*Syst. Biol.* 64(1):84–101, 2015

DOI:10.1093/sysbio/syu071
Advance Access publication October 3, 2014

# From Gene Trees to a Dated Allopolyploid Network: Insights from the Angiosperm Genus *Viola* (Violaceae)

THOMAS MARCUSSEN[1,2], LISE HEIER[1], ANNE K. BRYSTING[1], BENGT OXELMAN[2], AND KJETILL S. JAKOBSEN[1,*]

[1]*Department of Biosciences, Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, PO Box 1066 Blindern, NO-0316 Oslo, Norway and* [2]*Department of Biological and Environmental Sciences, University of Gothenburg, PO Box 461, 405 30 Gothenburg, Sweden*
**Correspondence to be sent to: Department of Biosciences, Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, PO Box 1066 Blindern, NO-0316; E-mail: k.s.jakobsen@ibv.uio.no.*

*Received 13 December 2012; reviews returned 5 September 2014; accepted 8 September 2014*
*Associate Editor: Mark Fishbein*

*Abstract*.—Allopolyploidization accounts for a significant fraction of speciation events in many eukaryotic lineages. However, existing phylogenetic and dating methods require tree-like topologies and are unable to handle the network-like phylogenetic relationships of lineages containing allopolyploids. No explicit framework has so far been established for evaluating competing network topologies, and few attempts have been made to date phylogenetic networks. We used a four-step approach to generate a dated polyploid species network for the cosmopolitan angiosperm genus *Viola* L. (Violaceae Batch.). The genus contains ca 600 species and both recent (neo-) and more ancient (meso-) polyploid lineages distributed over 16 sections. First, we obtained DNA sequences of three low-copy nuclear genes and one chloroplast region, from 42 species representing all 16 sections. Second, we obtained fossil-calibrated chronograms for each nuclear gene marker. Third, we determined the most parsimonious multilabeled genome tree and its corresponding network, resolved at the section (not the species) level. Reconstructing the "correct" network for a set of polyploids depends on recovering all homoeologs, i.e., all subgenomes, in these polyploids. Assuming the presence of *Viola* subgenome lineages that were not detected by the nuclear gene phylogenies ("ghost subgenome lineages") significantly reduced the number of inferred polyploidization events. We identified the most parsimonious network topology from a set of five competing scenarios differing in the interpretation of homoeolog extinctions and lineage sorting, based on (i) fewest possible ghost subgenome lineages, (ii) fewest possible polyploidization events, and (iii) least possible deviation from expected ploidy as inferred from available chromosome counts of the involved polyploid taxa. Finally, we estimated the homoploid and polyploid speciation times of the most parsimonious network. Homoploid speciation times were estimated by coalescent analysis of gene tree node ages. Polyploid speciation times were estimated by comparing branch lengths and speciation rates of lineages with and without ploidy shifts. Our analyses recognize *Viola* as an old genus (crown age 31 Ma) whose evolutionary history has been profoundly affected by allopolyploidy. Between 16 and 21 allopolyploidizations are necessary to explain the diversification of the 16 major lineages (sections) of *Viola*, suggesting that allopolyploidy has accounted for a high percentage—between 67% and 88%—of the speciation events at this level. The theoretical and methodological approaches presented here for (i) constructing networks and (ii) dating speciation events within a network, have general applicability for phylogenetic studies of groups where allopolyploidization has occurred. They make explicit use of a hitherto underexplored source of ploidy information from chromosome counts to help resolve phylogenetic cases where incomplete sequence data hampers network inference. Importantly, the coalescent-based method used herein circumvents the assumption of tree-like evolution required by most techniques for dating speciation events. [Dating; low-copy nuclear gene; polyploidy; species network; *Viola*; violaceae.]

Polyploidization, implying whole-genome duplication (WGD), is widely recognized as an important mechanism of speciation and evolution across eukaryotes (reviewed in Levin 2002; Gregory and Mable 2005; Tate et al. 2005; Wendel and Doyle 2005; Soltis et al. 2014). Polyploidy is particularly common in plants, and has been estimated to account for 15%–30% of the speciation events in angiosperms (Wood et al. 2009; Mayrose et al. 2011), and has been associated with major radiations (e.g., Fawcett et al. 2009; Jiao et al. 2011). Genome duplication is generally irreversible in the short term, and evidence of polyploidy may remain in the genome for hundreds of millions of years. Phylogenomic analyses suggest polyploidization events in the early history of seed plants and angiosperms, and in all of the major angiosperm lineages (Cui et al. 2006; Jaillon et al. 2007; Tang et al. 2010; Jiao et al. 2011). In animals there have been two polyploidizations early in the vertebrate lineage (Dehal and Boore 2005), and there are nearly 200 examples of polyploidy in insects and vertebrates (Otto 2007) and many more in other invertebrate groups (Gregory and Mable 2005). Polyploidy is often classified based on the relatedness of the duplicated genomes, where autopolyploidy implies that the duplicated genomes are identical (homologuous) or nearly so and stem from the same species, and allopolyploidy refers to conditions where the duplicated genomes are nonidentical (homoeologous) and have been brought together within a single organism by interspecific hybridization.

Young polyploids (neopolyploids) are as a rule identifiable by increased genome size, doubled chromosome numbers, and redundant gene content compared with their progenitors. Many neopolyploids are of postglacial origin, and deduction of polyploid relationships in such cases is often claimed to be intuitive and not requiring phylogenetics (e.g., Müntzing 1932; Ownbey 1950; Hedrén 1996). Over time, usually millions of years, the signatures of polyploidy will be eroded away by a suite of molecular so-called diploidization

mechanisms (Wolfe 2001; Leitch and Bennett 2004; Ma and Gustafson 2005) that will ultimately lead to the return of the lineage to an apparent "diploid" state. This includes the return of many genes to single copy, disomic chromosomal inheritance, and often reduced genome size and chromosome numbers. Thus, the deeper the phylogeny, the more obscure the identification of polyploidy events and the greater the requirement for evidence from numerous genes. By definition, mesopolyploid species are older polyploids whose parental subgenomes are only discernible by comparative (cyto)genetic and phylogenetic methods, and paleopolyploids are even older polyploids in which WGD events can only be uncovered by comparison of orthologous sequences (Mandáková et al. 2010).

In molecular systematics, species phylogenies are often estimated from individual gene phylogenies. Gene trees are contained within the species tree, and they may differ from it in relative branch lengths and topology. This owes to a number of factors at play at the gene level, including coalescent stochasticity (of alleles of orthologs), introgression (of xenologs), or duplication and loss of individual genes (paralogs) (Doyle 1992; Maddison 1997; Nichols 2001; Marcussen et al. 2014). The nodes in a gene tree are allele splits that may not directly reflect the speciation events that we wish to infer; speciation events are in fact more likely to happen along the branches *between* nodes rather than *at* nodes in the gene tree. Homoploid speciation times can be estimated from a set of allele split times using coalescent theory (Kingman 1982; Yang and Rannala 2003; Degnan and Rosenberg 2009). This theory makes the assumption that, for a genetic diploid, the difference in age of a species split and its associated allele splits is exponentially distributed with a rate parameter that is tied to the effective population size, and further assumes no selection or introgression. However, theoretical and methodological developments to estimate polyploid speciation times from allele splits are in their infancy (Bartoszek et al. 2013; Jones et al. 2013), and restricted to simple scenarios.

Species phylogenies involving allopolyploids are by definition networks, not trees, although the gene phylogenies are always treelike, except in cases of recombination. A polyploid inherits gene copies (homoeologs) from both its parental lineages, and the number of homoeologs is accordingly expected to directly reflect ploidy. Therefore, gene trees and genome trees for polyploids (i.e., the equivalent of a species tree for diploid taxa) are characterized by having more than one leaf labeled by the same taxon and are referred to as multilabeled trees (Huber et al. 2006). A universal method has been devised that estimates the species network with the fewest hybridizations from a set of multilabeled gene trees (Huber et al. 2006). However, successfully finding the "correct" network depends on whether the available set of multilabeled gene trees recovers all homoeologs, representing all subgenomes, in the polyploid. Incomplete sampling of homoeologs is especially likely in more ancient (meso- and paleo-) polyploids where redundant gene copies may have become extinct. In theory, this situation can be overcome by explicitly modeling alternative scenarios of putatively extinct homoeologs. Parsimony can then be used to identify a shortest network as the one that contains the fewest number of putative homoeolog extinction and polyploidization events (Marcussen et al. 2012). When ploidy levels of the study species are known or can be estimated, this can give independent information on how many homoeologs to expect in each taxon.

The large cosmopolitan genus *Viola* (Table 1) is an attractive candidate genus for studies of polyploid evolution. Ploidy among extant *Viola* lineages and species ranges from diploid ($2x$) to at least octadecaploid ($18x$), and the genus contains both young (neo-) and older (meso-) polyploids (Marcussen and Nordal 1998; van den Hof et al. 2008; Marcussen et al. 2011; 2012). The 580 to 620 species of *Viola* (Wahlert et al. 2014) are distributed among 16 morphologically, chromosomally, and geographically defined groups that we treat as tentative sections (Table 1). Only 3 out of the 16 sections have chromosome base numbers consistent with diploidy, i.e., sect. *Chamaemelanium* and sect. *Rubellium* with $x = 6$ and sect. *Andinium* with $x = 7$ (Table 1), but without phylogenetic or rooting information there can be no certainty which of these numbers is ancestral in *Viola*. Evidence from biogeography, karyology, and phylogenies suggests that *Viola* originated in South America and subsequently spread to the Northern Hemisphere and elsewhere (Clausen 1929; Ballard et al. 1999; Marcussen et al. 2012). Colonization of the northern hemisphere can be dated with the appearance of fossil seeds of *Viola* in Eurasian sediments 17–18 Ma ago (Kovar-Eder et al. 2001; Marcussen et al. 2012). Although a number of phylogenetic studies have been published on *Viola*, none has been comprehensive in terms of taxon sampling and exploring ancient reticulations at the genus level (Ballard et al. 1999; Ballard and Sytsma 2000; Yockteng et al. 2003; Yoo et al. 2005; van den Hof et al. 2008; Gong et al. 2010; Liang and Xing 2010; Yoo and Jang 2010; Marcussen et al. 2010, 2011, 2012; Nakamura et al. 2014).

In this study, we generated a comprehensive, evolutionary dated phylogenetic network for the 16 section lineages of *Viola*, using a four-step approach as outlined in Figure 1. (1) We first obtained sequence data with exhaustive sampling of homoeologs for three low-copy nuclear genes (*GPI*, *NRPD2a*, and *SDH*) and one chloroplast (*trnL-F*) region. (2) We then individually calibrated the nuclear gene phylogenies using primary (a >18 Ma old seed fossil) and secondary calibrations. (3) From the nuclear gene phylogenies we reconciled a set of five alternative multilabeled genome trees that differed in the interpretation of deep coalescence and gene loss for one deep node. To select among these alternative trees we used the following parsimony criteria: fewest gene losses, fewest polyploidizations, and minimum deviation from assumed ploidy levels as inferred from published chromosome numbers. (4) In order to calibrate the network, we used Bayesian modeling to estimate

TABLE 1. Subdivision of the genus *Viola* (Violaceae) into sections, showing number of species (altogether 583–620), distribution, chromosome numbers, base chromosome numbers and inferred ploidy

| Section | Species | Distribution | Chromosome numbers (haploid) | Estimated ploidy |
|---|---|---|---|---|
| Sect. *Andinium* W.Becker | 113 | S America | $n = \underline{7}^{e}$ | 2*x* |
| Sect. *Chamaemelanium* Ging. s.lat.[a] | 61 | N America, northeast Asia; *V. biflora* circumpolar | $n = \underline{6}$ 12 18 24 36 | 2*x* |
| Sect. *Chilenium* W.Becker | 8 | southern S America | – | 10*x* |
| Sect. *Delphiniopsis* W.Becker | 3 | western Eurasia: southern Spain; Balkans | $n = \underline{10}^{f}$ | 4*x* |
| Sect. *Erpetion* (Banks) W.Becker | 11–18 | eastern Australia; Tasmania | $n = \underline{30}$ $60^{g}$ | 10*x* |
| Sect. *Leptidium* Ging. | 19 | S America | $n = [\underline{14}]$ $27^{h}$ | 4*x* |
| Sect. *Melanium* Ging. | 125 | western Eurasia; *V. bicolor* in eastern N America | $n = 2$ 4–5 7–15 17–18 20–22 24–26 ~32 34 ~48 60 $\sim 64^{i}$ | 4*x* |
| Sect. *Nosphinium* W.Becker s.lat.[b] | 31–50 | N, C and northern S America; Beringia; Hawaii | $n = 27$ [28] 40 $51^{j}$ | 10*x* |
| Sect. nov. A (*V. abyssinica* group) | 1–3 | Africa: equatorial high mountains | $n \approx 36^{k}$ | 12*x* |
| Sect. nov. B (*V. spathulata* group) | 7–9 | western and central Asia: northern Iraq to Mongolia | – | 8*x* |
| Sect. *Plagiostigma* Godr.[c] | 120 | northern hemisphere; ca 5 spp. in Australasia | $n = 11$ 12 13 22 23 24 36 | 4*x* |
| Sect. *Rubellium* W.Becker | 3–6 | S America: Chile | $n = \underline{6}^{l}$ | 4*x* |
| Sect. *Sclerosium* W.Becker | 1–4 | northeastern Africa to southwestern Asia | $n = \underline{11}^{m}$ | 4*x* |
| Sect. *Tridens* W.Becker | 2 | southern S America | $n = \underline{40}^{n}$ | 12*x* |
| Sect. *Viola* s.str.[d] | 75 | northern hemisphere | $n = \underline{10}$ 20 29 30 | 4*x* |
| Sect. *Xylinosium* W.Becker | 3–4 | Mediterranean region; *V. decumbens* in South Africa | $n = \underline{26}^{o}$ | 8*x* |

Notes: The systematics is provisional, based on earlier treatments (Becker 1925; Clausen 1929; Brizicky 1961; Clausen 1964) and our own studies, published (Marcussen et al. 2010; 2011; 2012) and in progress. Known chromosome numbers (*n*) within each section are indicated, with species names if only a few species have been counted, and numbers interpreted to be base chromosome number (*x*) for each section are underlined. Inferred, but not observed base chromosome numbers are underlined and given within square brackets. No base number is inferrable for section *Melanium*, owing to dysploidy (but see Erben 1996; Yockteng et al. 2003).

[a]sensu Brizicky (1961), Clausen (1929, 1964) and Marcussen et al. (2012); i.e., including sect. *Dischidium* Ging. and grex *Orbiculares* Pollard.

[b]sensu Marcussen et al. (2012); i.e., including subsects. *Boreali-Americanae* (W.Becker) Gil-ad, *Langsdorffianae* (W.Becker) auct., *Mexicanae* (W.Becker) auct. and *Pedatae* (Pollard) auct.

[c]sensu Marcussen et al. (2012); i.e., excluding subsects. *Boreali-Americanae* (W.Becker) Gil-ad, *Langsdorffianae* (W.Becker) auct., *Mexicanae* (W.Becker) auct. and *Pedatae* (Pollard) auct.

[d]sensu Brizicky (1961), Clausen (1964; as sect. *Rostellatae* Boiss.) and Marcussen et al. (2012).

[e]*V. montagnei*, *V. roigii* (Sanso and Seo 2005).

[f]*V. delphinantha*, *V. cazorlensis* (Schmidt 1964; Leal Perez-chao et al. 1980; Diosdado et al. 1993).

[g]"*V. hederacea* complex" (Moore in Smith-White 1959).

[h]*V. dombeyana* (Heilborn 1926; as *V. humboldtii*).

[i]e.g., Yockteng et al. (2003).

[j]following interpretation in Marcussen et al. (2012).

[k]*V. abyssinica* (Morton 1993).

[l]*V. rubella* (Blaxland & Windham in Marcussen et al. 2012).

[m]*V. stocksii* (Khatoon and Ali 1993).

[n]*V. tridentata* (Moore 1967).

[o]*V. arborescens*, *V. saxifraga* (Arrigoni and Mori 1980; Galland 1985, 1988; Verlaque and Espeut 2007).

homoploid speciation times from allele splits in the three nuclear gene trees by use of a coalescent approach, and allopolyploid speciation times.

## MATERIALS AND METHODS

We have subdivided *Viola* into 16 sections that are well-defined phylogenetically, morphologically, chromosomally, and geographically (Table 1). A short account of the provisional taxonomy is given in Supplementary Appendix 1 (see online Appendix 1, http://dx.doi.org/10.5061/dryad.jc754). We sampled 42 species (Supplementary Table S1, http://dx.doi.org/10.5061/dryad.jc754) that represent all 16 tentative sections of *Viola* (Table 1). Where applicable, taxa were sampled so as to represent the morphological and taxonomic diversity of sections, and low ploidy was always preferred to simplify analyses and minimize cost. DNA was extracted using a CTAB extraction protocol (Doyle and Doyle 1987). In most cases, DNA working solutions were made by diluting extractions 1:20, of which 1 μL was used per PCR reaction. "Difficult" DNA preparations were further cleaned using the DNeasy Blood & Tissue Kit (Qiagen, Düsseldorf, Germany), following the manufacturer's guidelines except omitting the first two steps. We obtained sequences from the three low-copy nuclear genes *GPI* (glucose-6-phosphate isomerase; exons 12 to 18; ca 2000 bp), *NRPD2a* (the shorter of two paralogs encoding the second-largest subunit of plant RNA polymerase IV/V; exons 5 to 7; ca 1300 bp; Marcussen et al. 2010) and *SDH* (shikimate dehydrogenase; exons 5 to 10; ca 1200 bp), and from the chloroplast *trnL-F* region (*trnL* intron, *trnL* 3′ exon, and *trnL-trnF* spacer; ca 900 bp). *Allexis batangae* (*NRPD2a*), *A. cauliflora* (*trnL-F*), and *Noisettia orchidiflora* (*GPI*, *trnL-F*)

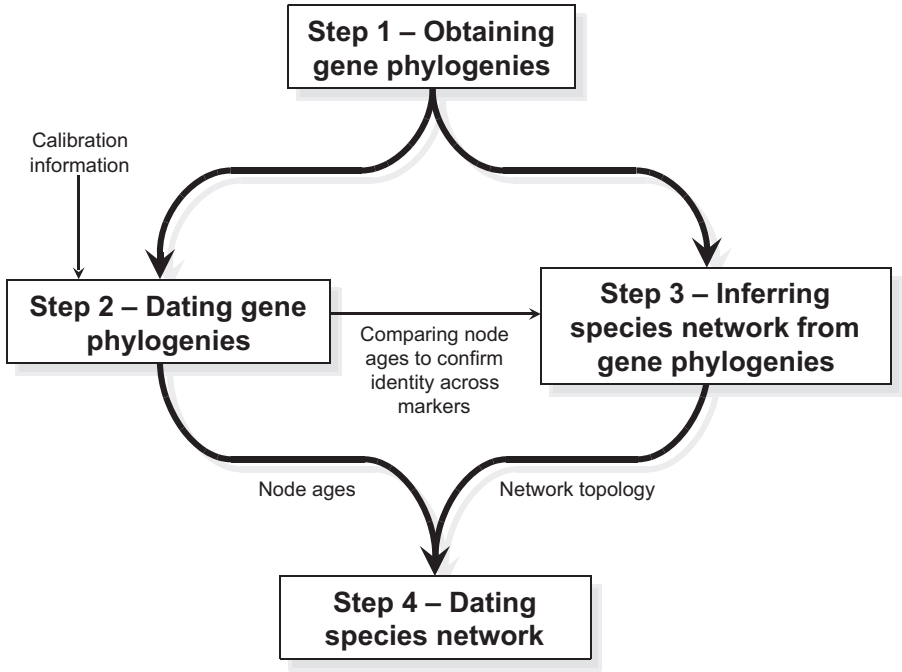


FIGURE 1. A flow diagram of the four-step approach used to estimate a dated network from individual gene trees, using *Viola* as an example. See main text for further details.

served as outgroups (Tokuoka 2008; Wahlert et al. 2014). For *SDH*, initial phylogenetic analysis that included exon sequences of *Ricinus communis* (XM002516810) and *Populus trichocarpa* (AC214213) identified paralogs in *V. congesta* and *V. tridentata* as out-paralogs with respect to *Viola*; these out-paralogs were subsequently used as outgroups in phylogenetic analysis of *SDH* for *Viola*.

We used a four-step approach (Fig. 1), detailed below, to estimate a dated network from individual gene trees. Step 1 entails obtaining sequence data and gene phylogenies from low-copy nuclear genes. It is crucial that all homoeologs are recovered for each gene, e.g., by subcloning and/or applying several PCR primer sets (Brysting et al. 2011; Scheen et al. 2012). Step 2 entails dating the gene phylogenies (Rutschmann 2006; Forest 2009; Parham et al. 2012; Sauquet et al. 2012). Since allele coalescence times may be deep and depart from the species phylogeny by several Ma in lineages with large effective population sizes, it is necessary to take into account the uncertainty associated with calibrating the *gene* phylogeny instead of the *species* phylogeny. Step 3 entails inferring a species network from gene phylogenies (Huber et al. 2006; Marcussen et al. 2012). The number of polyploidizations may be obscured by paralogs or loss of homoeologs, especially if few genes are considered over a large time scale. By obtaining additional information on ploidy, from e.g., chromosome counts, the most parsimonious species network can be recovered by scoring the total number of steps required for alternative scenarios of polyploidization, homoeolog loss, and deviation from expected ploidy. Step 4 entails dating the homoploid and polyploid speciations in the network. By compiling coalescence times for each gene for each node in the species network, speciation times can be calculated under the model of multispecies coalescent (Degnan and Rosenberg 2009; Doyle and Egan 2010; Bartoszek et al. 2013; Jones et al. 2013).

### *Step 1. Obtaining Multilabeled Gene Phylogenies*

Gene homologs in diploids were amplified by PCR using a single set of general primers. Gene homoeologs for higher-ploids were either *in vitro* cloned by single-molecule (sm) PCR (Marcussen et al. 2012), *in vivo* cloned (Marcussen et al. 2010), or amplified and sequenced using homoeolog-specific primers (Marcussen et al. 2012). PCR products were subject to Sanger sequencing. Further details about primer sequences, PCR protocols, and alignment procedures are given in Supplementary Appendix 2 see online Appendix 2, http://dx.doi.org/10.5061/dryad.jc754.

Phylogenies for *GPI*, *NRPD2a*, *SDH*, and *trnL-F* were reconstructed by Bayesian inference (BI) with MrBayes v3.2.0 (Ronquist et al. 2011). The alignments of the nuclear regions (*GPI*, *NRPD2a*, *SDH*) were each divided into three data partitions corresponding to exon, intron, and coded indels, whereas the *trnL-F* region was divided into two data partitions corresponding to nucleotides and coded indels. Nucleotide substitution models for each of the data partitions were proposed by Treefinder version of March 2008 (Jobb et al. 2004) based on the AICc model selection criterion. Introns of the three nuclear regions and the nucleotide partition of *trnL-F* were analyzed under the GTR + Γ model. Exons of the nuclear regions were analyzed under the HKY + Γ model.

All indel partitions were analyzed under the simple model for binary data implemented in MrBayes. The gamma distribution was simulated by four discrete rate categories. For each region five parallel MCMC chains were run to convergence (1 to 3 million generations), as indicated by the average standard deviation of split frequencies reaching below 0.01 and effective sample size (ESS) values reaching above 200, and the first 25% of the log and tree data was discarded as burn-in.

### *Step 2. Estimation of Divergence Times for the Gene Phylogenies*

We estimated divergence times for the three nuclear gene phylogenies separately (*GPI*, *NRPD2a*, and *SDH*), using one primary (fossil) and five secondary calibration points, using a Bayesian relaxed clock as implemented in BEAST 1.5.4 (Drummond et al. 2006; Drummond and Rambaut 2007). We applied the same substitution models as in the MrBayes analyses (see above). All dating analyses used a speciation model that followed a Yule tree prior, with rate variation across branches uncorrelated and lognormally distributed. Two MCMC chains were run for the *GPI* data, for 50 and 25 million generations and with the first 4 and 1 million generations discarded as burn-in, respectively. Two MCMC chains were run for the *NRPD2a* data, each for 30 million generations and with the first 3 million generations discarded as burn-in. Two MCMC chains were run for the *SDH* data, for 15 million generations each and with the first 1.5 and 3.5 million generations discarded as burn-in, respectively. For all analyses run in BEAST, parameters were sampled every 1000 steps. Burn-in was determined and discarded after visual inspection of each log file in Tracer v1.5 (Rambaut and Drummond 2009) and the two log files of each marker were controlled for similar convergence before they were combined. ESS values for all estimated parameters and node ages were above 200, as recommended.

The primary calibration was based on a 17–18 Ma old seed fossil from Austria (Kovar-Eder et al. 2001), which is the oldest (and most accurately dated) fossil out of several *Viola* seed morphotypes that more or less synchronously appeared in Lower Miocene sediments (Dorofeev 1963). This series of fossils has been interpreted as reflecting the colonization and rapid diversification of the genus in the northern hemisphere (Marcussen et al. 2012). The fossil was used to calibrate the most recent common ancestor (tmrca) of the Northern Hemisphere lineages which, owing to tetraploidy, corresponds to two nodes, the so-called "CHAM" crown node and the "MELVIO" crown node. The clades are named after ribotype (Marcussen et al. 2010): early analyses of rDNA (Ballard et al. 1999; Ballard and Sytsma 2000) revealed that the Northern Hemisphere species, irrespective of ploidy, had either a "CHAM" ribotype similar to sect. *Chamaemelanium* or a "MELVIO" ribotype similar to sects. *Melanium* and *Viola* (see Fig. 5 in Marcussen et al. 2012). Later it became clear that this pattern was a result of lineage sorting of ribotypes in the polyploids (Marcussen et al. 2010, 2011, 2012). We constrained the CHAM crown node and the MELVIO crown node to be at least 18 Ma, using identical lognormal priors with mean = standard deviation (SD) = 0.6 and offset = 18; the 95% highest probability density (HPD) of this distribution corresponds to a time interval of 5 Ma.

No appropriate fossils are available for reliable calibration of the rate-heterogeneous north-temperate lineages given our sampling scheme. Therefore, in order to fully work through our proposed procedure (Fig. 1) we applied secondary calibrations on internal nodes based on estimates from Marcussen et al. (2012) and on the basal nodes based on estimates from a multigene Violaceae/Malpighiales dataset (see below). Marcussen et al. (2012) applied fossil calibrations internally in sects. *Viola* (at >10 Ma and >5.2 Ma) and *Plagiostigma* (at >3.6 Ma) for the gene *GPI*, however for nodes not represented in the current three-gene dataset. For secondary calibration, we used the estimates of mean and 95% HPD from Marcussen et al. (2012), and applied corresponding normal priors on the sect. *Viola* crown node, set to 11.92 (SD 0.51) Ma, the sect. *Plagiostigma* crown node, set to 17.27 (SD 0.51) Ma, and the sect. *Nosphinium* s.lat. crown node, set to 9.48 (SD 0.43) Ma (sect. *Nosphinium* s.lat. has homoeologs nested within the sects. *Plagiostigma* and *Viola*). The close proximity (1–2 Ma) to the original fossil calibration points is the reason for the relatively high precision for these secondary calibrations. The secondary calibration of the root of *Viola* is based on a dated phylogeny that incorporates the same >18 Ma *Viola* fossil as our study, but here the distance to fossil-calibrated nodes is larger and so is the age uncertainty.

Finally, we applied normal-distributed secondary calibrations also on the basal nodes of the phylogenies: the root for *GPI* (tmrca of *Noisettia* + *Viola*) was set to 37.25 (SD 0.76) Ma, the root for *NRPD2a* (tmrca of *Allexis* + *Viola*) was set to 42.34 (SD 0.90) Ma, and no age prior was set on the *SDH* root because it was rooted with paralogs; the *Viola* crown node (split of the sect. *Andinium* lineage) was set to 29.50 (SD 1.11) Ma for all three analyses. The two first age priors (tmrca of *Noisettia* + *Viola* and tmrca of *Allexis* + *Viola*) equal the 95% HPD distributions obtained for the corresponding nodes in a fossil-calibrated Malpighiales/Violaceae dataset for *matK*, *atpB*, and 18S (Supplementary Table S2; see details below). The *Viola* crown node was not represented in the *matK*, *atpB*, and 18S dataset and had to be estimated in a secondary analysis of this subclade, with denser taxon sampling, using *trnL-F* (see details below). Because dating analyses are vulnerable to missing data and unequal taxon sampling between sister clades, which can distort node ages, this analysis design consisting of consecutive analyses of smaller complementary datasets was preferred to one single analysis. The *matK*, *atpB*, and 18S dataset (Supplementary Table S2, http://dx.doi.org/10.5061/dryad.jc754) was obtained for 70 taxa of Violaceae and related families (Tokuoka and Tobe 2006; Tokuoka 2008) and was calibrated using three

Malpighiales fossils: arrival of *Viola* in the northern hemisphere >18 Ma (same fossil as above), tmrca of *Adenia* and *Passiflora* (Passifloraceae) >37 Ma (Hearn 2006), and tmrca of *Idesia*, *Populus*, and *Salix* (Salicaceae) >65.5 Ma (Bell in Fawcett et al. 2009). These were given lognormal distributions, with offset ages as indicated and with both mean and standard deviation equalling 0.6, 0.7, and 0.8, respectively. In addition, we set normal-distributed secondary age constraints on the root being 113 (SD 2.4) Ma, and the crown age of Malpighiales being 102 (SD 3.1) Ma, based on estimates from analysis "BRC-2" of Wang et al. (2009), and in general agreement with the fossil record and newer estimates (Bell et al. 2010; Xi et al. 2012). The data matrix was partitioned with respect to region and each partition was analyzed under the nucleotide substitution model GTR + Γ. Two MCMC chains were run for the *matK*/*atpB*/18S data, for 30 million respective 35 million generations, checked for mixing and convergence, and the first 0.3 million generations of each were discarded as burn-in before the files were merged.

In order to obtain age priors for the *Viola* crown node (split of sect. *Andinium* lineage), which was not represented in the *matK*/*atpB*/18S dataset but necessary for calibrating the nuclear gene phylogenies (*GPI*, *NRPD2a*, and *SDH*), we used the posterior age distributions of the *matK*/*atpB*/18S analysis as normal-distributed priors for the two basal nodes in a *trnL-F* dataset with a denser sampling of this clade (Violaceae subtribe Violinae). These two nodes were tmrca of *Allexis* + *Noisettia* + *Schweiggeria* + *Viola* with 42.55 (SD 1.91) Ma and tmrca of *Noisettia* + *Schweiggeria* + *Viola* with 35.10 (SD 1.75) Ma. The MCMC chain for the *trnL-F* dataset was run for 10 million generations, checked for mixing and convergence, and the first 1 million generations were discarded as burn-in. We also ran BEAST analyses for the three nuclear genes (i) without sequence data to check for spurious effects of the prior combinations, and (ii) without calibrations, except on the *Viola* crown node; all other settings were unchanged.

### *Step 3. Inference of the Most Parsimonious Network from Multilabeled Gene Trees*

We chose to resolve the allopolyploid networks at section level rather than species level, considering that each of the 16 section lineages is represented in this study by a fraction of their species diversity, and that allopolyploidization is known to occur extensively also within sections. For each of the three nuclear gene phylogenies *GPI*, *NRPD2a*, and *SDH*, we identified subgenome lineages, i.e., clades consisting of co-homoeologs (i.e., orthologs of a homoeolog) resolved at the section lineage level, and used these as OTUs in the following analysis (Supplementary Figs. S1–S4, http://dx.doi.org/10.5061/dryad.jc754). Each subgenome lineage was named by section, as given in Table 1. By this approach paralogy within co-homoeolog clades, apparently present in some polyploids, was irrelevant to the analysis. All subgenome lineages were found in more than one marker and were therefore interpreted as homoeologs and not paralogs. Nodes differing in age by more than an arbitrarily chosen level of 5 Ma among gene trees were assumed to represent different speciation events; more than 5 Ma would perhaps require unrealistically large population sizes.

To explain topological incongruence concerning a single node, among the three gene trees, a set of five alternative genome trees was generated manually (Table 2 and Fig. 2). Each tree produced a unique network topology and assumed different events of deep coalescence and gene loss, while intuitively minimizing the number of such events. Three of the alternative genome trees invoked the presence of lineages that were not detected by homoeologs in any of the three nuclear gene phylogenies. We refer to such missing lineages as "ghost (subgenome) lineages" in the following.

We used the following three parsimony criteria for selecting which out of five scenarios gave the shortest network: (i) fewest possible ghost subgenome lineages, (ii) fewest possible polyploidization events, and (iii) least possible deviation from expected ploidy inferred from available chromosome counts for the involved polyploid taxa. The three parsimony criteria were given equal weights and the sum of steps for each scenario was compared. For calculating the minimum number of allopolyploidizations and constructing the polyploid network for each scenario, we used the general HOLM algorithm (Huber et al. 2006), as implemented in Dendroscope 3 (Huson and Scornavacca 2012). Each inferred homoeolog extinction and each inferred polyploidization equalled one step. Likewise, each 2*x* difference between expected and observed ploidy (i.e., chromosome number-inferred ploidy minus phylogeny-inferred ploidy) equalled one step. For instance, observing 8*x* instead of expected 12*x* gives a difference of 4*x* and two steps, which equals the number of residual polyploidizations (i.e., two) that have not been recovered by the gene phylogenies. The affected polyploids in this case were sects. *Chilenium*, *Erpetion*, and *Tridens*. Given the known base chromosome numbers for *Viola*, $x = 6$ and $x = 7$ (Table 1), sect. *Erpetion* ($2n = 60$) and sect. *Tridens* ($2n = 80$) are assumed to be 10*x* and 12*x*, respectively; no counts exist for sect. *Chilenium*.

In order to get a better resolution of the chloroplast CHAM clade, and to estimate the number of polyploidizations associated with it, we aligned relevant *trnL-F* sequences to an eight-gene chloroplast matrix obtained from GenBank, consisting of *atpB-rbcL*, *atpF-atpH*, *matK*, *psbA-trnH*, *psbK-psbI*, *rpl16*, *rpoC1*, and *trnL-F*. Data for a total of 55 species were included (Supplementary Table S3, http://dx.doi.org/10.5061/dryad.jc754), obtained mainly from the studies of Yoo and Jang (2010), Liang and Xing (2010), and Burgess et al. (2011). The phylogeny was reconstructed by Bayesian inference (BI) with MrBayes as above (under Step 1), except: each nucleotide region analyzed as separate partition under the GTR + Γ model; indels characters (coded as above) were analyzed using the

TABLE 2. Parsimony evaluation of five HOLM phylogenetic networks (A-E; Supplementary Table S4, http://dx.doi.org/10.5061/dryad.jc754) for *Viola*

| Scenario | Lineages | (1) Inferred ghost lineages | (2) Sum poly-ploidizations | Section *Erpetion* | | | Section *Tridens* | | | Parsimony score |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Inferred ploidy | Actual ploidy | (3) Undetected poly-ploidizations | Inferred ploidy | Actual ploidy | (3) Undetected poly-ploidizations | |
| A | 1 | 0 | 6 | 6*x* | 10*x* | 2 | 8*x* | 12*x* | 2 | 10 |
| B | 2 | 0 | 9 | 8*x* | 10*x* | 1 | 12*x* | 12*x* | 0 | 10 |
| C | 2 | 1: Erpetion1 | 9 | 10*x* | 10*x* | 0 | 12*x* | 12*x* | 0 | 10 |
| D | 2 | 1: Chilenium3 | 9 | 8*x* | 10*x* | 1 | 12*x* | 12*x* | 0 | 11 |
| E | 2 | 2: Chilenium3, Erpetion1 | 7 | 10*x* | 10*x* | 0 | 12*x* | 12*x* | 0 | 9 |

Notes: The corresponding multilabeled trees are shown in Figure 2. Parsimony criteria were (1) fewest possible "ghost" subgenome lineages (i.e., subgenome lineages that were not detected by the nuclear gene phylogenies), (2) fewest possible polyploidization events, and (3) least possible deviation in ploidy as inferred from available chromosome counts, i.e., 10*x* for sect. *Erpetion* and 12*x* for sect. *Tridens*. The three parsimony criteria were given equal weights and the sum of steps for each scenario was compared. Scenario E (nine steps) results in the most parsimonious network (Fig. 4). Lineages refer to whether the gene tree Clade I and Clade II (Fig. 2) are considered to represent one genome lineage (scenario A) under the assumption that the gene tree incongruence is due to deep coalescence, or two genome lineages (scenarios B–E) under the assumption of polyploidization followed by complementary loss of homoeologs in Clade I and Clade II.

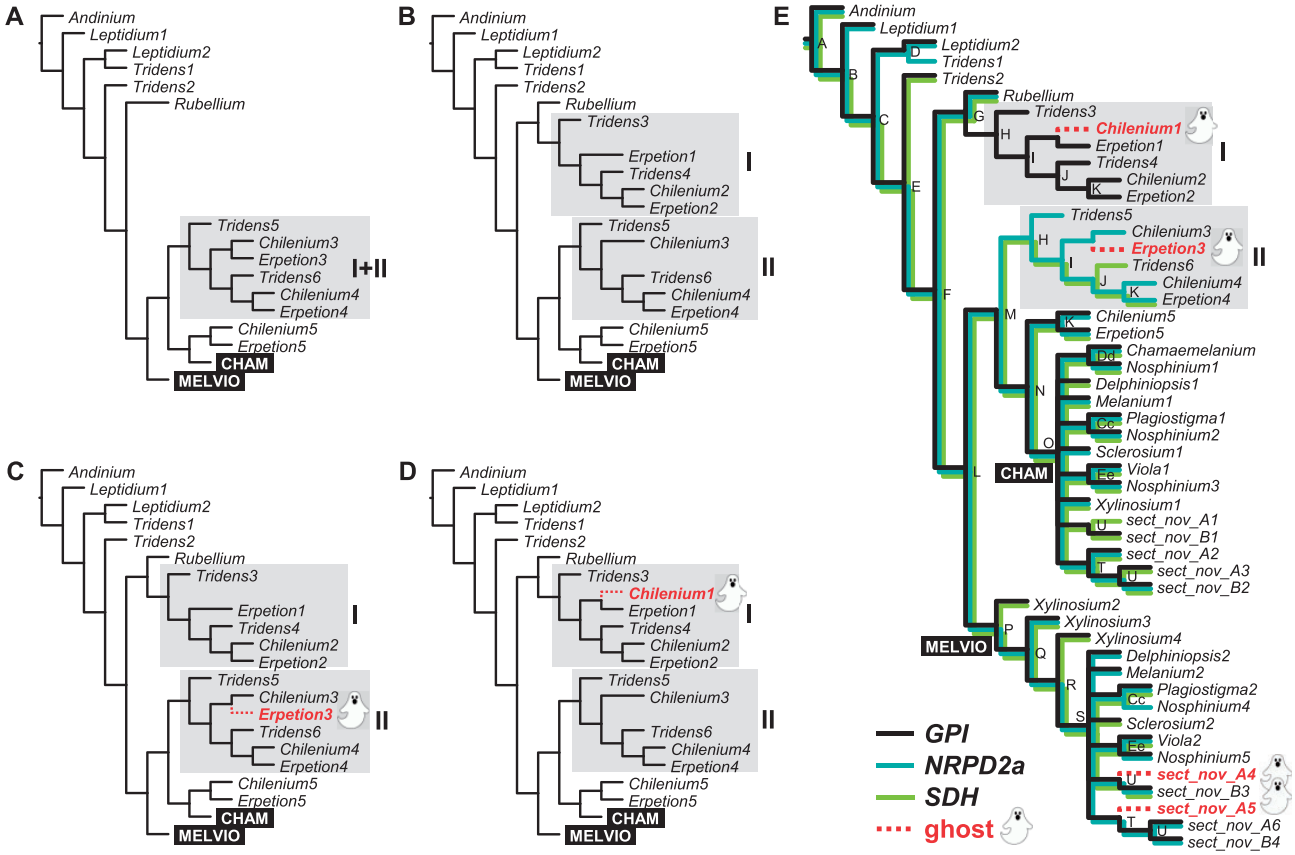


FIGURE 2. Multilabeled genome trees corresponding to the five competing network scenarios A–E in Table 2, reconciled from three low-copy nuclear gene phylogenies. The five competing network scenarios are explained in the text. The focal subclades with conflicting topology are shaded, and each inferred but not observed ghost lineage therein is indicated with a broken line and with a little ghost symbol. The subclades CHAM and MELVIO have identical topology under all scenarios and are shown only for scenario E (collapsed in trees A–D), which produces the most parsimonious network. The individual gene phylogenies (*GPI*, *NRPD2a*, *SDH*) are superimposed on the genome tree for scenario E, and node labels are given. The networks corresponding to scenario E is shown in Figure 4; networks for scenarios A–D are shown in Supplementary Table S4, http://dx.doi.org/10.5061/dryad.jc754. Homoeolog clades are indicated with numbers following section names.

simple model for binary data; and the analysis was run for 13 million generations.

Finally, for the unresolved nodes of CHAM and MELVIO, for which the number of CHAM × MELVIO tetraploidizations was unknown but theoretically between one and seven, we calculated the expected number of polyploidizations, assuming a binomial distribution. For the value of the parameter *p*, i.e.,

the probability that a speciation was polyploid, we used the minimum proportion of polyploid speciations estimated for *Viola*, omitting extinct lineages (see "Results" section).

*Step 4. Dating the Most Parsimonious Species Network*

Of the existing dating softwares that estimate species splits from multigene allele splits in a coalescent framework (e.g., *BEAST), none are able to deal with multilabeled gene/genome trees, except for very simple scenarios (Jones et al. 2013). Based on the shortest estimated network (scenario E; see "Results" section), we used a Bayesian model in the software WinBUGS 1.4.3 (Lunn et al. 2000) to estimate species split ages. These were estimated from the correspondent allele split ages extracted from the individual chronograms for *GPI*, *NRPD2a*, and *SDH* (see "Results" section). Assumptions were made that (i) for each species branch, the time from an allele split to the subsequent species split is exponentially distributed with rate $\mu$; that (ii) $\mu$ is constant throughout the network, justified by no prior information of how $\mu$ may vary through the tree; and that (iii) the topology of the species phylogeny is given (network scenario E; see "Results" section). This method is related to the one used in Marcussen et al. (2014). A model with separate $\mu_i$ for each branch $i$ was also tested, where $\mu_i$ was gamma distributed. Model selection, performed by use of latent variable modeling (George and McCulloch 1993), favored the simpler model, however.

Polyploid networks contain two types of internal branches (Fig. 3), those that have experienced a polyploidization event and those that have not. Polyploidization must have happened in the time interval between the younger of the two parental lineage splits ($t_{\text{parent2}}$)—because both parent lineages must be present at the time of polyploidization—and the subsequent homoploid speciation ($t_{\text{H}}$). Different estimates for $t_{\text{parent1}}$ and $t_{\text{parent2}}$ may reflect different coalescence due to sampling effects—the individuals chosen to represent the parental lineages and the polyploid are not identical with those originally involved in the allopolyploidization. We used a Bayesian model in the WinBUGS software (Lunn et al. 2000) to estimate the timing of polyploidization events $t_{\text{P}}$. A constant homoploid speciation rate ($\lambda_{\text{H}}$) and a constant polyploidization rate ($\lambda_{\text{P}}$) were assumed. The likelihood of the polyploidization and the subsequent homoploid speciation was

$$\lambda_{\text{P}}\exp(-\lambda_{\text{P}}(t_{\text{parent2}}-t_{\text{P}}))\cdot\lambda_{\text{H}}\exp(-\lambda_{\text{H}}(t_{\text{P}}-t_{\text{H}}))$$

where $t_{\text{parent2}}$ and $t_{\text{H}}$ had already been estimated as part of the dated genome network. Values of $t_{\text{H}}$ for the crown groups of sect. *Delphiniopsis* (5.33 Ma) and sect. *Sclerosium* (6.93 Ma), where only one species had been included by us, were taken from the literature (Herrera 1990) and from an unpublished *NRPD2a* phylogeny (Mohammadi Shahrestani 2013), respectively.

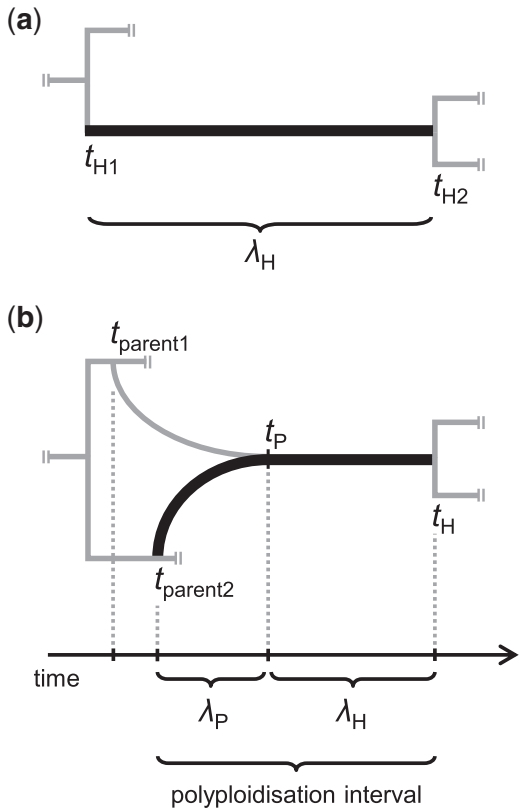


FIGURE 3. The two types of branches in a polyploid species network and their associated parameters. a) Species branch with no polyploid speciation event. The branch spans the time interval between two homoploid speciations occurring at time $t_{\text{H1}}$ and time $t_{\text{H2}}$ and is associated with a homoploid speciation rate $\lambda_{\text{H}}$. b) Species network containing one polyploid speciation event. The branch spans the time interval between the youngest split from the parental lineages at time $t_{\text{parent2}}$ and the homoploid speciation at time $t_{\text{H}}$, within which interval an allopolyploidization happened at time $t_{\text{P}}$. This branch is associated with a polyploid speciation rate $\lambda_{\text{P}}$ before polyploidization (i.e., between $t_{\text{parent2}}$ and $t_{\text{P}}$) and with the general homoploid speciation rate $\lambda_{\text{H}}$ after polyploidization (i.e., between $t_{\text{P}}$ and $t_{\text{H}}$). When $t_{\text{parent2}}$, $t_{\text{H}}$ and $\lambda_{\text{H}}$ are known, $\lambda_{\text{P}}$ and $t_{\text{P}}$ can be estimated from the data. The difference in the estimates for $t_{\text{parent1}}$ and $t_{\text{parent2}}$ is due to different coalescence and arises because the individuals representing the parental lineages and the polyploid are not identical with those originally involved in the allopolyploidization.

Some genome mergers involved more than two lineages. These represented unresolved, successive polyploidizations for which we failed to sample all lineages of "transient" ploidy, presumably due to extinction. In the case of unresolved, double polyploidization events, the three involved lineages could theoretically have been combined in three ways. We made the assumption that the first polyploidization happened between the two ancestral lineages with the earliest parental splits ($t_{\text{parent1}}$ and $t_{\text{parent2}}$), and that the second polyploidization involved this allopolyploid genome and the third ancestral lineage, splitting off at $t_{\text{parent3}}$. This assumption was based on the initial finding that single polyploidizations always seemed to occur early within the available time interval ($t_{\text{parent2}}$, $t_{\text{H}}$). The likelihood for the two polyploidizations P1 and P2 and

the subsequent homoploid speciation was

$$\lambda_P \exp[-\lambda_P(t_{parent2} - t_{P1})] \cdot$$
$$\lambda_P \exp\{-\lambda_P[\min(t_{P1}, t_{parent3}) - t_{P2}]\} \cdot$$
$$\lambda_H \exp[-\lambda_H(t_{P2} - t_H)].$$

The case of unresolved mergers of four genome lineages was more complex. These cases represent two or three subsequent polyploidizations and can be resolved in 15 different ways. Selection among these scenarios was obviously not possible, and instead we assumed the simplest model where genomes were successively "added" to the polyploid lineage by decreasing stem lineage age (i.e., $t_{parent1}$ and $t_{parent2}$ before $t_{parent3}$ before $t_{parent4}$).

The prior distributions for $\lambda_H$ and $\lambda_P$ were set to be gamma with shape parameter $\alpha = 0.5$ and rate parameter $\beta = 1$ for both speciation rates. This gives a prior mean of 0.5 and variance of 0.5. These values were chosen to maximize the variance without allowing a large skewness in the posterior distributions. Other values were also tested, and these gave similar estimates for $\lambda_H$ and $\lambda_P$. The priors for the polyploidization events were uniform. For the single polyploidizations, the prior started at $t_{parent2}$ and ended at $t_H$. For unresolved, double polyploidizations, the starting point for $t_{P1}$ was $t_{parent2}$, and the end-point was $t_{P2}$. For $t_{P2}$, the starting point was the younger of $t_{P1}$ and $t_{parent3}$, and the end point was $t_H$. Triple polyploidizations were treated accordingly.

The number of CHAM × MELVIO polyploidizations is uncertain, owing to polytomies in all phylogenetic markers. To account for this uncertainty, which affects estimates of polyploidization times, we analyzed the data separately for the minimum number (two; empirically inferred by chloroplast phylogeny) and maximum number (seven) of CHAM × MELVIO polyploidizations (i.e., in total in *Viola* 16 and 21, respectively).

## Results

### *Step 1. Multilabeled Gene Phylogenies*

For all four gene markers, *trnL-F*, *GPI*, *NRPD2a*, and *SDH*, the homologs of the diploid (2*x*) sect. *Andinium* were resolved as sister to the rest of the genus (Supplementary Figs. S1–S4, http://dx.doi.org/10.5061/dryad.jc754). The homoeologs of the remaining South American and Australian sections, i.e., *Chilenium*, *Erpetion*, *Leptidium*, *Rubellium*, and *Tridens*, formed a grade basal to two large homoeolog clades (CHAM and MELVIO; cf. Marcussen et al. 2010), which were exclusive to the Northern Hemisphere sections except for CHAM homoeologs present in sect. *Chilenium* and *Erpetion*. The three nuclear gene phylogenies were largely congruent, but one subclade consisting of homoeologs of sects. *Chilenium*, *Erpetion* and *Tridens* was differently placed in *GPI* (Supplementary Fig. S2, http://dx.doi.org/10.5061/dryad.jc754) compared with *NRPD2a* (Supplementary Fig. S3, http://dx.doi.org/10.5061/dryad.jc754) and *SDH* (Supplementary Fig. S4, http://dx.doi.org/10.5061/dryad.jc754), indicating either deep coalescence (Fig. 2, scenario A) or polyploidization followed by sorting of homoeologs (Fig. 2, scenarios B–E). Among the Northern Hemisphere taxa, sect. *Chamaemelanium*, being diploid, possessed only CHAM homologs, while all the polyploid sections possessed homoeologs from both the CHAM clade and the MELVIO clade. The crown topology of both the CHAM clade and the MELVIO clade was polytomous for the three nuclear markers (Supplementary Figs. S2–S4, http://dx.doi.org/10.5061/dryad.jc754), but contained numerous well-supported lineages.

Pronounced among-lineage rate heterogeneity was observed amongst the CHAM and MELVIO daughter lineages for the chloroplast and all three nuclear phylogenies (Supplementary Figs. S1–S4, http://dx.doi.org/10.5061/dryad.jc754). Short branches were found in the homoeolog lineages of sects. *Chamaemelanium*, *Plagiostigma* and *Viola*; these never formed a clade. Branches subtending sect. *Melanium* homoeologs were three to five times longer, and branches subtending homoeologs of the other lineages (*Delphiniopsis*, *Sclerosium*, *Xylinosium*, sects. nov. A, and B) were intermediate in length.

There was a general correspondence between the number of homoeologs and the putative ploidy levels of sections (Table 1). Diploid taxa generally possessed one homolog, sometimes with allelic variation (*GPI*: *V. uniflora*) or extra paralogs (*SDH*: *V. biflora*, *V. congesta*, *V. pubescens*, *V. pusilla*), occasionally pseudogenized (*NRPD2a*: *V. sheltonii*), and in one case (*NRPD2a*: *V. pusilla*) we repeatedly failed to amplify any homolog. Polyploids generally possessed several homoeologs but often fewer than expected from ploidy. In some cases, though, we observed a higher number of gene copies than expected, typically in high-polyploids for the markers *NRPD2a* and *SDH*. These "additional" copies were usually easy to identify as pseudogenized paralogs since they were confined to single homoeolog clades, single species, or single markers. Paralogs were discarded prior to the network analysis because their origins do not reflect speciation events.

### *Step 2. Divergence Times for Multilabeled Gene Phylogenies*

All dating analyses, performed in BEAST, were run also without sequence data (not shown) to check for possible spurious effects of the prior combinations, but no such effects were found. The dating analyses of the nuclear genes produced overlapping but slightly higher posterior ages on the basal node of *Viola* compared to the prior (mean 29.5 ± SD 1.11 Ma): mean (95% HPD) were 32.3 (30.3–34.2) Ma for *GPI*, 31.3 (29.4–33.2) Ma for *NRPD2a*, and 32.1 (30.1–34.0) Ma for *SDH* (Supplementary Figs. S5–S8, http://dx.doi.org/10.5061/dryad.jc754). At the base of the genus only

homoploid speciation nodes were present until ca 25 Ma. Following that, the chronograms confirmed an abrupt radiation of the Northern Hemisphere lineages between 20 and 15 Ma ago, as previously indicated by polytomous crown nodes for the CHAM and MELVIO clades in all three gene phylogenies (Supplementary Figs. S2–S4, http://dx.doi.org/10.5061/dryad.jc754). The topologies were similar to those obtained with MrBayes (Supplementary Figs. S1–S4, http://dx.doi.org/10.5061/dryad.jc754).

Dating analyses with calibration only on the basal node (Supplementary Fig. S9, http://dx.doi.org/10.5061/dryad.jc754), tmrca of *Viola*, produced phylogenies that differed considerably from the fossil-constrained ones (Supplementary Figs. S5–S8, http://dx.doi.org/10.5061/dryad.jc754) and the unconstrained MrBayes phylogenies (Supplementary Figs. S1–S4, http://dx.doi.org/10.5061/dryad.jc754): (i) The age of the polytomious CHAM and MELVIO clades were variably 5–10 Ma younger. Their variable age suggests the dating analysis was strongly affected by the rate heterogeneity observed within these clades and the uneven sampling of homoeologs (Supplementary Figs. S1–S4, http://dx.doi.org/10.5061/dryad.jc754). (ii) The obtained ages for sect. *Viola* (2.5–5 Ma) are in conflict with the known fossil age (at least 10 Ma; Marcussen et al. 2012). (iii) Homoeologs of the slow-rate taxa, sect. *Chamaemelanium*, *Plagiostigma* and *Viola*, formed in five out of six cases moderately to strongly supported (BI 0.54–0.95) clades not present in the MrBayes phylogenies (Supplementary Figs. S1–S4, http://dx.doi.org/10.5061/dryad.jc754). These "slow-rate" clades can be interpreted as artifacts resulting from a more parameter-rich model coupled with a possible poor fit with the clock model applied (the uncorrelated lognormal clock).

### *Step 3. The Most Parsimonious Network*

We reconciled five possible multilabeled genome trees (Fig. 2A–E) from the *GPI*, *NRPD2a*, and *SDH* gene trees. This was due to the conflicting placement of two clades (Fig. 2, clades I and II), which could be a result of either deep coalescence of the same homoeolog or allopolyploidy followed by complementary loss (sorting) of homoeologs. For *GPI* this clade was sister to sect. *Rubellium*, albeit with weak support, while for *NRPD2a* and *SDH* it was sister to the CHAM lineage (Supplementary Figs. S2–S4, http://dx.doi.org/10.5061/dryad.jc754).

Each of the five trees had a different topology and was associated with a different scenario of homoeolog loss and polyploidization. Under scenario A (Fig. 2A), the incongruent placement of clade I/II among gene trees was due to deep coalescence. Under scenarios B–E, the *GPI* lineage and the *NRPD2a*/*SDH* lineage represented different homoeologs. In order to reconcile identical topologies for the two gene tree topologies, we had to assume two ghost lineages ("Chilenium1", "Erpetion3") for which no homoeologs were detected in any of the three genes. We therefore made one scenario for each of the four combinations of presence/absence of these ghost lineages: Scenario B (Fig. 2B) assumed no homoeolog extinction (i.e., no ghost lineages were inferred). Scenario C (Fig. 2C) and scenario D (Fig. 2D) made the assumption of extinction of one homoeolog each (i.e., inferring one ghost lineage each). Finally, scenario E (Fig. 2E) assumed extinction of both homoeologs (i.e., inferring two ghost lineages).

Parsimony scores for each of the five scenarios (Fig. 2A–E) are listed in Table 2, and were calculated based on the following criteria, (i) fewest gene loss events, (ii) fewest polyploidizations as identified using the HOLM algorithm, and (iii) minimum deviation from expected ploidy levels in the affected polyploids. The corresponding HOLM networks (A–E) are shown in Supplementary Table S4, http://dx.doi.org/10.5061/dryad.jc754. Scenario E received the lowest parsimony score of 9 (for a dated network, see Fig. 4). Under this scenario, two ghost lineages were assumed, and sects. *Erpetion* and *Tridens* were inferred to have the expected 10*x* and 12*x* levels, respectively. The other scenarios received scores of 10 (scenarios A, B, and C) or 11 (scenario D). All five scenarios gave identical network topology and ploidy for the remaining sections (not shown): three sections were found to be diploid (2*x*; sects. *Andinium*, *Chamaemelanium*, *Rubellium*); six sections were found to be tetraploid (4*x*; sects. *Delphiniopsis*, *Leptidium*, *Melanium*, *Plagiostigma* s.str., *Sclerosium* and *Viola* s.str.); two sections were found to be probably octoploid (8*x*; sect. nov. B, sect. *Xylinosium*); two sections were found to be decaploid (10*x*; sects. *Chilenium*, *Erpetion*, *Nosphinium* s.lat.); and two were found to be dodecaploid (12*x*; sect. *Tridens*, sect. nov. A).

The HOLM network algorithm interpreted the unresolved radiation of CHAM × MELVIO polyploid lineages, deriving from nodes O and S in Figure 4, to be due to (a minimum of) seven independent allopolyploidizations (13–18 and 21; Fig. 4). At the other extreme, parsimony suggests a single polyploidization followed by homoploid speciation of the tetraploids. We assumed that the number of polyploid speciations was binomially distributed with parameter $p = 15/24$, where 15 was the minimum number of polyploidizations in *Viola*, and 24 was the total number of speciations (see next paragraph). Under these assumptions, the expected number of polyploidizations was 4.75, with low probabilities for the marginal one, two, and seven polyploidizations ($P = 0.0028$, $P = 0.028$, and $P = 0.06$, respectively). The number of CHAM × MELVIO polyploidizations was most likely to be between three and seven ($P = 0.97$). Empirical observations indicate that this number is at least two, based on phylogenetic analysis of our *trnL-F* sequence data together with an eight-gene chloroplast phylogeny (Supplementary Fig. S10; P1 and P2, http://dx.doi.org/10.5061/dryad.jc754). Here the CHAM clade was subdivided into two subclades. The first clade, strongly supported (posterior

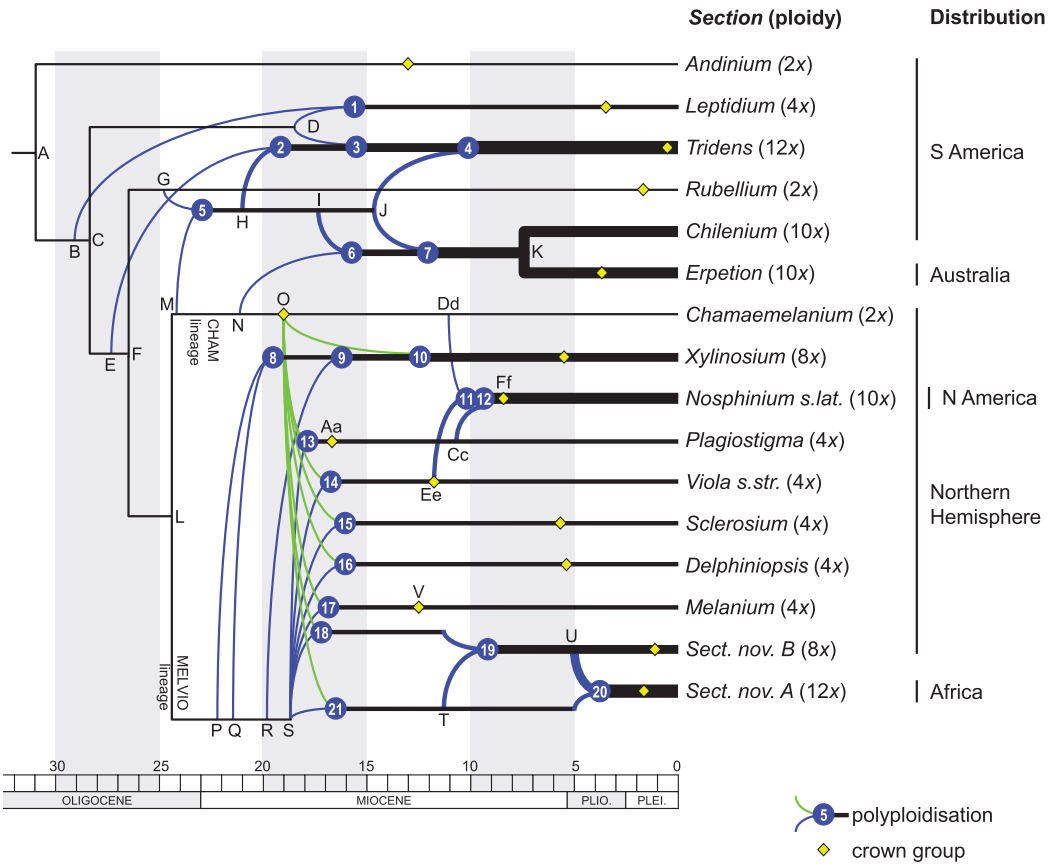


FIGURE 4. Most parsimonious HOLM network for the 16 provisional sections of *Viola* based on the multilabeled tree in Figure 2E. Node labels correspond to those in Figure 2 and Supplementary Table S5. Lineages immediately involved in allopolyploid speciation are drawn as curved lines. A total of between 15 and 21 polyploid speciations was inferred; all 21 are shown here. Whether events 13–18 and 21 are seven independent polyploidizations or homoploid segregates of a single polyploidization is unclear. A multigene chloroplast phylogeny (Yoo and Jang 2010) and the pattern of ITS ribotype segregation (Ballard et al. 1999, Yockteng et al. 2003) indicate more than one origin but these data are incomplete in terms of sampling and/or resolution. Estimated ages for homoploid and polyploid speciations are shown as means and medians, respectively. Estimated ploidy for section lineages ranges from diploidy (2*x*) to dodecaploidy (12*x*), as indicated after each section name and by line thickness for each lineage.

probability 0.97), contained all extant diploids (sect. *Chamaemelanium*) together with sects. *Sclerosium* and *Viola* (and sect. *Xylinosium* which, however, has a different MELVIO parent). The second clade, weakly supported (0.61), contained the remaining polyploids (sects. *Delphiniopsis*, *Melanium*, *Plagiostigma*, sects. nov. A and B) but no diploids. This clade was strongly supported (0.99) when taxa sequenced for *trnL-F* were omitted (i.e., all but *Plagiostigma* and *Melanium*; not shown).

Thus, the most parsimonious network for *Viola*, based on scenario E (Fig. 2) and incorporating the uncertainty concerning the CHAM × MELVIO polyploids, required a minimum of between 16 and 21 polyploid speciations to explain the origin of the extant 16 section lineages of *Viola* (Fig. 4). The minimum number of homoploid speciations was, if extinct lineages are omitted, between three and eight: two in the divergence of the diploid sects. *Andinium*, *Chamaemelanium* and *Rubellium*, one in the divergence of the decaploid *Chilenium* and *Erpetion*, and between zero (in the case of seven polyploidizations) and five (in the case of two polyploidizations) in the divergence of the seven CHAM × MELVIO allopolyploids. Thus, if the sample here, resolved at the section level, is an unbiased representative for the allopolyploid speciation rate in *Viola*, then allopolyploidy accounts for between 67% (= 16/24) and 88% (= 21/24) of the speciation events within the genus.

Prior to analysis, we omitted from the set of multilabeled trees the *NRPD2a* copy "MACULATA_chilerp_1" (Supplementary Fig. S3, http://dx.doi.org/10.5061/dryad.jc754), as it was one step more parsimonious to interpret it as a paralog than as a homoeolog (i.e., paralogization versus polyploidization + extinction in the second parental clade); frameshift mutations and premature stop codons suggested this was a pseudogene. Furthermore, the inference of two ghost lineages ("sect_nov_A4" and "sect_nov_A5") for the MELVIO lineage of *V. abyssinica* was necessary to accommodate the presence of their counterparts in the CHAM lineage.

*Step 4. The Dated Network*

The dated network based on genome tree scenario E is shown in Figure 4. The multigene coalescent model returned species node ages on average 0.48 Ma (SD 0.26, min 0.15, max 1.26) younger than the youngest corresponding allele split in the three gene trees (Supplementary Table S5, http://dx.doi.org/10.5061/dryad.jc754). The rate parameter $\mu$ was estimated to 0.817 (SD 0.0953). The crown group age of *Viola* was estimated to 30.9 Ma. Crown group ages for the individual sections varied between 19.0 Ma in sect. *Chamaemelanium* and 16.6 Ma in sect. *Plagiostigma* down to less than 5 Ma in several other sections (*Erpetion*, *Leptidium*, *Rubellium*, *Tridens*, *Xylinosium*, sect. nov. A, and sect. nov. B). Although our sampling was designed so as to cover the assumed diversity within each section and therefore should give a good approximation of the crown node ages, some of these ages may be underestimates as a result of incomplete sampling. The effect on estimation of speciation rates is small, however, given an uncertainty of up to 2 Ma in crown node ages, and we do not expect crown node estimates to be biased with respect to the ploidy of lineages.

Analyses of speciation rates and polyploidization times (Fig. 3) provided very similar results under the scenarios of 16 or 21 polyploidizations (Supplementary Fig. S11 and Supplementary Table S6, http://dx.doi.org/10.5061/dryad.jc754). Under 16/21 polyploidizations, the polyploid speciation rate $\lambda_P$ was estimated to be 0.61/0.54 (SD 0.53/0.46), and almost three times higher than the homoploid speciation rate $\lambda_H$, estimated to be 0.21/0.20 (SD 0.05/0.05). Median polyploid speciation time $t_P$ under the scenario of 16 polyploidizations was on average 0.27 Ma younger than under the scenario of 21 polyploidizations. The posterior distribution for $t_P$ was generally highly skewed towards old ages, usually with the mode within 1 Ma after the youngest $t_{parent}$. On average, the corresponding median of single/first polyploidizations (1–2, 5–6, 8, 11, 13–21, P1, P2) was 1.56/1.80 (SD 0.53/0.63, min 0.71/0.71, max 2.80/2.86) Ma after the youngest $t_{parent}$, while the medians for the second (3, 7, 9, 12) and third (4, 10) polyploidizations were higher, at 2.74/2.97 (SD 0.44/0.49) Ma and 4.85/5.61 (SD 1.44/1.38) Ma, respectively. The higher medians for second and third polyploidizations reflect increased uncertainty, which results in nearly flat probability posterior distributions, relaxing the estimate towards the middle of the interval.

## Discussion

*Generating a Dated Network from Gene Phylogenies*

Reconstruction of phylogenetic polyploid networks is currently a rapidly developing field of research in phylogenetics, but no general-purpose method or software has yet been developed (but see Jones et al. 2013). Herein, we reconstructed and dated the polyploid network of the angiosperm genus *Viola*, using a generally applicable approach consisting of four principal steps (Fig. 1), i.e., (i) obtaining gene phylogenies, (ii) dating individual gene phylogenies, (iii) identifying the most parsimonious multilabeled genome tree and its corresponding species network, and (iv) dating the species network using a coalescent approach. Below we will discuss the reliability of our approach, and theoretical, methodological, and practical shortcomings that need to be further developed by future research.

*Gene phylogenies.*—Reconstructing a polyploid phylogeny can only be done reliably by use of gene phylogenies that recover the polyploid signal. For this reason, low-copy nuclear markers are far better suited than the otherwise widely used rDNA and cpDNA markers. rDNA is a gene family consisting of thousands of paralogs, and the PCR-amplified sequence is often a variably weighted "consensus" of its paralogs (Álvarez and Wendel 2003). The chloroplast reproduces predominantly asexually. Different chloroplast lineages are therefore prone to accumulation of selective differences and to introgression, resulting in a genealogy incongruent to that of the nuclear genome (see e.g., Renoult et al. 2009 and references therein).

*Dating gene phylogenies.*—Phylogenetic dating is a difficult topic and best practices have been discussed in a number of papers (e.g., Rutschmann 2006; Forest 2009; Parham et al. 2012; Sauquet et al. 2012). Fossils represent species, not genes, and can therefore logically be used to calibrate species phylogenies only, not gene phylogenies. In the present study of *Viola*, we nevertheless had to apply fossil- and secondary calibration directly to the gene phylogenies, because current methodology does not allow age constraints on nodes within a phylogenetic species network, except in level-1 networks, i.e., networks where the hybrids are not allowed to hybridize (Jones et al. 2013). Our approximation is unlikely to have distorted ages noticeably because of the relatively rapid allele coalescence in *Viola* (0.48 Ma), and the comparatively high age of the calibrated nodes (10–30 Ma).

We also applied secondary calibrations on the root age of *Viola* and on the crown age of each of the ortholog clades of sect. *Viola* and sect. *Plagiostigma*. This was desirable because of the high rate heterogeneity observed among *Viola* polyploid lineages (Supplementary Figs. S2–S4, http://dx.doi.org/10.5061/dryad.jc754) and the failure to detect all homoeologs for all lineages (Fig. 2), which affected coalescence times to the extent of being in conflict with the fossil record, and also topology (Supplementary Fig. S9, http://dx.doi.org/10.5061/dryad.jc754). Secondary calibration can be problematic in the sense that sources of error generated by the first dating exercise remain and are propagated in subsequent analyses (Forest 2009; Sauquet et al. 2012). Such sources of error include taxonomic mistakes, incorrect interpretation of divergence times (typically

owing to a sparse fossil record), biased taxon sampling, or widely different gene coalescence ages for the genes used in the primary and in the secondary calibration. We tried to avoid these caveats by performing the primary calibrations with taxa carefully sampled to avoid bias (Marcussen et al. 2012), with each of the clades used for secondary calibration having been internally calibrated using fossils, and taking the uncertainty of the age estimate into account. The good correspondance between the abrupt differentiation of the Northern hemisphere lineages as seen in the gene trees and in the Eurasian fossil record, and the divergence of sect. *Viola* (Marcussen et al. 2012), suggest that our results are realistic.

In other taxa, the time difference between speciation and average coalescence of alleles may be much larger and a significant source of error. In such cases gene phylogenies may differ considerably from the species phylogeny both in topology and node ages (e.g., Escobar et al. 2011), but it is important to note that gene coalescence times are either as old or older than the speciation events. Using gene coalescence times to date species phylogenies will therefore result in bias.

*Network building and gene tree incongruences.*— Distinguishing between homoeologs and paralogs can be a problem when dealing with polyploid phylogenies, especially when few markers are investigated and if the polyploidization is old. Nevertheless, it is reasonable to assume that simultaneous duplication in unlinked markers results from whole genome duplication rather than from independent gene duplications. In a polyploid, redundant homoeologs may over time become extinct owing to diploidization, and also parental species lineages may become extinct. As a consequence, there will be fewer data to support ancient than recent polyploidizations. In cases where the data are insufficient to allow for inference of a single multilabeled genome tree and its corresponding network, it may be helpful to consider independent data on ploidy, typically inferred from chromosome counts (or alternatively from e.g., allozyme or microsatellite patterns, flow-cytometry, or pollen size). An assumption that has to be made is that chromosome counts directly reflect ploidy, which may be true only for neopolyploids and some mesopolyploids (e.g., Mandáková et al. 2010).

The *Viola* phylogenetic data contained topological conflicts among gene trees, hypothesized to result from sorting of either alleles or homoeologs, that prevented the reconciliation of a single multilabeled genome tree. A single genome tree (Fig. 2) could only be inferred by considering also independent data on ploidy. With the exception of sect. *Melanium*, which has been subject to considerable chromosomal remodeling (Table 1), chromosome number is generally tightly correlated with ploidy in *Viola* (see e.g., Marcussen et al. 2012 for evolution of a group of high-polyploid species). Thus for the involved lineages, we made the assumption that $2n = 60$ (sect. *Erpetion*) and $2n = 80$ (sect. *Tridens*) reflected the $10x$ and $12x$ level, respectively, based on the hypothetical base chromosome numbers of $x = 6$ and $x = 7$ (Table 1). The counts $2n = 60$ and $2n = 80$ have been reported from only a single study each (Moore in Smith-White 1959; Moore 1967), with incomplete supporting data at least for the former, so additional counts from these and related groups are unfortunately necessary to determine to which extent these numbers are representative for their lineages.

We inferred the most parsimonious network, counting as one step each inferred gene loss, each polyploidization, and each $2x$ deviation from expected ploidy as informed by chromosome counts (Table 2). Due to lack of proper information on the probability of each of these types of events, they were given equal weight in the parsimony test. Scenario E was chosen as it showed the lowest parsimony score (9), but the other scenarios were also possible, having scores of 10 and 11. Assuming a Poisson distribution with $\lambda = 9$ for the total number of events, the probability of observing 10 or 11 events is only slightly smaller (0.118 or 0.097) than the probability of observing nine events (0.132). That being said, the most parsimonious network (scenario E; Fig. 2) both minimizes the number of polyploidizations required to explain the data and agrees with the expected ploidy of sect. *Erpetion* and sect. *Tridens*. In any case, the parsimony reconstruction only concerns the ancestry of three out of the 16 section lineages of *Viola*, i.e., the high-polyploid sections *Chilenium*, *Erpetion* and *Tridens* (Figs. 2, 4 and Table 2). The origins of the remaining *Viola* polyploid lineages were unambiguously supported by the gene trees.

A peculiarity of the *Viola* data is the polytomy at the base of the seven CHAM × MELVIO polyploid lineages (Supplementary Figs. S1–S10, http://dx.doi.org/10.5061/dryad.jc754) which implies that the number of polyploidizations in this lineage could range from one (invoking parsimony) to seven (as reconstructed by the Holm algorithm; Huber et al. 2006). Empirical data from the chloroplast suggest at least two polyploidizations (Supplementary Fig. S10, http://dx.doi.org/10.5061/dryad.jc754), i.e., one tetraploidization putatively shared by sects. *Sclerosium* and *Viola* and another tetraploidization putatively shared by the others. However, when the high proportion of speciations by polyploidy in *Viola* is taken into account, the number of CHAM × MELVIO polyploidizations is likely between three and seven ($P = 0.97$), with an expectation of 4.75. In a wider context, multiple origins of allopolyploids are the rule rather than the exception (Soltis and Soltis 1993), so polyploidization is probably far from being so rare as being "limiting" in a phylogenetic perspective –unlike e.g., dispersal to remote islands. The persistence of polyploids may be determined largely by environmental factors and not by the rate by which polyploids arise. There may thus be good reasons to question the usefulness, or even adequacy, of applying parsimony thinking to the counting of polyploidizations.

*Dating the phylogenetic network.*—Under the multispecies coalescent, the expected time from speciation backwards to allele coalescence can be approximated by an exponential distribution (Kingman 1982; Yang and Rannala 2003; Degnan and Rosenberg 2009). We generated a dated species phylogeny by collecting allele split ages for each homoeolog across the three independently dated gene trees, and analyzing these jointly using a coalescent tree prior and a Bayesian hierarchical model. One limitation to this approach is that it does not incorporate the uncertainty in allele split ages in the final coalescent calculation of homoploid speciations (see "Results" section). In principle, the uncertainty could be included in the Bayesian model by use of the posterior distributions for the allele split ages and their correlations; this implementation, however, would require extensive programming. Modeling of the *Viola* data suggests that this underestimates the coalescence rate $\mu$ by about 50% and gives somewhat too young estimates for homoploid speciation times. However, since the speciation times were only on average 0.48 Ma younger than the youngest allele split time, this is not critical for a phylogeny that spans more than 30 Ma.

While coalescent theory (Kingman 1982) permits estimation of homoploid speciation times when associated allele split times and effective population sizes are known, theoretical models to estimate hybridization times are in their infancy (Bartoszek et al. 2013). Methods for detecting homoploid hybridization in the presence of lineage sorting have been proposed by Kubatko (2009), Gerard et al. (2011), Chen and Wang (2010, 2012), Yu et al. (2011, 2013), and Marcussen et al. (2014). Cai et al. (2012) devised a strategy where a backbone species tree of diploids is first made using *BEAST, and then each polyploid allele is analyzed individually to find its location on the backbone tree. A more general solution has been presented by Jones et al. (2013), where a parameter assigns the homeologs to the correct genome with a certain probability. A key component of this problem is that one is very unlikely to sample a parental genome or its direct descendant, with the consequence that a node deeper than the most recent common ancestor is typically scored for each parent. Therefore polyploidization times are likely to be overestimated by an unknown factor (e.g., Doyle and Egan 2010) that is expected to increase with the time since polyploidization as a result of extinction.

By definition, a branch in a species network delimits the time between two subsequent homoploid speciation events. We have proposed to estimate allopolyploidization time by assuming a constant homoploid speciation rate $\lambda_H$ that is independent of ploidy (cf. Soltis et al. 2014 for discussion). This implies that for branches sustaining polyploidization, whose lengths are known, the homoploid speciation rate $\lambda_H$ after polyploidization is known, and the time before allopolyploidization can therefore be estimated. This approach is probably more powerful for larger datasets with low rate heterogeneity, and where stochastic variation in branch length, owing to extinction, is evened out. Although the upper and lower bounds of polyploidization time were sometimes very wide (15 Ma or more) for the *Viola* dataset, this approach always found the highest probability density of polyploidization time $t_H$ within 1 Ma after the oldest bound. We interpret this as an indication that, at least in some cases within *Viola*, allopolyploid speciation may have happened immediately after its split from the ancestral lineages. It appears that the youngest $t_{parent}$ is the best general estimator of polyploidization time $t_H$ that we have currently. We can, however, still assume that it overestimates polyploidization times for older polyploidizations as a result of extinction.

Resolving successive polyploidization events leading to a merger of three or more lineages needed special consideration. For unresolved mergers of three lineages, i.e., producing a hexaploid, it is possible to infer the most likely sequence and timing of the two polyploidizations because single polyploidizations can be inferred to occur early within their available time intervals, and because only three possible ways of combining lineages are possible. However, the problem becomes more complex for unresolved mergers of four or more lineages; for instance, the four $2x$ genomes of an octoploid can be combined in as much as 15 ways and by either three or four polyploidizations. In general, where $s$ is the number of $2x$ genomes, the number of rooted networks is $(2s–3)!/(2^{s-2} \cdot (s–2)!)$. Probably, the only way to actually resolve mergers of more than two lineages to polyploidizations is by empirical study of subtle differences in genome downsizing, amount of homoeolog loss or pseudogenization, or inter-subgenome recombination within such a high-polyploid. Next-generation sequencing techniques are likely to generate the massive amount of data required for this kind of analysis.

### *The Phylogenetic Network for* Viola

Using the approach discussed above, we generated a dated phylogenetic network for the 16 sections of the genus *Viola*. The inferred age of *Viola*, ca 31 Ma (Fig. 4), coincides with the abrupt Early Oligocene cooling 34–27 Ma ago, during which global temperatures were lowered by 4°C (Liu et al. 2009). Numerous temperate angiosperm lineages diversified in this period, e.g., Brassicaceae s.str. (Couvreur et al. 2010), Campanulaceae s.str. (Roquet et al. 2009), and *Hypericum* (Meseguer et al. 2013). This may have facilitated the specialization and early diversification of the temperate *Viola* from a tropical ancestor within subtribe Violinae (Wahlert et al. 2014).

The inferred network for *Viola* required between 16 and 21 polyploid speciations and between three and eight homoploid speciations to explain the diversification into 16 extant section lineages (Fig. 4). If this is representative for the speciation process within the genus, this suggests that between 67% and 88% of the speciation events in *Viola* happen by polyploidy. Although this figure does not take into account speciation within sections, which cannot be

determined owing to a lack of chromosome counts and phylogenetic data, the incidence of speciation by polyploidization is considerably higher in *Viola* than the average of 15–30% estimated for angiosperms in general (Wood et al. 2009; Mayrose et al. 2011). For *Viola*, 81% of the sections (13 of 16) are polyploid, which translates into at least 75% of the species (assuming that all the species in sects. *Andinium* and *Rubellium* as well as half of the species in sect. *Chamaemelanium* are diploids; cf. Table 1). Also other temperate genera contain similar numbers, such as *Draba* (78%), *Festuca* (70%), and *Hedera* (60%) (Soltis et al. 2014), which suggests that *Viola* is not a special case. However, percentage polyploidy can only serve as a very rough guide when comparing genera: (i) polyploidy accumulates over time and older genera will therefore contain more polyploidy than younger ones; (ii) ploidy is often not specified; and (iii) the number of polyploids also does not have to reflect the number of polyploidizations.

We determined the oldest reticulations (i.e., homoeolog splits) in *Viola* to be about 29 Ma. If the presence of such old reticulations is even remotely representative for angiosperms in general, this would call into question the adequacy of using uniparental markers (i.e., cpDNA, mtDNA) for resolving organism phylogenies covering the same age—corresponding to the subfamily or tribe level in many angiosperm groups.

### *Future Directions and Concluding Remarks*

In this study, we generated a dated, allopolyploid species network for *Viola* (Fig. 4) in a series of four discrete steps (Fig. 1) that entailed obtaining gene phylogenies, dating each phylogeny, determining the species network, and estimating homoploid and polyploid speciation times. The biological mechanisms underlying these four steps are not independent, and all steps should logically be analyzed jointly in a single operation, if possible. Doing so would also better take into account the phylogenetic uncertainty present in the data. However, existing theory and software do not permit data to be analyzed in a single operation, except for very simple cases with few polyploidizations and taxa (Bartoszek et al. 2013; Jones et al. 2013). Considering the very complex polyploid relationships that exist for large eukaryotic groups, as exemplified here with *Viola*, refinements of existing theories, methods, and software to enable also phylogenetic analysis of networks are clearly needed.

Estimating the species network itself is a difficult task, however. This primarily owes to the confounding signatures of polyploidization as well as different gene-level processes such as duplication, extinction, and coalescent stochasticity. At the species level, the theoretical problem of resolving the sequence of polyploidizations in unresolved polyploid nodes appears to be hard and will require further study. On top of these factors come introgression and horizontal transfer of genetic material among nonsister lineages. Its signature is often difficult to distinguish from that of lineage sorting (Twyford and Ennos 2012; Yu et al. 2013), especially when the contribution is small, but this typically requires large numbers of genes being sampled (Rosenberg and Feldman 2002; Marcussen et al. 2014; Zwickl et al. 2014). Frequent observations of phylogenetic discordance among nuclear, plastid, and mitochondrial genomes may hint at introgression being more common than currently appreciated (e.g., Maureira-Butler et al. 2008, Abbott et al. 2013). Owing to the low number of genes studied here, we were unable to consider introgression, but the estimated short coalescence time of 0.48 Ma suggests that introgression is not prevalent in our data for *Viola*. It is expected that the accurate estimation of the species network topology will be greatly aided by the use of next generation sequence data, whose massive amount of data potentially eliminates gene-specific problems. Owing to the complexity of determining the network topology itself, it may be an acceptable approximation to first determine the species network topology and then use the topology as a constraint in the estimation of species split ages. In any case, coalescent-based method and software that integrate the estimation of both homoploid and polyploid speciations need to be developed.

## Supplementary Material

Supplementary material, including data files and online-only appendices, can be found in the Dryad data repository at http://dx.doi.org/10.5061/dryad.jc754. Alignment files and tree files can in addition be found in the Treebase data repository at http://purl.org/phylo/treebase/phylows/study/TB2:S15248.

## Funding

This work was supported by the Research Council of Norway (grant no. 170832 to K.S.J. and T.M.) and the Swedish Research Council (grant no. 2009-5202 to B.O.).

## Acknowledgments

We wish to thank editors Frank Anderson, Ron DeBry, and Mark Fishbein, and four anonymous reviewers, for valuable help and advise in improving the manuscript. The late Kim Blaxland is thanked for generously sharing her insights in *Viola* through numerous discussions and for obtaining plant material. We thank the following for plant material: Gerd Knoche, and Harvey E. Ballard, and the curators of EPS, UPS, and Le Jardin Botanique Alpin du Lautaret. Anna Skog and Flor-Inés Arias Sánchez are thanked for doing part of the molecular work. Trond Reitan is thanked for help and advice with the statistical analyses and John M. Watson for help with the language. Ana R. Flores, John M. Watson, Kevin Thiele and Harvey E. Ballard are thanked for sharing unpublished information on *Viola*. T.M. and K.S.J. conceived and designed the experiments. T.M. led the project, collected

the plant material, did the laboratory work, ran most analyses, and wrote most of the text. L.H. did the Bayesian modeling and estimation in WinBUGS. All authors have contributed to the preparation of the study and commented on and approved the final manuscript.

## References


Abbott R., Albach D., Ansell S., Arntzen J.W., Baird S.J.E., Bierne N., Boughman J., Brelsford A., Buerkle C.A., Buggs R., Butlin R.K., Dieckmann U., Eroukhmanoff F., Grill A., Cahan S.H., Hermansen J.S., Hewitt G., Hudson A.G., Jiggins C., Jones J., Keller B., Marczewski T., Mallet J., Martinez-Rodriguez P., Möst M., Mullen S., Nichols R., Nolte A.W., Parisod C., Pfennig K., Rice A.M., Ritchie M.G., Seifert B., Smadja C.M., Stelkens R., Szymura J.M., Väinölä R., Wolf J.B.W., Zinner D. 2013. Hybridization and speciation. J. Evol. Biol. 26:229–246.

Álvarez I., Wendel J.F. 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogen. Evol. 29:417–434.

Arrigoni P.V., Mori B. 1980. Numeri cromosomici per la Flora Italiana: 722-727. Infor. Bot. Ital. 12:145–148.

Ballard H.E., Sytsma K.J. 2000. Evolution and biogeography of the woody Hawaiian violets (*Viola*, Violaceae): arctic origins, herbaceous ancestry and bird dispersal. Evolution 54:1521–1532.

Ballard H.E., Sytsma K.J., Kowal R.R. 1999. Shrinking the violets: phylogenetic relationships of infrageneric groups in *Viola* (Violaceae) based on internal transcribed spacer DNA sequences. Syst. Bot. 23:439–458.

Bartoszek K., Jones G., Oxelman B., Sagitov S. 2013. Time to a single hybridization event in a group of species with unknown ancestral history. J. Theor. Biol. 322:1–6.

Becker W. 1925. *Viola* L. In: Engler A., editor. Die natürlichen Pflanzenfamilien. Parietales und Opuntiales. Leipzig: Wilhelm Engelmann. pp. 363–376.

Bell C.D., Soltis D.E., Soltis P.S. 2010. The age and diversification of the angiosperms re-revisited. Am. J. Bot. 97:1296–1303.

Brizicky G.K. 1961. The genera of Violaceae in the southeastern United States. J. Arnold Arboretum 42:321–333.

Burgess K.S., Fazekas A.J., Kesanakurti P.R., Graham S.W., Husband B.C., Newmaster S.G., Percy D.M., Hajibabaei M., Barrett S.C.H. 2011. Discriminating plant species in a local temperate flora using the *rbcL+matK* DNA barcode. Meth. Ecol. Evol. 2:333–340.

Brysting A.K., Mathiesen C., Marcussen T. 2011. Challenges in polyploid phylogenetic reconstruction: a case story from the arctic-alpine *Cerastium alpinum* complex. Taxon 60:333–347.

Cai D., Rodríguez F., Teng Y., Ané C., Bonierbale M., Mueller L.A., Spooner D.M. 2012. Single copy nuclear gene analysis of polyploidy in wild potatoes (*Solanum* section *Petota*). BMC Evol. Biol. 12:70.

Chen Z.-Z., Wang L. 2010. Hybridnet: a tool for constructing hybridization networks. Bioinformatics 26:2912–2913.

Chen Z.-Z., Wang L. 2012. Algorithms for reticulate networks of multiple phylogenetic trees. IEEE/ACM Trans. Comput. Biol. Bioinform. 9:372–384.

Clausen J. 1929. Chromosome number and relationship of some North American species of *Viola*. Ann. Bot. 63:741–764.

Clausen J. 1964. Cytotaxonomy and distributional ecology of western North American violets. Madroño 17:173–197.

Couvreur T.L.P., Franzke A., Al-Shehbaz I.A., Bakker F.T., Koch M.A., Mummenhoff K. 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol. Biol. Evol. 27:55–71.

Cui L., Wall P.K., Leebens-Mack J.H., Lindsay B.G., Soltis D.E., Doyle J.J., Soltis P.S., Carlson J.E., Arumuganathan K., Barakat A., Albert V.A., Ma H., dePamphilis C.W. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:738–749.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Dehal P., Boore J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:1701–1708.

Diosdado J.C., Ojeda F., Pastor J. 1993. IOPB chromosome data 5. Int. Organ. Plant Biosyst. Newsl. 20:6–7.

Dorofeev P.I. 1963. Tretichnye flory zapadnoi Sibiri (The Tertiary floras of western Siberia). Moscow: Izdatel'stvo Akademii Nauk SSSR.

Doyle J.J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17:144–163.

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

Doyle J.J., Egan A.N. 2010. Dating the origins of polyploidy events. New Phytol. 186:73–85.

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Erben M. 1996. The significance of hybridization on the forming of species in the genus *Viola*. Bocconea 5:113–118.

Escobar J.S., Scornavacca C., Cenci A., Guilhaumon C., Santoni S., Douzery E.J.P., Ranwez V., Glémin S., David J. 2011. Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). BMC Evol. Biol. 11:181–198.

Fawcett J.A., Maere S., Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–tertiary extinction event. Proc. Natl. Acad. Sci. USA 106:5737–5742.

Forest F. 2009. Calibrating the tree of life: fossils, molecules and evolutionary timescales. Ann. Bot. 104:789–794.

Galland N. 1985. Chromosome number reports LXXXVII. Taxon 34:346–351.

Galland N. 1988. Recherche sur l'origine de la flore orophile du Maroc, étude caryologique et cytogéographique. Trav. Inst. Sci., Univ. Mohammed V. Sér. Bot. 35:1–168.

George E.I., McCulloch R.E. 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88:881–889.

Gerard D., Gibbs H.L., Kubatko L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol. Biol. 11:291–303.

Gingins de la Sarraz F.C.J. 1823. Mémoire sur la famille des Violacées. Mém. Soc. Phys. Hist. Nat. Genève 2:1–27.

Gingins de la Sarraz F.C.J. 1824. *Viola*. In: De Candolle A.P., editor. Prodromus systematis naturalis regni vegetabilis. München: Michelsen.

Gong Q., Zhou J.S., Zhang Y.X., Liang G.X., Chen H.F., Xing F.W. 2010. Molecular systematics of genus *Viola* L. in China. J. Trop. Subtrop. Bot. 18:633–642.

Gregory T.R., Mable B.K. 2005. Polyploidy in animals. In: Gregory T.R., editor. The evolution of the genome. San Diego: Elsevier Academic Press. p. 428–501.

Hearn D.J. 2006. *Adenia* (Passifloraceae) and its adaptive radiation: phylogeny and growth form diversification. Syst. Bot. 31:805–821.

Hedrén M. 1996. Genetic differentiation, polyploidization and hybridization in northern European *Dactylorhiza* (Orchidaceae): evidence from allozyme markers. Plant Syst. Evol. 201:31–55.

Heilborn O. 1926. Bidrag til Violaceernas cytologi. Sven. Bot. Tidskr. 20:414–419.

Herrera C.M. 1990. The adaptedness of the floral phenotype in a relict endemic, hawkmoth-pollinated violet. 2. Patterns of variation among disjunct populations. Biol. J. Linn. Soc. 40:275–291.

Huber K.T., Oxelman B., Lott M., Moulton V. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. Mol. Biol. Evol. 23:1784–1791.

Huson D.H., Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61:1061–1067.

Jaillon O., Aury J.-M., Noel B., Policriti A., Clepet C., Casagrande A., Choisne N., Aubourg S., Vitulo N., Jubin C., Vezzi A., Legeai F., Hugueney P., Dasilva C., Horner D., Mica E., Jublot D., Poulain J., Bruyère C., Billault A., Segurens B., Gouyvenoux M., Ugarte E., Cattonaro F., Anthouard V., Vico V., Del Fabbro C., Alaux M., Di Gaspero G., Dumas V., Felice N., Paillard S., Juman I., Moroldo M., Scalabrin S., Canaguier A., Le Clainche I., Malacrida G., Durand

E., Pesole G., Laucou V., Chatelet P., Merdinoglu D., Delledonne M., Pezzotti M., Lecharny A., Scarpelli C., Artiguenave F., Pè M.E., Valle G., Morgante M., Caboche M., Adam-Blondon A.-F., Weissenbach J., Quétier F., Wincker P. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–468.

Jiao Y., Wickett N.L., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.

Jobb G., von Haeseler A., Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4:9 pp.

Jones G., Sagitov S., Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol. 62:1–12.

Khatoon S., Ali S.I. 1993. Chromosome atlas of the angiosperms of Pakistan. Karachi: Department of Botany, University of Karachi.

Kingman J.F.C. 1982. On the genealogy of large populations. J. Appl. Probab. 19A:27–43.

Kovar-Eder J., Kvaček Z., Meller B. 2001. Comparing Early to Middle Miocene floras and probable vegetation types of Oberdorf N Voitsberg (Austria), Bohemia (Czech Republic), and Wackersdorf (Germany). Rev. Palaeobot. Palynol. 114:83–125.

Kubatko L.S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.

Leal Perez-chao J., Valbuena A.O., Sotomayor S.P., Pascual M.L.R. 1980. Numeros cromosomicos para la flora Española 121–182. Lagascalia 9.

Leitch I.J., Bennett M.D. 2004. Genome downsizing in polyploid plants. Biol. J. Linn. Soc. 82:651–663.

Levin D.A. 2002. The role of chromosomal change in plant evolution. New York: Oxford University Press.

Liang G.X., Xing F.W. 2010. Infrageneric phylogeny of the genus *Viola* (Violaceae) based on *trnL-trnF*, *psbA-trnH*, *rpL16*, ITS sequences, cytological and morphological data. Acta Bot. Yunn. 32:477–488.

Liu Z., Pagani M., Zinniker D., DeConto R., Huber M., Brinkhuis H., Shah S.R., Leckie R.M., Pearson A. 2009. Global cooling during the Eocene-Oligocene climate transition. Science 323:1187–1190.

Lunn D.J., Thomas A., Best N., Spiegelhalter D. 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comp. 10:325–337.

Ma X.-F., Gustafson J.P. 2005. Genome evolution of allopolyploids: a process of cytological and genetic diploidization. Cytogen. Genome Res. 109:236–249.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mandáková T., Joly S., Krzywinski M., Mummenhoff K., Lysak M.A. 2010. Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. Plant Cell 22:2277–2290.

Marcussen T., Nordal I. 1998. *Viola suavis*, a new species in the Nordic flora, with analyses of the relation to other species in the subsection *Viola* (Violaceae). Nord. J. Bot. 18:221–237.

Marcussen T., Oxelman B., Skog A., Jakobsen K.S. 2010. Evolution of plant RNA polymerase IV/V genes: evidence of subneofunctionalization of duplicated *NRPD2/NRPE2*-like paralogs in *Viola* (Violaceae). BMC Evol. Biol. 10:45.

Marcussen T., Blaxland K., Windham M.D., Haskins K.E., Armstrong F. 2011. Establishing the phylogenetic origin, history and age of the narrow endemic *Viola guadalupensis* (Violaceae). Am. J. Bot. 98:1978–1988.

Marcussen T., Jakobsen K.S., Danihelka J., Ballard H.E., Blaxland K., Brysting A.K., Oxelman B. 2012. Inferring species networks from gene trees in high-polyploid North American and Hawaiian violets (*Viola*, Violaceae). Syst. Biol. 61:107–126.

Marcussen T., Sandve S.R., Heier L., Spannagl M., Pfeifer M., IWGSC, Jakobsen K.S., Wulff B.B.H., Steuernagel B., Mayer K.F.X., Olsen O.-A. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science 345:1250092.

Maureira-Butler I.J., Pfeil B.E., Muangprom A., Osborn T.C., Doyle J.J. 2008. The reticulate history of *Medicago* (Fabaceae). Syst. Biol. 57:466–482.

Mayrose I., Zhan S.H., Rothfels C.J., Magnuson-Ford K., Barker M.S., Rieseberg L.H., Otto S.P. 2011. Recently formed polyploid plants diversify at lower rates. Science 333:1257.

Meseguer A.S., Aldasoro J.J., Sanmartín I. 2013. Bayesian inference of phylogeny, morphology and range evolution reveals a complex evolutionary history in St. John's wort (*Hypericum*). Mol. Phylogen. Evol. 67:379–403.

Mohammadi Shahrestani M. 2013. Biosystematical investigation of sect. *Sclerosium* of genus *Viola* in Iran. M.Sc. thesis. Rasht: University of Rasht. 1–164 (In Persian, with English abstract).

Moore D.M. 1967. Chromosome numbers of Falkland islands angiosperms. Brit. Antarct. Surv. Bull. 14:69–82.

Morton J.K. 1993. Chromosome numbers and polyploidy in the flora of Cameroon mountain. Opera Bot. 121:159–172.

Müntzing A. 1932. Cyto-genetic investigations on synthetic *Galeopsis tetrahit*. Hereditas 16:105–154.

Nakamura K., Denda T., Kokubugata G., Huang C.-J., Peng C.-I., Yokota M. 2014. Phylogeny and biogeography of the *Viola iwagawae-tashiroi* species complex (Violaceae, section *Plagiostigma*) endemic to the Ryukyu Archipelago, Japan. Plant Syst. Evol.: in press.

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Otto S.P. 2007. The evolutionary consequences of polyploidy. Cell 131:452–462.

Ownbey M. 1950. Natural hybridization and amphiploidy in the genus *Tragopogon*. Am. J. Bot. 37:489–499.

Parham J.F., Donoghue P.C.J., Bell C.J., Calway T.D., Head J.J., Holroyd P.A., Inoue J.G., Irmis R.B., Joyce W.G., Ksepka D.T., Patané J.S.L., Smith N.D., Tarver J.E., van Tuinen M., Yang Z., Angielczyk K.D., Greenwood J.M., Hipsley C.A., Jacobs L., Makovicky P.J., Müller J., Smith K.T., Theodor J.M., Warnock R.C.M., Benton M.J. 2012. Best practices for justifying fossil calibrations. Syst. Biol. 61:346–359.

Rambaut A., Drummond A.J. 2009. Tracer v1.5. Available from: http://beast.bio.ed.ac.uk/Tracer.

Renoult J.P., Kjellberg F., Grout C., Santoni S., Khadari B. 2009. Cyto-nuclear discordance in the phylogeny of *Ficus* section *Galoglychia* and host shifts in plant-pollinator associations. BMC Evol. Biol. 9:248–266.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2011. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Roquet C., Sanmartín I., Garcia-Jacas N., Sáez L., Susanna A., Wikström N., Aldasoro J.J. 2009. Reconstructing the history of Campanulaceae with a Bayesian approach to molecular dating and dispersal-vicariance analyses. Mol. Phylogen. Evol. 52:575–587.

Rosenberg N.A., Feldman M.W. 2002. Chapter 9. The relationship between coalescence times and population divergence times. In: Slatkin M., Veuille M., editors. Modern Developments in Theoretical Population Genetics. Oxford: Oxford University Press, p. 130–164.

Rutschmann F. 2006. Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. Divers. Distrib. 12:35–48.

Sanso A.M., Seo M.N. 2005. Chromosomes of some Argentine angiosperms and their taxonomic significance. Caryologia 58:171–177.

Sauquet H., Ho S.Y.W., Gandolfo M.A., Jordan G.J., Wilf P., Cantrill D.J., Bayly M.J., Bromham L., Brown G.K., Carpenter R.J., Lee D.M., Murphy D.J., Sniderman J.M.K., Udovicic F. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales) Syst. Biol. 61:289–313.

Scheen A.C., Pfeil B.E., Petri A., Heidari N., Nylinder S., Oxelman B. 2012. Use of allele-specific sequencing primers is an efficient alternative to PCR subcloning of low-copy nuclear genes. Mol. Ecol. Resour. 12:128–135.

Schmidt A. 1964. Zur Systematischen Stellung von *Viola chelmea* Boiss. et Heldr. ssp. *chelmea* und *V. delphinantha* Boiss. Ber. deut. bot. Ges. 77:256–261.

Smith-White S. 1959. Cytological evolution in the Australian flora. Cold Spring Harb. Sym. Quant. Biol. 24:273–289.

Soltis D.E., Salcedo M.C.S., Thaden I.J., Majure L.M., Miles N.M., Mavrodiev E.V., Mei W., Cortez M.B., Soltis P.S., Gitzendanner M.A. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). New Phytol. 202:1105–1117.

Soltis D.E., Soltis P.S. 1993. Molecular data and the dynamic nature of polyploidy. Crit. Rev. Plant Sci. 12:243–273.

Tang H., Bowers J.E., Wang X., Paterson A.H. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc. Natl. Acad. Sci. USA 107:472–477.

Tate J., Soltis D.E., Soltis P.S. 2005. Polyploidy in plants. In: Gregory T.R., editor. The evolution of the genome. San Diego: Elsevier Academic Press. pp. 371–426.

Tokuoka T. 2008. Molecular phylogenetic analysis of Violaceae (Malpighiales) based on plastid and nuclear DNA sequences. J. Plant Res. 121:253–260.

Tokuoka T., Tobe H. 2006. Phylogenetic analyses of Malpighiales using plastid and nuclear DNA sequences, with particular reference to the embryology of Euphorbiaceae sens. str. J. Plant Res. 119:599–616.

Twyford A.D., Ennos R.A. 2012. Next-generation hybridization and introgression. Heredity 108:179–189.

van den Hof K., van den Berg R.G., Gravendeel B. 2008. Chalcone synthase gene lineage diversification confirms allopolyploid evolutionary relationships of European rostrate violets. Mol. Biol. Evol. 25:2099–2108.

Verlaque R., Espeut M. 2007. Violaceae. In: Marhold K., editor. IAPT/IOPT chromosome data 3. Taxon 56:E1.

Wahlert G.A., Marcussen T., de Paula-Souza J., Feng M., Ballard H.E. 2014. A phylogeny of the Violaceae (Malpighiales) inferred from plastid DNA sequences: implications for generic diversity and intrafamilial taxonomy. Syst. Bot. 39:239–252.

Wang H., Moore M.J., Soltis P.S., Belle C.D., Brockington S.F., Alexandre R., Davis C.C., Latvis M., Manchester S.R., Soltis D.E. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc. Natl. Acad. Sci. USA 106:3853–3858.

Wendel J.F., Doyle J.J. 2005. Polyploidy and evolution in plants. In: Henry R., editor. Diversity and evolution in plants. Oxon: CABI Publishing. pp. 97–117.

Wolfe K.H. 2001. Yesterday's polyploidization and mystery of diploidization. Nat. Rev. Genet. 2:333–341.

Wood T.E., Takebayashi N., Barker M.S., Mayrose I., Greenspoon P.B., Rieseberg L.H. 2009. The frequency of polyploid speciation in vascular plants. Proc. Natl. Acad. Sci. USA 106:13875–13879.

Xi Z., Ruhfel B.R., Schaefer H., Amorim A.M., Sugumarane M., Wurdack K.J., Endress P.K., Matthews M.L., Stevens P.F., Mathews S., Davis C.C. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc. Natl. Acad. Sci. USA 109:17519–17524.

Yang Z., Rannala B. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

Yockteng R., Ballard H.E., Mansion G., Dajoz I., Nadot S. 2003. Phylogenetic relationships among pansies (*Viola* section *Melanium*) investigated using ITS and ISSR markers. Plant Syst. Evol. 241:153–170.

Yoo K.-O., Jang S.-K. 2010. Infrageneric relationships of Korean *Viola* based on eight chloroplast markers. J. Syst. Evol. 48:474–481.

Yoo K.-O., Jang S.-K., Lee W.-T. 2005. Phylogeny of Korean *Viola* based on ITS sequences. Korean J. Plant Taxon. (Sigmul Bunryu Hag-hoeji) 35:7–23 (in Korean).

Yu Y., Than C., Degnan J., Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst. Biol. 60:138–149.

Yu Y., Barnett R.M., Nakhleh L. 2013. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Syst. Biol. 62:738–751.

Zwickl D.J., Stein J.C., Wing R.A., Ware D., Sanderson M.J. 2014. Disentangling methodological and biological sources of gene tree discordance on *Oryza* (Poaceae) chromosome 3. Syst. Biol. 63:645–659.