Repeatome-Based Phylogenetics in *Pelargonium* Section *Ciconium* (Sweet) Harvey

Floris C. Breman [1], Guangnan Chen[1], Ronald C. Snijder[2], M. Eric Schranz [1] and Freek T. Bakker [1,*]

[1]Biosystematics Group, Wageningen University & Research, Netherlands

[2]Syngenta Seeds BV Cornelis, Andijk, Netherlands

*Corresponding author: E-mail: freek.bakker@wur.nl.

Accepted: 22 November 2021

Abstract

The repetitive part of the genome (the repeatome) contains a wealth of often overlooked information that can be used to resolve phylogenetic relationships and test evolutionary hypotheses for clades of related plant species such as *Pelargonium*. We have generated genome skimming data for 18 accessions of *Pelargonium* section *Ciconium* and one outgroup. We analyzed repeat abundancy and repeat similarity in order to construct repeat profiles and then used these for phylogenetic analyses. We found that phylogenetic trees based on read similarity were largely congruent with previous work based on morphological and chloroplast sequence data. For example, results agreed in identifying a "Core *Ciconium*" group which evolved after the split with *P. elongatum*. We found that this group was characterized by a unique set of repeats, which confirmed currently accepted phylogenetic hypotheses. We also found four species groups within *P.* sect. *Ciconium* that reinforce previous plastome-based reconstructions. A second repeat expansion was identified in a subclade which contained species that are considered to have dispersed from Southern Africa into Eastern Africa and the Arabian Peninsula. We speculate that the Core *Ciconium* repeat set correlates with a possible WGD event leading to this branch.

Key words: repeatome, evolution, *Pelargonium*, speciation, phylogeny.

Significance

The repeatome of plants contains valuable phylogenetic information. It provides a different perspective on evolution compared with the more commonly used chloroplast-based markers. We have studied repeatome evolution in *Pelargonium* sect. *Ciconium*. We find that repeatome-based phylogenetic trees by and large confirm plastome-based trees. We find that there are repeats that are unique to the section as well as to clades and branches within the section, demonstrating the value of the repeatome for phylogeny reconstruction.

Introduction

A large part of the eukaryote nuclear genome consists of repetitive DNA sequences (discovered by Britten et al. [1974] and Flavell et al. [1974]) and the collective repetitive DNA fraction of the genome is referred to as the "repeatome" (Maumus and Quesneville 2014). In plants, the repeatome can make up >90% of the nuclear genome (Elliot and Gregory 2015; Novák et al. 2020; C value database by Leitch et al. (2019) at https://cvalues.science.kew.org/, Date accessed October, 2020).

The repeatome has been shown to be a useful resource for phylogenetic markers, especially when studying closely related species (Dodsworth et al. 2015; Weiss-Schneeweiss et al. 2015). By using both repeat abundance and repeat similarity comparisons, repeatome dynamics and evolution can be studied in greater detail (Vitales et al. 2020). The

repeatome is not homogenous but consists of different classes of repeats such as various classes of transposons, ribosomal genes, and tandem repeats, each with their own properties and dynamics (Wicker et al. 2007; Craig et al. 2015; Hartley and O'Neill 2019; Paço et al. 2019; Salim and Gerton 2019; Enriquez-Gasca et al. 2020; Louzada et al. 2020). Furthermore, repeats occur at positions throughout the chromosomes, can be variable in abundance, and can either be conserved, or rapidly evolving. Conserved repeats may have higher sequence similarity as measured across species, perhaps because they may be under functional constraints. Faster evolving repeat clusters on the other hand, are often transposable elements (TEs) that have "escaped" from purifying selection and these can be highly mutagenic (Morgante et al. 2007; Deniz et al. 2019), caused by gene disruption as a result of their insertion in the genome. Their potential to acquire new substitutions is stochastic and may result, when neutral, in an escape from purifying selection because it confers no (dis)advantage (Oliver et al. 2013). We would expect to see selective pressures to be reflected in both abundancies and sequence similarities of repeats. An example of TEs under purifying selection with subsequent duplications comes from so-called "pack MULE" TEs (Hanada et al. 2009). Using repeatome dynamics in a phylogenetic context can shed light on the evolution of the different constituting repeat-classes, that is, reconstruct "repeat demography" or the expansion and contraction of repeat clusters through time.

In this paper, we study repeatome evolution in the well-known clade *Pelargonium* (Geraniaceae) section *Ciconium* (Sweet) Harvey, which is the source of the iconic "garden geraniums" (*P. x hortorum*) and "hanging basket geraniums" (*P. x peltatum)*, both originating from species from this section (James et al. 2004; and see below). We present the first analysis in *Pelargonium* of how nuclear genomic repeat families emerge and expand during cladogenesis. In doing so, we gain insight in *Pelargonium* section *Ciconium* phylogenetic relationships from the repeatome perspective.

Existing *Pelargonium* phylogenetic trees are based on highly conserved plastid and nuclear genes (Bakker, Culham, and Gibby 1999; Bakker, Culham, et al. 1999; Bakker et al. 2004; Jones et al. 2009; Weng et al. 2012; Roeschenbleck et al. 2014). Two studies have incorporated rDNA Internal Transcribed Spacer (ITS) sequences as a proxy for the nuclear genomic perspective alongside chloroplast markers (Bakker et al. 2004; van de Kerke et al. 2019). Two studies also incorporated mitochondrial markers (Bakker et al. 2000, 2004), making the latter the only *Pelargonium* phylogenetic study to date based on the perspective of all three genomic compartments.

From these studies a consensus emerged that five main clades can be distinguished in *Pelargonium*: named A1, A2, B, C1, and C2 (Bakker et al. 2004; van de Kerke et al. 2019). These clades were associated with subg. *Pelargonium* (A), subg. *Parvipetala* (B), subg. *Magnipetala* (C1), and subg. *Paucisignata* (C2) by Roeschenbleck et al. 2014. While there are differences as to the exact placement of individual species (e.g., *P. nanum*) within these clades, the general phylogenetic hypothesis seems to be stable and has been the basis of current taxonomic classifications which recognizes 15 sections in the genus (Roeschenbleck et al. 2014).

Section *Ciconium*, which includes the previously recognized section *Eumorpha* (Roeschenbleck et al. 2014), is grouped in clade C2 (Bakker et al. 2004; van de Kerke et al. 2019) and has a base chromosome number of $x = 9$. It contains the ancestral, parental species of two of the most commonly sold *Pelargonium* cultivars: the "garden geranium" *P. x hortorum*, with as supposed ancestors *P. inquinans* and *P. zonale* (James et al. 2004), and the "hanging basket" or "ivy leaved geranium" derived from *P. peltatum* (James et al. 2004), with various contributions from *P. x hortorum*, Snijder RC, personal communication and Breman FC, personal observations). *Pelargonium*. sect. *Ciconium* represents a clade that proliferated ~5 Ma, according to plastome exon dating analysis by van de Kerke et al. (2019). The *Ciconium* clade likely evolved in the Cape Floristic Region (CFR) of South Africa (van de Kerke 2019), with several of its species occurring outside the CFR in the Eastern Cape (*P. aridum* and *P. peltatum*), Eastern Africa (*P. multibracteatum*, *P. quinquelobatum*, *P. alchemilloides*, *P. somalense*, and *P. insularis*), and the Arabian Peninsula (*P. yemenense* sp. nov., Gibby M et al., in preparation). These SW versus NE African *Pelargonium* (and *Ciconium*) disjunctions have been interpreted to reflect previous jump dispersal events, along the high-altitude rift mountain corridors (van de Kerke 2019). Despite the widespread occurrence of the section, many *Ciconium* species are restricted to specific niches (van der Walt and Vorster 1988, Vols I–III; Verboom et al. 2009) and may have gone through historical population bottlenecks.

We compared both abundancies and sequence similarities of nuclear genomic repeats in most species of *Pelargonium* section *Ciconium* in order to explore their evolution and utility as phylogenetic markers. We then combined these in a principal component analysis (PCA)-based approach to test if different repeats have different patterns of evolutionary change, which could be related to evolutionary age, chromosomal location, or specific class.

Results

Paired-end read libraries (insert size 350 bp) contained 5–7M reads of 150 bp length each. The multispecies library (MSL) was created by random sampling from each accession (see table 1). The final MSL, comprising reads from all 19 accessions consisted of 5.42×10^6 reads. Of these, 4.31×10^6 reads clustered whereas 1.11×10^6 reads were not assigned to any cluster and are therefore considered "singlets" (see table 1 and fig. 1).

Table 1
Plant Material Used in This Study

Voucher/Specimen Code /Herbarium Ref	*Pelargonium* Species	1C (pg)	Read Pairs Used	Reads in Top Clusters	Reads in Singlets/ Minor Clusters
S1002/STEU1243	*acetosum*	2.43	121,410	102,187	140,633
S1003/STEU1975	*acreaum*	2.44	121,934	116,451	127,417
S1010/STEU1885	*alchemilloides*	2.26	112,850	111,973	113,727
S1009/STEU1882	*alchemilloides (4×)*	4.14	206,872	212,131	201,613
S1088/WAG1972053	*aridum*	2.23	111,740	103,642	119,838
S1026/WAG1972055	*articulatum (4×)*	4.23	211,706	177,820	245,592
S1027/WAG1972061	*barklyi*	2.34	117,014	99,603	134,425
S1072/STEU1022	*elongatum*	1.30	64,930	31,862	97,998
S1087/WAG1972062	*frutetorum*	2.23	111,324	100,655	121,993
S1029/STEU0682	*inquinans*	2.32	115,768	99,017	132,519
S1319/STEU0621	*karooicum*	3.35	167,610	94,827	240,393
S1032/STEU2902	*multibracteatum*	3.14	157,040	186,939	127,141
S1034/STEU1263	*peltatum*	2.20	110,230	97,818	122,642
S1044/WAG1972049	*quinquelobatum*	4.13	206,494	268,654	144,334
S1045/MSUN A3651	*ranunculophyllum*	2.10	104,946	110,818	99,074
S1089/WAG1972045	*salmoneum*	2.45	122,294	108,640	135,948
S1046/STEU3074	*tongaense*	2.59	129,464	133,304	125,624
S1033/WAG1972037	*yemenense*[a]	6.07	303,702	354,143	253,261
S1056/STEU1896	*zonale*	2.27	113,440	91,146	135,734
PEZ-BD8517/ WAG1972048	*P. x hortorum*	2.33[b]	_		

NOTE.—Flowcytometry values 1C as measured in this study, total reads used per accession in the RE analysis, and percentage of clustered and nonclustered reads overall. STEU, Stellenbosch University, RSA; AL, Albers; MSUN, Münster and Bakker et al. (2004); WAG, National Herbarium of the Netherlands.
[a]sp. nov.
[b]Based on an average across three different measurements, this was the reference plant.

Analyzing the MSL on the Galaxy server using RE2 required ~3.5 days using default parameters for a "long run," except for the "RAM used by TAREAN," which was changed to 96,000,000 MB. This yielded a total of 316,059 superclusters (SCs) and 316,161 clusters of which 311 contained ≥542 reads, or 0.01% of the genome. After filtering out the 56 organelle-based clusters, the final number of clusters was reduced from 311 to 255. A comparative abundancy matrix was compiled consisting of 255 clusters for 19 accessions, representing 18 *P.* sect. *Ciconium* accessions and an outgroup species: *P. karooicum*.

Flow Cytometry

Flowcytometry values are listed in table 1. Values (1C) range from 1.3 pg for *P. elongatum* to 6.07 pg for *P. yemenense*. The largest diploid genome is *P. quinquelobatum* (4.13 pg) which is approximately the same size as the polyploid specimens of *P. articulatum* (4.23 pg) and *P. alchemilloides* (4.14 pg).

General Overview and Distribution of Reconstructed Read Clusters

About 20% of total clustered reads were in the top 20 clusters, except for *P. elongatum* and the outgroup. Within the Core *Ciconium* group (i.e., *Pelargonium* sect. *Ciconium* without *P. elongatum*) some clear differences in terms of the read abundance per cluster per accession could be observed. These will be discussed in more detail below. The clustering results in figure 1 were ranked by size (in terms of total reads per clusters). A cluster was considered "large" when it represented >1% of the total repeatome of an accession. It was considered "very large" when it represented >3% of an accession's repeatome. In the top 20 clusters #10 (*P. zonale*), #19 (*P. quinquelobatum*, *P. multibracteatum*, and *P. yemenense*) clusters fell into the "large or very large" category. Only *P. quinquelobatum*, *P. multibracteatum*, and *P. yemenense* contained autapomorphic clusters that made up >3% of their respective repeatomes (see Heatmap in supplementary material 1, Supplementary Material online).

In terms of absolute read contributions, *P. yemenense* sp. nov. was the largest -and *P. elongatum* the smallest contributor of overall accession reads in the MSL (fig. 1), which was in line with the flow cytometry measurements (see table 1). The diploid accessions *P. multibracteatum* and *P. quinquelobatum* were the largest contributors with ~175K and ~250K (fig. 1) reads respectively, per accession, which was comparable to that of the tetraploid accessions. This was not unexpected given their large flow cytometry-based 2C values we found for these accessions

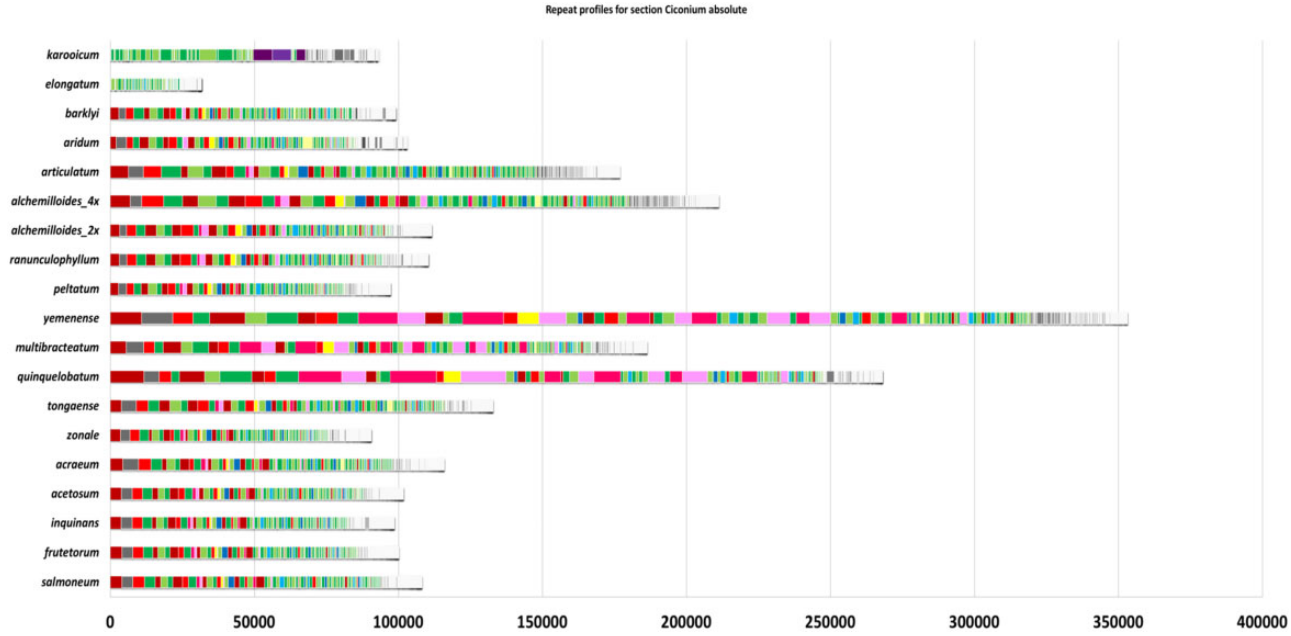

Fig. 1.—Stacked histogram of homologous and abundance-ranked repeat clusters. The *x* axis denotes cumulative read counts and the *y* axis denotes the accessions. The colors indicate the different cluster categories in terms of their occurrence on the *Ciconium* phylogenetic tree. Green denotes symplesiomorphic clusters, red core *Ciconium* clusters (*Pelargonium elongatum* and *Pelargonium karooicum* contribute <0.1% reads to the cluster), blue *Ciconium* synapomorphic cluster (outgroup contributes <0.1% reads), yellow autapomorphic clusters (one accession contributes >20% reads), pink; other synapomorphic clusters (two or more accession contribute >20% reads each), purple outgroup-specific autapomorphic clusters (*P. karooicum* contributes >50% reads), gray other, small, clusters.

(see table 1). *Pelargonium yemenense*, *P. multibracteatum*, and *P. quinquelobatum* further shared a number of potentially synapomorphic clusters in terms of abundance and in terms of similarity which is further discussed below.

Pelargonium karooicum and *P. elongatum* each had four clusters from the "large" category (#6, #14, #15, and #20 for *P. elongatum* and #47, #63, #88, and #95 for *P. karooicum*) in the top 100. For *P. karooicum* three of these clusters were autapomorphic with virtually no contributions from the other accessions reflecting its more ancient common ancestor with rest of the accessions. Even though *P. elongatum* does share these four aforementioned clusters, it is conspicuous for the fact that it, just as for *P. karooicum*, contributed virtually no reads to the other large clusters that occurred in the Core *Ciconium* accessions. The plotting of nonsymplesiomorphic clusters over the abundancy-based tree (fig. 2) yielded the following results: 174 *Ciconium* synapomorphic clusters, 17 of which are "Core *Ciconium*" synapomorphic clusters. We counted 22 synapomorphic clusters, and 17 autapomorphic clusters (see supplementary material 2, Supplementary Material online).

Superclusters

From the top 100 most abundant clusters, 26 belonged to SC 1 (SC1). SC1 contained virtually no reads from *P. elongatum* nor *P. karooicum* with respect to read contributions per cluster (<0.1% for both *P. elongatum* and *P. karooicum*, fig. 1) making it diagnostic for Core *Ciconium*. Two SCs are presented in figure 3. SC3 is a symplesiomorphic clusters and is added for contrast.

Abundancy-Based Approach

From the comparative abundancy matrix, we generated a cladogram based on the 19 accessions with 255 characters (supplementary material 3, Supplementary Material online), 29 of the characters represented "incomplete clusters." The cladogram contained 16 clades with Bootstrap support (BS) values ≥50. Nine of these had BS values >85.

Similarity-Based Approach

From the similarity matrix, we used the same clusters as for the comparative abundancies, but we removed one incomplete cluster which contained only reads for *P. karooicum*. From the resulting 254 neighbor-joining (NJ) trees, we generated a consensus network (similarity-based CN) with edge threshold set to 0.1 (supplementary material 4, Supplementary Material online). The similarity-based CNs for the overall data set (fig. 4) showed little conflict amongst the accessions, but was also poorly resolved and not informative with regards to the relations between the *Ciconium* accessions. Therefore, we opted to take an edge threshold of 0.05 (fig. 4) and in this CN, we

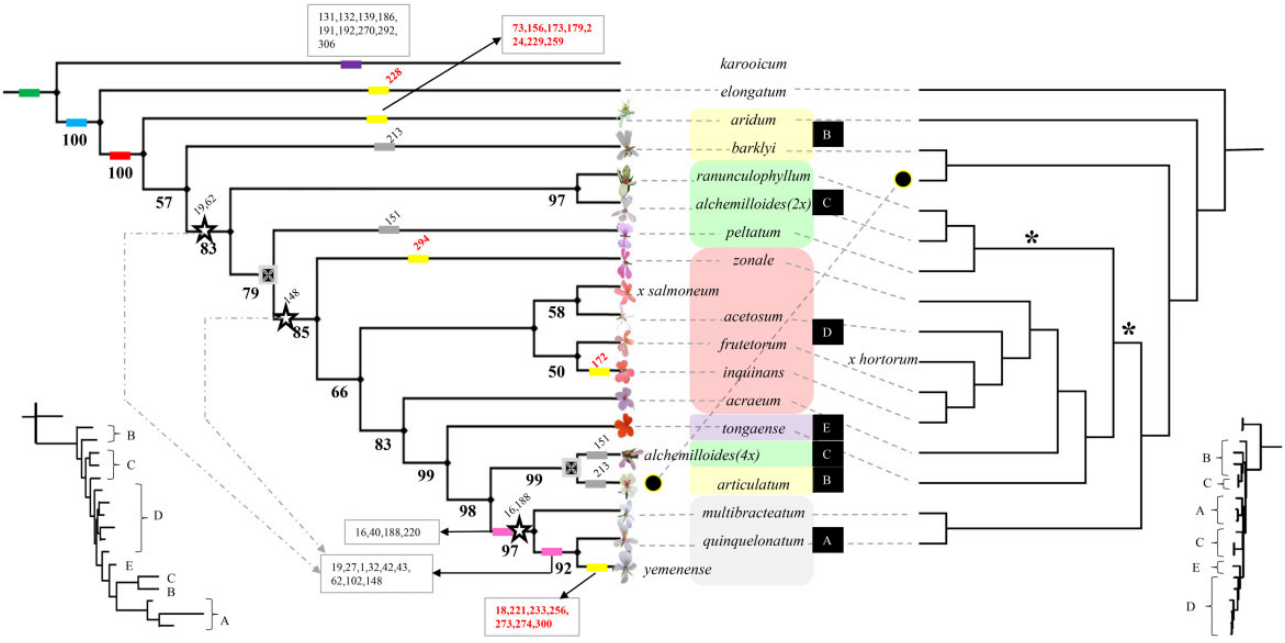

Fig. 2.—Repeat cluster abundancies. Abundancy-based cladogram compared with the most recently published plastome-based phylogenetic tree by van de Kerke et al. (2019). Repeat clusters are indicated by colored boxes. Symplesiomorphic clusters are indicated in green; autapomorphic clusters in yellow; synapomorphic clusters in blue (for *Ciconium*), in red for Core *Ciconium*, in pink for other clades, and polyphyletic clusters are in gray. For the corresponding clusters, see supplementary material 11, Supplementary Material online. Letters A–E and color groupings of the names refer to informal groups as indicated in figure 2. Nodes absent from the similarity-based consensus networks are indicated by "." Bootstrap support values are labeled in bold type on the branches. Small numbers indicate cluster synapomorphies for nodes with 0.1% threshold for read contribution (with clusters indicated). The dashed arrows indicate clusters that shifted when the threshold for synapomorphy was changed from 0.1% read contribution to 20%. Inset: the same tree as additive tree. Names of species absent in either this or van de Kerkes' study are shifted left or right dependent on the study that includes them; black circles indicate species placed incongruently among the two trees, which is possibly due to "Large Genome Attraction" (see text). *Indicates nodes with BS <90 in van de Kerke et al. (2019). Inset: same tree, now as additive tree.

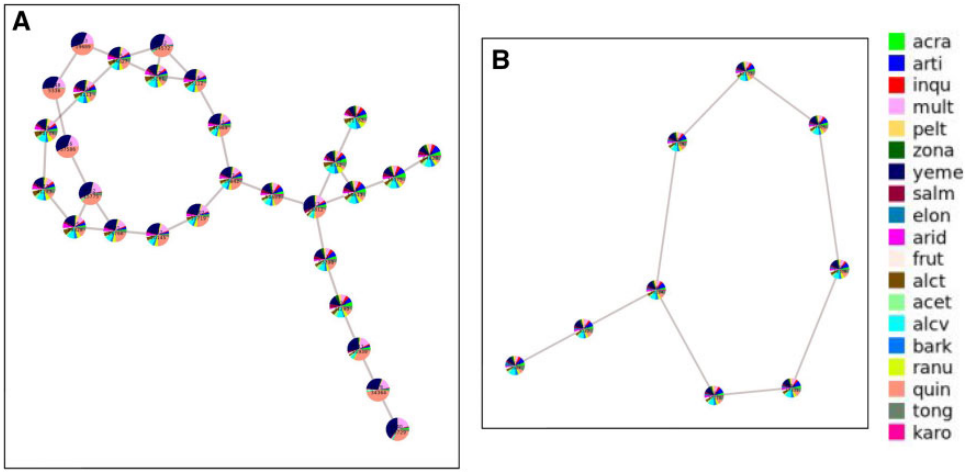

Fig. 3.—Superclusters. (*A*) SC1. This SC consists of the core *Ciconium* clusters; (*B*) SC3. This SC consists of the symplesiomorphic clusters with contributions from all accession including the outgroups. In both figures, the pie charts represent clusters connected through mates (read pairs). The colors in the pie charts indicate contributions from each accession to a particular cluster.

could still distinguish the following four informal groups in: a "Dispersing" group (A) containing *P. quinquelobatum*, *P. multibracteatum*, and *P. yemenense*. sp. nov..; a "Yellow-Flowered" group (B) containing *P. aridum*, *P. barklyi*, and *P. articulatum*; Creeping and Climbing group (C), which contains *P. ranunculophyllum*, *P. peltatum*, and *P. alchemilloides*

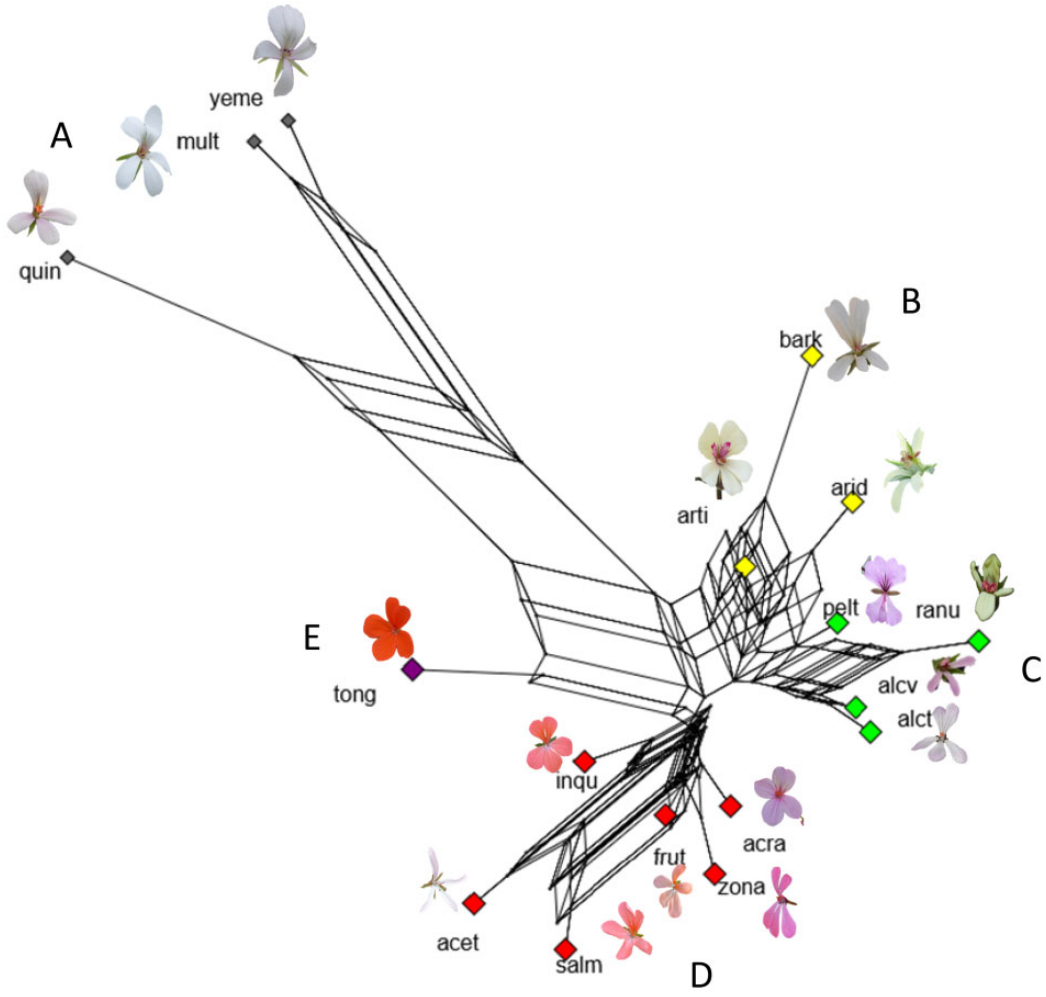

Fig. 4.—Repeat cluster similarities. Consensus network (with mean edge weights and conflict threshold 0.05) of 254 NJ trees based on all 255 read clusters (including incomplete ones) that contain ≥542 reads; organelle-based clusters are excluded (see text for further details). The length of the edges corresponds to the number of splits supporting it (Holland et al. 2004). Informal/putative groups are indicated, that is, the "dispersing" group (*A*), the "yellow-flowered" group (*B*), the "creeping and climbing" group (*C*), the "red-flowered" group (*D*), and *Pelargonium tongaense* (*E*) not placed in any group.

(2× and 4×); and finally, a "Red-Flowered" group (D) which contains *P. acraeum*, *P. zonale*, *P. frutetorum, P. inquinans, P. acetosum*, and *P. x salmoneum*. *Pelargonium tongaense* ("group" E) remained as a single branch, connected to the others via several splits. Thus, our similarity-based CN could be summarized in a tree as follows: (*P. karooicum* (*P. elongatum* (B(C(A(*P. tongaense*, D))))))). The Red-flowered and Dispersing groups seem to be clearly separated from the rest of the section with relatively few conflicts occurring under any of the two evaluated thresholds. Between the Yellow-flowered and Creeping and Climbing group there remained a number of unresolved conflicts.

Abundancy versus Similarity

The overall phylogenetic patterns were similar when comparing the abundance- and similarity-based approaches, with only two clades differing in position between the two. In the similarity-based analysis the dispersing group was clearly separated from the red-flowered group. However, in the abundancy cladogram, Core *Ciconium* Dispersing and Red-flowered groups (fig. 2) were nested in a larger clade (BS = 85) with BS = 97 for the Dispersing clade. This clade further contained *P. articulatum* and the tetraploid *P. alchemilloides*. Contrastingly, *P. articulatum* was located much closer to *P. barklyi* and *P. aridum* in the similarity analysis (fig. 4). In the abundancy cladogram, *P. peltatum* formed a single branch and *P. yemenense* was now sister to *P. quinquelobatum*, while in the similarity analysis it was sister to *P. multibracteatum*. Even though the Yellow-flowered group and the creeping and climbing group in the similarity-based CN (fig. 4) were not entirely resolved, the tetraploid *P. alchemilloides* grouped together with the diploid

P. alchemilloides and *P. ranunculophyllum*. The Creeping and Climbing group also contained *P. peltatum* and while the relationships with the other accessions were not entirely clear, this does not contradict either the abundancy or similarity analyses. Therefore, it seemed likely that polyploids "attract" each other by virtue of their large genome sizes (large genome attraction or LGA) in the abundancy analysis and this was not the case when analyzing the similarities.

Ciconium-Specific Patterns

We found that most of the diploid Core *Ciconium* accessions contributed comparable amounts of reads to the overall analysis, except for *P. multibracteatum* and *P. quinquelobatum*, whose accessions contributed ~70–150% more reads.

We represented relationships between repeat clusters using SCs. In *Ciconium*, SC1 is unique (fig. 3) to the Core *Ciconium* accessions whereas SC2 consisted solely of plastome-based clusters (not shown) and SC3 is the largest symplesiomorphic SC consisting of nine clusters (fig. 3). SC1 consisted of 31 clusters in total with 26 of them in the top 100 and eight in the top ten. SC1 is expanded in, and diagnostic for section *Ciconium* relative to *P. elongatum*. This SC is further expanded within *P. quinquelobatum*, *P. multibracteatum*, and *P. yemenense* relative to the other Core *Ciconium* accessions. These three accessions contain homologous clusters with the rest of the section, but six clusters (#16, #19, #32, #40, #43 and #148) from SC1 were unique to these three accessions (the other accessions contributing <0.1% reads). Taken together these made up ~10% of their respective genomes indicating expansion and possibly a relaxation of constraints on proliferation of these repeats. Therefore, it seemed that species with a high 2C value, that is, *P. quinquelobatum*, P. *multibracteatum*, and *P. yemenense*, had a different repeatome development as compared with diploid species, but similar to polyploid species.

Repeat Abundancies and Sequence Similarity Trends

Comparison with Existing Phylogenetic Trees for P. *sect* Ciconium

The same four groups and four single lineages that were recoverable from the similarity-based repeat analysis (see description above and fig. 4) were also present in our consensus network summary of published phylogenetic trees (plastome-CN) (supplementary material 5, Supplementary Material online). The single lineages, connected through multiple splits to other terminals, were *P. acraeum*, *tongaense*, *ranunculophyllum*, and *P. aridum*. The large number of unresolved splits between the "Creeping and Climbing" group and the "Yellow-flowered" group indicates that these should be merged. Experimental evidence from interspecific crosses also demonstrates that representatives of these groups yield fit, green plants albeit infertile (Breman FC et al., in preparation) indicating at least some incompatibility between the groups. Based on further evidence from plastome-based phylogenetic trees (van de Kerke et al. 2019) and morphology (Roeschenbleck et al. 2014), we treat them here as separate groups.

The abundancy-based tree (fig. 2) conflicted with the plastome-CN, caused chiefly by the "polyploid branch attraction" mentioned above. The similarity-based CN on the other hand approached the plastome-CN much closer, with *P. aridum* being included by us in the yellow-flowered group rather than being considered a single lineage. However, it did not contradict the plastome-CN directly as *P. aridum* could be considered a single lineage as well as it has a few splits setting it apart from *P. barklyi* and *P. articulatum*. There was agreement between the plastome-CN, the similarity-based CN and the abundancy base tree with 100% BS for the placement of *P. elongatum* in *Ciconium*, separate from the Core *Ciconium* accessions, which themselves formed a clade with 100% BS. Within the core *Ciconium*, BSs >85% are found for the dispersing group nested in the red-flowered group, effectively making it paraphyletic. The two accessions for *P. alchemilloides* were split in our abundancy-based tree. This was, again, in conflict with the similarity-based CN (fig. 4), but not with the most recent phylogenetic tree by van de Kerke et al. (2019) which also saw *P. alchemilloides* being grouped in different clades (fig. 2).

Pelargonium articulatum was sister to the tetraploid *alchemilloides* and this was also in conflict with the similarity-based CN, the plastome-CNs and most recently published phylogenetic tree. In the abundancy-based tree *P. aridum* and *P. barklyi* are also single branches. This was partly in conflict with both the similarity-based CN and the plastome-CN where *P. barklyi* was grouped with *P. articulatum*. *Pelargonium aridum* formed a single branch in the plastome-CN and the abundancy-based tree, but not in the similarity-based CN. *Pelargonium peltatum* formed a single branch (BS = 79) in the abundancy-based tree, but not in the similarity-based CN or the plastome-CN. The grouping of *P. inquinans, frutetorum, zonale, salmoneum* in one clade and of *P. multibracteatum* and *P. quinquelobatum* in another was supported by all three analyses. *Pelargonium yemenense* is a new species and was not included in the summary of previously published phylogenetic trees. *Pelargonium insularis* was not included in our study (but was in Bakker et al. [2004]), therefore their positions cannot be directly compared. However, they are likely closely related because in all analyses they are sister to *P. multibracteatum* and *P. quinquelobatum*. The positions of *P. acraeum* and *P. tongaense* vary, but all analyses agreed that *P. acraeum* is either a sister to, or part of the red-flowered group in the plastome-based CN (supplementary material 5, Supplementary Material online) and the similarity-based CN respectively (fig. 4). *P. tongaense* is a

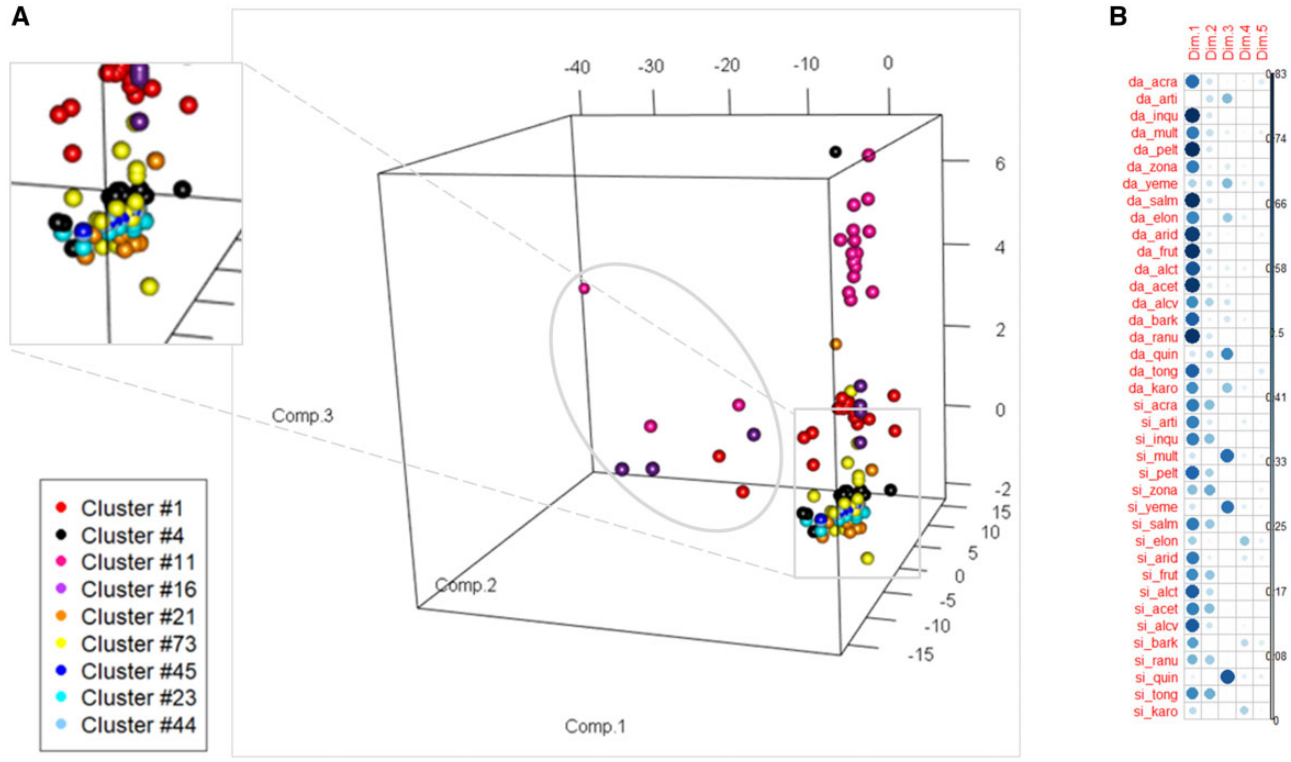

Fig. 5.—Principal component analysis on contrasting repeat clusters similarity and abundance; (*A*) six selected clusters (1, 4, 11, 16, 21, 73) along with three plastome-based clusters (23,44,45) for comparison (see the cutout). The ellipse and indicates clusters belonging to the dispersing group (*A*). The red dots denote a core *Ciconium* synapomorphic cluster, the orange dots represent a *Ciconium* synapomorphic cluster, the black dots represent a symplesiomorphic cluster, the pink dots represent a cluster synapomorphic for the dispersing group (*A*), the teal, light blue, and blue dots represent the three chloroplast-based clusters, and the yellow dots represent an autapomorphic cluster. (*B*) Contribution of each variable to the first 5 axis of the PCA. "Dim" refers to axis 1–5, respectively, "da_####" denote the pairwise abundancy differences. "si_####" denote the pairwise similarities and a four-letter species acronym.

single branch in all analyses and its position remains unresolved in the consensus networks (fig. 4).

Contrasting Repeat Abundance and Similarity Patterns

The primary axis of the PCA (fig. 5*A* and *B*) explains both abundancy and similarity differences among the *Ciconium* accessions. Repeat abundancy showed a positive correlation with the primary axis whereas the repeat similarities were negatively correlated (see the biplot in supplementary materials 6 and 7, Supplementary Material online). The abundancy differences were more pronounced, and they explained the largest part of the first axis (fig. 5*B*). The second axis is also explained by both the abundancy and similarity variables, but mainly for accessions from the dispersing clade, indicated by the dots in the ellipse in figure 5*A*. Interestingly, the third axis is mainly explained by the abundancy differences and similarities from the polyploid and large diploid genomic accessions, perhaps suggesting a different trend (fig. 5*A* and *B*). From our PCA, we see that clusters in the plastome, symplesiomorphic and the *Ciconium* category are driven both by differences in similarity and abundancy but there is no indication of expansion or contraction, see the cut-out in figure 5*A*, leaving sampling differences as main cause explaining the abundancy patterns. The Core *Ciconium* cluster (red dots) in contrast, showed increased correlation with the first axis and an increase in abundancy seemed to explain these clusters better. These are possibly younger clusters that have not yet gone through curbing/restraining of their expansions. The synapomorphic clusters (red, orange, and pink dots in fig. 5*A*) were variable with both similarity and abundancy being responsible for the observed variation. This may indicate a lack of constraints on expansion and substitutions, making these possibly even younger than the Core *Ciconium* clusters. These clusters corresponded to the clusters from SC1 that are synapomorphic for the clade of *P. yemenense*, *P. multibracteatum*, and *P. quinquelobatum* whose estimated age of about ~1.5 Myr (van de Kerke et al. 2019) is indeed considerably younger than the rest of *Ciconium*.

The autapomorphic clusters also showed contrasting patterns, especially in the case of cluster 16, which belongs to the clusters that have expanded in the abovementioned three accessions. Cluster 73 in contrast, which is autapomorphic for *P. aridum*, does not appear to be inflated or diverged, it

represents perhaps an older cluster that evolved uniquely in this species. Figure 5*A* further shows that the plastome-based clusters are highly correlated, with all showing virtually equal trends, in contrast to the five repeatome-based clusters added to the same analysis. This further suggests that different (or fewer) genomic constraints act on the six selected repeatome-based clusters. Testing for correlation between similarities and abundancy differences, for the selected six clusters and five accessions, revealed no correlation between the two types of data (not shown, but see supplementary material 8, Supplementary Material online).

Reproducibility of Cluster-Based Phylogenetic Patterns

We reconstructed similarity-based CNs and cladograms (abundancy-based trees) using the 10%, 25%, 50%, and 100% Multi Species sub libraries (supplementary material 9, Supplementary Material online). We have used the top 50 largest clusters of the MS sublibraries and these yielded 37/38/45/47 repeatome-based NJ trees from which we reconstructed the similarity-based CNs (fig. 6). The abundancy-based trees were based on 37/38/45/47 characters with repeatome-based abundancy scores, respectively (supplementary material 10, Supplementary Material online). When comparing the similarity-based CNs, we observed that the patterns obtained are largely consistent across the 10%, 25%, 50%, and 100% samplings. Overall, node resolution appears to increase with increasing sampling.

Data Consistency Analysis

We find that the groups distinguished by the analysis based on 255 clusters are also recovered by our cumulative trees. The informal groups (labeled a–e) are recovered using the first 50 clusters and adding the others does not change this (fig. 7*A*–*E*). We do see that the number of among-NJ tree conflicts in the consensus networks increases somewhat, especially for the dispersing clade in the cumulative network when using clusters 100–150 (fig. 7*C*), but this resolves when adding clusters 150–200 (fig. 7*D*). Adding these increases the conflicts in the red-flowered group, but these resolve when adding the last slice of 50 clusters (fig. 7*E*). Overall though, throughout the figures, the patterns of five groups remains discernable regardless of the amount of clusters used.

Discussion

Our study confirms earlier findings (Dodsworth et al. 2015, 2017) that the repeatome contains a treasure trove of information useful for studying phylogenetics and evolution. Combining sequence similarities and abundancies for a number of clusters appears to provide important information about changes in abundancy and sequence similarity for particular repeat classes. Vitales et al. (2020) left out clusters lacking edges between species as they provide no information of interspecific relationships. We decided to include the autapomorphic clusters in both the abundancy and similarity analysis as they could, potentially, shed more light on the evolutionary history of that particular accession. These types of clusters, by their very nature, reinforce the differences between accessions, some of which contribute reads to them and some do not. In the context of synapomorphic clusters, we expect the autapomorphic ones to add to better resolution in our comparative analysis.

Pelargonium Sect. *Ciconium* Phylogeny

Phylogenetic patterns are well studied in *Pelargonium* using mainly cpDNA-based characters (Bakker et al. 1998, 2000, 2004, 2005; Bakker, Culham, and Gibby 1999; Bakker, Culham, et al. 1999; James et al. 2004; Jones et al. 2009; Roeschenbleck et al. 2014; van de Kerke et al. 2019). Comparisons of chemical compounds (Lis-Balchin 1996, 1997) and karyology (Gibby et al. 1990) have also been used to assess relationships in *Pelargonium* in general and for section *Ciconium* in particular. Some nonchloroplast sequences have been used for phylogenetic reconstructions including nuclear genomic rDNA ITS sequences (Bakker et al. 2004; James et al. 2004; Jones et al. 2009; van de Kerke et al. 2019) and mitochondrial encoded *nad6* exons (Bakker et al. 2000).

Our phylogenetic trees were based on the nuclear genomic repeat landscape, which are considered to be independent from specific constraints acting on individual genes or genomic regions (Dodsworth et al. 2015). Our similarity-based CN (fig. 4) was comparable to previously published phylogenetic trees (e.g., van de Kerke et al. 2019, see fig. 2). Both the abundancy- and similarity-based analysis confirmed the close relationships between the accessions of the red-flowered group (fig. 4). The grouping of Core *Ciconium* (or the splitting of *P. elongatum*) was supported by all analyses. In contrast, the cluster abundancy-based analysis was in conflict with the previously published literature and our sequence similarity-based analysis. In this analysis, the dispersing group was nested in the red-flowered group together with *P. articulatum* and this was not found in any other analysis.

Abundancy versus Similarity

Abundancies encountered in the repeatome have been reported to be phylogenetically informative (Dodsworth et al. 2015), meaning that related accessions will have similar abundancies. Our results call this into question and contrast with what was found in *Nicotiana* (Dodsworth et al. 2017) where accessions with comparable genome size differences were included. However, in the case of *Nicotiana* no major autapomorphic repeat clusters were found, or these were excluded from the analysis, and the repeat characteristics were, by and large, comparable over the range of accessions

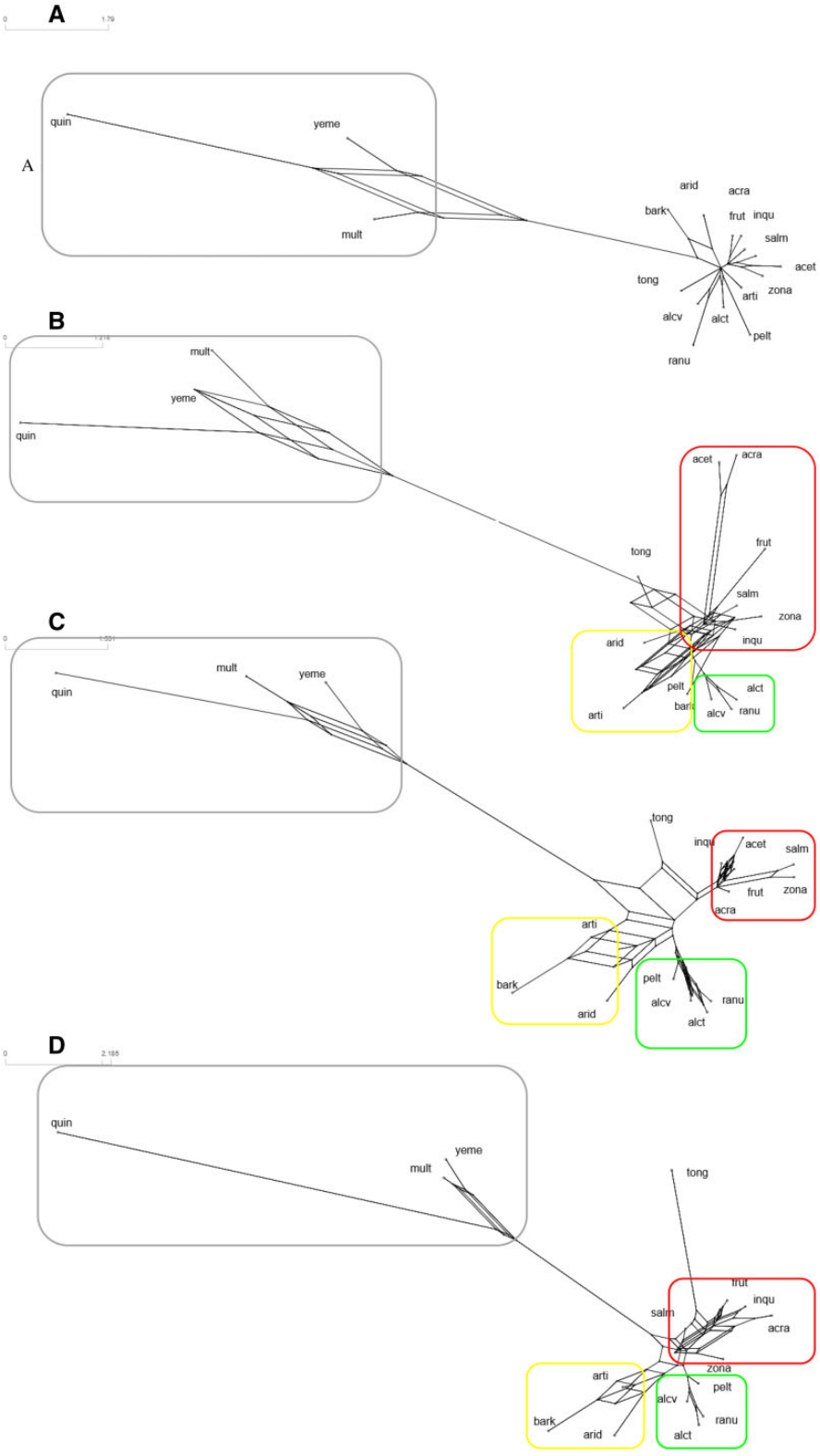

Fig. 6.—Pattern reproducibility. Consensus networks of NJ trees that are based on top 50 clusters (excluding organelle-based clusters including incomplete clusters) of the Repeat Explorer analysis using 10% (*A*), 25% (*B*), 50% (*C*), and 100% (*D*) of the multispecies read library (see text). We have excluded *Pelargonium karooicum* and *Pelargonium elongatum* as these species appeared on too long edges. Edge weights: mean; threshold 0.1. Colored boxes and group labeling (*A*–*E*) as in figure 2.

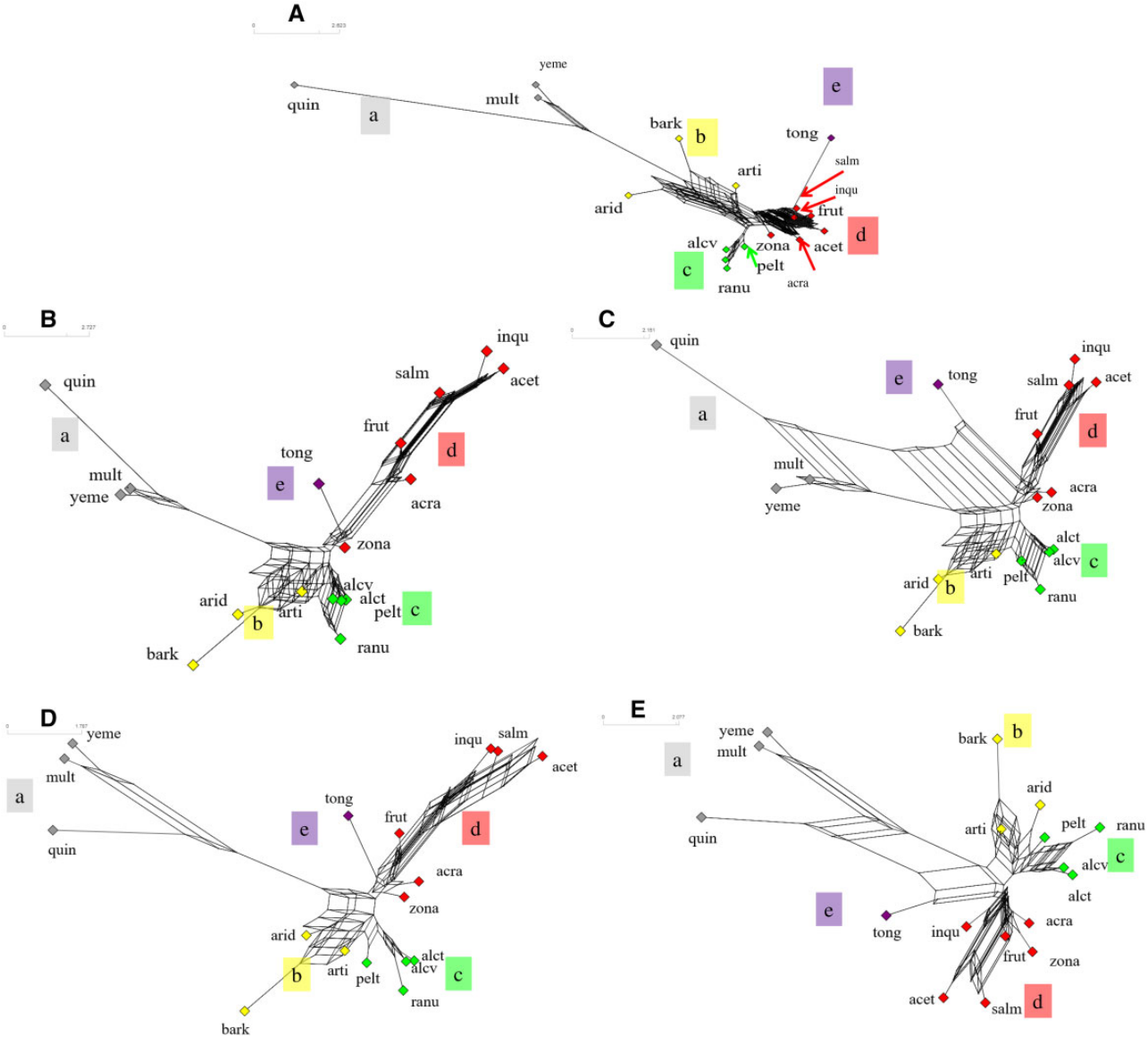

Fig. 7.—Repeat cluster similarities. Consensus network (with edge weights mean and threshold 0.05) of the first 50 (*A*), 100 (*B*), 150 (*C*), 200 (*D*), and 250 (*E*) NJ trees based on all clusters (including incomplete clusters) that contain ≥542 reads; with organelle-based clusters are excluded (see text for further details). The length of the branches corresponds to the number of splits supporting this branch (Holland et al. 2004). Informal/putative groups are indicated, that is, the dispersing group (a, gray nodes), the yellow-flowered clade (b, yellow nodes), the creeping and climbing group (c, green nodes), the red-flowered group (d, red nodes), and *Pelargonium tongaense* by itself (e, purple node). (*A*) shows the CN for trees 1–50. (*B*) show the CN for trees 1–100. (*C*) shows the CN for trees 1–150. (*D*) shows the CN for trees 1–200. (*E*) shows the CN for trees 1–250.

included. This, in itself, does not have to indicate that there is no phylogenetic signal in our abundancy data, but large genomes attract each other by virtue of their increased overall read contribution (LGA), confusing the pattern. To mitigate these effects, we applied a square root transformation, but it did not completely remove LGA. In particular those repeats belonging to the Core *Ciconium* repeat SC1 may be responsible for LGA, as these were found to be expanded across the Core *Ciconium* and even more so in the Creeping and Climbing group (fig. 4).

Another cause for these discrepancies if the abundancy data versus the similarity data is compared is the inclusion of autapomorphic clusters, which, if one or a few unrelated accessions carry some, could lead to long branch attraction purely based on these few, very abundant clusters. Therefore, the abundancy-based data, at least in our case, is probably more appropriately used to study repeatome evolution rather than phylogeny reconstruction. We found that the similarity-based approach provided a solid basis for studying the phylogenetic relationships between the accessions and avoids the

kind of abundancy-related artifacts mentioned. We are cautious in interpreting the relationships found based on the abundancies alone for some diploid accessions (e.g., *P. quinquelobatum*) as well because the relationships of the tetraploid accessions (*P. alchemilloides*, *P. articulatum*, and *P. yemenense*) are likely the result of LGA. Our suspicion was further reinforced by the observation that the abundancy-based tree does not correlate well with the previously published phylogenetic trees nor with our similarity-based CN.

Ciconium Phylogenetic Patterns

The Greater Cape Floristic Region (CFR) is a plant evolutionary model area, comprising plant species hyperdiversity. It contains a wide variety of microclimates, soil types, fire regimes on a relatively small geographical region, which is thought to have spurred speciation in a number of angiosperm clades (Linder 2003; Verboom et al. 2009; Warren et al. 2011) with *Pelargonium* being the seventh largest one of them. The GCFR includes an estimated 80% of known *Pelargonium* species (van der Walt and Vorster 1988, Vols. I–III; Bakker et al. 2005; Verboom et al. 2009; Roeschenbleck et al. 2014). As with other GCFR clades, *Pelargonium* species too (Verboom et al. 2009) generally display highly local, and endemic distributions (van der Walt and Vorster 1988, Vols. I–III; Marais 1994, 2005, 2014; Roeschenbleck et al. 2014; gbif.org, GBIF home page, https://www.gbif.org, https://www.gbif.org/occurrence/search?q=Pelargonium; 2020; Date accessed September, 2020). This may lead to very small effective population sizes and, combined with the large differences in ecosystem parameters (climate, soil, pollinators), this may have led in turn to the fixation of specific repeat clusters through genomic drift processes. As an example, especially in *P. aridum*, autapomorphic clusters occur (fig. 2) that may be the results of fixation due to historical population bottlenecks. The other example from this study came from the "dispersing group" containing *P. multibracteatum*, *P. quinquelobatum*, and *P. yemenense* which was supported by six synapomorphies. Further, *P. yemenense* has an additional seven autapomorphic clusters. These stand out even more because they appear to be the result of a dispersal event either out of the CFR via the "African track" (Linder et al. 1992) and then via Socotra to the Arabian Peninsula or the other way around. They also possibly went through quite a population size bottleneck, possibly causing the fixation of the large, unique repeat clusters inferred here. We may in fact have underestimated the occurrence of these autapomorphic clusters because we used the threshold of 20% of reads in a particular cluster that must originate from a single accession for it to be considered an autapomorphic cluster. This is $\sim$10$\times$ more than can be expected if the reads are distributed evenly per accession.

Pelargonium. Sect. *Ciconium* Repeatome Evolution Shaped by a Possible WGD Event?

WGDs are widespread in angiosperms, and these have had a profound effect on the evolution of flowering plants (Soltis et al. 2009, 2015; Schranz et al. 2012; Wendel 2015). One of the resulting effects is an increased speciation rate (Landis et al. 2018). Section *Ciconium* comprises a recently evolved group of species according to (van de Kerke et al. 2019) that emerged in the CFR $\sim$5 Ma. Whether we see an increased speciation rate in Core *Ciconium* is tempting to conclude but still difficult to say. Perhaps multiple $x = 4$ species went extinct already, and a speciation analysis using for instance "Lineages Trough Time" plots may not be applicable here.

Support for a WGD comes from the following evidence. We found >25 unique Core *Ciconium* clusters and the increased basic chromosome numbers of the Core *Ciconium* accessions (from $x = 4$ to $x = 9$, see Gibby et al. [1996], and Bakker et al. 2005) correlate with larger Core *Ciconium* Cx values (2.23–4.13 pg) compared with *P. elongatum* (1.3 pg, see table 1). Finally, the occurrence of polyploids in some of our accessions (*P. articulatum P. alchemilloides, and P. yemenense* sp. nov.) are local WGD events on within these three species. Polyploidization may further increase genome plasticity explaining the varied niches (Leitch and Leitch 2008) that *P.* section *Ciconium* species occupy. The occurrence of, "Core *Ciconium*" specific repeat clusters, especially those in SC1 (fig. 3*A*), could indicate incomplete genome reduction, which is common after a WGD event (Sankoff et al. 2010; Buggs et al. 2012). This is known to especially affect the noncoding, repetitive DNA (Eckardt 2001; Freeling et al. 2012). However, we do not see this effect in our repeat profiles when comparing them with *P. elongatum*. We would have expected more similarity between *P. elongatum* and the other species had this reduction progressed further. Previously, based on transcriptome analysis of one *Pelargonium* species in comparison with other angiosperms, Ren et al. (2018) deduced that there may have been a WGD ($\pm$10 Ma, HPD: 9.0–10.5 Ma) in the stem-lineage of *Pelargonium*. However, their data are based on one *P. x hortorum* specimen, a taxon that is often sold as a polyploid (Snijder RC, personal communication). Our data hint that this may have taken place, but rather after the split of the Core *Ciconium* species from *P. elongatum*, placed at $\pm$5 Ma (HPD 4.0–6.0 Ma) by van de Kerke et al. 2019. Perhaps more than one WGD event occurred in genus *Pelargonium* with one leading to the much younger lineage of section *Ciconium*. Finally, the occurrence of polyploids, derived from diploid interspecific crossings (Breman FC et al., unpublished data; Snijder RC and Breman FC, personal observations), shows that parental species generate polyploid pollen and have matching genomes. The pattern of the *Ciconium* proliferation driven by possible WGD (Leitch and Leitch 2008) is not unlike patterns observed in other angiosperm clades (at family-level) associated with WGD events such as Poaceae

(Leitch et al. 2010), Brassicaceae (Johnston et al. 2005; Koo et al. 2011), and Asteraceae (Vitales et al. 2019), but see Puttick et al. (2015) for a critical review.

Analyzing Patterns of Genome/Repeatome Evolution

PCAs are commonly used to analyze large genomic data sets especially when answering questions of trends in populations or groups of closely related species such as: heritability of selected SNP markers in *Citrus* (Ollitrault et al. 2012); to detect convergent evolution of a gene in humans (Galinsky et al. 2016); to detect direction of evolution after hybridization in *Vitis* (Miller et al. 2013); or to detect genetic divergence between closely related species of *Camellia* (Yang et al. 2016). All these analyses have in common that they aim to discover a trend in a large genomic data set (often SNP based) finding which combination of characters provides the most explanation for patterns observed. We aimed to do the same using the abundancy differences and similarities as characters. We proposed that those clusters that did not show obvious trends (positive or negative) for abundancy and similarities could be considered conserved or "evolutionary old," assuming some sort of compacting/streamlining through time. The symplesiomorphic clusters should especially display such a pattern. We expected that clusters that display other patterns (e.g., increased effects of abundancy or similarity) to be clusters that could be evolutionarily younger, having so far eluded constraints on their proliferation. Our results indicate that the PCA is indeed an appropriate method to explore the repeatome for clusters that have differing trends. Symplesiomorphic clusters do show highly similar trends, whereas those clusters (e.g., #11 and #12) that have expanded significantly, such as the syn- or autapomorphic ones in the dispersing group, show a trend for being especially affected by the expansion, but also by changes in similarities (therefore by substitutions) (see fig. 5*A*). The Core *Ciconium* clusters also display such a pattern, but the effect of similarity and abundancy is smaller compared with those of the clusters expanding in the dispersing group. Given that the grouping of points in the PCA for these clusters do not show obviously different trends from the symplesiomorphic clusters suggests that some constraints are already acting on these both with regards to abundancy and similarity.

To be able to also detect possible positional effects (centromeric or telomeric) we would, ideally have an annotated *Pelargonium* genome to be able to map our repeat clusters to a specific region. Since this is not available, we have used the plastome, which has tight and, presumably, equal selective pressures acting on it. Plastids are highly constrained in their function (Wicke et al. 2011) and we expected clusters based on the plastome to show a roughly equal similarity and abundancy patterns. In other words, these cp-cluster behave as an "old" repeat cluster would. Our evidence shows they do (fig. 5*A*). The use of the plastome as an approximation of a single genomic region does require some caution because different selective pressures, such as increased substitution rates for specific regions (e.g., in *Oenothera*, see Greiner et al. 2008, or in *Caragana*, see Jiang et al. 2018) or structural re-arrangements (e.g., for the plastome in Silene, Sloan et al. 2014) have been reported. *Pelargonium* plastomes are also subject to increased re-arrangements and increased nonsilent substitution rates (Weng et al. 2014, 2017; Ruhlman and Jansen 2018) thus we need to be cautious when interpreting the different plastome-based clusters patterns. Nevertheless they, presumably, all belong to the same chromosome and stem from a much more functionally constrained region than the repeatome. Therefore, we chose plastome-based evidence as a base-line to compare other clusters against.

Consistency of the Reconstructed Similarity-Based Patterns

There is no consensus yet in the literature on how many repeatome cluster characters are sufficient to reliably recover phylogenetic trees with stable groups. Obviously this will depend on the number of terminals included, and on the distribution of homoplasy across a range of clusters considered (see *P. x salmoneum* in the supplementary material 11*B–D*, Supplementary Material online). Dodsworth et al. (2017) used 1,000 clusters as characters in their abundancy-based analyses of six terminals. Vitales et al. (2020) settled for 100 clusters for six to nine terminals, but they employed and devised the similarity-based analysis. We have analyzed 255 nuclear genomic clusters and constructed NJ trees based on the read similarities found in these clusters. We found for our 17 terminals that the overall pattern is driven by ~the first 100 clusters (in order of size) with minor additions from the 155 smaller clusters.

Superclusters

We found most SCs to be composed of only a few clusters, connected by few paired-end reads. Some, however are composed of the largest clusters found and one of these (SC1, see fig. 3*A*) provides useful information on *P.* section *Ciconium*-specific repeat dynamics. In contrast, SC3 (fig. 3*B*) is symplesiomorphic for "Core Ciconium" and is not informative with regards to resolving intrasectional *Ciconium* relationships. It could however be employed for repeatome-based phylogeny reconstruction at higher taxonomic levels. We have little information from other plant groups, but if patterns like we obtained in *Ciconium* recur in sections from other plant groups, SCs could open the way for both higher level phylogeny reconstruction (using evolutionary "old" repeats), as well as repeatome-based phylogeny reconstruction at low taxonomic level (using "young" repeats).

Materials and Methods

Plants were grown from seed, in a climate-controlled greenhouse for 10 months after which leaf material was collected

for DNA extraction (see table 1). Current taxonomic opinion on *Pelargonium* sect. *Ciconium* (Roeschenbleck et al. 2014) recognizes 17 species. Of these, we included 15 in our study plus *P. karooicum* (section *Subsucculentia*) as an outgroup. For one species (*P. alchemilloides*), two accessions were included, as both ploidy levels and morphology were different (table 1 and Gibby et al. 1990), as well as their separate phylogenetic (polyphyletic) placement based on plastome exon sequence comparisons (van de Kerke et al. 2019). Furthermore, we included a soon to be recognized species from the Arabian Peninsula (*P. yemenense*, Gibby M et al., in preparation). Finally, we included *P. salmoneum*, whose species status is uncertain as it may be a hybrid (Breman FC et al., personal observations).

Genomic DNA was extracted from leaf material using the modified CTAB protocol described by Bakker et al. (1998), now including RNAse treatment (RQ1 Promega), followed by cleaning on a silica column (Nucleo Spin Machery Nagel). DNA extracts were sent to Novogene Inc. (Cambridge and Hong Kong) for Illumina HiSeq sequencing (0.5–1× coverage). Read libraries were generated from 1.0 μg genomic DNA using NEBNext DNA Library Prep Kit following the manufacturer's protocols, with genomic DNA randomly fragmented by shearing to ~350 bp. Fragments were subsequently subjected to end polishing, A-tailing, and ligation to the NEBNext adapter for Illumina HiSeq sequencing.

Flow Cytometry

Average total genomic content per cell (2C value expressed in pg) was determined using flow cytometry (Iribov SBW, the Netherlands) for all 19 accessions. As a reference for the size estimates, we used *P. x hortorum* PEZ-BD8517 with known ploidy (2×) and total genome size (2C = 2.33 pg, see table 1).

Multispecies Comparative Analysis

In order to perform a comparative analysis of genomic repeats from all 19 accessions, we combined a random subsample of reads from each of the Illumina read libraries into a combined multispecies read library (MSL), see below. Clustering of Illumina read pairs from the MSL was performed using the Repeat Explorer 2 (RE2, Novák et al. 2010, 2013) pipeline (version 2.3.7), implemented in the Galaxy server environment, using default settings (i.e., a minimum of 90% similarity over 55% of the read length will build a cluster). A cluster is a visual representation, using de Bruijn graphs, of relationships and overlap between reads. In these graphs, single reads are "nodes" and sequence overlaps (or relations) are "edges." In this way, a cluster allows for the visualization of read differences as well as relations between reads. The addition of reads to a cluster is stopped when no more reads match the abovementioned criterium of similarity. Repeat explorer also creates so-called "superclusters." These are constructed using information from the paired-end reads. When one read from a pair ends up in one cluster and another in another, clusters can be connected by virtue of the fact that they belong to a pair. These can be useful for inferring broader connections and relations within the genomes of the samples. For more details on the clustering process, see Novák et al. (2010, 2013).

As indicated above, we used read-subsampling for our MSL as implemented in Repeat Explorer 2's default settings for filtering out poor-quality reads. We then set the number of read pairs to be sampled for each accession to correspond to the 1C value (half the 2C value) for each accession, to obtain a set of read pairs corrected for genome size (see table 1). As a practical value, we used 100,000 read pairs per 1 pg of genomic DNA (table 1) which amounts to ~1.5% genomic coverage. This is slightly more than the 1% shown by Dodsworth et al. (2015) to be sufficient to confidently recover read clusters and their abundancies.

In *Pelargonium* section, *Ciconium* some species are polyploid, that is, *P. alchemilloides* ($2n = 2\times, 4\times, 6\times, 8\times$; Gibby and Westfold 1986), *P. articulatum* ($2n = 4\times$, this study) and *P. yemenense* sp. nov. ($2n = 4\times$, this study) (Gibby and Westfold 1986; Gibby et al. 1990). We did not reduce the number of read pairs selected from these polyploids in the manner outlined above, because we wanted to capture their genome dynamics postpolyploidization, which can be profound (reviewed in Wendel [2015]). It was shown for *Nicotiana* that postpolyploidization genomic variation can be captured by sampling the full size of the polyploid genome instead of reducing to the diploid level and that it has an impact on the reconstruction of phylogenetic relationships in that it gives insight into evolution postpolyploidization that would otherwise be missed (Dodsworth et al. 2015).

Comparative Analysis

To visualize the repeat content unique to section *Ciconium* and/or specific accessions, we used a cumulative, stacked histogram of read abundancies per cluster per accession (fig. 1). We optimized occurrence of these clusters on the abundancy-based tree for *Ciconium* (see below and fig. 2).

Phylogenetic Tree Reconstructions

Abundancies for selected clusters were recorded as counts per accession per cluster. Clusters containing ≥542 reads (or 0.01% of the genome) were retained for downstream analysis. For phylogeny reconstruction, we followed the approach of Dodsworth et al. (2017) and Vitales et al. (2020) who use cluster abundancy- and cluster similarity-based analysis, respectively, and outlined below.

To perform a character-based analysis of the cluster data, treating each cluster as one continuously distributed character, we used Tree analysis using New Technology (TNT, version 1.5; Goloboff and Catalano 2016). Cluster data were arranged into an "accession × cluster" matrix with cluster

abundancy as a continuous character state. The matrix was then cube-root transformed in order to reduce the effect of large abundancy differences and converted to TNT format using Mesquite (Maddison and Maddison 2019). The continuous range of character states was then "binned" by TNT into equally sized slices, the optimal amount of which was determined based on the distribution of reads across the cluster. In our case, TNT assigned 64 classes to the matrix, the maximal value for continuous character states (Goloboff et al. 2006, 2008). Tree inference and bootstrap resampling were performed using default "traditional search" settings in TNT. Resampling analysis involved "standard" bootstrapping with 1,000 replicates (fig. 2).

For the similarity-based approach to tree building, of the same clusters, we followed Vitales et al. (2020) with the addition that we also included clusters to which not all accessions contribute reads, referred to here as "incomplete clusters." In the case of such incomplete clusters a tree based on a NJ analysis yields a polytomy or zero-distance branch for those accessions that do not contribute reads to a comparison. However, for summarizing the NJ trees in a CN this is not a problem as here only splits (not branch lengths) are used.

We inferred pairwise sequence similarities of observed/expected frequencies of reads between the clusters and produced a distance matrix by inverting the values in the similarity matrix. This provided a measure of relatedness based on an all-to-all read comparison per cluster. For every cluster, the distance matrix is then converted into NJ trees using the R-package APE (Paradis and Schliep 2019). NJ tree topologies were then summarized in a consensus network (similarity-based CN) using SplitsTree v 4.14.6 (Huson and Bryant 2006) deploying split conflict thresholds of 5% (figs. 4 and 6 and supplementary material 2, Supplementary Material online).

Ciconium-Specific Patterns

To assess the stability of the inferred relationships based on our repeatome data, we assessed congruence between the similarity-based CN and the abundancy-based cladogram. We further compared the obtained cladogram and CN to test whether there are unique or synapomorphic groups of repeats or clusters. We also compared our repeatome-based patterns with those in previously published, mainly plastome-based, phylogenetic trees to assess possible incongruencies (see below).

To study the possible expansion or shrinking of repeat clusters over evolutionary time, that is, "cluster-demographics," we use a SC approach (supplementary material 5, Supplementary Material online). A SC is a "cluster of clusters" connected by edges based on reads from read pairs that occur in different clusters. Because these read pairs were the actual paired-end reads generated in the Illumina sequencing, they support the connection of these clusters (Novák et al. 2010, 2013). We identified SC1 which comprised eight out of the ten largest clusters from the entire repeatome analysis. A SC is therefore a useful object to evaluate the changes in contributions of all accessions included. Moreover, when you know the age of accessions, or can place accessions in a phylogenetic framework, a temporal context may be added as well. Changes of contributions per accession may then indicate expansion of reduction of a given repeat in one or more accessions.

Comparison with Previously Published Phylogenetic Trees for *Pelargonium* Sect. *Ciconium*

To be able to efficiently discuss the current and past phylogenetic hypotheses, we have summarized all available published phylogenetic hypotheses for *P.* sect. *Ciconium* (James et al. 2004 [their figures 1 and 5]; Jones et al. 2009 [which is a Bayesian version of Bakker et al. 2004]; Roeschenbleck et al. 2014; van de Kerke et al. 2019). These studies are mostly plastome-based, each differing in their taxonomic sampling, therefore making supertree analysis a better option over consensus tree analysis. "Best trees" from each of these studies, that is, parsimony consensus trees, best ML trees or Bayesian consensus trees, were collected and decomposed into a MRP (matrix representation using parsimony) matrix in PAUP* with subsequent concatenation and parsimony reconstruction of the resulting super-MRP. The resulting set of equally most parsimonious resolutions were then summarized (supplementary material 1, Supplementary Material online) in a Consensus Network (plastome-CN) using SplitsTree v 4.14.6.

Annotation of Repeat Clusters in *Pelargonium* Section *Ciconium*

Annotation of nuclear genomic repeat clusters in RE2 is based on existing hierarchical classifications of repeat classes (Wicker et al. 2007; Jurka et al. 2011, 2012; Llorens et al. 2011), but given the fast evolution and the limited knowledge of repeat classes across the plant kingdom many of our clusters may be *Pelargonium*-specific and could not be annotated. Therefore, we assigned names using the cluster numbering as they were assigned in the RE2 analysis, based on abundancy.

Contrasting Abundance- and Similarity-Based Patterns

For the purpose of discussing and describing the repeat profiles in a phylogenetic context, we defined "generic" or "symplesiomorphic" clusters shared by all accessions. In contrast, "Core *Ciconium*" synapomorphic clusters (occurring in *Ciconium* excluding *P. elongatum*) were defined as containing two or more species with each contributing at least 20% reads, and autapomorphic clusters (clusters containing $\geq$20% of the reads from one accession (for the full list, see supplementary material 11, Supplementary Material online). We subsequently plotted the nonsymplesiomorphic clusters

over the abundancy-based tree (fig. 2). Clades found in the abundancy-based tree were compared with the similarity-based CN (fig. 4) in order to determine conflicts between both approaches.

We performed a PCA to test if different trends of abundancies and sequence similarities are present in different repeat clusters. We took abundancy differences and sequence similarities as variables and performed a PCA on a selection of clusters. We used the three largest cp-based clusters, representing clusters with a comparable genomic location, and therefore presumably subject to comparable trends (#23, #44 and #45) and six selected clusters (#1, #4, #11, #16, #21, #73) that represent contrasting phylogenetic signals (fig. 5*A*). We selected clusters: symplesiomorphic (#4), synapomorphic for section *Ciconium* (#21), synapomorphic for Core *Ciconium* (#1), synapomorphic for a clade within section *Ciconium* (#16) and two that were autapomorphic (#16 and #73). We refer to figure 1 for the legend and to supplementary material 11, Supplementary Material online, for the cluster characterization. These clusters can also be found together with the clusters plotted over the abundancy-based phylogeny (fig. 2). All PCA analyses were carried out in R studio v.1.3.1073 using the libraries: "FactoMineR" (Le et al. 2008) and "factoextra v.1.0.7" (Kassambara and Mundt 2020). Plots were visualized using the ggplot2 package (Wickham 2016).

Pattern Reproducibility

Repeatome-based phylogenetic reconstruction is claimed to be congruent with other methods of phylogeny reconstruction, efficient, and reproducible (Dodsworth et al. 2015; Vitales et al. 2020). To explore the reproducibility and stability of the phylogenetic trees in relation to the percentage of genome representation that they are based on, we repeated the RE2 analyses using multispecies-sublibraries with (sub)-sample sizes of 10%, 25%, and 50% of the reads from the read libraries respectively (see supplementary material 4, Supplementary Material online). We reconstructed CNs with edge weights: mean; with threshold 0.1, of the NJ trees based on the top 50 clusters for each multispecies-sublibrary (fig. 6). We also compared trees obtained from overall repeat abundancies for each multispecies-sublibraries using the top 50 clusters (supplementary material 3, Supplementary Material online).

Data Consistency Analysis

We attempted to ascertain if the results of groups recovered in the NJ trees (summarized in our consensus network) would also occur with reduced cluster-character-sampling. To test this, we broke the data set into five groups of 50 trees, sorted by size, and constructed consensus networks cumulatively for the first 50, 100, 150, 200, and 250 characters using the same approach as described above for the reconstruction of the CN for the 255 trees (fig. 7*A*–*E*). We also created CNs for each slice of 50 trees and these are presented in the supplementary material 11, Supplementary Material online.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Petr Novák and Jirí Macas for help with the RE server. We thank Daniel Vitales for advice on the data analysis. We also thank an anonymous reviewer for a critical review of this study. This research was funded by the Dutch Foundations for applied scientific research (TTW). "*Pelargonium* genomics for overcoming cytonuclear incompatibility and bridging species barriers" (Grant No. 14531) of the Green Genetics program. R.C.S. is employed by Syngenta Seeds B.V., The Netherlands. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Conceived the study: F.C.B. and F.T.B. Carried out the analysis: F.C.B. Wrote the manuscript: F.C.B., F.T.B., and M.E.S. Informatics analyses: F.C.B. and G.C. All authors read the draft and gave feedback. We thank one anonymous reviewer for helpful and constructive feedback on the manuscript.

Data Availability

Illumina sequences have been uploaded to the sequence read archive at: https://dataview.ncbi.nlm.nih.gov/object/PRJNA782426?reviewer=780pogot3uofa1lgudvs06f2r5.

Literature Cited

Bakker FT, Culham A, Daugherty LC, Gibby M. 1999. A *trnL-F* based phylogeny for species of *Pelargonium* (Geraniaceae) with small chromosomes. Plant Syst Evol. 216(3–4):309–324.

Bakker FT, Culham A, Gibby M. 1999. Phylogenetics and diversification in *Pelargonium*. In: Hollingsworth P, Bateman R, Gornall R, editors. Molecular systematics and plant evolution. London: Chapman and Hall. p. 353–374.

Bakker FT, Culham A, Hettiarachi P, Touloumenidou T, Gibby M. 2004. Phylogeny of *Pelargonium* (Geraniaceae) based on DNA sequences from three genomes. Taxon 53(1):17–28.

Bakker FT, Culham A, Marais EM, Gibby M. 2005. Nested radiation in cape *Pelargonium*. In: Bakker FT, Chatrou LW, editors. Plant species-level systematics: new perspectives on pattern and process. Koenigstein: Koeltz Scientific Books. p. 75–100.

Bakker FT, Culham A, Pankhurst CE, Gibby M. 2000. Mitochondrial and chloroplast DNA based phylogeny of *Pelargonium* (Geraniaceae). Am J Bot. 87(5):727–734.

Bakker FT, Hellbrügge D, Culham A, Gibby M. 1998. Phylogenetic relationships within *Pelargonium* sect. *Peristera* (Geraniaceae) inferred from nrDNA and cpDNA sequence comparisons. Plant Syst Evol. 211(3–4):273–287.
Britten RJ, Graham DE, Neufeld BR. 1974. Analysis of repeating DNA sequences by reassociation methods. Enzymology 29:363–418.
Buggs RJA, et al. 2012. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. Curr Biol. 22(3):248–252.
Craig NL, et al. 2015. Mobile DNA III. 3rd ed. Washington: American Society for Microbiology.
Deniz O, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications. Nat Rev Genet. 20(7):417–431.
Dodsworth S, et al. 2015. Genomic repeat abundances contain phylogenetic signal. Syst Biol. 64(1):1112–1126.
Dodsworth S, et al. 2017. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). Plant Syst Evol. 303:1013–1020.
Eckardt NA. 2001. A sense of self: the role of DNA sequence elimination in allopolyploidization. Plant Cell 13(8):1699–1704.
Elliott TA, Gregory TR. 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc B. 370 (1678):20140331.
Enriquez-Gasca R, Gould PA, Rowe HM. 2020. Host gene regulation by transposable elements: the new, the old and the ugly. Viruses 12(10):1089.
Flavell RB, Bennett MD, Smith JB, Smith DB. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet. 12(4):257–269.
Freeling M, et al. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr Opin Plant Biol. 15(2):131–139.
Galinsky KJ, et al. 2016. Fast principal-component analysis reveals convergent evolution of *ADH*1B in Europe and East Asia. Am J Hum Genet. 98(3):456–472.
Gibby M, Albers F, Prinsloo B. 1990. Karyologieal studies in *Pelargonium* sectt. *Ciconium*, *Dibrachya*, and *Jenkinsonia* (Geraniaceae). Plant Syst Evol. 170(3–4):151–159.
Gibby M, Hinnah S, Marais EM, Albers F. 1996. Cytological variation and evolution within *Pelargonium* Section *Hoarea* (Geraniaceae). Plant Syst Evol. 203(1–2):111–114.
Gibby M, Westfold J. 1986. A cytological study of *Pelargonium* sect. Plant Syst Evol. 153(3–4):205–222.
Goloboff P, Catalano S. 2016. TNT version 1.5 with a full implementation of phylogenetic morphometrics. Cladistics 32(3):221–238.
Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylocladistics. Cladistic 24(5):774–786.
Goloboff PA, Mattoni CI, Quinteros AS. 2006. Continuous characters analyzed as such. Cladistics 22(6):589–601.
Greiner S, et al. 2008. The complete nucleotide sequences of the 5 genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: II. A microevolutionary view using bioinformatics and formal genetic data. Mol Biol Evol. 25(9):2019–2030.
Hanada K, et al. 2009. The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile. Plant Cell 21(1):25–38.
Hartley G, O'Neill RJ. 2019. Centromere repeats: hidden gems of the genome. Genes 10:223.
Holland BR, Huber KT, Moulton V, Lockhart PJ. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol Biol Evol. 21(7):1459–1461.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2):254–267.
James CM, Gibby M, Barrett JA. 2004. Molecular studies in Pelargonium (Geraniaceae). A taxonomic appraisal of section Ciconium and the origin of the "Zonal" and "Ivy-leaved" cultivars. Plant Syst Evol. 243(3-4):131–146.
Jiang M, et al. 2018. Sequencing, characterization, and comparative analyses of the plastome of *Caragana rosea* var. *rosea*. Int J Mol Sci. 19(5):1419.
Johnston JS, et al. 2005. Evolution of genome size in Brassicaceae. Ann Bot. 95(1):229–235.
Jones C, Bakker FT, Schlichting CD, Nicotra AB. 2009. Leaf shape evolution in the South African genus *Pelargonium* l'Hér. (Geraniaceae). Evolution 63(2):479–497.
Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure and the origin of species. Biol Direct. 6:44.
Jurka J, Bao W, Kojima KK, Kohany O, Yurka MG. 2012. Distinct groups of repetitive families preserved in mammals correspond to different periods of regulatory innovations in vertebrates. Biol Direct. 7:36.
Kassambara A, Mundt F. 2020. factoextra: extract and visualize the results of multivariate data analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.
Koo D-H, et al. 2011. Rapid divergence of repetitive DNAs in *Brassica* relatives. Genomics 97(3):173–185.
Landis JB, et al. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. Am J Bot. 105(3):348–363.
Le S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. J Stat Softw. 25:1–18.
Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid plants. Science 320(5875):481–483.
Leitch IJ, Beaulieu JM, Chase MW, Leitch AR, Fay MF. 2010. Genome size dynamics and evolution in monocots. J Bot. 2010:1–18.
Leitch IJ, Johnston E, Pellicer J, Hidalgo O, Bennett MD. 2019. Plant DNA C-values database (release 7.1, April 2019). Available from: https://cvalues.science.kew.org/.
Linder HP. 2003. The radiation of the Cape flora, southern Africa. Biol Rev Camb Philos Soc. 78(4):597–638.
Linder R, Meadows ME, Cowling RM. 1992. History of the Cape flora. In: Cowling RM, editor. The ecology of fynbos. Cape Town (South Africa): Oxford University Press.
Lis-Balchin M. 1997. A chemotaxonomic study of the *Pelargonium* (Geraniaceae) species and their modern cultivars. J Horticult Sci. 72(5):791–795.
Lis-Balchin MT. 1996. A chemotaxonomic reappraisal of the section *Ciconium Pelargonium* (Geraniaceae). S Afr J Bot. doi:10.1016/S0254-6299(15)30657-8.
Llorens C, et al. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. 39:70–74.
Louzada S, et al. 2020. Decoding the role of satellite DNA in genome architecture and plasticity—an evolutionary and clinical affair. Genes 11(1):72.
Maddison WP, Maddison DR. 2019. Mesquite: a modular system for evolutionary analysis. Version 3.61. Available from: http://www.mesquiteproject.org.
Marais EM. 1994. Taxonomic studies in *Pelargonium* section *Hoarea* (Geraniaceae) [dissertation]. South Africa: University of Stellenbosch.
Marais EM. 2005. Differences between *Pelargonium moniliforme* (Geraniaceae) and the closely related *P. vinaceum*. S Afr J Bot. 71(2):221–227.
Marais EM. 2014. One name change and three new species of *Pelargonium* section *Hoarea* (Geraniaceae) from Western Cape Province. S Afr J Bot. 90:118–127.
Maumus F, Quesneville H. 2014. Deep investigation of *Arabidopsis thaliana* junk DNA reveals a continuum between repetitive elements and genomic dark matter. PLoS One 9(4):e94101.
Miller AJ, et al. 2013. *Vitis* phylogenomics: hybridization intensities from a SNP array outperform genotype calls. PLoS One 8(11):e78680.

Morgante M, De Paoli E, Radovic S. 2007. Transposable elements and the plant pan-genomes. Curr Opin Plant Biol. 10(2):149–155.
Novák P, et al. 2020. Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat Plants. 6(11):1325–1329.
Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378.
Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29(6):792–793.
Oliver KR, McComb JA, Greene WK. 2013. Transposable elements: powerful contributors to angiosperm evolution and diversity. Genome Biol Evol. 5(10):1886–1901.
Ollitrault P, et al. 2012. SNP mining in *C. clementina* BAC end sequences; transferability in the *Citrus* genus (Rutaceae), phylogenetic inferences and perspectives for genetic mapping. BMC Genomics 13(13):13.
Paço A, Freitas R, Vieira-da-Silva A. 2019. Conversion of DNA sequences: from a transposable element to a tandem repeat or to a gene. Genes 10(12):1014.
Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35(3):526–528.
Puttick MN, Clark J, Donoghue PCJ. 2015. Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms. Proc Biol Sci. 282(1820):20152289.
Ren R, et al. 2018. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol Plant. 11(3):414–428.
Roeschenbleck J, Albers F, Müller K, Weinl S, Kudla J. 2014. Phylogenetics, character evolution and a subgeneric revision of the genus *Pelargonium* (Geraniaceae). Phytotaxa 159(2):31–76.
Ruhlman TA, Jansen RK. 2018. Aberration or analogy? The atypical plastomes of Geraniaceae. In: Chaw S-M, Jansen RK, editors. Advances in botanical research 85: plastid genome evolution. Amsterdam (Netherlands): Elsevier. p. 223–262.
Salim D, Gerton JL. 2019. Ribosomal DNA instability and genome adaptability. Chromosome Res. 27(1–2):73–87.
Sankoff D, Zhang C, Zhu Q. 2010. The collapse of gene complement following whole genome duplication. BMC Genomics 11:313.
Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. Curr Opin Plant Biol. 15(2):147–153.
Sloan DB, et al. 2014. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Mol Phylogenet Evol. 72:82–89.
Soltis DE, et al. 2009. Polyploidy and angiosperm diversification. Am J Bot. 96(1):336–348.
Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. Curr Opin Genet Dev. 35:119–125.
van de Kerke SJ. 2019. Exploring floral evolution in *Pelargonium* (Geraniaceae) linking shapes and macro-evolution [PhD thesis]. Wageningen (Netherlands): Wageningen UR.
van de Kerke SJ, et al. 2019. Plastome based phylogenetics and younger crown node age in *Pelargonium*. Mol Phylogenet Evol. 137:33–43.
van der Walt JJA, Vorster PJ. 1988. Pelargoniums of Southern Africa. Volumes I-III. Ann Kirtenbosch Bot Gardens.
Verboom GA, et al. 2009. Origin and diversification of the Greater Cape flora: ancient species repository, hot-bed of recent radiation, or both? Mol Phylogenet Evol. 51(1):44–53.
Vitales D, Fernández P, Garnatje T, Garcia S. 2019. Progress in the study of genome size evolution in Asteraceae: analysis of the last update. Database 2019:baz098.
Vitales D, Garcia S, Dodsworth S. 2020. Reconstructing phylogenetic relationships based on repeat sequence similarities. Mol Phylogenet Evol. 147:106766.
Warren BH, et al. 2011. Consistent phenological shifts in the making of a biodiversity hotspot: the Cape flora. BMC Evol Biol. 11:39.
Weiss-Schneeweiss H, Leitch AR, McCann J, Jang T-S, Macas J. 2015. Employing next generation sequencing to explore the repeat landscape of the plant genome. In: Hörandl E, Appelhans MS, editors. Next-generation sequencing in plant systematics. Chapter 5. Bratislava, Slovakia: International Association for Plant Taxonomy (IAPT). doi:10.14630/000006.
Wendel JF. 2015. The wondrous cycles of polyploidy in plants. Am J Bot. 102(11):1753–1756.
Weng ML, Blazier JC, Govindu M, Jansen RK. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Biol Evol. 31(3):645–659.
Weng ML, Ruhlman TA, Gibby M, Jansen RK. 2012. Phylogeny, rate variation, and genome size evolution in *Pelargonium* (Geraniaceae). Mol Phylogenet Evol. 64(3):654–670.
Weng M-L, Tracey A, Ruhlman TA, Jansen RK. 2017. Expansion of inverted repeat does not decrease substitution rates in *Pelargonium* plastid genomes. New Phytol. 214(2):842–851.
Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 76(3–5):273–297.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8(12):973–982.
Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag. Available from: https://ggplot2.tidyverse.org.
Yang H, et al. 2016. Genetic divergence between *Camellia sinensis* and its wild relatives revealed via genome-wide SNPs from RAD sequencing. PLoS One 11(3):e0151424.

Associate editor: Yves Van De Peer