



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 Hloušková¹, Xinyi Guo¹, Pieter Winter⁴, Zuzana Chumová⁵, Adriaan Van Niekerk⁶, Klaus Mummenhoff⁷, Ihsan A. Al-Shehbaz⁸, 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, Průhonice, 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 unigenic 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 (K_s) 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 (Kalyanamoorthy 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 Mummehoff 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 software¹.

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 = $0.1\times$) 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 *heatmap* package (Kolde and Kolde, 2015) in R. The abundance matrix was transformed into a distance matrix by *heatmap*. The clusters with genome proportion higher than 0.01% were used to construct the dendrogram relationship of 15 *Heliophila* and *C. circaeoides* species. *heatmap* 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 package² 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.

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 $2\times$ sodium saline citrate (SSC; $20\times$ 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 $2\times$ 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 $\mu\text{g}/\text{ml}$) 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.

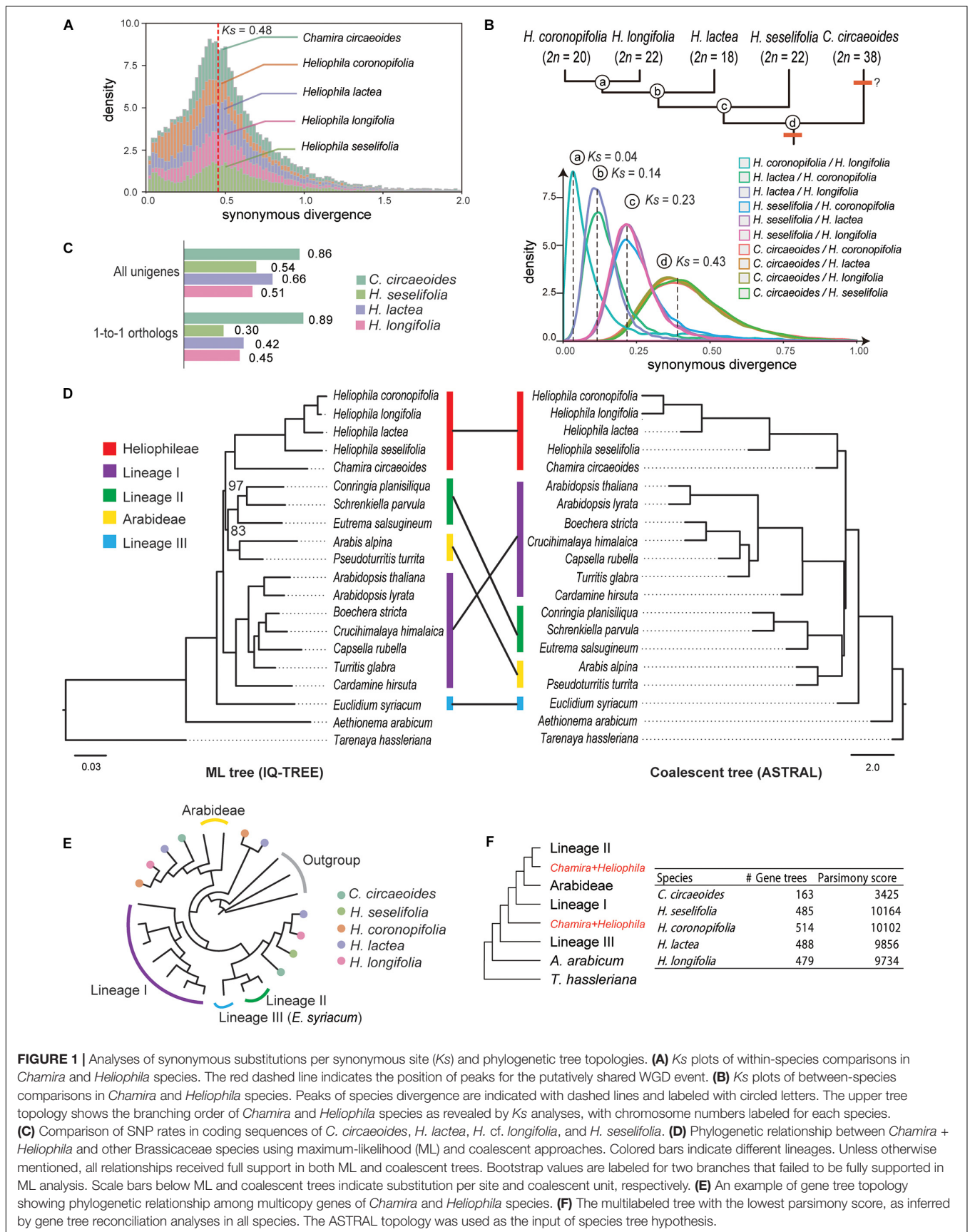
RESULTS

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 (*K_s*; **Supplementary Table 7**). The distribution of *K_s* 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

²<https://www.geneious.com>



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

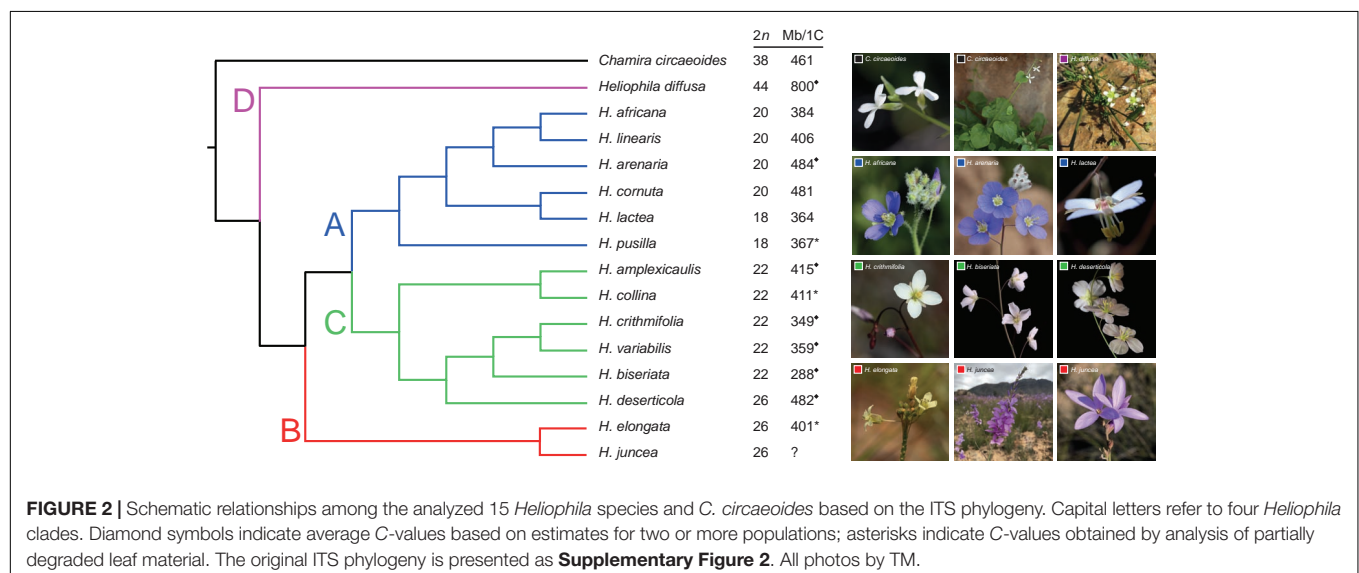
showed that it had a more ancestral position, being outside of lineage I + lineage II + Arabideae (Figure 1D). We also retrieved 130 homologous gene groups that consisted of mostly single-copy genes in diploid species and multicopy genes in *C. circaeoides* and *Heliophila* species. We observed that *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



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
<i>Heliophila africana</i>	A	30.07	511,217	0.19×	232
<i>H. arenaria</i> subsp. <i>arenaria</i>	A	30.18	901,412	0.26×	183
<i>H. cornuta</i> var. <i>cornuta</i>	A	31.38	652,717	0.19×	203
<i>H. lactea</i>	A	33.93	744,973	0.29×	177
<i>H. linearis</i> var. <i>linearis</i>	A	37.03	1,002,030	0.35×	209
<i>H. pusilla</i> var. <i>pusilla</i>	A	37.34	576,080	0.22×	164
<i>H. elongata</i>	B	30.73	662,141	0.23×	181
<i>H. juncea</i>	B	25.04	886,973	0.31×	173
<i>H. amplexicaulis</i>	C	37.21	961,390	0.32×	165
<i>H. collina</i>	C	38.90	774,496	0.31×	130
<i>H. crithmifolia</i>	C	34.06	853,487	0.29×	224
<i>H. deserticola</i> var. <i>micrantha</i>	C	38.64	1,304,122	0.38×	281
<i>H. biseriata</i>	C	32.96	962,154	0.47×	276
<i>H. variabilis</i>	C	33.55	1,056,742	0.41×	148
<i>H. diffusa</i> var. <i>diffusa</i>	D	38.67	911,217	0.16×	141
<i>Chamira circaeoides</i>	–	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).

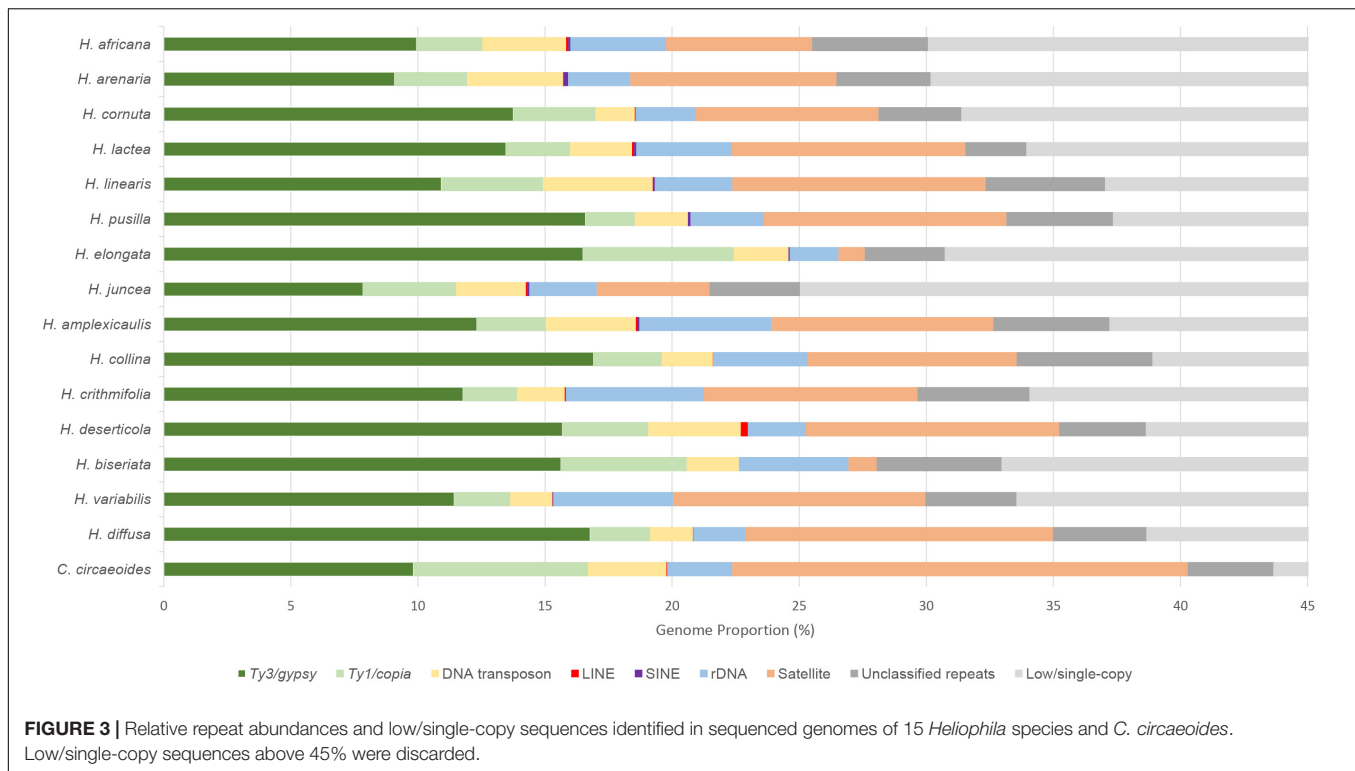
We tested possible correlations between the estimated abundances of identified repeat families (Table 2) and genome size (Mb/1C) of the analyzed *Heliophila* species. The total repeat content was positively correlated with genome size (p value = 0.0006, R^2 = 0.6446). A weak but significant positive 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



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 (%).

Repeat family	Clade															
	A						B			C				D		
	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
<i>Ty3/gypsy</i>																
<i>Athila</i>	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
<i>CRM</i>	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
<i>Tekay</i>	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
<i>Galadriel</i>	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
<i>Reina</i>	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
<i>Ogre/Tat</i>	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
<i>Ty1/copia</i>																
<i>Ale</i>	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
<i>Bianca</i>	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
<i>Angela</i>	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
<i>Ivana</i>	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
<i>TAR</i>	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
<i>Tork</i>	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
<i>SIRE</i>	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																
<i>Harbinger</i>	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
<i>Helitron</i>	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
<i>CACTA</i>	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
<i>Mariner</i>	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
<i>Mutator</i>	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
<i>hAT</i>	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

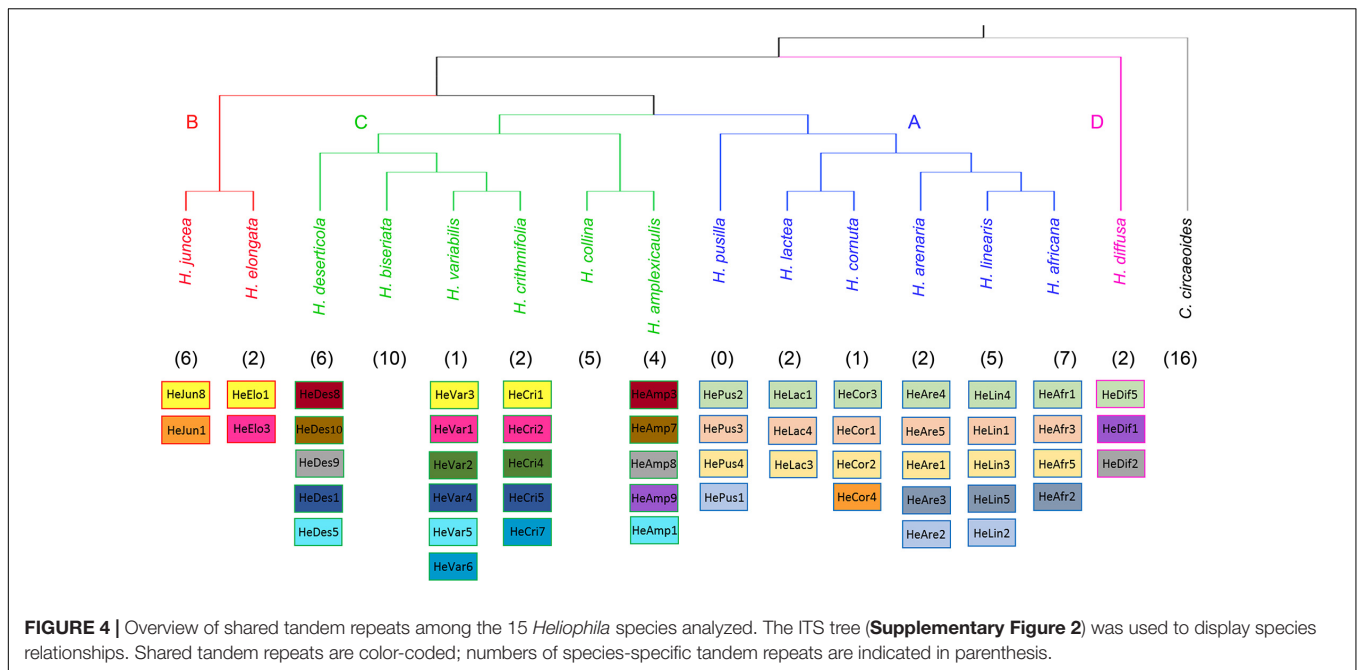
HeAfr, *Heliophila africana*; *HeAre*, *Heliophila arenaria*; *HeCor*, *Heliophila cornuta*; *HeLac*, *Heliophila lactea*; *HeLin*, *Heliophila linearis*; *HePus*, *Heliophila pusilla*; *HeElo*, *Heliophila elongata*; *HeJun*, *Heliophila juncea*; *HeAmp*, *Heliophila amplexicaulis*; *HeCol*, *Heliophila 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 (*HeAfr*2, *HeAre*3, and *HeLin*5). 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 (*HePus*2, *HePus*3, and *HePus*4)

or only with *H. arenaria* (*HeAre*2) and *H. linearis* (*HeLin*2). Interestingly, the 168-bp *HeCor*4 repeat in *H. cornuta* was found to be homologous to the *HeJun*1 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 (*HeElo*1 and *HeJun*8) which was also shared with two clade C species – *H. crithmifolia* (*HeCri*1) and *H. variabilis* (*HeVar*3). The 184-bp *HeElo*3 identified in the *H. elongata* genome was also detected in *H. crithmifolia* (*HeCri*2) and *H. variabilis* (*HeVar*1) (**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



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

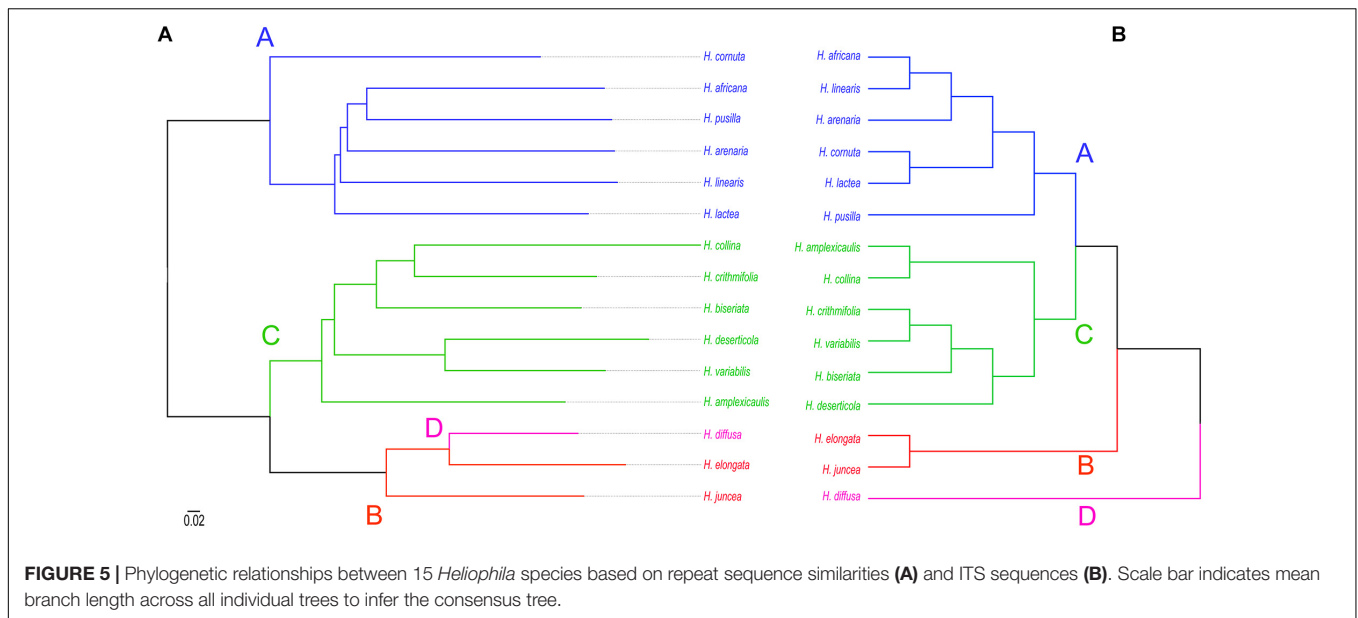
Consensus tree phylogeny was reconstructed using the ape package (Paradis and Schliep, 2019) in R based on pairwise genetic distances between all repeats in the 100 most abundant clusters retrieved from RepeatExplorer2 comparative clustering analysis. In the dataset which consisted of 15 *Heliophila* 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



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**).

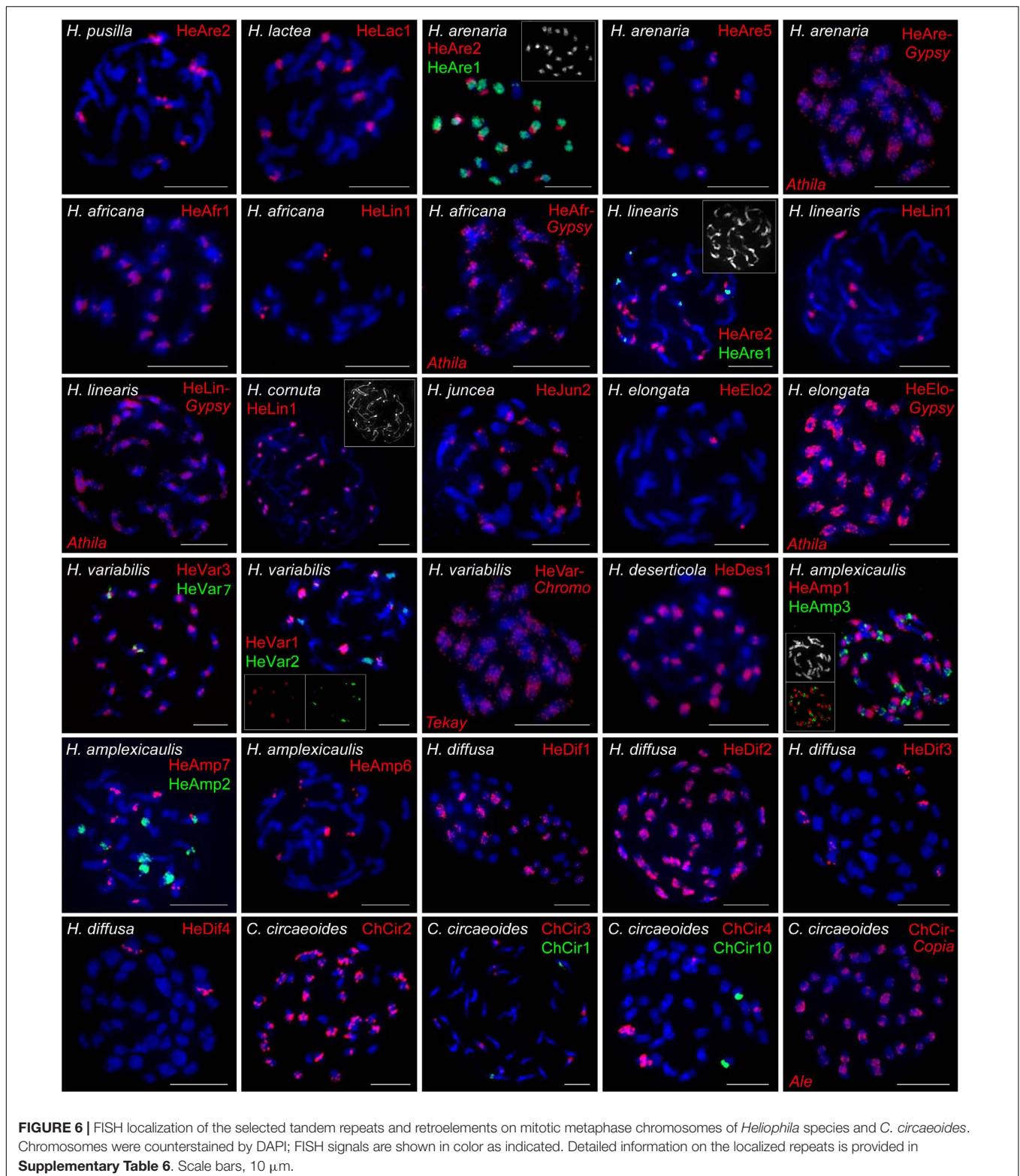
In *C. circaeoides*, three pericentromeric (294-bp ChCir2, 202-bp ChCir3, and 198-bp ChCir4) and two subtelomeric (249-bp ChCir1 and 1,427-bp ChCir10) tandem repeats were localized. The ChCir2 repeat was present in all pericentromeres, whereas ChCir3 and ChCir4 localized in centromeres to four and two 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



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 ($2n = 32, 36, 40, 44, 60, 64, 80, \text{ and } 88$; 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 ($2n = 16, 22, 26, 32, \text{ and } 64$; *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 Heliophleae.

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.

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, A. V., 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., Tomaszewicz, M., Arbeituber, 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.
- Dierckxens, 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.

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.

- 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 Jermin, 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 dioecious 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/tbj.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 Ttv1 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. Phylog. 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.

Copyright © 2021 Dogan, Pouch, Mandáková, Hloušková, Guo, Winter, Chumová, Van Niekerk, Mummenhoff, Al-Shehbaz, Mucina and Lysak. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.