Discordance in a South African Memecylon Clade (Melastomataceae): Evidence for Reticulate Evolution

Prabha Amarasinghe,1,∗+ Phuc Pham,‡ Robert Douglas Stone,§ and Nico Cellinese∗

∗Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA, and Biodiversity Institute, University of Florida, Gainesville, Florida 32611, USA; †Department of Biology, University of Florida, Gainesville, Florida 32611, USA, and Cooperative Agricultural Research Center, Prairie View A&M University, Prairie View, Texas 77446, USA; §Herbert Wertheim College of Engineering, University of Florida, Gainesville, Florida 32611, USA; and §School of Life Sciences, University of KwaZulu-Natal, South Africa

Editor: Jill S. Miller

Premise of research. Evergreen forests in eastern South Africa have high biodiversity but are limited in extent and have a highly fragmented distribution. Populations of forest plants are thus geographically isolated, and fine-scale evolutionary studies of these lineages might yield important insights into the history and assembly of the forests themselves. A prior study showed that, despite their morphological diversity, three South African Memecylon taxa in Melastomataceae (Memecylon natalense, M. bachmannii, and M. australissimum) had almost identical nuclear ribosomal spacer sequences. Our study investigates phylogenetic relationships within this clade using multiple samples collected across populations and a next-generation phylogenomic approach.

Methodology. We used 87 low-copy nuclear (LCN) loci to examine relationships among these taxa using both concatenated and coalescent methods. We further used LCN loci to estimate phylogenetic networks and single-nucleotide polymorphisms (SNPs) derived from LCN genes for STRUCTURE analysis of South African Memecylon populations. Finally, we employed two approaches (flow cytometry and SNPs) to infer the ploidy levels of these three taxa.

Pivotal results. Our investigations showed discordance among gene trees and the species tree and low statistical support for relationships, indicating that species monophyly cannot be recovered from this phylogenomic analysis. Phylogenetic networks and population structures showed that the South African Memecylon clade may be affected by gene flow and reticulate evolution. Flow cytometry and SNP-based estimations provided evidence for polyploidy within this group.

Conclusions. We found no evidence of monophyly for species within the South African Memecylon clade, which we infer to be the consequence of reticulation and recent and rapid evolution. More cytological studies and genomic data are needed to elucidate the evolutionary history of this group. Additionally, our study identifies priority populations for conservation within the South African Memecylon clade.

Keywords: flow cytometry, gene discordance, hybridization, Memecylon, phylogenomics, polyploidy.

Online enhancements: supplemental tables and figures. Dryad data: https://dx.doi.org/10.5061/dryad.8cz8w9gqw.

Introduction

An increasing number of investigations using genomic data reveal the complex patterns of relationships among and within populations. Reticulation events such as incomplete lineage sorting (ILS), introgression, hybridization, and polyploidy commonly complicate the reconstruction of phylogenies at the population level (Crowl et al. 2017; Folk et al. 2017; Dunning and Christin 2020; Stubbs et al. 2020). Here, we investigate the fine-scale evolution of a complex plant group consisting of South African members of the Buxifolia clade of Memecylon (Melastomataceae).

Extant members of the Buxifolia clade in southern Africa are evergreen shrubs to small trees generally found in the understory and subcanopy of densely forested blocks (Stone et al. 2017a, 2017b, 2019). The naturally fragmented distribution of extant forests in eastern South Africa (Eastern Cape, KwaZulu-Natal, Mpumalanga, and Limpopo Provinces) fits an island-like biogeographical model, with limited dispersal possibly favoring...
high genetic diversity among populations (Stone et al. 2017a; but see Hughes et al. 2005). South African members of the Buxifolia clade are sporadically distributed in forests of limited extent separated by wider expanses of grassland and savanna vegetation (fig. 1; Stone et al. 2017a, 2019). This pattern of distribution can result in limited opportunities to pollinate and disperse across plant populations (Hughes et al. 2005). Forest types inhabited by the Buxifolia clade include lowland, Afromontane, mist belt, sandstone, riverine (e.g., streamside and rocky riverbank forests, forests within deep river gorges), coastal, and coastal scarp areas (Stone et al. 2017a, 2019). Additionally, forests in South Africa have been greatly affected by paleoclimatic changes resulting in disjunctions and population size fluctuations among flora and fauna (Lawes 1990; Eeley et al. 1999; Hughes et al. 2005; Lawes et al. 2007). Given this environmental setting, studies of plant groups like Memecylon might yield important insights into the processes shaping the diversity of the woody flora in these areas and the evolutionary dynamics in the South African forest archipelago.

The Buxifolia clade, which currently includes 17 recognized species, has an overall distribution from East Africa (Kenya and Tanzania) to the eastern part of South Africa and a disjunct distribution in dry forests of western and northern Madagascar (Stone 2014; Stone et al. 2017a, 2019). In this study, we focused on three closely related South African taxa: Memecylon natalense, M. bachmannii, and M. australissimum (Stone et al. 2017a, 2019). Memecylon natalense and M. bachmannii have partly overlapping geographic ranges but are morphologically distinct with respect to each other (Stone et al. 2017a). The recently described M. australissimum combines certain morphological character states of M. natalense with others of M. bachmannii (Stone et al. 2019). For example, M. australissimum has relatively small and elliptic-ovate leaves similar to those of M. natalense, whereas the persistent floral bracteoles and ovoid fruits are similar to those of M. bachmannii (Stone et al. 2019). A previous molecular phylogenetic analysis of the Buxifolia clade was based on two nuclear ribosomal DNA (rDNA) spacer sequences, the internal transcribed spacer (ITS) and 5′ external transcribed spacer (ETS) of the 18S-26S rDNA cistron (Stone et al. 2017a). This rDNA-based study showed that ITS1, ITS2, and ETS sequences of closely related South African Memecylon taxa are identical or nearly so (Stone et al. 2017a). For example, variable sites of ITS1, ITS2,
and ETS regions of *M. natalense* and *M. bachmannii* were 1%, 2%, and 1%, respectively. The other members of *Memecylon* used for that study showed comparatively variable ITS and ETS regions (Stone et al. 2017a). In the resulting phylogeny of Stone et al. (2017a), *M. soutpansbergense* was strongly supported as sister to an internally unresolved clade consisting of *M. natalense*, *M. bachmannii*, and *M. australissimum* (the latter cited as *M. aff. bachmannii* in that work). Another recently described species, *M. kosiense* (Stone et al. 2019), was found to be more distantly related to the clade mentioned above and closer to *M. incisilobum* (of southern Mozambique) and *M. nubi genum* (of northern Mozambique and southern Malawi; Stone et al. 2017a).

Given that the previous ITS and ETS analyses of Stone et al. (2017a) were not able to resolve the phylogenetic relationships of *M. natalense*, *M. bachmannii*, and *M. australissimum*, further molecular studies are needed to determine their species boundaries. The purpose of this study is to unravel the evolutionary relationships within the South African *Memecylon* species complex through population-level sampling to understand whether *M. natalense*, *M. bachmannii*, and *M. australissimum* represent three distinct taxa or, alternatively, whether they represent a single morphologically variable clade.

As phylogenetic reconstructions can be obfuscated by hybridization and ILS, multiple low-copy nuclear (LCN) loci and coalescent-based methods are useful to elucidate evolutionary processes (Maddison and Knowles 2006). Additionally, phylogenetic reconstructions can also be affected by polyploidy and other processes (Linder and Rieseberg 2004; Vriesendorp and Bakker 2005). Although the available chromosome counts for other *Memecylon* taxa are limited (e.g., 2n = 14 for *M. liberiae* Gilx ex Engler, 2n = 28 for *M. capitellatum* L., and 2n = 24 for *M. caeruleum* Jack) and ploidal levels have not been recorded (Rice et al. 2015), it is important to verify whether results from the phylogenetic analysis of South African *Memecylon* are affected by polyploidy. Both cytological and base frequency distribution analyses are useful in determining polyploidy (Servick et al. 2015; Corrêa dos Santos et al. 2017; Weiß et al. 2018; Viruel et al. 2019). Also, to test whether this group consists of distinct taxa, concatenated and coalescent phylogenomic approaches can be useful. Therefore, to understand evolutionary processes within this group, here we employ phylogenomic reconstruction and network and concordance analytical approaches for South African *Memecylon* populations using 87 LCN genes from next-generation sequencing combined with cytological techniques.

### Material and Methods

#### Taxon Sampling

Silica-dried tissue samples of *Memecylon* were collected from representative populations spanning the geographic ranges of *Memecylon natalense* (seven populations), *M. bachmannii* (five populations), and *M. australissimum* (two populations) in eastern South Africa (KwaZulu-Natal, Eastern Cape, and Mpumalanga Provinces). Accessions and their collection details are provided in table A1 (tables A1–A4 are available online). Voucher specimens were deposited in either the University of KwaZulu-Natal Bews Herbarium (NU) or the Buffelskloof Nature Reserve Herbarium (BNRH). Additional samples of outgroups of *M. soutpansbergense* (one population) and *M. kosiense* (one population) were collected from Limpopo and KwaZulu-Natal Provinces. Further, plants and their habitats were photographed, and GPS and ecological data were recorded in the field. Two to five individuals per population were selected for the phylogenomic analysis, and a total of 37 samples from 11 populations representing major geographic areas were included (table A2).

#### Library Preparation, Target Capture, Sequence Cleaning, and Assembly

Target capture was employed to enrich genomic regions of interest for the population-level samples. We designed a probe set for phylogenomic analysis using a two-tiered approach for capturing LCN genes of two distantly related clades of Melastomataceae (Jantzen et al. 2020) to understand fine-scale evolutionary processes within South African *Memecylon*. Total genomic DNA was extracted following a modified CTAB extraction protocol for Melastomataceae (Jantzen et al. 2020) and quantified using a Qubit BR assay, and quality was evaluated using both NanoDrop 2000 (ThermoFisher Scientific, Waltham, MA) and gel electrophoresis. Preparation and quantification of double-barcoded genomic libraries and target enrichment were performed by Rapid Genomics (http://www.rapid-genomics.com) using the customized probe set (Jantzen et al. 2020). All accessions were processed as described in Jantzen et al. (2020). Cleaned reads were assembled with HybPiper version 1.3.1 using the template sequences used for probe design as references (Johnson et al. 2016). Potential paralogs identified by HybPiper scripts were removed (Jantzen et al. 2020). Cleaned reads have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (PRJNA576018), and NCBI accessions are listed in table A1. The alignments and raw phylogenetic trees are available from the Dryad Digital Repository (https://dx.doi.org/10.5061/dryad.8cz8w9gqw; Amarasinghe et al. 2021b).

#### Phylogenetic Analyses

The LCN sequences were individually aligned using MAFFT version 7.215 (Kato and Standley 2013) with the default gap opening penalty. The alignments were then trimmed using trimAl version 1.2 (Capella-Gutiérrez et al. 2009) with a gap threshold of 0.9. The gene alignments were reviewed using Geneious version 10.2 (Kearse et al. 2012) and evaluated for sequence length, missing data, and quality. Gene trees were generated for each of the individual gene alignments using maximum likelihood (ML) in RAxML version 8 (Stamatakis 2014), with 500 rapid bootstrap pseudoreplicates combined with an ML tree search using the GTRGAMMA model. The gene trees were manually reviewed, and uninformative genes were removed from the analysis. The uninformative genes included those having no variation, as well as those resulting in poor alignments or unresolved trees. The remaining genes were concatenated, and a partition scheme was generated using Phyx version 8.2.0 (Brown et al. 2017). ML analyses of the concatenated data set were repeated for supercontigs, exons, and introns using the RAxML parameters as described above, and the trees were compared. In total, 87 individual introns were informative at the population level and resulted in phylogenies with higher bootstrap support (BS) compared...
with the individual supercontig and exon data sets. Therefore, we used the 87-intron data set to infer the concatenated phylogenetic tree and conduct downstream analyses. In the resulting phylogenetic trees, we considered strong bootstrap support to be BS ≥ 90, moderate support to be 90 > BS ≥ 60, and poor support to be BS < 60.

To infer a coalescent species tree, 87 ML gene trees (derived from introns) generated from RAxML were input into ASTRAL-III version 5.0.3 (Zhang et al. 2018). These 87 gene trees were obtained from the most informative introns for the samples used in this study. In the resulting species tree, we considered strong posterior probability (PP) support to be PP ≥ 0.95, moderate support to be 0.95 > PP ≥ 0.75, and poor support to be PP < 0.75. We present PP values here because it has been shown that the PP values of clade support have more precision than multilocus bootstrapping (Sayyari and Mirarab 2016).

Topology comparisons between RAxML and ASTRAL-III trees were performed using the cophylo function of the R package phytools (Revell 2012). All gene trees were rooted using Phyx (Brown et al. 2017) and used in PhyParts, a program that calculates the proportion of unique, conflicting, and concordant nodes between gene trees and a reference tree (Smith et al. 2015). Here, PhyParts assesses gene tree congruence with the ASTRAL-III species tree.

Quartet Sampling and Concordance Analysis

We implemented quartet sampling (QS) to understand the nature of low BS support in the phylogeny (Pease et al. 2018). This Python-based program uses a quartet-based evaluation system (a statistical approach reflecting the number of alternative topologies that a resampled quartet tree recovers) that differentiates between strong conflict and weak support in large sequence data sets. The QS scripts were used for the ML topology to obtain the quartet concordance (QC), quartet differential (QD), and quartet informativeness (QI) scores at each node. The QC, QD, and QI values were used to understand the informativeness of the genes and the nature of the discordance in the phylogeny.

Topological concordance among loci was summarized using IQ-TREE (Nguyen et al. 2015). Here, we used the gene concordance factor (gCF) and the site concordance factor (sCF) to quantify genealogical concordance in the phylogenomic data sets (Nguyen et al. 2015; Minh et al. 2018). For every branch of a reference tree, gCF is the percentage of gene trees containing that branch, and sCF is the percentage of alignment sites supporting a branch in the reference tree. For input into IQ-TREE, we used the ASTRAL-III species trees and the RAxML gene trees, and 500 quartets were randomly sampled per branch.

Ploidy Estimation

A modified flow cytometry protocol (hand-chopped version; Roberts et al. 2009) was used to estimate the genome size and infer ploidy for two to five individuals per population of the South African Memecylon clade. Silica gel-dried samples were run with *Raphanus sativus* as an internal standard (Bennett and Leitch 2005) with the flow cytometry propidium iodide indicator on a BD Accuri C6 flow cytometer at the Interdisciplinary Center for Biotechnology Research, University of Florida. The quality of each data file was assessed using a variety of parameters (e.g., forward scatter-height, fluorescence signal 2-area, time, and cell counts), and data plots (e.g., histograms, dot plots, density plots, etc.) were generated with the FCS Express version 7 software (De Novo Software, Los Angeles, CA). Genome sizes were estimated using the histogram statistics of both the internal standard and Memecylon samples. Ploidy levels were estimated by comparing the flow cytometry results for Memecylon samples with mitotic chromosome count data (i.e., for *M. caeruleum* [2n = 24]; Rice et al. 2015). Here, both the genome estimation from flow cytometry and known chromosome count of *M. caeruleum* were used to infer the ploidy level of our samples.

Additionally, using cleaned sequenced data, read mapping and deduplication were performed with bwa, samtools, and bcftools (Li et al. 2009; Li 2011, 2013) and Picard version 2.21.2 (Broad Institute 2019), respectively. Then single-nucleotide polymorphisms (SNPs) obtained from the above approaches were input into plodidyNGS (Corrêa dos Santos et al. 2017), which derives ploidy information from base frequency distributions at variable sites by mapping each genome’s reads to its assembly.

Network Analysis

Phylogenetic networks for the South African Memecylon clade were estimated using species networks applying quartets (SNQ; Solís-Lemus and Ané 2016) implemented in the software PhyloNetworks’ Julia package (https://github.com/crs14/PhyloNetworks.jl). This program calculates the maximum pseudolikelihood of a network of four-taxon concordance factors. For the starting tree, a species tree estimated with ASTRAL-III was used. SNQ was run using a range of possible hybrid nodes from zero to three, allowing one (maximum allowed number of hybridizations [hmax] = 1), two (hmax = 2), and three (hmax = 3) hybridization events. For each number of hybrid nodes, 10 independent SNQ searches were run, and the result with the highest pseudolikelihood value was retained. Reticulations were visually inspected to assess whether the increase in hybrid nodes represented an additional reticulation.

Admixture Analysis

To explore further the population structure within this South African Memecylon clade, SNPs from each individual sample from the sequence reads were identified with the alleles workflow as used by Kates et al. (2018; scripts retrieved from https://github.com/mossmatters/phyloscripts/tree/master/alleles_workflow). Here, reads were mapped to previously assembled supercontigs using bwa and samtools (Li et al. 2009; Li 2011, 2013). Reads were deduplicated using Picard, and variant sites were called using GATK version 3.8.1 HaplotypCaller (McKenna et al. 2010; Poplin et al. 2018; Broad Institute 2019). This process was applied for reads of each sample separately, and the resulting individual vcf files were combined using bcftools (Li 2011). The combined vcf file, which contains SNPs, was converted to an input file for admixture analysis using PGDSpider version 2.1.1.5 (Lischer and Excoffier 2012). Genetic similarity among individuals was evaluated using principal component analysis (PCA) implemented in the R version 3.6.1 (R Core Team 2019) packages SNPRelate (Zheng et al. 2012) and vegan (Oksanen et al. 2017) for SNP data as described in Hodel et al. (2021). In a
preliminary PCA, outgroups were shown to have significantly diverged from the ingroup samples, and they were removed from the admixture analysis. To determine the optimal value of population number (K) and population structure among 35 accessions, the SNP genotyping information was analyzed in STRUCTURE version 2.3.4 (Pritchard et al. 2000). STRUCTURE uses a Bayesian algorithm to assign individuals to one or more populations/clusters (K). The likelihood that an individual belongs to a particular cluster is given by a Q value. Higher Q values indicate a greater PP that an individual belongs to that cluster. Runs were executed with 50,000 burn-in steps followed by 100,000 Markov chain Monte Carlo iterations, and 10 replicate runs were performed for K = 1 through K = 20. The parameters were set to allow for admixture between clusters and used the correlated allele frequency model. Ten replicates were used for each value of K, and the most likely number of clusters was estimated based on the highest value of ΔK where ln(ΔK) plateaued using STRUCTURE HARVESTER (Earl and vonHoldt 2012). Ten replicates of the best K were combined and visualized in CLUMPP (Kopelman et al. 2015) and visualized using both STRUCTURE PLOT version 2.0 (Ramasamy et al. 2014) and the R packages ggplot2 and gridExtra (Auguie 2015; Wickham 2016).

Results

Phylogenetic Analyses

Statistics regarding the alignments of the supercontigs, introns, and exons, including alignment lengths, variable sites, constant sites, phylogenetically informative sites, and proportions of missing data, are presented in table A.3. These statistics showed that the introns have the highest percentage of phylogenetically informative sites compared with exons and supercontigs. Further, introns contained the highest percentage of variable sites compared with exons and supercontigs. The concatenated supercontig, exon, and intron data sets recovered different relationships, and only intron sequence data were able to provide some resolution for this species complex. Probe names, numbers of samples per sequence, total sites, unique sites, phylogenetically informative sites, invariant sites, and constant sites for individual introns are provided in table A.4. Of the total LCN loci obtained for Memecylon from the two-tier approach (Jantzen et al. 2020), 89 were found to be uninformative for this group. Removal of paralogs and uninformative genes resulted in 87 phylogenetically informative introns. The length of the trimmed and concatenated 87-intron alignment of 37 samples was 42,087 bp. These introns were used in the phylogenomic analysis of South African Memecylon.

The concatenated ML species tree obtained from the phylogenomic analysis of intron regions showed six major clades with low support (fig. A1; figs. A1–A4 are available online). The ML analysis of the concatenated intron data set resolved most nodes as poorly supported (41 of 58). In general, the support values that we recovered were poor at both deeper and shallower nodes.

The concatenated ML analysis of the phylogenomic data set and the coalescence analysis using ASTRAL-III showed highly incongruent relationships. The ASTRAL-III phylogeny conformed to a geographical pattern (fig. 2). However, the local PP values for most of the clades were poor. Memecylon bachmannii clustered together in one clade (PP = 0.43; fig. 2B), but both M. natalense and M. australisium were not monophyletic in the ASTRAL-III topology (fig. 2). Moderate support was found for M. natalense from Barberton, Mpumalanga (PP = 0.78), Stainbank, Durban, KwaZulu-Natal (PP = 0.78), and Everton, KwaZulu-Natal (PP = 0.87). For some populations, samples collected from the same geographical localities were scattered within the phylogeny. For example, five M. natalense accessions from the Ongoye population included in this analysis do not form a monophyletic group.

When the RAxML and ASTRAL-III phylogenies were compared using a tanglegram (fig. 3), we found that 23 relationships are different in the RAxML and ASTRAL-III trees. Moreover, individual LCN gene trees showed high levels of phylogenetic discordance. When poorly supported nodes were collapsed in both RAxML and ASTRAL-III phylogenies, disagreement still remained about the placement of samples of M. natalense from Ongoye (samples O1 and O8) and Stainbank (samples S2 and S7). The PhyParts analysis (fig. A2) suggested conflict among gene trees for many relationships within the major clades of Memecylon at both deeper and shallower nodes. For example, 18 nodes showed that none of the gene trees are in agreement at those nodes. However, two M. natalense populations from Barberton (B2 and B4) and Stainbank (S2 and S7) and one M. bachmannii population from Everton (E3 and E5) showed more concordant gene trees in the PhyParts analysis than the rest of the nodes.

Quartet Sampling and Concordance Analysis

The QC, QD, and QI scores from the analysis are presented in figure A3. The QC value quantifies the relative support among the three possible resolutions of four taxa in a resampled quartet tree (Pease et al. 2018). In this phylogeny, the QC values were less than −0.5 for most of the nodes, and this indicates counter-support for alternative topologies. The QD score measures the disparity between the sampled proportions of two discordant topologies in a resampled quartet tree (Pease et al. 2018). In this phylogeny, the frequency of all possible topologies was similar because almost all QD values were closer to one. The QI score quantifies, for a given branch, the proportion of replicates where the best-likelihood quartet tree has a likelihood value that exceeds the quartet tree with the second-best likelihood value by a given differential cutoff (Pease et al. 2018). In this study, the proportion of informative quartets (QI) was <30% for the nodes. Overall, this phylogeny showed that only a weak majority of quartets support the focal branch.

The concordance analysis using IQ-TREE supported the highest amount of discordance seen in the QS analysis (figs. A3, A4). For many of the branches within the Memecylon phylogeny, the two concordance factors, gCF and sCF, are low.

Network Analysis

SNaQ analyses recovered a single hybridization event (fig. 2). The best network (log likelihood = 6.66) included a single hybrid edge between M. natalense (L1, from Lusikisiki) and M. bachmannii (LP2, from Luputana), indicating these as the parental lineages. These analyses estimated a λ value of 0.465
for *M. natalense* (Lusikisiki) and 0.535 for *M. bachmannii* (Luputana). The $\lambda$ value represents the proportion of the genome estimated to have been contributed by each parental lineage. No improvement in the likelihood was observed (log likelihood $= 6.66$) when we ran the analysis using $h_{\text{max}} = 2$. Our STRUCTURE analysis complemented the SNaQ results, indicating that *M. australissimum* may have originated through hybridization between *M. natalense* (Lusikisiki) and *M. bachmannii* (Luputana). An extra reticulation was found with $h_{\text{max}} = 1$ between two *M. natalense* populations (B2 from Barberton, Mpumalanga, and O4 from Ongoye, KwaZulu-Natal). The estimated $\lambda$ values are 0.369 for B2 and 0.651 for O4 (log likelihood $= 37.82$), without any improvement when rerun with the second hybridization event.

**Ploidy Estimation**

Flow cytometry signals were poor for most of the *Memecylon* samples, and this method was successful only in 18 samples (table 1). The genome sizes based on geometric statistics showed that average genome sizes (2C) of *M. natalense* L1, *M. bachmannii* LP2, and *M. australissimum* He2 were 2.96, 3.67, and 6.85 pg, respectively. Ploidal levels of *M. natalense*, *M. bachmannii*, and *M. australissimum*, obtained by comparing them with a known ploidal estimate of *M. caeruleum* (2n = 24), were approximately 4n, 6n, and 10n, respectively. Ploidal levels estimated using ploidyNGS were fairly concordant with flow cytometry for some samples (table 1). PloidyNGS also showed that some samples within the South African *Buxifolia* clade are hexaploid and heptaploid. Ploidal levels of *M. natalense*, *M. bachmannii*, and *M. australissimum*, estimated using ploidyNGS were 7n, 6n, and 7n, respectively. A comparison of these values with ploidal level estimation using flow cytometry is provided in table 1.

**Admixture Analysis**

In the STRUCTURE analysis, $\Delta K$ as a function of $K$ reached a peak at $K = 7$. When all recovered values of $K$ were examined, gene flow among a widespread set of individuals was identified. The STRUCTURE plot of optimal $K = 7$ (fig. 4) reflected the results from the genetic PCA to some extent (fig. 5). Samples from clusters 5 and 6 in the STRUCTURE plot (fig. 4) were congruent with the clusters in the genetic PCA (fig. 5). Samples in cluster 4 in the STRUCTURE plot (fig. 4) showed two samples near zero on PC1 and two samples with extreme negative values on PC1 and PC2 in the genetic PCA (fig. 5). However, clusters 1, 2, 3, and 7 in the STRUCTURE plot (fig. 4) were not clearly separated from each other in the PC space (fig. 5).

The following samples showed a high degree of admixture: *M. natalense* BR1, S2, S7, MO2, and MO3 and *M. bachmannii* C3, U5, G3, OM10, and OM12. In general, *M. australissimum* (Eastern Cape) showed contributions of genes from parental populations of *M. bachmannii* and *M. natalense*. The clusters...
Fig. 3  Tanglegram comparing the nuclear tree topologies inferred with RAxML (A; bootstrap support values of >70% are presented) and ASTRAL-III (B; local posterior probability support values of >0.7 are presented). A color version of this figure is available online.
that include these parental individuals (M. natalense L1 and M. bachmannii LP2) in the STRUCTURE plot shared an almost similar PC space closer to zero (fig. 5), which prevented a clear understanding of their genetic isolation. Samples that showed moderate support in the coalescent analysis (i.e., Barbston, Mpumalanga, PP = 0.78; Stainbank, Durban, KwaZulu-Natal, PP = 0.78; and Everton, KwaZulu-Natal, PP = 0.87) showed similar population structures with K = 7.

### Discussion

Although the main purpose of this study was to disentangle the South African Memecylon species complex, relationships within this group consistently exhibited low statistical support at both shallow and deep nodes in both the concatenated and coalescent species trees. Here, we used multiple approaches to highlight the discordance in our data set. Only introns were able to provide some resolution for this species complex, compared with supercontigs and exons from the LCN genes. Also, multiple studies have pointed out the suitability of introns for lower-level taxonomic questions (de Sousa et al. 2014; Folk et al. 2015; Govindarajulu et al. 2015). However, our concatenated RAxML analysis (fig. A2) based on intron regions indicated that Memecylon natalense and M. bachmannii are not monophyletic. We independently performed the same RAxML analysis excluding 12 taxa that seemed to have hybrid origins (fig. A2) and ASTRAL-III (fig. A2) analyses based on introns high-lighted gene and species tree discordance in this group. Low support values for shallow-level relationships are not always an indication of a lack of phylogenetic signal but can be a result of evolutionary processes that are not adequately captured by a bifurcating tree model (Dunning and Christin 2020; Stubbs et al. 2020). Therefore, the conflict observed between gene trees and the species tree, taken together with the lack of resolution in the species tree, may be considered evidence of gene flow occurring among geographically adjacent entities and/or ancient reticulation. Although it would be desirable to compare nuclear and plastome data, the percentage of off-target plastome reads recovered during sequencing was unfortunately low (<15%; Jantzen et al. 2020), which precluded the comparison of discrepancies in the nuclear and chloroplast phylogenies.

Both PhyParts and IQ-tree concordance factors (figs. A2, A4, respectively) further indicate gene tree conflicts within this clade. PhyParts recovered many gene trees in disagreement with each other and with the species tree. For some nodes, we observed 100% discordance in the gene trees. Given the significant level of incongruence, the three focal taxa should be considered a single unresolved clade.

Our SNaQ analysis suggested reticulate evolution within the South African Memecylon clade. We recovered a hybridization event that showed a nearly 50% contribution from putative parental lineages (fig. 2). Here, M. natalense from the Lusikisiki population (L1) and M. bachmannii from the Luputana population (LP2) were implicated as parental lineages of M. australissimum from Cwebe (He2). To understand the maternal inheritance, a plastid data set would be useful, but plastome sequence reads were too limited in our study. The observation that M. australissimum combines certain morphological character states of M. natalense with others from M. bachmannii could also be a result of hybridization. To confirm this, a more detailed morphological study of these populations would be necessary. The results from the SNaQ analysis (fig. 2) suggested two putative reticulation events within this clade. However, one of these putative events was not well supported (unequal contribution from

### Table 1

#### Nuclear DNA Contents of Memecylon Samples and a Comparison among Ploidal Levels Estimated Using Flow Cytometry, PloidyNGS, and NQuire

| Sample name, code | Average 2C DNA content (pg) and standard deviation from flow cytometry | Ploidal level approximate value from flow cytometry (x) | Ploidal level from ploidyNGS (x) |
|-------------------|-------------------------------------------------|--------------------------------------------------|----------------------------------|
| Memecylon australissimum, H3 | 4.48 (± 1.37) | 6 | 7 |
| M. australissimum, He2 | 6.79 (± 71) | 10 | 7 |
| M. bachmannii, E5 | 3.66 (± 47) | 5 | 7 |
| M. bachmannii, EM4 | 4.08 (± 87) | 6 | 7 |
| M. bachmannii, GM5 | 4.20 (± 2.64) | 6 | 6 |
| M. bachmannii, LP2 | 3.95 (± 55) | 6 | 6 |
| M. bachmannii, MK3 | 3.78 (± 86) | 5 | 6 |
| M. bachmannii, MK6 | 3.84 (± 39) | 5 | 6 |
| M. bachmannii, OM12 | 4.07 (± 86) | 6 | 6 |
| M. bachmannii, U5 | 3.71 (± 33) | 5 | 7 |
| M. bachmannii, 2785 | 4.05 (± 76) | 6 | 7 |
| M. natalense, B2 | 2.78 (± 54) | 4 | 7 |
| M. natalense, BR1 | 3.39 (± 12) | 5 | 6 |
| M. natalense, L1 | 2.89 (± 94) | 4 | 7 |
| M. natalense, O4 | 3.79 (± 32) | 5 | 7 |
| M. natalense, O7 | 4.44 (± 1.17) | 6 | 7 |
| M. natalense, OH1 | 3.35 (± 07) | 5 | 7 |
| M. caeruleum, SIN-TA01 | 1.42 (± 31) | 2 | 2 |
parental lineages), and this result may be due to either backcrossing or an artifact of limited sampling (of both genes and the number of individuals per population).

Estimations of genome sizes from flow cytometry (table 1) have provided further evidence for the hybridization event recovered in the SNaQ analysis. The low strength of the signal in flow cytometry prevented an understanding of the exact ploidal levels of the individuals used in this study. Even with a low signal, flow cytometry suggested a genome size nearly twice as high for *M. australissimum* (He2 sample: 6.79 pg) as for *M. natalense* (L1 sample: 2.89 pg) or *M. bachmannii* (LP2 sample: 3.95 pg). Therefore, the possibility that *M. australissimum* is a homoploid hybrid species is unlikely, and the available evidence favors allopolyploidy, which could have resulted from reticulation. However, to demonstrate further that *M. australissimum* is indeed an allopolyploid, chromosome counts of the standard group, that is, *M. caeruleum*, should be confirmed. We used silica-dried material of all samples (including the standard material) and are aware that rapid drying of leaf tissues has been found to contribute somewhat to variation in genome size estimates in plants because desiccated plant tissues sometimes preclude cell sorting (Suda and Trávníček 2006; Bainard et al. 2011). Therefore, our ploidal estimates might be affected by the use of dried material. In future studies, fresh materials of *Memecylon* would provide better confidence in the estimates of ploidal levels.

Flow cytometry, additionally, indicated polyploidy in all samples within the South African *Memecylon* clade (table 1). Reticulated patterns of evolution can also result from allopolyploid and homoploid hybrid speciation (Marhold and Lihová 2006), and, in such situations, their evolutionary patterns can be difficult to unravel using tree construction procedures (Linder...
and Rieseberg 2004; Vriesendorp and Bakker 2005). However, given the practical limitations of the methods of ploidal estimation used in this study, the type of polyploidy (i.e., allopolyploid, homoploid hybrids, etc.) of our samples is unclear. The tissue samples of the standard group used in this study were from *M. caeruleum*; however, this taxon is distantly related to the *Buxifolia* clade (Stone 2014), and there are no chromosome counts available from members of the *Buxifolia* clade. Among African *Memecylon* species, the only chromosome number and ploidal level known so far are for *M. liberiae* (2n = 14; Rice et al. 2015). Unfortunately, recently collected samples of *M. liberiae* were not available for use in this study. It is possible that a genome duplication event occurred in the ancestral lineage of the *Buxifolia* clade, which could explain the high genome sizes and ploidal levels recovered in South African *Memecylon*. From investigations in other plant groups, there is evidence that taxa may show distinct morphological variation but limited genetic diversity (Dubcovsky and Dvorak 2007; Hunt et al. 2014), which has been attributed to population bottlenecks due to polyploidy. Therefore, the populations with ploidal variations have the potential for bottlenecks. However, our ploidal level estimations from the ploidy-NGS analysis and flow cytometry are incongruent for some samples. Ploidy-NGS requires visual inspection of the allele frequency graph distribution of each sample or comparisons against simulated data to determine the ploidy of each sample, which limits the applicability of this method to high-throughput analysis (Corrêa dos Santos et al. 2017). Additionally, one of the assumptions of this method is that there is a sufficient amount of heterozygosity in the genome; otherwise, ploidy changes may go undetected (Corrêa dos Santos et al. 2017).

**Fig. 5** Principal component analysis (PCA) produced using single-nucleotide polymorphisms from each sample. The percentage of genetic variation is indicated as PC1 and PC2 on the axes. The samples are categorized as clusters that resulted from structure analysis.
Therefore, direct chromosomal counts obtained from members of the *Buxifolia* clade in southern Africa are important for verifying these results.

The STRUCTURE analysis suggested that gene flow has occurred within the South African *Memecylon* clade. This analysis (fig. 4) resulted in seven genetic clusters and showed admixture in the clusters. This genetic diversity might be maintained via a high degree of gene flow within and between populations. Some clusters from the STRUCTURE analysis were compatible with the coalescent-based phylogenies (figs. 2, 4). This observation provides further evidence of widespread gene flow within a single clade. However, our sampling density is not ideal for allowing a rigorous test of monophyly for these groups or for interpreting phylogenetic clusters.

With the available methods so far, it is not possible to connect the spatial and temporal dimensions of reticulate evolution with any degree of accuracy. However, our dating analysis of a larger *Memecylon* clade at a global scale (Amarasinghe et al. 2021a) suggests a late Miocene origin for the South African *Memecylon* clade. Therefore, the likely *M. natalense–M. bachmannii* hybridization event may have coincided with vegetation shifts due to climatic oscillations in the Pliocene or Pleistocene. During the Quaternary, the forests in South Africa expanded and contracted (van Zinderen Bakker 1978; Eeley et al. 1999). The forests in montane areas disappeared in this geological time period; however, they expanded at lower elevations such as the southern part of KwaZulu-Natal and the Eastern Cape (Lawes 1990). This forest expansion might have facilitated the range expansion of *M. natalense* and *M. bachmannii*, providing suitable conditions for interspecific hybridization. Even in their current distribution, *M. natalense* and *M. bachmannii* have overlapping geographic ranges in southern KwaZulu-Natal and the northern part of Eastern Cape. In such areas, interspecific hybridization may have naturally taken place.

Our ASTRAL-III topology (fig. 2) correlates with the geographic distribution of clades in eastern South Africa. However, low PP values for these clades indicate a lack of support. Additionally, in the STRUCTURE analysis, K = 7 does not show distinct clusters in terms of geography, indicating that the geographical isolation of these populations is unlikely (fig. 4G, 4D).

Herbarium specimen data show that the flowering times of *M. natalense* and *M. bachmannii* overlap to some degree in late November to early December (Stone et al. 2017a), and this was also observed during our fieldwork in southern KwaZulu-Natal and the Eastern Cape in November–December 2017. On the basis of limited data, *Memecylon* is pollinated by a variety of insects, including flying insects. For instance, Diptera (unidentified Syrphidae) and Hymenoptera (e.g., *Apis, Amegilla, Braunsapis*, etc.) are reported as some pollinators (Nayak and Davidar 2010; Yong et al. 2019). With this variety of pollinators, gene flow is possible even in fragmented forests. Unfortunately, no pollination study of South African *Memecylon* is available. Therefore, more studies on pollination biology and induced pollinations are needed to understand interspecific hybridization and other reticulation events, such as introgression.

A large amount of diversity in terms of ploidal levels, admixtures, and phylogenetic networks characterizes populations of *M. natalense* in the Ongoye region (KwaZulu-Natal) and populations of *M. bachmannii* in the Oribi Gorge (southern KwaZulu-Natal) and Luputana (Eastern Cape) regions. The maintenance of such high levels of diversity, in particular the number of divergent groups, indicates that the Ongoye, Oribi Gorge, and Luputana regions have been able to support sizable populations of these forest-dwelling plants throughout the recent period of climate change. Some of the collections of this species complex have been from small habitat remnants and degraded lands. For example, Luputana degraded habitat still harbors *M. natalense*. We hope that our results will help to inform conservation policies aimed at preserving the genetic diversity of plant groups in these regions.

The South African *Memecylon* clade has been recognized as a well-supported entity by Stone et al. (2017a) and confirmed by this study and Amarasinghe et al. (2021a). However, we recovered evidence for the nonmonophyly of taxa within this group, putative reticulation, and polyploidy within a single clade. Our results, therefore, suggest that this group represents a single internally unresolved and morphologically and cytologically variable clade with putative reticulate evolution.

**Conclusions**

This is the first in-depth phylogenomic study of the South African *Memecylon* species complex covering a wide range of forest types. We found no evidence for the monophyly of any taxa as currently defined within this complex. None of the methods used showed that accessions from these populations clustered together as single units. Also, multiple lines of evidence, such as STRUCTURE, flow cytometry, and network analyses, indicated the presence of reticulate evolution as the expected pattern inside this single clade. Likely, a recent and rapid origin of this group coincided with the time of forest expansion in southern KwaZulu-Natal and the Eastern Cape. Finally, some of the samples used in this study are from small, degraded habitats and areas containing a high amount of diversity in terms of ploidal levels, admixtures, and phylogenetic networks, suggesting the need to consider appropriate conservation measures to protect these areas.

**Acknowledgments**

We thank the BNRH and NU herbaria for curating our vouchers, John Burrows and Barbara Turpin of Buffelskloof Nature Reserve Herbarium and the Pondoland CREW (especially Kate and Graham Grieve) for fieldwork assistance, and Johanna Jantzen, Chathurani Ranathunga, Andrew Crowl, and R. D. Stone). We are also grateful to three anonymous reviewers and the editor Dr. Jill S. Miller for their constructive comments, which greatly improved the manuscript.
the origin of tetraploid broomcorn millet, P. milaeum. J Exp Bot 65:3163–3175.

Jantzen J, R Amarasinghe, RA Folk, M Reginato, FA Michelangeli, DE Solits, N Cellinese, PS Solits 2020 A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in Melastomataceae. Appl Plant Sci 8:e11345.

Johnson MG, EM Gardner, Y Liu, R Medina, B Goffinet, AJ Shaw, NJC Zerega, NJ Wickett 2016 HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl Plant Sci 4:1600016.

Kates HR, MG Johnson, EM Gardner, NJC Zerega, NJ Wickett 2018 Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of Artocarpus. Am J Bot 105:404–416.

Katoh K, DM Standley 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780.

Kearse M, R Moir, A Wilson, S Stones-Havas, M Cheung, S Sturrock, S Buxton, et al 2012 Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Kopelman NM, J Mayzel, M Jakobsson, NA Rosenberg, I Mayrose 2015 CoMap: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour 15:1179–1191.

Lawes MJ 1990 The distribution of the Samango monkey (Cercopithecus mitis erythrarchus Peters, 1852 and Cercopithecus mitis labiatus I. Geoffroy, 1843) and forest history in southern Africa. J Biogeogr 26:595–599.

Li H 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

——— 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 13033997.

Li H, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, 1000 Genome Project Data Processing Subgroup 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Linder CR, LH Rieseberg 2004 Reconstructing patterns of reticulate evolution in plants. Am J Bot 91:1700–1708.

Lischer HEL, L Excoffier 2012 PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28:298–299.

Maddison WP, LL Knowles 2006 Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55:21–30.

Marhold K, J Lihová 2006 Polyplody, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Syst Evol 259:143–174.

McKenna A, M Hanna, F Banks, A Sivachenko, K Cibulskis, A Kernytsky, K Garimella, et al 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303.

Minh BQ, M Hahn, R Lanfear 2018 New methods to calculate concordance factors for phylogenetic datasets. bioRxiv 487801. https://doi.org/10.1101/487801.

Nayak KG, P Davidar 2010 Pollination and breeding systems of woody plant species in tropical dry evergreen forests, southern India. Flora 205:745–753.

Nguyen LT, HA Schmidt, A von Haeseler, B Gugger 2015 IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274.
Memecylon (Melastomataceae: section Micranthes Jack (Melastomataceae) in Singapore: Phylogenetic analysis. INTERNATIONAL JOURNAL OF PLANT SCIENCES

R Core Team 2019 R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/.

Revell LJ 2012 phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3:217–223.

Rice A, L Glick, S Abadi, M Einhorn, NM Kopelman, A Salman-Minkov, J Mayzel, O Chay, I Mayrose 2015 The Chromosome Counts Database (CCDB): a community resource of plant chromosome numbers. New Phytol 206:19–26. http://ccdb.tau.ac.il/.

Roberts AV, T Gladis, H Brumme 2009 DNA amounts of roses (Rosa L.) and their use in attributing ploidy levels. Plant Cell Rep 28:61–71.

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

Servick S, CJ Visger, MA Gitzendanner, PS Soltis, DE Soltis 2015 Population genetic variation, geographic structure, and multiple origins of autopolyploidy in Galax urceolata. Am J Bot 102:973–982.

Smith SA, MJ Moore, JW Brown, Y Yang 2015 Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol Biol 15:150.

Solís-Lemus C, C Ane 2016 Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet 12:e1005896.

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

Stone RD 2014 The species-rich, paleotropical genus Memecylon (Melastomataceae): molecular phylogenetics and revised infrageneric classification of the African species. Taxon 63:539–561.