EVOLUTION AND BIODIVERSITY OF WILD POLYPLOIDS

EDITED BY: Elvira Hörandl, Natascha D. Wagner, Karol Marhold and Christoph Oberprieler

PUBLISHED IN: Frontiers in Plant Science

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-373-8
DOI 10.3389/978-2-88971-373-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

EVOLUTION AND BIODIVERSITY OF WILD POLYPLOIDS

Topic Editors:
Elvira Hörandl, University of Göttingen, Germany
Natascha D. Wagner, University of Göttingen, Germany
Karol Marhold, Center of Plant Science and Biodiversity, Slovakia
Christoph Oberprieler, University of Regensburg, Germany

Citation: Hörandl, E., Wagner, N. D., Marhold, K., Oberprieler, C., eds. (2021). Evolution and Biodiversity of Wild Polyploids. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-373-8

Table of Contents

frontiers in Plant Science

EDITORIAL
published: 20 July 2021
doi: 10.3389/fpls.2021.723439

Editorial: Evolution and Biodiversity of Wild Polyploids

Elvira Hörandl[1], Christoph Oberprieler[2], Karol Marhold[3,4] and Natascha D. Wagner[1]*

[1] Department of Systematics, Biodiversity and Evolution of Plants, University of Göttingen, Göttingen, Germany, [2] Institute of Plant Sciences, Evolutionary and Systematic Botany, University of Regensburg, Regensburg, Germany, [3] Institute of Botany, Plant Science and Biodiversity Centre, Slovak Academy of Sciences, Bratislava, Slovakia, [4] Department of Botany, Faculty of Science, Charles University, Prague, Czechia

Keywords: polyploidy, evolution, biogeography, biodiversity, repeatomes, transcriptomes

Editorial on the Research Topic

Evolution and Biodiversity of Wild Polyploids

Polyploidy is a major driver of evolution in plants. Angiosperms have undergone several polyploidization events in their evolutionary history (Leebens-Mack et al., 2019, Van De Peer et al., 2017) and polyploidy is an important factor causing speciation and influencing diversification of flowering plants (Abbott et al., 2013, Soltis et al., 2014b). However, the processes of evolutionary origin, establishment, speciation and post-origin evolution of polyploid plants are still not well-understood. Furthermore, most research on polyploids is performed on few model taxa or on crop plants. Overcoming the gap between cultivated model organisms and natural systems is crucial for a complete understanding of polyploid evolution. Here we present a symposium on polyploid evolution in wild plants including 10 case studies that can be grouped around three major themes: first, origins and evolutionary histories of polyploids; second, aspects of genome evolution in polyploids; and third, the relationships between polyploidy and biogeography and ecology.

Natural origins of polyploid species, their phylogenetic relationships, and the impact of polyploidy on diversification are major topics of biodiversity research. Melichárková et al. demonstrate multiple and recurrent origins of autopolyploids and inter-cytotype matings in the genus *Cardamine*. This neopolyploid complex diversified further *via* allopatry and ecological differentiation. Reconstructing phylogenies of genera including polyploids have been notoriously difficult due to the frequent origin of polyploid species *via* hybridization (allopolyploidy). In the genus *Salix*, Wagner et al. elucidate multiple polyploidization events and post-origin genome evolution of mesopolyploid species by using phylogenomic approaches. Their results supported allopolyploid origins and their establishment as distinct species. Paule et al. reconstruct genome size evolution in the big plant family *Bromeliaceae* based on the phylogeny of this family. Results suggest that polyploidy and dysploidy have played a role in the survival of early diverged lineages under extreme climatic conditions, but did not enhance net diversification. Finally, Meudt et al. analyze polyploid frequencies in floras of oceanic islands and use phylogenetic path analysis to estimate effects of polyploidy on diversification. The results suggest that endemic diversification and speciation of the analyzed island floras were shaped by polyploidy in many cases. These papers seem to support also the general pattern that young polyploid complexes tend more to increase diversification rates than older ones (Landis et al., 2018).

The evolution of polyploid genomes is highly dynamic (Comai, 2005; Chen, 2007), but many molecular consequences of whole genome duplication are still not well-understood. Zagorski et al. analyze the repeatome of *Hieracium* species and their synthetic F1 hybrids by NGS analyses and FISH. They find a marked influence of hybridization on satellite and rDNA dynamics, whereas transposable elements showed no significant bursts but rather dosage-dependent patterns.

OPEN ACCESS

Edited by:
Marcial Escudero, Sevilla University, Spain

Reviewed by:
João Loureiro, University of Coimbra, Portugal
Yves Van de Peer, Ghent University, Belgium
Martin A. Lysak, Masaryk University, Czechia

****Correspondence:***
Elvira Hörandl
elvira.hoerandl@biologie.uni-goettingen.de

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *10 June 2021*
Accepted: *23 June 2021*
Published: *20 July 2021*

Citation:
Hörandl E, Oberprieler C, Marhold K and Wagner ND (2021) Editorial: Evolution and Biodiversity of Wild Polyploids. Front. Plant Sci. 12:723439. doi: 10.3389/fpls.2021.723439

Analyses of the repeatome in the mesopolyploids of *Heliophila* by Dogan et al. reveal a highly dynamic evolution of retrotransposons and tandem repeats, reflecting the post-origin cladogenesis in the genus. Boatwright et al. analyze gene expression patterns of homoeologs in mesopolyploid populations of *Tragopogon* compared to their parental species and reveal similar and convergent expression changes, but also stochastic silencing of loci. Moreover, the authors found significantly more non-additive expression in mesopolyploids compared to recently formed neopolyploids.

Polyploids are thought to have advantages for ecological niche shifts (Marchant et al., 2016, Soltis et al., 2014b) and for range expansions, specifically to more northern regions (Rice et al., 2019). Buono et al. conducted phylogeographical analyses, niche modeling and crossing experiments on two European polyploid *Veronica* species and found that despite some introgression species boundaries are maintained by spatial and ecological isolation. Duchoslav et al. analyze the niche dynamics of various cytotypes of *Allium oleraceum* and find niche expansions in tetraploids, but not in higher ploidy levels. High ploidies rather showed unfilled niches and a tendency toward synanthropic habitats. Spoelhof et al. use a modeling approach to understand the establishment of polyploids in natural habitats by including spatial factors and relationships between organisms, habitat shape, and population density. Results suggest that narrow, constrained habitats such as roadsides and coastlines may enhance polyploid establishment.

The symposium highlights the diversity of evolutionary processes that are connected with polyploidy. Several papers point at the importance of post-origin evolution and diversification of polyploids, a topic that is hitherto been neglected in biodiversity research. We are at the infancies of understanding genome evolution and functionality in natural polyploid model systems. Here we present insights into repeatome evolution and gene expression studies. Whole genome sequencing and building reference genomes for wild model systems are highly needed to understand the consequences of polyploidization on genome structure and function. The symposium further shows various biogeographical and ecological scenarios for polyploids. Altogether, the case studies in this symposium refute the idea that polyploidization would be a dead end of evolution, as proposed by some researchers based on simple diversification statistics (Mayrose et al., 2011). The bias in such studies by undersampling, by neglecting reticulate evolution and different ages of diploids vs. polyploids was reviewed before (Soltis et al., 2014a). Careful case studies as presented here will provide a better understanding of the realized ecological and evolutionary potentials of polyploids.

AUTHOR CONTRIBUTIONS

EH wrote the Editorial with contributions from all co-authors. All authors contributed to the article and approved the submitted version.

REFERENCES

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., et al. (2013). Hybridization and speciation. *J. Evol. Biol.* 26, 229–246. doi: 10.1111/j.1420-9101.2012.02599.x

Chen, Z. J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Ann. Rev. Plant Biol.* 58, 377–406. doi: 10.1146/annurev.arplant.58.032806.103835

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Landis, J. B., Soltis, D. E., Li, Z., Marx, H. E., Barker, M. S., Tank, D. C., et al. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* 105, 348–363. doi: 10.1002/ajb2.1060

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574:679. doi: 10.1038/s41586-019-1693-2

Marchant, D. B., Soltis, D. E., and Soltis, P. S. (2016). Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *New Phytol.* 212, 708–718. doi: 10.1111/nph.14069

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., et al. (2011). Recently formed polyploid plants diversify at lower rates. *Science* 333, 1257–1257. doi: 10.1126/science.1207205

Rice, A., Smarda, P., Novosolov, M., Drori, M., Glick, L., Sabath, N., et al. (2019). The global biogeography of polyploid plants. *Nat. Ecol. Evol.* 3:265. doi: 10.1038/s41559-018-0787-9

Soltis, D. E., Segovia-Salcedo, M. C., Jordon-Thaden, I., Majure, L., Miles, N. M., Mavrodiev, E. V., et al. (2014a). Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytol.* 202, 1105–1117. doi: 10.1111/nph.12756

Soltis, D. E., Visger, C. J., and Soltis, P. S. (2014b). The polyploidy revolution then… and now: Stebbins revisited. *Am. J. Bot.* 101, 1057–1078. doi: 10.3732/ajb.1400178

Van De Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 17 July 2020
doi: 10.3389/fpls.2020.01077

Phylogenomic Relationships and Evolution of Polyploid Salix Species Revealed by RAD Sequencing Data

Natascha D. Wagner[1]*, *Li He*[2] *and Elvira Hörandl*[1]

[1] Department of Systematics, Biodiversity and Evolution of Plants (with Herbarium), University of Goettingen, Göttingen, Germany, [2] College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China

Polyploidy is common in the genus *Salix*. However, little is known about the origin, parentage and genomic composition of polyploid species because of a lack of suitable molecular markers and analysis tools. We established a phylogenomic framework including species of all described sections of Eurasian shrub willows. We analyzed the genomic composition of seven polyploid willow species in comparison to putative diploid parental species to draw conclusions on their origin and the effects of backcrossing and post-origin evolution. We applied recently developed programs like SNAPP, HyDe, and SNiPloid to establish a bioinformatic pipeline for unravelling the complexity of polyploid genomes. RAD sequencing revealed 23,393 loci and 320,010 high quality SNPs for the analysis of relationships of 35 species of Eurasian shrub willows (*Salix* subg. *Chamaetia/Vetrix*). Polyploid willow species appear to be predominantly of allopolyploid origin. More ancient allopolyploidization events were observed for two hexaploid and one octoploid species, while our data suggested a more recent allopolyploid origin for the included tetraploids and identified putative parental taxa. SNiPloid analyses disentangled the different genomic signatures resulting from hybrid origin, backcrossing, and secondary post-origin evolution in the polyploid species. Our RAD sequencing data demonstrate that willow genomes are shaped by ancient and recent reticulate evolution, polyploidization, and post-origin divergence of species.

Keywords: phylogenomics, polyploidy, restriction associated DNA sequencing, shrub willows, SNiPloid

OPEN ACCESS

Edited by:
Marcial Escudero,
Universidad de sevilla, Spain

Reviewed by:
Ales Kovarik,
Academy of Sciences of the Czech Republic (ASCR), Czechia
Michael R. McKain,
University of Alabama, United States
Yingnan Chen,
Nanjing Forestry University, China

***Correspondence:**
Natascha D. Wagner
natascha.wagner@uni-goettingen.de

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *14 April 2020*
Accepted: *30 June 2020*
Published: *17 July 2020*

Citation:
Wagner ND, He L and Hörandl E (2020) Phylogenomic Relationships and Evolution of Polyploid Salix Species Revealed by RAD Sequencing Data. Front. Plant Sci. 11:1077. doi: 10.3389/fpls.2020.01077

INTRODUCTION

Although the impact of polyploidy in plant evolution is evident (Soltis and Soltis, 2009; Wood et al., 2009; Jiao et al., 2011; Mayrose et al., 2011; Barker et al., 2016), there is much to discover about the impact of polyploidy on evolution of plant genomes. Polyploids may originate from combining genomes via hybridization (allopolyploids) or from intraspecific genome duplication (autopolyploids). Post-origin evolution and introgression may further leave genomic signatures (Peralta et al., 2013). This information, however, is crucial to understand the evolution of species. The potential offered by the rise of high-throughput sequencing tools to analyze the origin of natural polyploid plant species has not yet been fully exploited. The main reasons might be the difficulties in variant calling, especially the need to distinguish different alleles derived from the same parent

(homologs) from alleles originating from different parents (homeologs; Clevenger et al., 2015). Other difficulties are the necessity to filter paralogs (Hirsch and Buell, 2013), the uncertainty of allelic dosage as well as the lack of appropriate bioinformatic tools and models (Meirmans et al., 2018). Given a sufficient coverage, high-throughput sequencing methods are generally suited to detect all different alleles at a specific locus. However, to use a massive amount of short read data for the analysis of polyploids is still challenging, especially, if no reference genome is available.

Restriction Associated DNA (RAD) sequencing (Baird et al., 2008) is a widely used method that can harvest ten-thousands of informative SNPs representing the whole genome. The huge amount of biallelic SNPs is well suited to study evolutionary processes, resolve reticulate relationships, reconstruct species trees, and to analyze polyploid species in a phylogenetic framework (Bryant et al., 2012; Dufresne et al., 2014; Clevenger et al., 2015; Clevenger et al., 2018). However, there are only few studies using RAD sequencing data when dealing with polyploid non-model plants (Chen et al., 2013; Mastretta-Yanes et al., 2014; Qi et al., 2015; Brandrud et al., 2020; He et al., 2020).

The genus *Salix* L. (Salicaceae) comprises about 400-450 species of trees and shrubs that are mainly distributed in the Northern Hemisphere. About two thirds of the species occur in Eurasia, while about 140 species are distributed in America. Classical taxonomy and systematics in *Salix* have proven to be extremely difficult because of dioecy, simple, reduced flowers, common natural formation of hybrids, high intraspecific phenotypic variation, and the presence of polyploid species (Skvortsov, 1999; Hörandl et al., 2012; Cronk et al., 2015). Phylogenetic analyses using traditional Sanger sequencing and standard markers (like ITS) were able to separate subgenus *Salix* from a big clade uniting the two subgenera *Chamaetia* and *Vetrix*, but lacked any interspecific resolution (Leskinen and Alström-Rapaport, 1999; Chen et al., 2010; Percy et al., 2014; Lauron-Moreau et al., 2015; Wu et al., 2015). Especially in the case of polyploidization and frequent hybridization, the usage of uniparentally inherited plastid markers is useful for phylogenetic reconstruction. However, plastid markers show only little variation within willows (see Leskinen and Alstrom-Rapaport, 1999; Chen et al., 2010; Savage and Cavender-Bares, 2012; Percy et al., 2014; Lauron-Moreau et al., 2015; Wu et al., 2015). More recently, the complete plastomes of twelve *Salix* species were reconstructed and showed only 0.9% variability (N.D.Wagner, unpublished data). The *Chamaetia/Vetrix* clade comprises three quarters of species diversity in *Salix* ranging from creeping arctic-alpine dwarf shrubs to medium-sized trees (Wagner et al., 2018). In contrast to many other woody plant lineages about 40% of the species are polyploid (Suda and Argus, 1968), ranging from diploid to octoploid species (rarely deca-/dodecaploid), based on the chromosome number $x = 19$ (Argus, 1997). Previous studies proposed that ancient genome duplications occurred several times in *Salix* (Dorn, 1976; Leskinen and Alström-Rapaport, 1999) and in the Salicoid lineage (Tuskan et al., 2006). Paleopolyploid species are more difficult to recognize, because their cytological behaviour has undergone 'diploidization', i.e. they have returned to regular chromosome pairing at meiosis as typical for diploid sexual species (Comai, 2005). Cold climatic regimes during the Pleistocene may have triggered spontaneous autopolyploidization by unreduced gamete formation (Ramsey and Schemske, 1998). Recent allopolyploidization ('young polyploidy') may have happened during the Pleistocene, when climatic oscillations caused range fluctuations of plant species and enhanced secondary contact hybridization of previously isolated species (Hewitt, 2004; Abbott et al., 2013; Kadereit, 2015). While homoploid hybridization may result in introgressive hybridization with the parental species (Hardig et al., 2010; Gramlich and Hörandl, 2016; Gramlich et al., 2018), allopolyploidy can establish a rapid crossing barrier against the parents and hence may result in the evolution of an independent lineage. In willows, natural hybridization is a frequent phenomenon (Argus, 1997; Skvortsov, 1999; Hörandl et al., 2012) and may occur even between distantly related species (e.g. Hardig et al., 2000; Gramlich et al., 2018). While in the 19th century experimental crosses in *Salix* were popular and even triple or multiple hybrids have been produced (Wichura, 1865), the origin of most natural polyploid species is still unknown. Population genetic markers like SSRs and AFLPs have been used so far to study polyploids (Barcaccia et al., 2014; Guo et al., 2016), but in the *Chamaetia/Vetrix* clade the polyploid species have never been analyzed with molecular markers. We expect to trace signatures of ancient and/or recent hybrid origin, secondary lineage-specific evolution, backcrossing, and introgression in their genomes.

RAD sequencing (Baird et al., 2008) was recently used to overcome the lack of phylogenetic information within the genus *Salix* (Gramlich et al., 2018; Wagner et al., 2018; He et al., 2020). Wagner et al. (2018) published the first well-resolved phylogeny of diploid European members of the *Chamaetia/Vetrix* clade, while He et al. (2020) recognized a recent adaptive radiation of willows in the Hengduan mountains. Gramlich et al. (2018) successfully used RAD sequencing to analyze recent homoploid hybridization patterns and introgression between two *Salix* species on alpine glacier forefields. RAD loci in *Salix* represent almost exclusively the nuclear genome, and provide 10-thousands of SNPs from both conservative and rapidly evolving non-coding genomic regions (Gramlich et al., 2018). Therefore, this method has the potential to resolve reticulate relationships and to analyze the origin of the polyploid species in *Salix*.

In this study we want to test the utility of RAD sequencing data for the analysis of polyploid willow species. We will use a comprehensive approach of different traditional and recently developed analytical tools to elucidate the evolutionary scenarios of polyploid origin in Eurasian shrub willows. By using various approaches, we will 1) establish a phylogenomic framework of the *Vetrix/Chamaetia* clade, 2) test for recent allopolyploidization by hybridization of two (or more) related diploid parents versus more ancient allopolyploidization of distantly related progenitors of extant species, and 3) test for autopolyploidization. Based on this information, we 4) try to get insights into post-origin evolution of polyploid species by analyzing their genomic composition.

MATERIAL AND METHODS

Sampling

For this study, we sampled 28 diploid species (incl. two subspecies), one triploid, three tetraploid, two hexaploid and one octoploid species representing 19 sections sensu Skvortsov (1999) of the *Chamaetia*/*Vetrix* clade. *Salix triandra* (subg. *Salix*) was included to serve as outgroup (Wu et al., 2015). Hence, the complete sample set consists of 36 *Salix* species. The samples were collected in Central and Northern Europe as well as in China and determined after Fang et al. (1999); Skvortsov (1999) and Hörandl et al. (2012). The diploid species represent all main European lineages (Rechinger and Akeroyd, 1993). Leaves were dried in silica gel and herbarium voucher specimens were deposited in the herbarium of the University of Goettingen (GOET). Two to five accessions per species were included in the analyses given a total of 133 samples (**Supplement Table S1**).

Ploidy Determination Using Flow Cytometry

The ploidy levels of seven species (23 samples) with unknown chromosome counts were estimated using flow cytometric analyses based on DAPI (4', 6-diamidino-2-phenylindole) fluorochrome applying a modified protocol of Suda and Trávníček (2006) as described in He et al. (2020). *Salix caprea* with known ploidy level (2x=2n=38) was used as an external standard.

Molecular Treatment and Analyses

The DNA of all samples was extracted using the Qiagen DNeasy Plant Mini Kit following the manufacturer´s instructions (Valencia, CA). After quality check, the DNA was sent to Floragenex, Inc. (Portland, Ore., USA) where the sequencing library preparation was conducted after Baird et al. (2008) using the restriction enzyme *Pst*I (see Wagner et al., 2018 for details). Polyploids require an increased depth of coverage based on the bigger genome size and the higher number of alleles (Hirsch and Buell, 2013; Clevenger et al., 2015). Thus, we sequenced polyploid and diploid taxa on different plates to avoid loss of coverage.

There are several tools available to analyse RAD sequencing data. While e.g. STACKS (Catchen et al., 2013) is more suited to population genetic analyses, ipyrad is predominantly used for phylogenetic approaches (see Eaton et al., 2017). Ipyrad conducts quality filtering and de novo locus identification and genotyping, with the advantage that it can handle insertion-deletion variation among alleles, and it is therefore better suited for studies of a broader taxonomic scale (Andrews et al., 2016). The quality of the resulting single-end 100bp long sequence reads was checked using FastQC v.0.10.1 (Andrews, 2010). After de-multiplexing, the reads were used to run ipyrad v.0.7.28 (Eaton and Overcast, 2020) with a clustering threshold of 85% and a minimum depth of eight reads for base calling. The clustering was done with VSEARCH as implemented in ipyrad v.0.7.28. The maximum number of SNPs per locus was set to 20, the maximum number of indels to 8. We set a threshold of maximal four alleles per site in the final cluster filtering. Ipyrad summarizes the underlying allelic information into a consensus sequence in the form of ambiguous sites at heterozygous positions, which is a common strategy to face the problem of more than two alleles (Blischak et al., 2018b). The resulting sequence information can be used for downstream analyses. Initial settings for the minimum number of samples sharing a locus (m) ranged from m4 (average number of accessions per species) to m120 (loci present in 90% of accessions) and were optimized as described in Wagner et al. (2018). Statistics for the tested settings are summarized in **Supplement Table S2**. The m40 dataset (29.06% missing data) was used for phylogenetic analyses and the m100 dataset (9.19% missing data) for the genetic structure analysis. Additionally, we performed tree reconstruction with ipyrad for each clade separately that contains one or more polyploid samples applying the same settings.

We inferred phylogenetic relationships on concatenated alignments of the complete dataset as well as for each clade separately by using the GTR+ Γ model of nucleotide substitution implemented in RAxML v.8.2.4 (Stamatakis, 2014). We conducted for each ML analysis a rapid bootstrapping (BS) analysis with 100 replicates. Next to BS we applied *quartetsampling* (QS; Pease et al., 2018) with default settings to test statistical support of a given topology. We ran 300 replicates using the –L option (minimum likelihood differential). QS is able to distinguish between conflicting signals and poor phylogenetic information. The Quartet Concordance score [QC, (1,-1)] gives the support of the current topology, the Quartet Differential score [QD, (0,1)] is an indicator of introgression and the Quartet Informativeness score [QI, (0,1)] quantifies the informativeness of each branch. For each phylogeny shown here, the observed QS values (QC/QD/QI) were visualized along with the BS values above and below branches, respectively.

Species Tree Estimation Using SNAPP

The SNAPP method (Bryant et al., 2012) estimates species trees directly from biallelic markers (e.g., SNP data), and bypasses the necessity of sampling the gene trees at each locus. This is specifically advantageous for RAD sequencing data sets based on short loci. For species tree calculation with SNAPP v1.4.2 we used the unlinked SNPs output (one SNP per locus, uSNPs) of the clade-specific ipyrad pipeline. We created the input file using BEAUti v2.5.1 (Bouckaert et al., 2014). The multiple accessions of a species were assigned to one taxon. As priors we used default settings for speciation rate (lambda of 1/X). Non-polymorphic sites were excluded. We ran the analysis in BEAST2 (Bouckaert et al., 2014). The number of mcmc generations was up to 10,000,000 and we sampled every thousandth tree. The results were checked with TRACER 1.7.1 (Rambaut et al., 2018). The resulting trees were summarized in the SNAPP tree analyzer tool. The trees of all RAxML and SNAPP analyses were obtained in FigTree v1.4.3 (Rambaut, 2014).

Genetic Structure and Network Analyses

To test for an influence of reticulate evolution on the genetic composition of the included species, we explored the genetic structure of each sample. We simplified the data to be able to combine diploids and polyploids in the same STRUCTURE analyses by summarizing the observed allelic information of a polyploid to a "diploid" consensus sequence (e.g. the four alleles "AATT" will be reduced to "AT"). We used the uSNPs of the

m100 data set (< 10% missing data) for the complete dataset to avoid bias by too much missing data. The clade specific analyses were performed using the uSNP data of the clade specific ipyrad runs. We chose a burn-in of 10,000 and a MCMC of 100,000 replicates, with three iterations of each value of K (K=number of genotypic groups). After an initial test, the range of K was set from 2 to 7 in the final analysis. For the subclades, K was set from two up to the number of taxa included. The optimal K value was estimated by inferring the Evanno test revealing the optimal delta K value in Structure Harvester (Earl and vonHoldt, 2012).

Reticulate relationships resulting from hybridization or allopolyploidy are not well represented by bifurcating tree topologies (Huson and Bryant, 2006; McBreen and Lockhart, 2006). To overcome this problem, we utilized SplitsTree4 (Huson and Bryant, 2006) in order to reconstruct possible network-like evolutionary relationships among the species for each clade. Based on informative SNP data (see Feng et al., 2018), we generated the split network by implementing NeighbourNet analysis with variance of ordinary least squares complemented by a bootstrapping with 1000 replicates to test for statistical support. Missing data were treated as unknown in all analyses.

Hybridization Detection

HyDe (Hybridization Detection; Blischak et al., 2018a) allows testing for hybridization and introgression at a population or species level based on D-Statistics by estimating the amount of admixture (γ). We applied the "run_hyde_mp.py" script to test for putative parent-hybrid combinations with respect to the polyploid individuals. HyDe uses p-values to test for the significance of results. While a 50:50 hybrid is characterized by a γ-value of about 0.5, very low levels of admixture (e.g. 0.01 = close to parent P1; 0.99 = close to parent P2) may be indicators for several processes such as ILS and more ancient hybridization. After initial tests, we used a range of γ=0.4-0.6 to identify recent hybridization events in the data set, and intermediate ranges of γ= 0.1-0.4 and 0.6-0.9 for older events. We excluded significant values <0.1 and >0.9. The complete sequences as well as the uSNP data of the clade-specific analyses were used as input data, as the authors stated this datatype as the "most appropriate input" (Blischak et al., 2018a). We tested two approaches: first, we assigned all individuals of a species to one taxon, and second, we tested all individuals as separate entities. The latter approach is better to detect individual admixture. To underline the usage of HyDe in RAD sequencing data we performed a preliminary test on six diploid naturally formed F1 hybrids between *S. helvetica* and *S. purpurea* (Gramlich et al., 2018). The results revealed γ-values of 0.4-0.5 (complete sequence data) and 0.3-0.4 (uSNPs). These values are in accordance with recent studies which showed similar γ-values for modelled and empirical hybrid data (Blischak et al., 2018a; Zhang et al., 2019).

Categorization of SNPs Using the SNiPloid Pipeline

SNiPloid (Peralta et al., 2013) is a tool developed to analyze the SNP composition of recently established allotetraploid taxa emerging from diploid parental lineages. Originally, this approach is based on RNA sequencing data. By mapping the sequencing reads of an allotetraploid individual to a reference (diploid parent 1) and comparing the observed SNPs and read depth with the results of diploid parent 2, conclusions on the SNP composition of the tetraploid species can be drawn. SNiPloid is able to distinguish five different SNP categories (Peralta et al., 2013, **Figure 5A**). Category 1/2 = "inter-specific" SNPs describes that the observed allele in the polyploid is identical with only one of the parental species. Category 3/4 = "derived" SNPs is attributed, when the variation observed in the tetraploid is not identified between parental genomes. This hints to a mutation that occurred after the polyploidization event. Finally, category 5 corresponds to putative homeo-SNPs, i.e., the tetraploid is heterozygous for homeologous alleles of both parental genomes (**Figure 5A**). SNPs that do not fall into the categories mentioned above are categorized as "other".

To use the advantages of this pipeline and to endeavour the SNP composition of the tetraploids in our dataset we adapted the SNiPloid pipeline to RAD sequencing data (**Figure 1**). For our study, we used the concatenated consensus RAD loci of one putative diploid parental species as pseudoreference (similar as in Chen et al., 2013) for each test. The obtained RAD loci for every single individual (SAMPLE.consens file) were transferred into FASTA format. The indexing of the reference was done in BWA/0.7.12 (Li and Durbin, 2009). The sequence dictionary was created with PICARD/2.10.5 using the CreateSequenceDictionary tool (http://broadinstitute.github.io/picard/). The trimmed and filtered reads of the tetraploid species as well as of the second putative diploid parental species were mapped to the indexed reference applying the BWA/0.7.12 mapping tool using the MEM algorithm. The option -M (mark shorter split hits) for compatibility with PICARD was used, and 12 threads were applied for calculations (-t 12). The PICARD suite was used to sort the SAM file and to add read groups to the mapped alignment ('AddOrReplaceReadGroups'). The resulting BAM file was indexed with SAMTOOLS v. 1.8 (Li et al., 2009) and used as input for the Genome Analysis Toolkit [GATK 3.8; McKenna et al. (2010)] to analyze the read depth ("DepthOfCoverage") and the observed SNPs ("HaplotypeCaller"). For the tetraploid samples the—ploidy argument of HaplotypeCaller was set to 4. The generated read depth and the vcf files were eventually used as input for the executable perl script SNiPloid.pl (http://sniplay.southgreen.fr/cgi-bin/sniploid.cgi). The results of each tested combination were summarized in a table and pie diagram, given the percentage of each SNP category, and were then compared to each other. Recently established allotetraploid hybrids should show a high proportion of homeo-SNPs, because they should contain half of the alleles of diploid "parent1" and the other half of diploid "parent2" and should therefore be heterozygous for most loci. With ongoing speciation, we assume a higher proportion of cat 3/4 SNPs that are specific for the polyploid and originated after the polyploidization event. Backcrossing with one or both parents might lead to an increase of interspecific SNPs (cat 1/2), but these shared SNPs might also be the result of incomplete lineage sorting. The remaining "other" SNPs contain all other SNP combinations that cannot be assigned to any of the defined categories. This includes, e.g., SNPs that are heterozygous in the parental taxa and SNPs that show more than

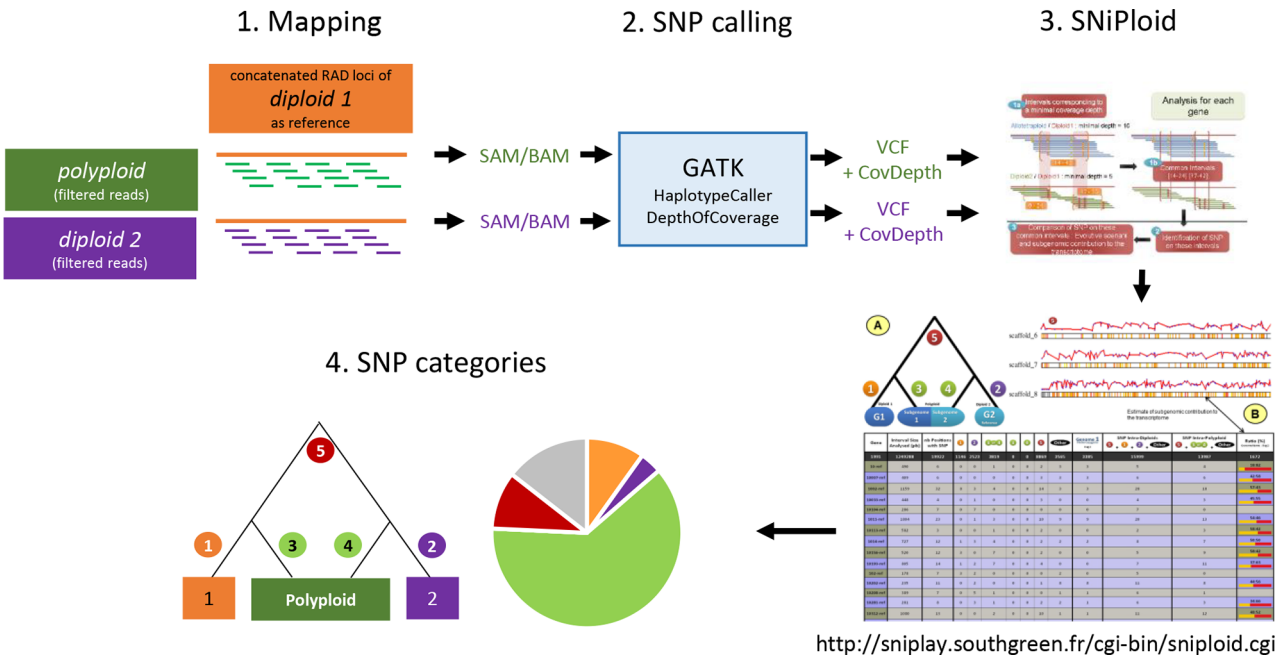

FIGURE 1 | SNiPloid workflow adapted from (Peralta et al., 2013). The filtered RAD sequencing reads of the polyploid and one putative parental species (*Diploid 2*) are mapped to the concatenated RAD loci of the second putative parental species (*Diploid 1*) that act as a "pseudo-reference". SNPs and coverage depth are determined with GATK tools "HaplotypeCaller" and "DepthOfCoverage". The resulting "VCF" and "CovDepth" files of both samples serve as input for SNiPloid. The output of SNiPloid contain the assigned SNP categories 1-5 for each compared SNP. Those data are finally summarized in a pie chart that shows the proportions of the observed categories.

two alleles. They partly represent shared ancient polymorphisms of the whole group.

We initially narrowed down the number of combinations by choosing the most likely putative parental species/lineages according to the results of RAxML, SNAPP, NeighbourNet, and STRUCTURE. These parental combinations were tested with SNiPloid (see **Table 1**).

RESULTS

Ploidy Determination

For 29 species of our sample set we could derive the ploidy level from chromosome counts available in the literature (**Supplement Table S1**). To determine the ploidy level of the remaining samples, we analyzed seven species via flow cytometry. Six of them, *S. integra, S. nummularia, S. pyrenaica, S. pyrolifolia, S. rehderiana*, and *S. schwerinii*, were analyzed for the first time. All analyzed species were diploid (**Table S1**). Selected histograms are available in **Supplement Figure S1**.

TABLE 1 | Allotetraploid species and their putative diploid parental species according to results of the phylogeny and the combination of NeighborNet, HyDe, STRUCTURE, and SNiPloid analyses.

Polyploid	Parent 1	Parent 2
Salix caesia (4x)	*S. purpurea*	*S. repens*
Salix cinerea (4x)	*S. aurita*	*S. appendiculata*
Salix laggeri (4x)	*S. appendiculata*	*S. caprea*

Phylogenetic Relationships

The average number of raw reads obtained from the RAD sequencing was 8.15 million reads per sample. The length after removal of adapters and barcode was 86 bp. After filtering, an average of 8.08 (+/-5.81) million reads per sample were used for clustering. An average of 161,233 clusters per sample was generated with an average depth of 52 reads per cluster. After optimization, we used a data set that consists of 23,393 RAD loci shared by at least 40 individuals and contained 320,010 variable sites of which 191,615 are parsimony informative. The concatenated alignment had a length of 1,931,205 bp with 29.06% missing data (**Supplement Table S2**). The analysis with a clustering threshold of loci shared by at least 100 individuals (m100) revealed 1,660 loci with 16,761 SNPs (**Supplement Table S2**). The amount of missing data was only 9.1%, however, the dataset was not suitable to resolve the phylogenetic relationships (see **Supplement Figure S5**). Therefore, we used this dataset only for a genetic structure analysis with STRUCTURE, which is sensitive to missing data. The HyDe results of the m40 dataset revealed 4,724 significant hybridization events of 21,420 tested combination in the taxon-assigned test based on the uSNP data. Of these, 38.6% belonged to combinations that treated a polyploid taxon as "hybrid". The average γ-value of all observed events was 0.52.

The RAxML phylogeny included 133 accessions representing 35 species of the *Chamaetia/Vetrix* clade from Europe and Asia as well as four accessions of *S. triandra* as outgroup (**Figure 2**, **Supplement Figure S2**). The phylogeny revealed that all species are clearly monophyletic. The topology showed *S. reticulata* in sister position to four well-supported clades. For consistency, we used the same Roman numerals for the main clades as in Wagner et al. (2018). Clade I comprised ten species, including two well supported subclades. Clade II contained all members of section *Vetrix* as well as *S. eleagnos* and *S. serpyllifolia*, clade III comprised all members of sections *Villosae* and *Vimen* as well as triploid *S. bicolor*, and clade IV contained *S. hastata*, *S. herbacea*, *S. pyrenaica* and *S. lanata*. The tetraploid species fell into clades I and II, the triploid *S. bicolor* into clade III. The octoploid *S. glaucosericea* was in sister position to clade III and IV while hexaploid *S. glabra* was sister to *S. eleagnos* and *S. serpyllifolia* in clade II. Hexaploid *S. myrsinifolia* was in sister position to the remaining accessions of clade II. The genetic structure analysis of the complete sampling based on 1,655 uSNPs (m100) reflected the four observed main clades (I-IV) and indicated genetic admixture of up to four clusters for the high polyploids *S. glaucosericea* and *S. myrsinifolia*. Hexaploid *S. glabra* showed admixture of "clade II" and "clade IV" (**Supplement Figure S4**).

Each clade that contained one or more polyploid samples was analyzed separately with the minimum number of accessions

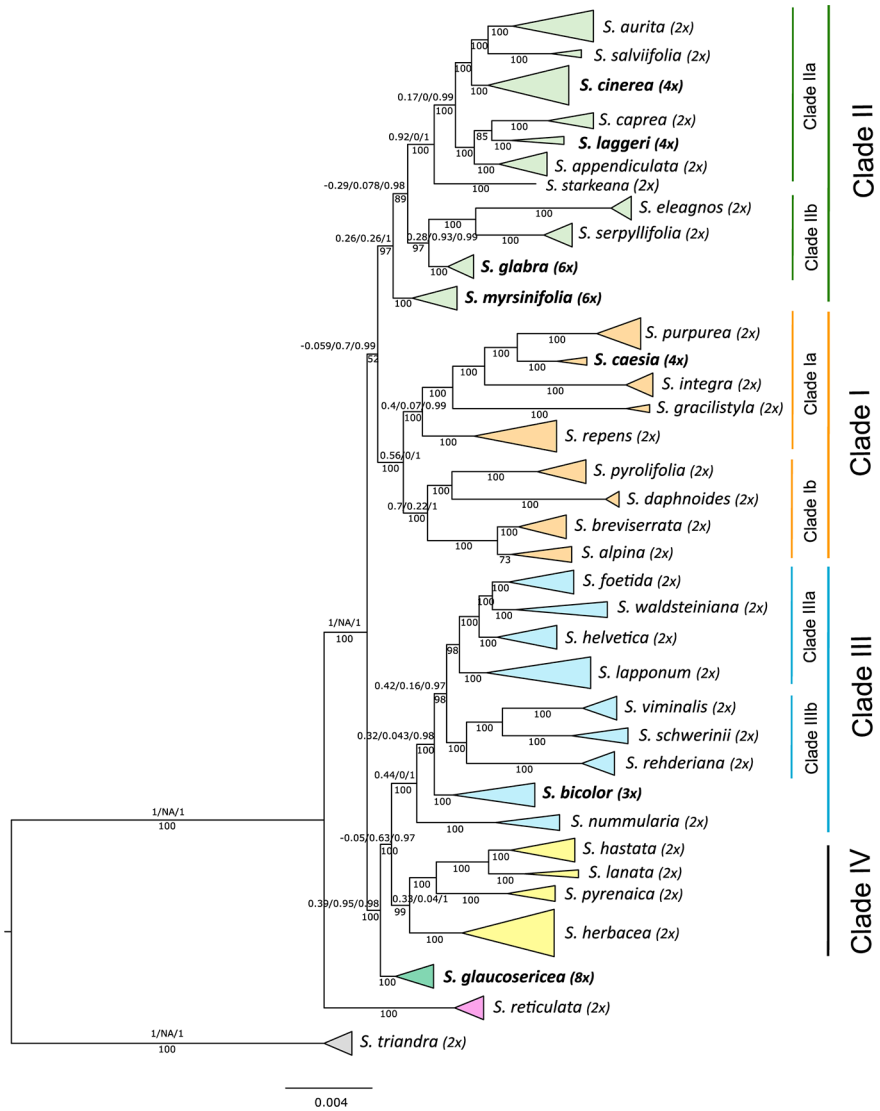

FIGURE 2 | Simplified RAxML phylogeny of 133 accessions representing 35 species of *Salix* subg. *Chamaetia/Vetrix* based on 23,393 RAD loci. Four clades (numbering follows Wagner et al., 2018) are highlighted in different colours, subclades for detailed analyses are also designated. Ploidy level is indicated behind species name, polyploid samples are highlighted in bold. *Salix triandra* (subg. *Salix*) was used to root the tree. BS values below branches, QS values of selected clades and subclades above branches. A detailed phylogeny including QS support values for all branches is supplied in **Supplement Figure S2**.

sharing a locus (m) dependent on the different overall number of accessions in each clade. These were m15 for clade Ia, m23 for clade II and m10 for clade III, respectively.

Relationships of Polyploids in Clade I

In the RAxML phylogeny, clade I was divided into two subclades (**Figure 2**). The supported monophyletic subclade Ia (BS 100, QS 0.42/0.073/0.99) comprised the species *S. purpurea, S. caesia, S. integra* which belong to section *Helix* sensu Skvortsov (1999), as well as *S. gracilistyla.* They were in sister position to the closely related species *S. repens* (incl. *S. repens ssp. rosmarinifolia;* sect. *Incubaceae* sensu Skvortsov (1999)). Subclade Ib comprised the dwarf shrub species *S. breviserrata* and *S. alpina*, as well as the medium-sized species *S. pyrolifolia* and *S. daphnoides*. The tetraploid *S. caesia* appeared in a well-supported sister position to *S. purpurea* (BS 100, QS 0.18/0/0.99). The analysis of the 18 accessions of subclade Ia with two accessions of *S. reticulata* as outgroup based on loci shared by at least 15 accessions, contained 38,603 shared RAD loci with 238,614 SNPs. The SNAPP (**Figure 3A**) and RAxML topology (**Figure 3B**) show strong statistical support for the sister relationship of *S. purpurea* and *S. caesia*. Using 37,867 uSNPs as input and designating *S. reticulata* as outgroup, HyDe performed a test of 2,448 combinations in the individual approach of which 164 were detected as significant hybridization events. Only six of 30 tested events in the assigned taxa approach were significant. HyDe showed no significant hybridization event with *S. caesia* as hybrid. Thus, we repeated the analysis with the complete sequence data of the clade specific analysis comprising 238,614 SNPs. Here, HyDe detected 606 significant results of 2,448 tested combinations in the individual approach with an average γ-value of 0.59. In 85 hybridization events (14%) *S. caesia* was counted as "hybrid". For these events the average observed γ-value was 0.65. The average γ-value was 0.8 in 20 significant combinations with *S. purpurea* and *S. repens* as parental species. In the m15 dataset, the 18 accessions of clade I (without outgroup) shared 35,078 RAD loci comprising 191,211 SNPs. The genetic structure analysis based on 34,071 uSNPs for the most likely K-value of four revealed genetic admixture in *S. caesia* with about 50% of each the genetic partition of *S. purpurea* and *S. repens* (**Figure 3D**). In the NeighbourNet analysis (**Figure 3C**), *S. caesia* was situated between *S. purpurea* and *S. repens*. To test a hypothesis of allopolyploid origin of *S. caesia* in more detail, we conducted a SNiPloid analysis using *S. purpurea* and *S. repens* as putative diploid parents of putative allopolyploid *S. caesia*. The results revealed 5.78% cat 1 and 12.65% cat 2 interspecific SNPs, 44.52% homeo-SNPs (cat 5), and 19.17% cat 3/4 SNPs. About 17.89% of SNPs did not fall into the given categories and were treated as "others" (**Figure 5B**).

Relationships of Polyploids in Clade II

Clade II contained hexaploid *S. myrsinifolia* and *S. glabra*, both monophyletic. The latter was in sister position to *S. eleagnos* and *S. serpyllifolia* forming subclade IIb. *Salix myrsinifolia* was in sister position to the remainders of clade II, with good BS but a skew to an alternative topology (BS 97, QS 0.28/0.93/0.99; **Figure 2**). All included members of section *Vetrix* s.l. [sensu Skvortsov (1999)] formed a well-supported monophyletic group (subclade IIa; BS 100, QS 0.92/0/1, **Figure 2**) including two tetraploid species, *S. laggeri* and *S. cinerea*. The clade-specific analysis of subclade IIa included 25 accessions and was based on 33,515 RAD loci comprising 254,819 SNPs shared by at least 23 samples (m23). The species were all monophyletic and well supported (**Figure 4B**). One accession each of *S. eleagnos* and *S. serpillifolia* served as outgroup. *Salix starkeana* was in sister position to the remaining accessions of clade IIa (BS100, 1/NA/1). The backbone of the tree topology showed some less supported branches, especially with regard to the QS values that indicate incongruencies of different branching topologies.

In the taxon-assigned approach, HyDe tested 104 combinations and yielded only four significant events. When we treated the samples as individuals, 5,313 combinations were tested and revealed 106 significant hybridization events. In 24 events (23%), *S. cinerea* was treated as 'hybrid'. The average γ-value for these events was 0.46. For the parental combination of *S. appendiculata* and *S. aurita*, the observed nine significant hybrid combinations showed an average γ-value of 0.49. No significant hybrid event was observed for tetraploid *S. laggeri*.

The tetraploid species *S. cinerea* was in well-supported sister-position to *S. aurita* and *S. salviifolia* in both the complete and the clade specific RAxML phylogeny (**Figures 2** and **4B**), while *S. appendiculata* was sister to this clade. *Salix cinerea* was in sister position to *S. aurita* in the SNAPP species tree (**Figure 4A**) and in close relationship to *S. aurita* in the NeighbourNet analysis (**Figure 4C**). The genetic structure analysis of 39,746 uSNPs (without outgroup) for the most likely K-value of three revealed genetic admixture for all samples (**Figure 4D**). *Salix cinerea* showed a similar structure as *S. aurita* and *S. salviifolia*. The SNiPloid results for *S. cinerea* using *S. aurita* and *S. appendiculata* as potential diploid parental species, showed about 5.6% homeo-SNPs and 5.6% cat 1 and 5.7% cat 2 SNPs, respectively, whereas about 47.5% were heterozygous sites sharing one allele of only one parent (cat 3/4 SNPs); 35.3 % of SNPs did not fall into the given categories (**Figure 5D**).

Tetraploid *S. laggeri* appeared in moderately supported sister position to *S. caprea* in both the clade-specific and in the overall RAxML phylogeny (**Figures 2** and **4B**). The NeighbourNet showed *S. laggeri* in close relationship to *S. appendiculata* and *S. caprea* (**Figure 4C**). The genetic structure analysis of subclade IIa for K=3 revealed a similar genetic composition as *S. appendiculata*, with some admixture of the *S. caprea* specific partition (**Figure 4D**). The SNiPloid analysis with *S. caprea* and *S. appendiculata* as putative diploid parents revealed 5.6% homeoSNPs and about 12.2% interspecific SNPs (5.3% cat 1, 6.9% cat 2). About 36.6% of SNPs were shared with one allele of one parent (cat 3/4), while 45.9% of SNPs were not categorized (**Figure 5C**).

Relationships of Polyploids in Clade III

Nine species grouped into clade III, which is monophyletic and well supported (BS 98, QS 0.43/0/1). The clade-specific analysis was based on 37 accessions with a minimum of 30 individuals sharing a locus and revealed 16,463 RAD loci containing 97,576 SNPs. *Salix reticulata* was used to root the RAxML tree. As

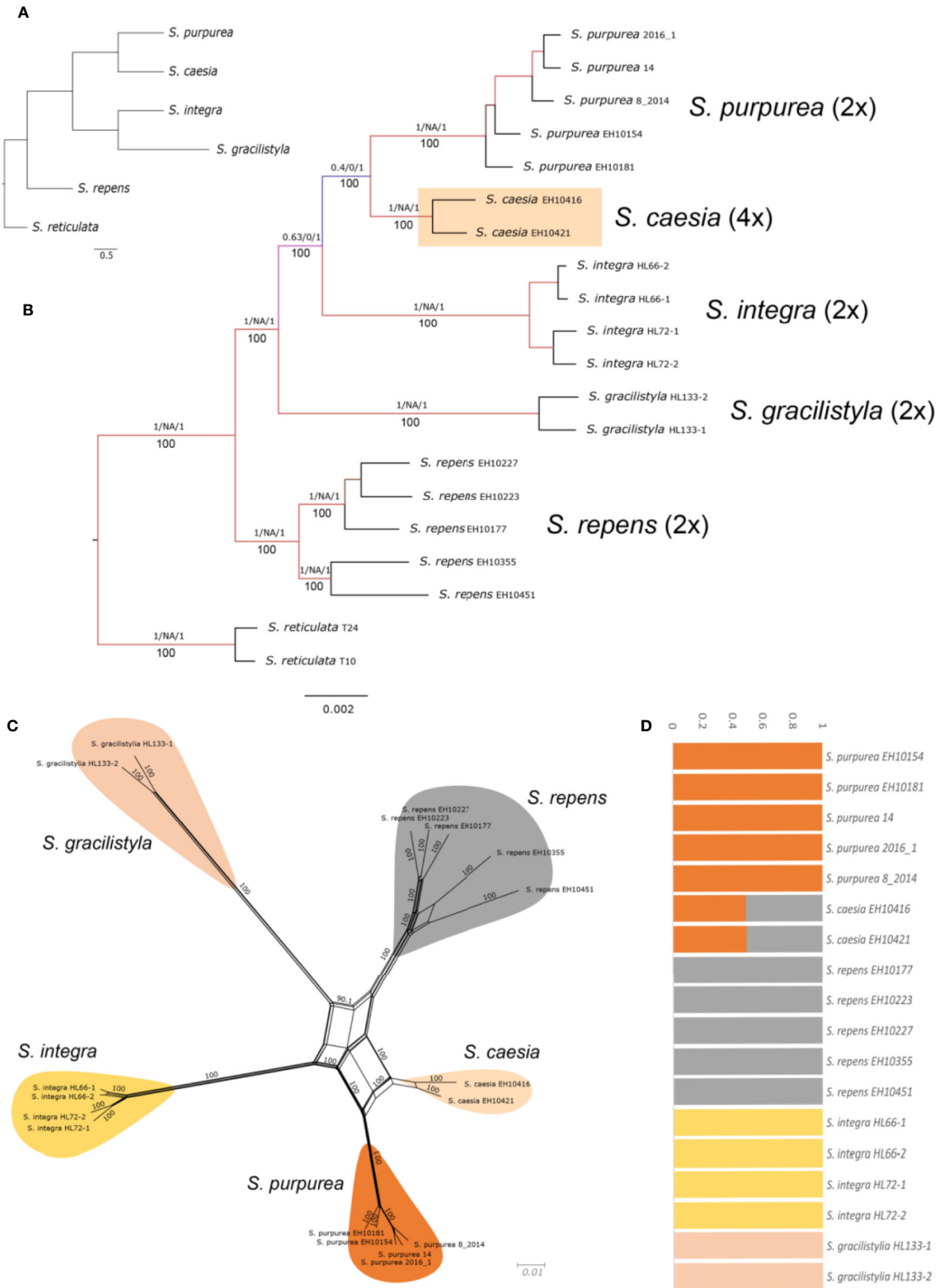

FIGURE 3 | Relationships of tetraploid *S. caesia* in clade Ia. **(A)** SNAPP species tree. **(B)** RAxML phylogeny based on 38,603 RAD loci and 238,614 SNPs, respectively. Bootstrap support values below, QS values above branches. **(C)** Resulting Splitsgraph of the NeighbourNet analysis of subclade Ia. Bootstrap values indicated at branches. **(D)** Genetic structure analysis for the subclade for the most likely value of K=4 based on 37,867 unlinked SNPs.

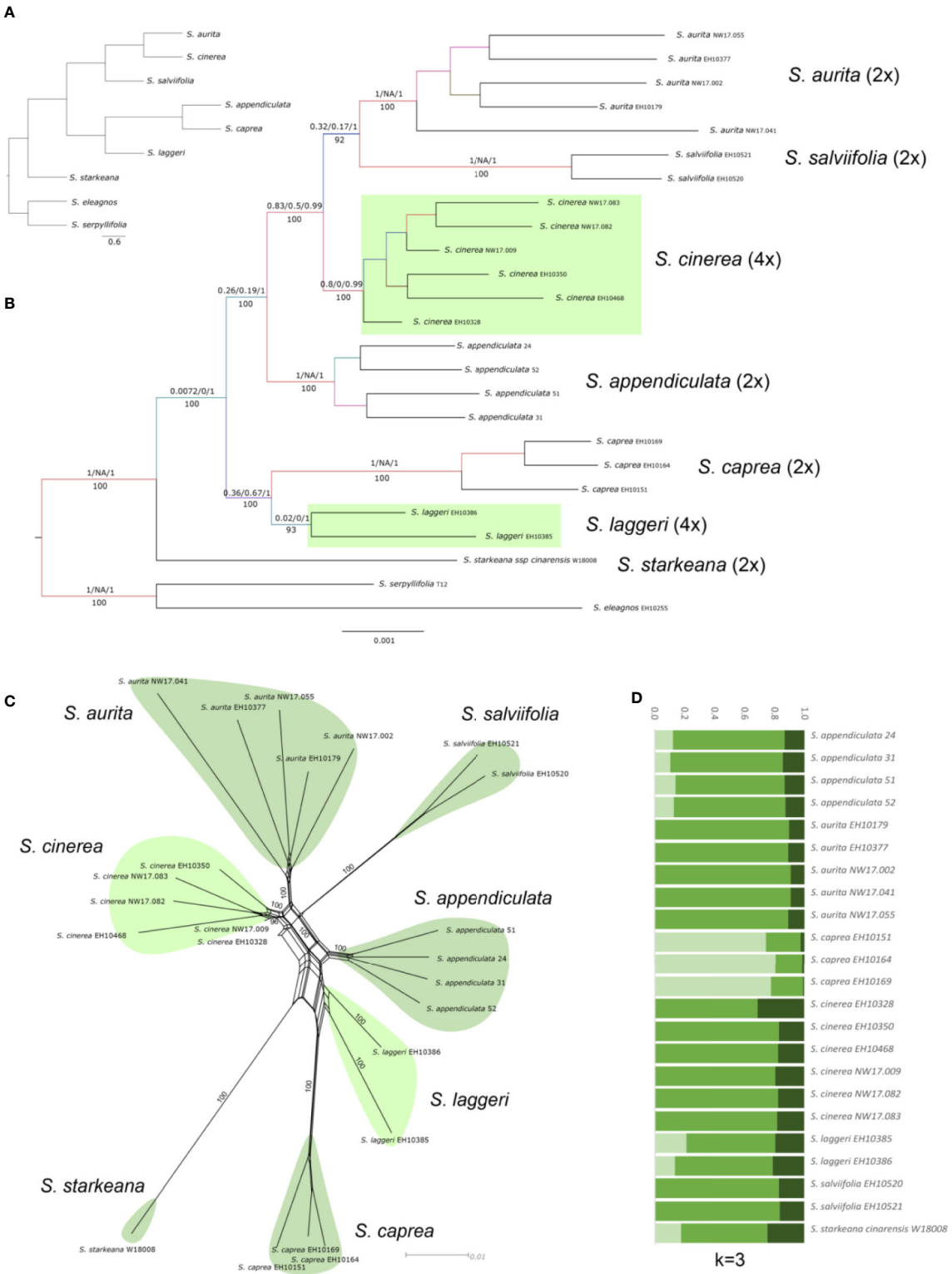

FIGURE 4 | Relationships of polyploids in subclade IIa. **(A)** most abundant SNAPP species tree. **(B)** RAxML phylogeny based on 33,515 RAD loci, with QS support values above and BS values below branches Tetraploid species indicated with light green boxes. **(C)** Splitsgraph of the NeighbourNet analysis of clade IIa, BS values given at branches. **(D)** Results of the genetic structure analysis for subclade IIa (K=3) based on 39,746 unlinked SNPs.

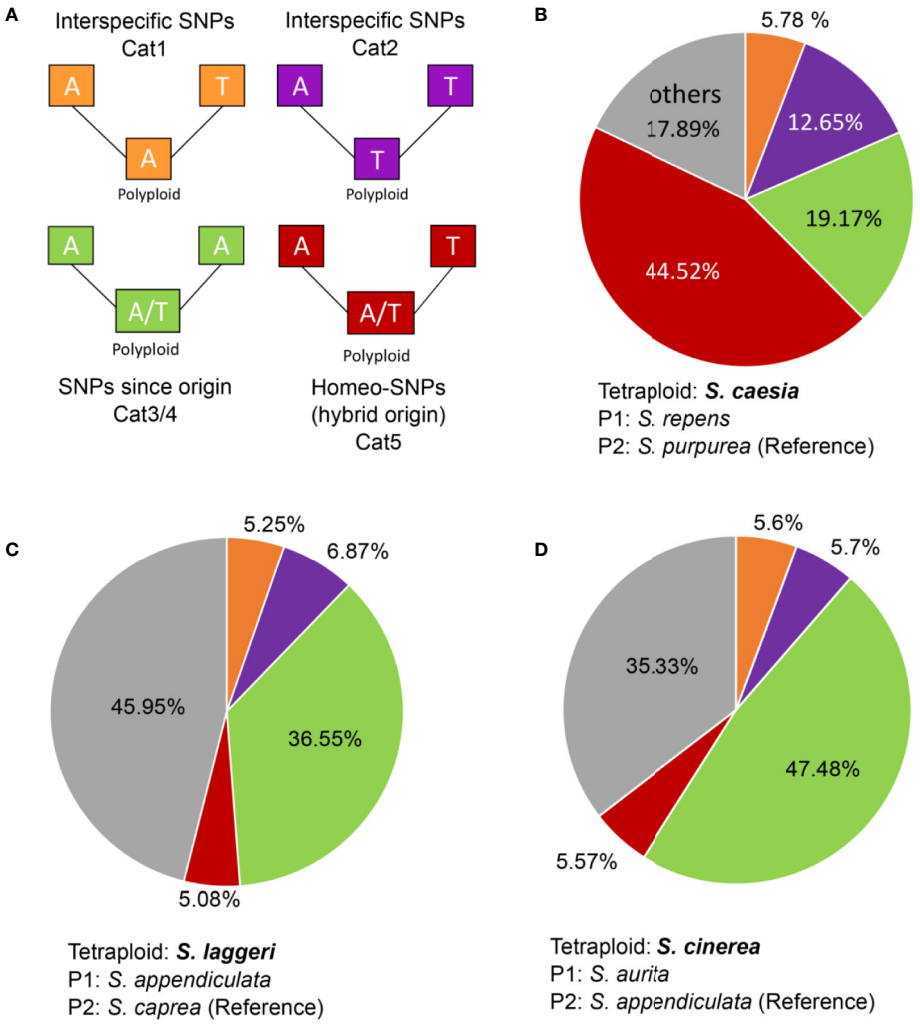

FIGURE 5 | SNiPloid results for three tetraploid species. The colours in the pie diagrams represent the proportions of the different observed SNP categories: Cat1, orange, Cat2, lilac (both interspecific SNPs), Cat3/4, green (post-origin SNPs), Cat5, red (homeo-SNPs). Grey indicates the proportion of observed SNPs not falling into the five specified categories. The legend **(A)** shows an example of SNP categorization by using the same color (after Peralta et al., 2013) code. The parental samples (parent1 and parent2) are diploids. The polyploid SNPs are categorized by comparing the SNP composition with both parents. **(B)** *S. caesia*, with *S. repens* and *S. purpurea* as putative parental species, **(C)** *S. laggeri*, with *S. appendiculata* and *S. caprea* as putative parental species, and **(D)** *S. cinerea* with *S. aurita* and *S. appendiculata* as putative parental species.

illustrated in **Figure S3**, *S. nummularia* was in sister position to all remaining members of this clade followed by *S. bicolor*, which is triploid and the only polyploid species in this clade. Two subclades diverged: subclade IIIa contained the species *S. helvetica*, *S. waldsteiniana*, *S. foetida*, and *S. lapponum*, while subclade IIIb consisted of *S. viminalis*, *S. schwerinii*, and *S. rehderiana*. Each species was clearly monophyletic, and morphological sister relationships were confirmed. In the clade-specific structure analysis for the most likely K-value of five (**Supplement Figure S3**), *S. bicolor* showed genetic admixture between *S. nummularia* and subclade IIIa. The amount of admixture differed between the two samples from the Eastern Alps in Austria (about two thirds shared with subclade IIIa) and the two samples from the Harz (Brocken) in Germany (about half shared with subclade IIIa). The deep split between the two geographical regions is also reflected in the NeighbourNet network. SNiPloid was not used to analyze the putative parenthood of *S. bicolor*, for the tool is optimized for tetraploid species (containing two subgenomes) and therefore not suitable for triploid species.

DISCUSSION

RAD Sequencing Data and Analysis Pipelines for Polyploids

Reduced representation libraries like RAD sequencing are frequently used to analyze intraspecific population structure,

closely related species groups and to infer phylogenetic relationships of diverged lineages based on high numbers of SNPs (e.g. Cariou et al., 2013; Andrews et al., 2016; Eaton et al., 2017). However, the use of RAD sequencing for polyploid species is still hampered by the lack of suitable tools and the statistical difficulties of dealing with more than two alleles [reviewed in Clevenger et al. (2015)]. Thus, only few studies using RAD sequencing on polyploid species are published so far (Mastretta-Yanes et al., 2014; Qi et al., 2015; Brandrud et al., 2017; Feng et al., 2018; Brandrud et al., 2020). Here we present a phylogenomic study on diploid and polyploid willow species by using a de-novo assembly of RAD sequencing data.

Our assembly and concatenation of short sequenced fragments representing the whole genome resulted in a robust, fully resolved phylogeny for the *Chamaetia/Vetrix* clade. The method of concatenation has been often criticized, especially in case of conflicting signal among genomic regions. However, Rivers et al. (2016) showed that the data output of the concatenated RAD loci including thousands of unlinked SNPs is able to recover a robust tree. Our findings support the suitability of RAD sequencing for interspecific phylogenetic inference and are in accordance with many studies on different levels of divergence (e.g. Hipp et al., 2014; Eaton et al., 2017; Wagner et al., 2018; He et al., 2020). The allelic information of the polyploid samples can be condensed to a single consensus sequence to circumvent the challenges of dealing with more than two alleles for phylogenetic approaches. However, the loss of information caused by this simplification might lead to wrong placement of the polyploid accessions in the phylogeny (Eriksson et al., 2018; Andermann et al., 2019). Our results showed that the RAD sequencing data contain enough information to resolve the phylogenetic relationships in *Salix* and that condensed allelic information was appropriate for the included allopolyploid species. While for phylogenetic approaches complete sequence information is preferable (Leaché et al., 2015), many unlinked loci are the preferable source of information for analyses, where the sites are treated as independently evolving genetic entities. Thus, we used the unlinked SNPs for network and genetic structure analyses. The uSNPs were also successfully used to reconstruct the species tree of each subclade using a coalescence-based method implemented in SNAPP (Bryant et al., 2012). Our results confirmed that the coalescent approach using RAD sequencing data is suitable for species delimitation, as recently shown by Brandrud et al. (2020).

HyDe (Blischak et al., 2018a) is a tool to detect hybridization and we showed that it is possible to use it with RAD sequencing data. We used HyDe to draw conclusions on the putative parental species of allopolyploids. First, we used the uSNP data, as suggested by Blischak et al. (2018a). The complete alignment using all SNPs, however, revealed a higher amount of significant results. Therewith we confirm the findings of Blischak et al. (2018a) that the number of significant hybridization events depends on the size of the input data and more input data reveal more accurate results. HyDe detected 4,724 significant hybridization events in the complete dataset. About 40% of them were combinations including a polyploid as "hybrid". That means that 60% were indicated as hybridization events between diploid samples. The high amount of natural hybridization in *Salix* is well known (Skvortsov, 1999; Argus, 2010), and homoploid hybridization even between distantly related species has been documented (Hardig et al., 2000; Gramlich et al., 2018). However, in this study we observe clearly distinct monophyletic groups for all included species and the sampled accessions represent morphologically definite individuals. We assume that HyDe uncovers a high amount of ancient gene flow that resulted in a shared polymorphism in the whole phylogeny and its subclades. We further suppose that non-visible introgression or, that the huge amount of input data cause significant number of "false" hybrid combinations.

SNiPloid (Peralta et al., 2013) was used to reconstruct the genomic constitution of the tetraploid species. SNiPloid was originally developed for transcriptomic data (Peralta et al., 2013). The application of this tool to RAD sequencing data provided us valuable information about the contributions of putative parental species to the tetraploid genome. In combination with genetic structure, HyDe and NeighborNet analyses, we could test for allopolyploid origin and revealed some insights into genome evolution after the polyploidization events. Unfortunately, the tool is only suitable for tetraploid species that consist of not more than two subgenomes (Peralta et al., 2013). However, for RAD sequencing data SNiPloid provides an alternative approach by using biallelic SNPs instead of sequence data to reveal potential parenthood of allotetraploids without previous knowledge. For the interpretation of the SNiPloid data it must be considered that in most cases, we did not test the "real" parental accessions. Our results are instead based on putative parental species/lineages, and we therefore expect a certain amount of SNPs that do not fall into the narrowly defined SNP categories. This will lead to an increase of proportions of non-categorized "other" SNPs. Furthermore, proportions of SNP categories must be interpreted with caution as also the intra-specific sampling can influence the actual amounts of detected SNPs for the respective categories. Nevertheless, we can present here a first overview of the origin and evolution of polyploid species within a phylogenetic framework. This increases the potential to analyze polyploid samples with reduced representation methods.

Phylogenetic Relationships and Origins of Polyploids in *Salix*

The RAD sequencing data revealed a well-resolved phylogeny of the *Chamaetia/Vetrix* clade including 35 Eurasian species plus *S. triandra* as outgroup and thus provide a sufficient framework for reconstructing the origin of the polyploid species. The presented phylogeny included members of all sections sensu Skvortsov (1999) and hence covered the morphological diversity and biogeographical range of Eurasian willows. All of the 36 included species are monophyletic. The isolated position of *S. reticulata* and the observed four clades are in accordance with a

former study on European diploid species by Wagner et al. (2018). The samples from Asia fall into the four clades.

About 40% of *Salix* species are polyploid (Suda and Argus, 1968) ranging from tetraploid to octoploid, rarely to decaploid. Our flow cytometry results showed a diploid level for the six included species from Asia that were previously not investigated (**Supplement Table 1**, **Supplement Figure S1**). Although some studies are published about tetraploid species in subgenus *Salix* s.l. (Triest et al., 2000; Triest, 2001; Barcaccia et al., 2014) no molecular studies existed so far on the origin of polyploid species of the *Chamaetia/Vetrix* clade. In this study we included seven polyploids from Europe with different ploidy levels to study different scenarios of their evolutionary origin. For the polyploid species included here, no diploid cytotypes have ever been reported, and they all exhibit very distinct morphologies and a genetic structure that is composed of several genetic clusters. These features make autopolyploid origins unlikely. The included polyploid species are all monophyletic. However, the species appear scattered over the phylogeny, indicating multiple independent origins of polyploids within the genus resulting from different parental combinations. Thereby *Salix* differs from plant genera in which a single allopolyploidization event resulted in post-origin adaptive radiation and speciation (e.g., the Hawaiian Silversword alliance, Seehausen, 2004). Willows rather resemble plant genera with repeated independent polyploidization events, like in *Nicotiana* (Chase et al., 2003), *Achillea* (Guo et al., 2013), *Ranunculus* (Baltisberger and Hörandl, 2016), and *Dactylorhiza* (Brandrud et al., 2020), among others.

The higher polyploid willows are in sister position to major clades or on basal branches. This could be explained by an old allopolyploid origin of these species. We would expect this pattern if two or more ancestral lineages of the extant major subclades (I-IV) had contributed to the polyploid genome. However, these ancestral lineages are probably not conspecific to any of the extant species. According to the crown group age of the *Chamaetia/Vetrix* clade (23.76 Ma, Wu et al., 2015), their origins may date back to the late Miocene/ early Pliocene. However, next to that, the hybrid nature of their origin might be responsible for the placement of the polyploids on basal branches of the subclades in the phylogeny. Our structure analyses indicated that the hexa- and octopolyploids, but also triploid *S. bicolor*, harbor two to four genetic partitions that characterize otherwise the main subclades I-IV (see **Figure S4**). Hence, we assume allopolyploid origins from divergent lineages. However, since the genetic structure analyses cannot disentangle ancient and recent hybridization or introgression events, further studies on the origin on the high polyploid willows are necessary.

Origin and Evolution of the Allotetraploid Species

The tetraploids fall into the observed clades (**Figure 2**). Since all included tetraploid species are monophyletic and do form well-supported entities in our analysis, we do not assume multiple origins with different parental combinations. Instead, we assume a single or few hybrid (allopolyploid) origin(s) from a pool of related individuals (Abbott et al., 2013), followed by speciation. We are aware that extant species might not represent the true parental species, and that extinct taxa might have been also involved. Nevertheless, by testing different combinations of putative diploid parents that are the next extant relatives in our observed trees, we analyzed potential parental lineages that might have contributed to allotetraploid species formation. The comparison of the two respective subclades, however, suggests different evolutionary scenarios.

Clade Ia comprised species assigned to *Salix* sect. *Helix* sensu Skvortsov (1999), except for *S. gracilistyla*. However, all these species share the morphological character of connate filaments. The SNAPP species tree revealed a close relationship of *S. caesia* and S. *purpurea*, supporting the RAxML analyses. Based on our genetic structure analyses, tetraploid *S. caesia* combined the genetic partitions that occur in *S. purpurea* and *S. repens*, with about equal proportions, while the two Asian species *S. gracilistyla* and *S. integra* showed each a different genetic partition. The NeighbourNet analysis confirmed these findings and placed *S. caesia* between the *S. purpurea* and *S. repens* (**Figure 3C**). These two species are sympatric with *S. caesia* (Skvortsov, 1999). HyDe detected 14% significant hybridization events for *S. caesia* based on the complete sequence data. In case of a comparatively young hybridization event we would expect γ-values between 0.4 and 0.6. The high observed average γ-value of 0.8 for the combinations of *S. purpurea* and *S. repens* supports the hypothesis of an older event or other processes like incomplete lineage sorting. In so far, the HyDe analysis corroborates results of our SNiPloid analysis that revealed a high proportion of homeo-SNPs (44.52%) derived from the two parental species, whereas proportions of SNPs from post-origin interspecific events were much lower (18.53% for cat 1+2). However, the observed amounts of interspecific SNPs (cat1, cat2) may also be the result of incomplete lineage sorting. Thus, the observed results support the hypothesis of an allotetraploid origin from *S. purpurea* and *S. repens* s.l., with rather few post-origin backcrossing events. Indeed, no extant hybrids of *S. caesia* with the putative parental species have been reported so far, and hybridization with other species is extremely rare (Hörandl et al., 2012). The considerable proportion of post-origin SNPs (Cat 3/4, 19.17%) specific for the polyploid *S. caesia* indicates an independent evolution of the lineage. Occupation of niches at higher elevations may have contributed to reproductive isolation of *S. caesia* from the two parental species that occur in lowlands (Hörandl et al., 2012).

The tetraploid species *S. cinerea* and *S. laggeri* belong to subclade IIa that contain all included species of section *Vetrix* sensu Skvortsov (1999). The members of this section are medium sized shrubs with hairy leaves occurring from lowland to montane regions. They grow often in sympatry with other species in mixed populations and hybridize frequently, also with species of other sections (Hörandl et al., 2012). A big, shared gene pool is visible in high proportions of "other", unassigned SNPs in SNiPloid. High ongoing gene flow may

explain that the clade-specific genetic structure analysis revealed no clear species-specific partitions within subclade IIa. However, both *Salix laggeri* and *S. cinerea* showed high proportions of species-specific SNPs (36.6% and 47.5%, respectively), indicating independent evolutionary lineages. The most likely parental combinations for their origin revealed in both cases around 5% of homeo-SNPs derived from hybrid origin, and also low proportions of post-origin hybridization with both parents, supporting allopolyploid origin. We cannot rule out that the true parental lineages were not included here – either due to extinction or to involvement of species that occur nowadays outside Europe (e.g., in adjacent Russia). A long evolutionary history with post-origin hybridization/ introgression events further affected the genomic composition of the hybrids. Interestingly, the two tetraploids differ strongly in the results of HyDe analyses: *S. cinerea* is responsible for 23% of significant hybridization events in the range of a more recent hybrid, while no significant hybrid event including *S. laggeri* was detected. Extant hybridization of *S. cinerea* with other co-occurring lowland species (*S. aurita, S. caprea, S. eleagnos, S. myrsinifolia, S. repens*, and a hybrid series with *S. viminalis*, Hörandl et al., 2012) may explain this result, while *S. laggeri* is a subalpine species that is reproductively better isolated from the other species and hybridizes only occasionally with the subalpine, sympatric *S. appendiculata* (Lautenschlager-Fleury and Lautenschlager-Fleury, 1986; Hörandl et al., 2012).

Our results on tetraploids confirmed an allopolyploid origin and a dynamic post-origin evolution of genomes, indicating speciation and evolution of independent lineages. Distribution ranges and ecological niches of the parental species, however, could have fluctuated from the origin of the clade onwards and may have caused various secondary contact hybridizations in different time periods of the Cenozoic (Hewitt, 2004; Hewitt, 2011). The relatively high age of the whole *Chamaetia/Vetrix* clade with a crown group age in the late Miocene, and the lack of a dated phylogeny makes it difficult to pinpoint hybridization events to certain geological time periods. According to extant hybridization patterns, isolation of the polyploid willows appears to be strongly influenced by the strength of habitat differentiation as discussed by Martini and Paiero (1988); Hörandl et al. (2012) and, Gramlich et al. (2016) for Central European species, and by Huang et al. (2015) on species from Taiwan. These findings support the notion that occupation of a separate niche is important for the establishment of a newly formed polyploid willow lineage.

CONCLUSIONS

Our data demonstrate that high-quality RAD sequencing data allow for the reconstruction of phylogenetic frameworks and give insights into origin and evolution of polyploid species. In willows, polyploidization appears to be predominantly connected to hybridization, i.e. to allopolyploid origin, as hypothesized by Skvortsov (1999). Our data suggest that polyploids harbor considerable proportions of lineage-specific SNPs and managed to establish stable, self-standing evolutionary lineages after allopolyploid origin.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA433286.

AUTHOR CONTRIBUTIONS

NW and EH planned and designed research. NW, LH and EH conducted fieldwork. NW and LH performed experiments, NW analyzed data, and established new pipelines. NW and EH wrote the manuscript with contributions from LH.

FUNDING

This study was financially supported by the German research foundation DFG (Ho 5462/7-1 to EH), the National Natural Science Foundation of China (31800466 to LH), and the Natural Science Foundation of Fujian Province of China (2018J01613 to LH). LH was sponsored by China Scholarship Council for his research stay at the University of Goettingen (201707870015).

ACKNOWLEDGMENTS

We thank Jennifer Krüger and Susanne Gramlich for extracting the DNA and support in lab work. We also thank Susanne Gramlich as well as Andrea Danler, Katrin Scheufler, Claudia Pätzold, Jun Zhao, Wan-cheng Hou, and Gong-ru Zhou for collecting plant material and field assistance. We thank Ovidiu Paun for initial help with bioinformatic pipelines. We thank anonymous reviewers for their supportive comments.
This manuscript has been released as a pre-print at *BioRxiv*, (Wagner et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.01077/full#supplementary-material

REFERENCES

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., et al. (2013). Hybridization and speciation. *J. Evol. Biol.* 26, 229–246. doi: 10.1111/j.1420-9101.2012.02599.x

Andermann, T., Fernandes, A. M., Olsson, U., Töpel, M., Pfeil, B., Oxelman, B., et al. (2019). Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Syst. Biol.* 68, 32–46. doi: 10.1093/sysbio/syy039

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17, 81–92. doi: 10.1038/nrg.2015.28

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Argus, G. W. (1997). Infrageneric classification of Salix (Salicaceae) in the New World. *Syst. Bot. Monogr.* 52, 1–121. doi: 10.2307/25096638

Argus, G. W. (2010). "Salix," in *Flora of North America vol. 7 Magnoliophyta: Salicaceae to Brassicaceae*. Ed. E. Flora of North America Editorial Committee (New York: Oxford University Press), 23–51.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* 3, 1–7. doi: 10.1371/journal.pone.0003376

Baltisberger, M., and Hörandl, E. (2016). Karyotype evolution supports the molecular phylogeny in the genus Ranunculus (Ranunculaceae). *Perspect. Plant Ecol. Evol. Syst.* 18, 1–14. doi: 10.1016/j.ppees.2015.11.001

Barcaccia, G., Meneghetti, S., Lucchin, M., and de Jong, H. (2014). Genetic segregation and genomic hybridization patterns support an allotetraploid structure and disomic inheritance for Salix species. *Diversity* 6, 633–651. doi: 10.3390/d6040633

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210, 391–398. doi: 10.1111/nph.13698

Blischak, P. D., Chifman, J., Wolfe, A. D., and Kubatko, L. S. (2018a). HyDe: A python package for genome-scale hybridization detection. *Syst. Biol.* 67, 821–829. doi: 10.1093/sysbio/syy023

Blischak, P. D., Kubatko, L. S., and Wolfe, A. D. (2018b). SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics* 34, 407–415. doi: 10.1093/bioinformatics/btx587

Bouckaert, R., Vaughan, T., Wu, C., Xie, D., Suchard, M. A., Rambaut, A., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10 (4), e1003537. doi: 10.1371/journal.pcbi.1003537

Brandrud, M. K., Paun, O., Lorenzo, M. T., Nordal, I., and Brysting, A. K. (2017). RADseq provides evidence for parallel ecotypic divergence in the autotetraploid Cochlearia officinalis in Northern Norway. *Sci. Rep.* 7, 1–13. doi: 10.1038/s41598-017-05794-z

Brandrud, M. K., Baar, J., Lorenzo, M. T., Athanasiadis, A., Bateman, R. M., Chase, M. W., et al. (2020). Phylogenomic relationships of diploids and the origins of allotetraploids in dactylorhiza (Orchidaceae). *Syst. Biol.* 69 (1), 91–109. doi: 10.1093/sysbio/syz035

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and Roychoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29, 1917–1932. doi: 10.1093/molbev/mss086

Cariou, M., Duret, L., and Charlat, S. (2013). Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecol. Evol.* 3, 846–852. doi: 10.1002/ece3.512

Catchen, J., Hohenlohe, P., Bassham, S., Amores, A., and Cresko, W. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354

Chase, M. W., Knapp, S., Cox, A. V., Clarkson, J. J., Butsko, Y., Joseph, J., et al. (2003). Molecular systematics, GISH and the origin of hybrid taxa in Nicotiana (Solanaceae). *Ann. Bot.* 92, 107–127. doi: 10.1093/aob/mcg087

Chen, J. H., Sun, H., Wen, J., and Yang, Y. P. (2010). Molecular phylogeny of Salix L. (Salicaceae) inferred from three chloroplast datasets and its systematic implications. *Taxon* 59, 29–37.

Chen, X., Li, X., Zhang, B., Xu, J., Wu, Z., Wang, B., et al. (2013). Detection and genotyping of restriction fragment associated polymorphisms in polyploid crops with a pseudo-reference sequence: a case study in allotetraploid Brassica napus. *BMC Genomics* 14, 346. doi: 10.1186/1471-2164-14-346

Clevenger, J., Chavarro, C., Pearl, S. A., Ozias-Akins, P., and Jackson, S. A. (2015). Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. *Mol. Plant* 8, 831–846. doi: 10.1016/j.molp.2015.02.002

Clevenger, J. P., Korani, W., Ozias-Akins, P., and Jackson, S. (2018). Haplotype-Based Genotyping in Polyploids. *Front. Plant Sci.* 9, 1–6. doi: 10.3389/fpls.2018.00564

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Cronk, Q., Ruzzier, E., Belyaeva, I., and Percy, D. (2015). Salix transect of Europe: latitudinal patterns in willow diversity from Greece to arctic Norway. *Biodivers. Data J.* 3, e6258. doi: 10.3897/BDJ.3.e6258

Dorn, R. D. (1976). A synopsis of American Salix. *Can. J. Bot.* 54, 2769–2789. doi: 10.1139/b76-297

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol.* 23, 40–69. doi: 10.1111/mec.12581

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Eaton, D. A. R., and Overcast, I. (2020). iPYRAD: interactive assembly and analysis of RADseq data sets. *Bioinformatics* 36 (8), 2592–2594. doi: 10.1093/bioinformatics/btz966

Eaton, D. A. R., Spriggs, E. L., Park, B., and Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Syst. Biol.* 66, 399–412. doi: 10.1093/sysbio/syw092

Eriksson, J. S., De Sousa, F., Bertrand, Y. J. K., Antonelli, A., Oxelman, B., and Pfeil, B. E. (2018). Allele phasing is critical to revealing a shared allopolyploid origin of Medicago arborea and M. strasseri (Fabaceae). *BMC Evol. Biol.* 18, 1–12. doi: 10.1186/s12862-018-1127-z

Fang, C., Zhao, S., and Skvortsov, A. (1999). "Salicaceae," in *Flora of China*, vol. 4 . Eds. Z. Y. Wu and P. H. Raven (St. Louis: Science Press, Beijing & Missouri Botanical Garden Press), 139–274.

Feng, J. Y., Li, M., Zhao, S., Zhang, C., Yang, S. T., Qiao, S., et al. (2018). Analysis of evolution and genetic diversity of sweetpotato and its related different polyploidy wild species I. trifida using RAD-seq. *BMC Plant Biol.* 18, 1–12. doi: 10.1186/s12870-018-1399-x

Gramlich, S., and Hörandl, E. (2016). Fitness of natural willow hybrids in a pioneer mosaic hybrid zone. *Ecol. Evol.* 6, 7645–7655. doi: 10.1002/ece3.2470

Gramlich, S., Sagmeister, P., Dullinger, S., Hadacek, F., and Hörandl, E. (2016). Evolution in situ: Hybrid origin and establishment of willows (Salix L.) on alpine glacier forefields. *Heredity (Edinb.)* 116, 531–541. doi: 10.1038/hdy.2016.14

Gramlich, S., Wagner, N. D., and Hörandl, E. (2018). RAD-seq reveals genetic structure of the F2-generation of natural willow hybrids (Salix L.) and a great potential for interspecific introgression. *BMC Plant Biol.* 18, 317. doi: 10.1186/s12870-018-1552-6

Guo, Y. P., Tong, X. Y., Wang, L. W., and Vogl, C. (2013). A population genetic model to infer allotetraploid speciation and long-term evolution applied to two yarrow species. *New Phytol.* 199, 609–621. doi: 10.1111/nph.12262

Guo, W., Hou, J., Yin, T., and Chen, Y. (2016). An analytical toolkit for polyploid willow discrimination. *Sci. Rep.* 6, 1–8. doi: 10.1038/srep37702

Hörandl, E., Florineth, F., and Hadacek, F. (2012). *Weiden in Österreich und angrenzenden Gebieten (willows in Austria and adjacent regions)* (Vienna: University of Agriculture, Vienna).

Hardig, T. M., Brunsfeld, S. J., Fritz, R. S., Morgan, M., and Orians, C. M. (2000). Morphological and molecular evidence for hybridization and introgression in a willow (Salix) hybrid zone. *Mol. Ecol.* 9, 9–24. doi: 10.1046/j.1365-294X.2000.00757.x

Hardig, T. M., Anttila, C. K., and Brunsfeld, S. J. (2010). A phylogenetic analysis of Salix (Salicaceae) Based on matK and ribosomal DNA sequence data. *J. Bot.* 2010, 1–12. doi: 10.1155/2010/197696

He, L., Wagner, N. D., and Hörandl, E. (2020). RAD sequencing data reveal a radiation of willow species (*Salix* L., Salicaceae) in the Hengduan Mountains and adjacent areas. *J. Systematics Evol.* 00, 1–14. doi: 10.1111/jse.12593

Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philos. Trans. R. Soc. B Biol. Sci.* 359, 183–195. doi: 10.1098/rstb.2003.1388

Hewitt, G. M. (2011). Quaternary phylogeography: the roots of hybrid zones. *Genetica* 139, 617–638. doi: 10.1007/s10709-011-9547-3

Hipp, A. L., Eaton, D. A. R., Cavender-Bares, J., Fitzek, E., Nipper, R., and Manos, P. S. (2014). A framework phylogeny of the American oak clade based on sequenced RAD data. *PloS One* 9 (7), e102272. doi: 10.1371/journal.pone.0093975

Hirsch, C. N., and Buell, C. R. (2013). Tapping the promise of genomics in species with complex, nonmodel genomes. *Annu. Rev. Plant Biol.* 64, 89–110. doi: 10.1146/annurev-arplant-050312-120237

Huang, C. L., Chang, C.-T., Huang, B. H., Chung, J.-D., Chen, J. H., Chiang, Y. C., et al. (2015). Genetic relationships and ecological divergence in Salix species and populations in Taiwan. *Tree Genet. Genomes* 11, 39. doi: 10.1007/s11295-015-0862-1

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916

Kadereit, J. W. (2015). The geography of hybrid speciation in plants. *Taxon* 64, 673–687. doi: 10.12705/644.1

Lauron-Moreau, A., Pitre, F. E., Argus, G. W., Labrecque, M., and Brouillet, L. (2015). Phylogenetic relationships of American Willows (Salix L., Salicaceae). *PloS One* 10, e0121965. doi: 10.1371/journal.pone.0121965

Lautenschlager-Fleury, D., and Lautenschlager-Fleury, E. (1986). Salix laggeri Wimmer - Monographie einer wenig bekannten Weide. *Bauhinia* 8 (3), 149–156.

Leaché, A. D., Banbury, B. L., Felsenstein, J., De Oca, A. N. M., and Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Syst. Biol.* 64, 1032–1047. doi: 10.1093/sysbio/syv053

Leskinen, E., and Alström-Rapaport, C. (1999). Molecular phylogeny of Salicaceae and closely related Flacourtiaceae: evidence from 5.8 S, ITS 1 and ITS 2 of the rDNA. *Plant Syst. Evol.* 215, 209–227. doi: 10.1007/BF00984656

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows – wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Martini, F., and Paiero, P. (1988). I Salici d'Italia. Ed. Lint: Trient.

Mastretta-Yanes, A., Zamudio, S., Jorgensen, T. H., Arrigo, N., Alvarez, N., Pinero, D., et al. (2014). Gene duplication, population genomics, and species-level differentiation within a tropical mountain shrub. *Genome Biol. Evol.* 6, 2611–2624. doi: 10.1093/gbe/evu205

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., et al. (2011). Recently formed polyploid plants diversify at lower rates. *Science* (333), 1257. doi: 10.1126/science.1207205

McBreen, K., and Lockhart, P. J. (2006). Reconstructing reticulate evolutionary histories of plants. *Trends Plant Sci.* 11, 398–404. doi: 10.1016/j.tplants.2006.06.004

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meirmans, P. G., Liu, S., and Van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *J. Hered.* 109, 283–296. doi: 10.1093/jhered/esy006

Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E., and Smith, S. A. (2018). Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. *Am. J. Bot.* 105, 385–403. doi: 10.1002/ajb2.1016

Peralta, M., Combes, M. C., Cenci, A., Lashermes, P., and Dereeper, A. (2013). SNiPloid: A utility to exploit high-throughput SNP data derived from RNA-Seq in allopolyploid species. *Int. J. Plant Genomics* 2013, 890123. doi: 10.1155/2013/890123

Percy, D. M., Argus, G. W., Cronk, Q. C., Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., et al. (2014). Understanding the spectacular failure of DNA barcoding in willows (Salix): does this result from a trans-specific selective sweep? *Mol. Ecol.* 23, 4737–4756. doi: 10.1111/mec.12837

Qi, Z. C., Yu, Y., Liu, X., Pais, A., Ranney, T., Whetten, R., et al. (2015). Phylogenomics of polyploid Fothergilla (Hamamelidaceae) by RAD-tag based GBS - insights into species origin and effects of software pipelines. *J. Syst. Evol.* 53, 432–447. doi: 10.1111/jse.12176

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rambaut, A. (2014). FigTree v1.4.2, A graphical viewer of phylogenetic trees. Available from: https://github.com/rambaut/figtree/releases.

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Rechinger, K. H., and Akeroyd, J. R. (1993). "Salicaceae," in *Flora Europaea*, vol. 1 . Eds. T. G. Tutin, N. A. Burges, A. O. Chater, J. R. Edmondson, V. H. Heywood, D. M. Moore, D. H. Valentine, S. M. Walters and D. A. Webb (Cambridge: Cambridge University Press), 53–66.

Rivers, D. M., Darwell, C. T., and Althoff, D. M. (2016). Phylogenetic analysis of RAD-seq data: examining the influence of gene genealogy conflict on analysis of concatenated data. *Cladistics* 32, 672–681. doi: 10.1111/cla.12149

Savage, J. A., and Cavender-Bares, J. (2012). Habitat specialization and the role of trait lability in structuring diverse willow (genus Salix) communities. *Ecology* 9, 138–150. doi: 10.1890/11-0406.1

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends Ecol. Evol.* 19, 198–207. doi: 10.1016/j.tree.2004.01.003

Skvortsov, A. K. (1999). *Willows of Russia and adjacent countries. Taxonomical and geographical revision* (Joensuu, Finland: University of Joensuu).

Soltis, P. S., and Soltis, D. E. (2009). The Role of Hybridization in Plant Speciation. *Annu. Rev. Plant Biol.* 60, 561–588. doi: 10.1146/annurev.arplant.043008.092039

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

Suda, Y., and Argus, G. W. (1968). Chromosome numbers of some North American Salix. *Brittonia* 20, 191–197. doi: 10.2307/2805440

Suda, J., and Trávníček, P. (2006). Estimation of relative nuclear DNA content in dehydrated plant tissues by flow cytometry. *Curr. Protoc. Cytom.* 38, 7.30.1–7.30.14. doi: 10.1002/0471142956.cy0730s38

Triest, L., De Greef, B., De Bondt, R., and Van Slycken, J. (2000). RAPD of controlled crosses and clones from the field suggests that hybrids are rare in the Salix alba-Salix fragilis complex. *Heredity (Edinb.)* 84, 555–563. doi: 10.1046/j.1365-2540.2000.00712.x

Triest, L. (2001). Hybridization in staminate and pistillate Salix alba and S. fragilis (Salicaceae): Morphology versus RAPDs. *Plant Syst. Evol.* 226, 143–154. doi: 10.1007/s006060170062

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* (313), 1596–1604. doi: 10.1126/science.1128691

Wagner, N. D., Gramlich, S., and Hörandl, E. (2018). RAD sequencing resolved phylogenetic relationships in European shrub willows (Salix L. subg. Chamaetia and subg. Vetrix) and revealed multiple evolution of dwarf shrubs. *Ecol. Evol.* 8, 8243–8255. doi: 10.1002/ece3.4360

Wagner, N. D., He, L., and Hörandl, E. (2019). Relationships and genome evolution of polyploid Salix species revealed by RAD sequencing data. Preprint at BioRxiv. doi: 10.1101/864504

Wichura, M. (1865). *Die Bastardbefruchtung im Pflanzenreich erläutert an den Bastarden der Weiden* (Breslau: Morgenstern).

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci.* 106, 13875–13879. doi: 10.1073/pnas.0811575106

Wu, J., Nyman, T., Wang, D.-C., Argus, G. W., Yang, Y.-P., and Chen, J.-H. (2015). Phylogeny of Salix subgenus Salix s.l. (Salicaceae): delimitation, biogeography, and reticulate evolution. *BMC Evol. Biol.* 15, 31. doi: 10.1186/s12862-015-0311-7

Zhang, B.-W., Xu, L.-L., Li, N., Yan, P.-C., Jiang, X.-H., Woeste, K. E., et al. (2019). Phylogenomics reveals an ancient hybrid origin of the Persian walnut. *Mol. Biol. Evol.*, 36 (11), 2451–2461. doi: 10.1093/molbev/msz112

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 09 September 2020
doi: 10.3389/fpls.2020.01295

Early Diverging and Core Bromelioideae (Bromeliaceae) Reveal Contrasting Patterns of Genome Size Evolution and Polyploidy

Juraj Paule[1,2]*, *Sascha Heller*[2], *Jefferson Rodrigues Maciel*[3], *Raquel F. Monteiro*[4], *Elton M. C. Leme*[5,6] *and Georg Zizka*[1,2]

[1] Department of Botany and Molecular Evolution, Senckenberg Research Institute and Natural History Museum, Frankfurt am Main, Germany, [2] Institute of Ecology, Evolution and Diversity, Goethe University, Frankfurt am Main, Germany, [3] Jardim Botânico do Recife, Prefeitura da Cidade do Recife, Recife, Brazil, [4] Department of Botany, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [5] Marie Selby Botanical Gardens, Sarasota, FL, United States, [6] Rio de Janeiro Botanical Garden, Rio de Janeiro, Brazil

OPEN ACCESS

Edited by:
Karol Marhold,
Slovak Academy of Sciences, Slovakia

Reviewed by:
Carolina Granados Mendoza,
National Autonomous University of Mexico, Mexico
Daniel Blaine Marchant,
Stanford University, United States

***Correspondence:**
Juraj Paule
jpaule@senckenberg.de

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *22 April 2020*
Accepted: *07 August 2020*
Published: *09 September 2020*

Citation:
Paule J, Heller S, Maciel JR, Monteiro RF, Leme EMC and Zizka G (2020) Early Diverging and Core Bromelioideae (Bromeliaceae) Reveal Contrasting Patterns of Genome Size Evolution and Polyploidy. Front. Plant Sci. 11:1295. doi: 10.3389/fpls.2020.01295

The subfamily Bromelioideae is one of the most diverse groups among the neotropical Bromeliaceae. Previously, key innovations have been identified which account for the extraordinary radiation and species richness of this subfamily, especially in the so-called core Bromelioideae. However, in order to extend our understanding of the evolutionary mechanisms, the genomic mechanisms (*e.g.* polyploidy, dysploidy) that potentially underlie this accelerated speciation also need to be tested. Here, using PI and DAPI staining and flow cytometry we estimated genome size and GC content of 231 plants covering 30 genera and 165 species and combined it with published data. The evolutionary and ecological significance of all three genomic characters was tested within a previously generated dated phylogenetic framework using ancestral state reconstructions, comparative phylogenetic methods, and multiple regressions with climatic variables. The absolute genome size (2C) of Bromelioideae varied between 0.59 and 4.11 pg, and the GC content ranged between 36.73 and 41.43%. The monoploid genome sizes (Cx) differed significantly between core and early diverging lineages. The occurrence of dysploidy and polyploidy was, with few exceptions, limited to the phylogenetically isolated early diverging tank-less lineages. For Cx and GC content Ornstein–Uhlenbeck models outperformed the Brownian motion models suggesting adaptive potential linked to the temperature conditions. 2C-values revealed different rates of evolution in core and early diverging lineages also related to climatic conditions. Our results suggest that polyploidy is not associated with higher net diversification and fast radiation in core bromelioids. On the other hand, although coupled with higher extinction rates, dysploidy, polyploidy, and resulting genomic reorganizations might have played a role in the survival of the early diverging bromelioids in hot and arid environments.

Keywords: bromeliads, chromosome number, climate, C-value, GC content, flow cytometry, phylogenetic signal, ploidy

INTRODUCTION

Bromelioideae is the second most species-rich Bromeliaceae subfamily comprising 986 species as well as the most diverse subfamily concerning the number of genera (Gouda et al., 2020). The members of this subfamily occupy terrestrial, lithophytic, and epiphytic habitats in subtropical and tropical biomes in the Neotropics (Smith and Downs, 1979). The taxonomy of the subfamily is challenging as the present concepts still recognize polyphyletic groups as genera, especially in the case of the largest genus *Aechmea* (Schulte et al., 2009; Silvestro et al., 2014; Aguirre-Santoro et al., 2016). Moreover, the lack of resolution in phylogenetic studies and recent diversification (crown age of the extant lineages: 10.9 Ma; Silvestro et al., 2014) still limits our understanding of inter- and infrageneric evolutionary relationships (*e.g.* Schulte et al., 2005; Horres et al., 2007; Evans et al., 2015).

Nonetheless, the subfamily can be divided into two major groups: several informal early diverging tank-less lineages and mostly tank-forming core bromelioids (Schulte et al., 2009; Silvestro et al., 2014). The tank-less lineages are terrestrial or lithophytic usually characterized by succulent leaves with considerable water storage tissue and CAM photosynthesis, except the earliest diverging *Greigia*, *Ochagavia*, *Fascicularia*, as well as several other species from the Cryptanthoid complex and the genus *Fernseea*, which are C3 (Horres and Zizka, 1995; Silvestro et al., 2014; Crayn et al., 2015; Leme et al., 2017). In the tank-forming lineages of the core Bromelioideae, one large and/or several small tanks are formed by the tightly clasping leaf sheaths, which can thus impound up to several liters of water in larger plants (*e.g.* Cogliatti-Carvalho et al., 2010). The externally stored water is absorbed by leaf scales (tank-absorbing trichomes; Benzing, 2000; Males, 2016). The water bodies (phytotelmata) are also microhabitats and water source for a considerable range of animals and microorganisms (*e.g.* Wheeler, 1942; Frank and Lounibos, 2009). Interestingly, very few species among the core Bromelioideae lack the tank or have a rudimentary one (*e.g. Acanthostachys strobilacea, Araeococcus flagellifolius*), and only a few have been reported to have C3 instead of CAM photosynthesis (especially from genera *Nidularium* and *Ronnbergia*; Crayn et al., 2015).

Tank-less Bromelioideae of the early diverging lineages includes over 200 species, whereas the core group comprises over 600 species (Gouda et al., 2020) and these two groups revealed contrasting evolutionary histories. It has been shown that the presence of a tank in core bromeliads is coupled with increased net diversification and five times lower extinction rates when compared to early diverging tank-less clades (Silvestro et al., 2014). Hence, tank habit is considered an important key innovation and of great importance for the diversification of the species-rich epiphytic core group in the Atlantic and Amazonian rainforests. On the other hand, the majority of tank-less species are found in dry habitats of Brazilian Cerrado and Caatinga or in the Andes as terrestrial or lithophytic xerophytes, and the higher extinction rate could be attributed to the absence of external water storage (tank) as well as to the periods of higher aridification during the Pleistocene climatic oscillations (Simon et al., 2009; Silvestro et al., 2014).

Genome size varies enormously among land plants, and the variation range is approximately 2,400-fold, which demonstrates the importance of ploidy level changes, the abundance of transposable elements, repetitive DNA, and chromosomal rearrangements in the plant evolution (Pellicer et al., 2018). This variation is considered to be the result of a combination of ecological, physiological, and morphological selection processes at the molecular level as genome size can impose phenotypic constraints on plant development, phenology, and ecological performance (Knight et al., 2005). In particular, monocot genomic diversity includes striking variation at many levels. Although the ancestral genome size of all monocots was small (1C = 1.9 pg), there have been notable increases in the rates of genome size-evolution particularly in Poales, to which the family Bromeliaceae belongs (Leitch et al., 2010). An increase in genome size by means of polyploidy has been repeatedly suggested for higher adaptive potential, which could be attributed to *e.g.* the increased genetic variability of polyploids, masking of mutations, gene redundancy, heterosis (*e.g.* Comai, 2005; Finigan et al., 2012). Indirect evidence for adaptive ecological significance of polyploidy was recently shown in *Fosterella* (Bromeliaceae: Pitcairnioideae) (Paule et al., 2017) and can be possibly assumed also for *Tillandsia* subg. *Diaphoranthema* (Till, 1992). Moreover, the GC content was considered to have an adaptive role due to differences in the physical properties of GC and AT base pairs (Šmarda and Bureš, 2012) and/or due to mode of endoreplication (Trávníček et al., 2019). In monocots, it has been demonstrated that higher GC contents are favored in cold and dry climates (Šmarda et al., 2014).

Within the subfamily Bromelioideae the dominant chromosome number is 2n = 2x = 50, similar to the whole family (Gitaí et al., 2014). Deviating chromosome numbers (2n = 32/34/36) were observed only in the genus *Cryptanthus* and *Hoplocryptanthus* and were attributed to dysploidy (Ramírez-Morillo and Brown, 2001; Gitaí et al., 2014; Cruz et al., 2020). Genome size (2C-values, Greilhuber et al., 2005) is small within the subfamily, ranging from 0.61 to 2.19 pg (Gitaí et al., 2014; Müller et al., 2019), and for GC content rather a broad interval ranging between 33.95 and 44.00% was reported (Favoreto et al., 2012; Šmarda et al., 2014). Polyploidy was so far documented only in two species of core Bromelioideae (*Aechmea eurycorymbus*, *Neoglaziovia variegata*; Cotias-de-Oliveira et al., 2000; Gitaí et al., 2014), but in several genera of early diverging Bromelioideae (*Ananas*: Collins and Kerns, 1936; Brown et al., 1997; *Bromelia*: Cotias-de-Oliveira et al., 2000; Gitaí et al., 2014; *Deinacanthon*: Gitaí et al., 2005). Nevertheless, the data on the genomic characters are scarce as chromosome numbers are known for *ca.* 16%, genome size for 13%, and GC content for 1% of the species only.

By adding substantial new data, we aim here 1) to analyze the distribution and potential evolutionary consequences of genome size and GC content variation as well as of the occurrence of polyploidy in the subfamily Bromelioideae. More specifically, using previously published phylogenetic framework combined

with multiple regression analyses of genomic and climatic variables and model tests of genomic character evolution, we would like to examine whether 2a) previously confirmed fast radiation in the core tank-forming Bromelioideae (Silvestro et al., 2014) could be coupled with polyploidy and genomic rearrangements (assessed by divergent GC content and monoploid genome size (Cx); Trávníček et al., 2019) or, alternatively 2b) if the higher extinction rate in early diverging tank-less Bromelioideae is associated with these processes.

MATERIALS AND METHODS

Plant Material

Altogether, 230 individual plants were sampled, covering 158 out of 986 (16.1%) currently recognized Bromelioideae species, 30 out of 37 Bromelioideae genera, and an additional 5 *Puya* outgroup species. We provide for the first time genome size for 127 and GC content estimates for 152 Bromelioideae species. The sampling covered 101 (including one published record) out of 114 species of the Bromelioideae phylogeny by Silvestro et al. (2014) and aimed representatively to cover the whole subfamily including so far underrepresented genera such as *Bromelia*. The studied material included 1 to 6 accessions per species. Taxonomic assignments and nomenclature are based on the Encyclopaedia of Bromeliads v4 (Gouda et al., 2020), except for the *Ananas* group, for which we applied the taxonomy as used by Matuszak-Renger et al. (2018). A list of all studied samples, their geographic origin, collection history, and herbarium voucher information is provided in **Supplementary Table 1**.

Relative Genome Size, Absolute Genome Size, and GC Content Estimation

Relative genome sizes (RGS) were estimated by flow cytometric analyses of fresh leaves using a CyFlow space (Partec, Münster, Germany) fitted with a high power UV LED (365 nm). Leaf tissues of the analyzed sample and internal standard *Glycine max* cv. Polanka (2C = 2.50 pg) or *Pisum sativum* cv. Ctirad (2C = 9.09 pg) (Doležel et al., 1994; Doležel et al., 1998) were homogenized using a razor blade in a plastic Petri-dish containing 1 ml of ice-cold Otto I buffer (0.1 M citric acid, 0.5% Tween 20; Otto, 1990) to extract the nuclei. The suspension was filtered through Partec CellTrics® 30 μm to remove tissue debris and incubated for at least 5 min at room temperature. Isolated nuclei in filtered suspension were stained with 1 ml of Otto II buffer (0.4 M $Na_2HPO_4 \times 12H_2O$) containing the AT-specific fluorochrome 4′,6-diamidino-2-phenylindole (DAPI, 4 $\mu g \cdot ml^{-1}$) and β-mercaptoethanol (2 $\mu g \cdot ml^{-1}$). The relative fluorescence intensity was recorded for 3,000 particles (nuclei) with one to three replicates per accession. RGSs expressed as sample/standard ratios were calculated from the means of fluorescence histograms visualized using the FloMax v2.4d software (Partec, Münster, Germany). Only histograms with coefficients of variation (CVs) < 5% for the G0/G1 peak of the sample were considered. In the case of CVs exceeding this threshold, the measurement was discarded, and the sample was reanalyzed.

Absolute genome sizes (2C-values) were estimated using the identical protocol as for RGS except for the staining solution, which consisted of 1 ml of Otto II buffer and intercalating propidium iodide (PI) and RNase IIA (both at final concentrations of 50 $\mu g \cdot ml^{-1}$). Fluorescence was induced by 30 mW green solid-state laser (532 nm) and fluorescence intensities of 5,000–10,000 nuclei per measurement were recorded. Two to eight replicate measurements of each sample were carried out on different days. The between-day variation caused by random instrument drift and/or non-identical sample preparation was low, and the difference between the maximum and minimum values of replicates did not exceed a 4% threshold.

The 2C-values were calculated by multiplying the sample/standard ratios with a known genome size of the standard *G. max*. *Pisum sativum* was used as a standard in five cases (**Supplementary Table 1**), in which the sample signal overlapped with the signal of *G. max*. The 2C-values estimated using the internal standard *P. sativum* were adjusted to those using *G. max* by multiplying the values by the coefficient of 3.772 which was based on 12 repeats of ratios among the two standards. Monoploid genome sizes (Cx-values; Greilhuber et al., 2005) were calculated for the species for which the chromosome number or ploidy level was known by dividing the 2C-values with respective ploidy. The estimation of GC content was based on a comparison of nuclei fluorescence stained with the DNA intercalating propidium iodide (PI ratio) and AT-specific DAPI (RGS) using the protocols and the GC content calculation tool by Šmarda et al. (2008). The base content of standard *G. max* (0.636) was extracted from Barow and Meister (2002).

Literature Review

For the review of previously published chromosome numbers, genome size, and GC content estimates, taxa were critically assessed, and the nomenclature of Gouda et al. (2020) was followed. We retrieved 71 chromosome counts, which could be related to new or published genome size estimates, 54 chromosome counts for species present in the underlying phylogeny (**Supplementary Table 2**), and 43 genome size and 5 GC content estimates, which were included in further analyses (**Supplementary Table 3**).

Data Analyses

Statistical analyses were performed in R v4.0.2 (R Core Team, 2020). The relationship between chromosome counts and 2C-values was assessed by Pearson's correlation coefficient as well as linear regression. The differences between all 2C-values, monoploid genome sizes, and GC content estimates of early diverging and core bromeliads were tested using the non-parametric Wilcoxon rank-sum test (due to departure from normality assessed by Shapiro–Wilk normality test). For the comparison of the monoploid genome sizes, *Cryptanthus* and *Hoplocryptanthus* were excluded due to diverging presumably dysploid genomes. If several accessions per species were measured, those with higher 2C-value for each ploidy level were analyzed further using comparative methods. If a new and published 2C estimate was available, a newly generated data point was taken.

Comparative analyses were performed in R using phytools v0.6-60 (Revell, 2012) and OUwie v2.3 (Beaulieu et al., 2012) and maximum clade credibility chronogram of Silvestro et al. (2014) for all three genomic characters (2C, Cx, GC content). The phylogenetic signal was calculated as Pagel's λ (Pagel, 1999) with a test of departure from $\lambda = 0$ by a likelihood-ratio test using function phylosig in phytools. To determine whether rates of genomic characters' evolution differ among the early diverging and core clades, the fit of two Brownian motion (BM) and five Ornstein–Uhlenbeck (OU) models was compared using OUwie. Both BM and OU models estimate the rate of stochastic motion (σ^2). OU process allows the trait to fluctuate around an optimum value (θ) in parameter space with a strength of attraction (α) towards that optimum, while BM allows the trait to move equally to any parameter space. Models BM1 and BMS assign single and multiple rates (σ^2) of random drift. OU1 and OUM model single and multiple optima (θ) for different clades with a single α and σ^2. The remaining models assume either multiple σ^2 (OUMV), multiple α (OUMA), or both (OUMVA) among clades. When fitting models using OUwie, the starting value θ_0 was dropped from the model and assumed to be distributed according to the stationary distribution of the OU process (default setting). The performance of each model was assessed by 1) confirming that the eigenvalues of the Hessian matrix were positive (Beaulieu et al., 2012) and 2) checking that the estimated optima (θ) of traits were not outside a plausible range. Only models passing these criteria were retained. The best-fitting model was selected using AIC weights based on the sample size-corrected Akaike information criterion (AICc) using the function aic.w in phytools.

To assess the environmental correlates of the genomic characters, 19 climatic variables were extracted from the WorldClim v1.4 database (Hijmans et al., 2005) downloaded at the 30 seconds resolution for all Bromelioideae species based on the recently published species distribution ranges (Zizka et al., 2020). The shape of the refined distributional polygons was used as a 'zonal feature' to extract values of each Bioclim variable for all species using ArcGIS v10.0. Then, the mean of each climatic variable was calculated for every species. These mean values were used to explore climatic preferences of core and early diverging lineages by principal component analysis (PCA) implemented in function dudi.pca from the R package ade4 v1.4-14 (Chessel et al., 2004). For variable pairs with absolute correlation coefficients higher than 0.75 (**Supplementary Table 4**) only one, biologically more significant variable, was kept in order to not overemphasize the contribution of a particular climatic factor. Statistical differences between climatic niches of both groups were tested using the Wilcoxon rank-sum test of principal components. Relationships between genomic characters and climatic variables were further evaluated by multiple phylogenetic generalized least-squares (PGLS) (Grafen, 1989) based on a reduced genomic dataset (89 taxa) selected to match the phylogenetic tree and available distribution data. The PGLS analyses were carried out using the R package caper v1.0.1 (Orme et al., 2018) with the λ value estimated by maximum likelihood. Due to possible violation of PGLS assumptions, relationships between 2C and climatic variables were further evaluated by multiple linear regressions based on the whole genomic dataset matching the extracted distributions ranges (172 taxa).

Ancestral chromosome number reconstruction was carried out to assess the temporal dimension of polyploidization events. Estimates of ancestral chromosome numbers were inferred using maximum likelihood (ML) under the Markov k-state 1 (Mk1) parameter model, using the software Mesquite v2.74 (Maddison and Maddison, 2010). Mk1 implements a single parameter for the rate of change among any character state (Lewis, 2001). Different chromosome numbers were coded as categorical characters and proportional likelihood (PL) values were used to determine which ancestral state was the most likely. The ancestral states of monoploid genome size (Cx) and GC content were reconstructed using the maximum likelihood estimation (function fastAnc) and visualized on the phylogenetic tree using the function contMap in phytools.

RESULTS

Genomic Characters

The genomes of the analyzed Bromelioideae species (**Supplementary Table 1**) were relatively small, ranging from 0.59 pg (*Orthophytum disjunctum* var. *viridiflorum*) to 4.11 pg (*Deinacanthon urbanianum*) for 2C and 0.27 pg (*Orthophytum compactum*) to 0.78 pg (*Greigia sphacelata*) for Cx-values. The genomes of five *Puya* species varied from 2C = 1 pg in *Puya densiflora* to 2C = 1.30 pg in *Puya ferruginea*. The differences between new and previously published 2C-values or among several accessions per species were low and did not exceed 5%, except cases in which the determination was provisional ("cf.") or the differences could be attributed to polyploidy. The GC content ranged from 36.73 (*Deinacanthon urbanianum*) to 41.43% (*Aechmea filicaulis*), which represents a smaller range than previously reported by Favoreto et al. (2012). Hence, the published outlier GC content values were excluded from further analyses.

Polyploidy and Dysploidy

To explore if genome size can be used as a proxy for ploidy level in Bromelioideae, published chromosome numbers were combined with available 2C-value data (**Figure 1**). After the exclusion of *Cryptanthus* and *Hoplocryptanthus* due to deviating chromosome numbers and presumed dysploid genomes, the correlation analysis showed a strong positive linear relationship and presented a robust linear model (Pearson's r = 0.85, df = 61, $CI_{95\%}$ = 0.77, 0.91). Accordingly, when omitting the outliers for 2n = 50 (*Greigia sphacelata*, 2C = 1.56 pg; *Aechmea filicaulis*, 2C = 1.95 pg), we can assume diploidy for the majority of bromelioids with 2C < 1.37 pg (**Supplementary Figure 1**), although *Orthophytum* certainly represents a critical group with possible tetraploid 2Cs close to this threshold. In the case of polyploidy, a specific ploidy level was attributed based on the multiplication factor within a particular evolutionary lineage as shown *e.g.* for genus *Bromelia*, *Neoglaziovia variegata* or *Pseudananas sagenarius*, for which chromosome

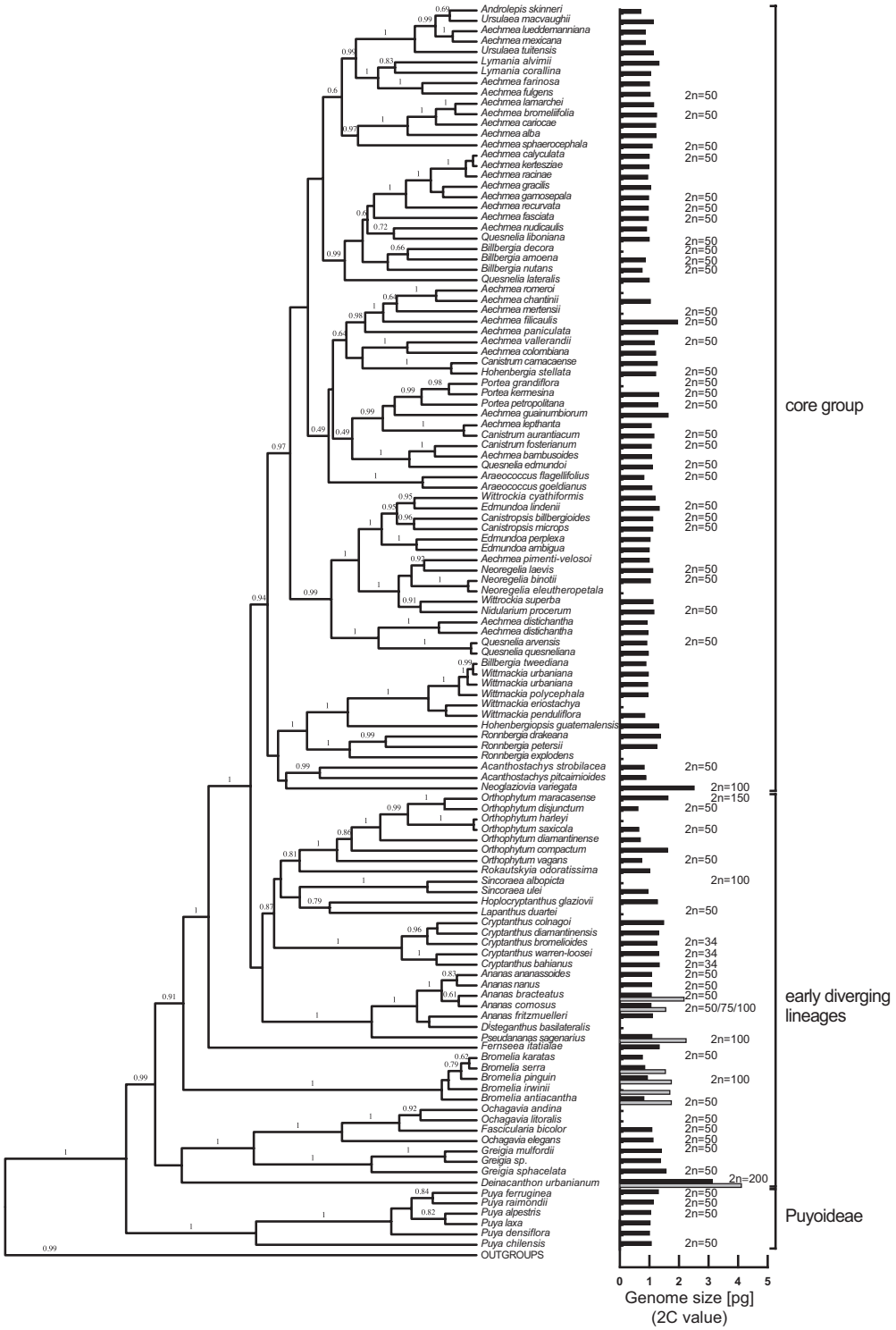

FIGURE 1 | Phylogenetic tree of the subfamily Bromelioideae with the posterior probabilities estimated by BEAST above branches (adapted from Silvestro et al., 2014). Genome size (2C-value) is shown to the right of the tree as a black bar. The presence of a gray bar indicates intraspecific variation in genome size (*i.e.* ploidy variation).

numbers are known (**Figure 1**, **Supplementary Table 1**). We identified several cases of infrageneric and intraspecific ploidy variation in the early diverging lineages such as *Ananas*, *Bromelia*, *Deinacanthon*, and *Pseudananas*, while diploids with only two exceptions (*Aechmea eurycorymbus*, *Neoglaziovia variegata*) were found in the core group (**Figure 1**, **Supplementary Table 1**). In three cases ploidy assignment was questionable (*Aechmea filicaulis*, *Aechmea guainumbiorum*, *Ronnbergia deleonii* (**Supplementary Table 1**) due to higher but not yet a multitude value of 2C within a presumably diploid lineage.

As mentioned above, dysploid chromosome numbers (2n = 32/34/36) were previously found in the genera *Cryptanthus* and *Hoplocryptanthus* of the early diverging group (Gitaí et al., 2014; Cruz et al., 2020). The genome size within the genus *Cryptanthus* is in our dataset relatively conserved (2C = 1.36 ± 0.09 pg) and we assume that also for the *Cryptanthus* and *Hoplocryptanthus* species without chromosome counts dysploid chromosome numbers can be expected based on previous chromosome count calibrations (Gitaí et al., 2014; Cruz et al., 2020).

Ancestral Chromosome Number

Using 2C-values as a proxy for ploidy level (see above) together with previously published chromosome numbers a reconstruction of ancestral chromosome numbers was carried for internal nodes of the Bromelioideae phylogenetic tree (**Figure 2**). The ancestral chromosome number inferred for the subfamily was 2n = 50 with a PL value of 0.99. For all but nine nodes a diploid ancestral chromosome number of 2n = 50 was inferred (PL = 0.99–1.00) as well. Nodes with a PL of a dysploid/polyploid ancestor were present in several lineages from the early diverging clades, namely *Bromelia*, *Cryptanthus*, *Sincoraea* (**Figure 2**).

Genome Size Evolution

For the 2C-values of early diverging and core bromelioids insignificant differences were recovered ($W = 7551.5$, $p = 0.4075$), although the variance in the early diverging lineages was 6.5 times higher than in the core group (0.399 *vs* 0.061, **Figure 3A**). For the Cx-values ($W = 3209$, $p < 0.001$) (**Figure 3B**) as well as GC content ($W = 4462$, $p < 0.001$) (**Figure 3C**) significant differences were recovered between early diverging and core bromelioids. The phylogeny explains the distribution of 2C-values ($\lambda = 0.842$, $p < 0.001$), Cx-values ($\lambda = 0.974$, $p < 0.001$) as well as GC content ($\lambda = 0.762$, $p < 0.01$) (**Table 1**). When fitting multi-regime models of trait evolution, different models were suggested for studied three genomic characters. For 2C-values BMS model was best-fitting revealing 12 times higher stochastic motion around optimum in the early diverging lineages. For Cx-values OUMV model revealed diverging optima in both groups as well as three times higher stochastic motion in the early diverging lineages. For the GC content, the phylogenetic signal was confirmed by fitting the OU1 model (AICc weight = 0.61) suggesting adaptive potential towards a common optimum of both groups. However, AICc weights favored also OUM with a conditional probability of 0.29, but the difference in estimated optima was rather small (core 39.27 vs early diverging 39.11), which is in line with the OU1 as the best-fitting model (**Table 1**). The performance of fitted models that passed the criteria mentioned above is summarized in **Supplementary Table 5**. The Bromelioideae ancestral monoploid genome size was 0.56 pg, and the ancestral GC content was 39.48% (**Supplementary Figures 2** and **3**).

Climatic Niches and Regression Analyses

Climatic niches of early diverging and core lineages were compared by PCA (**Supplementary Figure 4**). Together, the first two axes (PC1, PC2) explained 61.90% of the total variance. PC1 (37.50% of the variance) corresponds to a gradient in temperature seasonality (bio4), the minimum temperature of the coldest month (bio6) and the precipitation of the driest month (bio14). Loadings of variables for PC2 (24.40% of the variance) correspond to variation in precipitation of the warmest quarter (bio18), the maximal temperature of the warmest month (bio5), and precipitation of the wettest month (bio13). A shift of 95% inertia ellipses of early diverging and core lineages recovered by the PCA suggested climatic differentiation (**Supplementary Figure 4A**), which was confirmed by significant differences for both PC1 ($W = 81203$, $p < 0.001$) and PC2 ($W = 58594$, $p < 0.001$). Hence, the species from early diverging lineages tend to occupy drier and colder habitats with higher temperature seasonality than species from core lineages.

Only temperature related variables were significantly associated with the genomic characters when considering multiple PGLS regressions. All temperature related variables (bio2, bio4, bio5, bio6) were significantly associated with Cx, and the model explained 20.26% of the variation ($p < 0.05$), similarly as for GC content (except bio4; multiple $R^2 = 0.2737$, $p < 0.001$). No climatic variables were significantly associated with 2C-values using PGLS (**Supplementary Table 6**); however, non-phylogenetic multiple regression revealed a significant model ($p = 0.004$) explaining 11.65% of the variation.

DISCUSSION

Our study revealed strong contrasts between the genomes of core and early diverging lineages. This was supported by pairwise comparisons, by the strong phylogenetic signal as well as best-fitting models of evolution for all three studied genomic characters. A strong phylogenetic signal for 2C in Bromelioideae was also previously observed by Müller et al. (2019).

In the core group the vast majority of the studied species was considered diploid (2n = 2x = 50). The incidence of polyploidy was low, with only two polyploid species and three cases of ambiguous ploidy level among all studied species and published records. 2C-values showed smaller variance when compared to the early diverging lineages suggesting that the genomes of the core group are much more conserved. Thus, we conclude that polyploidy did not play a role in the fast radiation and increased diversification of the core group identified previously (Silvestro et al., 2014). Additionally, diploidy and genomic uniformity may play a role in gene flow among core lineages. It has been shown that interspecific, as well as intergeneric experimental hybridization

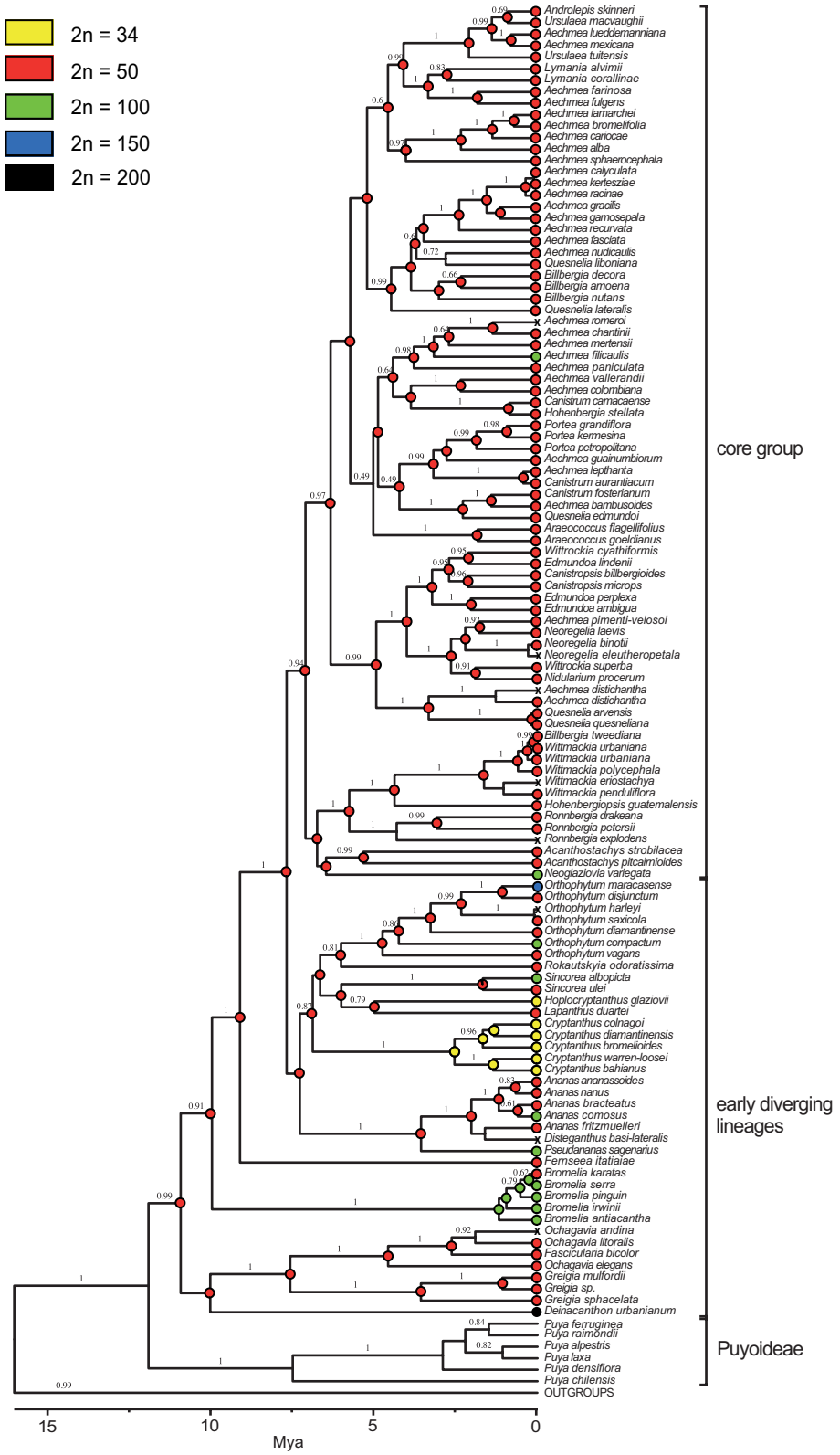

FIGURE 2 | Ancestral chromosome number reconstructions of the subfamily Bromelioideae (adapted from Silvestro et al., 2014). Chromosome numbers of each accession (tips) and reconstructed proportional likelihood at each node are indicated by colored circles. "X", missing data. The scale bar (Mya) below the tree represents time as reconstructed by Silvestro et al., 2014 and the values above branches are the posterior probabilities estimated by BEAST.

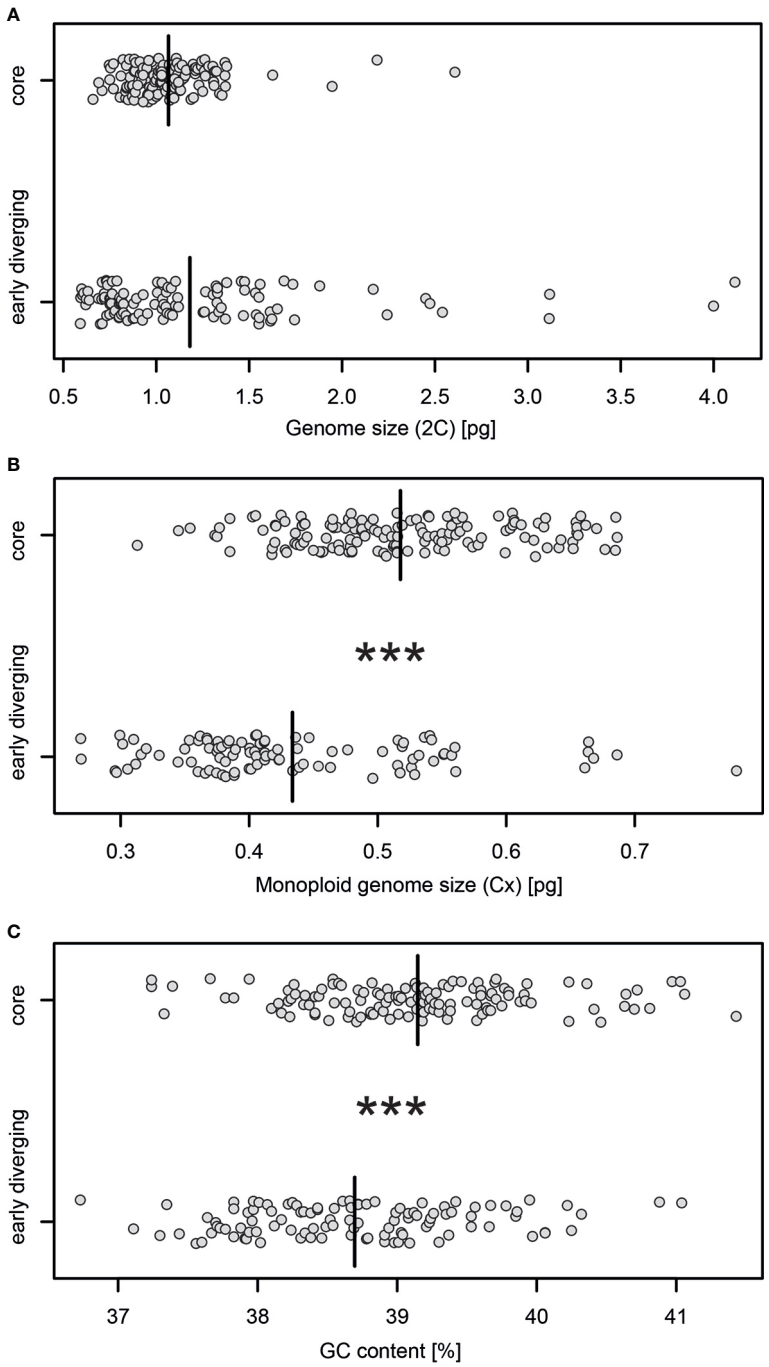

FIGURE 3 | Stripchart comparisons of the **(A)** genome size estimates (2C), **(B)** monoploid genome size (Cx), **(C)** GC content between early diverging and core Bromelioideae. Asterisks indicate statistical significance based on the non-parametric Wilcoxon rank-sum test (***$p \leq 0.001$), black bars indicate mean values of a particular variable and ploidy.

within core Bromelioideae, is possible (*e.g.* Zhang et al., 2012; Souza et al., 2017), and rare natural hybridization events (Furtado et al., 2003; Wendt et al., 2008; Matallana et al., 2016) are important drivers of evolution in the group (Goetze et al., 2016). Hence, uniform genomes in terms of genome size and chromosome number could be considered favorable for gene flow as chromosomal and genomic uniformity usually facilitates chromosome pairing and recombination (*e.g.* Rieseberg, 2001).

On the contrary, in the early diverging group 12 dysploid and 21 polyploid individuals (out of 116 samples) and several cases of intraspecific ploidy variation (*e.g. Bromelia laciniosa*, *Deinacanthon urbanianum*, *Pseudananas sagenarius*) were observed. The

TABLE 1 | Parameters estimated using best-fitted models of trait evolution for 2C, Cx, and GC content.

Genomic character/model	groups	α	σ^2	θ (S.E.)
2C/BMS	early diverging	NA	0.1188	1.6834 (0.4511)
	core	NA	0.0133	1.6834 (0.4511)
Cx/OUMV	early diverging	2.0411	0.1297	0.5067 (0.0374)
	core	2.0411	0.0397	0.5287 (0.0124)
GC/OU1	both	1.644	2.122	39.2234 (0.0838)

α, strength of attraction towards optimum; σ^2, rate of stochastic motion; θ, optimum value; S.E., standard error.

genomes from the early diverging clades showed higher Cx as well as 2C variation when compared to the core group. Polyploidy evolved at least eight times independently, between the recent past and 1.67 Ma (95% HPD 0.44–3.52), whereas for dysploidy an origin at 2.53 Ma (95% HPD 1.27–4.22) was inferred (**Figure 2**; Silvestro et al., 2014). A second dysploidy event can be assumed for the genus *Hoplocryptanthus*. Interestingly, for some *Cryptanthus* and *Hoplocryptanthus* species (Ramírez-Morillo and Brown, 2001; Cruz et al., 2020) also "half-sized" genomes (*e.g.* 2C = 0.75-0.90 pg vs 2C = 1.28–1.66 pg) related to 2n = 34 were reported (**Supplementary Table 2**, **Supplementary Table 3**). This implies that within dysploid lineages further diverging chromosomal speciation processes and large genome rearrangements may be assumed.

It is striking that the taxa from the clade with more variable genomes are distributed mainly in hot and dry habitats of Brazilian Cerrado and Caatinga as terrestrial or lithophytic xerophytes (**Supplementary Figure 4**). Concerning the Cx-values, the OUMV model assuming different stochastic motion around different optima was suggested as the best-fitting (**Supplementary Table 5**), although both groups experienced a genomic contraction compared to the ancestrally reconstructed estimate (**Supplementary Figure 2**). The selection of OUMV supported significant differences in Cx between core and early diverging lineages, differential evolution of Cx in both groups as well as adaptive scenario driven by temperature as revealed by significant associations with temperature related climatic variables using PGLS (**Figure 3**, **Table 1**). Gain or loss of single chromosomes as well as deletion or proliferation of DNA are considered the main sources of Cx variation. So far there is no evidence for widespread aneuploidy in Bromelioideae, and we excluded dysploid lineages from the analysis. Hence, deletion or proliferation of DNA may be considered the sole source of recovered Cx variation most probably caused by an altered abundance of transposable elements and repetitive DNA. Higher stress-induced retrotransposon activity as a response to increased drought has been identified previously on a microscale in *Hordeum* (Kalendar et al., 2000) and retrotransposons in *Crucihimalaya himalaica*, which likely contributed to the adaptation to high altitude, proliferated shortly after the uplift and climatic change of the Himalayas from the Late Pliocene to Pleistocene (Zhang et al., 2019). In the early diverging lineages three times higher rate around the optimum Cx (σ^2) was revealed (**Table 1**), and we, therefore, assume differential and/or elevated activity of transposable elements either as a response to current or historical climate conditions.

Regarding GC content, our data revealed small but significant differences in the GC content of both groups, with slightly higher mean GC content in the core group. Interestingly, both groups reach similar maxima, and the phylogenetic signal was weaker when compared with 2C and Cx, indicating a weaker link with the separation by the group. The common evolution of GC in both groups was also confirmed by the best fit of the OU1 model with a single evolutionary optimum and adaptive consequences. Similarly, as for Cx, PGLS regression revealed significant associations of GC content with temperature related climatic variables, predicting higher GC for higher mean diurnal range (bio2) as well as temperature seasonality (bio4). Higher GC content was in monocots associated with increased tolerance and ability to grow in regions of extremely cold winters or experiencing at least some seasonal water deficiency (Šmarda et al., 2014). Although our dataset does not represent sharp seasonal climatic variation and the interval of the GC content estimates is relatively small, our results would be in line with the general pattern observed in monocots explained by the higher thermal stability of GC base pairs (Vinogradov, 2003; Šmarda and Bureš, 2012).

We could not relate the variability of Cx and GC content with any studied precipitation related variables and confirm previously suggested water deficiency as a potential evolutionary driver of genomic characters. In terrestrial tank bromelioids such as *Aechmea bromeliifolia*, *A. nudicaulis*, or *Neoregelia cruenta* the water could be stored even over longer dry periods (Lopez and Rios, 2001). Accordingly, relationships with precipitation variables could suffer from a certain bias as low precipitation does not necessarily point towards drought stress due to the water storage capacity in the tank, which buffers against temperature-limited soil water availability (Benzing, 2000).

As revealed by the best-fitting BMS model, early diverging lineages revealed 12 times higher rate of stochastic motion around optimum 2C than the core lineages, which supports differential 2C evolution of both groups. Previously, the best fit to a single-optimum OU model was suggested for Bromelioideae (Müller et al., 2019). However, only single regime models have been considered in that study, and it has been shown that the OU model can be incorrectly favored over simpler models (Cooper et al., 2015). Interestingly, no significant associations of 2C with the climatic data were recovered, which suggests that the current climate is not a good predictor for 2C. Due to polyploidy, 2Cs of closely related species are not necessarily more similar, which might disturb some of the PGLS assumptions. Using a non-phylogenetic multiple regression (**Supplementary Table 6**), we recovered a significant multiple regression model ($p < 0.01$) with 11.65% explained variation. Accordingly, together with the prevalence of polyploids in early diverging group and 12 times

higher rate of evolution, we assume that polyploidy in early diverging Bromelioideae plays a role in surviving in hot and dry climatic conditions as often assumed also in other plant groups (*e.g.* Manzaneda et al., 2012; Pyšek et al., 2018). However, to further test the associations with current climate a more comprehensive phylogenetic framework and/or additional distribution data would be advantageous as several polyploid 2C-values were missing in the analyses.

The relationship between polyploidy and harsher habitats has been hypothesized in many plant groups (Finigan et al., 2012; Manzaneda et al., 2012; Pyšek et al., 2018) and can be explained by increased heterozygosity, higher levels of diversity (*i.e.* higher number of alleles), and tolerance against increased levels of selfing or robustness of genomes against mutations (Comai, 2005). Here we additionally assessed Cx and GC content, which revealed patterns compatible with those hypothesized for polyploids. Interestingly, an analogical pattern as revealed for the whole subfamily was already observed within the early diverging genus *Orthophytum*. The polyploid *Orthophytum* species are closely associated with xeric microhabitats and possess xeromorphic traits (*e.g.* coriaceous, densely lepidote leaves) while the majority of diploid taxa were found in more mesic to wet microhabitats (Louzada et al., 2010). *Orthophytum* seems particularly interesting as it represents the strongest genomic contraction from the ancestral monoploid genome size 0.56 pg (**Supplementary Figure 1**), showing the smallest monoploid genomes in the subfamily (**Supplementary Table 1**) as well as polyploidy. Associations between genome size and different habitat preferences were also hypothesized for dysploid Cryptanthoid complex, relating bigger genomes of the genus *Cryptanthus* to moist Atlantic rainforest and smaller genomes of *Hoplocryptanthus* to dry campos rupestres (Cruz et al., 2020). Moreover, cytotypes in the genus *Fosterella* (Bromeliaceae: Pitcairnioideae) were found to be ecologically differentiated, showing that polyploids preferentially occupy colder habitats with high temperature seasonality (Paule et al., 2017) and members of polyploid *Tillandsia* subgen. *Diaphoranthema* (Tillandsioideae) were considered for high stress tolerance as they are found in more isolated and climatically extreme areas than any other *Tillandsia* species (Till, 1992).

The Quaternary climatic oscillations are considered key events in biodiversity diversification. Particularly in South America, the climatic oscillations led to a series of contractions and expansions of forest and non-forest vegetation as well as periods of aridification resulting in nowadays Cerrado and Caatinga (*e.g.* Simon et al., 2009; Werneck et al., 2012). Due to the estimated origin of the dysploidy in *Cryptanthus* at app. 2.53 Ma as well as of the earliest polyploidization events in *Bromelia* at app. 1.18 Ma most of the identified genomic alterations in the early diverging lineages could be associated with the Pleistocene climatic changes. Accordingly, the following processes or their combination can be assumed in the Quaternary history of Bromelioideae: 1) Climatic changes may have promoted repeated secondary contacts among otherwise geographically isolated Bromelioideae lineages resulting in hybridization and allopolyploidy. The combination of different genomes in these lineages might have further added towards an adaptive potential (Barker et al., 2016) and supported the survival of the early diverging bromelioids in the arid environments. A similar scenario can be assumed for other polyploid taxa from Cerrado such as *Mimosa* (Fabaceae; Morales et al., 2018) and *Eriotheca* (Malvaceae; Mendes-Rodrigues et al., 2019). However, the abundance of polyploidy and climatic preferences of polyploids in Cerrado and Caatinga remain to be tested across several taxa. 2) Climatic oscillations and/or other environmental factors could have triggered the formation of unreduced gametes as previously shown in *e.g. Solanum* (McHale, 1983). 3) Allopatric and parapatric speciation was triggered by dysploidy and polyploidy leading to reproductive isolation and the origin of polyploid and dysploid taxa such as *e.g. Deinacanthon urbanianum* or genus *Cryptanthus*, respectively. 4) Finally, it has been demonstrated that polyploids and plants with larger genomes are selectively disadvantaged under limited availability of environmental nitrogen and phosphorus (Šmarda et al., 2013; Guignard et al., 2016). Limited availability of both nutrients was exemplified in tank-forming epiphyte *Werauhia sanguinolenta* (Tillandsioideae) (Wanek and Zotz, 2011). Hence, nutrient shortage in almost exclusively tank-forming core Bromelioideae could be considered another factor contributing towards dominant diploidy in this group.

CONCLUSION

The genomes in the core Bromelioideae were revealed to be strikingly uniform concerning both ploidy as well as monoploid genome size. Hence, polyploidy and genomic reorganizations are not associated with higher net diversification and speciation in core bromelioids. On the contrary, the early diverging lineages revealed a higher incidence of polyploidy, presumed dysploidy as well as higher variation in the monoploid genome size. For Cx and GC content Ornstein–Uhlenbeck models outperformed the Brownian motion models suggesting adaptive potential linked to the temperature conditions. 2C-values revealed different rates of evolution in core and early diverging lineages also related to climatic conditions. The origins of polyploidy in the subfamily could be followed back to the Pleistocene and could be most probably attributed to the dynamic climatic conditions in the areas of today's Cerrado and Caatinga. Accordingly, although coupled with higher extinction rates, polyploidy and genomic reorganizations might have played a role in the survival of the early diverging bromelioids in the arid environments.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

JP and GZ conceived the study. EL, RM, and JM provided plant material and taxonomic determinations. JP, SH, RM, and JM

carried out the experiments. JP, SH, and JM analyzed the data and produced the figures. JP drafted the manuscript with contributions from all authors. All authors approved the final version of the manuscript.

FUNDING

This work was supported by the research funding program "LOEWE—Landesoffensive zur Entwicklung wissenschaftlich-ökologischer Exzellenz" of Hesse's Ministry of higher education; the "Freunde und Förderer" of Goethe-University Frankfurt; Paul Ungerer-Stiftung; German Research Foundation [DFG Zi 557/7-1, Schu2426/1-1]; DAAD (Deutscher Akademischer Austauschdienst), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico); CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) in scope of the programs PROTAX and PROBRAL, respectively; and FAPERJ (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro).

ACKNOWLEDGMENTS

The authors are grateful to the staff of the Botanical Gardens of the University of Heidelberg (A. Franzke, T. Stolten), the University of Vienna (W. Till, M.H.J. Barfuss), the University of Göttingen (M. Schwerdtfeger), Palmengarten Frankfurt (C. Bayer, M. Jacobi, M. Schmidt), Utrecht Botanic Gardens (E. Gouda), Berlin-Dahlem Botanical Garden and Botanical Museum (N. Köster) and Rio de Janeiro Botanical Garden for providing plant material as well as to F. Lappe (Goethe Uni, Frankfurt am Main) for lab support. The authors thank M. Forrest (Senckenberg BiK-F, Frankfurt am Main) for checking the English of the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.01295/full#supplementary-material

REFERENCES

Aguirre-Santoro, J., Michelangeli, F. A., and Stevenson, D. W. (2016). Molecular phylogenetics of the *Ronnbergia* Alliance (Bromeliaceae, Bromelioideae) and insights into their morphological evolution. *Mol. Phylogenet. Evol.* 100, 1–20. doi: 10.1016/j.ympev.2016.04.007

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210, 391–398. doi: 10.1111/nph.13698

Barow, M., and Meister, A. (2002). Lack of correlation between AT frequency and genome size in higher plants and the effect of nonrandomness of base sequences on dye binding. *Cytom. Part A* 47, 1–7. doi: 10.1002/cyto.10030

Beaulieu, J. M., Jhwueng, D.-C., Boettiger, C., and O'Meara, B. C. (2012). Modeling stabilizing selec-tion: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* 66, 2369–2383. doi: 10.1111/j.1558-5646.2012.01619.x

Benzing, D. H. (2000). *Bromeliaceae: profile of an adaptive radiation* (Cambridge: Cambridge University Press).

Brown, G. K., Palací, C. A., and Luther, H. (1997). Chromosome numbers in Bromeliaceae. *Selbyana* 18, 85–88. doi: 10.2307/2444413

Chessel, D., Dufour, A. B., and Thioulouse, J. (2004). The ade4 Package – I: One-Table Methods. *R News* 4, 5–10.

Cogliatti-Carvalho, L., Rocha-Pessôa, T. C., Nunes-Freitas, A. F., and Duarte Rocha, C. F. (2010). Volume de água armazenado no tanque de bromélias, em restingas da costa brasileira. *Acta Bot. Bras.* 24, 84–95. doi: 10.1590/S0102-33062010000100009

Collins, J. L., and Kerns, K. R. (1936). Origin and nature of tetraploid pineapples. *Am. Nat.* 70, 45.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Cooper, N., Thomas, G. H., Venditti, C., Meade, A., and Freckletom, R. P. (2015). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biol. J. Linn. Soc* 118, 64–77. doi: 10.1111/bij.12701

Cotias-de-Oliveira, A. L. P., Assis, J. G. A., Bellintani, M. C., Andrade, J. C., and Guedes, M. L. S. (2000). Chromosome numbers in Bromeliaceae. *Genet. Mol. Biol.* 23, 173–177. doi: 10.1590/S1415-47572000000100032

Crayn, D. M., Winter, K., Schulte, K., and Smith, J. A. C. (2015). Photosynthetic pathways in Bromeliaceae: phylogenetic and ecological significance of CAM and C3 based on carbon isotope ratios for 1893 species. *Bot. J. Linn. Soc* 178, 169–221. doi: 10.1111/boj.12275

Cruz, G. A. S., Mendonça Filho, J. R., Vasconcelos, S., Gitaí, J., Salabert De Campos, J. M., Facio Viccini, L., et al. (2020). Genome size evolution and chromosome numbers of species of the cryptanthoid complex (Bromelioideae, Bromeliaceae) in a phylogenetic framework. *Bot. J. Linn. Soc* 192, 887–899. doi: 10.1093/botlinnean/boz103

Doležel, J., Doleželová, M., and Novák, F. J. (1994). Flow cytometric estimation of nuclear DNA amount in diploid bananas (*Musa acuminata* and *M. balbisiana*). *Biol. Plant* 36, 351–357. doi: 10.1007/BF02920930

Doležel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysák, M. A., Nardi, L., et al. (1998). Plant genome size estimation by flow cytometry: Inter-laboratory comparison. *Ann. Bot.* 82 (Suppl. A), 17–26. doi: 10.1093/oxfordjournals.aob.a010312

Evans, T. M., Jabaily, R. S., Gelli de Faria, A. P., Sousa, L. O. F., Wendt, T., and Brown, G. K. (2015). Phylogenetic relationships in Bromeliaceae subfamily Bromelioideae based on chloroplast DNA sequence data. *Syst. Bot.* 40, 116–128. doi: 10.1600/036364415X686413

6Favoreto, F. C., Carvalho, C. R., Lima, A. B. P., Ferreira, A., and Clarindo, W. R. (2012). Genome size and base composition of Bromeliaceae species assessed by flow cytometry. *Plant Syst. Evol.* 298, 1185–1193. doi: 10.1007/s00606-012-0620-x

Finigan, P., Tanurdzic, M., and Martienssen, R. A. (2012). "Origins of novel phenotypic variation in polyploids," in *Polyploidy and genome evolution*. Eds. P. S. Soltis and D. E. Soltis (Heidelberg: Springer), 57–76.

Frank, J. H., and Lounibos, L. P. (2009). Insects and allies associated with bromeliads: a review. *Terr. Arthropod. Rev.* 1, 125–153. doi: 10.1163/187498308X414742

Furtado, S. L. O., Rezende, S. B., and Sampaio, S. R. C. O. (2003). *xHohenmea*, a new natural intergeneric hybrid in the Bromelioideae. *J. Bromeliad Soc* 53, 71–77.

Gitaí, J., Horres, R., and Benko-Iseppon, A. M. (2005). Chromosome features and evolution of Bromeliaceae. *Plant Syst. Evol.* 253, 65–80. doi: 10.1007/s00606-005-0306-8

Gitaí, J., Paule, J., Zizka, G., Schulte, K., and Benko-Iseppon, A. M. (2014). Chromosome numbers and DNA content in Bromeliaceae: additional data and critical review. *Bot. J. Lin. Soc* 176, 349–368. doi: 10.1111/boj.12211

Goetze, M., Schulte, K., Palma-Silva, C., Zanella, C. M., Büttow, M. V., Capra, F., et al. (2016). Diversification of Bromelioideae (Bromeliaceae) in the Brazilian Atlantic rainforest: A case study in *Aechmea* subgenus *Ortgiesia*. Mol. *Phylogenet. Evol.* 98, 346–357. doi: 10.1016/j.ympev.2016.03.001

Gouda, E. J., Butcher, D., and Gouda, C. S. (2020). *Encyclopaedia of Bromeliads, Version 4* (University Utrecht Botanic Gardens). Available online at: http://bromeliad.nl/encyclopedia. (Accessed on July 31, 2020).

Grafen, A. (1989). The phylogenetic regression. *Philos. Trans. R. Soc Lond. B Biol. Sci.* 326, 119–157. doi: 10.1098/rstb.1989.0106

Greilhuber, J., Doležel, J., Lysak, M. A., and Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. *Ann. Bot.* 95, 255–260. doi: 10.1093/aob/mci019

Guignard, M. S., Nichols, R. A., Knell, R. J., Macdonald, A., Romila, C.-A., Trimmer, M., et al. (2016). Genome size and ploidy influence angiosperm species' biomass under nitrogen and phosphorus limitation. *New Phytol.* 210, 1195–1206. doi: 10.1111/nph.13881

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978. doi: 10.1002/joc.1276

Horres, R., and Zizka, G. (1995). Untersuchungen zur Blattsukkulenz bei Bromeliaceae. *Beitr. Biol.* 69, 43–76.

Horres, R., Schulte, K., Weising, K., and Zizka, G. (2007). Systematics of Bromelioideae (Bromeliaceae) - evidence from molecular and antomical studies. *Aliso* 23, 27–43. doi: 10.5642/aliso.20072301.05

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., and Schulman, A. H. (2000). Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6603–6607. doi: 10.1073/pnas.110587497

Knight, C. A., Molinari, N. A., and Petrov, D. A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* 95, 177–190. doi: 10.1093/aob/mci011

Leitch, I. J., Beaulieu, J. M., Chase, M. W., Leitch, A. R., and Fay, M. F. (2010). Genome size dynamics and evolution in monocots. *J. Bot.* 2010:862516. doi: 10.1155/2010/862516

Leme, E. M., Heller, S., Zizka, G., and Halbritter, H. (2017). New circumscription of *Cryptanthus* and new Cryptanthoid genera and subgenera (Bromeliaceae: Bromelioideae) based on neglected morphological traits and molecular phylogeny. *Phytotaxa* 318, 1–88. doi: 10.11646/phytotaxa.318.1.1

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925. doi: 10.1080/106351501753462876

Lopez, L. C. S., and Rios, R. I. (2001). Phytotelmata faunal communities in sun-exposed versus shaded terrestrial bromeliads from Southeastern Brazil. *Selbyana* 22, 219–224.

Louzada, R. B., Palma-Silva, C., Corrêa, A. M., Kaltchuk-Silva, E., and Wanderley, M. G. L. (2010). Chromosome numbers of *Orthophytum* species (Bromeliaceae). *Kew Bull.* 65, 53–58. doi: 10.1007/s12225-010-9175-6

Maddison, W. P., and Maddison, D. R. (2010). *Mesquite: a modular system for evolutionary analysis. Version 2.74.* Available online at: http://mesquiteproject.org.

Males, J. (2016). Think tank: water relations of Bromeliaceae in their evolutionary context. *Bot. J. Linn. Soc* 181, 415–440. doi: 10.1111/boj.12423

Manzaneda, A. J., Rey, P. J., Bastida, J. M., Weiss-Lehman, C., Raskin, E., and Mitchell-Olds, T. (2012). Environmental aridity is associated with cytotype segregation and polyploidy occurrence in *Brachypodium distachyon* (Poaceae). *New Phytol.* 193, 797–805. doi: 10.1111/j.1469-8137.2011.03988.x

Matallana, G., Oliveira, P. E., Silva, P. R. R., and Wendt, T. (2016). Post-pollination barriers in an assemblage of Bromeliaceae in south-eastern Brazil. *Bot. J. Linn. Soc.* 181, 521–531. doi: 10.1111/boj.12406

Matuszak-Renger, S., Paule, J., Heller, S., Leme, E. M. C., Steinbeisser, G. M., Barfuss, M. H. J., et al. (2018). Phylogenetic relationships among *Ananas* and related taxa (Bromelioideae, Bromeliaceae) based on nuclear, plastid and AFLP data. *Plant Syst. Evol.* 304, 841–851. doi: 10.1007/s00606-018-1514-3

McHale, N. A. (1983). Environmental induction of high frequency of 2n pollen formation in diploid *Solanum*. *Can. J. Genet. Cytol.* 25, 609–615. doi: 10.1139/G07-011

Morales, M., Fradkin, M., Bessega, C., Poggio, L., and Fortunato, R. H. (2018). Cytogeography and morphological characterisation of a taxonomic, polyploid complex of *Mimosa* (Leguminosae) from subtropical South America. *Aust. Syst. Bot.* 31, 190–208. doi: 10.1071/SB16033

Mendes-Rodrigues, C., Marinho, R. C., Balao, F., Artista, M., Ortiz, P. L., Carmo-Oliveira, R., et al. (2019). Reproductive diversity, polyploidy, and geographical parthenogenesis in two *Eriotheca* (Malvaceae) species from Brazilian Cerrado. *Perspect. Plant Ecol. Syst.* 36, 1–12. doi: 10.1016/j.ppees.2018.11.001

Müller, L. L. B., Zotz, G., and Albach, D. C. (2019). Bromeliaceae subfamilies show divergent trends of genome size evolution. *Sci. Rep.* 9, 5136. doi: 10.1038/s41598-019-41474-w

Orme, D., Freckleton, R. P., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., et al. (2018). *Caper: comparative analyses of phylogenetics and evolution in R. R package version 1.0.1*. Available online at: http://cran.r-project.org/web/packages/caper/.

Otto, F. (1990). DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. *Methods Cell Biol.* 33, 105–110. doi: 10.1016/S0091-679X(08)60516-6

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* 401, 877–884. doi: 10.1038/44766

Paule, J., Wagner, N. D., Weising, K., and Zizka, G. (2017). Ecological range shift in the polyploid members of the South American genus *Fosterella* (Bromeliaceae). *Ann. Bot.* 120, 233–243. doi: 10.1093/aob/mcw245

Pellicer, J., Hidalgo, O., Dodsworth, S., and Leitch, I. J. (2018). Genome size diversity and its impact on the evolution of land plants. *Genes* 9, 88. doi: 10.3390/genes9020088

Pyšek, P., Skálová, H., Čuda, J., Guo, W.-Y., Suda, J., Doležal, J., et al. (2018). Small genome separates native and invasive populations in an ecologically important cosmopolitan grass. *Ecology* 99, 79–90. doi: 10.1002/ecy.2068

R Core Team (2020). *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing).

Ramírez-Morillo, I. M., and Brown, G. K. (2001). The origin of the low chromosome number in *Cryptanthus* (Bromeliaceae). *Syst. Bot.* 26, 722–726. doi: 10.1043/0363-6445-26.4.722

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi: 10.1111/j.2041-210X.2011.00169.x

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16, 351–358. doi: 10.1016/S0169-5347(01)02187-5

Schulte, K., Horres, R., and Zizka, G. (2005). Molecular phylogeny of Bromelioideae and its implications on biogeography and the evolution of CAM in the family (Poales, Bromeliaceae). *Senckenb Biol.* 85, 113–125.

Schulte, K., Barfuss, M. H. J., and Zizka, G. (2009). Phylogeny of Bromelioideae (Bromeliaceae) inferred from nuclear and plastid DNA loci reveals the evolution of the tank habit within the subfamily. *Molec. Phylogenet. Evol.* 51, 327–339. doi: 10.1016/j.ympev.2009.02.003

Silvestro, D., Zizka, G., and Schulte, K. (2014). Disentangling the effects of key innovations on the diversification of Bromelioideae (Bromeliaceae). *Evolution* 68, 163–175. doi: 10.1111/evo.12236

Simon, M. F., Grether, R., de Queiroz, L. P., Skema, C., Pennington, R. T., and Hughes, C. E. (2009). Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proc. Natl. Acad. Sci. U. S. A.* 106, 20359–20364. doi: 10.1073/pnas.0903410106

Šmarda, P., and Bureš, P. (2012). "The variation of base composition in plant genomes," in *Plant genome diversity*, Vol. 1. Eds. J. F. Wendel, J. Greilhuber, J. Doležel and I. J. Leitch (Vienna: Springer), 209–235.

Šmarda, P., Bureš, P., Horová, L., Foggi, B., and Rossi, G. (2008). Genome size and GC content evolution of *Festuca*: ancestral expansion and subsequent reduction. *Ann. Bot.* 101, 421–433. doi: 10.1093/aob/mcm307

Šmarda, P., Hejcman, M., Brezinová, A., Horová, L., Steigerová, H., Zedek, F., et al. (2013). Effect of phosphorus availability on the selection of species with different ploidy levels and genome sizes in a long-term grassland fertilization experiment. *New Phytol.* 200, 911–921. doi: 10.1111/nph.12399

Šmarda, P., Bureš, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., et al. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc. Natl. Acad. Sci. U. S. A.* 111, E4096–E4102. doi: 10.1073/pnas.1321152111

Smith, L. B., and Downs, R. J. (1979). "Flora Neotropica Monograph No. 14 Part 3: Bromelioideae (Bromeliaceae)," in *Flora Neotropica*. Ed. C. T. Rogerson (New York, NY: Organization for Flora Neotropica, New York Botanical Garden), 1493–2142.

Souza, E. H., Versieux, L. M., Souza, F. V. D., Rossi, M. L., Costa, M. A. P. C., and Martinelli, A. P. (2017). Interspecific and intergeneric hybridization in Bromeliaceae and their relationships to breeding systems. *Sci. Hortic.* 223, 53–61. doi: 10.1016/j.scienta.2017.04.027

Till, W. (1992). Systematics and evolution of the tropical-subtropical *Tillandsia* subgenus *Diaphoranthema* (Bromeliaceae). *Selbyana* 13, 88–94.

Trávníček, P., Čertner, M., Ponert, J., Chumová, Z., Jersáková, J., and Suda, J. (2019). Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. *New Phytol.* 224, 1642–1656. doi: 10.1111/nph.15996

Vinogradov, A. E. (2003). DNA helix: the importance of being GC-rich. *Nucleic Acids Res.* 31, 1838–1844. doi: 10.1093/nar/gkg296

Wanek, W., and Zotz, G. (2011). Are vascular epiphytes nitrogen or phosphorus limited? A study of plant ^{15}N fractionation and foliar N:P stoichiometry with the tank bromeliad *Vriesea sanguinolenta*. *New Phytol.* 192, 462–470. doi: 10.1111/j.1469-8137.2011.03812.x

Wendt, T., Coser, T. S., Matallana, G., and Guilherme, F. A. G. (2008). An apparent lack of prezygotic reproductive isolation among 42 sympatric species of Bromeliaceae in southeastern Brazil. *Plant Syst. Evol.* 275, 31–41. doi: 10.1007/s00606-008-0054-7

Werneck, F. P., Nogueira, C., Colli, G. R., Sites, J. W. Jr., and Costa, G. C. (2012). Climatic stability in the Brazilian Cerrado: Implications for biogeographical connections of South American savannas, species richness and conservation in a biodiversity hotspot. *J. Biogeogr.* 39, 1695–1706. doi: 10.1111/j.1365-2699.2012.02715.x

Wheeler, W. M. (1942). Studies of Neotropical ant plants and their ants. *Bull. Mus. Comp. Zool.* 90, 1–262. doi: 10.5281/zenodo.25273

Zhang, F., Ge, Y.-Y., Wang, W.-Y., Shen, X.-L., and Yu, X.-Y. (2012). Assessing genetic divergence in interspecific hybrids of *Aechmea gomosepala* and *A. recurvata* var. *recurvata* using inflorescence characteristics and sequence-related amplified polymorphism markers. *Genet. Mol. Res.* 11, 4169–4178. doi: 10.4238/2012.September.27.2

Zhang, T., Qiao, Q., Novikova, P. Y., Wang, Q., Yue, J., Guan, Y., et al. (2019). Genome of *Crucihimalaya himalaica*, a close relative of *Arabidopsis*, shows ecological adaptation to high altitude. *Proc. Natl. Acad. Sci. U. S. A.* 116, 7137–7146. doi: 10.1073/pnas.1817580116

Zizka, A., Azevedo, J., Leme, E., Neves, B., da Costa, A. F., Caceres, D., et al. (2020). Biogeography and conservation status of the pineapple family (Bromeliaceae). *Divers. Distrib.* 26, 183–195. doi: 10.1111/ddi.13004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 02 November 2020
doi: 10.3389/fpls.2020.591053

Characterization and Dynamics of Repeatomes in Closely Related Species of *Hieracium* (Asteraceae) and Their Synthetic and Apomictic Hybrids

Danijela Zagorski*, Matthias Hartmann, Yann J. K. Bertrand, Ladislava Paštová, Renata Slavíková, Jiřina Josefiová and Judith Fehrer

Institute of Botany, Czech Academy of Sciences, Průhonice, Czechia

OPEN ACCESS

Edited by:
Elvira Hörandl,
University of Göttingen, Germany

Reviewed by:
Sònia Garcia,
Botanical Institute of Barcelona, Spain
Ales Kovarik,
Academy of Sciences of the Czech Republic (ASCR), Czechia

****Correspondence:***
Danijela Zagorski
danijela.zagorski@gmail.com

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *03 August 2020*
Accepted: *09 October 2020*
Published: *02 November 2020*

Citation:
Zagorski D, Hartmann M, Bertrand YJK, Paštová L, Slavíková R, Josefiová J and Fehrer J (2020) Characterization and Dynamics of Repeatomes in Closely Related Species of Hieracium (Asteraceae) and Their Synthetic and Apomictic Hybrids. Front. Plant Sci. 11:591053. doi: 10.3389/fpls.2020.591053

The repetitive content of the plant genome (repeatome) often represents its largest fraction and is frequently correlated with its size. Transposable elements (TEs), the main component of the repeatome, are an important driver in the genome diversification due to their fast-evolving nature. Hybridization and polyploidization events are hypothesized to induce massive bursts of TEs resulting, among other effects, in an increase of copy number and genome size. Little is known about the repeatome dynamics following hybridization and polyploidization in plants that reproduce by apomixis (asexual reproduction via seeds). To address this, we analyzed the repeatomes of two diploid parental species, *Hieracium intybaceum* and *H. prenanthoides* (sexual), their diploid F1 synthetic and their natural triploid hybrids (*H. pallidiflorum* and *H. picroides*, apomictic). Using low-coverage next-generation sequencing (NGS) and a graph-based clustering approach, we detected high overall similarity across all major repeatome categories between the parental species, despite their large phylogenetic distance. Medium and highly abundant repetitive elements comprise ∼70% of *Hieracium* genomes; most prevalent were Ty3/Gypsy chromovirus Tekay and Ty1/Copia Maximus-SIRE elements. No TE bursts were detected, neither in synthetic nor in natural hybrids, as TE abundance generally followed theoretical expectations based on parental genome dosage. Slight over- and under-representation of TE cluster abundances reflected individual differences in genome size. However, in comparative analyses, apomicts displayed an overabundance of pararetrovirus clusters not observed in synthetic hybrids. Substantial deviations were detected in rDNAs and satellite repeats, but these patterns were sample specific. rDNA and satellite repeats (three of them were newly developed as cytogenetic markers) were localized on chromosomes by fluorescence *in situ* hybridization (FISH). In a few cases, low-abundant repeats (5S rDNA and certain satellites) showed some discrepancy between NGS data and FISH results, which is due partly to the bias of low-coverage sequencing and partly to low amounts of the

satellite repeats or their sequence divergence. Overall, satellite DNA (including rDNA) was markedly affected by hybridization, but independent of the ploidy or reproductive mode of the progeny, whereas bursts of TEs did not play an important role in the evolutionary history of *Hieracium*.

Keywords: apomixis, hybridization, polyploidization, RepeatExplorer, next-generation sequencing, repeatome, hawkweed

INTRODUCTION

Repetitive elements, collectively known as the 'repeatome' (Maumus and Quesneville, 2014), comprise DNA fragments that are present in multiple copies throughout the genome. Due to their fast-evolving nature and tendency toward accumulation, they represent one of the most important factors contributing to the remarkable variation of genome size in plants (Bennetzen et al., 2005). Based on their genome organization, repetitive elements fall into two main categories: tandem repeats and transposable elements (TEs).

Tandem repeats are composed of multiple copies of the same DNA sequence (monomers) arrayed in a head-to-tail fashion. They are usually classified based on the length of the monomer into microsatellite (up to 10 bp), minisatellite (10–60 bp), and satellite repeats (from tens of base pairs up to several kilobases (Stupar et al., 2002; Ávila Robledillo et al., 2018). Their amount, chromosomal distribution, and homology may be genus-, species-, genome- or chromosome-specific, which consequently make them useful markers for inheritance studies of interspecific hybrids (Hemleben et al., 2007).

TEs are DNA sequences that can copy and insert themselves to different locations within a genome. TEs can be classified according to their mode of transposition: Class I (retrotransposons) that replicate and transpose via an RNA intermediate, and Class II (DNA transposons) that directly excise themselves and insert into a new location without any intermediate. Each class is divided into further subcategories based on the presence and structure of protein domains and different non-coding sequences (Wicker et al., 2007). In plants, TE content is highly variable, e.g., from 10% in *Arabidopsis thaliana*, to over 85% in the maize genome (Dubin et al., 2018).

TEs play many important roles such as shaping the genome architecture (reviewed in Galindo-González et al., 2017). Furthermore, TEs can alter gene expression if inserted into genes or promoter regions. TEs are controlled by various silencing mechanisms (Okamoto and Hirochika, 2001), therefore factors decreasing silencing efficacy can trigger TE activity.

McClintock (1984) hypothesized that hybridization in plants may represent a 'genomic shock' inducing bursts of TEs, which may lead to massive genome reorganization in newly formed hybrids. Similar effects of TE activation and transposition have been suggested for polyploidization (Comai, 2000). In line with this hypothesis, several investigations into the effects of hybridization and polyploidization on TEs (mostly in synthetic or recent natural hybrids and allopolyploids) showed that some TEs responded with increased transcriptional activity (e.g., Kashkush et al., 2002; Madlung et al., 2005), methylation changes (Xu et al., 2009; Kraitshtein et al., 2010; Yaakov and Kashkush, 2010), and induction of changes in genome structure (Madlung et al., 2005). However, only few studies have detected bursts of TEs that resulted in an increase of copy number (Petit et al., 2010; Ben-David et al., 2013; An et al., 2014; Hake et al., 2018) and transposon mobility (Shan et al., 2005) in response to hybridization and polyploidization events, whereas a number of studies showed no or only limited TE transposition (e.g., Madlung et al., 2005; Beaulieu et al., 2009; Parisod et al., 2009; Mestiri et al., 2010; Sarilar et al., 2013; see also reviews of Parisod and Senerchia, 2012 and Vicient and Casacuberta, 2017). Some studies reported the increased amplification of certain TEs in natural hybridogenous and polyploid species compared to their diploid ancestors (e.g., Sarilar et al., 2011; Piednoël et al., 2015). Thus, TE response to hybridization and polyploidization differs between genomes and types of TEs, indicating that TE bursts are not a general consequence of these processes (Shan et al., 2005; Parisod and Senerchia, 2012; Belyayev, 2014; Vicient and Casacuberta, 2017). Also, the influence of other factors that trigger TE activity such as environmental stress (Casacuberta and González, 2013) in the study of old hybrids and polyploids cannot be excluded.

The mode of reproduction may constitute one of the major factors affecting plant genome size and TE content. Theoretical predictions about the direction of TE content change (and consequently the increase or decrease of genome size) in selfers and asexuals in comparison to their sexual relatives, remain unclear (Glémin and Galtier, 2012). In the simplest scenario, sexual reproduction represents the most favorable means for TEs to spread horizontally to all lineages within the population (Hickey, 1982). Despite being considered as selfish elements that have mostly adverse effects on the host's fitness, TEs can develop maximum transposition rates in sexual populations (Charlesworth and Langley, 1986). Conversely, in asexual populations, the propagation of TEs is limited to vertical (within-lineage) transmission. As asexuals are usually derived from sexual progenitors, they inherit their parental TE load, including active elements. In order to survive the detrimental proliferation of TEs, asexuals should purge and inactivate their TE load by means of self-regulatory mechanisms. Ultimately, over longer evolutionary time, those asexual lineages with a more efficiently inactivated TE load would be selected over lineages with higher TE proliferation rates. In this case, it is expected that asexuals would have smaller TE contents and genome sizes than sexuals (Dolgin and Charlesworth, 2006; Ågren et al., 2015). On the other hand, as the purging can take a very long time, in small asexual populations, due to genetic drift, a Muller's ratchet-like process will drive the accumulation of

TEs (Dolgin and Charlesworth, 2006; Glémin and Galtier, 2012). Under this scenario, it is expected that asexuals would have a higher TE content and genome size than their sexual relatives. Furthermore, asexually reproducing plants often emerge through the process of hybridization of (predominantly) sexual species, which is almost always followed by polyploidization. These events may be accompanied by massive genome re-patterning, including bursts of TEs, which additionally complicates the study of the influence of the reproductive mode on repeatome dynamics in asexual plants. It is therefore of invaluable interest to disentangle the consequences of hybridization and/or polyploidization from the consequences of the transition to the asexual mode of reproduction.

The theoretical predictions are mostly based on computer simulations, and empirical studies of repeatomes in asexually reproducing plants are scarce (e.g., Docking et al., 2006; Ågren et al., 2015; Ferreira de Carvalho et al., 2016). Until recently, technologies suitable for studying TEs in plants were not available and repeatome studies were mostly focused on a small number of model plant species. The advent of next-generation sequencing (NGS) and bioinformatics tools such as the graph-based clustering approach (Novák et al., 2010, 2013) have enabled the comprehensive characterization of repeatomes of non-model plant species at low cost and without a need for an assembled genome.

The genus *Hieracium* L. (Asteraceae) represents a model system for the study of apomixis, which is asexual reproduction via seeds. Concerted events of hybridization, polyploidization, and shifts to apomixis have played a major role in *Hieracium* evolution (Fehrer et al., 2009). This genus consists of ca. 25 sexual diploids ($x = 9$) and 500–5000 (Majeský et al., 2017) polyploid taxa (mostly triploid, $2n = 3x = 27$, and tetraploid, $2n = 4x = 36$); the latter reproduce almost exclusively apomictically (Mráz and Zdvořák, 2019). Diploid species are well differentiated morphologically, and together with several apomictic polyploid taxa they form so-called 'basic' species (altogether ca. 45 species) and are considered as main units of species evolution in the genus *Hieracium* (Chrtek et al., 2009). The rest of polyploid taxa are considered 'intermediate' species as they share morphological characters of two or more basic species and are supposed to be of hybridogenous origin. The division on 'basic' and 'intermediate' species is still largely based on morphology.

In a recent study, Chrtek et al. (2020) used complementary phylogenetic and cytogenetic approaches to study one such system consisting of two parental 'basic' species, *H. intybaceum* (Int) and *H. prenanthoides* (Pre), and two 'intermediate' triploid species, *H. pallidiflorum* (Pal) and *H. picroides* (Pic). The authors demonstrated that the two 'intermediates' indeed originated through hybridization of the parental pair as suggested by their morphology. The genome dosage of *H. pallidiflorum* (morphologically closer to *H. intybaceum*) is 2 Int : 1 Pre, and that of *H. picroides* (morphologically closer to *H. prenanthoides*) is 1 Int : 2 Pre. The two apomictic lineages have each originated independently multiple times (polytopic speciation). In addition to these natural polyploid hybrids, experimental crosses between the parental species successfully produced synthetic F1 diploid sexual hybrids (genome dosage 1 Int : 1 Pre) with intermediate morphology. Therefore, this system–consisting of the diploid parental species, old apomictic allopolyploids, and synthetic F1 sexual diploid hybrids–provides an excellent model to investigate and compare the short- and long-term consequences of hybridization and polyploidization events, followed by the shift to apomixis.

Chrtek et al. (2020) carried out basic repeatome comparisons of the parental species *H. intybaceum* and *H. prenanthoides*, and genomic *in situ* hybridization (GISH) on their hybrids. The analyses revealed a surprising repeatome similarity of the parental species, that is not reflected by their large phylogenetic divergence based on several molecular markers (Fehrer et al., 2007, 2009; Krak et al., 2013). The same study showed small, but stable differences in genome size between the parental species. Genomes of allopolyploid *H. pallidiflorum* samples were slightly larger than the sum of the two parental genomes (theoretical expectation) while the genomes of allopolyploid *H. picroides* samples showed more diversity, with both higher or lower sizes compared to the theoretical expectation.

In the present contribution, we perform a detailed repeatome analysis of the plant model system introduced by Chrtek et al. (2020), using the graph-based clustering approach (Novák et al., 2010, 2013). We aim to: 1) characterize the repeatomes of the parental *Hieracium* species and their natural apomictic and synthetic F1 hybrids, 2) compare diploid synthetic hybrids and allopolyploid apomicts to their parental species to detect changes in the repeatome following hybridization and polyploidization events, and 3) compare the repeatomes of synthetic hybrids and allopolyploid apomicts in order to discover repeatome patterns or compositions that would be apomict-specific. As we discovered a significant deviation in the amount of rDNA and satellite repeats in apomicts and synthetic hybrids, we localized these loci on the chromosomes by fluorescence *in situ* hybridization (FISH).

MATERIALS AND METHODS

Plant Material, Sampling, and Genome Size Estimation

In total, eleven accessions have been included in this study. We sampled two individuals for each of the diploid parental species (*H. intybaceum* [Int], *H. prenanthoides* [Pre]) and their natural triploid hybrids (*H. pallidiflorum* [Pal], *H. picroides* [Pic]) and three for the synthetic F1 diploid sexual hybrids (Hyb). Samples of the triploids were chosen from a larger set of polytopic populations, based on their intraspecific differences in genome size and maternal origins (Pic) or different geographic origins (Pal). Pic was represented with two accessions that showed the highest (accession no.: H1613, ID: PicF) and the lowest genome size (accession no.: H1615, ID: PicB). Their genome sizes were slightly outside of the range of genome sizes that would be theoretically expected based on summation of the genome sizes of parental species. Samples of Pal (accession no.: H1609 [ID: PalA] and accession no.:H1614 [ID: PalF]) had slightly higher genome size than theoretically expected (data from Chrtek et al., 2020). Synthetic hybrids (accession no.: 17038_2, 17038_3 and 17038_4 [IDs: Hyb2, Hyb3 and Hyb4]) were obtained from the same

homoploid cross between two of the parental accessions included in this study (accession no.: int_1531/8 [IntA] and pre_6/8/5 [PreC]). Details about the accessions' origins and genome sizes are provided in **Table 1**.

DNA Isolation, NGS Sequencing, and Reads' Pre-processing

DNA was extracted from fresh or silica-gel dried leaf tissue using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Library preparation and low-coverage Illumina NGS were performed at GATC Biotech (Konstanz, Germany) / Eurofins Genomics (Ebersberg, Germany) using a standardized protocol that produced 150 bp paired-end reads with an insert size of ∼450 bp. For selected parental accessions (IntA, IntC and PreA), two independent libraries were prepared and sequenced in order to test for potential bias at the level of library preparation. The sequencing resulted in datasets of 25–54 million individual pair-end reads (i.e., 12.5–27 million of pairs) per library (**Table 1**). The raw Illumina datasets have been submitted to the European Nucleotide Archive (ENA) under study no. PRJEB35856. Quality filtering was performed using the pre-processing tool included in the public Galaxy server running the RepeatExplorer pipeline,[1] with a Phred quality score 10 over at least 95% of the bases in a sequence, with no ambiguous bases (Ns) allowed. Only proper pairs of reads of the same length (150 bp) were retained. Reads were scanned for overlap, and only non-overlapping pairs of reads were kept for further analysis in order to increase the representation of the genome. In the case of datasets used for comparative analyses (see below), prior filtering of plastid and mitochondrial reads was included because of the high variability in the amount of chloroplast sequences among samples (3–13% of the total reads) that may distort the quantification of TEs. The filtering was done with the script 'bbsplit.sh' from BBTools v.37.44[2] using the most closely related genomes currently available, i.e., the chloroplast genome of *Lactuca sativa* (DQ383816.1) and the mitochondrial genome of *Helianthus annuus* (MG770607.2).

Graph-Based Clustering Approach

The repeatomes of all accessions were analyzed using the graph-based clustering approach, as described in Novák et al. (2010, 2013) employing the RepeatExplorer (RE) pipeline implemented within the Galaxy server. This approach allows *de novo* identification of repetitive elements from unassembled reads obtained from low-coverage genome sequencing. The pipeline performs all-to-all BLAST comparisons of the input reads (using a threshold of 90% or higher similarity over at least 55% of the read length), which is then used to construct a graph wherein reads are represented as vertices and their sequence similarities as edges. The graph is further partitioned into highly connected communities of vertices (called 'clusters'), which tends to cluster reads from the same family of repetitive elements. The clusters of reads are quantified, annotated and reads within clusters are

[1] https://repeatexplorer-elixir.cerit-sc.cz/galaxy/

[2] http://sourceforge.net/projects/bbmap/

TABLE 1 | List of samples, genome size, and initial NGS datasets.

Species, accession labels	ID	Locality[a]	Ploidy level	Genome size 2C (pg)	No. of raw NGS pair-end reads
H. intybaceum					
int_1531/8	IntA	Austria, Tirol, Arlberg Massif: Arlbergpass	2x	7.540 [a]	45,992,578 27,990,512 [d]
int_6/14/25	IntC	France, Savoie: Col du Petit Saint-Bernard	2x	7.590[a]	37,998,498 35,974,348[d]
H. prenanthoides					
pre_6/5/5	PreA	Italy, Piedmont: Claviere	2x	7.220[a]	52,429,354 45,062,158 [d]
pre_6/8/5	PreC	Italy, Piedmont: Claviere	2x	7.180[b]	26,260,212
***H. pallidiflorum* (2 Int × 1 Pre)**					
H1609	PalA	Austria, Bundesland Salzburg: Muhr	3x	11.334[a]	24,970,582
H1614	PalF	France, Savoie: Col du Petit Saint-Bernard	3x	11.584[a]	28,335,496
***H. picroides* (1 Int × 2 Pre)**					
H1613	PicF	France, Hautes-Alpes: Col du Lautaret	3x	11.412[a]	30,698,202
H1615	PicB	France, Savoie: Col du Petit Saint-Bernard	3x	10.659[a]	26,850,644
Synthetic F1 hybrids *H. intybaceum* × *H. prenanthoides* (IntA × PreC)					
17038_2	Hyb2	experimental garden, Průhonice, Czechia	2x	7.360[c]	26,922,000
17038_3	Hyb3	experimental garden, Průhonice, Czechia	2x	7.360[c]	24,944,800
17038_4	Hyb4	experimental garden, Průhonice, Czechia	2x	7.341[a]	26,225,800

[a] *For details of accession origin, voucher information and genome sizes, see Chrtek et al. (2020).*
[b] *mean value of all genome size measurements for diploid H. prenanthoides from Chrtek et al. (2009), Chrtek et al. (2020).*
[c] *values are calculated based on the average of genome sizes of parental accessions.*
[d] *for samples IntA, IntC and PreA two independent libraries were sequenced.*

assembled into contigs. In addition to clusters, the pipeline also produces superclusters, which represent groups of clusters that share split pair-end reads. Superclusters occur when the reads from the same read pair (which originate from the same insert in the NGS library) end up in two separated clusters during the clustering phase, and therefore such clusters are assumed to originate from a single repeat. Thus, information about shared pair-end reads can be used in annotation of otherwise unannotated clusters, if they are connected to annotated clusters within the supercluster. Following the criteria described in McCann et al. (2018), clusters were considered to belong to the same supercluster, if the ratio of the number of pair-end reads shared between clusters to the sum of the total number of unpaired reads in each cluster was higher than 0.1.

Testing the Bias Due to Library Preparation and Similarity Between Parental Accessions

Library preparation for NGS can be a source of significant bias (van Dijk et al., 2014), especially when dealing with closely related/highly similar samples. We determined how this bias affects cluster quantification based on different libraries from the same accession. We sequenced two independently prepared libraries for three selected parental accessions (IntA, IntC and PreA; **Table 1**). The bias was investigated using the functionality of RE comparative clustering, where we performed a set of 'intra-library' comparative runs (comparison of two independent subsamples of 2 million reads originating from the *same* library) and 'inter-library' comparative runs (comparison of independent subsamples of 2 million reads originating from *different* libraries of the same accession). The read datasets were created using a random sampling tool on the Galaxy server. The variation in cluster sizes in those comparative runs was statistically assessed using a pairwise Wilcoxon test. To this extent, the absolute difference in read proportion across all clusters was used as response. The results of intra-library and inter-library comparisons were further assessed with intra-specific and inter-specific comparative runs, which were prepared in the same way. All statistical analyses in this study were carried out in R (R Core Development Team, 2018).

Individual Clustering Analysis and Repeat Identification

For annotation purposes, all accessions were first analyzed individually using the maximum number of pair-end reads (**Table 2**) allowed by the pipeline, given the available computational resources (max. 112 GB of RAM) and depending on the repeat content of the analyzed genome (i.e., the more repetitive a genome, the smaller the number of reads that can be used). Using the highest possible number of reads ensures the best accuracy of the annotation and maximizes the possibility of detecting repeats with lower genomic proportions. Clusters were quantified, and those containing at least 0.01% of the total read input were automatically annotated using the DNA and reference TEs domain database available within RE (Viridiplantae version 2.2). The annotation results were manually

TABLE 2 | Number of reads used by RepeatExplorer in individual analyses and main clustering results.

Sample ID and ploidy level	Genome size 1C (Gbp)	No. of analyzed reads[a]	Coverage	Reads in all clusters including small uncharacterized clusters (%)	Reads in clusters above the threshold of 0.01% (%)	Singlets (%)	No. of clusters above the threshold of 0.01%	No. of superclusters above the threshold of 0.01%
IntA (2x)	3.686	4,349,101	0.18 x	82.87	71.54	17.13	246	157
IntC (2x)	3.711	4,270,305	0.17 x	81.69	69.87	18.31	224	145
PreA (2x)	3.531	4,212,792	0.18 x	82.55	71.11	17.45	230	150
PreC (2x)	3.511	4,368,530	0.19 x	81.42	69.48	18.58	228	149
PalA (3x)	5.543	4,278,270	0.12 x	81.39	70.53	18.61	236	149
PalF (3x)	5.664	4,467,725	0.12 x	82.19	70.92	17.81	242	158
PicF (3x)	5.580	3,767,752	0.10 x	80.74	69.96	19.26	209	138
PicB (3x)	5.212	4,338,941	0.12 x	81.31	69.60	18.69	260	173
Hyb2 (2x)	3.600	4,457,370	0.20 x	82.01	71.22	17.99	241	160
Hyb3 (2x)	3.600	4,636,407	0.20 x	81.82	71.06	18.18	251	161
Hyb4 (2x)	3.590	4,465,263	0.19 x	81.83	71.11	18.17	246	157

[a]*After exclusion of plastid and mitochondrial clusters.*

checked and corrected. Clusters from plastid and mitochondrial origins were identified and excluded from further analysis. Following McCann et al. (2018), clusters were annotated if at least 5% of their reads produced a BLAST hit to one or more protein domains that belong to the same lineage of transposable elements. Additionally, in order to improve cluster annotation, their contigs were subjected to a RepeatMasker search against the Viridiplantae database (Jurka et al., 2005) and to BLASTn and BLASTX searches against public databases.[3] Tandem Repeat Finder (Benson, 1999) and the YASS genomic similarity tool[4] (Noé and Kucherov, 2005) were used for the discovery of potential tandem repeats. Finally, annotated contigs of parental species were used to double-check the annotation of clusters of natural and synthetic hybrids. Annotation was done primarily at the level of superclusters, but also at the level of clusters, and both results were compared.

Comparative Clustering Analysis

Besides individual analyses, read datasets of parental species and their natural or synthetic hybrids were subjected to comparative clustering analyses, which precisely detect differences in the abundance of specific sequence variants of repetitive elements. In order to maximize the number of reads per sample, we performed three sets of analyses with a maximum number of three samples in each analysis (**Table 3**): comparisons between (1) parental species; (2) natural hybrids and their parental species; and (3) synthetic F1 hybrids and their actual parents. As for comparisons in (1) and (2), we created 'in silico' parents (IntX and PreX) by pooling reads from all sequenced libraries of both individuals of *H. intybaceum* and *H. prenanthoides*. This approach aims to increase sampling diversity of the parental species since the actual parental genotypes involved in the origins of Pic and Pal accessions are unknown. Read datasets were created by random subsampling and ensuring equal coverage of all samples within each analysis, i.e., taking into account genome size differences and ploidy level (**Table 3**). Like in individual analyses, only clusters exceeding the threshold of 0.01% were annotated at the supercluster level. Cluster abundances were analyzed in two ways (see Macas et al., 2011; Renny-Byfield et al., 2013; McCann et al., 2018): (1) by comparing their absolute sizes (in read numbers), and (2) by calculating a ratio of cluster abundances between samples, which removes cluster size effect. In the case of (1), we statistically assessed whether overall cluster abundances of each natural and synthetic hybrid are closer to either parent. For this analysis, we first recalculated abundances of natural triploids to correspond to the monoploid genome size (1Cx) in order to make them comparable to homoploid genome size (1C) of diploid parents. Then, we used the absolute difference in the number of reads between parent 1 and hybrid and compared it to the absolute difference in the number of reads between parent 2 and hybrid. We applied a pairwise non-parametric Wilcoxon test on all differences across all clusters, separately for each parental-hybrid combination. We assessed the degree of 'intermediacy' ($\text{Difference}_{\text{Parent1}-\text{Hybrid}} - \text{Difference}_{\text{Parent2}-\text{Hybrid}}$) across all parent1-hybrid-parent2 combinations and determined which hybrid type (natural or synthetic) has a more 'intermediate position' using a Wilcoxon test. In the case of (2) we compared the actual cluster abundances of natural polyploid apomictic and synthetic diploid F1 hybrids with the theoretically expected abundances, which were produced by summing up the number of reads of the parental samples in each cluster and taking into account the actual genome dosage (i.e., 2 Int : 1 Pre for *H. pallidiflorum*, 1 Int : 2 Pre for *H. picroides* and 1 Int : 1 Pre for F1 hybrids). For this purpose, deviation scores for all clusters were calculated following Renny-Byfield et al. (2013), using the formula: (observed cluster size/expected cluster size) - 1. Cumulative deviation scores were calculated for each sample and compared among them. In order to check for a significant difference between natural and synthetic hybrids, we compared the deviation scores' means and the magnitude of variation between groups of samples using ANOVA and Levene test. Finally, individual clusters showing the highest deviation from the expected values were identified and evaluated for patterns of deviation across samples that could be group-specific. The reliability of their abundance was double-checked by BLASTn searches (version BLAST+ 2.6.0; Altschul et al., 1990) of the initial, quality- and chloroplast/mitochondrial-filtered NGS datasets using their contigs as queries. The BLASTn parameters followed those in RE for assembling clusters.

[3] https://blast.ncbi.nlm.nih.gov/Blast.cgi

[4] http://bioinfo.lifl.fr/yass/yass.php

TABLE 3 | Combinations of samples subjected to comparative RepeatExplorer analysis.

Sample IDs and numbers of reads	Total no. of reads	Coverage[a]
a) Comparison of the representatives of parental species		
IntX[b] (1,622,720) + PreX[c] (1,544,840)	3,167,560	0.066
b) Comparison of natural apomictic hybrids (3x) and representatives of parental species		
PalA (2,438,920) + IntX[b] (1,622,720) + PreX[c] (1,544,840)	5,606,480	0.066
PalF (2,492,160) + IntX[b] (1,622,720) + PreX[c] (1,544,840)	5,659,720	0.066
PicF (2,455,200) + IntX[b] (1,622,720) + PreX[c] (1,544,840)	5,622,760	0.066
PicB (2,293,280) + IntX[b] (1,622,720) + PreX[c] (1,544,840)	5,460,840	0.066
c) Comparison of synthetic hybrids (2x) and their actual parents		
Hyb2 (1,920,000) + IntA (1,968,000) + PreC (1,872,000)	5,760,000	0.080
Hyb3 (1,920,000) + IntA (1,968,000) + PreC (1,872,000)	5,760,000	0.080
Hyb4 (1,914,668) + IntA (1,968,000) + PreC (1,872,000)	5,754,668	0.080

[a] *Coverage calculated using 1C-values.*

[b] *IntX: sample created by pooling reads from both H. intybaceum individuals.*

[c] *PreX: sample created by pooling reads from both H. prenanthoides individuals.*

Development and Application of New Satellite Probes

Altogether, besides rDNAs, six satellite repeats were detected in *Hieracium* genomes using RE. Three of them were

already described and used as FISH probes in Belyayev et al. (2018). We used contigs of the remaining three new satellites (**Supplementary Figure 1**) as templates for development of new FISH probes and tested their usability as cytogenetic markers.

The first novel satellite repeat, originally detected as cluster (CL) 229 in IntA, is a perfect inverted repeat of 46 bp, a sequence unsuitable for PCR amplification. Instead, half of it was synthetized as a 23 bp modified oligonucleotide (AAGACTTATACACATCCAAGAAG) labeled with Cyanine3 (Cy3) at the 5'-end (Eurofins Genomics, Ebersbach, Germany); FISH probe Cy3-CL229.

For the second satellite repeat, CL217 (a 126 bp monomer), contigs of 378 bp from nine *Hieracium* samples as well as all repeats contained within each contig were aligned in BioEdit (Hall, 1999). Primers int_126F (CTAAATGTTGC ATCATGTTCG) and int_126R (TGTATGATCCA CGGAATGC) were designed toward conserved motifs and employed for PCR amplification under the following conditions: 25-µl reaction mixtures contained 2.5 µl of 10 × PCR Blue buffer T059, 2.5 mM $MgCl_2$, 0.2 mM of each primer, 200 µM of each dNTP, 20 ng DNA of IntA and 0.5 U *Taq* DNA polymerase (Top-Bio, Vestec, Czechia). The temperature profile included an initial denaturation step at 95°C for 5 min, followed by 35 cycles consisting of 95°C for 30 s, 50°C for 30 s, 72°C for 1 min and a final extension step at 72°C for 15 min. Amplification produced a ladder typical for tandem repeats. The longest bands were cut, purified, cloned, re-amplified and sequenced as described in Belyayev et al. (2018). Cloned sequences were aligned with the contigs, and one clone of 243 bp length (GenBank accession number MN784126) corresponding to almost two full repeats was used for FISH probe preparation (probe int126X2).

For the third satellite repeat, CL201 (a 172 bp monomer), aligned repeats of the longest contig (642 bp) of PreA were used for primer design; primers Hpre201f (ACTGGTCTCAAAT GCTTAGG) and Hpre201r (AAGCATTTGAGACCAGTAGG) were used for PCR amplification under the same conditions like above with the following alterations: 1.5 mM $MgCl_2$ and DNA of PreA were used, annealing temperature was 52°C, and extension at 72°C was done for 2 min. Subsequent procedures were the same as described above, and one clone (GenBank accession number MN784127) consisting of almost three full repeats (533 bp) was chosen for FISH probe preparation (Hpre201X1).

Chromosome preparation and *in situ* hybridization procedures were done as described in Belyayev et al. (2018). Hpre201X1 was found to co-localize with 5S rDNA in previously tested accessions of *H. prenanthoides* (data not shown), therefore we sequenced the non-transcribed spacer (5S-NTS) of two samples of each *H. intybaceum* and *H. prenanthoides* following Kaplan et al. (2013) to see if they may correspond with the tandem repeat. 5S-NTS sequences (GenBank accession no. MN784128-MN784131) were different from the repeat probe, therefore Hpre201X1 is localized near the 5S locus of those accessions, but not a part of it.

All accessions included in this study were tested with FISH probes of the following tandem repeats: 45S rDNA, 5S rDNA, CL82 (369 bp), and newly developed CL229 (46 bp), CL217 (126 bp) and CL201 (172 bp). The satellites CL18 (23+21 bp) and CL160 (89 bp) were excluded from the FISH experiments, because in preliminary trials they were found to be uninformative (data not shown). The results were compared with the results of RE and BLAST searches of the entire NGS datasets.

RESULTS

Bias Due to Library Preparation and Similarity Between Parental Accessions

The results of the analysis are presented in **Figure 1**. The average subsampling bias (intra-library) was the lowest (mean: 1.92%), followed by the library preparation bias (inter-library; mean: 2.56%) the comparison of different accessions of the same species (intra-specific; mean: 3.60%) and the bias due to parental accessions (inter-specific, mean: 11.78%). All those differences were highly significant ($p < 0.001$). As for the inter-library comparisons, IntC (mean: 3.12%) showed significantly higher ($p < 0.001$) bias than the libraries of IntA (mean: 2.15%) and PreA (mean: 2.41%), which, in turn, had the same magnitude as the intra-specific comparison of Pre (mean 3.47%). Overall, despite being significantly different, the magnitude of library preparation bias falls within the magnitude of subsampling bias and the magnitude of intraspecific variation. Therefore, we assume that library preparation does not affect conclusions for our interspecific comparisons.

Individual Clustering and Annotation of Genomes

In the individual clustering analyses, the pipeline used 3.7–4.8 million reads per sample, which corresponds to a genomic coverage of 0.17–0.20 x for diploids and 0.10–0.12 x for triploids (**Table 2**). On average, the majority of reads were grouped into ca. 140,000 clusters composed of 2 reads and more, corresponding to ca. 82% of the genomes (**Table 2**), while singlets represented the remaining 18% of the genomes. About 70% of all reads belonged to clusters containing at least 0.01% of the total input reads, representing medium and highly abundant repetitive elements (in total 209–260 clusters). The annotation at the level of superclusters assigned on average 70% of clusters (total of 55–62% of reads) to specific types of repetitive elements while the rest of clusters containing 9–15% of the input reads remained unclassified (**Figure 2A** and **Table 4**). The repeatomes of all genomes were dominated by LTR retrotransposons, constituting 53–60% of the whole genome (**Figure 2A**, **Table 4**, and **Supplementary Tables 1, 2**). The LTR Ty1/Gypsy elements occupied 33–39% of the genomes and were represented by four clades: Tekay and CRM (chromovirus lineages) and Athila and Ogre Tat (non-chromovirus lineages). The Tekay clade dominated, ranging from 32 to 36% of the genome. In all accessions, the majority of Tekay clusters were connected into one dominant supercluster (no. 1; consisting of 29–45 clusters, which on average contained 25% of all input reads (**Supplementary Figure 2**). Athila occurred at 1.35–1.80%, while CRM and Ogre Tat were

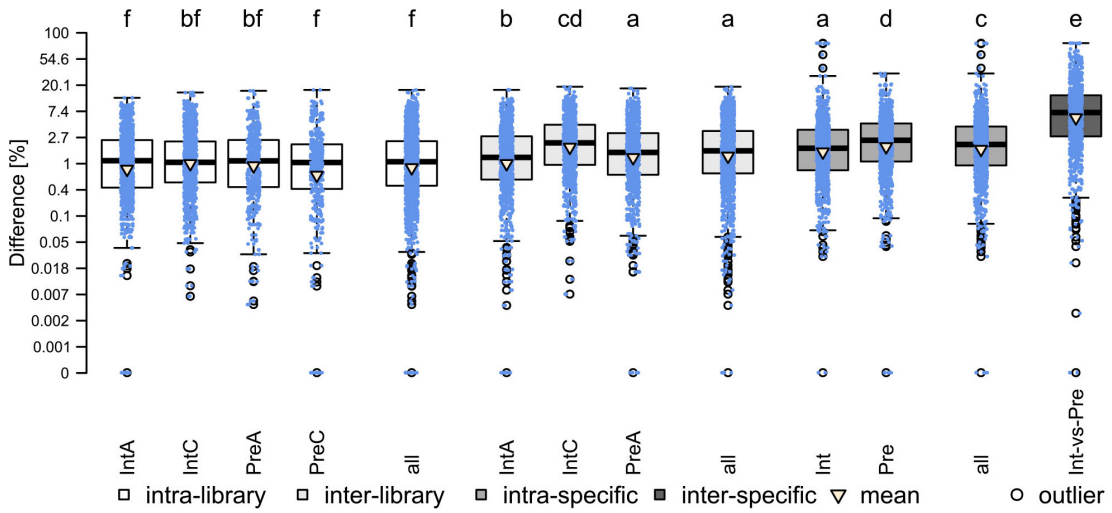

FIGURE 1 | Results of testing the bias due to library preparation and similarity between parental accessions (paired Wilcoxon signed-rank test). Different letters indicate significant differences between groups. Mean values (triangle) from left to right: 1.89, 1.96, 1.97, 1.83, 1.92, 2.15, 3.12, 2.41, 2.56, 3.72, 3.47, 3.60, 11.78.

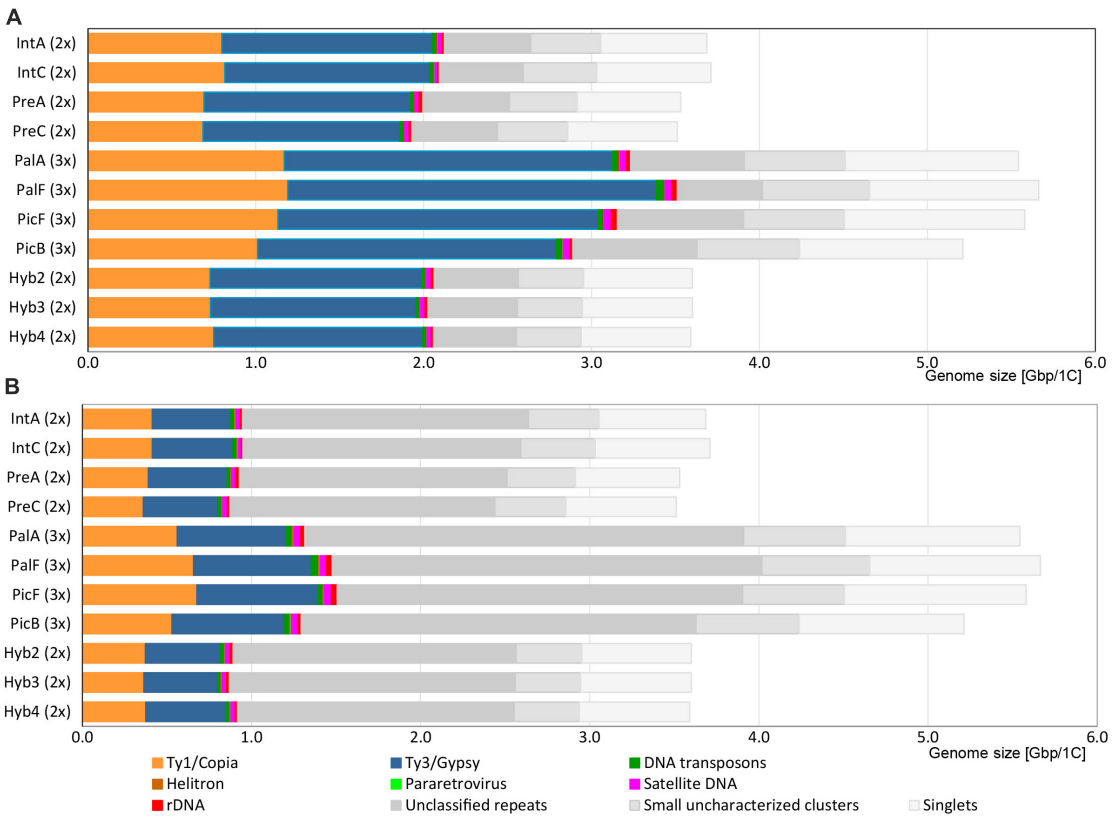

FIGURE 2 | Genomic composition of *Hieracium* species. **(A)** Results of annotation at the supercluster level. **(B)** Results of annotation at the cluster level. Repetitive elements like pararetrovirus and Helitrons are present in amounts too low to be visible on the graphs.

present only in trace amounts. The LTR Ty1/Copia elements belonged to seven families and occupied 19–22% of the genomes. Among them, the most abundant was Maximus/SIRE, and with 14–17% represented the second most abundant repeatome component. The Angela family of Ty1/Copia was moderately abundant, and its content was constant across all samples (average 4.3%). Other Ty1/Copia families–Ale, Bianca, Ikeros, TAR and Tork constituted all together only ca. 1% of the

TABLE 4 | Results of annotation at the supercluster and cluster level.

Sequence type	Repeat family	Genomic proportion (%)[a]										
		IntA	**IntC**	**PreA**	**PreC**	**PalA**	**PalF**	**PicF**	**PicB**	**Hyb2**	**Hyb3**	**Hyb4**
Ty1/Copia retrotransposons	Ale	0.149	0.149	0.177	0.176	0.147	0.151	0.192	0.152	0.159	0.160	0.154
	Angela	4.199	4.202	4.474	4.409	4.202	4.334	4.408	4.397	4.351	4.319	4.363
		(2.740)	*(2.829)*	*(3.055)*	*(2.864)*	*(3.196)*	*(3.319)*	*(3.426)*	*(3.050)*	*(2.799)*	*(2.951)*	*(2.845)*
	Bianca	0.163	0.163	0.155	0.164	0.165	0.169	0.160	0.172	0.164	0.166	0.169
	Ikeros	0.032	0.030	0.015	0.000	0.015	0.017	0.000	0.018	0.017	0.016	0.017
	SIRE	16.399	16.776	14.043	14.151	15.940	15.734	14.873	14.022	14.874	14.929	15.535
		(7.442)	*(7.328)*	*(6.926)*	*(6.466)*	*(5.981)*	*(7.335)*	*(7.719)*	*(6.134)*	*(6.571)*	*(6.114)*	*(6.654)*
	TAR	0.337	0.277	0.257	0.254	0.258	0.267	0.292	0.276	0.265	0.266	0.255
	Tork	0.367	0.359	0.470	0.346	0.359	0.351	0.354	0.351	0.360	0.431	0.356
	total Ty1/Copia	**21.646**	**21.956**	**19.591**	**19.500**	**21.086**	**21.023**	**20.279**	**19.388**	**20.190**	**20.287**	**20.849**
		(11.230)	*(11.135)*	*(11.055)*	*(10.270)*	*(10.120)*	*(11.609)*	*(12.143)*	*(10.153)*	*(10.335)*	*(10.104)*	*(10.450)*
Ty3/Gypsy retrotransposons	chromovirus/Tekay	31.687	30.830	33.006	31.413	33.114	36.522	32.350	32.104	32.849	31.884	32.571
		(10.227)	*(10.843)*	*(11.456)*	*(10.583)*	*(9.570)*	*(10.049)*	*(10.997)*	*(10.980)*	*(10.156)*	*(9.983)*	*(11.269)*
	chromovirus/CRM	0.014	0.029	0.032	0.064	0.023	0.027	0.000	0.063	0.026	0.095	0.039
	non-chromovirus/OTA/Athila	1.773	1.472	1.355	1.502	1.668	1.699	1.380	1.528	1.667	1.567	1.622
		(1.756)	*(1.457)*	*(1.335)*	*(1.483)*	*(1.652)*	*(1.699)*	*(1.363)*	*(1.510)*	*(1.667)*	*(1.548)*	*(1.602)*
	non-chromovirus/ OTA/Ogre_Tat/TatV	0.571	0.527	0.463	0.441	0.482	0.512	0.445	0.453	0.500	0.482	0.491
		(0.571)	*(0.527)*	*(0.375)*	*(0.441)*	*(0.398)*	*(0.512)*	*(0.445)*	*(0.239)*	*(0.500)*	*(0.482)*	*(0.409)*
	total Ty3/Gypsy	**34.045**	**32.858**	**34.856**	**33.420**	**35.287**	**38.760**	**34.175**	**34.148**	**35.042**	**34.028**	**34.723**
		(12.568)	*(12.856)*	*(13.198)*	*(12.571)*	*(11.643)*	*(12.287)*	*(12.805)*	*(12.792)*	*(12.349)*	*(12.108)*	*(13.319)*
pararetrovirus		0.069	0.055	0.000	0.000	0.050	0.031	0.017	0.042	0.021	0.029	0.031
DNA transposons	EnSpm_CACTA	0.175	0.178	0.145	0.177	0.163	0.265	0.284	0.168	0.151	0.189	0.160
	hAT	0.068	0.056	0.068	0.063	0.070	0.060	0.060	0.075	0.058	0.034	0.060
	MuDR_Mutator	0.152	0.159	0.149	0.184	0.159	0.195	0.090	0.073	0.160	0.092	0.089
		(0.132)	*(0.159)*	*(0.148)*	*(0.167)*	*(0.159)*	*(0.195)*	*(0.090)*	*(0.073)*	*(0.160)*	*(0.092)*	*(0.089)*
	PIF_Harbinger	0.262	0.284	0.286	0.284	0.241	0.253	0.123	0.283	0.279	0.256	0.249
	total DNA transposons	**0.657**	**0.677**	**0.648**	**0.708**	**0.633**	**0.773**	**0.557**	**0.599**	**0.648**	**0.571**	**0.558**
		(0.636)	*(0.677)*	*(0.647)*	*(0.691)*	*(0.633)*	*(0.773)*	*(0.557)*	*(0.599)*	*(0.648)*	*(0.571)*	*(0.558)*
Helitron		0.160	0.155	0.105	0.133	0.143	0.125	0.101	0.146	0.126	0.124	0.128
satellite DNA	satellite CL18 [b] (23+21 bp)	0.405	0.349	0.369	0.424	0.471	0.414	0.477	0.485	0.417	0.452	0.405
	satellite CL160 (89 bp)	0.129	0.125	0.118	0.089	0.113	0.119	0.088	0.114	0.108	0.121	0.124
	satellite CL82 (369 bp)	0.086	0.000	0.163	0.200	0.077	0.103	0.120	0.106	0.192	0.112	0.108
	satellite CL229 (46bp)	0.042	0.000	0.000	0.000	0.023	0.014	0.021	0.021	0.024	0.025	0.025
	satellite CL201 (172bp)	0.000	0.000	0.026	0.028	0.000	0.000	0.031	0.013	0.025	0.014	0.027
	satellite CL217 (126bp)	0.023	0.029	0.000	0.011	0.016	0.021	0.013	0.018	0.014	0.017	0.023
	total satellite DNA	**0.685**	**0.503**	**0.676**	**0.752**	**0.700**	**0.671**	**0.749**	**0.758**	**0.781**	**0.741**	**0.712**
45S rDNA		0.331	0.225	0.576	0.380	0.394	0.561	0.619	0.297	0.458	0.442	0.331
5S rDNA		0.027	0.000	0.052	0.024	0.029	0.028	0.021	0.024	0.031	0.030	0.023
unclassified repeats		13.918	13.445	14.613	14.562	12.211	8.952	13.447	14.202	13.918	14.836	13.755
		(45.831)	*(44.267)*	*(44.806)*	*(44.658)*	*(46.821)*	*(44.839)*	*(42.954)*	*(44.793)*	*(46.465)*	*(46.939)*	*(45.557)*
small uncharacterized clusters		11.336	11.817	11.436	11.938	10.856	11.264	10.779	11.707	10.797	10.733	10.717
singlets		17.128	18.311	17.450	18.582	18.612	17.811	19.256	18.690	17.988	18.179	18.172

[a] *Values in regular font represent results of annotation at the supercluster level; results at the cluster level are indicated in italics and in parentheses. Annotations at the cluster level are only reported when they differ from the ones at the supercluster level.*
[b] *complex structure with nested repeats including two minisatellites (see Belyayev et al., 2018).*

Hieracium genomes. DNA transposons on average covered only 0.65% of the genomes with four types: EnSpm_CACTA, hAT, MuDR_Mutator and PiF_Harbinger. Other dispersed repetitive elements included Helitrons and pararetroviruses. The latter elements were not detected at the 0.01% threshold in *H. prenanthoides*.

The results of the annotation at the supercluster level were then compared with the annotation at the level of clusters, which left many clusters unannotated (**Table 4**, values in parentheses). The cluster level approach annotated less than half of the clusters (total of 25–27% of input reads), while the remaining unannotated ones containing ca. 43–47% of reads had few or completely lacked protein domain BLAST hits (**Figure 2B** and **Supplementary Table 2**). The main drivers for the difference in annotation between supercluster and cluster level were Tekay, Maximus/SIRE and Angela. These were highly abundant TEs that formed several large superclusters composed of both annotated and unannotated clusters (**Table 4** and **Supplementary Tables 1, 2**). Ty3/Gypsy Athila, Ogre/Tat and MuDR_Mutator type of DNA transposons also formed several smaller superclusters. Other TEs did not form superclusters, and therefore, their genomic proportion remained the same in both approaches.

Comparative Clustering Analysis of Parental Species, Natural, and Synthetic Hybrids

The comparative analysis of the parental species *H. intybaceum* (IntX) and *H. prenanthoides* (PreX) revealed a high similarity in abundance across all major repeat types (**Figure 3A**). The largest clusters that contributed the highest to the genome size variation between species contained mostly chromovirus Tekay (slightly more abundant in *H. prenanthoides*) and Maximus/SIRE (more abundant in *H. intybaceum*). However, after removing the effect of cluster size (**Figure 3C**), the highest variability occurred among small clusters that contribute the least to the total genome size variation (**Figure 3C**). The most variable elements were unclassified repeats. Overall, the number of annotated clusters decreased with cluster size (**Figure 3B**).

In the comparative analyses of parental species and their natural and synthetic hybrids, we compared firstly their cluster abundances in read numbers (**Supplementary Figure 3**). In all hybrids, the majority of clusters showed an intermediate position between the clusters of parental species. However, across all comparisons, each hybrid was significantly closer to one of the parents (**Figure 4**). Cluster abundances of PalA and PalF were significantly more similar to Int, while PicB and PicF were significantly more similar to Pre. This observation reflects well the parental genome dosages of 2:1 and 1:2, respectively. The position of the clusters of synthetic hybrids was more intermediate when compared to natural hybrids. No cluster was specific for any of the hybrids or parental species except for the occasional presence/absence of a certain satellite repeat.

Theoretically, in a newly formed hybrid, we would expect a direct inheritance of the repeat content from the parents, in proportions that correspond to the genome dosage of the hybrid. Significant deviation from this initial content in synthetic and natural hybrids would indicate a burst (or loss) of TEs. We compared the observed cluster abundances of F1 hybrids and natural allopolyploids with the expected values that were calculated based on the parental abundances. As a measure of departure from expectations, we used deviation scores, by which the cluster size effect is removed (**Figure 5** and **Supplementary Figure 4**). A positive deviation score implies a larger cluster size than expected, and vice versa for a negative deviation score (**Figure 5**).

In general, the majority of clusters of all hybrids follow the theoretical expectation. The deviation tends to be the smallest in big clusters and increases as clusters get smaller. In PalF, PicF and Hyb3, the majority of clusters were slightly more abundant than expected (**Supplementary Figure 4**, ascending curve). In PicB, clusters were more dispersed, and the largest clusters (the first hundred clusters) were slightly smaller than expected (**Supplementary Figure 4**, descending curve). The clusters of PalA, Hyb2 and Hyb4 were the closest to the expected values. The statistical assessment of overall deviation scores' means and the magnitude of variation of deviation scores in two groups of accessions (natural vs. synthetic) showed no statistically significant difference between natural apomictic and synthetic hybrids (ANOVA: $F = 0.561$, $p > 0.05$; variation of deviation scores - Levene test: $F = 0.8221$, $p > 0.05$).

Nevertheless, 45S rDNA, 5S rDNA and several satellite repeats deviated substantially from the general trend (**Figure 5**). Consistent deviation was found for 45S rDNA and 5S rDNA across all samples except Hyb4. However, they displayed sample specific trends with either under- or over-representation. The over-representation of 45S rDNA is the strongest in PalF and PicF, which contained the highest amount of 45S rDNA in all hybrids in the individual clustering analysis (**Table 4**). Consistent deviation was also found for satellite CL201 (172bp), which deviated from expectation in Pic and in synthetic hybrids (in Pal this satellite was not detected). Other satellite repeats did not show any strong pattern of deviation across groups of samples (natural hybrids vs. synthetic hybrids; Pal vs. Pic; Hyb vs. Pal or Pic). An occasional deviation of a few unclassified repeats was detected, which also did not show any consistent pattern.

In contrast, all apomicts displayed a higher deviation (an increase) of pararetrovirus clusters than synthetic hybrids. This represents the only apomict-specific finding among clusters of repetitive elements. In all comparative analyses, two pararetrovirus clusters were detected. In PalA and PicF accessions, both clusters were considerably larger than expected (deviation scores in PalA: 0.263 and 0.278; PicF: 0.230 and 0.486), whereas in PalF and PicB, only one of the two clusters showed a prominent deviation (PalF: 0.289; PicB: 0.328). Statistical analyses (**Supplementary Figure 5**) showed the differences between natural and synthetic hybrids to be highly significant (mean deviation scores: synthetic hybrids −0.04, natural hybrids 0.26; ANOVA: $F = 27.2$; $p < 0.001$).

Abundance and Cytogenetic Analyses of rDNA and Satellite Repeats

45S rDNA content varied substantially between accessions (0.225–0.619%). The average content of 5S rDNA was only 0.029%, and it was not detected at all in IntC (but see below). According to the individual RE clustering, satellite DNA represented only a tiny fraction of *Hieracium* genomes (from 0.503% in IntC to 0.781% in Hyb2) (**Table 4**). In addition to

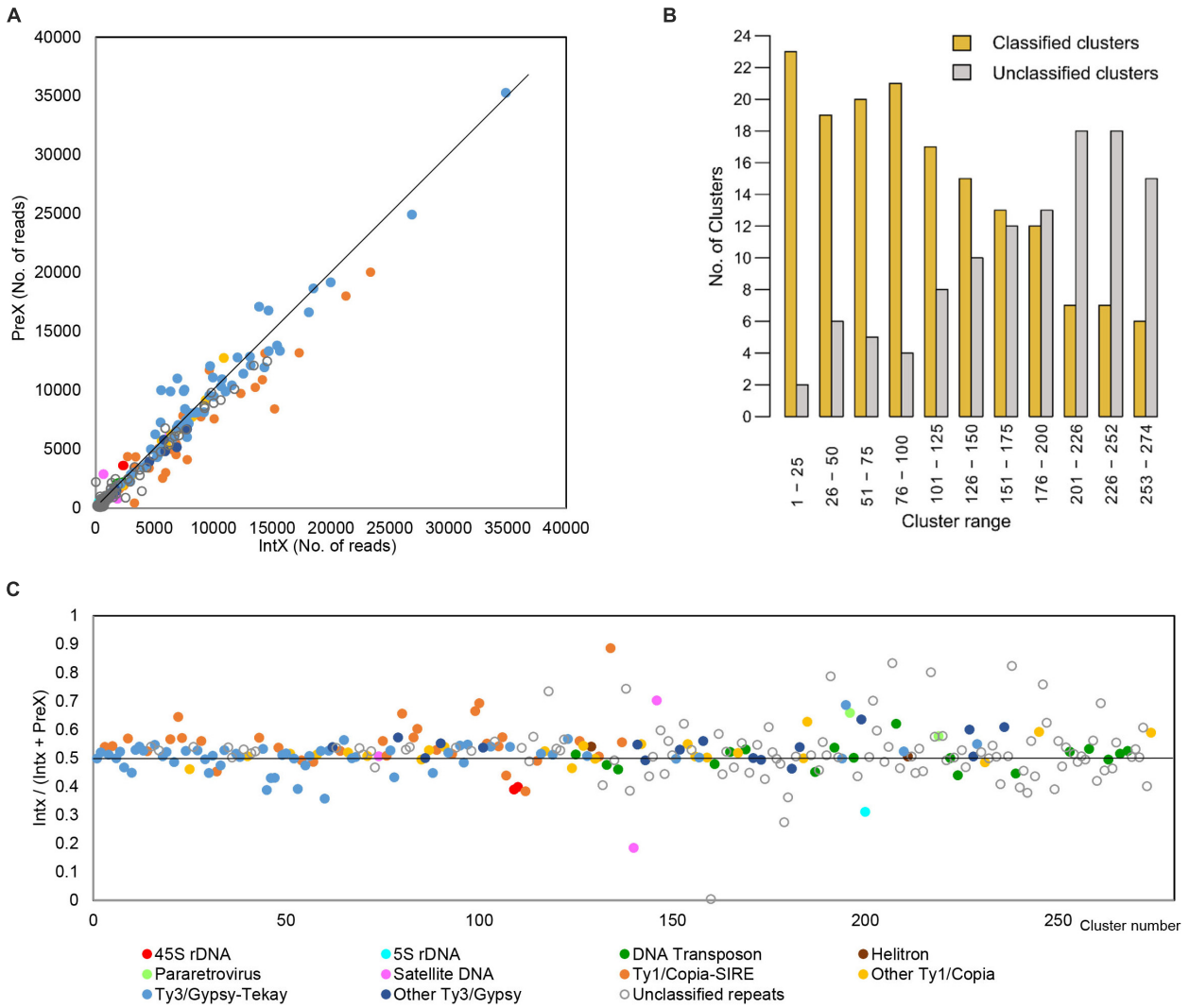

FIGURE 3 | Results of comparative analysis of the two parental species. **(A)** Comparison of the number of reads from *H. intybaceum* (IntX) and *H. prenanthoides* (PreX) across all clusters. Dots represent clusters of repetitive elements. Clusters placed on the line represent repeats with the same genomic proportion in both species. **(B)** Proportion of annotated (classified) and unannotated (unclassified) clusters. Clusters are arranged by decreasing size. **(C)** Comparison of cluster abundances in *H. intybaceum* (IntX) and *H. prenanthoides* (PreX) irrespective of cluster size. Clusters on the black line (value of 0.5) have the same proportion in both species; clusters above are more abundant in *H. intybaceum*; clusters below are more abundant in *H. prenanthoides*.

the three described satellite repeats in *Hieracium* [CL18, CL82 and CL160 (Belyayev et al., 2018)], we detected three novel repeats, CL229, CL201 and CL217 (**Supplementary Figure 1**). CL18 and CL160 were detected by RE in all accessions, while CL82 was detected in all accessions except in IntC (**Table 5**). Satellite CL229 (46 bp) was only detected in IntA and in all natural and synthetic F1 hybrids, but not in IntC and both Pre accessions. CL201 (172 bp) was detected in both Pre accessions, but not in any of Int; it was also found in Pic and synthetic hybrids, but not in Pal accessions. CL217 (126 bp) was detected in all accessions except PreA.

The results of cytogenetic experiments are shown in **Figure 6** and **Table 5**. In short, FISH signals corresponding to 45S and 5S rDNA were detected in all accessions. CL82 was detected in all accessions except in IntC. The tandem structure of the three new satellites was confirmed cytogenetically; the probe CL229 produced signals in IntA and all natural and synthetic hybrids, while no signal was detected in IntC and both Pre accessions. CL201 was detected in Pre and all natural and synthetic hybrids except in PalF, and was not found in Int accessions. Lastly, CL217 was observed in both Int accessions and in all natural and synthetic hybrids, but not in any Pre accessions.

Comparison of RE, BLAST, and Cytogenetic Results

rDNA and satellite repeats showed the highest deviation from the expected values in natural and synthetic hybrids in RE. In order

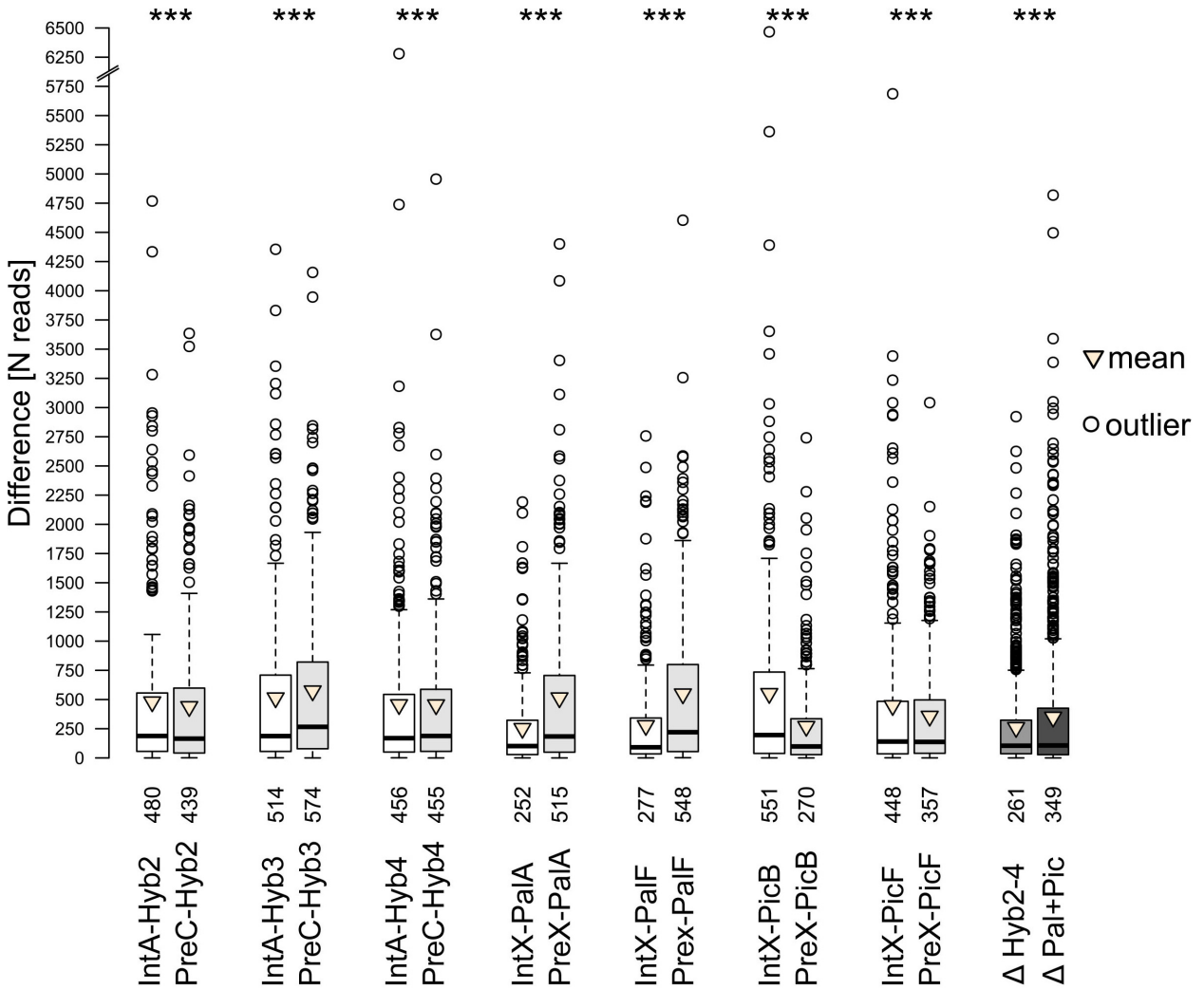

FIGURE 4 | Difference in number of reads between parental and progeny accessions obtained from comparative analyses. The paired Wilcoxon signed-rank test was, separately for each parent-progeny combination, employed to access whether the progeny is closer to either parent. Results are shown above the boxes. Numbers below the boxes indicate mean differences in read numbers for the respective comparison. Δ – Degree of 'intermediacy' of progeny (ΔParent1-Progeny – ΔParent2-Progeny). Significance level *** – $p < 0.001$.

to verify the genuine absence of satellites and rDNA, that were not identified with RE, we searched for them with BLASTn on the whole filtered NGS datasets (**Supplementary Table 3**) and compared the outcome with the results of the FISH experiments (see **Table 5**).

FISH for 45S rDNA showed four chromosomal loci in *H. intybaceum* (two per haploid genome) and six loci in *H. prenanthoides* (**Figures 6A–C**). However, the genomic proportions of 45S rDNA obtained by RE were different among individuals within each parental species, despite having the same number of loci (IntA: 0.331%, IntC: 0.225%; PreA: 0.576%; PreC: 0.380%). While all tested accessions of *H. pallidiflorum* and *H. picroides* bear the same number of 45S rDNA loci (7), the accessions differed in their genomic proportion detected by RE. Similarly, the three synthetic hybrids possessed the same number of loci (5), but they also differed in their proportions obtained by RE. The results of BLASTn searches for 45S rDNA were in a good agreement with the RE results.

As of 5S rDNA, all tested accessions had one FISH locus per haploid genome (i.e., all diploids had 2, whereas triploids had 3 loci) (**Figure 6** and **Table 5**). Contrary to the FISH results, the proportion of 5S rDNA in PreA was twofold higher than in PreC (0.052% vs. 0.024%), and similarly, IntA had 0.027% (and a slightly higher result in BLASTn search of the entire NGS dataset) while IntC had only 0.13% of 5S rDNA, which might explain why it was not detected by RE (**Figure 6A**). In contrast to the parental species, both groups of natural and synthetic hybrids had relatively similar genomic proportions of 5S rDNA as detected by RE. The RE and BLASTn results showed slightly less congruence than other tested repeats.

Only one CL82 locus was detected by FISH in IntA, whereas IntC does not possess this tandem repeat according to FISH

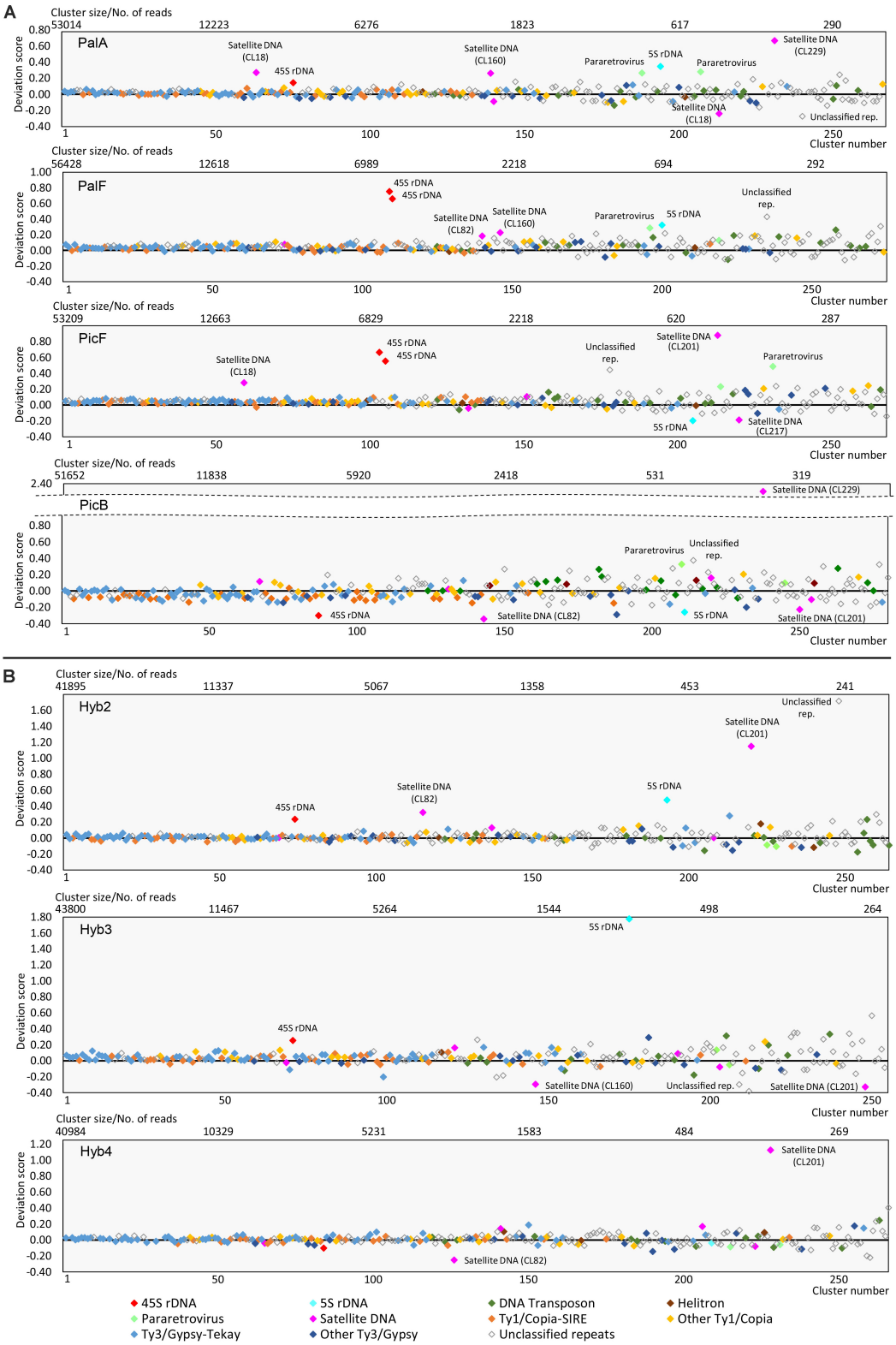

FIGURE 5 | Deviation of clusters from their expected size in: **(A)** natural triploid apomicts *H. pallidiflorum* (PalA, PalF) and *H. picroides* (PicF, PicB) and **(B)** synthetic diploid hybrids (Hyb2, Hyb3, Hyb4). Clusters are arranged by decreasing size. The symbols on the black line (value zero) indicate clusters of expected cluster size. Clusters above or below this line are larger or smaller compared to the expected values calculated from the parental species.

TABLE 5 | rDNA and satellite repeats as detected by RepeatExplorer, BLASTn and FISH.

	45 rDNA			5S rDNA			CL82 (369 bp)			CL229 (46bp)			CL217 (126 bp)			CL201 (172bp)[e]		
ID	**RE (+/−)**	**BLAST (+/−)**	**No. of FISH loci**	**RE (+/−)**	**BLAST (+/−)**	**No. of FISH loci**	**RE (+/−)**	**BLAST (+/−)**	**No. of FISH loci**	**RE (+/−)**	**BLAST (+/−)**	**No. of FISH loci**	**RE (+/−)**	**BLAST (+/−)**	**No. of FISH loci**	**RE (+/−)**	**BLAST (+/−)**	**No. of FISH loci**
IntA	+	+	4	+	+	2	+	+	1	+	+	2	+	+	2	−	−	0
IntC	+	+	4	−	+	2	−	(+)[b]	0	−	(+)[c]	0	+	+	2	−	−	0
PreA	+	+	6	+	+	2	+	+	7	−	(+)[c]	0	−	+	0	+	+	2
PreC	+	+	6	+	+	2	+	+	7	−	(+)[c]	0	+	+	0	+	+	2
PalA	+	+	7	+	+	3	+	+	2	+	+	1	+	+	2	−	(+)[c]	1
PalF	+	+	7	+	+	3	+	+	6	+	+	1	+	+	3+6[d]	−	(+)[c]	0
PicF	+	+	7[a]	+	+	3	+	+	9	+	+	1	+	+	4+8[d]	+	+	2
PicB	+	+	7[a]	+	+	3	+	+	6	+	+	1	+	+	2	+	+	2
Hyb2	+	+	5	+	+	2	+	+	4	+	+	1	+	+	1	+	+	2
Hyb3	+	+	5	+	+	2	+	+	4	+	+	1	+	+	1	+	+	1
Hyb4	+	+	5	+	+	2	+	+	6	+	+	1	+	+	1	+	+	2

ID = Sample ID; RE/BLAST (+/−) = RepeatExplorer/BLAST (cluster detected/cluster not detected).
[a] The numbers of loci for 45S rDNA in PicF and PicB were reported in Chrtek et al. (2020).
*[b] CL82 was detected in a very low amount in IntC (see **Supplementary Table 3**) and a closer inspection showed that the corresponding reads displayed a rather low similarity to the sequence of satellite CL82 (see section "Discussion").*
*[c] satellite repeat detected only in trace amounts (see **Supplementary Table 3**).*
*[d] besides major loci found on all metaphase plates (PalF 3, PicF 4), additional loci (PalF 6, PicF 8) of CL217 occurred in these samples (see section "Results" and **Figure 6**).*
[e] CL201 was present as two hemizygous loci on non-homologous chromosomes in Pre.

(**Figures 6A,B**) and RE; but the BLASTn search found 0.007% of reads corresponding to this repeat in IntC (below the RE threshold). The accessions of *H. prenanthoides* had 7 loci each (**Figure 6C**), however, their total copy numbers detected by RE were only about twice as high as in IntA. In natural apomicts, the number of CL82 loci was variable (**Table 5** and **Supplementary Figure 6**), and the genomic proportions obtained by RE approximately followed the pattern detected by FISH (the smallest amount in PalA, the highest in PicF, approximately the same amount in PalF and PicB). The picture was not entirely consistent in synthetic hybrids; the genomic abundances detected by RE did not follow the number of FISH loci. RE and BLASTn results were generally in agreement.

The novel satellite repeat CL229 (46 bp monomer) was only found in one accession of *H. intybaceum* (IntA) and in all hybrids. In IntA, it was present in 2 loci on a pair of homologous chromosomes (**Figure 6B**). While the satellite was not detected in IntC nor in both *H. prenanthoides* by RE and FISH, BLASTn searches of their NGS datasets detected trace amounts of this repeat in these three accessions. All natural and synthetic hybrids had one CL229 locus each, with similar genomic proportions detected by both RE and BLASTn.

The novel satellite repeat CL217 (126 bp monomer) was detected by FISH in both accessions of *H. intybaceum* (2 loci on homologous chromosomes, **Figure 6D**), but not in *H. prenanthoides* (**Figure 6E**). However, RE and BLASTn of *H. prenanthoides* detected it in very small amounts: the NGS dataset of PreA contained 0.008% of the reads (BLASTn hits to this satellite repeat, no cluster detected by RE), whereas PreC contained 0.01% and a cluster detected by RE. Each of the synthetic hybrids contained one FISH signal. *Hieracium pallidiflorum* accessions had 2 or 3 major loci (**Figures 6F,G**), whereas *H. picroides* had 2 or 4 major loci. In one accession of each (PalF and PicF), a proliferation of this satellite was detected, resulting in 6 (**Figure 6G**) and 8 additional loci (**Table 5**), respectively. RE and BLASTn results were congruent.

Regarding the third novel satellite CL201 (172 bp monomer), two FISH loci were detected in each *H. prenanthoides* accession, but they did not constitute a homologous pair. Instead, they were hemizygous as indicated by their presence on chromosomes bearing or not bearing also 5S rDNA loci (**Figures 6H,I**). Furthermore, CL201 co-localized with 45S rDNA. Both accessions had similar genomic proportions according to RE and BLASTn. CL201 was not detected in *H. intybaceum* accessions by any of the methods (**Figures 6J,K** and **Table 5**). RE analyses did not detect CL201 in *H. pallidiflorum* accessions, nevertheless, BLASTn searches revealed trace amounts. Also, despite a low proportion of 0.005% in PalA, it produced a FISH signal; in PalF, the FISH locus was not detected. *Hieracium picroides* accessions possessed two loci each. The synthetic hybrids had one or two loci with correspondingly proportional genomic abundances.

DISCUSSION

Characterization of *Hieracium* Repeatomes

This study represents the first characterization of repeatomes in *Hieracium* s.str. Transposable elements represent the dominant component of the investigated genomes, which is a general pattern in plants (Kejnovsky et al., 2012; Dubin et al., 2018). With 70% of TEs representing a genomic proportion of at least 0.01%, the *Hieracium* genomes can be considered as highly repetitive. The result is in accordance with a comparative study of the family Asteraceae (Staton and Burke, 2015) where the average

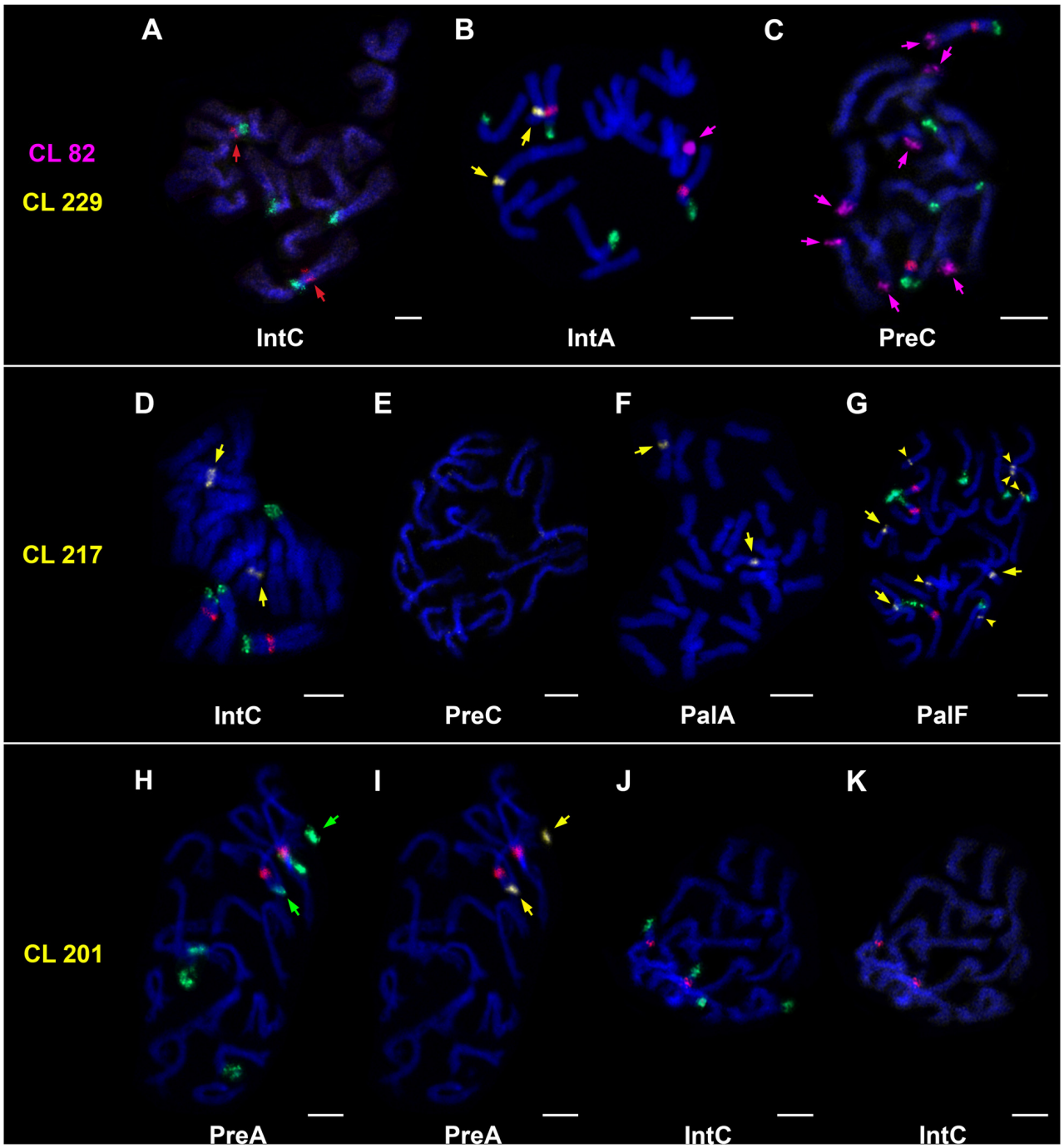

FIGURE 6 | Chromosomal localization of clusters/satellites by FISH. 5S rDNA (red signal), 45S rDNA (green signal). **(A–C)** Presence/absence of 5S rDNA (**A**, red arrows), CL82 (**B,C**, magenta arrows) and CL229 (**B**, yellow arrows). **(D–G)** Presence/absence of CL217 (**D,F,G** – major loci marked by yellow arrows; **G**–additional loci marked by yellow arrowheads). **(H–K)** Presence/absence of CL201. The same metaphase plates are used in H/I to show co-localization of 45S rDNA (**H**, green arrows) and CL201 (**I**, yellow arrows), and J/K (absence of CL201). Chromosomes were counterstained with DAPI (blue). Bars = 5 μm.

proportion of TEs for the family was estimated to 69.9 ± 5.3% and the average proportions of LTRs and DNA transposons to 53 ± 19.1% and 0.60 ± 0.7%, respectively. *Hieracium* with an average of 55% for LTR retrotransposons and 0.65% for DNA transposons matches those results.

Staton and Burke (2015) revealed a linear increase in the abundance of Ty3/Gypsy elements from the most ancestral Asteraceae to the most derived subfamily, Asteroideae. Among the latter, species of the genus *Helianthus* and *Phoebanthus tenuifolius* exhibit the highest content of Ty3/Gypsy (62.4 ± 2.7%

and 67.5 ± 5.6%, respectively). In contrast to Ty3/Gypsy, Ty1/Copia showed the opposite pattern: basal species in the family have proportionally more Ty1/Copia compared to species of Asteroideae. *Hieracium* species of the subfamily Cichorioideae fit this picture; Ty3/Gypsy (33%–39%) dominate over Ty1/Copia (19–22%), in a ratio of ca. 1.5–1.8 : 1. The closest representative of the subfamily Cichorioideae, the diploid *Taraxacum kok-saghyz*, exhibits a similar pattern for these two groups. However, other repeatome studies performed on *Helianthus* species (subfamily Asteroideae) reported a much higher variability in the ratio of Ty3/Gypsy and Ty1/Copia (1.45–5.91 : 1, Mascagni et al., 2017). Also, it seems that the dominance of Ty3/Gypsy over Ty1/Copia cannot be generalized to the whole subfamily Asteroideae; recent studies of *Melampodium* (McCann et al., 2018), *Anacyclus* and *Heliocauta* (Vitales et al., 2019) showed that some of their species are dominated by Ty1/Copia over Ty3/Gypsy. Those findings confirm the widely recognized trend of remarkable variation in repeat composition among related taxa, and with future studies, we expect to see even higher diversity in the repeat composition of Asteraceae.

In *Hieracium*, only four families of LTRs (Ty3/Gypsy chromovirus/Tekay and non-chromovirus/Athila, Ty1/Copia Angela and SIRE) were represented in amounts higher than 1%. Among those, Ty3/Gypsy chromovirus/Tekay and Ty1/Copia SIRE were by far the most abundant, present in a ratio of about 2:1. Again, these two lineages seem to prevail in the Asteraceae, although the ratios between them may differ (e.g., Mascagni et al., 2017; Mlinarec et al., 2019).

Among all LTRs detected, only these four most abundant families plus Ty3/Gypsy non-chromovirus/Ogre Tat (genomic proportion ca 0.5%) formed superclusters. The annotation at the supercluster level is clearly important, because it captured the proliferation and complexity of highly abundant TE families that together constitute about 54% of *Hieracium* genomes. The comparison of annotations at the supercluster and the cluster level showed that roughly one third of all clusters that lacked a BLAST match in the TE protein domain database were additionally annotated at the supercluster level due to pair-end reads shared with annotated clusters. Such a result is expected for large and repetitive genomes like *Hieracium*. Superclusters represent highly proliferated LTR families that comprise both active and inactive copies at various stages of degradation due to accumulated mutations: such a high sequence divergence causes a lack of BLAST similarity between copies of the same element and thus creates separate clusters (Novák et al., 2010). Since the NGS reads from different regions within the genome (containing both active and inactive copies) cluster together, it remains unclear to which extent such clusters and superclusters represent active elements. The splitting into clusters and superclusters is also a consequence of the exceptionally large size of all five LTRs (Neumann et al., 2019) and low coverage sequencing; low read depth and gaps in the coverage of particular sequences cause a lack of overlapping NGS reads that may otherwise fuse two or more clusters into one (Novák et al., 2010). Therefore, it is possible that an increase of the sequencing depth would facilitate the connection of clusters and foster additional annotations. Also, it should be emphasized that some of the clusters contained in superclusters likely comprise sequences that do not necessarily represent TEs (e.g., TE's flanking regions). Their proportion is difficult to estimate. Taking this and other factors mentioned above into account, it means that an exact level of TE annotation cannot be achieved with low coverage sequencing. Therefore, genomic proportions of TE categories obtained by individual clustering that have similar values (**Figure 2** and **Table 4**) in closely related/highly similar accessions cannot be strictly compared. Nevertheless, the intention of the method is to give a general estimation of repeatome components, and the supercluster approach ensures a good overview of the extent of proliferation of highly abundant TEs.

TE Dynamics Following Hybridization and Polyploidization

Comparative clustering analysis evaluates efficiently the repeatome variability in parental taxa and their derived allopolyploids (Renny-Byfield et al., 2013; Dodsworth et al., 2017; McCann et al., 2018). In *Hieracium*, repeat abundances for synthetic and natural hybrids were remarkably similar to the expected values calculated from parental genomes. Overall, clusters of synthetic hybrids expectedly showed a high degree of adherence to the expected values. Rather unexpectedly, natural hybrids followed the same pattern, only with slight over- or under-representation of cluster abundances across all repeat categories that corresponded to individual differences in their genome sizes. Among TE clusters, there were no large bursts or eliminations of any particular TE that could be interpreted as a response to hybridization and/or polyploidization events. However, apomicts displayed evident overabundance of pararetrovirus clusters not observed in synthetic hybrids, and this is the only detectable difference between natural apomictic and synthetic F1 hybrids (**Figure 5**). The differences between these two groups were statistically highly significant (**Supplementary Figure 5**). This finding may point to the accumulation of pararetrovirus sequences following polyploidization. Pararetroviruses are a type of retroelements that have double-stranded DNA and use reverse transcription for their replication. They seem to be ubiquitous in plants, but it is not clear if they represent a neutral component of plant genomes (e.g., due to a potentially lost function) or if they represent pathogens or even contribute to virus resistance of the host (Staginnus et al., 2007). The dynamic behavior of pararetroviruses was also observed in a study of allotetraploid *Nicotiana tabacum*; one of the pararetrovirus families was more abundant in one of its progenitors, *N. tomentosiformis*, than in *N. tabacum*. Such a pattern is a result of either preferential elimination from the polyploid genome or specific accumulation in the diploid progenitor genome following polyploidization (Matzke et al., 2004). However, the significant overabundance of pararetrovirus clusters we detected in apomicts does not necessarily imply a causal relationship. Also, the actual parents of the natural hybrids are unknown and may have shown higher abundances of pararetrovirus sequences than the diploids used in this study (although to assume that just these two repetitive

elements out of hundreds were affected in the observed way may not be very likely). Besides pararetroviruses, an occasional deviation of a few different unclassified repeats was observed, without any obvious pattern among accessions. However, both pararetrovirus and unclassified clusters are located among small clusters with increasing deviation from the expected values that creates the appearance of a 'widening tail' in the graph, which is a consequence of low-coverage sequencing. While it is possible that the higher variability of small clusters is a sign of ongoing evolutionary processes, the pronounced bias of low-coverage sequencing undermines definitive interpretations of these results.

Generally, our results corroborate the findings in two studies of Asteraceae genera using the graph-based clustering method, which also did not reveal differential amplification of TEs: allopolyploid species of *Melampodium* that most likely originated in the Pleistocene (McCann et al., 2018), and *Anacyclus*, which comprises diploid species of hybridogenous origin (Vitales et al., 2019). These findings are consistent with recent observations according to which amplification of TEs after hybridization and polyploidization seem to be rare (Parisod and Senerchia, 2012). It is possible that the hybridization event in our synthetic hybrids triggered TE activity, which resulted in transcriptional, epigenetic or structural changes that did not result in a net increase of TE copy number, and these aspects would be worth further investigation. In the case of *H. pallidiflorum* and *H. picroides*, it is not possible to predict possible long-term results of TE activity triggered by historical polyploidization events. Most probably, the observed repetitive profiles also depend on the age of the allopolyploids. It has been suggested that neopolyploidization events may be accompanied by transiently increased activity of TEs (release of TE silencing and possibly transposition) in the first generations of polyploids, after which TE silencing is restored (Vicient and Casacuberta, 2017), however, the response may vary between genomes and TE families. Also, polyploidization events that formed *H. pallidiflorum* and *H. picroides* might be immediately followed by a switch to apomixis so that no further generations could facilitate such restoration by meiosis. *Hieracium pallidiflorum* and *H. picroides* are assumed to be at least hundreds and up to several thousands of years old, and there is no evidence of a recent formation of polyploids by hybridization between diploid sexual taxa (Mráz and Zdvořák, 2019). So far, attempts of experimental crossings failed to produce triploid hybrids from the parental combination.

Concerning the mode of reproduction, there was no statistically significant difference between the groups of natural apomictic and synthetic *Hieracium* hybrids in the overall deviation pattern of their clusters from the expected values. This result is in agreement with a study of Ågren et al. (2015) on genus *Oenothera*, which comprises sexual species as well as species utilizing functionally asexual reproduction (permanent translocation heterozygosity); there was no evidence that the mode of reproduction was responsible for an almost twofold variation in genome size between the species. However, it should be emphasized that apomixis is considered a young evolutionary trait (Van Dijk and Vijverberg, 2005) and therefore it is possible that the consequences of asexuality in *Hieracium* are not yet detectable. This is in line with the study of Docking et al. (2006) where the authors did not find evidence of relaxed selection in three investigated types of TEs in four asexual taxa, among which were three presumably young Asteraceae: *Antennaria parlinii*, *Taraxacum officinale* and closely related *Hieracium aurantiacum* (recently reclassified as *Pilosella aurantiaca*; Bräutigam and Greuter, 2007).

The interesting diversity of genome sizes in natural apomicts (slight net increase in both *H. pallidiflorum* accessions; slight net increase or decrease in *H. picroides*), might be indeed a consequence of their different age and the independent evolution of TEs in each apomictic lineage after the polyploidization events, since all tested accessions were demonstrated to be of independent origin. However, other potential factors involved in the origin of these apomictic species that could produce the observed genome size differences cannot be ruled out. They include involvement of unknown (even extinct) diploid parental genotypes with higher genome size diversity, polyploid parental genotypes introgressed by other *Hieracium* species, and involvement of polyploid hybrids since they produce partially fertile pollen (Chrtek et al., 2020). In order to obtain more conclusive answers about the trends in genome size, measurements will have to be performed on a larger number of accessions of each apomictic and parental species. Overall, our results indicate a relatively low turnover of repetitive DNA during the formation of apomictic lineages and suggest that bursts of TEs do not play an important role in the evolution of our system.

Dynamics of rDNA and Satellite Repeats in *Hieracium*

In contrast to TEs, rDNAs and satellite repeats showed substantial deviations from the expected values in both natural and synthetic hybrids, although because of their low abundance they do not significantly contribute to genome size differences. These deviations did not show any correlation with the accession origin, and are therefore explained by the hybridization event as such, independent of the ploidy and reproductive mode of the plants. Higher deviation of rDNA and satellite repeats over TEs was also reported in McCann et al. (2018) for allopolyploid species of *Melampodium*.

The RE findings for rDNA and satellite repeats were at first cross-checked with BLASTn searches of the entire NGS datasets. For the majority of satellite repeats and rDNAs, the genomic proportions detected by RE and BLASTn were congruent, but in a few occasions, different results were obtained. For example, 5S rDNA in IntC was not detected by RE, but BLASTn detected a genomic proportion of 0.013%. In contrast, the novel satellite repeat CL217 with its similar genomic proportion in PreC (0.01%) was detected by both RE and BLASTn. It should be noted that the low proportion of 5S rDNA in IntC is most likely not only a consequence of low-coverage sequencing bias as a similarly low proportion of this repeat was found in two independently sequenced libraries.

The overall picture became more complex when FISH results were compared with the results of RE and BLASTn: genomic abundance among accessions was not proportional

to the number of FISH loci for the majority of repeats. This was especially pronounced in the case of 45S rDNA. That the number of loci does not necessarily correspond to the amount of tandemly repeated sequences contained in a locus was already shown in Rosato et al. (2016).

FISH experiments for the very low abundant satellite repeats (detected by BLASTn only) in several accessions also produced different results. The FISH loci were not detected for CL229 in IntC, PreA and PreC, and BLASTn found only trace amounts of this repeat. Probably its abundance may not have been high enough to be detectable by FISH. CL201 was present in a proportion of only 0.005% in PalA, but was sufficient to generate a FISH signal. In contrast, the satellite CL217 did not produce FISH signals despite its slightly higher genomic proportions in the genomes of PreA and PreC (0.008 and 0.01%, respectively). At this moment, it is not possible to answer why CL217 in PreA and PreC was not observed by FISH: most probably its abundance was below the discrimination level of the FISH protocol used (Valárik et al., 2004). CL82 in IntC had a similar genomic proportion detected by BLASTn (0.007%), but no FISH signal was detected. In contrast to other repeats, a closer inspection of this BLASTn finding showed that the corresponding reads displayed a rather low similarity to the sequence of satellite CL82 (maximum of 90% similarity over 60% of the read length only), which, in addition to its low abundance, possibly explains the lack of FISH signals.

The number of loci of the satellite repeats in hybrids (synthetic as well as natural) were in the range of expectation. In contrast to the genomic proportion of the satellites, the number of loci was not altered by hybridization and polyploidization events. The only exception was CL217: in one accession of each *H. pallidiflorum* (PalF) and *H. picroides* (PicF), a proliferation of this satellite was detected as additional minor loci. Moreover, additional minor loci were observed in the selfed progeny of one *H. intybaceum* accession (IntA; data not shown), which could be an indication that already at the intraspecific level, without the action of hybridization and polyploidization, satellites have a dynamic behavior. On the other hand, it could be explained as a consequence of differential chromatin condensation, which prevents the exposure of minor loci during FISH experiments (Krishnan et al., 2001).

Two satellite repeats displayed a hemizygous nature. Belyayev et al. (2018) have already shown CL82 as occurring in a hemizygous state in both parental species. Similarly, two non-homologous FISH loci of the novel marker CL201 were detected in *H. prenanthoides*; one locus was found on a chromosome bearing one 45S rDNA locus and the second on a chromosome bearing both 45S and 5S rDNA loci. Comparison of RE clusters of CL201 and rDNAs revealed that these repeats do not overlap. Furthermore, CL201 displayed intraspecific variation in its localization; in the case of two *H. prenanthoides* accessions, both CL201 loci co-localized with 45S only, but in repeated experiments on other accessions (data not shown), the repeat co-localized with 5S rDNA loci. The hemizygous pattern of CL201 in *H. prenanthoides* corresponds to the pattern subsequently observed in hybrids: their occurrence is in keeping with the genome dosage of *H. prenanthoides* in all natural and synthetic hybrids. This also explains why *H. pallidiflorum* had no (PalF) or one (PalA) locus. As this satellite presents two hemizygous loci, it can easily fail to get inherited, if the chromosome lacking it is passed on to the progeny. Hemizygous loci were reported in some apomictic species of *Pilosella* (Okada et al., 2011) and *Cenchrus* (Akiyama et al., 2005). In our case, the hemizygosity observed was not restricted to apomicts; it was already detected at the level of diploid sexual species and therefore, was not necessarily involved in the emergence of apomixis or facilitating a switch to this reproductive mode. Hemizygous loci can be caused by a hybridogenous origin of diploids (Myburg et al., 2003). Both parental species are in fact ancient hybrids; *H. intybaceum* most likely originated by wide hybridization between *Hieracium* and an unknown genus whereas *H. prenanthoides* shows signatures of a merger between Western and Eastern European lineages of genus *Hieracium* (Fehrer et al., 2007, 2009). The hybrid origin of many diploids in this genus may trigger polyploidization (linked with the apomictic mode of reproduction) as a means to stabilize the diploid hybrid genomes.

CONCLUSION

Our results contribute to the understanding of plant genome evolution following hybridization and polyploidization and add to the knowledge about the genomic landscapes of polyploids known to undergo an asexual mode of reproduction. No evidence for a massive genome reorganization in TE abundance was found by studying repeatomes of parents, their natural apomictic and synthetic diploid hybrids, but a signal for a 'genomic shock' may still be present at the transcription or epigenetic level. The only elements showing deviations from expected abundances were two small pararetrovirus clusters that were significantly overrepresented in all apomicts. Whether there exists any causal relation to the mode of reproduction remains unclear, but this finding is worth further investigation. In contrast to TEs, satellite and rDNAs showed substantial deviations in all hybrids, independent of their ploidy, and were therefore a consequence of hybridization as such. Our study also highlights the need to study low-abundant repeats with a combination of approaches. We hypothesize that some deviations detected by different bioinformatic approaches may be a consequence of the rapidly evolving nature of these elements. Additional discrepancies between bioinformatic and cytogenetic evidence may in some cases be caused by bias in low coverage NGS or by a lack of sensitivity of *in situ* hybridization experiments. Therefore, results concerning low-abundant repeats should be interpreted with caution.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/ena, PRJEB35856; https://www.ncbi.nlm.nih.gov/genbank/, MN784126–MN784131.

AUTHOR CONTRIBUTIONS

JF and DZ conceived of the study. DZ, JF, and YB designed bioinformatics experiments. DZ and MH analyzed the data. JF designed probes. LP and RS did cytogenetic analyses. JJ did molecular labwork. DZ wrote the manuscript. JF and YB suggested on structure and content. All authors contributed to the drafts and gave final approval for publication.

FUNDING

This work was supported by the Czech Science Foundation (17–14620S to JF and P. Mráz) and the long–term research development project no. RVO 67985939 of the Czech Academy of Sciences.

ACKNOWLEDGMENTS

We are grateful to A. Belyayev for the support in RepeatExplorer and cytogenetic analyses and to J. Chrtek, P. Mráz, and J. Pinc for their collaboration in field sampling and crossing experiments. We are also thankful to J. Macas and P. Novák for valuable comments regarding RepeatExplorer analyses. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures. Computational resources were also provided by the ELIXIR–CZ project (LM2015047), part of the international ELIXIR infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.591053/full#supplementary-material

Supplementary Figure 1 | Graphical representation of the newly detected satellite repeats. **(A)** RepeatExplorer graph layouts of the respective clusters showing a circular shape, typical for tandem repeats. **(B)** Self-similarity dot-plots of the cluster contigs created by the YASS genomic similarity search tool. The tandemly repeated motifs are displayed as green, parallel, diagonal lines, while the distance between lines equals lengths of the motifs. The red lines perpendicular to the main diagonal lines indicate inverted repeats.

Supplementary Figure 2 | An example of summary histogram of RE clustering analysis in Hieracium (accession IntA). Superclusters are represented as columns in the histogram; the width of the columns corresponds to the size of superclusters. The horizontal lines within columns represent clusters in descending order of their sizes. The largest supercluster (no. 1) contains 23,14% of input reads and the first five superclusters contain 49% of all input reads.

Supplementary Figure 3 | Comparative analyses of parental species and their natural and synthetic hybrids. **(A–D)** Comparative analyses of natural triploid hybrids and representatives of parental species (IntX - created by pooling reads from both individuals of *H. intybaceum*; PreX - created by pooling reads from both individuals of *H. prenanthoides*. **(E–G)** Comparative analyses of synthetic diploid hybrids and their actual parents. In **(A–D)**, the numbers of reads of triploids were recalculated using the monoploid genome size (1Cx), in order to make them comparable to homoploid genome size (1C) of diploid parents.

Supplementary Figure 4 | Cumulative deviation scores of natural and synthetic *Hieracium* hybrids.

Supplementary Figure 5 | Deviation scores of pararetrovirus clusters. Gray triangles show deviation scores of both pararetrovirus clusters detected in the seven comparative Repeat Explorer analyses. For combinations of samples subjected to comparative analyses, see text and **Table 3**. Deviation scores were grouped by hybrid origin (boxplots) and compared via ANOVA. The deviation score in pararetrovirus clusters was significantly higher in natural hybrids (mean deviation scores: synthetic hybrids −0.04, natural hybrids 0.26; ANOVA: $F = 27.2$; $p < 0.001$). $^{***}p < 0.01$; syn – synthetic hybrids; nat – natural hybrids.

Supplementary Figure 6 | Chromosomal localization of CL82 by FISH in synthetic and natural hybrids. **(A–C)** Synthetic F1 diploid hybrids. **(D–G)** Natural triploid hybrids. FISH signals: CL82 (purple signal, magenta arrows), 45S rDNA (green signal), 5S rDNA (red signal). Chromosomes were counterstained with DAPI (blue). Bars = 5 μm.

Supplementary Table 1 | Individual clustering analyses – detailed annotation of individual clusters of all samples.

Supplementary Table 2 | Summarized annotation at the supercluster (SCL) and cluster (CL) level, including numbers of clusters and superclusters.

Supplementary Table 3 | Comparison of the most important RE individual clustering results with BLASTn searches of NGS datasets, in all accessions.

REFERENCES

Ågren, J. A., Greiner, S., Johnson, M. T. J., and Wright, S. I. (2015). No evidence that sex and transposable elements drive genome size variation in evening primroses. *Evolution* 69, 1053–1062. doi: 10.1111/evo.12627

Akiyama, Y., Hanna, W. W., and Ozias-Akins, P. (2005). High-resolution physical mapping reveals that the apospory-specific genomic region (ASGR) in *Cenchrus ciliaris* is located on a heterochromatic and hemizygous region of a single chromosome. *Theor. Appl. Genet.* 111, 1042–1051. doi: 10.1007/s00122-005-0020-5

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

An, Z., Tang, Z., Ma, B., Mason, A. S., Guo, Y., Yin, J., et al. (2014). Transposon variation by order during allopolyploidisation between *Brassica oleracea* and *Brassica rapa*. *Plant Biol.* 16, 825–835. doi: 10.1111/plb.12121

Ávila Robledillo, L., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., et al. (2018). Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Sci. Rep.* 8:5838. doi: 10.1038/s41598-018-24196-3

Beaulieu, J., Jean, M., and Belzile, F. (2009). The allotetraploid *Arabidopsis thaliana–Arabidopsis lyrata* subsp. *petraea* as an alternative model system for the study of polyploidy in plants. *Mol. Genet. Genomics* 281, 421–435. doi: 10.1007/s00438-008-0421-7

Belyayev, A. (2014). Bursts of transposable elements as an evolutionary driving force. *J. Evol. Biol.* 27, 2573–2584. doi: 10.1111/jeb.12513

Belyayev, A., Paštová, L., Fehrer, J., Josefiová, J., Chrtek, J., and Mráz, P. (2018). Mapping of *Hieracium* (Asteraceae) chromosomes with genus-specific satDNA elements derived from next-generation sequencing data. *Plant Syst. Evol.* 304, 387–396. doi: 10.1007/s00606-017-1483-y

Ben-David, S., Yaakov, B., and Kashkush, K. (2013). Genome-wide analysis of short interspersed nuclear elements SINES revealed high sequence conservation, gene association and retrotranspositional activity in wheat. *Plant J.* 76, 201–210. doi: 10.1111/tpj.12285

Bennetzen, J. L., Ma, J., and Devos, K. M. (2005). Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* 95, 127–132. doi: 10.1093/aob/mci008

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bräutigam, S., and Greuter, W. (2007). A new treatment of *Pilosella* for the Euro-Mediterranean flora [Notulae ad floram euro-mediterraneam pertinentes 24]. *Willdenowia* 37, 123–137. doi: 10.3372/wi.37.37106

Casacuberta, E., and González, J. (2013). The impact of transposable elements in environmental adaptation. *Mol. Ecol.* 22, 1503–1517. doi: 10.1111/mec.12170

Charlesworth, B., and Langley, C. H. (1986). The evolution of self-regulated transposition of transposable elements. *Genetics* 112:359.

Chrtek, J. Jr., Zahradníček, J., Krak, K., and Fehrer, J. (2009). Genome size in *Hieracium* subgenus *Hieracium* (Asteraceae) is strongly correlated with major phylogenetic groups. *Ann. Bot.* 104, 161–178. doi: 10.1093/aob/mcp107

Chrtek, J., Mráz, P., Belyayev, A., Paštová, L., Mrázová, V., Caklová, P., et al. (2020). Evolutionary history and genetic diversity of apomictic allopolyploids in *Hieracium* s.str.: morphological versus genomic features. *Am. J. Bot.* 107, 66–90. doi: 10.1002/ajb2.1413

Comai, L. (2000). "Genetic and epigenetic interactions in allopolyploid plants," in *Plant Gene Silencing*, eds M. A. Matzke and A. J. M. Matzke (Dordrecht: Springer Netherlands), 267–279. doi: 10.1007/978-94-011-4183-3_19

Docking, T. R., Saadé, F. E., Elliott, M. C., and Schoen, D. J. (2006). Retrotransposon sequence variation in four asexual plant species. *J. Mol. Evol.* 62, 375–387. doi: 10.1007/s00239-004-0350-y

Dodsworth, S., Jang, T.-S., Struebig, M., Chase, M. W., Weiss-Schneeweiss, H., and Leitch, A. R. (2017). Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). *Plant Syst. Evol.* 303, 1013–1020. doi: 10.1007/s00606-016-1356-9

Dolgin, E. S., and Charlesworth, B. (2006). The fate of transposable elements in asexual populations. *Genetics* 174:817. doi: 10.1534/genetics.106.060434

Dubin, M. J., Mittelsten Scheid, O., and Becker, C. (2018). Transposons: a blessing curse. *Curr. Opin. Plant Biol.* 42, 23–29. doi: 10.1016/j.pbi.2018.01.003

Fehrer, J., Gemeinholzer, B., Chrtek, J., and Bräutigam, S. (2007). Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in *Pilosella* hawkweeds (*Hieracium*, Cichorieae, Asteraceae). *Mol. Phylogenet. Evol.* 42, 347–361. doi: 10.1016/j.ympev.2006.07.004

Fehrer, J., Krak, K., and Chrtek, J. (2009). Intra-individual polymorphism in diploid and apomictic polyploid hawkweeds (*Hieracium*, Lactuceae, Asteraceae): disentangling phylogenetic signal, reticulation, and noise. *BMC Evol. Biol.* 9:239. doi: 10.1186/1471-2148-9-239

Ferreira de Carvalho, J., de Jager, V., van Gurp, T. P., Wagemaker, N. C. A. M., and Verhoeven, K. J. F. (2016). Recent and dynamic transposable elements contribute to genomic divergence under asexuality. *BMC Genomics* 17:884. doi: 10.1186/s12864-016-3234-9

Galindo-González, L., Mhiri, C., Deyholos, M. K., and Grandbastien, M.-A. (2017). LTR-retrotransposons in plants: engines of evolution. *Gene* 626, 14–25. doi: 10.1016/j.gene.2017.04.051

Glémin, S., and Galtier, N. (2012). "Genome evolution in outcrossing versus selfing versus asexual species," in *Evolutionary Genomics: Statistical and Computational Methods*, Vol. 1, ed. M. Anisimova (Totowa, NJ: Humana Press), 311–335. doi: 10.1007/978-1-61779-582-4_11

Hake, A. A., Shirasawa, K., Yadawad, A., Nadaf, H. L., Gowda, M. V. C., and Bhat, R. S. (2018). Genome-wide structural mutations among the lines resulting from genetic instability in peanut (*Arachis hypogaea* L.). *Plant Gene* 13, 1–7. doi: 10.1016/j.plgene.2017.11.001

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.

Hemleben, V., Kovarik, A., Torres-Ruiz, R. A., Volkov, R. A., and Beridze, T. (2007). Plant highly repeated satellite DNA: molecular evolution, distribution and use for identification of hybrids. *Syst. Biodiv.* 5, 277–289. doi: 10.1017/S147720000700240X

Hickey, D. A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101:519.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kaplan, Z., Jarolímová, V., and Fehrer, J. (2013). Revision of chromosome numbers of Potamogetonaceae: a new basis for taxonomic and evolutionary implications. *Preslia* 85, 421–482.

Kashkush, K., Feldman, M., and Levy, A. A. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160:1651.

Kejnovsky, E., Hawkins, J. S., and Feschotte, C. (2012). "Plant transposable elements: biology and evolution," in *Plant Genome Diversity Volume 1: Plant Genomes, their Residents, and their Evolutionary Dynamics*, eds J. F. Wendel, J. Greilhuber, J. Dolezel, and I. J. Leitch (Vienna: Springer Vienna), 17–34. doi: 10.1007/978-3-7091-1130-7_2

Kraitshtein, Z., Yaakov, B., Khasdan, V., and Kashkush, K. (2010). Genetic and epigenetic dynamics of a retrotransposon after allopolyploidization of wheat. *Genetics* 186:801. doi: 10.1534/genetics.110.120790

Krak, K., Caklová, P., Chrtek, J., and Fehrer, J. (2013). Reconstruction of phylogenetic relationships in a highly reticulate group with deep coalescence and recent speciation (*Hieracium*, Asteraceae). *Heredity* 110, 138–151. doi: 10.1038/hdy.2012.100

Krishnan, P., Sapra, V. T., Soliman, K. M., and Zipf, A. (2001). FISH mapping of the 5S and 18S-28S rDNA loci in different species of *Glycine*. *J. Heredity* 92, 295–300. doi: 10.1093/jhered/92.3.295

Macas, J., Kejnovský, E., Neumann, P., Novák, P., Koblížková, A., and Vyskot, B. (2011). Next generation sequencing-based analysis of repetitive DNA in the model dioceous plant *Silene latifolia*. *PLoS One* 6:e27335. doi: 10.1371/journal.pone.0027335

Madlung, A., Tyagi, A. P., Watson, B., Jiang, H., Kagochi, T., Doerge, R. W., et al. (2005). Genomic changes in synthetic *Arabidopsis* polyploids. *Plant J.* 41, 221–230. doi: 10.1111/j.1365-313X.2004.02297.x

Majeský, Ľ, Krahulec, F., and Vašut, R. J. (2017). How apomictic taxa are treated in current taxonomy: a review. *TAXON* 66, 1017–1040. doi: 10.12705/665.3

Mascagni, F., Giordani, T., Ceccarelli, M., Cavallini, A., and Natali, L. (2017). Genome-wide analysis of LTR-retrotransposon diversity and its impact on the evolution of the genus *Helianthus* (L.). *BMC Genomics* 18:634. doi: 10.1186/s12864-017-4050-6

Matzke, M., Gregor, W., Mette, M. F., Aufsatz, W., Kanno, T., Jakowitsch, J., et al. (2004). Endogenous pararetroviruses of allotetraploid *Nicotiana tabacum* and its diploid progenitors, *N. sylvestris* and *N. tomentosiformis*. *Biol. J. Linnean Soc.* 82, 627–638. doi: 10.1111/j.1095-8312.2004.00347.x

Maumus, F., and Quesneville, H. (2014). Deep investigation of *Arabidopsis thaliana* junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PLoS One* 9:e94101. doi: 10.1371/journal.pone.0094101

McCann, J., Jang, T.-S., Macas, J., Schneeweiss, G. M., Matzke, N. J., Novák, P., et al. (2018). Dating the species network: allopolyploidy and repetitive DNA evolution in american daisies (*Melampodium* sect. *Melampodium*, Asteraceae). *Syst. Biol.* 67, 1010–1024. doi: 10.1093/sysbio/syy024

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* 226, 792–801. doi: 10.1126/science.15739260

Mestiri, I., Chagué, V., Tanguy, A.-M., Huneau, C., Huteau, V., Belcram, H., et al. (2010). Newly synthesized wheat allohexaploids display progenitor-dependent meiotic stability and aneuploidy but structural genomic additivity. *New Phytol.* 186, 86–101. doi: 10.1111/j.1469-8137.2010.03186.x

Mlinarec, J., Skuhala, A., Jurković, A., Malenica, N., McCann, J., Weiss-Schneeweiss, H., et al. (2019). The repetitive DNA composition in the natural pesticide producer *Tanacetum cinerariifolium*: interindividual variation of subtelomeric tandem repeats. *Front. Plant Sci.* 10:613. doi: 10.3389/fpls.2019.00613

Mráz, P., and Zdvořák, P. (2019). Reproductive pathways in *Hieracium* s.s. (Asteraceae): strict sexuality in diploids and apomixis in polyploids. *Ann. Bot.* 123, 391–403. doi: 10.1093/aob/mcy137

Myburg, A. A., Griffin, A. R., Sederoff, R. R., and Whetten, R. W. (2003). Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F1 hybrid based on a double pseudo-backcross mapping approach. *Theor. Appl. Genet.* 107, 1028–1042. doi: 10.1007/s00122-003-1347-4

Neumann, P., Novák, P., Hoštáková, N., and Macas, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA* 10:1. doi: 10.1186/s13100-018-0144-1

Noé, L., and Kucherov, G. (2005). YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* 33(suppl_2), W540–W543. doi: 10.1093/nar/gki478

Novák, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinform.* 11:378. doi: 10.1186/1471-2105-11-378

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 29, 792–793. doi: 10.1093/bioinformatics/btt054

Okada, T., Ito, K., Johnson, S. D., Oelkers, K., Suzuki, G., Houben, A., et al. (2011). Chromosomes carrying meiotic avoidance loci in three apomictic eudicot *Hieracium* subgenus *Pilosella* species share structural features with two monocot apomicts. *Plant Physiol.* 157:1327. doi: 10.1104/pp.111.181164

Okamoto, H., and Hirochika, H. (2001). Silencing of transposable elements in plants. *Trends Plant Sci.* 6, 527–534. doi: 10.1016/S1360-1385(01)02105-7

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M.-A., and Ainouche, M. (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. *New Phytol.* 184, 1003–1015. doi: 10.1111/j.1469-8137.2009.03029.x

Parisod, C., and Senerchia, N. (2012). "Responses of Transposable Elements to Polyploidy," in *Plant Transposable Elements: Impact on Genome Structure and Function*, eds M.-A. Grandbastien and J. M. Casacuberta (Berlin: Springer Berlin), 147–168. doi: 10.1007/978-3-642-31842-9_9

Petit, M., Guidat, C., Daniel, J., Denis, E., Montoriol, E., Bui, Q. T., et al. (2010). Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytol.* 186, 135–147. doi: 10.1111/j.1469-8137.2009.03140.x

Piednoël, M., Sousa, A., and Renner, S. S. (2015). Transposable elements in a clade of three tetraploids and a diploid relative, focusing on Gypsy amplification. *Mobile DNA* 6:5. doi: 10.1186/s13100-015-0034-8

R Core Development Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Renny-Byfield, S., Kovarik, A., Kelly, L. J., Macas, J., Novak, P., Chase, M. W., et al. (2013). Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *Plant J.* 74, 829–839. doi: 10.1111/tpj.12168

Rosato, M., Kovařík, A., Garilleti, R., and Rosselló, J. A. (2016). Conserved organisation of 45S rDNA sites and rDNA gene copy number among major clades of early land plants. *PLoS One* 11:e0162544. doi: 10.1371/journal.pone.0162544

Sarilar, V., Marmagne, A., Brabant, P., Joets, J., and Alix, K. (2011). BraSto, a Stowaway MITE from *Brassica*: recently active copies preferentially accumulate in the gene space. *Plant Mol. Biol.* 77, 59–75. doi: 10.1007/s11103-011-9794-9

Sarilar, V., Palacios, P. M., Rousselet, A., Ridel, C., Falque, M., Eber, F., et al. (2013). Allopolyploidy has a moderate impact on restructuring at three contrasting transposable element insertion sites in resynthesized *Brassica napus* allotetraploids. *New Phytol.* 198, 593–604. doi: 10.1111/nph.12156

Shan, X., Liu, Z., Dong, Z., Wang, Y., Chen, Y., Lin, X., et al. (2005). Mobilization of the active MITE transposons *mPing* and *Pong* in rice by introgression from wild rice (*Zizania latifolia* Griseb.). *Mol. Biol. Evol.* 22, 976–990. doi: 10.1093/molbev/msi082

Staginnus, C., Gregor, W., Mette, M. F., Teo, C. H., Borroto-Fernández, E. G., Machado, M. L. D. C., et al. (2007). Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. *BMC Plant Biol.* 7:24. doi: 10.1186/1471-2229-7-24

Staton, S. E., and Burke, J. M. (2015). Evolutionary transitions in the Asteraceae coincide with marked shifts in transposable element abundance. *BMC Genomics* 16:623. doi: 10.1186/s12864-015-1830-8

Stupar, R. M., Song, J., Tek, A. L., Cheng, Z., Dong, F., and Jiang, J. (2002). Highly condensed potato pericentromeric heterochromatin contains rDNA-related tandem repeats. *Genetics* 162, 1435.

Valárik, M., Bartoš, J., Kováóová, P., Kubaláková, M., De Jong, J. H., and Doležel, J. (2004). High-resolution FISH on super-stretched flow-sorted plant chromosomes. *Plant J.* 37, 940–950. doi: 10.1111/j.1365-313X.2003.02010.x

van Dijk, E. L., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* 322, 12–20. doi: 10.1016/j.yexcr.2014.01.008

Van Dijk, P. J., and Vijverberg, K. (2005). "The significance of apomixis in the evolution of the angiosperms: a reappraisal," in *Plant Species—Level Systematics: New Perspectives on Pattern and Process*, eds L. C. F. Bakker, B. Gravendeel, and P. B. Pelser (Ruggell: Gantner Verlag), 101–116.

Vicient, C. M., and Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* 120, 195–207. doi: 10.1093/aob/mcx078

Vitales, D., Álvarez, I., Garcia, S., Hidalgo, O., Nieto Feliner, G., Pellicer, J., et al. (2019). Genome size variation at constant chromosome number is not correlated with repetitive DNA dynamism in *Anacyclus* (Asteraceae). *Ann. Bot.* 125, 611–623. doi: 10.1093/aob/mcz183

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982. doi: 10.1038/nrg2165

Xu, Y., Zhong, L., Wu, X., Fang, X., and Wang, J. (2009). Rapid alterations of gene expression and cytosine methylation in newly synthesized *Brassica napus* allopolyploids. *Planta* 229, 471–483. doi: 10.1007/s00425-008-0844-8

Yaakov, B., and Kashkush, K. (2010). Massive alterations of the methylation patterns around DNA transposons in the first four generations of a newly formed wheat allohexaploid. *Genome* 54, 42–49. doi: 10.1139/G10-091

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

HYPOTHESIS AND THEORY
published: 16 November 2020
doi: 10.3389/fpls.2020.592356

Habitat Shape Affects Polyploid Establishment in a Spatial, Stochastic Model

Jonathan P. Spoelhof[1,2], Douglas E. Soltis[1,2] and Pamela S. Soltis[1]*

[1] Florida Museum of Natural History, University of Florida, Gainesville, FL, United States, [2] Department of Biology, University of Florida, Gainesville, FL, United States

Polyploidy contributes massively to the taxonomic and genomic diversity of angiosperms, but certain aspects of polyploid evolution are still enigmatic. The establishment of a new polyploid lineage following whole-genome duplication (WGD) is a critical step for all polyploid species, but this process is difficult to identify and observe in nature. Mathematical models offer an opportunity to study this process by varying parameters related to the populations, habitats, and organisms involved in the polyploid establishment process. While several models of polyploid establishment have been published previously, very few incorporate spatial factors, including spatial relationships between organisms, habitat shape, or population density. This study presents a stochastic, spatial model of polyploid establishment that shows how factors such as habitat shape and dispersal type can influence the fixation and persistence of nascent polyploids and modulate the effects of other factors. This model predicts that narrow, constrained habitats such as roadsides and coastlines may enhance polyploid establishment, particularly in combination with frequent clonal reproduction, limited dispersal, and high population density. The similarity between this scenario and the growth of many invasive or colonizing species along disturbed, narrow habitats such as roadsides may offer a partial explanation of the prevalence of polyploidy among invasive species.

Keywords: dispersal, genome duplication, invasive species, polyploid establishment, spatial, stochastic model

OPEN ACCESS

Edited by:
Karol Marhold, Slovak Academy of Sciences, Slovakia

Reviewed by:
Eric Baack, Luther College, United States
Santiago Ramírez Barahona, National Autonomous University of Mexico, Mexico

****Correspondence:***
Jonathan P. Spoelhof
spoelhof.jon@ufl.edu

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *06 August 2020*
Accepted: *22 October 2020*
Published: *16 November 2020*

Citation:
Spoelhof JP, Soltis DE and Soltis PS (2020) Habitat Shape Affects Polyploid Establishment in a Spatial, Stochastic Model. Front. Plant Sci. 11:592356. doi: 10.3389/fpls.2020.592356

INTRODUCTION

Polyploid evolution is impossible to understand without knowing how and under what conditions new polyploid populations become established. Polyploid establishment is ephemeral and difficult to observe, so researchers have generally investigated this process through a combination of trait comparisons between diploids and recent polyploids, phylogenetic inference of trait evolution and ploidy, and mathematical models of the establishment process. Many traits associated with polyploidy (e.g., self-compatibility and perenniality) may be inherited from diploid progenitors, result from whole-genome duplication (WGD), or evolve after establishment (discussed below). Knowing the sequence of evolution in these traits is critical to understanding their potential

contribution to establishment. Inferring that order is difficult, however, and consistent patterns may be clade-specific or dependent on other stochastic factors.

Polyploidy may facilitate the breakdown of self-incompatibility, particularly through genotype-specific effects (Mable, 2004). However, broad comparisons of diploid and polyploid taxa show minor or inconsistent associations between self-incompatibility, mating system and ploidy (Mable, 2004; Barringer, 2007), and phylogenetic studies of mating system and polyploidy in Solanaceae (Robertson et al., 2011; Zenil-Ferguson et al., 2019) have shown that self-compatibility is more likely to evolve prior to polyploidy than in concert with it. This trait is also frequently variable within species or even populations, which further complicates any inference performed at the species level. Polyploidy may also promote the expression of clonality (Van Drunen and Husband, 2018) or serve as a precursor for apomictic reproductive pathways (Hörandl and Hojsgaard, 2012; Hojsgaard et al., 2014; Hojsgaard and Hörandl, 2019). There is a strong association between polyploidy and clonality, yet phylogenetic analyses of evolutionary order (i.e., whether polyploidy precedes clonality or vice versa) have produced inconsistent results, particularly among clades (Herben et al., 2017; Van Drunen and Husband, 2019). Finally, polyploidy may spur the evolution of perenniality, possibly through a decrease in the rate of growth and development (as noted in Te Beest et al., 2012), but this is certainly not the rule. As with clonality, which is highly correlated with perenniality, inferring the order of evolution between these two traits yields inconclusive results (Van Drunen and Husband, 2019).

Mathematical models may provide the most general predictions about the factors and traits that facilitate polyploid establishment, at least in lieu of repeated, direct observations or tests of establishment itself. The model presented by Levin (1975) predicted that polyploid establishment is very unlikely for organisms that primarily outcross, while self-fertilization and clonal reproduction mitigate the mating disadvantage experienced by polyploids as the minority cytotype due to reproductive interference from diploids (minority cytotype exclusion, or MCE). This pattern recurs in other establishment models with variable parameters representing reproductive assurance, either through selfing (Rodriguez, 1996; Baack, 2005; Rausch and Morgan, 2005; Oswald and Nuismer, 2011; Fowler and Levin, 2016) or clonality (Chrtek et al., 2017), and the prevalence of at least some form of reproductive assurance across plant lineages with frequent recent and ancient polyploidy suggests that it is a critical driver of polyploid success (Spoelhof et al., 2019).

However, one very neglected area of polyploidy establishment modeling is spatial dynamics. Non-spatial models assume random spatial interactions between individuals in a population, which is almost necessarily violated in sessile organisms. Outcrossing, sessile organisms are more likely to mate with nearby individuals, and dispersal probability, while variable, will generally decrease proportionally with distance from that individual (Levin and Kerster, 1974). To our knowledge, one published model has focused on the interactions among polyploid and diploid plants in a fully spatial context (but see Li et al., 2004): Baack (2005) modeled the effects of selfing rate, polyploid advantage, and dispersal distance of pollen and seeds on the persistence of polyploids in a model parameterized on the species *Ranunculus adoneus*, which comprises diploid and autotetraploid cytotypes. One novel conclusion of this study was that the spatial effects of short seed dispersal distances increased the probability of polyploid persistence. Still, this model leaves many questions about the spatial aspects of polyploid establishment unanswered. How much do spatial effects contribute to polyploid establishment in a spatial model compared to a null (i.e., non-spatial) model? Do these effects interact with other parameters such as reproductive assurance or perenniality? How do habitat size and shape influence polyploid persistence?

This paper describes a new spatial, stochastic polyploid establishment model that addresses these questions. This model includes parameters for population size (K), reproductive assurance (Ra) through selfing or clonality, lifespan (annual vs. perennial), and habitat shape (square vs. narrow). The model simulates the introduction of a single autopolyploid to an otherwise diploid population, then tracks the length of time that cytotype polymorphism persists in the population and whether or not the polyploid cytotype is eventually fixed. Importantly, the model does not incorporate any trait or fitness differences between cytotypes (aside from the fact that they are reproductively isolated from each other, which can lead to reproductive interference when mating occurs between cytotypes), and each simulation is compared to a non-spatial control with the same starting conditions.

METHODS

Model Construction

A stochastic, simulation-based model with parameters for habitat length and width (Hl and Hw, respectively), population size, individual lifespan (Ls), and reproductive assurance was developed in R (R Core Team, 2019). Two versions of the model were made to reflect different modes of reproductive assurance: a "clonal" model, and a "selfing" model. Both models begin by randomly populating a discrete spatial habitat of width Hw and length Hl with K individuals. For example, a habitat of $Hw = Hl = 100$ that contains $K = 200$ individuals would have a square shape, 10,000 discrete cells, and a population density of 2%. All individuals are self-compatible and able to act as either maternal (uniparentally or biparentally) or paternal parents of offspring. The spatial habitat is also occupied by a number of randomly placed, non-interactive individuals, such that the sum of K and the number of non-interactive individuals occupies 95% of the total cells in the habitat (this ratio was held constant in all simulations). Non-interactive individuals do not interact with simulated individuals (as described below) and are included to better represent a habitat that is co-occupied by other species and experiences a reasonable amount of disturbance (i.e., 5% of the habitat is available to new recruits in any given season). For example, if no non-interactive individuals were included, the model would represent a scenario where

the habitat is completely open and none of the space has been occupied by other species. This is not realistic, even in frequently disturbed habitats. Individuals have staggered ages, so the habitat contains an equal number of individuals of each possible age (1 through *Ls*) during each simulated season. At the beginning of each season, all individuals of age *Ls* are removed from the population. Next, a maternal individual is randomly sampled from the population and produces a single offspring uniparentally with probability *Ra*. In the clonal model, this offspring can only disperse to an adjacent, empty position in the habitat. In the selfing model, the offspring disperses to an empty position in the habitat with a probability proportional to the inverse square of the distance between the maternal individual and the empty position. If uniparental reproduction does not occur (probability = $1 - Ra$), a paternal individual is sampled from the rest of the population with a probability proportional to the inverse square of the distance between the maternal individual and the paternal individual. If the maternal and paternal individuals share the same cytotype, an outcrossed offspring is produced and dispersed to an empty position in the habitat with a probability proportional to the inverse square of the distance between the maternal individual and the empty position. If the maternal and paternal individuals do not share the same cytotype, they do not produce offspring. This process is repeated until the population reaches *K* individuals again, at which point the age of all individuals advances by one, the locations of the non-interactive individuals are randomized within the remaining cells, and the next season begins.

Initially, the simulated population is entirely diploid, and seasons proceed until the population turns over 50 times (i.e., $50 \times Ls$ seasons). This period allows the initial diploid population to reach a spatial equilibrium. Visual plotting of the populations during this stage showed that the qualitative spatial distributions (e.g., highly clumped and evenly dispersed) that emerged from a given set of model parameters usually became stable before this point. Next, a single individual in the population is replaced with an autopolyploid, and the model proceeds until either diploids or polyploids reach fixation. Aside from their inability to mate with each other, diploids and polyploids are treated identically by the model. If fixation does not occur after the population turns over 1,000 times ($1{,}000 \times Ls$ seasons after the polyploid is introduced), cytotype polymorphism is considered stable and the simulation halts. For each simulation, the final ploidy of the population (diploid, polyploid, or polymorphic) is recorded along with the total number of seasons required to reach fixation, if it occurs.

For each parameter combination used in this study, a control simulation was performed using identical parameters and starting conditions. In control simulations, the probability of mating between any two individuals is not dependent on the distance between those individuals, and the probability of dispersal to any unoccupied cell is not dependent on the distance between the maternal individual and that cell. Therefore, control simulations are functionally non-spatial (even though they occur within the same spatial framework as the primary model) and simulate fully random mating between all individuals in a population. As most prior models of polyploid establishment assume random mating in a non-spatial framework, we included these simulations to identify the specific contributions of spatial factors to the predictions of this model.

Simulation Conditions

A variety of parameter combinations was used to assess the model (**Table 1**). All possible combinations of these variables were simulated in the selfing model. In the clonal model, only the perennial lifespan ($Ls = 10$) was used because clonality is limited to perennial species. Two habitat shapes were used in this model: Square ($Hw{:}Hl = 1{:}1$) and narrow ($Hw{:}Hl = 1{:}64$). Square habitats were meant to represent relatively unconstrained, contiguous regions (e.g., a large forest tract), whereas narrow habitats were meant to represent highly constrained habitats such as road margins and coastlines. The spatial parameters were specified so that the effects of population density, habitat shape, and habitat size could be controlled (**Table 1**). For example, a population of 200 individuals in a narrow habitat (8 cells × 512 cells) could be compared to a population of 50 individuals in a habitat with the same shape and population density (4 cells × 256 cells).

We have included the code for this model in the **Supplementary Material**. Although R is not the most efficient environment for stochastic modeling, it is accessible to a wider academic audience in biology than other languages (e.g., C++). As a relatively simple model, this code could be used as an introductory resource for stochastic modeling, or to test hypotheses not considered in this paper. For example, any habitat that can be specified as a matrix with habitable and uninhabitable cells can be used, so users could test hypotheses about habitat fragmentation or specify habitats based on geographical locations of interest.

RESULTS

Fixation Time

The diploid cytotype becomes fixed very quickly in most simulations (**Figures 1–3**) because the initial proportion of polyploids is low [$1/(K - 1)$]. Higher reproductive assurance and perenniality consistently increased time to fixation (**Figure 5**). The effects of population size, population density, and habitat

TABLE 1 | Establishment model parameters and values.

Parameter	Description	Values
Hw	Habitat width	Square: 32, 64 Narrow: 4, 8
Hl	Habitat length	Square: 32, 64 Narrow: 256, 512
K	Population carrying capacity	Small: 50 Large: 200
Ls	Lifespan (reproductive seasons + 1)	Annual: 2 Perennial: 10
Ra	Reproductive assurance (probability of uniparental reproduction)	Outcrossing: 0.05 Mostly outcrossing: 0.25 Mixed mating: 0.5 Mostly uniparental: 0.75 Uniparental: 0.95

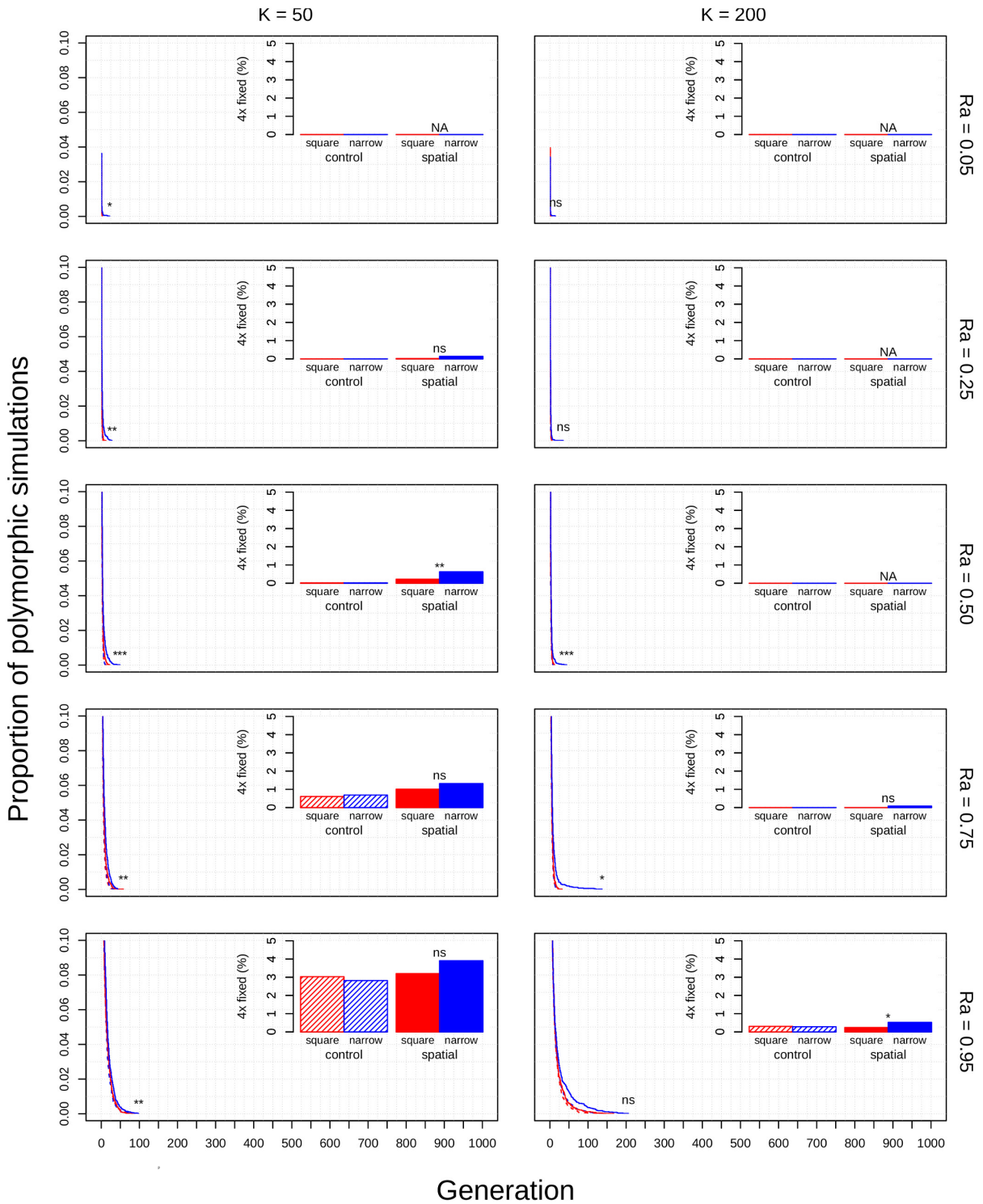

FIGURE 1 | Polyploid persistence and fixation in simulated annual populations with selfing reproductive assurance. Values of *K* (columns) and *Ra* (rows) are noted in the margins. Each panel contains a survival plot showing the proportion of simulations that contained multiple cytotypes at each generation, and a bar plot showing the percentage of simulations that fixed as polyploid for each parameter combination. Results from non-spatial control simulations are shown as dashed lines and shaded bars, results from spatial simulations are shown as solid lines and solid bars, results from simulations in square habitats are shown in red, and results from simulations in narrow habitats are shown in blue. Asterisks denote the significance of log-rank tests (in survival plots) and Fisher tests (in bar plots) of differences between simulations in square and narrow habitats. Only results from simulations at a population density of 4.9% are shown. Significance: *ns* = $p > 0.05$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. See **Supplementary Figure 1** in the **Supplementary Material** for a magnified view of the results from these simulations.

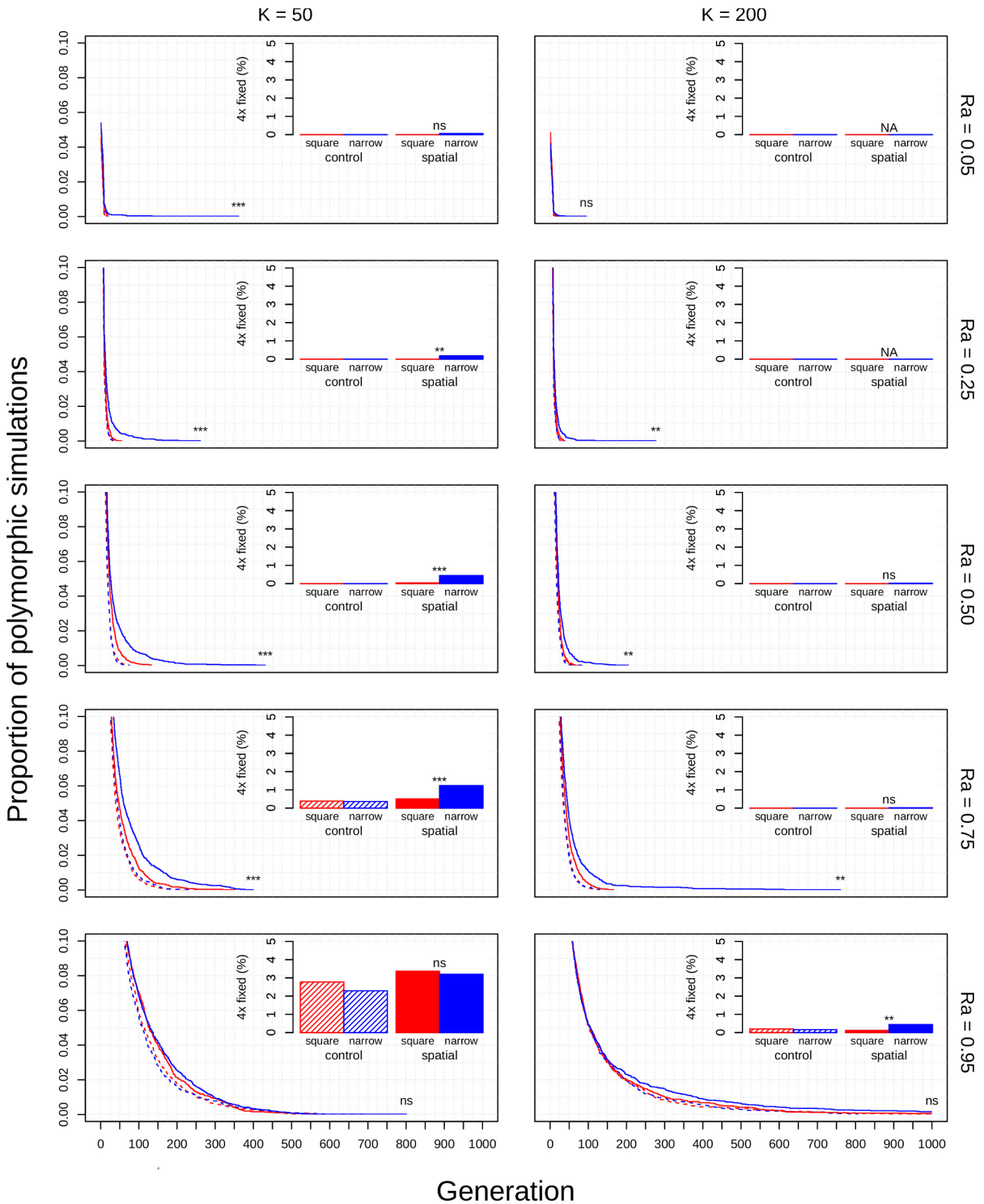

FIGURE 2 | Polyploid persistence and fixation in simulated perennial populations with selfing reproductive assurance. Values of *K* (columns) and *Ra* (rows) are noted in the margins. Each panel contains a survival plot showing the proportion of simulations that contained multiple cytotypes at each generation, and a bar plot showing the percentage of simulations that fixed as polyploid for each parameter combination. Results from non-spatial control simulations are shown as dashed lines and shaded bars, results from spatial simulations are shown as solid lines and solid bars, results from simulations in square habitats are shown in red, and results from simulations in narrow habitats are shown in blue. Asterisks denote the significance of log-rank tests (in survival plots) and Fisher tests (in bar plots) of differences between simulations in square and narrow habitats. Only results from simulations at a population density of 4.9% are shown. Significance: *ns* = $p > 0.05$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

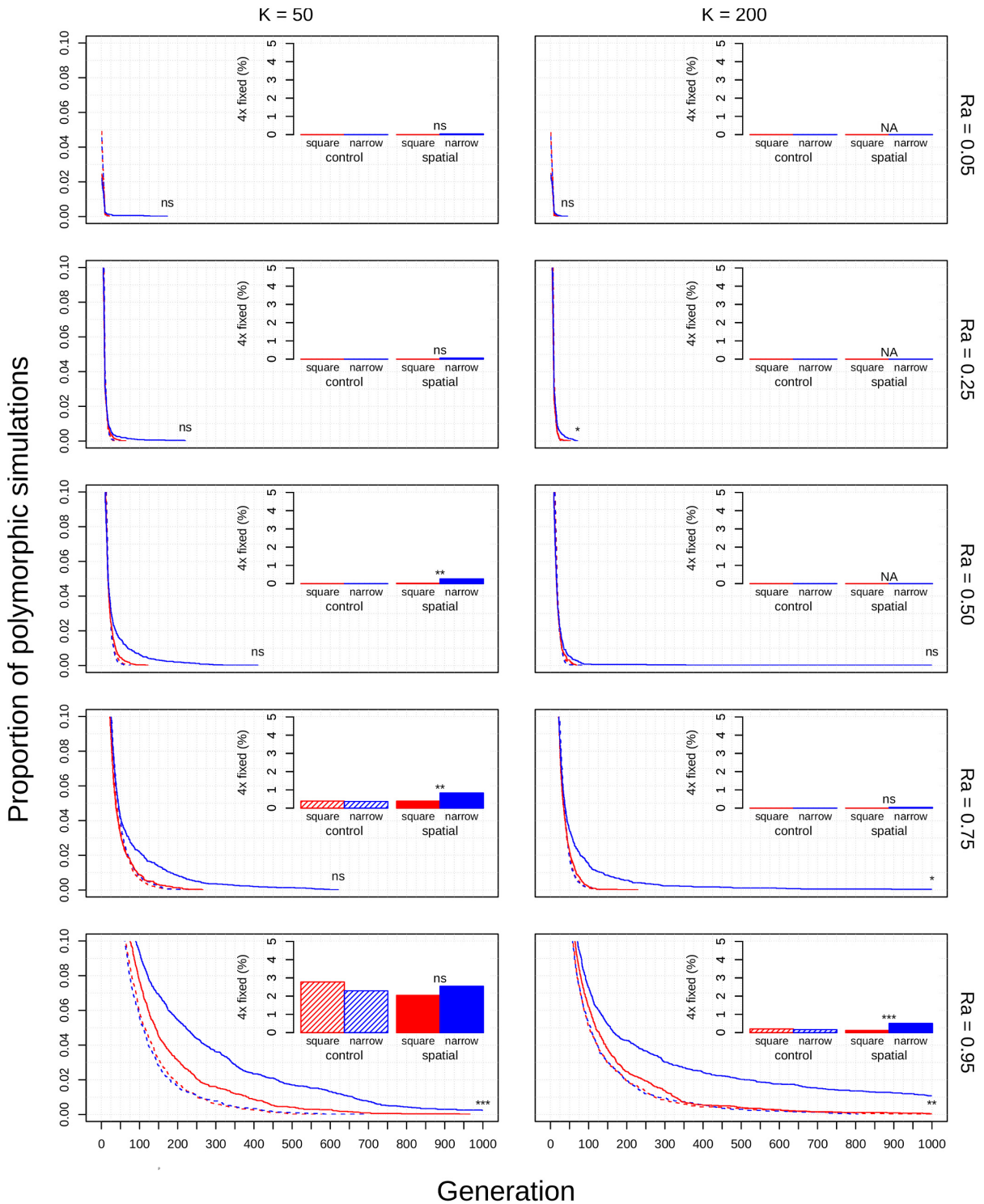

FIGURE 3 | Polyploid persistence and fixation in simulated perennial populations with clonal reproductive assurance. Values of *K* (columns) and *Ra* (rows) are noted in the margins. Each panel contains a survival plot showing the proportion of simulations that contained multiple cytotypes at each generation, and a bar plot showing the percentage of simulations that fixed as polyploid for each parameter combination. Results from non-spatial control simulations are shown as dashed lines and shaded bars, results from spatial simulations are shown as solid lines and solid bars, results from simulations in square habitats are shown in red, and results from simulations in narrow habitats are shown in blue. Asterisks denote the significance of log-rank tests (in survival plots) and Fisher tests (in bar plots) of differences between simulations in square and narrow habitats. Only results from simulations at a population density of 4.9% are shown. Significance: *ns* = $p > 0.05$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

shape on time to fixation were more complex. Higher population size resulted in faster rates of fixation in earlier generations, but higher probabilities of polymorphism persistence in later generations (**Figures 1–3**), particularly in the clonal model and in simulations with higher reproductive assurance. Narrow habitat shape also generally increased time to fixation compared to square habitat shape and non-spatial controls (**Figures 1–3**). This effect appears to increase in proportion with population density in the clonal model, but diminishes at the highest level of reproductive assurance in the selfing model (**Figure 4**).

For simulations that resulted in polyploid fixation, time to fixation was still largely dependent on population size, reproductive assurance, and lifespan (**Figure 5**), but the results varied based on model type. The times to polyploid fixation were more variable in the selfing and clonal models than in the control model, and fixation times were much higher overall in simulations where $K = 200$ in the clonal model (**Figure 5**).

Polyploid Fixation Probability

Polyploid fixation was highest in simulations with low population size and high reproductive assurance (**Figures 1–3**). In general, polyploid fixation was also higher in spatial simulations with narrow habitat shape. One exception to this trend is that spatial differences in polyploid fixation probability weaken or disappear with low population size and high reproductive assurance, particularly in perennial simulations of the selfing model (**Figures 1–3**).

DISCUSSION

Caveats

The simulations used in this model do not include intercytotype gene flow or the formation of triploids from intercytotype crosses. These opportunities were not included primarily to simplify the model. The production of unreduced gametes and triploid fertility are both highly variable factors that are likely to be inconsistent across taxa and even throughout the timeline of establishment (Ramsey and Schemske, 1998; Köhler et al., 2010; Mason and Pires, 2015; Kreiner et al., 2017). This is also why each simulation included a single WGD, as the rate of spontaneous polyploid formation will depend on the same, highly variable

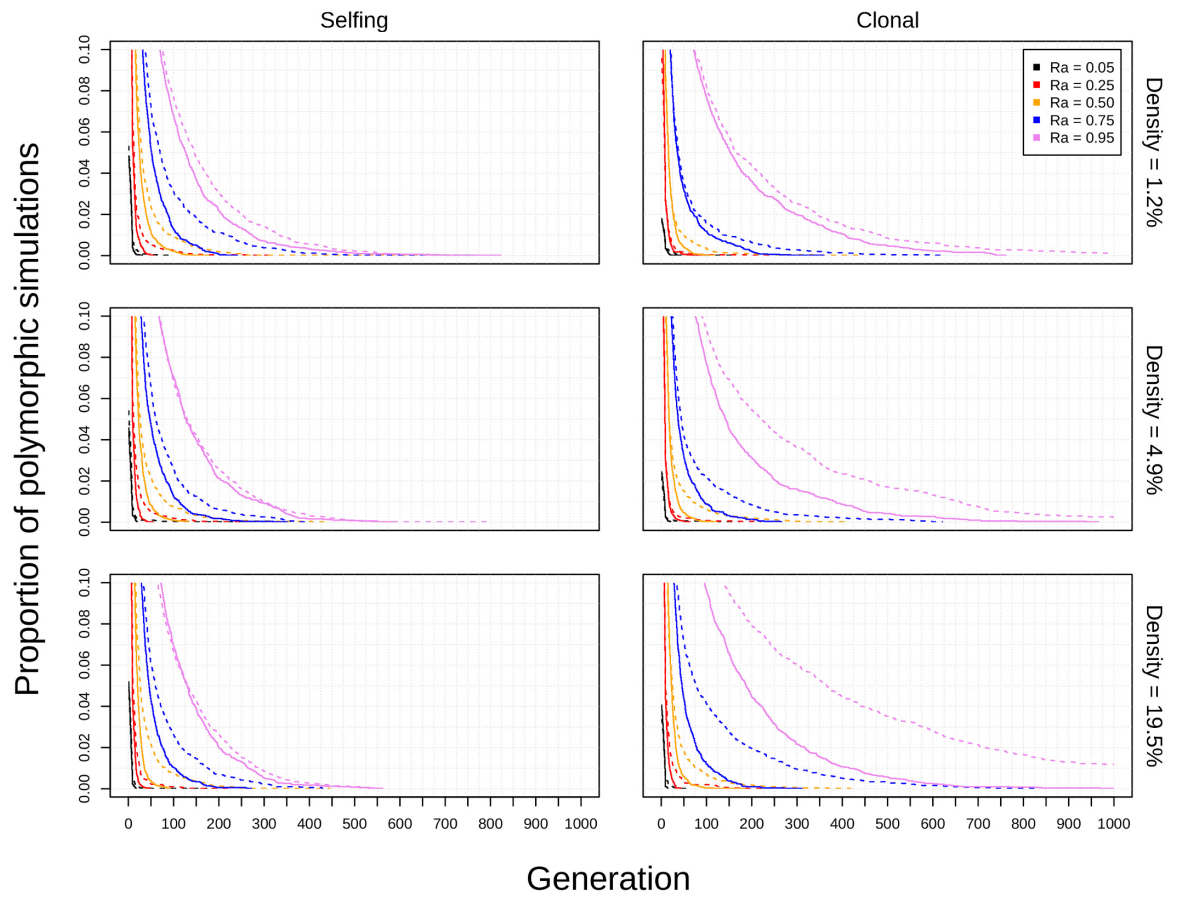

FIGURE 4 | Effects of population density on polyploid persistence in spatial simulations ($K = 50$). Model type (columns) and population density (rows) are noted in the margins. Each panel contains a survival plot showing the proportion of simulations that contained multiple cytotypes at each generation, with varying values *Ra* represented by different colors (see legend). Results from simulations in square habitats are shown as solid lines, and results from simulations in narrow habitats are shown as dotted lines.

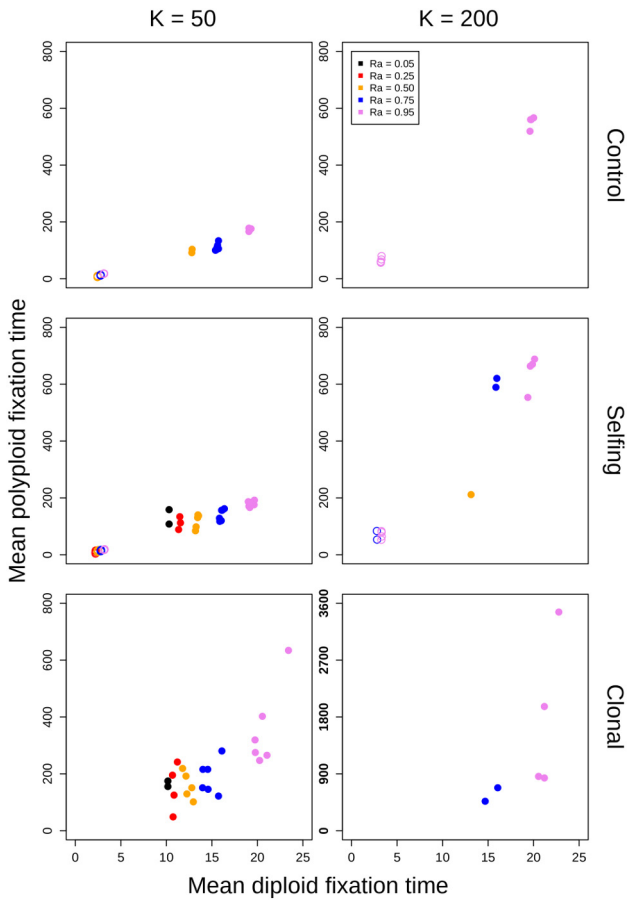

FIGURE 5 | Polyploid fixation time vs. diploid fixation time. Values of *K* (columns) and model type (rows) are noted in the margins. In each panel, the geometric means of polyploid fixation time for all parameter combinations are plotted against the geometric means of diploid fixation time for all parameter combinations. Open circles represent simulations of annual populations, and closed circles represent simulations of perennial populations. Note the different *y*-axis scaling in the plot showing clonal simulations at $K = 200$.

factors. Both processes would act to favor the persistence and fixation of polyploids in the population, so measures of polyploid success are estimated conservatively in this study.

The inclusion of only a single WGD was also intentional for another reason. WGDs are rare events (Ramsey and Schemske, 1998), but established polyploid populations and species are typically the result of multiple formations (Soltis and Soltis, 1999). The long-term success of a polyploid lineage likely depends on multiple formations or gene flow from diploids (through unreduced gametes or odd-ploidy intermediates) as a means of developing genetic diversity within a population, which is vital for adaptation and response to selective pressures. By determining how long cytotype polymorphism persists in a population following a single WGD, these models can also estimate the relative likelihood that WGDs will overlap in the same population (or that rare, intercytotype gene flow will occur). In other words, conditions that favor longer polyploid persistence favor the accumulation of successive WGDs and genetic diversity among polyploids. While polyploid fixation is another measure of polyploid success in this study, rapid fixation will likely favor the production of genetically uniform polyploid populations that may be less adaptable in the long term.

Additionally, this model did not consider the effects of dispersal outside of the defined habitat. Offspring lost to dispersal outside of the suitable habitat would not affect the size of the population, as reproduction would still proceed until the population reached carrying capacity during each generation. Still, individuals along the periphery of their habitat would be less likely to achieve successful dispersal than those in the interior. The spatial difference between central and peripheral individuals is larger in square habitats, but any difference in this effect between square and narrow habitats would be dependent on average dispersal distance, which was not manipulated in this model. Further modeling could examine this effect by varying dispersal distance, as well as the geometric gradation between square and narrow habitats.

Finally, we did not include other important factors such as inbreeding depression, variable fitness or lifespan between cytotypes, prezygotic barriers between cytotypes, or differences in habitat disturbance. Our primary goal was to determine whether spatial dynamics (particularly habitat shape) could affect polyploid establishment in a simple, controlled framework. We acknowledge that these omissions limit the generality of our conclusions, and, as we noted in the methods, we encourage other researchers to build on this model framework and incorporate new factors or specify parameters based on their system of interest.

Model Assessment

Each simulation of this model includes three critical phases. In the first phase, the initial polyploid must reproduce at least once during its lifespan. If that happens, the simulation reaches the second phase, where there are multiple polyploids in the population that can reproduce uniparentally or biparentally. If polyploids are able to reproduce until they reach half of the population carrying capacity, the population reaches the third phase: a tipping point where there is an equal chance that polyploids and diploids will reach fixation because they are treated neutrally by the model. The first and second phases are critical in determining the likelihood of polyploid fixation. Factors that act to reduce MCE during these phases increase the probability that polyploids will reach parity with diploids in the population.

The non-spatial parameters of this model all had substantial impacts on the rate of polyploid fixation or the time to fixation. High reproductive assurance directly reduces polyploid MCE by decreasing the probability of reproductive interference (**Figures 1–3**). Reproductive assurance is especially critical during the first phase of each simulation; without the possibility of intercytotype crossing, the first polyploid must rely on uniparental reproduction to establish a breeding population. Simulations with higher reproductive assurance allow polyploids to reproduce more quickly in phase two and lose fewer reproductive opportunities to intercytotype crosses. These effects are equally present in the third phase, where they reduce

the chance that either cytotype will go extinct by reducing the proportional effects of MCE. This stabilizing effect is reflected in the longer times to fixation observed in simulations with high reproductive assurance (**Figure 5**). Alternatively, low reproductive assurance caused nearly all populations to fix as diploid within a short period of time due to the unmitigated effects of MCE (**Figures 1–3**). Low population size also directly reduces polyploid MCE, but through a different mechanism: by increasing the initial proportion of polyploids in the population. In small populations, less reproduction is required for the initial polyploid to reach parity with diploids and accelerates the process of cytotype drift in phase three. Rates of polyploid fixation are therefore much higher in simulations with low population sizes (**Figures 1–3**). While annual lifespan did not greatly impact the rate of polyploid fixation, it did greatly decrease the time to fixation (**Figure 5**). This makes intuitive sense: annual populations turn over much more quickly, so the minimum time required for either diploids or polyploids to reach fixation is reduced. This distinction may be important in explaining the differential frequency of occurrence of polyploids among annuals and perennials in nature—polyploidy appears to be much more common in herbaceous perennials than in annuals (Stebbins, 1938; Husband et al., 2013; Van Drunen and Husband, 2019). If rapid fixation in annuals reduces the probability of overlapping WGDs and intercytotype gene flow within the population, then annual polyploid populations may be more prone to extinction because of a lack of genetic diversity and eventual deterioration through inbreeding depression or the accumulation of deleterious alleles.

The theoretical impacts of reproductive assurance and population size on polyploid establishment have been examined in previously published deterministic and stochastic models (Rodriguez, 1996; Baack, 2005; Rausch and Morgan, 2005; Oswald and Nuismer, 2011; Fowler and Levin, 2016; Chrtek et al., 2017), and the results presented here are qualitatively similar to the results of those studies. However, these factors have not been studied extensively within a spatial context (aside from Baack, 2005). The selfing and clonal models generally relaxed constraints on polyploid fixation, particularly regarding reproductive assurance. In the control model, no polyploid fixation was seen at $Ra < 0.5$ in populations of 50 individuals and at $Ra < 0.95$ in populations of 200 individuals. In the selfing and clonal models, polyploid fixation was observed at all values of Ra in populations of 50 individuals and at $Ra > 0.5$ in populations of 200 individuals. This effect is likely due to the spatial dynamics of mating and dispersal—assortative mating among cytotypes is more likely when dispersal is limited (Li et al., 2004; Baack, 2005), which mitigates the effects of MCE that polyploids experience in the initial phases of establishment. However, this effect varied with habitat shape in both spatial models. Most of the increase in polyploid fixation in spatial models came from simulations in narrow habitats; square habitats often had rates of polyploid fixation similar to or slightly larger than non-spatial control simulations (**Figures 1–3**). This effect likely results from an amplification of assortative mating in habitats that are more spatially restricted, especially if dispersal is limited, too. Narrow habitats also generally increased time to fixation, with particularly pronounced effects in the clonal model (**Figure 3**). Because the offspring of simulated clonal organisms could only disperse to adjacent cells, each cytotype could effectively block an entire portion of the narrow habitat. This spatial organization stabilizes polymorphism within the population, and the effect increased with greater population density within the clonal model (**Figure 4**).

Significance

Spatial effects may relax constraints on polyploid establishment by causing non-random mating and dispersal. In the selfing and clonal models described above, factors that limited the spatial probabilities of mating and dispersal generally counteracted the effects of MCE. These factors included a clonal mode of reproductive assurance and narrow habitat shape, in particular.

Both clonal reproduction and selfing directly counteract MCE, but clonal reproduction results in more closely spaced groups of plants. If these plants are also capable of outcrossing, they are much more likely to mate with their own clonal offspring than with other individuals in the population. In the context of a polycytotypic population, this results in assortative mating among diploids and polyploids, respectively. It also decreases the probability that the offspring of another cytotype will disperse into a "clump" of plants and begin to displace them. Polyploidy and clonality are very strongly associated in angiosperms (Herben et al., 2017; Van Drunen and Husband, 2019). While there is a possibility that WGD may induce or enhance clonality (Van Drunen and Husband, 2019), this study demonstrates that preexisting clonal traits may facilitate polyploid establishment, even when these traits do not differ between diploids and polyploids. Unlike selfing, clonality could allow polyploids to persist in a population without experiencing the effects of inbreeding depression (which was not considered in this study), possibly increasing the relative chance that they will avoid extinction long enough to overlap with subsequent WGDs in the population.

The effects of habitat shape have not been examined in previous models of polyploid establishment, but it varies widely in both natural and disturbed areas. Coastlines, riverbanks, cliff sides, roadsides, railways, and agricultural margins are all examples of relatively linear habitats whose communities tend to differ from those in surrounding areas. In these cases, suitable habitat may be restricted in one dimension, so only dispersal along a narrow corridor will be successful. The results of this study suggest that narrow habitats can increase the probability of polyploid fixation and delay the fixation of any cytotype following WGD, and that these effects may be particularly pronounced in highly clonal, high-density populations. Populations of many invasive angiosperm species often share these characteristics, and they are also more likely to be polyploid (Pandit et al., 2011; Te Beest et al., 2012). The association between polyploidy and invasiveness has usually been supported with adaptive explanations, including that polyploids will have broader ecological niches, different reproductive traits, or increased vigor when compared to their diploid progenitors (Prentis et al., 2008; Te Beest et al., 2012). While these explanations certainly hold true in some cases, there may

be an additional, neutral explanation for the enrichment of polyploidy among invasive species. The disturbance associated with roads and railways provides invasive species with exploitable habitat, and the linear, interconnected nature of these habitats also allows these species to spread to new areas. Both of these factors are critical in the establishment of many invasive species (e.g., Mortensen et al., 2009; Joly et al., 2011; Dar et al., 2015). We have shown that long, narrow habitats may enhance polyploid establishment and fixation without invoking any difference between the characteristics of diploid and polyploid cytotypes. If this model is representative of nature, one would expect invasive plants that exploit narrow habitats to be (1) more frequently polyploid or polycytotypic, and (2) more frequently polyploid in their invaded habitat than in their natural habitat (assuming their natural habitat is not often "narrow"). Both of these predictions are supported by observational studies in several species (Pyšek et al., 2009; Pandit et al., 2011; Te Beest et al., 2012). Furthermore, *Tragopogon miscellus* and *Senecio cambrensis,* some of the best-studied examples of recently formed polyploid species, arose through hybridization of introduced species and were discovered along disturbed, often narrow habitats such as railway margins and roadsides (Ownbey, 1950; Novak et al., 1991; Abbott and Lowe, 2004). Another well-studied and recently formed polyploid, *Spartina anglica,* also arose through hybridization of an introduced species and spread along a different type of narrow habitat—coastlines in the United Kingdom (Gray et al., 1991). Notably, each of these species is capable of reproductive assurance (via self-fertilization in *T. miscellus* and *S. cambrensis* and extensive clonal spread in *S. anglica*), which, according to this model and others, can also facilitate polyploid establishment. Therefore, disentangling the contributions of habitat shape, reproductive traits, and disturbance to polyploid establishment is difficult. One way to isolate these factors would be to intensively measure ploidy variation (e.g., via flow cytometry) in plant communities occupying natural, linear habitats and comparing the results to nearby, less-linear communities. For example, the effects of habitat shape observed in our model would be indirectly corroborated if (1) polyploidy is more common in a riverside community than in an adjacent forest community, and (2) both communities experience minimal human disturbance and have similar distributions of reproductive traits. In general, increased efforts to identify ploidy variation and recently formed polyploids in nature will be essential for testing the predictions of this study and other polyploid establishment models.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JS developed the model, analyzed model output, and prepared the manuscript. DS and PS provided critical revisions to the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank S. F. McDaniel, P. R. Muñoz, and E. B. Sessa for feedback and comments on the manuscript. This study also appears in JS's doctoral dissertation (Spoelhof, 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.592356/full#supplementary-material

REFERENCES

Abbott, R. J., and Lowe, A. J. (2004). Origins, establishment and evolution of new polyploid species: *Senecio cambrensis* and *S. eboracensis* in the British Isles. *Biol. J. Linnean Soc.* 82, 467–474. doi: 10.1111/j.1095-8312.2004.00333.x

Baack, E. J. (2005). To succeed globally, disperse locally: effects of local pollen and seed dispersal on tetraploid establishment. *Heredity* 94, 538–546. doi: 10.1038/sj.hdy.6800656

Barringer, B. C. (2007). Polyploidy and self-fertilization in flowering plants. *Am. J. Bot.* 94, 1527–1533. doi: 10.3732/ajb.94.9.1527

Chrtek, J., Herben, T., Rosenbaumová, R., Münzbergová, Z., Dočkalová, Z., Zahradnčèek, J., et al. (2017). Cytotype coexistence in the field cannot be explained by inter-cytotype hybridization alone: linking experiments and computer simulations in the sexual species *Pilosella echioides* (Asteraceae). *BMC Evol. Biol.* 17:87. doi: 10.1186/s12862-017-0934-y

Dar, P. A., Reshi, Z. A., and Shah, M. A. (2015). Roads act as corridors for the spread of alien plant species in the mountainous regions: a case study of Kashmir Valley, India. *Trop. Ecol.* 56, 183–190.

Fowler, N. L., and Levin, D. A. (2016). Critical factors in the establishment of allopolyploids. *Am. J. Bot.* 103, 1236–1251. doi: 10.3732/ajb.1500407

Gray, A. J., Marshall, D. F., and Raybould, A. F. (1991). "A century of evolution in *Spartina anglica*," in *Advances in Ecological Research*, Vol. 21, eds M. Begon, A. H. Fitter, and A. Macfadyen (London: Academic Press), 1–62. doi: 10.1016/S0065-2504(08)60096-3

Herben, T., Suda, J., and Klimešová, J. (2017). Polyploid species rely on vegetative reproduction more than diploids: a re-examination of the old hypothesis. *Ann. Bot.* 120, 341–349. doi: 10.1093/aob/mcx009

Hojsgaard, D., Greilhuber, J., Pellino, M., Paun, O., Sharbel, T. F., and Hörandl, E. (2014). Emergence of apospory and bypass of meiosis via apomixis after sexual hybridisation and polyploidisation. *New Phytol.* 204, 1000–1012. doi: 10.1111/nph.12954

Hojsgaard, D., and Hörandl, E. (2019). The rise of apomixis in natural plant populations. *Front. Plant Sci.* 10:358. doi: 10.3389/fpls.2019.00358

Hörandl, E., and Hojsgaard, D. (2012). The evolution of apomixis in angiosperms: a reappraisal. *Plant Biosyst. An Int. J. Deal. Aspects Plant Biol.* 146, 681–693. doi: 10.1080/11263504.2012.716795

Husband, B. C., Baldwin, S. J., and Suda, J. (2013). "The incidence of polyploidy in natural plant populations: major patterns and evolutionary processes," in *Plant Genome Diversity*, Vol. 2, eds I. J. Leitch, J. Greilhuber, J. Dolezel, and J. Wendel (Vienna: Springer), 255–276. doi: 10.1007/978-3-7091-1160-4_16

Joly, M., Bertrand, P., Gbangou, R. Y., White, M. C., Dubé, J., and Lavoie, C. (2011). Paving the way for invasive species: road type and the spread of common ragweed (*Ambrosia artemisiifolia*). *Environ. Manag.* 48, 514–522. doi: 10.1007/s00267-011-9711-7

Köhler, C., Scheid, O. M., and Erilova, A. (2010). The impact of the triploid block on the origin and evolution of polyploid plants. *Trends Genet.* 26, 142–148. doi: 10.1016/j.tig.2009.12.006

Kreiner, J. M., Kron, P., and Husband, B. C. (2017). Frequency and maintenance of unreduced gametes in natural plant populations: associations with reproductive mode, life history and genome size. *New Phytol.* 214, 879–889. doi: 10.1111/nph.14423

Levin, D. A. (1975). Minority cytotype exclusion in local plant populations. *Taxon* 24, 35–43. doi: 10.2307/1218997

Levin, D. A., and Kerster, H. W. (1974). "Gene flow in seed plants," in *Evolutionary Biology*, eds T. Dobzhansky, M. T. Hecht, and W. C. Steere (Boston, USA: Springer), 139–220. doi: 10.1007/978-1-4615-6944-2_5

Li, B. H., Xu, X. M., and Ridout, M. S. (2004). Modelling the establishment and spread of autotetraploid plants in a spatially heterogeneous environment. *J. Evol. Biol.* 17, 562–573. doi: 10.1111/j.1420-9101.2004.00700.x

Mable, B. K. (2004). Polyploidy and self-compatibility: is there an association? *New Phytol.* 162, 803–811. doi: 10.1111/j.1469-8137.2004.01055.x

Mason, A. S., and Pires, J. C. (2015). Unreduced gametes: meiotic mishap or evolutionary mechanism? *Trends Genet.* 31, 5–10. doi: 10.1016/j.tig.2014.09.011

Mortensen, D. A., Rauschert, E. S., Nord, A. N., and Jones, B. P. (2009). Forest roads facilitate the spread of invasive plants. *Invasive Plant Sci. Manag.* 2, 191–199. doi: 10.1614/IPSM-08-125.1

Novak, S. J., Soltis, D. E., and Soltis, P. S. (1991). Ownbey's Tragopogons: 40 years later. *Am. J. Bot.* 78, 1586–1600. doi: 10.1002/j.1537-2197.1991.tb11438.x

Oswald, B. P., and Nuismer, S. L. (2011). Unified model of autopolyploid establishment and evolution. *Am. Nat.* 178, 687–700. doi: 10.1086/662673

Ownbey, M. (1950). Natural hybridization and amphiploidy in the genus *Tragopogon*. *Am. J. Bot.* 37, 487–499. doi: 10.2307/2438023

Pandit, M. K., Pocock, M. J., and Kunin, W. E. (2011). Ploidy influences rarity and invasiveness in plants. *J. Ecol.* 99, 1108–1115. doi: 10.1111/j.1365-2745.2011.01838.x

Prentis, P. J., Wilson, J. R., Dormontt, E. E., Richardson, D. M., and Lowe, A. J. (2008). Adaptive evolution in invasive species. *Trends Plant Sci.* 13, 288–294. doi: 10.1016/j.tplants.2008.03.004

Pyšek, P., Jarošík, V., Pergl, J., Randall, R., Chytrý, M., Kühn, I., et al. (2009). The global invasion success of Central European plants is related to distribution characteristics in their native range and species traits. *Divers. Distribut.* 15, 891–903. doi: 10.1111/j.1472-4642.2009.00602.x

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Ann. Rev. Ecol. Syst.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Rausch, J. H., and Morgan, M. T. (2005). The effect of self-fertilization, inbreeding, depression, and population size on autopolyploid establishment. *Evolution* 59, 1867–1875. doi: 10.1111/j.0014-3820.2005.tb01057.x

Robertson, K., Goldberg, E. E., and Igić, B. (2011). Comparative evidence for the correlated evolution of polyploidy and self-compatibility in Solanaceae. *Evolution* 65, 139–155. doi: 10.1111/j.1558-5646.2010.01099.x

Rodriguez, D. J. (1996). A model for the establishment of polyploidy in plants. *Am. Nat.* 147, 33–46. doi: 10.1086/285838

Soltis, D. E., and Soltis, P. S. (1999). Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.* 14, 348–352. doi: 10.1016/S0169-5347(99)01638-9

Spoelhof, J. P. (2020). *Neo to Paleo: Examining Factors that Affect Polyploid Success Across Timescales*. Dissertation, University of Florida, Gainesville, FL.

Spoelhof, J. P., Keeffe, R., and McDaniel, S. F. (2019). Does reproductive assurance explain the incidence of polyploidy in plants and animals? *New Phytol.* 227, 14–21. doi: 10.1111/nph.16396

Stebbins, G. L. (1938). Cytological characteristics associated with the different growth habits in the dicotyledons. *Am. J. Bot.* 25, 189–198. doi: 10.2307/2436589

Te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubešová, M., et al. (2012). The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.* 109, 19–45. doi: 10.1093/aob/mcr277

Van Drunen, W. E., and Husband, B. C. (2018). Whole genome duplication decreases clonal stolon production and genet size in the wild strawberry *Fragaria vesca*. *Am. J. Bot.* 105, 1712–1724. doi: 10.1002/ajb2.1159

Van Drunen, W. E., and Husband, B. C. (2019). Evolutionary associations between polyploidy, clonal reproduction, and perenniality in the angiosperms. *New Phytol.* 224, 1266–1277. doi: 10.1111/nph.15999

Zenil-Ferguson, R., Burleigh, J. G., Freyman, W. A., Igić, B., Mayrose, I., and Goldberg, E. E. (2019). Interaction among ploidy, breeding system and lineage diversification. *New Phytol.* 224, 1252–1265. doi: 10.1111/nph.16184

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 09 December 2020
doi: 10.3389/fpls.2020.591137

Intricate Distribution Patterns of Six Cytotypes of *Allium oleraceum* at a Continental Scale: Niche Expansion and Innovation Followed by Niche Contraction With Increasing Ploidy Level

Martin Duchoslav[1]*, *Michaela Jandová*[1,2], *Lucie Kobrlová*[1], *Lenka Šafářová*[1], *Jan Brus*[3] *and Kateřina Vojtěchová*[1]

[1] *Plant Biosystematics and Ecology RG, Department of Botany, Faculty of Science, Palacký University, Olomouc, Czechia,* [2] *Institute of Botany, Czech Academy of Sciences, Pruhonice, Czechia,* [3] *Department of Geoinformatics, Faculty of Science, Palacký University, Olomouc, Czechia*

OPEN ACCESS

Edited by:
Karol Marhold,
Slovak Academy of Sciences, Slovakia

Reviewed by:
Karl Hülber,
University of Vienna, Austria
Božo Frajman,
University of Innsbruck, Austria

***Correspondence:**
Martin Duchoslav
martin.duchoslav@upol.cz

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *03 August 2020*
Accepted: *06 November 2020*
Published: *09 December 2020*

Citation:
Duchoslav M, Jandová M, Kobrlová L, Šafářová L, Brus J and Vojtěchová K (2020) Intricate Distribution Patterns of Six Cytotypes of Allium oleraceum at a Continental Scale: Niche Expansion and Innovation Followed by Niche Contraction With Increasing Ploidy Level. Front. Plant Sci. 11:591137. doi: 10.3389/fpls.2020.591137

The establishment and success of polyploids are thought to often be facilitated by ecological niche differentiation from diploids. Unfortunately, most studies compared diploids and polyploids, ignoring variation in ploidy level in polyploids. To fill this gap, we performed a large-scale study of 11,163 samples from 1,283 populations of the polyploid perennial geophyte *Allium oleraceum* with reported mixed-ploidy populations, revealed distribution ranges of cytotypes, assessed their niches and explored the pattern of niche change with increasing ploidy level. Altogether, six ploidy levels (3*x*−8*x*) were identified. The most common were pentaploids (53.6%) followed by hexaploids (22.7%) and tetraploids (21.6%). Higher cytotype diversity was found at lower latitudes than at higher latitudes (>52° N), where only tetraploids and pentaploids occurred. We detected 17.4% of mixed-ploidy populations, usually as a combination of two, rarely of three, cytotypes. The majority of mixed-ploidy populations were found in zones of sympatry of the participating cytotypes, suggesting they have arisen through migration (secondary contact zone). Using coarse-grained variables (climate, soil), we found evidence of both niche expansion and innovation in tetraploids related to triploids, whereas higher ploidy levels showed almost zero niche expansion, but a trend of increased niche unfilling of tetraploids. Niche unfilling in higher ploidy levels was caused by a contraction of niche envelopes toward lower continentality of the climate and resulted in a gradual decrease of niche breadth and a gradual shift in niche optima. Field-recorded data indicated wide habitat breadth of tetraploids and pentaploids, but also a pattern of increasing synanthropy in higher ploidy levels. Wide niche breadth of tetra- and pentaploids might be related to their multiple origins from different environmental conditions, higher "age", and retained sexuality, which likely preserve their adaptive potential. In contrast, other cytotypes with narrower niches are mostly asexual, probably originating from a limited range of contrasting environments. Persistence of local ploidy mixtures could be enabled by the perenniality of *A. oleraceum* and its prevalence

of vegetative reproduction, facilitating the establishment and decreasing exclusion of minority cytotype due to its reproductive costs. Vegetative reproduction might also significantly accelerate colonization of new areas, including recolonization of previously glaciated areas.

Keywords: cytogeography, chromosome numbers, ecological niche, flow cytometry, geophytes, ploidy coexistence, polyploidy

INTRODUCTION

Polyploidy is a widespread phenomenon among flowering plants (e.g., Wendel, 2000; Van de Peer et al., 2017). Several lines of evidence suggest that all flowering plants have experienced several polyploid events at some points in their ancestry (Wood et al., 2009; Jiao et al., 2011), and polyploidization is also an active evolutionary process in many lineages (Soltis and Soltis, 2012; Soltis et al., 2016; Levin, 2020). Before becoming evolutionarily successful, newly formed polyploids often have to overcome numerical inferiority, mating incompatibility, and competition with parents (Levin, 1975). Polyploidization has profound consequences for the physiological and ecological behavior of plants (Levin, 2002; Soltis et al., 2004; Ramsey, 2011; Ramsey and Ramsey, 2014) as well as their genetic diversity (Soltis and Soltis, 2000). Interactions between nucleotypic effect (Bennett and Smith, 1972), increased genetic buffering, and changes in gene expression in polyploids (Adams and Wendel, 2005; Yoo et al., 2014; Gallagher et al., 2016) may drive phenotypic changes which not only immediately affect the ecology of polyploids (Ramsey, 2011), but also their potential for novel adaptive responses to selection (Bretagnole and Thompson, 1996; Otto and Whitton, 2000; Levin, 2002; Soltis et al., 2004; Balao et al., 2011; Visger et al., 2016). Niche separation is highly important in the establishment and further spread of neopolyploids (Fowler and Levin, 2016), decreasing intercytotype reproductive contacts and competition, thus increasing neopolyploid population growth (Ramsey and Schemske, 1998; Rieseberg and Willis, 2007). Where (two) divergent genomes merge, allopolyploids may exhibit even greater rates of population establishment, persistence, and exploitation of novel habitats than autopolyploids (Arrigo et al., 2016; Barker et al., 2016; Solhaug et al., 2016) due to exceeding parental niches caused by great flexibility in gene expression (Doyle et al., 2008; Leitch and Leitch, 2008; Yoo et al., 2014). Indeed, it has been evidenced repeatedly that polyploids (prevalently tetraploids) have a broader niche or may differ in their niche optima from their diploid progenitors (Soltis and Soltis, 1995, 2000; Levin, 2002; Weiss-Schneeweiss et al., 2013; Ramsey and Ramsey, 2014; Baniaga et al., 2020). They are reported to have a stronger colonization ability (Treier et al., 2009), including invasion potential (Pandit et al., 2006, 2011), and increased ability to cope with environmental extremes better, especially at higher latitudes and elevations, and in arid or artificially disturbed habitats (Grant, 1981; Stebbins, 1984; Brochmann et al., 2004; Wu et al., 2010; Ramsey, 2011; Manzaneda et al., 2012; te Beest et al., 2012; Muñoz-Pajares et al., 2018; Rice et al., 2019; Castro et al., 2020; Decanter et al., 2020).

The study of niche changes has recently made rapid progress incorporating readily available large-scale climatic data (Fick and Hijmans, 2017) and developed new statistical tools (Warren et al., 2008; Broennimann et al., 2012; Guisan et al., 2014). Several studies examining ecological (usually climatic) differentiation between polyploids and their progenitors within groups of closely related taxa do, however, not always support niche innovation in polyploids (e.g., Broennimann et al., 2007; Godsoe et al., 2013; Theodoridis et al., 2013; Glennon et al., 2014; Arrigo et al., 2016; Marchant et al., 2016). These results suggest substantial controversy in this area, which might be explainable partly by methodological incongruities between studies due to the use of a too coarse resolution of studied abiotic variables (Kirchheimer et al., 2016), ignorance of other axes of a taxon's niche (Guisan et al., 2014; Guignard et al., 2016; Segraves and Anneberg, 2016) or absence of assessments of niche differentiation of cytotypes across more spatial scales (Treier et al., 2009; Laport et al., 2013; Čertner et al., 2019). Apart from the methodological issues, niche evolution in polyploids may also be affected by an ancestral niche breadth of their progenitors and "age", order or number of cytotype origins within polyploid series (Theodoridis et al., 2013; López-Jurado et al., 2019). Higher ploidy levels, in contrast to lower ones, may show a lower probability of niche expansion (Brittingham et al., 2018; López-Jurado et al., 2019) because they may attain their ecological limits during the filling of remaining available unoccupied niche space (Araújo et al., 2013) and their larger genome sizes might constrain their adaptive ability (Pandit et al., 2014). On the other hand, multiple origins of polyploids (Soltis and Soltis, 1999) might increase the genetic and physiological diversity of cytotypes which subsequently enhance their ecological tolerances (McIntyre, 2012; Karunarathne et al., 2018; López-Jurado et al., 2019).

Existence of diverse global and local distribution patterns of cytotypes within polyploid complexes, ranging from sympatry through more common parapatry with cytotype-mixed populations over contact zones to allopatry (Lewis, 1980; reviewed by Stebbins, 1985; Thompson and Lumaret, 1992; Petit et al., 1999; Levin, 2002; Martin and Husband, 2009; Husband et al., 2013; Kolář et al., 2017) suggest that cytotype distributions are the results of complex processes and that certain distribution patterns (e.g., local sympatry, large-scale allopatry) can be generated by processes other than niche evolution (Šingliarová et al., 2019; Wos et al., 2019) or that some of these processes (e.g., reproductive isolation) could reinforce niche differentiation between cytotypes (Rausch and Morgan, 2005; Rojas-Andrés et al., 2020). Modification of the reproductive system toward autogamy, apomixis, or vegetative propagation

to secure reproduction in polyploids (Barringer, 2007; Paule et al., 2011; Herben et al., 2017) can be a key advantage in their local establishment irrespective of their initial minority status (Kao, 2007, 2008). Asexual reproduction drastically affects the dispersal abilities of polyploids, allowing their rapid colonization and establishment in new areas, which are particularly important when resource competition with an ancestor excludes them from local populations (Baker, 1967; Kearney, 2005; Kirchheimer et al., 2016). If residual sexuality is retained in the polyploid, it also preserves its adaptive potential (Cosendai et al., 2013). However, the same distributional patterns may also be explained as the results of stochastic processes, e.g., founder and drift effects (Lewis, 1967; Kliber and Eckert, 2005; Kolář et al., 2016). The present distribution of cytotypes may also mirror the position of past cytotype refuges in relation to sites available for colonization by single cytotypes, e.g., after ice retreat during the Quaternary (van Dijk et al., 1992; Mandáková and Münzbergová, 2006; Godsoe et al., 2013).

With the application of flow cytometry (FCM), which allows us to order more sample analyses, it has also become clear that mixed-ploidy populations are much more frequent than previously anticipated ones (e.g., Halverson et al., 2008; Cires et al., 2010; Marhold et al., 2010; Trávníček et al., 2010, 2011a,b; Čertner et al., 2017, 2019). In the past, cases of locally coexisting cytotypes were assumed to represent transient situations following the frequent generation or, in the case of secondary contacts, immigration of an alternative cytotype (Kao, 2007). However, commonly occurring cytotype-mixed populations intermixed with single-cytotype populations extending over large areas have been detected in several species (Husband et al., 2013; Kolář et al., 2017 and references therein). The knowledge of cytotype composition within mixed-cytotype populations and their spatial context with cytotype-uniform populations could serve as a basis for addressing questions of the frequency of polyploid formation, niche differentiation of cytotypes, and polyploid evolution (Husband et al., 2013). However, only little is known about many polyploid complexes because they have not been studied systematically throughout their entire distribution range (Afonso et al., 2020).

An example of a mixed-ploidy plant with a complex cytogeographic pattern at various spatial scales is *Allium oleraceum* L., consisting of tri-, tetra-, penta-, hexa-, hepta-, and octoploid cytotypes ($2n = 24, 32, 40, 48, 56, 64$) of presumably allopolyploid origin (Duchoslav et al., 2010, 2013). Detailed screening of its cytotype distribution in Central Europe has demonstrated complex spatial patterns, ranging from parapatry to sympatry at the landscape scale, with frequent occurrence of mixed-ploidy populations representing mostly secondary contacts between cytotypes (Duchoslav et al., 2010; Šafářová and Duchoslav, 2010; Šafářová et al., 2011), This complex cytogeographic pattern was explained by the interaction of several mechanisms, including (i) observational (Duchoslav et al., 2010; Šafářová et al., 2011) and experimental evidence (Ježilová et al., 2015; Duchoslav et al., 2017) for slight niche differentiation among cytotypes, (ii) the prevalence of asexual reproduction via aerial bulbils (Fialová and Duchoslav, 2014; Fialová et al., 2014), facilitating founder events and escape from minority cytotype exclusion effect (Levin, 1975) at mixed-cytotype sites, supported also by (iii) unstable environmental conditions at many sites caused by periodic disturbances and local patterns of migration in strongly human-influenced landscapes (Duchoslav et al., 2010). However, there are no robust data on cytotype diversity, distribution patterns and niche differentiation at a large spatial (continental) scale which could differ from those at a regional scale due to an interaction of large-scale (climate) and small-scale environmental factors (Treier et al., 2009; Laport et al., 2013; Čertner et al., 2019) and/or historical causes (te Beest et al., 2012). *Allium oleraceum* is distributed (Meusel et al., 1965) from the northern part of the Mediterranean basin, a refugium of thermophilic and mesic plant species during Glacial periods (Médail and Diadema, 2009), where it is currently in contact with its supposed diploid progenitors of the *Allium paniculatum* L. group (Pastor and Valdés, 1983; Brullo et al., 1996a; Salmeri et al., 2016), to Northern Europe, which was glaciated during the Last Glacial Maximum (LGM; Vandenberghe et al., 2014). Therefore, the current distributional pattern of cytotypes may also have been affected by the interaction between the effects of Pleistocene climatic oscillations, which had a profound effect on range expansions, retraction, and melting of differentiated lineages in species of the European flora (Hewitt, 1999; Stewart and Lister, 2001; Stewart et al., 2010), on the one hand and ecological attributes of cytotypes on the other. Knowledge of cytotype composition of populations and cytotype distribution, especially over contact zones between *A. oleraceum* and its presupposed progenitors, may also allow for inferences about the evolutionary history of polyploidy in *A. oleraceum*.

Specifically, we addressed the following questions: (1) What is the diversity of *A. oleraceum* cytotypes and the pattern of their geographic distribution over the species range? Is there a different cytotype composition in the contact zones with presumed progenitors in comparison to northern, previously glaciated regions without such contacts? (2) Is the current cytotype distribution a consequence of niche divergence? Is there a different pattern of niche shift with increasing ploidy level? (3) How frequent are mixed-ploidy populations, which cytotypes participate in their composition, and what processes stand behind their existence?

MATERIALS AND METHODS

Studied Species

Allium oleraceum is a geophyte with prolific asexual propagation by aerial bulbils formed within an inflorescence and partly also by daughter bulbs produced belowground by the mother bulb, while sexual seeds are formed less frequently and usually in small numbers (Fialová and Duchoslav, 2014; Fialová et al., 2014). It belongs to *Allium* section *Codonoprasum* Reichenb., an evolutionarily young group consisting of a set of diploid and polyploid species (Friesen et al., 2006). This section is distributed in Northern Africa and Europe, extending to Iran and southwestern Siberia (Vvedenskii, 1935; Meusel et al., 1965; Stearn, 1980; Brullo et al., 1996a,b; Brullo et al., 2001). *Allium oleraceum* shows morphological similarity to species of the informal *A. paniculatum* complex, reaching its northern range

limit in the southern parts of Europe (Stearn, 1980; Pastor and Valdés, 1983; Brullo et al., 1996a, 1997, 2003; Dobrotchaeva et al., 1999; Ciocârlan, 2000; Jauzein and Tison, 2001; Bogdanović et al., 2008; Aedo, 2013; Tison and de Foucault, 2014; Ghendov, 2015; Salmeri et al., 2016; Brullo and Guarino, 2017), and characterized by plants with ribbed and glabrous leaves with a semicylindrical to flat outline, spathe valves with a long appendage, and a campanulate perigon with stamens included or just slightly exerted (Brullo et al., 1996a,b, 2001, 2003, 2008; Salmeri et al., 2016). The origin of *A. oleraceum* is still puzzling; nevertheless, an alloploid origin is the most probable (Levan, 1937; Vosa, 1976; Duchoslav et al., 2010).

Plant Material and Field Sampling

For the chromosome number survey, data on chromosome counts of *A. oleraceum* was extracted from 33 publications resulting in 399 localities (**Supplementary Table 1**, including complete reference list). The dataset was supplemented with unpublished chromosome data of 20 populations sampled by Fialová (1996) throughout the Czech Republic. Unfortunately, only limited information concerning the number of analyzed individuals and habitat conditions of sampled sites was available from most published papers. For this reason, one individual per ploidy was considered within a population.

Plant material was sampled from 2004 to 2019 throughout Europe covering most of the range of *A. oleraceum*. Additional samples of seeds or bulbils originating from 20 natural populations were obtained via "Index seminum" from foreign botanical gardens. In total, material from 446 populations of *A. oleraceum* was collected. Simultaneously, data from the detailed ploidy-level screening of 418 populations of the Czech Republic and Slovakia (Duchoslav et al., 2010; Šafářová and Duchoslav, 2010; Šafářová et al., 2011) was extracted and incorporated into subsequent analyses. Altogether, data on the cytotype composition of 1,283 populations was collected (**Supplementary Table 1**).

Field sampling followed our previous studies (Duchoslav et al., 2010; Šafářová and Duchoslav, 2010; Šafářová et al., 2011) allowing for joint analyses of all datasets. Specifically, sampling was usually carried out during the spring season (March to early June) when both non-flowering and flowering plants are present aboveground (Duchoslav, 2009), and plants were collected from a wide spectrum of habitats to attain a maximum representation of the species' niche. Each population sample consisted of mostly 3–30 plants depending on population size. An effort was made to avoid collection of individuals growing close together and to cover the entire population. During sampling, the sampling area expressed in square meters was estimated. Plants were cultivated in the experimental garden of Palacký University Olomouc, Czech Republic. Herbarium specimens are deposited in the Herbarium of Palacký University Olomouc (OL).

Flow Cytometry and Chromosome Counts

DNA ploidy levels (Suda et al., 2006) were determined using FCM with *Triticum aestivum* cv. *Saxana* (2C DNA = 34.24 pg, Šafářová and Duchoslav, 2010) as an internal standard. Measurements were conducted on the following flow cytometers using two different fluorochromes: (i) BD Accuri C6 (BD Biosciences, San Jose, CA, USA)—propidium iodide (PI); (ii) Partec PAS (Partec GmbH, Münster, Germany)—PI and DAPI (4,6-diamidino-2-phenylindole); (iii) Partec Cy Flow ML (Partec GmbH)—PI and DAPI staining. Samples were prepared following the simplified protocol with LB01 isolation buffer and stained with either PI or DAPI (Doležel et al., 2007). Fresh leaves or meristematic buds of bulbil were used for FCM. For each run, the fluorescence intensity of at least 3,000 particles was recorded. Mostly, each individual was analyzed separately, but sometimes 2–3 individuals from the same population were analyzed together. Generally, histograms (using both PI and DAPI) with a coefficient of variation (CV) <5.0% were accepted. The ploidy level of the sample was determined by the position of its G_0/G_1 peak relative to the G_0/G_1 peak of an internal standard (i.e., relative fluorescence). Ploidy level was assessed based on calibration using plants for which chromosome numbers were previously (Duchoslav et al., 2010, 2013) as well as newly counted, covering the studied geographic range. Chromosome numbers were counted following the protocol by Duchoslav et al. (2010). Several individuals were analyzed with both PI and DAPI to calibrate the position of peaks if a different dye was used. Several samples with a DNA content on the margin of variation of the respective ploidy level were additionally subjected to chromosome counting to detect possible aneuploidy.

Coarse-Grained Environmental Data Extracted From Climatic and Soil Layers

Environmental data related to different eco-physiological constraints of the studied species were selected and downloaded from various open-source databases. Since climate is often seen as the main factor driving species distributions at large scales (Guisan and Thuiller, 2005; Rice et al., 2019), annual trends and extreme limiting conditions related to precipitation and temperature (BIO 1-19 variables) and elevation were extracted from WorldClim 2.1 (Fick and Hijmans, 2017). Also mean annual solar radiation ($kW.m^{-2}$) was downloaded from this database. Two variables related to evapotranspiration processes and rainfall deficit for potential vegetative growth (Global Aridity Index and Potential Evapotranspiration) were downloaded from Global Aridity Index and Potential Evapotranspiration Climate Database v2 (Trabucco and Zomer, 2019). Since edaphic conditions might play an important part in a polyploid's niche (Šmarda et al., 2013; Guignard et al., 2016), available quantitative physical and chemical soil variables were downloaded from the SoilGrid database (Hengl et al., 2014, 2017). All downloaded variables had a resolution of 30 arcseconds (∼1 km).

Since collinearity is a common feature in any descriptive ecological data set and can be a problem for parameter estimation potentially leading to the wrong identification of relevant predictors in a statistical model (Dormann et al., 2013),

all downloaded environmental variables were examined for pairwise correlations based on extracted values of environmental variables for all established locations of *A. oleraceum*. After evaluation, 13 not highly correlated variables (Pearson's correlations ≤ 0.70) were retained and used in further analyses (**Supplementary Table 2**).

Field-Recorded, Local-Scale Environmental Data

At each sampled site the following set of local environmental variables was recorded, adopting and revising those used in the previous study by Duchoslav et al. (2010): (i) Habitat type was assessed in the field and classified using the recent vegetation classification of Europe consistent with the Braun-Blanquet approach (Mucina et al., 2016). Because of the large geographic range and consecutively great vegetation diversity of the sampled area, ecologically similar vegetation types were merged into ten aggregated habitat types (rocky outcrop, alpine grassland, dry grassland, mesic grassland, seminatural dry forest, seminatural mesic forest, seminatural alluvial forest, shrub, *Robinia pseudacacia* forest, arable land & field margin), which were used for subsequent analyses ("habitat type"). Correspondence between aggregated habitat types and observed vegetation types is explained in the **Supplementary Table 3**; (ii) Habitat naturalness, i.e., the vegetation at the site was classified according to decreasing degree of anthropic impact (synanthropy) into one of three levels: (1) "highly human-affected" (HHA; vegetation strongly influenced by man, typically with a relatively high representation or dominance of ruderal species, e.g., intensively managed agricultural habitats and their margins, road ditches, urbanized areas), (2) "extensively human-affected" (EHA; extensively cultivated landscapes, cultivated and plantation-like forests and shrubs, partly synanthropic, and extensively cultivated locations) and (3) "natural" (NAT; natural and seminatural vegetation without strong anthropic influence, e.g., semi-natural forests, semi-natural grasslands, relict sites). (iii) Habitat heterogeneity, i.e., the number of aggregated habitat types inhabited by the local *Allium* population. (iv) Presence of arable land: the effect of agricultural practices on the potential dispersion of *Allium* propagules into a site was approximated by the distance of the population to the nearest arable field and transformed into two categories ($0 = >20$ m; $1 = \leq 20$ m). (v) Rock proximity: populations were classified into two categories according to their proximity to rocky outcrops as follows: 0/1 = absence/presence of rocky outcrops within circle of 100 m radius around the population. (vi) Light conditions were assessed in the field according to the visually estimated proportion of full sunlight falling on the ground during late spring according to the following ordinal scale: 1 = strong shade, 2 = half-shade, 3 = light shade, 4 = full insolation. (vii) Heat load, i.e., a unitless index estimating the amount of heat absorbed by local site from solar radiation. The heat load was calculated from the slope, aspect and latitude of the site, using equation 2 in McCune and Keon (2002). The final values were converted to an arithmetic scale with the exp(x) function. (viii) Elevation, recorded *in situ* with a GPS instrument.

Calculation of Niche Characteristics in Environmental Space (E-Space) and Local-Scale Environmental Differences Between Cytotypes

Niche characteristics of the cytotypes were estimated with two groups of variables: (i) coarse-grained environmental (climatic and soil) data downloaded from open-source databases and (ii) field-recorded, local-scale environmental data. As the first step, georeferenced location data were spatially stratified to avoid discrepancies caused by highly unequal sampling in different parts of the species range, using R package "spThin" (Aiello-Lammens et al., 2015). Since octoploids were only found at two sites, these were excluded from all ecological analyses. Considering the extremely dense sampling in Central Europe (Czech Republic, Slovakia; Duchoslav et al., 2010; Šafářová et al., 2011), Finland (Åström et al., 2015), and Latvia (Karpavičienė, 2012), a 40-km threshold distance was used for each cytotype found in those regions. In the remaining parts of Europe, a 15-km threshold distance was used for each cytotype. This resulted in 560 localities, which were used for subsequent ecological analyses (**Supplementary Table 1**).

Environmental niche space occupied by each cytotype and quantification of niche overlap, equivalence, and similarity between cytotypes were accessed with an ordination technique (PCA-env) which applies kernel smoothers to the cytotype presences in environmental space for the selection, combination, and weighting of environmental variables (Broennimann et al., 2012). As to environmental data, 13 environmental variables from climatic and soil open-source databases were used (**Supplementary Table 2**). We specified a division of the environmental space of PCA (the first two axes) into a grid of 200 $\times$ 200 cells, in which each cell corresponds to a unique vector of the available environmental conditions in the background area, i.e., available environmental conditions from which each cytotype is presumed to select its habitat. It is recommended that the background area is taken from buffer zones around known occurrences or from range maps (McCormack et al., 2010; Warren et al., 2010). Therefore, we used a 20-km buffer zone around the occurrence points of each cytotype for the definition of a background area. The number of background (random) points per cytotype equalled a hundred-fold of the number of the respective occurrence points.

The niche overlap between each pair of cytotypes was computed employing the Schoener's D statistic (D) directly from environmental niche space (Schoener, 1968; Warren et al., 2008). The value of D ranges from 0, when two cytotypes have no overlap in the environmental space, to 1, when two cytotypes share the same environmental space. For testing niche conservatism vs. evolution, the niche equivalency (identity) test and the niche similarity test (Warren et al., 2008) were computed for each pair of cytotypes. The niche equivalency test determines whether niches of two cytotypes are equivalent, i.e., whether the niche overlap is constant when randomly reallocating cytotype identities over compared cytotype ranges (Broennimann et al., 2012). Specifically, occurrences of compared cytotypes were pooled and randomly split into two datasets, with the same

sizes as the two original datasets, after which D was calculated. This process was repeated 1,000 times and generated a null distribution of D (Warren et al., 2008). We determined the non-equivalence or conservatism of environmental niches if the observed D of the cytotypes being compared were within the lower or upper 2.5% quantile, respectively, of a null distribution of simulated values of D.

The niche similarity test establishes whether the observed niche overlap between two cytotypes is different from the overlap between the observed niche of one cytotype and niches randomly selected from the available environmental background of the other cytotype within the environmental space defined by PCA axes (Broennimann et al., 2012). Specifically, points were randomly selected from the background area of one cytotype and the niche of this random sample was then compared to the observed niche of another cytotype using D. This process was repeated 1,000 times and generated a null distribution of D (Warren et al., 2008). If the observed D was within the upper or lower 2.5% quantile of a null distribution of simulated values of D, the cytotypes were considered to be more similar or more dissimilar, respectively, than expected by chance. This test was applied in both directions, i.e., by resampling the occurrences of the first cytotype and then those of the other cytotype.

Schoener's D statistic does not contain information on optima and breadths of niches (Glennon et al., 2014). Therefore, we used the procedure described in Theodoridis et al. (2013) and Kirchheimer et al. (2016) for comparing niches in terms of optima and breadths. Specifically, 100 cells were randomly resampled in the niche of each cytotype and their scores were extracted along the first two PCA axes. The niche optimum and the niche breadth were calculated as the mean and the variance of the sampled scores along the first and the second PCA axes. This procedure was repeated 1,000 times. Distributions of values of niche optimum and breadth for each PCA axis were compared between cytotypes.

Additionally, the following indices of niche change were computed relative to the niche of lower ploidy level only (Petitpierre et al., 2012; Guisan et al., 2014): niche expansion (E), i.e., proportion of the niche space of the higher ploidy level non-overlapping the niche of the lower ploidy level; niche unfilling (U), i.e., proportion of the niche space of the lower ploidy level non-overlapping the niche of the higher ploidy level; and niche stability (S_n, S_e), i.e., proportion of the niche of either lower (S_n) or higher ploidy level (S_e), shared with the other ploidy level. To decrease the effect of rare (marginal, extreme) environments on the estimation of the indices of niche change, analyses were performed in two settings, i.e., at the intersection of the 75th quantile and 95th quantile, respectively, of both compared environmental densities. All analyses were performed in the R platform (R Development Core Team, 2014), using the packages ecospat (Di Cola et al., 2017), raster (Hijmans et al., 2017), and ENMTools (Warren et al., 2010).

Ecological differentiation among cytotypes was also tested using eight field-recorded, local-scale environmental variables. Firstly, each environmental variable was used as the dependent variable and the ploidy level (with five cytotypes, $3x-7x$) was used as the independent variable in separate univariate analyses. Mixed populations were duplicated according to the cytotype composition and duplicates were assigned in each group of participating cytotypes. Log-linear models were used for the analyses of categorical environmental variables, whereas the non-parametric Kruskal-Wallis test followed by multiple comparisons Dunn's test was used for the analyses of quantitative and ordinal variables (Zar, 1996). Subsequently, eight field-recorded environmental variables were subjected to constrained principal coordinate analysis (db-RDA; Legendre and Anderson, 1999) using the Gower coefficient of dissimilarity for mixed numeric/ordinal/categorical data (Legendre and Legendre, 1998). Due to the poor habitat description of 72 sites analyzed for niche characteristics in E-space, these sites were excluded from db-RDA, resulting in 488 analyzed populations (**Supplementary Table 1**). Cytotype composition of populations (**Table 2**) was used as the explanatory variable and visualized in an ordination diagram. Additionally, pairwise tests between cytotypes for environmental differences were performed using a reduced set of data matrices only consisting of populations of the respective pairs of cytotypes. Mixed populations were duplicated according to cytotype composition and duplicates were assigned to each group of participating cytotypes. Environmental differences between groups with different cytotype compositions were tested with a Monte Carlo permutation test using 999 permutations. The Bonferroni correction of α (at $\alpha = 0.05$) for multiple tests was applied. Univariate analyses were performed using NCSS 9 (Hintze, 2013), whereas multivariate analyses were performed using CANOCO 5 (ter Braak and Šmilauer, 2012).

RESULTS

Cytotype Diversity and Population Composition

Six ploidy levels ranging from tri- to octoploid ($2n = 24$, 32, 40, 48, 56, 64) were identified in our dataset using chromosome counting (**Supplementary Figure 1**). This was fairly consistent with the FCM analysis of 3,931 individuals from 446 European populations, newly analyzed in this study. Relative fluorescence of different cytotypes differed [Linear Models, PI dye: $F_{(5,6.9)} = 1630.5$, $P < 0.001$; DAPI dye: $F_{(2,4.7)} = 97.2$, $P < 0.001$; **Table 1**]. Large intracytotypic variation in relative fluorescence was observed in tetra- and pentaploids. This resulted in an almost continuous distribution of relative fluorescence values between tetra- and pentaploids (**Supplementary Figure 2**). However, both chromosome counts of several plants with marginal values of relative fluorescence (**Supplementary Table 1**) and spatial correlation of relative fluorescence values in tetra- and pentaploids (**Supplementary Figure 2**; see Duchoslav et al., 2013 for the identical pattern found for absolute DNA content) allowed us to infer the ploidy level of such plants with certainty. Moreover, no aneuploid counts were identified, not even in plants with marginal relative fluorescence values within particular cytotypes. This is in line with published data on karyologically examined plants from 399 localities (**Supplementary Table 1**) where only euploid chromosome counts ($2n = 3x$, $4x$, $5x$, $6x$) were reported. The ploidy evaluation of a total of 11,163

TABLE 1 | Relative fluorescence (RF) of the *A. oleraceum* cytotypes assessed using flow cytometry with the stain propidium iodide (PI) or 4,6-diamidino-2-phenylindole (DAPI).

Ploidy level	PI							DAPI						
	Npop	Nind	Mean RF	SD	Min	Max	Variation (%)	Npop	Nind	Mean RF	SD	Min	Max	Variation (%)
3*x*	16	132	1.25[a]	0.03	1.19	1.29	8.4	–	–	–	–	–	–	–
4*x*	134	2,087	1.56[b]	0.08	1.41	1.71	21.3	6	29	2.45[a]	0.09	2.36	2.59	9.7
5*x*	345	4,249	1.84[c]	0.05	1.66	1.98	19.3	36	286	2.84[b]	0.10	2.51	3.04	21.1
6*x*	89	1,858	2.10[d]	0.05	2.01	2.24	11.4	3	15	3.20[c]	0.06	3.16	3.27	3.4
7*x*	10	87	2.46[e]	0.06	2.36	2.52	6.8	–	–	–	–	–	–	–
8*x*	2	23	2.65[f]	0.05	2.62	2.69	2.7	–	–	–	–	–	–	–

*All values of relative fluorescence are calculated relative to the internal standard (Triticum aestivum cv. Saxana). Mean RF values within a column with different letters are significantly different at P ≤ 0.05 (Bonferroni multiple comparison test). All available data of RF measured by us (**Supplementary Table 1**) were used for calculations. Npop, number of populations analyzed; Nind, number of individuals analyzed; SD, standard deviation; Variation (%) = (Max-Min) × 100/Min.*

TABLE 2 | Cytotype composition of 858 populations of *Allium oleraceum* with at least two sampled individuals, and odds of uniform-ploidy populations (Odds), expressed as the ratio of the number of uniform-ploidy populations to the number of mixed-ploidy populations containing a particular cytotype.

Population cytotype composition (Odds)	Number of populations	Percent of total
3*x* (4.0)	12	1.40%
4*x* (1.6)	114	13.28%
5*x* (3.3)	450	52.45%
6*x* (1.3)	123	14.33%
7*x* (4.5)	9	1.05%
8*x* (1.0)	1	0.12%
3*x* + 4*x*	1	0.12%
3*x* + 5*x*	2	0.23%
4*x* + 5*x*	51	5.94%
4*x* + 6*x*	10	1.17%
4*x* + 7*x*	1	0.12%
4*x* + 8*x*	1	0.12%
5*x* + 6*x*	74	8.62%
5*x* + 7*x*	1	0.12%
4*x* + 5*x* + 6*x*	8	0.93%

Populations are grouped according to the number of cytotypes per population (one, two, or three). Frequencies are shown both in absolute numbers and in percentages of the total number of populations.

individuals from 1,283 populations (incl. all previously published data; **Supplementary Table 1**) revealed that the most common cytotype was pentaploid (53.6%) followed by hexaploid (22.7%) and tetraploid (21.6%). Other ploidy levels, i.e., triploids (1.2%), heptaploids (0.8%), and octoploids (0.2%), were extremely rare.

Considering populations with at least two individuals analyzed (858 populations, 97.1% of all populations sampled by us; the number of analyzed plants per population: mean ± SD, 13 ± 14, min. = 2, max. = 213), a high diversity in cytotype composition was found within populations (**Table 2**). The majority of populations comprised a single cytotype (83.6%), much fewer populations contained two (16.5%), and just eight populations (0.9%) contained three cytotypes. Among populations consisting of a single ploidy level, pentaploids were found to be the most common while octoploids were the rarest (**Table 2**).

More individuals were analyzed from populations revealed as mixed-ploidy (median = 15) than uniform-ploidy (median = 8) by FCM (Mann-Whitney one-sided test, $z = 8.9$, $P < 0.001$). Also a sampling area of mixed-ploidy populations (median = 300 m^2) was larger than that of uniform-ploidy populations (median = 80 m^2; $z = 7.5$, $P < 0.001$). However, habitat heterogeneity was similar between uniform and mixed-ploidy populations ($z = 1.1$, $P = 0.286$) and most populations inhabited a single habitat.

The proportion of mixed-ploidy populations was cytotype-dependent. Populations containing odd ploidy level(s) were uniform two to three times more likely than those containing even ploidy level(s) (**Table 2**). Among populations consisting of two cytotypes, $4x + 5x$ and $5x + 6x$ combinations were most common. On the other hand, only three mixed-ploidy populations consisting of triploids were found, once with tetraploids and twice with pentaploids. Heptaploids also rarely formed mixed-ploidy populations, once with tetraploids and once with pentaploids. Only one mixed-ploidy population consisting of tetraploids and octoploids was found. Mixed-ploidy populations comprising three cytotypes withal had the combination of $4x + 5x + 6x$. Populations containing four or more cytotypes were not found (**Table 2**).

Geographic Distribution of Cytotypes

Cytotypes differed considerably in their distribution patterns at both large and small spatial scales and also in intensity of spatial intermingling (**Figure 1**, **Table 2**, **Supplementary Figures 3, 4**). Cytotype diversity over the studied area was found to be higher at lower latitudes than at higher latitudes (>52°), where only tetra- and pentaploids occurred. This pattern roughly traces the zones without and with continuous glaciation during the LGM along with areas with and without the current occurrence of diploid taxa of the *A. paniculatum* complex (**Figure 1**, **Supplementary Figures 3, 4**).

Triploids were found to be restricted to the south-eastern part of Central Europe (Pannonian Basin and Transylvanian Basin)

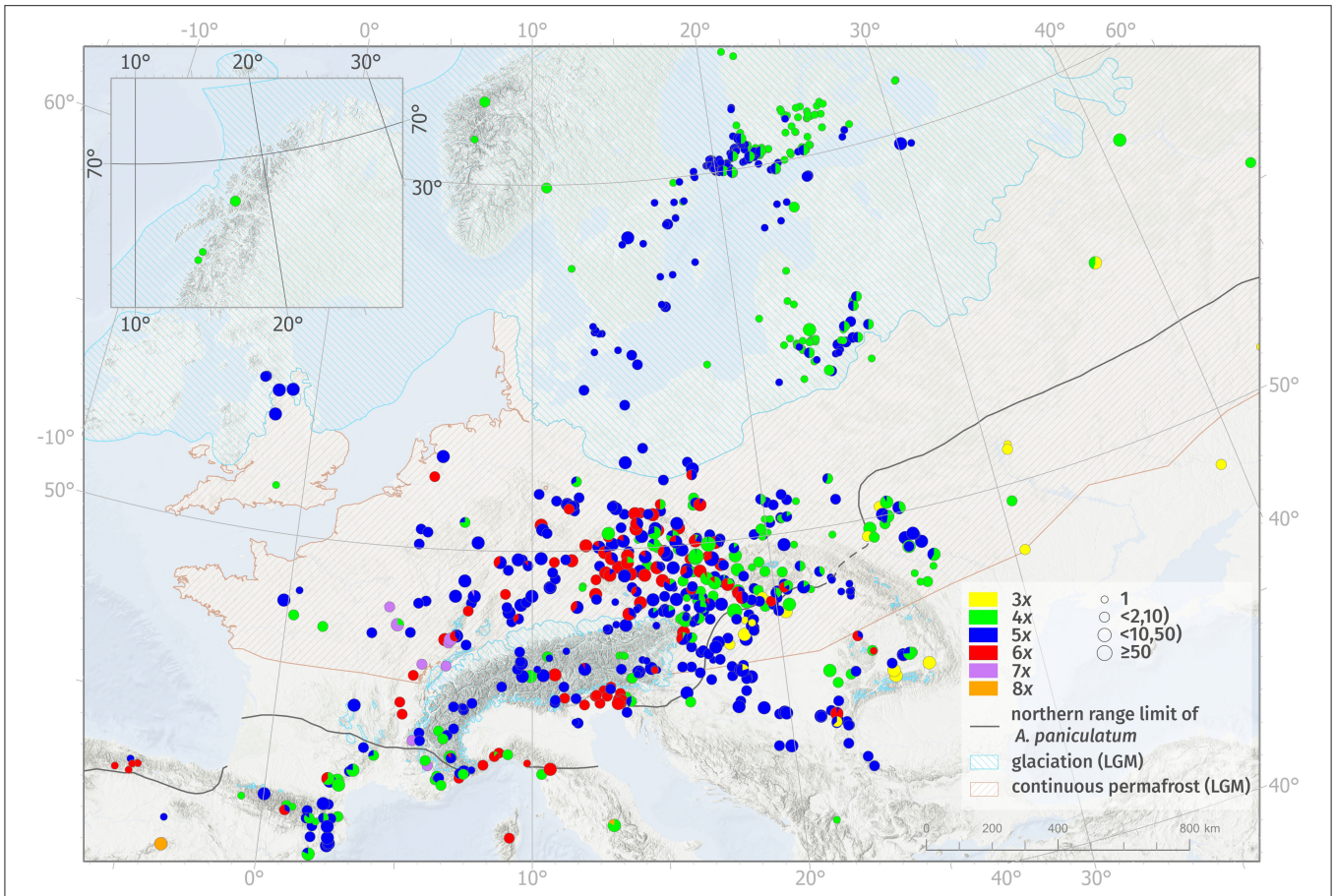

FIGURE 1 | Geographic distribution of ploidy levels of *Allium oleraceum*, based on all available data (**Supplementary Table 1**). Different ploidy levels are distinguished by different colors: 3*x*—yellow, 4*x*—green, 5*x*—blue, 6*x*—red, 7*x*—purple, 8*x*—orange. Ploidy-mixed populations are depicted as a pie chart showing the local frequency of cytotypes. The size of the circle represents the number of analyzed individuals per population. For the sake of clarity, populations from the Czech Republic and Slovakia were thinned. Lines with hatching indicate the limits of the continuous permafrost (light brown line) and glaciation (light blue line or polygons) during the last permafrost maximum according to Lindgren et al. (2016). The inset map shows the distribution of cytotypes in northern Norway. The black line shows the northern range limit of diploids of the *Allium paniculatum* complex, compiled from various sources (see Materials and Methods).

and Eastern Europe (Ukraine, southern part of European Russia) where they co-occurred in a mosaic-like pattern with tetra- and pentaploids. Specifically, two contact zones consist of (i) uniform tri- and pentaploid populations and mixed $3x + 5x$ populations in Hungary, and (ii) uniform tri- and tetraploid populations and one mixed $3x + 4x$ population in Russia, respectively. Moreover, two triploid populations were found in the contact zone with tetra-, penta- and hexaploids, one in southern Slovakia (for details see Šafářová et al., 2011), and another in southwestern Romania (**Figure 1**, **Supplementary Figure 4**).

Both tetra- and pentaploids were widespread and their distributions largely overlapped. Tetra- and pentaploid cytotypes formed a diffuse mosaic-like contact zone with both single-cytotype and mixed-ploidy populations over most of their ranges. Tetraploids had the largest geographic range of all cytotypes, from northern Spain to the coastal areas of northern Norway and Iceland (here as a non-native plant; Åström et al., 2015) and the European part of Russia. Pentaploids occurred more frequently in the western and central parts of Europe, reaching their eastern distribution limit along a line from western Ukraine to the eastern coastal areas of the Baltic Sea. At the landscape scale, however, we found several regions where tetra- and pentaploids were spatially segregated and formed ploidy-uniform areas (i.e., large-scale parapatry) with more or less pronounced contact zones with ploidy mixtures (e.g., the western part of Hungary, continuing southwards to Croatia and Serbia vs. Slovakia; Germany; Sweden vs. Finland; **Figure 1**).

Hexaploids were geographically restricted to Central Europe and northern parts of the west-Mediterranean area. A striking boundary of hexaploid distribution runs alongside the outer/inner ranges of the Western Carpathians and continues along the outer part of the Eastern Alps to southern Austria and northeastern Italy. Three isolated hexaploid populations (one single-cytotype and two mixed-ploidy populations) were found in Romania. Considering mixed populations consisting of hexaploids, only $5x + 6x$ mixed populations were found frequently and occurred in the contact zone of penta- and hexaploids, while $4x + 6x$ populations were rare, and with two

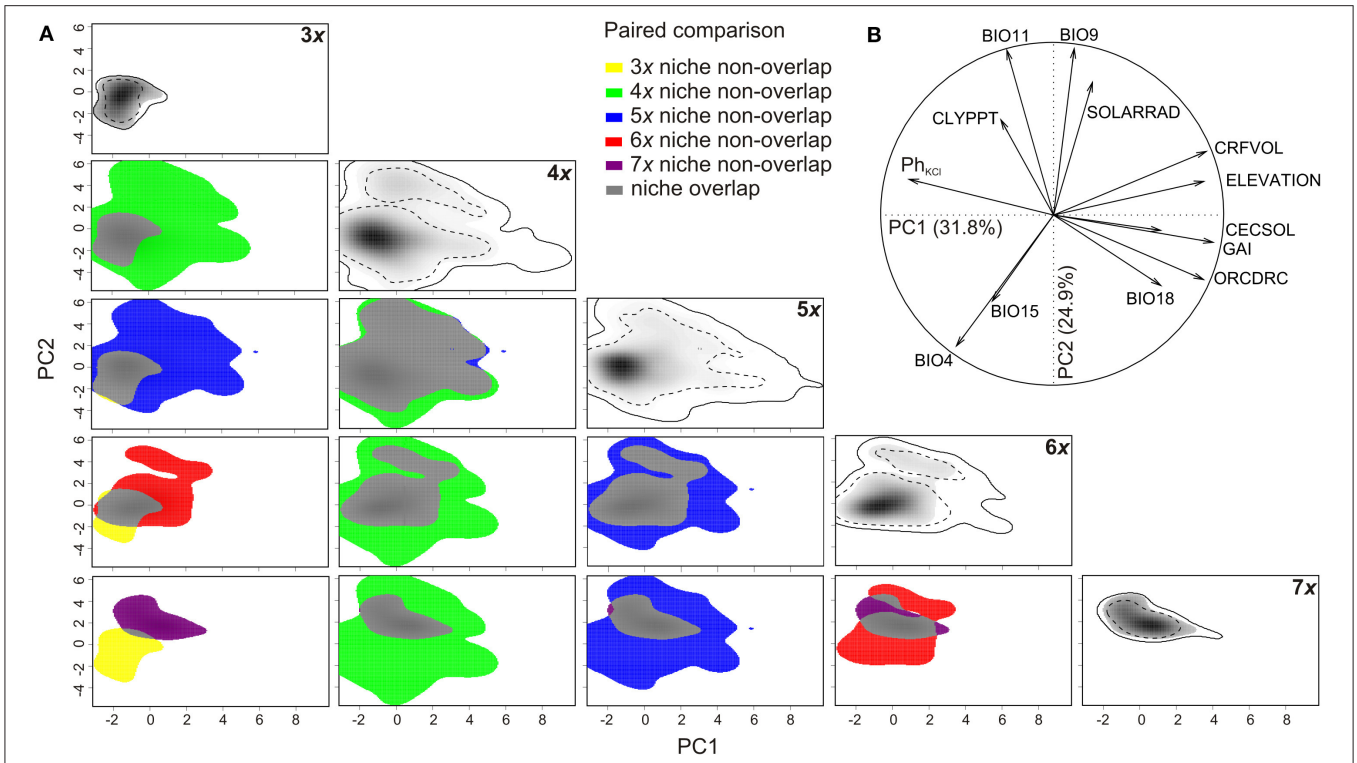

FIGURE 2 | Niches of five cytotypes of *Allium oleraceum* in the environmental space of coarse-grained variables (climate, soil) along the first two axes of PCA (PCAenv). **(A)** The panels along the diagonal represent the niches of the respective cytotypes (3*x*–7*x*). The panels below the diagonal compare niches of all pairs of cytotypes. Niches of tri-, tetra-, penta-, hexa-, and heptaploids not overlapping the niche of the other cytotypes in pairwise comparisons are shown in yellow, green, blue, red, and purple, respectively. Niche overlap is shown in gray. Color shading shows the density of the occurrences of the cytotype. Full and dashed contour lines illustrate 100 and 50%, respectively, of available (background) environments delimited by a 20-km buffer zone around the occurrence points of each cytotype. **(B)** The correlation circle shows the loadings of the individual environmental variables to the first two PCA axes. BIO4, temperature seasonality; BIO9, mean temperature of driest quarter; BIO11, mean temperature of coldest quarter; BIO15, precipitation seasonality; BIO18, precipitation of warmest quarter; CECSOL, cation exchange capacity of soil [mmol(c)/kg]; CLYPPT, weight percentage of clay particles (<0.0002 mm); CRFVOL, volumetric percentage of coarse fragments (>2 mm); GAI, global aridity index; ORCDRC, soil organic carbon content; Ph$_{KCl}$, soil acidity measured in KCl solution; SOLARRAD, mean annual solar radiation (kW.m^{-2}).

exceptions, occurred only in Central Europe. Mixed 4*x* + 5*x* + 6*x* populations were found only in Central Europe, where tetra-, penta-, and hexaploid cytotypes co-occur. At the landscape scale, however, we detected several regions where hexaploids occupied ploidy-uniform areas (e.g., part of southern Bohemia in the Czech Republic, lowlands of the Friuli-Venezia Giulia and Veneto regions in northern Italy; **Figure 1**).

Heptaploids (incl. uniform and mixed populations 4*x* + 7*x* and 5*x* + 7*x*) were found only in the western foothills of the Western Alps and Jura in eastern and southeastern France and western Switzerland. Only two octoploid populations were discovered: one, ploidy-uniform, in central Spain, and the other, mixed with tetraploids, in central Italy (**Figure 1**).

Distribution of Cytotypes Along Coarse-Grained Ecological Gradients

The variation in available environmental conditions in the species range was summarized by two PCA axes, explaining 31.7% and 24.9%, respectively, of the total variation in the environmental space (**Figure 2**). The PC1 axis mirrored a general humidity gradient (in terms of increasing humidity expressed by the global aridity index and increasing precipitation in the warmest period of the year), associated with more acid, coarse-grained soils containing more soil organic carbon and a higher cation exchange capacity at higher elevations. The axis constrained the distribution of all cytotypes in the most humid climates with acid and coarse-grain soils, usually at higher elevations (**Figure 2B**). The triploids occupied an extreme position on the PC1 axis, corresponding to habitats with alkaline, fine-grained soils in more arid conditions, while other cytotypes were slightly shifted into more humid conditions, with heptaploids being found in the relatively most humid conditions. Both tetra- and pentaploids occupied a wide range of environmental conditions along PC1. The PC2 axis mirrored a general seasonality (continentality) gradient, in terms of increasing mean temperatures of both the coldest and the driest periods of the year, and decreasing temperature and precipitation seasonalities along PC2. Considering all cytotypes together, almost all available environmental conditions along PC2 were covered. However, the axis mainly constrained the distribution of the tri- and heptaploid cytotypes, for which it showed the highest differences in temperature and precipitation seasonalities. On the

TABLE 3 | Niche overlap metric (Schoener's D) and results of niche equivalency tests, niche similarity tests, and indices of niche change in E-space (PCAenv) for five cytotypes of *A. oleraceum*.

***Allium* cytotypes**		**Niche overlap (D)**	**Niche equivalency**	**Niche similarity**		**Indices of niche change (95th percentile)**			
1	**2**			**1→ 2**	**2→ 1**	**Expansion (E)**	**Stability (S_e)**	**Unfilling (U)**	**Stability (S_n)**
3*x*	4*x*	0.161	ns	More similar	More similar	0.101	0.899	0.000	1.000
	5*x*	0.227	ns	ns	ns	0.068	0.932	0.005	0.995
	6*x*	0.264	ns	ns	ns	0.077	0.923	0.107	0.893
	7*x*	0.056	Less equivalent	ns	ns	0.000	1.000	0.509	0.491
4*x*	5*x*	0.806	More equivalent	More similar	More similar	0.001	0.999	0.039	0.961
	6*x*	0.439	ns	ns	ns	0.000	1.000	0.232	0.768
	7*x*	0.224	ns	ns	ns	0.001	0.999	0.254	0.746
5*x*	6*x*	0.540	ns	More similar	More similar	0.000	1.000	0.145	0.855
	7*x*	0.251	ns	ns	ns	0.012	0.988	0.258	0.742
6*x*	7*x*	0.371	ns	ns	ns	0.234	0.766	0.243	0.757

The niche similarity test was repeated in both directions, i.e., by resampling first the occurrences of higher ploidy level and then those of lower ploidy level. Expansion (E) and unfilling (U) indices represent proportions of the niche of higher ploidy level which is not occupied by the lower ploidy level and vice versa. Niche stability (S_n, S_e) represents the proportion of the niche of either lower (S_n) or higher ploidy level (S_e) shared with the other cytotype. Indices of niche change were calculated at the intersection of the 95th percentile of both compared environmental densities. We determined significance of niche equivalency and niche similarity tests if the observed D of the cytotypes being compared was within the lower ("less equivalent/similar") or upper 2.5% ("more equivalent/similar") quantile of a null distribution of simulated values of D, respectively (ns, non-significant).

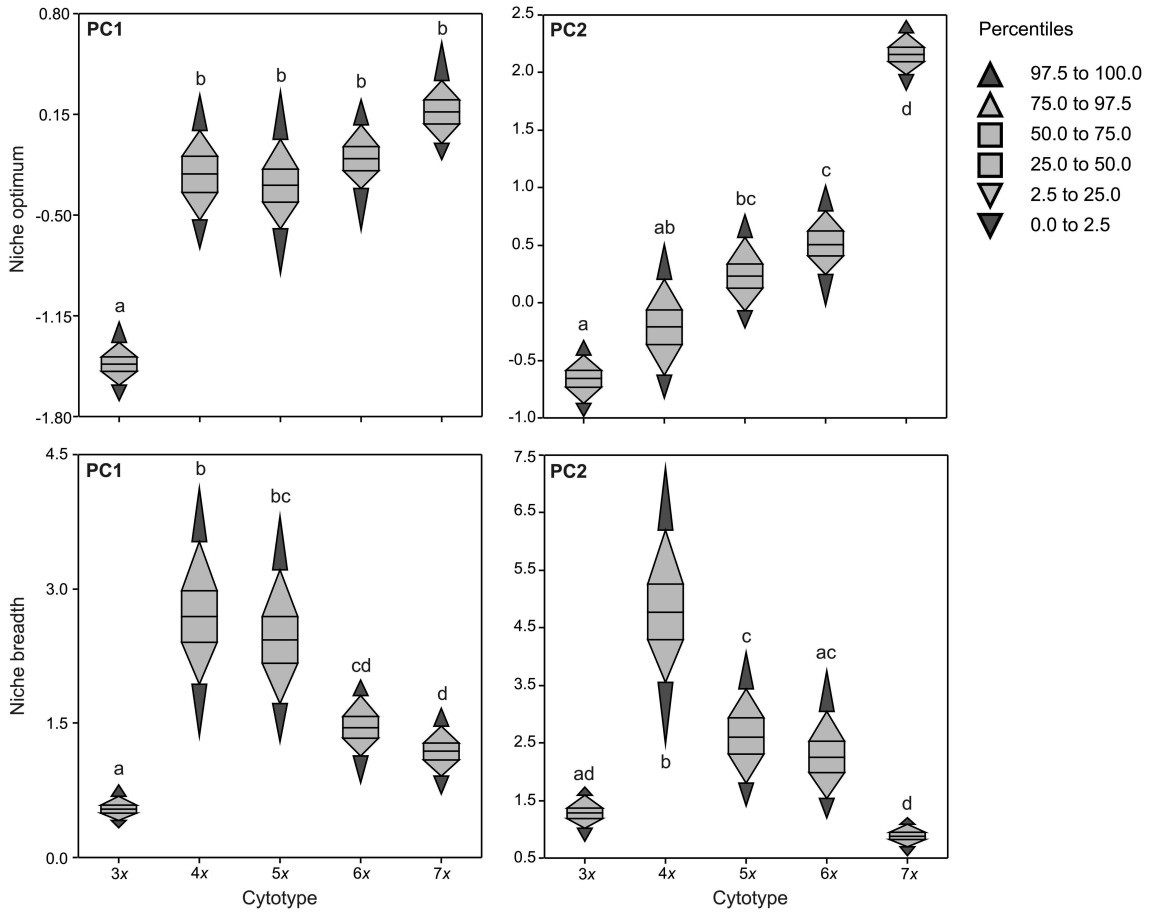

FIGURE 3 | Niche optima and niche breadths of five cytotypes of *Allium oleraceum* along the first two PCA axes (PCAenv). The same letters in each panel indicate non-significant differences in niche traits between respective cytotypes (i.e. overlaps of 95% central ranges of cytotypes).

other hand, tetraploids, and to a lesser extent also penta- and hexaploids, covered a wide range of environmental conditions along PC2, with a tendency for two optima (dominant higher and less apparent lower climate seasonality) (**Figure 2A**).

Values of niche overlap (Schoener's D) between cytotypes ranged from 0.056 for tri- and heptaploids to 0.806 for tetra- and pentaploids, and usually exceeded 0.250, suggesting a moderate to high niche overlap. Results of niche equivalency tests were inconclusive in most cases but suggested that most of the niches are equivalent (**Table 3**; **Figure 2A**). For some pairs of cytotypes, the niches were significantly less (3*x*−7*x*) or more equivalent (4*x*−5*x*) than expected by chance. In the analysis of niche similarity, the majority of tests indicated insufficient power to make inferences regarding niche differentiation. For some pairs of cytotypes (3*x*−4*x*, 4*x*−5*x*, 5*x*−6*x*), the niche similarities were significantly higher than expected by chance, irrespective of the direction of the test.

Indices of niche change comparing higher vs. lower ploidy levels showed complex patterns (**Table 3**, **Figure 2A**). Particularly tetraploids showed the highest niche expansion and complete filling of the niche of triploids of all cytotypes. Higher ploidy levels showed decreasing niche expansion but slightly increasing niche unfilling of the triploid niche when compared with the pattern observed in the tetraploids. When comparing with the niches of tetraploids or pentaploids, niches of higher ploidy levels showed almost zero expansion but an increase in niche unfilling. Only heptaploids showed both higher niche expansion and higher niche unfilling when compared with the niche of hexaploids. Using a more stringent setting (i.e., removing more marginal environments; 75th percentile) resulted in qualitatively the same conclusions as in the case of a less stringent setting (95th percentile) (data not shown).

The niche optimum of triploids along PC1 was significantly different from that of other cytotypes, being shifted to the left. The niche optimum of cytotypes along PC2 significantly increased with increasing ploidy level, but there were also some overlaps in niche optima between nearby cytotypes (e.g., 3*x*−4*x*, 4*x*−5*x*, 5*x*−6*x*; **Figure 3**). Cytotypes differed in niche breadth along both PCA axes in a complex manner. Tetra- and pentaploids, and tetraploids showed the largest niche breadth along PC1 and PC2, respectively. Both lower (3*x*) and higher (5*x* in the case of PC2, 6*x*, 7*x*) ploidy levels had lower niche breadths than tetraploids, with triploids and tri- and heptaploids having the narrowest niches along PC1 and PC2, respectively (**Figure 3**).

Local-Scale Environmental Differences Between Cytotypes

Cytotypes showed significantly different associations with particular habitat types although their habitat requirements overlapped (**Table 4**, **Figure 4**). Both tetra- and pentaploids were found in a wide spectrum of habitats, whereas tri- and heptaploids occupied a narrower spectrum of habitat types. Lower ploidy levels (3*x*−5*x*) were frequently found in open semi-natural habitats such as grassland and rocky outcrops, whereas higher ploidy levels (6*x*, 7*x*) were frequently observed in human-affected habitats such as arable land and field margins as well as

TABLE 4 | Summary of the associations between cytotypes and field-recorded local-scale environmental variables in populations of *Allium oleraceum*.

Variable	Test	DF	Test statistics	*P*
Habitat type	LLM	36	65.34	**0.002**
Presence of arable land	LLM	4	32.08	**<0.001**
Habitat naturalness	LLM	4	51.73	**<0.001**
Habitat heterogeneity	LLM	8	4.49	0.810
Rock proximity	LLM	4	13.04	0.011
Light conditions	K-W	4	5.66	0.226
Heat load	K-W	4	22.36	**<0.001**
Elevation	K-W	4	19.70	**<0.001**

Differences were tested either by log-linear models (LLM) or Kruskal–Wallis test (K-W). P-values in bold are significant after Bonferroni correction (P < 0.006).

Robinia pseudoacacia forests. Populations of higher ploidy levels occupied preferentially extensively to highly human-affected sites in close contact with arable land (Armitage test for trend in proportions, $z = -4.81$, $P < 0.001$; **Figure 4**). Populations of different cytotypes did not differ in habitat heterogeneity, light conditions of occupied habitats, and proximity of rocky habitats. Tetra- and pentaploids showed the widest elevation range from lowlands (0 m) to the alpine belt (c. 2,100–2,300 m), while tri- and heptaploids were found only at elevations below ca 600 m a s l. Hexaploids occupied slightly higher elevations than tetra- and pentaploids. Sites with the occurrence of tetraploids had a significantly lower heat load than sites of other cytotypes (**Table 4**, **Supplementary Figure 5**).

Both tests on the first (pseudo$F = 1.2$, $P = 0.001$) and all canonical axes (pseudo$F = 1.9$, $P = 0.001$) of constrained principal coordinate analysis revealed non-random differences in local-scale environmental conditions between groups of different cytotype compositions (**Figure 5**). However, all canonical axes collectively explained only 4.6% of the total variation among cytotype groups (PC1 = 2.9%, PC2 = 0.6%). The first canonical axis represented a gradient of synanthropy, with natural and semi-natural sites (rocky outcrop, grassland, dry, and mesic forests) with longer distances to the nearest arable land on the left to ruderalised or even synanthropic sites (*Robinia* forest, shrub, arable land & field margin) on the right of the ordination diagram. Except for triploid populations, single-ploidy populations were ordered along the first canonical axis from lower ploidy levels on the left to the higher ones on the right. Most mixed-cytotype groups of sites with the participation of pentaploids were usually placed in intermediate positions between uniform-ploidy sites of participating cytotypes. Other mixed-cytotype groups were usually placed near the uniform-cytotype sites of one of the participating cytotypes. The second canonical axis represents a gradient of increasing habitat heterogeneity. Cytotype groups placed in the upper part of the ordination diagram were usually mixed-cytotype sites in addition to uniform triploid populations. Paired tests between cytotypes revealed significant differences (after Bonferroni correction) in local-scale environmental conditions between lower (3*x*, 4*x*) and

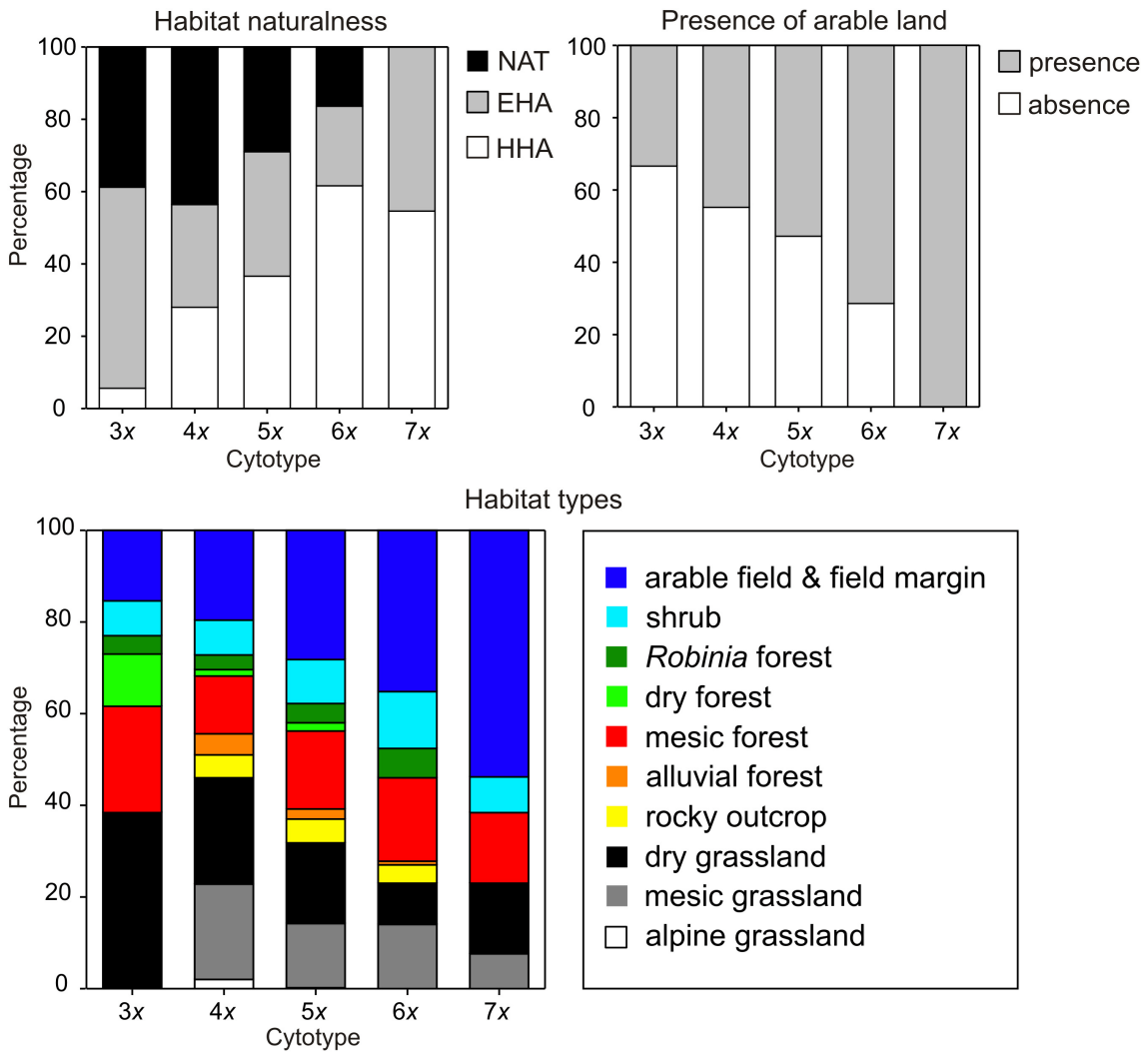

FIGURE 4 | Frequency distribution (%) of categories of field-recorded environmental variables at sites of different cytotypes of *Allium oleraceum*: habitat naturalness (HHA, highly human-affected; EHA, extensively human-affected; NAT, natural habitats), presence of arable land (presence/absence: distance to the nearest arable field ≤/>20 m), and habitat types.

higher ploidy levels (6*x*, 7*x*), and between penta- and hexaploids (**Supplementary Table 4**).

DISCUSSION

Cytotype Diversity

We found a great diversity of ploidy levels in *A. oleraceum*, ranging from triploids to octoploids. Our data on the distribution of lower ploidy levels (i.e., $2n = 3x–6x$) correspond well to published chromosome counts, but significantly extends their distributional ranges (**Figure 1**). Moreover, two uncommon cytotypes corresponding to hepta- and octoploids are the first such counts within *A.* sect. *Codonoprasum* and represent extremely rare ploidy levels within *Allium* (Hanelt et al., 1992; Goldblatt and Johnson, 2010; Rice et al., 2015; Peruzzi et al., 2017), which is considered an example of a genus with extraordinary high intraspecific variation in ploidy levels (Han et al., 2020). Such substantial intraspecific ploidy variation is only rarely seen in plants, e.g., in *Cardamine yezoensis* Maxim. with six (Marhold et al., 2010) and *Oxalis obtusa* Jacq. with seven (Krejčíková et al., 2013) cytotypes, respectively.

The majority of analyzed plants and populations were pentaploid. Higher frequencies of odd-ploidy levels (especially >3*x*) are rare within ploidy-polymorphic species (Husband et al., 2013; Kolář et al., 2017) and odd-ploidy dominance has only exceptionally been documented in mixed-ploidy species (Mock et al., 2012; Šingliarová et al., 2019). High frequencies of odd-ploidy cytotypes might indicate frequent intercytotype hybridization between even-ploidy levels (Husband et al., 2013; Hanušová et al., 2019; Šingliarová et al., 2019), and odd cytotypes generate mostly aneuploids due to unequal meiotic division resulting in unbalanced chromosome segregation and

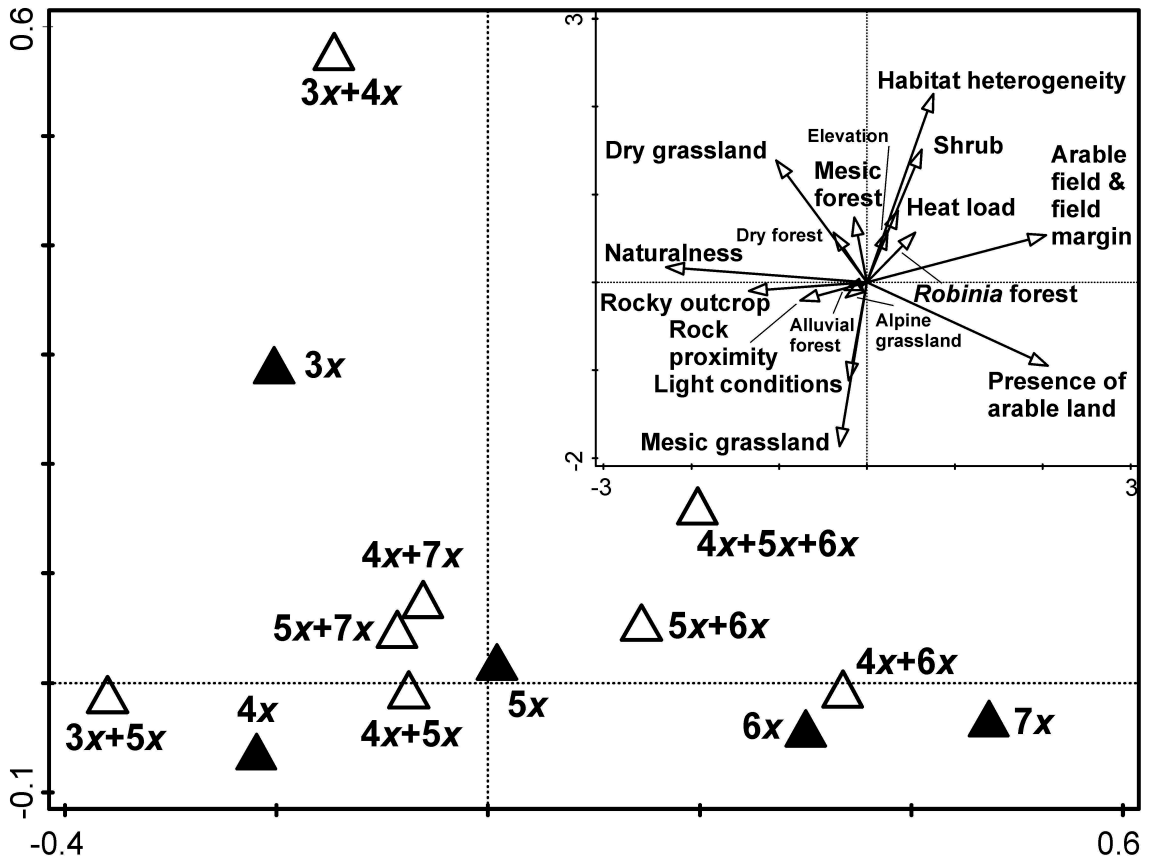

FIGURE 5 | The first two canonical axes of constrained principal coordinate analysis (db-RDA). Cytotype composition of populations was used as an explanatory variable and field-recorded environmental factors as dependent variables. The inset diagram shows correlations of field-recorded environmental factors with the first two canonical axes. Vectors of the categorical variables can be interpreted as the probability of their occurrence. Environmental variables with names in large font size are significantly correlated with at least one of the first two canonical axes ($P < 0.001$).

chromosome elimination (Ramsey and Schemske, 1998, 2002). In contrast, odd-ploidy cytotypes of *A. oleraceum* (incl. pentaploids) formed cytotype-uniform populations more frequently than even-ploidy cytotypes (**Table 2**), and no aneuploids were found in natural populations despite karyological screening of more than 500 adult plants (Levan, 1933; Karpavičienė, 2007, 2012; Åström et al., 2015; this study). We suggest that the strong asexual mode of reproduction of *A. oleraceum* cytotypes (Fialová et al., 2014) and non-viability of aneuploid seeds are crucial factors behind the observed patterns. Indeed, other polyploid species with higher frequencies of odd-ploidy cytotypes exhibit almost exclusively asexual reproduction (e.g., Krahulcová and Jarolímová, 1993; Kao, 2007; Mock et al., 2012; Kolář et al., 2017).

Drivers of Large-Scale Cytotype Distributions

The large-scale distribution pattern of different cytotypes throughout the species range was found to be remarkably complex. While tetra- and pentaploids were common across almost the entire species range and sympatric with other cytotypes at lower latitudes, other cytotypes occupied lower latitudes (below 52°) and were distributed parapatrically (tri- vs. hexaploids) or allopatrically (tri- vs. hepta- and octoploids; **Figure 1**). The current distribution of all cytotypes also overlaps the range of diploid, sexually reproducing hypothetical progenitors of the *A. paniculatum* complex (Stearn, 1980; Pastor and Valdés, 1983; Brullo et al., 1996a; Salmeri et al., 2016), but exclusively at lower latitudes. Considering both polyploid *A. oleraceum* and diploids of *A. paniculatum* complex, most regions with a high cytotype diversity are confined to unglaciated areas during the LGM while only polyploids are currently distributed at both higher latitudes and elevations, with tetra- and pentaploids being the only cytotypes represented in northern areas which were glaciated during the LGM (Huntley and Birks, 1983). A somewhat similar scenario has been described within several sexual–asexual taxa (e.g., *Antennaria* L., Bayer and Stebbins, 1987; *Ranunculus auricomus* group, Hörandl, 2006; *Crataegus* L., Lo et al., 2013; *Ranunculus kuepferi* Greuter & Burdet, Kirchheimer et al., 2016; Schinkel et al., 2016; *Paspalum intermedium* Munro ex Morong, Karunarathne et al., 2018), where polyploid apomicts tend to have larger ranges and occupy colder areas (i.e., higher latitudes and/or elevations) than diploid sexuals, suggesting evolutionary advantages of polyploidy

associated with apomixis in the colonization of deglaciated areas (i.e., geographical parthenogenesis; Bierzychudek, 1985; Hörandl, 2009; Tilquin and Kokko, 2016; Hojsgaard and Hörandl, 2019). Below, we discuss the roles of several processes which might explain the current distribution patterns of *A. oleraceum* cytotypes, such as niche evolution, asexuality, and non-adaptive processes (e.g., biogeographic history, dispersal).

Role of Niche Differentiation

There has been mixed support for the role of niche differentiation in polyploid establishment and evolution (Levin, 2002; Martin and Husband, 2009; Parisod et al., 2010; te Beest et al., 2012; Glennon et al., 2014; Visser and Molofsky, 2015; Marchant et al., 2016), although recent research using biogeographic data across vascular plants has proved that ecological niche differentiation is an important component of polyploid speciation and that polyploids have a faster niche differentiation than their diploid relatives (Baniaga et al., 2020). Unfortunately, our study does not directly compare niches between polyploids and diploid progenitors to test the phylogenetic niche conservatism (Wiens et al., 2010), since the diploid progenitors of *A. oleraceum* are not known with certainty (Duchoslav et al., 2010). Putative diploid ancestors of *A. oleraceum* (e.g., *A. paniculatum*, *A. podolicum* Blocki ex Racib. & Szafer, and *A. fuscum* Waldst. & Kit.; Levan, 1937) however share two characteristics: (i) their niches are restricted to a narrow range of natural habitats (rocky outcrops and dry grasslands) in (ii) dry and warm regions of southeastern Europe (Stearn, 1980; Brullo et al., 1996a, 2001; Dobrotchaeva et al., 1999; Ciocârlan, 2000; Ghendov, 2015; Salmeri et al., 2016). In contrast, the niche of polyploid *A. oleraceum* also comprises more mesic (colder and moister) climatic conditions (**Figures 1**–**3**) and a wide spectrum of habitat types, including fertile weedy habitats (**Figure 4**). The stronger synanthropic affinity of polyploids, in contrast to their diploid congeners, was recently found in several polyploid complexes (Zozomová-Lihová et al., 2014; Chung et al., 2015; Němečková et al., 2019; Rejlová et al., 2019). Exact niche comparison between polyploid *A. oleraceum* and its diploid progenitors is needed, as soon as molecular analyses reveal the diploid progenitors of *A. oleraceum*.

Using coarse-grained environmental variables, we identified high niche overlap in most pairwise comparisons between *A. oleraceum* cytotypes (**Table 3**, **Figure 2**). Even though the niche equivalency and similarity tests do not allow us to confirm significant differentiation or similarity in most cases, we identified two patterns of niche change within this polyploid series. Whereas, we found evidence of both niche expansion and innovation in tetraploids compared to triploids, higher ploidy levels showed (almost) zero niche expansion, but a trend of increasing unfilling of tetraploid niche. Niche unfilling in higher ploidy levels was primarily caused by the contraction of niche envelopes toward a lower continentality of the climate and resulted in a gradual decrease of niche breadth and a gradual shift in niche optima in higher ($>4x$) ploidy levels. Consequently, the geographic ranges of most cytotypes overlapped to various extents, with wider ranges of tetra- and pentaploids due to their wider niche breadth. Also analysis of local-scale environmental variables demonstrated that higher ploidy levels ($6x$, $7x$) tend to occupy a narrower segment of habitats occupied by lower ploidy levels ($4x$, $5x$), but without any sign of niche novelty. In agreement with some recent studies in other polyploid complexes (e.g., Brittingham et al., 2018; Gaynor et al., 2018; Hanušová et al., 2019), the observed divergent patterns of niche changes among *A. oleraceum* cytotypes seem to contradict the classical adaptive evolutionary scenario (Levin, 1975, 2002; Fowler and Levin, 1984, 2016), assuming shift in ecological niches of higher polyploids (i.e., niche innovation or niche expansion) compared to their lower ploidy level relatives.

One likely explanation of the observed niche patterns in *A. oleraceum* polyploids is linked to the type of polyploidy because autopolyploidy, in contrast to allopolyploidy, does not inevitably produce transgressive traits to fuel adaptive ecological divergence (Hegarty and Hiscock, 2008; Parisod et al., 2010). Additionally, multiple origins of polyploids (Soltis et al., 2010) might affect the genetic variation and consequently niche breadth of polyploids as was recently shown in e.g., tetraploids of *Dianthus broteri* Boiss. & Reut. (López-Jurado et al., 2019) and *Jasione maritima* (Duby) Merino (Castro et al., 2020). Our previous study of genome size variation in *A. oleraceum* cytotypes revealed a remarkably complex pattern both within and between cytotypes (Duchoslav et al., 2013). Specifically, $4x$–$7x$ cytotypes from (south-)western Europe had similarly sized monoploid genome sizes (1Cx), which might suggest their (recent) autopolyploid origin (Balao et al., 2009), whereas differently sized 1Cx of the eastern $3x$–$5x$ cytotypes could be the result of independent polyploidization events, potentially involving crosses between different diploids of the *A. paniculatum* complex (Duchoslav et al., 2013). Moreover, populations of two cytotypes ($4x$, $5x$), whose distributions cover both western and eastern parts of Europe (**Figure 1**), also diverged in genome sizes in a west-east direction (Duchoslav et al., 2013; **Supplementary Figure 2**). Consequently, we suppose that both tetra- and pentaploids are assemblages of different evolutionary lineages arising from different environmental conditions (Duchoslav et al., 2013), i.e., their niches are actually a composition of narrower niches of lineages with different origins. If the lineages are considered as one unit (= cytotype), they show a large niche breadth whereas other cytotypes distributed in either western ($6x$, $7x$) or eastern ($3x$) parts of the species range just show a subset of broad niches of tetra- and pentaploids (**Figures 2**, **3**). Two additional pieces of evidence support outlined explanations. Firstly, common garden experiments and physiological measurements on Central European populations of tetra-, penta-, and hexaploids have generally shown similar competitive ability (Fialová and Duchoslav, 2014) and photosynthetic capacity (Ježilová et al., 2015) of cytotypes, but also significantly lower interpopulation variation of the studied parameters in hexaploids than in tetra- and pentaploids. Secondly, a study of allozymes revealed more genetically differentiated populations of Central European tetra- and pentaploids, with two-fold higher total and within-population diversity, in comparison with genetically uniform hexaploids (Duchoslav and Staňková, 2015).

Triploids had the narrowest niche of all cytotypes with a niche optimum resembling the ecological characteristics of putative diploid progenitors (Čeřovský et al., 1999; Ciocârlan,

2000; Ghendov, 2015; Salmeri et al., 2016). Moreover, triploids only slightly exceed the geographic range of putative diploid progenitors. The narrow niche of triploids may be explained by their complete sterility (Jírová, 2007), a factor that hinders the activation of natural selection (see below; Chung et al., 2015).

The narrower range sizes and niches of higher ploidy levels (6*x*, 7*x*) could also be explained by their more recent origin. Despite several competing hypotheses as to how species (cytotype) "age" may affect range size (Johnson et al., 2014; Brittingham et al., 2018; Sheth et al., 2020), and the unknown evolutionary age of *A. oleraceum* cytotypes, newly arising higher ploidy levels (6*x*, 7*x*) certainly face competition from already established lower ploidy levels (4*x*, 5*x*). Niche contraction in higher ploidy levels has previously been observed in several polyploid complexes (Theodoridis et al., 2013; Gaynor et al., 2018). Furthermore, niche-contraction in higher ploidy levels to fertile sites rich in nitrogen and phosphorus (i.e., field margins and road ditches, *Robinia pseudoacacia* forests) might be affected by their larger genome sizes (Duchoslav et al., 2013). Both cytotypes have very large genomes (*sensu* Leitch et al., 1998), which has been suggested to have increased expenses of building and maintaining nucleic acids and associated proteins which may, under certain circumstances, act as selection pressure (Šmarda et al., 2013; Guignard et al., 2016, 2017). Indeed, the concentration of available phosphorus was found to be significantly higher in soils of Central-European sites of hexaploids than in soils of sites of tetra- and pentaploids (Duchoslav et al., 2010). Additionally, plants of higher ploidy levels with large genomes might also be constrained by the threshold for minimum stomatal size and therefore be less tolerant of water stress due to a larger stomatal length (Veselý et al., 2012). Consequently, their spread might be constrained in more arid climates.

Asexuality

Asexuality has long been hypothesized as a viable mechanism (Gustafsson, 1948; Yamauchi et al., 2004; Husband et al., 2013; Herben et al., 2017) by which newly emerging or immigrating polyploids might establish at a local site. It enables them to persist in face of strongly reduced fitness either due to mating with their more common progenitors or relatives of other ploidy levels (minority cytotype exclusion; Levin, 1975) or to meiotic irregularities caused by the formation of various meiotic configurations causing a reduction of the seed set (Ramsey and Schemske, 2002). Asexuality may also play a decisive role in the colonization of new areas, e.g., during the recolonization of previously glaciated habitats (Hörandl, 2009). Due to the ability to reproduce without mates or pollinators, such as asexually formed seeds in apomicts (Hörandl, 2006), a new population can establish from a single propagule ("Baker's Law," Baker, 1967). However, an application of this scenario on polyploid *A. oleraceum* plants is puzzling. Even though *A. oleraceum* is not an apomictic plant (*sensu* Asker and Jerling, 1992), its cytotypes predominantly reproduce uniparentally by huge and similar numbers of asexual propagules in the form of bulbils formed in the inflorescence (Fialová et al., 2014). Bulbils are ecologically similar to seeds (Ronsheim, 1994), but have a higher amount of stored resources (Karpavičienė and Karanauskaitė, 2010) and higher germination rates (Fialová et al., 2014), expressed in a higher survival and growth rate of plants originating from bulbils in comparison with seedlings (Fialová and Duchoslav, 2014). Consequently, bulbils increase the probability of successful establishment of a new population in various types of habitats differing in intensity of competition (e.g., open early-successional vegetation, dense grasslands, shaded forests; Fialová and Duchoslav, 2014; Duchoslav et al., 2017). The abovementioned traits favor *A. oleraceum* cytotypes over their putative sexual diploid progenitors lacking the ability of aerial bulbil formation (Stearn, 1980; Salmeri et al., 2016), but cannot favor one cytotype over others. However, two cytotypes with the widest niche breadth (tetra- and pentaploids) produce an order of magnitude more sexual seeds than other cytotypes (Karpavičienė, 2012; Fialová et al., 2014). Such a combination of sexual and asexual reproduction potentially advantages tetra- and pentaploids over other cytotypes due to their better ability to adapt, together with the conservation of such genetically and ecologically different genotypes via asexual propagules (= "Frozen Niche Variation Model"; Vrijenhoek, 1994). In contrast, cytotypes with scarce sexual reproduction may have a restricted niche because local adaptation by recombination will occur at lower rates, as recently shown in apomictic tetraploids of *Ranunculus kuepferi* (Kirchheimer et al., 2016).

Historical Factors

Despite high niche overlap between cytotypes and spatial intermingling of most cytotypes at a landscape scale, we detected some often quite large, nearly or entirely ploidy-uniform areas in different parts of the species' range (**Figure 1**). Most of them were formed by either penta- or tetraploids, although these two cytotypes have identical niches. One possible explanation could be that there are other environmental factors not explored here which may be responsible for niche differentiation between cytotypes (e.g., Zozomová-Lihová et al., 2014). However, we rather suggest that movements of cytotypes during and following post-glacial migration may have resulted in founder effects (te Beest et al., 2012; McAllister et al., 2015). Some of such founder effects might be of anthropogenic origin owing, for example, to the strong connection of *A. oleraceum* to Iron Age habitation and ancient settlements in the Nordic countries (Hæggström and Åström, 2005). For example, the dominance of pentaploids in southwestern Finland but prevalence of tetraploids in southeastern Finland was associated with sailing routes from Sweden dominated by pentaploids and old trade routes of pedlars from Russia dominated by tetraploids (Åström et al., 2015). Other ploidy-uniform areas, dominated by hexaploids, were found in e.g., the lowlands of the Friuli-Venezia Giulia and Veneto regions in northern Italy and the hills of southern Bohemia in the Czech Republic (**Figure 1**). Both regions are, however, strongly man-influenced landscapes with low habitat heterogeneity, dominated by intensively cultivated arable fields. We suggest that the interaction between affinity of hexaploids to fertile ruderal habitats and founder effect might be a viable explanation.

In addition, there is one large ploidy-uniform region dominated by pentaploids with rare occurrence of $3x + 5x$ mixed populations, ranging from the northwestern part of Hungary to northeastern Croatia and central Serbia. This region (Pannonian basin) has a subcontinental climate and large areas covered by loess, with degraded forest-steppe landscapes, surrounded by mountain ranges (Zólyomi and Fekete, 1994). Currently, we can only speculate about the causes of the observed pattern but we hypothesize that these pentaploids represent a lineage adapted to these specific environmental conditions.

The surprising absence of triploids from other southern regions of Europe (Iberian Peninsula, France, Italy) is hardly explainable as a sampling bias (we sampled more intensively in these regions than in Eastern Europe). More likely, their absence or perhaps rareness coincides with the apparently low diversity of putative diploid progenitors in the western part of the Mediterranean (Spain, France) than in the eastern part (e.g., Brullo et al., 1997; Jauzein and Tison, 2001; Bogdanović et al., 2008; Aedo, 2013; Tison and de Foucault, 2014; Brullo and Guarino, 2017), which may decrease the probability of triploid formation by fusion of reduced and unreduced gametes in crosses between diploid progenitors or by crosses between diploid progenitors and tetraploids of *A. oleraceum*. The latter explanation is supported by newly discovered high-level polyploids (hepta- and octoploids) in *A. oleraceum* found in southwestern Europe. Such high ploidy levels could have originated via various pathways including auto- and/or allopolyploidy (see also below), with participation of polyploid species of the *A. paniculatum* complex (e.g., tetraploid *A. oporinanthum* Brullo, Pavone & Salmeri), which are reported from the northwestern Mediterranean (Brullo et al., 1997).

A sharp boundary of the continuous distribution of hexaploids was observed in the Western Carpathians, where hexaploids meet triploids of eastern origin (**Figure 1**). This contact zone could be considered as secondary, with hexaploids migrating from the west. Concordant patterns of cytotype contact zones have been detected in several different polyploid complexes in this region (Mráz and Ronikier, 2016). However, the discovery of rare hexaploids in the Transylvanian basin, separated from the easternmost localities of the continuous range by a c. 300 km disjunction, could be explained as the result of independent origin of this cytotype (see also below).

Mixed-Ploidy Populations: Their Origin and Persistence

A high frequency of mixed-ploidy populations has been recently observed in several plant species (Kolář et al., 2017 and references therein; Hanušová et al., 2019) despite theoretical arguments that the coexistence of multiple cytotypes within populations may be unstable and represent only a transient stage following frequent generation or immigration of a divergent cytotype (Levin, 1975). We found 17.4% of populations comprising two or three cytotypes within *A. oleraceum*, which is within the mean estimates at the population level for mostly sexual and mostly asexual mixed-ploidy species as reviewed by Kolář et al. (2017). However, the frequency of mixed-ploidy populations in *A. oleraceum* observed in our study likely undervalues its real estimate because of a positive correlation between sampling effort (area, number of individuals) and within-population cytotype richness (see also Sonnleitner et al., 2010; Šingliarová et al., 2019).

We revealed mixed-ploidy populations across most of the studied species range, confirmed all previously reported combinations of *A. oleraceum* cytotypes composing mixed-ploidy populations (Duchoslav et al., 2010), and discovered five new cytotype combinations (**Table 2**). We regard the origin and spatial distribution of the majority of these mixtures to be most likely explained by secondary contacts between cytotypes (*sensu* Petit et al., 1999). The following arguments support the "secondary contact hypothesis" in *A. oleraceum.* Firstly, most mixed-ploidy populations occur exclusively in contact zones of participating cytotypes, in a diffuse, mosaic-like pattern with uniform-cytotype populations, not as rare cases of cytotype mixtures in otherwise ploidy-uniform areas. Furthermore, all cytotypes also form their own populations, suggesting a rather long term stability of all cytotypes and not a scenario of ongoing polyploidisation via the production of unreduced gametes, as proposed, for example, for hexaploids and higher ploidy levels of *Cardamine yezoensis* (Marhold et al., 2010), or a scenario of repeated interploidy hybridization such as in the case of the assumed origin of pentaploids in mixed $4x + 6x$ populations of *Cystopteris fragilis* (L.) Bernh. (Hanušová et al., 2019). In these examples, newly generated polyploids do not spread from mixed populations and do not develop their own populations. Secondly, the origin of some mixed-ploidy populations in *A. oleraceum* (e.g., $4x + 5x$, $4x + 7x$; $5x + 7x$) is difficult to explain as a result of interploidy crosses or of primary formation of a higher cytotype in lower-ploidy level populations (Ramsey and Schemske, 1998, 2002). Thirdly, the observed moderate to strong overlaps of both niches and geographic distributions between most cytotypes may result in local spatial contacts, which are significantly enhanced by the dispersion of propagules (especially aerial bulbils) via human-mediated transport of hay and cereals, soil movement during soil preparation, or collection and spreading of bulbils by small rodents (Duchoslav, 2001). Fourthly, short-time cytotype coexistence in *A. oleraceum* was experimentally evidenced in the form of survival of foreign ploidy levels at home-ploidy sites in a 5-year experiment with aerial bulbils transplanted reciprocally to cytotype-uniform sites of tetra-, penta-, and hexaploids in central Europe (Duchoslav et al., 2017). Lastly, asexual reproduction as a dominant reproductive mode of all cytotypes (Fialová et al., 2014) facilitates their establishment (Fialová and Duchoslav, 2014) and allows established plants to avoid reproduction costs due to their minority status (Yamauchi et al., 2004; McAllister et al., 2015) and potential lack of compatible partners for mating in this allogamous species. The decisive role of asexuality for local coexistence of different cytotypes was proved by Kao (2007, 2008), who observed no reduction in fitness of rare cytotypes in mixed-cytotype populations in *Arnica cordifolia* Hook, a facultatively asexual species.

Nevertheless, we cannot rule out that some mixed populations represent the primary zone of contacts between cytotypes or that gene flow between cytotypes occurs. An allozyme study of Czech populations of *A. oleraceum* (Duchoslav and Staňková, 2015)

demonstrated complex patterns of genetic similarity between tetra-, penta-, and hexaploid plants in mixed-ploidy populations, ranging from close to weak relatedness between locally coexisting cytotypes. The identical banding pattern of tetra- and hexaploids observed in two $4x + 6x$ populations was interpreted as hexaploids having originated *in-situ de novo* via fusion of unreduced and reduced gametes of tetraploids (Duchoslav and Staňková, 2015). However, we believe that a neohexaploid origin in tetraploid populations should be a rare process given the fact that no hexaploid plants were found in tetraploid populations outside the continuous range of hexaploids (**Figure 1**). On the other hand, the discovery of rare octoploids in a tetraploid population observed in central Italy (No. 452; **Figure 1**) might be interpreted as a spontaneous polyploidization event.

Higher odd-ploidy cytotypes (especially pentaploids) have a relatively higher fitness than triploids due to an increased production of viable euploid gametes ranging from $1x$ to $5x$ (Ramsey and Schemske, 1998; Costa et al., 2014). These may thus participate in both intra- and interploidal crosses (Sutherland and Galloway, 2017). However, no evidence of the occurrence of adult plants of other-than-mother cytotypes was observed in populations in ploidy-uniform regions (**Figure 1**), even though some of those regions were dominated by pentaploids, potentially producing karyologically variable progeny. Cytotype uniformity within such populations could be explained by a weak competitive ability and consequently lower survival of seedlings in comparison with young plants sprouted from bulbils, i.e., asexual copies of the maternal cytotype (Fialová and Duchoslav, 2014). Further studies into the cytotype variation of generative offspring produced by each cytotype and inter-cytotype crossing experiments are needed to evaluate the role of gene flow between cytotypes as a potential source of genetic and ecological diversity in *A. oleraceum*.

CONCLUSIONS

Our study has provided detailed insight into the diversity and ploidy distribution patterns in populations of a geophytic species at a continental scale. Large ploidy diversity (six ploidy levels; $3x$—$8x$) ranks *A. oleraceum* among the most cytotype-diverse mixed-ploidy plant species (Kolář et al., 2017), and the dominance of pentaploids (50.2%) makes *A. oleraceum* a unique case. All types of spatial arrangements of cytotypes at the landscape scale (i.e., allopatry, parapatry, and sympatry) were detected in *A. oleraceum*, together with relatively frequent coexistence of multiple ploidy levels within populations, resulting in various cytotype combinations. The widespread distribution of tetraploids and pentaploids contrasted with narrower ranges of both lower (triploid) and higher ploidy levels (hexa- to octoploids), which also differed from each other in geographic distributions.

We further focused on possible evolutionary drivers of the observed spatial patterns of cytotypes. In contrast to the classical adaptive evolutionary scenario, assuming niche innovation or niche expansion in higher ploidy levels, we found niche expansion in tetraploids compared to triploids followed by gradual niche contraction in higher ploidy levels, with higher ploidy levels occupying mainly synanthropic habitats. The wide niche breadth of tetraploids and pentaploids might be explained by their multiple origins from different environmental conditions, higher "age," and retained sexuality, which probably preserves their adaptive potential. By contrast, both lower and higher ploidy levels with narrower niches are mostly asexual and probably originated in a limited range of contrasting environments. Newly evolved higher ploidy levels have also faced competition from existing cytotypes, which might have affected a shift of their niche. Persistence of local ploidy mixtures could be enabled by the perenniality of *A. oleraceum*, and intensive vegetative reproduction of all cytotypes, which facilitates establishment and allows plants to avoid reproductive costs due to their minority status. Vegetative reproduction might also significantly accelerate colonization of new areas, including recolonization of previously glaciated habitats. Further study should primarily focus on molecular analyses to reveal the relationships between cytotypes and to identify their origin. Besides that, comparison of the niches of polyploids with those of identified diploid progenitors could also be tested.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

MD designed the research and conducted field collection with the participation of MJ, LŠ, and KV. MJ, LK, and LŠ conducted laboratory analyses. JB extracted geoinformatics data and draw maps. MD performed statistical analyses and wrote the manuscript with help of JB and LK. All authors participated in the discussion and approved the submitted version.

FUNDING

Population sampling and FCM analyses were supported by the Grant Agency of the Czech Republic (grant number 206/04/P115) and final completion was supported by the Internal Grant Agency of Palacký University (IGA PrF-2020-003). MJ was supported by long-term research development Project No. RVO 67985939 of the Czech Academy of Sciences.

ACKNOWLEDGMENTS

We acknowledge all the people (incl. staff of several Botanic Gardens) mentioned in the Electronic Appendix 1 for their help in collecting samples of *Allium oleraceum*. We thank J. Ohryzek, M. Fialová, A. Lepší, J. Duchoslavová, V. Koláčková, and R. Slípková for their help with fieldwork and maintenance of samples in the experimental garden. Gabrielle Casazza helped with niche analyses. MD thanks František Krahulec, who inspired him to study the *Allium oleraceum* complex, helped a lot with

fieldwork, and was always keen to discuss the topic. The authors are grateful to two reviewers for all the constructive and helpful comments that greatly improved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.591137/full#supplementary-material

Supplementary Figure 1 | Microphotographs of mitotic plates of *Allium oleraceum* cytotypes ($2n = 3x - 8x$). Scale bar = 10 μm. Plant locations: 3*x*: Ukraine (Population No. 234), 4*x*: Czech Republic (No. 1118), 5*x*: Czech Republic (No. 1091), 6*x*: Czech Republic (No. 1147), 7*x*: France (No. 80), 8*x*: Spain (No. 239).

Supplementary Figure 2 | (A) Box-plots of mean relative fluorescence (ratio to the internal standard *Triticum aestivum* cv. *Saxana*) of 596 populations of six ploidy levels of *Allium oleraceum* measured by FCM (PI stain; see **Table 1**). **(B, C)** Geographic pattern of the mean relative fluorescence of *Allium oleraceum* cytotypes along latitude **(B)** and longitude **(C)**, respectively. The median smoothed line was fitted separately for each cytotype by moving average procedure where running medians are used instead of running means. Chromosome counts were done for selected populations of tetra- and pentaploids with marginal relative fluorescence. Such populations are marked by asterisks (∗) in the plots B and C.

Supplementary Figure 3 | Detailed distribution of ploidy levels of *Allium oleraceum* over its range, based on all available data (**Supplementary Table 1**). Separate maps were drawn for each dominant cytotype including ploidy-mixed populations. Different ploidy levels are distinguished by different colours: 3*x* - yellow, 4*x* - green, 5*x* - blue, 6*x* - red, 7*x* - purple, 8x - orange. Ploidy-mixed populations are depicted as a pie chart showing the local frequency of cytotypes. The size of the circle represents the number of analysed individuals per population. The lines with corresponding hatching indicate the limits of the continuous permafrost (light brown line) and glaciation (light blue line or polygons) during the last permafrost maximum according to Lindgren et al. (2016). The inset map shows the distribution of cytotypes in northern Norway. The black line shows the northern range limit of diploids of *Allium paniculatum* complex, compiled from various sources (see Material and Methods).

Supplementary Figure 4 | Detailed distribution of ploidy levels of *Allium oleraceum* in Central Europe, based on all available data (**Supplementary Table 1**). Different ploidy levels are distinguished by different colours: 3*x* - yellow, 4*x* - green, 5*x* - blue, 6*x* - red, 7*x* - purple, 8x - orange. Ploidy-mixed populations are depicted as a pie chart showing the local frequency of cytotypes. The size of the circle represents the number of analysed individuals per population. The lines with corresponding hatching indicate the limits of the continuous permafrost (light brown line) and glaciation (light blue line or polygons) during the last permafrost maximum according to Lindgren et al. (2016). The inset map shows the distribution of cytotypes in northern Norway. The black line shows the northern range limit of diploids of *Allium paniculatum* complex, compiled from various sources (see Material and Methods).

Supplementary Figure 5 | Frequency distribution (%) of categories of the following field-recorded, fine-scale environmental variables at sites of different cytotypes of *Allium oleraceum*: habitat heterogeneity (number of local habitats occupied by the local population: white = 1 habitat, grey = 2 habitats, black = 3 habitats); Light conditions (white = deep shade, light grey = half shade, dark grey = light shade, black = full insolation). The bottom panels represent percentile-plots of elevation and heat load at sites of different cytotypes of *A. oleraceum*.

Supplementary Table 1 | *Allium oleraceum* populations analysed in this study, including extracted data from other sources. For each population, ID code, locality description, geographic coordinates (WGS-84), sample size (Nind = number of individuals analysed; in the case of literature sources without sample size information, just one analysed individual per ploidy was considered for each site), ploidy level (Ploidy composition, '+' means the case of mixed-cytotype populations), counted chromosome numbers, mean, SD, and sample size (Nindi) of the relative fluorescence intensity of identified cytotypes (flow-cytometry (FCM); *Triticum aestivum* cv. *Saxana* as an internal standard); dye used for FCM estimation (PI, DAPI), field-recorded, local-scale ecological variables and extracted coarse-grained environmental variables (see **Supplementary Table 3** for explanation), and sources of data are given. Populations, which were used for ecological analyses, are marked with an asterisk within the column named "Niche analyses".

Supplementary Table 2 | Survey of coarse-grained environmental variables used in ecological niche analysis (PCAenv). For each variable, its abbreviation, units, scaling factor, and source of data are presented. Additionally, descriptive statistics (mean and SD) of each coarse-grained environmental variable for each cytotype, using all localities in **Supplementary Table 1**, are presented.

Supplementary Table 3 | Survey of local-scale environmental and population variables used in the study. Vegetation types according to Mucina et al. (2016), which were included within each habitat type used in the study, are given at the bottom.

Supplementary Table 4 | Summary of the pairwise multivariate analyses between cytotypes using db-RDA applied to local-scale environmental variables recorded in populations of *A. oleraceum*. The significance level of the db-RDA was assessed by Monte Carlo permutation test using 999 permutations. *P*-values in bold are significant after Bonferroni correction ($P < 0.006$).

REFERENCES

Åström, H., Hæggström, C. A., and Hæggström, E. (2015). Geographical distribution of *Allium oleraceum* cytotypes in Finland and Sweden. *Nord. J. Bot.* 33, 120–125. doi: 10.1111/njb.00521

Adams, K. L., and Wendel, J. F. (2005). Novel patterns of gene expression in polyploid plants. *Trends Genet.* 21, 539–543. doi: 10.1016/j.tig.2005.07.009

Aedo, C. (2013). "*Allium* L.," in *Flora Ibérica. Vol. XX. Liliaceae-Agavaceae,* eds E. Rico, M. B. Crespo, A. Quintanar, A. Herrero, and C. Aedo (Madrid: Consejo Superior de Investigaciones Cientificas), 220–273.

Afonso, A., Loureiro, J., Arroyo, J., Olmedo-Vicente, E., and Castro, S. (2020). Cytogenetic diversity in the polyploid complex *Linum suffruticosum* s.l. (Linaceae). *Bot. J. Linn. Soc.* boaa060. doi: 10.1093/botlinnean/boaa060

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* 38, 541–545. doi: 10.1111/ecog.01132

Araújo, M. B., Ferri-Yáñez, F., Bozinovic, F., Marquet, P. A., Valladares, F., and Chown, S. L. (2013). Heat freezes niche evolution. *Ecol. Lett.* 16, 1206–1219. doi: 10.1111/ele.12155

Arrigo, N., de La Harpe, M., Litsios, G., Zozomová-Lihová, J., Španiel, S., Marhold, K., et al. (2016). Is hybridization driving the evolution of climatic niche in *Alyssum montanum*? *Am. J. Bot.* 103, 1348–1357. doi: 10.3732/ajb.1500368

Asker, S. E., and Jerling, L. (1992). *Apomixis in Plants.* Florida, FL: CRC Press, Boca Raton.

Baker, H. G. (1967). Support for Baker's law–as a rule. *Evolution* 21, 853–856. doi: 10.1111/j.1558-5646.1967.tb03440.x

Balao, F., Casimiro-Soriguer, R., Talavera, M., Herrera, J., and Talavera, S. (2009). Distribution and diversity of cytotypes in *Dianthus broteri* as evidenced by genome size variations. *Ann. Bot.* 104, 965–973. doi: 10.1093/aob/mcp182

Balao, F., Herrera, J., and Talavera, S. (2011). Phenotypic consequences of polyploidy and genome size at the microevolutionary scale: a multivariate morphological approach. *New Phytol.* 192, 256–265. doi: 10.1111/j.1469-8137.2011.03787.x

Baniaga, A. E., Marx, H. E., Arrigo, N., and Barker, M. S. (2020). Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecol. Lett.* 23, 68–78. doi: 10.1111/ele.13402

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210, 391–398. doi: 10.1111/nph.13698

Barringer, B. C. (2007). Polyploidy and self-fertilization in flowering plants. *Am. J. Bot.* 94, 1527–1533. doi: 10.3732/ajb.94.9.1527

Bayer, R. J., and Stebbins, G. L. (1987). Chromosome numbers, patterns of distribution, and apomixis in *Antennaria* (Asteraceae: Inuleae). *Syst. Bot.* 12, 305–319. doi: 10.2307/2419326

Bennett, M. D., and Smith, J. B. (1972). The effects of polyploidy on meiotic duration and pollen development in cereal anthers. *Proc. Roy. Soc. Biol. Sci. Ser. B* 181, 81–107. doi: 10.1098/rspb.1972.0041

Bierzychudek, P. (1985). Patterns in plant parthenogenesis. *Experientia* 41, 1255–1264. doi: 10.1007/BF01952068

Bogdanović, S., Brullo, S., Mitić, B., and Salmeri, C. (2008). A new species of *Allium* (Alliaceae) from Dalmatia, Croatia. *Bot. J. Linn. Soc.* 158, 106–114. doi: 10.1111/j.1095-8339.2008.00790.x

Bretagnole, F., and Thompson, J. D. (1996). An experimental study of ecological differences in winter growth between sympatric diploid and autotetraploid *Dactylis glomerata*. *J. Ecol.* 84, 343–351. doi: 10.2307/2261197

Brittingham, H. A., Koski, M. H., and Ashman, T.-L. (2018). Higher ploidy is associated with reduced range breadth in the *Potentilleae* tribe. *Am. J. Bot.* 105, 700–710. doi: 10.1002/ajb2.1046

Brochmann, C., Brysting, A. K., Alsos, I. G., Borgen, L., Grundt, H. H., Scheen, A.-C., et al. (2004). Polyploidy in arctic plants. *Biol. J. Linn. Soc.* 82, 521–536. doi: 10.1111/j.1095-8312.2004.00337.x

Broennimann, O., Fitzpatrick, M. C., Pearman, P. B., Petitpierre, B., Pellissier, L., Yoccoz, N. G., et al. (2012). Measuring ecological niche overlap from occurrence and spatial environmental data. *Glob. Ecol. Biogeogr.* 21, 481–497. doi: 10.1111/j.1466-8238.2011.00698.x

Broennimann, O., Treier, U., Müller-Schärer, H., Thuiller, W., Peterson, A., and Guisan, A. (2007). Evidence of climatic niche shift during biological invasion. *Ecol. Lett.* 10, 701–709. doi: 10.1111/j.1461-0248.2007.01060.x

Brullo, S., and Guarino, R. (2017). "Chapter 7: *Allium* L. -Aglio (incl. cipolla, porro)," in *Flora d'Italia, Seconda Edizione*, eds S. Pignatti, R. Guarino, and M. La Rosa (Bologna: Edagricole di New Business Media), 238–269.

Brullo, S., Guglielmo, A., Pavone, P., and Salmeri, C. (2008). Taxonomic study on *Allium dentiferum* Webb & Berthel. (Alliaceae) and its relations with allied species from the Mediterranean. *Taxon* 57, 243–253. doi: 10.2307/25065965

Brullo, S., Guglielmo, A., Pavone, P., Scelsi, F., and Terrasi, M. C. (1996a). Cytotaxonomic consideration of *Allium fuscum* Waldst. et Kit. (Liliaceae), a critical species of the European flora. *Folia Geobot. Phytotax.* 31, 465–472. doi: 10.1007/BF02812087

Brullo, S., Pavone, P., and Salmeri, C. (1996b). Considerazioni citotassonomische su *Allium pallens* L. (Alliaceae), specie critica del Mediterraneo. *Inform. Bot. Ital.* 27:309.

Brullo, S., Pavone, P., and Sameri, C. (1997). *Allium oporinanthum* (Alliaceae), a new species from the NW Mediterranean area. *Anales Jard. Bot. Madrid* 55, 297–302. doi: 10.3989/ajbm.1997.v55.i2.276

Brullo, S., Guglielmo, A., Pavone, P., and Salmeri, C. (2001). Osservazioni tassonomiche e cariologiche sulle specie del ciclo di *Allium paniculatum* L. in Italia. *Inform. Bot. Ital.* 33, 500–506.

Brullo, S., Guglielmo, A., Pavone, P., and Salmeri, C. (2003). Cytotaxonomical remarks on *Allium pallens* and its relationships with *A. convallarioides* (Alliaceae). *Bocconea* 16, 557–571.

Castro, M., Loureiro, J., Figueiredo, A., Serrano, M., Husband, B. C., and Castro, S. (2020). Different patterns of ecological divergence between two tetraploids and their diploid counterpart in a parapatric linear coastal distribution polyploid complex. *Front. Plant Sci.* 11:315. doi: 10.3389/fpls.2020.00315

Čeřovský, J., Feráková, V., Holub, J., Maglocký, Š., and Procházka, F. (eds.). (1999). *Červená kniha ohrožených a vzácných druhu rostlin a Živočichů ČR a SR. Vol. 5. Vyšší rostliny*. Bratislava. Príroda.

Čertner, M., Fenclová, E., Kúr, P., Kolář, F., Koutecký, P., Krahulcová, A., et al. (2017). Evolutionary dynamics of mixed-ploidy populations in an annual herb: dispersal, local persistence and recurrent origins of polyploids. *Ann. Bot.* 120, 303–315. doi: 10.1093/aob/mcx032

Čertner, M., Kúr, P., Kolář, F., and Suda, J. (2019). Climatic conditions and human activities shape diploid–tetraploid coexistence at different spatial scales in the common weed *Tripleurospermum inodorum* (Asteraceae). *J. Biogeogr.* 46, 1355–1366. doi: 10.1111/jbi.13629

Chung, M. Y., López-Pujol, J., Chung, J. M., Kim, K.-J., Park, S. J., and Chung, M. G. (2015). Polyploidy in *Lilium lancifolium*: evidence of autotriploidy and no niche divergence between diploid and triploid cytotypes in their native ranges. *Flora* 213, 57–68. doi: 10.1016/j.flora.2015.04.002

Ciocârlan, V. (2000). *Flora ilustrată a României.* Bucureşti: Edit Ceres.

Cires, E., Cuesta, C., Revilla, M. A., and Prieto, J. A. D. (2010). Intraspecific genome size variation and morphological differentiation of *Ranunculus parnassifolius* (Ranunculaceae), an alpine–pyrenean–cantabrian polyploid group. *Biol. J. Linn. Soc.* 101: 251–271. doi: 10.1111/j.1095-8312.2010.01517.x

Cosendai, A. C., Wagner, J., Ladinig, U., Rosche, C., and Hörandl, E. (2013). Geographical parthenogenesis and population genetic structure in the alpine species *Ranunculus kuepferi* (Ranunculaceae). *Heredity* 110, 560–569. doi: 10.1038/hdy.2013.1

Costa, J., Ferrero, V., Louriero, J., Castro, M., Navarro, L., and Castro, S. (2014). Sexual reproduction of the pentaploid, short-styled *Oxalis pes-caprae* allows the production of viable offspring. *Plant Biol.* 16: 208–214. doi: 10.1111/plb.12010

Decanter, L., Colling, G., Elvinger, N., Heiðmarsson, S., and Matthies, D. (2020). Ecological niche differences between two polyploid cytotypes of *Saxifraga rosacea*. *Am. J. Bot.* 107, 423–435. doi: 10.1002/ajb2.1431

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., D'Amen, M., Randin, C., et al. (2017). ecospat: an R package to support spatial analyses and modeling of species niches and distributions. *Ecography* 40, 774–787. doi: 10.1111/ecog.02671

Dobrotchaeva, D. N., Kotov, M. I., and Prokudin, J. N. (eds.). (1999). *Opredetitel' Vyssich Rastenij Ukrainy*. Kiev: Institute of Botany.

Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2, 2233–2244. doi: 10.1038/nprot.2007.310

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46. doi: 10.1111/j.1600-0587.2012.07348.x

Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P. S., et al. (2008). Evolutionary genetics of genome merger and doubling in plants. *Ann. Rev. Genet.* 42, 443–461. doi: 10.1146/annurev.genet.42.110807.091524

Duchoslav, M., Fialová, M., and Jandová, M. (2017). The ecological performance of tetra-, penta- and hexaploid geophyte *Allium oleraceum* in reciprocal transplant experiment may explain the occurrence of multiple-cytotype populations. *J. Plant Ecol.* 10: 569–580. doi: 10.1093/jpe/rtw039

Duchoslav, M., Šafářová, L., and Jandová, M. (2013). Role of adaptive and non-adaptive mechanisms forming complex patterns of genome size variation in six cytotypes of polyploid *Allium oleraceum* (Amaryllidaceae) on a continental scale. *Ann. Bot.* 111, 419–431. doi: 10.1093/aob/mcs297

Duchoslav, M., Šafářová, L., and Krahulec, F. (2010). Complex distribution patterns, ecology and coexistence of ploidy levels of *Allium oleraceum* (Alliaceae) in the Czech Republic. *Ann. Bot.* 105, 719–735. doi: 10.1093/aob/mcq035

Duchoslav, M., and Staňková, H. (2015). The population genetic structure and clonal diversity of *Allium oleraceum* (Amaryllidaceae), a polyploid geophyte with common asexual but variable sexual reproduction. *Folia Geobot.* 50, 123–136. doi: 10.1007/s12224-015-9213-0

Duchoslav, M. (2001). *Allium oleraceum* and *A. vineale* in the Czech Republic: distribution and habitat differentiation. *Preslia* 73, 173–184.

Duchoslav, M. (2009). Effects of contrasting habitats on the phenology, seasonal growth, and dry-mass allocation pattern of two bulbous geophytes (Alliaceae) with partly different geographic ranges. *Pol. J. Ecol.* 57, 15–32.

Fialová, M., and Duchoslav, M. (2014). Response to competition of bulbous geophyte *Allium oleraceum* differing in ploidy levels. *Plant Biol.* 16, 186–196. doi: 10.1111/plb.12042

Fialová, M., Jandová, M., Ohryzek, J., and Duchoslav, M. (2014). Biology of the polyploid geophyte *Allium oleraceum* (Amaryllidaceae): variation in size, sexual and asexual reproduction and germination within and between tetra-, penta- and hexaploid cytotypes. *Flora* 209, 312–324. doi: 10.1016/j.flora.2014.04.001

Fialová, R. (1996). *Polyploid complexes in the genus Allium* (dissertation). Palacký University Olomouc, Olomouc, Czech Republic.

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi: 10.1002/joc.5086

Fowler, N., and Levin, D. (1984). Ecological contrasts on the establishment of a novel polyploid in competition with its diploid progenitor. *Am. Nat.* 124, 703–711. doi: 10.1086/284307

Fowler, N. L., and Levin, D. A. (2016). Critical factors in the establishment of allopolyploids. *Am. J. Bot.* 103, 1236–1251. doi: 10.3732/ajb.1500407

Friesen, N., Fritsch, R. M., and Blattner, F. (2006). Phylogeny and new intragenetic classification of *Allium* (Alliaceae) based on nuclear ribosomal DNA ITS sequences. *Aliso* 22, 372–395. doi: 10.5642/aliso.20062201.31

Gallagher, J. P., Grover, C. E., Hu, G., and Wendel, J. F. (2016). Insights into the ecology and evolution of polyploid plants through network analysis. *Molec. Ecol.* 25, 2644–2660. doi: 10.1111/mec.13626

Gaynor, M. L., Marchant, D. B., Soltis, D. E., and Soltis, P. S. (2018). Climatic niche comparison among ploidal levels in the classic autopolyploid system, *Galax urceolata*. *Am. J. Bot.* 105, 1631–1642. doi: 10.1002/ajb2.1161

Ghendov, V. (2015). Notes on *Allium paniculatum* L. s. l. (Alliaceae Juss.) in the flora of Republic of Moldova. *J. Bot.* 7, 101–105.

Glennon, K. L., Ritchie, M. E., and Segraves, K. A. (2014). Evidence for shared broad-scale climatic niches of diploid and polyploid plants. *Ecol. Lett.* 17, 574–582. doi: 10.1111/ele.12259

Godsoe, W., Larson, M. A., Glennon, K. L., and Segraves, K. A. (2013). Polyploidization in *Heuchera cylindrica* (Saxifragaceae) did not result in a shift in climatic requirements. *Am. J. Bot.* 100, 496–508. doi: 10.3732/ajb.1200275

Goldblatt, P., and Johnson, D. E. (2010). *Index to Plant Chromosome Numbers 2004-2006*. Monographs in systematic botany. St. Louis, MO: Missouri Botanical Garden.

Grant, V. (1981). *Plant Speciation*. New York, NY: Columbia University Press. doi: 10.7312/gran92318

Guignard, M. S., Leitch, A. R., Acquisti, C., Eizaguirre, C., Elser, J. J., Hessen, D. O., et al. (2017). Impacts of nitrogen and phosphorus: from genomes to natural ecosystems and agriculture. *Front. Ecol. Evol.* 5:70. doi: 10.3389/fevo.2017.00070

Guignard, M. S., Nichols, R. A., Knell, R. J., Macdonald, A., Romila, C. A., Trimmer, M., et al. (2016). Genome size and ploidy influence angiosperm species' biomass under nitrogen and phosphorus limitation. *New Phytol.* 210, 1195–1206. doi: 10.1111/nph.13881

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C., and Kueffer, C. H. (2014). Unifying niche shift studies: insights from biological invasions. *Trends Ecol. Evol.* 29, 260–269. doi: 10.1016/j.tree.2014.02.009

Guisan, A., and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8, 993–1009. doi: 10.1111/j.1461-0248.2005.00792.x

Gustafsson, A. (1948). Polyploidy, life-form and vegetative reproduction. *Hereditas* 34, 1–22. doi: 10.1111/j.1601-5223.1948.tb02824.x

Hæggström, C. A., and Åström, H. (2005). *Allium oleraceum* (Alliaceae) in Finland: distribution, habitats and accompanying vascular plant species. *Mem. Soc. Fauna Flora Fenn.* 81, 1–18.

Halverson, K., Heard, S. B., Nason, J. D., and Stireman, J. O. (2008). Origins, distribution, and local co-occurrence of polyploid cytotypes in *Solidago altissima* (*Asteraceae*). *Am. J. Bot.* 95, 50–58. doi: 10.3732/ajb.95.1.50

Han, T. S., Zheng, Q. J., Onstein, R. E., Rojas-Andrés, B. M., Hauenschild, F., Muellner-Riehl, A. N., et al. (2020). Polyploidy promotes species diversification of *Allium* through ecological shifts. *New Phytol.* 225, 571–583. doi: 10.1111/nph.16098

Hanelt, P., Schultze-Motel, J., Fritsch, R., Kruse, J., Maaß, H. I., Ohle, H., et al. (1992). "Infrageneric grouping of *Allium* – the gatersleben approach," in The Genus *Allium* – Taxonomic Problems and Genetic Resources. eds P. Hanelt, K. Hammer, and H. Knüpffer, *Proceedings of the International Symposium* (Gatersleben: Institute of Plant Genetics and Crop Plant Research Gatersleben), 107–123.

Hanušová, K., Čertner, M., Urfus, T., Koutecký, P., Košnar, J., Rothfels, C. J., et al. (2019). Widespread co-occurrence of multiple ploidy levels in fragile ferns (*Cystopteris fragilis* complex; Cystopteridaceae) probably stems from similar ecology of cytotypes, their efficient dispersal and inter-ploidy hybridization. *Ann. Bot.* 123, 845–855. doi: 10.1093/aob/mcy219

Hegarty, M. J., and Hiscock, S. J. (2008). Genomic clues to the evolutionary success of polyploid plants. *Curr. Biol.* 18, 435–444. doi: 10.1016/j.cub.2008.03.043

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* 12:e0169748. doi: 10.1371/journal.pone.0169748

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., et al. (2014). SoilGrids1km—global soil information based on automated mapping. *PLoS ONE* 9:e105992. doi: 10.1371/journal.pone.0105992

Herben, T., Suda, J., and Klimešová, J. (2017). Polyploid species rely on vegetative reproduction more than diploids: a re-examination of the old hypothesis. *Ann. Bot.* 120, 341–349. doi: 10.1093/aob/mcx009

Hewitt, G. M. (1999). Postglacial recolonization of European biota. *Biol. J. Linn. Soc.* 68, 87–112. doi: 10.1111/j.1095-8312.1999.tb01160.x

Hijmans, R. J., van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., et al. (2017). *Package 'raster'*. Available online at: https://cran.r-project.org/package=raster (accessed October 06, 2020).

Hintze, J. (2013). NCSS 9. NCSS, LLC. Kaysville, UT. Available online at: www.ncss.com (accessed October 06, 2020).

Hojsgaard, D., and Hörandl, E. (2019). The rise of apomixis in natural plant populations. *Front. Plant Sci.* 10, 358–358. doi: 10.3389/fpls.2019.00358

Hörandl, E. (2006). The complex causality of geographical parthenogenesis. *New Phytol.* 171, 525–538. doi: 10.1111/j.1469-8137.2006.01769.x

Hörandl, E. (2009). A combinational theory for maintenance of sex. *Heredity* 103, 445–457. doi: 10.1038/hdy.2009.85

Huntley, B., and Birks, H. J. B. (1983). *An atlas of Past and Present Pollen Maps for Europe*. Cambridge: Cambridge University Press.

Husband, B. C., Baldwin, S. J., and Suda, J. (2013). "The incidence of polyploidy in natural plant populations: major patterns and evolutionary processes," in *Flow Cytometry With Plant Cells: Analysis of Genes, Chromosomes and Genome*s, eds J. Doležel, J. Greilhuber, and J. Suda (Weinheim: Wiley-VCH), 255–276. doi: 10.1007/978-3-7091-1160-4_16

Jauzein, P., and Tison, J. M. (2001). Étude analytique du genre *Allium* L. sous-genre *Codonoprasum* (Reichenb) zahar section *Codonoprasum* Rechenb en France. *J. Bot. Soc. Bot. France* 15, 29–50.

Ježilová, E., Nožková-Hlaváčková, V., and Duchoslav, M. (2015), Photosynthetic characteristics of of three ploidy levels of *Allium oleraceum* L. (Amaryllidaceae) differing in ecological amplitude. *Plant Spec. Biol.* 30, 212–224. doi: 10.1111/1442-1984.12053

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916

Jírová, A. (2007). *Reproductive biology and phenology of allium oleraceum polyploid complex* (MSc thesis). Palacký University Olomouc, Olomouc, Czech Republic.

Johnson, A. L., Govindarajulu, R., and Ashman, T.-L. (2014). Bioclimatic evaluation of geographical range in *Fragaria* (Rosaceae): consequences of variation in breeding system, ploidy and species age. *Bot. J. Linn. Soc.* 176, 99–114. doi: 10.1111/boj.12190

Kao, R. H. (2007). Asexuality and coexistence of cytotypes. *New Phytol.* 175, 764–772. doi: 10.1111/j.1469-8137.2007.02145.x

Kao, R. H. (2008). Origins and widespread distribution of co-existing polyploids in *Arnica cordifolia* (Asteraceae). *Ann. Bot.* 101, 145–152. doi: 10.1093/aob/mcm271

Karpavičienė, B. (2007). Chromosome numbers of *Allium* from Lithuania. *Ann. Bot. Fenn.* 44, 345–352.

Karpavičienė, B. (2012). Morphological, reproductive and karyological variability in *Allium oleraceum* in Lithuania. *Biologia* 67, 278–283. doi: 10.2478/s11756-012-0003-3

Karpavičienė, B., and Karanauskaitė, D. (2010). Variation in reproductive modes of *Allium oleraceum*, *A. scorodoprasum* and *A. vineale* in field collection. *Acta Biologica Universitatis Daugavpiliensis* 10, 1–9.

Karunarathne, P., Schedler, M., Martínez, E. J., Honfi, A. I., Novichkova, A., and Hojsgaard, D. (2018). Intraspecific ecological niche divergence and reproductive shifts foster cytotype displacement and provide ecological opportunity to polyploids. *Ann. Bot.* 121, 1183–1196. doi: 10.1093/aob/mcy004

Kearney, M. (2005). Hybridization, glaciation and geographical parthenogenesis. *Trends Ecol. Evol.* 20, 495–502. doi: 10.1016/j.tree.2005.06.005

Kirchheimer, B., Schinkel, C. C. F., Dellinger, A. S., Klatt, S., Moser, D., Winkler, M., et al. (2016). A matter of scale: apparent niche differentiation of diploid and

tetraploid plants may depend on extent and grain of analysis. *J. Biogeogr.* 43, 716–726. doi: 10.1111/jbi.12663

Kliber, A., and Eckert, C. G. (2005). Interaction between founder effect and selection during biological invasion in an aquatic plant. *Evolution* 59, 1900–1913. doi: 10.1554/05-253.1

Kolář, F., Čertner, M., Suda, J., Schönswetter, P., and Husband, B. C. (2017). Mixed-ploidy species: progress and opportunities in polyploid research. *Trends Plant Sci.* 22, 1041–1055. doi: 10.1016/j.tplants.2017.09.011

Kolář, F., Lučanová, M., Záveská, E., Fuxová, G., Mandáková, T., Španiel, S., et al. (2016). Ecological segregation does not drive the intricate parapatric distribution of diploid and tetraploid cytotypes of the *Arabidopsis arenosa* group (Brassicaceae), *Biol. J. Linn. Soc*. 119, 673–688. doi: 10.1111/bij.12479

Krahulcová, A., and Jarolímová, V. (1993). Ecology of two cytotypes of *Butomus umbellatus* I. *Karyology and breeding behaviour. Folia Geobot. Phytotax*. 28, 385–411. doi: 10.1007/BF02853305

Krejčíková, J., Sudová, R., Lučanová, M., Trávníček, P., Urfus, T., Vít, P., et al. (2013). High ploidy diversity and distinct patterns of cytotype distribution in a widespread species of *Oxalis* in the greater cape floristic region. *Ann. Bot.* 111, 641–649. doi: 10.1093/aob/mct030

Laport, R. G., Hatem, L., Minckley, R. L., and Ramsey, J. (2013). Ecological niche modeling implicates climatic adaptation, competitive exclusion, and niche conservatism among *Larrea tridentata* cytotypes in North American deserts. *J. Torrey Bot. Soc.* 140, 349–364. doi: 10.3159/TORREY-D-13-00009.1

Legendre, P., and Anderson, M. J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* 69, 1–24. doi: 10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2

Legendre, P., and Legendre, L. (1998). *Numerical Ecology. 2nd Edn*. Amsterdam: Elsevier.

Leitch, A. R., and Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science* 320, 481–483. doi: 10.1126/science.1153585

Leitch, I. J., Chase, M. W., and Bennett, M. D. (1998). Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. *Ann. Bot.* 82, 85–94. doi: 10.1006/anbo.1998.0783

Levan, A. (1933). Cytological studies in *Allium*, III *Allium carinatum* and *Allium oleraceum*. *Hereditas* 18, 101–114. doi: 10.1111/j.1601-5223.1933.tb02602.x

Levan, A. (1937). Cytological studies in *Allium paniculatum* group. *Hereditas* 23, 317–370. doi: 10.1111/j.1601-5223.1937.tb02671.x

Levin, D. A. (1975). Minority cytotype exclusion in local plant populations. *Taxon* 24, 35–43. doi: 10.2307/1218997

Levin, D. A. (2002). *The Role of Chromosomal Change in Plant Evolution*. Oxford: Oxford University Press.

Levin, D. A. (2020). Has the polyploid wave ebbed? *Front. Plant Sci.* 11:251. doi: 10.3389/fpls.2020.00251

Lewis, W. H. (1967). Cytocatalytic evolution in plants. *Bot. Rev.* 33, 105–115. doi: 10.1007/BF02858665

Lewis, W. H. (1980). *Polyploidy, Biological Relevance*. New York, NY: Plenum Press. doi: 10.1007/978-1-4613-3069-1

Lindgren, A., Hugelius, G., Kuhry, P., Christensen, T. R., and Vandenberghe, J. (2016). GIS-based maps and area estimates of northern hemisphere permafrost extent during the last glacial maximum. *Permafrost Periglac. Process.* 27, 6–16. doi: 10.1002/ppp.1851

Lo, E. Y. Y., Stefanović, S., and Dickinson, T. A. (2013). Geographical parthenogenesis in Pacific Northwest hawthorns (*Crataegus*; Rosaceae). *Botany* 91, 107–116. doi: 10.1139/cjb-2012-0073

López-Jurado, J., Mateos-Naranjo, E., and Balao, F. (2019). Niche divergence and limits to expansion in the high polyploid *Dianthus broteri* complex. *New Phytol.* 222, 1076–1087. doi: 10.1111/nph.15663

Mandáková, T., and Münzbergová, Z. (2006). Distribution and ecology of *Aster amellus* aggregates in the Czech republic. *Ann. Bot.* 98, 845–856. doi: 10.1093/aob/mcl165

Manzaneda, A. J., Rey, P. J., Bastida, J. M., Weiss-Lehman, C., Raskin, E., and Mitchell-Olds, T. (2012). Environmental aridity is associated with cytotype segregation and polyploidy occurrence in *Brachypodium distachyon* (Poaceae). *New Phytol.* 193, 797–805. doi: 10.1111/j.1469-8137.2011.03988.x

Marchant, D. B., Soltis, D. E., and Soltis, P. S. (2016). Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *New Phytol.* 212, 708–718. doi: 10.1111/nph.14069

Marhold, K., Kudoh, H., Pak, J.-H., Watanabe, K., Španiel, S., and Lihová, J. (2010). Cytotype diversity and genome size variation in eastern Asian polyploid *Cardamine* (Brassicaceae) species. *Ann. Bot.* 105, 249–264. doi: 10.1093/aob/mcp282

Martin, S. L., and Husband, B. (2009). Influence of phylogeny and ploidy on species ranges of North American angiosperms. *J. Ecol.* 97, 913–922. doi: 10.1111/j.1365-2745.2009.01543.x

McAllister, C., Blaine, R., Kron, P., Bennett, B., Garrett, H., Kidson, J., et al. (2015). Environmental correlates of cytotype distribution in *Andropogon gerardii* (Poaceae). *Am. J. Bot.* 102, 92–102. doi: 10.3732/ajb.1400296

McCormack, J. E., Zellmer, A. J., and Knowles, L. L. (2010). Does niche divergence accompany allopatric divergence in *Aphelocoma* Jays as predicted under ecological speciation?: insights from tests with niche models. *Evolution* 64, 1231–1244. doi: 10.1111/j.1558-5646.2009.00900.x

McCune, B., and Keon, D. (2002). Equations for potential annual direct incident radiation and heat load. *J. Veg. Sci.* 13, 603–606. doi: 10.1111/j.1654-1103.2002.tb02087.x

McIntyre, P. J. (2012). Polyploidy associated with altered and broader ecological niches in the *Claytonia perfoliata* (Portulacaceae) species complex. *Am. J. Bot.* 99, 655–662. doi: 10.3732/ajb.1100466

Médail, F., and Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *J. Biogeogr.* 36, 1333–1345. doi: 10.1111/j.1365-2699.2008.02051.x

Meusel, H., Jäger, E., and Weinert, E. (1965). *Vergleichende Chorologie der zentraleuropäischen Flora*. Jena: Gustav Fischer Verlag.

Mock, K. E., Callahan, C. M., Islam-Faridi, M. N., Shaw, J. D., Rai, H. S., et al. (2012). Widespread triploidy in western north american aspen (*Populus tremuloides*). *PLoS ONE* 7:e48406. doi: 10.1371/journal.pone.0048406

Mráz, P., and Ronikier, M. (2016). Biogeography of the Carpathians: evolutionary and spatial facets of biodiversity. *Biol. J. Linn. Soc.* 119, 528–559. doi: 10.1111/bij.12918

Mucina, L., Bültmann, H., Dierßen, K., Theurillat, J.-P., Raus, T., Carni, A., et al. (2016). Vegetation of Europe: hierarchical floristic classification system of vascular plant, bryophyte, lichen, and algal communities. *Appl. Veg. Sci.* 19, 3–264. doi: 10.1111/avsc.12257

Muñoz-Pajares, A. J., Perfectti, F., Loureiro, J., Abdelaziz, M., Biella, P., Castro, M., et al. (2018). Niche differences may explain the geographic distribution of cytotypes in *Erysimum mediohispanicum*. *Plant Biol.* 20, 139–147. doi: 10.1111/plb.12605

Němečková, H., Krak, K., and Chrtek, J. (2019). Complex pattern of ploidal and genetic variation in *Seseli libanotis* (Apiaceae). *Ann. Bot. Fenn.* 56, 57–77. doi: 10.5735/085.056.0111

Otto, S. P., and Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.* 34, 401–437. doi: 10.1146/annurev.genet.34.1.401

Pandit, M. K., Pocock, M. J. O., and Kunin, W. E. (2011). Ploidy influences rarity and invasiveness in plants. *J. Ecol.* 99, 1108–1115. doi: 10.1111/j.1365-2745.2011.01838.x

Pandit, M. K., Tan, H. T. W., and Bisht, M. S. (2006). Polyploidy in invasive plant species of Singapore. *Bot. J. Linn. Soc.* 151, 395–403. doi: 10.1111/j.1095-8339.2006.00515.x

Pandit, M. K., White, S. M., and Pocock, M. J. O. (2014). The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: a global analysis. *New Phytol.* 203, 697–703. doi: 10.1111/nph.12799

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

Pastor, J., and Valdés, B. (1983). *Revisión del género Allium (Liliaceae) en peninsula Ibérica e Islas Baleares*. Sevilla: Universidad de Sevilla Press.

Paule, J., Sharbel, T. F., and Dobeš, C. (2011). Apomictic and sexual lineages of the *Potentilla argentea* L. group (Rosaceae) – cytotype and molecular genetic differentiation. *Taxon* 60, 721–732. doi: 10.1002/tax.603008

Peruzzi, L., Carta, A., and Altinordu, F. (2017). Chromosome diversity and evolution in *Allium* (Allioideae, Amaryllidaceae). *Plant Biosyst.* 151, 212–220. doi: 10.1080/11263504.2016.1149123

Petit, C., Bretagnolle, F., and Felber, F. (1999). Evolutionary consequences of diploid-polyploid hybrid zones in wild species. *Trends Ecol. Evol.* 14, 306–311. doi: 10.1016/S0169-5347(99)01608-0

Petitpierre, B., Kueffer, C., Broennimann, O., Randin, C., Daehler, C., and Guisan, A. (2012). Climatic niche shifts are rare among terrestrial plant invaders. *Science* 335, 1344–1348. doi: 10.1126/science.1215933

R Development Core Team (2014). *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramsey, J. (2011). Polyploidy and ecological adaptation in wild yarrow. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7096–7101. doi: 10.1073/pnas.1016631108

Ramsey, J., and Ramsey, T. S. (2014). Ecological studies of polyploidy in the 100 years following its discovery. *Philos. Trans. R. Soc. B Biol. Sci.* 369:20130352. doi: 10.1098/rstb.2013.0352

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Ramsey, J., and Schemske, D. W. (2002). Neopolyploidy in flowering plants. *Annu. Rev. Ecol. Evol. Syst.* 33, 589–639. doi: 10.1146/annurev.ecolsys.33.010802.150437

Rausch, J., and Morgan, M. T. (2005). The effects of self-fertilization, inbreeding depression, and population size on autopolyploid establishment. *Evolution* 59, 1867–1875. doi: 10.1554/05-095.1

Rejlová, L., Chrtek, J., Trávníček, P., Lučanová, M., Vít, P., and Urfus, T. (2019). Polyploid evolution: the ultimate way to grasp the nettle. *PLoS ONE* 14:e0218389. doi: 10.1371/journal.pone.0218389

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N. M., Salman-Minkov, A., et al. (2015). The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytol.* 206, 19–26. doi: 10.1111/nph.13191

Rice, A., Šmarda, P., Novosolov, M., Drori, M., Glick, L., Sabath, N., et al. (2019). The global biogeography of polyploid plants. *Nat. Ecol. Evol.* 3, 265–273. doi: 10.1038/s41559-018-0787-9

Rieseberg, L. H., and Willis, J. H. (2007). Plant speciation. *Science* 317, 910–914. doi: 10.1126/science.1137729

Rojas-Andrés, B. M., Padilla-García, N., de Pedro, M., López-González, N., Delgado, L., Albach, D. C., et al. (2020). Environmental differences are correlated with the distribution pattern of cytotypes in *Veronica* subsection *Pentasepalae* at a broad scale. *Ann. Bot.* 125, 471–484. doi: 10.1093/aob/mcz182

Ronsheim, M. L. (1994). Dispersal distances and predation rates of sexual and asexual propagules of *Allium vineale*. *Am. Midl. Nat.* 131, 55–64. doi: 10.2307/2426608

Šafářová, L., Duchoslav, M., Jandová, M., and Krahulec, F. (2011). *Allium oleraceum* in Slovakia: cytotype distribution and ecology. *Preslia* 83, 513–527.

Šafářová, L., and Duchoslav, M. (2010). Cytotype distribution in mixed populations of polyploid *Allium oleraceum* measured at a microgeographic scale. *Preslia* 82, 107–126.

Salmeri, C., Brullo, C., Brullo, S., Giusso del Galdo, G., and Moysiyenko, I. I. (2016). What is *Allium paniculatum*? Establishing taxonomic and molecular phylogenetic relationships within *A.* sect. *Codonoprasum*. *J. Syst. Evol.* 5, 123–135. doi: 10.1111/jse.12170

Schinkel, C. C. F., Kirchheimer, B., Dellinger, A. S., Klatt, S., Winkler, M., Dullinger, S., et al. (2016). Correlations of polyploidy and apomixis with elevation and associated environmental gradients in an alpine plant. *AoB Plants* 8:plw064. doi: 10.1093/aobpla/plw064

Schoener, T. W. (1968). The anolis lizards of bimini: resource partitioning in a complex fauna. *Ecology* 49: 704–726. doi: 10.2307/1935534

Segraves, K. A., and Anneberg, T. J. (2016). Species interactions and plant polyploidy. *Am. J. Bot.* 103, 1326–1335. doi: 10.3732/ajb.1500529

Sheth, S. N., Morueta-Holme, N., and Angert, A. L. (2020). Determinants of geographic range size in plants. *New Phytol.* 226, 650–665. doi: 10.1111/nph.16406

Šingliarová, B., Zozomová-Lihová, J., and Mráz, P. (2019). Polytopic origin and scale-dependent spatial segregation of cytotypes in primary diploid–autopolyploid contact zones of *Pilosella rhodopea* (Asteraceae). *Biol. J. Linn. Soc.* 126, 360–379. doi: 10.1093/biolinnean/bly199

Šmarda, P., Hejcman, M., Brezinová, A., Horová, L., Steigerová, H., Zedek, F., et al. (2013). Effect of phosphorus availability on the selection of species with different ploidy levels and genome sizes in a long-term grassland fertilization experiment. *New Phytol.* 200, 911–921. doi: 10.1111/nph.12399

Solhaug, E. M., Ihinger, J., Jost, M., Gamboa, V., Marchant, B., Bradford, D., et al. (2016). Environmental regulation of heterosis in the allopolyploid *Arabidopsis suecica*. *Pl. Physiol.* 170, 2251–2263. doi: 10.1104/pp.16.00052

Soltis, D. E., Buggs, R. J. A., Doyle, J. J., and Soltis, P. S. (2010). What we still don't know about polyploidy. *Taxon* 59, 1387–1403. doi: 10.1002/tax.595006

Soltis, D. E., and Soltis, P. S. (1999). Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.* 14, 348–352. doi: 10.1016/S0169-5347(99)01638-9

Soltis, D. E., Soltis, P. S., and Tate, J. A. (2004). Advances in the study of polyploidy since Plant speciation. *New Phytol.* 161, 173–191. doi: 10.1046/j.1469-8137.2003.00948.x

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103, 1146–1166. doi: 10.3732/ajb.1500501

Soltis, P. S., and Soltis, D. E. (1995). The dynamic nature of polyploid genomes. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8089–8091. doi: 10.1073/pnas.92.18.8089

Soltis, P. S., and Soltis, D. E. (2000). The role of genetic and genomic attributes in the success of polyploids. *Proc. Natl. Acad. Sci. U.S.A.* 97, 7051–7057. doi: 10.1073/pnas.97.13.7051

Soltis, P. S., and Soltis, D. E. (eds.). (2012). *Polyploidy and Genome Evolution*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-31442-1

Sonnleitner, M., Flatscher, R., García, P. E., Rauchová, J., Suda, J., Schneeweiss, G. M., et al. (2010). Distribution and habitat segregation on different spatial scales among diploid, tetraploid and hexaploid cytotypes of *Senecio carniolicus* (Asteraceae) in the Eastern Alps. *Ann. Bot.* 106, 967–977. doi: 10.1093/aob/mcq192

Stearn, W. T. (1980). "*Allium* L.," in *Flora Europaea Vol. 5*, eds T. G. Tutin, V. H. Heywood, N. A. Burges, D. M. Moore, D. H. Valentine, S. M. Walters (Cambridge: Cambridge University Press), 49–69.

Stebbins, G. L. (1984). Polyploidy and the distribution of the arctic-alpine flora: new evidence and a new approach. *Bot. Helv.* 94, 1–13.

Stebbins, G. L. (1985). Polyploidy, hybridization, and the invasion of new habitats. *Ann. Missouri Bot.* 72, 824–832. doi: 10.2307/2399224

Stewart, J. R., and Lister, A. M. (2001). Cryptic northern refugia and the origins of modern biota. *Trends Ecol. Evol.* 16, 608–613. doi: 10.1016/S0169-5347(01)02338-2

Stewart, J. R., Lister, A. M., Barnes, I., and Dalén, L. (2010). Refugia revisited: individualistic responses of species in space and time. *Proc. Roy. Soc. B Biol. Sci.* 277, 661–671. doi: 10.1098/rspb.2009.1272

Suda, J., Krahulcová, A., Trávníček, P., and Krahulec, F. (2006). Ploidy level versus DNA ploidy level: an appeal for consistent terminology. *Taxon* 55, 447–450. doi: 10.2307/25065591

Sutherland, B. L., and Galloway, L. F. (2017). Postzygotic isolation varies by ploidy level within a polyploid complex. *New Phytol.* 213, 404–412. doi: 10.1111/nph.14116

te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubešová, M., et al. (2012). The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.*109, 19–45. doi: 10.1093/aob/mcr277

ter Braak, C. J. F., and Šmilauer, P. (2012). *CANOCO Reference Manual and User's Guide: Software for Ordination (Version 5.0)*. Wageningen: Biometris.

Theodoridis, S., Randin, C., Broennimann, O., Patsiou, T., and Conti, E. (2013). Divergent and narrower climatic niches characterize polyploid species of European primroses in *Primula* sect. *Aleuritia*. *J. Biogeogr.* 40, 1278–1289. doi: 10.1111/jbi.12085

Thompson, J., and Lumaret, R. (1992). The evolutionary dynamics of polyploid plants: origins, establishment and persistence. *Trends Ecol. Evol.* 7, 302–307. doi: 10.1016/0169-5347(92)90228-4

Tilquin, A., and Kokko, H. (2016). What does the geography of parthenogenesis teach us about sex? *Phil. Trans. R. Soc. B* 371:20150538. doi: 10.1098/rstb.2015.0538

Tison, J. M., and de Foucault, B. (2014). *Flora Gallica. Flore de France*. Mèze: Biotope.

Trabucco, A., and Zomer, R. (2019). *Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2.* doi: 10.6084/m9.figshare.7504448.v3

Trávníček, P., Dočkalová, Z., Rosenbaumová, R., Kubátová, B., Szelag, Z., and Chrtek, J. (2011a). Bridging global and microregional scales: ploidy distribution in *Pilosella echioides* (Asteraceae) in central Europe. *Ann. Bot.* 107, 443–454. doi: 10.1093/aob/mcq260

Trávníček, P., Kubátová, B., Čurn, V., Rauchová, J., and Krajníková, E., Jersáková, J., et al. (2011b). Remarkable coexistence of multiple cytotypes of the *Gymnadenia conopsea* aggregate (the fragrant orchid): evidence from flow cytometry. *Ann. Bot.* 107, 77–87. doi: 10.1093/aob/mcq217

Trávníček, P., Eliášová, A., and Suda, J. (2010). The distribution of cytotypes of *Vicia cracca* in Central Europe: the changes that have occurred over the last four decades. *Preslia* 82, 149–163.

Treier, U. A., Broennimann, O., Normand, S., Guisan, A., Schaffner, U., Steinger, T., et al. (2009). Shift in cytotype frequency and niche space in the invasive plant *Centaurea maculosa*. *Ecology* 90: 1366–1377. doi: 10.1890/08-0420.1

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26

van Dijk, P., Hartog, M., and van Delden, W. (1992). Single cytotype areas in autopolyploid *Plantago media* L. *Biol. J. Linn. Soc.* 46, 315–331. doi: 10.1111/j.1095-8312.1992.tb00867.x

Vandenberghe, J., French, H. M., Gorbunov, A., Marchenko, S., Velichko, A. A., Jin, H., et al. (2014). The last permafrost maximum (LPM) map of the Northern hemisphere: permafrost extent and mean annual air temperatures, 25-17 ka BP. *Boreas* 43, 652–666. doi: 10.1111/bor.12070

Veselý, P., Bureš, P., Šmarda, P., and Pavlíček, T. (2012). Genome size and DNA base composition of geophytes: the mirror of phenology and ecology? *Ann. Bot.* 109, 65–75. doi: 10.1093/aob/mcr267

Visger, C. J., Germain-Aubrey, C. C., Patel, M., Sessa, E. B., Soltis, P. S., and Soltis, D. E. (2016). Niche divergence between diploid and autotetraploid *Tolmiea*. *Am. J. Bot.* 103, 1–11. doi: 10.3732/ajb.1600130

Visser, V., and Molofsky, J. (2015). Ecological niche differentiation of polyploidization is not supported by environmental differences among species in a cosmopolitan grass genus. *Am. J. Bot.* 102, 36–49. doi: 10.3732/ajb.1400432

Vosa, C. G. (1976). Heterochromatic banding patterns in *Allium*. II. Heterochromatin variation in species of the paniculatum group. *Chromosoma* 57, 119–133. doi: 10.1007/BF00292911

Vrijenhoek, R. C. (1994). Unisexual fish: model systems for studying ecology and evolution. *Annu. Rev. Ecol. Syst.* 25, 71–96. doi: 10.1146/annurev.es.25.110194.000443

Vvedenskii, A. (1935). "Genus *Allium* L.," in *Flora U.S.S.R., Vol.4, Liliiflorae and Microspermae*, ed V. Komarov (Leningrad: Izdatelstvo Akademii Nauk SSSR), 112–280.

Warren, D. L., Glor, R. E., and Turelli, M. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62, 2868–2883. doi: 10.1111/j.1558-5646.2008.00482.x

Warren, D. L., Glor, R. E., and Turelli, M. (2010). ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography* 33, 607–611. doi: 10.1111/j.1600-0587.2009.06142.x

Weiss-Schneeweiss, H., Emadzade, K., Jang, T. S., and Schneeweiss, G. M. (2013). Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenet. Genome Res.* 140, 137–150. doi: 10.1159/000351727

Wendel, J. (2000). Genome evolution in polyploids. *Pl. Molec. Biol.* 42, 225–249. doi: 10.1023/A:1006392424384

Wiens, J. J., Ackerly, D. D., Allen, A. P., Anacker, B. L., Buckley, L. B., Cornell, H. V., et al. (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecol. Lett.* 13, 1310–1324. doi: 10.1111/j.1461-0248.2010.01515.x

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13875–13879. doi: 10.1073/pnas.0811575106

Wos, G., Morkovská, J., Bohutínská, M., Šrámková, G., Knotek, A., Lučanová, M., et al. (2019). Role of ploidy in colonization of alpine habitats in natural populations of *Arabidopsis arenosa*. *Ann. Bot.* 124, 255–268. doi: 10.1093/aob/mcz070

Wu, L.-L., Cui, X.-K., Milne, R. I., Sun, Y.-S., and Liu, J.-Q. (2010). Multiple autopolyploidizations and range expansion of *Allium przewalskianum* Regel. (Alliaceae) in the Qinghai-Tibetan plateau. *Molec. Ecol.* 19, 1691–1704. doi: 10.1111/j.1365-294X.2010.04613.x

Yamauchi, A., Hosokawa, A., Nagata, H., and Shimoda, M. (2004). Triploid bridge and role of parthenogenesis in the evolution of autopolyploidy. *Am. Nat.* 164, 101–112. doi: 10.1086/421356

Yoo, M.-J., Liu, X., Pires, J. C., Soltis, P. S., and Soltis, D. E. (2014). Nonadditive gene expression in polyploids. *Annu. Rev. Genet.* 48, 485–517. doi: 10.1146/annurev-genet-120213-092159

Zar, J. H. (1996). *Biostatistical Analysis. 4th Edn*. New Jersey, NJ: Prentice Hall.

Zólyomi, B., and Fekete, G. (1994). The pannonian loess steppe: differentiation in space and time. *Abstr. Bot.*, 18, 29–41.

Zozomová-Lihová, J., Krak, K., Mandáková, T., Shimizu, K. K., Španiel, S., Vít, P., et al. (2014). Multiple hybridization events in *Cardamine* (Brassicaceae) during the last 150 years: Revisiting a textbook example of neoallopolyploidy. *Ann. Bot.* 113, 817–830. doi: 10.1093/aob/mcu012

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 18 December 2020
doi: 10.3389/fpls.2020.588856

So Closely Related and Yet So Different: Strong Contrasts Between the Evolutionary Histories of Species of the *Cardamine pratensis* Polyploid Complex in Central Europe

Andrea Melichárková[1†], Marek Šlenker[1†], Judita Zozomová-Lihová[1†], Katarína Skokanová[1], Barbora Šingliarová[1], Tatiana Kačmárová[1], Michaela Caboňová[1], Matúš Kempa[1], Gabriela Šrámková[2], Terezie Mandáková[3,4], Martin A. Lysák[3,5], Marek Svitok[6,7], Lenka Mártonfiová[8] and Karol Marhold[1,2]*

[1] Institute of Botany, Plant Science and Biodiversity Centre, Slovak Academy of Sciences, Bratislava, Slovakia, [2] Department of Botany, Faculty of Science, Charles University, Prague, Czechia, [3] Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czechia, [4] Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czechia, [5] National Centre for Biomolecular Research (NCBR), Faculty of Science, Masaryk University, Brno, Czechia, [6] Department of Biology and General Ecology, Faculty of Ecology and Environmental Sciences, Technical University in Zvolen, Zvolen, Slovakia, [7] Department of Ecosystem Biology, Faculty of Science, University of South Bohemia, České Budějovice, Czechia, [8] Botanical Garden, Pavol Jozef Šafárik University, Košice, Slovakia

OPEN ACCESS

Edited by:
Carl J. Rothfels,
University of California, Berkeley,
United States

Reviewed by:
Katherine Waselkov,
California State University, Fresno,
United States
Rebecca Stubbs,
University of Zürich, Switzerland

***Correspondence:**
Karol Marhold
karol.marhold@savba.sk

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: 29 July 2020
Accepted: 19 November 2020
Published: 18 December 2020

Citation:
Melichárková A, Šlenker M, Zozomová-Lihová J, Skokanová K, Šingliarová B, Kačmárová T, Caboňová M, Kempa M, Šrámková G, Mandáková T, Lysák MA, Svitok M, Mártonfiová L and Marhold K (2020) So Closely Related and Yet So Different: Strong Contrasts Between the Evolutionary Histories of Species of the Cardamine pratensis Polyploid Complex in Central Europe. Front. Plant Sci. 11:588856. doi: 10.3389/fpls.2020.588856

Recurrent polyploid formation and weak reproductive barriers between independent polyploid lineages generate intricate species complexes with high diversity and reticulate evolutionary history. Uncovering the evolutionary processes that formed their present-day cytotypic and genetic structure is a challenging task. We studied the species complex of *Cardamine pratensis*, composed of diploid endemics in the European Mediterranean and diploid-polyploid lineages more widely distributed across Europe, focusing on the poorly understood variation in Central Europe. To elucidate the evolution of Central European populations we analyzed ploidy level and genome size variation, genetic patterns inferred from microsatellite markers and target enrichment of low-copy nuclear genes (Hyb-Seq), and environmental niche differentiation. We observed almost continuous variation in chromosome numbers and genome size in *C. pratensis* s.str., which is caused by the co-occurrence of euploid and dysploid cytotypes, along with aneuploids, and is likely accompanied by inter-cytotype mating. We inferred that the polyploid cytotypes of *C. pratensis* s.str. are both of single and multiple, spatially and temporally recurrent origins. The tetraploid *Cardamine majovskyi* evolved at least twice in different regions by autopolyploidy from diploid *Cardamine matthioli*. The extensive genome size and genetic variation of *Cardamine rivularis* reflects differentiation induced by the geographic isolation of disjunct populations, establishment of triploids of different origins, and hybridization with sympatric *C. matthioli*. Geographically structured genetic lineages identified in the species under study, which are also ecologically divergent, are interpreted as descendants from different source populations in multiple glacial refugia. The postglacial range expansion was accompanied by substantial genetic admixture between the lineages of *C. pratensis* s.str., which is reflected by diffuse borders in

their contact zones. In conclusion, we identified an interplay of diverse processes that have driven the evolution of the species studied, including allopatric and ecological divergence, hybridization, multiple polyploid origins, and genetic reshuffling caused by Pleistocene climate-induced range dynamics.

Keywords: Brassicaceae, environmental niche, genome size, hybridization, microsatellites, phylogeography, polyploidy, target enrichment

INTRODUCTION

Polyploidy is a widespread evolutionary phenomenon and a major mechanism of sympatric speciation in plants (Soltis et al., 2003; Coyne and Orr, 2004; Soltis and Soltis, 2009; Soltis et al., 2014). Diploids and their polyploid progeny often coexist at least in initial phases, although later some geographical or ecological shifts may evolve (Soltis and Soltis, 2009; Parisod et al., 2010; Soltis et al., 2014; see e.g., Hülber et al., 2015; Arrigo et al., 2016). Although reproductive isolation is assumed between diploids and related polyploids, allowing for independent evolution and speciation of polyploid lineages even in sympatry, it may be incomplete and permits some inter-ploidy gene flow (Sonnleitner et al., 2013; Kolář et al., 2017; Sutherland and Galloway, 2017; Baduel et al., 2018). Recurrent polyploid formation and weak reproductive barriers between cytotypes or polyploid lineages generate intricate polyploid species complexes with reticulate evolutionary histories (e.g., Bardy et al., 2010; Ma et al., 2010; Frajman et al., 2016; Španiel et al., 2017; Mandák et al., 2018; Padilla-García et al., 2018; Melichárková et al., 2019). They frequently show weak genetic separation among polyploid species, discrepancies between morphological and genetic patterns, and shallow, largely unresolved phylogenetic structuring. Several studies of polyploid species complexes in Europe indicate their recent diversification, which has been dated to the Pliocene and, especially, the Pleistocene, driven by repeated cycles of glaciation-induced range shifts, and population isolation in refugia followed by range expansion and secondary contact (Franzke and Hurka, 2000; Bardy et al., 2010; Pachschwöll et al., 2015; Frajman et al., 2016; Dauphin et al., 2018; Melichárková et al., 2019; Rojas-Andrés et al., 2020). High species and genetic diversity has repeatedly been observed in Southern Europe, reflecting allopatric long-term survival in stable glacial refugia and only small-scale range shifts (Nieto Feliner, 2014), whereas a highly dynamic glacial and postglacial history can be expected in Central Europe, shaped by colonization of different lineages from southern refugia, their admixture in contact zones, as well as population survival and expansion from cryptic northern refugia (Hewitt, 2001; Birks and Willis, 2008; Stewart et al., 2010). Therefore, understanding the evolutionary processes that generated the present-day variation patterns, cytotypic and genetic structure in European polyploid species complexes is a challenging task.

To uncover evolutionary processes underlying high diversity in polyploid complexes, we focus on the *Cardamine pratensis* species complex, which is widespread in Europe and exhibits extensive but poorly resolved variation in Central Europe. It is one of the most complicated polyploid complexes of the family Brassicaceae, comprising sexually reproducing, allogamous perennials capable of vegetative propagation (Lövkvist, 1956) but with no reports of apomixis (which is extremely rare in this family, Brukhin et al., 2019). It includes a number of species and genetic lineages from the diploid to the dodecaploid level, with both aneuploids and dysploids (arisen via chromosomal rearrangements, Mandáková et al., 2013) documented, which makes it an excellent model system for addressing the reticulate evolutionary histories of polyploid complexes (Franzke and Hurka, 2000; Lihová and Marhold, 2006; Marhold et al., 2018 and references therein). Several well-defined and genetically distinct diploid endemic species occur in Southern Europe (Franzke and Hurka, 2000; Lihová et al., 2003; Lihová et al., 2004). Conversely, Central and Northern Europe harbor a series of less differentiated diploid and polyploid populations and species with unresolved relationships, putatively of postglacial origin (see Franzke and Hurka, 2000). Their taxonomic treatment is based on detailed morphological, chromosomal, and ecological analyses, as well as on hybridization experiments (Lövkvist, 1956; Urbanska-Worytkiewicz and Landolt, 1974; Marhold and Záborský, 1986; Marhold, 1994a,b; reviewed by Lihová and Marhold, 2006; Marhold et al., 2018; see also **Table 1**). In Central Europe, they include diploid *Cardamine matthioli*, tetraploid *Cardamine majovskyi*, highly polyploid *Cardamine dentata,* and the most complicated species *Cardamine pratensis* s.str., that includes diploid to heptaploid populations growing across most of Europe from lowlands to the alpine belt. Several attempts have been made to split this polymorphic species into more entities. Some authors recognized *Cardamine nemorosa* and *Cardamine udicola* in lowland to montane areas (Urbanska-Worytkiewicz and Landolt, 1974), and diploid populations from the foothills of the Eastern Carpathians were informally treated as a morphotype 'ucranica' (Marhold, 1994b, 1995). In addition, high-elevation Alpine and Eastern Carpathian populations were attributed to *C. rivularis* by Lövkvist (1956) and Urbanska-Worytkiewicz and Landolt (1974; see **Table 1**). *Cardamine rivularis*, however, is a morphologically and genetically different species occurring only in the Southern Carpathians (Romania) and in Bulgarian mountains (Marhold, 1995; Franzke and Hurka, 2000). Thus, the erroneously classified high-elevation Alpine and Eastern Carpathian populations later became informally referred to as *C. rivularis* auct. non Schur (e.g., Zozomová-Lihová et al., 2014a). Despite the observed ploidy level variation, morphological, and ecological diversity of these species, all genetic markers applied so far (i.e., allozymes, sequences of nrDNA, cpDNA, RAPDs, and AFLP markers; Franzke and Hurka, 2000; Lihová et al., 2003, 2004) have yielded only low resolution and failed to resolve relationships among Central

European populations of the *C. pratensis* complex and within *C. pratensis* s.str. in particular.

The aim of this paper is to disentangle the evolution of the *C. pratensis* complex in Central Europe (**Figure 1**), where much variation among the populations is observed, but with so far unresolved structure and relationships. We analyzed chromosome number, ploidy level and genome size variation, explored genetic patterns based on microsatellite markers and high-throughput target enrichment of low-copy nuclear genes (Hyb-Seq), and reconstructed species' environmental niches. Microsatellites are efficient, high-resolution genetic markers for both population- and species-level studies (Hodel et al., 2016; Li et al., 2017), and have also been used with success when addressing hybridization events and polyploid evolution (e.g., Sampson and Byrne, 2012; Zozomová-Lihová et al., 2014a, 2015; Feulner et al., 2019; Hu et al., 2019). The Hyb-Seq approach, which captures target exons with flanking intronic and intergenic regions (Weitemier et al., 2014), allows to resolve relationships at various evolutionary levels from the population level up (Villaverde et al., 2018; Tomasello et al., 2020), including polyploids (e.g., Carter et al., 2019). Our specific aims were to (1) resolve the cytotypic and genetic structure of the species under study, (2) identify intraspecific genetic lineages, potentially reflecting descendants from distinct glacial refugia and postglacial colonization routes, (3) explore if and to what extent the lineages correlate with the cytotypic variation and display different environmental niches, (4) resolve single or polytopic polyploid origins, (5) explore the patterns and extent of hybridization between species and lineages, and, finally, (6) to determine how the structure resolved corresponds with previous taxonomic treatments.

MATERIALS AND METHODS

Plant Material

The *C. pratensis* complex occupies wet meadows and pastures, stream banks, springs, and floodplain forests from lowlands up to the alpine belt. We sampled all species recognized in Central Europe (except for *C. dentata*, which was beyond the scope of the present study due to its high polyploidy), and representatively covered their geographic, elevational and cytotypic variation (**Figure 1** and **Supplementary Data 1**). They were identified based on plant morphology and ecology of the sites (mainly elevation), following the concepts of previous authors (see Introduction), with the aid of ploidy level estimation if necessary. In addition, *C. rivularis* growing in Southeastern Europe was also sampled and analyzed here for proper interpretation of the high-elevation Central European populations previously attributed to this species (**Table 1**), as well as due to its hybridization with *C. matthioli* reported by Ančev et al. (2013, see below).

Cardamine matthioli is a diploid ($2n = 16$) distributed in Central Europe and extending to Southeastern Europe (Marhold, 1994b, 2000). Aneuploids with supernumerary chromosomes ranging from 17 to 22 have been observed sporadically (Kučera et al., 2005). The species grows mainly in lowlands and up to the montane belt, and only exceptionally reaches elevations above

TABLE 1 | Overview of present and previous taxonomic treatments of the *Cardamine pratensis* species complex in Central Europe.

Marhold et al. (2018) and present study	Ploidy level (Kučera et al., 2005)	Distribution range	Lövkvist, 1956	Urbanska-Worytkiewicz and Landolt, 1974	Landolt, 1984	Marhold, 1994a,b	Franzke and Hurka, 2000
C. matthioli Moretti	2*x*	Central and SE Europe	*C. matthioli*	*C. matthioli*	*C. matthioli*	*C. matthioli*	*C. matthioli*
C. majovskyi Marhold and Záborský	4*x*	Central and SE Europe	–	–	–	*C. majovskyi*	*C. majovskyi*
C. pratensis L.	2*x*–7*x*	Europe, N Africa, N Asia, introduced to N America	*C. pratensis*	*C. pratensis*	*C. pratensis*	*C. pratensis*	*C. pratensis*
C. pratensis	2*x*	Germany, Switzerland, France (lower altitudes)	*C. pratensis*	*C. nemorosa* Lej.	*C. pratensis*	–	*C. pratensis*
C. pratensis	2*x*, 4*x*	Germany, Switzerland (lower altitudes)	*C. pratensis*	*C. udicola* Jord.	*C. udicola*	*C. pratensis*	*C. udicola*
C. pratensis	2*x*	E Carpathians (lower altitudes)	*C. pratensis*	–	–	*C. pratensis* "ucranica" type	–
C. pratensis	2*x*, 4*x*	Carpathians and Alps [(sub-)alpine belt]	*C. rivularis* Schur	*C. rivularis* Schur	*C. rivularis* Schur	*C. pratensis* "rivularis auct." type	*C. rivularis* auct. non Schur
C. rivularis Schur *	2*x*, 3*x*	S Carpathians and Bulgaria [(sub-)alpine belt]	–	–	–	*C. rivularis*	*C. rivularis*
C. dentata Schult. **	(7*x*)8*x* – 12*x*	Central and NW Europe	*C. palustris* Petermann	–	*C. palustris*	*C. dentata*	*C. dentata*

**A species growing only in Southeastern Europe, but previously confused with (sub-)alpine populations from the Eastern Carpathians and Alps; sampled and analyzed in the present study. **A species not included in the present study due to its high polyploidy.*

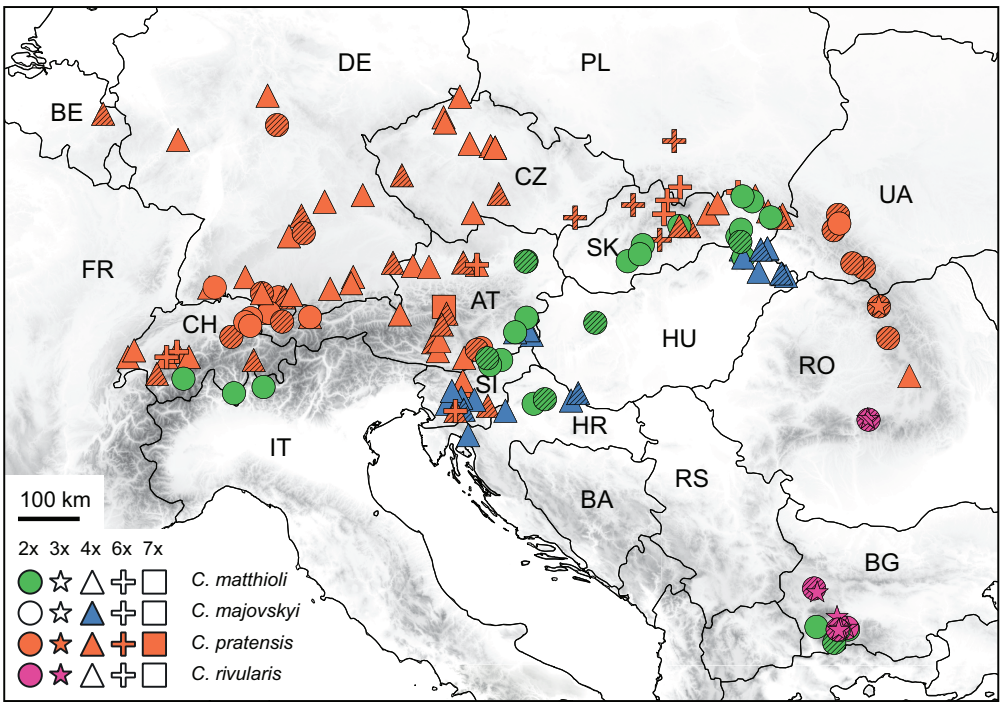

FIGURE 1 | Distribution and ploidy levels of populations of *Cardamine matthioli*, *C. majovskyi*, *C. pratensis*, and *C. rivularis* sampled in the present study. All population samples were used in flow cytometric and microsatellite analyses, while a selection was used in Hyb-Seq analyses (symbols highlighted by hatching). For details on the localities, see **Supplementary Data 1**.

1,000–1,500 m (Marhold, 1994b; Ančev, 2006). In total, 223 plants from 28 populations were collected.

Cardamine majovskyi is a tetraploid ($2n = 32$) so far reported from two disjunct areas in Central Europe, growing in lowlands and the montane belt up to 600 m above sea level (Marhold et al., 2018 and references therein). Its autotetraploid origin from *C. matthioli* was inferred (Franzke and Hurka, 2000; Lihová and Marhold, 2006; Marhold et al., 2018). Aneuploids with $2n = 34$ have been rarely observed (Marhold, 1994a). In total, 208 plants from 21 populations were collected.

Cardamine pratensis s.str. (further referred to without 's.str.') is widely distributed throughout most of Europe, except its southernmost areas (Marhold, 1994b; Lihová and Marhold, 2006). It grows across a large elevational span, from lowlands up to the alpine belt in the Alps, Carpathians, and the central Pyrenees. It comprises a number of cytotypes ranging from diploid to heptaploid, including also aneuploids and dysploids (Kučera et al., 2005; Mandáková et al., 2013). We sampled 959 plants from 87 populations originating in Central Europe and adjacent areas. These also included populations previously attributed to *C. nemorosa*, *C. udicola*, *C. rivularis* auct., and the 'ucranica' morphotype. In the case of these putative taxa, we specifically sampled the same populations or in regions referred to by previous authors (see **Supplementary Data 1** for details).

Cardamine rivularis occurs in the Southern Carpathians in Romania and in high mountain ranges in Bulgaria, from the montane to the alpine belt (Marhold, 1994b, 1995, 1996; Ančev et al., 2013). The diploid cytotype has so far been recorded in Bulgaria, and both diploids and less frequently also triploids were found in Romania (Marhold, 1994a; Kučera et al., 2005; Ančev et al., 2013). In addition, a triploid hybrid of *C. matthioli* and *C. rivularis* (*Cardamine* × *rhodopaea*) was described from the Rhodope Mountains in Bulgaria (Ančev et al., 2013). We collected 91 plants from nine populations, which also included putative hybrids with *C. matthioli*.

Altogether, we sampled 145 populations, which were used in flow cytometric (1,443 individuals) and microsatellite analyses (1,104 individuals). Chromosome numbers were counted from 58 populations (219 plants). A selection of 48 accessions, representing the observed overall taxonomic, geographic, ecological, and cytotypic variation, was subjected to Hyb-Seq analyses (**Supplementary Data 1** and **Figure 1**).

Chromosome Counting and Estimation of Relative Genome Size

Mitotic chromosomes were counted from young anthers or root tips fixed in an ethanol:acetic acid (3:1) mixture and stored in 70% ethanol at −20°C. Chromosome spreads were prepared following Marhold et al. (2002) or Mandáková et al. (2019). Chromosomes were stained using 4′,6-diamidino-2-phenylindole (2 mg/ml DAPI in Vectashield antifade; fluorescence signals were analyzed and photographed using a Zeiss Axioimager epifluorescence microscope with a CoolCube camera by MetaSystems) or using the Giemsa stain.

Relative nuclear DNA content was measured from fresh leaf tissues by flow cytometry (FCM) using the AT-selective DAPI fluorochrome (Doležel et al., 2007a,b). Each individual was analyzed separately or up to three individuals were pooled prior to measurements. Due to a wide range of ploidy levels and expected DNA content in the studied species, three internal standards with different 2C values were used: *Solanum lycopersicum* 'Stupické polní rané' (2C = 1.96 pg; Doležel et al., 1992), *Solanum pseudocapsicum* (2C = 2.59 pg; Temsch et al., 2010), and *Bellis perennis* (2C = 3.38 pg; Schönswetter et al., 2007). The DNA content of *Bellis* and *S. pseudocapsicum* was calibrated against *Solanum lycopersicum* based on three repeated analyses performed on different days. DNA ploidy level was inferred on the basis of DNA content measured in plants with counted chromosome numbers. Sample preparation followed the two-step procedure using Otto buffers (Doležel et al., 2007a,b) and the protocol described by Španiel et al. (2011). The fluorescence intensity of 5,000 particles (stained nuclei) was measured using a Partec CyFlow ML flow cytometer (Partec GmbH, Münster, Germany). The resulting flow cytometric histograms were evaluated using Partec FloMax software (v. 2.7d; Partec GmbH). Relative nuclear DNA content (relative genome size, 2C value given in arbitrary units; a.u.) was determined from the ratio between the positions of the G1 peaks of the sample and the standard. The coefficients of variation (CV) were calculated for both the internal standard and the sample to assess the reliability of measurements. Only histograms with CV values below the 5% threshold were accepted.

Amplification and Scoring of Microsatellite Loci

Total genomic DNA was extracted from silica gel-dried young leaves using the DNeasy Plant Kit (Qiagen, Hilden, Germany).

The microsatellite markers used in this study were developed from genomic sequence data of two individuals of *Cardamine amara* and *C. pratensis* (an individual referred to as *C. rivularis* auct.) following Zozomová-Lihová et al. (2014a). Out of 50 SSR loci with di-, tri-, and tetranucleotide repeat motifs tested for amplification and allelic variation, 18 loci (**Supplementary Table 1**) proved successful and were used for the final analyses. Amplifications were performed in five multiplex assays using the Multiplex PCR kit (Qiagen) and fluorescently labeled primers following the protocols detailed in Zozomová-Lihová et al. (2014a). The Applied Biosystems 3130xl Genetic Analyzer (DNA Sequencing Laboratory, Faculty of Science, Charles University, Prague, Czechia) was used to separate and visualize the alleles. To check the consistency of amplification and allele determination, 86 replicate samples were included.

Allele sizes were recorded using Geneious R7.1 and R10 software (Kearse et al., 2012) with the microsatellite plugin version 1.4.4. Taking into account the high ploidy level of many samples (up to 7*x*), we did not estimate the allele copy number in partial heterozygotes from peak heights, as described in Esselink et al. (2004). Alleles were recorded as either present in or absent from each sample, regardless of the number of allele copies, inferring 'marker phenotypes' or 'allelic phenotypes' (following the terminology by Meirmans and van Tienderen, 2004; Dufresne et al., 2014; Meirmans et al., 2018) instead of complete genotypes.

Target Enrichment Probe Design

Probes targeting orthologous low-copy nuclear loci were developed using Sondovač 0.99[1] (Schmickl et al., 2016), which is based on a combination of transcriptome and genome skim data. For the input, paired-end genome skim raw data of *Cardamine parviflora* (NCBI accession number: SRR11977919), a mitochondrial sequence of *Arabidopsis thaliana* (NCBI accession number: NC_001284.2), a plastome sequence of *C. impatiens* (NCBI accession number: KJ136821.1), and transcriptome sequences of *C. amara* (NCBI accession number: SRR11977918) were used. These sequences of *C. parviflora* and *C. amara* were generated as part of this study on the HiSeq2500 Illumina platform. The transcriptome of *C. amara* was assembled using Trinity 2.0.6 (Grabherr et al., 2011). Sondovač was executed with the above-mentioned input data, using default values, except for the 'Maximum overlap length expected in approximately ≥ 90% of read pairs' that was set to 200. Transcriptome sequences that were at least 90% similar and genome reads mapping to mitochondrial or chloroplast genomes were removed. Genome skim reads were matched to transcripts, filtered, and *de novo* assembled in Geneious 7.1.9 (Kearse et al., 2012). Assembled contigs were filtered for length (retaining contigs >120 bp, total length of all contigs for a transcript >1,200 bp) and uniqueness (with a threshold value of 90%). Sequences sharing more than 90% similarity with the plastome sequence were removed. Retained contigs were compiled as target sequences for probe synthesis. In total, 14,464 120-mer probes (biotinylated RNA baits) were synthesized by MYcroarray (now Arbor Biosciences, MI, United States), targeting 2,246 exons from 1,235 genes.

Illumina Library Preparation for Hyb-Seq

Genomic DNA (400 ng per accession) was fragmented with Covaris S220 or M220 sonicators (Woburn, MA, United States) to a target fragment size of 500 bp. Sequencing libraries were prepared using the NEBNext® Ultra™ DNA Library Prep Kit for Illumina® following the manufacturer's protocol (New England Biolabs, MA, United States). After cleanup of adaptor-ligated DNA with the QIAquick PCR Purification Kit (Qiagen), size selection was performed by SPRIselect beads (Beckman Coulter, MA, United States) to produce 500–600 bp long fragments. Amplification was performed with 8 cycles of PCR, using index primers from NEBNext® Multiplex Oligos for Illumina®. The amplified libraries were cleaned up with AMPure XP beads (Beckman Coulter) and pooled equimolarly (24 accessions/pool). The pooled libraries were enriched by hybridization with synthesized RNA baits (at 65°C for 26 h) using the MYbaits kit, following the protocol v. 3.02 (Arbor Biosciences). The target-enriched libraries were PCR amplified for nine cycles with the KAPA HiFi HotStart mix (Kapa Biosystems, Cape Town, South Africa), purified with the QIAquick PCR Purification Kit and submitted for sequencing with 150 bp paired end reads on an Illumina MiSeq system at BIOCEV or CEITEC, Czechia.

[1]https://github.com/V-Z/sondovac

Environmental Data

Environmental data were acquired for each of the sampling sites of the studied species, complemented by published records (Marhold, 1992, 1994a; Kučera et al., 2005; Ančev et al., 2013) that could be georeferenced and unequivocally attributed to the species and here recognized genetic lineages. Overall, 346 sites (see **Supplementary Data 2**) were included. The environmental data consisted of 72 habitat characteristics (**Supplementary Data 2**). Climatic and terrain variables were preprocessed by GeoModel Solar (Bratislava, Slovakia), as described in detail by Zozomová-Lihová et al. (2015). The temperature data (30 arc-sec resolution) consisted of average air temperature values for the period 1990–2009 (Saha et al., 2010). The precipitation data (having 2 arc-min resolution) covered a period from 1951 to 2000 (Schneider et al., 2014). Solar radiation was calculated using the SolarGIS model (15 arc-sec resolution) based on monthly long-term averages (1994 to 2013). Global tilted irradiation represents the thermal regime of the soil and vegetation cover whereas photosynthetically active radiation quantifies radiation available for plant photosynthesis. A digital elevation model was derived from SRTM3 data (3 arc-sec resolution). We also added bioclimatic variables obtained from the WorldClim database at 30 arc-sec resolution (Fick and Hijmans, 2017).

Data Analyses

Flow Cytometry

Differences in relative DNA content were tested in R version 3.3.2 (R Core Team, 2019) using the non-parametric Mann-Whitney test. The spatial segregation of cytotypes was tested using the Mantel test (Mantel, 1967) as implemented in the R package ade4 (Dray and Dufour, 2007). A pairwise distance matrix calculated from the geographic coordinates of populations was compared with a pairwise binary matrix of ploidy levels coding cytotypic identity. Mixed ploidy-level populations were excluded. Significance levels were estimated by random rearrangements (9,999 replicates) of the original matrices.

Microsatellite Datasets

Genotypic data from SSR loci were exported from Geneious in the GeneMapper format, which was loaded in the POLYSAT 1.5-0 package (Clark and Jasieniuk, 2011). The POLYSAT package was used for generating input data files for STRUCTURE (Pritchard et al., 2000) and GenoDive (Meirmans and van Tienderen, 2004), for creating partial data matrices (see below), and for calculating Lynch distance (Lynch, 1990) between individuals or lineages. This distance is a simple measure of band-sharing similarities, which best suits our data. The error rate was calculated from the binary matrix exported from POLYSAT, as the ratio of mismatches over matches of the alleles determined in the profiles of the duplicated samples. The following data matrices were assembled: (1) Complete Matrix: It consisted of all 1,104 samples, and was employed to infer the overall genetic structure, to identify potential interspecific hybrids, and to calculate population diversity parameters. (2) MajMatRiv Matrix: This partial matrix comprised *C. matthioli*, *C. majovskyi*, and *C. rivularis*. It was employed to elucidate the deeper structure within these species, and to identify potential gene flow and the occurrence of hybrids in the contact zone of *C. matthioli* and *C. rivularis* in Bulgaria. (3) Prat Matrix: This matrix comprised samples of *C. pratensis*, and was used to uncover the deeper structure within this species. The latter two data matrices corresponded to the clusters identified by STRUCTURE analyses (see below) of the Complete Matrix. Hybrid populations (see below for details on their identification) were generally omitted from the partial datasets but retained in cases when both parental clusters were present in the partial dataset.

Microsatellite Data Analyses

Because allele dosage could not be reliably estimated, standard population genetic diversity statistics that require genotype or allele frequencies (such as F_{ST} and expected heterozygosity, H_E) could not be calculated (Bruvo et al., 2004; Obbard et al., 2006). Therefore, we used GenoDive 2.0b27 (Meirmans and van Tienderen, 2004) and POLYSAT 1.5-0 (Clark and Jasieniuk, 2011), which can handle genetic data from polyploids and mixed-ploidy datasets, and correct for unknown dosage of alleles in partial heterozygotes. Genetic diversity was evaluated through the following descriptive statistics: total number of alleles (A), number of population- and species-specific (private) alleles (priv.; both calculated manually), average number of alleles per locus (A′), effective number of alleles (Ae), total heterozygosity (H_T), all three calculated in GenoDive, and Shannon diversity index (ShDI) calculated in POLYSAT. In addition, frequency down-weighted marker indices for individuals and populations (Rarity1, Rarity2; Schönswetter and Tribsch, 2005) were calculated using the R script AFLPdat (Ehrich, 2006). The diversity statistics for each lineage (or species) were calculated as the average of values obtained for populations, subjected to rarefaction correction to account for unequal number of populations. Only populations that unequivocally belonged to the lineage/species (omitting those identified as hybrid, see below) and with at least seven samples were included. In addition, allelic richness (α) was estimated for each species or lineage following the generalized rarefaction approach in ADZE v. 1.0 (Szpiech et al., 2008).

Bayesian clustering of individuals was performed to infer homogeneous genetic clusters and detect genetic admixture with STRUCTURE 2.3.4 (Pritchard et al., 2000). This model-based clustering has proven to provide unbiased inference from mixed-ploidy data with unknown allele dosage, even when population differentiation was weak (Stift et al., 2019). Ten replicates for each K = 1–10 were run with the settings and result processing as described in Zozomová-Lihová et al. (2014a). The approach of Evanno et al. (2005) was used to determine the optimal K value.

Hybrid identification followed the STRUCTURE results. As an operational unit we chose the level of population. Thus, the putatively hybrid populations were either kept or removed as a whole from the partial datasets and further analyses, instead of removing single, potentially hybrid individuals. A population was considered hybrid if genetic admixture of at least a half of its samples exceeded the threshold of 25%, and the geographic proximity of the relevant species or lineages also favored gene flow among them.

A neighbor-joining (NJ) tree was calculated using the package ape (Paradis et al., 2004) in R 3.3.2 (R Core Team, 2019). A pairwise distance matrix of individuals was calculated using Lynch distances (Lynch, 1990), as implemented in the R package POLYSAT. Trees were visualized using FigTree software version 1.4.4[2]. The genetic similarity between the taxa or lineages was also expressed by pairwise Lynch similarity indices. A hierarchical analysis of molecular variance (AMOVA) as described in Excoffier et al. (1992) was performed using the R package poppr 2.8.2 (Kamvar et al., 2014, 2015). The groups defined in this analysis followed the clustering results obtained with STRUCTURE and NJ tree. Potential hybrid populations, as described above, were omitted from the input matrices. The significance of estimated genetic partitioning was tested using 999 permutations.

Hyb-Seq Data Analyses

Demultiplexed reads were trimmed of adapters and quality-filtered using Trimmomatic-0.36 (Bolger et al., 2014). Bases at read ends with quality <Q20 were discarded, and the remaining part of the read was trimmed beyond read ends, if average quality in a 4 bp sliding window was <Q15. Finally, any reads trimmed to under 50 bp were discarded. PCR duplicates were removed using the Clumpify command of BBTools[3].

Two approaches were employed for further data processing: the first was based on the assembled genes (consensus or allele sequences obtained by read-backed phasing), while the second one utilized single-nucleotide polymorphism (SNP) calling and Bayesian clustering of the SNP datasets.

Consensus sequences were assembled using HybPiper version 1.3 (Johnson et al., 2016). Reads were mapped to the target probe sequences using BWA (Li and Durbin, 2009), assembled into contigs using SPAdes (Bankevich et al., 2012), and coding sequences were identified using Exonerate (Slater and Birney, 2005). The 'supercontigs' (targeted exons and flanking sequences) were recovered using the script intronerate.py. The supercontig sequences were aligned with MAFFT (Katoh and Toh, 2008). Flanks and sites with gaps in more than 25% of sequences were removed using the trimEnds and deleteGaps functions of the ips package (Heibl, 2008 onward) in R 3.3.2 (R Core Team, 2019). Alignment statistics were calculated with AMAS (Borowiec, 2016), and alignments displaying extreme values in descriptive statistics were inspected visually; misassemblies were removed. Finally, 1,487 supercontigs which passed inspection were concatenated into 963 genes.

Maximum-likelihood gene trees were constructed for 499 genes containing at least 32 phylogenetically informative sites (calculated with AMAS; Borowiec, 2016) using RAxML-NG v. 0.9.0 (Kozlov et al., 2019). The best-fit model of substitution for each gene was estimated using the IQ-TREE's ModelFinder function (Kalyaanamoorthy et al., 2017) under the Bayesian information criterion. Bootstrap analyses were performed using 1,000 replicates. Internal branches with bootstrap support ≤20% were collapsed with Newick-Utilities v. 1.6 (Junier and Zdobnov, 2010). To summarize the gene trees as well as to visualize potential incongruences among them, a supernetwork was inferred with SuperQ v.1.1 applying JOptimizer scaling and Gurobi optimizer (Grünewald et al., 2013; Bastkowski et al., 2018). In this analysis, the gene trees are decomposed into weighted quartets and then stitched together to a split network using the QNet algorithm (Grünewald et al., 2013).

Allele sequences were derived using the scripts and the workflow available at https://github.com/mossmatters/phyloscripts/tree/master/alleles_workflow, described in detail by Kates et al. (2018), but using the latest versions of GATK and WhatsHap (Martin et al., 2016) enabling to call and phase also polyploid variants. If the phased sequences were divided to multiple blocks, only the longest phase block for each individual was retained and the remaining variant sites were replaced with ambiguity characters. Phased alleles obtained from all 1,487 supercontigs were concatenated using AMAS (Borowiec, 2016) and a phylogenetic tree was constructed using the maximum likelihood approach after automatic model selection for each charset of a best-fit partitioning scheme using ModelFinder (Chernomor et al., 2016; Kalyaanamoorthy et al., 2017) and RAxML-NG v. 0.9.0 (Kozlov et al., 2019). Bipartition support was computed from 100 bootstrap replicates.

Bayesian clustering of the SNP datasets was obtained using STRUCTURE 2.3.4 (Pritchard et al., 2000). Variant calling and subsequent parsing to STRUCTURE format was performed using the snipStrup pipeline, which takes into account different ploidy levels of the analyzed samples (Šlenker et al., in prep.[4]). This pipeline uses BWA (Li and Durbin, 2009) to align reads to reference sequences, samtools (Li et al., 2009) for sorting and indexing, GATK (McKenna et al., 2010) for variant calling and filtering, and the R packages ape (Paradis et al., 2004) and vcfR (Knaus and Grünwald, 2016; Knaus and Grünwald, 2017) for allele extraction. As a reference, target sequences (used for probe synthesis) of 1,487 exons (those whose alignments passed inspection) were used. Exons were concatenated to genes and, to ensure no linkage existed between sites, 500 datasets were produced by drawing a single random SNP site from each gene. The datasets were piped to STRUCTURE 2.3.4 (Pritchard et al., 2000), which was run as with the microsatellite data above. The results of 500 analyses were averaged using the program CLUMPP (Jakobsson and Rosenberg, 2007) and drawn with distruct (Rosenberg, 2004). To uncover the deeper genetic structure within the species, SNP variation of *C. pratensis* (Prat dataset) and the remaining samples (MajMatRiv dataset) were analyzed separately.

Environmental Niche Differentiation

The environmental niches of the species under study and their genetic lineages were defined using 72 GIS-derived habitat characteristics (**Supplementary Data 2**), rescaled to zero mean and unit standard deviation. To quantify differences in niche positions and niche breadths, the environmental heterogeneity was summarized in a matrix of standardized Euclidean distances among the sites.

[2]http://tree.bio.ed.ac.uk/software/figtree

[3]https://jgi.doe.gov/data-and-tools/bbtools

[4]https://github.com/MarekSlenker/snipStrup

We used redundancy analysis (RDA; Rao, 1964) to test for overall differences in habitat characteristics between species and genetic lineages (sites with equivocally classified, putatively hybrid populations were removed). Randomization tests with 10,000 unrestricted permutations were used to calculate probabilities. Results of RDAs were plotted in ordination space to visualize environmental niche shifts among the groups.

A matrix of standardized Euclidean distances was used for principal coordinate analysis (PCoA) and niche breadth was quantified using the dispersion of sites from their spatial median in ordination space. A distance-based test of homogeneity of multivariate dispersions (Anderson, 2006) was employed to compare environmental niche breadths among the species and genetic lineages. Probabilities for the test statistic F_m were calculated from 10,000 unrestricted permutations of the least-absolute-deviation residuals.

Finally, multinomial generalized linear models (GLMs; McCullagh and Nelder, 1989) were used to identify subsets of environmental variables accurately discriminating habitats of the defined groups. Because the environmental dataset was truly multidimensional (72 habitat characteristics) and many variables were strongly correlated, which would pose problems related to unstable estimates of coefficients in a standard GLM (Dormann et al., 2013), we used penalized GLMs with the least absolute shrinkage and selection operator penalty (LASSO; Tibshirani, 1996). The LASSO is a regularization technique shrinking the coefficients toward zero and able to set some coefficients exactly to zero, that is, to conduct variable selection. The amount of penalization is controlled by the regularization parameter λ; as the λ penalty becomes large, fewer coefficients are left in the model. Finding an optimal value of the regularization parameter is subject to tuning. We used a 10-fold cross-validation procedure to select the final model with the λ penalty yielding the lowest misclassification rate.

Analyses were performed in R (R Core Team, 2019) using the libraries caret (Kuhn, 2019), ggplot2 (Wickham, 2016), glmnet (Friedman et al., 2010), and vegan (Oksanen et al., 2019).

RESULTS

Cytotypic Composition and Chromosome Counts

Chromosome numbers were counted for 219 plants from 58 populations (**Supplementary Data 1** and **Supplementary Figure 1**). We found extensive variation in chromosome numbers, consisting of the counts of $2n$ = 16, 17, 18, 19, 20, 22, 28, ca. 28–30, 30, 31, 32, ca. 33, 34, 36, 37, 38, ca. 34–40, ca. 42–44, 44, 48, 52, which represent diploid, tetraploid and hexaploid cytotypes (with the base number x = 8) along with frequent occurrence of probable aneuploids and dysploids. Although the chromosomes did not differ significantly in size, the presence of B chromosomes cannot be ruled out either. In addition, triploids and heptaploids were inferred from relative DNA content (not confirmed by chromosome counting) and based on the number of alleles recorded per microsatellite locus.

Flow cytometry analyses mostly yielded high-resolution histograms, with an average CV for the samples of 2.09% (range 1.1–4.9%) and 1.89% (range 0.94–4.62%) for the standards. In several populations, some individuals showed minor differences in their relative 2C values, which was most likely caused by aneuploidy and dysploidy (indeed, in a number of cases confirmed by direct chromosome counting, see **Supplementary Data 1**). Simultaneous flow cytometric analyses of such individuals yielded histograms with clearly separated or bifurcated peaks and confirmed that these differences reflected real variation (**Figure 2**).

Cardamine matthioli

All samples of *C. matthioli* analyzed for this study were confirmed to be diploid. Nevertheless, extensive intraspecific variation in nuclear DNA content was found, which exceeded 40% (**Supplementary Table 2** and **Figure 3**). At the level of populations, the variation ranged from negligible to as much as 37.58%, which was caused either by aneuploidy or the presence of B chromosomes ($2n$ = 16–20, 22; **Supplementary Data 1** and **Supplementary Figure 1**). Genome sizes derived from population averages for the two lineages defined by genetic data (denoted as the Central and the Widespread lineage, see below) did not differ significantly (Mann-Whitney test, $Z = -1.47$, $p = 0.142$).

Cardamine majovskyi

All analyzed samples of *C. majovskyi* were tetraploid. Intraspecific variation in nuclear DNA content reached 22.65% (omitting the potentially hybrid population Maj_LOG, see below), which is, as in the case of *C. matthioli*, attributable to aneuploidy ($2n$ = 32, 34; **Supplementary Data 1** and **Supplementary Figure 1**). The intrapopulational variation reached up to 12.34%. Similarly as in *C. matthioli*, relative genome sizes of *C. majovskyi* did not differ significantly between the Central and the Widespread lineage ($Z = -0.05$, $p = 0.96$). In the population Maj_LOG, where plants with $2n$ = 31 and $2n$ = mosaic of 36 and 40 were recorded (**Supplementary Figure 1**), and where hybridization with *C. pratensis* is assumed, the extent of intrapopulational variation was 22.24% (**Supplementary Data 1**).

Cardamine rivularis

Samples of *C. rivularis* (and its putative hybrids with *C. matthioli*) exhibited large variation in relative DNA content (**Figure 3** and **Supplementary Figure 2**). Although precise chromosome numbers were not determined in the present study (due to the lack of living material), chromosome number records were previously published from the same populations as examined here (Marhold, 1994a; Ančev et al., 2013). Thus, the DNA content measured herein could be unequivocally attributed to two ploidy levels as previously published, diploids and triploids. Different variation patterns were observed in Romanian and Bulgarian populations, the latter being affected by hybridization with *C. matthioli* (see below).

The two Romanian populations of *C. rivularis* analyzed were cytologically uniform, one consisting of diploids and the other

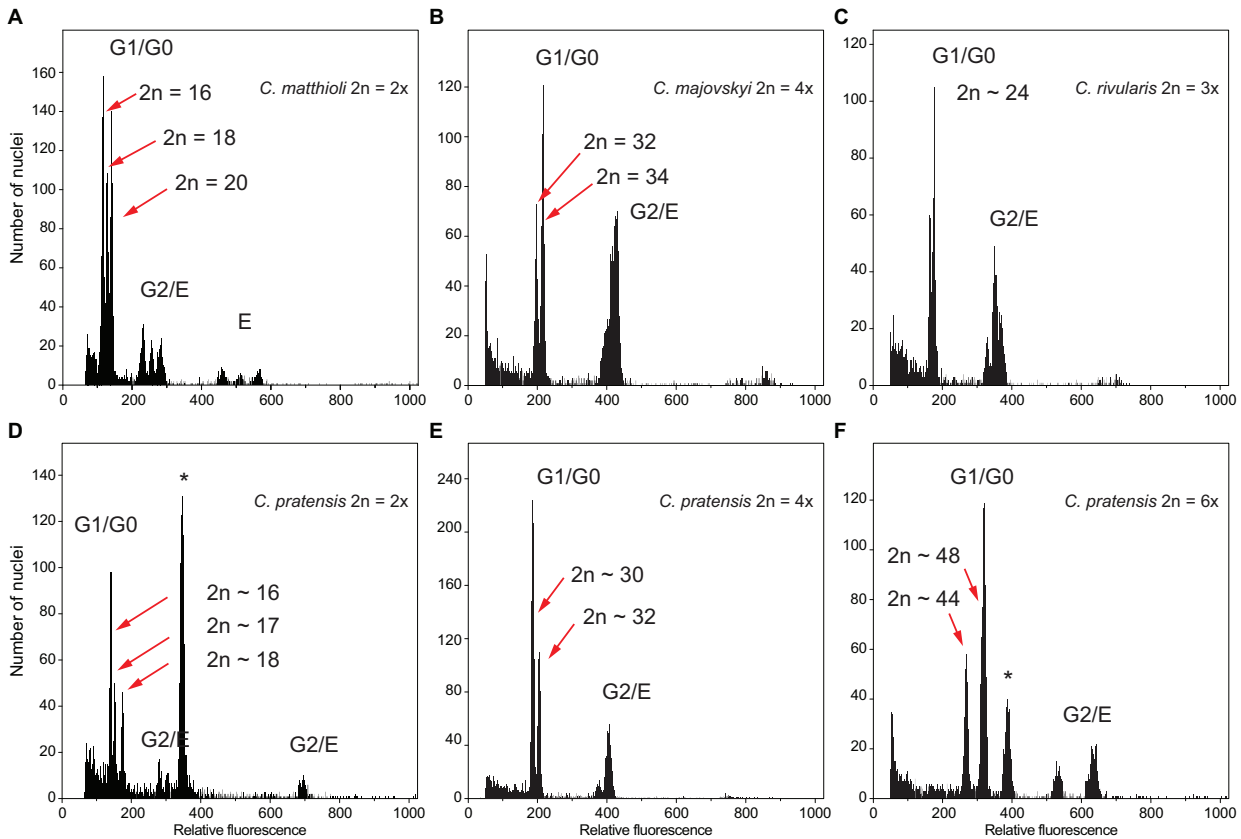

FIGURE 2 | Flow cytometric histograms of DAPI-stained nuclei obtained from simultaneous analyses of multiple samples from the same populations, documenting intraspecific divergence in nuclear DNA content in the *Cardamine* species under study caused by both aneuploidy and dysploidy. **(A)** Diploid *C. matthioli*, population Mat_CAJ; **(B)** tetraploid *C. majovskyi*, population Maj_KIS; **(C)** triploid *C. rivularis*, population Riv_KAT; **(D)** diploid *C. pratensis*, population Prat_DE5; **(E)** tetraploid *C. pratensis*, population Prat_AIS; **(F)** hexaploid *C. pratensis*, population Prat_GOR. Letters denote peaks of nuclei corresponding to different phases of the cell cycle (G0–G2) and/or levels of endopolyploidy **(E)**; the internal standards *Solanum pseudocapsicum* and *Bellis perennis* used in analyses **(D,F)**, respectively, are marked by an asterisk. For locality codes see **Supplementary Data 1**.

of triploids. The variation in relative DNA content in both populations was less than 5%, and their monoploid relative genome sizes were almost identical (C*x* = 0.262 and 0.264 for 2*x* and 3*x*, respectively; **Supplementary Table 2**).

Much greater variation was revealed in Bulgarian populations. Diploids were recorded in three out of the seven analyzed populations, and their genome size variation was relatively low (7.09%). Their monoploid genome size, however, was larger than that of diploids from Romania (C*x* = 0.28 in Bulgarian vs. C*x* = 0.262 in Romanian diploids, **Supplementary Table 2** and **Figure 3A**). The rest of the samples, putatively assigned to triploids, exhibited large variation in genome size (2C = 0.681–0.865; 27.06%; **Figure 3**). Individuals with the highest 2C values (2C ∼ 0.84; **Supplementary Figure 2B**) may represent triploids of *C. rivularis* because their monoploid genome size corresponds to that of sympatric diploids. On the other hand, the triploids with the lower 2C values most likely represent hybrids with *C. matthioli* (which has significantly smaller monoploid genome size than diploid *C. rivularis*; **Supplementary Table 2** and **Figure 3A**), where the unreduced gamete was provided either by *C. matthioli* (expected mean 2C ∼ 0.678) or *C. rivularis* (expected mean 2C ∼ 0.759). The co-occurrence of diploids with triploids was observed in one population, whereas in four populations, only triploids were recorded (**Supplementary Data 1**). The number of plants analyzed per population, however, was low (8–10), so we assume that diploid-triploid co-occurrence within populations may actually be more frequent.

Cardamine pratensis

The analyzed samples of *C. pratensis* comprised several cytotypes ranging from diploid to heptaploid. The majority of individuals were tetraploids (approximately 67%) whereas diploids and hexaploids made up 20 and 12% of the samples, respectively (although the assignment of some aneuploid individuals to ploidy levels might be equivocal). In addition, two rare odd-ploidy level cytotypes were inferred from FCM analyses: one heptaploid population and a single triploid individual in an otherwise diploid population. Almost 90% of the populations sampled consisted exclusively of one ploidy level. Sympatric occurrences of two cytotypes were encountered in only nine populations (**Supplementary Data 1**).

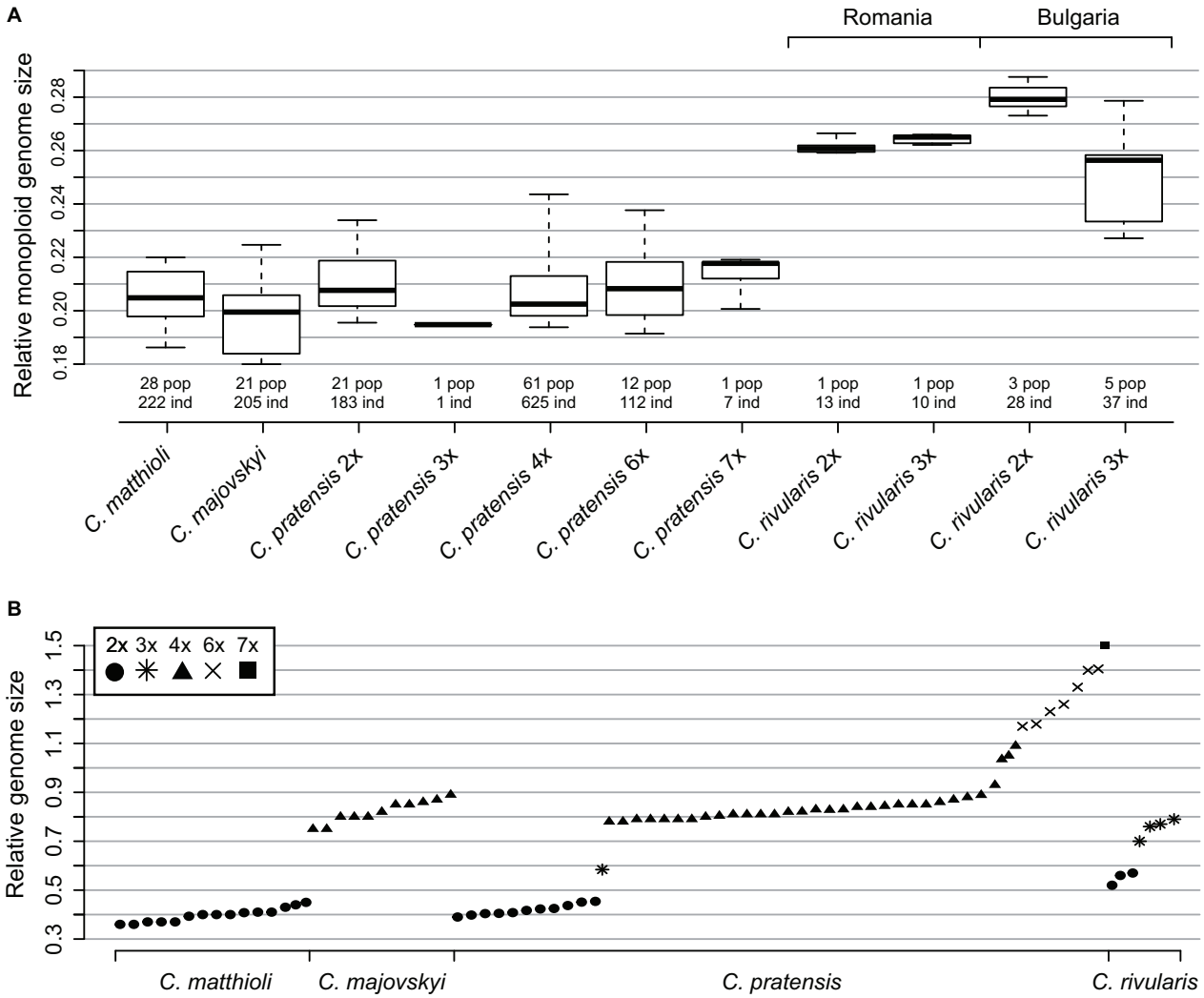

FIGURE 3 | Genome size variation among and within *Cardamine matthioli*, *C. majovskyi*, *C. pratensis*, and *C. rivularis*. **(A)** Relative monoploid genome size (C*x* value, following the terminology of Greilhuber et al., 2005) of the species and cytotypes studied. Boxes show interquartile range, and whiskers are extended to the 5th and 95th percentiles. **(B)** Relative genome size (2C) of the species under study (population mean values). Only representative populations are plotted. The fluorescence intensity of the internal standard *Solanum lycopersicum* (2C = 1.96 pg) was set to a unit value. Values given under *C. rivularis* 3*x* from Bulgaria most likely include both autotriploids of *C. rivularis* and hybrids between *C. rivularis* and *C. matthioli* (see the text for further explanations). Detailed population-, cytotype-, and species-level values are listed in **Supplementary Data 1** and **Supplementary Table 2**.

The distribution of ploidy levels in Central Europe departed from a random pattern (Mantel $r = 0.23$, $p < 0.001$). Diploid populations occupied three geographically distinct areas: in the Eastern Carpathians, the Alps, and central and southern Germany. The distribution of tetraploids overlapped with that of diploids and in general covered the whole distributional range of *C. pratensis*. Hexaploid populations were predominantly found in the Western Carpathians and adjacent areas. Several spatially isolated hexaploid samples were found in four mixed-ploidy populations in Switzerland, Slovenia, and Lower Austria. The single heptaploid population was recorded in Styria, Austria (**Supplementary Data 1** and **Figure 1**).

The intra-cytotypic genome size variation amounted to 35.54% in diploids, 50.37% in tetraploids, and 29.95% in hexaploids (**Supplementary Table 2**). This large variation in nuclear DNA content is congruent with the detected spectrum of chromosome numbers (i.e., aneuploidy and dysploidy). Monoploid relative DNA content of dominant cytotypes was almost identical among all main ploidy levels (C*x* = 0.21, 0.208, and 0.211 for 2*x*, 4*x*, and 6*x*; **Figure 3**).

Within the diploid cytotype, chromosome numbers ranged from $2n = 16$ to 19. At the tetraploid level, they varied from $2n = 28$ to 38, reflecting most likely both aneuploidy and dysploidy. Apart from supernumerary chromosomes counted sporadically (i.e., aneuploidy), tetraploid *C. pratensis* possessed two cytotypes: regular tetraploid ($2n = 4x = 32$) and hypotetraploid caused by chromosome fusion (dysploidy, $2n = 4x = 30$). Both cytotypes are widely distributed across the

sampled range. At the hexaploid level, $2n$ = ca. 42–44, 44, 48, 52 chromosomes were confirmed (**Supplementary Figure 1**). Similarly to the case of tetraploids, hexaploids comprise two main cytotypes, of which the hypohexaploid one (dysploidy, $2n = 6x = 44$) is more common than the regular hexaploid one ($2n = 6x = 48$; **Supplementary Data 1**).

Relative genome size values (2C) of diploid *C. matthioli* overlap with those of diploids of *C. pratensis*, as do also 2C values of tetraploid *C. majovskyi* with the tetraploids of *C. pratensis* (**Figure 3A**, **Supplementary Figure 2**, and **Supplementary Table 2**). By contrast, diploids of *C. rivularis* (both from Bulgaria and Romania) had markedly larger genome size than *C. matthioli* and diploid *C. pratensis*.

Genetic Structure Based on Microsatellite Variation

The 18 microsatellite loci used yielded 394 alleles. The total number of alleles and phenotypes detected per locus ranged significantly, from five alleles (eight 'marker phenotypes') in the locus Card12 to 58 alleles (528 'marker phenotypes') in the locus Card19 (**Supplementary Table 1**). The error rate calculated from the replicates was 0.001 on average, and ranged from 0 to 0.02.

Complete Matrix

The STRUCTURE analysis of the Complete Matrix comprising all 1,104 accessions and 394 alleles showed the following optimal genetic partitioning into three clusters (**Figures 4A,C**): (1) populations of *C. matthioli* and *C. majovskyi*, (2) *C. rivularis*, and (3) *C. pratensis*. The neighbor-joining clustering was consistent with this structure (**Figure 4B**). AMOVA suggested that 18.34% of genetic variation can be explained by differences between these three clusters, 31.1% by differences between populations, and 50.56% of the variation was present within populations (**Table 2**). The pairwise Lynch similarity index showed that the clusters *C. pratensis* and *C. matthioli* – *C. majovskyi* are genetically the most similar (74.83%) and that *C. rivularis* is genetically closer to *C. matthioli* – *C. majovskyi* than to *C. pratensis* (58% vs. 49.41%, respectively).

This clustering corresponded to the taxonomic classification; however, individuals in several populations were assigned to more than one group, indicating potential gene flow between taxa. Applying the above defined criterion to the determination of hybrids (i.e., the threshold of 25% genetic admixture in at least half of individuals and geographic proximity), hybrid origins were indicated for 11 populations: four populations of *C. majovskyi* and seven populations of *C. pratensis* from the contact zone between these two species (marked by asterisks, see **Figure 4C**). In addition, several samples of *C. rivularis* showed genetic admixture of *C. matthioli*; however, these did not exceed half of the population sample size.

Partial Matrix of *C. matthioli*, *C. majovskyi,* and *C. rivularis*

The two genetic clusters of *C. rivularis* and *C. matthioli* with *C. majovskyi* (see above) were merged into a single dataset (MajMatRiv Matrix) comprising 389 samples from 54 populations and 254 alleles. STRUCTURE analysis of this matrix at $K = 2$ resulted in the same clustering pattern (results not shown) as the analysis of the complete matrix above. With increasing K to its optimal value ($K = 3$), individuals of *C. matthioli* and *C. majovskyi* were separated into two clusters irrespective of their ploidy level (and thus also taxonomy), which are clearly differentiated also in the NJ tree. These two lineages (denoted as Central and Widespread) showed a non-random spatial structure. The Widespread lineage is widely distributed across the species' distribution range, including its western areas (border area of Italy and Switzerland), northern areas (Lower Austria, Slovakia, and northern Hungary) and southeastern areas (Bulgaria). By contrast, the Central lineage is localized in a relatively small area of Slovenia, northern Croatia, extending also to southeastern Austria (Styria and Burgenland) and western Hungary (**Figure 5**).

AMOVA revealed that the differentiation between the Central and the Widespread lineage explained 14.81% of variance (**Table 2**), while most variation is present within populations (57.49%), and the rest among populations (27.7%).

Populations belonging to the *C. rivularis* cluster are genetically different from the grouping of *C. matthioli* – *C. majovskyi*; however, gene flow between *C. rivularis* and *C. matthioli* in their contact zone is clearly manifested. Four populations of *C. rivularis* from Bulgaria included some individuals showing genetic admixture with the *C. matthioli* – *C. majovskyi* clusters (see **Figure 5C**). There was, however, no simple relationship between genetic admixture and ploidy level. Genetically admixed individuals were all triploid, but not all triploids exhibited signs of genetic admixture.

Partial Matrix of *C. pratensis*

The partial matrix of *C. pratensis* (Prat matrix) comprised 635 samples from 80 populations and 368 alleles. STRUCTURE analyses indicated that a division into three clusters best explained the genetic structure of this dataset (**Figure 6**). The three lineages are spatially segregated in the longitudinal direction, and these are denoted here as the Western ('Yellow'), Alpine-Bohemian-E Carpathian ('Orange'), and the W Carpathian ('Red') lineage; for simplicity in the following text we refer to them by the colors as depicted in **Figure 6**. The Yellow lineage (Western) extends from Belgium and the lowlands of Germany to the Swiss Alps, and encompasses diploid and tetraploid cytotypes (plus a single hexaploid plant in one tetraploid population). Diploids of this lineage span over a wide elevational range, from foothills of the Alps (ca. 500 m) to alpine meadows (ca. 1,650 m). Tetraploids of this lineage are common in lowlands north of the Alps and in the montane zone (below 1,000 m), and only exceptionally reach higher elevations. Populations of the Yellow lineage partially overlap with those of the Orange lineage. The Orange lineage (Alpine-Bohemian-E Carpathian) comprises diploids and tetraploids occurring both in the Western and Eastern Alps, and expanding northwards to lower elevations of Austria, Germany (mainly Eastern Bavaria), and Czechia. This lineage also encompasses spatially isolated populations from the Eastern Carpathians and surrounding lowland areas. Diploids of the Orange lineage occupy montane and subalpine zones (ca. 900–1,900 m) in the

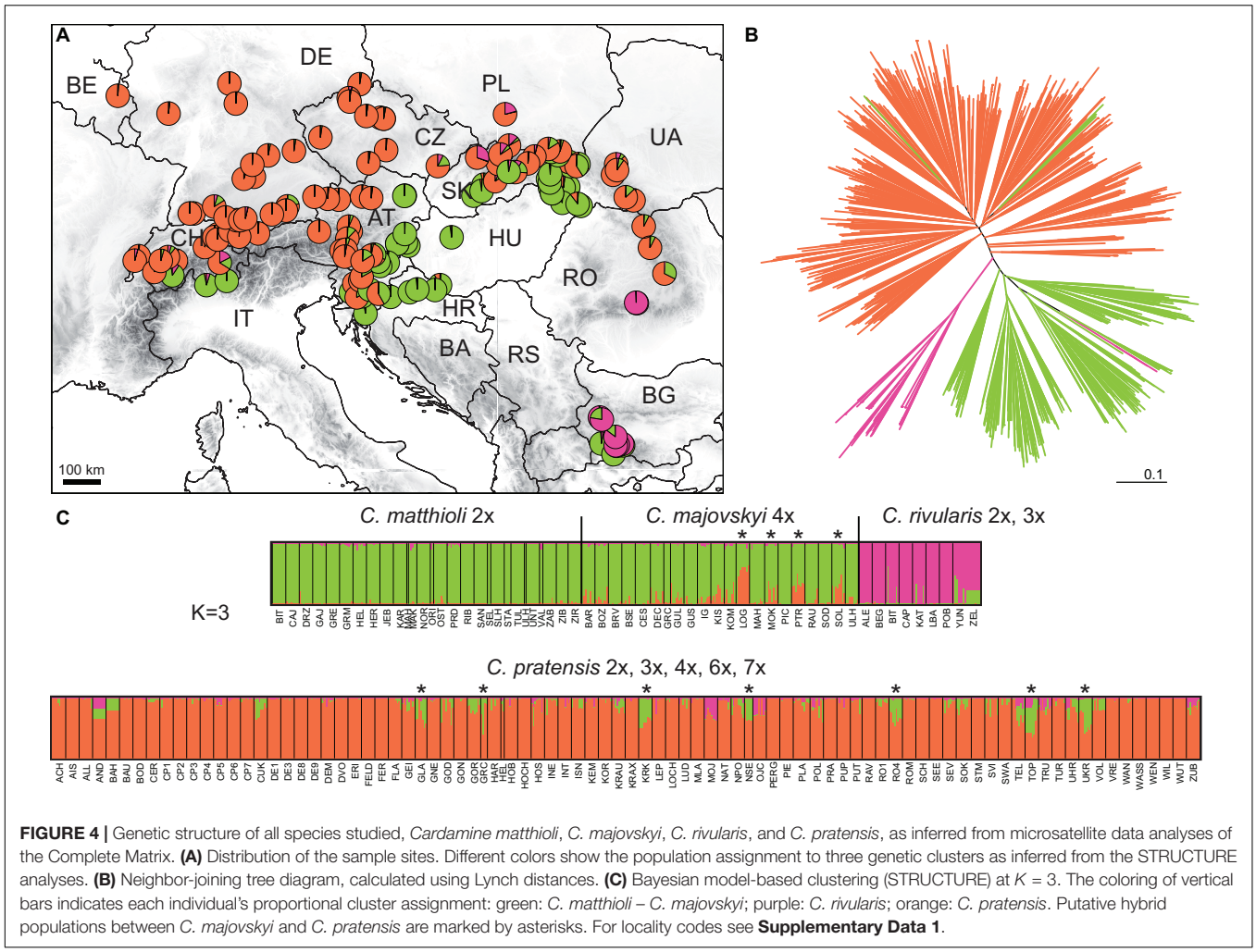

FIGURE 4 | Genetic structure of all species studied, *Cardamine matthioli*, *C. majovskyi*, *C. rivularis*, and *C. pratensis*, as inferred from microsatellite data analyses of the Complete Matrix. **(A)** Distribution of the sample sites. Different colors show the population assignment to three genetic clusters as inferred from the STRUCTURE analyses. **(B)** Neighbor-joining tree diagram, calculated using Lynch distances. **(C)** Bayesian model-based clustering (STRUCTURE) at $K = 3$. The coloring of vertical bars indicates each individual's proportional cluster assignment: green: *C. matthioli – C. majovskyi*; purple: *C. rivularis*; orange: *C. pratensis*. Putative hybrid populations between *C. majovskyi* and *C. pratensis* are marked by asterisks. For locality codes see **Supplementary Data 1**.

TABLE 2 | Results of AMOVA demonstrating partitioning of genetic variation based on microsatellite markers.

Source of variation	d. f.	Sum of squares	Variance components	Percentage of variation	Φ – statistics
(A) *Cardamine pratensis* complex					
Between clusters	2	2725.494	4.817	18.34	ΦCT = 0.183***
Between populations within clusters	133	10070.431	8.171	31.1	ΦSC = 0.381***
Within populations	904	12009.773	13.285	50.56	ΦST = 0.494***
(B) Genetic lineages of *C. matthioli* and *C. majovskyi*					
Between lineages	1	460.168	2.626	14.81	ΦCT = 0.148***
Between populations within lineages	43	1947.602	4.909	27.7	ΦSC = 0.325***
Within populations	278	2832.936	10.19	57.49	ΦST = 0.425***
(C) Genetic lineages of *C. pratensis*					
Between lineages	2	885.793	2.634	10.37	ΦCT = 0.104***
Between populations within lineages	53	4390.685	7.846	30.87	ΦSC = 0.344***
Within populations	388	6195.837	14.93	58.76	ΦST = 0.412***

****p-value < 0.001. (A) Between genetic clusters identified within the Cardamine pratensis complex (C. matthioli + C. majovskyi, C. pratensis, C. rivularis), (B) between genetic lineages identified within the C. matthioli + C. majovskyi cluster (Widespread and Central lineages), and (C) between genetic lineages identified within C. pratensis (Red, Orange, and Yellow lineages). The delimitation of genetic clusters and lineages follows the results of STRUCTURE analyses.*

Alps, while in the Eastern Carpathians they span from lowlands (350 m) up to the subalpine zone (ca. 1,900 m). Tetraploids of this lineage are dominant across Germany, Austria, and Czechia, and only scarcely occur in the Eastern Carpathians. They grow in lowland and montane belts, and, in contrast to tetraploids of the Yellow lineage, they also occupy (sub)alpine

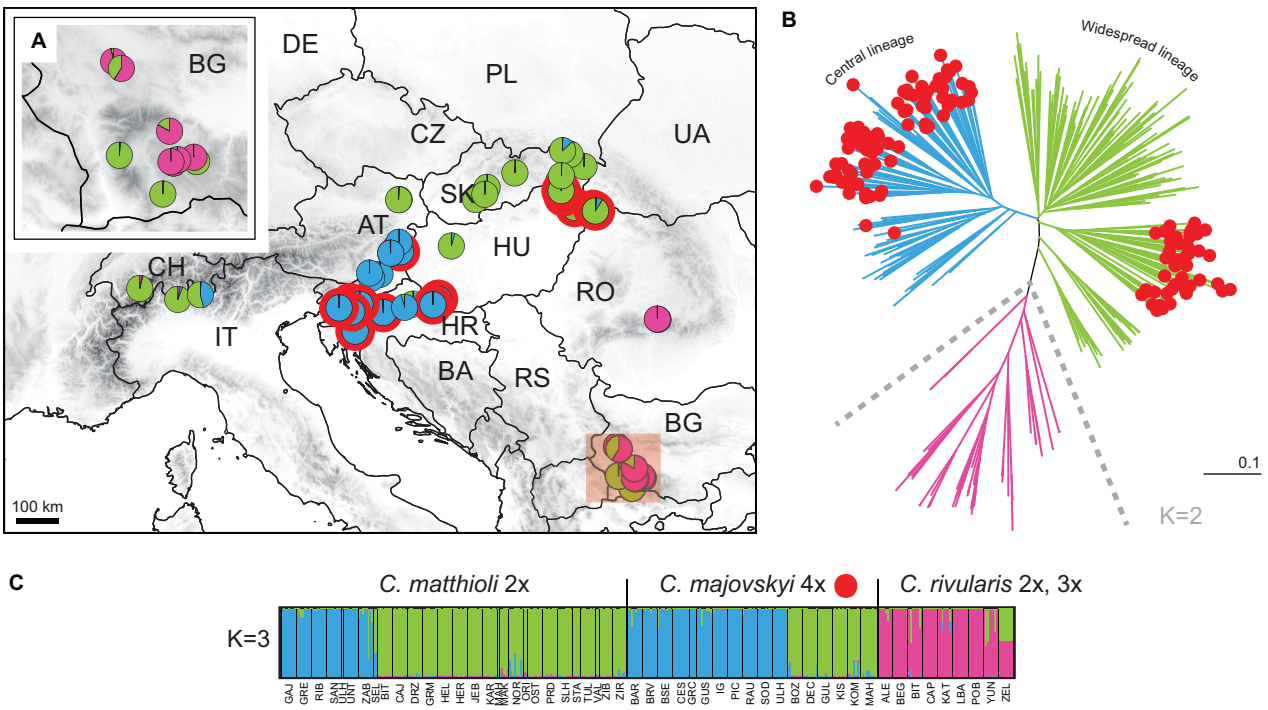

FIGURE 5 | Genetic structure of *Cardamine matthioli*, *C. majovskyi*, and *C. rivularis* inferred from microsatellite data analyses of the MajMatRiv Matrix. **(A)** Distribution of the sample sites. Different colors indicate the population assignment to the three genetic clusters as inferred from STRUCTURE analyses. Circles with red outlines represent populations of *C. majovskyi.* **(B)** Neighbor-joining tree diagram, calculated using Lynch distances. Red dots highlight individuals of *C. majovskyi*. **(C)** Bayesian model-based clustering (STRUCTURE) at $K = 3$. Coloring of vertical bars indicates each individual's proportional cluster assignment: blue: Central lineage; green: Widespread lineage; purple: *C. rivularis*. For locality codes see **Supplementary Data 1**.

environment (ca. 1,000–1,800 m). Finally, the Red lineage (W Carpathian) occurs predominantly in the Western Carpathians and southern Poland. It encompasses tetraploids, hexaploids and one heptaploid population. The hexaploid cytotype is the most common and occupies the whole elevational gradient in this lineage (300–1,300 m). On the other hand, tetraploid and heptaploid populations are restricted to a narrow elevational range, from 600 to 800 m.

Nevertheless, considerable admixture among the three above-described lineages of *C. pratensis* is indicated by our STRUCTURE analyses (**Figure 6C**). The NJ clustering, too, indicates that genetic differentiation among the three lineages is weak and, in addition, suggests a split within the Red lineage (following the cytotypes) and subdivision within the Orange lineage (**Figure 6B**). AMOVA showed that only 10.37% of the variation was attributable to differences between the three lineages, and the majority of variation occurred within population (58.76%) (**Table 2**).

Populations previously classified as *C. nemorosa* and *C. udicola* (see **Table 1** and **Supplementary Data 1**) are both included, among other populations, within the Yellow lineage. Not even the subalpine populations from the Alps and Eastern Carpathians, commonly referred to as *C. rivularis* auct., were identified by our analyses as a distinct entity, as they are placed both within the Yellow and the Orange lineage along with adjacent lowland populations. The morphotype 'ucranica,' reported from the foothills of the Eastern Carpathians (**Table 1** and **Supplementary Data 1**), is placed within the Orange lineage.

Genetic Diversity

Genetic diversity within the populations analyzed varied largely (**Supplementary Data 1**). After rarefaction correction, the highest diversity was recorded in *C. pratensis* and the lowest in *C. rivularis*. At the diploid level, diploid populations of *C. pratensis* were more diverse than those of *C. matthioli*. Similarly, at the tetraploid level, tetraploid populations of *C. pratensis* appeared more diverse than *C. majovskyi* (**Table 3**).

In *C. matthioli*, patterns of genetic diversity exhibited a non-random spatial structure. The highest values of the Shannon diversity and population rarity were detected in populations sampled in Slovenia, Croatia, and Slovakia (i.e., from both the Central and Widespread lineages) whereas peripheral Bulgarian populations appeared genetically depauperate (**Supplementary Data 1**). Swiss and Italian populations were omitted due to low sample sizes. When comparing the two lineages (comprising both *C. matthioli* and *C. majovskyi*), they displayed very similar genetic diversity values; the Central lineage appeared more diverse at the level of populations, but the overall estimated allelic richness was slightly higher in the Widespread lineage (**Table 3**). The percentage of genetic similarity between the two lineages was 72.16%.

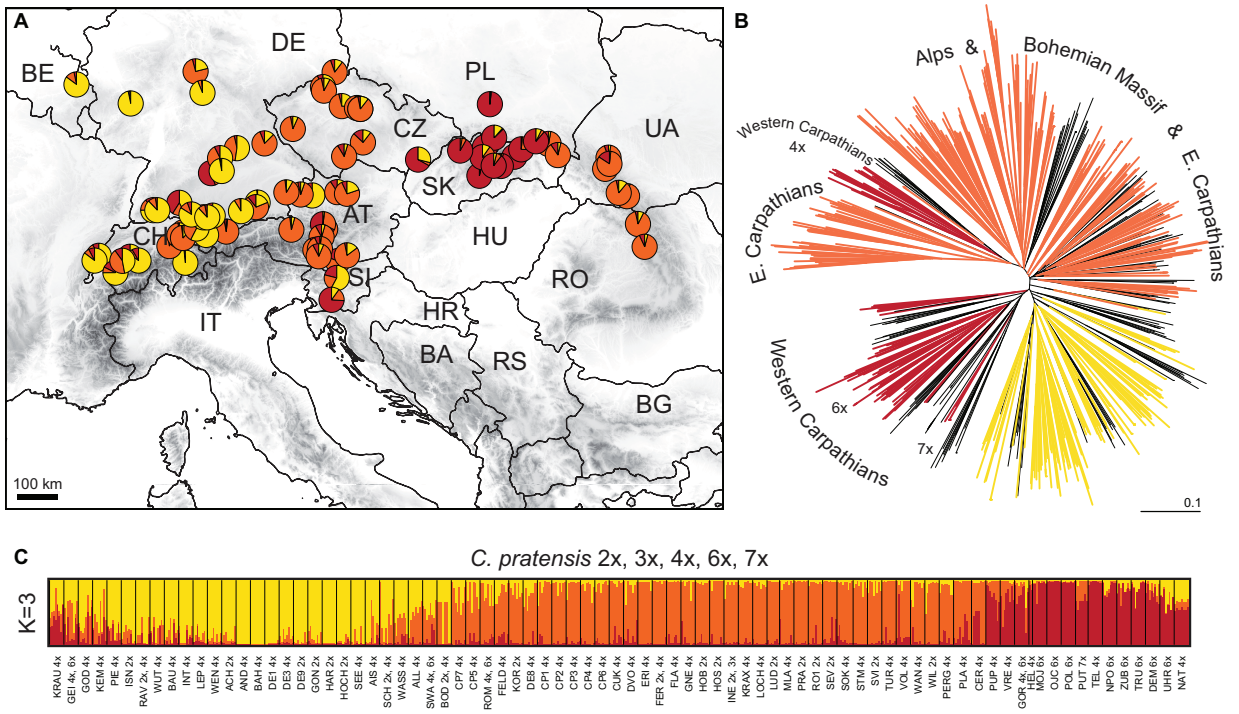

FIGURE 6 | Genetic structure of *Cardamine pratensis* inferred from microsatellite data analyses of the Prat Matrix. **(A)** Distribution of sample sites. Different colors indicate the population assignment to three genetic clusters as inferred from STRUCTURE analyses. **(B)** Neighbor-joining tree diagram based on Lynch distances. Branches are colored according to the population assignment to three genetic clusters as inferred from STRUCTURE analyses; populations with equivocal (admixed) assignments are in black. **(C)** Bayesian model-based clustering (STRUCTURE) at $K = 3$. The coloring of vertical bars indicates each individual's proportional cluster assignment: yellow: Western lineage; orange: Alpine-Bohemian-E Carpathian lineage; red: W Carpathian lineage. The ploidy level of each population is also indicated. For locality codes see **Supplementary Data 1**.

In *C. pratensis*, the Yellow lineage was the least diverse, whereas the Red lineage was the most diverse, as shown by all parameters (**Table 3**). Some spatial structure in the diversity distribution was observed also within the lineages; however, most of this structure could be attributed to the distribution of different cytotypes, as is obvious from the positive correlation between ploidy level and detected diversity (**Supplementary Data 1**). Parameters of genetic rarity indicated high values especially in the Alps and Carpathians (the Red lineage and southern margins of both the Yellow and Orange lineages), while lower values were concentrated in northern regions of both the Yellow and Orange lineages (**Supplementary Data 1**). The three lineages showed high genetic similarity (pairwise Lynch similarity index), in congruence with their spatial distribution. Geographically overlapping lineages, the Orange and Red (82.76% similarity) and the Orange and Yellow (81.83% similarity) were less divergent than the geographically separated Yellow and Red lineages (78.85% similarity).

Genetic Structure Based on Hyb-Seq

Sequencing process yielded, on average, 1.25 million reads per sample. Adapter trimming, quality filtering, and deduplication resulted in an average loss of 15.5% of reads. Of the remaining reads, 79.46% on average were mapped to target sequences. Out of the 2,246 exons from 1,235 genes, targeted by the designed RNA baits, 1,868 (82.8%) exons were recovered from all 48 samples. After inspection, 1,487 supercontig sequences were kept and concatenated to 963 genes of the overall length of 1,186,798 bases. The gene length ranged from 110 to 10,705 bp, with a mean of 1,232 bases. The proportion of parsimony informative sites ranged from 0.5 to 22.7%, with a mean of 3.8%. The variant calling approach (the snipStrup pipeline) utilized 936 genes (out of 963), which contained at least one variant for each sample. The genes harbored between one and 231 SNPs (omitting SNPs that occurred only in one sample), with an average of 41.21 and 47.86 variants per gene for the MajMatRiv and the Prat datasets, respectively.

The different approaches that we employed for the Hyb-Seq data analyses, an ML tree based on phased allele sequences of all genes in concatenation, a supernetwork based on the most informative genes, and STRUCTURE analyses of SNP data drawn from the genes, provided highly congruent results (**Figure 7**). *C. rivularis*, *C. pratensis,* and *C. matthioli* intermingled with *C. majovskyi*, were identified as genetically clearly distinct. The extensive intraspecific variation observed within the latter three species is geographically structured, which is largely concordant with the patterns inferred from microsatellites. The same two genetic lineages, Central and Widespread, were identified within the clade of *C. matthioli–C. majovskyi*. Within *C. pratensis*, the same three lineages, denoted as Yellow, Orange, and Red,

TABLE 3 | Indices of genetic diversity based on microsatellite markers in the species, lineages, and cytotypes of the *Cardamine pratensis* complex.

Group	N_{ind}	N_{pop}	*A*	Priv.	*A′*	Ae	H_T	Shan. div. ± SD	α ± *SE*
***Cardamine pratensis* complex**									
C. matthioli + *C. majovskyi*	296	37	228	12	3.293	2.311	0.496	1.01 ± 0.083	7.484 ± 0.665
C. pratensis	621	78	366	136	4.967	3.232	0.635	1.262 ± 0.118	10.786 ± 1.282
C. rivularis	72	9	119	9	2.321	1.835	0.368	0.477 ± 0	5.919 ± 0.705
Diploids									
C. matthioli	167	21	153	25	2.709	2.01	0.429	0.865 ± 0.117	6.319 ± 0.640
C. pratensis	134	17	231	92	3.419	2.403	0.53	1.051 ± 0.069	9.145 ± 1.242
C. rivularis	24	3	68	13	1.759	1.492	0.249	0.342 ± 0	3.778 ± 0.698
Tetraploids									
C. majovskyi	129	16	208	24	4.077	2.709	0.582	1.193 ± 0	7.560 ± 0.650
C. pratensis	361	45	320	136	5.174	3.294	0.648	1.298 ± 0.081	10.291 ± 1.228
Hexaploid									
C. pratensis	79	10	276	–	6.811	4.179	0.724	1.61 ± 0	10.335 ± 1.093
Genetic lineages of *C. matthioli* and *C. majovskyi*									
Widespread	167	21	175	55	3.218	2.297	0.485	0.996 ± 0.037	6.827 ± 0.675
Central	129	16	173	53	3.413	2.333	0.507	1.031 ± 0	6.427 ± 0.576
Genetic lineages of *C. pratensis*									
Yellow	128	16	231	14	4.206	2.858	0.58	1.037 ± 0.085	8.682 ± 1.046
Orange	247	31	294	41	4.557	2.943	0.611	1.248 ± 0.076	9.956 ± 1.235
Red	79	10	256	26	6.133	3.897	0.705	1.452 ± 0	10.088 ± 1.047

N_{ind}, number of analyzed individuals; N_{pop}, number of populations included in the diversity calculations, i.e., without hybrid populations and those with low sample size (N_{ind} < 7). A, total number of alleles; priv., number of species- and lineage-specific (private) alleles; A′, average number of alleles per locus; Ae, effective number of alleles; H_T, total heterozygosity; Shan. div., Shannon's diversity index; α, allelic richness estimated with generalized rarefaction, calculated as the number of distinct alleles expected in a random subsample of size g, here g = 60. Values in A′, Ae, H_T, and Shan. div. are population averages after rarefaction correction; α are values calculated at the species or lineage levels.

were resolved by STRUCTURE at $K = 3$ (**Figure 7C**), albeit with admixture observed especially within the Red lineage, besides some further subdivisions suggested by ML tree and supernetwork (**Figures 7A,B**). The Yellow lineage is most coherent, formed by western diploid and tetraploid populations from lowlands up to the subalpine belt in the Western Alps. Within the Orange lineage, the same differentiation between the Alpine range and the Eastern Carpathian range is indicated, both comprising populations from lowland to subalpine sites. The Red lineage from the Western Carpathians exhibits a subdivision based on ploidy levels; the hexaploid populations showed affinity to the Yellow lineage, whereas the tetraploids to the Orange one, which is also concordant with the microsatellite results. Only three populations from contact zones between the Yellow and Orange lineages were assigned differently by microsatellite and Hyb-Seq data (marked by asterisks, see **Figure 7**), which confirms the diffuse borders between these intraspecific lineages. Finally, two samples from Slovenia (Cprat_KRK7, Cprat_GOR3) were identified as potential hybrids between *C. pratensis* and *C. majovskyi* in the Hyb-Seq analysis, which was partly confirmed by microsatellite data (**Figure 7**). In conclusion, the Hyb-Seq data generated for a subset of 47 populations revealed a genetic structure that was highly congruent with that inferred from the microsatellite data of 145 populations.

Environmental Niche Differentiation

Redundancy analysis revealed significant differences in environmental niches of the four *Cardamine* species (pseudo-$F = 50.5$, $p < 0.0001$) (**Figure 8A**). The niches were separated along the main climatic gradient represented by elevation and accompanied changes in temperature and solar radiation. The niche of *C. rivularis* was typical for higher elevations, *C. pratensis* occupied milder mid-elevation sites, while *C. majovskyi* and *C. matthioli* occurred in lower elevations (**Supplementary Data 2**). The environmental niches of *C. matthioli* and *C. majovskyii* showed more overlap than the other species, but still differed, especially in the amount of photosynthetically active radiation and temperature seasonality (both higher in *C. majovskyi*). The multinomial LASSO model identified 42 habitat characteristics (**Supplementary Table 3**) that discriminate among species with relatively high overall predictive performance (79% of correctly classified species in cross-validation). The model showed excellent prediction of *C. rivularis* occurrence (0% misclassification). The rate of misclassification was greater in the case of *C. majovskyi* (28%), *C. matthioli* (29%), and *C. pratensis* (22%). The environmental niches of those species were of significantly different breadths ($F_m = 18.7$, $p < 0.0001$). *C. pratensis* and *C. rivularis* have broader niches than *C. majovskyi* ($p = 0.0033$ and $p < 0.0001$) and *C. matthioli* ($p = 0.0004$ and $p < 0.0001$). On the other hand, we found comparable niche breadths between *C. pratensis* and *C. rivularis* ($p = 0.2541$), and between *C. majovskyi* and *C. matthioli* ($p = 0.9630$).

The environmental niches of genetic lineages within *C. majovskyi* and *C. matthioli* (Central, Widespread, i.e., four groups in total), were significantly different (pseudo-$F = 15.8$,

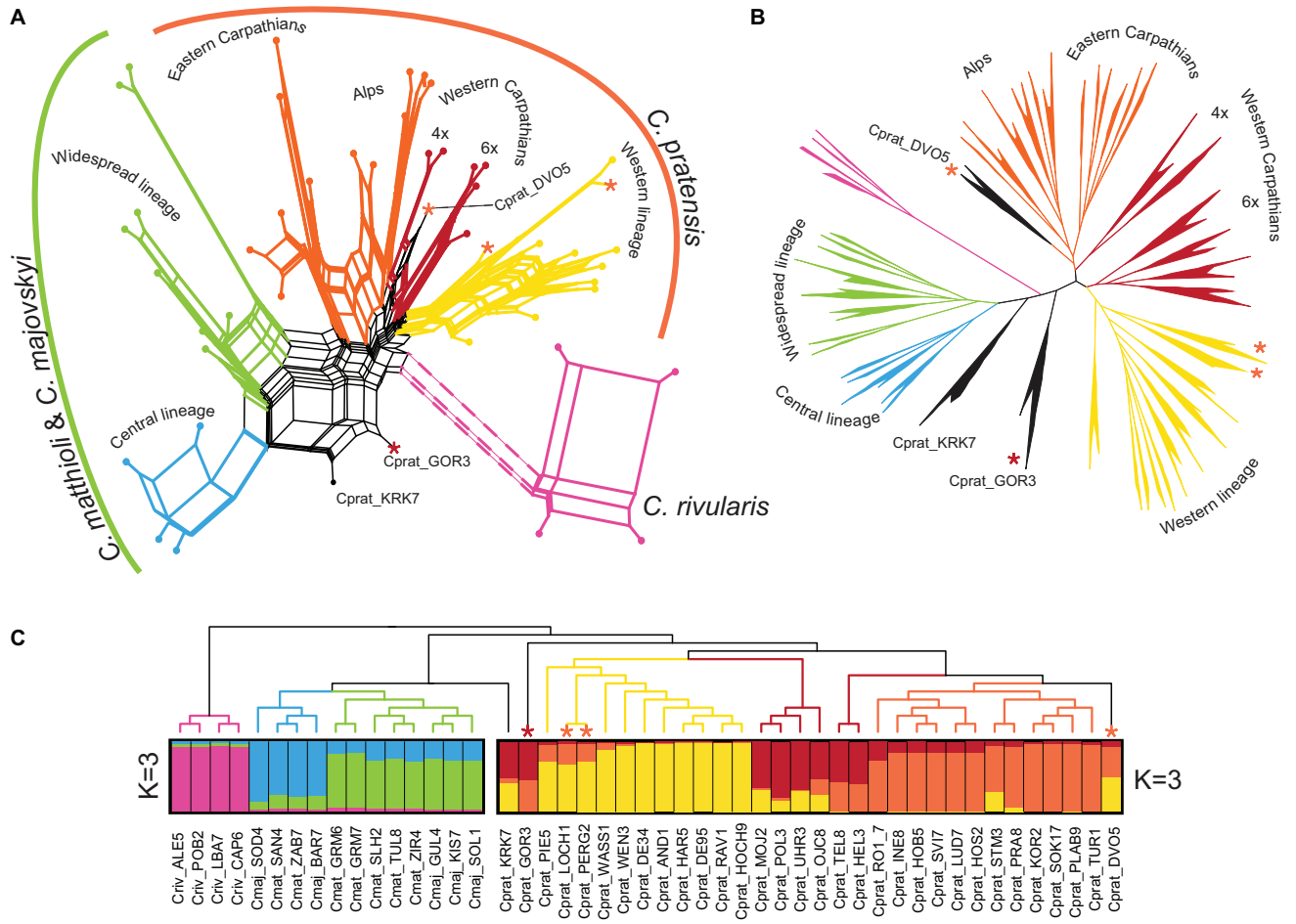

FIGURE 7 | Genetic structure of *Cardamine matthioli*, *C. majovskyi, C. rivularis*, and *C. pratensis* inferred from Hyb-Seq data analyses. The coloring indicates the sample assignment to genetic clusters as inferred from STRUCTURE analyses: blue and green represent the Central and Widespread lineages of *C. matthioli* and *C. majovskyi*, respectively; purple represents *C. rivularis*; yellow, orange, and red represent the Western, Alpine-Bohemian–E Carpathian and W Carpathian lineages of *C. pratensis*, respectively. **(A)** Supernetwork representation of quartets generated in SuperQ, derived from 499 most informative maximum likelihood (ML) gene trees. **(B)** ML tree constructed from concatenated allele sequences obtained by read-backed phasing from all 963 genes in RAxML-NG. All major clades received bootstrap support greater than 95%. **(C)** Bayesian model-based clustering (STRUCTURE) based on SNP data from two partial datasets of *C. matthioli*, *C. majovskyi, C. rivularis* (on the left), and *C. pratensis* (on the right); ML tree constructed using the concatenated genes (as in **B**) is displayed as a cladogram on top. Samples with significant admixture (hybrid origins) between the species or lineages are shown with black branches. Asterisks denote samples differently placed in microsatellite analyses, and their colors indicate the assignment based on microsatellite data. For accession codes see **Supplementary Data 1**.

$p < 0.0001$) and each was statistically distinguishable from the others in pair-wise comparisons (**Figure 8B**). The main habitat-related difference was observed between the geographic areas along the precipitation gradient (greater precipitation sums in the more restricted Central region). Within the geographic areas, *C. majovskyi* typically occupies habitats with higher temperatures and lower seasonality of precipitation than *C. matthioli*. The final LASSO model consisted of 19 variables (**Supplementary Table 3**) which predicted the occurrence of the groups with good accuracy (the cross-validated overall classification accuracy equalled 78%). The classification error rate was quite high for *C. majovskyi* of the Central lineage (38%), but the other groups reached acceptable rates of misclassification (*C. matthioli*, Central – 12%; *C. majovskyi,* Widespread – 14%; *C. matthioli*, Widespread – 25%). We found the niche breadths of the four defined groups to be unequal ($F_m = 4.1$, $p = 0.0075$), but the only significant difference was between the two species within the Widespread lineage (*C. matthioli* > *C. majovskyi*, $p = 0.01$).

Genetic lineages within *C. pratensis* exhibited significantly different environmental niches (pseudo-$F = 2.9$, $p = 0.0066$). In the pairwise-comparison, habitat characteristics of the Yellow (Western) lineage differed significantly from both the Red and Orange lineages, while the latter two lineages were statistically indistinguishable (**Figure 8C**). In contrast to the Red and Orange lineages, the niche of the Yellow lineage is shifted toward more humid conditions with higher amounts of precipitation. The multinomial LASSO model involving 26 habitat characteristics (**Supplementary Table 3**) relatively accurately predicted the occurrence of the Yellow (25% cross-validated misclassification

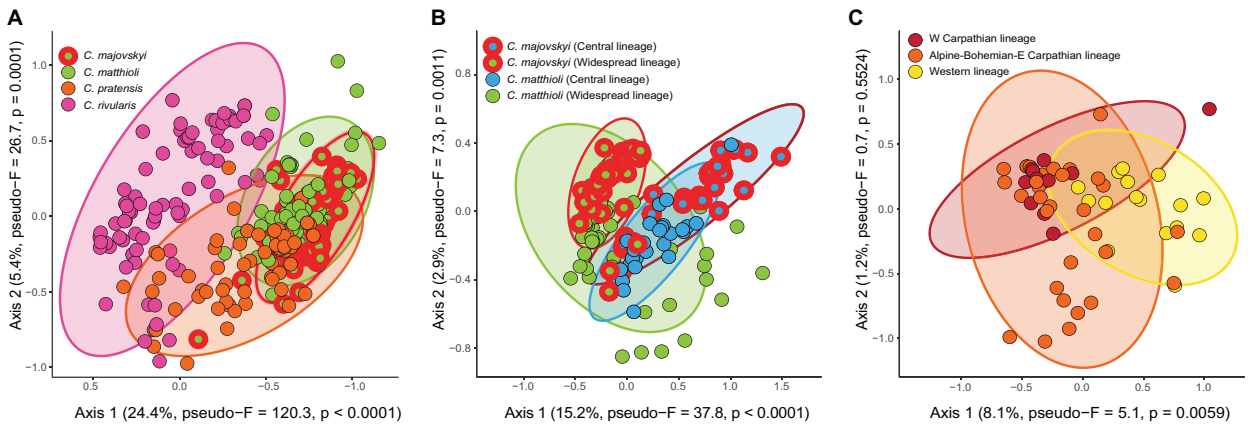

FIGURE 8 | Results of RDA showing significant differences in the environmental niches of the four analyzed *Cardamine* species **(A)**, genetic lineages within *C. matthioli* and *C. majovskyi* **(B)**, and three genetic lineages within *C. pratensis* **(C)**. Ellipses define regions expected to contain 95% of all populations of the species or lineages. Vectors of habitat characteristics contributing to the ordination axes are shown in a full version of the RDA ordination graphs in **Supplementary Figure 3**. Variance explained by the ordination axes and results of randomization tests are given in parentheses. The scaling of the ordination plots is focused on habitat similarity among sites.

error rate) and Orange lineages (27%), however, the rate of misclassification was quite high in the Red lineage (46%). We found only marginally significant differences in niche breadths between the lineages (F_m = 3.3, p = 0.0436); the Orange lineage has a broader niche than the Red lineage (p = 0.0335), while the other pairs are statistically comparable.

DISCUSSION

Patterns and Sources of High Cytotypic and Genome Size Variation

The results of the extensive flow cytometric screening of genome size and chromosome counting presented here support previous chromosome number reports (Kučera et al., 2005; Paule et al., 2017; Marhold et al., 2018) and describe the distribution patterns of ploidy levels and cytotypes in great detail. They underscore the extraordinarily large cytotypic variation within the *C. pratensis* complex in Central Europe, present both among and within the species. Apart from the euploid numbers, based on the primary base chromosome number x = 8, numerous aneuploids and dysploids were found. Aneuploidy was recorded in the diploid *C. matthioli*, tetraploid *C. majovskyi*, and across all ploidy levels of *C. pratensis*. In addition, descending dysploidy occurs at the tetraploid and higher levels in *C. pratensis*. The hypotetraploids with $2n$ = 30 possess one pair of markedly longer chromosomes, which were noticed already by Lawrence (1931) and Lövkvist (1956). Comparative chromosome painting revealed that the two long chromosomes underwent a nested chromosome insertion (fusion), a translocation event involving the 'insertion' of one chromosome into the pericentromere of the second chromosome, resulting in a chromosome number reduction from $2n$ = 32 to 30 (Mandáková et al., 2013). The hypotetraploids were most likely involved in the origin of hypohexaploid plants with $2n$ = 44 that were noted already by Lövkvist (1956) and detected also by the present study. Descending dysploidy, which has been recognized as a major route of diploidization of polyploid genomes and a significant mechanism of chromosome number evolution (Mandáková and Lysak, 2018; Udall et al., 2019), has been reported from several genera, including a pseudotriploid *Cardamine* species (Mandáková et al., 2016).

The co-occurrence of euploid and dysploid cytotypes in *C. pratensis*, along with aneuploids, may be facilitated by the perennial life-form and the capability for vegetative reproduction, and is probably accompanied by inter-cytotype mating, which results in almost continuous variation in chromosome numbers and genome size at and above the tetraploid level. Indeed, pollination and hybridization experiments by Lövkvist (1956) revealed weak reproduction barriers between the polyploids within the *C. pratensis* complex. Still, we observed that mixed-ploidy populations were relatively rare (constituting about 10% of populations). Based on their distribution patterns we assume that they have arisen both by secondary contacts between the two cytotypes (in the case of diploids and tetraploids) and by multiple *in situ* polyploidization events (in the case of tetraploids and hexaploids). Populations with the co-occurrence of diploids and tetraploids were found only in the western part of the study area, where these ploidy levels grow largely intermingled, enabling and furthering their secondary contacts. By contrast, tetraploid-hexaploid populations were found scattered in areas occupied by tetraploid populations, suggesting that the hexaploids originated multiple times within those populations. A similar scenario of recurrent hexaploid formation within tetraploid populations has been inferred in ferns based on their Cx value variation (Hanušová et al., 2019).

In congruence with the phylogenetically distinct position of *C. rivularis* (Franzke and Hurka, 2000), this species had significantly larger monoploid genome size than the other

species analyzed here, namely *C. matthioli*, *C. majovskyi*, and *C. pratensis*, which, in turn, possessed highly similar monoploid genome sizes. Still, we found large genome size variation within *C. rivularis*, which apparently reflects various evolutionary processes. First, some genome size divergence was observed between the Romanian and Bulgarian range of this (sub)alpine species, which was probably facilitated by their long-term disjunction (postglacial at least) and evolution in allopatry. As is known from other studies, differential evolution of repetitive DNA and/or selection can generate genome size differences across populations and species, which can be fixed in the absence of gene flow (Bennetzen et al., 2005). As the next cause of the observed genome size variation, triploids have been revealed in the Romanian range (Făgăraş Mts), in agreement with earlier reports (Marhold, 1994a). The formation of unreduced gametes in natural populations is a significant pathway by which triploid cytotypes originate (Mason and Pires, 2015), and it is proposed here also for the case of *C. rivularis*. This species is capable of vegetative propagation (Marhold, 1994a; Ančev, 2006), and, indeed, genetic clones were inferred from the lack of microsatellite variation in both populations from the Făgăraş Mts analyzed here. Vegetative reproduction may therefore explain the successful establishment of triploids and their long-term persistence despite the cytotype minority exclusion phenomenon and sterility issues (Levin, 1975; Ramsey and Schemske, 1998; Husband, 2004; Köhler et al., 2010), similarly as was observed in the triploid *Cardamine* × *insueta* in the Alps (Zozomová-Lihová et al., 2014b). Much more complex genome size patterns are seen in the Bulgarian range, where apparently both unreduced gamete formation and hybridization with *C. matthioli* have resulted in a mixture of different auto- and allotriploids persisting along with diploids. The substantial genome size variation revealed here within as well as among populations may be due to multiple hybrid formation and even crossing between triploids of different origins, as they seem to be partially fertile (Ančev et al., 2013).

Multiple Polyploid Origins and Cases of Interspecific Hybridization

Previous authors suggested an autopolyploid origin of *C. majovskyi* based on its occurrence within the area of the diploid *C. matthioli*, morphological as well as molecular data (Marhold, 1994a, 1996; Franzke and Hurka, 2000; Lihová et al., 2003). The genetic patterns resolved here clearly support this scenario, and, in addition, show that the tetraploid *C. majovskyi* arose at least twice. We are thus adding to an ample body of evidence that multiple origins of polyploids are rather the rule than the exception (e.g., Doyle et al., 1990; Soltis et al., 2004; Servick et al., 2015; Rešetnik et al., 2016; Novikova et al., 2017, 2018). Sympatric occurrence of diploids and tetraploids at the same site was rare (**Supplementary Data 1**). Thus, even if autopolyploids may be recurrently formed within diploid populations, a frequency-dependent mating disadvantage (the principle of minority cytotype exclusion; Levin, 1975) probably hampers common and long-term cytotype co-existence (Kolář et al., 2017; see e.g., Castro et al., 2019), unless they reach a different site or habitat. Even though *C. majovskyi* has not expanded beyond the range of its diploid progenitor so far, it appears to have a significantly different environmental niche. Still, the magnitude of niche divergence is greater between the two genetically and geographically defined lineages spanning both species (i.e., Central and Widespread) than between the two species. Previous studies on other species groups have indicated that autopolyploids may exhibit lower rates of niche evolution than allopolyploids and may consequently persist in geographically close areas and occupy similar niches as their ancestors (e.g., Arrigo et al., 2016; Castro et al., 2019).

Multiple evolutionary scenarios need to be considered to resolve the origins of polyploids of *C. pratensis*. Both diploids and polyploids of *C. pratensis* occur throughout the distribution range of this species (Kučera et al., 2005), as well as within our focus area of Central Europe. Central European diploids grow in several disjunct areas, they are ecologically and genetically diverse, and this heterogeneity provides great potential for multiple polyploidization events. The nested chromosome fusion detected in the hypotetraploid cytotype ($2n = 30$; Mandáková et al., 2013) represents an evolutionarily unique event, which strongly suggests a single origin of hypotetraploids and their subsequent range expansion. The occurrence of this cytotype in all genetic lineages of *C. pratensis*, supposedly descendants from different glacial refugia (see below), also indicates that the polyploidization event predated at least the last glacial period. Unlike *C. majovskyi*, these hypotetraploids have spread widely within as well as beyond Central Europe (Lövkvist, 1956; Lihová et al., 2003, 2004) and typically occupy lowland to montane sites. Hypohexaploids with $2n = 44$ belonging to the Red lineage, which likely originated from these hypotetraploids, form a stabilized dysploid, genetically and ecologically defined lineage putatively of a single origin (see also Marhold, 1994a,b), although this cytotype has been recorded also in other parts of Europe (Kučera et al., 2005). On the contrary, we can expect multiple origins of other cytotypes at the tetraploid and hexaploid level. Multiple polyploid origins are for instance, supported by a few cases of tetraploid-hexaploid mixed populations, in which the hexaploids apparently originated only recently and *in situ* (discussed above). Tetraploids with $2n = 32$ were recorded in both the Yellow and Orange lineages, and it is reasonable to assume that they have originated by polyploidization independently in each of them. In contrast to populations with $2n = 30$, those with $2n = 32$ grow at higher (sub)alpine sites in the Alps and were recorded also in northern Spain (Lihová et al., 2003). We therefore suggest that the polyploids of *C. pratensis* may have arisen via both single and multiple origins, and probably in earlier (glacial or interglacial) as well as in later (postglacial) periods, as has been documented also in other polyploid complexes (e.g., Brandrud et al., 2020; Rojas-Andrés et al., 2020). Nevertheless, details regarding polyploid origins, such as their source geographic areas, specific diploid or lower-ploidy ancestral lineages and the time of their origin, remain largely unresolved. Franzke and Hurka (2000) proposed that polyploids of the phylogenetic lineage delimited in Central and Northern Europe evolved postglacially. Here we infer that its postglacial origin is highly improbable and that it probably originated in earlier times of Pleistocene glacial-interglacial periods. In fact, the recently published divergence time estimates

in the Cardamineae tribe suggests that the entire *C. pratensis* species complex originated at the turn of the Pliocene-Pleistocene (Huang et al., 2020).

The present results have also revealed that, under favorable conditions, the species studied may hybridize. *Cardamine pratensis* is common and widespread in Central Europe, and locally in lowland to montane areas it may grow in close proximity to *C. majovskyi*. A relatively high proportion of genetically admixed populations in their contact zone (found mainly in Slovenia and eastern Slovakia) indicates that gene flow occurs between them. In some of these populations, hybridization was manifested also by increased genome size variation.

Another case of interspecific gene flow was traced between *C. matthioli* and *C. rivularis* in Bulgarian mountains. Even though *C. matthioli* occurs predominantly at lower elevations and only rarely reaches the upper montane belt, and *C. rivularis* is generally a (sub)alpine species (Marhold, 1994b; Ančev et al., 2013), the two species meet in the West Rhodope Mts in Bulgaria. Their co-occurrence and hybridization were recorded at multiple localities within a relatively small area (Ančev et al., 2013), which is confirmed by the present results on three sites in the West Rhodope Mts, and in addition, also in the Vitosha Mts. Ančev et al. (2013) reported morphological intermediacy and triploid chromosome numbers, as well as disrupted meiosis and unbalanced embryological processes in the hybrids. They even suggested apomictic embryo development, but further investigations are needed. It has been assumed that the species contacts and hybridizations occurred relatively recently, as a result of mountain flora disturbance by human activities and because of climatic oscillations (Ančev et al., 2013). Another case of very recent interspecific hybridization in *Cardamine*, triggered by land use changes, has been inferred in the Western Alps (Urbanska et al., 1997; Mandáková et al., 2013; Zozomová-Lihová et al., 2014a).

Inference of Phylogeographic History

During glacial periods in the Pleistocene, most currently occupied areas in Central Europe were climatically unfavorable for the large-scale survival of temperate species, especially those adapted to moist lowland and montane sites (Hewitt, 2001; Birks and Willis, 2008; Janská et al., 2017). The genetic patterns presented here were presumably shaped by allopatric differentiation when the species ranges became fragmented and their populations survived in more restricted glacial refugia (Hewitt, 2004; Birks and Willis, 2008).

The large genetic variation found in *C. matthioli* and its differentiation into two lineages, strongly suggests that these plants survived the last glaciation in at least two separate refugia. The Central lineage, distributed in a relatively small region spanning from the southeastern margin of the Eastern Alps to northern Croatia, overlaps with an area commonly recognized as a potential glacial refugium and a source area for recolonization by several temperate trees and herbs (e.g., Heuertz et al., 2004, Heuertz et al., 2006; Magri et al., 2006; Bardy et al., 2010; Slovák et al., 2012; Rešetnik et al., 2016; Šrámková-Fuxová et al., 2017; Skokanová et al., 2019). The great diversity and rarity observed currently in populations belonging to the Central lineage support the scenario of *in situ* glacial survival without major population displacement. In contrast to the studies listed above, however, populations of Central lineage did not spread significantly in postglacial times, which resembles the case of *Silene hayekiana* (Durović et al., 2017). The distribution of many species is evidently limited by slow postglacial spreading rather than by environmental limitations (e.g., Svenning and Skov, 2007; Willner et al., 2009; Baer and Maron, 2019). The restricted distribution of the Central lineage stands in strong contrast to the widely extended Widespread lineage. The present-day genetic patterns do not suggest any straightforward hypothesis concerning the glacial survival and postglacial spreading of the latter lineage. Populations in glacial refugia commonly have unique haplotypes and their level of genetic diversity is high because of diversity accumulation and random allele fixation during their long-term persistence (Hewitt, 2000). Large diversity along with high to moderate rarity values in the Western Carpathians suggest that some populations may have survived in microclimatically favorable sites in the foothills of this mountain chain, in so-called 'cryptic' or 'northern' refugia (Stewart et al., 2010). The Carpathians were only scarcely glaciated (Ronikier, 2011) and hosted fragmented forest communities throughout the last glacial maximum (Jankovská and Pokorný, 2008). The existence of such 'northern' refugia has been documented for several temperate species (e.g., Těšitel et al., 2009; Schmickl et al., 2012; Slovák et al., 2012; Kolář et al., 2016; Stachurska-Swakoń et al., 2020). The locations of other potential refugial areas for *C. matthioli* remain unclear, but they may have been patchily scattered across the present range.

The occurrence of *C. majovskyi* in two disjunct areas overlaps with the glacial refugia proposed for its parental diploid lineages, so it seems very likely that *C. majovskyi* originated independently in the same areas where it is currently distributed and that it did not spread significantly afterward. An open question remains if this tetraploid originated during the glacial period or if it has arisen postglacially, which may be supported by its ecological optimum shifted toward higher precipitation and temperatures.

The three longitudinally correlated and ecologically differentiated genetic lineages detected within *C. pratensis* indicate the existence of three separate colonization routes from different refugia, but at the same time substantial genetic admixture among them is suggesting also major postglacial shuffling. In addition, each of them consists of at least two ploidy levels, multiple cytotypes of different origins (e.g., tetraploids with 30 and 32 chromosomes) as well as lowland to subalpine populations, which complicates the reconstruction of their phylogeographic history. Genetic diversity patterns do not indicate any clear geographic trends that are expected for simple scenarios of out-of-refugia colonization routes (Hewitt, 1999; Hewitt, 2000; Nieto Feliner, 2014), but are largely governed by cytotypic distribution patterns. Only the geographically restricted Red lineage (Western Carpathians) appears genetically highly diverse and some populations also exhibited high rarity values, which supports their glacial survival *in situ*, in cryptic microrefugia in the Western Carpathians, similarly as assumed and discussed for *C. matthioli*. Both the more widely distributed

Orange and Yellow lineages tend to possess greater rarity values in the Alps and the Carpathians, suggesting that some refugial populations may have survived in their foothills providing favorable humid habitats, followed by northward colonization only postglacially.

Finally, populations of the (sub)alpine taxon *C. rivularis* from the Southern Carpathians and Bulgarian mountains have probably responded to glaciations only by small-scale and elevational range shifts. Mountainous regions in Southern Europe have offered multiple favorable sites for long-term plant survival during climatic oscillations, and the extinction of genotypes and populations was minimized here (Gómez and Lunt, 2007; Nieto Feliner, 2011, 2014). In addition, elevational range shifts may have caused past contacts between *C. rivularis* with *C. matthioli*. Bulgarian populations of *C. matthioli* are genetically slightly differentiated from the rest of the species and exhibit greater affinity to *C. rivularis*. This indicates that contacts and gene flow between *C. matthioli* and *C. rivularis* probably do not occur only at present, but may have happened already in much earlier times.

Taxonomic Implications

The highly polymorphic species *C. pratensis* has been extensively investigated by many authors since the 1950s, who attempted to split it into more homogeneous entities (e.g., Lövkvist, 1956; Dersch, 1969; Urbanska-Worytkiewicz and Landolt, 1974; Landolt, 1984; Marhold, 1994b; reviewed by Marhold et al., 2018). The genetic patterns revealed in the present study, however, stand in strong contrast with previous taxonomic concepts, as they support none of the putative taxa or entities such as *C. nemorosa*, *C. udicola, C. rivularis* auct. or the 'ucranica' morphotype. For instance, morphologically similar (sub-)alpine populations commonly attributed to *C. rivularis* auct. are not genetically closely related and have probably originated multiple times from adjacent lower-elevation populations. Similar scenarios of independent colonization of alpine habitats by several distinct genetic lineages from foothill areas were recently inferred also in *Arabidopsis arenosa* (Kolář et al., 2017) and in the formerly broadly conceived species *Alyssum cuneifolium* (Španiel et al., 2019). In both these cases, the overall phenotypic similarity of alpine populations contrasting with their genetic heterogeneity has been caused by a similar (alpine) environment (Kolář et al., 2017; Monnahan et al., 2019; Španiel et al., 2019; Wos et al., 2019).

The observed genetic structure in *C. pratensis* indicates spatial segregation in the longitudinal direction, but it is not strongly pronounced, with more than one quarter of populations considered inter-lineage hybrids. Genetic differentiation is likely maintained by spatially restricted gene flow, as geographically neighboring lineages are genetically more similar than spatially separated ones. We therefore conclude that there are no strict borders among the lineages, which could potentially serve as basis for a new taxonomic concept, and that the splitting of *C. pratensis*, despite its high genetic, cytotypic and ecological heterogeneity, into segregate taxa cannot be supported.

Although *C. matthioli* and *C. majovskyi* are genetically closely related, which reflects the autotetraploid origin of the latter, their morphological differentiation (Marhold and Záborský, 1986; Lihová and Marhold, 2003), the ecological niche divergence documented here, and rare records of species co-occurrence suggest that the tetraploid represents an independent evolutionary unit, supporting its separate taxonomic treatment (Soltis et al., 2007).

In conclusion, the present study provides detailed insights into the complicated cytotypic and genetic structure of the *C. pratensis* species complex in Central Europe. We identified an interplay of diverse processes that have driven the evolution of the species under study, including allopatric and ecological divergence, hybridization, and multiple polyploid origins in different times, as well as genetic reshuffling caused by Pleistocene climate-induced range dynamics. It is fascinating to see to what extent different evolutionary patterns – regarding cytotypic variation and polyploid evolution as well as phylogeographic scenarios – have evolved in these closely related species.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://data.mendeley.com/datasets/z8tstt9prd/1 and https://www.ncbi.nlm.nih.gov/, PRJNA638616.

AUTHOR CONTRIBUTIONS

KM and JZ-L conceived and designed the study. BŠ, KS, TK, MŠ, and KM sampled plant material. AM, TK, MŠ, BŠ, KS, MS, GŠ, JZ-L, MC, TM, ML, and LM generated data and performed data analyses. MŠ, JZ-L, KM, and MS wrote the manuscript. MK designed the Hyb-Seq probes. All authors have read, revised, and approved the final manuscript.

FUNDING

This work was supported by research grants from the Czech Science Foundation (grants GAČR 16-10809S, 19-06632S and 19-03442S), the Slovak Research and Development Agency (APVV; grant APVV-17-0616), and the CEITEC 2020 project (grant LQ1601). MS was supported by the Operational Programme Integrated Infrastructure (OPII), funded by the ERDF (ITMS 313011T721).

ACKNOWLEDGMENTS

We thank Roswitha Schmickl (Prague) for her advice on the design of Hyb-Seq probes, for sharing the Hyb-Seq laboratory protocol, and valuable advice regarding its optimization. We also thank Vojtěch Zeisek (Prague) for his advice and discussion on Hyb-Seq data processing. We appreciate the assistance of all that contributed to plant sampling: Iva Hodálová (Bratislava), Filip Holič (Prague), Filip Kolář (Prague, Oslo), Charly Rey, Sabine

Rey-Carron (both Conthey), Stanislav Španiel (Bratislava), and Eliška Záveská (Prague). Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.588856/full#supplementary-material

REFERENCES

Ančev, M. (2006). Polyploidy and hybridization in Bulgarian Brassicaceae (Cruciferae): distribution and evolutionary role. *Phytol. Balcan.* 12, 357–366.

Ančev, M., Yurukova-Grancharova, P., Ignatova, P., Goranova, V., Stoyanov, S., Yankova-Tsvetkova, E., et al. (2013). *Cardamine* × *rhodopaea* (Brassicaceae), a triploid hybrid from the West Rhodope Mts: Morphology, distribution, relationships and origin. *Phytol. Balcan.* 19, 323–338.

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* 62, 245–253. doi: 10.1111/j.1541-0420.2005.00440.x

Arrigo, N., de La Harpe, M., Litsios, G., Zozomová-Lihová, J., Španiel, S., Marhold, K., et al. (2016). Is hybridization driving the evolution of climatic niche in *Alyssum montanum*? *Amer. J. Bot.* 103, 1348–1357. doi: 10.3732/ajb.1500368

Baduel, P., Bray, S., Vallejo-Marin, M., Kolář, F., and Yant, L. (2018). The "Polyploid Hop": shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Front. Ecol. Evol.* 6:117. doi: 10.3389/fevo.2018.00117

Baer, K. C., and Maron, J. L. (2019). Declining demographic performance and dispersal limitation influence the geographic distribution of the perennial forb *Astragalus utahensis* (Fabaceae). *J. Ecol.* 107, 1250–1262. doi: 10.1111/1365-2745.13086

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bardy, K. E., Albach, D. C., Schneeweiss, G. M., Fischer, M. A., and Schönswetter, P. (2010). Disentangling phylogeography, polyploid evolution and taxonomy of a woodland herb (*Veronica chamaedrys* group, Plantaginaceae s.l.) in southeastern Europe. *Molec. Phylogen. Evol.* 57, 771–786. doi: 10.1016/j.ympev.2010.06.025

Bastkowski, S., Mapleson, D., Spillner, A., Wu, T., Balvociute, M., and Moulton, V. (2018). SPECTRE: a suite of phylogenetic tools for reticulate evolution. *Bioinformatics* 34, 1056–1057. doi: 10.1093/bioinformatics/btx740

Bennetzen, J. L., Ma, J., and Devos, K. M. (2005). Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* 95, 127–132. doi: 10.1093/aob/mci008

Birks, H. J. B., and Willis, K. J. (2008). Alpines, trees, and refugia in Europe. *Pl. Ecol. Div.* 1, 147–160. doi: 10.1080/17550870802349146

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ.* 4:e1660. doi: 10.7717/peerj.1660

Brandrud, M. K., Baar, J., Lorenzo, M. T., Athanasiadis, A., Bateman, R. M., Chase, M. W., et al. (2020). Phylogenomic relationships of diploids and the origins of allotetraploids in *Dactylorhiza* (Orchidaceae). *Syst. Biol.* 69, 91–109. doi: 10.1093/sysbio/syz035

Brukhin, V., Osadtchiy, J. V., Florez-Rueda, A. M., Smetanin, D., Bakin, E., Nobre, M. S., et al. (2019). The *Boechera* genus as a resource for apomixis research. *Front. Plant Sci.* 10:392. doi: 10.3389/fpls.2019.00392

Bruvo, R., Michiels, N. K., D'Souza, T. G., and Schulenburg, H. (2004). A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molec. Ecol.* 13, 2101–2106. doi: 10.1111/j.1365-294X.2004.02209.x

Carter, K. A., Liston, A., Bassil, N. V., Alice, L. A., Bushakra, J. M., Sutherland, B. L., et al. (2019). Target capture sequencing unravels *Rubus* evolution. *Front. Plant Sci.* 10:1615. doi: 10.3389/fpls.2019.01615

Castro, M., Loureiro, J., Serrano, M., Tavares, D., Husband, B. C., Siopa, C., et al. (2019). Mosaic distribution of cytotypes in a mixed-ploidy plant species, *Jasione montana*: nested environmental niches but low geographical overlap. *Biol. J. Linn. Soc.* 190, 51–66. doi: 10.1093/botlinnean/boz007

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* 65, 997–1008. doi: 10.1093/sysbio/syw037

Clark, L. V., and Jasieniuk, M. (2011). POLYSAT: an R package for polyploid microsatellite analysis. *Molec. Ecol. Resour.* 11, 562–566. doi: 10.1111/j.1755-0998.2011.02985.x

Coyne, J. A., and Orr, H. A. (2004). *Speciation*. Sunderland: Sinauer Associates.

Dauphin, B., Grant, J. R., Farrar, D. R., and Rothfels, C. J. (2018). Rapid allopolyploid radiation of moonwort ferns (*Botrychium*; Ophioglossaceae) revealed by PacBio sequencing of homologous and homeologous nuclear regions. *Molec. Phylogen. Evol.* 120, 342–353. doi: 10.1016/j.ympev.2017.11.025

Dersch, G. (1969). Über das Vorkommen von diploidem Wiesenschaumkraut (*Cardamine pratensis* L.) in Mitteleuropa. *Ber. Deutsch. Bot. Ges.* 82, 201–207.

Doležel, J., Greilhuber, J., and Suda, J. (2007a). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2, 2233–2244. doi: 10.1038/nprot.2007.310

Doležel, J., Greilhuber, J., and Suda, J. (2007b). "Flow cytometry with plants: an overview," in *Flow cytometry with plant cells, analysis of genes, chromosomes and genomes*, eds J. Doležel, J. Greilhuber, and J. Suda (Weinheim: Wiley–VCH), 41–65. doi: 10.1002/9783527610921

Doležel, J., Sgorbati, S., and Lucretti, S. (1992). Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol. Pl.* 85, 625–631. doi: 10.1111/j.1399-3054.1992.tb04764.x

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46. doi: 10.1111/j.1600-0587.2012.07348.x

Doyle, J. J., Doyle, J. L., Brown, A. H., and Grace, J. P. (1990). Multiple origins of polyploids in the *Glycine tabacina* complex inferred from chloroplast DNA polymorphism. *Proc. Natl. Acad. Sci. U S A.* 87, 714–717. doi: 10.1073/pnas.87.2.714

Dray, S., and Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* 22, 1–20. doi: 10.18637/jss.v022.i04

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molec. Ecol.* 23, 40–69. doi: 10.1111/mec.12581

Durović, S., Schönswetter, P., Niketić, M., Tomović, G., and Frajman, B. (2017). Disentangling relationships among the members of the *Silene saxifraga* alliance (Caryophyllaceae): Phylogenetic structure is geographically rather than taxonomically segregated. *Taxon* 66, 343–364. doi: 10.12705/662.4

Ehrich, D. (2006). AFLPdat: a collection of R functions for convenient handling of AFLP data. *Molec. Ecol. Notes* 6, 603–604. doi: 10.1111/j.1471-8286.2006.01380.x

Esselink, G. D., Nybom, H., and Vosman, B. (2004). Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting—peak ratios) method. *Theor. Appl. Genet.* 109, 402–408. doi: 10.1007/s00122-004-1645-5

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molec. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131, 479–491.

Feulner, M., Weig, A., Voss, T., Schott, L. F., and Aas, G. (2019). Central European polyploids of *Sorbus* subgenus *Aria* (Rosaceae) recurrently evolved from diploids of central and south-eastern Europe: evidence from microsatellite data. *Bot. J. Linn. Soc.* 191, 315–324. doi: 10.1093/botlinnean/boz053

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi: 10.1002/joc.5086

Frajman, B., Rešetnik, I., Niketić, M., Ehrendorfer, F., and Schönswetter, P. (2016). Patterns of rapid diversification in heteroploid *Knautia* sect. *Trichera* (Caprifoliaceae, Dipsacoideae), one of the most intricate taxa of the European flora. *BMC Evol. Biol.* 16:204. doi: 10.1186/s12862-016-0773-2

Franzke, A., and Hurka, H. (2000). Molecular systematics and biogeography of the *Cardamine pratensis* complex (Brassicaceae). *Pl. Syst. Evol.* 224, 213–234. doi: 10.1007/BF00986344

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Gómez, A., and Lunt, D. H. (2007). "Refugia within Refugia: Patterns of Phylogeographic Concordance in the Iberian Peninsula," in *Phylogeography of Southern European Refugia*, eds S. Weiss and N. Ferrand (Dordrecht: Springer), 155–188. doi: 10.1007/1-4020-4904-8_5

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Greilhuber, J., Dolezel, J., Lysák, M. A., and Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. *Ann. Bot.* 95, 255–260. doi: 10.1093/aob/mci019

Grünewald, S., Spillner, A., Bastkowski, S., Bögershausen, A., and Moulton, V. (2013). SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 151–160. doi: 10.1109/TCBB.2013.8

Hanušová, K., Čertner, M., Urfus, T., Koutecký, P., Košnar, J., Rothfels, C. J., et al. (2019). Widespread co-occurrence of multiple ploidy levels in fragile ferns (*Cystopteris fragilis* complex; Cystopteridaceae) probably stems from similar ecology of cytotypes, their efficient dispersal and inter-ploidy hybridization. *Ann. Bot.* 123, 845–855. doi: 10.1093/aob/mcy219

Heibl, C. (2008). *PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages*. URL: Available online at: http://www.christophheibl.de/Rpackages.html

Heuertz, M., Carnevale, S., Fineschi, S., Sebastiani, F., Hausman, J. F., Paule, L., et al. (2006). Chloroplast DNA phylogeography of European ashes, *Fraxinus* sp. (Oleaceae): roles of hybridization and life history traits. *Molec. Ecol.* 15, 2131–2140. doi: 10.1111/j.1365-294X.2006.02897.x

Heuertz, M., Fineschi, S., Anzidei, M., Pastorelli, R., Salvini, D., Paule, et al. (2004). Chloroplast DNA variation and postglacial recolonization of common ash (*Fraxinus excelsior* L.) in Europe. *Molec. Ecol.* 13, 3437–3452. doi: 10.1111/j.1365-294X.2004.02333.x

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature* 405, 907–913. doi: 10.1038/35016000

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biol. J. Linn. Soc.* 68, 87–112. doi: 10.1006/bijl.1999.0332

Hewitt, G. M. (2001). Speciation, hybrid zones and phylogeography—or seeing genes in space and time. *Molec. Ecol.* 10, 537–549.

Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philos. Trans. Roy. Soc. London B Biol. Sci.* 359, 183–195. doi: 10.1098/rstb.2003.1388

Hodel, R. G., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., et al. (2016). The report of my death was an exaggeration: a review for researchers using microsatellites in the 21st century. *Appl. Pl. Sci.* 4:1600025. doi: 10.3732/apps.1600025

Hu, Y. N., Zhao, L., Buggs, R. J., Zhang, X. M., Li, J., and Wang, N. (2019). Population structure of *Betula albosinensis* and *Betula platyphylla*: evidence for hybridization and a cryptic lineage. *Ann. Bot.* 123, 1179–1189. doi: 10.1093/aob/mcz024

Huang, X. C., German, D. A., and Koch, M. A. (2020). Temporal patterns of diversification in Brassicaceae demonstrate decoupling of rate shifts and mesopolyploidization events. *Ann. Bot.* 125, 29–47. doi: 10.1093/aob/mcz123

Hülber, K., Sonnleitner, M., Suda, J., Krejčíková, J., Schönswetter, P., Schneeweiss, G. M., et al. (2015). Ecological differentiation, lack of hybrids involving diploids, and asymmetric gene flow between polyploids in narrow contact zones of *Senecio carniolicus* (syn. *Jacobaea carniolica, Asteraceae)*. *Ecol. Evol.* 5, 1224–1234. doi: 10.1002/ece3.1430

Husband, B. C. (2004). The role of triploid hybrids in the evolutionary dynamics of mixed-ploidy populations. *Biol. J. Linn. Soc.* 82, 537–546. doi: 10.1111/j.1095-8312.2004.00339.x

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Jankovská, V., and Pokorný, P. (2008). Forest vegetation of the last full-glacial period in the Western Carpathians (Slovakia and Czech Republic). *Preslia* 80, 307–324.

Janská, V., Jiménez-Alfaro, B., Chytrý, M., Divíšek, J., Anenkhonov, O., Korolyuk, A., et al. (2017). Palaeodistribution modelling of European vegetation types at the Last Glacial Maximum using modern analogues from Siberia: Prospects and limitations. *Quat. Sci. Rev.* 159, 103–115. doi: 10.1016/j.quascirev.2017.01.011

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., et al. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Pl. Sci.* 4:1600016. doi: 10.3732/apps.1600016

Junier, T., and Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* 26, 1669–1670. doi: 10.1093/bioinformatics/btq243

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kamvar, Z. N., Brooks, J. C., and Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front. Genet.* 6:208. doi: 10.3389/fgene.2015.00208

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ.* 2:e281. doi: 10.7717/peerj.281

Kates, H. R., Johnson, M. G., Gardner, E. M., Zerega, N. J., and Wickett, N. J. (2018). Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *Amer. J. Bot.* 105, 404–416. doi: 10.1002/ajb2.1068

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* 9, 286–298. doi: 10.1093/bib/bbn013

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Knaus, B. J., and Grünwald, N. J. (2016). VcfR: a package to manipulate and visualize VCF format data in R. *bioRxiv* 041277. doi: 10.1101/041277

Knaus, B. J., and Grünwald, N. J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Molec. Ecol. Resour.* 17, 44–53. doi: 10.1111/1755-0998.12549

Köhler, C., Mittelsten Scheid, O., and Erilova, A. (2010). The impact of the triploid block on the origin and evolution of polyploid plants. *Trends Genet.* 26, 142–148. doi: 10.1016/j.tig.2009.12.006

Kolář, F., Čertner, M., Suda, J., Schönswetter, P., and Husband, B. C. (2017). Mixed-ploidy species: progress and opportunities in polyploid research. *Trends Pl. Sci.* 22, 1041–1055. doi: 10.1016/j.tplants.2017.09.011

Kolář, F., Fuxová, G., Záveská, E., Nagano, A. J., Hyklová, L., Lučanová, M., et al. (2016). Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model *Arabidopsis arenosa*. *Molec. Ecol.* 25, 3929–3949. doi: 10.1111/mec.13721

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi: 10.1093/bioinformatics/btz305

Kučera, J., Valko, I., and Marhold, K. (2005). On-line database of the chromosome numbers of the genus *Cardamine* (Brassicaceae). *Biologia* 60, 473–476.

Kuhn, M. (2019). *caret: Classification and Regression Training. R package version 6.0-84*. URL: https://CRAN.R-project.org/package = caret.

Landolt, E. (1984). Über die Artengruppe der *Cardamine pratensis* L. in der Schweiz [Some remarks on the species group of *Cardamine pratensis* L. s. l. in Switzerland]. *Diss. Bot.* 72, 481–497.

Lawrence, W. J. C. (1931). The chromosome constitution of *Cardamine pratensis* and *Verbascum phoeniceum*. *Genetica* 13, 183–208. doi: 10.1007/BF01725043

Levin, D. A. (1975). Minority cytotype exclusion in local plant populations. *Taxon* 24, 35–43. doi: 10.2307/1218997

Li, F.-W., Rushworth, C. A., Beck, J. B., and Windham, M. D. (2017). *Boechera* microsatellite website: an online portal for species identification and determination of hybrid parentage. *Database* 217:baw169. doi: 10.1093/database/baw169

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lihová, J., and Marhold, K. (2003). Taxonomy and distribution of the *Cardamine pratensis* group (Brassicaceae) in Slovenia. *Phyton* 43, 241–261.

Lihová, J., and Marhold, K. (2006). "Phylogenetic and diversity patterns in *Cardamine* (Brassicaceae) - a genus with conspicuous polyploid and reticulate evolution," in *Plant genome: biodiversity and evolution, vol. 1C: Phanerogams (Angiosperms - Dicotyledons)*, eds A. K. Sharma and A. Sharma (Enfield: Science Publishers, Inc), 149–186.

Lihová, J., Tribsch, A., and Marhold, K. (2003). The *Cardamine pratensis* (Brassicaceae) group in the Iberian Peninsula: taxonomy, polyploidy and distribution. *Taxon* 52, 783–802. doi: 10.2307/3647352

Lihová, J., Tribsch, A., and Stuessy, T. F. (2004). *Cardamine apennina*: a new endemic diploid species of the *C. pratensis* group (Brassicaceae) from Italy. *Pl. Syst. Evol.* 245, 69–92. doi: 10.1007/s00606-003-0119-6

Lövkvist, B. (1956). The *Cardamine pratensis* complex. *Symb. Bot. UPS.* 14, 1–131.

Lynch, M. (1990). The similarity index and DNA fingerprinting. *Molec. Biol. Evol.* 7, 478–484. doi: 10.1093/oxfordjournals.molbev.a040620

Ma, J. X., Li, Y. N., Vogl, C., Ehrendorfer, F., and Guo, Y. P. (2010). Allopolyploid speciation and ongoing backcrossing between diploid progenitor and tetraploid progeny lineages in the *millefolium* species complex: Analyses of single-copy nuclear genes and genomic AFLP. *BMC Evol. Biol.* 10:100. doi: 10.1186/1471-2148-10-100

Magri, D., Vendramin, G. G., Comps, B., Dupanloup, I., Geburek, T., Gömöry, D., et al. (2006). A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytol.* 171, 199–221. doi: 10.1111/j.1469-8137.2006.01740.x

Mandák, B., Krak, K., Vít, P., Lomonosova, M. N., Belyayev, A., Habibi, F., et al. (2018). Hybridization and polyploidization within the *Chenopodium album* aggregate analysed by means of cytological and molecular markers. *Molec. Phylogen. Evol.* 129, 189–201. doi: 10.1016/j.ympev.2018.08.016

Mandáková, T., and Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Pl. Biol.* 42, 55–65. doi: 10.1016/j.pbi.2018.03.001

Mandáková, T., Gloss, A. D., Whiteman, N. K., and Lysak, M. A. (2016). How diploidization turned a tetraploid into a pseudotriploid. *Amer. J. Bot.* 103, 1187–1196. doi: 10.3732/ajb.1500452

Mandáková, T., Kovařík, A., Zozomová-Lihová, J., Shimizu-Inatsugi, R., Shimizu, K. K., Mummenhoff, K., et al. (2013). The more the merrier: recent hybridization and polyploidy in *Cardamine*. *Pl. Cell* 25, 3280–3295. doi: 10.1105/tpc.113.114405

Mandáková, T., Zozomová-Lihová, J., Kudoh, H., Zhao, Y., Lysak, M. A., and Marhold, K. (2019). The story of promiscuous crucifers: origin and genome evolution of an invasive species, *Cardamine occulta* (Brassicaceae), and its relatives. *Ann. Bot.* 124, 209–220. doi: 10.1093/aob/mcz019

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, 209–220.

Marhold, K. (1992). *Rod Cardamine v Karpatoch a Panónii (The Genus Cardamine in the Carpathians and Pannonia)*. PhD thesis, Bratislava: Comenius University.

Marhold, K. (1994a). Chromosome numbers of the genus *Cardamine* L. (Cruciferae) in the Carpathians and in Pannonia. *Phyton* 34, 19–34.

Marhold, K. (1994b). Taxonomy of the genus *Cardamine* L. (Cruciferae) in the Carpathians and Pannonia. *Folia Geobot. Phytotax.* 29, 335–374.

Marhold, K. (1995). *Cardamine rivularis* auct. non Schur in the Eastern Alps. *Carinthia II* 53, 101–102.

Marhold, K. (1996). Multivariate morphometric study of the *Cardamine pratensis* group (Cruciferae) in the Carpathian and Pannonian area. *Pl. Syst. Evol.* 200, 141–159. doi: 10.1007/BF00984932

Marhold, K. (2000). Chromosome numbers of the *Cardamine pratensis* group in Austria with taxonomic remarks. *Fl. Austr. Novit.* 6, 1–6.

Marhold, K., and Záborský, J. (1986). A new species of *Cardamine pratensis* agg. from Eastern Slovakia. *Preslia* 58, 193–198.

Marhold, K., Lihová, J., Perný, M., Grupe, R., and Neuffer, B. (2002). Natural hybridization in *Cardamine* (Brassicaceae) in the Pyrenees: evidence from morphological and molecular data. *Bot. J. Linn. Soc.* 139, 275–294. doi: 10.1046/j.1095-8339.2002.00066.x

Marhold, K., Šlenker, M., and Zozomová-Lihová, J. (2018). Polyploidy and hybridization in the Mediterranean and neighbouring areas towards the north: examples from the genus *Cardamine* (Brassicaceae). *Biol. Serbica* 40, 47–59. doi: 10.5281/zenodo.1406320

Martin, M., Patterson, M., Garg, S., Fischer, S., Pisanti, N., Klau, G. W., et al. (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv* 085050. doi: 10.1101/085050

Mason, A. S., and Pires, J. C. (2015). Unreduced gametes: meiotic mishap or evolutionary mechanism? *Trends Genet.* 31, 5–10. doi: 10.1016/j.tig.2014.09.011

McCullagh, P., and Nelder, J. A. (1989). *Generalized linear models*, 2nd Edn. Boca Raton, FL: Chapman & Hall/CRC.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meirmans, P. G., and van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molec. Ecol. Notes* 4, 792–794. doi: 10.1111/j.1471-8286.2004.00770.x

Meirmans, P. G., Liu, S., and van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *J. Heredity* 109, 283–296. doi: 10.1093/jhered/esy006

Melichárková, A., Španiel, S., Marhold, K., Hurdu, B. I., Drescher, A., and Zozomová-Lihová, J. (2019). Diversification and independent polyploid origins in the disjunct species *Alyssum repens* from the Southeastern Alps and the Carpathians. *Amer. J. Bot* 106, 1499–1518. doi: 10.1002/ajb2.1370

Monnahan, P., Kolář, F., Baduel, P., Sailer, C., Koch, J., Horvath, R., et al. (2019). Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nat. Ecol. Evol.* 3, 457–468. doi: 10.1038/s41559-019-0807-4

Nieto Feliner, G. (2011). Southern European glacial refugia: a tale of tales. *Taxon* 60, 365–372. doi: 10.1002/tax.602007

Nieto Feliner, G. (2014). Patterns and processes in plant phylogeography in the Mediterranean Basin. *A Rev. Perspect. Pl. Ecol. Syst.* 16, 265–278. doi: 10.1016/j.ppees.2014.07.002

Novikova, P. Y., Hohmann, N., and Van de Peer, Y. (2018). Polyploid *Arabidopsis* species originated around recent glaciation maxima. *Curr. Opin. Plant Biol.* 42, 8–15. doi: 10.1016/j.pbi.2018.01.005

Novikova, P. Y., Tsuchimatsu, T., Simon, S., Nizhynska, V., Voronin, V., Burns, R., et al. (2017). Genome sequencing reveals the origin of the allotetraploid *Arabidopsis suecica*. *Molec. Biol. Evol.* 34, 957–968. doi: 10.1093/molbev/msw299

Obbard, D. J., Harris, S. A., and Pannell, J. R. (2006). Simple allelic-phenotype diversity and differentiation statistics for allopolyploids. *Heredity* 97, 296–303. doi: 10.1038/sj.hdy.6800862

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2019). *vegan: Community Ecology Package. R Package Version 2.5-5*. URL: http://CRAN.R-project.org/package = vegan

Pachschwöll, C., Garcia, P. E., Winkler, M., Schneeweiss, G. M., and Schönswetter, P. (2015). Polyploidisation and geographic differentiation drive diversification in a European high mountain plant group (*Doronicum clusii* aggregate, Asteraceae). *PLoS One* 10:e0118197. doi: 10.1371/journal.pone.0118197

Padilla-García, N., Rojas-Andrés, B. M., López-González, N., Castro, M., Castro, S., Loureiro, J., et al. (2018). The challenge of species delimitation in the diploid-polyploid complex *Veronica* subsection *Pentasepalae*. *Molec. Phylogen. Evol.* 119, 196–209. doi: 10.1016/j.ympev.2017.11.007

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

Paule, J., Gregor, T., Schmidt, M., Gerstner, E. M., Dersch, G., Dressler, S., et al. (2017). Chromosome numbers of the flora of Germany—A new online database of georeferenced chromosome counts and flow cytometric ploidy estimates. *Pl. Syst. Evol.* 303, 1123–1129. doi: 10.1007/s00606-016-1362-y

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Evol. Syst.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya A* 26, 329–358.

Rešetnik, I., Frajman, B., and Schönswetter, P. (2016). Heteroploid *Knautia drymeia* includes *K. gussonei* and cannot be separated into diagnosable subspecies. *Amer. J. Bot.* 103, 1300–1313. doi: 10.3732/ajb.1500506

Rojas-Andrés, B. M., Padilla-García, N., de Pedro, M., López-González, N., Delgado, L., Albach, D. C., et al. (2020). Environmental differences are correlated with the distribution pattern of cytotypes in *Veronica* subsection *Pentasepalae* at a broad scale. *Ann. Bot.* 125, 471–484. doi: 10.1093/aob/mcz182

Ronikier, M. (2011). Biogeography of high mountain plants in the Carpathians: an emerging phylogeographical perspective. *Taxon* 60, 373–389. doi: 10.1002/tax.602008

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. *Molec. Ecol. Notes* 4, 137–138. doi: 10.1046/j.1471-8286.2003.00566.x

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., et al. (2010). The NCEP climate forecast system reanalysis. *Bull. Amer. Meteorol. Soc.* 91, 1015–1058. doi: 10.1175/2010BAMS3001.1

Sampson, J. F., and Byrne, M. (2012). Genetic diversity and multiple origins of polyploid *Atriplex nummularia* Lindl. (Chenopodiaceae). *Biol. J. Linn. Soc.* 105, 218–230. doi: 10.1111/j.1095-8312.2011.01787.x

Schmickl, R., Liston, A., Zeisek, V., Oberlander, K., Weitemier, K., Straub, S. C., et al. (2016). Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molec. Ecol. Resour.* 16, 1124–1135. doi: 10.1111/1755-0998.12487

Schmickl, R., Paule, J., Klein, J., Marhold, K., and Koch, M. A. (2012). The evolutionary history of the *Arabidopsis arenosa* complex: diverse tetraploids mask the Western Carpathian center of species and genetic diversity. *PLoS One* 7:e42691. doi: 10.1371/journal.pone.0042691

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., et al. (2014). GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theor. Appl. Climatol.* 115, 15–40. doi: 10.1007/s00704-013-0860-x

Schönswetter, P., and Tribsch, A. (2005). Vicariance and dispersal in the alpine perennial *Bupleurum stellatum* L. *(Apiaceae). Taxon* 54, 725–732. doi: 10.2307/25065429

Schönswetter, P., Suda, J., Popp, M., Weiss-Schneeweiss, H., and Brochmann, C. (2007). Circumpolar phylogeography of *Juncus biglumis* (Juncaceae) inferred from AFLP fingerprints, cpDNA sequences, nuclear DNA content and chromosome numbers. *Molec. Phylogen. Evol.* 42, 92–103. doi: 10.1016/j.ympev.2006.06.016

Servick, S., Visger, C. J., Gitzendanner, M. A., Soltis, P. S., and Soltis, D. E. (2015). Population genetic variation, geographic structure, and multiple origins of autopolyploidy in *Galax urceolata*. *Amer. J. Bot.* 102, 973–982. doi: 10.3732/ajb.1400554

Skokanová, K., Šingliarová, B., Kochjarová, J., and Paule, J. (2019). Nuclear ITS and AFLPs provide surprising implications for the taxonomy of *Tephroseris longifolia* agg. and the endemic status of *T. longifolia* subsp. *moravica*. *Pl. Syst. Evol.* 305, 865–884. doi: 10.1007/s00606-019-01624-z

Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* 6:31. doi: 10.1186/1471-2105-6-31

Slovák, M., Kučera, J., Turis, P., and Zozomová-Lihová, J. (2012). Multiple glacial refugia and postglacial colonization routes inferred for a woodland geophyte, *Cyclamen purpurascens*: patterns concordant with the Pleistocene history of broadleaved and coniferous tree species. *Biol. J. Linn. Soc.* 105, 741–760. doi: 10.1111/j.1095-8312.2011.01826.x

Soltis, D. E., Soltis, P. S., and Tate, J. A. (2003). Advances in the study of polyploidy since plant speciation. *New Phytol.* 161, 173–191. doi: 10.1046/j.1469-8137.2003.00948.x

Soltis, D. E., Soltis, P. S., Pires, J. C., Kovařík, A., Tate, J. A., and Mavrodiev, E. (2004). Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biol. J. Linn. Soc.* 82, 485–501. doi: 10.1111/j.1095-8312.2004.00335.x

Soltis, D. E., Soltis, P. S., Schemske, D. W., Hancock, J. F., Thompson, J. N., Husband, B. C., et al. (2007). Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* 56, 13–30. doi: 10.2307/25065732

Soltis, P. S., and Soltis, D. E. (2009). The role of hybridization in plant speciation. *Ann. Rev. Pl. Biol.* 60, 561–588. doi: 10.1146/annurev.arplant.043008.092039

Soltis, P. S., Liu, X., Marchant, D. B., Visger, C. J., and Soltis, D. E. (2014). Polyploidy and novelty: Gottlieb's legacy. *Philos. Trans. Roy. Soc. London B Biol. Sci.* 369:20130351. doi: 10.1098/rstb.2013.0351

Sonnleitner, M., Weis, B., Flatscher, R., García, P. E., Suda, J., Krejčíková, J., et al. (2013). Parental ploidy strongly affects offspring fitness in heteroploid crosses among three cytotypes of autopolyploid *Jacobaea carniolica* (Asteraceae). *PLoS One* 8:e78959. doi: 10.1371/journal.pone.0078959

Španiel, S., Marhold, K., and Zozomová-Lihová, J. (2017). The polyploid *Alyssum montanum-A. repens* complex in the Balkans: a hotspot of species and genetic diversity. *Pl. Syst. Evol.* 303, 1443–1465. doi: 10.1007/s00606-017-1470-3

Španiel, S., Marhold, K., and Zozomová-Lihová, J. (2019). Polyphyletic *Alyssum cuneifolium* (Brassicaceae) revisited: Morphological and genome size differentiation of recently recognized allopatric taxa. *J. Syst. Evol.* 57, 287–301. doi: 10.1111/jse.12464

Španiel, S., Marhold, K., Filová, B., and Zozomová-Lihová, J. (2011). Genetic and morphological variation in the diploid-polyploid *Alyssum montanum* in Central Europe: taxonomic and evolutionary considerations. *Pl. Syst. Evol.* 294, 1–25. doi: 10.1007/s00606-011-0438-y

Šrámková-Fuxová, G., Záveská, E., Kolář, F., Lučanová, M., Španiel, S., and Marhold, K. (2017). Range-wide genetic structure of *Arabidopsis halleri* (Brassicaceae): glacial persistence in multiple refugia and origin of the Northern Hemisphere disjunction. *Bot. J. Linn. Soc.* 185, 321–342. doi: 10.1093/botlinnean/box064

Stachurska-Swakoń, A., Cieślak, E., Ronikier, M., Nowak, J., and Kaczmarczyk, A. (2020). Genetic structure of *Doronicum austriacum* (Asteraceae) in the Carpathians and adjacent areas: toward a comparative phylogeographical analysis of tall-herb species. *Pl. Syst. Evol.* 306:14. doi: 10.1007/s00606-020-01652-0

Stewart, J. R., Lister, A. M., Barnes, I., and Dalén, L. (2010). Refugia revisited: individualistic responses of species in space and time. *Proc. Roy. Soc. B* 277, 661–671. doi: 10.1098/rspb.2009.1272

Stift, M., Kolář, F., and Meirmans, P. G. (2019). STRUCTURE is more robust than other clustering methods in simulated mixed-ploidy populations. *Heredity* 123, 429–441. doi: 10.1038/s41437-019-0247-6

Sutherland, B. L., and Galloway, L. F. (2017). Postzygotic isolation varies by ploidy level within a polyploid complex. *New Phytol.* 213, 404–412. doi: 10.1111/nph.14116

Svenning, J. C., and Skov, F. (2007). Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? *Ecol. Lett.* 10, 453–460. doi: 10.1111/j.1461-0248.2007.01038.x

Szpiech, Z. A., Jakobsson, M., and Rosenberg, N. A. (2008). ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* 24, 2498–2504. doi: 10.1093/bioinformatics/btn478

Temsch, E. M., Greilhuber, J., and Krisai, R. (2010). Genome size in liverworts. *Preslia* 82, 63–80.

Těšitel, J., Malinová, T., Štech, M., and Herbstová, M. (2009). Variation in the *Melampyrum sylvaticum* group in the Carpathian and Hercynian region: two lineages with different evolutionary histories. *Preslia* 81, 1–22.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. Series B Stat. Methodol.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tomasello, S., Karbstein, K., Hodač, L., Paetzold, C., and Hörandl, E. (2020). Phylogenomics unravels Quaternary vicariance and allopatric speciation patterns in temperate-montane plant species: a case study on the *Ranunculus auricomus* species complex. *Molec. Ecol.* 29, 2031–2049. doi: 10.1111/mec.15458

Udall, J. A., Long, E., Ramaraj, T., Conover, J. L., Yuan, D., Grover, C. E., et al. (2019). The genome sequence of *Gossypioides kirkii* illustrates a descending dysploidy in plants. *Front. Plant Sci.* 10:1541. doi: 10.3389/fpls.2019.01541

Urbanska, K. M., Hurka, H., Landolt, E., Neuffer, B., and Mummenhoff, K. (1997). Hybridization and evolution in *Cardamine* (Brassicaceae) at Urnerboden, Central Switzerland: biosystematic and molecular evidence. *Pl. Syst. Evol.* 204, 233–256. doi: 10.1007/BF00989208

Urbanska-Worytkiewicz, K., and Landolt, E. (1974). Biosystematics investigations in *Cardamine pratensis* L. *s.l.* 1. Diploid taxa from Central Europe and their fertility relationships. *Ber. Geobot. Inst. ETH. Stiftung Rübel* 42, 43–139.

Villaverde, T., Pokorny, L., Olsson, S., Rincón-Barrado, M., Johnson, M. G., Gardner, E. M., et al. (2018). Bridging the micro-and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytol.* 220, 636–650. doi: 10.1111/nph.15312

Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., et al. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Appl. Pl. Sci.* 2:1400042. doi: 10.3732/apps.1400042

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Willner, W., Di Pietro, R., and Bergmeier, E. (2009). Phytogeographical evidence for post-glacial dispersal limitation of European beech forest species. *Ecography* 32, 1011–1018. doi: 10.1111/j.1600-0587.2009.05957.x

Wos, G., Mořkovská, J., Bohutínská, M., Šrámková, G., Knotek, A., Lučanová, M., et al. (2019). Role of ploidy in colonization of alpine habitats in natural populations of *Arabidopsis arenosa*. *Ann. Bot.* 124, 255–268. doi: 10.1093/aob/mcz070

Zozomová-Lihová, J., Krak, K., Mandáková, T., Shimizu, K. K., Španiel, S., Vít, P., et al. (2014a). Multiple hybridization events in *Cardamine* (Brassicaceae) during the last 150 years: revisiting a textbook example of neoallopolyploidy. *Ann. Bot.* 113, 817–830. doi: 10.1093/aob/mcu012

Zozomová-Lihová, J., Malánová-Krásná, I., Vít, P., Urfus, T., Senko, D., Svitok, M., et al. (2015). Cytotype distribution patterns, ecological differentiation, and genetic structure in a diploid-tetraploid contact zone of *Cardamine amara*. *Amer. J. Bot.* 102, 1380–1395. doi: 10.3732/ajb.1500052

Zozomová-Lihová, J., Mandáková, T., Kovaříková, A., Mühlhausen, A., Mummenhoff, K., Lysak, M. A., et al. (2014b). When fathers are instant losers: homogenization of rDNA loci in recently formed *Cardamine* × *schulzii* trigenomic allopolyploid. *New Phytol.* 203, 1096–1108. doi: 10.1111/nph.12873

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 12 January 2021
doi: 10.3389/fpls.2020.607893

Evolution of Tandem Repeats Is Mirroring Post-polyploid Cladogenesis in *Heliophila* (Brassicaceae)

*Mert Dogan[1,2], Milan Pouch[1,2], Terezie Mandáková[1,3], Petra Hloušková[1], Xinyi Guo[1], Pieter Winter[4], Zuzana Chumová[5], Adriaan Van Niekerk[6], Klaus Mummenhoff[7], Ihsan A. Al-Shehbaz[8], Ladislav Mucina[6,9] and Martin A. Lysak[1,2]**

[1] CEITEC, Masaryk University, Brno, Czechia, [2] NCBR, Faculty of Science, Masaryk University, Brno, Czechia, [3] Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czechia, [4] South African National Biodiversity Institute (SANBI), Kirstenbosch, Cape Town, South Africa, [5] Institute of Botany, Czech Academy of Sciences, Prùhonice, Czechia, [6] Department of Geography & Environmental Studies, Stellenbosch University, Stellenbosch, South Africa, [7] Department of Biology, Botany, Osnabrück University, Osnabrück, Germany, [8] Missouri Botanical Garden, St. Louis, MO, United States, [9] Harry Butler Institute, Murdoch University, Perth, WA, Australia

OPEN ACCESS

Edited by:
Christoph Oberprieler,
University of Regensburg, Germany

Reviewed by:
J. Chris Pires,
University of Missouri, United States
Ales Kovarik,
Academy of Sciences of the Czech Republic (ASCR), Czechia

***Correspondence:**
Martin A. Lysak
martin.lysak@ceitec.muni.cz;
lysak@sci.muni.cz

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: 18 September 2020
Accepted: 16 November 2020
Published: 12 January 2021

Citation:
Dogan M, Pouch M, Mandáková T, Hloušková P, Guo X, Winter P, Chumová Z, Van Niekerk A, Mummenhoff K, Al-Shehbaz IA, Mucina L and Lysak MA (2021) Evolution of Tandem Repeats Is Mirroring Post-polyploid Cladogenesis in Heliophila (Brassicaceae). Front. Plant Sci. 11:607893. doi: 10.3389/fpls.2020.607893

The unigeneric tribe Heliophileae encompassing more than 100 *Heliophila* species is morphologically the most diverse Brassicaceae lineage. The tribe is endemic to southern Africa, confined chiefly to the southwestern South Africa, home of two biodiversity hotspots (Cape Floristic Region and Succulent Karoo). The monospecific *Chamira* (*C. circaeoides*), the only crucifer species with persistent cotyledons, is traditionally retrieved as the closest relative of Heliophileae. Our transcriptome analysis revealed a whole-genome duplication (WGD) ~26.15–29.20 million years ago, presumably preceding the *Chamira*/*Heliophila* split. The WGD was then followed by genome-wide diploidization, species radiations, and cladogenesis in *Heliophila*. The expanded phylogeny based on nuclear ribosomal DNA internal transcribed spacer (ITS) uncovered four major infrageneric clades (A–D) in *Heliophila* and corroborated the sister relationship between *Chamira* and *Heliophila*. Herein, we analyzed how the diploidization process impacted the evolution of repetitive sequences through low-coverage whole-genome sequencing of 15 *Heliophila* species, representing the four clades, and *Chamira*. Despite the firmly established infrageneric cladogenesis and different ecological life histories (four perennials vs. 11 annual species), repeatome analysis showed overall comparable evolution of genome sizes (288–484 Mb) and repeat content (25.04–38.90%) across *Heliophila* species and clades. Among *Heliophila* species, long terminal repeat (LTR) retrotransposons were the predominant components of the analyzed genomes (11.51–22.42%), whereas tandem repeats had lower abundances (1.03–12.10%). In *Chamira*, the tandem repeat content (17.92%, 16 diverse tandem repeats) equals the abundance of LTR retrotransposons (16.69%). Among the 108 tandem repeats identified in *Heliophila*, only 16 repeats were found to be shared among two or more species; no tandem repeats

were shared by *Chamira* and *Heliophila* genomes. Six "relic" tandem repeats were shared between any two different *Heliophila* clades by a common descent. Four and six clade-specific repeats shared among clade A and C species, respectively, support the monophyly of these two clades. Three repeats shared by all clade A species corroborate the recent diversification of this clade revealed by plastome-based molecular dating. Phylogenetic analysis based on repeat sequence similarities separated the *Heliophila* species to three clades [A, C, and (B+D)], mirroring the post-polyploid cladogenesis in *Heliophila* inferred from rDNA ITS and plastome sequences.

Keywords: repetitive DNA, repeatome, whole-genome duplication (WGD), rDNA ITS, plastome phylogeny, Cruciferae, Cape flora, South Africa

INTRODUCTION

Geographically and phylogenetically well-defined groups are ideal study objects to analyze the evolution of diverse genomic parameters during long periods of isolation that prevented gene flow with other species groups. Although Brassicaceae (mustard family, Cruciferae) occur on all continents, except for Antarctica, and several weedy and crop species have a worldwide distribution, some crucifer clades are restricted to (sub)continents or smaller geographic regions (Lysak and Koch, 2011; Al-Shehbaz, 2012). For instance, tribes of the CES clade (i.e., Cremolobeae, Eudemeae, and Schizopetaleae), as well as Halimolobeae, Physarieae, and all but one Thelypodieae species, are endemic to the New World, while Microlepidieae occur only in Australia and New Zealand. In Africa, the family has a reduced species and generic diversity, with the largest endemic clade confined to southern Africa (South Africa, Lesotho, eSwatini, and Namibia). The tribe Heliophileae includes some 104 species (compilation by I. A. Al-Shehbaz) concentrated chiefly in the winter-rainfall region of the southwestern South Africa, home to two global biodiversity hotspots – Cape Floristic Region and Succulent Karoo. *Heliophila* ranks among the largest crucifer genera, such as *Alyssum*, *Boechera*, *Cardamine*, *Draba*, *Erysimum*, *Lepidium*, and *Physaria* (Al-Shehbaz, 2012). The genus is often regarded as morphologically the most diverse Brassicaceae lineage (Mummenhoff et al., 2005). *Heliophila* varies from small ephemeral annual to perennial herbs (incl. one lianella), subshrubs, and tall shrubs (e.g., *Heliophila brachycarpa*). The species vary particularly in foliage (entire to variously dissected); petal length (1.2–30 mm) and color (white, pink, mauve, purple, blue, or yellow); number and presence vs. absence of petal and stamen appendages; presence vs. absence of paired glands at the bases of pedicels and/or leaves; ovule number (1–80); fruit length (2–120 mm long), shape (linear, lanceolate, oblong, ovate, elliptic, orbicular), constriction (moniliform or not), type (silique, silicle, samara, schizocarp), and flattening (terete, quadrangular, latiseptate, angustiseptate); gynophore length (obsolete to 12 mm long); style length (0.3–20 mm long) and shape (linear, filiform, conical, clavate, ovoid, globose); seed length (0.6–9 mm long), shape, and development of wing; and cotyledonary type (diplecolobal, spirolobal) (Marais, 1970; Mummenhoff et al., 2005; Mandáková et al., 2012; unpublished data).

Despite *Heliophila* species being a frequent and sometimes dominating element of some southern African plant communities, there is limited knowledge of the phylogeographic origin of the genus, interspecies relationships, and genome evolution of *Heliophila* species. Mummenhoff et al. (2005) published a pioneering study, laying foundations for follow-up phylogenomic analyses, demonstrating monophyly of the tribe Heliophileae with South Africa's endemic *Chamira circaeoides* as the sister species to *Heliophila*, finding support for rapid diversification against a background of aridification in the Pliocene/Pleistocene, and showing massive parallel evolution of fruit characters traditionally used in the classification of Heliophileae. Further, ecological optimization analysis allowed preliminary insights into the ecogeographical evolution in Heliophileae.

The last phylogenetic study of c. 57 *Heliophila* species based on internal transcribed spacer (ITS) sequences suggested basal polytomy involving three clades, all sister to *Chamira* (Mandáková et al., 2012). The latter authors showed that two ITS clades are dominated by two chromosome numbers ($2n = 20$ and $2n = 22$), whereas the third clade mainly contained shrubby species with chromosome numbers known only for two species at that time. Chromosome numbers in 27 analyzed *Heliophila* species ranged from $2n = 16$ to $2n =$ c. 88, presumably due to polyploidy and dysploidal chromosomal rearrangements. Interestingly, comparative chromosome painting analyses, revealing the duplicated nature of *Heliophila* genomes, suggested the existence of an allohexaploid ancestor preceding the divergence of *Heliophila* lineages (Mandáková et al., 2012). This was supported by an analysis of synonymous substitution rates (*Ks*) of paralogous and orthologous genes in *Heliophila* cf. *longifolia* (Mandáková et al., 2017).

The high species diversity (>100 species) and extraordinary ecomorphological variability of *Heliophila* impacted by ancient and more recent whole-genome duplication (WGD) events and following post-polyploid diploidization (PPD), confined to one of the most remarkable biodiversity hotspots, make the genus an intricate but attractive phylogenomic model. In this study, based on the previous results and by including a broader spectrum of species, we aim at providing new insight to the WGD–PPD process, test the robustness of the inferred infrageneric relationships (Mummenhoff et al., 2005; Mandáková et al., 2012),

and analyze the evolution of repetitive DNA sequences. Based on the updated ITS phylogeny, we selected 15 *Heliophila* species, representing the major infratribal ITS clades, for low-coverage whole-genome sequencing (lcWGS). Using lcWGS data, we reconstructed a dated whole-plastome phylogeny and characterized the most abundant repetitive sequences (repeatomes) of 15 *Heliophila* species and the sister *C. circaeoides*. We tested whether the ITS-based infrageneric clades are congruent with the plastome phylogeny and phylogenetic relationships inferred from repeat sequence similarities (Vitales et al., 2020). Further, we analyzed the repeat diversity and abundances in relation to the post-polyploid cladogenesis in *Heliophila*. The inclusion of *C. circaeoides*, the only crucifer species with persistent cotyledons, allowed us to get a first insight into its genome.

MATERIALS AND METHODS

Plant Material

The list of all analyzed *Heliophila* and *Chamira* accessions, and outgroup species, is provided in **Supplementary Table 1**. Errors in the determination of species names for accessions used in previous phylogenetic analyses (Mummenhoff et al., 2005; Mandáková et al., 2012) were investigated and revisited where necessary. Selected 15 *Heliophila* species and *C. circaeoides* were used for detailed phylogenetic, repeatome, and cytogenetic analyses (**Supplementary Table 2**).

Genome Size Estimation

Holoploid genome sizes were estimated by flow cytometry in species from which we had seeds and could grow plants in a greenhouse (**Supplementary Table 2**). One sepal (if available) or a fully developed intact leaf was prepared according to Doležel et al. (2007), and isolated nuclei were stained using propidium iodide + RNase IIA (both 50 μg/ml) solution, for 5 min at room temperature, and analyzed using a Partec CyFlow cytometer. A fluorescence intensity of 5,000 particles was recorded. *Solanum pseudocapsicum* (1C = 1.30 pg; Temsch et al., 2010) served as the primary reference standard. One individual of each species measured on three consecutive days was analyzed.

Transcriptome Sequencing and Analyses of Whole-Genome Duplication

Total RNA was extracted from *H. lactea*, *H.* cf. *longifolia*, *H. seselifolia* subsp. *nigellifolia*, and *C. circaeoides* (**Supplementary Table 3**) using RNeasy Plant Mini Kit (Qiagen). Strand-specific library preparation (Illumina Truseq Stranded mRNA) and RNA-Seq (Illumina MiSeq, paired-end reads, 2 × 300 bp) were performed at the Oklahoma Medical Research Foundation (Oklahoma City, United States). Raw reads were corrected with Rcorrector v1.0.4 (Song and Florea, 2015) and trimmed with Trimmomatic v0.36 (Bolger et al., 2014) to remove low-quality reads and potential adapters. *De novo* assembly of transcriptomes was carried out with Trinity v2.5.1 (Haas et al., 2013) with default settings. Assembly summary statistics can be found in **Supplementary Table 3**. We excluded low-quality transcripts detected by Transrate v1.0.3 (Smith-Unna et al., 2016), removed chimeric transcripts, and clustered the remaining transcripts with Corset v1.07 (Davidson and Oshlack, 2014) after mapping RNA-Seq reads with Salmon v0.9.1 (Patro et al., 2017). Coding sequences (CDS) were predicted from the longest sequence of each cluster by TransDecoder v5.0.2 (Haas and Papanicolaou, 2016). Potentially redundant sequences (identity higher than 99%) were further removed with CD-HIT v4.7 (Fu et al., 2012). Gene completeness was then assessed by BUSCO v4.1.2 (Simão et al., 2015). For comparative purposes, we also included publicly available genome of *H.* aff. *coronopifolia* (Kiefer et al., 2019) into the downstream analyses.

To investigate the timing of speciation and potential WGD events in *Chamira* and *Heliophila*, we analyzed synonymous substitutions per synonymous site (*Ks*) for paralogous and orthologous gene pairs identified from within- and between-species comparisons, respectively, using the wgd pipeline (Zwaenepoel and Van de Peer, 2019). We also estimated the heterozygosity of coding genes by detecting SNPs with the GATK v4.0.1.0 pipeline (Poplin et al., 2017). Base Quality Score Recalibration built in GATK was used to detect systematic errors in accuracy of each base call during sequencing. The following filters were applied in GATK when detecting SNPS: QD < 2.0, FS > 60.0, MQ < 40.0, and SOR > 4.0. For each species, RNA-Seq reads were mapped to their respective CDS by Bowtie2-2.3.0 (Langmead and Salzberg, 2012). To allow a more direct comparison, we used OrthoFinder pipeline (Emms and Kelly, 2015) to identify 1-to-1 orthologs shared by *C. circaeoides* and *Heliophila* species. We excluded *H.* aff. *coronopifolia* (Kiefer et al., 2019) from the heterozygosity analysis because there were no RNA-Seq reads available.

Nuclear Gene Phylogeny and Phylogenetic Reconciliation

For phylogenetic analyses, we complemented our CDS dataset (five species mentioned above) with 15 additional species of Brassicales which had public genomic data available (Kiefer et al., 2019; **Supplementary Table 4**), including *Tarenaya hassleriana* (Cleomaceae) as an outgroup. Following Yang and Smith (2014), we inferred sequence homology by all-against-all BLASTn search and filtered the output with a hit fraction of 0.3. We employed MCL v14-137 (Van Dongen and Abreu-Goodger, 2012), with parameters "-tf 'gq(5)' -I 1.4," to obtain putative homologous gene clusters. The clusters with a minimum of 15 taxa were aligned using MAFFT v7.450 (Katoh and Standley, 2013) with the settings –genafpair and –maxiterate 1,000. The alignment columns with more than 90% missing data were removed using the Phyx software (Brown et al., 2017). We built a maximum-likelihood tree using a concatenated alignment of 37 single-copy genes with IQ-TREE v1.6.10 (Nguyen et al., 2014), with 1,000 rapid bootstrap replicates. ModelFinder (Kalyaanamoorthy et al., 2017) was used to identify the best fitted substitution model. We also built gene trees separately and inferred coalescent-based phylogeny with ASTRAL v5.7.3 (Zhang et al., 2018). For homologous gene groups with multiple copies in one species, we

explicitly selected for those with higher copy number (>= 2) in *Chamira* + *Heliophila* species and single copy in the remaining species. To test for the mode of WGD, we converted the gene trees to multilabeled ones and performed phylogenetic reconciliation using GRAMPA v1.3 (Thomas et al., 2017). The ASTRAL topology was used as the input of species tree hypothesis.

Ribosomal Internal Transcribed Spacer Phylogeny

The ITS1 and ITS2 regions were newly sequenced in 102 *Heliophila* accessions, and the obtained sequences (**Supplementary Table 1**; GenBank accession numbers MW216680–MW216783 for ITS1 and MW216784–MW216887 for ITS2) were combined with data published earlier (Mandáková et al., 2012). Methods for DNA extraction, PCR amplification, and ITS sequencing followed Mummenhoff et al. (2004). Multiple alignment of ITS sequences was generated using MAFFT v7.450 and then manually checked and trimmed. Bayesian inference from ITS alignment was performed using MrBayes XSEDE v3.2.7a (Ronquist et al., 2012) at CIPRES Science Gateway (Miller et al., 2010; Towns et al., 2014). Two independent Markov chain Monte Carlo (MCMC) analyses under the GTR+I+G model were run for 200 million generations, chains sampling every 5,000 generations, and burn-in 0.25. Convergence diagnostics for MCMC were conducted by Tracer v1.7.1 (Rambaut et al., 2018).

Low-Coverage Whole-Genome Sequencing

NucleoSpin Plant II kit (Macherey-Nagel) was used to extract the genomic DNA from fresh or silica-dried leaves. DNA sequencing libraries were prepared and sequenced at the sequencing core facility of the Oklahoma Medical Research Foundation (Oklahoma City, United States). The Illumina MiSeq platform, generating 151-bp paired-end reads, was used for sequencing.

Chloroplast Genome Assembly and Divergence Time Estimated Phylogeny

We assembled complete chloroplast (cp) genomes for 15 *Heliophila* species, *C. circaeoides*, and *Subularia aquatica* using NOVOPlasty v3.2 (Dierckxsens et al., 2016), using the *ndhF* gene of *Arabidopsis thaliana* (GenBank: NC_000932.1) as the seed (**Supplementary Table 5**). The cp genomes were annotated by plann v1.1.2 (Huang and Cronk, 2015) with *A. thaliana* as the reference genome, which was followed by manual curation using the Sequin software[1].

We retrieved cp genomes of additional 42 Brassicaceae species from GenBank, representing all major Brassicaceae lineages, to investigate the maternal phylogeny of *Heliophila* within the whole family. A total of 103 genic and 102 intergenic regions were extracted from multiple-sequence alignment generated by MAFFT v7.450 with the L-INS-i mode. Gblocks v0.91b (Talavera and Castresana, 2007) was used to remove poorly aligned regions with a minimum block length of 2 bp. We subsequently concatenated the alignments and selected the best partitioning scheme with PartitionFinder v2.1.1 (Lanfear et al., 2016). A maximum-likelihood (ML) tree was reconstructed using IQ-TREE v1.6.10 with three *Aethionema* species as outgroup. Following Guo et al. (2017), we performed molecular dating with MCMCTREE (Rannala and Yang, 2007). The root age was set to 31.2 million years ago (Mya) according to Hohmann et al. (2015). The burn-in period was set to 2,000,000 cycles, and the MCMC run was sampled every 800 cycles until a total of 10,000 samples were collected. Diagnostics for MCMC were performed by Tracer v1.7.1.

Preprocessing and Cluster Analysis of Repetitive DNA From Next-Generation Sequencing Data

Quality checks were performed with FastQC v0.11.7 (Andrews, 2010). Illumina adapter removal, quality filtering with 90% of bases equal to or above the quality cutoff value of 20, and trimming procedures were performed with Trimmomatic v0.36. All reads were quality filtered and trimmed down to 140 bp. Using Bowtie2-2.3.0 aligner software, organelle DNA that originated from chloroplast and mitochondria were filtered out prior to the analysis. Characterization and analysis of repetitive DNA were conducted using the graph-based clustering pipeline RepeatExplorer2 (Novák et al., 2013) as described by Novák et al. (2010, 2017).

Clustering was performed using 90% similarity over 55% of the read length as default settings. This analysis resulted in the construction of clusters that represent different repetitive DNA families. All sequences that built the clusters were in the form of contigs. Clusters with a genome proportion higher than 0.01% were annotated in detail. The maximum number of reads was used to perform detailed annotations in individual species to identify all repetitive sequences. Comparative clustering analysis was performed with concatenated next-generation sequencing (NGS) reads of 15 *Heliophila* and the sister species *C. circaeoides* (**Supplementary Table 2**). To avoid the coverage bias in the comparative repeatome analysis, preprocessed paired-end reads were randomly sampled in order to represent 10% of a genome (i.e., coverage = 0.1×) based on (1C) genome sizes (Novák et al., 2010). The same RepeatExplorer2 settings were used in the comparative analysis with individual clustering analysis.

Repetitive DNA cluster annotations were done by RepeatExplorer2 pipeline using DNA and protein similarity searches on clusters with known protein domains. Clusters which could not be classified by the pipeline were manually annotated using BLAST (Altschul et al., 1990) searches against the GenBank sequence and Censor (Kohany et al., 2006) databases. Clusters which were annotated as tandem repeats (directly or manually from the shape of the cluster graph) were further tested with Tandem Repeat Finder software (Benson, 1998) and similarity dot-plots with Dotter (Sonnhammer and Durbin, 1995). Tandem Repeat Analyzer (TAREAN, Novák et al., 2017) which is implemented in RepeatExplorer2 pipeline was used to reconstruct consensus monomers of the tandem repeats. All annotations were revised and corrected if necessary. Subsequently, all identified tandem repeats from all species were

[1]https://www.ncbi.nlm.nih.gov/Sequin

compared with each other using BLASTn searches to detect shared tandem repeats.

Phylogeny Based on Repeatome Similarity

The novel phylogeny inference method using repeatome similarities as a source of phylogenetic marker was performed as introduced by Vitales et al. (2020). This method is based on the pairwise genetic distances between repeatomes of closely related species. By calculating the observed/expected number of edges (of similarity) between all species for each cluster from the output of RepeatExplorer2 comparative analysis, a similarity matrix is generated and transformed into distance matrices by calculating the inverse of the values (Vitales et al., 2020). Three datasets were created: 15 *Heliophila* species from all clades (A–D); 9 species from clades A, B, and D; and 9 species from clades B, C, and D. Subsequently, neighbor-joining trees were constructed for the clusters which included repeats that were present in all species out of the first 100 clusters using R (R Core Team, 2013) and *ape* package (Paradis and Schliep, 2019). The trees were then used to construct a consensus tree for each dataset with the SplitsTree4 v4.14.6 software (Huson and Bryant, 2006). Lastly, consensus tree including all *Heliophila* species was transformed to a dendrogram for a better representation.

Further, using RepeatExplorer2 comparative analysis, read abundance matrix, hierarchical cluster analysis was performed using *pheatmap* package (Kolde and Kolde, 2015) in R. The abundance matrix was transformed into a distance matrix by *pheatmap*. The clusters with genome proportion higher than 0.01% were used to construct the dendrogram relationship of 15 *Heliophila* and *C. circaeoides* species. *pheatmap* package in R was used to construct the heatmap.

Chromosome Preparations

Young inflorescences were collected from plants in the field. Inflorescences were fixed in freshly prepared fixative (ethanol:acetic acid, 3:1) overnight, transferred into 70% ethanol, and stored at −20°C until used. Chromosome spreads from young fixed flower buds, containing immature anthers, were prepared according to the published protocol (Mandáková and Lysak, 2016a).

DNA Probes

The list of all the designed probes and primers specific to repetitive elements is provided in **Supplementary Table 6**. Synthetic oligonucleotide probes were used for tandem repeats with shorter monomers (<500 bp). Target sequences (60 nt) with GC content 30–50% were selected from DNA alignments using Geneious v11.1.5 software package[2] to minimize self-annealing and formation of hairpin structures. DNA probe preparation and labeling followed the published protocol (Mandáková and Lysak, 2016b). For satellites with longer monomers, PCR primers were designed to face outward from the monomer; therefore, PCR amplification was performed only between monomers tandemly arrayed. For retrotransposons, PCR primers were designed to the GAG domain which is generally the most variable domain among different retrotransposon families. PCR products were purified using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and labeled by nick translation.

[2]https://www.geneious.com

Fluorescence *in situ* Hybridization, Microscopy, and Image Processing

Twenty microliters of the hybridization mix containing 100 ng of the labeled probe dissolved in 50% formamide and 10% dextran sulfate in 2× sodium saline citrate (SSC; 20× SSC: 3M sodium chloride, 300 mM trisodium citrate, pH 7.0) was pipetted on a suitable chromosome-containing slide and immediately denatured on a hot plate at 80°C for 2 min. In some experiments, two differentially labeled probes (100 ng of each) were pooled. Hybridization was carried out in a moist chamber at 37°C for 24 h. Post-hybridization washing was performed in 20% formamide in 2× SSC at 42°C. The immunodetection of hapten-labeled probes was performed as described by Mandáková and Lysak (2016b) as follows: biotin-dUTP was detected by avidin–Texas Red (Vector Laboratories) and amplified by goat anti-avidin–biotin (Vector Laboratories) and avidin–Texas Red, and digoxigenin-dUTP was detected by mouse anti-digoxigenin (Jackson ImmunoResearch) and goat anti-mouse–Alexa Fluor 488 (Invitrogen). Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI, 2 μg/ml) in Vectashield. The preparations were photographed using a Zeiss Axio Imager 2 epifluorescence microscope with a CoolCube camera (MetaSystems). Images were acquired separately for two or three fluorochromes using appropriate excitation and emission filters (AHF Analysentechnik). Individual monochromatic images were pseudocolored and merged and cropped using Adobe Photoshop CS.

RESULTS

Chamira and *Heliophila* Have Most Likely Undergone a Shared WGD in Oligocene

Transcriptomes were assembled for *C. circaeoides* and three *Heliophila* species (*H. lactea*, *H.* cf. *longifolia*, and *H. seselifolia*), from which we predicted 16,671 to 30,264 protein-CDS. Compared to the publicly available genome of *H.* aff. *coronopifolia* (Kiefer et al., 2019), which showed 53.2% gene completeness, these transcriptomes had more than 70% of the 1,440 conserved BUSCO genes complete (**Supplementary Figure 1**). In addition, more than 10% genes were still identified as duplicated ones in all transcriptome-derived CDS after removing potential isoforms (**Supplementary Figure 1**).

From within-species comparisons of CDS, we identified 1,711 to 3,838 paralogous gene pairs and calculated their rates of synonymous site changes per synonymous site (*Ks*; **Supplementary Table 7**). The distribution of *Ks* showed a clear peak between 0.43 and 0.48 in all *Heliophila* species (**Figure 1A**), which can indicate a lineage-specific mesopolyploidy event as proposed by Mandáková et al. (2012, 2017). Interestingly, a

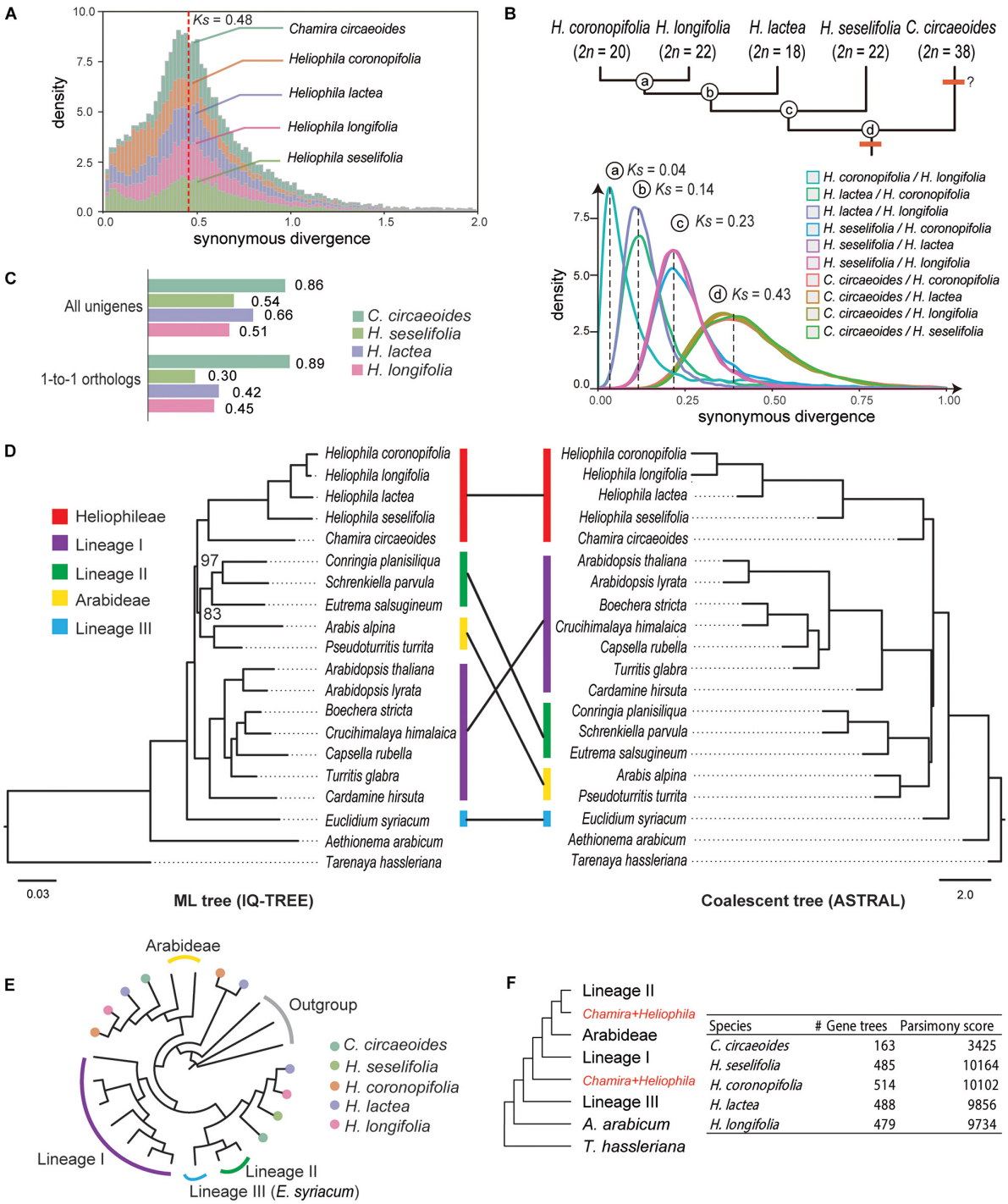

FIGURE 1 | Analyses of synonymous substitutions per synonymous site (*Ks*) and phylogenetic tree topologies. **(A)** *Ks* plots of within-species comparisons in *Chamira* and *Heliophila* species. The red dashed line indicates the position of peaks for the putatively shared WGD event. **(B)** *Ks* plots of between-species comparisons in *Chamira* and *Heliophila* species. Peaks of species divergence are indicated with dashed lines and labeled with circled letters. The upper tree topology shows the branching order of *Chamira* and *Heliophila* species as revealed by *Ks* analyses, with chromosome numbers labeled for each species. **(C)** Comparison of SNP rates in coding sequences of *C. circaeoides*, *H. lactea*, *H.* cf. *longifolia*, and *H. seselifolia*. **(D)** Phylogenetic relationship between *Chamira* + *Heliophila* and other Brassicaceae species using maximum-likelihood (ML) and coalescent approaches. Colored bars indicate different lineages. Unless otherwise mentioned, all relationships received full support in both ML and coalescent trees. Bootstrap values are labeled for two branches that failed to be fully supported in ML analysis. Scale bars below ML and coalescent trees indicate substitution per site and coalescent unit, respectively. **(E)** An example of gene tree topology showing phylogenetic relationship among multicopy genes of *Chamira* and *Heliophila* species. **(F)** The multilabeled tree with the lowest parsimony score, as inferred by gene tree reconciliation analyses in all species. The ASTRAL topology was used as the input of species tree hypothesis.

Ks peak at the same location was observed in *C. circaeoides* (**Figure 1A**). To assess whether the WGD event(s) occurred before or after the divergence of *Chamira* and *Heliophila*, we retrieved 7,681 to 16,901 orthologous gene pairs and compared *Ks* peaks from between-species comparisons. We found a *Ks* peak at 0.43 in all comparisons between *Heliophila* species and *C. circaeoides*, which represented the oldest divergence in our comparisons (**Figure 1B** and **Supplementary Table 8**). Thus, the WGD event(s) likely occurred before the *Chamira*/*Heliophila* split and might be shared by the two genera. Considering a mutation rate of 8.22×10^{-9} substitutions/synonymous site per year (Kagale et al., 2014), the time of WGD or subgenome divergence was estimated between 26.15 and 29.20 Mya, and the *Chamira*/*Heliophila* split around 26.16 Mya.

In addition to the WGD peak, we detected a minor *Ks* peak between 0 and 0.1 in all analyzed species (**Figure 1A**). By mapping RNA-Seq reads to the assembled transcriptomes, we observed that the heterozygosity in *C. circaeoides* was two times higher than in *Heliophila* species (**Figure 1C** and **Supplementary Table 9**). This, along with relatively high chromosome number in *C. circaeoides* ($2n = 38$), may suggest that the minor *Ks* peak in this species represents an additional WGD post-dating the *Chamira*–*Heliophila* divergence (**Figure 1B**).

Transcriptome Phylogeny Corroborates the Sistership of *Chamira* and *Heliophila* and Suggests Their Allopolyploid Origin

After including 15 available genomes from major Brassicaceae lineages as well as the outgroup *T. hassleriana* (Cleomaceae), we retrieved 37 strictly single-copy genes that are shared by all species. Our phylogenetic analyses corroborated the sistership of *Chamira* and *Heliophila* (**Figure 1D**). However, maximum-likelihood (IQ-TREE) and coalescent-based (ASTRAL) methods recovered different topologies regarding the placement of *Chamira* + *Heliophila*. Whereas the ML tree suggested that this clade was sister to lineage II + Arabideae, coalescent analysis showed that it had a more ancestral position, being outside of lineage I + lineage II + Arabideae (**Figure 1D**). We also retrieved 130 homologous gene groups that consisted of mostly single-copy genes in diploid species and multicopy genes in *C. circaeoides* and *Heliophila* species. We observed that *C. circaeoides* and *Heliophila* genes frequently formed sister clades that were sister to different Brassicaceae lineages (see **Figure 1E** for an example), which suggested that the mesopolyploidy event(s) involved distant hybridization(s). Despite the number of multicopy genes varying across species, gene tree reconciliation analyses focusing on individual species recovered the same source of potential parental genomes for both *Chamira* and *Heliophila* (**Figure 1F**).

The Updated ITS Phylogeny Revealed Four Major Clades in *Heliophila*

A Bayesian 50% majority-rule consensus ITS tree (**Supplementary Figure 2**) was inferred from sequences of 198 *Heliophila* accessions and five outgroup species (**Supplementary Table 1**). Four major ITS clades were identified in *Heliophila*. The largest clade A contained 88 accessions, clade B 32 accessions, clade C 73 accessions, and clade D grouped only five accessions. Clade D was newly identified as compared to the previous analyses based on a less extensive taxon sampling (Mummenhoff et al., 2005; Mandáková et al., 2012). All major clades were well supported (posterior probability, $pp \geq 0.98$) except for clade A ($pp = 0.69$). Among the four major clades, clade D was sister to the other three clades; clades A and C showed a sister relationship. *C. circaeoides* was confirmed as the sister genus of *Heliophila*/Heliophileae.

Dated Plastome Phylogeny Suggested a Middle Miocene Origin of *Heliophila*

Using the basally resolved ITS tree, 15 *Heliophila* species, proportionally representing the four clades, and *C. circaeoides* were selected for lcWGS (**Figure 2** and **Supplementary Table 2**). The next-generation sequence data was used to construct

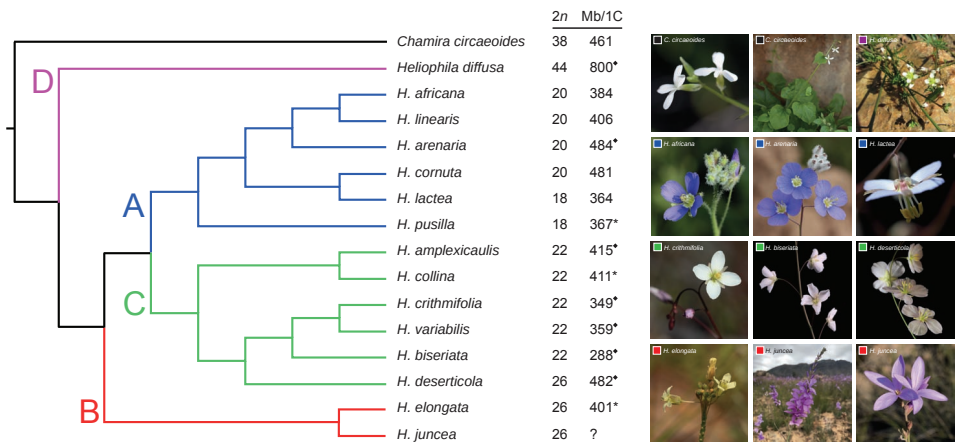

FIGURE 2 | Schematic relationships among the analyzed 15 *Heliophila* species and *C. circaeoides* based on the ITS phylogeny. Capital letters refer to four *Heliophila* clades. Diamond symbols indicate average *C*-values based on estimates for two or more populations; asterisks indicate *C*-values obtained by analysis of partially degraded leaf material. The original ITS phylogeny is presented as **Supplementary Figure 2**. All photos by TM.

whole-plastome phylogeny and analyze nuclear repeatomes of the 16 genomes.

We assembled complete cp genomes for all 16 sequenced species, ranging in length from 152,794 (*C. circaeoides*) to 154,300 bp (*H. pusilla* var. *pusilla*) (**Supplementary Table 10**). All the cp genomes showed a typical quadripartite structure in which a large single-copy (LSC) region (82,736–83,766 bp) and a short single-copy (SSC) region (17,429–17,958 bp) are separated by two inverted repeat (IR) copies (26,235–26,413 bp). All analyzed genomes encoded 131 genes, including 86 protein-coding genes, 37 tRNA genes, and eight rRNA genes (**Supplementary Table 10**). The GC content of the assembled cp genomes ranged between 36.1 and 36.7%.

After excluding unalignable or ambiguous regions and sites, a supermatrix with 100,707 nucleotide sites was generated, of which 15,637 (15.5%) were parsimony informative. *Heliophila* species were retrieved as a monophyletic clade sister to *C. circaeoides*. The maternal phylogeny was largely congruent with the above-described ITS phylogeny, except for *H. diffusa* var. *diffusa* (clade D) clustering with clade B species (**Supplementary Figure 3**). Based on the plastome phylogeny, we estimated that the split between (Heliophileae + *Chamira*) species and their closest relative, *S. aquatica*, occurred (15.95) 20.26 (24.64) Mya, at the Oligocene–Miocene boundary. The divergence between *Chamira* and *Heliophila* was dated to (13.77) 18.53 (23.33) Mya, followed by the diversification of the four *Heliophila* clades c. 16 to 8 Mya.

TABLE 1 | Characteristics of NGS data used for repeatome analysis.

Species	Clade	Total repeats (%)	No. of reads	Genome coverage	No. of clusters
Heliophila africana	A	30.07	511,217	0.19×	232
H. arenaria subsp. *arenaria*	A	30.18	901,412	0.26×	183
H. cornuta var. *cornuta*	A	31.38	652,717	0.19×	203
H. lactea	A	33.93	744,973	0.29×	177
H. linearis var. *linearis*	A	37.03	1,002,030	0.35×	209
H. pusilla var. *pusilla*	A	37.34	576,080	0.22×	164
H. elongata	B	30.73	662,141	0.23×	181
H. juncea	B	25.04	886,973	0.31×	173
H. amplexicaulis	C	37.21	961,390	0.32×	165
H. collina	C	38.90	774,496	0.31×	130
H. crithmifolia	C	34.06	853,487	0.29×	224
H. deserticola var. *micrantha*	C	38.64	1,304,122	0.38×	281
H. biseriata	C	32.96	962,154	0.47×	276
H. variabilis	C	33.55	1,056,742	0.41×	148
H. diffusa var. *diffusa*	D	38.67	911,217	0.16×	141
Chamira circaeoides	–	43.66	690,569	0.21×	163

Repeatome Analysis

The RepeatExplorer2 pipeline was used to analyze and compare the repeatomes of 15 *Heliophila* species and *C. circaeoides*. Maximum number of reads was used for the detailed repeatome analysis with the genome coverage from 0.16× to 0.47× (**Table 1**). The total repeat content of the analyzed species ranged from 25.04% to 43.66%, whereas single- or low-copy sequences made up the remainder of the genome sequences (**Figure 3**). In all *Heliophila* genomes, the predominant repeat type was long terminal repeat (LTR) retrotransposons, ranging from 11.51% (*H. juncea*) to 22.42% (*H. elongata*) (**Table 2** and **Figure 3**). The most abundant repeat type of the *C. circaeoides* genome was tandem repeats (17.92%), whereas among the 15 *Heliophila* genomes, tandem repeat abundances varied from 1.03% (*H. elongata*) to 12.10% (*H. diffusa*) (**Table 2** and **Figure 3**). In all the analyzed genomes, DNA transposon abundances were lower compared with LTR retrotransposons, ranging from 1.54% (*H. cornuta* var. *cornuta*) to 4.31% (*H. linearis* var. *linearis*) (**Table 2** and **Figure 3**).

We tested possible correlations between the estimated abundances of identified repeat families (**Table 2**) and genome size (Mb/1C) of the analyzed *Heliophila* species. The total repeat content was positively correlated with genome size (p value = 0.0006, R^2 = 0.6446). A weak but significant positive correlation was found between tandem repeat content and genome size (p value = 0.0397, R^2 = 0.3071).

Transposable Elements

Ty3/gypsy was the most abundant superfamily of LTR retrotransposons in *Heliophila* species, ranging from 7.84% (*H. juncea*) to 16.91% (*H. collina*), while *Ty1/copia* retrotransposons were less prominent, ranging from 1.94% (*H. pusilla*) to 5.92% (*H. elongata*). In the *C. circaeoides* genome, LTR retrotransposons represent 16.69% of the genome (9.82% of *Ty3/gypsy* and 6.87% of *Ty1/copia* elements) (**Figure 3**).

Analyzing the *Ty3/gypsy* superfamily, Chromovirus-type elements were represented by *CRM*, *Tekay*, *Galadriel*, and *Reina* lineages (ordered by their abundances), whereas non-Chromovirus-type elements were represented by *Athila* and *Ogre/Tat* lineages (**Table 2**). In all analyzed genomes, *Athila* was the predominant lineage. The abundance of *Athila* elements ranged from 5.87% in *H. juncea* to 13.99% in *H. pusilla*. From Chromovirus lineage elements, *CRM* was found to be the most abundant, ranging from 0.45% in *H. arenaria* subsp. *arenaria* to 5.31% in *H. deserticola* var. *micrantha*. In *C. circaeoides*, *Athila* lineage was the most abundant *Ty3/gypsy* element (6.57%), followed by *CRM* (2.27%).

Ty1/copia superfamily consisted of seven lineages: *Bianca*, *Ale*, *Tork*, *TAR*, *Ivana*, *Angela*, and *SIRE* (ordered by their abundances) (**Table 2**). *Bianca* was identified as the most abundant lineage among the *Heliophila* species, ranging from 0.67% in *H. lactea* to 2.35% in *H. elongata*. In *C. circaeoides*, *Ale* lineage was the most abundant *Ty1/copia* element (3.31%), followed by *Bianca* (1.23%). The amplification of the *Ale* elements differentiated the *C. circaeoides* genome from those of *Heliophila* species. The diversity and abundances of the identified LTR

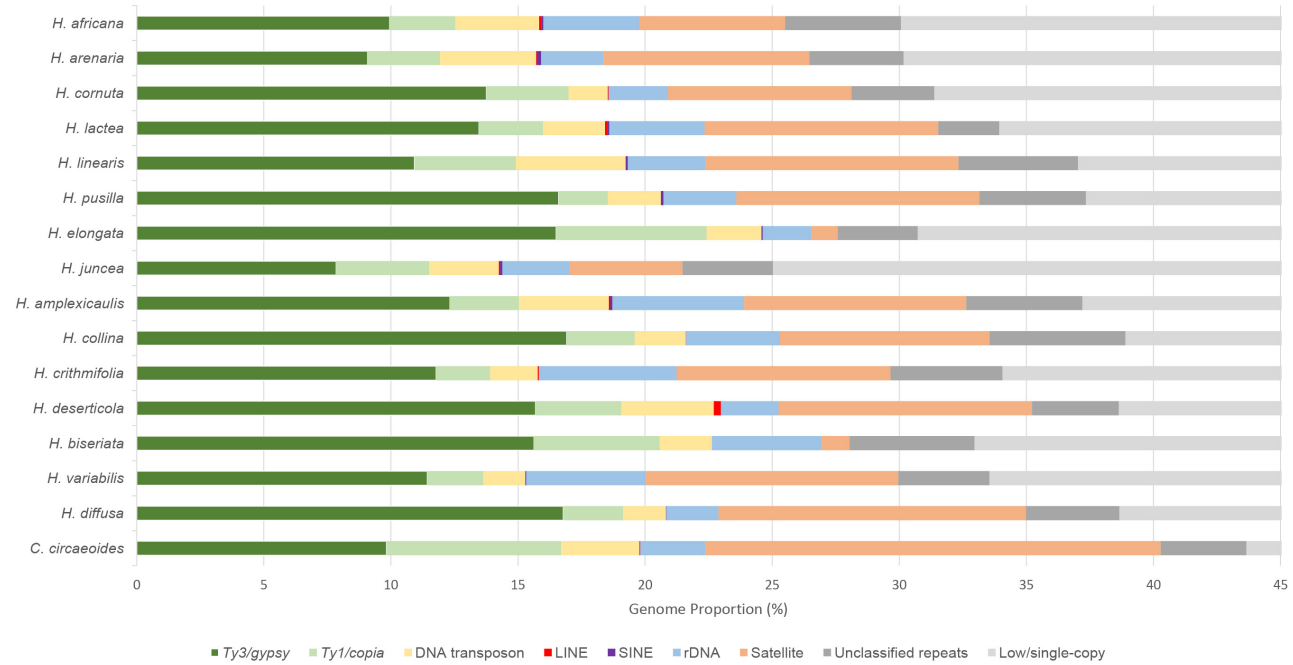

FIGURE 3 | Relative repeat abundances and low/single-copy sequences identified in sequenced genomes of 15 *Heliophila* species and *C. circaeoides*. Low/single-copy sequences above 45% were discarded.

retrotransposons have not followed the infrageneric groupings in *Heliophila*, and LTR retroelement abundance (in Mb) has not been found correlated with genome size (p value = 0.0569, R^2 = 0.2700).

Non-LTR retrotransposons, *LINE* and *SINE* elements, were found at very low abundances or not detected in the 16 analyzed genomes; the highest abundances were encountered in *H. deserticola* (0.28%) for *LINE* and *H. arenaria* (0.17%) for *SINE* (**Table 2**).

In *Heliophila* species, DNA transposons were represented by *Mutator*, *CACTA*, *hAT*, *Helitron*, *Harbinger*, and *Mariner* lineages (**Table 2**). *Mutator* (0.42% in *H. crithmifolia* to 1.12% in *H. juncea*) and *CACTA* (0.17% in *H. crithmifolia* to 1.61% in *H. deserticola*) were the more abundant elements. In *C. circaeoides*, *CACTA* lineage was the most abundant DNA transposon (1.85%), followed by *hAT* (0.50%). The diversity and abundance of DNA transposons did not correspond to the infrageneric *Heliophila* clades, but the amounts of identified DNA transposons (in Mb) were found to be weakly correlated with genome size (p value = 0.0184, R^2 = 0.3823).

Tandem Repeats

In total, 124 tandem repeats were identified in the analyzed *Heliophila* and *Chamira* genomes. The identified tandem repeats varied in monomer lengths (e.g., 27-bp HeJun6 in *H. juncea* and 4,034-bp ChCir9 in *C. circaeoides*), numbers (four in *H. pusilla* and *H. elongata* up to 16 in *C. circaeoides*), and abundances (from 1.03% in *H. elongata* to 17.92% in *C. circaeoides*) (**Supplementary Table 11**). The high tandem repeat content in *H. diffusa* (12.10%) differentiates this genome from genomes of the other three *Heliophila* clades. Tandem repeats of species in clades A, B, and C ranged from 4.44% to 9.97%, except for ∼1% in *H. elongata* (clade B) and *H. biseriata* (clade C) (**Figure 3**).

No apparent correspondence between the diversity of tandem repeats and their genomic proportion was observed. For example, 10 tandem repeats identified in *H. biseriata* represented only 1.11% of its genome, whereas only four tandem repeats built up 9.56% of the *H. pusilla* genome (**Supplementary Table 11**). In all *Heliophila* species, one or two tandem repeats were dominating their tandem repeatomes [e.g., HeAfr1: 4.72% (out of 5.75%), HeAmp1: 5.9% (8.74%), HeJun1: 2.1% and HeJun2: 1.73% (4.44%)].

The genome of *C. circaeoides* exhibited the highest number of identified tandem repeats among all the sequenced species. The monomer length of the 16 tandem repeats varied from 180 to 4,034 bp, whereby seven and four repeats were longer than 1,000 and 3,000 bp, respectively (**Supplementary Table 11**). The abundances of these repeats ranged from 0.074% to 0.59%. Seven tandem repeats with monomers longer than 1,000 bp were also identified in four *Heliophila* genomes from sister clades A and C species (*H. africana*, *H. biseriata*, *H. crithmifolia*, and *H. linearis*) at very low abundances (<0.1%, except for HeBis2: 0.24%).

Shared Tandem Repeats

Our analyses have not identified any homologous tandem repeats between *C. circaeoides* and *Heliophila* species. In *Heliophila* genomes, among the 108 tandem repeats identified, 16 repeats were found to be shared among two or more species (**Figure 4** and **Supplementary Table 11**). Monomer

TABLE 2 | Detailed classification of repetitive elements and their genome proportions (%).

		Clade															
		A						B		C						D	
Repeat family		**HeAfr**	**HeAre**	**HeCor**	**HeLac**	**HeLin**	**HePus**	**HeElo**	**HeJun**	**HeAmp**	**HeCol**	**HeCri**	**HeDes**	**HeBis**	**HeVar**	**HeDif**	**ChCir**
LTR retrotransposons		12.54	11.93	17.00	15.99	14.93	18.54	22.42	11.51	15.03	19.59	13.90	19.07	20.58	13.63	19.14	16.69
Ty3/gypsy	*Athila*	7.62	8.21	8.23	10.81	6.95	13.99	12.79	5.87	9.53	11.56	7.67	6.38	9.36	8.13	12.77	6.57
	CRM	0.75	0.45	1.56	1.43	2.15	1.64	1.85	1.63	2.26	4.51	0.88	5.31	1.76	1.12	2.94	2.27
	Tekay	0.84	0.28	3.60	0.75	1.78	0.67	0.54	0.07	0.39	0.78	3.07	3.72	3.58	2.13	0.59	0.00
	Galadriel	0.06	0.00	0.16	0.26	0.00	0.00	0.10	0.01	0.04	0.00	0.07	0.12	0.31	0.00	0.00	0.48
	Reina	0.02	0.02	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.06	0.02	0.24	0.00	0.00	0.29
	Ogre/Tat	0.00	0.00	0.04	0.00	0.00	0.17	0.32	0.17	0.00	0.00	0.00	0.09	0.25	0.00	0.45	0.00
	Unclassified	0.66	0.14	0.15	0.23	0.03	0.12	0.90	0.10	0.10	0.06	0.03	0.06	0.12	0.05	0.03	0.21
	Total	9.94	9.09	13.76	13.47	10.92	16.60	16.50	7.84	12.32	16.91	11.77	15.69	15.63	11.43	16.78	9.82
Ty1/copia	*Ale*	0.21	0.20	0.14	0.11	0.35	0.21	1.43	0.81	0.10	0.22	0.44	0.54	1.55	0.17	0.30	3.31
	Bianca	1.14	1.26	1.86	0.67	1.29	1.28	2.35	0.76	0.96	1.20	1.26	1.06	1.28	1.17	0.69	1.23
	Angela	0.27	0.34	0.04	0.18	0.39	0.13	0.25	0.18	0.13	0.33	0.03	0.35	0.34	0.11	0.03	0.07
	Ivana	0.29	0.36	0.20	0.84	0.55	0.09	0.85	0.21	0.64	0.31	0.09	0.31	0.10	0.24	0.86	0.66
	TAR	0.40	0.45	0.40	0.33	0.79	0.10	0.29	0.42	0.41	0.51	0.12	0.43	0.53	0.22	0.01	1.14
	Tork	0.23	0.15	0.56	0.31	0.59	0.10	0.67	1.15	0.40	0.05	0.16	0.63	0.93	0.18	0.16	0.40
	SIRE	0.05	0.04	0.03	0.08	0.03	0.01	0.04	0.10	0.02	0.03	0.01	0.05	0.13	0.09	0.24	0.04
	Unclassified	0.01	0.03	0.01	0.01	0.02	0.03	0.03	0.04	0.05	0.04	0.01	0.01	0.08	0.01	0.06	0.01
	Total	2.60	2.84	3.24	2.52	4.00	1.94	5.92	3.67	2.71	2.68	2.12	3.37	4.95	2.20	2.36	6.87
DNA transposons	*Harbinger*	0.53	0.58	0.04	0.17	0.38	0.11	0.08	0.04	0.23	0.07	0.05	0.11	0.21	0.03	0.06	0.19
	Helitron	0.38	0.36	0.04	0.12	0.11	0.10	0.24	0.17	0.42	0.28	0.24	0.20	0.11	0.05	0.09	0.00
	CACTA	0.67	0.56	0.57	0.49	1.05	0.35	0.53	0.48	0.96	0.60	0.17	1.61	0.48	0.38	0.29	1.85
	Mariner	0.14	0.15	0.00	0.18	0.15	0.09	0.00	0.01	0.07	0.05	0.07	0.07	0.05	0.03	0.18	0.00
	Mutator	0.45	0.82	0.50	0.59	0.92	0.91	0.86	1.12	0.88	0.51	0.42	0.86	0.87	0.60	0.72	0.43
	hAT	0.81	0.61	0.08	0.43	0.89	0.24	0.23	0.45	0.82	0.11	0.24	0.59	0.17	0.23	0.11	0.50
	Unclassified	0.33	0.72	0.31	0.45	0.81	0.28	0.22	0.46	0.17	0.40	0.69	0.20	0.17	0.34	0.22	0.12
	Total	3.29	3.79	1.54	2.44	4.31	2.08	2.16	2.74	3.54	2.02	1.88	3.64	2.06	1.65	1.69	3.09
LINE		0.08	0.02	0.03	0.07	0.01	0.00	0.00	0.05	0.05	0.01	0.04	0.28	0.00	0.02	0.00	0.03
SINE		0.08	0.17	0.00	0.09	0.07	0.10	0.06	0.09	0.09	0.00	0.00	0.00	0.00	0.02	0.01	0.00
rDNA		3.76	2.45	2.34	3.75	3.05	2.87	1.91	2.65	5.18	3.69	5.42	2.27	4.31	4.71	2.05	2.56
Tandem repeats		5.75	8.12	7.21	9.19	9.96	9.56	1.03	4.44	8.74	8.25	8.42	9.97	1.11	9.93	12.10	17.92
Unclassified repeats		4.56	3.70	3.26	2.40	4.70	4.18	3.16	3.56	4.57	5.34	4.40	3.42	4.91	3.58	3.67	3.37
Low/single-copy sequences		69.93	69.82	68.62	66.07	62.97	62.66	69.27	74.96	62.79	61.10	65.94	61.36	67.04	66.45	61.33	56.34
All repeats total		30.07	30.18	31.38	33.93	37.03	37.34	30.73	25.04	37.21	38.90	34.06	38.64	32.96	33.55	38.67	43.66

HeAfr, Heliophila africana; HeAre, Heliophila arenaria; HeCor, Heliophila cornuta; HeLac, Heliophila lactea; HeLin, Heliophila linearis; HePus, Heliophila pusilla; HeElo, Heliophila elongata; HeJun, Heliophila juncea; HeAmp, Heliophila amplexicaulis; HeCol, Heliophila collina; HeCri, Heliophila crithmifolia; HeDes, Heliophila deserticola; HeBis, Heliophila biseriata; HeVar, Heliophila variabilis; HeDif, Heliophila diffusa; ChCir, Chamira circaeoides.

lengths of the shared tandem repeats varied between 158 and 184 bp, and overall pairwise sequence homologies ranged from 82.5% to 100% (**Supplementary Tables 11,12**). Dot-plot comparison of consensus monomer sequences of shared tandem repeats is shown in **Supplementary Figure 4**, and multiple and pairwise alignments of the 16 shared repeats are presented in **Supplementary Figure 5**.

In clade A species, three tandem repeats were shared among all the six genomes, whereas one repeat was shared only by three species of the *H. africana* subclade (HeAfr2, HeAre3, and HeLin5). Whereas five clade A species have unique tandem repeats, all four tandem repeats detected in *H. pusilla* were shared either among all clade A species (HePus2, HePus3, and HePus4) or only with *H. arenaria* (HeAre2) and *H. linearis* (HeLin2). Interestingly, the 168-bp HeCor4 repeat in *H. cornuta* was found to be homologous to the HeJun1 in *H. juncea* from clade B (**Figure 4** and **Supplementary Tables 11,12**).

In clade B, *H. elongata* and *H. juncea* shared one tandem repeat (HeElo1 and HeJun8) which was also shared with two clade C species – *H. crithmifolia* (HeCri1) and *H. variabilis* (HeVar3). The 184-bp HeElo3 identified in the *H. elongata* genome was also detected in *H. crithmifolia* (HeCri2) and *H. variabilis* (HeVar1) (**Figure 4** and **Supplementary Tables 11,12**). Among the six clade C genomes analyzed, two genomes (*H. biseriata* and *H. collina*) possessed only species-specific repeats, while 10 repeats were shared by at least two of the four remaining

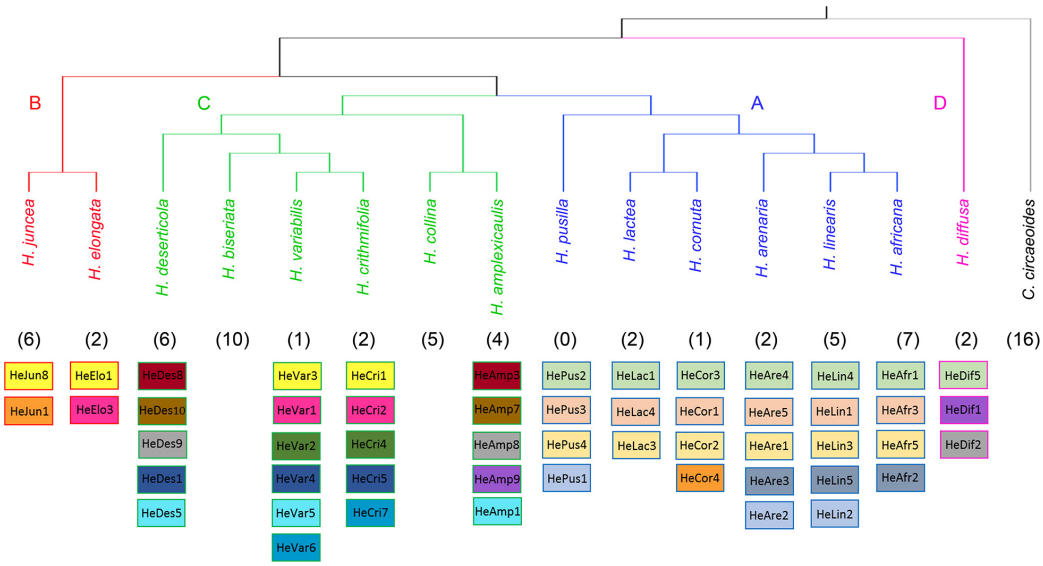

FIGURE 4 | Overview of shared tandem repeats among the 15 *Heliophila* species analyzed. The ITS tree (**Supplementary Figure 2**) was used to display species relationships. Shared tandem repeats are color-coded; numbers of species-specific tandem repeats are indicated in parenthesis.

species. Sister species *H. crithmifolia* and *H. variabilis* shared five different repeats, whereby two were also shared by *H. deserticola* (HeDes1, HeDes5) and the other two were identified in two clade B species (see above). Three other repeats were shared between *H. amplexicaulis* (HeAmp3, HeAmp7, and HeAmp8) and *H. deserticola* (HeDes8, HeDes9, and HeDes10) species without a sister relationship (**Figure 4** and **Supplementary Tables 11,12**). *H. diffusa* shared one repeat (HeDif5) with all clade A genomes and two repeats (HeDif1, HeDif2) with clade C species *H. amplexicaulis* (HeAmp8, HeAmp9) and *H. deserticola* (HeDes9) (**Figure 4** and **Supplementary Tables 11,12**).

In summary, the identified tandem repeats shared among *Heliophila* species, but not with *Chamira*, corroborates the monophyletic origin of the former genus. The three repeats shared among all clade A genomes may reflect younger age of speciation events in this group (**Supplementary Figure 3**). Tandem repeatomes in clade C genomes show high evolutionary dynamism, manifested by (i) high diversity of shared satellites, (ii) some repeats being shared with more ancestral clades B and D, and (iii) accelerated evolution or elimination of shared repeats in two species (*H. biseriata* and *H. collina*).

Phylogenetic Analysis of the Identified Repeats

Consensus tree phylogeny was reconstructed using the ape package (Paradis and Schliep, 2019) in R based on pairwise genetic distances between all repeats in the 100 most abundant clusters retrieved from RepeatExplorer2 comparative clustering analysis. In the dataset which consisted of 15 *Heliophila* species, 25 clusters included sequence overlaps (similarities) between reads from all species to generate sequence similarity matrices. In 75 clusters, sequence reads shared by all the analyzed species were lacking, indicating that those repeats are either species- or clade-specific. The consensus tree, reconstructed from 25 clusters with complete similarity matrices, separated repeatomes of clade A and C species, whereas clade B and D genomes formed a third clade (**Figure 5A**). This topology is congruent with the plastome (**Supplementary Figure 2**) and ITS tree (**Figure 5B**) in retrieving clades A and C, but differs by grouping clade B and D genomes into one clade. To test whether the number of retrieved similarity-based clusters may change, two alternative sub-datasets with either clade A or clade C genomes excluded were analyzed. The exclusion of either clade C or clade A genomes resulted in the separation of *H. diffusa* from the two clade B species (**Supplementary Figure 6**), similar to the ITS-based phylogeny (**Figure 5B** and **Supplementary Figure 2**). While the repeat-based analysis identified three major infrageneric clades in *Heliophila* (**Figure 5A**), the interspecies relationships in clades A and C differed from those in the ITS-based tree (**Figure 5B**) and plastome phylogeny (**Supplementary Figure 3**).

RepeatExplorer2 comparative analysis read abundance matrix was transformed to distance matrix and used to reconstruct the hierarchical clustering relationship of *Heliophila* species and *Chamira* (**Supplementary Figure 7**). Clade A, B, and C species formed separated clusters in the reconstructed dendrogram, and clade D species *H. diffusa* together with *C. circaeoides* was retrieved as sister to the remaining *Heliophila* genomes. This clustering was incongruent with interclade relationships in the plastome (**Supplementary Figure 3**), ITS, and repeat sequence similarity-based (**Figure 5**) phylogenies.

Chromosomal Localization of the Identified Repeats

Chromosomal distribution of selected identified repeats was determined by fluorescence *in situ* hybridization (FISH) in six

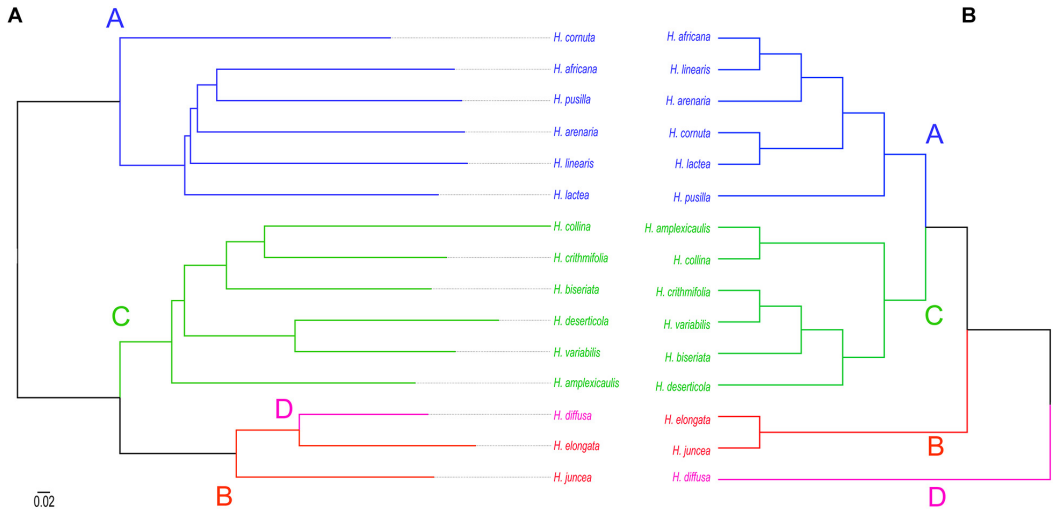

FIGURE 5 | Phylogenetic relationships between 15 *Heliophila* species based on repeat sequence similarities **(A)** and ITS sequences **(B)**. Scale bar indicates mean branch length across all individual trees to infer the consensus tree.

clade A species (*H. africana*, *H. arenaria*, *H. cornuta*, *H. lactea*, *H. linearis*, *H. pusilla*), two clade B species (*H. elongata*, *H. juncea*), three clade C species (*H. amplexicaulis*, *H. deserticola*, *H. variabilis*), in clade D species *H. diffusa*, and *C. circaeoides* (**Supplementary Table 6** and **Figure 6**).

In clade A, FISH of DNA probe for the 172-bp repeat HeAre2 identified species-specific chromosomal distribution of the satellite in three *Heliophila* genomes. HeAre2 was identified in pericentromeric heterochromatin of four chromosome pairs in *H. pusilla*, subtelomeric region of five chromosome pairs in *H. arenaria*, and at terminal heterochromatic knobs of seven chromosome pairs in *H. linearis*. In *H. arenaria* and *H. linearis*, the 174-bp HeAre1 repeat localized to pericentromeric regions of all and three chromosome pairs, respectively. In *H. africana*, *H. cornuta*, and *H. linearis*, the 171-bp HeLin1 tandem repeat showed localization at one, all (c. 24), and three chromosome pairs, respectively. In *H. lactea*, the 177-bp HeLac1 tandem repeats localized to pericentromeric regions of four chromosome pairs. The 177-bp HeAfr1 and 171-bp HeAre5 tandem repeats localized to all pericentromeres in *H. africana* and six chromosome termini in *H. arenaria*, respectively (**Figure 6**). In clade B genomes, the 167-bp HeJun2 and 383-bp HeElo2 repeats were present in subtelomeric regions of c. 20 chromosomes in *H. juncea* and pericentromeres of one chromosome pair in *H. elongata*, respectively (**Figure 6**).

In clade C species *H. variabilis*, four major tandem repeats formed pericentromeric chromatin. The 177-bp HeVar3 repeat localized to all chromosome pairs, the 168-bp HeVar2 provided hybridization signals on five chromosome pairs, the 184-bp HeVar1 localized to four chromosome pairs, and the 832-bp HeVar7 repeat gave hybridization signal on one chromosome pair. The 178-bp HeDes1 tandem repeat was located at all but two pericentromeres in *H. deserticola*. In *H. amplexicaulis*, 172-bp HeAmp2, 175-bp HeAmp3, and 184-bp HeAmp7 tandem repeats localized to pericentromeric heterochromatin of four, 11, and five chromosome pairs. The 188-bp HeAmp6 tandem repeat localized to subtelomeric regions of four chromosome pairs. Finally, the 162-bp HeAmp1 provided a strong hybridization signal at all interstitial and terminal heterochromatic knobs (**Figure 6**).

In clade D species *H. diffusa*, four major tandem repeats formed pericentromeres. The 177-bp HeDif2, 178-bp HeDif1, 184-bp HeDif3, and 171-bp HeDif4 repeats gave hybridization signals in all 22, c. 11, three, and one chromosome pair, respectively (**Figure 6**).

In *C. circaeoides*, three pericentromeric (294-bp ChCir2, 202-bp ChCir3, and 198-bp ChCir4) and two subtelomeric (249-bp ChCir1 and 1,427-bp ChCir10) tandem repeats were localized. The ChCir2 repeat was present in all pericentromeres, whereas ChCir3 and ChCir4 localized in centromeres to four and two chromosome pairs, respectively. ChCir1 and ChCir10 showed localization at chromosome termini of two different chromosome pairs (**Figure 6**).

In the investigated *Heliophila* and *Chamira* species, retrotransposons were mostly accumulated in pericentromeric heterochromatin; however, to a lesser extent, they were also distributed on chromosome arms (distribution of *Ty3/gypsy* in *H. africana*, *H. arenaria* subsp. *arenaria*, *H. linearis*, *H. elongata*, and *H. variabilis* and of *Ty1/copia* in *C. circaeoides* are shown in **Figure 6**).

Repeatome of *C. circaeoides* vs. Repeatomes of *Heliophila* Species

Detailed repeat analysis showed that *C. circaeoides* contained about 5% more repetitive elements (43.66%) in its genome compared with *H. diffusa* which exhibited the highest repeat content (38.67%) among *Heliophila* genomes (**Figure 3**), despite the genome size difference between the two species (461 and 800 Mb, respectively, **Figure 2**).

C. circaeoides showed minor differences in its overall repeatome composition compared with the 15 *Heliophila* species

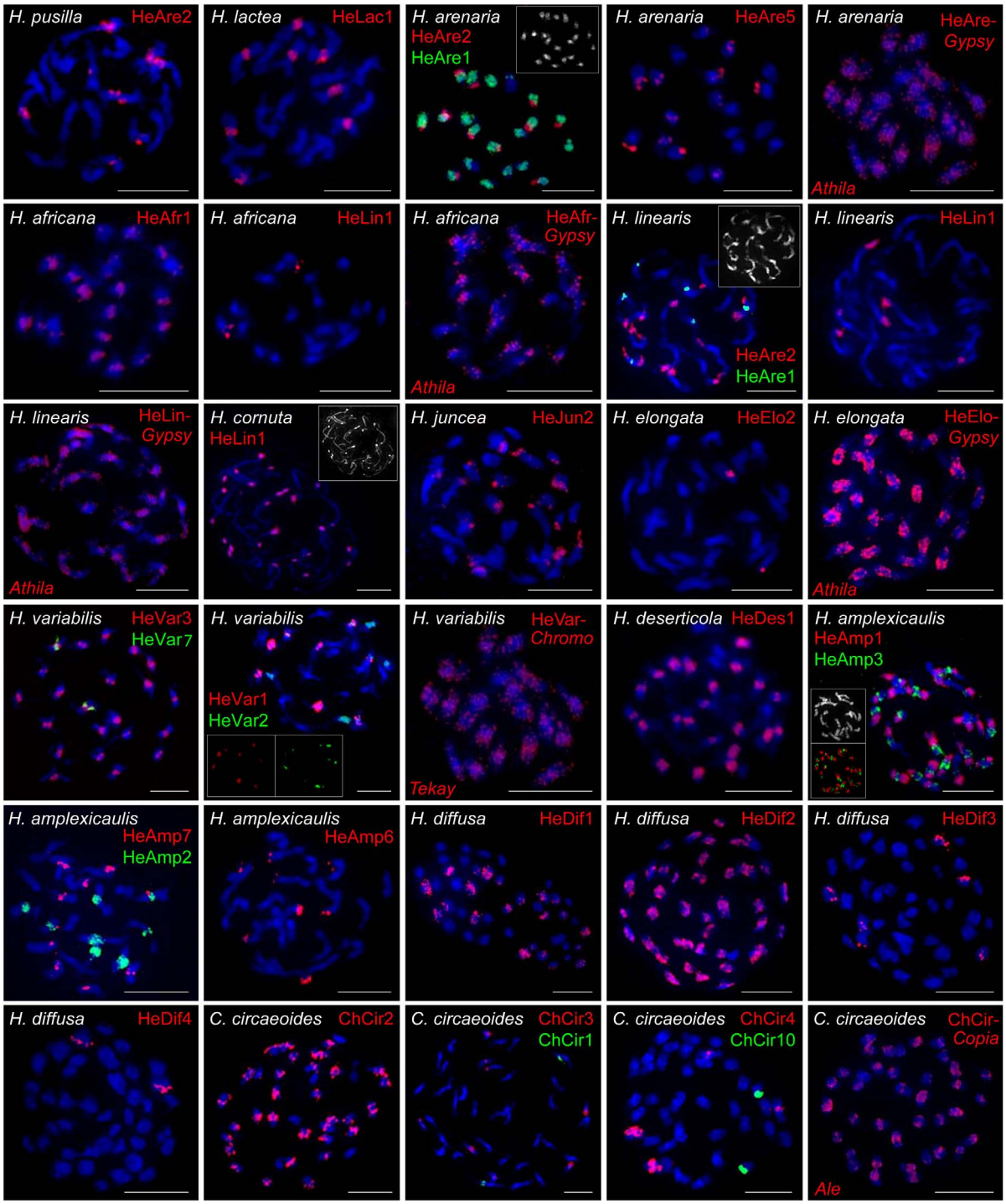

FIGURE 6 | FISH localization of the selected tandem repeats and retroelements on mitotic metaphase chromosomes of *Heliophila* species and *C. circaeoides*. Chromosomes were counterstained by DAPI; FISH signals are shown in color as indicated. Detailed information on the localized repeats is provided in **Supplementary Table 6**. Scale bars, 10 μm.

analyzed (**Figure 3**). Total LTR retrotransposon abundances in *C. circaeoides* were comparable with those observed in *Heliophila* genomes (16.69% in *C. circaeoides* vs. 11.51% to 22.42% in *Heliophila*, **Table 2**). Whereas *Ty3/gypsy* abundance was similar in *Chamira* and *Heliophila* genomes, *Ty1/copia* abundance in *C. circaeoides* was observed to be the highest (6.87%) among all the sequenced genomes (the highest proportion of *Ty1/copia* elements was detected in *H. elongata* – 5.92%). Unlike in

Heliophila species, where *Bianca* is the predominant *Ty1/copia* lineage, *Ale* lineage was the most abundant *Ty1/copia* element in *C. circaeoides* (3.31%) (**Table 2**).

The most distinct feature of the *C. circaeoides* genome is the high accumulation of tandem repeats (17.92%, **Figure 3**). In *Heliophila*, the highest genomic proportion of tandem repeats was found in *H. diffusa* (12.10%). In contrast to *Heliophila* genomes, long monomer satellites constitute a significant portion of the *C. circaeoides* tandem repeatome, such as ChCir5: 3,388 bp – 0.59%, ChCir6: 3,342 bp – 0.42%, ChCir7: 3,558 bp – 0.37%, and ChCir9: 4,034 bp – 0.17% (**Supplementary Table 11**).

DISCUSSION

The Origin of *Chamira* and *Heliophila* Was Preceded by a WGD

By analyzing transcriptomes of four *Heliophila* species (Mandáková et al., 2017; and this study), we corroborated the earlier conclusion based on chromosome painting data that the genus has undergone a mesopolyploid WGD (Mandáková et al., 2012, 2017). The occurrence of similarly positioned *Ks* peaks in *Heliophila* and *Chamira* genomes, along with the repeatedly retrieved sister relationship of both genera and their sympatry in the Greater Cape Floristic Region, suggests that either the WGD predated the *Chamira*/*Heliophila* divergence or the ancestors of both mesopolyploid genera were closely related. Whereas transcriptome-based divergence time estimates dated the WGD between 26 and 29 Mya and the *Chamira*/*Heliophila* split to 26 Mya, plastome-based dating yielded somewhat younger dates of the *Chamira*/*Heliophila* divergence (c. 21 Mya) and dated the diversification of the four *Heliophila* clades to c. 14–16 Mya. Nevertheless, transcriptome as well as plastome data congruently date the WGD to Oligocene or Miocene and major infrageneric cladogenesis in *Heliophila* to Middle Miocene. Although a much younger origin of *Heliophila* was previously proposed by Mandáková et al. (2012), we reason that those estimates were affected by the use of questionable fossil records and secondary calibration points (Franzke et al., 2016).

Chromosome number of $2n = 38$ established for *Chamira* (Mandáková et al., 2015; and this study) is similar to those of *Heliophila* neopolyploids ($2n = 32$, 36, 40, 44, 60, 64, 80, and 88; Mandáková et al., 2012) and suggests that the mesopolyploid WGD might have been followed by an additional genome duplication in *Chamira*. Only genome sequences of *C. circaeoides* can shed more light into its genome history and phylogenomic relationship to *Heliophila*. Despite the overall rarity of (neo)polyploidy in Cape flora (Oberlander et al., 2016), ancient WGDs, such as that documented in *Chamira* and *Heliophila*, are probably awaiting their discovery in other southern African angiosperm lineages.

Major Clades of the *Heliophila* Phylogeny

The monophyly of *Heliophila* and its sister position to *Chamira* were established by Mummenhoff et al. (2005) based on analysis of rDNA ITS sequences. That study retrieved three main clades in *Heliophila* which were confirmed as a basal trichotomy in a follow-up ITS-based study including more species (Mandáková et al., 2012). Herein, by further expanding our taxon sampling, we recovered four well-resolved ITS clades, with clade D (*H. diffusa*, *H. pendula*, and a putative hybrid, aff. *H. macra*) being sister to the three remaining clades (**Supplementary Figure 2**). The plastome phylogeny was largely congruent with the ITS tree, although it indicated a closer relationship between clades D and B (**Supplementary Figure 3**). The overall congruence between the two phylogenies further corroborates ITS as a reliable marker for inferring infrageneric relationships in *Heliophila* and other eukaryotic lineages (e.g., García-Robledo et al., 2013; Wang et al., 2015; Minamoto et al., 2017; Yang et al., 2018).

Future analyses of more unplaced species (particularly the *H. concatenata* species complex, *H. astyla*, *H. meyeri*, *H. obibensis*, *H. patens*, *H. scandens*, and some undescribed species) should clarify whether the basal D group could be expanded or whether further clades will be revealed. Altogether, ITS and plastome phylogenies corroborated a minimum of three to four major clades in *Heliophila*. At least two major clades (corresponding to ITS clades A+D and clades B+C, respectively) were retrieved based on pollen types (Kumwenda, 2003), and chromosome number variation in *Heliophila* also supports such cladogenesis. The chromosome number $2n = 44$ is repeated in the two most morphologically related species of clade D and is rare elsewhere. While species of clade A have mostly chromosome number of $2n = 20$, $2n = 22$ is prevalent in clade C species. The known chromosome numbers of B clade species are more variable ($2n = 16$, 22, 26, 32, and 64; *H. dregeana*, *H. elongata*, and *H. juncea*; Mandáková et al., 2012 and this study). This pattern is congruent with the sister relationship of clades A and C, as well as with the more ancestral position of clade B in both nuclear and plastome phylogenies (**Supplementary Figures 2,3**). While the occurrence of truly monocarpic species is limited to clades A, C, and D, the apparent woodiness and the presence of an intercalary type of inflorescence even in short-lived perennial species are diagnostic characters of clade B species.

Repeatome Diversity Is Reflecting Infrageneric Cladogenesis in *Heliophila*

A substantial fraction of nuclear plant genomes is composed of repeated DNA. These highly abundant genomic elements are influencing the function and evolution of plant genomes (e.g., Macas et al., 2011; Garrido-Ramos, 2015, 2017), and their diversity and abundance patterns can reflect phylogenetic distances (Dodsworth et al., 2014, 2017; Bolsheva et al., 2019; Vitales et al., 2020). Here, we sequenced and analyzed repetitive elements of 15 *Heliophila* species proportionally representing four major infrageneric clades. As transposable elements (TEs) are usually conserved across closely related species groups (Moisy et al., 2014; Wicker et al., 2018), we did not expect to identify clade-specific TEs in *Heliophila*. TAREAN analysis detected 108 tandem repeats in the sequenced *Heliophila* genomes. Fifty-four percent of all tandem repeats identified in *Heliophila* had a monomer length between 170 and 190 bp; the remaining 46% ranged widely in length from 27 to 2,012 bp in *H. juncea* and *H. biseriata*, respectively. Out of the 108 tandem repeats, 56

(51.9%) were species-specific, 32 (29.6%) shared among species of the same clade, and 20 (18.5%) were shared across the clades.

Most within-clade shared repeats were identified among clade A species – 43% of the repeats were shared among all six species and 14.3% were shared among three species of the clade. Tandem repeatomes of clade C species are more divergent, with 12% of shared repeats being shared among three out of six species analyzed and 31% of repeats shared by only two species. The two clade B species share only a single tandem repeat (one out of 12 identified). No shared repeats were found between any clade A and clade C species, and only 4 tandem repeats were homologous between clade B and clade D species.

According to the dated plastome phylogeny (**Supplementary Figure 3**), clades B+D split from clades A+C ∼12 Mya and all four clades diverged between 10 and 11 Mya; based on the ITS tree (**Supplementary Figure 2**), clade D was the first to diverge from the remaining three clades. While the major diversification within clades A and C occurred around 7 Mya, a number of speciation events in clade A seem to be younger than species diversifications in clade C (though the phylogeny suffers from species under-representation). As tandem repeats are evolving rapidly in most cases (Henikoff et al., 2001; Melters et al., 2013), their sequence conservation can be observed only on short evolutionary distances (Henikoff et al., 2001; Meraldi et al., 2006; Koukalova et al., 2010; Renny-Byfield et al., 2013; Dodsworth et al., 2014). Hence, the highest number of shared repeats among clade A species may reflect their close relationships and more recent origins. Similarly, in clade C, the highest number of shared repeats was identified in the species pair *H. crithmifolia–H. variabilis* representing the youngest (3.4 Mya) speciation within this clade. Some identified tandem repeats had a relic character, linking distantly related lineages, such as the two repeats shared between *H. elongata* and *H. juncea* from clade B and *H. crithmifolia* and *H. variabilis* from clade C. The most basal species *H. diffusa* (clade D) shares one repeat (HeDif5) with all clade A genomes and two with clade C species (HeDif1: *H. amplexicaulis*, HeDif2: *H. amplexicaulis* and *H. deserticola*). The three repeats shared between clade C and D genomes, and the only tandem repeat (HeCor4) shared between clade A (*H. cornuta*) and clade B (*H. juncea*) remained conserved for 12 million years since the divergence of these clades.

The Use of Tandem Repeats to Infer Phylogenetic Relationships Among Plant Genomes

While low-pass genome skimming of plant genomes is not adequate to analyze their gene space, repetitive sequences present in thousands of copies are sufficiently represented in this data. Repeat analysis using graph-based clustering methods allowed for computationally efficient and robust characterization of repetitive elements and provided much deeper insights into repeatome structure and evolution (Harkess et al., 2016; Doronina et al., 2017; McCann et al., 2020). Moreover, abundances of *de novo* identified repetitive elements were found to carry phylogenetic signals (Dodsworth et al., 2014, 2016, 2017). If assuming that repeat abundances are evolving through random genetic drift (Jurka et al., 2011), the abundances can be analyzed as continuous characters for phylogeny inference (Dodsworth et al., 2014). When using a genome proportion of 0.1% or higher, this method proved to be highly congruent with phylogenies inferred using other nuclear or plastome markers (Dodsworth et al., 2014, 2017; Bolsheva et al., 2019). Recently, Vitales et al. (2020) reported a novel approach of phylogenetic inference using repeats as markers. They utilized the RepeatExplorer2 similarity matrices and generated derived matrices which consist of the observed/expected read similarity values by considering the number of reads of each taxon that are represented in clusters. Thus, the matrices consist of pairwise sequence similarities, disregarding the number of reads for each species. By transforming these similarity matrices to distance matrices, they were able to build consensus networks for each dataset. Similar to the abundance-based method, the lineage-specific differences between homologous repeats are regarded to be regulated by random genetic drift in diversification, thus expected to carry phylogenetic signals (Jurka et al., 2011; Dodsworth et al., 2014; Vitales et al., 2020). However, it should be noted that tandem repeats undergo rapid turnover in plant (e.g., Koukalova et al., 2010; Renny-Byfield et al., 2013) and animal genomes (e.g., Sinha and Siggia, 2005; Cechova et al., 2019) and that their phylogenetic signals can be erased during long-term reproductive isolation and independent evolution of initially closely related genomes. As approaches using repeats as phylogenetic markers are still in their infancy, these phylogenetic inferences should be applied cautiously, along with other marker gene sets (Vitales et al., 2020), as done here for inferring phylogenetic relationships in Heliophileae.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repositories and accession numbers can be found in the **Supplementary Tables 1–12**.

AUTHOR CONTRIBUTIONS

ML and TM conceived the experiments. MD, MP, TM, PH, XG, PW, ZC, and IA-S conducted the study and/or processed the data. MD, XG, MP, TM, PW, IA-S, AV, KM, LM, and ML wrote the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the Czech Science Foundation (project no. 19-07487S), the National Geographic Society (grant no. 9345-13), the RVO 67985939 project, and the CEITEC 2020 project (grant no. LQ1601). LM acknowledges the

logistic support of the Iluka Chair in Vegetation Science and Biogeography at the Murdoch University.

ACKNOWLEDGMENTS

We thank Syd Ramdhani for assisting in the field work, Anže Žerdoner Čalasan for MrBayes analyses on CIPRES, and Sheng Zuo for helping with the phylogenetic analyses using repetitive elements. The Core Facility Plants Sciences of CEITEC MU is acknowledged for obtaining the scientific data presented in this paper. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program "Projects of Large Research, Development and Innovations Infrastructures". RepeatExplorer analyses were supported by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.607893/full#supplementary-material

Supplementary Table 1 | List of *Heliophila* and *Chamira* accessions used in our study.

Supplementary Table 2 | Chromosome numbers and genome sizes of *Heliophila* and *Chamira* species used for repeatome and plastome assembly.

Supplementary Table 3 | Summary statistics of transcriptome assemblies.

Supplementary Table 4 | List of species used in transcriptome analyses and phylogeny reconstruction based on transcriptome data.

Supplementary Table 5 | List of selected chloroplast genomes and their respective GenBank accession numbers used in maternal phylogeny inference.

Supplementary Table 6 | Repetitive elements investigated by FISH.

Supplementary Table 7 | Inferred meso-WGD Ks peaks and estimated ages of WGD in the sampled *Heliophila* species and *C. circaeoides*.

Supplementary Table 8 | Mixture modeling of synonymous divergence between *Heliophila* and *Chamira* species.

Supplementary Table 9 | Summary of SNP frequency analysis based on transcriptome data.

Supplementary Table 10 | Chloroplast genome assembly statistics of analyzed *Heliophila* and *Chamira* species.

Supplementary Table 11 | Tandem repeats identified by RepeatExplorer2 and manual curation.

Supplementary Table 12 | Pairwise sequence similarities between shared tandem repeats.

Supplementary Figure 1 | BUSCO analysis of gene completeness.

Supplementary Figure 2 | Fifty percent majority-rule consensus tree of the Bayesian inference of the ITS region.

Supplementary Figure 3 | Time-calibrated plastome phylogeny of 59 Brassicaceae species.

Supplementary Figure 4 | Dot-plot pairwise comparison of monomer consensus sequences of 16 identified tandem repeats shared among *Heliophila* species.

Supplementary Figure 5 | Multiple and pairwise sequence alignments of 16 tandem repeats shared among 13 *Heliophila* species.

Supplementary Figure 6 | Consensus network relationships of *Heliophila* species based on repeatome sequence similarities.

Supplementary Figure 7 | Hierarchical clustering relationships of 15 *Heliophila* and *C. circaeoides* species produced from read abundance matrix produced by RepeatExplorer2 pipeline.

REFERENCES

Al-Shehbaz, I. A. (2012). A generic and tribal synopsis of the Brassicaceae (Cruciferae). *Taxon* 61, 931–954. doi: 10.1002/tax.615002

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Benson, G. (1998). "An algorithm for finding tandem repeats of unspecified pattern size," in *Proceedings of the Second Annual International Conference on Computational Molecular Biology*, (New York, NY: ACM), 20–29. doi: 10.1145/279069.279079

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bolsheva, N. L., Melnikova, N. V., Kirov, I. V., Dmitriev, A. A., Krasnov, G. S., Amosova, ÀV., et al. (2019). Characterization of repeated DNA sequences in genomes of blue-flowered flax. *BMC Evol. Biol.* 19:49. doi: 10.1186/s12862-019-1375-6

Brown, J. W., Walker, J. F., and Smith, S. A. (2017). Phyx: phylogenetic tools for unix. *Bioinformatics* 33, 1886–1888. doi: 10.1093/bioinformatics/btx063

Cechova, M., Harris, R. S., Tomaszkiewicz, M., Arbeithuber, B., Chiaromonte, F., and Makova, K. D. (2019). High satellite repeat turnover in great apes studied with short-and long-read technologies. *Mol. Biol. Evol.* 36, 2415–2431. doi: 10.1093/molbev/msz156

Davidson, N. M., and Oshlack, A. (2014). Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.* 15, 1–14.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2016). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45:e18.

Dodsworth, S., Chase, M. W., Kelly, L. J., Leitch, I. J., Macas, J., Novák, P., et al. (2014). Genomic repeat abundances contain phylogenetic signal. *Syst. Biol.* 64, 112–126. doi: 10.1093/sysbio/syu080

Dodsworth, S., Chase, M. W., Särkinen, T., Knapp, S., and Leitch, A. R. (2016). Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae). *Biol. J. Linn. Soc.* 117, 96–105. doi: 10.1111/bij.12612

Dodsworth, S., Jang, T.-S., Struebig, M., Chase, M. W., Weiss-Schneeweiss, H., and Leitch, A. R. (2017). Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). *Plant Syst. Evol.* 303, 1013–1020. doi: 10.1007/s00606-016-1356-9

Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2:2233. doi: 10.1038/nprot.2007.310

Doronina, L., Churakov, G., Kuritzin, A., Shi, J., Baertsch, R., Clawson, H., et al. (2017). Speciation network in Laurasiatheria: retrophylogenomic signals. *Genome Res.* 27, 997–1003. doi: 10.1101/gr.210948.116

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.

Franzke, A., Koch, M. A., and Mummenhoff, K. (2016). Turnip time travels: age estimates in Brassicaceae. *Trends Plant Sci.* 21, 554–561. doi: 10.1016/j.tplants.2016.01.024

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

García-Robledo, C., Erickson, D. L., Staines, C. L., Erwin, T. L., and Kress, W. J. (2013). Tropical plant–herbivore networks: reconstructing species interactions using DNA barcodes. *PLoS One* 8:e52967. doi: 10.1371/journal.pone.0052967

Garrido-Ramos, M. A. (2015). Satellite DNA in plants: more than just rubbish. *Cytogenet. Genome Res.* 146, 153–170. doi: 10.1159/000437008

Garrido-Ramos, M. A. (2017). Satellite DNA: an evolving topic. *Genes* 8:230. doi: 10.3390/genes8090230

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Wang, X., et al. (2017). Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* 18:176. doi: 10.1186/s12864-017-3555-3

Haas, B., and Papanicolaou, A. (2016). *TransDecoder (Find Coding Regions Within Transcripts)*. Available online at: https://sourceforge.net/projects/transdecoder/

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Harkess, A., Mercati, F., Abbate, L., McKain, M., Pires, J. C., Sala, T., et al. (2016). Retrotransposon proliferation coincident with the evolution of dioecy in *Asparagus*. *G3* 6, 2679–2685. doi: 10.1534/g3.116.030239

Henikoff, S., Ahmad, K., and Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293, 1098–1102. doi: 10.1126/science.1062939

Hohmann, N., Wolf, E. M., Lysak, M. A., and Koch, M. A. (2015). A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* 27, 2770–2784.

Huang, D. I., and Cronk, Q. C. B. (2015). Plann: a command-line application for annotating plastome sequences. *Appl. Plant Sci.* 3:1500026. doi: 10.3732/apps.1500026

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Jurka, J., Bao, W., and Kojima, K. K. (2011). Families of transposable elements, population structure and the origin of species. *Biol. Direct* 6:44. doi: 10.1186/1745-6150-6-44

Kagale, S., Robinson, S. J., Nixon, J., Xiao, R., Huebert, T., Condie, J., et al. (2014). Polyploid evolution of the Brassicaceae during the Cenozoic era. *Plant Cell* 26, 2777–2791. doi: 10.1105/tpc.114.126391

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

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

Kiefer, C., Willing, E.-M., Jiao, W.-B., Sun, H., Piednoël, M., Hümann, U., et al. (2019). Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. *Nat. Plants* 5, 846–855. doi: 10.1038/s41477-019-0486-9

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinform.* 7:474. doi: 10.1186/1471-2105-7-474

Kolde, R., and Kolde, M. R. (2015). *Package 'Pheatmap.' R Package 1, 790*. Available online at: https://www.rdocumentation.org/packages/pheatmap/versions/1.0.12/topics/pheatmap

Koukalova, B., Moraes, A. P., Renny-Byfield, S., Matyasek, R., Leitch, A. R., and Kovarik, A. (2010). Fall and rise of satellite repeats in allopolyploids of *Nicotiana* over c. 5 million years. *New Phytol.* 186, 148–160. doi: 10.1111/j.1469-8137.2009.03101.x

Kumwenda, M. W. (2003). *A Palynological Study of Heliophila (Brassicaceae) in Southern Africa*. M. Sc. Thesis, University of Stellenbosch, Stellenbosch.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2016). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–773.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Meth.* 9:357. doi: 10.1038/nmeth.1923

Lysak, M. A., and Koch, M. A. (2011). "Phylogeny, genome, and karyotype evolution of crucifers (Brassicaceae)," in *Genetics and Genomics of the Brassicaceae*, eds R. Schmidt and I. Bancroft (New York, NY: Springer), 1–31. doi: 10.1007/978-1-4419-7118-0_1

Macas, J., Kejnovský, E., Neumann, P., Novák, P., Koblížková, A., and Vyskot, B. (2011). Next generation sequencing-based analysis of repetitive DNA in the model dioceous plant *Silene latifolia*. *PLoS One* 6:e27335. doi: 10.1371/journal.pone.0027335

Mandáková, T., Mummenhoff, K., Al-Shehbaz, I. A., Mucina, L., Mühlhausen, A., and Lysak, M. A. (2012). Whole-genome triplication and species radiation in the southern African tribe Heliophileae (Brassicaceae). *Taxon* 61, 989–1000. doi: 10.1002/tax.615006

Mandáková, T., Winter, P., Al-Shehbaz, I. A., Mucina, L., Mummenhoff, K., Lysak, M. A., et al. (2015). "Brassicaceae. IAPT/IOPB chromosome data 19," in *Taxon*, Vol. 64, ed. K. Marhold (Hoboken, NJ: Wiley), 1068–1074.

Mandáková, T., and Lysak, M. A. (2016a). Chromosome preparation for cytogenetic analyses in Arabidopsis. *Curr. Protoc. Plant Biol.* 1, 43–51. doi: 10.1002/cppb.20009

Mandáková, T., and Lysak, M. A. (2016b). Painting of Arabidopsis chromosomes with chromosome-specific BAC clones. *Curr. Protoc. Plant Biol.* 1, 359–371. doi: 10.1002/cppb.20022

Mandáková, T., Li, Z., Barker, M. S., and Lysak, M. A. (2017). Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *Plant J.* 91, 3–21. doi: 10.1111/tpj.13553

Marais, W. (1970). "Cruciferae," in *Flora of Southern Africa*, Vol. 13, eds L. E. Codd, B. De Winter, D. J. Killick, and H. B. Rycroft (Pretoria: Government Printer), 1–118.

McCann, J., Macas, J., Novák, P., Stuessy, T. F., Villaseñor, J. L., and Weiss-Schneeweiss, H. (2020). Differential genome size and repetitive DNA evolution in diploid species of *Melampodium* sect. *Melampodium* (Asteraceae). *Front. Plant Sci.* 11:362. doi: 10.3389/fpls.2020.00362

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., et al. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* 14:R10. doi: 10.1186/gb-2013-14-1-r10

Meraldi, P., McAinsh, A. D., Rheinbay, E., and Sorger, P. K. (2006). Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. *Genome Biol.* 7:R23. doi: 10.1186/gb-2006-7-3-r23

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES science gateway for inference of large phylogenetic trees," in *Proceedings of the 2010 Gateway Computing Environments Workshop (GCE)*, (New Orleans, LA: IEEE), 1–8.

Minamoto, T., Uchii, K., Takahara, T., Kitayoshi, T., Tsuji, S., Yamanaka, H., et al. (2017). Nuclear internal transcribed spacer−1 as a sensitive genetic marker for environmental DNA studies in common carp *Cyprinus carpio*. *Mol. Ecol. Resour.* 17, 324–333. doi: 10.1111/1755-0998.12586

Moisy, C., Schulman, A. H., Kalendar, R., Buchmann, J. P., and Pelsy, F. (2014). The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years. *Theor. Appl. Genet.* 127, 1223–1235. doi: 10.1007/s00122-014-2293-z

Mummenhoff, K., Al-Shehbaz, I. A., Bakker, F. T., Linder, H. P., and Mühlhausen, A. (2005). Phylogeny, morphological evolution, and speciation of endemic Brassicaceae genera in the Cape flora of southern Africa. *Ann. Missouri Bot. Garden* 92, 400–424.

Mummenhoff, K., Linder, P., Friesen, N., Bowman, J. L., Lee, J., and Franzke, A. (2004). Molecular evidence for bicontinental hybridogenous genomic constitution in *Lepidium* sensu stricto (Brassicaceae) species from Australia and New Zealand. *Am. J. Bot.* 91, 254–261. doi: 10.3732/ajb.91.2.254

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Novák, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinform.* 11:378. doi: 10.1186/1471-2105-11-378

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and MacAs, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 29, 792–793. doi: 10.1093/bioinformatics/btt054

Novák, P., Robledillo, L. Á, Koblížková, A., Vrbová, I., Neumann, P., and Macas, J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* 45:e111. doi: 10.1093/nar/gkx257

Oberlander, K. C., Dreyer, L. L., Goldblatt, P., Suda, J., and Linder, H. P. (2016). Species-rich and polyploid-poor: insights into the evolutionary role of whole-genome duplication from the Cape flora biodiversity hotspot. *Am. J. Bot.* 103, 1336–1347. doi: 10.3732/ajb.1500474

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Meth.* 14, 417–419. doi: 10.1038/nmeth.4197

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* 201178. doi: 10.1101/201178

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Available online at: https://www.R-project.org/

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67:901. doi: 10.1093/sysbio/syy032

Rannala, B., and Yang, Z. (2007). Inferring speciation times under an episodic molecular clock. *Syst. Biol.* 56, 453–466. doi: 10.1080/10635150701420643

Renny-Byfield, S., Kovarik, A., Kelly, L. J., Macas, J., Novak, P., Chase, M. W., et al. (2013). Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *Plant J.* 74, 829–839. doi: 10.1111/tpj.12168

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Sinha, S., and Siggia, E. D. (2005). Sequence turnover and tandem repeats in cis-regulatory modules in *Drosophila*. *Mol. Biol. Evol.* 22, 874–885. doi: 10.1093/molbev/msi090

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., and Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* 26, 1134–1144. doi: 10.1101/gr.196469.115

Song, L., and Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience* 4:48.

Sonnhammer, E. L. L., and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167, GC1–GC10.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577. doi: 10.1080/10635150701472164

Temsch, E. M., Greilhuber, J., and Krisai, R. (2010). Genome size in liverworts. *Preslia* 82, 63–80.

Thomas, G. W. C., Ather, S. H., and Hahn, M. W. (2017). Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Syst. Biol.* 66, 1007–1018. doi: 10.1093/sysbio/syx044

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., et al. (2014). XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* 16, 62–74.

Van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. *Methods Mol. Biol.* 804, 281–295. doi: 10.1007/978-1-61779-361-5_15

Vitales, D., Garcia, S., and Dodsworth, S. (2020). Reconstructing phylogenetic relationships based on repeat sequence similarities. *Mol. Phylogen. Evol.* 147:106766. doi: 10.1016/j.ympev.2020.106766

Wang, X., Liu, C., Huang, L., Bengtsson-Palme, J., Chen, H., Zhang, J., et al. (2015). ITS 1: a DNA barcode better than ITS 2 in eukaryotes? *Mol. Ecol. Resour.* 15, 573–586. doi: 10.1111/1755-0998.12325

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R. H., et al. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* 19, 1–18.

Yang, R.-H., Su, J.-H., Shang, J.-J., Wu, Y.-Y., Li, Y., Bao, D.-P., et al. (2018). Evaluation of the ribosomal DNA internal transcribed spacer (ITS), specifically ITS1 and ITS2, for the analysis of fungal diversity by deep sequencing. *PLoS One* 13:e0206428. doi: 10.1371/journal.pone.0206428

Yang, Y., and Smith, S. A. (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* 31, 3081–3092. doi: 10.1093/molbev/msu245

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* 19:153. doi: 10.1186/s12859-018-2129-y

Zwaenepoel, A., and Van de Peer, Y. (2019). wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* 35, 2153–2155. doi: 10.1093/bioinformatics/bty915

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 01 February 2021
doi: 10.3389/fpls.2020.588354

Comparative Phylogeography of *Veronica spicata* and *V. longifolia* (Plantaginaceae) Across Europe: Integrating Hybridization and Polyploidy in Phylogeography

*Daniele Buono[1], Gulzar Khan[1], Klaus Bernhard von Hagen[1], Petr A. Kosachev[2], Eike Mayland-Quellhorst[1], Sergei L. Mosyakin[3] and Dirk C. Albach[1]**

[1] Institute for Biology and Environmental Sciences, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany, [2] Faculty of Biology, Altai State University, Barnaul, Russia, [3] M.G. Kholodny Institute of Botany, National Academy of Sciences of Ukraine, Kyiv, Ukraine

OPEN ACCESS

Edited by:
Karol Marhold,
Institute of Botany, Slovak Academy of Sciences, Slovakia

Reviewed by:
Gonzalo Nieto Feliner,
Real Jardín Botánico (RJB, CSIC), Spain
Ines Carrasquer,
Natural History Museum of Geneva, Switzerland

***Correspondence:**
Dirk C. Albach
dirk.albach@uni-oldenburg.de

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: 28 July 2020
Accepted: 19 November 2020
Published: 01 February 2021

Citation:
Buono D, Khan G, von Hagen KB, Kosachev PA, Mayland-Quellhorst E, Mosyakin SL and Albach DC (2021) Comparative Phylogeography of Veronica spicata and V. longifolia (Plantaginaceae) Across Europe: Integrating Hybridization and Polyploidy in Phylogeography. Front. Plant Sci. 11:588354. doi: 10.3389/fpls.2020.588354

Climatic fluctuations in the Pleistocene caused glacial expansion-contraction cycles in Eurasia and other parts of the world. Consequences of these cycles, such as population expansion and subsequent subdivision, have been studied in many taxa at intraspecific population level across much of the Northern Hemisphere. However, the consequences for the potential of hybridization and polyploidization are poorly understood. Here, we investigated the phylogeographic structure of two widespread, closely related species, *Veronica spicata* and *Veronica longifolia*, across their European distribution ranges. We assessed the extent and the geographic pattern of polyploidization in both species and hybridization between them. We used genome-scale SNP data to clarify phylogenetic relationships and detect possible hybridization/introgression events. In addition, crossing experiments were performed in different combination between *V. spicata* and *V. longifolia* individuals of two ploidy levels and of different geographic origins. Finally, we employed ecological niche modeling to infer macroclimatic differences between both species and both ploidy levels. We found a clear genetic structure reflecting the geographical distribution patterns in both species, with *V. spicata* showing higher genetic differentiation than *V. longifolia*. We retrieved significant signals of hybridization and introgression in natural populations from the genetic data and corroborated this with crossing experiments. However, there were no clear phylogeographic patterns and unequivocal macroclimatic niche differences between diploid and tetraploid lineages. This favors the hypothesis, that autopolyploidization has happened frequently and in different regions. The crossing experiments produced viable hybrids when the crosses were made between plants of the same ploidy levels but not in the interploidy crosses. The results suggest that hybridization occurs across the overlapping areas of natural distribution ranges of both species, with apparently directional introgression from *V. spicata* to *V. longifolia*. Nevertheless, the two species maintain their species-level separation due to their adaptation to different habitats and spatial isolation rather than reproductive isolation.

Keywords: hybridization, polyploidy, comparative phylogeography, steppe, Europe, macroclimatic niche

INTRODUCTION

The past 30 years have seen an enormous increase in knowledge of how species migrated and evolved, thanks to the analyses of intraspecific genetic variation and progress in the field of phylogeography (Avise, 2000, 2009). Pleistocene refugia and migration corridors have been inferred (e.g., Hewitt, 2000; Harrison and Noss, 2017) and phylogeographic patterns among species have been compared to infer commonalities in dispersal history and differences, indicative of changes in community composition over time (e.g., Brunsfeld et al., 2001; Muellner-Riehl, 2019). Based on such information, hypotheses about the effect of climate change on biodiversity decline have been established (Keppel et al., 2012; Carvalho et al., 2019). A special situation occurs when species interact, i.e., they hybridize or form polyploids. In most cases, such studies examined the origin of hybrids or polyploids on a narrow spatial scale (e.g., Schönswetter et al., 2007; Balao et al., 2015). Undetected hybrids or structure caused by intraspecific lineages with different ploidy levels can potentially mislead analyses to infer genetically distinct groups or can inflate genetic diversity of particular groups of populations. Further, hybridization and polyploidy have potential large effects on a variety of processes shaping the distribution range of a species (Nieto Feliner, 2014; Arrigo et al., 2016; Paule et al., 2017), for example allowing species to exploit new ecological niches. Likewise, the advantages of polyploids over diploid progenitors are the increased number of alleles to mask the deleterious recessive mutations and lead to heterosis in allopolyploids (Gu et al., 2003), and neo-functionalization or sub-functionalization of duplicated genes copies. These factors are considered important for ecological niche expansion and provision of high flexibility toward environmental challenges (Adams and Wendel, 2005; Birchler et al., 2010), facilitating colonization of new and extreme habitats. In consequence, populations of a species affected by hybridization might have survived in places where the non-hybrid populations of that species normally would not be able to occur (Kadereit, 2015). These factors may be especially prevalent in phylogeographic analyses, in which both diploids and polyploids are widespread and sympatric. Here, we compared phylogeographic patterns of two species able to hybridize and assessed how hybridization and polyploidization might have affected these patterns considering past climatic changes. We have studied two species of the genus *Veronica* L., *Veronica spicata* L., and *Veronica longifolia* L. from *Veronica* subgenus *Pseudolysimachium* (W.D.J.Koch) Buchenau, both are widely sympatric across Eurasia and known to have two different ploidy levels, the polyploids (in our case, tetraploids) being derived by autopolyploidy (Bardy et al., 2011; Albach, unpubl.). The two species, however, differ in ecology. *Veronica spicata* occurs on nutrient-poor, open and sunny dry meadows, dunes, or rocks. It is considered to have been a member of the late-glacial steppe-tundra-community and occurs nowadays in the remnants of this vegetation (Pigott and Walters, 1954). It is highly intolerant of shade and competition (Pigott and Walters, 1954), which makes it endangered in large parts of its western distribution area. In contrast, *V. longifolia* is a species inhabiting nutrient-rich, mesic to moist habitats along rivers, lakes or even ditches or in swamps (Winter et al., 2008).

Veronica subgenus *Pseudolysimachium* includes ∼30 species distributed across Eurasia, with several of them used in horticulture, especially cultivars of *V. spicata* and *V. longifolia* (Albach et al., 2004; Kosachev et al., 2019). These are also the most common and widespread species of the group. The Altai region, with 13 reported species, is considered the center of its diversity; also, seven hybrids have been described there based on morphology (Kosachev et al., 2015 and 2016). The northern Balkan Peninsula is the second center of diversity of the group, with six species (Albach and Fischer, 2003). The origin of the subgenus has been estimated to have occurred 12.5 Mya (±4 My), likely caused by a polyploidization event (Meudt et al., 2015), thus what we consider diploids in this study are in fact ancient tetraploids. Nevertheless, more recent polyploidization events have occurred within the last 3 My (confidence interval 4–0.5 Mya) within the subgenus (Meudt et al., 2015). Besides polyploidization, introgressive hybridization events between more or less diverged lineages in the subgenus may have occurred. Hybridization has been reported repeatedly in the subgenus. For example, Kosachev et al. (2016), studied 18 species of *V.* subg. *Pseudolysimachium*, reported an unresolved phylogenetic structure based on cpDNA and ITS1, with most species being polyphyletic, and multiple samples of *V. spicata* and *V. longifolia* showing signs of introgression from the other species. Nevertheless, the ITS1 region at least resolved alleles into six groups with most alleles from *V. longifolia* in one group and those of *V. spicata* in another. Kosachev et al. (2016) also suggested an East Asian origin followed by repeated westwards range expansions across the Pleistocene Eurasian steppes with frequent interspecific hybridization events. Gene flow between *Veronica* species (*V. spicata, V. orchidea* Crantz, and *V. barrelieri* H.Schott) has been shown to be common in the Balkan Peninsula, with genetic isolation restricted to few geographic regions (Bardy et al., 2011). Potential hybridization between species in the subgenus has also been detected in the Altai Mountains using SRAP (Sequence-related amplified polymorphism) markers, albeit with weak support (Kosachev et al., 2019). Bardy et al. (2011) suggested that complete homogenization of the identified groups is prevented by ecological divergence, ploidy differences, and geographic isolation. While not being responsible for speciation, hybridization may have caused the present wide morphological variation by generating several variants, which are morphologically intermediate between the recognized species.

Hybridization in the subgenus, though not specifically between *V. spicata* and *V. longifolia,* has been reported in a number of places across Eurasia, such as in the Balkan Mountains, Eastern Alps, Southern Carpathians (Bardy et al., 2011), Altai Mountains (Kosachev et al., 2019; Khan et al., submitted), and possibly other regions (Kosachev et al., 2016). However, it is still unclear whether hybridization occurred more widely across the Eurasian steppes and between *V. spicata* and *V. longifolia* or not. The intermediate plants between both species inferred to be hybrids have taxonomically been called *V.* × *media* Schrad. Based on crossing experiments between several species of the subgenus, Härle (1932) and Graze (1933)

reported that hybridization between plants with the same ploidy level is possible and the fertile hybrids are able to backcross with their parents. Both species, *V. spicata* and *V. longifolia*, have diploid ($2n = 34$) and tetraploid ($2n = 68$) populations (Albach et al., 2008). The two different ploidy levels found in *V. spicata* and *V. longifolia* (2× and 4×), however, do not seem to be related to differences in either geographical distribution, ecological requirements, or morphology (Trávníček et al., 2004). However, in earlier publications, Raitanen (1967) reported that morphologically determined samples of *V. spicata* and *V. longifolia* in Fennoscandia seem to hybridize despite differences in ploidy levels (with *V. spicata* occurring in the region only as 4× and *V. longifolia* mainly as 2×). According to Härle (1932), *V. spicata* and *V. longifolia* are self-incompatible, but Wilson et al. (2000) reported *V. spicata* as self-compatible. Similarly, Scalone et al. (2013) suggested that both species are facultatively outcrossing. Pollination in the subgenus is mostly performed by insects (entomophily) and is not species-specific (i.e., performed by non-specialized and mainly generalist pollinators), providing the possibility of hybridization in natural habitats. Bees (Hymenoptera: Apoidea, mainly small- to medium sized species) seem to be the most frequent visitors of the species, but Lepidoptera (Müller, 1881; Knuth, 1909; Ellis and Ellis-Adam, 1994; Albach and Kosachev, unpublished), flies (Diptera: small-sized Syrphidae and various other groups) and small-sized beetles (Coleoptera: Mordellidae, Nitidulidae, and some other non-specialized anthophilous groups) are also observed as occasional flower visitors (Kampny, 1995; Mosyakin, unpublished). It looks like the guilds of pollinators of *V. spicata* and *V. longifolia* are recruited from among the local faunas of non-specialist (both nectar-collecting and pollen-eating) anthophilous insects and thus may differ considerably in various regions. At the same time, such pollinator flexibility allows opportunities for cross-pollination within and between the two plant species, when they occur in close proximity. In addition, both species have rhizomes and can reproduce vegetatively by underground rhizome systems (Wilson et al., 2000), a trait considered important for the establishment and local-scale spread of hybrid and polyploid populations (Herben et al., 2017).

Here, we investigated the phylogeographic structure of *V. spicata* and *V. longifolia* species in parallel and estimated the extent of hybridization between both species. We employed genome wide SNPs (single nucleotide polymorphism) to analyze genetic diversity and hybridization/introgression in both *V. spicata* and *V. longifolia*. Furthermore, we conducted crossing experiments to test for the possibility of generating viable offspring from crosses between *V. spicata* and *V. longifolia* individuals of the same as well as different ploidy levels and geographic origins. Finally, we assessed the geographic distribution of ploidy levels in both species. We based our initial hypotheses on the works of Härle (1932) and Graze (1933), and the observation in the wild by Borsos (1967) and Trávníček et al. (2004). Our initial hypotheses were, first, that hybridization between *V. longifolia* and *V. spicata* is restricted to individuals having the same ploidy levels [e.g., *V. longifolia* (2×) × *V. spicata* (2×) and *V. longifolia* (4×) × *V. spicata* (4×)], while different ploidy levels [*V. longifolia* (2×) × *V. spicata* (4×) and *V. longifolia* (4×) × *V. spicata* (2×)] represent an important crossing barrier. Second, we also expected a significant genetic intraspecific structure across the native distribution ranges of the species reflecting ecological and spatial isolation in the wild. Lastly, we hypothesized that ecological niches of the two species did not differ based on the wide occurrence but considered it likely that macroclimatic differentiation between diploid and tetraploid individuals at both intra- and interspecific levels occurred in both species.

MATERIALS AND METHODS

Sampling

We collected 177 accessions: 81 of *V. spicata,* 74 of *V. longifolia,* one putative hybrid (*V.* × *media*), and *Veronica schmidtiana* as an outgroup for the molecular analyses. *V. schmidtiana* has been used as outgroup since it is the only species unequivocally positioned within the subgenus and sister to the clade containing both ingroup species (**Figure 1** and **Supplementary Table 1**). The accessions were collected on various field excursions during 2000–2019 in different areas across the distribution range of the species (**Figure 1B**) and identified using the morphological key provided in Albach and Fischer (2003). During the excursion, fresh leaf samples were collected and dried in silica gel (Voucher information; **Supplementary Table 1**). For crossing experiments, living plants of various accessions have been cultivated in the greenhouse of the Botanical Garden of Carl von Ossietzky-University (see details below). Vouchers for all crosses are deposited in the herbarium of Carl von Ossietzky University of Oldenburg (OLD).

For the analysis of ecological niches, we generated a dataset of 404 specimens with known ploidy levels based on chromosome counting or flow cytometry and sufficiently precise geographical coordinates (about one arc minute resolution or better; **Figure 2** and **Supplementary Table 2**) from literature data and our own estimations. New records of ploidy level directly measured by us constitute 38% of the data. In the first step, we excluded data points with less than 25 km linear distance from another point in each group (**Table 1**) using the geosphere package (Hijmans et al., 2019) for R (R Core Team, 2019). This reduces the chance that individuals might stem from the same (meta-) population and it is also convenient for statistical purposes when there is only one data point per grid square or less at the same time. This step also reduces a bias in geographical sampling, which can lead to distorted results (Fourcade et al., 2014). After thinning, we used a total of 248 data points. The threshold of using 25 km linear distance (and not an even greater distance) for exclusion was based on the assumption that a lower sample number is negatively affecting the statistical confidence. Therefore, we sticked to our 248 data set. Following our genetic analysis about 4% on average of all existing individuals in both species (or data points in the macroclimatic analysis) might have been compromised by introgression. It will become evident further below that the variation within each group and the overlap between groups was that high that these potential 4% could not have perturbed the results in any way. Plants for which potential

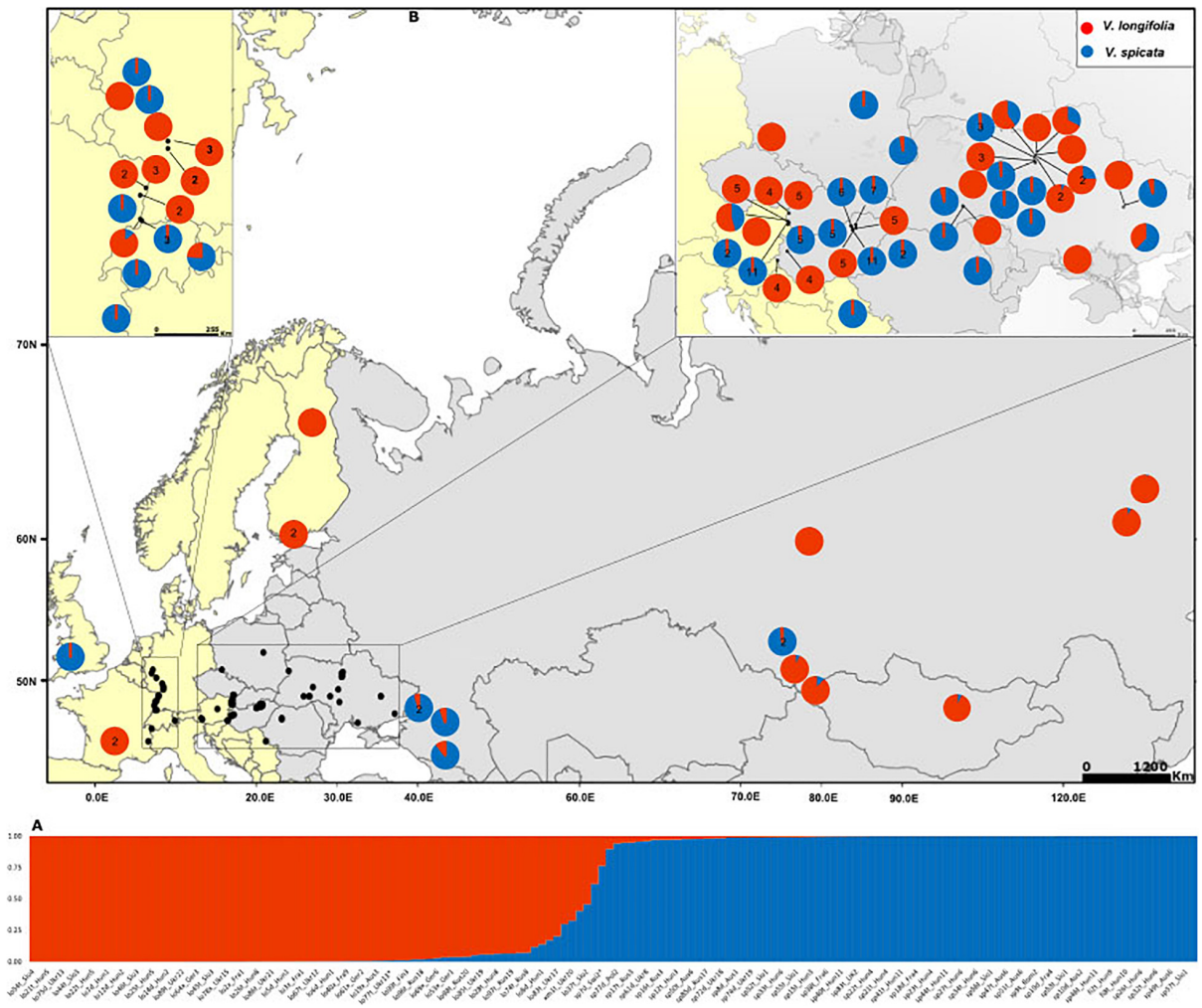

FIGURE 1 | (A) Assignment of individuals to $K = 2$ clusters based on Bayesian statistics as implemented in STRUCTURE. Blue color represents the probability of *Veronica spicata* ancestry and red for *Veronica longifolia* ancestry. Complete details of all individuals with the same order are in **Supplementary Table 1**. **(B)** Distribution map of the sampled individuals with genetic composition based on $K = 2$. Pie charts without number represent only one individual whereas those with numbers give the number of individuals included from that particular locality. Genetic composition of the localities with more than one individual has been averaged accept where the difference among the individual's genetic composition exceeded 10%.

hybrid influence was detected in our genetic analyses were not part of the macroclimatic analyses except for two individuals. One of these individuals was included in tetraploid *V. longifolia* and one in tetraploid *V. spicata* based on their typical morphology and the majority of the genetic markers found. Excluding both would not make any difference because their climatic parameters were typical for their respective groups.

Flow Cytometry and DNA Extraction

Ploidy information of all samples was assessed by flow cytometry following Meudt et al. (2015) on a CyFlow SL (Partec, Germany). Overall, we report here 156 new measurements of ploidy for the two species. Since intra-population ploidy variation occurs rarely in *V.* subg. *Pseudolysimachium* (∼1.8% of the cases; Bardy et al., 2011), we measured for the majority of populations only one sample after confirming the finding of Bardy et al. (2011) for 12 populations in which we measured three individuals (**Supplementary Table 2**). DNA was extracted from about 20 mg of dried leaf material using innuPREP Plant DNA Kit (Analytic Jena AG, Jena, Germany) following the manufacturer's instructions. The quality of DNA was checked using gel electrophoresis (1% agarose gel). Photometric DNA concentration and A260/A280 ratio was measured using TECAN infinite F200 Pro (Tecan Group AG, Männedorf, Switzerland).

Genotyping by Sequencing

Several datasets from different years and sequencing runs were combined for the analysis in this study. Combinability of runs was checked in preliminary analyses including the same individual sequenced in different runs. All GBS (genotype by sequencing) libraries were prepared following Elshire et al.

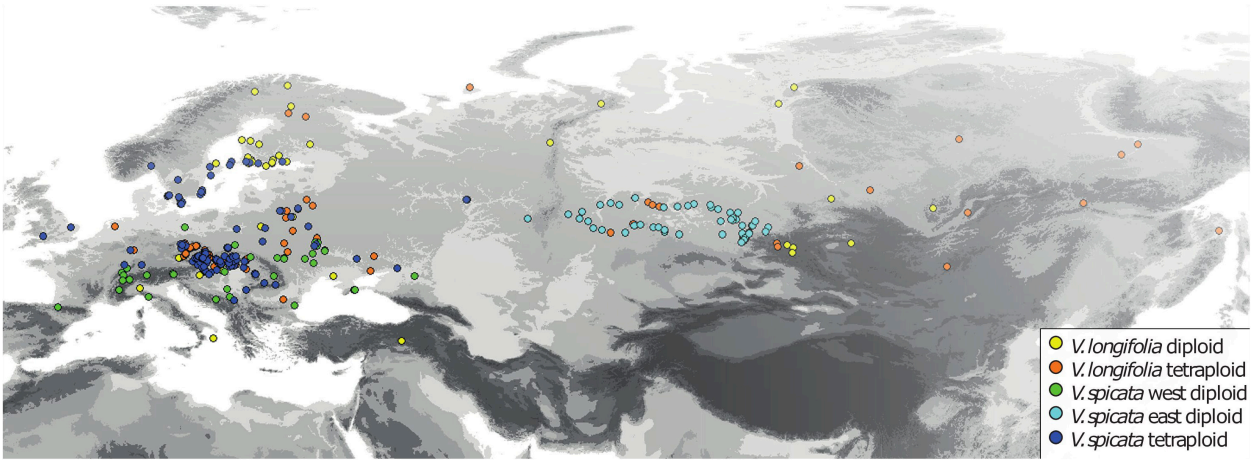

FIGURE 2 | Geographical distribution of all *V. longifolia* and *V. spicata* individuals with known ploidy level (*N* = 404).

TABLE 1 | Number of points for which georeferenced data and ploidy level was known.

Veronica groups	All available	After thinning	Subsets
Diploid *V. longifolia*	48	40	
Tetraploid *V. longifolia*	102	65	
Diploid *V. spicata*	118	74	33 west, 41 east
Tetraploid *V. spicata*	136	69	
Total	404	248	

In the second column only the points with more than 25 km linear distance from each other were included. Only diploid V. spicata was divided in two subsets for some analyses.

(2011), using the methylation-insensitive restriction enzyme *Msl*I (5′. . .CAYNN^NNRTG. . .3′) with modifications as described by Siadjeu et al. (2018). Briefly, 200 ng genomic DNA were included in restriction-digestion with one unit *Msl*I (New England Biolabs, NEB) in 30 µl (1× NEB4 buffer) at 37°C for 1 h and heat inactivated at 80°C for 20 min. Adaptor ligation was conducted with 15 µl of restriction product and 3 µl ligation adapters (Ovation Rapid DR Multiplex System, Nugen Technologies, Leek, Netherlands) in 12 µl of master mix (4.6 µl D1 water, 6 µlL1 ligation buffer mix, and 1.5 µlL3 ligation enzyme) at 25°C for 15 min and heat inactivated at 65°C for 10 min. The ligation product was finalized in 20 µl of the kits 'final repair' master mix at 72°C for 3 min which was then purified with magnetic beads (homemade Serapure beads). The purified ligated products were amplified with a low number of PCR-cycles (10 cycles) using 4 µl of 5× MyTaq buffer (Bioline), 0.2 µl polymerase and 1 µl (10 pmol/µl) of standard Illumina TrueSeq amplification primers. To remove small fragments from the libraries, all amplicons were purified again with magnetic beads. In this way, libraries were prepared for all individuals, normalized, and pooled. An additional step of purification was performed to remove the PCR polymerase through Qiagen MinElute Columns (Hilden, Germany) from the pooled library. The genomic library was sent to LGC genomics (Berlin, Germany) for sequencing.

RAD-Seq *de novo* Assembly Into Loci and Genotyping of SNPs

Sequencing quality was checked using FASTQC version 0.11.9 (Andrews, 2010) and all the results in HTML format were aggregated in MULTIQC version 1.9 (Ewels et al., 2016). At the end, we included the reads with mean Phred quality score >28, without any adapter contamination and a per base *N* content close to 0. In addition, to ensure and avoid any type of false positive/negative errors in SNP calling, we employed process_radtags step from the STACKS version 2.4 (Catchen et al., 2013). The settings for process_radtags were: trimming after 135 bp; remove all uncalled bases with the option -c and low quality scores with the option -q. After confirming upon the quality of the sequencing, we used STACKS version 2.4 (Rochette et al., 2019) pipeline to assemble the raw reads into *de novo* loci and to genotype the SNPs. The pipeline consists of the sequential execution of different tools as (I) USTACKS, (II) CSTACKS, (III) SSTACKS, (IV) TSV2BAM, (V) GSTACKS, and (VI) POPULATIONS. USTACKS was used to construct a set of putative loci and detect SNP at each locus following: maximum distance allowed between stacks = 4; minimum depth of coverage to create a stack = 5; maximum number of stacks at a single *de novo* locus = 5. Some of these settings were dictated by having tetraploid individuals in the dataset. Subsequently, the tool CSTACKS was executed to build a catalog (a set of consensus loci) from the set of processed samples, allowing 4 mis-matches between sampled loci when building the catalog. SSTACKS was then used to search a set of stacks against the catalog produced. We used TSV2BAM to transpose the R2 reads, by orienting it to their corresponding loci instead by sample associated with each single-end locus assembled *de novo* in the previous steps. Following TSV2BAM, we assembled the paired-end reads into a contig, merging the contig with the single-end locus, and aligned reads from individual samples to the locus within GSTACKS. Lastly, the tool POPULATIONS from STACKS was used to filter the data and produce VCF (Variant Call Format, ver. 4.2) files. The minimum percentage of individuals across

populations required to process a locus (option −R) was selected as 0.5 after inspection of several phylogenetic trees obtained with various values of −R option (0.9, 0.75, 0.5; results not shown) since this setting has important effect on the maximum missing data (for example with R = 0.5, 50% of missing data). At the end, we generated two different VCF files: one with more than one SNP per locus (concatenated), and one with a single (the first) SNP per locus (using the populations option –write-single-snp), assuming that loci are independent (e.g., unlinked). In the downstream population genomics, we only used the biallelic SNPs assumed to be unlinked to ensure that allele frequencies are not correlated. We included the post process filtering step to remove the variants with low (<8) and high coverages (>800). This step is necessary as the variants sequenced at very low coverage may represent only one of four alleles in tetraploids and the one with high coverage may result from repetitive regions. Subsequently, we removed the variants and individuals with more than 50% missing data. This post-process step was performed using the R package *VCFR* (Knaus and Grünwald, 2017) and self-written R script (**Supplementary Material 1**). The tool *PARALOG-FINDER* implementing the method of McKinney et al. (2017) was applied to remove paralogs in our dataset before the downstream analyses, however, no appreciable differences in the results were found and we kept the original datasets. Our final dataset for population genomics included 27% missing data for concatenated loci (290114 SNPs; 153 individuals) and 25.73% missing data on unlinked loci (8723 SNPs; 157 individuals).

The two datasets contain different numbers of individuals as the post process filtering step removed samples using the same limit of missing information admitted, which differs slightly between samples from unlinked and concatenated dataset (since on unlinked dataset missing data is related with only the first SNP per locus; **Supplementary Table 1**).

Phylogeographic Structure and Admixture

To assess the genetic structure involving all individuals of both species sampled across the sampling area, we firstly adopted a non-hierarchical approach in STRUCTURE version 2.3 (Pritchard et al., 2000) using the unlinked dataset. STRUCTURE groups the individuals originating from different genetic clusters and identifies migrants. It is also useful in studies involving hybridization using Bayesian statistics by applying Markov Chain Monte Carlo (MCMC) and multiple repetition (Pritchard et al., 2000; Porras-Hurtado et al., 2013). We included all individuals of *V. longifolia* and *V. spicata* in STRUCTURE analysis to assign them to their respective genetic clusters, find the role of geography in their structuring and investigate admixed individuals resulted from the contribution of alleles from both species. The analysis was run for 100000 MCMC (Markov chain Monte Carlo) burning the initial 10% chains, with the admixture model and correlated allele frequencies between populations, from K one to four each with 20 iterations, considering that migration is still occurring. This analysis included morphologically defined *V. spicata* (81 individuals), *V. longifolia* (74 individuals) and *V.* × *media* (one individual). To investigate the role of geography in sub-structuring of both species, we ran an additional second-level hierarchical analysis as well including only samples of *V. spicata* and *V. longifolia* with the same parameters except K = 13. We used the program CLUMPP version 1.1.2 (Jakobsson and Rosenberg, 2007) to clump and align all the 20 iteration for the optimal K value following the Evanno approach as employed in STRUCTURE HARVESTER version 0.6.94 (Evanno et al., 2005; Earl and von Holdt, 2012) selecting the Full-Search option. To further explore the role of geography, we used TESS3R package in R (Caye et al., 2016; **Supplementary Material 2**) as well. TESS3R determines genetic variation in natural populations considering both genetic and geographic data simultaneously. We used the output results of the second level hierarchical analyses on *V. spicata* and *V. longifolia* from STRUCTURE (results not shown) and combined them with geographic coordinates of all the included individuals.

To check the relationship among the individuals from different geographical localities and complement the STRUCTURE/TESS3R results, we further constructed the maximum likelihood trees as well using the concatenated SNPs data. For this analysis, we used two different datasets, one including all individuals and another one excluding admixed individuals (based on STRUCTURE results). We considered the individuals admixed which had admixture proportion more than 20%. The best substitution model of evolution was assessed with JMODELTEST-NG (Darriba et al., 2019), which provided GTR + I as best model with both BIC and AIC criteria. The actual analysis was then carried out in RAxML version 8.2.12 (Stamatakis, 2014). To inform RAxML about the nature of the dataset containing only concatenated SNPs without invariable sites, we applied the ASC correction as suggested (Stamatakis, 2014) to minimize the impact on branch lengths. To check the robustness of our results, we ran the analysis with 100 bootstrap replicates. In addition to phylogenetic reconstruction, pairwise F_{ST} statistics were also calculated (Wright, 1949) following Weir and Cockerham (1984). This analysis was implemented using the function stamppFst of the R package STAMPP (Pembleton et al., 2013; **Supplementary Material 3**) on the concatenated dataset with 100 bootstrap replicates and a confidence interval of α = 0.05.

Characterization of Hybrids Classes and Introgression

After getting the admixing proportion through Bayesian approach between *V. longifolia* and *V. spicata* in STRUCTURE, we allocated the samples to distinct hybrid or pure classes, specifically, we tested to identify the presence of parent species individuals, F1 or F2 hybrids and backcrossed individuals to both species [parental *V. longifolia* (A) or *V. spicata* (B)]. We only included the individuals for which no single class had $Q \geq 0.80$ probabilities to belong to one parent species. We implemented this analysis using the program NEWHYBRID (Anderson and Thompson, 2002). NEWHYBRIDS assign individuals to specific hybrid classes by looking at the patterns of gene inheritance

within each locus. Since NEWHYBRIDS cannot handle large datasets, we created subsets of 200 unlinked most informative and differentiated SNP loci (method = *AvgPIC*) using the function *gl.nhybrids* included in the R package *DARTR* (Gruber and Georges, 2019; **Supplementary Material 3**). This function identifies loci that exhibit a fixed difference by comparing two sets of parental populations and creates an input file for the program NewHybrids. To avoid any further bias caused by the heterozygous loci of parental populations, we only included the pure individuals, those showing less than 1% of admixture from STRUCTURE analysis. Additionally, to get the robust results, we run the program with no prior information regarding the status or class of individuals. The analysis was run for 10^5 MCMC iterations with a burn-in threshold of 10^4. To run the program NEWHYBRID, we used the R package PARALLELNEWHYBRID (Wringe et al., 2016, 2017).

To check the possibility of introgression between both species, we measured genomic clines in the admixed individuals shared by the parental species as implemented in INTROGRESS (Gompert and Buerkle, 2009, 2010, 2012). INTROGRESS analyzes introgression of genotypes between divergent lineages and hybridizing species. In addition, INTROGRESS estimates genomic clines from co−dominant, dominant, and haploid marker data, without any requirement of fixed allelic differences between parental populations for the sampled genetic markers. The package tests for deviations of loci from neutral expectations as well. INTROGRESS is ideal in cases with high number of informative markers in which the potentially admixed individuals cover the full range of variation between parental species (Gompert and Buerkle, 2010). Since our data is based on continuous sampling, without any spatial correspondence, INTROGRESS suits well to our model. Briefly, we divided the data into three groups, two homozygotes (pure parental) and one heterozygote genotypes (admixed individuals), based on the NEWHYBRID and STRUCTURE results. These groups included 63 individuals of *V. longifolia*, 79 individuals of *V. spicata* and 13 admixed individuals ($Q \geq 0.2$). In the actual analysis, firstly, we estimated hybrid indices and then genomic clines using multinomial regression. We measured introgression for each individual locus in admixed individuals relative to clines using their genotype frequency along an admixture gradient (Buerkle, 2005; Gompert and Buerkle, 2010) by comparing likelihood of regression models to that of a neutral model. The significance of likelihood was accessed using 1000 permutations. Lastly, we counted the introgressed loci from the observed hybrid genotypes by comparing with the probability densities of homozygotes and heterozygotes genotypes as suggested (Nolte et al., 2009). We adjusted the neutral expectation of genomic clines through multiple comparisons using the false discovery rate (FDR) as suggested (Benjamin and Hochberg, 1995; examples in Harrison et al., 2017; Khan et al., 2020).

Crossing and Germination Experiments

We conducted crossing experiments between individuals of *V. spicata* and *V. longifolia* of both ploidy levels and geographic origins to assess experimentally reproductive barriers between the two species. The experiments were conducted in the greenhouses of the Carl von Ossietzky-University (Oldenburg, Germany) in spring and summer 2019. When enough inflorescences were available on a single plant, three treatments were applied for each replicate beside the controls (Ctr): H (hybridization), C (negative control) and CS (cage + selfing). To avoid any unwanted pollination from insects, all floral spikes were bagged using tea-bag filters. In treatment H self-pollination was tried to avoid by emasculation of the flowers on daily basis before anthesis. When the flowers opened, the freshly dehiscent anthers of the donor species were removed with clean tweezers and their pollens were transferred by direct contact with stigma on the receptor species. Negative controls (treatment C) were used to determine the minimum amount of pollen contamination acceptable after applying only emasculation. Negative controls were performed in parallel with H on the same plant. Spikes for treatment CS were bagged before anthesis till fruit maturation. Further, the controls (Ctr) were used to determine the germination rate of seeds pollinated with pollen of the same species. They were handled in a similar way to H with the exception that the pollen donor was from the same species and ploidy level as the control.

Seeds obtained from the crossing experiment were sterilized following the protocol suggested by Pickens et al. (2003): (1) 70% ethanol for 2 min; (2) rinse in distilled water for 5 min; (3) 2.6% NaOCl + Tween 20 for 40 min; (4) rinse in sterile water twice for 5 min each. After sterilization, seeds were placed in Petri dishes on filter paper, adding sterilized distilled water. A climate chamber was used to ensure controlled temperature and light conditions, with the settings: light 12/12 h; and temperature 21°C. The position of the single Petri dishes was randomized inside the chamber in order to account for small variation of temperature and light related to the position. Germination rate was defined as the number of seeds germinated after 30 days divided by the total number of seeds plated.

Polyploidy and Ecological Differentiation

All niche analyses are based on the WorldClim data set (Hijmans et al., 2005) with a resolution of 2.5 min. The 19 bioclim variables were attached to the occurrence data of 248 individuals (see sampling) using DIVA-GIS (Hijmans et al., 2012).

In a preliminary principal component analysis (PCA) including all variables, we found that many of the variables were highly correlated with each other. To avoid collinearity and type II errors caused by high correlation, we applied pairwise Pearson product correlations in R (R Core Team, 2019). This Pearson correlations test retrieved bioclim variables with mod $\geq$ 0.8 such as: bio4 and bio7; bio5 and bio10; bio1, bio6, bio9, and bio11; bio12, bio13, bio14, bio16, bio17, bio18, and bio19, and the remaining four variables, i.e., bio2, bio3, bio8, and bio15 were not correlated. In the actual analysis we used eight variables, excluding the remaining 11 variables. The choice of the four variables that showed correlation with other variables was based on criteria such as longer time period over shorter period in the definition of the variables and more or less subjective hypotheses about the biological importance in our species. The eight variables included were: bio2 mean diurnal range, bio3 isothermality, bio7 temperature

annual range, bio8 mean temperature of wettest quarter, bio9 mean temperature of driest quarter, bio10 mean temperature of warmest quarter, bio15 precipitation seasonality, and bio17 precipitation of driest quarter.

The reduced data set with 248 specimens and eight bioclimatic variables included was subjected to a PCA using NTSYSPC version 2.2 (Rohlf, 2009). It turned out that the Siberian samples of diploid *V. spicata* (>50° East) grow under rather distinctive climatic conditions. Tetraploid *V. spicata* do not occur there. To accommodate for this, we repeated several of our analyses separating the Siberian samples (>50° East) from diploid *V. spicata* from the West (**Table 1**). Such a problem did not arise in *V. longifolia*, because its diploid and tetraploid groups occur in the same general geographical regions or there were only very few data points in the one group when far distant from the other group.

To qualitatively compare climatic niche models (or in other words: the macroclimatic envelopes of a group projected on a geographical map) between conspecific diploid and tetraploid groups, we used MAXENT version 3.4.1 (Phillips et al., 2006, 2019) on the data set described above (248 data points divided in four or in five groups). For the application of MAXENT, we followed Merow et al. (2013) and Fourcade et al. (2014), who recommended minimizing the correlation among predictors. Therefore, we again used only the eight bioclim variables as described above. For purposes as in our study, they recommended to use the relevant and complete geographical space for background selection. We therefore first limited the bioclim layers in DIVA-GIS (Hijmans et al., 2012) to 5°W, 135°E, 35°N, and 75°N (**Figure 2**) and used this restricted rectangle of the eight climatic layers in MAXENT. It was not possible to account for sampling effort in our data, because the data was opportunistically gathered from different times and sources. However, a facultative exclusion of the densely sampled Siberian *V. spicata* and the thinning of all data to 25 km linear distance follow the suggestions of Fourcade et al. (2014) to minimize geographically biased sampling. We tested different output options in MAXENT and chose the logistic output because the results seemed less diffuse (steeper and therefore clearer gradients) than with the other options.

We investigated differentiation in macroclimatic niches between diploids and tetraploids among individual variables using the analysis of variance (ANOVA) tool in EXCEL 2016 (Microsoft Corp., Redmond; significance considered $p < 0.05$). All variables were tested for both taxon combinations and this might lead to statistical type I errors. Therefore, we additionally applied Bonferroni corrections for all 19 parallel comparisons, which leads to $p < 0.003$ as a much more conservative threshold for significance. Since the differences found between European and Siberian samples of diploid *V. spicata* could be misleading (see below), we repeated the ANOVA for *V. spicata* including only the western samples (west of 50° East).

Finally, to get a quantitative idea about climatic niche differences between the ploidy levels or between species, we calculated Schoener's *D*, a frequently used index for niche overlap, which ranges between 0 and 1 (the latter stands for complete niche overlap) from the original MAXENT results. We used the precompiled version 1.3 for windows of ENMtools (Warren et al., 2008, 2010) for this and also for the following analysis. To test whether the MAXENT scores exhibit statistically significant ecological differences between several groups, we performed a niche identity test using Schoener's *D* as a test variable. In this test the empirical occurrence points were pooled, their identity randomized, subjected to a MAXENT analysis, and Schoener's *D* was calculated. We did 100 pseudo-replicates per comparison. The resulting distribution of Schoener's *D* represents the null hypotheses that niches differ by chance only and it can be compared with the *D* score of the original data to judge about the probability that niches are identical.

RESULTS

Phylogeographic Structure and Admixture

The Bayesian modeling to assess the genetic structure clustered all individuals in two groups (Evanno method; $K = 2$; **Figure 1A**) corresponding to *V. spicata* and *V. longifolia*. This analysis clearly showed that the genetic structure of both species is in accordance with their morphology-based classification. However, our results also recovered individuals with admixture higher than 20% (**Figure 1A** and **Supplementary Table 1**). The threshold of 20% was subsequently used to infer hybrids since individuals with more than 20% did not group according to geography in other analyses. The second level hierarchical clustering on species alone found optimal intraspecific clustering in *V. spicata* at $K = 2$ (result not shown). However, after averaging the results of STRUCTURE between individuals in the populations and using the R package TESS3R to plot those averages on a map, we recovered a reasonable geographical pattern until $K = 5$ (**Figure 3**). The Evanno method indicated an optimal $K = 2$ for the second level hierarchical clustering analysis using the dataset of *V. longifolia* in STRUCTURE and TESS3R was used to plot the results. At $K = 2$, the individuals were structured into one group composed of individuals from Asia and the other including those from Europe (**Figure 3**) except for one individual from Siberia [Rus18 (lo96t_Rus18)] grouping with European individuals. Admixture between the two species based on the STRUCTURE analysis, considering individuals with $Q \geq 0.20$ for the minor ancestry as admixed, occurred in seven individuals. Of these, four individuals were *a priori* identified as *V. longifolia*, two as *V. spicata*, and one considered a hybrid. Admixture occurred in both ploidy levels, six in tetraploids and one in a diploid. It was found widely from Switzerland to Ukraine. The divergence between *V. spicata* and *V. longifolia* was lower, with an F_{ST} value of 0.296 ± 0.002, than between *V. longifolia* and *V. schmidtiana* ($F_{ST} = 0.625$), and between *V. spicata* and *V. schmidtiana* ($F_{ST} = 0.632$; **Supplementary Table 3**), while highly significant for all ($P = \sim 0.0$).

Similarly, the phylogenetic analysis based on maximum likelihood also clustered all individuals into two main groups with high support corresponding to *V. spicata* and *V. longifolia*. Since individuals admixed according to STRUCTURE appeared basally branching and not in geographical meaningful clusters,

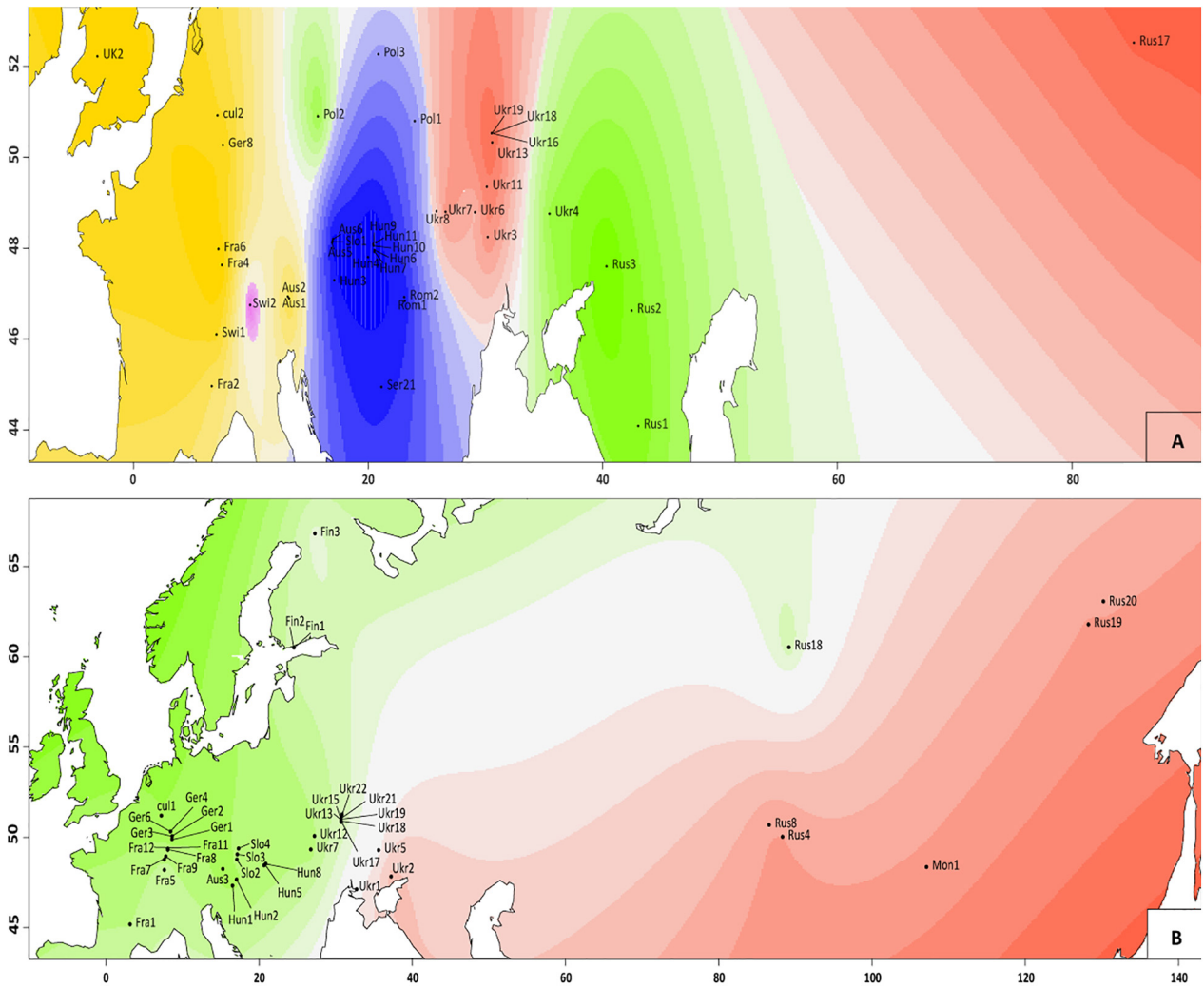

FIGURE 3 | Genetic clusters identified with TESS3R analysis in *V. spicata* **(A)** and *V. longifolia* **(B)** populations. Different colors indicate different clusters inferred (two for *V. longifolia* and five for *V. spicata*) and different shading indicates different probabilities to belong to the respective cluster. The graphs are based on the results of STRUCTURE between individuals in the populations, as employed in R package tess3r (details are in main text).

we consider in the following only the analysis without admixed individuals (**Figure 4** and see **Supplementary Figure 1** for results including hybrids). The phylogenetic analysis also recovered sub-clustering representing Asian/East European individuals and more western European individuals in both species, except lo96t_Rus18. This individual was cultivated in a Russian botanical garden with an unknown geographic origin and placed in a group that contains individuals from Central and Western Europe (France, Germany, Switzerland, and Austria), which suggests that the garden did not cultivate a local genotype.

Putative Hybrid Individuals and Introgression

Probabilities to be parental species A or B (*PA* for *V. longifolia* and *PB* for *V. spicata*) obtained in NEWHYBRIDS followed the same pattern as explored with Bayesian statistics in STRUCTURE (**Figure 5A**). The NEWHYBRID results did not recover any individual to be an F_1 hybrid. The estimated posterior probability to be an F_2 hybrid was high (>0.88) for lo37t_Slo2, lo15t_Ukr2, and xm1t_Ukr20, while for sp7d_Swi2 and lo83t_Ukr17 the probability to be F_2 was lower (0.51 and 0.32, respectively). Sample sp7d_Swi2 had a posterior probability of 0.45 to be a back-cross with *V. longifolia* parental group. Similarly, samples sp73t_Ukr19, lo18x_Fra5, lo83t_Ukr17, lo84t_Ukr18, and lo74t_Rus8 showed high probabilities (>0.6) to be backcrossed individuals of *V. longifolia,* whereas lo8d_Hun1, lo58t_Slo4, lo20t_Mon1, and lo24t_Hun5 had lower probabilities (<0.6) to be in the same category (**Figure 5A**). Similarly, the hybrid indices <0.78 or >0.18 recovered with INTROGRESS with very low heterozygosity ∼0.02, suggest that the hybrid individuals are fertile (**Figure 5B**).

The genomic clines approach implemented in INTROGRESS supported the hypothesis of higher probability of homozygotes

FIGURE 4 | Phylogenetic analysis based on maximum likelihood statistics as implemented in RAxML. Two main groups with high support correspond to *V. spicata* (upper part) and *V. longifolia* (lower part). The IDs assigned to the samples (e.g., lo82t_Ukr17) are constructed using the first two letter to identify the species, e.g., lo for *V. longifolia*, sp for *V. spicata*, and sc for *V. schmidtiana*, followed by a unique number for every individual and a letter to indicate the ploidy level (*d* for 2×, *t* for 4×, and × for unknown ploidy). Underscore separates the information related with population constituted by 3 letters to identify the country (e.g., CzR for Czechia) and a unique number in that species (see complete detail in **Supplementary Table 1**).

in admixed individuals than expected under neutral expectations. The homozygosity inferred was 41% from *V. longifolia* and 51% from *V. spicata* (**Figure 5C**). We recovered only one heterozygous locus in all admixed individuals, reflecting high homozygosity. The results infer a higher tendency of homozygotes clines from *V. spicata* with 216 more SNPs than *V. longifolia*. This tendency of 18% more homozygotes from *V. spicata* suggests asymmetrical introgression. For this analysis we only used the non-neutral loci,

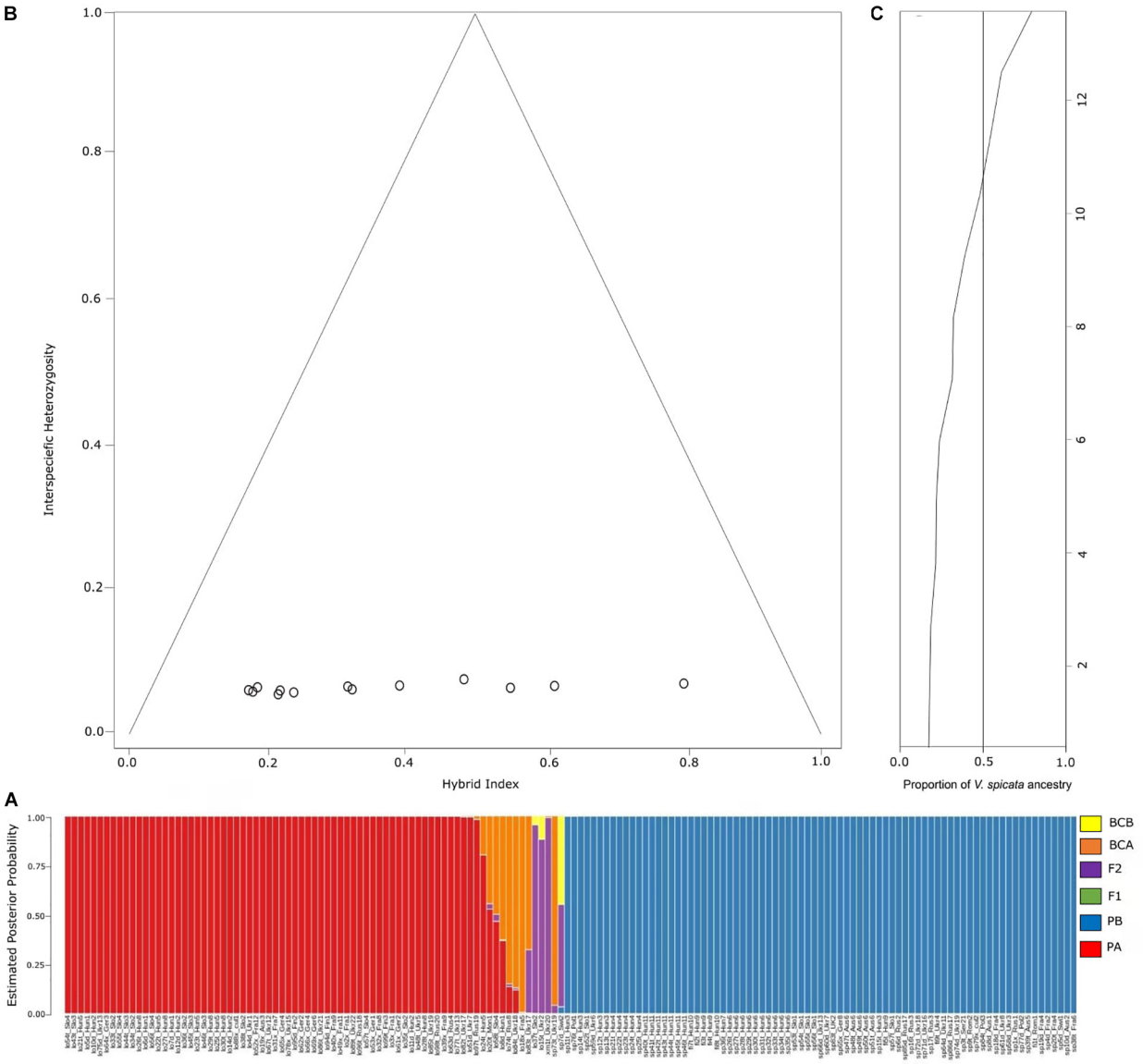

FIGURE 5 | (A) Probability assignment into different hybrid categories according to NEWHYBRIDS analysis. PA, parental population A (*V. longifolia*); PB, parental population B (*V. spicata*); BCA, backcross with parental population A; BCB, backcross with parental population B; F1, F1 hybrid between populations A and P; F2, F2 hybrid between two F1 individuals. **(B)** Triangle plot detailing the relationship between interspecific heterozygosity against hybrid index of *V. longifolia*, *V. spicata* and admixed individuals. The hybrid index was calculated as proportion of *V. longifolia* and *V. spicata* alleles (h = 0 indicates individuals with alleles from *V. longifolia* ancestry, and h = 1 denotes *V. spicata* ancestry). We considered those individuals from the STRUCTURE presenting $Q \leq 0.80$ as putative hybrid individuals; and **(C)** ancestry plot showing probability based on hybrid indices of *V. longifolia* (Hybrid index = 0) and *V. spicata* (Hybrid index = 1) ancestry across all 13 admixed individuals.

showing deviation from a model of neutral introgression based on genome-wide admixture ($P \leq 0.025$; significance following FDR correction; **Table 2**).

Crossing and Germination Experiments

We have used 32 individuals during the crossing experiment (11 of *V. spicata* and 21 from *V. longifolia*). Overall, 24 crossings were performed in which 13 were based on the same ploidy and 11 on different ploidy levels (**Table 3**). In almost all cases, the negative controls (n = 20) resulted in the production of zero seeds except in two cases, in which the number of seeds was 18 and 4. In successful crosses, we got usually more than 100 seeds per spike. We considered the crossing as successful only when it led to production of more than 20 seeds per spike, attributing the eventual production of small amounts of seeds to contamination by small arthropods, which were able to reach the spikes through the lower opening of the bag. Crossings performed between plants of the same ploidy level (2× × 2× and 4× × 4×; **Table 3**), even from very distant geographical origin, resulted in production of viable seeds in all cases, with

TABLE 2 | Patterns of introgression based on all admixed individuals.

Patterns of Introgression relative to neutral expectations	No of Loci
Non-neutral	1793
V. longifolia (Parent A) homozygotes to admixed individuals	472
V. spicata (Parent B) homozygotes to admixed individuals	688
Heterozygous excess in admixed individuals	1
Heterozygous deficiency in admixed individuals	1603
Homozygous excess in admixed individuals	1160
Homozygous deficiency in admixed individuals	2
Neutral	6930

The results are based on pure parental individuals of V. longifolia (63 individuals with 0.1 ≥ Q ≥ 0.9), V. spicata (79 individuals with 0.1 ≥ Q ≥ 0.9), and admixed individuals (13 with Q ≥ 0.2) following the NEWHYBRID, and STRUCTURE results. The total number of SNPs used were 8,723 with 25.73% missing data. Prediction of introgression at each locus was compared through likelihood of regression model to that of a neutral model and its significance by using 1,000 permutations. Non-neutral loci were selected by adjusting the neutral expectation of genomic clines through multiple comparisons (P < 0.025; significance following FDR correction).

germination rates ranging from 18 to 100%. However, the crossings between different ploidy levels (2× × 4× and 4× × 2×; **Table 3**), produced almost no seed in crosses with the tetraploid mother, and in four out of seven crosses when the mother was diploid (**Table 3**). Plants of crosses, which produced less than 10 seeds, did not grow to fully developed plants, except for a single plant. This plant was determined to be tetraploid (Albach S372 × Albach S634). All five plants of the cross Albach 1450 and Albach S854 analyzed flow cytometrically were tetraploids. Six of the plants from the cross between Albach 1450 and Albach S485 were analyzed with flow cytometry and all were triploid. The controls between plants of the same species and ploidy level showed a high production of seeds and germination rates (>93%) except for the crossing of diploid *V. spicata*, which could be due to sterility or self-incompatibility reactions.

Polyploidy and Ecological Differentiation

Based on our literature search and our own results, we demonstrate that diploids and tetraploids are about equally common in *V. spicata* (115 vs. 138 measurements, respectively), whereas diploids are less common in *V. longifolia* (49 vs. 102 measurements, respectively). However, these numbers are biased by our sampling and the accessibility of populations.

The ANOVA results for the diploid/tetraploid comparison of eight climatic variables are given in **Table 4** (results for all 19 variables are given in **Supplementary Table 4**). When Siberian diploid samples of *V. spicata* were excluded, mainly the variables dealing somehow with precipitation (e.g., bio17 in **Table 4**) were significantly different between diploids and tetraploids. However, when all *V. spicata* diploids were included, variables somehow dealing with temperature or seasonality (e.g., bio2, 3, 7 in **Table 4**) were significant. This discrepancy is a clear indication and consequence of the distinctly continental climate in Siberia in contrast to the sub-oceanic climate in western and central Europe. The difference between diploids and tetraploids within *V. longifolia* seemed to be less pronounced and was indicated by five significant variables (three after Bonferroni correction). These variables (e.g., bio8 and 10 in **Table 4**) basically refer to different summer temperatures.

The first four axes of the PCA incorporated 52, 23, 11, and 7% of the total variance. The first axis was mainly determined by variables based on seasonality (bio2, bio3, bio7, and bio15) and precipitation (bio9 and bio17), while the second axis was mainly determined by temperature variables (bio8 and bio10). The 41 eastern samples of diploid *V. spicata* (cyan blue triangle in **Figure 6**) were separate from all other *V. spicata* data points in the PCA. However, the western remainder of diploid *V. spicata* was largely overlapping with tetraploid *V. spicata*, and the same was true for most of the diploid and tetraploid individuals of *V. longifolia*, respectively.

The variables that contributed more than 10% to MAXENT models were: for diploid *V. longifolia* bio7 (43%) and bio17 (28%); for tetraploid *V. longifolia* bio8 (48%), bio 17 (21%), and bio7 (11%); for diploid *V. spicata* (all data) bio17 (41%), bio10 (22%), and bio9 (12%); for diploid *V. spicata* west bio17 (63%) and bio3 (21%), for diploid *V. spicata* east bio7 (34%), bio10 (25%), bio17 (20%), and bio8 (16%); for tetraploid *V. spicata* bio7 (36%) and bio17 (34%). The maps produced by the logistic output of MAXENT revealed slight differences between the conspecific diploid and tetraploid groups of each species (**Supplementary Figure 2**). The main difference within *V. longifolia* was that the plot for diploids indicated suitable climate also in northern Europe (Scandinavia and northeast of the Baltic Sea), but less so for the tetraploids. Within *V. spicata*, the areas suitable for tetraploids and for the western diploids are very similar. This changed when eastern individuals were also included, adding a large corridor from eastern Europe to western Asia to the potentially suitable conditions of the diploids.

The results of the quantitative niche overlap analysis (Schoener's *D*) and the niche identity test for all pairwise comparisons indicated significant niche differentiation (**Table 5**). These analyses recovered the least differentiation between *V. longifolia* tetraploid and *V. spicata* diploid and highest difference between *V. spicata* diploid and *V. spicata* tetraploid.

DISCUSSION

In this study, we investigated the phylogeographic structure using genome-wide SNPs considering hybridization, introgression and polyploidization of two congeneric species, *V. spicata* and *V. longifolia* across their vast European distribution ranges. Though *V. longifolia* exhibited lower geographic structure, our results demonstrated clear geographical patterns for both species. Hybridization and introgression through backcrossing and polyploidization was detected in several natural populations across the European distribution of both species, but apparently these cases of hybridization have not lead to large-scale merger of the species, neither such hybridization obscured the species boundaries. Finally, experimental crossings revealed high level of compatibility between the two species in cases in which individuals have the same ploidy levels. We have not recovered any clear geographical or macroclimatic patterns related to polyploidy. Thus, the two species do not form hybrid zones on a

TABLE 3 | Crossing experiment results.

Hybrids (H)										
Ploidy	♂		♀						**ID**	
(♂ × ♀)	**Species**	**Origin**	**Species**	**Origin**	**Seeds obt.**	**Seeds plated**	**Germ.30 d**	**Germ.%**	♂	♀
	V. spicata	Switzerland	*V. longifolia*	Altai Mount.	>100	50	47	94	Albach S526	Albach S625
	V. longifolia	Altai Mount.	*V. spicata*	France	>100	50	50	100	Albach S625	Albach S758
	V. longifolia	Altai Mount.	*V. spicata*	Switzerland	>100	50	48	96	Albach S625	Albach S526
2× × 2×	*V. longifolia*	Finland	*V. spicata*	Romania	>300	50	50	100	Albach S848	Albach S376
	V. longifolia	Finland	*V. spicata*	Hungary	>100	50	46	92	Albach S849	Albach S748
	V. longifolia	Finland	*V. spicata*	Poland	>100	50	9	18	Albach S848	Albach S485
	V. longifolia	Finland	*V. spicata*	Poland	>100	50	13	26	Albach S849	Albach S485
	V. spicata	France	*V. longifolia*	Finland	>50	50	12	24	Albach S758	Albach S849
	V. longifolia	Finland	*V. spicata*	Hungary	0	0	0	0	Albach S848	Albach S869
2× × 4×	*V. spicata*	Altai Mount.	*V. longifolia*	Ukraine	∼10	0	0	0	Albach S634	Boiko 020
	V. spicata	Poland	*V. longifolia*	Czechia	0	0	0	0	Albach S485	Albach S499
	V. longifolia	Yakutia	*V. spicata*	France	4	4	4	100	Albach S372	Albach S758
	V. longifolia	Yakutia	*V. spicata*	Altai Mount.	9	9	6	67	Albach S372	Albach S634
4× × 2×	*V. longifolia*	Ukraine	*V. spicata*	Poland	>50	28	19	68	Albach 1450	Albach S485
	V. longifolia	Ukraine	*V. spicata*	Poland	∼100	50	1	2	Albach 1450	Albach S485
	V. longifolia	Ukraine	*V. spicata*	Altai Mount.	0	0	0	0	Boiko 020	Albach S634
	V. longifolia	Ukraine	*V. spicata*	Hungary	>100	50	49	98	Boiko 020	Albach S869
4× × 4×	*V. spicata*	Hungary	*V. longifolia*	Czechia	>100	50	48	96	Albach S869	Albach S499
	V. spicata	Hungary	*V. longifolia*	Ukraine	>100	50	50	100	Albach S869	Albach 1450
	V. longifolia	Ukraine	*V. spicata*	Hungary	>100	50	40	80	Boiko 020	Albach S869
Positive Control (Ctr)										
2× × 2×	*V. longifolia*	Finland	*V. longifolia*	Finland	∼30	33	31	94	Albach S849	Albach S849
	V. spicata	Austria	*V. spicata*	Austria	>50	50	50	100	Albach S749	Albach S749
4× × 4×	*V. spicata*	Austria	*V. spicata*	Austria	>100	50	50	100	Albach S749	Albach S749
	V. longifolia	Czechia	*V. longifolia*	Czechia	∼37	35	34	97	Albach S499	Albach S499
	V. longifolia	Czechia	*V. longifolia*	Czechia	>50	50	50	100	Albach S499	Albach S499

The number of seeds obtained refers to single floral spikes. Germ. 30d refers to seed germinated 30 days after plating. ID refers to vouchers in herbarium OLD. Hybrids indicate crossings between different species, while positive controls crossings between same species and ploidy level. The IDs assigned to the samples (e.g., sp30aSwi_d1) were constructed using the first two letter to identify the species (sp for V. spicata, and lo for V. longifolia), followed by a unique letter e.g., 30a representing an individual of population 30, and 3 letter code of country of origin (e.g., Swi for Switzerland). An underscore separates the information related with polyploidy e.g., t for tetraploid and d for diploid.

TABLE 4 | One-way ANOVA calculations comparing bioclim variables (Hijmans et al., 2005) of diploid (2) and tetraploid (4) *Veronica longifolia* and of diploid (2) and tetraploid (4) *Veronica spicata*.

Bioclim variable	***V. longifolia* 2/4**			***V. spicata* 2 total/4**			***V. spicata* 2 west/4**	
	Mean 2 (*N* = 40)	**Mean 4 (*N* = 65)**	***p*-value**	**Mean 2 (*N* = 74)**	**Mean 4 (*N* = 69)**	***p*-value**	**Mean 2 (*N* = 33)**	***p*-value**
bio2	8.99	9.58	**0.047**	9.65	8.62	**<0.001***	8.42	0.431
bio3	25.20	25.61	0.629	25.36	29.00	**<0.001***	29.10	0.894
bio7	36.28	38.89	0.185	39.35	29.81	**<0.001***	29.17	0.383
bio8	13.39	16.90	**<0.001***	16.20	16.32	0.861	13.68	**0.006**
bio9	−5.18	−7.18	0.267	−7.10	0.11	**<0.001***	0.93	0.292
bio10	15.27	17.26	**<0.001***	17.72	17.40	0.365	16.92	0.332
bio15	39.48	42.85	0.334	36.48	32.40	**0.031**	27.24	**0.009**
bio17	85.88	77.75	0.258	98.34	105.32	0.406	153.31	**<0.001***

The latter was tested with and without the diploid Siberian samples (2 total and 2 west, respectively). The V. spicata mean 4 column is valid for both comparisons. Only the eight uncorrelated bioclim variables are shown here but the test was done for all 19 variables (see ***Supplementary Table 4****). Significant results ($p < 0.05$) are in bold. Significant results after Bonferroni correction for multiple tests ($p < 0.003$) have an asterisk.*

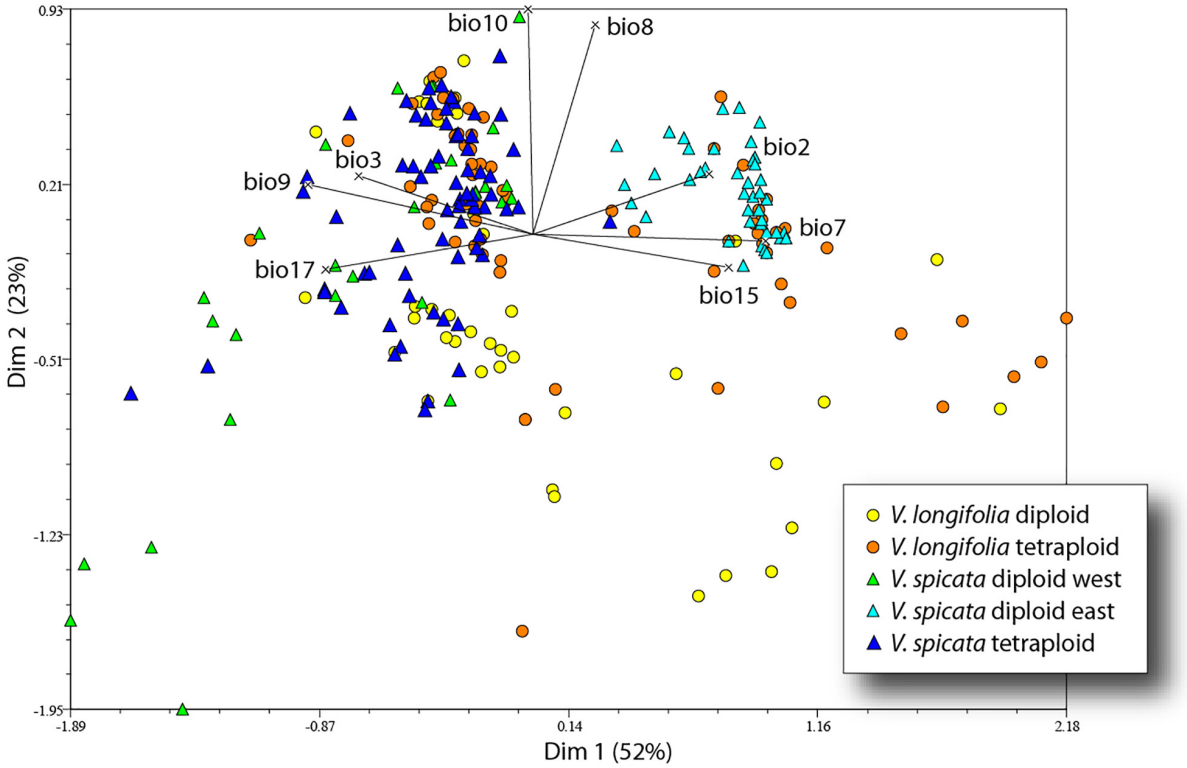

FIGURE 6 | Principal component analysis of *V. longifolia* and *V. spicata* based on 248 data points and eight bioclim variables.

TABLE 5 | The results of the niche overlap (Schoener's *D*) and niche identity tests.

	Schoener's *D*	Niche identity test mean *D* ± *SD*	*p*-value
V. longifolia tetraploid vs. diploid	0.61	0.79 ± 0.04	<0.01
V. spicata tetraploid vs. diploid	0.39	0.84 ± 0.03	<0.01
V. spicata tetraploid vs. diploid (west)	0.66	0.81 ± 0.04	<0.01
V. longifolia tetraploid vs. *V. spicata* diploid	0.69	0.82 ± 0.03	<0.01
V. longifolia diploid vs. *V. spicata* diploid	0.51	0.81 ± 0.04	<0.01
V. longifolia tetraploid vs. *V. spicata* tetraploid	0.40	0.81 ± 0.04	<0.01
V. longifolia diploid vs. *V. spicata* tetraploid	0.41	0.78 ± 0.05	<0.01
V. longifolia diploid vs. *V. spicata* diploid (west)	0.46	0.78 ± 0.06	<0.01
V. longifolia tetraploid vs. *V. spicata* diploid (west)	0.40	0.77 ± 0.04	<0.01

The mean, the standard deviation (SD) of the D value, and the p-values were derived from the distribution of D values from 100× pseudoreplicates.

continental scale but likely have formed hybrid zones of different extent and temporal duration on a local scale ever since the two species split about 1.5 million years ago (Khan et al., submitted).

Phylogeographic Structure of *V. longifolia* and *V. spicata* Across Eurasia

Both the population genetic structure and maximum likelihood tree indicated a basal split between European and Central to East Asian (plus two European) samples of *V. longifolia* after removal of admixed ($Q > 0.2$ for the minor ancestry) individuals. The higher amount of inferred admixture for these individuals (4–20%; **Supplementary Table 1**) is not considered to be real evidence for introgression but rather an artifact of STRUCTURE based on genetic divergence of these samples from the majority of (European) individuals. Currently, we lack evidence for this hypothesis but this is currently being analyzed in more detail. Such a split between European and Central to East Asian populations has previously been found in *Clausia* (Franzke et al., 2004) with a split between groups found around 70° East and further west along the Ural Mountains in *Camelina microcarpa* (Žerdoner Čalasan et al., 2019). This split along the Ural Mountains is in line with the division of periglacial steppes in two large ecosystems east and west of the Ural Mountains based on paleo-environmental reconstructions (Hurka et al., 2019). Considering the postulated East Asian origin of the subgenus (Kosachev et al., 2016), we hypothesize that *V. longifolia* originated in Eastern Asia and subsequently dispersed westward toward Europe and further differentiated there. A split between the European and Asian populations in *V. longifolia* has already been suggested by Trávníček (2000),

who considered the differences between Asian and European plants substantial enough to treat them as two different species. However, given our sparse sampling among Asian populations and the lack of clear morphological differentiation, we currently refrain from adopting this view.

In contrast to *V. longifolia*, the earliest branching samples in *V. spicata* are from the South Russian – North Caucasian region (**Figure 4**). This argues against the East Asian origin but rather indicates early differentiation along the southern border of the extant distribution, though not necessarily an origin within that region. The only Asian samples of that species included here were nested among samples from Ukraine, which suggests that *V. spicata* migrated eastwards to the Altai Mountains either from Southern Russia or from a widespread Eurasian steppe group. The results of the STRUCTURE analysis support the latter view that the samples of *V. spicata* from the Altai Mountains and Eastern Europe belong to a group of widespread Eurasian steppe genotypes and the Southern Russian populations belong to a separate population group that differentiated early in the evolution of the species. As in *V. longifolia*, these results show higher amounts of admixture, which we again attribute more to intraspecific genetic divergence than to real introgression from *V. longifolia* (see above). In this case, it is noteworthy that sample sp1x_Rus1 morphologically belongs to *V. spicata* subsp. *transcaucasica* Bordz., which is morphologically differentiated from typical *V. spicata*. The existence of widespread Eurasian steppe genotypes is somehow counter to the more general assumption that many species during glacial phases of the Pleistocene survived in southern refugia, e.g., the three southern European peninsulas with a subsequent spread across Europe and eastwards to Siberia (Taberlet et al., 1998; Hewitt, 2001). This generalized phylogeographic pattern, being in general true for many species, is, however, not considered a universal key suitable for explaining all phylogeographic patterns of European and Eurasian plant species anymore. Several other patterns for taxa of vascular plants have been proposed as well (e.g., Suchan et al., 2019; Friesen et al., 2020; Volkova et al., 2020, and references therein), such as survival in more northern refugia (even cryptic ones: see Stewart and Lister, 2001), ecological shifts with possible survival in habitats and plant communities not now peculiar to the species, etc. (see a paleobotanical overview in Bezusko et al., 2011). These patterns reflect that each species reacted to the Pleistocene and Holocene environmental and climate changes in an idiosyncratic way reflecting that many characters and ecological adaptations are specific to that species. Phylogeographic patterns of Eurasian steppe plants should also be considered against the background of the general development and transformations of the Eurasian steppe zone (see an overview in Hurka et al., 2019, and references therein), especially during the Pleistocene and Holocene climatic and environmental cycles. Thus, an analysis with better sampling across the Eurasian steppe is necessary to identify refugia of these two species of *Veronica*.

The ML phylogram (**Figure 4**) further suggests that *V. spicata* expanded westwards to Eastern Europe (southern Poland and Ukraine) from the Southern to Central Russian refugium. All individuals west of Ukraine form two geographically distinct groups, one constitutes a Western European (Germany to United Kingdom) and another a Pannonian-Balkanian group. Our TESS3R (**Figure 3A**) results detected five genetic clusters in *V. spicata*, which form geographically delimited groups mostly in congruence with the ML phylogram. These are (I) the Western European group with individuals from United Kingdom, France, Germany, Switzerland, and Western Austria; (II) Pannonian-Balkanian group that includes Serbian, Romanian, Hungarian, Slovenian, Eastern Austrian, and Southern Polish individuals; (3) the South Russian group with an outlying individual from Poland; (4) a widespread Eurasian group composed by Ukrainian populations including the samples from the Russian Altai; and (5) a rather isolated Swiss individual. Consequently, our results agree with previous analyses and support the existence of a western, a southeastern European, a Caucasian, and an Eastern refugium (Stewart and Lister, 2001).

The single diploid individual (sp7d_Swi2) sampled from population Swi2 constitutes a separate Swiss group in the STRUCTURE analysis. STRUCTURE analysis indicates slight admixture for this *V. spicata* individual, while it shows no morphological similarities with *V. longifolia*. Our NewHybrids analysis indicates that sp7d_Swi2 has a partial probability to be a F_2 hybrid or a backcross with *V. longifolia*, explaining the uniqueness of this Swiss population. Another alternative for this uniqueness might be that sp7d_Swi2 originates from a separate refugial area. This is in concordance with Trávníček et al. (2004) who mentioned *V. spicata* diploid relicts in Switzerland and other areas of Europe. During the last Pleistocene glaciations, the Alps were covered with glaciers, and many species found refuge in the present-day territory of Italy, which was then free of ice (Lugon-Moulin and Hausser, 2002). After glaciations, many species recolonized the Alps from the South following different migration pathways (e.g., through the Rhodanian pathway; Parisod, 2008). A phylogeographic analysis of *V. spicata* in the Alps seems necessary to resolve this question. However, our study warns that hybridization and introgression from *V. longifolia* needs to be considered in such an analysis.

Intraspecific differentiation as in *V. spicata* is not as clear in *V. longifolia*. This may partly be due to our sampling, but also to the higher connectivity of suitable habitats (*V. longifolia* is considered mainly a river corridor plant; Winter et al., 2008; **Figure 3B**), which allows high gene flow. Nevertheless, phylogenetic analysis, though hampered by low sampling points in western Asia, indicates that *V. longifolia* spread westward across the Russian steppe, reached first Ukraine (**Figure 4**), from where it seems to have spread to Fennoscandia independently from the rest of Europe. The remaining samples form two clades, one in Western Europe (France, and Germany) and another from Central to Eastern Europe (from Austria to Ukraine).

Effects of Polyploidization

The genetic structure of both species seems to be largely unrelated to ploidy levels, since both geographical groups in *V. longifolia* and the European populations in *V. spicata* contain both ploidy levels with some geographically coherent clusters but no clear pattern. Thus, recurrent polyploidization events across the distribution area best explain the pattern, although

we cannot locate specific areas where tetraploids originated. The Balkan group of *V. spicata* is constituted only by 4× individuals and the East European group only by 2× individuals, while the other groups have both ploidy levels present. Diploid *V. spicata* have been reported from Bosnia (Fischer, 1974), Bulgaria (Graze, 1933), Serbia (Pustahija et al., 2013), Hungary (Löve, 1977), and Romania (**Supplementary Table 2**) but they have not been analyzed genetically and could potentially be misidentified or, in the case of our Romanian sample, could be the westernmost extension of the Ukrainian group. It is intriguing that there are no tetraploid individuals of *V. spicata* known in Asia. This can have at least three different explanations somewhat depending on the unknown frequency of polyploidization. First, it might be that tetraploids have not originated there by chance; second, they might not have migrated there by chance; or, third, they might have originated there with similar frequency as in Europe, but the habitats were not suitable. Unfortunately, the ANOVA analysis is inconclusive (**Table 4**), because significant niche differences within Europe seem mainly due to less precipitation in habitats of tetraploids and higher precipitation in diploids but when looking at the total range of the diploids, the significant differences indicate different temperature regime only. The same pattern can be seen in the PCA (**Figure 6**) in which the diploid and tetraploid *V. spicata* groups were largely overlapping and only the Siberian diploids were separate from all other *V. spicata* points. Also, the analysis of niche overlap (**Table 5**) is not a good or valid indicator for putative differences between ploidy levels in our study. Admittedly, there were significant differences between the ploidy levels. This test statistic, however, seems overly sensitive to unequal geographical sampling and we could not adjust our sample to accommodate for this. An example for this problem was, that the measured niche overlap between diploid and tetraploid *V. spicata* was smaller than between any other comparison, even between the different species. In conclusion, Schoener's *D* test statistic may not be suitable for such a wide and therefore partly incomplete geographical sampling as in our study.

Since tetraploid *V. spicata* have been found mainly in the Pannonian-Balkan region and diploids have rarely been found in the region, it should be investigated in more detail whether the tetraploid *V. spicata* from the region constitute a genetically differentiated infraspecific taxon with higher drought resistance (**Table 4**), a trait reported for other autopolyploids, although this is not a universal pattern (e.g., Yang et al., 2014). In that respect, the connection between the Pannonian-Balkan tetraploids and the other tetraploids needs to be studied further. Are these tetraploids of different origin, which could not spread as much as the Balkan populations? Despite this question, tetraploid *V. spicata* fits the common scenario found in many other European polyploids that were better able to colonize North European areas after glacier retreats (see **Figure 2**; blue vs. green points), possibly by being better adapted to cold conditions (Brochmann et al., 2004; but see Tkach et al., 2008), although this is not directly reflected in the comparison of climatic variables shaping ecological niches (e.g., similar bio10 values in **Table 4**). An alternative explanation for the success of polyploids applicable here is that due to fixed heterozygosity polyploids are able to cope better with inbreeding depression after colonization events in new areas (Birchler, 2012). In contrast, in *V. longifolia* the diploids spread further north, with the tetraploids requiring higher temperature (e.g., bio10 in **Table 4**). Thus, the macroclimatic niche signal we found is weak and partly inconsistent, but if there was any, it seemed that polyploidization had different effects in the two related species. Finally, our genetic data showed that there was and is a recurrent evolution of polyploids in both species, which produces many unrelated and unconnected autopolyploid lineages. Thus, our explorations of climatic niches fit well to the theory, that the *Veronica* polyploids still carry the macroclimatic niche requirements of their individuals and immediate diploid ancestral population rather than a common, "typical" and detectable polyploid niche requirement.

Hybridization

The genetic divergence between *V. spicata* and *V. longifolia* was found to be low (F_{ST} = 0.29), indicating structured populations with connectivity (**Supplementary Table 3**). This suggests that the amount of evolutionary divergence between the parental species is not high enough to cause significant genetic incompatibilities (Moyle and Nakazato, 2010). Further, the possibility for hybridization to occur was revealed in our crossing experiments, which also demonstrated that plants of the two species can form viable seeds independent of their geographic origin, but only when both parents have the same ploidy level (**Table 3**). Third, macroclimatic differences between the two species appear minor (**Table 5**), which is consistent with the coexistence of the two species in close proximity (e.g., Hun6, Ukr13, and Ukr16-20).

Hybridization due to secondary contacts is considered to be a common phenomenon following recolonization after retreat of Pleistocene glaciers (Hewitt, 2001). During these glaciations (glacial phases, in contrast to interglacials), communities of steppe vegetation and its ecological analogs (such as cold and dry periglacial steppes or similar dry grassland plant communities) were favored by predominantly continental climate in most part of Europe, thus possibly providing large and connected open habitats (Mai, 1995; Bezusko et al., 2011; Hurka et al., 2019); such habitats were most probably suitable for *V. spicata* and ensured its survival and dispersal during Pleistocene cycles. On the other hand, interglacial phases of the Pleistocene were normally warmer and more humid in most of Europe and thus promoted such habitats as riparian meadows and floodplain grasslands developing on fluvioglacial deposits, which were more favorable for survival and dispersal of *V. longifolia*. Thus, most probably these two species reacted in different manners to the environmental changes of the Pleistocene and Holocene but shifts in environmental conditions brought them into frequent contact.

Unfortunately, direct paleobotanical (paleopalynological) data on the history of *Veronica* in the Late Quaternary are limited because pollen grains of those entomophilous species are rare in fossil and subfossil pollen samples (see Bezusko et al., 2011, 2018); also, distinguishing species of *Veronica* based solely on their pollen is problematic (Tsymbalyuk and Mosyakin, 2013). However, some paleobotanical (in fact, paleoecological)

evidence on our two species can be obtained indirectly, through reconstructions of the possible distribution of ancient plant communities suitable for these species. In such assumptions we should, however, take into considerations possible habitat shifts of species in the past, as well as the existence of plant communities of the past that were only roughly corresponding to modern plant communities of the region concerned (Bezusko et al., 2011). In any case, such paleoreconstructions should be done in parallel and in agreement with experimental and molecular results, such as those presented here.

Our results indicate that hybridization events can be found between *V. spicata* and *V. longifolia* across the whole distribution range in Europe, with tetraploid and diploid hybrid individuals identified in populations from Ukraine, Slovakia, and Switzerland and likely happened for a long time. However, reliable identification of hybrids can be difficult. Here, we adopted a genetic definition based on results of the admixture analysis in combination with phylogenetic evidence. Individuals with higher admixture levels branched earlier among the respective species. Thus, the genetic divergence appears continuous from intraspecific variation to hybrids (**Supplementary Table 1**), and geographic coherence was added to justify the cut-off.

NEWHYBRIDS results are in line with STRUCTURE (**Figures 1**, **5**). We observed that individuals with high probability to be hybrids or back-crosses with one of the two parental species in the STRUCTURE analysis showed a significant introgression probability proportional to the genotype frequency class expected (Anderson and Thompson, 2002). Overall, 13 individuals had a significant posterior probability to be one of the four hybrid classes (**Figure 5**). Backcrossing was mainly inferred toward *V. longifolia*, with only one exception, suggesting that introgression is asymmetrical from *V. spicata* to *V. longifolia*. Introgression seems to be not uniform inside natural populations, as indicated by the fact that some individuals resulted to be highly introgressed, while other individuals from the same population (e.g., Ukr19) were not admixed. Furthermore, in natural populations hybridization seems to be a phenomenon not restricted only to tetraploid individuals, even though it seems to occur less frequently in diploids (in only one case the identified hybrid was a diploid compared to five cases among tetraploids). Our crossing results indicate that the vitality of diploid and tetraploid hybrid individuals is comparable, while crossings between different ploidy levels produce seeds with a lower seed development, germination success and survivability, suggesting ploidy level as a crossing barrier. In line with this, hybridization in the two species was only observed in the populations in which *V. spicata* and *V. longifolia* had the same ploidy (Ukr17–20) and not in the populations where the two species differed in their ploidy (Hun6, Ukr13, and Ukr16–18). This result is in accordance with Härle (1932), who suggested within-ploidy but not between-ploidy cross-compatibility. Stebbins (1971) already hypothesized that inter-cytotype gene flow is mostly unidirectional from diploids to tetraploids and this has been demonstrated in the last years with molecular tools (e.g., Jørgensen et al., 2011; Zohren et al., 2016). Whereas documented cases of introgression in the wild involving the same ploidy level are frequent, less common examples between different ploidy levels are known. This is due to the fact that triploid hybrids are not produced (triploid block), and if produced, they are highly sterile (Ramsey and Schemske, 1998). In fact, despite thousands of flow cytometric analyses in *Veronica*, uneven ploidy levels have never been reported in the wild (Bardy et al., 2011, **Supplementary Table 1** for *V.* subg. *Pseudolysimachium*, but see also Bardy et al., 2010; Padilla-García et al., 2018; Rojas-Andrés et al., 2020; López-González et al., 2020). However, we produced some triploids in our greenhouse from interploidal crosses (**Table 3**), although we did not yet grow them to maturity to check for fertility. Thus, in some rare cases, triploid individuals may exist in the wild and functional gametes may be produced by triploid hybrids. This can break ploidy barrier and, through backcross with one of the parental species, result in gene transfer (Chapman and Abbott, 2010).

Thus, as suggested by Bardy et al. (2011), hybridization may have generated a great morphological variation by generating several morphologically intermediates between species. Some hybrids seem to have intermediate morphology, while others seem to be more similar to one of the parents morphologically and/or in term of ecological niche. For example, the individual *V.* × *media* (xm1t_Ukr20), classified as F_2 hybrid, showed intermediate morphological characters between *V. spicata* and *V. longifolia*, while the other hybrids detected, even those inferred to be F_2 such as lo37t_Slo2 and lo15t_Ukr2, did not. The predominance of backcrossing to *V. longifolia* combined with the majority of hybrids *a priori* identified as *V. longifolia* suggests that there is selective pressure favoring a *V. longifolia*-phenotype for the hybrids and, thus, backcrossing with *V. longifolia*. However, this selection does not occur up to the seedling stage since we did not observe maternal effects in our crossings (**Table 3**).

CONCLUSION

The genus *Veronica*, with its high species richness and diversification, as well as a great ecological amplitude, constitutes an important model study system which allows investigating several aspects and factors of plant evolution. In the present study, further evidence is provided to support the hypothesis, already proposed for *V.* subg. *Pseudolysimachium* in other regions (Balkans – Bardy et al., 2011; Altai – Kosachev et al., 2019), that hybridization between *V. spicata* and *V. longifolia* is common but these two species remain separate and distinct due to their different ecological adaptations and partial geographical or spatial isolation, with hybrids being probably less competitive than their parents. Furthermore, as shown in the crossing experiment, ploidy can constitute a further reproductive barrier, preventing gene flow between natural populations of different ploidy levels. Our analyses further demonstrate that, despite similar distribution ranges and flower biology, the two species differ markedly in their phylogeographic structure. Dry grasslands appear to have occurred more scattered than riparian habitats with their high connectivity, leading to more elaborate population structure in *V. spicata* as compared to that in *V. longifolia*. Increasing the number of populations sampled and performing sampling in a more geographically homogeneous way across Eurasia may provide interesting and more solid data on the

origins and phylogeographic histories of the two species and their evolutionary interactions. Further ecological studies in different regions would be necessary to evaluate the effective ecological competitiveness of the hybrids compared with their parents and to identify which ecological niches they can occupy. The ecology of the hybrids is totally unknown; it can be intermediate between parents, but also one of the two extremes.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: 10.5061/dryad.kkwh70s3.

AUTHOR CONTRIBUTIONS

DA, PK, SM, and GK conceived the idea. DB and EM-Q performed the experiments. DB, KBH, and GK conducted formal analyses. PK, EM-Q, and DA contributed reagents, materials, and analysis tools. DA, GK, DB, PK, and SM authored the drafts of the article. All authors contributed, made multiple revisions, and approved the final draft.

ACKNOWLEDGMENTS

We acknowledge the support from the VW-Foundation (grant number 90256), state task (Russian Federation) project number FZMW-2020-0003), and the German Research Foundation (DFG) project AL 632/15-1 within the priority program SPP-1991 Taxon-Omics. This work benefited from the sharing of expertise within the DFG priority program SPP 1991. Computations were performed at the HPC Cluster HERO (High End Computing Resources Oldenburg) funded by the DFG through its Major Research Instrumentation Program (INST 184/108-1 FUGG) and the Ministry of Science and Culture (MWK) of the state of Lower Saxony. We are grateful for material from France and Germany by Laurent Hardion and from Austria by Monika Kriechbaum. Finally, we thank Silvia Kempen, Imke Notholt, and Sabrina Schöngart for help in the lab and in the greenhouse and Jan Reimuth for help in the niche modeling.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.588354/full#supplementary-material

Supplementary Figure 1 | Phylogenetic analysis based on maximum likelihood statistics as implemented in RAxML including hybrids. Coloring and labeling follow **Figure 4**. The IDs assigned to the samples (e.g., lo82t_Ukr17) are constructed using the first two letter to identify the species, e.g., lo for *Veronica longifolia*, sp for *V. spicata*, and sc for *V. schmidtiana*, followed by a unique number for every individual and a letter to indicate the ploidy level (*d* for 2×, *t* for 4×, and × for unknown ploidy). Underscore separates the information related with population constituted by 3 letters to identify the country (e.g., CzR for Czechia) and a unique number in that species (see complete detail in **Supplementary Table 1**).

Supplementary Figure 2 | Macroclimatic maps resulting from the logistic output of MAXENT (Phillips et al., 2019). Macroclimatic suitability of regions for the groups analyzed ranges from very high (red) to poor (blue). White squares indicate points used for training the model. **(A)** Diploid *V. longifolia*. **(B)** Tetraploid *V. longifolia*. **(C)** All of diploid *V. spicata*. **(D)** Tetraploid *V. spicata*. **(E)** Only western individuals of diploid *V. spicata*. **(F)** Only eastern individuals of diploid *V. spicata*.

REFERENCES

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141.

Albach, D., and Fischer, M. (2003). AFLP-and genome size analyses: contribution to the taxonomy of *Veronica* subg. *Pseudolysimachium* sect. *Pseudolysimachion* (*Plantaginaceae*), with a key to the European taxa. *Phyt. Balc.* 9, 401–424.

Albach, D. C., Martínez-Ortega, M. M., Delgado, L., Weiss-Schneeweiss, H., Özgökce, F., and Fischer, M. A. (2008). Chromosome numbers in *Veroniceae* (*Plantaginaceae*): review and several new counts. *Ann. Missouri. Bot. Gard.* 95, 543–567. doi: 10.3417/2006094

Albach, D. C., Martínez-Ortega, M. M., Fischer, M. A., and Chase, M. W. (2004). A new classification of the tribe *Veroniceae*-problems and a possible solution. *Taxon* 53, 429–452. doi: 10.2307/4135620

Anderson, E. C., and Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* 160, 1217–1229.

Andrews, S. (2010). *FASTQC, a Quality Control Tool for High Throughput Sequence Data*. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed February 1, 2020).

Arrigo, N., de La Harpe, M., Litsios, G., Zozomová-Lihová, J., Španiel, S., Marhol, K., et al. (2016). Is hybridization driving the evolution of climatic niche in *Alyssum montanum*. *Am. J. Bot.* 103, 1348–1357. doi: 10.3732/ajb.1500368

Avise, J. C. (2000). *Phylogeography: the History and Formation of Species*. Cambridge, MA: Harvard University Press.

Avise, J. C. (2009). Phylogeography: retrospect and prospect. *J. Biogeogr.* 36, 3–15. doi: 10.1111/j.1365-2699.2008.02032.x

Balao, F., Casimiro-Soriguer, R., García-Castaño, J. L., Terrab, A., and Talavera, S. (2015). Big thistle eats the little thistle: does unidirectional introgressive hybridization endanger the conservation of *Onopordum hinojense*? *New Phytol.* 206, 448–458. doi: 10.1111/nph.13156

Bardy, K. E., Albach, D. C., Schneeweiss, G. M., Fischer, M. A., and Schönswetter, P. (2010). Disentangling phylogeography, polyploid evolution and taxonomy of a woodland herb (*Veronica chamaedrys* group, *Plantaginaceae* s.l.) in southeastern Europe. *Mol. Phyl. Evol.* 57, 771–786. doi: 10.1016/j.ympev.2010.06.025

Bardy, K. E., Schönswetter, P., Schneeweiss, G. M., Fischer, M. A., and Albach, D. C. (2011). Extensive gene flow blurs species boundaries among *Veronica barrelieri*, *V. orchidea* and *V. spicata* (*Plantaginaceae*) in southeastern Europe. *Taxon* 60, 108–121. doi: 10.1002/tax.601010

Benjamin, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bezusko, L. G., Mosyakin, S. L., and Bezusko, A. G. (2011). *Patterns and Trends of Development of the Plant Cover of Ukraine in the Late Pleistocene and Holocene*. Kiev: Alterpress.

Bezusko, L. G., Tsymbalyuk, Z., and Mosyakin, S. (2018). Distribution of some representatives of *Scrophulariaceae sensu* lato on the plain part of Ukraine during the Holocene: paleofloristic and palynomorphological aspects. *Visnyk Lviv. Univ. Ser. Biol.* 78, 96–99. doi: 10.30970/vlubs.2018.78.19

Birchler, J. A. (2012). "Genetic consequences of polyploidy in plants," in *Polyploidy and Genome Evolution*, eds P. S. Soltis and D. E. Soltis (Berling: Springer), 21–32. doi: 10.1007/978-3-642-31442-1_2

Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R. A. (2010). Heterosis. *Plant Cell* 22, 2105–2112.

Borsos, O. (1967). Über einige *Rorippa*- und *Veronica*-Arten. *Acta Bot. Acad. Sci. Hung.* 13, 1–10. doi: 10.1007/978-3-319-23534-9_1

Brochmann, C., Brysting, A. K., Alsos, I. G., Borgen, L., Grundt, H. H., Scheen, A. C., et al. (2004). Polyploidy in arctic plants. *Biol. J. Linn. Soc.* 82, 521–536. doi: 10.1111/j.1095-8312.2004.00337.x

Brunsfeld, S. J., Sullivan, J., Soltis, D. E., and Soltis, P. S. (2001). "A comparative phylogeography of northwestern North America: a synthesis," in *Integrating Ecology and Evolution in a Spatial Context*, eds J. Silvertown and J. Antonovics (Oxford: Blackwell Science), 319–340.

Buerkle, C. A. (2005). Maximum-likelihood estimation of a hybrid index based on molecular markers. *Mol. Ecol. Notes* 5, 684–687. doi: 10.1111/j.1471-8286.2005.01011.x

Carvalho, S. B., Torres, J., Tarroso, P., and Velo-Antón, G. (2019). Genes on the edge: a framework to detect genetic diversity imperiled by climate change. *Glob. Chang. Biol.* 25, 4034–4047. doi: 10.1111/gcb.14740

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354

Caye, K., Deist, T. M., Martins, H., Michel, O., and François, O. (2016). TESS3: fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Res.* 16, 540–548. doi: 10.1111/1755-0998.12471

Chapman, M., and Abbott, R. (2010). Introgression of fitness genes across a ploidy barrier. *New Phytol.* 186, 63–71. doi: 10.1111/j.1469-8137.2009.03091.x

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. (2019). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37, 291–294. doi: 10.1093/molbev/msz189

Earl, D. A., and von Holdt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Res.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Ellis, W. N., and Ellis-Adam, A. C. (1994). To make a meadow it takes a clover and a bee: the entomophilous flora of N.W. Europe and its insects. *Bijdragen Dierkunde* 63, 193–220. doi: 10.1163/26660644-06304001

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294x.2005.02553.x

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Fischer, M. (1974). Beitrag zu einer systematischen neubearbeitung der gruppe um *Pseudolysimachion spicatum* (L.) Opiz (*Veronica spicata* L.). *Phyton.* 16, 29–47.

Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PLoS One* 9:e97122. doi: 10.1371/journal.pone.0097122

Franzke, A., Hurka, H., Janssen, D., Neuffer, B., Friesen, N., Markov, M., et al. (2004). Molecular signals for late tertiary/early quaternary range splits of an Eurasian steppe plant: *Clausia aprica* (Brassicaceae). *Mol. Ecol.* 13, 2789–2795. doi: 10.1111/j.1365-294x.2004.02272.x

Friesen, N., Žerdoner Čalasan, A., Neuffer, B., German, D. A., Markov, M., and Hurka, H. (2020). Evolutionary history of the Eurasian steppe plant *Schivereckia podolica* (Brassicaceae) and its close relatives. *Flora* 268:151602. doi: 10.1016/j.flora.2020.151602

Gompert, Z., and Buerkle, C. A. (2009). A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Mol. Ecol.* 18, 1207–1224. doi: 10.1111/j.1365-294x.2009.04098.x

Gompert, Z., and Buerkle, C. A. (2010). INTROGRESS, A software package for mapping components of isolation in hybrids. *Mol. Ecol. Res.* 10, 378–384. doi: 10.1111/j.1755-0998.2009.02733.x

Gompert, Z., and Buerkle, C. A. (2012). BGC; Software for Bayesian estimation of genomic clines. *Mol. Ecol. Res.* 12, 1168–1176.

Graze, H. (1933). Die chromosomalen verhältnisse in der sektion *Pseudolysimachia* Koch der gattung *Veronica*. *Jahrb. Wiss. Bot.* 77, 507–559.

Gruber, B., and Georges, A. (2019). *dartR: Importing and Analysing SNP and Silicodart Data Generated by Genome-Wide Restriction Fragment Analysis. R Package Version 1.1.11*. Available online at: https://CRAN.R-project.org/package=dartR (accessed February 1, 2020).

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., and Li, W. H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66. doi: 10.1038/nature01198

Härle, A. (1932). Die arten und formen der *Veronica* - Sektion *Pseudolysimachia* K OCH auf Grund systematischer und experimenteller Untersuchungen. *Biblioth. Bot.* 104, 1–86.

Harrison, H. B., Berumen, M. L., Saenz-Agudelo, P., Salas, E., Williamson, D. H., and Jones, G. P. (2017). Widespread hybridization and bidirectional introgression in sympatric species of coral reef fish. *Mol. Ecol.* 26, 5692–5704. doi: 10.1111/mec.14279

Harrison, S., and Noss, R. (2017). Endemism hotspots are linked to stable climatic refugia. *Ann. Bot.* 119, 207–214. doi: 10.1093/aob/mcw248

Herben, T., Suda, J., and Klimešová, J. (2017). Polyploid species rely on vegetative reproduction more than diploids: a re-examination of the old hypothesis. *Ann. Bot.* 120, 341–349. doi: 10.1093/aob/mcx009

Hewitt, G. M. (2000). The genetic legacy of the Quaternary ice ages. *Nature* 405, 907–913. doi: 10.1038/35016000

Hewitt, G. M. (2001). Speciation, hybrid zones and phylogeography-or seeing genes in space and time. *Mol. Ecol.* 10, 537–549. doi: 10.1046/j.1365-294x.2001.01202.x

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978. doi: 10.1002/joc.1276

Hijmans, R. J., Guarino, L., and Mathur, P. (2012). *DIVA-GIS Version 7.3.0*.

Hijmans, R. J., Williams, E., and Vennes, C. (2019). *CRAN. R-Project.org/Package=Geosphere, v. 1.5-10*.

Hurka, H., Friesen, N., Bernhardt, K. G. G., Neuffer, B., Smirnov, S. V., Shmakov, A. I., et al. (2019). The Eurasian steppe belt. Status quo, origin and evolutionary history. *Turczaninowia* 22, 5–71. doi: 10.14258/turczaninowia.22.3.1

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Jørgensen, M. H., Ehrich, D., Schmickl, R., Koch, M. A., and Brysting, A. K. (2011). Interspecific and interploidal gene flow in Central European *Arabidopsis* (Brassicaceae). *BMC Evol. Biol.* 11:346. doi: 10.1186/1471-2148-11-346

Kadereit, J. W. (2015). The geography of hybrid speciation in plants. *Taxon* 64, 673–687. doi: 10.12705/644.1

Kampny, C. M. (1995). Pollination and flower diversity in Scrophulariaceae. *Bot. Rev.* 61, 350–366. doi: 10.1007/bf02912622

Keppel, G., Van Niel, K. P., Wardell-Johnson, G. W., Yates, C. J., Byrne, M., Mucina, L., et al. (2012). Refugia: identifying and understanding safe havens for biodiversity under climate change. *Glob. Ecol. Biogeogr.* 21, 393–404. doi: 10.1111/j.1466-8238.2011.00686.x

Khan, G., Franco, F. F., Silvaa, G. A. R., Bombonato, J. R., Machadoc, M., Alonso, D. P., et al. (2020). Maintaining genetic integrity with high promiscuity: frequent hybridization with low introgression in multiple hybrid zones of *Melocactus* (*Cactaceae*). *Mol. Phyl. Evol.* 142:106642. doi: 10.1016/j.ympev.2019.106642

Knaus, B. J., and Grünwald, N. J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Res.* 17, 44–53. doi: 10.1111/1755-0998.12549

Knuth, P. (1909). *Handbook of Flower Pollination*, Vol. 3. Oxford: Clarendon Press.

Kosachev, P. A., Albach, D., and Ebel, A. L. (2015). Check-list of *Veronica* subg. *Pseudolysimachium* (*Plantaginaceae*) of siberia. *Turczaninowia* 18, 84–95. doi: 10.14258/turczaninowia.18.3.4

Kosachev, P. A., Behçet, L., Mayland-Quellhorst, E., and Albach, D. C. (2016). Analyzing reticulate relationships using cpDNA and pyrosequenced ITS1 as exemplified by *Veronica subgen. Pseudolysimachium* (*Plantaginaceae*). *Syst. Bot.* 41, 105–119. doi: 10.1600/036364416x690697

Kosachev, P. A., Mayland-Quellhorst, E., and Albach, D. C. (2019). Hybridization among species of *Veronica* subg. *Pseudolysimachium* in the Altai detected by SRAP markers. *Nord. J. Bot.* 37, e02209.

López-González, N., Bobo-Pinilla, J., Padilla-García, N., Loureiro, J. C. M., Castro, S., Rojas-Andrés, B. M., et al. (2020). Genetic similarities versus morphological resemblance: unraveling a polyploid complex in a Mediterranean biodiversity hotspot. *Molec. Phylogenet. Evol*. doi: 10.1016/j.ympev.2020.107006. [Epub ahead of print].

Löve, A. (1977). IOPB chromosome number reports LVI. *Taxon* 26, 257–274. doi: 10.1002/j.1996-8175.1977.tb04167.x

Lugon-Moulin, N., and Hausser, J. (2002). Phylogeographical structure, postglacial recolonization and barriers to gene flow in the distinctive Valais chromosome race of the common shrew (*Sorex araneus*). *Mol. Ecol.* 11, 785–794. doi: 10.1046/j.1365-294x.2002.01469.x

Mai, D. H. (1995). *Tertiäre Vegetationsgeschichte Europas: Methoden und Ergebnisse*. Heidelberg: Spektrum Akademischer Verlag.

McKinney, G. J., Waples, R. K., Seeb, L. W., and Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol. Ecol. Res.* 17, 656–669. doi: 10.1111/1755-0998.12613

Merow, C., Smith, J. A., and Silander, J. R. (2013). A practical guide to MAXENT for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* 36, 1058–1069. doi: 10.1111/j.1600-0587.2013.07872.x

Meudt, H. M., Rojas-Andrés, B. M., Prebble, J. M., Low, E., Garnock-Jones, P. J., and Albach, D. C. (2015). Is genome downsizing associated with diversification in polyploid lineages of *Veronica*? *Bot. J. Linn. Soc.* 178, 243–266. doi: 10.1111/boj.12276

Moyle, L. C., and Nakazato, T. (2010). Hybrid incompatibility "snowballs" between *Solanum species*. *Science* 329, 1521–1523. doi: 10.1126/science.1193063

Muellner-Riehl, A. N. (2019). Mountains as evolutionary arenas: patterns, emerging approaches, paradigm shifts, and their implications for plant phylogeographic research in the Tibeto-Himalayan Region. *Front. Plant Sci.* 10:195. doi: 10.3389/fpls.2019.00195

Müller, H. (1881). *Alpenblumen, ihre Befruchtung Durch Insekten und ihre Anpassungen an dieselben*. Leipzig: Wilhelm Engelmann.

Nieto Feliner, G. (2014). Patterns and processes in plant phylogeography in the Mediterranean Basin. A review. *Perspect. Plant Ecol. Evol. Syst.* 16, 265–278. doi: 10.1016/j.ppees.2014.07.002

Nolte, A. W., Gompert, Z., and Buerkle, C. A. (2009). Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. *Mol. Ecol.* 18, 2615–2627. doi: 10.1111/j.1365-294x.2009.04208.x

Padilla-García, N., Rojas-Andrés, B. M., López-González, N., Castro, M., Castro, S., Loureiro, J., et al. (2018). The challenge of species delimitation in the diploid-polyploid complex *Veronica* subsection *Pentasepalae*. *Mol. Phyl. Evol.* 119, 196–209. doi: 10.1016/j.ympev.2017.11.007

Parisod, C. (2008). Postglacial recolonisation of plants in the western Alps of Switzerland. *Bot. Helv.* 118, 1–12. doi: 10.1007/s00035-008-0825-3

Paule, J., Wagner, N. D., Weising, K., and Zizka, G. (2017). Ecological range shift in the polyploid members of the South American genus *Fosterella* (Bromeliaceae). *Ann. Botany* 120, 233–243. doi: 10.1093/aob/mcw245

Pembleton, L. W., Cogan, N. O. I., and Forster, J. W. (2013). StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Mol. Ecol. Res.* 13, 946–952. doi: 10.1111/1755-0998.12129

Phillips, S. J., Anderson, R., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecol. Model.* 190, 231–259. doi: 10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., and Schapire, R. E. (2019). *Maxent Software for Modeling Species Niches and Distributions (Version 3.4.0)*. Available online at: http://biodiversityinformatics.amnh.org/open_source/maxent/ (accessed February 1, 2020).

Pickens, K. A., Affolter, J. M., Wetzstein, H. Y., and Wolf, J. H. (2003). Enhanced seed germination and seedling growth of *Tillandsia eizii* in vitro. *Hort. Sci.* 38, 101–104. doi: 10.21273/hortsci.38.1.101

Pigott, C. D., and Walters, S. M. (1954). On the interpretation of the discontinuous distributions shown by certain british species of open habitats. *J. Ecol.* 42, 95–116. doi: 10.2307/2256981

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., and Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Front. Genet.* 29:98. doi: 10.3389/fpls.2019.00098

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Pustahija, F., Brown, S. C., Bogunić, F., Bašić, N., Muratović, E., Ollier, S., et al. (2013). Small genomes dominate in plants growing on serpentine soils in West Balkans, an exhaustive study of 8 habitats covering 308 taxa. *Plant Soil* 373, 427–453. doi: 10.1007/s11104-013-1794-x

R Core Team (2019). *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raitanen, P. R. (1967). Taxonomic studies on *Veronica longifolia* and *V. spicata* in Eastern *Fennoscandia*. *Ann. Bot. Fenn.* 4, 471–485.

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Evol. Syst.* 29, 467–501.

Rochette, N. C., Rivera-Colón, A. G., and Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADSeq-based population genomics. *Mol. Ecol.* 28, 4737–4754. doi: 10.1111/mec.15253

Rohlf, F. J. (2009). *NTSYSpc: Numerical Taxonomy System. ver. 2.21c*. Setauket, NY: Exeter Software.

Rojas-Andrés, B. M., Padilla-García, N., Pedro, M. D., López-González, N., Delgado, L., Albach, D. C., et al. (2020). Environmental differences are correlated with the distribution pattern of cytotypes in *Veronica* subsection pentasepalae at a broad scale. *Ann. Bot.* 125, 471–484.

Scalone, R., Kolf, M., and Albach, D. C. (2013). Mating system variation in veronica (Plantaginaceae): inferences from pollen/ovule ratios and other reproductive traits. *Nord. J. Bot.* 31, 372–384. doi: 10.1111/j.1756-1051.2012.01706.x

Schönswetter, P., Lachmayer, M., Lettner, C., Prehsler, D., Rechnitzer, S., Sonnleitner, M., et al. (2007). Sympatric diploid and hexaploid cytotypes of *Senecio carniolicus* (Asteraceae) in the Eastern Alps are separated along an altitudinal gradient. *J. Pl. Res.* 120, 721–725. doi: 10.1007/s10265-007-0108-x

Siadjeu, C., Mayland-Quellhorst, E., and Albach, D. C. (2018). Genetic diversity and population structure of trifoliate yam (*Dioscorea dumetorum* Kunth) in Cameroon revealed by genotyping-by-sequencing (GBS). *BMC Plant Biol.* 18:359. doi: 10.1186/s12870-018-1593-x

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

Stebbins, G. L. (1971). *Chromosomal Evolution in Higher Plants*. London: Edward Arnold.

Stewart, J. R., and Lister, A. M. (2001). Cryptic northern refugia and the origins of the modern biota. *Trends Ecol. Evol.* 16, 608–613. doi: 10.1016/s0169-5347(01)02338-2

Suchan, T., Malicki, M., and Ronikier, M. (2019). Relict populations and Central European glacial refugia: the case of *Rhododendron ferrugineum* (Ericaceae). *J. Biogeogr.* 46, 392–404. doi: 10.1111/jbi.13512

Taberlet, P., Fumagalli, L., Wust-Saucy, A. G., and Cosson, J. F. (1998). Comparative phylogeography and postglacial colonization routes in Europe. *Mol. Ecol.* 7, 453–464. doi: 10.1046/j.1365-294x.1998.00289.x

Tkach, N. V., Hoffmann, M. H., Röser, M., Korobkov, A. A., and von Hagen, K. B. (2008). Parallel evolutionary patterns in multiple lineages of Arctic *Artemisia* L. (Asteraceae). *Evolution* 62, 184–198.

Trávníček, B. (2000). Notes on the taxonomy of *Pseudolysimachion longifolium* complex (Scrophulariaceae). *Thaiszia* 10, 1–26. doi: 10.1007/978-3-319-23534-9_1

Trávníček, B., Lysák, M. A., Číhalíková, J., and Doležel, J. (2004). Karyo-taxonomic study of the genus *Pseudolysimachion* (Scrophulariaceae) in the Czech Republic and Slovakia. *Folia Geobot. Phytotax.* 39, 173–203. doi: 10.1007/bf02805245

Tsymbalyuk, Z. M., and Mosyakin, S. L. (2013). *Atlas of Pollen Grains of Representatives of Plantaginaceae and Scrophulariaceae*. Kyiv: Nash Format Publisher.

Volkova, P. A., Burlakov, Y. A., and Schanzer, I. A. (2020). Genetic variability of *Prunus padus* (Rosaceae) elaborates "a new Eurasian phylogeographical paradigm". *Plant Syst. Evol.* 306:1. doi: 10.1007/978-3-319-77088-8_110-2

Warren, D. L., Glor, R. E., and Turelli, M. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62, 2868–2883. doi: 10.1111/j.1558-5646.2008.00482.x

Warren, D. L., Glor, R. E., and Turelli, M. (2010). ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography* 33, 607–611.

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.2307/2408641

Wilson, G. B., Houston, L., Whittington, W. J., and Humphries, R. N. (2000). *Veronica spicata* L. ssp. *spicata* and ssp. *hybrida* (L.) Gaudin

(*Pseudolysimachium spicatum* (L.) Opiz). *J. Ecol.* 88, 890–909. doi: 10.1046/j.1365-2745.2000.00501.x

Winter, C., Lehmann, S., and Diekmann, M. (2008). Determinants of reproductive success: a comparative study of five endangered river corridor plants in fragmented habitats. *Biol. Conserv.* 141, 1095–1104. doi: 10.1016/j.biocon.2008.02.002

Wright, S. (1949). The genetical structure of populations. *Ann. Eugen.* 15, 323–354. doi: 10.1111/j.1469-1809.1949.tb02451.x

Wringe, B. F., Stanley, R. R., Jeffery, N. W., Anderson, E. C., and Bradbury, I. R. (2016). *Parallelnewhybrid: An R Package for the Parallelization of Hybrid Detection Using NEWHYBRIDS. R Package Version 0.0.0.9002*. Available online at: https://github.com/bwringe/parallelnewhybrid (accessed February 1, 2020).

Wringe, B. F., Stanley, R. R., Jeffery, N. W., Anderson, E. C., and Bradbury, I. R. (2017). parallelnewhybrid: an R package for the parallelization of hybrid detection using new hybrids. *Mol. Ecol. Res.* 17, 91–95. doi: 10.1111/1755-0998.12597

Yang, P. M., Huang, Q. C., Qin, G. Y., Zhao, S. P., and Zhou, J. G. (2014). Different drought-stress responses in photosynthesis and reactive oxygen metabolism between autotetraploid and diploid rice. *Photosynthetica* 52, 193–202. doi: 10.1007/s11099-014-0020-2

Žerdoner Čalasan, A., Seregin, A. P., Hurka, H., Hofford, N. P., and Neuffer, B. (2019). The Eurasian steppe belt in time and space: phylogeny and historical biogeography of the false flax (*Camelina* Crantz, Camelineae, Brassicaceae). *Flora* 260:151477. doi: 10.1016/j.flora.2019.151477

Zohren, J., Wang, N., Kardailsky, I., Borrell, J. S., Joecker, A., Nichols, R. A., et al. (2016). Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Mol. Ecol.* 25, 2413–2426. doi: 10.1111/mec.13644

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with one of the authors SM.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 04 March 2021
doi: 10.3389/fpls.2021.637214

Polyploidy on Islands: Its Emergence and Importance for Diversification

Heidi M. Meudt[1], Dirk C. Albach[2], Andrew J. Tanentzap[3], Javier Igea[3], Sophie C. Newmarch[4], Angela J. Brandt[5], William G. Lee[5] and Jennifer A. Tate[4]**

[1]Museum of New Zealand Te Papa Tongarewa, Wellington, New Zealand, [2]Institute of Biology and Environmental Sciences, University of Oldenburg, Oldenburg, Germany, [3]Ecosystems and Global Change Group, Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom, [4]School of Fundamental Sciences, Massey University, Palmerston North, New Zealand, [5]Manaaki Whenua – Landcare Research, Dunedin, New Zealand

OPEN ACCESS

Edited by:
Natascha D. Wagner, University of Göttingen, Germany

Reviewed by:
Daniel Crawford, University of Kansas, United States
Marc S. Appelhans, Georg-August-University Goettingen, Germany

***Correspondence:**
Heidi M. Meudt
heidim@tepapa.govt.nz
Jennifer A. Tate
j.tate@massey.ac.nz

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: 03 December 2020
Accepted: 11 February 2021
Published: 04 March 2021

Citation:
Meudt HM, Albach DC, Tanentzap AJ, Igea J, Newmarch SC, Brandt AJ, Lee WG and Tate JA (2021) Polyploidy on Islands: Its Emergence and Importance for Diversification. Front. Plant Sci. 12:637214. doi: 10.3389/fpls.2021.637214

Whole genome duplication or polyploidy is widespread among floras globally, but traditionally has been thought to have played a minor role in the evolution of island biodiversity, based on the low proportion of polyploid taxa present. We investigate five island systems (Juan Fernández, Galápagos, Canary Islands, Hawaiian Islands, and New Zealand) to test whether polyploidy (i) enhances or hinders diversification on islands and (ii) is an intrinsic feature of a lineage or an attribute that emerges in island environments. These island systems are diverse in their origins, geographic and latitudinal distributions, levels of plant species endemism (37% in the Galapagos to 88% in the Hawaiian Islands), and ploidy levels, and taken together are representative of islands more generally. We compiled data for vascular plants and summarized information for each genus on each island system, including the total number of species (native and endemic), generic endemicity, chromosome numbers, genome size, and ploidy levels. Dated phylogenies were used to infer lineage age, number of colonization events, and change in ploidy level relative to the non-island sister lineage. Using phylogenetic path analysis, we then tested how the diversification of endemic lineages varied with the direct and indirect effects of polyploidy (presence of polyploidy, time on island, polyploidization near colonization, colonizer pool size) and other lineage traits not associated with polyploidy (time on island, colonizer pool size, repeat colonization). Diploid and tetraploid were the most common ploidy levels across all islands, with the highest ploidy levels (>8*x*) recorded for the Canary Islands (12*x*) and New Zealand (20*x*). Overall, we found that endemic diversification of our focal island floras was shaped by polyploidy in many cases and certainly others still to be detected considering the lack of data in many lineages. Polyploid speciation on the islands was enhanced by a larger source of potential congeneric colonists and a change in ploidy level compared to overseas sister taxa.

Keywords: colonization, diversification, endemism, island floras, ploidy level, phylogenetic path analysis, polyploidy, whole genome duplication

INTRODUCTION

Since the time of Darwin (1839), the study of island biotas has provided major advances that have profoundly influenced ecological, evolutionary, and biogeographical theory (MacArthur and Wilson, 1967; Mayr, 1967; Carlquist, 1974; Emerson, 2002). Island biotas are generally the net outcome of immigration (dispersal and establishment), local diversification, and extinction (Carlquist, 1974), and these processes are known to be influenced by specifics of the island, such as age, area, distance from nearest potential source floras (Rosenzweig, 1995), and local habitat (abiotic and biotic) conditions (Savolainen et al., 2006; Vamosi et al., 2018). Islands may also pose selective filters that may be apparent in the intrinsic traits of species, including those increasing their ability to disperse, colonize, and establish in novel habitats (Crawford et al., 2009; Vargas et al., 2015). Among those traits, whole genome duplication or polyploidy has been suggested to be central to facilitating long-distance dispersal (Linder and Barker, 2014), the survival of small populations (Rodriguez, 1996), and evolution of novel traits (Soltis et al., 2014), all features of many island floras. Polyploids are broadly categorized as autopolyploids, when formed within a species, or allopolyploids, when formed between genetically distinct species (Stebbins, 1947). Polyploidy, especially allopolyploidy, may also be a mechanism for increasing the genetic diversity of the colonizer (Crawford and Stuessy, 1997; Carr, 1998), which is often low for island colonizers and lineages (e.g., Frankham, 1997; but see García-Verdugo et al., 2015). Thus, polyploidy could have multiple advantages, particularly in island floras.

Despite its potential benefits, polyploidy has been historically suggested to play a minor role in diversification of island floras, with many groups showing "chromosomal stasis" (Carr, 1998; Stuessy and Crawford, 1998). For the oceanic island systems that inform this perspective (Hawaiian, Juan Fernández, Galápagos, and Bonin Islands), paleopolyploidy was suggested to have helped some lineages establish, with little polyploidization thereafter (Carr, 1998; Stuessy and Crawford, 1998), such as in *Gossypium* (Malvaceae; **Figure 1**). Thus, while polyploid, these oceanic island lineages remained chromosomally static. By contrast, chromosomal variation was found in two island systems near their continental source, the Queen Charlotte Islands and Canary Islands (Stuessy and Crawford, 1998). Although Crawford et al. (2009) began to address the question of the evolution of polyploidy in island systems and its role in diversification (for Asteraceae; Crawford et al., 2009), others discounted its impact (Stuessy et al., 2014; Crawford and Archibald, 2017). However, in those studies (Carr, 1998; Stuessy and Crawford, 1998; Crawford et al., 2009), chromosome counts from only a small percentage of native island species were available, so interpretations were based on island origin (continental vs. oceanic) and island age rather than polyploidy itself. Other island systems, in particular New Zealand, have not been included in these larger comparative studies, despite chromosome numbers being widely available for many native plant species (see Dawson, 2000, 2008). Furthermore, a phylogenetic context was lacking in previous studies, which is important because the phylogeny will indicate the number of origins on the island as well as patterns of diversification on the island related to polyploidy or other factors.

As robust (and increasingly, dated) phylogenetic hypotheses have amassed over the last several decades, along with continued efforts to document chromosome numbers, we now have the capability to test the role of polyploidy in contributing to species diversification on islands in a phylogenetic context (Kellogg, 2016; Crawford and Archibald, 2017). As species in many different island floras are known to have diverse chromosome numbers and form species-rich groups, they may not be chromosomally static lineages as previously thought (Soltis et al., 2009; New Zealand: Murray and de Lange, 2011; Canary Islands: Caujapé-Castells et al., 2017). A recent analysis of the global distribution of polyploids indicated polyploid frequency is highest in temperate areas rich in perennial herbs, including mountainous areas (Rice et al., 2019). Analyzed from a global perspective, several island systems are polyploid-rich, including the Hawaiian Islands (50% of analyzed species are polyploid), New Zealand (46%), and Galápagos Islands (46%), whereas others have lower polyploid frequencies – such as the Canary Islands (32%) – or insufficient data (Juan Fernández: 0 of 2 species included were polyploid; Rice et al., 2019). However, that study was focused primarily on assessing external drivers of polyploid plant distribution globally, not on island systems, and did not utilize dated phylogenies. A dated phylogenetic context allows a more precise determination of when diversification and polyploidization have occurred within each lineage (i.e., as separate events), and puts the focus on (multiple) lineage ages rather than a single island age. Such an approach can generate multiple independent data points that can be analyzed together to address the timing and role of polyploidy on islands. Dated phylogenies are especially important in the context of island diversification because they make it possible to estimate *in situ* diversification rates and to consider in the analyses the geological processes (e.g., volcano eruptions, inundations) that may affect lineages at different evolutionary time scales.

We hypothesize that polyploidy has played an important role in the diversification of island floras by facilitating dispersal and establishment of plants to islands, and/or by generating additional diversity through varying ploidy levels. To test a conceptual model for how polyploidy influences species diversification on islands, we synthesized published chromosome and divergence time data for 150 lineages representing 1,805 endemic species across the Juan Fernández, Canary, Hawaiian, and New Zealand archipelagos (**Figure 1**). All these island systems, except New Zealand, have been included in previous comparative studies or reviews of polyploidy and diversification (see above), but without the time-calibrated phylogenetic context we add here. By using phylogenetic path analysis, we simultaneously tested the strength and direction of causal associations to explain the diversity of endemic island species in these four archipelagos. We predicted that island lineages would be more diverse (i.e., have more endemic species) if they (**Figure 2**): (P1) contained multiple ploidy levels; (P2)

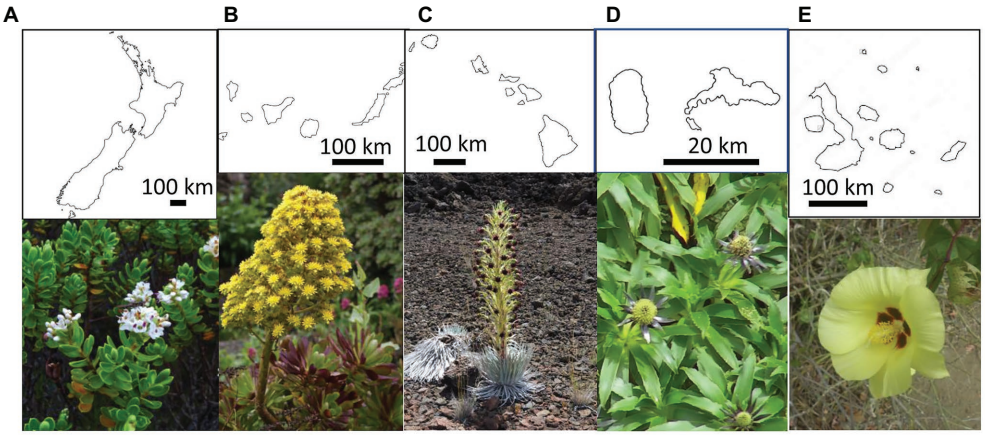

FIGURE 1 | Outline and representative polyploid plants from the five analyzed island systems. From left to right: **(A)** New Zealand: *Veronica topiaria* (Plantaginaceae) © Phil Garnock-Jones; **(B)** Canary Islands: *Aeonium arboreum* (Crassulaceae) © Dirk Albach; **(C)** Hawaiian Islands: *Argyroxiphium sandwicense* (Asteraceae) © Marc Appelhans; **(D)** Juan Fernández Islands: *Eryngium bupleuroides* (Apiaceae) © Lukas Mekis; **(E)** Galápagos Islands: *Gossypium darwinii* (Malvaceae) © A. Emmerson.

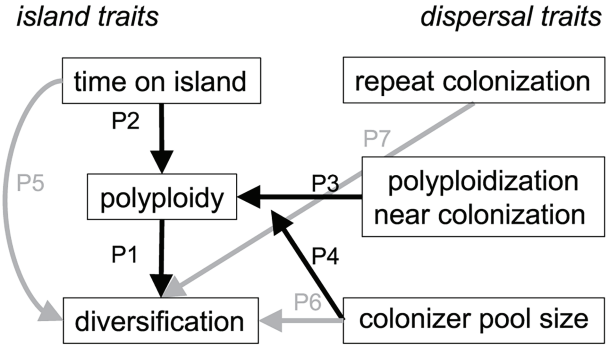

FIGURE 2 | Conceptual model for how species diversity in island lineages varies with polyploidy and extrinsic colonization history. Arrows show predicted linkages between direct and indirect effects of polyploidy (P1–P4, shown in black) and non-polyploidy effects (P5–P7, shown in gray) on diversification of island floras. Left-hand of the plot shows traits intrinsic to island lineages whereas the right-hand shows traits associated with island dispersal and colonization.

had more time to generate ploidy levels once on these islands, as indicated by the length of time they were present on an island; (P3) had a different (i.e., higher) ploidy level relative to their sister lineage, indicating polyploidization immediately before or after colonization of the island; (P4) were derived from a large pool of overseas congeners and thus the likelihood of polyploidization events that could enhance diversity was greater. We also tested other direct explanations that did not involve ploidy and predicted island lineages would be more diverse if they (**Figure 2**); (P5) were older, because they have simply had more time to undergo speciation and isolate populations across a greater availability of niches (e.g., Lee et al., 2012; Tanentzap et al., 2015); (P6) were derived from a large pool of overseas congeners and thus the probability of colonization from that pool would be higher; and (P7) repeatedly colonized islands with different ancestral species, indicated by a lack of monophyly. Hypotheses P1–P4 test effects of ploidy that are both direct (P1) or indirect (P2–P4) and mediated by time (P2), changes to the ploidy levels themselves (P3), or source pool size (P4). Alternatively, hypotheses P1, P2, and P5 test polyploidy as a trait important in diversification, whereas P3, P4, P6, and P7 test polyploidy as a trait important in dispersal or establishment.

MATERIALS AND METHODS

Data Collection

We compiled data of indigenous species, subspecies, varieties, and forms (hereafter: species), and noted whether they were native or endemic, for vascular plants on five island systems using published references for the Canary Islands (CI; Arechavaleta et al., 2010), Galápagos Islands (GI: Jaramillo Díaz et al., 2017, 2018), Hawaiian Islands (HI: Wagner et al., 2005 and electronic updates; Imada, 2012), Juan Fernández Islands (JF: Table 5.1 in Stuessy et al., 2017), and New Zealand (NZ: Breitwieser et al., 2012; Schönberger et al., 2019). We then added data for each species (where known and when from a sample from that island system) on chromosome number (Index to Plant Chromosome Numbers,[1] Chromosome Counts Database,[2] and additional literature search), genome size (Pellicer and Leitch, 2020[3]) and ploidy level. We summarized this information for each genus on each island system, including the total number of indigenous and endemic species, whether the genus was endemic to the island system, chromosome number, genome size, and ploidy level (**Table 1**). For each genus with at least one native species on a particular island group, we identified

[1]http://legacy.tropicos.org/Project/IPCN
[2]http://ccdb.tau.ac.il/
[3]https://cvalues.science.kew.org/

TABLE 1 | Summary of island systems for all genera and species[1] of native vascular plants.

	Archipelago				
Variable	**New Zealand[a]**	**Canary Islands[b]**	**Hawaiian Islands[c]**	**Juan Fernández[d]**	**Galápagos[e]**
No. of islands	3 main islands plus ~600 smaller islands/islets	7 main islands plus 6 smaller islands/islets	8 main islands, plus 129 smaller islands/islets	3 islands	18 main islands, plus 110 smaller islands/islets/rocks
Area of islands	268,021 km^2	7,493 km^2	28,311 km^2	100.2 km^2	7,985 km^2
Shortest distance from nearest continental neighbor	1,490 km E of Australia	96 km W of Morocco	3,500 km SW of United States	580 km W of Chile	930 km W of mainland Ecuador
Main reference for data in rows below	Schönberger et al. (2019)[2]	Arechavaleta et al. (2010)	Wagner et al. (2005 and electronic updates); Imada (2012)	Stuessy et al. (2017)	Jaramillo Díaz et al. (2017, 2018)
No. native species	2,551	1,677[3]	1,233	209	550
No. endemic species (%)	2,096 (82%)	647 (39%)	1,082 (88%)	136 (65%)	204 (37%)
No. species with chromosome numbers (%)	1,962 (77%)	1,171 (70%)	414 (34%)	87 (42%)	39 (7%)
No. species with genome size estimates (%)	245 (12.5%)	237 (14%)	29 (2%)	0	0
No. native genera	430	466	272	103	277
No. endemic genera (%)	48 (11%)[4]	22 (5%)	30 (11%)	12 (11%)	7 (3%)
No. genera with multiple ploidies on island (% of total/% of genera with at least two native species)	88 (21%/35%)	96 (21%/40%)	27 (10%/19%)	5 (5%/13%)	3 (1%/3%)
Range of ploidy levels in one genus	1–6	1–4	1–2	1–2	1–2
Most common ploidy levels (range)	2*x*, 4*x*, 6*x* (2*x*–20*x*)	2*x*, 4*x* (2*x*–12*x*)	2*x*, 4*x* (4*x*–6*x*)	2*x*, 4*x* (2*x*–8*x*)	2*x*, 4*x* (2*x*–4*x*)
No. genera with two or more native species on island (%)	256 (60%)	239 (51%)	139 (51%)	39 (38%)	105 (38%)
No. genera[5] for which a molecular phylogeny is available (% of native genera)	205 (48%)	170 (36%)	86 (32%)	31 (30%)	71 (26%)
No. genera[5] for which a dated molecular phylogeny is available (% of native genera)	112 (26%)	41 (9%)	27 (10%)	8 (8%)	4 (1%)
No. total genera (%) / no. genera[5] (%) for which there is at least one known chromosome number[6]	389 (92%)/250 (98%)	406 (87%)/219 (92%)	155 (57%)/92 (67%)	58 (56%)/25 (64%)	21 (8%)/13 (12%)
No. species per genus range/ average/median	1–143/6.0/2	1–33/3.6/2	1–80/4.5/2	1–12/2.0/1	1–19/2.0/1
No. gymnosperm species (% endemic)	22 (100%)	6 (33%)	0	0	0
No. gymnosperm genera (% endemic)	10 (30%)	3 (0%)	0	0	0
No. fern species (% endemic)	210 (46%)	50 (6%)	167 (75%)	57 (46%)	125 (9%)
No. fern genera (% endemic)	57 (4%)	23 (0%)	51 (4%)	24 (2%)	52 (0%)

Island geographic data from:
[a]*McGlone et al. (2001).*
[b]*Carracedo and Troll (2016).*
[c]*Price (2004).*
[d]*Stuessy et al. (1998).*
[e]*Rivas-Torres et al. (2018).*
[1]*Includes species, subspecies, varieties, and forms; collectively called "species" for simplicity.*
[2]*Most data generally agree with Breitwieser et al. (2012).*
[3]*The species numbers here include 288 subspecies, which means they will be higher than in Arechavaleta et al. (2010) which only counts species.*
[4]*But see Garnock-Jones (2014) who suggested the actual number of endemic genera is lower, ca. 28–44 (7–10%).*
[5]*With two or more native species on the island.*
[6]*From a specimen of at least one native species from the island.*

the most recent phylogenetic study (that included at least one native species sampled from that island) to determine if the genus was monophyletic on the island system and whether there was a change in ploidy level compared to the sister group. For lineages with dated phylogenies, we extracted the mean time of their divergence from both sister lineages and their most recent common ancestor (i.e., stem and crown ages, respectively). Values were either given in papers or estimated using WebPlotDigitizer[4] on dated tree figures. We also estimated the variance for each crown and stem age using the standard deviation (SD) or assuming normality of the longer of the two tails of the 95% highest posterior density or confidence intervals (after Lee et al., 2012; Tanentzap et al., 2015; Brandt et al., 2016). For genera that were not monophyletic

[4]https://automeris.io/WebPlotDigitizer/

within the archipelago, we used dates of the earliest-derived species or lineage on the island system. From these data (see **Supplementary Files**), we calculated several comparative statistics (**Table 2**), as described below.

Statistical Analyses

We tested our conceptual model (i.e., P1–P7) by fitting two separate phylogenetic least squares (PGLS) regression models using the *gls* function (implemented in the nlme package; Pinheiro et al., 2014) in R v.3.6, which were united into a single path analysis with the *piecewiseSEM* package (Lefcheck, 2016). As we were interested in the effect of ploidy on diversification, we analyzed lineages that had at least two native species on the island system, a dated phylogeny available with at least one native island species, and at least one chromosome count of a native island specimen (**Supplementary File 1**). Lineages that fit our criteria were included from New Zealand (n = 98), Canary Islands (n = 23), Hawaiian Islands (n = 23), and the Juan Fernández Islands (n = 6), but data from Galápagos Islands were insufficient for downstream analysis. These lineages were always genera aside from four exceptions that were closely related genera that formed a monophyletic group on the archipelago (**Supplementary File 1**). The models were fitted to the number of endemic species and number of ploidy levels for each lineage on each island system. To predict species number, we modeled observations based on both log-transformed stem age and number of ploidy levels of the island lineage, total number of accepted and unresolved species in the lineage outside of the island system according to The Plant List v1.1 (2013) , and separate binary variables for whether the lineage was monophyletic on the island system and its ploidy level was different from its closest sister lineage outside of the island system. We let the effects of ploidy levels vary among the four island systems and accounted for differences in the mean endemic diversity of lineages among island systems. Doing so also ensured that island systems with more data points, such as New Zealand, did not bias the estimated effects towards themselves. We used stem ages to represent lineage age because they were better sampled than crown ages (n = 150 vs. 116 lineages, respectively), were highly correlated with crown age (Pearson correlation between mean ages, r = 0.84, df = 110, p < 0.001), and may better reflect the entire evolutionary history of clades (Scholl and Wiens, 2016). The model of archipelago ploidy levels was identical to that for endemic diversity except without including ploidy levels and archipelago monophyly as predictors and letting the stem age effect vary with the island system. Following standard practice, the models were fitted assuming the expected covariance in the responses between any two lineages was proportional to their shared evolutionary history along a phylogenetic tree, i.e., a Brownian motion (Symonds and Blomberg, 2014). Distances were derived by pruning the largest time-calibrated phylogenetic tree available for vascular plants, which contained 74,531 species and was generated with hierarchical clustering analysis of GenBank data and a backbone provided by Open Tree of Life version 9.1 (Smith and Brown, 2018; Jin and Qian, 2019). On average, sister branches in this phylogeny had an overlap of 1,792 base pairs, which corresponds to roughly one or two gene regions. Although the responses were counts and so could also be modeled with other approaches (e.g., phylogenetic generalized linear models), log-normal transformations made the responses normally distributed, as expected for some Poisson distributed variables. The *gls* function also had the advantage that it could be used to fit a path analysis and incorporate uncertainty in the responses unlike these other regression approaches. We specifically accounted for uncertainty in divergence time estimates and different levels of data completeness by weighting observations of species counts and ploidy levels with the inverse square root of divergence time SDs and the proportion of species with chromosome counts, respectively, after Garamszegi and Møller (2010).

RESULTS

Comparison of Island Groups

The five island systems (New Zealand, Canary Islands, Hawaiian Islands, Juan Fernández, and Galápagos Islands) differ regarding number and total area of island system, distance to nearest

TABLE 2 | Summary of variables used for the statistical analyses.

	Island archipelago				
Variable	**New Zealand**	**Canary Islands**	**Hawaiian Islands**	**Juan Fernández**	**All islands**
No. of lineages	98	23	23	6	150
Number of monophyletic island lineages	33	13	18	3	67
Number of lineages with different ploidy level than closest sister outside of island	8	2	6	2	18
Mean (SD) number of island endemic species per lineage	11.3 (17.6)	13.7 (15.0)	16.4 (19.0)	4.5 (3.7)	12.2 (17.1)
Mean (SD) stem age per lineage in millions of years	15.63 (17.30)	8.48 (6.19)	9.21 (6.23)	5.93 (5.86)	13.16 (14.81)
Mean (SD) number of ploidy levels per lineage	1.7 (1.2)	1.5 (0.7)	1.2 (0.4)	1.0 (0.0)	1.6 (1.1)
Mean (SD) % of endemic island species with chromosome counts	86.0 (22.8)	74.3 (24.0)	50.9 (31.9)	71.8 (36.1)	78.2 (27.8)
Median number of species in lineage outside of island system	34	77	58	384	51

Means and standard deviations (SD) were calculated for each non-count variable. The data set included the subset of 150 lineages of native vascular plants with the following criteria: dated phylogeny available, at least one native species sampled in the phylogeny, at least two native species on the island system, and at least one known chromosome count from the island system.

continent, number (and percentage) of native and endemic species and genera, and data availability for phylogenies, dated phylogenies, chromosome numbers, genome size, and ploidy information (**Table 1**). The island systems have a five-fold difference in the number of indigenous genera (103 in JF to 466 in CI), a 10-fold difference in the number of native species (209 in JF to 2,551 in NZ), and a 20-fold difference in the number of endemic species (136 in JF to 2,056 in NZ). For simplicity, we chose to lump sub-specific ranks under species, so our determination of absolute species numbers may differ from other sources and we acknowledge that there is much taxonomic uncertainty in island radiations. We do not think that this approach has affected our analyses other than minor differences in species numbers.

On all island systems, the majority of indigenous vascular plant species are angiosperms, with a small percentage of ferns and even fewer gymnosperms. The average number of native species per genus is 2.0 (JF, GI), 3.6 (CI), 4.5 (HI), and 6.0 (NZ). The percentage of endemic genera was low for all island systems (3–11%), as was the availability of genome size data (up to 14% in CI). The availability of chromosome numbers was highest for CI and NZ (70–77%) and lowest for GI (7%), with HI (34%) and JF (42%) also having relatively few counts. The number of native non-monotypic genera on each island system ranged from 39 (38%, JF) to 256 (60%, NZ). On all island systems, molecular phylogenies have been published for the majority of these non-monotypic genera, ranging from 62% in HI to 80% in NZ, but dated phylogenies are less prevalent (4% in GI to 44% in NZ).

Of all the island systems studied here, New Zealand has the largest area, the lowest number of main islands (but the highest number of total islands including smaller islands), the largest flora, and the highest percentage of data available for its species (**Table 1**). Ferns comprise about 9% of species (210 species, 46% endemic) and 13% of genera in the native plant vascular flora [57 genera, of which only monotypic *Leptolepia* (Dennstaedtiaceae) and *Loxsoma* (Loxsomaceae) are endemic]. With respect to gymnosperms, all 22 species and one-third of the 10 native genera are endemic. Of the 430 NZ vascular plant genera, 256 have at least two native species, and the rest are monotypic on the archipelago. Of these 256, 78% have a phylogeny but only 46% have dated phylogenies. About 70% of genera with a phylogeny (144 genera) have 50% or more species included in phylogenies. Roughly one-third of these genera are monophyletic or nearly so in NZ (i.e., one NZ origin likely), one-third are not monophyletic in NZ (i.e., more than one origin in NZ), and about 40% are unknown due to lack of phylogeny or sampling. Ninety percent of the genera have at least one species with a chromosome count, but there are few published genome size estimates (12.5%, see **Table 1**). Fifty-seven genera have 10 or more native species, and two of these have over 100 species, i.e., *Veronica* (Plantaginaceae, 143 species, 96% endemic) and *Carex* (Cyperaceae, 118 species, 88% endemic). Ninety-four percent of genera with phylogenies have at least 50% of the species with chromosome counts. Genera have between 1 and 6 ploidy levels represented on the archipelago, ranging from 2*x* to 20*x*. Of the 199 NZ genera with phylogenies, 40% have multiple ploidies (the majority with two or three ploidy levels).

The Canary Islands rank second after NZ in terms of number of native species and data availability, third in terms of area, and first in terms of proximity to the nearest mainland (**Table 1**). They have the second-lowest percentage of endemic species (39%). There are 27 genera with 10 or more species, including several adaptive radiations, two of which comprise larger lineages of multiple genera, i.e., the *Aeonium* alliance (Crassulaceae; 62 endemic species in four genera) and the *Sonchus* alliance (Asteraceae; 62 endemic species in six genera; Kim et al., 2008). Gymnosperms are rare in CI (only six native species, two of which are endemic), and of the 50 fern species, only 6% are endemic; there are no endemic genera of gymnosperms or ferns. Of the 466 genera, 217 have between 2 and 33 species, and the rest are monotypic on the archipelago. Of these 217, 78% have a molecular phylogeny but only one-third are dated phylogenies. Between 1 and 27 native CI species per genus (17–100%) are included in the phylogenies; over half have 50% or more species included in phylogenies. Roughly 20% of these genera with phylogenies are monophyletic or nearly so in CI (i.e., one CI origin likely), 30% are not monophyletic in CI (i.e., more than one origin in CI), and about half are unknown due to lack of phylogeny or sampling. Eighty-seven percent of the genera have at least one species with a chromosome count, and there are genome size estimates known for about 14% of species. Ninety-four percent of genera with phylogenies have at least 50% of the species with chromosome counts. Genera have between 1 and 4 ploidy levels represented on the islands, ranging from 2*x* to 12*x*, with the most common ploidies being diploid and tetraploid. Of the 170 genera with phylogenies, 40% have multiple ploidies (the majority with two ploidy levels).

The Hawaiian Islands have the second-largest land mass (after New Zealand) of the five island systems and the highest level of species endemicity (88%; **Table 1**). There are 272 native vascular genera, 139 (51%) have between 2 and 80 species, and the remainder are represented by one species on the archipelago. Ferns represent 14% of the vascular plant species diversity and show a high level of endemicity at the species level (75%). There are no native gymnosperm taxa in HI. Phylogenies are available for about one-third of the genera, and of these, one-third are dated. Overall, we identified 142 phylogenies that included Hawaiian taxa, and 39% of these were for groups in which only one species occurs in HI. The flora is not well characterized chromosomally at the species level overall (34% of species with known chromosome numbers), but there is broad representation at the generic level, with 57% of genera having at least one species known. Ploidy levels range primarily from diploid to octoploid, with diploid and tetraploid representing the majority of known levels. Overall, polyploid series with multiple ploidy levels are lacking but there are true dysploid series (e.g., x = 13, 14) that occur in several genera of the Hawaiian silversword alliance (Asteraceae, Carr, 1998). The most common scenario for HI taxa (with known chromosome numbers) is that these lineages are tetraploid

relative to their overseas congeners and these show chromosomal stasis on the island system [e.g., angiosperms: the six genera of Campanulaceae, *Bidens* (Asteraceae), *Stenogyne* (Lamiaceae); ferns: *Deparia* (Athyriaceae)]. Among angiosperm genera with two or more species, 60 (43%) are the result of a single colonization event and 11 are considered to have diversified *in situ*. Representative adaptive radiations include the genera of the Lobelioideae (Campanulaceae), those in the Lamioideae (Lamiaceae), and the aforementioned silversword alliance (Asteraceae; Price and Wagner, 2018). Twenty-seven genera are the result of two colonization events and 10 have three or more introductions to HI.

The Juan Fernández Islands have the smallest flora and the smallest area, have the same low number of only three main islands as New Zealand, and are probably second to last in terms of data availability (**Table 1**). Of the 103 JF genera, 39 have between 2 and 12 species, and the rest are monotypic on the archipelago. Only three genera have more than 10 native species, i.e., *Carex* (Cyperaceae, 11 species including seven endemic), *Dendroseris* (Asteraceae, 12 species, all endemic), and *Hymenophyllum* (Hymenophyllaceae, 11 species, only one endemic). Overall, species endemicity is high (65%), placing JF as third among the island systems in terms of species-level endemicity, and equal with New Zealand and the Hawaiian Islands (11%) at the generic-level. Ferns represent about one-quarter of native vascular plant species (57 species, 46% endemic) and genera (24 genera, of which only monotypic *Thyrsopteris* is endemic); there are no native gymnosperms. There are phylogenies for 33 genera but only eight of these are dated. Between 1 and 7 native JF species (20–100%) are included in the phylogenies; 21 genera (63% of those with a phylogeny) have 50% or more species included in phylogenies. Roughly 20% of these genera are monophyletic or nearly so in JF (i.e., one JF origin likely), 20% are not monophyletic in JF (i.e., more than one origin in JF), and the remaining 60% are unknown due to lack of phylogeny or sampling. Only about half of the JF genera have at least one species with a chromosome count, and there are no genome size estimates known. JF genera have either one or two ploidy levels represented on the archipelago, ranging from 2*x* to 8*x*, with diploids and tetraploids the most common. Of the 33 JF genera with phylogenies, about 70% have at least one species with a chromosome count and half have at least 50% species with a chromosome count. Of these, almost all have only one ploidy level.

The Galápagos Islands have the highest number of main islands with an area only slightly greater than the Canary Islands (**Table 1**). Native vascular genera total 277, seven are endemic (3%); 105 (38%) of these have between 2 and 19 species, and the remaining are monotypic on the archipelago. Ferns represent ca. 23% of vascular plant species (9% of these are endemic); there are no endemic fern genera. Like the Hawaiian and Juan Fernández Islands, there are no native gymnosperms in the Galápagos. Molecular-based phylogenies are available for 42 genera, and only four of these are dated phylogenies. Between 1 and 13 GI species (11–87%) are included in the phylogenies; 21 genera have 50% or more species included in the phylogenies. Chromosome numbers are available for only 21 genera, and of these 13 genera have two or more species. Data for chromosome numbers for native taxa and dated phylogenies were only available for one lineage, which was not included in the statistical analyses. Where chromosome numbers are available, the species are primarily 4*x* or 6*x* with just one ploidy level in those lineages in most cases. In two cases, multiple ploidy levels exist on the archipelago and these were both in fern genera [*Adiantum* (Pteridaceae) and *Polypodium* (Polypodiaceae)].

Statistical Analysis of Island System Data

Table 2 summarizes the lineages from the four island systems that fitted the criteria for the statistical analyses (i.e., at least two native species on the island system, dated phylogeny available with at least one native island species included, and at least one chromosome count of a native island specimen). New Zealand lineages comprised 66% of the data set, followed by the Hawaiian and Canary Islands with 15% each, and Juan Fernández at 3% (no data were available for the Galápagos Islands). Just under half (45%) of the genera were monophyletic on the island system. For the majority of lineages where the ploidy level of the sister group could be determined, the lowest ploidy level on the island was the same as the sister group (JF 60% n = 6, HI 73% n = 23, CI 92% n = 23, NZ 92% n = 98; **Table 2**). The remaining lineages had a polyploidization event that occurred sometime along the stem lineage immediately before or after colonization of the island and thus may represent island neopolyploidization. New Zealand had the oldest mean stem age (15.63 vs. 4.18 million years in Juan Fernández) and the highest number of ploidy levels per lineage (1.7 vs. 1.0 in Juan Fernández). The Hawaiian Islands had the highest mean number of island endemic species per lineage (16.1 vs. 4.8 in Juan Fernández) but also the lowest mean percentage of species with chromosome counts (50 vs. 86% in New Zealand).

Our statistical analysis supported the hypothesis that polyploidy shapes the diversification of island floras (**Figure 2**). We found that greater levels of polyploidy directly promoted endemic diversity on island systems (P1; $t_{150} = 3.11$, $p = 0.002$). Over the range of observed ploidy levels (1–6), the estimated number of species in island genera increased from a mean of 1 to 4 [95% confidence interval (CI) for increase: 0.7–9.9]. Polyploidy itself was enhanced by a larger source of potential congeneric colonists (P4; $t_{150} = 5.36$, $p < 0.001$) and a change in ploidy level from overseas sister taxa (P3; $t_{150} = 4.04$, $p < 0.001$). Lineages that changed in ploidy near the time of island colonization had, on average, 4.4 ploidy levels as compared with 2.5 levels where these changes were absent (95% CI for difference: 0.83–3.22). In these same lineages, as the size of the potential colonist pool increased from the 25th to 75th percentile of observed values (5–275 species in the source pool), estimated ploidy levels increased from a mean of 3.5 to 5.4 (95% CI for increase: 1.0–2.9).

Lineage age also affected diversification outcomes. Older lineages were more diverse ($t_{150} = 3.69$, $p < 0.001$), as expected if they had more time to diversify (P5) and had slightly more ploidy levels across island systems (P2; $t_{150} = 1.98$, $p = 0.049$).

Over the interquartile range of observed stem ages, ploidy levels were estimated to increase from a mean of 4.1 to 4.7 (95% CI for increase: 0.4–2.1). The effect of stem age on ploidy levels was, however, reversed (negative and statistically significant) for the Hawaiian Islands (**Figure 3**, "island effect"), resulting in fewer ploidy levels in the older Hawaiian lineages ($t_{150} = -2.60$, $p = 0.010$). Endemic diversity was also higher, on average, for Hawaiian lineages ($t_{150} = 2.11$, $p = 0.036$) and lower in Juan Fernández lineages ($t_{150} = -2.10$, $p = 0.038$; **Figure 3**, "island effect"). Whether lineages were monophyletic on the island system (P7) or had more potential congeneric colonists outside of the island system (P6) had no direct effect on endemic diversity ($t_{150} = 1.62$, $p = 0.107$ and $t_{150} = 1.62$, $p = 0.108$, respectively). Tests of directed separation indicated a missing path in our analysis from the lineages being monophyletic ("repeat colonization") to ploidy levels ("polyploidy," **Figure 2**). Including this path in our final model made no difference to our results ($t_{150} = 0.15$, $p = 0.885$). Overall, the model predicted both endemic diversity and the number of ploidy levels reasonably well ($R^2 = 0.35$ and 0.38, respectively; **Figure 3**).

DISCUSSION

Understanding evolutionary processes on islands has improved with molecular phylogenetic analyses and especially the development of time-calibrated phylogenies, which provide a temporal framework for island colonization. Using this information from five island systems, we highlight the important role of polyploidy in the dispersal, colonization, and diversification of island lineages. Our results overturn the perception that chromosomal stasis is a feature of island systems, and instead provide evidence that demonstrates the importance of polyploidy in promoting both colonization and species diversification.

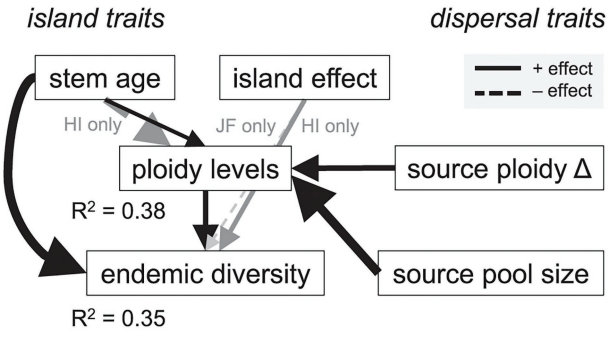

FIGURE 3 | Estimated path analysis to explain the influence of polyploidy on the diversity of endemic island plant species. The model was fitted to 150 lineages across four archipelagos for which we had published estimates of their stem age and chromosome numbers. Arrows are scaled proportional to standardized effect sizes and only statistically significant pathways ($p < 0.05$) are shown (e.g., direct effects of repeat colonization and source pool size on endemic diversity were tested but not significant). Gray text and arrows refer to pathways that were statistically significant only on the Hawaiian (HI) or the Juan Fernández (JF) Islands.

Previous authors have already commented that polyploidy may assist colonization of islands by facilitating the establishment of species (Linder and Barker, 2014) and this also parallels findings that polyploidy is more common in invasive plants (Te Beest et al., 2012). Furthermore, polyploids seem to be preadapted to island colonization. One factor providing polyploids an advantage over related diploids in the colonization of new habitats is higher genetic diversity, which leads to lower inbreeding depression of polyploids (Rosche et al., 2017) and improved adaptability (Scarrow et al., 2020). Other factors include improved environmental tolerance (Moura et al., 2020), higher frequency of vegetative reproduction, and selfing (in part due to a breakdown of self-incompatibility; Robertson et al., 2011), which provide a means to colonize and survive longer at low population size (Baker, 1967; Herben et al., 2017). Thus, there are several arguments for considering polyploids to be superior colonizers but only very few that support the idea that polyploids are better dispersers (e.g., Kuo et al., 2016 for ferns). For example, Harbaugh (2008) suggested that polyploid *Santalum* (Santalaceae) have smaller seeds and fruits with a thicker endocarp, making them easier to disperse by birds. However, easy dispersal does not necessarily translate to frequent establishment and consequently more species on an island.

Our analyses demonstrate that polyploidy contributes significantly to diversification of established island lineages. Endemic richness on islands increases over time in colonizing lineages, reflecting the accumulation of new species through local speciation on islands. The results of the path analysis indicate that polyploidy influences species diversity on islands through colonizing species belonging to large lineages, apparently able to speciate more than others independent of the geographic setting. Thus, the source pool size is one factor that drives polyploidization, which in turn affects endemicity on the archipelago. Additionally, ploidy levels also increase over time and thereby contribute to increasing the number of endemic taxa. According to our path analysis (**Figure 3**), lineages are more likely to produce new polyploid species on the archipelago, when the lineage already underwent polyploidization after diverging from its mainland sister lineage.

Together, our results agree with previous conclusions that speciation on islands is similar to elsewhere (Takayama et al., 2018) and supports our hypothesis that polyploidy is a diversification trait on islands, with this trait being taxonomically more frequent in some lineages than others (Wood et al., 2009), e.g., Asteraceae (**Figure 1**). For example, Meudt et al. (2015) demonstrated that diversification in the hexaploid lineage of *Veronica* in New Zealand, the largest endemic lineage in New Zealand (**Figure 1**), is related to its decrease in genome size. This genome downsizing is also related to diversification in mainland lineages of *Veronica*. A connection of genome downsizing in polyploid lineages has similarly been found in polyploid *Cheirolophus* (Asteraceae) on the Canary Islands (Hidalgo et al., 2017). More generally, Kapralov and Filatov (2011) demonstrated a correlation of low genome size with species richness in endemic island genera. Thus, the relevance of polyploidy for island diversification involves both lineage

features, such as the tendency to form polyploids, and aspects of island characteristics, such as habitat heterogeneity and availability of pollinators. Islands also have many intrinsic features that may facilitate polyploid differentiation and persistence, including small size, habitat heterogeneity and proximity, and nutrient-rich soils associated with volcanic activity or maritime animals.

It remains to be studied which aspects of polyploidy are critical for a given taxon and island and whether there is a more general aspect that allows polyploids to disperse, colonize, and/or diversify. Given advances in molecular biology, it seems feasible in the future to determine a propensity to form polyploid species (Bomblies, 2020) and to detect ancient rounds of polyploidy (Tiley et al., 2018). At the moment, polyploidy *per se* outside the island systems could not be considered in our analyses because of the difficulty to establish this in many cases (i.e., where the chromosome number of a sister lineage was unknown) and our observations that many of the lineages colonizing the island systems are already polyploid lineages of mainland taxa [e.g., *Coprosma* (Rubiaceae)].Where there is high species richness within a genus, it seems the polyploid lineage is the one that colonizes the island systems, but overall these examples remain few. One excellent example is the Hawaiian silversword alliance, a recent radiation that includes ~30 species in three endemic genera, which exhibit a variety of growth forms and exploit diverse habitats (Baldwin and Sanderson, 1998). The endemic Hawaiian species are all allotetraploid derivatives from diploid continental ancestors (Carr, 1998; Barrier et al., 1999). Our result of increased diversity on the island systems based on a higher ploidy level of the island lineage compared to the mainland sister lineage, nevertheless, allows us to focus on the question of whether polyploidy is a dispersal or colonization trait. Crawford et al. (2009) already queried whether the colonizers themselves were polyploid or if polyploidy evolved *in situ*, which motivated our first prediction (P1) in our conceptual model. Our literature survey provides many examples of such cases to be studied in the future under different evolutionary scenarios, with and without further diversification, with or without further changes in ploidy level, and in different geological and temporal settings.

Polyploidy and Diversification on Islands – Generalities

Islands have complex geological histories, and generally New Zealand is considered to be a "continental island" as compared to the other "oceanic islands" that we studied here (Weigelt et al., 2013). However, our analyses showed that this was not a meaningful distinction for determining factors influencing diversification as results from all island systems were generally similar (few island effects, and none that distinguished New Zealand from the other four archipelagos). Thus, despite the fact that New Zealand has not been included with other studies of island polyploid speciation, we find similarities among all the island systems included here in terms of the contribution of polyploidy to species diversification (**Tables 1** and **2**). Also, the age of New Zealand's flora, as estimated based on mean stem age per lineage (**Table 2**), does not distinguish New Zealand from others due to the large variation in age. This large variation in ages of island lineages (**Table 2**) demonstrates that colonization has been successful mostly independent of time of arrival. Thus, even old islands are dispersal- or establishment-limited not niche-limited. Based on Carvajal-Endara et al. (2017) the establishment of a lineage seems to be the bigger hurdle than dispersal. While our study focused on particular island systems that have been well studied, the patterns of species diversification on islands that have emerged may be more general to other island systems (e.g., López-Alvarado et al., 2020). Other island systems would be interesting to evaluate for these larger-scale patterns, as studies on their specific flora are becoming available (e.g., Sardinia – López-Alvarado et al., 2020; Balearic Islands – Rosselló and Castro, 2008). Indeed, Rice et al. (2019), in their global analysis of polyploid biogeography, find high levels of polyploidy on several island systems, which mainly relate to the climatic conditions and predominance of perennial taxa.

Polyploidy and Species Diversification on the Individual Islands

New Zealand is considered a continental island, because it separated from Gondwana 80 million years ago. Nevertheless, much of its flora, especially non-woody taxa, is considered to have arrived by long distance-dispersal following large-scale marine transgression during the Oligocene (Heenan and McGlone, 2019). Still, lineages on New Zealand included in our analyses are on average older than those of the other island systems (**Table 2**). Sanmartín et al. (2007) inferred most of the colonization events to have occurred from Australia, and our survey of available phylogenies shows that many New Zealand genera are part of larger Southern Hemisphere lineages. New Zealand has the largest flora of those studied here with a high number of endemics and several large radiations involving polyploid formation (e.g., *Veronica*; Meudt et al., 2015, **Table 1**). The flora is well-studied phylogenetically and chromosomally with genome size measurements increasing recently, which makes New Zealand an excellent example for studies on the importance of polyploidy.

The Canary Islands are characterized among the five archipelagos as the one closest to a continent, which may explain its low endemicity despite having the highest diversity per square kilometer of all island systems except Juan Fernández (**Table 1**). Indeed, nearby mainland Europe or Africa seems to be the ancestral areas for most Canary Island genera and species (Sanmartín et al., 2008). This diversity and the ease to access the island system explains why the Canary Islands have become such an important natural laboratory for evolutionary botanists with studies investigating patterns of dispersal and evolution on the archipelago, such as the evolution of woodiness (Böhle et al., 1996; Schüßler et al., 2019), breeding system (Soto-Trejo et al., 2013), and photosynthetic pathways (Mort et al., 2007). Most important in our context is the intensive karyological study of the archipelago (e.g., Suda et al., 2005; **Table 1**). Interestingly, a number of studies

(e.g., Allan et al., 2004) demonstrate that Canary Island taxa occupy similar habitats as compared to their continental relatives, suggesting that their ancestors might have been pre-adapted to occupying particular habitats.

Polyploidy seems to have a larger impact on diversification on New Zealand and the Canary Islands compared to the other island systems, for example in the largest radiation on the Canary Islands, the *Aeonium* alliance (**Figure 1**), or *Sideritis* (Lamiaceae; Raskina et al., 2008). The suggestion that polyploidy seems to be less common in Asteraceae on the Canary Islands compared to other island systems (Crawford et al., 2009) only holds for the two largest genera of Asteraceae on the island system (*Argyranthemum*, *Sonchus* alliance), not generally across the family. More detailed analysis on the distribution of polyploids on New Zealand and the Canary Islands will be necessary to determine whether, for example, the larger altitudinal range on these two archipelagos promote diversification by polyploidy. This should be analyzed in connection with the report of the low genome size of the Macaronesian flora in relation with the rest of the world (Suda et al., 2005).

The Hawaiian Islands are the island system farthest away from a continent, and thus its genera are mostly considered to have colonized the island system only once (Funk and Wagner, 1995; Keeley and Funk, 2011). Following from this it also has the highest number of endemics (**Table 1**). Ancestors of Hawaiian taxa are derived from diverse geographical regions including North America, Australia, Asia, Africa, and other island systems in the Pacific (e.g., New Zealand; Keeley and Funk, 2011; Price and Wagner, 2018). The considerable age of some of its lineages is likely caused by the influence of early colonization of now succumbed islands of the archipelago (García-Verdugo et al., 2019), while most others represent more recent arrivals (Price and Clague, 2002). Despite great interest in the flora of the island system, it is only poorly investigated karyologically (**Table 1**). Thus, the hypothesis of chromosomal stasis of lineages on the Hawaiian Islands (Stuessy and Crawford, 1998) is founded on few, though important, cases. Nevertheless, the data available suggest a lower importance of polyploidy than in New Zealand and the Canary Islands, chromosomal stasis thus being peculiar to the Hawaiian Islands. One notable pattern for the Hawaiian Islands is the formation of dysploid series (Carr, 1998), which has also been noted on rare occasions in the Canary Islands (*Sideritis*, Barber et al., 2000) and New Zealand (*Veronica*, Wagstaff and Garnock-Jones, 1998).

The Juan Fernández Islands are the smallest archipelago but the flora has been intensively investigated floristically (Greimler et al., 2013) and evolutionarily (Takayama et al., 2018), including a number of karyological studies but so far no measurements of genome sizes (**Table 1**). The closest continental source is South America, which seems to be the ancestral area for several Juan Fernández genera and species (Stuessy et al., 2017). Despite the fact that some genera have species with different ploidy levels on different islands, there is no clear case of polyploid origin there because those cases are either based on separate colonization events or have not been studied phylogenetically. Also, Stuessy et al. (2017) state only two possible cases of polyploid origins on this island system. The high number of polyploids in the ancestral lineages of some Juan Fernández endemics, such as *Dendroseris* and *Erigeron* (Asteraceae), therefore suggests that polyploidy is a dispersal or colonization trait rather than a diversification trait for the flora of Juan Fernández. Two factors could explain this lack of polyploid speciation on Juan Fernández Islands: their small size providing little space for the avoidance of inter-cytotype gene flow or the young age of the lineages compared to that of the other islands (**Table 2**). Given more time the intraspecific variation in ploidy of several species on the islands [Stuessy et al., 2017; e.g., in *Eryngium bupleuroides* (Asteraceae); **Figure 1**] could translate into new polyploid species.

The Galápagos have not been included in our analyses, despite the importance of this island system for the history of evolution and biogeography, because its flora has been poorly studied. Its flora, like its fauna, is considered to be derived from South America (Darwin, 1839; Elisens, 1992), although exceptions occur (e.g., Andrus et al., 2009). The Galápagos flora is only a third the size of the Canary Islands flora despite similar area, possibly due to its dryness, which could also explain the low number of endemic fern species (**Table 1**). Polyploid speciation, however, has been shown to occur in the genus *Pectis* (Asteraceae, Hansen et al., 2016). In *Scalesia* (Asteraceae), the largest radiation on the archipelago, the species are considered to be tetraploid but have not further diversified in ploidy level (Eliasson, 1974; Fernández-Mazuecos et al., 2020).

The discussion of these five island systems already demonstrates that there are as many idiosyncratic patterns as there are island systems, but polyploidy seems to play a role in many of them. For example, Rosselló and Castro (2008) found a large frequency of neopolyploidization events among the endemic plants of the Balearic Islands, whereas Sun and Stuessy (1998) recorded chromosomal stasis in the flora of Ullung Islands. Studies of additional island systems that vary in their locations and floristic complexities would be beneficial for enhancing our understanding of endemic polyploid diversification.

OUTLOOK AND FUTURE DIRECTIONS

We see the study of polyploidy on islands as an exciting avenue to further our understanding of polyploid species diversification. Our comparative study was limited to data that were available for the island systems under study and we were surprised that some of these "classic" island systems remain poorly known chromosomally and phylogenetically, especially the Galápagos. While genomic studies have revolutionized our ideas of polyploid genome dynamics (e.g., Bomblies, 2020), there is much value in continuing to generate chromosome numbers and dated phylogenies for native island endemics. Because island radiations are often young and/or rapid on an evolutionary timescale, in many cases acquiring a well-resolved phylogeny for a group of species based on single or few gene sequences has been challenging (e.g., Meudt and Simpson, 2006; Knope et al., 2012; Vitales et al., 2014). Newer methodologies that take advantage of next-generation

sequencing methods should be helpful in this regard (Larridon et al., 2020). Similarly, overseas sister lineages need to be investigated to understand the context of island diversification. One additional limitation to our study was the lack of data for genome size estimates, either to help resolve ploidy levels or to analyze its effect on polyploid species diversification on the island systems. Given that genome downsizing is considered to be an important part of polyploid evolution (Soltis et al., 2016), we see this factor as a potentially important player in facilitating establishment on islands, as mentioned previously for *Veronica* (Meudt et al., 2015). Additionally, future studies should further compare the effect of polyploidy on colonization and diversification among woody vs. herbaceous, dry vs. fleshy fruited, and selfing vs. outcrossing lineages (Vamosi et al., 2007) for a more complete picture of island diversification.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

HMM, JAT, AJT, DCA, and WGL designed the study and compiled the data, with assistance from SCN, AJB and JI. AJT and JI completed statistical analyses. All authors contributed to drafting the manuscript and developing the final version and are accountable for the contents of the work.

FUNDING

We thank the Marsden Fund, Royal Society Te Apārangi (LCR1702) for research funding focused on polyploidy in the native New Zealand flora to WL, HM, and JT; the Gatsby Charitable Foundation (Grant Number GAT2962) to AT for funding focused on evolutionary diversification, and the German Science Foundation (DFG; project AL632/16-1) for travel support to New Zealand for DA.

ACKNOWLEDGMENTS

Thanks to Clyde Imada and Cliff Morden for making data from Hawaiian species available, and to Patrick Brownsey for providing information regarding some fern taxa.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.637214/full#supplementary-material

Supplementary File 1 | The data file analyzed in this study, which includes the following information from 150 island lineages from New Zealand (n = 98), Canary Islands (n = 23), Hawaiian Islands (n = 23), and Juan Fernández Islands (n = 6): lineage (genus) name, number of endemic species, stem age, number of ploidy levels, whether ploidy level differs from closest non-island sister group, monophyly on the island, and number of species in closest non-island sister group.

Supplementary File 2 | References for the 150 lineages included in the statistical analyses (with dated phylogenies) grouped by island system.

REFERENCES

Allan, G. J., Francisco-Ortega, J., Santos-Guerra, A., Boerner, E., and Zimmer, E. A. (2004). Molecular phylogenetic evidence for the geographic origin and classification of Canary Island *Lotus* (Fabaceae: Loteae). *Mol. Phylogenet. Evol.* 32, 123–138. doi: 10.1016/j.ympev.2003.11.018

Andrus, N., Tye, A., Nesom, G., Bogler, D., Lewis, C., Noyes, R., et al. (2009). Phylogenetics of *Darwiniothamnus* (Asteraceae: Astereae) – molecular evidence for multiple origins in the endemic flora of the Galápagos Islands. *J. Biogeogr.* 36, 1055–1069. doi: 10.1111/j.1365-2699.2008.02064.x

Arechavaleta, M., Rodríguez, S., Zurita, N., and García, A. (2010). *Lista de especies silvestres de Canarias. Hongos, plantas y animales terrestres. 2009*. Tenerife: Gobierno de Canarias.

Baker, H. G. (1967). Support for Baker's law - as a rule. *Evolution* 21, 853–856. doi: 10.1111/j.1558-5646.1967.tb03440.x

Baldwin, B. G., and Sanderson, M. J. (1998). Age and rate of diversification of the Hawaiian silversword alliance (Compositae). *Proc. Natl. Acad. Sci.* 95, 9402–9406. doi: 10.1073/pnas.95.16.9402

Barber, J. C., Ortega, J. F., Santos-Guerra, A., Marrero, A., and Jansen, R. K. (2000). Evolution of endemic *Sideritis* (Lamiaceae) in Macaronesia: insights from a chloroplast DNA restriction site analysis. *Syst. Bot.* 25, 633–647. doi: 10.2307/2666725

Barrier, M., Baldwin, B. G., Robichaux, R. H., and Purugganan, M. D. (1999). Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplications. *Mol. Biol. Evol.* 16, 1105–1113. doi: 10.1093/oxfordjournals.molbev.a026200

Böhle, U. R., Hilger, H. H., and Martin, W. F. (1996). Island colonization and evolution of the insular woody habit in *Echium* L. (Boraginaceae). *Proc. Natl. Acad. Sci.* 93, 11740–11745. doi: 10.1073/pnas.93.21.11740

Bomblies, K. (2020). When everything changes at once: finding a new normal after genome duplication. *Proc. R. Soc. B Biol. Sci.* 287:20202154. doi: 10.1098/rspb.2020.2154

Brandt, A., Tanentzap, A., Leopold, D., Heenan, P., Fukami, T., and Lee, W. (2016). Precipitation alters the strength of evolutionary priority effects in forest community assembly of pteridophytes and angiosperms. *J. Ecol.* 104, 1673–1681. doi: 10.1111/1365-2745.12640

Breitwieser, I., Brownsey, P. J., Garnock-Jones, P. J., Perrie, L. R., and Wilton, A. D. (2012). *Phylum Tracheophyta: Vascular plants*. Christchurch: Canterbury University Press.

Carlquist, S. (1974). *Island biology*. New York: Columbia University Press.

Carr, G. D. (1998). "Chromosome evolution and speciation in Hawaiian flowering plants" in *Evolution and speciation of island plants*. eds. T. F. Stuessy and M. Ono (Cambridge: Cambridge University Press), 5–47.

Carracedo, J. C., and Troll, V. R. (2016). *The geology of the Canary Islands*. London: Elsevier.

Carvajal-Endara, S., Hendry, A. P., Emery, N. C., and Davies, T. J. (2017). Habitat filtering not dispersal limitation shapes oceanic island floras: species assembly of the Galápagos archipelago. *Ecol. Lett.* 20, 495–504. doi: 10.1111/ele.12753

Caujapé-Castells, J., García-Verdugo, C., Marrero-Rodríguez, Á., Fernández-Palacios, J. M., Crawford, D. J., and Mort, M. E. (2017). Island ontogenies, syngameons, and the origins and evolution of genetic diversity in the Canarian endemic flora. *Perspect. Plant Ecol. Evol. Syst.* 27, 9–22. doi: 10.1016/j.ppees.2017.03.003

Crawford, D. J., and Archibald, J. K. (2017). Island floras as model systems for studies of plant speciation: prospects and challenges. *J. Syst. Evol.* 55, 1–15. doi: 10.1111/jse.12234

Crawford, D. J., Lowrey, T. K., Anderson, G. J., Bernardello, G., Santos-Guerra, A., and Stuessy, T. F. (2009). "Genetic diversity in Asteraceae endemic to ocean islands: Baker's law and polyploidy" in *Systematics, evolution, and biogeography of compositae*. eds. V. A. Funk, T. F. Stuessy, A. Susanna and R. J. Bayer (Vienna: International Association of Plant Taxonomists), 101–113.

Crawford, D. J., and Stuessy, T. F. (1997). "Plant speciation on oceanic islands" in *Evolution and diversification of land plants*. eds. K. Iwatsuki and P. H. Raven (Tokyo: Springer-Verlag), 249–267.

Darwin, C. R. (1839). *Narrative of the surveying voyages of His Majesty's Ships Adventure and Beagle between the years 1826 and 1836, describing their examination of the southern shores of South America, and the Beagle's circumnavigation of the globe. Journal and remarks. 1832-1836. Vol. 3*. London: Henry Colburn.

Dawson, M. I. (2000). Index of chromosome numbers of indigenous New Zealand spermatophytes. *N. Z. J. Bot.* 38, 47–150. doi: 10.1080/0028825X.2000.9512673

Dawson, M. I. (2008). Index of chromosome numbers of indigenous New Zealand vascular plants [Online]. Landcare Research, New Zealand. Available at: http://www.landcareresearch.co.nz (Accessed February 23, 2021).

Eliasson, U. (1974). Studies in Galápagos plants XIV. The genus *Scalesia* Arn. *Opera Bot.* 36, 1–117.

Elisens, W. J. (1992). Genetic divergence in *Galvezia* (Scrophulariaceae): evolutionary and biogeographic relationships among south American and Galápagos species. *Am. J. Bot.* 79, 198–206. doi: 10.1002/j.1537-2197.1992.tb13638.x

Emerson, B. (2002). Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. *Mol. Ecol.* 11, 951–966. doi: 10.1046/j.1365-294x.2002.01507.x

Fernández-Mazuecos, M., Vargas, P., McCauley, R. A., Monjas, D., Otero, A., Chaves, J. A., et al. (2020). The radiation of Darwin's giant daisies in the Galápagos Islands. *Curr. Biol.* 30, 4989–4998.e7. doi: 10.1016/j.cub.2020.09.019

Frankham, R. (1997). Do island populations have less genetic variation than mainland populations? *Heredity* 78, 311–327. doi: 10.1038/hdy.1997.46

Funk, V. A., and Wagner, W. L. (1995). *Hawaiian biogeography: Evolution on a hot spot archipelago*. Washington, DC: Smithsonian Institution Press.

Garamszegi, L. Z., and Møller, A. P. (2010). Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. *Biol. Rev.* 85, 797–805. doi: 10.1111/j.1469-185X.2010.00126.x

García-Verdugo, C., Caujapé-Castells, J., and Sanmartín, I. (2019). Colonization time on island settings: lessons from the Hawaiian and Canary Island floras. *Bot. J. Linn. Soc.* 191, 155–163. doi: 10.1093/botlinnean/boz044

García-Verdugo, C., Sajeva, M., La Mantia, T., Harrouni, C., Msanda, F., and Caujapé-Castells, J. (2015). Do island plant populations really have lower genetic variation than mainland populations? Effects of selection and distribution range on genetic diversity estimates. *Mol. Ecol.* 24, 726–741. doi: 10.1111/mec.13060

Garnock-Jones, P. J. (2014). Evidence-based review of the taxonomic status of New Zealand's endemic seed plant genera. *N. Z. J. Bot.* 52, 163–212. doi: 10.1080/0028825X.2014.902854

Greimler, J., López-Sepulveda, P., Reiter, K., Baeza, C., Peñailillo, P., Ruiz, E., et al. (2013). Vegetation of Alejandro Selkirk island (Isla Masafuera), Juan Fernández archipelago. Chile. *Pac. Sci.* 67, 267–282. doi: 10.2984/67.2.9

Hansen, D. R., Jansen, R. K., Sage, R. F., Villaseñor, J. L., and Simpson, B. B. (2016). Molecular phylogeny of *Pectis* (Tageteae, Asteraceae), a C_4 genus of the Neotropics, and its sister genus *Porophyllum*. *Lundellia* 19, 6–38. doi: 10.25224/1097-993X-19.1.6

Harbaugh, D. (2008). Polyploid and hybrid origins of Pacific island sandalwoods (*Santalum*, Santalaceae) inferred from low-copy nuclear and flow cytometry data. *Int. J. Plant Sci.* 169, 677–685. doi: 10.1086/533610

Heenan, P. B., and McGlone, M. S. (2019). Cenozoic formation and colonisation history of the New Zealand vascular flora based on molecular clock dating of the plastid *rbcL* gene. *N. Z. J. Bot.* 57, 204–226. doi: 10.1080/0028825X.2019.1632356

Herben, T., Suda, J., and Klimešová, J. (2017). Polyploid species rely on vegetative reproduction more than diploids: a re-examination of the old hypothesis. *Ann. Bot.* 120, 341–349. doi: 10.1093/aob/mcx009

Hidalgo, O., Vitales, D., Vallès, J., Garnatje, T., Siljak-Yakovlev, S., Leitch, I. J., et al. (2017). Cytogenetic insights into an oceanic island radiation: the dramatic evolution of pre-existing traits in *Cheirolophus* (Asteraceae: Cardueae: Centaureinae). *Taxon* 66, 146–157. doi: 10.12705/661.8

Imada, C. T. (2012). Hawaiian Native and Naturalized Vascular Plants Checklist (December 2012 update). *Bishop Musem Technical Report 60*.

Jaramillo Díaz, P., Guézou, A., Mauchamp, A., and Tye, A. (2017). "CDF checklist of Galapagos ferns and related groups - FCD Lista de especies de Helechos y grupos relacionados Galápagos" in *Charles Darwin Foundation Galapagos species checklist - Lista de Especies de Galápagos de la Foundation Charles Darwin*. eds. F. Bungartz, H. Herrera, P. Jaramillo, N. Tirado, G. Jiménez-Uzcátegui, D. Ruiz, et al (Puerto Ayora, Galapagos: Charles Darwin Foundation / Fundación Charles Darwin).

Jaramillo Díaz, P., Guézou, A., Mauchamp, A., and Tye, A. (2018). "CDF checklist of Galapagos flowering plants - FCD Lista de especies de Plantas con flores Galápagos" in *Charles Darwin Foundation Galapagos species checklist - Lista de Especies de Galápagos de la Foundation Charles Darwin*. eds. F. Bungartz, H. Herrera, P. Jaramillo, N. Tirado, G. Jiménez-Uzcátegui, D. Ruiz, et al (Puerto Ayora, Galapagos: Charles Darwin Foundation / Fundación Charles Darwin).

Jin, Y., and Qian, H. (2019). V.PhyloMaker: an R package that can generate very large phylogenies for vascular plants. *Ecography* 42, 1353–1359. doi: 10.1111/ecog.04434

Kapralov, M. V., and Filatov, D. A. (2011). Does large genome size limit speciation in endemic island floras? *J. Bot.* 2011:458684. doi: 10.1155/2011/458684

Keeley, S. C., and Funk, V. A. (2011). "Origin and evolution of Hawaiian endemics: new patterns revealed by molecular phylogenetic studies" in *The biology of island floras*. eds. D. Bramwell and J. Caujapé-Castells (Cambridge: Cambridge University Press), 57–88.

Kellogg, E. A. (2016). Has the connection between polyploidy and diversification actually been tested? *Curr. Opin. Plant Biol.* 30, 25–32. doi: 10.1016/j.pbi.2016.01.002

Kim, S. -C., McGowen, M. R., Lubinsky, P., Barber, J. C., Mort, M. E., and Santos-Guerra, A. (2008). Timing and tempo of early and successive adaptive radiations in Macaronesia. *PLoS One* 3:e2139. doi: 10.1371/journal.pone.0002139

Knope, M. L., Morden, C. W., Funk, V. A., and Fukami, T. (2012). Area and the rapid radiation of Hawaiian *Bidens* (Asteraceae). *J. Biogeogr.* 39, 1206–1216. doi: 10.1111/j.1365-2699.2012.02687.x

Kuo, L. -Y., Ebihara, A., Shinohara, W., Rouhan, G., Wood, K. R., Wang, C. -N., et al. (2016). Historical biogeography of the fern genus *Deparia* (Athyriaceae) and its relation with polyploidy. *Mol. Phylogenet. Evol.* 104, 123–134. doi: 10.1016/j.ympev.2016.08.004

Larridon, I., Villaverde, T., Zuntini, A. R., Pokorny, L., Brewer, G. E., Epitawalage, N., et al. (2020). Tackling rapid radiations with targeted sequencing. *Front. Plant Sci.* 10:1655. doi: 10.3389/fpls.2019.01655

Lee, W., Tanentzap, A., and Heenan, P. (2012). Plant radiation history affects community assembly: evidence from the New Zealand alpine. *Biol. Lett.* 8, 558–561. doi: 10.1098/rsbl.2011.1210

Lefcheck, J. S. (2016). piecewiseSEM: piecewise structural equation modeling in R for ecology, evolution, and systematics. *Methods Ecol. Evol.* 7, 573–579. doi: 10.1111/2041-210X.12512

Linder, H. P., and Barker, N. P. (2014). Does polyploidy facilitate long-distance dispersal? *Ann. Bot.* 113, 1175–1183. doi: 10.1093/aob/mcu047

López-Alvarado, J., Mameli, G., Farris, E., Susanna, A., Filigheddu, R., and Garcia-Jacas, N. (2020). Islands as a crossroad of evolutionary lineages: a case study of *Centaurea* sect. *Centaurea* (Compositae) from Sardinia (Mediterranean Basin). *PLoS One* 15:e0228776. doi: 10.1371/journal.pone.0228776

MacArthur, R. H., and Wilson, E. O. (1967). *The theory of island biogeography*. Princeton, New Jersey: Princeton University Press.

Mayr, E. (1967). The challenge of island faunas. *Australian Nat. Hist.* 15, 369–374.

McGlone, M. S., Duncan, R. P., and Heenan, P. B. (2001). Endemism, species selection and the origin and distribution of the vascular plant flora of New Zealand. *J. Biogeogr.* 28, 199–216. doi: 10.1046/j.1365-2699.2001.00525.x

Meudt, H. M., Rojas-Andrés, M., Prebble, J. M., Low, E., Garnock-Jones, P. J., and Albach, D. C. (2015). Is genome downsizing associated with diversification in polyploid lineages of *Veronica*? *Bot. J. Linn. Soc.* 178, 243–266. doi: 10.1111/boj.12276

Meudt, H. M., and Simpson, B. B. (2006). The biogeography of the austral, subalpine genus *Ourisia* (Plantaginaceae) based on molecular phylogenetic evidence: south American origin and dispersal to New Zealand and Tasmania. *Biol. J. Linn. Soc.* 87, 479–513. doi: 10.1111/j.1095-8312.2006.00584.x

Mort, M. E., Solits, D. E., Soltis, P. S., Santos-Guerra, A., and Francisco-Ortega, J. (2007). Physiological evolution and association between physiology and growth form in *Aeonium* (Crassulaceae). *Taxon* 56, 453–464. doi: 10.1002/tax.562016

Moura, R. F., Queiroga, D., Vilela, E., and Moraes, A. P. (2020). Polyploidy and high environmental tolerance increase the invasive success of plants. *J. Plant Res.* 134, 105–114. doi: 10.1007/s10265-020-01236-6

Murray, B. G., and de Lange, P. J. (2011). "Chromosomes and evolution in New Zealand endemic angiosperms and gymnosperms" in *The biology of island floras*. eds. D. Bramwell and J. Caujapé-Castells (Cambridge: Cambridge University Press), 265–283.

Pellicer, J., and Leitch, I. J. (2020). The plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* 226, 301–305. doi: 10.1111/nph.16261

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2014). nlme: linear and nonlinear mixed effects models. R package version 3.1-117. *R Core Team (2014)*. Available at: https://CRAN.R-project.org/package=nlme (Accessed February 23, 2021).

Price, J. P. (2004). Floristic biogeography of the Hawaiian islands: influences of area, environment and paleogeography. *J. Biogeogr.* 31, 487–500. doi: 10.1046/j.0305-0270.2003.00990.x

Price, J. P., and Clague, D. A. (2002). How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proc. R. Soc. Lond. B Biol. Sci.* 269, 2429–2435. doi: 10.1098/rspb.2002.2175

Price, J. P., and Wagner, W. L. (2018). Origins of the Hawaiian flora: phylogenies and biogeography reveal patterns of long-distance dispersal. *J. Syst. Evol.* 56, 600–620. doi: 10.1111/jse.12465

Raskina, O., Barber, J. C., Nevo, E., and Belyayev, A. (2008). Repetitive DNA and chromosomal rearrangements: speciation-related events in plant genomes. *Cytogenet. Genome Res.* 120, 351–357. doi: 10.1159/000121084

Rice, A., Šmarda, P., Novosolov, M., Drori, M., Glick, L., Sabath, N., et al. (2019). The global biogeography of polyploid plants. *Nat. Ecol. Evol.* 3, 265–273. doi: 10.1038/s41559-018-0787-9

Rivas-Torres, G. F., Benítez, F. L., Rueda, D., Sevilla, C., and Mena, C. F. (2018). A methodology for mapping native and invasive vegetation coverage in archipelagos: an example from the Galápagos Islands. *Prog. Phys. Geogr.* 42, 83–111. doi: 10.1177/0309133317752278

Robertson, K., Goldberg, E. E., and Igić, B. (2011). Comparative evidence for the correlated evolution of polyploidy and self-compatibility in Solanaceae. *Evolution* 65, 139–155. doi: 10.1111/j.1558-5646.2010.01099.x

Rodriguez, D. J. (1996). A model for the establishment of polyploidy in plants. *Am. Nat.* 147, 33–46.

Rosche, C., Hensen, I., Mráz, P., Durka, W., Hartmann, M., and Lachmuth, S. (2017). Invasion success in polyploids: the role of inbreeding in the contrasting colonization abilities of diploid versus tetraploid populations of *Centaurea stoebe* s.l. *J. Ecol.* 105, 425–435. doi: 10.1111/1365-2745.12670

Rosenzweig, M. L. (1995). *Species diversity in space and time*. New York: Cambridge University Press.

Rosselló, J. A., and Castro, M. (2008). Karyological evolution of the angiosperm endemic flora of the Balearic Islands. *Taxon* 57, 259–273. doi: 10.2307/25065967

Sanmartín, I., Van Der Mark, P., and Ronquist, F. (2008). Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands. *J. Biogeogr.* 35, 428–449. doi: 10.1111/j.1365-2699.2008.01885.x

Sanmartín, I., Wanntorp, L., and Winkworth, R. C. (2007). West wind drift revisited: testing for directional dispersal in the southern hemisphere using event-based tree fitting. *J. Biogeogr.* 34, 398–416. doi: 10.1111/j.1365-2699.2006.01655.x

Savolainen, V., Anstett, M. -C., Lexer, C., Hutton, I., Clarkson, J. J., Norup, M. V., et al. (2006). Sympatric speciation in palms on an oceanic island. *Nature* 441, 210–213. doi: 10.1038/nature04566

Scarrow, M., Wang, Y., and Sun, G. (2020). Molecular regulatory mechanisms underlying the adaptability of polyploid plants. *Biol. Rev.* doi:10.1111/brv.12661

Scholl, J. P., and Wiens, J. J. (2016). Diversification rates and species richness across the tree of life. *Proc. R. Soc. B Biol. Sci.* 283:20161334. doi: 10.1098/rspb.2016.1334

Schönberger, I., Wilton, A. D., Boardman, K. F., Breitwieser, I., de Lange, P. J., de Pauw, B., et al. (2019). *Checklist of the New Zealand Flora – Seed plants*. Manaaki Whenua-Landcare Research: Lincoln.

Schüßler, C., Bräuchler, C., Reyes-Betancort, J. A., Koch, M. A., and Thiv, M. (2019). Island biogeography of the Macaronesian *Gesnouinia* and Mediterranean *Soleirolia* (Parietarieae, Urticaceae) with implications for the evolution of insular woodiness. *Taxon* 68, 537–556. doi: 10.1002/tax.12061

Smith, S. A., and Brown, J. W. (2018). Constructing a broadly inclusive seed plant phylogeny. *Am. J. Bot.* 105, 302–314. doi: 10.1002/ajb2.1019

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C. F., et al. (2009). Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348. doi: 10.3732/ajb.0800079

Soltis, P. S., Liu, X., Marchant, D. B., Visger, C. J., and Soltis, D. E. (2014). Polyploidy and novelty: Gottlieb's legacy. *Philos. Trans. R. Soc. Lond. B* 369:20130351. doi: 10.1098/rstb.2013.0351

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103, 1146–1166. doi: 10.3732/ajb.1500501

Soto-Trejo, F., Kelly, J. K., Archibald, J. K., Mort, M. E., Santos-Guerra, A., and Crawford, D. J. (2013). The genetics of self-compatibility and associated floral characters in *Tolpis* (Asteraceae) in the Canary Islands. *Int. J. Plant Sci.* 174, 171–178. doi: 10.1086/668788

Stebbins, G. L. (1947). Types of polyploids: their classification and significance. *Adv. Genet.* 1, 403–429. doi: 10.1016/s0065-2660(08)60490-3

Stuessy, T. F., and Crawford, D. J. (1998). "Chromosomal stasis during speciation in angiosperms of oceanic islands" in *Evolution and speciation of island plants*. eds. T. F. Stuessy and M. Ono (Cambridge: Cambridge University Press), 307–324.

Stuessy, T. F., Crawford, D. J., López-Sepúlveda, P., Baeza, C. M., and Ruiz, E. A. (eds.) (2017). *Plants of oceanic islands: Evolution, biogeography, and conservation of the flora of the Juan Fernández (Robinson Crusoe) archipelago*. (Cambridge: Cambridge University Press).

Stuessy, T. F., Crawford, D. J., Marticorena, C., and Rodríguez, R. (1998). "Island biogeography of angiosperms of the Juan Fernandez archipelago" in *Evolution and speciation of island plants*. eds. T. F. Stuessy and M. Ono (Cambridge: Cambridge University Press), 121–140.

Stuessy, T. F., Takayama, K., López-Sepúlveda, P., and Crawford, D. J. (2014). Interpretation of patterns of genetic variation in endemic plant species of oceanic islands. *Bot. J. Linn. Soc.* 174, 276–288. doi: 10.1111/boj.12088

Suda, J., Kyncl, T., and Jarolímová, V. (2005). Genome size variation in Macaronesian angiosperms: forty percent of the Canarian endemic flora completed. *Plant Syst. Evol.* 252, 215–238. doi: 10.1007/s00606-004-0280-6

Sun, B. -Y., and Stuessy, T. F. (1998). "Preliminary observations on the evolution of endemic angiosperms" in *Evolution and speciation of island plants*. eds. T. F. Stuessy and M. Ono (Cambridge: Cambridge University Press), 181–202.

Symonds, M. R. E., and Blomberg, S. P. (2014). "A primer on phylogenetic generalised least squares" in *Modern phylogenetic comparative methods and their application in evolutionary biology*. ed. L. Z. Garamszegi (Berlin, Heidelberg: Springer), 105–130.

Takayama, K., Crawford, D. J., López-Sepúlveda, P., Greimler, J., and Stuessy, T. F. (2018). Factors driving adaptive radiation in plants of oceanic islands: a case study from the Juan Fernández archipelago. *J. Plant Res.* 131, 469–485. doi: 10.1007/s10265-018-1023-z

Tanentzap, A. J., Brandt, A. J., Smissen, R. D., Heenan, P. B., Fukami, T., and Lee, W. G. (2015). When do plant radiations influence community assembly? The importance of historical contingency in the race for niche space. *New Phytol.* 207, 468–479. doi: 10.1111/nph.13362

Te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubesova, M., et al. (2012). The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.* 109, 19–45. doi: 10.1093/aob/mcr277

Tiley, G. P., Barker, M. S., and Burleigh, J. G. (2018). Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biol. Evol.* 10, 2882–2898. doi: 10.1093/gbe/evy200

The Plant List (2013). Published on the Internet. Available at: http://www.theplantlist.org/ (Accessed February 23, 2021).

Vamosi, J. C., Goring, S. J., Kennedy, B. F., Mayberry, R. J., Moray, C. M., Neame, L. A., et al. (2007). Pollination, floral display, and the ecological correlates of polyploidy. *Funct. Ecosyst. Communities* 1, 1–9.

Vamosi, J. C., Magallón, S., Mayrose, I., Otto, S. P., and Sauquet, H. (2018). Macroevolutionary patterns of flowering plant speciation and extinction. *Annu. Rev. Plant Biol.* 69, 685–706. doi: 10.1146/annurev-arplant-042817-040348

Vargas, P., Arjona, Y., Nogales, M., and Heleno, R. H. (2015). Long-distance dispersal to oceanic islands: success of plants with multiple diaspore specializations. *AoB PLANTS* 7:plv073. doi: 10.1093/aobpla/plv073

Vitales, D., Garnatje, T., Pellicer, J., Vallès, J., Santos-Guerra, A., and Sanmartín, I. (2014). The explosive radiation of *Cheirolophus* (Asteraceae, Cardueae) in Macaronesia. *BMC Evol. Biol.* 14:118. doi: 10.1186/1471-2148-14-118

Wagner, W. L., Herbst, D. R., and Lorence, D. H. (2005). *Flora of the Hawaiian Islands website* [Online]. Available at: http://botany.si.edu/pacificislandbiodiversity/hawaiianflora/index.htm (Accessed January 1, 2020).

Wagstaff, S. J., and Garnock-Jones, P. J. (1998). Evolution and biogeography of the *Hebe* complex (Scrophulariaceae) inferred from ITS sequences. *N. Z. J. Bot.* 36, 425–437. doi: 10.1080/0028825X.1998.9512581

Weigelt, P., Jetz, W., and Kreft, H. (2013). Bioclimatic and physical characterization of the world's islands. *Proc. Natl. Acad. Sci.* 110, 15307–15312. doi: 10.1073/pnas.1306309110

Wood, T., (2009). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci.* 106, 13875–13879. doi: 10.1073/pnas.0811575106

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 June 2021
doi: 10.3389/fpls.2021.679047

Trajectories of Homoeolog-Specific Expression in Allotetraploid *Tragopogon castellanus* Populations of Independent Origins

J. Lucas Boatwright[1]*, *Cheng-Ting Yeh*[2], *Heng-Cheng Hu*[2,3], *Alfonso Susanna*[4], *Douglas E. Soltis*[5,6,7,8,9], *Pamela S. Soltis*[6,7,8,9], *Patrick S. Schnable*[2] and *William B. Barbazuk*[5]

[1] Advanced Plant Technology Program, Clemson University, Clemson, SC, United States, [2] Department of Agronomy, Iowa State University, Ames, IA, United States, [3] Covance Inc., Indianapolis, IN, United States, [4] Botanic Institute of Barcelona, Consejo Superior de Investigaciones Científicas, ICUB, Barcelona, Spain, [5] Department of Biology, University of Florida, Gainesville, FL, United States, [6] Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, FL, United States, [7] Florida Museum of Natural History, University of Florida, Gainesville, FL, United States, [8] Genetics Institute, University of Florida, Gainesville, FL, United States, [9] Biodiversity Institute, University of Florida, Gainesville, FL, United States

OPEN ACCESS

Edited by:
Elvira Hörandl, University of Göttingen, Germany

Reviewed by:
Malika Ainouche, University of Rennes 1, France
Tom A. Ranker, University of Hawaii at Manoa, United States

***Correspondence:**
J. Lucas Boatwright
jboatw2@clemson.edu

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *10 March 2021*
Accepted: *20 May 2021*
Published: *23 June 2021*

Citation:
Boatwright JL, Yeh C-T, Hu H-C, Susanna A, Soltis DE, Soltis PS, Schnable PS and Barbazuk WB (2021) Trajectories of Homoeolog-Specific Expression in Allotetraploid Tragopogon castellanus Populations of Independent Origins. Front. Plant Sci. 12:679047. doi: 10.3389/fpls.2021.679047

Polyploidization can have a significant ecological and evolutionary impact by providing substantially more genetic material that may result in novel phenotypes upon which selection may act. While the effects of polyploidization are broadly reviewed across the plant tree of life, the reproducibility of these effects within naturally occurring, independently formed polyploids is poorly characterized. The flowering plant genus *Tragopogon* (Asteraceae) offers a rare glimpse into the intricacies of repeated allopolyploid formation with both nascent (< 90 years old) and more ancient (mesopolyploids) formations. Neo- and mesopolyploids in *Tragopogon* have formed repeatedly and have extant diploid progenitors that facilitate the comparison of genome evolution after polyploidization across a broad span of evolutionary time. Here, we examine four independently formed lineages of the mesopolyploid *Tragopogon castellanus* for homoeolog expression changes and fractionation after polyploidization. We show that expression changes are remarkably similar among these independently formed polyploid populations with large convergence among expressed loci, moderate convergence among loci lost, and stochastic silencing. We further compare and contrast these results for *T. castellanus* with two nascent *Tragopogon* allopolyploids. While homoeolog expression bias was balanced in both nascent polyploids and *T. castellanus*, the degree of additive expression was significantly different, with the mesopolyploid populations demonstrating more non-additive expression. We suggest that gene dosage and expression noise minimization may play a prominent role in regulating gene expression patterns immediately after allopolyploidization as well as deeper into time, and these patterns are conserved across independent polyploid lineages.

Keywords: *Tragopogon*, allopolyploid, additive expression, homoeologs, expression bias, homoeolog-specific expression, non-model system, RNA-Seq

1. INTRODUCTION

The consequences of plant polyploidization have been a subject of intense interest for several decades (reviewed in Wendel, 2000, 2015; Doyle et al., 2008; Leitch and Leitch, 2008; Van de Peer et al., 2009; Barker et al., 2012; Soltis et al., 2016). Polyploidization results in broad-scale genomic changes that serve as potentially novel avenues upon which evolution may act (reviewed in Otto and Whitton, 2000; Flagel and Wendel, 2009). Many changes occur in the generations immediately after polyploidization including changes in genome size (reviewed in Soltis et al., 2003; Leitch et al., 2008; Leitch and Leitch, 2013) spanning the extremes in both gain (e.g., *Paris japonica* Pellicer et al., 2010) and loss (e.g., *Utricularia gibba* Ibarra-Laclette et al., 2013), expression (Chen and Pikaard, 1997; reviewed in Adams and Wendel, 2005b; Chaudhary et al., 2009; Hu et al., 2015), epigenetic modifications (Shaked et al., 2001; Salmon et al., 2005; reviewed in Chen, 2007; Madlung and Wendel, 2013; Cheng et al., 2016), transposon activity (reviewed in Woodhouse et al., 2014; Vicient and Casacuberta, 2017; Wendel et al., 2018) as well as changes in protein folding and dosage (reviewed in Birchler and Veitia, 2010, 2012; Pires and Conant, 2016). These changes are variable across lineages (Anssour and Baldwin, 2010; reviewed in Soltis et al., 2016) and may occur in repeated cycles (Soltis and Soltis, 1999; Buggs et al., 2012; reviewed in Wendel, 2015; Soltis et al., 2016). In some paleopolyploids, these changes appear to largely converge over time, at least within closely related lineages (Blanc and Wolfe, 2004; reviewed in Barker et al., 2008; Edger and Pires, 2009; Freeling, 2009).

Polyploids are categorized as either autopolyploids, which are formed from a whole-genome duplication within a single species (reviewed in Otto and Whitton, 2000; Spoelhof et al., 2017), or allopolyploids, which are generated by the combination of entire genomes from two different species (Kihara and Ono, 1926). However, these definitions represent an oversimplification of the dynamic range of variability that polyploids may cover (reviewed in Stebbins, 1947; Ramsey and Schemske, 1998) and the various mechanisms by which they are formed (reviewed in Mason and Pires, 2015). Allopolyploid formation results in duplicated gene copies originating from each parent known as homoeologs. Immediately after polyploidization, homoeologs are expected to be functionally redundant and as such, one copy may be altered without deleterious effect or conserved in duplicate (reviewed in Conant et al., 2014; Pires and Conant, 2016). Whole-genome duplication in an organism can impose unfavorable dosage effects upon cellular functions unless gene balance is maintained (Freeling, 2009; Birchler and Veitia, 2010, 2012). These dosage effects likely represent one aspect of a larger framework that directs genome evolution after polyploidization (Conant et al., 2014). As such, duplicate loci in allopolyploids may experience a number of possible fates. Genomes may experience silencing or loss of one homoeologous copy via fractionation over time. Homoeolog functions may diverge from the parentally inherited state such that functions are partitioned between homoeologs (subfunctionalization), or copies may develop novel functionality (neofunctionalization) (reviewed in Edger and Pires, 2009; Freeling, 2009). Homoeologs may also interact via convergent evolution, homoeologous recombination or gene conversion (Langham et al., 2004; Doyle et al., 2008).

Expression patterns may also vary in the polyploid such that loci demonstrate spatiotemporally divergent expression from the progenitors (Pires et al., 2004b; Wang et al., 2006a; Buggs et al., 2010b; Baldauf et al., 2016), homoeolog-specific expression (HSE) (Buggs et al., 2010a; reviewed in Grover et al., 2012; Yoo et al., 2013; Woodhouse et al., 2014) or additive expression (Guo et al., 2006; Stupar and Springer, 2006; Wang et al., 2006b; reviewed in Yoo et al., 2014). HSE occurs when the polyploid expresses one parental homoeolog over the other (Woodhouse et al., 2014; Boatwright et al., 2018). HSE is similar to allele-specific expression in that both refer to expression differences that are caused by *cis*- and *trans*-regulatory variation (Bell et al., 2013), and each has been a topic of interest in hybrid and polyploid studies (Wright et al., 1998; Adams and Wendel, 2005a; Aguilar-Rangel et al., 2017). HSE differs from allele-specific expression in that HSE examines expression across homoeologous chromosomes in contrast to allele-specific expression, which examines expression between homologous chromosomes. Homoeolog expression may also diverge in an additive manner where expression in the polyploid is the arithmetic mean of the two diploid progenitors (reviewed in Yoo et al., 2014).

It is worth noting that the degree of similarity/dissimilarity in expression between parents of a polyploid and the polyploid itself, also known as parental legacy (Buggs et al., 2014), may have a significant effect upon the fate of homoeolog expression in the polyploid (Conant et al., 2014). Similarly, differences among polyploids and their diploid progenitors may derive from numerous processes such as divergent evolution of the lineages after polyploidization, effects of whole-genome duplication (i.e., larger cells and stomata, higher photosynthetic rates and gas exchange, and different stress tolerance) (Hegarty et al., 2006; Sémon and Wolfe, 2007; De Smet and Van de Peer, 2012), or hybridization (resulting in heterosis, increased genetic variation and additive expression) (Mallet, 2007; Bell et al., 2013; Soltis et al., 2016). While the fates of homoeologs after polyploidization are convergent within some lineages (Blanc and Wolfe, 2004; reviewed in Edger and Pires, 2009; and Freeling, 2009), establishing a paradigm has proved elusive (reviewed in Soltis et al., 2016).

The evolutionary model organism *Tragopogon* serves as a prominent example of repeated, naturally occurring allopolyploidization. The *Tragopogon* system includes synthetic lines, nascent (< 90 years) and meso- (∼2.6 million years) polyploids (Mavrodiev et al., 2015; Soltis et al., 2016). While most species of *Tragopogon* have chromosome numbers of $2n = 12$, there are several well-studied allopolyploids ($2n = 24$), *Tragopogon mirus*, *T. miscellus*, and *T. castellanus*. Both *T. mirus* and *T. miscellus* represent neoallotetraploids that formed recently in the northwestern United States after their three diploid progenitors (*T. dubius*-*T. porrifolius* and *T. dubius*-*T. pratensis*, respectively) were introduced from Europe in the early 1900s (Ownbey, 1950; Soltis et al., 2004). These allopolyploids never formed in Europe due to their geographic isolation but

have formed repeatedly in the USA since the diploids were brought into close proximity. These polyploids are estimated to be approximately 45 generations old (80–90 years for these biennials) (Soltis et al., 1995; Symonds et al., 2010).

Similarly, *T. castellanus* has formed multiple times from independent allopolyploidization events (Mavrodiev et al., 2015). *Tragopogon castellanus* is endemic to Spain and occurs only along the northern half of the Iberian Peninsula (Blanca and de la Guardia, 1996). Morphological, cytological, and molecular phylogenetic analyses support *T. lamottei* and *T. crocifolius* as putative parents. *Tragopogon castellanus* is morphologically variable and somewhat similar to parental *T. crocifolius*; as a result *T. castellanus* was once considered a subspecies of *T. crocifolius* (Willkomm, 1893; de la Guardia Guerrero and López, 1989). The parentage of *T. castellanus* was validated using phylogenetic analyses of external transcribed spacers, internal transcribed spacers, *Adh*, and plastid datasets, fluorescence *in situ* hybridization, and genome *in situ* hybridization (Mavrodiev et al., 2008, 2015). *Tragopogon castellanus* may have formed before the last glacial maximum that would date the formation of this polyploid species to perhaps as long ago as 2.6 million years (Mavrodiev et al., 2015). As such, the multiple, independent occurrences of *Tragopogon* allopolyploid formation in young US species and the older *T. castellanus* permits the assessment of the fate of homoeologs in both neo- and mesopolyploids.

Previous studies have demonstrated that duplicate gene fates after polyploidization are non-random within the newly formed allopolyploid species of *Tragopogon*. That is, many gene loss and expression changes were repeated across polyploid populations of independent origins (Buggs et al., 2012; Soltis et al., 2012). However, these studies were primarily small-scale and the fates of duplicated gene copies do not generalize across all polyploid plants (reviewed in Soltis et al., 2016). Here, we examine multiple, independently formed allopolyploid *T. castellanus* lineages estimated to have formed as long as 2.6 mya (Mavrodiev et al., 2015). We show that not only are expression patterns similar, but duplicate loss is largely convergent across independent lineages of *T. castellanus*. We further compare duplicate fates in populations of this mesopolyploid from Spain to the neopolyploids from the US, *T. mirus* and *T. miscellus* (based on earlier studies; Buggs et al., 2010a,b, 2012; Boatwright et al., 2018) in which identical methods were used so that we may examine changes due to natural allopolyploidization over a large evolutionary time scale of perhaps several million years.

2. MATERIALS AND METHODS

2.1. Sample Processing

Leaf tissue was collected from plants grown in controlled conditions as described by Tate et al. (2006), and RNA was extracted as described in Tate et al. (2006). Three individuals of the diploid *T. crocifolius* were sampled from the P-B lineage along with five individuals of diploid *T. lamottei* composed of two and three individuals/lineages from lineage P-I and P-II lineages, respectively (Mavrodiev et al., 2015). Both parental species are phylogenetically distinct and appeared as members of two distinct clades based on ITS phylogeny as estimated in Mavrodiev et al. (2005); namely, clade Majores s. l. [incl. clade Hebecarpus] (*T. crocifolius*) and clade Tragopogon (*T. lamottei*) (Mavrodiev et al., 2008). Sample localities and voucher information for all samples are given in **Supplementary Table 1**. Additional information is provided in Mavrodiev et al. (2015) and vouchers are deposited at the University of Florida herbarium (FLAS). We sequenced 12 allotetraploid *T. castellanus* individuals representing three bioreplicates for four independent polyploidization events (**Supplementary Figure 1** and **Supplementary Table 1**). RNA-Seq samples were bar-coded and processed using the Illumina TruSeq kit.

2.2. Sequencing, Assembly and Annotation

Samples were sequenced using the Illumina MiSeq sequencing platform resulting in 2 X 300 paired-end reads (**Supplementary Table 2**). Adapters were removed using CutAdapt (Martin, 2011), and sequences were trimmed using Trimmomatic (Bolger et al., 2014). RNA reads were pooled from all individuals of each diploid species and assembled using the Trinity *de novo* assembler (Grabherr et al., 2011), resulting in one assembly per species. Redundant isoforms were removed from our assemblies using a previously described pipeline (Boatwright et al., 2018). The final assemblies were annotated using Trinotate (Altschul et al., 1990; Ashburner et al., 2000; Krogh et al., 2001; Lagesen et al., 2007; Finn et al., 2011; Grabherr et al., 2011; Kanehisa et al., 2011; Petersen et al., 2011; Powell et al., 2011; Punta et al., 2011) with default parameters (https://github.com/jlboat/Tragopogon_castellanus).

2.3. Ortholog Identification and Common Orthologous Regions

Putative orthologs were identified between the *T. crocifolius* and *T. lamottei* assemblies using a reciprocal best-hit approach (Moreno-Hagelsieb and Latimer, 2008) as described in Boatwright et al. (2018). Common Orthologous REgions (COREs) were identified between orthologous pairs using the local alignment provided by WU-BLAST (Gish, 2005) and a custom CPython script (https://github.com/BBarbazukLab/papers/). This resulted in BED files containing CORE s that were used to filter BAM files after aligning reads to complete assemblies (Boatwright et al., 2018).

2.4. Poisson-Gamma Model

As in Boatwright et al. (2018), parental RNA-Seq reads were mapped to both complete, diploid references independently using Bowtie v0.12.9, -m1,-v 3] (Langmead et al., 2009) and Last [v531, -l 25] (Frith et al., 2010; Graze et al., 2012; Munger et al., 2014). The BED files containing CORE s were used to filter the resulting SAM files for respective references. Parental reads that mapped uniquely within CORE s were isolated, and the reads were subsequently identified as mapping equally well to both references or better to one of the two parents. A Bayesian Poisson-Gamma model (León-Novelo et al., 2014), which provides conservative estimates of the type I error (Fear et al., 2016), was used to identify CORE s biasedly mapping reads from the alternative parent. CORE s demonstrated expression

bias if the credible interval did not overlap 0.5 for all priors (0.4, 0.5, 0.6). Polyploid reads were mapped following the same procedure, and the biased COREs—as determined by diploid read mapping—were filtered out after processing, leaving the remaining set of unbiased COREs for inference. Within the set of unbiased COREs, remaining expression bias corresponded to loci demonstrating HSE. Overlapping gene sets were visually displayed using UpSet plots generated using R (Team, 2014) and the UpSetR package (Lex et al., 2014).

2.5. Additively Expressed Genes

Reads mapping to both references within COREs were used to generate an expression matrix for diploids and all independent polyploids. Loci were filtered from the expression matrix that did not contain at least 10 counts-per-million based upon the average library size in 11 samples. Differentially expressed genes were identified in R (Team, 2014) using the empirical Bayesian analysis pipeline within the limma package (Ritchie et al., 2015) after using voom (Law et al., 2014) to apply precision weights to account for the mean-variance trend. Loci were considered differentially expressed at a false discovery rate of 0.05 (Benjamini and Hochberg, 1995). Contrasts were performed between *T. lamottei* and *T. crocifolius* to determine when parental expression was the same or different. To test for additivity, contrasts were performed between each population of the polyploid *T. castellanus* and its two parents where polyploid expression is expected to be the arithmetic mean of the two parental expression levels. Overlapping gene sets were visually displayed using UpSet plots generated using R (Team, 2014) and the UpSetR package (Lex et al., 2014).

2.6. Homoeolog Loss and Silencing

Orthologs were used to design probes for NimbleGen sequence capture to isolate genomic reads from allopolyploid *T. castellanus* individuals. Each probe was designed to target unique regions of each contig with 1-3 probes along each contig. These probes were used to isolate genomic DNA corresponding to expressed transcripts (**Supplementary Table 3**). Polyploid DNA reads obtained from sequence capture were aligned to diploid references in the same manner as polyploid RNA reads, and homoeolog loss and silencing were assessed within COREs using the unbiased homoeolog set. Homoeologs with mapped DNA and mapped RNA reads represent genes that are both present and expressed. Homoeologs with DNA reads and no RNA reads represent putative silencing events. Homoeologs with RNA reads but no DNA reads likely represent a failed capture or mismapped reads. Homoeologs with neither DNA nor RNA reads represent putative loss. Overlapping gene sets were visually displayed using UpSet plots generated using R (Team, 2014) and the UpSetR package (Lex et al., 2014).

2.7. Functional Protein Association Network and GO Enrichment

Loci common to all independently formed polyploid populations that were lost, exhibited additive or non-additive expression as well as loci that demonstrated HSE were individually tested for interaction enrichment. *Arabidopsis thaliana* orthologs were identified for *Tragopogon* contigs using WU-BLAST blastx with an *A. thaliana* protein database. The *E*-value cutoff was set at 1E-75, and the high-scoring segment pair had to represent 70% of either the total query or subject length. The resulting lists of *A. thaliana* genes were used to construct functional protein association networks using STRING10 (Szklarczyk et al., 2014). The resulting networks used only high-confidence, experimentally validated protein-protein interactions with disconnected nodes in the networks hidden, and the edge thickness represented confidence of data supporting interaction. Protein-protein interaction enrichment *p*-values were FDR corrected (Benjamini and Hochberg, 1995). All gene sets were further checked for GO enrichment using GOSeq (Young et al., 2012). The background for HSE and lost loci was the set of unbiased COREs. The background for additively expressed genes included all loci tested for additivity. Functional network details and GO annotations are available at (https://github.com/jlboat/Tragopogon_castellanus).

3. RESULTS

3.1. Assembly, Annotation and Ortholog Identification

The assemblies of the diploids, *T. crocifolius* and *T. lamottei*, contained 113,865 and 155,600 contigs, respectively. Assemblies were annotated using Trinotate, and putative orthologs and domains were identified. For each of the diploid species, over 7,000 entries hit *Arabidopsis thaliana* sequences, and approximately 500 of the remainder hit *Oryza sativa* ssp. *japonica* using NCBI-BLAST against the SwissProt database. We also identified approximately 4,800 unique eggNOG hits for each diploid, where eggNOG hits represent hierarchical orthologous groups and provide functional annotations for homologous sequences. We identified 14,388 orthologs between the diploid genomes and delimited COREs for downstream processing (Gish, 2005; Moreno-Hagelsieb and Latimer, 2008). COREs were assessed for similarity in both length and GC content (**Supplementary Figures 1, 2**) and were found to be highly similar between the two species, with length differences never exceeding 16 bases and GC content differences primarily falling under 2%.

3.2. Additive Expression

Additivity was assessed by first performing a contrast between diploid parents to identify loci where parental expression deviated or was the same (**Table 1** and **Supplementary Figure 4**). The matrix of counts used to estimate additive expression was subjected to multi-dimensional scaling, where samples that lie in close proximity exhibit more similar expression patterns, and plotted. The clustering of *T. castellanus* individuals is consistent with the known lineages as samples for Cast_2 and Cast_10 both come from lineage I and cluster together (**Figures 1**, **2**, and **Supplementary Figure 1**). Similarly, the *T. lamottei* individuals Lam_1 and Lam_2 come from the same lineage, P-I, and are adjacent. Of the 5,806 loci remaining after filtering, parental expression was the same at 4,533 loci and different at 1,273. We found that polyploid expression is primarily non-additive where

TABLE 1 | Test for additivity in polyploid expression.

	Not additive	Consistent with additive
Cast_2		
Parents same	2,762 (50.6%)	1,516 (27.8%)*
Parents different	787 (14.4%)	391 (7.2%)**
Cast_10		
Parents same	2612 (47.9%)	1,666 (30.5%)*
Parents different	757 (13.9%)	421 (7.7%)**
Cast_13		
Parents same	2,515 (46.1%)	1,763 (32.3%)*
Parents different	710 (13.0%)	468 (8.6%)**
Cast_13		
Parents same	2,466 (45.2%)	1,812 (33.2%)*
Parents different	744 (13.6%)	434 (8.0%)**

*Counts represent loci where parental expression is not significantly different or is significantly different and polyploid expression is either additive or non-additive in the T. castellanus individuals of independent origin. Percentages are based on per-individual totals such that the expression categories of each individual sum to 100%. * These loci are not strictly additive as T. castellanus expression could deviate from mid-parent expression and yet be consistent with additive when parental expression is the same. ** These loci have power issues because the hybrid mean expression falls within the diploid mean expression levels.*

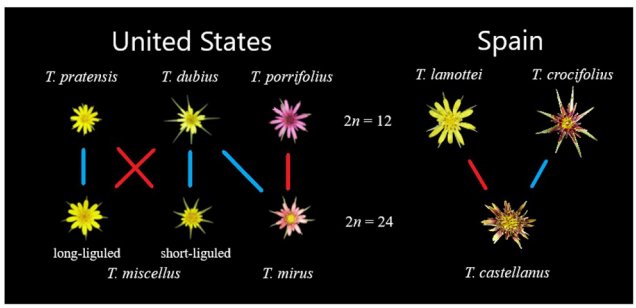

FIGURE 1 | Relationships among the *Tragopogon* diploids and polyploids. US species are left of the chromosome counts and Spanish species are on the right. Diploids are aligned along the top row and polyploid offspring are along the bottom row. Colored lines indicate whether the diploid serves as the maternal or paternal parent for the corresponding polyploid, where blue is paternal and red is maternal.

additivity was examined with respect to parental expression (**Table 1**), and overlap of each additive/non-additive category was assessed (**Figure 3**). Approximately 65% (2,155) of the loci were not additive in all four of the independent polyploids, whereas only 43% (909) of the loci consistently exhibited additivity over the four polyploids. There was no significant (FDR < 0.05) GO enrichment for shared additively expressed loci.

3.3. Homoeolog-Specific Expression

HSE was assessed using the PG model and, similar to the additivity assessment, was examined in light of parental expression using unbiased COREs (**Table 2**). The number of polyploid loci exhibiting homoeolog expression bias toward each parent was similar, with a moderate, consistent bias toward *T. crocifolius* of about 50 loci, which accounts for ∼7% of

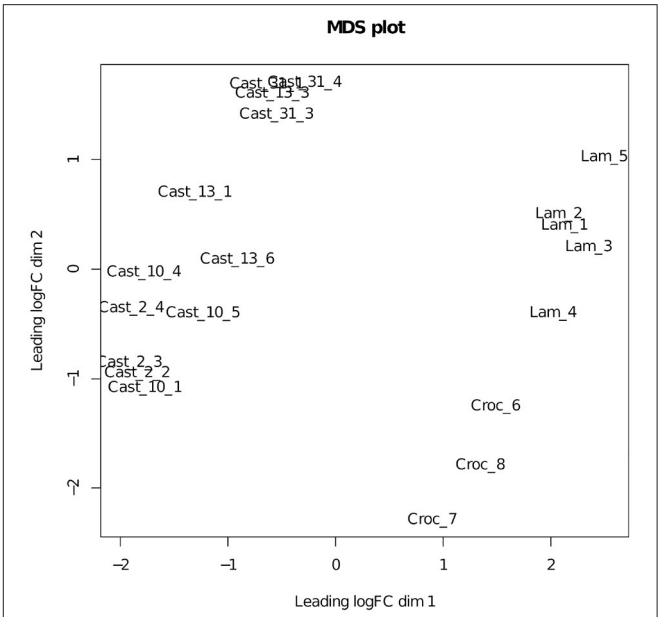

FIGURE 2 | MDS plot of the additive expression matrix for *T. castellanus* and its diploid progenitors. Lam represents *T. lamottei*, Croc represents *T. crocifolius*, and Cast represents *T. castellanus*.

loci in which parental gene expression is the same but ∼23% of loci which exhibit significantly non-equal expression in the parents. The percent of loci overlapping between independent polyploids was ∼60% when parental expression was the same and ∼64% when parental expression differed (**Figure 4**). There was no significant (FDR < 0.05) difference in GO enrichment for common loci demonstrating HSE.

3.4. Homoeolog Silencing and Loss

After orthologs were identified between diploid assemblies, exon-capture probes were designed so that genomic data could be used to distinguish between loci lost vs. silenced after polyploidization. As seen with both additivity and HSE, the number of loci expressed, silenced or lost is highly consistent across all polyploids of independent origin (**Table 3**). However, the degree of overlap varies among expressed, silenced or lost loci. For expressed loci, approximately 95% of the same parentally derived homoeologs (4,113 for *T. lamottei* and 4,054 for *T. crocifolius*) overlap among the four polyploids (**Figure 5**). Of those few homoeologs demonstrating loss, approximately 66% overlap, again, for both *T. lamottei* (92) and *T. crocifolius* (99) homoeologs, independently (**Figure 6**).

Silenced homoeologs showed the most variability even though a similar number of loss events occurred across all independently formed polyploids. Only 14 *T. lamottei*-derived homoeologs (∼10% of all *T. lamottei* homoeolog silencing events) and 35 *T. crocifolius*-derived homoeologs (∼25% of all *T. crocifolius* homoeolog silencing events) were silenced in all four polyploids. In fact, the majority of silencing events were unique to each polyploid, suggesting that silencing is likely a much more stochastic process than

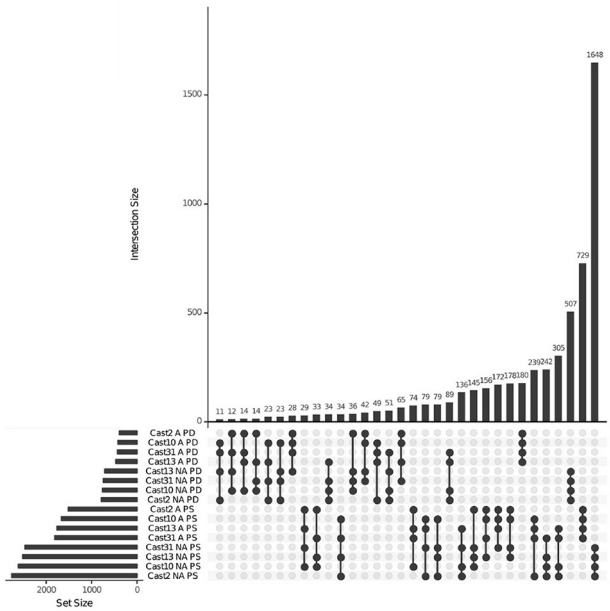

FIGURE 3 | Additive and non-additive expression overlap across *T. castellanus* individuals. Set IDs represent samples that are additive (A) or non-additive (NA) where parents are different (PD) or the same (PS). Sample sets with common loci are indicated by filled circles with connecting lines, and the number of loci within that intersection may be seen directly above in the bar chart with corresponding size over each bar. The total sample sizes are found in the left bar chart and correspond to the adjacent sample.

TABLE 2 | Homoeolog-Specific Expression.

	T. lamottei	*T. crocifolius*	No HSE
(A)			
Parents same			
Cast_2	774	856	1,400
Cast_10	827	881	1,346
Cast_13	789	824	1,429
Cast_31	805	858	1,397
(B)			
Parents different			
Cast_2	225	286	324
Cast_10	233	293	310
Cast_13	218	285	328
Cast_31	219	304	310
(C)			
Ignoring parents			
Cast_2	1,274	1,429	2,360
Cast_10	1,365	1,481	2,279
Cast_13	1,291	1,387	2,397
Cast_31	1,327	1,447	2,344

*Counts represent total number of loci demonstrating expression bias toward a particular parental homoeolog. Homoeolog expression biases are examined in light of **(A)** loci expression levels being the same in the diploid parents, **(B)** loci expression levels being different in the diploid parents, and **(C)** HSE if parental patterns are ignored; these numbers are higher due to filtering constraints used to determine differences in parental expression.*

homoeolog loss (**Figures 7**, **8**). There was no significant (FDR < 0.05) GO enrichment for loci expressed, lost or silenced.

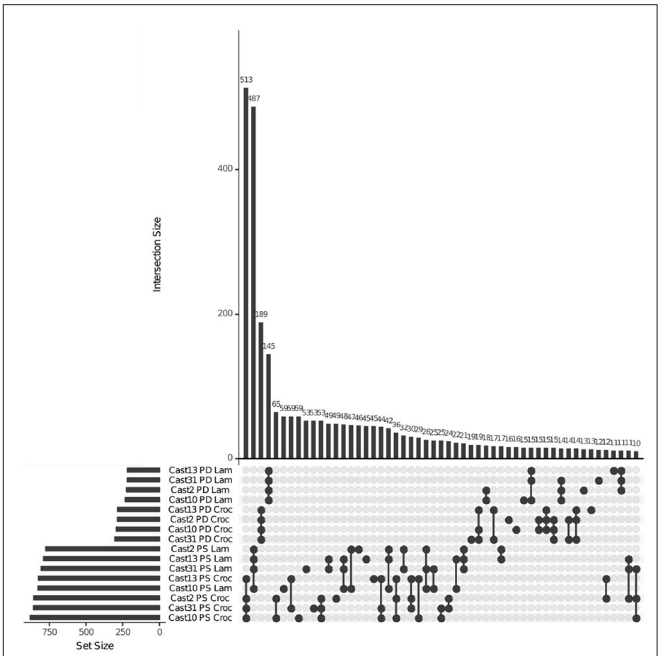

FIGURE 4 | Homoeolog-specific expression overlap across *T. castellanus*. Set IDs represent samples where parents are different (PD) or the same (PS) and the direction of homoeolog-specific expression (i.e. toward higher expression of the *T. lamottei* (Lam) or *T. crocifolius* (Croc) homoeolog). Overlapping sets are indicated by filled circles, and the number of loci within that intersection may be seen directly above in the bar chart with corresponding size over each bar. The total sample sizes are found in the left bar chart and correspond to the adjacent sample.

3.5. Functional Protein Association Network

The only group of loci that demonstrated significantly more interactions than expected based on chance included the additively expressed genes common to all independently formed allopolyploids (**Supplementary Figure 5**). The resulting network included 787 nodes with 185 edges with an expected edge count of 101. The average node degree was 0.47 with an average local clustering coefficient of 0.132. The FDR-corrected *q*-value was 2.75e-13.

4. DISCUSSION

4.1. Assembly and Annotation

Transcript assembly sizes were similar in these *Tragopogon* diploids from Spain (113,865 and 155,600 contigs for *T. crocifolius* and *T. lamottei*, respectively) to those seen in the US diploid parental species (105,282, 116,777, and 122,024 for *T. dubius*, *T. porrifolius*, and *T. pratensis*, respectively) (Boatwright et al., 2018). The number of orthologous pairs identified between diploid progenitors was also similar, with 14,389 pairs identified in this study between *T. lamottei* and *T. crocifolius*, while US species were represented by 15,493 pairs between *T. dubius*-*T. pratensis* and 15,587 between *T. dubius*-*T. porrifolius*. Differences in CORE lengths were similar between studies with no difference

TABLE 3 | Expression states of homoeologs derived from *T. lamottei* or *T. crocifolius*.

	T. lamottei				*T. crocifolius*			
	Expressed	**Silenced**	**Lost**	**Failed**	**Expressed**	**Silenced**	**Lost**	**Failed**
Cast_2	4,292	142	134	109	4,225	160	139	153
Cast_10	4,360	121	138	105	4,291	139	146	148
Cast_13	4,310	129	136	112	4,248	137	161	141
Cast_31	4,352	121	145	105	4,287	135	148	153

Counts represent total number of loci exhibiting the specified expression state. T. lamottei- and T. crocifolius-derived homoeologs may be present and expressed, silenced, lost or have failed either to be isolated using sequence capture probes or spuriously mismapped reads.

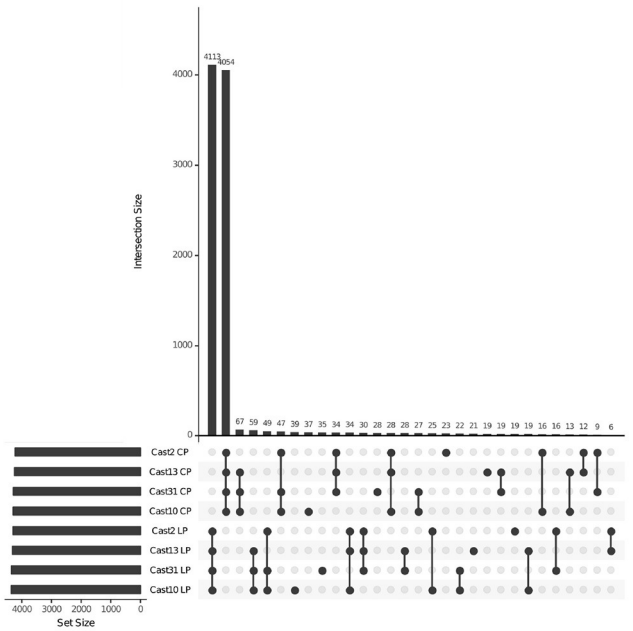

FIGURE 5 | UpSet plot showing homoeologs mapping both DNA and RNA reads across *T. castellanus* individuals. Set IDs represent samples where homoeologs from *T. crocifolius* (C) or *T. lamottei* (L) are present (P) based upon both DNA and RNA alignment. Overlapping sets are indicated by filled circles, and the number of loci within that intersection may be seen directly above in the bar chart with corresponding size over each bar. The total sample sizes are found in the left bar chart and correspond to the adjacent sample. As expected, there should be no overlap across *T. crocifolius* and *T. lamottei* homoeologs.

over 16 bp, while the %GC difference never exceeded 5% (**Supplementary Figures 2, 3**). The number of hits against the SwissProt database was nearly identical for all diploid assemblies (∼7,000 to *A. thaliana* and ∼500 to *O. sativa* for both US and Spanish species) as were the numbers of unique eggNOG hits (∼4,800). These metrics are significant in that they demonstrate that these studies contain large, similarly sized and comparable data. Thus, differences between the studies should largely be due to biological differences and not methodological differences.

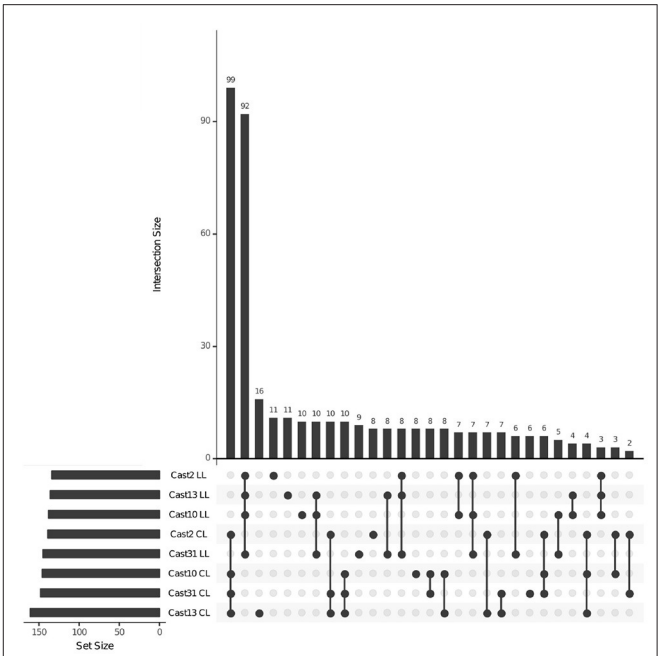

FIGURE 6 | Loci lost across *T. castellanus individuals*. Set IDs represent samples where either *T. crocifolius* (C) or *T. lamottei* (L) homoeologs are lost (L). Overlapping sets are indicated by filled circles, and the number of loci within that intersection may be seen directly above in the bar chart with corresponding size over each bar. The total sample sizes are found in the left bar chart and correspond to the adjacent sample. As expected, there should be no overlap across *T. crocifolius* and *T. lamottei* homoeologs.

4.2. Additive Expression

Additive and non-additive gene expression patterns are commonly studied in hybrid and polyploid plants (Guo et al., 2006; Stupar and Springer, 2006; Swanson-Wagner et al., 2006, 2009; Wang et al., 2006a,b; Baldauf et al., 2016). Synthetic *Brassica napus* exhibits proteome additivity where differential regulation was not related to protein function (Albertin et al., 2007). Additive protein expression has been previously described in the neopolyploid *Tragopogon mirus* (Koh et al., 2012), and additive gene expression in both neopolyploids *T. mirus* and *T. miscellus* (Boatwright et al., 2018).

However, expression in the diploid parents of *T. castellanus* is significantly different from that seen in the parents of the nascent polyploids. For the neopolyploids, the expression of the diploids *T. dubius* and *T. porrifolius* was the same for 5,806 loci and different for 4,706 loci; *T. dubius* and *T. pratensis* expression was the same for 5,026 and different for 5,121 loci (Boatwright et al., 2018). While expression was different between diploid parents for the neopolyploids about 50% of the time, the values seen here for *T. crocifolius* and *T. lamottei* indicate that parental expression is primarily the same. Similarly, whereas the homoeolog expression was consistent with additivity for the majority of loci within the neopolyploids (Boatwright et al., 2018), plants of the mesopolyploid *T. castellanus* exhibit more non-additive expression.

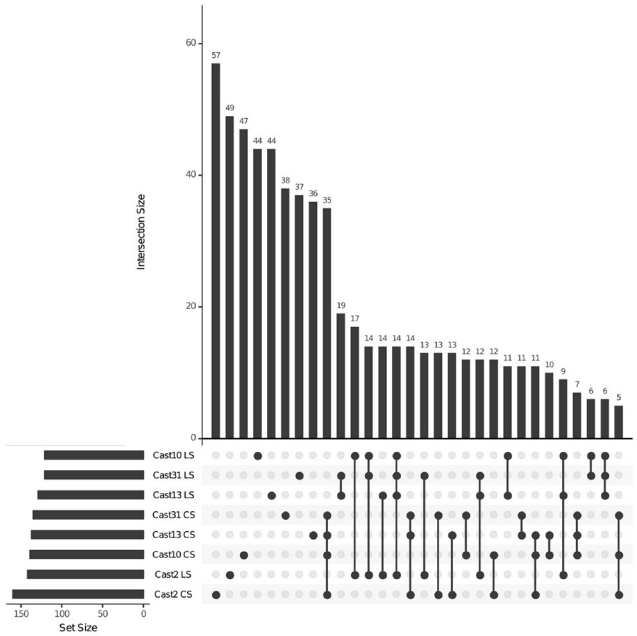

FIGURE 7 | Loci silenced across *T. castellanus* individuals. Set IDs represent samples where either *T. crocifolius* (C) or *T. lamottei* (L) homoeologs are silenced (S). Overlapping sets are indicated by filled circles, and the number of loci within that intersection may be seen directly above in the bar chart with corresponding size over each bar. The total sample sizes are found in the left bar chart and correspond to the adjacent sample. As expected, there should be no overlap across *T. crocifolius* and *T. lamottei* homoeologs.

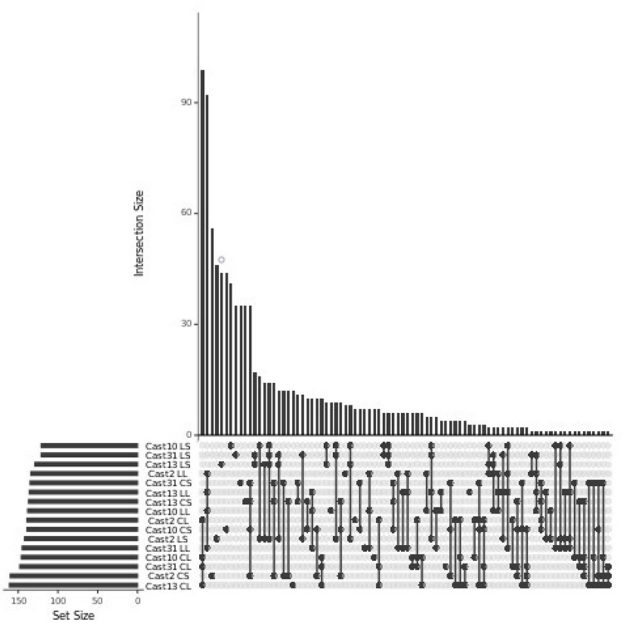

FIGURE 8 | Loci lost or silenced across *T. castellanus* individuals. Set IDs represent samples where either *T. crocifolius* (C) or *T. lamottei* (L) homoeologs are lost (L) or silenced (S). Overlapping sets are indicated by filled circles, and the number of loci within that intersection may be seen directly above in the bar chart. The total sample sizes are found in the left bar chart and correspond to the adjacent sample. As expected, there should be no overlap across *T. crocifolius* and *T. lamottei* homoeologs.

A recent publication in *Spartina* (Giraud et al., 2021) has demonstrated that a high degree transcriptome repatterning (52% of genes deviated from parental additivity) occurs following neopolyploidy (within the last 170 years), and long-term, divergent transcriptome evolution is evident between the mesohexaploid parents that diverged 2-3 MYA (with 36% genes deviating from parental additivity).

One potential reason for this difference in additive expression may be that neopolyploid survival is dependent upon reduction in gene expression noise, as expression noise can have negative impacts upon fitness (Barkai and Leibler, 2000; Rao et al., 2002; Fraser et al., 2004; Pires and Conant, 2016). Thus, shrinkage toward mean parental expression within neopolyploids may alleviate the effects of transcriptomic shock, especially for polyploids with sub-genome *trans*-acting factors that are largely interchangeable. Over longer periods of time, mutation and selection may then optimize expression of genes, resulting in more non-additive expression. Both noise reduction and gene dosage are expected to play a large role after polyploidization (reviewed in Conant et al., 2014; Pires and Conant, 2016). Interestingly, dosage effects are seen in numerous additively expressed genes within polyploids (Guo et al., 1996; Chen, 2007). These dosage effects are expected to primarily affect genes that function in protein complexes or biological pathways (reviewed in Freeling, 2009; Birchler and Veitia, 2010, 2012). This explanation appears to be the case for additively expressed genes conserved across these independently formed polyploids in that our functional protein association network was significantly enriched for protein-protein interactions.

There is evidence that members of protein complexes within yeast, fruit flies, and humans all exhibit reduced expression noise (Ohno, 1970; Lemos et al., 2004; Schuster-Böckler et al., 2010; reviewed in Pires and Conant, 2016). As such, finding protein-protein interaction enrichment among additively expressed genes may be further evidence that noise reduction and dosage play a significant role in expression changes after allopolyploidization. The degree of dissimilarity between parental expression may also significantly affect homoeolog expression fate between neo- and mesopolyploids (Conant et al., 2014). Environmental differences may also select for different expression patterns over time (Otto and Whitton, 2000). As such, there is likely a complex interplay among the processes governing expression patterns after polyploidization.

4.3. Homoeolog-Specific Expression

HSE, also sometimes called homoeolog expression bias, has been observed in neopolyploids such as *Senecio* (Hegarty et al., 2012), mesopolyploids such as *Gossypium* (Adams et al., 2004; Chaudhary et al., 2009; Yoo et al., 2013), and even more broadly across polyploid plants (Buggs et al., 2010b; Schnable et al., 2011; Grover et al., 2012; Woodhouse et al., 2014; Yang et al., 2016). Notably, we observed numerous loci demonstrating HSE, but

overall, we see a similar proportion of loci exhibiting homoeolog expression bias toward each parent (Grover et al., 2012). This balanced proportion of HSE in *Tragopogon* is interesting in that numerous other allopolyploid plants have exhibited substantial subgenome expression bias (Chen and Pikaard, 1997; Wang et al., 2006b; Flagel et al., 2008; Chaudhary et al., 2009; Akhunova et al., 2010; Schnable and Freeling, 2011; Schnable et al., 2011). However, both neoallopolyploid *Tragopogon* (Boatwright et al., 2018) and the mesoallopolyploid *T. castellanus* demonstrate similar proportions of homoeolog expression bias toward their corresponding parents. HSE in resynthesized *Brassica* neoallopolyploids is established soon after the initial genome merger and allopolyploidization (Yang et al., 2016). So, HSE is potentially yet another ameliorative response to whole-genome duplication and/or hybridization (Pires and Conant, 2016). The cause of these expression patterns in *Tragopogon* is unclear, but numerous genetic and epigenetic mechanisms have been proposed to affect expression in polyploids (Chen, 2007).

The maintenance of dosage balance is not likely to occur indefinitely after whole-genome duplication (Conant et al., 2014; McGrath et al., 2014). HSE is believed to allow duplicated copies to undergo subfunctionalization, neofunctionalization or fractionation, but it is possible that recurrent gene conversion between duplicated copies may maintain sequence identity between them (Pires and Conant, 2016). Biased sub-genome expression dominance has been observed following whole-genome duplication in maize where biased expression occurs within neofunctionalized regulatory genes, and non-regulatory neofunctionalized genes incrementally acquire sub-genome dominance during development (Hughes et al., 2014). Epigenetic regulation has been shown to facilitate sub-genome dominance after whole-genome triplication in *B. rapa* where a biased distribution of transposable elements among sub-genomes as well as small targeting RNAs are responsible for expression dominance at a sub-genome scale (Cheng et al., 2016). It is also possible that HSE reconciles problems arising from heterologous protein complexes for proteins that function more efficiently as homopolymers or require precise binding affinities, stoichiometry or product ratios (Birchler and Veitia, 2010, 2012; Boatwright et al., 2018).

4.4. Homoeolog Silencing and Loss

The process of genome evolution after polyploidization is characterized by alterations in methylation, transposable element activity, expression and function changes as well as genome rearrangement and downsizing (reviewed in Van de Peer et al., 2009; Wendel, 2015; Soltis et al., 2016; Wendel et al., 2018). While these changes have been observed in mesopolyploids (Wang et al., 2011) and paleopolyploids (Schnable et al., 2011), they also occur in neopolyploids where a wide spectrum of genomic changes may occur soon after genome merger and duplication (Madlung and Wendel, 2013), indicating that neopolyploid genomes are not necessarily additive or static (Leitch et al., 2008). Stochastic silencing has been proposed to play an important role in the formation of new species and diploidization after polyploidization. Polyploid species are notable in their tendency to preserve duplicate gene copies, which could be a result of gene dosage effects (Lynch and Conery, 2000; Conant et al., 2014). Dynamic silencing likely serves as a damage-control mechanism to temper potentially adverse effects of polyploidization on gene dosage to improve chances of establishment and adaptation of nascent polyploids (Wendel, 2000; Chaudhary et al., 2009; Buggs et al., 2011). In this study, the silencing of specific homoeologs was more inconsistent across independent polyploids than were loss events or expressed genes. In fact, the majority of silencing events were unique to each polyploid, which seems to support the role of stochastic silencing in polyploid plants. However, it is notable that even though silencing appeared to be stochastic, the homoeologs that were lost were more consistent. This may suggest that the mechanisms governing fractionation are more systematic.

Tragopogon castellanus was previously shown to exhibit a nearly additive genome size of its parents, and the degree of loss seen here (∼3% of loci examined) is consistent with that finding (Mavrodiev et al., 2015). Neopolyploid *Tragopogon* species from the US also exhibited very little putative gene loss (Boatwright et al., 2018) and exhibit an additive genome size (Pires et al., 2004a). Long-term gene loss and retention after whole-genome duplication has demonstrated what appears to be a non-random progression in previous studies (Barker et al., 2008; Freeling, 2009; Birchler and Veitia, 2010; Schnable et al., 2011; Severin et al., 2011; De Smet et al., 2013; Soltis et al., 2016). These observations may also be consistent with the biased fractionation hypothesis, where genome dominance is expected when the subgenomes are highly diverged but not when the subgenomes are similar (Garsmeur et al., 2014; Zhao et al., 2017). While the exact divergence between for diploid parents of *T. castellanus* has not been thoroughly investigated, the P-derived parental genetic divergence index, the ratio between parental divergence and the average genetic divergence in the respective genus, is 1.14 (Paun et al., 2009), indicating that the balanced expression may be justified by the low parental divergence. This biased fractionation theory is also supported by the contrasting case of recently formed *Mimulus peregrinus* allopolyploids (Edger et al., 2017) where subgenome expression dominance occurs immediately following the hybridization of divergent genomes and increases significantly over subsequent generations and results from *Ephedra* allotetraploids whose subgenomes are approximately 8MY diverged, where it has been shown that the rapid formation of large genomes could be attributed to even and slow fractionation following polyploidization (Wu et al., 2021).

Tragopogon seems to be yet another case of convergent homoeolog loss after multiple, independent polyploidization events similar to recent results from *Capsella* allotetraploids have demonstrated predictable patterns of gene retention and loss following polyploidization (Douglas et al., 2015). We further checked for gene ontology enrichment within our retained, lost and silenced genes but found no significant enrichment. Differential regulation of proteome additivity was not related to protein function in *Brassica napus* allotetraploids (Albertin et al., 2007). So, a lack of enrichment within additively expressed genes may be expected. While the lack of enrichment within lost genes contrasts with studies that found binding proteins, protein kinases, transcription factors, and transferases are usually retained in duplicate (Jiao et al., 2011), and photosynthesis and cell cycle genes typically drop to singleton status (De Smet et al.,

2013), it is the same result found in the neopolyploid *Tragopogon* species (Boatwright et al., 2018). It may be that loss is not predominantly determined by functional category but rather by some other genetic or epigenetic characteristic such as noise reduction or dosage, at least within *Tragopogon*.

5. FINAL REMARKS

The short- and long-term effects of *cis*- and *trans*-acting interactions are sure to have a significant, if not dynamically different, effect on duplicate gene fate within allopolyploid species. Studies of these processes lack duplication but are certain to identify broader physiological, ecological, and evolutionary implications of polyploidization (Soltis et al., 2016). Here, we compared both homoeolog fate convergence within independently formed mesoallopolyploid populations (*T. castellanus*) and how those compare to neoallopolyploids within the same genus using the same methodology. While homoeolog expression bias was balanced in both the two neopolyploids and in the mesopolyploid, the degree of additive expression was significantly different, with populations of the mesopolyploid demonstrating more non-additive expression. We found that homoeologs that are retained or lost seem to be strongly convergent across independently formed allopolyploids, while silencing tends to occur stochastically. Further, this non-random trend in long-term homoeolog retention and loss is not unique to *Tragopogon* but may be selectively advantageous for polyploid speciation and survival (Barker et al., 2008; Freeling, 2009; Birchler and Veitia, 2010; Severin et al., 2011; Schnable et al., 2012; De Smet et al., 2013; Soltis et al., 2016). While there was no GO enrichment among the studied gene sets, additively expressed genes demonstrated enrichment for protein-protein interactions within a functional network. It may be that gene dosage and noise minimization play leading roles in regulating gene expression patterns after allopolyploidization, and these patterns are conserved across independent lineages.

DATA AVAILABILITY STATEMENT

The data presented in this study are deposited in the NCBI's Sequence Read Archive (SRA) under BioProject PRJNA728143. Scripts used to run the analyses are available on GitHub at https://github.com/jlboat/Tragopogon_castellanus.

AUTHOR CONTRIBUTIONS

JB, DS, PSo, PSc, and WB designed the experiments. H-CH and AS generated data. JB and C-TY performed analyses. JB, WB, PSo, and DS wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Department of Biological Sciences at the University of Florida, the University of Florida College of Liberal Arts and Sciences, and the Florida Genetics Institute and National Science Foundation grant IOS-1146065 (DS, PSo, WB, PSc). The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the NSF.

ACKNOWLEDGMENTS

We would like to acknowledge the University of Florida's High-Performance Computing Center for providing computational resources and support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.679047/full#supplementary-material

REFERENCES

Adams, K. L., Percifield, R., and Wendel, J. F. (2004). Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* 168, 2217–2226. doi: 10.1534/genetics.104.033522

Adams, K. L., and Wendel, J. F. (2005a). Allele-specific, bidirectional silencing of an alcohol dehydrogenase gene in different organs of interspecific diploid cotton hybrids. *Genetics* 171, 2139–2142. doi: 10.1534/genetics.105.047357

Adams, K. L., and Wendel, J. F. (2005b). Novel patterns of gene expression in polyploid plants. *Trends Genet.* 21, 539–543. doi: 10.1016/j.tig.2005.07.009

Aguilar-Rangel, M. R., Montes, R. A. C., González-Segovia, E., Ross-Ibarra, J., Simpson, J., and Sawers, R. J. (2017). Allele specific expression analysis identifies regulatory variation associated with stress-related genes in the mexican highland maize landrace palomero toluqueño. *bioRxiv* 152397. doi: 10.7717/peerj.3737

Akhunova, A. R., Matniyazov, R. T., Liang, H., and Akhunov, E. D. (2010). Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics* 11:505. doi: 10.1186/1471-2164-11-505

Albertin, W., Alix, K., Balliau, T., Brabant, P., Davanture, M., Malosse, C., et al. (2007). Differential regulation of gene products in newly synthesized brassica napus allotetraploids is not related to protein function nor subcellular localization. *BMC Genomics* 8:56. doi: 10.1186/1471-2164-8-56

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Anssour, S., and Baldwin, I. T. (2010). Variation in antiherbivore defense responses in synthetic nicotiana allopolyploids correlates with changes in uniparental patterns of gene expression. *Plant Physiol.* 153, 1907–1918. doi: 10.1104/pp.110.156786

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Baldauf, J., Marcon, C., Paschold, A., and Hochholdinger, F. (2016). Nonsyntenic genes drive tissue-specific dynamics of differential, nonadditive and allelic expression patterns in maize hybrids. *Plant Physiol.* 171, 1144–1155. doi: 10.1104/pp.16.00262

Barkai, N. and Leibler, S. (2000). Biological rhythms: Circadian clocks limited by noise. *Nature* 403, 267–268. doi: 10.1038/35002258

Barker, M. S., Baute, G. J., and Liu, S. L. (2012). "Duplications and turnover in plant genomes," in *Plant Genome Diversity Volume 1* (Vienna: Springer), 155–169.

Barker, M. S., Kane, N. C., Matvienko, M., Kozik, A., Michelmore, R. W., Knapp, S. J., et al. (2008). Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* 25, 2445–2455. doi: 10.1093/molbev/msn187

Bell, G. D., Kane, N. C., Rieseberg, L. H., and Adams, K. L. (2013). Rna-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biol. Evol.* 5, 1309–1323. doi: 10.1093/gbe/evt072

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 67, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Birchler, J. A., and Veitia, R. A. (2010). The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62. doi: 10.1111/j.1469-8137.2009.03087.x

Birchler, J. A., and Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14746–14753. doi: 10.1073/pnas.1207726109

Blanc, G., and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. *Plant Cell* 16, 1679–1691. doi: 10.1105/tpc.021410

Blanca, G., and de la Guardia, C. D. (1996). Sinopsis del género *Tragopogon* L. (Asteraceae) en la Península Ibérica. *Anales del Jardin Botanico de Madrid,* Vol. 54, 358–363

Boatwright, J. L., McIntyre, L. M., Morse, A. M., Chen, S., Yoo, M.-J., Koh, J., et al. (2018). A robust methodology for assessing differential homeolog contributions to the transcriptomes of allopolyploids. *Genetics* 210, 883–894. doi: 10.1534/genetics.118.301564

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Buggs, R. J., Chamala, S., Wu, W., Gao, L., May, G. D., Schnable, P. S., et al. (2010a). Characterization of duplicate gene evolution in the recent natural allopolyploid tragopogon miscellus by next-generation sequencing and sequenom iplex massarray genotyping. *Mol. Ecol.* 19, 132–146. doi: 10.1111/j.1365-294X.2009.04469.x

Buggs, R. J., Chamala, S., Wu, W., Tate, J. A., Schnable, P. S., Soltis, D. E., et al. (2012). Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr. Biol.* 22, 248–252. doi: 10.1016/j.cub.2011.12.027

Buggs, R. J., Elliott, N. M., Zhang, L., Koh, J., Viccini, L. F., Soltis, D. E., et al. (2010b). Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid tragopogon mirus. *New Phytol.* 186, 175–183. doi: 10.1111/j.1469-8137.2010.03205.x

Buggs, R. J., Wendel, J. F., Doyle, J. J., Soltis, D. E., Soltis, P. S., and Coate, J. E. (2014). The legacy of diploid progenitors in allopolyploid gene expression patterns. *Philos. Trans. R. Soc. B Biol. Sci.* 369, 20130354. doi: 10.1098/rstb.2013.0354

Buggs, R. J., Zhang, L., Miles, N., Tate, J. A., Gao, L., Wei, W., et al. (2011). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Curr. Biol.* 21, 551–556. doi: 10.1016/j.cub.2011.02.016

Chaudhary, B., Flagel, L., Stupar, R. M., Udall, J. A., Verma, N., Springer, N. M., et al. (2009). Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). *Genetics* 182, 503–517. doi: 10.1534/genetics.109.102608

Chen, Z. J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Ann. Rev. Plant Biol.* 58:377. doi: 10.1146/annurev.arplant.58.032806.103835

Chen, Z. J., and Pikaard, C. S. (1997). Transcriptional analysis of nucleolar dominance in polyploid plants: biased expression/silencing of progenitor rrna genes is developmentally regulated in brassica. *Proc. Natl. Acad. Sci. U.S.A.* 94, 3442–3447. doi: 10.1073/pnas.94.7.3442

Cheng, F., Sun, C., Wu, J., Schnable, J., Woodhouse, M. R., Liang, J., et al. (2016). Epigenetic regulation of subgenome dominance following whole genome triplication in brassica rapa. *New Phytol.* 211, 288–299. doi: 10.1111/nph.13884

Conant, G. C., Birchler, J. A., and Pires, J. C. (2014). Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* 19, 91–98. doi: 10.1016/j.pbi.2014.05.008

de la Guardia Guerrero, C. D., and López, G. B. (1989). "Tragopogon castellanus" Levier = "T. crocifolius" subsp. "badalii" Willk. *Anales del Jardin Botanico de Madrid,* Vol. 47, 253–256.

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2898–2903. doi: 10.1073/pnas.1300127110

De Smet, R., and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events. *Curr. Opin. Plant Biol.* 15, 168–176. doi: 10.1016/j.pbi.2012.01.003

Douglas, G. M., Gos, G., Steige, K. A., Salcedo, A., Holm, K., Josephs, E. B., et al. (2015). Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid capsella bursa-pastoris. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2806–2811. doi: 10.1073/pnas.1412277112

Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P. S., et al. (2008). Evolutionary genetics of genome merger and doubling in plants. *Ann. Rev. Genet.* 42, 443–461. doi: 10.1146/annurev.genet.42.110807.091524

Edger, P. P., and Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* 17, 699–717. doi: 10.1007/s10577-009-9055-9

Edger, P. P., Smith, R., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., et al. (2017). Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* 29, 2150–2167. doi: 10.1105/tpc.17.00010

Fear, J. M., Leon-Novelo, L. G., Morse, A. M., Gerken, A. R., Van Lehman, K., Tower, J., et al. (2016). Buffering of genetic regulatory networks in drosophila melanogaster. *Genetics* 203, 1177–1190. doi: 10.1534/genetics.116.188797

Finn, R. D., Clements, J., and Eddy, S. R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Flagel, L., Udall, J., Nettleton, D., and Wendel, J. (2008). Duplicate gene expression in allopolyploid gossypium reveals two temporally distinct phases of expression evolution. *BMC Biol.* 6:16. doi: 10.1186/1741-7007-6-16

Flagel, L. E., and Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytol.* 183, 557–564. doi: 10.1111/j.1469-8137.2009.02923.x

Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J., and Eisen, M. B. (2004). Noise minimization in eukaryotic gene expression. *PLoS Biol.* 2:e137. doi: 10.1371/journal.pbio.0020137

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Ann. Rev. Plant Biol.* 60, 433–453. doi: 10.1146/annurev.arplant.043008.092122

Frith, M. C., Wan, R., and Horton, P. (2010). Incorporating sequence quality data into alignment improves dna read mapping. *Nucleic Acids Res.* 38, e100–e100. doi: 10.1093/nar/gkq010

Garsmeur, O., Schnable, J. C., Almeida, A., Jourda, C., D'Hont, A., and Freeling, M. (2014). Two evolutionarily distinct classes of paleopolyploidy. *Mol. Biol. Evol.* 31, 448–454. doi: 10.1093/molbev/mst230

Giraud, D., Lima, O., Rousseau-Gueutin, M., Salmon, A., and Aïnouche, M. (2021). Gene and transposable element expression evolution following recent and past polyploidy events in spartina (poaceae). *Front. Genet.* 12:589160. doi: 10.3389/fgene.2021.589160

Gish, W. (2005). Wu-blast 2.0. *Person. Commun*.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Graze, R., Novelo, L., Amin, V., Fear, J., Casella, G., Nuzhdin, S., and McIntyre, L. (2012). Allelic imbalance in drosophila hybrid heads: exons, isoforms, and evolution. *Mol. Biol. Evol.* 29, 1521–1532. doi: 10.1093/molbev/msr318

Grover, C., Gallagher, J., Szadkowski, E., Yoo, M., Flagel, L., and Wendel, J. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.* 196, 966–971. doi: 10.1111/j.1469-8137.2012.04365.x

Guo, M., Davis, D., and Birchler, J. A. (1996). Dosage effects on gene expression in a maize ploidy series. *Genetics* 142, 1349–1355. doi: 10.1093/genetics/142.4.1349

Guo, M., Rupe, M. A., Yang, X., Crasta, O., Zinselmeier, C., Smith, O. S., et al. (2006). Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theor. Appl. Genet.* 113, 831–845. doi: 10.1007/s00122-006-0335-x

Hegarty, M. J., Abbott, R. J., and Hiscock, S. J. (2012). "Allopolyploid speciation in action: the origins and evolution of Senecio cambrensis," in *Polyploidy and Genome Evolution* (Berlin; Heidelberg: Springer), 245–270.

Hegarty, M. J., Barker, G. L., Wilson, I. D., Abbott, R. J., Edwards, K. J., and Hiscock, S. J. (2006). Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. *Curr. Biol.* 16, 1652–1659. doi: 10.1016/j.cub.2006.06.071

Hu, G., Koh, J., Yoo, M.-J., Chen, S., and Wendel, J. F. (2015). Gene-expression novelty in allopolyploid cotton: a proteomic perspective. *Genetics* 200, 91–104. doi: 10.1534/genetics.115.174367

Hughes, T. E., Langdale, J. A., and Kelly, S. (2014). The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res.* 24, 1348–1355. doi: 10.1101/gr.172684.114

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T.-H., et al. (2013). Architecture and evolution of a minute plant genome. *Nature* 498, 94. doi: 10.1038/nature12132

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2011). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Kihara, H., and Ono, T. (1926). Chromosomenzahlen und systematische gruppierung der rumex-arten. *Z. Zellforsch. Mikroskopisc. Anatom.* 4, 475–481. doi: 10.1007/BF00391215

Koh, J., Chen, S., Zhu, N., Yu, F., Soltis, P. S., and Soltis, D. E. (2012). Comparative proteomics of the recently and recurrently formed natural allopolyploid tragopogon mirus (asteraceae) and its parents. *New Phytologist* 196, 292–305. doi: 10.1111/j.1469-8137.2012.04251.x

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Lagesen, K., Hallin, P., Rødland, E., Stærfeldt, H., Rognes, T., and Ussery, D. (2007). Rnammer: consistent annotation of rrna genes in genomic sequences. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160

Langham, R. J., Walsh, J., Dunn, M., Ko, C., Goff, S. A., and Freeling, M. (2004). Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166, 935–945. doi: 10.1093/genetics/166.2.935

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol.* 10:R25. doi: 10.1186/gb-2009-10-3-r25

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol.* 15:R29. doi: 10.1186/gb-2014-15-2-r29

Leitch, A., and Leitch, I. (2008). Genomic plasticity and the diversity of polyploid plants. *Science* 320, 481–483. doi: 10.1126/science.1153585

Leitch, I., Hanson, L., Lim, K., Kovarik, A., Chase, M., Clarkson, J., et al. (2008). The ups and downs of genome size evolution in polyploid species of nicotiana (solanaceae). *Ann. Bot.* 101, 805–814. doi: 10.1093/aob/mcm326

Leitch, I. J., and Leitch, A. R. (2013). *Genome Size Diversity and Evolution in Land Plants* (Springer), 307–322.

Lemos, B., Meiklejohn, C. D., and Hartl, D. L. (2004). Regulatory evolution across the protein interaction network. *Nat. Genet.* 36, 1059–1060. doi: 10.1038/ng1427

León-Novelo, L. G., McIntyre, L. M., Fear, J. M., and Graze, R. M. (2014). A flexible bayesian method for detecting allelic imbalance in RNA-seq data. *BMC Genomics* 15:920. doi: 10.1186/1471-2164-15-920

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). Upset: visualization of intersecting sets. *IEEE Trans. Visual. Comput. Graph.* 20, 1983–1992. doi: 10.1109/TVCG.2014.2346248

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Madlung, A., and Wendel, J. F. (2013). Genetic and epigenetic aspects of polyploid evolution in plants. *Cytogenet Genome Res.* 140, 270–285. doi: 10.1159/000351430

Mallet, J. (2007). Hybrid speciation. *Nature* 446, 279–283. doi: 10.1038/nature05706

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Mason, A. S., and Pires, J. C. (2015). Unreduced gametes: meiotic mishap or evolutionary mechanism? *Trends Genet.* 31, 5–10. doi: 10.1016/j.tig.2014.09.011

Mavrodiev, E. V., Chester, M., Suárez-Santiago, V. N., Visger, C. J., Rodriguez, R., Susanna, A., et al. (2015). Multiple origins and chromosomal novelty in the allotetraploid tragopogon castellanus (asteraceae). *New Phytol.* 206, 1172–1183. doi: 10.1111/nph.13227

Mavrodiev, E. V., Soltis, P. S., and Soltis, D. E. (2008). Putative parentage of six old world polyploids in tragopogon l. (asteraceae: Scorzonerinae) based on its, ets, and plastid sequence data. *Taxon* 57, 1215–1222E. doi: 10.1002/tax.574014

Mavrodiev, E. V., Tancig, M., Sherwood, A. M., Gitzendanner, M. A., Rocca, J., Soltis, P. S., et al. (2005). Phylogeny of tragopogon l. (asteraceae) based on internal and external transcribed spacer sequence data. *Int. J. Plant Sci.* 166, 117–133. doi: 10.1086/425206

McGrath, C. L., Gout, J.-F., Doak, T. G., Yanagi, A., and Lynch, M. (2014). Insights into three whole-genome duplications gleaned from the paramecium caudatum genome sequence. *Genetics* 197, 1417–1428. doi: 10.1534/genetics.114.163287

Moreno-Hagelsieb, G., and Latimer, K. (2008). Choosing blast options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24, 319–324. doi: 10.1093/bioinformatics/btm585

Munger, S. C., Raghupathy, N., Choi, K., Simons, A. K., Gatti, D. M., Hinerfeld, D. A., et al. (2014). Rna-seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics* 198, 59–73. doi: 10.1534/genetics.114.165886

Ohno, S. (1970). *Evolution by Gene Duplication.* New York, NY: Springer Science & Business Media.

Otto, S. P., and Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.* 34, 401–437. doi: 10.1146/annurev.genet.34.1.401

Ownbey, M. (1950). Natural hybridization and amphiploidy in the genus tragopogon. *Am. J. Bot.* 37, 487–499. doi: 10.1002/j.1537-2197.1950.tb11033.x

Paun, O., Forest, F., Fay, M. F., and Chase, M. W. (2009). Hybrid speciation in angiosperms: parental divergence drives ploidy. *New Phytol.* 182, 507–518. doi: 10.1111/j.1469-8137.2009.02767.x

Pellicer, J., Fay, M. F., and Leitch, I. J. (2010). The largest eukaryotic genome of them all? *Bot. J. Linnean Soc.* 164, 10–15. doi: 10.1111/j.1095-8339.2010.01072.x

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Pires, J. C., and Conant, G. C. (2016). Robust yet fragile: expression noise, protein misfolding, and gene dosage in the evolution of genomes. *Ann. Rev. Genet.* 50, 113–131. doi: 10.1146/annurev-genet-120215-035400

Pires, J. C., Lim, K. Y., Kovarík, A., Matyásek, R., Boyd, A., Leitch, A. R., et al. (2004a). Molecular cytogenetic analysis of recently evolved tragopogon (asteraceae) allopolyploids reveal a karyotype that is additive of the diploid progenitors. *Am. J. Bot.* 91, 1022–1035. doi: 10.3732/ajb.91.7.1022

Pires, J. C., Zhao, J., Schranz, M., Leon, E. J., Quijada, P. A., Lukens, L. N., et al. (2004b). Flowering time divergence and genomic rearrangements in resynthesized brassica polyploids (brassicaceae). *Biol. J. Linnean Soc.* 82, 675–688. doi: 10.1111/j.1095-8312.2004.00350.x

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., et al. (2011). eggnog v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* 40, D284–D289. doi: 10.1093/nar/gkr1060

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2011). The pfam protein families database. *Nucleic Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Ann. Rev. Ecol. System.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Rao, C. V., Wolf, D. M., and Arkin, A. P. (2002). Control, exploitation and tolerance of intracellular noise. *Nature* 420, 231–237. doi: 10.1038/nature01258

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Salmon, A., Ainouche, M. L., and Wendel, J. F. (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in spartina (poaceae). *Mol. Ecol.* 14, 1163–1175. doi: 10.1111/j.1365-294X.2005.02488.x

Schnable, J. C., and Freeling, M. (2011). Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLoS ONE* 6:e17855. doi: 10.1371/journal.pone.0017855

Schnable, J. C., Freeling, M., and Lyons, E. (2012). Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol. Evol.*, 4, 265–277. doi: 10.1093/gbe/evs009

Schnable, J. C., Springer, N. M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4069–4074. doi: 10.1073/pnas.1101368108

Schuster-Böckler, B., Conrad, D., and Bateman, A. (2010). Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS ONE* 5:e9474. doi: 10.1371/journal.pone.0009474

Sémon, M., and Wolfe, K. H. (2007). Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17, 505–512. doi: 10.1016/j.gde.2007.09.007

Severin, A. J., Cannon, S. B., Graham, M. M., Grant, D., and Shoemaker, R. C. (2011). Changes in twelve homoeologous genomic regions in soybean following three rounds of polyploidy. *Plant Cell* 23, 3129–3136. doi: 10.1105/tpc.111.089573

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and Levy, A. A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* 13, 1749–1759. doi: 10.1105/TPC.010083

Soltis, D. E., Buggs, R. J., Barbazuk, W. B., Chamala, S., Chester, M., Gallagher, J. P., et al. (2012). *The Early Stages of Polyploidy: Rapid and Repeated Evolution in Tragopogon* (Springer), 271–292.

Soltis, D. E., and Soltis, P. S. (1999). Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.* 14, 348–352. doi: 10.1016/S0169-5347(99)01638-9

Soltis, D. E., Soltis, P. S., Bennett, M. D., and Leitch, I. J. (2003). Evolution of genome size in the angiosperms. *Am. J. Bot.* 90, 1596–1603. doi: 10.3732/ajb.90.11.1596

Soltis, D. E., Soltis, P. S., Pires, J. C., Kovarik, A., Tate, J. A., and Mavrodiev, E. (2004). Recent and recurrent polyploidy in tragopogon (asteraceae): cytogenetic, genomic and genetic comparisons. *Biol. J. Linnean Soc.* 82, 485–501. doi: 10.1111/j.1095-8312.2004.00335.x

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103, 1146–1166. doi: 10.3732/ajb.1500501

Soltis, P. S., Plunkett, G. M., Novak, S. J., and Soltis, D. E. (1995). Genetic variation in tragopogon species: additional origins of the allotetraploids t. mirus and t. miscellus (compositae). *Am. J. Bot.* 82, 1329–1341. doi: 10.1002/j.1537-2197.1995.tb12666.x

Spoelhof, J. P., Soltis, P. S., and Soltis, D. E. (2017). Pure polyploidy: closing the gaps in autopolyploid research. *J. Syst. Evol.* 55, 340–352. doi: 10.1111/jse.12253

Stebbins Jr, G. L. (1947). Types of polyploids: their classification and significance. *Adv. Genetics* 1, 403–429.

Stupar, R. M., and Springer, N. M. (2006). Cis-transcriptional variation in maize inbred lines b73 and mo17 leads to additive expression patterns in the f1 hybrid. *Genetics* 173, 2199–2210. doi: 10.1534/genetics.106.060699

Swanson-Wagner, R. A., DeCook, R., Jia, Y., Bancroft, T., Ji, T., Zhao, X., et al. (2009). Paternal dominance of trans-eqtl influences gene expression patterns in maize hybrids. *Science* 326, 1118–1120. doi: 10.1126/science.1178294

Swanson-Wagner, R. A., Jia, Y., DeCook, R., Borsuk, L. A., Nettleton, D., and Schnable, P. S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize f1 hybrid and its inbred parents. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6805–6810. doi: 10.1073/pnas.0510430103

Symonds, V. V., Soltis, P. S., and Soltis, D. E. (2010). Dynamics of polyploid formation in tragopogon (asteraceae): recurrent formation, gene flow, and population structure. *Evol. Int. J. Org. Evol.* 64, 1984–2003. doi: 10.1111/j.1558-5646.2010.00978.x

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). String v10: proteinprotein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Tate, J. A., Ni, Z., Scheen, A.-C., Koh, J., Gilbert, C. A., Lefkowitz, D., et al. (2006). Evolution and expression of homeologous loci in tragopogon miscellus (asteraceae), a recent and reciprocally formed allopolyploid. *Genetics* 173, 1599–1611. doi: 10.1534/genetics.106.057646

Team, R. C. (2014). *R: a Language and Environment for Statistical Computing*. Vienna: R foundation for statistical computing 2012.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10:725. doi: 10.1038/nrg2600

Vicient, C. M., and Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* 120, 195–207. doi: 10.1093/aob/mcx078

Wang, J., Tian, L., Lee, H.-S., and Chen, Z. J. (2006a). Nonadditive regulation of fri and flc loci mediates flowering-time variation in arabidopsis allopolyploids. *Genetics* 173, 965–974. doi: 10.1534/genetics.106.056580

Wang, J., Tian, L., Lee, H.-S., Wei, N. E., Jiang, H., Watson, B., et al. (2006b). Genomewide nonadditive gene regulation in arabidopsis allotetraploids. *Genetics* 172, 507–517. doi: 10.1534/genetics.105.047894

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The genome of the mesopolyploid crop species brassica rapa. *Nat. Genet.* 43, 1035–1039. doi: 10.1038/ng.919

Wendel, J. F. (2000). Genome evolution in polyploids. *Plant Mol. Evol.* 42, 225–249.

Wendel, J. F. (2015). The wondrous cycles of polyploidy in plants. *Am. J. Bot.* 102, 1753–1756. doi: 10.3732/ajb.1500320

Wendel, J. F., Lisch, D., Hu, G., and Mason, A. S. (2018). The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr. opin. Genet. Dev.* 49, 1–7. doi: 10.1016/j.gde.2018.01.004

Willkomm, M. (1893). *Supplementum prodromi florae hispanicae: sive, Enumeratio el'descriptio omnium plantarum inde ab anno 1862 usque ad annum 1893 in Hispania detectarum quae innotuerunt auctori, adjectis locis novis specierum jam notarum. E.* Schweizerbart.

Woodhouse, M. R., Cheng, F., Pires, J. C., Lisch, D., Freeling, M., and Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5283–5288. doi: 10.1073/pnas.1402475111

Wright, R. J., Thaxton, P. M., El-Zik, K. M., and Paterson, A. H. (1998). D-subgenome bias of xcm resistance genes in tetraploid gossypium (cotton) suggests that polyploid formation has created novel avenues for evolution. *Genetics* 149, 1987–1996. doi: 10.1093/genetics/149.4.1987

Wu, H., Yu, Q., Ran, J.-H., and Wang, X.-Q. (2021). Unbiased subgenome evolution in allotetraploid species of ephedra and its implications for the evolution of large genomes in gymnosperms. *Genome Biol. Evol.* 13:evaa236. doi: 10.1093/gbe/evaa236

Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., et al. (2016). The genome sequence of allopolyploid brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* 48, 1225–1232. doi: 10.1038/ng.3657

Yoo, M., Szadkowski, E., and Wendel, J. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* 110, 171–180. doi: 10.1038/hdy.2012.94

Yoo, M.-J., Liu, X., Pires, J. C., Soltis, P. S., and Soltis, D. E. (2014). Nonadditive gene expression in polyploids. *Ann. Rev. Genet.* 48, 485–517. doi: 10.1146/annurev-genet-120213-092159

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2012). goseq: gene ontology testing for RNA-seq datasets. *R Bioconductor* 8, 1–25.

Zhao, T., Holmer, R., de Bruijn, S., Angenent, G. C., van den Burg, H. A., and Schranz, M. E. (2017). Phylogenomic synteny network analysis of mads-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. *Plant Cell* 29, 1278–1292. doi: 10.1105/tpc.17.00312

Conflict of Interest: PSc is a Changjiang Scholar at China Agriculture University, and co-founder and managing partner of Data2Bio, LLC; Dryland Genetics, LLC; and EnGeniousAg, LLC. He is a member of the scientific advisory board and a shareholder of Hi-Fidelity Genetics, Inc. and a member of the scientific advisory boards of Kemin Industries and Centro de Tecnologia Canavieira. H-CH was employed by the company Covance Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership