referred to simply as *Calosphace*), includes ca. 600 species and is the most species-rich lineage of *Salvia* (Epling, 1939; Santos, 1995). This subgenus attains its greater species richness in the Neotropics, and most of its diversity is found in Mexico (ca. 307 spp.; Martínez-Gordillo et al., 2013). Other important diversity centers are the Andes (155 spp.), eastern South America (60 spp.) and the Greater Antilles (45 spp.) (Jenks et al., 2013). The phylogenetic relationships among the species of *Calosphace* have been explored with DNA sequences of the nuclear-ribosomal internal transcribed spacers (ITS) and the non-coding chloroplast regions *psbA-trnH* and *trnL-trnF*. The phylogenies resulting from these combined "conventional" markers were poorly resolved, and many clades lacked support due to low sequence variation among closely related species. However, those markers permitted the identification of six main *Calosphace* lineages (Jenks et al., 2013). One such lineage is "core *Calosphace*," which contains most of the species that have been sampled in prior phylogenetic studies and is believed to represent a recent radiation, given its short internal branches and the lack of resolution along its backbone (Jenks et al., 2013, 2011; Walker, 2006).

Core *Calosphace* includes species from all four main diversity centers of the subgenus, and exhibits a wide variety of floral attributes related to pollination syndromes. Mexico and Central America have been inferred as the ancestral distribution area of the subgenus, and the ancestral pollination syndrome for the lineage is melittophily (Jenks et al., 2013; Wester and Claßen-Bockhoff, 2011). However, numerous events of migration seem to have occurred leading to the current distribution of the subgenus, and multiple origins of ornithophily have likewise been suggested. To explore further the evolutionary history of this lineage of Neotropical sages, a better resolved and supported phylogeny is required.

The present study uses 12 species of *Salvia* subgenus *Calosphace*, including 10 representatives core *Calosphace*, and a species of the closely related genus *Lepechinia*, to test the performance of phylogenomic data obtained through Anchored Hybrid Enrichment coupled with next generation sequencing, in resolving phylogenetic relationships within this plant group. The aims of this study are to: 1) evaluate the effectiveness of gene capture and high-throughput sequencing for *Calosphace*; 2) test the performance of the captured loci in resolving and supporting internal relationships in the group; and 3) compare the effect of different data-filtering schemes on phylogenetic reconstruction.

2. Material and methods

2.1 Sampling

2.1.1 Loci sampled

The nuclear loci sequenced correspond to those selected by Buddenhagen et al. (2016) for their Angiosperm v. 1 enrichment kit. As starting point, these authors used the 959 APVO genes (Duarte et al., 2010), which are single-copy nuclear genes shared by *Arabidopsis* Heynh., *Populus* L., *Vitis* L. and *Oryza* L. The corresponding exons of *A. thaliana* (L.) Heynh. were mined and only those long enough for probe design (≥150 bp) were selected, yielding 3,050 exons. Homologous regions from *Oryza* were extracted and compared to those of *A. thaliana* in order to filter exons with less than 55% similarity between these two species. Using the resulting 1,721 exons, orthologous regions were mined among 42 angiosperm reference genomes. The final 499 regions were selected based on their low average copy number (≤1.2) and their presence in at least 85% of the genomes used. Additionally, the authors included 18 selenium-tolerance loci, yielding a total of 517 regions. Finally, to make a universal flowering plant capture kit, the probe set was designed based on a selection of 25 angiosperm reference species, which are distributed across the angiosperm tree (Buddenhagen et al., 2016). Probes were designed to be ~120 bp long and tiled across the complete sequence of all the loci, overlapping by ~43 bp (2.8× tiling density).

2.1.2 Taxon sampling

The sample of taxa consisted of 12 species belonging to different clades of subgenus *Calosphace* and one species of *Lepechinia* Willd., the sister genus of *Salvia* s.l., as outgroup. Clade representation was based on a previous phylogenetic study (Jenks et al., 2013), with emphasis on species of core *Calosphace*. To evaluate the power of the AHE loci to resolve relationships at different time scales, the sister species of the remaining *Calosphace* species (*Salvia axillaris* Moc. & Sessé ex Benth.) and one species of the genus

Lepechinia were included, to represent ancient divergence events. Two species closely related to each other (*Salvia protracta* Benth. and *S. ramamoorthyana* Espejo Serna *emend.* J.G.González) were included to represent recent divergences. Sampled species and voucher information are provided in Table 1.

2.2 Molecular methods

2.2.1 DNA extraction

DNA was extracted from silica gel-dried leaf tissue using a modification of the 2× CTAB method (Doyle and Doyle, 1987) described in Salazar et al. (2003), adding incubation with RNase A (Qiagen) and Proteinase K (Promega). In some cases, DNA was precipitated using a 3M sodium acetate solution to increase purity. DNA concentration and purity ratios were measured using a Nanodrop 2000 spectrophotometer (Thermo Scientific).

2.2.2 Library preparation, enrichment and sequencing

The following steps were carried out at the Center for Anchored Phylogenomics at Florida State University (http://anchoredphylogeny.com/) and followed Buddenhagen et al. (2016). Genomic DNA was fragmented using a Covaris E220 Focused-ultrasonicator, to a fragment size of ~300–800 bp. The adapters and indexes were linked to the fragmented DNA using a Beckman-Coulter Biomek FXp liquid-handling robot, using a modification of the protocol of Meyer and Kircher (2010). Once indexed, samples were pooled and one solution-based enrichment reaction was carried out using the Angiosperm v. 1 kit (Agilent Technologies Custom SureSelect XT kit). Enriched DNA was separated from non-hybridized DNA using streptavidin magnetic beads. The enrichment reaction was sequenced in one PE150 Illumina HiSeq 2500 lane at the Translational Science Laboratory in the College of Medicine at Florida State University, Tallahassee, Florida, USA.

2.3 Bioinformatic data processing

2.3.1 Raw data processing

Raw reads were processed with the CASAVA v. 1.8 pipeline with the high-chastity setting to filter low-quality reads. Reads were demultiplexed using 13 in-house developed indexes which are 8 bp long and with at least two different base pairs. Reads that failed to match any of the expected indexes were discarded.

2.3.2 Pair-read merging and assembly

Reads were merged based on the method proposed by Rokyta et al. (2012), to increase their accuracy and length. Unmerged reads were also used for assembly. Read assembly followed Prum et al. (2015) and Buddenhagen et al. (2016). The read assembler employs a two-level strategy to achieve sequence assembly. The first level is the divergent reference assembly, which uses a selection of the species (*Arabidopsis thaliana*, *Billbergia nutans* H.Wendl. and *Carex lurida* Wahlenb.) used during probe design. These species are distantly related to the study group and are employed as references for the assembly to map the reads to the conserved regions. The second level is a quasi-de novo assembly that uses a preliminary consensus sequence from the study group, generated in the previous step, to extend the mapping to less conserved flanking regions (Prum et al., 2015). Both levels of assembly were used to traverse repeatedly the read files until no additional reads were mapped. Once assembly clusters were formed, consensus bases were called. If a polymorphism could not be explained as a sequencing error, ambiguity bases were called, and if the coverage was lower than ten, the bases called were marked as N. Consensus sequences with less than 30 reads were discarded to prevent cross contamination or sequencing errors in index reads.

2.3.3 Orthology assessment

Each locus should produce a single cluster in the absence of contamination, low coverage or gene duplication. To identify loci with duplications and to discern between potential loci copies, orthologous sequences were determined following Prum et al. (2015) and Buddenhagen et al. (2016). The sequences were grouped by locus and a pairwise distance between two sequences was calculated as the percent of 20-mers found in both sequences. A distance matrix was constructed with this information and, using the neighbour-joining algorithm (Saitou and Nei, 1987), sequences were clustered, allowing a single sequence per species per locus. When more than one cluster was detected within a region, clusters of orthologs were separated and considered as different loci. When gene duplication occurs within a clade a subset of taxa will lack the locus, resulting in missing data. To reduce the effect of missing data, clusters including less than 50% of the species were discarded.

2.3.4 Alignment and alignment trimming

Each locus was aligned using MAFFT v. 7.023b (Katoh and Standley, 2013) with "-genafpair" and "maxiterate 1000" flags. Alignments were trimmed/masked in three steps: 1) Each column of the alignment was considered "conserved" if >40% of the sequences presented the most commonly observed character. 2) Regions of 20 bp in each sequence were scanned to identify those that lacked at least 10 characters matching the common base at a conserved site; these regions were not well aligned and were masked. 3) Sites were removed from the alignments when they presented fewer than 12 unmasked bases. Alignments for all loci were retrieved, including those missing the outgroup.

2.4 Evaluation of locus performance in phylogenetic resolution and support

The performance of captured loci in resolving internal relationships in *Calosphace* was evaluated with two different criteria: 1) phylogenetic reconstruction (concatenated and coalescent-based analyses); and 2) phylogenetic informativeness (Townsend, 2007). The results yielded by the original dataset were compared to the ones obtained from the filtered datasets under four different scenarios (see Data screening and matrix filtering, below).

2.4.1 Concatenated analyses

The trimmed matrices of the captured loci were concatenated and analysed with maximum likelihood (ML) using the programme RAxML-HPC2 on XSEDE v. 8.2.8 (Stamatakis, 2014), implemented in the CIPRES Science Gateway (Miller et al., 2010). The phylogenetic trees were inferred under the GTRGAMMA model, using the "rapid bootstrapping and search for the best-scoring ML tree" algorithm (Stamatakis et al., 2008), with 1,000 bootstrap replicates. Two partition schemes were employed for the concatenated matrix: 1) partitions per locus (448 partitions) and 2) partitions selected by PartitionFinder v. 1.1.1 (Lanfear et al., 2012), which permitted to explore the effect of model assignment on tree resolution and support. A concatenated matrix of the conventional markers (ITS, *trnL-trnF* and *trnH-psbA* IGS) from Fragoso-Martínez et al. (in review) was analysed under the same parameters for comparison purposes.

2.4.2 Coalescent-based analyses

Unrooted gene trees were inferred using RAxML-VI-HPC (Stamatakis, 2006) through a workflow in Geneious v.8.1.8 (http://www.geneious.com, Kearse et al., 2012) using the same model and algorithm used in the concatenated analyses, but with 100 bootstrap perturbations. To infer the species tree we employed ASTRAL (Mirarab et al., 2014) and support was calculated using local posterior probabilities, which have shown to be more precise than multi-locus bootstrapping (Sayyari and Mirarab, 2016). The quartet approach implemented in the ASTRAL algorithm makes it time-efficient and has proven to provide more accurate results than other coalescent-based and concatenated methods (Mirarab et al., 2014). Additionally, due to the small number of taxa sampled, we were able to find the globally optimal species trees using the exact version of ASTRAL (Mirarab et al., 2014) and, because ASTRAL does not require rooted gene trees, we could use all captured loci in the species tree inference, instead of only alignments that contained the outgroup.

2.4.3 Phylogenetic informativeness analyses

To evaluate the performance of the captured loci in the phylogenetic reconstruction of *Calosphace*, we performed a Phylogenetic Informativeness (PI) analysis (Townsend, 2007). For this analysis, the ML tree from the concatenated analysis was made ultrametric with a relative time scale of 0 at the tips and 1 at the root, using the non-parametric rate smoothing method (Sanderson, 1997), implemented in TreeEdit v 1.0a10 (Rambaut, 2002). The partitioned concatenated matrices and the reference ultrametric tree were uploaded to PhyDesign, an online application for profiling phylogenetic informativeness (López-Giráldez and Townsend, 2011; http://phydesign.townsend.yale.edu/). Substitution rates were estimated in HyPhy (Pond et al., 2005) using the GTR model with empirical base frequencies. A first PI analysis was performed using the alignments obtained after the bioinformatic data processing and the ML tree from those aligments, and a subsequent PI analyses were performed over filtered concatenated matrices (see next section), and the resulting ML trees.

2.5 Data screening and matrix filtering

Sites with unusually high substitution rates were visualized as “phantom” spikes on the individual PI profiles (Granados Mendoza et al., 2013; López-Giráldez and Townsend, 2011; http://phydesign.townsend.yale.edu/). To identify the specific positions in the alignments that could be introducing phylogenetic noise due to their unusually high substitution rate, we developed a R (R Core Team, 2014) script. This script (Appendix A) takes as input the spreadsheets with the estimated substitution rate per locus from HyPhy (Pond et al., 2005; obtained from the PhyDesign tool (López-Giráldez and Townsend, 2011; http://phydesign.townsend.yale.edu/) and constructs a new data frame including the name of the locus, position of the sites and their substitution rate. These data frames were filtered by removing sites with substitution rates lower than a given value. In this study, we tested four different substitution rate thresholds (i.e. deleting sites with rate values higher than 5, 10, 15 and 20), but these values can be modified by the user after observing the behaviour of the dataset in question. The output of the script is a spreadsheet of positions with high substitution rates and scatterplot files per locus that can be used to explore the behaviour of each alignment. Sites included in the spreadsheets were manually removed from the alignments from the AHE bioinformatic process using Geneious v.8.1.8 (http://www.geneious.com, Kearse et al., 2012). The scatterplots of the substitution rates of each site per locus and the list of the removed sites from each scenario are provided in Appendix A. The concatenated matrices, including the original matrix and the four filtered matrices are available in Mendeley data (doi:10.17632/z6zy58vrmt.1).

3. Results

3.1 Gene capture and high-throughput sequencing for *Calosphace*

From the 517 loci targeted by the plant capture probe set, 399 were captured. From these, 38 genes had 2 to 7 copies that were separated into different matrices, yielding 448 alignments, 370 of which included the outgroup. *Salvia ramamoorthyana* had the best capture (441 loci), whereas *S. axillaris* was the ingroup species for which the fewest loci were captured (378). The length of the alignments ranged from 158 to 1891 bp, with an average length of 704 bp. Sixty four percent of the alignments included all 13 taxa and only in 21% of the alignments there were more than one missing species. The concatenated matrix comprised 316,701 bp.

3.2 Performance of the AHE data

3.2.1 Phylogeny estimation within subgenus *Calosphace*

The phylogenetic hypothesis derived from the concatenated AHE data (Fig. 1a) showed increased support and resolution compared to the one obtained with conventional markers (Appendix A, Fig. A.1). In the tree resulting from the AHE dataset, most of the internal branches were supported by the highest bootstrap value (100%), including the core *Calosphace* clade. Three short branches had lower support and only one had <85% bootstrap support (BS). These branches are: branch 1 (B1; 92% BS) and branch 2a (B2a; 58% BS), which include *S. helianthemifolia* and *S. hispanica*, and these two species plus *S. connivens,* respectively. Branch 3 (B3; 86% BS) is a short deep branch within core *Calosphace*, which includes *S. melissodora* plus the eight remaining sampled species of core *Calosphace* (Fig. 1a). The BS values of the concatenated analysis using the partition schemes estimated by PartitionFinder (Lanfear et al., 2012) were similar to those using the 448 partitions, albeit slightly lower (Appendix A, Figs. A.7–A.9). In the tree resulting from conventional markers (Appendix A, Fig. A.1) most of the internal branches were supported by >85% BS and only the clade formed by *S. protracta* and *S. ramamoorthyana* had 100% BS. Four branches had <85% BS, two of them along the backbone of the core *Calosphace* clade, one of them equivalent to B3 of the concatenated AHE analysis. The other two branches with low support included the same species than B1 and B2a. There are topological incongruences between the phylogenetic trees from the analyses of both concatenated AHE and conventional markers, all involving the species from the above-mentioned branches (Appendix A, Fig. A.1).

The exact ASTRAL search of the 448 gene trees resulted in a species tree with most of its internal branches showing high local posterior probabilities (1 LPP) and three short branches with low LPP values (Fig. 1b). The branches that did not receive the highest support were the same in both the concatenated and the coalescent-based analyses from the AHE data. The branches B1 and B3 contain clades consisting of the same species than in the concatenated analyses, and B2b, in addition to *S. connivens* (B2a), included as well *S. mexicana*. This node was the only topological incongruence between both methods; however, this relationship lacked support in the coalescent-based tree (Fig. 1a–b).

3.2.2 Phylogenetic informativeness

The amount of net phylogenetic informativeness varied among the captured loci, ranging from 2.8 to 303.5, with an average of 84.9. Most loci (65%) reached their maximum value at the time interval of 0.11–0.23, which includes B2a and the branch below B3 in Fig. 2a. All loci peaked at time 0.36, prior to the divergence of the core *Calosphace* clade. Most of the PI curves showed a steady increase in phylogenetic informativeness until the maximum value was reached. However, the PI curves of several of loci showed “phantom” spikes towards the present, as a result of a disproportionate increase in the substitution rate of a few characters relative to the remaining sites (Fig. 2a).

Regarding the conventional markers, ITS had higher net phylogenetic informativeness values than the two chloroplast markers. However, ITS reached its maximum value (168.2) before time 0.06, after the divergence of the core *Calosphace* clade. Both plastid markers had a similar behaviour, with lower PI values, that decreased slowly through time, and showed “phantom” spikes towards the present (Appendix A, Fig. A.2)

3.3 Comparison of the impact of different filtering schemes

3.3.1 Phylogenetic reconstruction

After the inspection of the PI curves (Fig. 2a) and the substitution rate scatterplots of each locus (Appendix A), the presence of a few loci with “outlier” sites was evident. Due to the great variation in the substitution rate values across loci, we decided to exclude sites with rates higher than an arbitrary threshold value. We evaluated the impact on phylogenetic reconstruction of four different filtering scenarios (Table 2), i.e. removing sites with substitution rates higher than 5, 10, 15 and 20.

In most filtering scenarios, the trees obtained from the concatenated and coalescent-based analyses had some topological incongruences, albeit without high support. Trees derived from the same type of analysis but subjected to a different filtering scenario varied only in branch support. In the analyses of the concatenated matrices the gradual removal of sites with high substitution rates resulted in an increase of bootstrap support for the shallow branches (B1 and B2a, Fig. 2). The BS values showed a tendency to increase as more sites were removed. In the branch B3, while support values increased with site removal, this amount was not substantial and the BS values remained between 83–89%, decreasing below 85% in the strictest scheme, i.e. removal of substitution rates >5 (Fig. 3a–b; Appendix A, Figs. A.3–A.9). Support values for those three branches were higher in the scheme where sites with substitution rates higher than 10 were removed (Fig. 3a–b), than in the unfiltered dataset.

In the coalescent-based analyses, although the Local Posterior Probability (LPP) values increased in some scenarios, none of the changes were significant, unlike the analyses of the concatenated matrices. The LPP values of B3 were the highest among the branches of the three problematic nodes, and these values were stable across different schemes, behaving similarly to the concatenated analyses. In turn, the LPP values of B1 and B2b fluctuated strongly. The support value of B2b showed a tendency to increase, while that of B1 tended to decrease towards the strictest scheme (Fig. 3c; Appendix A, Figs. A.3–A.6).

3.3.2 Phylogenetic informativeness profiles

In general, the gradual removal of sites with high substitution rates (SR) resulted in smoother PI curves than those from the original matrix (Fig. 2; Appendix A, Figs. A.10–A.12). However, in the less strict scheme (SR> 20 removed) "phantom" spikes can still be distinguished in the profiles, although with lower peaks than those exhibited by the original dataset (Appendix A, Fig. A.10). In the remaining scenarios, no "phantom" spikes are visible and all the curves are smooth (Fig. 2b; Appendix A, Figs. A.11–A.12).

4. Discussion

4.1 Efficiency of AHE data for resolving relationships in *Salvia* subgenus *Calosphace*

As noted earlier, the probe design of the Angiosperm v. 1 kit (Buddenhagen et al., 2016) was based on the genomes of 25 angiosperms distributed along the flowering plant phylogeny to make the probes useful at a broad scale. Among those 25 species, the closest relative to *Salvia* is *Mimulus guttatus* DC. (Phrymaceae), also within order Lamiales. The fact that we were able to recover 399 loci (plus 49 additional orthologs) for our species of Lamiaceae, even though probe design was not based on a species of this family, confirms the universality of the Angiosperm v. 1 kit and suggests that this method could work for other families of the Lamiales. Our capture success suggests that the Angiosperm v. 1 kit, in combination with the AHE method, may allow to standardize loci sampling across major angiosperm lineages to produce large-scale phylogenomic trees.

4.2 Performance of captured phylogenomic data

4.2.1 Phylogenetic relationships

The phylogenetic tree from the concatenated matrix of the AHE data was better, in terms of support and resolution, than the tree obtained from the analysis of the concatenated matrix of conventional markers. In the tree derived from phylogenomic data all of the deep internal branches and most of the shallow ones were highly supported. The tree obtained represents a substantial advance in our understanding of phylogenetic relationships within core *Calosphace*, as phylogenetic resolution along the backbone of the clade had not been achieved previously. These results suggest that AHE data are a promising resource for phylogenetic reconstruction of this, and potentially other recently derived groups, even in the presence of short branches.

It is worth noting that the branches that did not achieve the highest levels of support using the AHE data involve the same clades that exhibited weak support with conventional markers. The branches B1, B2 and B3 received low support, even when using different datasets and inference methods (Fig. 1a–b; Appendix A, Fig. A.1, A.7). In the ML tree from the concatenated analysis of AHE data (Fig. 1a), the branches above B1–B3 are all 7–11 times longer than the extremely short problematic branches. These short branches are particularly difficult to resolve due to the combination of a low probability of synapomorphies to arise in such short time window involved, and the higher probability of homoplastic characters (e.g. by saturation resulting from multiple, independent changes to the same base) in the long branches above them, which could mislead phylogenetic reconstruction (Rokas and Carroll, 2006; Townsend et al., 2012).

Usually, short branches with low support below long branches have been interpreted as a rapid radiation (Rokas and Carroll, 2006). This could be a plausible explanation for the pattern observed here, since all the "problematic" branches of the tree are within the core *Calosphace* clade. Moreover, the lack of support on B1–B3 in the coalescent-based analysis suggests that there are incongruences among the gene trees at those nodes, which could be due to the small taxon sample analyzed or to incomplete lineage sorting, or reflect hybridization events. To resolve these problematic short branches, additional sampling will be needed, not only in characters, but also including more taxa, as suggested by Townsend and Leuenberger (2011). The latter could have a greater impact since many of the autapomorphies in the matrix could turn into synapomorphies with an expanded taxon sampling.

4.2.1 Phylogenetic informativeness

The combination of the difference in the amount of phylogenetic informativeness and the difference in the time at which each locus peaks led to a robust phylogenetic hypothesis for *Calosphace*, with most of the branches having strong support. When similar PI values are obtained, loci that peak at deeper phylogenetic levels than a time of interest are preferred for phylogenetic reconstruction; this choice minimizes the probability of including characters that could have evolved to convergent states (Townsend and Leuenberger, 2011). In this case, the fact that 21 loci peaked deeper than the core *Calosphace* crown node further supports the potential of the AHE loci for resolving the backbone of this clade.

During substitution rate estimation, the rates of certain sites in the alignments (usually indels or ambiguous base calls) cannot be accurately estimated by ML. This lack of accuracy causes the substitution rate values for these sites to be estimated as fast evolving sites (i.e. with high substitution rates), which are visualized in the PI profiles as “phantom” spikes towards the tips of the tree (López-Giráldez and Townsend, 2011; http://phydesign.townsend.yale.edu/). The presence of these spikes in the curves of some loci suggests that certain sites of those loci might introduce phylogenetic noise (Granados Mendoza et al., 2013). Because these spikes occur shallower than B1 and B2a, they may be introducing phylogenetic in them, which could account for the low support values obtained, compared with the highly supported adjacent branches. Regarding B3, “phantom” spikes are not affecting its resolution and support, but the obtained low BS values could depend on a combination of factors. First, loci with high peaks in recent times are likely to lead to noise at deeper phylogenetic levels. As most of the loci peaked before B3, it is expected for the curves to decrease deeper in the phylogeny, with a concomitant signal decrease and noise increase (Townsend and Leuenberger, 2011). This effect is due to the decrease in the number of sites evolving at an optimal rate for PI after the maximum PI is reached, which in turn increases the probability of introducing phylogenetic noise (Townsend and Leuenberger, 2011). Second, the short length of B3 might result in greater sensitivity to phylogenetic noise, as the latter may outweigh the signal in the data (Townsend et al., 2012; Townsend and Leuenberger, 2011).

An effort to resolve these conflicts could involve increasing the number of loci sampled, in particular loci that peak deeper than conflicting nodes, which have greater potential to resolve polytomies while introducing minimum noise (Townsend et al., 2012; Townsend and Leuenberger, 2011). However, taxon sampling could have a greater impact on resolution and support (Townsend and Leuenberger, 2011). If taxon sampling were to be increased, it is recommended to base it on previous phylogenetic hypotheses such that added taxa represent branches deeper than the clade of interest, as sampling recently derived taxa will have little effect on phylogenetic informativeness (Townsend and Lopez-Giráldez, 2010). Furthermore, because many loci have unusually fast substitution rates in the sample of taxa analysed here, adding more taxa will be a more productive approach in using this kind of data for phylogenetic reconstruction (Townsend and Leuenberger, 2011). Increasing taxon sampling has been suggested as a strategy to reduce or eliminate “phantom” spikes, as the addition of new taxa can aid in the estimation of substitution rates of sites with otherwise poor estimations (López-Giráldez and Townsend, 2011; http://phydesign.townsend.yale.edu/). The latter suggests that some of the sites here classified as having high substitution rates could be classified differently in the context of an expanded taxonomic sample, as calculations of substitution rates depend on the alignments but also on the branch lengths in the reference phylogeny. The addition of taxa would modify the branch lengths, which in turn will result in different substitution rate estimates for the sites in the alignment.

Increased support for “problematic” branches found here will require two distinct sampling approaches. Branches 1 and 2 are shallow, and occur in a time interval characterized by high informativeness, where poor resolution could result from close divergences (Townsend, 2007), but they are also affected by the phylogenetic noise introduced by the “phantom” spikes. In turn, B3 is a deeper short branch, in a phylogenetic level where data has low informativeness. In these cases, the addition of more data (i.e. loci or taxa) is recommended (Townsend 2007). In particular, an increased taxon sampling would have a greater impact on both kinds of branches, potentially decreasing the noise and resolving the relationships. Therefore, a future study of phylogenetic relationships of *Calosphace* should ideally include a denser sampling of the deepest lineages within the core *Calosphace* clade to better resolve B3, whereas a denser sampling within the clade with *S. atrocyanea* as its deepest-branching lineage could improve support for B1 and B2, or alternatively, would confirm that these nodes are the result of a rapid radiation.

4.3 Impact of character removal on phylogenetic hypotheses

As mentioned earlier, the “phantom” spikes that appear in some of the PI curves are the result of sites usually having either indels or ambiguous base calls. Although it is difficult to appropriately estimate substitution rates for these sites, they are estimated to evolve at very fast rates, for which their removal is advisable (López-Giráldez and Townsend, 2011; http://phydesign.townsend.yale.edu/). In the concatenated analyses, the impact of removing sites with high substitution rates was positive, resulting in an increase in clade support for recent divergences. The support values for B1 and B2 showed a tendency to increase when more data were removed, which is also consistent with the disappearance of “phantom” spikes on most of the filtering schemes. Particularly, support of B2a showed a dramatic increase (from 58% to 97% BS) when the less strict scheme (i.e. removal of substitution rates >20) was applied to the dataset (Fig. 3a–b; Appendix A, Fig. A.3–A.9). Such increases in BS values when the “phantom” spikes were removed confirm that those sites were introducing noise into the phylogenetic inference, affecting directly B1 and B2a (Fig. 2). The fact that BS values for these branches improved upon character removal suggests that, if the clades associated with B1 and B2a represent a rapid radiation as hypothesized before, processes such as sequence saturation or incomplete lineage sorting could be affecting phylogenetic reconstruction. Support values for B3 also increased with character removal, though slightly, and decreased with the strictest scheme, which involved the exclusion of most of the sites. The latter suggests that the removal of certain amount of data (from 0.08–0.16% of the data from the original matrix, Table 2, Fig. 3a–b) can improve resolution and support on this branch by reducing the slopes in the profiles, and causing the curves to decrease in a steadier fashion, resulting in a slight decrease in noise at B3 (Fig. 2b, Appendix A, Figs. A.10–A.12). However, the tendency of BS values to decrease in B3 when data removal increases (removal of 0.33% of the data from the original matrix, i.e.

strictest scheme; Appendix A, Figs. A.12) agrees with the hypothesis discussed earlier, regarding the need to add more data to resolve this node (Townsend, 2007). Since there is already a lack of data in the dataset to resolve B3, the removal of significant amounts of data (as in the strictest scheme, Table 2) can result in removing also informative characters, i.e. synapomorphies, for B3. An additional interpretation could be that B3 is only supported by noise. If this were the case, removal of sites with faster rates would decrease its support.

In the coalescent-based analyses, although there was fluctuation among the LPP values across different removal schemes (especially in B1 and B2b), none of these changes were sufficient to make these probability values significant (Fig. 3c). These results agree with the fact that coalescent-based methods are more robust to high substitution rates than concatenated methods (Xi et al., 2014). Low LPP values suggest incongruences among individual gene trees that affect those branches, which could result from lack of data to support certain parts of the phylogeny, or from incomplete lineage sorting, which is expected given the recent history of *Calosphace* (~19.1–11.9 Ma since its estimated split from subgenus *Audibertia*; Walker et al., 2015).

Although all the exclusion schemes applied resulted in higher BS values on at least two of the problematic branches than the original dataset, none produced the highest BS values for all three branches. This means that there is not a single scheme that on itself will increase BS support on all branches, and it is highly recommended the exploration of several filtering schemes. Most of the sites with high substitution rates that were removed contained unresolved (N) or missing information for at least one taxon, which resulted in lack of information. However, sites involving transversions also had high substitution rates, sometimes indistinguishable from those caused by lack of information, suggesting that during character removal some informative sites could be excluded along with “noisy” characters. It would be desirable to develop an algorithm to distinguish between high substitution rates resulting from missing information or transversions, to test the effect of the removal of these two kinds of data.

5. Conclusion

Loci capture using the Angiosperm v. 1 kit (Buddenhagen et al., 2016) with the AHE method in a New World sage lineage (*Salvia* subgenus *Calosphace*), was successful. We were able to capture 78% of the 517 genes targeted by the kit, confirming its universality. We obtained 448 alignments, over one half of them including sequences for all the sampled taxa, and the remaining missing less than 30% of the taxa. The success of the AHE method in our study group is encouraging and suggests that other lineages from Lamiales could potentially recover similar amounts of data.

The AHE data allowed the recovery of strongly supported phylogenetic trees for *Calosphace*, with strong support for most of the internal branches. These phylogenetic trees were superior in terms of support and resolution over those based on conventional markers. However, problematic branches with low support were still found. In both the concatenated and coalescent-based phylogenetic reconstruction methods applied here, these extremely short branches with limited support subtended groups that included the species: *S. hispanica*, *S. helianthemifolia*, *S. connivens*, *S. mexicana* and *S. melissodora*. The phylogenetic informativeness analysis showed that the recovered loci varied greatly in their net phylogenetic informativeness and its temporal distribution, suggesting that these data are promising for the reconstruction of the relationships of our model group. Nevertheless, noise in some loci is suggested by “phantom” spikes in the recent portion of phylogenetic informativeness curves.

The removal of sites with high substitution rates to remove noise resulted in the elimination of the “phantom” spikes and an increase in bootstrap support in the shallower problematic branches. Although the support values increased for the problematic deep branch, progressive removal of greater amounts of data had a negative effect on its support values. We suggest that an exploration of the PI profiles to detect “phantom” spikes, and a subsequent filtering of the data testing different rate substitution removal scenarios should be a step prior to phylogenetic inference using this kind of phylogenomic data. Also, increased taxon sampling is likely to aid in the resolution of short branches and the reduction or elimination of “phantom” spikes.

Acknowledgements

I.F.M. acknowledges the support from the Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México to conduct her doctoral studies. Financial support for this study through a Graduate Student Research Grant from the American Society of Plant

Taxonomists (ASPT), the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (UNAM-DGAPA-PAPIIT, project IG200316). A scholarship for doctoral studies from Consejo Nacional de Ciencia y Tecnología (CONACyT, Mexico) to I.F.M (grant 324065) and a postdoctoral grant from the Dirección General de Asuntos del Personal Académico (DGAPA-UNAM, 2014–2016) to C.G.M., are gratefully acknowledged. The authors thank Miguel Castañeda, Daniel Sandoval, Guadalupe Cornejo, Guillermo Ibarra, Gabriel González, Andrea Cocucci and Alicia Sérsic for providing tissue samples and/or assistance during fieldwork; the staff of the herbaria CORD, FCME and MEXU for access to the collections in their charge; María del Rosario García-Peña for suggestions on the taxonomic sampling; Adriana Benítez and Lidia Cabrera for assistance with laboratory work; and Alejandra Moreno-Letelier and two anonymous reviewers for suggestions that improved the manuscript.

References

Barrett, C.F., Bacon, C.D., Antonelli, A., Cano, Á., Hofmann, T., 2016. An introduction to plant phylogenomics with a focus on palms. Bot. J. Linn. Soc. 182, 234–255. doi:10.1111/boj.12399.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., Good, J.M., 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13, 403. doi:10.1186/1471-2164-13-403.

Blaimer, B.B., Brady, S.G., Schultz, T.R., Lloyd, M.W., Fisher, B.L., Ward, P.S., 2015. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. BMC Evol. Biol. 15, 271. doi:10.1186/s12862-015-0552-5.

Brandley, M.C., Bragg, J.G., Singhal, S., Chapple, D.G., Jennings, C.K., Lemmon, A.R., Lemmon, E.M., Thompson, M.B., Moritz, C., 2015. Evaluating the performance of Anchored Hybrid Enrichment at the tips of the tree of life: a phylogenetic analysis of Australian Eugongylus group scincid lizards. BMC Evol. Biol. 15, 62. doi:10.1186/s12862-015-0318-0.

Buddenhagen, C., A. R. Lemmon, E. C. Lemmon, J. Bruhl, J. Cappa, W.L. Clement, M.J. Donoghue, E.J. Edwards, A.L. Hipp, M. Kortyna, N. Mitchell, A. Morre, C.J. Prichid, M. C. Segovia-Salcedo, M.P. Simmons. P.S. Soltis, S. Wanke, A. Mast. 2016. Anchored phylogenomics of angiosperms I: assessing the robustness of phylogenetic estimates. bioRxiv http://dx.doi.org/10.1101/086298

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V., Udall, J., 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99, 291–311. doi:10.3732/ajb.1100356.

Doyle, J., Doyle, J., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.

Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, J.C., Leebens-Mack, J., dePamphilis, C.W., 2010. Identification of shared single copy nuclear genes in *Arabidopsis, Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10, 61. doi:10.1186/1471-2148-10-61.

Epling, C., 1939. A revision of *Salvia* subgenus *Calosphace*. Repert. Specierum Nov. Regni Veg. 110, 380.

Eytan, R.I., Evans, B.R., Dornburg, A., Lemmon, A.R., Lemmon, E.M., Wainwright, P.C., Near, T.J., 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. BMC Evol. Biol. 15, 113. doi:10.1186/s12862-015-0415-0.

Faircloth, B.C., Branstetter, M.G., White, N.D., Brady, S.G., 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Resour. 15, 489–501. doi:10.1111/1755-0998.12328.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–26. doi:10.1093/sysbio/sys004.

Folk, R. a., Mandel, J.R., Freudenstein, J. V., 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: a phylogenomic example from *Heuchera* (Saxifragaceae). Appl. Plant Sci. 3, 1500039. doi:10.3732/apps.1500039.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S., Nusbaum, C., 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–9. doi:10.1038/nbt.1523.

Goremykin, V.V., Nikiforova, S.V., Cavalieri, D., Pindo, M., Lockhart, P., 2015. The root of flowering plants and total evidence. Syst. Biol. 64, 879–891. doi:10.1093/sysbio/syv028.

Goremykin, V.V., Nikiforova, S.V., Bininda-Emonds, O.R.P., 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol. 71, 319–331. doi:10.1007/s00239-010-9398-z.

Granados Mendoza, C., Wanke, S., Salomo, K., Goetghebeur, P., Samain, M.-S., 2013. Application of the phylogenetic informativeness method to chloroplast markers: a test case of closely related species in tribe Hydrangeeae (Hydrangeaceae). Mol. Phylogenet. Evol. 66, 233–242. doi:10.1016/j.ympev.2012.09.029.

Grover, C.E., Gallagher, J.P., Jareczek, J.J., Page, J.T., Udall, J.A., Gore, M.A., Wendel, J.F., 2015. Re-evaluating the phylogeny of allopolyploid *Gossypium* L. Mol. Phylogenet. Evol. 92, 45–52. doi:10.1016/j.ympev.2015.05.023.

Grover, C.E., Salmon, A., Wendel, J.F., 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. Am. J. Bot. 99, 312–319. doi:10.3732/ajb.1100323.

Hamilton, C.A., Lemmon, A.R., Lemmon, E.M., Bond, J.E., 2016. Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. BMC Evol. Biol. 16, 212. doi: 10.1186/s12862-016-0769-y.

Harley, R.M., Atkins, S., Budanstev, A.L., Cantino, P.D., Conn, B.J., Grayer, R., Harley, M.M., de Kok, R., Krestovskaja, T., Morales, R., Paton, A.J., Ryding, O., Upson, T., 2004. Labiatae. in: Kubitzki, K., Kadereit, J.W. (Eds.) The Families and Genera of

Vascular Plants 7, Lamiales (except Acanthaceae including Avicenniaceae). Springer-Verlag., Berlin, pp. 167–275.

Heyduk, K., Trapnell, D.W., Barrett, C.F., Leebens-Mack, J., 2015. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. Biol. J. Linn. Soc. 117, 106–120. doi:10.1111/bij.12551.

Jenks, A.A., Walker, J.B., Kim, S.-C., 2013. Phylogeny of New World *Salvia* subgenus *Calosphace* (Lamiaceae) based on cpDNA (psbA-trnH) and nrDNA (ITS) sequence data. J. Plant Res. 126, 483–496. doi:10.1007/s10265-012-0543-1.

Jenks, A.A., Walker, J.B., Kim, S.-C., 2011. Evolution and origins of the mazatec hallucinogenic sage, *Salvia divinorum* (Lamiaceae): a molecular phylogenetic approach. J. Plant Res. 124, 593–600. doi:10.1007/s10265-010-0394-6.

Jones, M.R., Good, J.M., 2015. Targeted capture in evolutionary and ecological genomics. Mol. Ecol. 25, 185–202. doi:10.1111/mec.13304.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., Drummond, A., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28, 1647–1649.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. doi:10.1093/molbev/mst010.

Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–44. doi:10.1093/sysbio/sys049.

Lemmon, E.M., Lemmon, A.R., 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44, 99–121. doi:10.1146/annurev-ecolsys-110512-135822.

López-Giráldez, F., Townsend, J.P., 2011. PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evol. Biol. 11, 152. doi:10.1186/1471-2148-11-152.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J., 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118. doi:10.1038/nmeth.1419.

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., 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl. Plant Sci. 2, 1300085. doi:10.3732/apps.1300085.

Martínez-Gordillo, M., Fragoso-Martínez, I., García-Peña, M., Montiel, O., 2013. Géneros de Lamiaceae de México, diversidad y endemismo. Rev. Mex. Biodivers. 84, 30–86. doi:10.7550/rmb.30158.

McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22, 746–54. doi:10.1101/gr.125864.111.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66, 526–38. doi:10.1016/j.ympev.2011.12.007.

Meyer, M., Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 5, 1–11. doi:10.1101/pdb.prot5448.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees, in: Proceadings of the Gateway Computing Environments Workshop (GCE). New Orleans, pp. 1–8. doi:10.1109/GCE.2010.5676129.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., S. Swenson, M., Warnow, T., 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 30, 541–548. doi:10.1093/bioinformatics/btu462.

Mitchell, N., Lewis, P.O., Moriarty Lemmon E., Lemmon, A.R., Holsinger, K.E., 2017. Anchored phylogenomics resolves the evolutionary relationships in the rapid radiation of *Protea* L. (Proteaceae). Am. J. Bot. 104, 102–115.

Parks, M., Cronn, R., Liston, A., 2012. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). BMC Evol. Biol. 12, 100. doi:10.1186/1471-2148-12-100.

Parks, M., Cronn, R., Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7, 84. doi:10.1186/1741-7007-7-84.

Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: Hypothesis testing using phylogenies. Bioinformatics 21, 676–679. doi:10.1093/bioinformatics/bti079.

Portik, D.M., Smith, L.L., Bi, K., 2016. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). Mol. Ecol. Resour. 16, 1069–1083. doi:10.1111/1755-0998.12541.

Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569–572. doi:10.1038/nature15697.

Rambaut, A., 2002. TreeEdit, Version 1.0a10. Available at: http://tree.bio.ed.ac.uk/software/treeedit/ (accessed 08/2016).

Rokas, A., Carroll, S.B., 2006. Bushes in the tree of life. PLoS Biol. 4, 1899–1904. doi:10.1371/journal.pbio.0040352.

Rokyta, D.R., Lemmon, A.R., Margres, M.J., Aronow, K., 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). BMC Genomics 13, 312. doi:10.1186/1471-2164-13-312.

Ruane, S., Raxworthy, C.J., Lemmon, A.R., Lemmon, E.M., Burbrink, F.T., 2015. Comparing species tree estimation with large anchored phylogenomic and small

Sanger-sequenced molecular datasets: an empirical study on Malagasy pseudoxyrhophiine snakes. BMC Evol. Biol. 15, 221. doi:10.1186/s12862-015-0503-1.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evo 4, 406–425.

Salazar, G.A., Chase, M.W., Soto Arenas, M.A., Ingrouille, M., 2003. Phylogenetics of Cranichideae with emphasis on Spiranthinae (Orchidaceae, Orchidoideae): evidence from plastid and nuclear DNA sequences. Am. J. Bot. 90, 777–795. doi:10.3732/ajb.90.5.777.

Sanderson, M.J., 1997. A nonparametric approach of rate constancy to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14, 1218–1231.

Santos, E.P., 1995. Estudo das inflorescências no gênero *Salvia* L. subgênero *Calosphace* (Benth.) Benth. (Lamiaceae). Bradea 6, 372–380.

Sass, C., Iles, W.J.D., Barrett, C.F., Smith, S.Y., Specht, C.D., 2016. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. PeerJ 4, e1584. doi:10.7717/peerj.1584.

Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33, 1654–1668. doi:10.1093/molbev/msw079.

Schmickl, R., Liston, A., Zeisek, V., Oberlander, K., Weitemier, K., Straub, S.C.K., Cronn, R.C., Dreyer, L.L., 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). Mol. Ecol. Resour. 16, 1124–1135. doi:10.1111/1755-0998.12487.

Sousa, F. De, Bertrand, Y.J.K., Nylinder, S., Oxelman, B., Eriksson, J.S., Pfeil, B.E., 2014. Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. PLoS One 9, e109704. doi:10.1371/journal.pone.0109704.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Stamatakis, A., 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690. doi:10.1093/bioinformatics/btl446.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57, 758–71. doi:10.1080/10635150802429642.

Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., Malmberg, R.L., 2015a. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. Mol. Phylogenet. Evol. 85, 76–87. doi:10.1016/j.ympev.2015.01.015.

Stephens, J.D., Rogers, W.L., Mason, C.M., Donovan, L.A., Malmberg, R.L., 2015b. Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. Am. J. Bot. 102, 910–920. doi:10.3732/ajb.1500031.

Stout, C.C., Tan, M., Lemmon, A.R., Lemmon, E.M., Armbruster, J.W., 2016. Resolving Cypriniformes relationships using an anchored enrichment approach. BMC Evol. Biol. 16, 244. doi: 10.1186/s12862-016-0819-5.

Straub, S.C.K., Moore, M.J., Soltis, P.S., Soltis, D.E., Liston, A., Livshultz, T., 2014. Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. Mol. Phylogenet. Evol. 80, 169–185. doi:10.1016/j.ympev.2014.07.020.

Stull, G.W., Moore, M.J., Mandala, V.S., Douglas, N.A., Kates, H.-R., Qi, X., Brockington, S.F., Soltis, P.S., Soltis, D.E., Gitzendanner, M., 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. Appl. Plant Sci. 1, 1200497. doi:10.3732/apps.1200497.

Sun, K., Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2014.

The evolution of peafowl and other taxa with ocelli (eyespots): a phylogenomic approach. Proc. Biol. Sci. 281, 20140823. doi:10.1098/rspb.2014.0823.

Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222–231. doi:10.1080/10635150701311362.

Townsend, J.P., Leuenberger, C., 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol. 60, 358–365. doi:10.1093/sysbio/syq097.

Townsend, J.P., Lopez-Giráldez, F., 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst. Biol. 59, 446–457. doi:10.1093/sysbio/syq025.

Townsend, J.P., Su, Z., Tekle, Y.I., 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61, 835–849. doi:10.1093/sysbio/sys036.

Tucker, D.B., Colli, G.R., Giugliano, L.G., Blair Hedges, S., Hendry, C.R., Lemmon, E.M., Lemmon, A.R., Sites, J.W., Pyron, R.A., 2016. Methodological congruence in phylogenomic analyses with morphological support for teiid lizards (Sauria: Teiidae). Mol. Phylogenet. Evol. 75–84. doi:10.1016/j.ympev.2016.07.002.

Turner, E.H., Ng, S.B., Nickerson, D. A, Shendure, J., 2009. Methods for genomic partitioning. Annu. Rev. Genomics Hum. Genet. 10, 263–284. doi:10.1146/annurev-genom-082908-150112.

Walker, J., 2006. Systematics of the genus *Salvia* (Lamiaceae). PhD dissertation, University of Wisconsin-Madison. Wisconsin, 194 pp.

Walker, J.B., Sytsma, K.J., Treutlein, J., Wink, M., 2004. *Salvia* (Lamiaceae) is not monophyletic: implications for the systematics, radiation, and ecological specializations of *Salvia* and tribe Mentheae. Am. J. Bot. 91, 1115–1125.

Walker, J.B., Drew, B.T., Sytsma, K.J., 2015. Unravelling species relationships and diversification within the iconic California Floristic Province sages (*Salvia* subgenus *Audibertia*, Lamiaceae). Syst. Bot. 40, 826–844. doi:10.1600/036364415X689285.

Weitemier, K., Straub, S.C.K., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A., Liston, A., 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2, 1400042. doi:10.3732/apps.1400042.

Wen, J., Xiong, Z., Nie, Z.L., Mao, L., Zhu, Y., Kan, X.Z., Ickert-Bond, S.M., Gerrath, J., Zimmer, E.A., Fang, X.D., 2013. Transcriptome sequences resolve deep relationships of the grape Family. PLoS One 8, e74394. doi:10.1371/journal.pone.0074394.

Wenzel, J.W., Siddall, M.E., 1999. Noise. Cladistics 15, 51–64. doi:10.1111/j.1096-0031.1999.tb00394.x.

Wester, P., Claßen-Bockhoff, R., 2011. Pollination syndromes of New World *Salvia* species with special reference to bird pollination. Ann. Missouri Bot. Gard. 98, 101–155. doi:10.3417/2007035.

Xi, Z., Liu, L., Rest, J.S., Davis, C.C., 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. 63, 919–932. doi:10.1093/sysbio/syu055.

Young, A.D., Lemmon, A.R., Skevington, J.H., Mengual, X., Ståhls, G., Reemer, M., Jordaens, K., Kelso, S., Lemmon, E.M., Hauser, M., De Meyer, M., Misof, B., Wiegmann, B.M., 2016. Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). BMC Evol. Biol. 16, 143. doi:10.1186/s12862-016-0714-0.

Zhong, B., Deusch, O., Goremykin, V. V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S. V., Lockhart, P.J., 2011. Systematic error in seed plant phylogenomics. Genome Biol. Evol. 3, 1340–1348. doi:10.1093/gbe/evr105.

Table 1. Voucher specimens of the sampled species. Taxa marked with an asterisk (*) belong to the core *Calosphace* clade.

Species	Country, collector (Herbarium)
Lepechinia caulescens (Ortega) Epling	Mexico, *M. Castañeda-Zarate 841* (MEXU)
Salvia atrocyanea Epling*	Argentina, *A. Cocucci 5516* (CORD)
S. axillaris Moc. & Sessé *ex* Benth.	Mexico, *M. Castañeda-Zarate 1084* (MEXU)
S. connivens Epling*	Mexico*, F. Sazatornil 18* (MEXU)
S. fulgens Cav.*	Mexico*, F. Sazatornil 5* (MEXU)
S. helianthemifolia Benth.*	Mexico*, F. Sazatornil 9* (MEXU)
S. hispanica L.*	Cultivated, *I. Fragoso-Martínez 4* (FCME)
S. laevis Benth.	Mexico, *I. Fragoso-Martínez 284* (MEXU)
S. melissodora Lag.*	Mexico, *D. Sandoval 1264* (MEXU)
S. mexicana L.*	Mexico, *M. Castañeda-Zarate 840* (MEXU)
S. mocinoi Benth.*	Mexico, *G. Salazar-Chávez 9231* (MEXU)
S. protracta Benth.*	Mexico, *G. Salazar-Chávez 9463* (MEXU)
S. ramamoorthyana Espejo Serna *emend*. J.G.González*	Mexico, *G. Salazar-Chávez 9237* (MEXU)

Table 2. Characteristics of the concatenated matrices from each filtering scheme evaluated.

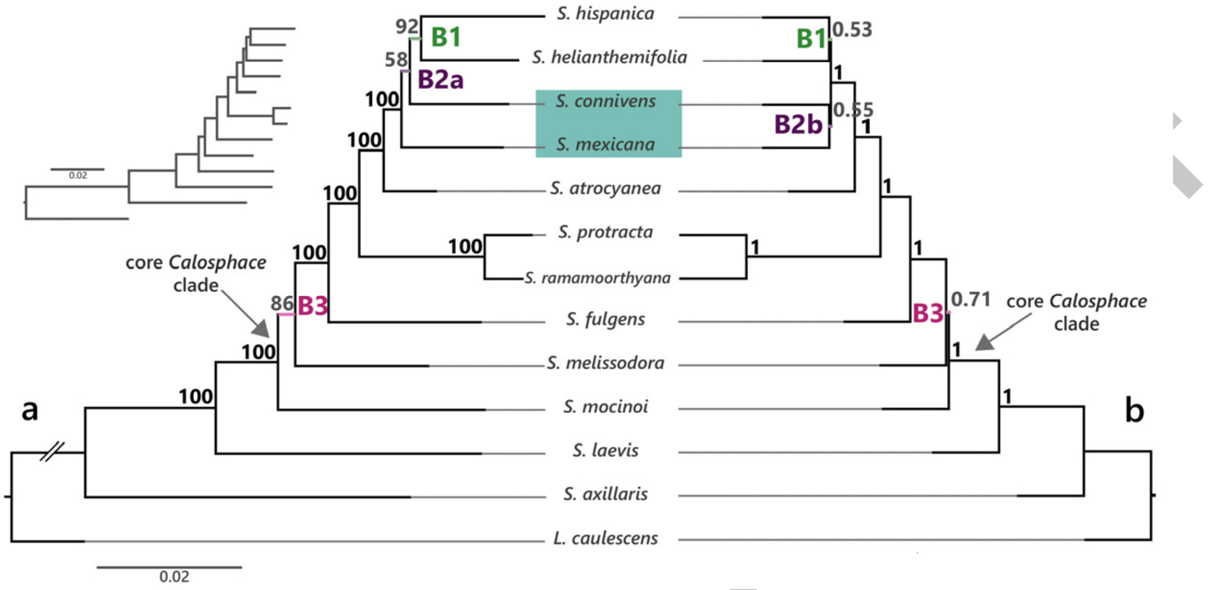
	Matrix	**Total number of characters (% removed from the original matrix)**	**Loci filtered (% from total)**
	Original	316,701 (0%)	0 (0%)
Less restrictive	**SR>20**	316,432 (0.08%)	160 (35.7%)
	SR> 15	316,332 (0.11%)	198 (44.1%)
	SR> 10	316,169 (0.16%)	251 (56%)
Strictest	**SR> 5**	315,626 (0.33%)	349 (77.9%)

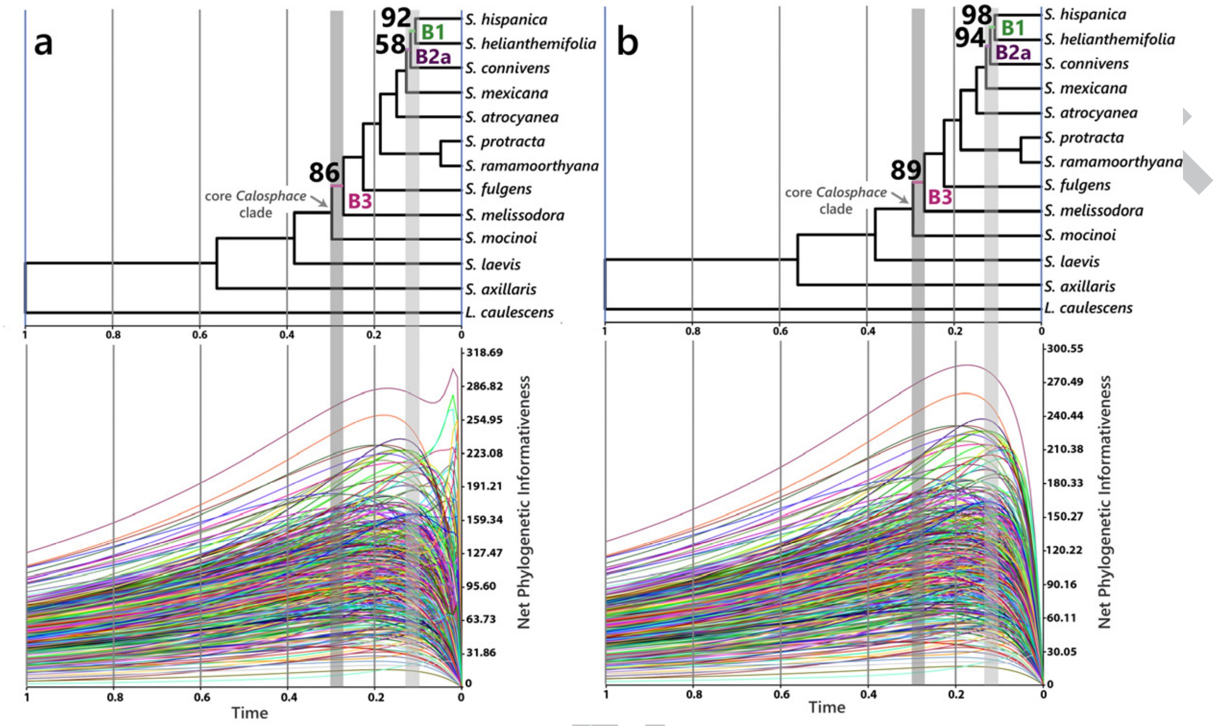
SR= substitution rate

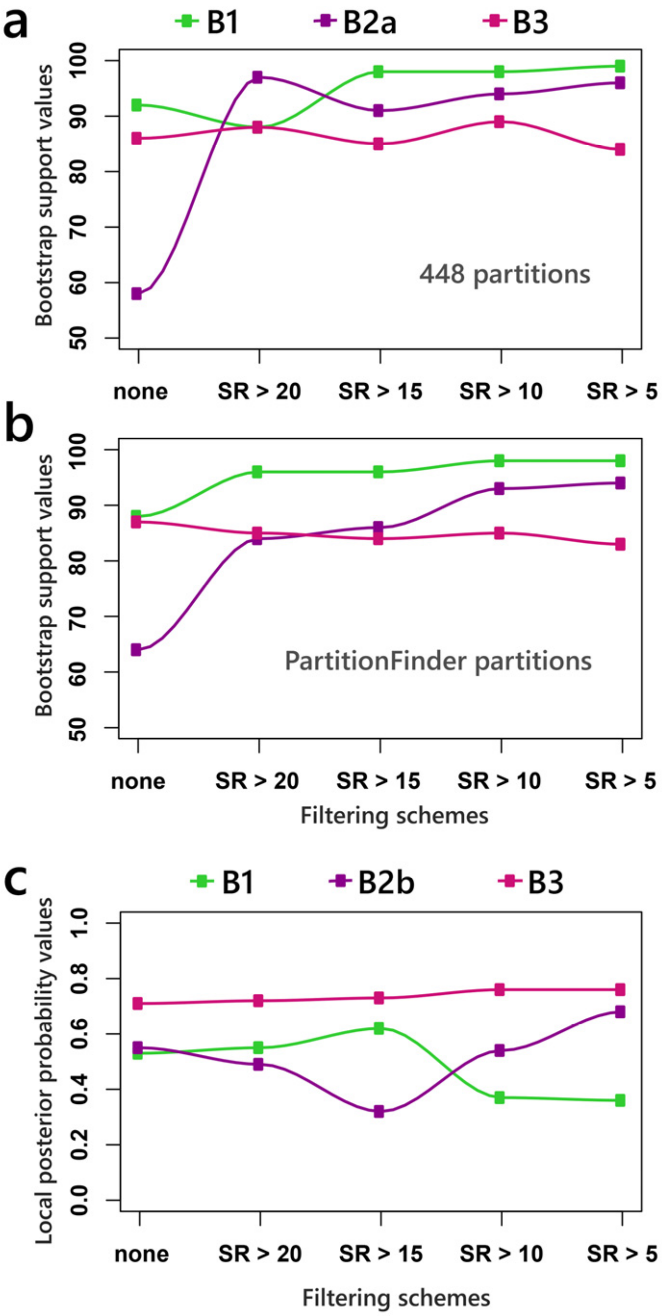
Fig. 1. Trees inferred from the AHE data: (a) maximum likelihood tree from the analysis of the concatenated matrix, (b) species tree from the coalescent-based method. Bootstrap support (BS) or local posterior probability (LPP) values are shown over the branches. Branches with low support are labeled as B1–B3, and only the branches B1 and B3 contain the same clades in both trees. The topological incongruence between the trees from different reconstruction methods is highlighted. In the concatenated analysis, only B2a lacked significant BS values, while in the coalescent-based analysis B1–B3 have low LPP values.

Fig. 2. Net phylogenetic informativeness of the AHE data through an arbitrary time scale: (a) profiles from the original, unfiltered dataset, (b) profiles from one filtered scenario where the sites with substitution rates >10 were removed. Branches that lacked 100% BS in the ML analysis are labeled as B1–B3. Most loci peak at/or deeper than B1 and B2a, and the slopes of most loci have decreased by B3, which is related to the decrease in PI for this branch. The curves from a number of loci of (a) show “phantom” spikes that suggest the presence of phylogenetic noise, whereas the curves from the loci of the filtered matrix (b) lack such spikes, which is related to the increased support values at B1–B3.

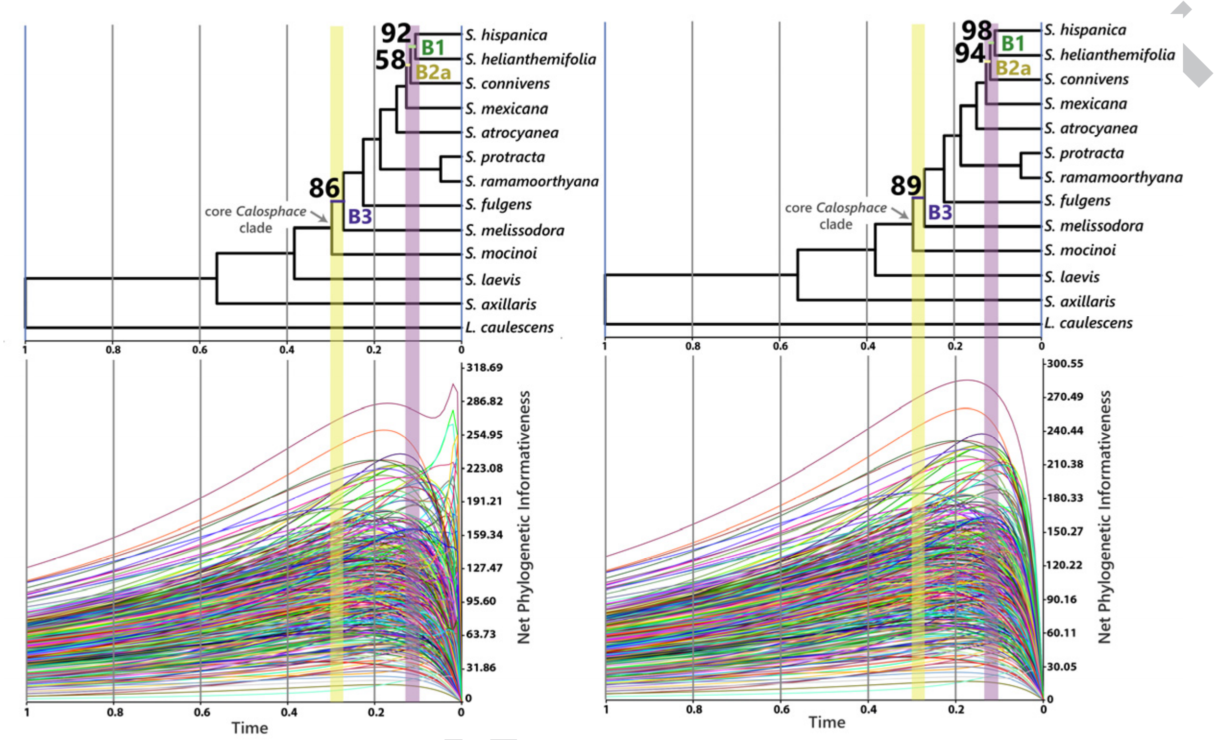
Fig. 3. Changes in support values in the different filtering schemes: bootstrap values on the analysis of the concatenated matrices (a) using 448 partitions, (b) using PartitionFinder schemes (7–17 partitions), (c) local posterior probability values on the coalescent-based analyses. Substitution rate removal scenarios depicted from less restrictive to the strictest: none = original dataset, sites removed with substitution rates higher than (SR>): 20, 15, 10 and 5. In general, for the concatenated analyses, in B1 and B2a the removal of sites with high substitution rates resulted in increased BS values, whereas for B3, removal of the sites increased slightly support values. In the coalescent-based analyses although LPP values increased in some scenarios, a significant value was never reached, and support values for B1 and B2b fluctuated between different scenarios, while the LPP values for B3 remained almost constant.


0.02
92
B1
58
B2a
100
100
100
100
100
86
B3
100
100
a
core *Calosphace* clade
S. hispanica
S. helianthemifolia
S. connivens
S. mexicana
S. atrocyanea
S. protracta
S. ramamoorthyana
S. fulgens
S. melissodora
S. mocinoi
S. laevis
S. axillaris
L. caulescens
0.02
B1
0.53
1
B2b
0.55
1
1
1
1
B3
0.71
1
1
core *Calosphace* clade
b


a
92
58
B1
B2a
86
B3
core Calosphace clade
S. hispanica
S. helianthemifolia
S. connivens
S. mexicana
S. atrocyanea
S. protracta
S. ramamoorthyana
S. fulgens
S. melissodora
S. mocinoi
S. laevis
S. axillaris
L. caulescens
1
0.8
0.6
0.4
0.2
0
318.69
286.82
254.95
223.08
191.21
159.34
127.47
95.60
63.73
31.86
0
Net Phylogenetic Informativeness
Time
b
98
94
B1
B2a
89
B3
core Calosphace clade
S. hispanica
S. helianthemifolia
S. connivens
S. mexicana
S. atrocyanea
S. protracta
S. ramamoorthyana
S. fulgens
S. melissodora
S. mocinoi
S. laevis
S. axillaris
L. caulescens
300.55
270.49
240.44
210.38
180.33
150.27
120.22
90.16
60.11
30.05
0
Net Phylogenetic Informativeness
Time


a
B1
B2a
B3
Bootstrap support values
50
60
70
80
90
100
448 partitions
none
SR > 20
SR > 15
SR > 10
SR > 5
b
Bootstrap support values
50
60
70
80
90
100
PartitionFinder partitions
none
SR > 20
SR > 15
SR > 10
SR > 5
Filtering schemes
c
B1
B2b
B3
Local posterior probability values
0.0
0.2
0.4
0.6
0.8
1.0
none
SR > 20
SR > 15
SR > 10
SR > 5
Filtering schemes

Graphical abstract



Highlights

- The Anchored Hybrid Enrichment method was tested in a New World lineage of sages.
- Capture and sequencing using the AHE method was successful, yielding 448 loci.
- AHE data improved support and resolution compared to conventional markers.
- Removal of sites with unusually high substitution rates resulted in increased support.