Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: Diplostephium)

Oscar M. Vargas1,2, Edgardo M. Ortiz1 and Beryl B. Simpson1
1Integrative Biology and Plant Resources Center, The University of Texas at Austin, Austin, TX 78712, USA; 2Department of Ecology and Evolutionary Biology, University of Michigan, 830 N. University Ave, Ann Arbor, MI 48109, USA

Key words: ddRAD, Diplostephium, genome skimming, hybridization, introgression, phylogenomics, rapid diversification, reticulate evolution.

Summary
• High-throughput sequencing is helping biologists to overcome the difficulties of inferring the phylogenies of recently diverged taxa. The present study analyzes the phylogenetic signal of genomic regions with different inheritance patterns using genome skimming and ddRAD-seq in a species-rich Andean genus (Diplostephium) and its allies.
• We analyzed the complete nuclear ribosomal cistron, the complete chloroplast genome, a partial mitochondrial genome, and a nuclear-ddRAD matrix separately with phylogenetic methods. We applied several approaches to understand the causes of incongruence among datasets, including simulations and the detection of introgression using the D-statistic (ABBA-BABA test).
• We found significant incongruence among the nuclear, chloroplast, and mitochondrial phylogenies. The strong signal of hybridization found by simulations and the D-statistic among genera and inside the main clades of Diplostephium indicate reticulate evolution as a main cause of phylogenetic incongruence.
• Our results add evidence for a major role of reticulate evolution in events of rapid diversification. Hybridization and introgression confound chloroplast and mitochondrial phylogenies in relation to the species tree as a result of the uniparental inheritance of these genomic regions. Practical implications regarding the prevalence of hybridization are discussed in relation to the phylogenetic method.

Introduction
Rapid diversifications usually occur in landscapes like archipelagos and mountain ranges that provide disjunct terranes suitable for speciation via isolation and ecological divergence (Givnish, 1997). The Andes Cordillera is one of these landscapes in which numerous plant groups have experienced high rates of speciation during and after its recent uplift (Madriñán et al., 2013; Luebert & Weigend, 2014; Hughes & Atchison, 2015). Phylogenies of Andean taxa based on Sanger sequencing often suffer from lack of resolution and support, especially in crown clades where accelerated speciation occurred (e.g. Emshwiller, 2002; Rauscher, 2002; Sánchez-Baracaldo, 2004; Bell & Donoghue, 2005; Hughes & Eastwood, 2006; Nürk et al., 2013; Zapata, 2013). Poorly resolved relationships within Andean taxa confirm the prediction that phylogenies resulting from rapid radiations are difficult to estimate because of low and conflicting signal caused by short internodes (Whitfield & Lockhart, 2007). High-throughput sequencing, a technology capable of producing orders of magnitude more data than obtained by Sanger sequencing, has created new opportunities for overcoming the difficulties of working with recently diversified taxa (Bock et al., 2014; Ma et al., 2014; Mort et al., 2015). Here, we implement genome skimming (Straub et al., 2012) and double-digest restriction site-associated DNA sequencing (ddRAD; Peterson et al., 2012) to infer the phylogenetic patterns of one of the most species-rich genera of Andean plants and its relatives.

Diplostephium is a main component of the tropical high Andean flora. The genus traditionally comprises 111 species (Vargas, 2011) characterized by a woody habit (from 10 cm decumbent subshrubs to 10 m tall trees) and radiate capitula with white to purple rays (Blake, 1928; Cuatrecasas, 1969; Vargas & Madriñán, 2006). Diplostephium inhabits the high elevations of the Talamanca Cordillera (Costa Rica), the Northern Andes (Venezuela–Colombia–Ecuador), and the Central Andes (Peru–Bolivia). Most species of the genus (c. 60) inhabit the páramo, a northern Andean ecosystem known for its high plant diversity (Luteyn, 1999), island-like geography (Simpson, 1974), and a large number of species-rich genera (Madriñán et al., 2013; Luebert & Weigend, 2014). In addition to the páramo, some Diplostephium species inhabit the Central Andean puna and the upper limit of the high Andean forest. Diplostephium belongs to
Astereae, where it has traditionally been classified as part of the *Chiliotrichum* group, a subset of the subtribe Hinterhuberinae (Nesom, 1994; Nesom & Robinson, 2007). Molecular phylogenies have shown that Astereae subtribes and their subdivisions are polyphyletic (Noyes & Rieseberg, 1999; Sancho & Karaman-Castro, 2008; Brouillet et al., 2009; Karaman-Castro & Urbatsch, 2009; Sancho et al., 2010; Vargas & Madriñán, 2012) and need to be recircumscribed. *Diplostephium* is positioned in the ‘South American lineages’ group (Brouillet et al., 2009), but its ambiguous position in the Astereae phylogeny (Noyes & Rieseberg, 1999; Brouillet et al., 2009; Karaman-Castro & Urbatsch, 2009; Vargas & Madriñán, 2012) has obscured its place among its closest genera. The low molecular variation found by Vargas & Madriñán (2012) in *Diplostephium* highlighted the problems of inferring the phylogeny of a rapidly diversified taxon with limited sampling using Sanger sequencing.

Genome skimming, or shallow shotgun sequencing, is an approach in which whole genomic DNA is sequenced in order to recover high-copy DNA regions suitable for phylogenetic analyses (Straub et al., 2012). The biparentally inherited (Volkov et al., 2007) complete nuclear ribosomal DNA has proved to be a useful marker for inferring species-level phylogenies (Linder et al., 2000; Straub et al., 2012; Bock et al., 2014). The mostly nonrecombinant and uniparentally inherited chloroplast DNA (Birky, 1995; Jansen & Ruhlman, 2012) has been used to reconstruct the phylogeny of Astereae at the tribal, generic, and species levels (Kim et al., 2005; Panero & Funk, 2008; Bock et al., 2014; Panero et al., 2014). Finally, the mostly nonrecombinant uniparentally inherited mitochondrial DNA (Birky, 1995) has been employed to infer the evolutionary history of angiosperms at the family and order levels (Qiu et al., 2010; Sun et al., 2015) and, more recently, to study phylogenetic patterns among species (Bock et al., 2014). Restriction site-associated DNA sequencing (RAD-Seq, e.g. ddRAD, GBS, etc.), on the other hand, is a technique that surveys hundreds or thousands of loci mainly from the nuclear genome (Peterson et al., 2012).

The comparison among markers with different inheritance patterns has the potential to elucidate hybridization and introgression events (Rieseberg & Solomon, 1991; Hardig et al., 2000; Bock et al., 2014; Sun et al., 2015) hypothesized to play an important role in cases of evolutionary bursts of speciation (Anderson & Stebbins, 1954; Seehausen, 2004). Because rapid divergence happens during a small window of time, it is expected that the descendants of a diversification burst are prone to interbreed before reproductive barriers develop. This expectation has been hypothesized as beneficial because hybridization among nascent lineages can result in genotypes potentially adapted to unexploited niches (Anderson & Stebbins, 1954; Seehausen, 2004). *Diplostephium* is thus an excellent model to test such ideas, as the genus appeared to have undergone a recent diversification associated with the uplift of the Andes (Vargas & Madriñán, 2012).

In this study, we focused on uncovering the phylogenetic patterns of *Diplostephium* and its allied genera. The aims of this paper are to compare the phylogenetic signal among nuclear-ddRAD, nuclear ribosomal, chloroplast, and mitochondrial DNA; and to test for infra- and intergeneric introgressive hybridization in *Diplostephium* and its Andean relatives.

**Materials and Methods**

Assembly of the genome skimming datasets

A total of 91 samples were sequenced. The ingroup contained 74 samples of *Diplostephium* (69 species, c. 62% of the total number of species) and 14 samples from 13 allied genera in Astereae (Supporting Information Table S1). We chose ingroup and outgroup genera based on their phylogenetic positions in Astereae inferred by Brouillet et al. (2009). The majority of the samples (63) were collected in the field where leaf tissue was dried using silica gel. The remaining samples were taken from herbarium specimens deposited in ANDES, F, FMB, HUA, HUSA, TEX, US and USM (Table S1).

We performed total genomic DNA extractions with the DNeasy Plant Mini Kit (Qiagen) following the manufacturer’s protocol. To increase the yield of DNA from herbarium material, we added 50 µl of proteinase K (Qiagen; activity > 600 mAU ml⁻¹) to the lysis solution and incubated it overnight at 45°C after the initial 10 min incubation. Standard genomic DNA Illumina paired-end libraries with an average of c. 400 bp length fragment size were prepared at the Genomic Sequencing and Analysis Facility (GSAF) at The University of Texas at Austin (UT) and then sequenced using an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA). The sequencing targeted 10 million paired-end reads (2 x 100 bp length) per sample. We inspected the quality of the reads with the program FASTQC v.0.10.1 (Andrews, 2010) and filtered the raw data with the program ‘process_shortreads’ of the software package STACKS v.1.20 (Catchen et al., 2011). The command discarded any reads with uncalled bases and trimmed terminal nucleotides that averaged a Phred score of ≤ 10 (90% of confidence in base calls) over a sliding window of 10 bp.

We employed a two-step strategy to create three subsets of reads per sample: nuclear ribosomal, chloroplast, and mitochondrial. By including only the necessary subsets of genomic data to perform the assemblies, we aimed to reduce noise and increase the accuracy and computational efficiency in our downstream analyses. Our first step consisted of performing a *de novo* assembly of each sample using all the genomic data. The software Ray v.2.3.1 (Boisvert et al., 2012) performed the assembly using three different k-mer values visually selected from the report graph provided by KMERGENIE (Chikhi & Medvedev, 2014). The contigs resulting from the Ray assembly were parsed with the ‘-search’ function of Ray using *Helianthus annuus* L. nuclear ribosomal, chloroplast, and mitochondrial reference sequences (GenBank accession HM638217.1, NC_007977.1, and KF815390.1, respectively); this step provided a subset of Ray contigs. The second step consisted in separating the genomic reads into nuclear ribosomal, chloroplast, and mitochondrial subsets by mapping all reads against the Ray contig produced after the first step. For example, whole-genome reads of *Diplostephium haenkei* were mapped against their own Ray chloroplast contigs to obtain a
chloroplast subset of reads. Bowtie v.2.2.3 (Langmead & Salzberg, 2012) was used to perform the mapping. Then, we analyzed with different pipelines the three subsets of reads obtained per sample to obtain the corresponding sequence alignments (Fig. S1).

Owing to the intraindividual polymorphic nature of the nuclear ribosomal DNA, we assembled these sequences in two steps. First, we created a de novo draft assembly with the nuclear ribosomal reads subset in each sample using SPAdes v.3.5.0 (Bankevich et al., 2012) with the ‘-careful’ option, and 21, 33, 55, 77 and 89 as k-mer sizes. The nuclear ribosomal contigs obtained by SPAdes were merged in Geneious v.7.1.4 (Kearse et al., 2012), allowing overlapping regions of two contigs to assemble with some mismatches by using the ‘medium sensitivity’ setting. Taking into account the fact that the nuclear ribosomal DNA is arranged in tandem repeats, the complete nuclear ribosomal contig produced by Geneious was treated as circular, aiming to increase the accuracy of the back-mapping. The first nucleotide position of the circular contig was set at the first base right (5’ to 3’ direction) of the TATAGGGGG promoter found at the end of the nontranscribed spacer (NTS) (Linder et al., 2000). Second, we back-mapped the nuclear ribosomal reads to their circular draft assembly in Geneious using a minimum overlap identity of 95%. From the back-mapping obtained for each sample, we calculated three consensus sequences per sample using different threshold percentages (50%, 75% and 90%; Fig. S1). By doing this, we were able to account for the different levels of intraindividual polymorphisms found in the nuclear ribosomal tandem repeats and evaluate their effects on phylogenetic estimation. The 50% consensus only called a nucleotide in a polymorphic position if the nucleotide was present in > 50% of the reads containing that position. Therefore, the 50% consensus sequence set had fewer ambiguities than the 75% and 90% consensus sequences. All the 273 nuclear ribosomal sequences (three per sample) were aligned with MAFFT version 7.017 (Katoh et al., 2002) into a master alignment. We corrected the master alignment by hand using Geneious. Finally, we extracted from the master alignment the three matrices corresponding to the 50%, 75% and 90% consensus sequences. Each matrix was analyzed separately.

A de novo assembly for each sample’s chloroplast reads subset was performed with SPAdes with the same options used for the nuclear ribosomal assembly. First, we annotated the chloroplast genome of D. haenkei using DOGMA (Wyman et al., 2004) with subsequent manual correction in Geneious using Guzotia abyssinica (L.F.) Cass. as a reference (GenBank accession EU549769.1). Then, we used the chloroplast of D. haenkei to merge and annotate the chloroplast-SPAdes contigs of the remaining 90 samples using Geneious (Fig. S1). We codified the gaps produced by nonoverlapping contigs as missing data. We employed MAFFT to create the chloroplast genome alignment that was later corrected by hand in Geneious. For the phylogenetic analysis, we only included one inverted repeat in our alignment and manually removed three unalignable regions of c. 2300 sites in total.

Owing to the high degree of rearrangements and the discontinuity found in the mitochondrial assemblies performed with SPAdes, it was not feasible to make a direct alignment of our mitochondrial genomes. Instead, we de novo assembled the mitochondrial genome of Diplastephium hartwegii and then used this genome as a reference for assembly by mapping the remaining 90 samples. We chose D. hartwegii based on the continuity and the high coverage of its mitochondrial contig obtained by Ray. We expected all the mitochondrial assemblies with the exception of D. hartwegii to have missing data at the boundaries of DNA blocks where rearrangements occurred relative to D. hartwegii. With this strategy, we intended to obtain a gapped mitochondrial genome assembly for each sample in which the order of the mitochondrial blocks matched that of D. hartwegii, making their alignment feasible. We employed SPAdes (with the same parameters used for the nuclear ribosomal assemblies) and MITOFY v.1.3.1 (Alverson et al., 2010) to assemble and annotate, respectively, the mitochondrial genome of D. hartwegii. We corrected by hand our MITOFY annotation using Geneious based on that of Cucumis sativus L. (Alverson et al., 2011). Before mapping, we filtered the mitochondrial reads of each sample because we detected a small proportion of chloroplast reads mixed in. The intraindividual presence of chloroplast reads in the mitochondrial subset is explained by the similarity of some tRNAs and the transference of DNA between these two genomes (Alverson et al., 2011). To filter out the chloroplast reads from the mitochondrial reads subsets, we mapped every sample’s mitochondrial reads subset against its de novo chloroplast genome assembled in this study (e.g. the mitochondrial reads of D. colombianum were mapped to the chloroplast genome of D. colombianum) using a 97% minimum overlap identity in Geneious. The reads not mapped to the chloroplast genome were mapped to the mitochondrial reference (D. hartwegii) to create a consensus sequence per sample (Fig. S1) using the ‘medium sensitivity’ setting in Geneious. MAFFT was used to perform the alignment of the 91 mitochondrial sequences. We visually inspected the matrix in Geneious, and regions difficult to align, or with significant amounts of missing data, were excluded from the matrix. We also removed the mitochondrial tRNA genes rrn5, rrnL, and rrsS from the alignment because we found bacterial DNA matching some hyperconserved regions of these genes. We suspect that the source of bacterial DNA in our dataset came from bacteria living on the leaf surfaces of our plant samples.

Assembly of the nuclear-ddRAD dataset

The RAD-Seq protocol known as ddRAD (Peterson et al., 2012) was used to prepare a reference library that included a subset of 44 samples (Table S1) to which we could map all the unused nuclear reads that resulted from genome skimming. The restriction enzymes EcoRI and SphI were used to digest total genomic DNA, with a size selection range for the resulting fragments (excluding adapters) of 354–414 bp. Library sequencing, carried out using an Illumina HiSeq 4000 at the UT-GSaf, produced c. 83 million of 2 × 150 bp paired-end reads. The reference library
was demultiplexed with the software deML v.1 (Renaud et al., 2014) which allowed us to confidently assign > 99% of the reads to their corresponding samples while tolerating up to two mismatches in the barcodes. The sequence quality assessed using FASTQC v.0.11.5 (Andrews, 2010) revealed low quality over the restriction overhang regions and a large portion of the R2 reads. Therefore, we excluded R2 reads from subsequent analyses and trimmed the restriction overhang of the R1 files, as well as 40 nucleotides from the end of the reads, to obtain sequences of 100 bp using the script ‘reformat.sh’ from the toolkit BBTOOLS v.36.38 (Bushnell, 2016). To guarantee that the assembly was done using only nuclear reads, we used the script ‘bsplit.sh’ from BBTOOLS to create sets of reads that did not match the ‘CLUSTS.GZ’ format produced by step 3 (within-sample clustering) of nrRAD v.3.0.66 (Eaton, 2014). From that point, SHOTGUN2RAD runs pyRAD steps 4–7 to produce loci alignments across species and a variety of matrix formats for phylogenetic analyses. The parameters used for SHOTGUN2RAD and pyRAD are indicated in Methods S2 and S3. Two datasets were derived from our ddRAD pipeline, a loci dataset (nuclear-ddRAD-loci) and a matrix containing only variable sites (nuclear-ddRAD). All computational processes related to the assembly of matrices were carried out at the Texas Advance Computing Center at UT (http://www.tacc.utexas.edu).

Phylogenetic analyses

We performed independent phylogenetic analyses for each genome skimming dataset using Bayesian inference (BI) and maximum likelihood (ML). We evaluated a nonpartitioned (M0) vs a partitioned model (M1). The M1 of the chloroplast and mitochondrial matrices contained a coding and a noncoding partition. The M1 of the nuclear ribosomal DNA dataset contained three partitions: rRNA, transcribed spacers (the external transcribed spacer (ETS), the ITS), and the nontranscribed spacer (NTS). We compared M0 and M1 for each dataset by calculating the stepping stone marginal likelihood (Xie et al., 2011) of both model schemes with MrBayes v.3.2.2 (Ronquist et al., 2012) on the CIPRES Science Gateway Server (Miller et al., 2010) using 10 million generations, two runs, four chains per run, and assuming a GTR+Γ model. The stepping stone marginal likelihood values were compared using Bayes factors (Fan et al., 2011). To calculate the model of evolution of the partitions, we employed a mixed strategy. First, we inferred the use +Γ and +I parameters in our partitions using the corrected Akaike information criterion (AICc) (Hurvich & Tsai, 1989) employed in jModelTest v.2.1.7 (Guindon & Gascuel, 2003; Darriba et al., 2012). Then, we calculated the substitution parameters among nucleotides with reversible jump Markov chain Monte Carlo (rjMCMC) simulations (Huelsenbeck et al., 2004) using MrBayes with 10 million generations, two runs, four chains per run, and the +Γ and/or +I parameters if suggested by the AICc. A final MrBayes analysis with 10 million generations, two runs, and four chains per run was performed using the best partition model along with the substitution parameters inferred for each dataset. We performed the ML analyses with RAxML v.8.1.11 (Stamatakis, 2014) in the CIPRES portal using the partitioned schemes favored, 100 rapid bootstrap replicates, and the GTR+Γ model of evolution (as recommended by RAxML manual Stamatakis, 2015). For all Bayesian analyses we used a burn-in fraction of 0.25. We employed TRACER v.1.6.0 (Rambaut et al., 2014) to confirm the convergence of parallel MCMC runs.

We analyzed the nuclear-ddRAD matrix with ML using the RAxML-HPC2 Workflow available in CIPRES (Miller et al., 2010). Two separate analyses were performed, one consisting of 100 independent runs and one consisting of 100 thorough bootstraps, both run with the Ascertainment Bias option enabled. The 100 independent runs were summarized into a majority consensus rule tree with ‘SUMTREES.PY’ v.4.10 (Sukumaran & Holder, 2010a) available in the package DENDROPHY v.4.10 (Sukumaran & Holder, 2010b). Support was obtained from the 100 RAxML thorough bootstraps using ‘SUMTREES.PY’. The same matrix was analyzed with SVDQUARTETS (Chifman & Kubatko, 2014) implemented in PAUP* v.4.0a150 (Swofford, 2002) using all quartets and performing 100 bootstraps. We excluded outgroups from both analyses because of their high amounts of missing data. FigTree v.1.4.2 (Rambaut, 2014) allowed us to inspect, compare, and export trees to image editors.

Convergence assessment and visualization

We visually inspected the congruence of the topologies obtained by different matrices with the multidimensional scaling of tree space using the Robinson–Foulds distance implemented in TreeSetViz v.3.0 (Amenta & Klingner, 2002) as a package for Mesquite v.3.04 (Maddison & Maddison, 2015). The first comparison was made among the results obtained by the BI and ML analysis of the three nuclear ribosomal datasets (nr50, nr75, and nr90). The second comparison contrasted the trees from the
analyses of the nuclear ribosomal 90%, chloroplast, mitochondrial, and nuclear-ddRAD datasets. From the BI analyses, we sampled the maximum clade credibility tree (MCCT from sumtrees.py) and 50 random Bayesian topologies from each dataset after a 0.25 burn-in. From the ML and SVDq analyses, we sampled the best tree obtained and 50 random bootstrap replicates from each subset. A hierarchical likelihood-ratio congruence test among the nuclear ribosomal 90%, chloroplast, and mitochondrial datasets was performed with CONCATENATOR v.1.8a (Leigh et al., 2008) coupled with RAxML v.7.2.8 (Stamatakis, 2006). We identified rogue taxa by using the Procrustean Approach to Cophylogeny (PACo; Balbuena et al., 2013) with the pipeline designed by Pérez-Escobar et al. (2016), specifically aimed at identifying incongruent taxa between nuclear and chloroplast phylogenies. The approach of Pérez-Escobar et al. (2016) assumes that the chloroplast phylogeny is dependent on the nuclear phylogeny and transforms the topologies into matrices of patristic distances, which are then transformed into Euclidean principal coordinate matrices using the method proposed by de Vienne et al. (2011). taxa with significant incongruent phylogenetic positions are identified by comparing the observed distances with those produced by permutations. Because this approach performs best with phylogenies (Pérez-Escobar et al., 2016), we used the BI trees obtained from the nuclear ribosomal 90% and chloroplast matrices that have low amounts of missing data and therefore a more accurate branch estimation.

Hybridization assessment

To distinguish between incomplete lineage sorting (ILS) and hybridization, we employed the method described by Joly et al. (2009) implemented in the software JML v.1.3.0 (Joly, 2012). This method uses the posterior distribution of coalescent species trees, estimated population sizes, and branch length from a *BEAST v.2.3.0 (Heled & Drummond, 2010) output to simulate a coalescent scenario with no migration. Hybridization is identified by comparing the minimum pairwise distance between the simulated and the empirical datasets. If an observed distance is significantly smaller than the simulated one, then ILS can be rejected, suggesting hybridization (Joly et al., 2009; Joly, 2012). Joly et al.'s (2009) approach assumes that the marker used to calculate genetic distances does not recombine and its power increases with longer sequences (Joly, 2012); this makes our chloroplast matrix an excellent candidate to simulate and compare genetic distances. We calculated a coalescent tree in *BEAST using the nuclear ribosomal and chloroplast matrices. We ran five independent analyses in *BEAST using 200 million generations, sampling every 1000, and with a GTR+Γ model. For the Jml input, we randomly sampled 2000 trees from the five *BEAST runs using LogCombiner v.2.3 (Drummond et al., 2012) after a burn-in of 0.25. With Jml, we carried out 2000 simulations of chloroplast genome pairwise distance comparisons. To visualize the results, we created a matrix map in R (R Core Team, 2016) indicating significant pairwise comparisons after the Benjamini–Hochberg correction.

To detect introgression, we calculated the Patterson’s D-statistic (ABBA-BABA test; Huson et al., 2005; Green et al., 2010; Durand et al., 2011) on selected quartets of taxa with the nuclear-ddRAD-loci dataset employing the software pyRAD. We chose the quartets based on the Jml results and the comparison between nuclear-ddRAD and chloroplast topologies. The D-statistic calculates the proportion ‘ABBA’ and ‘BABA’ patterns in a four-taxon phylogeny. This proportion should be equally frequent in a scenario with ILS and no gene flow; if the proportion between the two patterns is significantly different, then introgression between two of the taxa is inferred. Significant patterns were identified using a Z-score > 3, which correspond to a conservative \( \alpha = 0.01 \) (Eaton & Ree, 2013) after the Benjamini–Hochberg correction.

Time calibration of the phylogeny

We could not implement a primary calibration for our dataset because of the poor fossil record for Astereae. Therefore, in order to obtain a calibration point for our ingroup, we extracted the ITS region (c. 700 bp) of our nuclear ribosomal alignment and combined it with that of Strijk et al. (2012), which contains sequences from Heliantheae, Gnaphalieae, and Anthemideae providing external nodes for the time calibration. We aligned the ITS matrix with MAFFT and employed MrBayes to obtain a topology using 30 million generations, two runs, four chains, a temperature of 0.001, and GTR+Γ+I model. We calculated the chronogram with BEAST v.2.3.0 (Drummond et al., 2012) using the MrBayes topology (fixed), a lognormal relaxed clock, a Yule model of speciation, 100 million generations, a sampling frequency of 40 thousand generations, and the same calibration priors used by Strijk et al. (2012). We produced the control file using BEAUI v.2.3 (Drummond et al., 2012). The sets of trees from the two runs were combined after a burn-in of 0.25 using LogCombiner v.2.3 (Drummond et al., 2012) before calculating the chronogram with TREEANNOTATOR 2.3 (Drummond et al., 2012).

Because the ITS matrix (c. 700 bp) did not provide enough phylogenetic information to produce a robust and resolved topology in the part of the tree corresponding to our ingroup, we calibrated our Bayesian ddRAD tree (fixed) using the complete nuclear ribosomal 90% matrix (c. 13 kb), employing secondary calibration points derived from our ITS chronogram and the literature. We aligned to our matrix the nuclear ribosomal citron of Helianthus annuus (GenBank accession KF767534) using MAFFT and employed the following calibration points on the tree: a mean of 32.4 million yr (Myr) based on Kim et al. (2005) with a normal distribution and a sigma of 11 (95% CI: 14.4–50.6) to the root of the tree representing the crown clade of Asteroideae; and a mean of 11.21 Myr with a normal distribution and a sigma of 3.1 (95% CI: 6.11–16.3) to the most recent common ancestor of the ‘South American Lineages’ (inferred in our ITS chronogram). To calculate the chronogram, we employed BEAST with the same parameters used in the calibration of the ITS dataset and the model and partition scheme calculated for the phylogenetic analysis of the nuclear ribosomal matrix.
Consensus sequences were deposited in GenBank (Table S1). Subsets of reads, aligned matrices, and control files are available at Dryad (https://doi.org/10.5061/dryad.cn74h).

Results

Characteristics of the datasets

The nuclear ribosomal alignments comprise 13 362 characters, of which 1203–1425 (9.00–10.66%) are parsimony-informative characters (PICs) depending on the consensus threshold (the number of PICs reported here always exclude the outgroup; Table S2). While the nr50 matrix contains only 25 ambiguities, the nr75 and the nr90 have 3201 and 5012 ambiguities, respectively. The chloroplast genome of *D. haenkei* contains 85 genes plus 36 tRNAs and eight rRNAs in 152 292 bp (Fig. S2). No significant rearrangements or gene losses were found across the sampled species relative to *Guizotia abyssinica* with the exception of the loss of the *rps19* gene for *Baccharis genistelloides* and *trnT-GGU* for *B. genistelloides*, *B. tricuneata*, and *Llerasia caucana*. Fifty-six of the 91 genomes have missing data in their assemblies as a result of the low coverage of reads obtained in regions with a high number of repeats. The chloroplast alignment is 135 440 characters long, of which 2169 (1.69%) are PICs (Table S2). The *de novo* mitochondrial genome of *D. hartwegii* has a total length of 277 718 bp, containing 56 genes plus three chloroplast-like tRNAs (Fig. S3). The final mitochondrial matrix has a length of 209 392 and contains 1730 (0.83%) PICs. The nuclear-ddRAD-loci dataset used for introgression analyses has a total of 19 987 loci. The nuclear-ddRAD matrix containing only variable SNPs, employed for the phylogenetic analyses, has a total of 244 255 total sites with 96 421 (39.5%) PICs (Table S2).

Phylogenetic analysis

In all cases, the Bayes factors favor a partitioned matrix analysis over a nonpartitioned one (Tables S3, S4). Overall, all the phylogenetic topologies obtained are well resolved with highly supported nodes (Figs 1, S4–S11). The two topologies obtained by BI or SVDq and ML for each of the matrices analyzed (e.g. the BI-chloroplast tree compared with the ML-chloroplast tree) are consistent excluding clades or taxa with low support (bootstrap support (BS) < 80%, Bayesian posterior probability < 0.9).

Nuclear ribosomal topologies

The BI and ML backbones of the nr50, nr75 and nr90 trees are all highly congruent, the exception being the position of *Diplostephi um meyenii*, which has low support on all of the trees (Figs 1b, S5–S9). The visualization of the Robinson–Foulds tree distances of the six nuclear ribosomal tree subsets (BI-nr50, BI-nr75, BI-nr90, ML-nr50, ML-nr75 and ML-nr90) shows four clusters of topologies (Fig. S12). None of the four clouds is formed exclusively by the trees of one subset. This visualization emphasizes the fact that despite the general agreement among the nuclear ribosomal topologies, there are some incongruences located towards the tip of the trees (Figs 1b, S5–S9). For example, *Diplostephi um schultzei*, a species represented by two samples, is polyphyletic in the BI-nr50 and ML-nr50 topologies (Figs S5, S7), whereas it is monophyletic in the BI-nr75, BI-nr90, ML-nr75 and ML-nr90 topologies (Figs 1b, S6, S8, S9). Because the nr50, nr75 and nr90 backbones are congruent and the nr90 matrix is the most conservative in a phylogenetic context, capturing a considerable amount of intraindividual polymorphisms (informative to MrBayes and RAxML; Ronquist et al., 2011; Stamatakis, 2015), we selected the nr90 dataset to make comparisons with the chloroplast and mitochondrial datasets.

For the remaining part of the paper we will therefore refer to the nuclear ribosomal 90% dataset simply as ‘nuclear ribosomal.’

Incongruence among the genomic datasets

To avoid confusion in our results and discussion, we only use the BI MCCT topologies from the genome skimming matrices and the ML topology from the nuclear-ddRAD matrix when comparing all dataset topologies, as alternative topologies from the same matrices are almost identical.

Incongruence is significant among the topologies obtained from the nuclear-ddRAD, nuclear ribosomal, chloroplast, and mitochondrial genomic regions with the four trees recovering different numbers of clades of *Diplostephi um* taxa and contrasting generic relationships (Fig. 1). The topology in Fig. 1(a) shows two labeled groups which we refer to as *Diplostephi um* A and B. While in the nuclear-ddRAD and nuclear ribosomal topologies species of group A comprise a monophyletic group that is sister to a clade containing non-*Diplostephi um* genera and species of *Diplostephi um* group B, in the chloroplast and mitochondrial topologies, species of group A (except for *D. inesianum* in the mitochondrial topology) are nested within a clade containing *Diplostephi um* group B species and the rest of Astereae genera sampled. Topological incongruence among our datasets is also obvious in the visualization of the Robinson–Foulds tree distances (Fig. S13), depicting four independent clouds of topologies corresponding to the nuclear-ddRAD, nuclear ribosomal, chloroplast, and mitochondrial tree subsets. The chloroplast and mitochondrial clouds are positioned more closely together relative to the nuclear-ddRAD and nuclear ribosomal trees. The hierarchical likelihood ratio test rejects the concatenation of the most congruent genome skimming dataset combination, chloroplast plus mitochondrial DNA (*P*<0.0001). Normalized squared residuals values obtained by PACo identify 35 (38%) of the 91 tips as rogue taxa (Fig. 2).

Hybridization assessment

A strong signal of hybridization in our data is revealed by JmI (Fig. 2); 1647 (40%) out of 4095 pairwise chloroplast distances are significantly smaller than expected in a scenario with ILS and no migration (*P*<0.05) after the Benjamini–Hochberg correction. The Patterson’s *D*-statistic (ABBA-BABA) performed in selected quartets of taxa shows a signal of introgression among genera and inside *Diplostephi um* groups A and B for many of the quartets evaluated (Tables 1, S5).
Chronograms

The ITS chronogram shows poor support and resolution in nodes positioned in the ‘South American Lineages’ (Fig. S14); nevertheless, it provides a time estimate for the origin of the ‘South American lineages’ (node R, mean = 11.2 Myr, 95% CI: 6.2–16.4) employed to calibrate the nuclear-ddRAD topology with the complete nuclear ribosomal matrix. The time calibration of our nuclear topology using the nuclear ribosomal matrix (Fig. 3) provides a rough estimation of the age of Diplostephium.

Fig. 1 Phylogenies obtained from the different datasets: (a) nuclear double-digest restriction site-associated DNA sequencing (ddRAD) – maximum likelihood; (b) nuclear ribosomal 90% – Bayesian inference (BI) maximum clade credibility tree (MCCT); (c) chloroplast – BI-MCCT; and (d) mitochondrial – BI-MCCT. Dots indicate highly supported nodes (bootstrap support > 90, Bayesian posterior probability > 0.9). Shaded areas indicate the distribution of Diplostephium taxa according to the map on the right (e).
groups A (mean = 6.5 Myr, 95% CI: 1.3–11.4) and B (mean = 7.3 Myr, 95% CI: 2.3–12.5), and the divergence of South American genera (mean = 10.17 Myr, 95% CI: 3.2–16.2).

Discussion

Phylogenetic incongruence and hybridization

Incongruence among the nuclear-ddRAD, nuclear ribosomal, chloroplast, and mitochondrial datasets is found at the generic and species levels, especially between the nuclear vs organellar trees. The only clade recovered consistently across phylogenies (excluding D. inesianum for the mitochondrial topology) is a clade that contains a group of Diplostephium species restricted mostly to the Northern Andes, labeled group A in the nuclear-ddRAD topology (Figs 1, 3). However, despite this Northern Andean clade being recovered in the four genomic phylogenies, its position relative to other genera and the rest of species of Diplostephium is incongruent when nuclear and organellar topologies are compared. Incongruence among different markers can be produced by phylogenetic uncertainty, incomplete lineage sorting, and/or hybridization (Rieseberg & Soltis, 1991; Maddison, 1997; Huelsenbeck et al., 2000). To rule out phylogenetic uncertainty, we performed a PACo analysis to identify rogue taxa. Because PACo performs simulations incorporating phylogenetic uncertainty, the presence of significant outliers provides evidence for the phylogenies being affected by ILS.
or hybridization. The result that 38% of our samples are rogue indicates extensive ILS or hybridization in our dataset.

To distinguish between ILS and hybridization, we employed the method proposed by Joly et al. (2009), which is incorporated into the software JML (Joly, 2012). The significantly small pairwise distances of c. 40% of all possible comparisons among chloroplast genomes suggest high degrees of hybridization. Although JML results can be difficult to interpret because they only show species pairs affected by hybridization and do not incorporate a model to detect ancient hybridization, our visualization of the JML results mapped next to a phylogeny (Fig. 2) provides a new approach to formulate hybridization hypotheses. This visualization shows a strong signal of ancient hybridization because significant comparisons are present among different genera and among almost all comparisons between species of Diplostephium groups A and B.

We employed a nuclear-ddRAD-loci dataset to calculate empirically the amount of introgression among selected taxa using the Patterson’s D-statistic. Our results (Table 1) show that introgression has occurred among genera and inside Diplostephium groups A and B. Even though we did not test for every possible hybridization event in our D-statistic framework, the significant tests reported (Table 1) suggest at least 12 reticulate events (different tests may represent the same event). In the specific case of Diplostephium sp. nov. CAJ2, D. meyenii, and Parastrephia quadrangularis, the D-statistic identifies a strong signal of hybridization, also supported by morphological data. We hypothesize that Diplostephium sp. nov. CAJ2 is the product of a cross between D. meyenii and P. quadrangularis because Diplostephium sp. nov. CAJ2 has scale-like leaves similar to P. quadrangularis but heterogamous capitula with ray flowers like D. meyenii and group B species (Fig. 4a–c). Similarly, a high Z-score suggests that D. cinereum (Fig. 4d) has hybridized with D. meyenii or P. quadrangularis (Table 1), all of which occur in the Peruvian Andes in parapatry, and morphological hybrids have been observed in the field (V. Quipuscoa, pers. comm.).

The incongruence among genomic phylogenies, the JML results, and the signal of introgression inferred by the D-statistic suggests a complex pattern of reticulate evolution in our ingroup.

| Table 1 Patterson’s D-statistic showing the tests for which introgression between P3 and either P1 or P2 was detected using a Z-score threshold of 3 (see Supporting Information Table S1 for full species details) |
|---|---|---|---|---|---|---|
| P1 | P2 | P3 | Outgroup | Z | BABA | ABBA | nloci |
| D. ochraceum | D. violaceum | B. bartsifolia | D. haenkei | 4.99 | 8.5 | 35.25 | 198 |
| B. genistelloides | L. sophiophilica | D. ericoides | D. ochraceum | 4.8 | 30.75 | 84 | 159 |
| B. genistelloides | B. tricuneata | D. espinosae | D. frontinense | 5.06 | 77.25 | 169.5 | 291 |
| B. genistelloides | L. sophiophilica | D. espinosae | D. frontinense | 4.35 | 50.38 | 106.88 | 204 |
| B. genistelloides | B. tricuneata | D. ochraceum | D. haenkei | 4.58 | 184 | 89.5 | 303 |
| B. genistelloides | B. tricuneata | D. violaceum | D. haenkei | 3.55 | 172.5 | 96.5 | 309 |

To distinguish between ILS and hybridization, we employed the method proposed by Joly et al. (2009), which is incorporated into the software JML (Joly, 2012). The significantly small pairwise distances of c. 40% of all possible comparisons among chloroplast genomes suggest high degrees of hybridization. Although JML results can be difficult to interpret because they only show species pairs affected by hybridization and do not incorporate a model to detect ancient hybridization, our visualization of the JML results mapped next to a phylogeny (Fig. 2) provides a new approach to formulate hybridization hypotheses. This visualization shows a strong signal of ancient hybridization because significant comparisons are present among different genera and among almost all comparisons between species of Diplostephium groups A and B.

We employed a nuclear-ddRAD-loci dataset to calculate empirically the amount of introgression among selected taxa using the Patterson’s D-statistic. Our results (Table 1) show that introgression has occurred among genera and inside Diplostephium groups A and B. Even though we did not test for every possible hybridization event in our D-statistic framework, the significant tests reported (Table 1) suggest at least 12 reticulate events (different tests may represent the same event). In the specific case of Diplostephium sp. nov. CAJ2, D. meyenii, and Parastrephia quadrangularis, the D-statistic identifies a strong signal of hybridization, also supported by morphological data. We hypothesize that Diplostephium sp. nov. CAJ2 is the product of a cross between D. meyenii and P. quadrangularis because Diplostephium sp. nov. CAJ2 has scale-like leaves similar to P. quadrangularis but heterogamous capitula with ray flowers like D. meyenii and group B species (Fig. 4a–c). Similarly, a high Z-score suggests that D. cinereum (Fig. 4d) has hybridized with D. meyenii or P. quadrangularis (Table 1), all of which occur in the Peruvian Andes in parapatry, and morphological hybrids have been observed in the field (V. Quipuscoa, pers. comm.). The incongruence among genomic phylogenies, the JML results, and the signal of introgression inferred by the D-statistic suggests a complex pattern of reticulate evolution in our ingroup.

The high number of instances of hybridization seen in our dataset explains to a great extent the incongruence obtained among our genomic phylogenies. Numerous authors have proposed that horizontal gene transfer via introgression could cause deviation of the chloroplast and mitochondrial phylogenies from the species tree (Rieseberg & Solis, 1991; Moore, 1995; Hardig et al., 2000; Sun et al., 2015). Because organellar DNA inheritance is mostly uniparental and there is generally no recombination following fertilization (Birky, 1995; Jansen & Ruhlman, 2012), an event of hybridization could completely replace the original chloroplast and mitochondrial DNA of a lineage with an alien one, confounding their phylogenies relative to the species tree (Rieseberg & Solis, 1991). Chloroplast and mitochondrial genomes have the potential to be fixed rapidly because their effective population size (Ne) is one-quarter that of a nuclear autosomal gene (Moore, 1995), making organellar genomes less prone to ILS. Additionally, the discrepancies among the phylogenies of chloroplast and mitochondrial DNA suggest a decoupling of inheritance of these two genomes. The evidence presented here
leads us to conclude that the trees obtained from the chloroplast and mitochondrial DNA unequivocally deviate from the species tree.

The nuclear-ddRAD tree is, to our knowledge, the best species tree hypothesis, taking into account the numerous loci sampled from the nuclear genome and its phylogenetic robustness.
This topology fits better the morphological classification and biogeography of our ingroup than all the other phylogenies obtained in this study. The nuclear-ddRAD topology indicates that *Diplostephium* (sensu Cuatrecasas, 1969) is biphyletic (if *Parastrephia* is included in group B species) with both clades diverging approximately in the late Miocene. Because the nomenclatural type of the genus, *D. ericoides*, is a member of group B species (*Diplostephium* s.s.), group A must be circumscribed as a different taxon (O. M. Vargas, unpublished).

From the genome skimming dataset, the nuclear ribosomal DNA is better at capturing signal to infer the species tree than the chloroplast and mitochondrial DNA, because the nuclear ribosomal cistron is biparentally inherited (Volkov *et al.*, 2007), it recombines (Hughes & Petersen, 2001; Ambrose & Crease, 2011), and its copies are located on multiple chromosomes (Phillips *et al.*, 1971; Álvarez & Wendel, 2003). The nuclear ribosomal tree is also largely congruent with the nuclear-ddRAD tree (Fig. 1a,b).

Extensive introgression in South American Astereae is probably a result of lack of reproductive barriers. Even though the pollination biology of our ingroup is still understudied, *Diplostephium* and its allies lack obvious pollinator specialization (Cuatrecasas, 1969) probably because of the low diversity of pollinators in the high Andes (Berry & Calvo, 1989). Additionally, the complex topography of the Andes might promote speciation via geographic isolation while allowing gene flow in the early stages of differentiation before reproductive barriers develop.

Available reports of chromosome numbers (Chromosome Counts Database, http://ccdb.tau.ac.il; Fig. 2) suggest that the ancient hybridization identified in our data did not result in ploidy in the majority of cases because the only sample with a chromosome count report deviating from the base chromosome number in Astereae (*n* = 9; Nesom & Robinson, 2007) is *Diplostephium jaramilloi* (*n* = 18). However, more data about karyotypes are needed to confirm these statements, especially for *Diplostephium* group A species.

**Practical considerations**

Our results demonstrate that hybridization is prevalent in *Diplostephium* and its allies deviating the chloroplast and mitochondrial phylogenies from the species tree. In conjunction with documented cases of reticulate evolution in other organisms like heliconine butterflies (Mallet *et al.*, 2007) and cichlid fishes (Genner & Turner, 2012), our results support the hypothesis that hybridization might be a common process in events of rapid evolution (Anderson & Stebbins, 1954; Seehausen, 2004). Because hybridization might be frequent during evolutionary bursts of speciation, ancient and recent diversifications (and radiations) are especially prone to the phylogenetic biases explained here. We urge systematists to avoid concatenating data with different inheritance patterns (e.g. chloroplast and nuclear markers)
without testing for congruence, and to apply caution when using chloroplast and mitochondrial phylogenies as the basis for taxonomic classification, time calibrations, historical biogeographic reconstructions, and comparative phylogenetic analyses.

Phylogenetic methods have focused on a bifurcating model of evolution, but dichotomous trees cannot illustrate processes like hybridization and horizontal gene transfer (Doolittle, 1999; Huson & Bryant, 2006; Rieppel, 2010). Recent advances in phylogenetic explicit network estimation (Solé-Lemus & Ané, 2016; Wen et al., 2016) promise better models to understand the evolutionary history of taxa that have undergone hybridization and lateral gene transfer, yet these approaches are still in their infancy and are only sufficient in cases of recent hybridization with few reticulate events.

Approaches that identify rogue terminals (e.g. PACo and PARAfit; Legendre et al., 2002; Balbuena et al., 2013; Pérez-Escobar et al., 2016) are often used to remove taxa from phylogenetic analysis and decrease incongruence among datasets (Philippe et al., 2009; Regier et al., 2013; Salichos & Rokas, 2013; Reginato & Michelangeli, 2016). While PACo identified 38% of our taxa as rogue, our hybridization analysis suggests that c. 94% of our sampled taxa are affected by ancient or recent hybridization. Therefore, removing taxa identified as rogue from our analysis would not eliminate topological incongruence produced by hybridization. We encourage biologists to uncover the reason for incongruence (instead of removing rogue taxa) as the answer to this question could elucidate noteworthy biological processes.

Our study demonstrates that ancient hybridization is worthy of attention and should be modeled in phylogenetic inference. Future research should focus on understanding the role of hybridization in diversification and on developing methods to quantify ancient introgression and accurately infer reticulate networks. The evidence presented in our study builds on numerous reports (Sessa et al., 2012; Sochor et al., 2015; Sun et al., 2015; Wu et al., 2015) that emphasize a central role of reticulation in plant evolution.

Acknowledgements
We thank Fernando Alzate, Felipe Cardona, Michael Dillon, Alvaro Idarraga, Santiago Madriñan, Haydeé Montoya, Victor Quipuscoa, Adriana Prieto, Rusty Russell, Katya Romolerox, Luis Sánchez, and Tom Wendt for providing access to their herbaria. Hamilton Beltrán provided the pictures of Diplostephium cinnereum and Diplostephum sp. nov. CAJ 2. Jordon Bemmel, Joseph Brown, David Cannatella, Amalia Diaz, Robert Jansen, Thomas Juenger, Teofil Nakov, Juan Palacio, Sarah Sussman, Ning Wang, and Mao-Lun Weng provided help and comments during the development of this study. We also want to thank the three anonymous reviewers for their constructive comments. This project was funded by the University of Texas at Austin (Plant Biology Program Award, the C. L. Lundell Chair of Systematic Botany, and the Linda Escobar Award), the Garden Club of America (2012 Award in Tropical Botany) and the National Science Foundation (postdoctoral support FESD 1338694 and DEB 1240869). Research permits: resolution 0145 of ‘El Ministerio de Ambiente y Desarrollo Sostenible’ and permit 37 of ‘Sistema Ambiental Sina’ (Colombia); authorization 10-2012-IC-FLO-DPAP-MA of ‘Ministerio del Medio Ambiente’ (Ecuador); and resolution 0369-2011-AG-DGFFS-DGEFFS of ‘Dirección de Gestión Forestal y de Fauna Silvestre’ (Peru).

Author contributions
O.M.V. and E.M.O. performed the research, collected the samples, and analyzed the data. B.B.S. and O.M.V. designed the research and interpreted the results. The writing was primarily carried out by O.M.V.

References


**Supporting Information**

Additional Supporting Information may be found online in the Supporting Information tab for this article:

**Fig. S1** Diagram representing the sequence assembly pipeline for the genome skimming dataset.

**Fig. S2** Circular representation of the chloroplast genome of Diplostephium haenkei.

**Fig. S3** Circular representation of the mitochondrial genome of Diplostephium hartwegii.

**Fig. S4** Nuclear ddRAD tree obtained with SVD quartets.

**Fig. S5** Nuclear ribosomal maximum clade credibility tree obtained by Bayesian inference with the 50% consensus matrix.

**Fig. S6** Nuclear ribosomal maximum clade credibility tree obtained by Bayesian inference with the 75% consensus matrix.

**Fig. S7** Nuclear ribosomal best tree obtained by maximum likelihood with the 50% consensus matrix.

**Fig. S8** Nuclear ribosomal best tree obtained by maximum likelihood with the 75% consensus matrix.

**Fig. S9** Nuclear ribosomal best tree obtained by maximum likelihood with the 90% consensus matrix.

**Fig. S10** Chloroplast best tree obtained by maximum likelihood.

**Fig. S11** Mitochondrial best tree obtained by maximum likelihood.

**Fig. S12** Visualization of Robinson–Foulds tree distances among the nuclear ribosomal 50%, 75% and 90% consensus datasets.

**Fig. S13** Visualization of Robinson–Foulds tree distances among the ddRAD, nuclear ribosomal 90%, chloroplast, and mitochondrial (red) datasets.

**Fig. S14**. Chronogram of the ITS matrix.

**Table S1** List of specimens with their voucher and GenBank information

**Table S2** Descriptive statistics of the matrices obtained

**Table S3** Bayesian factor comparison between the two partition models calculated with MrBAYES in each genome skimming dataset

**Table S4** Models of evolution inferred by jMODELTEST and MrBAYES for the genome skimming datasets

**Table S5** Complete list of Patterson’s D-statistic calculations

**Methods S1** Parameters file used for assembly of the reference ddRAD library in IPYRAD.

**Methods S2** Parameters file used for assembly of the nuclear-ddRAD dataset in SHOTGUN2RAD.

**Methods S3** Parameters file used for assembly of the nuclear-ddRAD dataset in PYRAD.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.