MASARYKOVA
UNIVERZITA

FACULTY OF SCIENCE AND CEITEC

Bioinformatic approaches for structural analysis of plant genomes

Ph.D. Thesis

Petra Hloušková

Supervisor: Prof. Mgr. Martin Lysák, Ph.D., DSc.

NCBR and CEITEC

Brno 2020

Bibliographic Entry

Author:	Mgr. Petra Hloušková
	Faculty of Science, Masaryk University National Centre for Biomolecular Research Laboratory of Functional Genomics and Proteomics
	CEITEC – Masaryk University Mendel Centre for Plant Genomics and Proteomics Martin Lysák Research group

Title of Thesis:	Bioinformatic approaches for structural analysis of plant genomes
Degree programme:	Biochemistry
Field of Study:	Genomics and Proteomics
Supervisor:	Prof. Mgr. Martin Lysák, Ph.D., DSc.
Academic Year:	2019/2020
Number of Pages:	186
Keywords:	Plant genome; Genome size variability; Repetitive DNA; Transposable elements; Retrotransposons; Tandem repeats; Chloroplast DNA; Phylogenetics; Assembly; Repeat identification; Next generation sequencing

Bibliografický záznam

Autor:	Mgr. Petra Hloušková Přírodovědecká fakulta, Masarykova univerzita Národní centrum pro výzkum biomolekul Laboratoř funkční genomiky a proteomiky CEITEC – Masarykova univerzita Mendelovo centrum genomiky a proteomiky rostlin Výzkumná skupina Martina Lysáka

Název práce:	Aplikace bioinformatických přístupů pro analýzu struktury rostlinných genomů
Studijní program:	Biochemie
Studijní obor:	Genomika a proteomika
Vedoucí práce:	prof. Mgr. Martin Lysák, Ph.D., DSc.
Akademický rok:	2019/2020
Počet stran:	186
Klíčová slova:	Rostlinný genom; Variabilita ve velikosti genomu; Repetitivní DNA; Transpozibilní elementy; Retrotranspozony; Tandemové repetice; Chloroplastová DNA; Fylogenetika; Assembly; Identifikace repetic; Sekvenování nové generace

Abstract

Land plants are well known for their extensive genome size variation. The genome size range is primarily caused by polyploidy and/or proliferation of repetitive sequences. In addition to the nuclear genome, plant cells contain extranuclear, chloroplast and mitochondrial genomes. In particular, chloroplast sequences are useful to resolve the phylogenetic relationships at different taxonomic levels.

The first aim of the thesis was to conduct the phylogenetic analyses to obtain and resolve intratribal relationships within the *Hesperis* clade, one of the major lineages in the mustard family (Brassicaceae), and to date the diversification in this clade using whole-chloroplast sequences retrieved from low-pass Illumina sequence data. It was confirmed that the *Hesperis* clade is a well-supported monophyletic lineage with Miocene tribal diversification.

The second aim of the thesis was to explain the cause(s) of genome size variation in the diploid representatives of the *Hesperis* clade, containing the largest nuclear genomes in the mustard family. Using low-pass NGS data and bioinformatics tools for repeat identification, we aimed to identify qualitative and quantitative differences in the repeat content between the analyzed *Hesperis*-clade genomes. In the absence of recent whole-genome duplication events, we wanted to know whether these genomes are composed of a large number of different repetitive sequences, or of a single or few repeat families amplified to high copy numbers. Our results show that genome obesity has been caused by proliferation of LTR retrotransposons.

We were also interested in the identification of tandem repeats, mainly centromeric-specific satellites. Assembly of centromeric regions still represents a difficult and challenging bioinformatic task, however, newly developed approaches using low-pass sequence data have made the identification of satellite DNA feasible. We were able to *in silico* identify putative satellite sequences, which were used to identify (peri)centromeric chromosome regions in the Arabideae species (Brassicaceae).

Abstrakt

Vyšší rostliny jsou známé svou extrémní variabilitou ve velikosti genomu, způsobenou buďto polyploidií a/nebo proliferací repetitivních sekvencí. Kromě jaderného genomu rostlinné buňky obsahují extranukleární, chloroplastové a mitochondriální genomy. Zejména chloroplastové sekvence představují užitečný zdroj informací pro řešení fylogenetických vztahů na různých taxonomických úrovních.

Prvním cílem této dizertační práce bylo provedení fylogenetických analýz za účelem získání a vyřešení fylogenetických vztahů v rámci kladu *Hesperis*, jedné z hlavních vývojových linií čeledi brukvovitých (*Brassicaceae*), a datovat diverzifikaci této linie pomocí chloroplastových sekvencí získaných z low-pass Illumina sekvenačních dat. Bylo prokázáno, že klad *Hesperis* představuje monofyletickou linii, která se diverzifikovala v období miocénu.

Druhým cílem této dizertační práce bylo vysvětlit podstatu největších diploidních jaderných genomů v kladu *Hesperis*. S využitím dat sekvenování nové generace a bioinformatických nástrojů pro identifikaci repetitivních sekvencí jsme se snažili odpovědět na to, jak se kvalitativně a kvantitativně liší obsah repetitivních sekvencí mezi analyzovanými genomy kladu *Hesperis*. Jelikož u analyzovaných genomů nedošlo k recentní duplikaci celého genomu, zajímalo nás, zda se tyto genomy skládají z velkého počtu různých repetitivních sekvencí nebo z několika málo rodin repetic mnohonásobně amplifikovaných. Výsledky práce prokazatelně ukazují, že „obezita" genomů v kladu *Hesperis* byla způsobena proliferací LTR retrotranspozonů.

Zaměřili jsme se také na identifikaci tandemových repetic, zejména centromerických satelitů. Sestavení centromerických oblastí stále představuje obtížný a náročný bioinformatický úkol, avšak nově vyvinuté přístupy využívající sekvenační data nové generace umožňují snadnější identifikaci satelitní DNA. Podařilo se nám *in silico* identifikovat satelitní sekvence, které byly následně použity k identifikaci (peri)centromerických chromozomových oblastí u druhů tribu Arabideae (čeleď brukvovitých).

Acknowledgements

I would like to thank my supervisor Professor Martin A. Lysák for his guidance, support, extraordinary patience throughout my long studies and for the opportunity to work in his research group.

I would like to sincerely thank Dr. Terezie Mandáková for her support and fruitful discussions. Her passion for work was a great inspiration for me to continue and complete this work. I would also like to thank Milan Pouch for his helpful remarks and his help with probe design.

I am grateful to all my colleagues from the Martin Lysak Research group for a great working and friendly atmosphere.

I wish to express my gratitude to my family and friends, they have been a great support for me. In particular, I wish to thank Řehoř Šiška for his endless and incredible inspiration in life.

This work was supported by Czech Ministry of Education, Youth and Sports within the program INTER-EXCELLENCE (project no. LTAUSA17002), by the CEITEC 2020 project (grant no. LQ1601).

Declaration

I hereby declare that I worked on this thesis independently and I used only the literature stated in the list of references.

Date: Signed:

AUTHOR'S PUBLICATIONS

Description of the author's scientific contribution to the individual publications (in chronological order):

Publication 1:

Mandáková T, **Hloušková P**, German DA, Lysak MA. (2017). Monophyletic origin and evolution of the largest crucifer genomes. *Plant Physiology*. 174(4), 2062-2071.

PH performed *de novo* assembly of 13 chloroplast genomes, phylogenetic analysis based on chloroplast sequences, and the divergence time estimates. PH wrote the respective parts of Materials and Methods and Results.

Publication 2:

Hloušková P, Mandáková T, Pouch M, Trávnícek P, Lysak MA. (2019). The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing. *Annals of Botany*. 124(1), 103–120.

PH performed ancestral genome size reconstruction, all presented statistical analyses, repetitive DNA analysis including the qualitative and quantitative characterization, participated in designing oligoprobes and primers for cytogenetic experiments. PH interpreted data, made the story and wrote, reviewed and edited the manuscript.

Publication 3:

Mandáková T, **Hloušková P**, Koch MA, Lysak MA. (2020). Genome evolution in Arabideae was marked by frequent centromere repositioning. *Plant Cell*. 32(3), 650-665.

PH performed the bioinformatics analyses of the NGS data, identified repetitive sequences, design the oligoprobes and primers further used as cytogenetic probes, performed comparative genome analysis, and participated on writing of the respective parts of the manuscript.

TABLE OF CONTENTS

ABBREVIATIONS

bp	Base pair
cp	Chloroplast
cpDNA	Chloroplast DNA
DBG	De Bruijn graph
DNA	Deoxyribonucleic acid
ENC	Evolutionary new centromere
FISH	Fluorescent *in situ* hybridization
GS	Genome size
GUI	Graphical user interface
HMM	Hidden Markov models
INT	Integrase
IR	Inverted repeat
Kb	Kilo base pairs
LCR	Low-copy repeat
LINE	Long Interspersed Nuclear Element
LSC	Large single copy
LTR	Long terminal repeat
Mb	Mega base pairs
MCMC	Markov chain Monte Carlo
MITE	Miniature Inverted-repeat Transposable Element
ML	Maximum likelihood
mt	Mitochondrial
mtDNA	Mitochondrial DNA
NGS	Next generation sequencing
OLC	Overlap-Layout-Consensus
ORF	Open reading frame
PacBio	Pacific Biosciences
PBS	Primer binding site
PPT	Polypurine tract

PROT	Proteinase
QC	Quality control
rDNA	Ribosomal DNA
RH	Ribonuclease H
rRNA	Ribosomal RNA
RT	Reverse transcriptase
satDNA	Satellite DNA
SD	Segmental duplication
SINE	Short Interspersed Nuclear Element
SMRT	Single Molecule Real Time
SSC	Small single copy
SSR	Simple sequence repeat
TAREAN	Tandem repeat analyzer
TDS	Target duplication site
TE	Transposable element
TIR	Terminal inverted repeat
TR	Tandem repeat
TRF	Tandem repeat finder
WGD	Whole genome duplication

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The rise of next-generation sequencing (NGS) approaches changed the way we are exploring genomes. Rapid development of new sequencing technologies and their decreasing costs have led to the increase of genome sequencing, and this has gone hand in hand with development of the new and more effective computational and bioinformatic approaches. Bioinformatics has become an integral and essential part of modern genomics.

The genome comprises all the genetic material of an organism. Plant cells have three genomes: nuclear, chloroplast (cp) and mitochondrial (mt). The nuclear genome includes coding DNA, represented by genes, and non-coding DNA. The substantial proportion of non-coding DNA is composed of repetitive DNA. There are two categories of repetitive DNA: tandem repeats and dispersed/interspersed repeats, known as transposable elements (TEs).

Plant genomes are known for their variation in structure, complexity, heterozygosity and genome size. In many cases, genome size, the total number of DNA base pairs in one copy of haploid genome (1C value), does not reflect the complexity of plants, primarily land plants. This phenomenon is known as C-value enigma (Gregory, 2001; previously known as C-value paradox; Thomas, 1971). The land plants are known for their extensive genome size (GS) variation. Four major land plant groups are nowadays recognized: (i) bryophytes (non-vascular plants), (ii) lycophytes, (iii) monilophytes and (iv) seed plants represented by two lineages, gymnosperms and angiosperms. Both the smallest and the largest nuclear genomes have been so far found in the most diverse and abundant land plant group, angiosperms (flowering plants; ca. 352,000 species); from 61 Mb in the carnivorous dicot plant *Genlisea tuberosa* (Fleischmann *et al.*, 2014) to the extremely large genome (148,852 Mb) in the monocot *Paris japonica* (Pellicer *et al.*, 2010). Generally, GS distribution across the angiosperms is skewed towards smaller genomes (modal 1C-value = 587 Mb; **Table 1**). On the other hand, gymnosperms tend to have larger genomes, but GS ranges only by a 16-fold (Pellicer *et al.*, 2018; **Table 1**).

Table 1 Genome size variation for each land plant group. Minimum, maximum, mean, modal, median 1C-values, and the GS ranges are shown in Mb (adopted from Pellicer *et al.*, 2018). Approximate number of species recognized (Pellicer *et al.*, 2018) in a given group is shown together with number of species with known genome size.

	Min. (Mb)	Max. (Mb)	Mean (Mb)	Modal (Mb)	Median (Mb)	Range	Approx. no. of species recognized/No. of species with known GS
Non-vascular plants							
Hornworts	156	714	244	176	205	4-fold	250/23
Liverworts	206	20,010	1,844	740	751	97-fold	5,000/102
Mosses	170	2,004	504	442	433	12-fold	12,000/184
Vascular plants							
Lycophytes	78	11,704	1,165	117	127	150-fold	900/57
Monilophytes	748	147,297	14,320	12,073	11,110	196-fold	11,000/246
Seed plants							
Gymnosperms	2,201	35,208	17,947	21,614	21,614	16-fold	1,026/421
Angiosperms	61	148,852	5,020	587	1,663	2,440-fold	352,000/10,768

The astonishing 2440-fold GS variation in the land plants (Pellicer *et al.*, 2018) is to some extent caused by polyploidy, however proliferation of repetitive sequences, especially TEs, is also largely responsible for the existing GS variation (Piegu *et al.*, 2006). Repetitive DNA can represent up to 85% of the barley (The International Barley Genome Sequencing Consortium, 2012) and maize genomes (Schnable *et al.*, 2009). A three-fold GS increase in three *Gossypium* (cotton, Malvaceae) species was found to be due to the accumulation of LTR retroelements (Hawkins *et al.*, 2006). The sequencing of the cotton genomes uncovered an expansion of various TE families, and a massive amplification of one particular group of LTR retrotransposons, *Gorge3*, in two of them, accounting for a major fraction of genome-size change. A study by Estep *et al.* (2013) indicated that in seven panicoid grass species (Poaceae), TEs, especially LTR-retrotransposons, represent the major factors of (>5-fold variation in GS) variation. Macas *et al.* (2015) conducted large-scale comparative analysis of repetitive sequences in 23 Fabaceae species, having a 7.6-fold variation in GS. Repeats represented from 55 to 83% of the Fabaceae genomes and, interestingly, GS variation was found to be driven by a single lineage of LTR retrotransposons, Ogre element. Tetreault and Ungerer (2016) explored the

contribution of LTR retrotransposons to GS variation among eight diploid *Helianthus* (sunflower, Asteraceae) species and found out that the largest analyzed genome had the highest genomic repetitive fraction, and has experienced species-specific amplification of LTR retrotransposons. The amplification of repetitive DNAs as an important determinant of the GS variation has been shown in many other, and still increasing, studies of plant genomes (e.g. Zuccolo *et al.*, 2007; Ambrožová *et al.*, 2011; Nystedt *et al.*, 2013; Renny-Byfield *et al.*, 2013; Kelly *et al.*, 2015; Piednoël *et al.*, 2015).

The classical Sanger sequencing has been replaced by modern NGS methods that spurred the rate at which new plant genome sequences are being generated. Even at low genome coverage, NGS can greatly help with identification of the repetitive composition of a genome. Low-pass NGS data has been shown to be very effective in identifying repeat content in many plant species analyzed (e.g. Kelly and Leitch, 2011; Weiss-Schneeweiss *et al.*, 2015; Macas *et al.*, 2015). The low-coverage genome sequencing approach allows us not only to retrieve repetitive DNA sequences, but also other sequences presented in high abundances, such as rDNA, organelle genome sequences (cpDNA and mtDNA) and high-copy genes (Straub *et al.*, 2012). Moreover, this cost-effective method has become popular and widely used in phylogenetic studies based on whole chloroplast sequences (e.g. Weitemier *et al.*, 2014; Vitaceae: Zhang *et al.*, 2015; Chrysobalanaceae: Malé *et al.*, 2014; Clauseneae from Rutaceae family: Shivakumar *et al.*, 2017; Meliaceae: Mader *et al.*, 2018). The fact that chloroplast genomes are highly conserved compared to nuclear ones make them useful in phylogenetic studies at different taxonomic levels.

1. 1 PLANT GENOME

The genome includes coding DNA, represented by genes (unique sequences), and non-coding DNA. The substantial portion of non-coding DNA in a nuclear genome (**Figure 1**) is composed of repetitive DNA sequences which occur multiple times in the genome. There are two categories of repetitive DNA: tandem repeats and dispersed/interspersed repeats (or transposable elements, TEs).

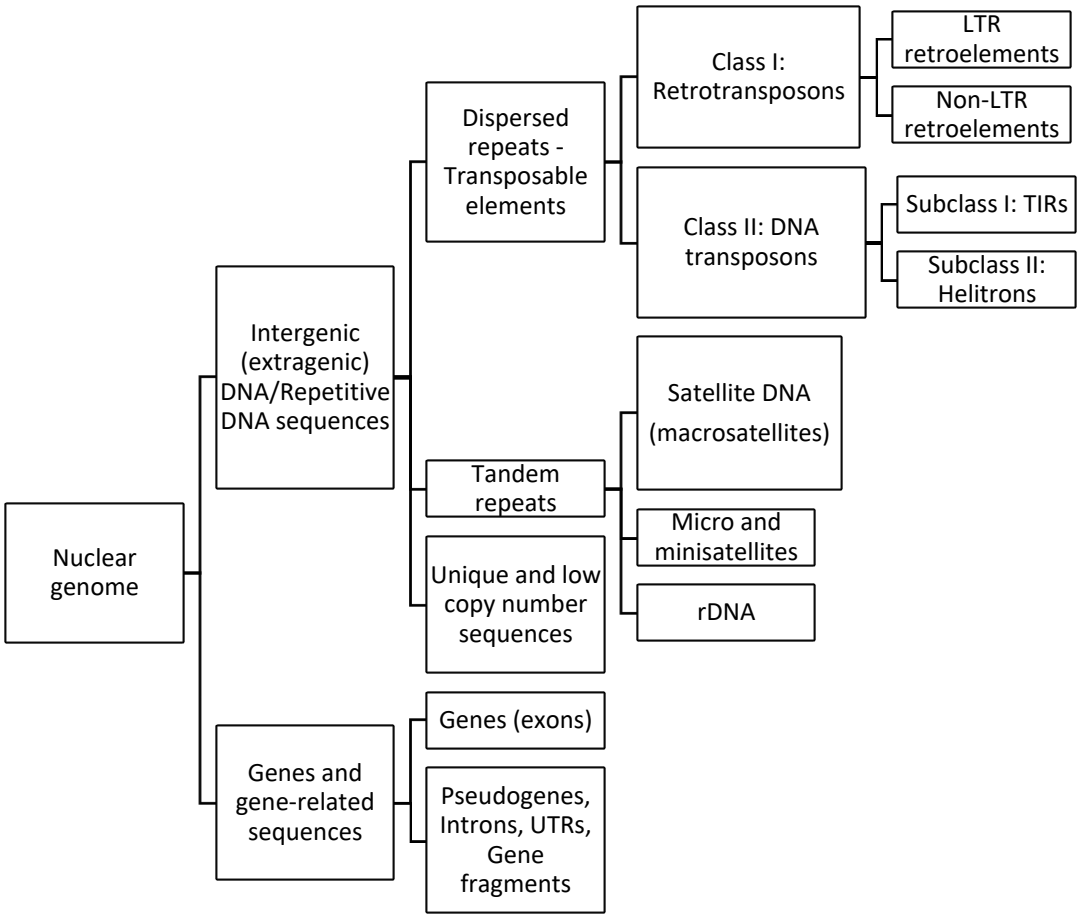


Figure 1 Nuclear genome composition (Richard *et al.*, 2008).

The number of coding genes is relatively similar in all eukaryotes, although it can vary to some extent, especially in relation to polyploidy. Polyploidy (or whole-genome duplication, WGD), i.e. the presence of multiple copies of identical (autopolyploidy) or

different (allopolyploidy) chromosome sets in one composite genome (Bennetzen, 2002), is common in the plant kingdom, particularly in angiosperms and pteridophytes, but rare in gymnosperms and bryophytes (Van Straalen and Roelofs, 2011). All angiosperm genomes are of paleopolyploid origin (Soltis *et al.*, 2009; Jiao *et al.*, 2011), whereby angiosperm genome evolution is based on cyclic alternation of polyploidization events and diploidization processes, known as post-polyploid genome diploidization (Mandáková and Lysak, 2018). This process involves gradual loss of duplicated genes. Thus, a large diploid genomes do not necessary possess more genes. Estimated number of protein-coding genes in land plants heads toward a plateau between 20,000 and 50,000 genes (Van Straalen and Roelofs, 2011). In particular, the relatively small 157-Mb genome of *Arabidopsis thaliana* (Bennett *et al.*, 2003) has 27,655 estimated protein-coding genes, *Arabidopsis lyrata* genome (207 Mb, Hu *et al.*, 2011) has 32,667 genes, *Vitis vinifera* genome (~500 Mb; Jaillon *et al.*, 2007; Velasco *et al.*, 2007) has 29,971 genes, *Glycine max* genome (1,115 Mb; Arumuganathan and Earle, 1991) has 55,897 genes, and monocot *Zea mays* genome (~2,600 Mb; Bennett and Smith, 1991) has 39,591 genes (numbers of coding genes for each species from https://plants.ensembl.org). Nystedt *et al.* (2013) showed in the Norway spruce genome (20,000 Mb) that the number of genes in gymnosperms is similar to the much smaller angiosperm genome of *Arabidopsis thaliana* (157 Mb). Having more coding genes does not stand for greater genomic complexity (Pray, 2008) or larger genome size.

In the absence of polyploidy, the repetitive DNA is undoubtedly the main cause responsible for the C-value paradox, referring to the tremendous genome size variation among higher plants and to the fact that genome size does not reflect complexity of organisms. This is in contrast to prokaryotic and lower eukaryotic genomes in which GS positively correlates with the morphological complexity of organism. So far, the monocots are showing the most extensive GS diversity among angiosperm species (from 196 to 148,852 Mb), followed by eudicots (from 61 to 77,555 Mb; Leitch *et al.*, 2010). The large genome size seems to result from the accumulation of TEs. Prior to the discovery of the immense genome of *Paris japonica*, for a long time the tetraploid *Fritillaria assyriaca* (Liliaceae) was considered as the largest known genome of 124,600

Mb (Bennett and Smith, 1976). Kelly *et al.* (2015) studied giant genomes of *Fritillaria* (Liliaceae) and their results showed that a lack of DNA removal and low turnover of repetitive DNA are major contributors to the evolution of extremely large genomes.

Together with polyploidization, the rapid proliferation of TEs, and associated lack of an efficient elimination mechanism of repetitive DNA, belongs to the major mechanisms of GS growth in plant genomes.

1.1.1 REPETITIVE ELEMENTS IN GENOMES

Repetitive DNA represents sequences repeated hundreds or thousands times in the genome and makes up the substantial portion of all genomes. Especially in plant genomes, the complexity of genomes is usually not correlated with gene content but rather with the abundance of repetitive elements (Devos *et al.*, 2002). Repeats can be classified based on several criteria, like size, frequency of occurrence, distribution pattern, biological function or replication mechanism.

Beside the main two groups of repetitive elements (tandem repeats and transposable elements), low-copy repeats (LCRs), also known as segmental duplications (SDs; Eichler, 2001), are another type of highly homologous (>90% sequence identity) sequence elements within eukaryotic genomes. SDs are blocks of long DNA sequences (1 to 400 kb) that can occur tandemly or interspersed in multiple locations in a genome as result of duplication events (Eichler, 2001).

1.1.1.1 Transposable elements

Transposable elements, or jumping genes, are defined as DNA sequences capable of movement from one chromosomal location to another inside the same genome, either through replicative, copy-and-paste, or conservative, cut-and-paste, mechanism. Barbara McClintock's discovery of the transposable elements (summarized in McClintock, 1950) in maize (Activator/Dissociation system) more than 70 years ago has changed the view of the dynamic nature of chromosomes. She noticed that activities of

these elements can lead to a change in the color of maize kernels or unusual color patterns on the leaves. Today, the TEs are well-known for their ability to rearrange genomes, through transposition, insertion, excision, chromosome breakage, and ectopic recombination (Bennetzen, 2000). The TEs create genetic variation and diversity by insertion mutations, modification of individual gene structure, expression and function, e.g. by reprogramming of gene expression by insertion into repressors and enhancers (Lisch, 2013). They can generate novel genes through transposition and contribute to adaptation (Oliver and Greene, 2009). TEs can serve as templates to repair double strand breaks (DSBs) in DNA (Vu *et al.*, 2015; Schubert and Vu, 2016). Plant genomes have developed defense mechanisms to epigenetically silence TE activity with RNA interference (Roessler *et al.*, 2018) or through DNA methylation (Hollister and Gaut, 2009).

The first TE classification system, suggested by Finnegan (1989), was based only on TE mechanism of replication (transposition). Transposition of TE can be mediated through an RNA or DNA intermediate (**Figure 2**). Class I transposable elements, retroelements, mobilize through a copy-and-paste, replicative, mechanism. The RNA intermediate of the element is reverse-transcribed into a cDNA copy that is integrated elsewhere in the genome. Class II elements, DNA transposons, are mobilized through a DNA intermediate directly through a cut-and-paste, conservative, mechanism (Bourque *et al.*, 2018).

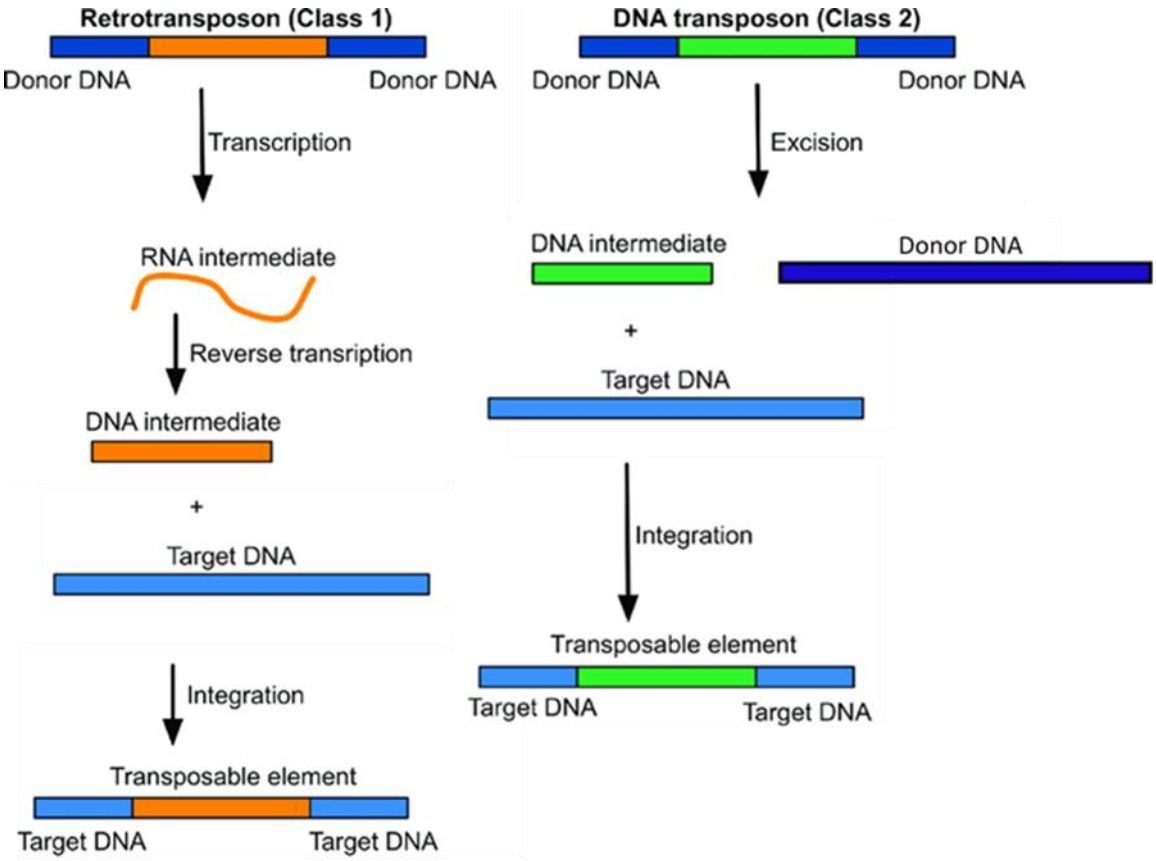


Figure 2 TE mechanism of replication, mobilization happens either in presence of an RNA or a DNA intermediate. Transposition of Class I elements involves amplification of the element by copying through transcription (RNA intermediate) followed by reverse transcription (DNA intermediate). The new copy of element is inserted elsewhere in the genome, the donor element stays at the original position. During transposition of Class II elements, the element is excised from the donor DNA (DNA intermediate) and integrates into a new position in the genome. (Figure adapted from Ågren and Clark, 2018)

This classification criterion had to be adjusted when elements that could move without intermediates (non-autonomous TEs) were discovered. Studies based on transposition mechanisms, sequence similarities, structural relationships and phylogenetic studies were carried out by Wicker *et al.* (2007) and resulted in the first unified hierarchical classification of TEs (**Figure 3**). In addition to the classical division into two classes, the TEs were further subdivided into subclasses, orders and superfamilies. The subclasses are used to distinguish elements that copy themselves from those that reincorporate elsewhere and reflect the number of DNA strands that are cut at the TE donor site (Wicker et al, 2007). The order level characterizes major differences in the insertion

mechanism (Wicker *et al.*, 2007). Superfamilies share a replication strategy, but are distinguished by the structure of protein or non-coding domains, and presence and size of the target site duplication (TSD). The TSDs are short, 4 to 6 bp long, identical duplicated sequences, which are characteristic for majority of superfamilies and are generated on both sides of insertion site of the target DNA and can be found in most elements of both TE classes. A single large genome can comprise hundreds or thousands of diverse TE families (Wicker *et al.*, 2007). The TEs can be further described as autonomous or non-autonomous, depending on whether they encode enzymes required for their own transposition. The autonomous TEs are capable of self-mobilization because they encode proteins which are essential for transposition from one chromosomal location to another. The non-autonomous TEs (e.g. Miniature Inverted-repeat Transposable Elements, MITE) do not encode such proteins and are dependent on autonomous TEs and their transposition mechanism; Wicker *et al.*, 2007).

Llorens *et al.* (2007, 2009, 2010) established the Gypsy Database (GyDB), an open editable database focusing on the evolutionary relationship of viruses, mobile genetic elements and the genomic repeats. Kapitonov and Jurka (2008) implemented a universal classification of eukaryotic TEs in Repbase, a database of repetitive DNA (http://www.girinst.org/) which is now the most commonly used database of repetitive DNA elements.

As the number of sequenced genomes increases, more sequences are now available. Modern genomics and bioinformatics approaches can be applied to characterize diverse TEs from different eukaryotic organisms and to understand the way how the TEs can influence host organisms. In plant genomes, we can find almost all the TEs known (for now) but they can significantly differ both in quantitative and qualitative level among species, even among populations (e.g. in maize; Díez *et al.*, 2013). Elliott and Gregory (2015) addressed question whether larger genomes contain more diverse TEs. They used all eukaryote data available but did not observe any linear relationship between diversity of TE superfamilies and genome size. Interestingly, they showed that land plants display much lower overall TE diversity as compared to animals, even though plants exhibit significantly more extensive GS variation (Elliott and Gregory, 2015).

Similarly, Kelly *et al.* (2015) have proven that genomic expansion can take place through the accumulation of highly heterogeneous, relatively low-abundant, repeat-derived DNA in large genomes of *Fritillaria* species (Liliaceae). Macas *et al.* (2015) found out that most of the inter-species GS variation in the tribe Fabeae (Fabaceae) was caused by accumulation of a single lineage of Ty3/*gypsy* LTR retrotransposons, the Ogre elements.

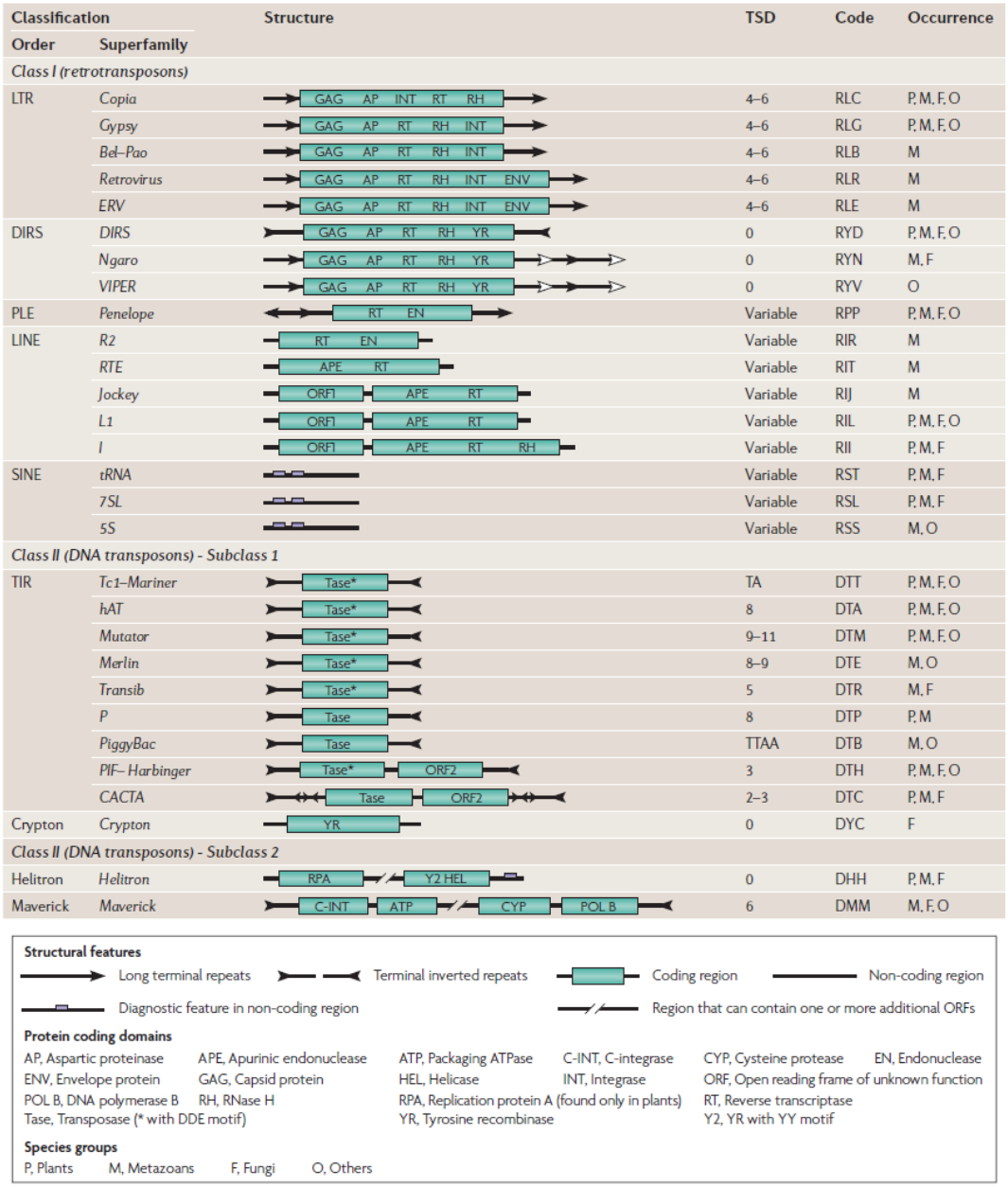


Figure 3 The hierarchical classification of eukaryotic TEs, adopted from Wicker *et al.* (2007).

CLASS I – RETROTRANSPOSONS

Retrotransposons encode the enzyme reverse transcriptase and spread via the process of retrotransposition, known as copy-and-paste mechanism (**Figure 2**). An RNA intermediate inferred from the DNA template is incorporated as a copy of the template into a new genome location. Because their copy-and-paste replication mechanism, retrotransposons are often the major contributors to the repetitive sequence content in plant genomes. According to Wicker *et al.* (2007) eukaryotic retrotransposons can be categorized into five subclasses (**Figure 3**): (i) Long terminal repeat (LTR) retrotransposons, (ii) Long Interspersed Nuclear Elements (LINEs), (iii) DIRS-retrotransposons (Dictyostelium intermediate repeat sequence), (iv) Penelope-like retrotransposons (PLE), and (v) Short Interspersed Nuclear Elements (SINEs).

Two major groups are known in plants, LTR retrotransposons and non-LTR retrotransposons (including LINE, SINE).

LTR retrotransposons

In plants, LTR retrotransposons often make up the largest portion of the TEs (Piednoël *et al.*, 2012; Macas *et al.*, 2015; Dodsworth *et al.*, 2017; Hloušková *et al.*, 2019). Although the LTR retroelements are highly diverse in their nucleotide sequences, their overall structure is highly conserved. The common feature of LTR retroelements is the presence of long terminal repeats (LTRs) at both ends (the 5’ and 3’ LTR) in direct orientation, containing promoter sequences necessary for transcription of the element. Most LTR retroelements have a primer binding site (PBS) downstream of the 5’ LTR and a polypurine tract (PPT) upstream of the 3’ LTR (**Figure 4**; Neumann *et al.*, 2019). The central part contains two open reading frames (ORFs) for GAG gene (encoding structural protein) and polyprotein POL. POL encodes proteinase (PROT), reverse transcriptase (RT), ribonuclease H (RH) and integrase (INT) enzymes. Length of the LTR elements ranges up to 25 kb, where LTRs occupy from few hundreds bp to 5 kb (Wicker *et al.*, 2007).

Plant LTR retrotransposons are classified into two superfamilies, Ty1/*copia* and Ty3/*gypsy*, according to internal order of genes encoded by POL (**Figure 3**). Neumann *et al.* (2019) established and improved plant LTR-retroelement classification system based on phylogenetic analyses of the three most conserved polyprotein domains (RT, RH and INT) in 80 species. New lineages were discovered and some lineages were reclassified in comparison to Wicker *et al.* (2007) and Llorens *et al.* (2010). This led to division of Ty1/*copia* into 16 lineages (Ale, Alesia, Angela, Bianca, Bryco, Lyco, Gymco-I, Gymco-II, Gymco-III, Gymco-IV, Ikeros, Ivana, Osser, SIRE, TAR, and Tork) and Ty3/*gypsy* into two major lineages (chromovirus and non-chromovirus). The vast majority of chromovirus lineage elements differs from other Ty3/*gypsy* lineages by presence of the chromodomain at the C-terminal region of POL. Both Ty3/*gypsy* major lineages comprising together 14 lineages (chromovirus|CRM, chromovirus|Chlamyvir, chromovirus|Galadriel, chromovirus|Tcn1, chromovirus|Tekay, chromovirus|Reina, non-chromovirus|OTA|Athila, non-chromovirus|OTA|Tat|TatI-III, non-chromovirus|OTA|Tat|Ogre, non-chromovirus|OTA|Tat|Retand, non-chromovirus|Phygy, non-chromovirus|Selgy). These findings were brought together into a comprehensive database of retrotransposon protein domains (REXdb, Neumann *et al.*, 2019) implemented in the RepeatExplorer web server (https://repeatexplorer-elixir.cerit-sc.cz/).

The Ty3/*gypsy* lineages seem to be predominant over Ty1/*copia* elements in plant genomes. One of the best known lineages of plant Ty3/*gypsy* retrotransposons is Athila lineage found in heterochromatic regions of the *A. thaliana* genome (Pélissier *et al.*, 1995). Some Athila retrotransposons are structurally identical to simple retroviruses and have a third ORF that may encode an envelope (*env*) protein (Wright and Voytas, 1998; **Figure 4**). The *env* protein was also found in SIRE lineage from Ty1/*copia* superfamily in the soybean genome (Laten *et al.*, 1998).

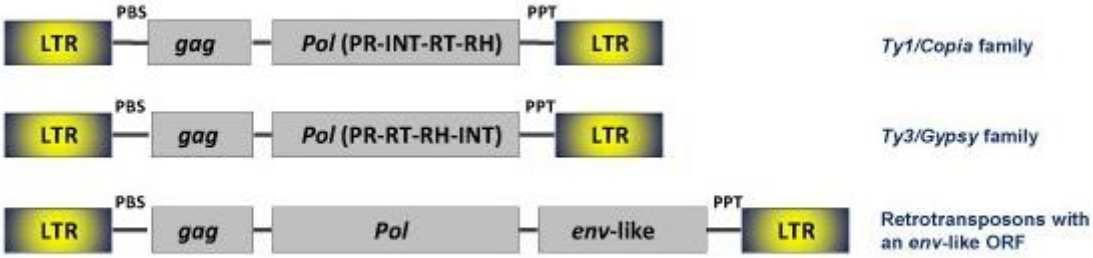


Figure 4 Detailed structure of LTR retrotransposon with and without env-like domain. (Adopted from http://gydb.org/index.php/LTR_retroelements)

The largest plant LTR retroelements are Ogre elements which were found in legume genomes and their size can reach up to 25 kb (Macas and Neumann, 2007). The smallest known LTR retrotransposons identified are non-autonomous TRIMs (terminal repeat retrotransposons in miniature; Witte *et al.*, 2001) which have terminal direct repeats (100-250 bp) as other LTR retrotransposons but lack the coding domains required for element's mobility. Their internal domain (about 100-300 bp) contains only PBS and PPT motifs. Kalendar *et al.* (2008) described TRIM element Cassandra, which can be found in almost all vascular plants and carries 120-bp conserved 5S RNA domains with well conserved RNA polymerase III promoters in their LTRs (Kalendar *et al.*, 2008). Another non-autonomous LTR retrotransposon derivatives found in Triticeae (Poaceae) are called the large retrotransposon derivatives (LARD; Kalendar *et al.*, 2004). They have long, conserved ends and their coding region is replaced by a large (usually >4 kb) conserved non-coding DNA sequence.

Non-LTR retrotransposons

Non-LTR transposons (also known as retroposons) do not contain LTRs and use a target-primed reverse transcriptase.

LINEs (Long Interspersed Nuclear Elements) contain RT, endonuclease and either (A)n tail or just A-rich region. In the human genome, LINEs built up one fifth of genome as

the most abundant repeat (International Human Genome Sequencing Consortium, 2001). In plants, they seem to be less abundant compared to LTR retrotransposons (except for Del-2, a LINE-like element found in *Lilium speciosum*; Leeton and Smyth, 1993). These findings could have been influenced by extreme heterogeneity of LINEs caused most likely by the error-prone RT and accumulation of mutations in LINE copies over long evolutionary periods (Schmidt, 1999) and subsequent difficult *in silico* identification. Although LINEs are usually dispersed along chromosomes, including pericentromeric regions, in banana (Musa, family) their preferential localizations are centromeric regions, and together with other retrotransposons are the main components of banana centromeres (Čížková *et al.*, 2013).

SINEs (Short Interspersed Nuclear Elements) are non-autonomous elements, as they do not have their own mechanism for retrotransposition. They borrow a RT from LINE-like active element which can recognize the sequence at the 3' end of the SINE (Wicker *et al.*, 2007).

Although LINEs and SINEs make only a small contribution to the plant genome size variation, they are widespread in all plant genomes.

Penelope-like retrotransposons were first found in the fly *Drosophila virilis* (Evgen'ev *et al.*, 1997), and since then have been detected also in animals, fungi and plants (Evgen'ev and Arkhipova, 2005). Their RT is more related to telomerase than to the RT of LTR-retrotransposons or LINEs, and their LTR-like sequences can be in a direct or an inverse orientation (Wicker *et al.*, 2007).

DIRS-retrotransposons (Cappello *et al.*, 1985) do not have INT and therefore are not able to form TSDs. They use a different mechanism of integration than the LTR retroelements, as they contain a tyrosine recombinase instead of INT. DIRS-retrotransposons either have split direct repeats or inverted repeats, and have been found in green algae, fungi and animals (Goodwin *et al.*, 2004).

CLASS II - DNA TRANSPOSONS

DNA transposons can move without any RNA intermediate as they cut themselves from one chromosomal location and integrate into another one. The DNA transposons are usually less abundant in plants than other TEs.

The DNA transposons are divided into two subclasses (Wicker *et al.*, 2007): (i) subclass I is spread through conservative classic, cut-and-paste, transposition, and (ii) subclass II elements (known as Helitrons) whose replication takes place without DNA DSBs, via a single-stranded DNA intermediate, and they spread through a rolling-circle replicative transposition mechanism using a rolling circle replication protein and a helicase (Kapitonov and Jurka, 2001).

Autonomous subclass I elements mostly encode only one protein, transposase, which is flanked by terminal inverted repeats (TIRs) of variable length. The TIRs are recognized by the transposase, this leads to double stranded DNA excision of element, and its integration to another location in the same genome where short TSD, 2–10 bp, are generated (Wicker *et al.*, 2007). Subclass I DNA transposons have been grouped into several superfamilies, such as CACTA (En/Spm), Mutator (MuDR), PIF-Harbinger, hAT, and Tc1 Mariner (Stowaway and Pogo). Superfamilies are characterized by specific transposases and specific length of TSDs (**Figure 2**).

MITEs (Miniature Inverted-Repeat Transposable Elements; MITE Tourist element firstly discovered in maize by Bureau and Wessler, 1992) are non-autonomous, short (<500 bp) DNA transposons, which do not encode their own transposase. They have arisen from related autonomous elements and their classification has been based on the similarity of TIRs and TSDs.

Another subclass II elements, self-synthesizing transposons Mavericks (also known as Polintons; Pritham *et al.*, 2007) with an average size of ∼15-20 kb are widespread in animals (mainly in invertebrates and non-mammalian vertebrates) but have not been detected in plants.

1.1.1.2 Tandem repeats

Tandem repeats (TRs) form long arrays of tandemly, head to tail, arranged highly conserved motifs (repeat units, monomers). They vary in repeat unit length, nucleotide sequence composition, genomic abundances and copy numbers. Monomer length is the most frequently used feature to classify tandem repeats. The TRs are classically classified as (i) microsatellites (also called simple sequence repeats, SSRs), (ii) minisatellites, and (iii) macrosatellites or satellite DNA (satDNA). There is no consensus about precise definition of micro- and minisatellites (Richard *et al.*, 2008). Microsatellites can be defined as the TRs of less than 10 bp in length or 1-6 bp in length in arrays less than 1 kb. Some authors do not consider mononucleotide repeat as microsatellites (Richard *et al.*, 2008), and classify the microsatellites as repeats with monomer size between 2 and 5 with an array size of the order of 10 to 100 repeat units (Mehrotra and Goyal, 2014). The microsatellites usually show high levels of polymorphism, mutation rate increases with an increasing number of repeat units (Ellegren, 2004). The minisatellites are characterized by longer repeats (>6 bp or 10 bp to 100 bp, with an array size of 0.5–30 kb; Richard *et al.* 2008; Garrido-Ramos, 2017). The macrosatellite monomers are larger than hundreds of bp, typically organized in long arrays that can occupy up to several million nucleotides of a genome.

The satDNAs, composed of families of TRs, are located preferentially at heterochromatic regions, which are found mostly in (peri)centromeric and subtelomeric regions of chromosomes (Mehrotra and Goyal, 2014), while the micro- and minisatellites have capability to cluster both in euchromatin and heterochromatin (Garrido-Ramos, 2015, 2017) and it was found that microsatellites are more abundant in the neighborhood of the TEs and it was suggested that they probably serve as targets for insertions of the TEs (Kejnovský *et al.*, 2013).

Centromeric satDNA

Centromeric satDNA sequences are not conserved among organisms and are the most rapidly evolving DNA sequences in eukaryotic genomes (Henikoff *et al.*, 2001; Melters

et al., 2013). The amount of the centromeric satDNA varies greatly from chromosome to chromosome (Jiang *et al.*, 2003). Furthermore, the presence of tandem repeats in (peri)centromeric regions suggest an important functional role in the centromere formation. Frequently, monomer length of the satDNA tends to have the length about the size of the DNA sequence wrapped around one nucleosome, from 160 to 180 bp (Biscotti *et al.*, 2015a; Garrido-Ramos, 2015). Length of most plant centromeric repeats is within this range (e.g. 178 bp in *Arabidopsis thaliana*; Heslop-Harrison *et al.*, 1999; 176 bp in *Brassica napus*; Xia *et al.*, 1993; 155 bp in rice; Cheng *et al.*, 2002; 156 bp in maize; Ananiev *et al.*, 1998). But with more centromeric satellite discovered either *in silico* or by wet-lab techniques, the assumption about constraint monomer length seems not to be a universal rule as repeat monomers were found to be variable in length (Melters *et al.*, 2013; Garrido-Ramos, 2015). A pericentromeric satellite repeat with long monomers of 5.9 kb (2D8; Stupar *et al.*, 2002) and 4.7 kb (Sobo; Tek *et al.*, 2005) was isolated from the *Solanum bulbocastanum* genome. Centromeric satellite repeats of short monomer length from 44 to 50 bp, as well as tandem repeats with monomer length up to 2 kb was found in *Vicia faba* (Ávila Robledillo *et al.*, 2018). Gill *et al.* (2009) identified and isolated centromeric satellite repeat in soybean of 92-bp length. The satellite sequences are also highly variable in sequence composition, abundance and number of tandem repeat families within one individual species (e.g. more than 30 satellites found in Fabaceae, Macas *et al.*, 2015).

Telomeric repeats

Telomeres are regions of simple repetitive sequences normally located at the end of chromosomes that are crucial for genome stability. However, they can be also found as short tandem arrays in non-terminal regions of chromosomes, e.g. at centromeres (*Arabidopsis thaliana*: Uchida *et al.*, 2002; Solanum species: He *et al.*, 2013), and are called interstitial telomeric repeats.

The telomere repeat array in most plants is composed of *Arabidopsis*-type 7-bp telomeric repeat (TTTAGGG)n. But some sequence variants were found. Asparagales

possess human-type 6-bp motif (TTAGGG)n (Sykorova *et al.*, 2003) and in *Allium* unusual sequence motif of (CTCGGTTATGGG)n was found (Fajkus *et al.*, 2016). In *Cestrum* (Solanaceae) telomeric repeat (TTTTTTAGGG)n was discovered (Peška *et al.*, 2015), and in some species from the carnivorous genus *Genlisea* two variants of telomere repeat, (TTCAGG)n and (TTTCAGG)n, were reported (Tran *et al.*, 2015).

Together with telomeric repeats, subtelomeric tandem repeats or telomere-associated sequences can be found at the ends of chromosomes (Garrido-Ramos, 2015). Monomer length of subtelomeric satDNA is variable, e.g. 352-bp satellite pBrSTR in *Brassica rapa* (Koo *et al.*, 2011), and two subtelomeric satellites of length 182- and 339-bp in *Solanum tuberosum* (Torres *et al.*, 2011).

1.1.1.3 Role of repetitive DNA sequences in plant genomes

In early days of genomic research, repeats were considered as junk, selfish or parasitic DNA (Doolittle and Sapienza, 1980; Orgel and Crick, 1980), with no benefits for their hosts, and with an unknown function. Although the significance of repeats is still not completely understood, with advances in molecular biology and growing knowledge of sequenced genomes, many studies have shown that repetitive DNA represents an important structural, functional and regulatory part of all living organisms. Repeats play a crucial role in genome evolution, plasticity and speciation (Lisch, 2013).

Many studies have been carried out in recent years with the aim to determine and clarify the biological role and importance of repetitive DNA sequences in eukaryotic genomes. Albeit majority of the TEs in plants are inactivated or silenced (Okamoto and Hirochika, 2001; Fultz *et al.*, 2015), the functional consequences on the genome are indisputable. The TEs can be expressed under stress conditions (e.g., Tnt1 elements, LTR retroelements in the Solanaceae family; Grandbastien *et al.*, 2005). The capability of the TEs to move and insert within a genome generates variability of repetitive components (Biscotti *et al.*, 2015a). Lisch (2013) reviewed the kinds of changes that the TEs can cause and discussed evidence how those changes have contributed to plant adaptation and evolution. TE insertions can have effect on gene structure and function, including

knockout of gene function, introduction of new functions, changes in the structure of genes, mobilization and rearrangement of gene fragments and epigenetic silencing of genes (Lisch, 2013). The connection between Ac/Ds transposons and DSBs was found in maize and rice (Zhang and Peterson, 2004; Xuan *et al.*, 2012).

Centromeres are functional chromosome domains that are necessary for chromosome segregation during the cell division and usually contain large blocks of satDNA and dispersed repeats (Heslop-Harrison *et al.*, 2003). Centromere-specific retrotransposons were found to be active component of centromeres in many angiosperms (Neumann *et al.*, 2011). Together with satDNAs they may play the key role in plant centromere evolution and function (Neumann *et al.*, 2011). Biscotti *et al.* (2015b) described transcriptional activity of tandem repeats and that satDNA-derived transcripts play a structural function in the heterochromatin formation and maintenance at both centromere and telomere. But there are still limitations in our understanding of the function of satDNA, as centromeres seem to be defined epigenetically and DNA sequences are not specifically required for centromere specification (McKinley and Cheeseman, 2016).

1.1.1.4 Mechanisms contributing to genome size expansion and reduction

Polyploidy and LTR retrotransposon amplification through burst of retrotransposition (Piegu *et al.*, 2006) are two key mechanisms responsible for genome growth. satDNA can as well as TEs greatly increase genome size of plant genomes. The mechanism that led to the satDNA accumulation is DNA replication errors and unequal crossing-over. About 25% of the *Vicia sativa* genome is represented by satDNA (Macas *et al.*, 2000) and up to 36% of the *Fritillaria falcata* genome (Ambrožová *et al.*, 2011) is made up by satDNA.

On the contrary, the counterbalancing mechanism to genome growth is process of elimination, genome size decrease, which prevents genomes from uncontrolled expansion (Pellicer *et al.*, 2018). Without a mechanism of TE sequences removing, plants may end up with a 'one-way ticket' to genomic obesity (Bennetzen and Kellogg, 1997).

Study by Devos *et al.* (2002) and Hu *et al.* (2011) indicated that illegitimate recombination represented the key mechanism of genome size decrease in *Arabidopsis thaliana*. Intra-element unequal recombination between LTRs of the same element and inter-element unequal intrastrand recombination between LTRs of different retroelements belonging to the same lineage can result in a DNA loss (Devos *et al.*, 2002). Recombination between two TEs results in deletion of sequence region between these two copies. Deletions caused by recombination between identical LTR sequences of of the same TE, removing the internal transposon sequence, lead to the formation of solo-LTRs. The existence of solo-LTRs has been so far found in many species, e.g. in barley (Shirasu *et al.*, 2000), *Arabidopsis thaliana* (Devos *et al.*, 2002), and rice (Vitte and Panaud, 2003). Macas *et al.* (2015) have developed a bioinformatic approach to estimate proportion of solo-LTRs by quantifying the number of sequencing reads containing the junction between the LTR 3' end and the internal retrotransposon region versus the reads containing just the LTR 3' end alone. While Vu *et al.* (2015) supported deletion-biased DSB repair (review in Schubert and Vu, 2016) as the main mechanism of genome shrinkage by non-specific DNA loss in the genome of carnivorous plant *Genlisea nigrocaulis* and that this mechanism played a more significant role in an overall evolution of plant genome size.

1.1.1.5 Identification of repetitive sequences

Knowledge of the repeat content and composition in genome enables us to estimate the genome complexity. Even with the advance of NGS, the detection, identification and annotation of repetitive sequences are still very challenging and not trivial bioinformatics tasks, as DNA repeats represent technical problems for sequence alignment and genome assembly algorithms. Genome assembly refers to the process where sequencing reads are put together to reconstruct the complete original genome sequence. Because many repeats have been present in genome for a long time, copies of repeat from the same family are not fully identical as the evolutionary mechanisms cause point mutations, indels and even rearrangements (Lerat, 2010). These mechanisms often result in fragmented and diverse copies of repeats that are difficult

to identify by similarity-based approaches (Lerat, 2010). Many repeats are preferentially inserted into intergenic regions and even into other repeats, forming nested repeats (SanMiguel *et al.*, 1996; Jiang and Wessler, 2001; Gao *et al.*, 2012; Wei *et al.*, 2013).

Repeats can be analyzed by similarity-based, signature-based or *de novo* (*ab initio*) methods (Bergman and Quesneville, 2007). Methods are applicable to either unassembled reads ("assembly-free" methods) or to an assembled genome. Many bioinformatics methods and tools have been invented over past two decades. Widely used methods, especially for the TE identification and annotation, were evaluated and compared in several reviews (Bergman and Quesneville, 2007; Saha *et al.*, 2008a, 2008b; Lerat, 2010; Janicki *et al.*, 2011; Ewing, 2015; Goerner-Potvin and Bourque, 2018).

Similarity-based or homology-based methods

This group of computational methods is based on similarity search and comparison of input sequence(s) with database of already known repetitive consensus sequences (references), so called repeat libraries (library-based methods).

The most commonly used repeat database is RepBase (RepBase Update; Jurka *et al.*, 2005, Bao *et al.*, 2015) which is a well-curated reference database of eukaryotic repetitive DNA sequences. The RepBase database includes prototypic sequences of the repeats and basic annotations of reference sequences. A tool Censor was designed to facilitate screening of the content of the RepBase and to identify repetitive elements by comparison with known repeats (Kohany *et al.*, 2006). It uses AB-BLAST (formerly known as WU-BLAST; https://blast.advbiocomp.com/) as a sequence search engine (heuristic alignment algorithm). Output includes a map of repeats present in the query sequence, masked query sequence (repetitive nucleotides of query sequence are replaced by Ns), repeat sequences found in the query, and alignments (Kohany *et al.*, 2006). It is available as both web-based service (https://www.girinst.org/censor/index.php) and downloadable version. For creating, formatting and annotating of new Repbase entries, a java-based interface RepbaseSubmitter was developed (Kohany *et al.*, 2006).

The most popular library-based program is RepeatMasker (Smit *et al.*, 2013–2015). It uses several search engines, e.g. cross_match (http://www.phrap.org), and AB-BLAST. The output includes a detailed annotation of the repeat(s) presented in the query sequence as well as a masked query sequence(s). RepeatMasker uses curated libraries of repeats and supports the RepBase databases.

The similarity-based approach has its limitations in detection of repeats which do not share conserved sequences between species or are composed entirely of non-coding sequences (e.g. non-autonomous MITEs, SINEs). As this approach is solely based on homology to already known existing sequences, it cannot detect completely novel repetitive elements. However, both Censor and RepeatMasker can use a repeat library created by a *de novo* method.

An alternative approach is software HMMER (Eddy, 2009; Finn *et al.*, 2011) that scans predicted ORFs using probabilistic methods (hidden Markov models, HMMs) from the PFAM database (Finn *et al.*, 2010) and InterPro (Hunter *et al.*, 2008) to detect common TE protein domains.

Signature-based methods

Signature-based methods search a query sequence for particular structural features that are characteristic for a given class of repeat. This approach can be used to find new elements, but not a new class of elements with unknown structures (Lerat, 2010).

Program LTRharvest (Ellinghaus *et al.*, 2008) was developed for a *de novo* detection of full length LTR retrotransposons in large genomes based on known features like length, distance, LTR or TSD. Other LTR prediction software tools are LTR_STRUC (McCarthy and McDonald, 2003), LTR_par (Kalyanaraman and Aluru, 2006), LTR_FINDER (Xu and Wang, 2007) and many more.

Programs for detecting non-LTR retrotransposons are e.g. TSDFINDER (Szak *et al.*, 2002), SINEDR (Tu *et al.*, 2004) and RTANALYZER (Lucier *et al.*, 2007). They search for the presence of a polyA tail in 3’ end of the sequence, TSDs, PBSs, PPTs, and ORFs.

The signature-based methods are also used for identification of DNA transposons. FINDMITE (Tu, 2001) and MITE-Hunter (Han and Wessler, 2010) were developed for identification of MITEs, which lack a transposase gene and thus similarity-based methods are not applicable for their identification. HelitronFinder (Du *et al.*, 2008) and HelitronScanner (Xiong *et al.*, 2014) facilitate the identification of Helitrons which lack the typical structural features of other DNA transposons and for long time were challenging to identify.

De novo methods

De novo methods work without using prior information about repeat structure or similarity to known repeat sequences. These methods were invented mainly to discover new repeats in newly sequenced genomes without any knowledge about their repeat content.

These methods are based on two strategies: (i) self-comparison (self-alignment) approach and (ii) k-mer frequency approach (Lerat, 2010).

The self-comparison approach employs self genome comparison to detect similar sequences and then to cluster these sequences to get repeat families for which consensus sequences are then built (Bergman and Quesneville, 2007). These methods are e.g. implemented in software RECON (Bao and Eddy, 2002) and PILER (Edgar and Myers, 2005).

Algorithms using k-mer frequencies count the occurrence of short identical motives (k-mers) presented multiple times in a genome and use seed extension strategy. They can work with an entire genome or unassembled reads. There are many softwares using these approaches, e.g. RePuter (Kurtz and Schleiermacher, 1999), ReAS (Li *et al.*, 2005), RepeatScout (Prince *et al.*, 2005), Tallymer (Kurtz *et al.*, 2008), RepARK (Koch *et al.*, 2014), Tedna (Zytnicki *et al.*, 2014), MixTaR (Fertin *et al.*, 2015) , REPdenovo (Chu *et al.*, 2016), and DeviaTE (Weilguny and Kofler, 2019). New software RepLong (Guo *et al.*, 2018) is the first *de novo* repeat identification method using long PacBio reads.

RepeatModeler (http://www.repeatmasker.org/RepeatModeler/; Smit and Hubley, 2008-2015) combines two *de novo* repeat finding programs, RECON (Bao and Eddy, 2002) and RepeatScout (Price *et al.*, 2005), that employ complementary computational methods for identifying repeat boundaries and family relationships from sequence data.

RepeatExplorer (Novák *et al.*, 2013) is a graph-based read clustering method publicly available on Galaxy server (https://galaxy-elixir.cerit-sc.cz/) using unassembled reads as input. This tool analyzes similarities between reads and builds graphs that represent repeat families. The use of this method is described in detail in the Methods section of this thesis.

De novo identification of tandem repeats can be carried out by program Tandem Repeat Finder (TRF, Benson, 1999) without a need to a priori specify either the sequence or length or copy number of tandem repeats. The algorithm is based on the detection of k-tuple matches (clusters of small matching words) separated by a common distance (O'Dushlaine and Shields, 2006). Detection of TRs is done by probabilistic models, combinatorial or heuristic approaches (Benson, 1999).

TAREAN (Novák *et al.*, 2017) is a novel computational pipeline that detects satellite repeats directly from unassembled short reads. The pipeline employs graph-based sequence clustering method, as in RepeatExplorer (Novák et al., 2013), to identify groups of reads that represent the repetitive elements. Putative satellite repeats are afterwards detected by the presence of peculiar circular structures in their cluster graphs. Consensus sequences of tandem repeat monomers are reconstructed from the most frequent k-mers in corresponding clusters (Novák *et al.*, 2017). RepeatExplorer and TAREAN pipelines provide information on both structure and abundance of repetitive sequences.

Repeat classification programs

A classification information about taxonomic position is needed after new repeat family discovery. The classification system of repetitive DNA is still developing (Wicker *et al.*, 2007, Lorens *et al.*, 2011, Neumann *et al.*, 2019). Unknown and uncharacterized TEs can

be automatically classified by programs TEclass (Abrusán *et al.*, 2009) or REPCLASS (Feschotte *et al.*, 2009) that reconstruct also the consensus sequences of repeats. However, some of newly identified repeats remains unclassified due to their incompleteness or they have not been described yet.

1.1.2 CHLOROPLAST GENOME

Chloroplasts (also known as plastomes) are cytoplasmic semi-autonomous organelles that originated through endosymbiosis from a cyanobacterium, and contain their own genomes, chloroplast DNA (cpDNA). They occur in high copy numbers in plant cells and are responsible for the process of photosynthesis.

Length of most land plant cpDNA typically ranges between 120 and 160 kb (Palmer, 1985; Green, 2011), mainly due to duplication of genes, small repeats and the size of intergenic spacers. Greater genome size range has been found in green algae - from 85 to 292 kb, and up to 2 Mb in *Acetabularia acetabulum* which is believed to be the largest cpDNA (Palmer, 1985; Tymms and Shweiger, 1985) but the genome has yet to be completely sequenced (de Vries *et al.*, 2013). Chloroplast genome contains 100-120 genes, which are primarily involved in photosynthesis, transcription, and translation. Usually there are four copies of rRNA genes, a number of tRNA genes, at least three subunits of prokaryotic RNA polymerases (**Figure 5**; Palmer, 1985; Green 2011).

The chloroplast genomes of most angiosperms seem to exhibit a conserved structure with four major segments (**Figure 5**): large single copy (LSC) and small single copy (SSC) regions separated by two inverted repeat (IR; IRa and IRb) regions (Palmer, 1985).

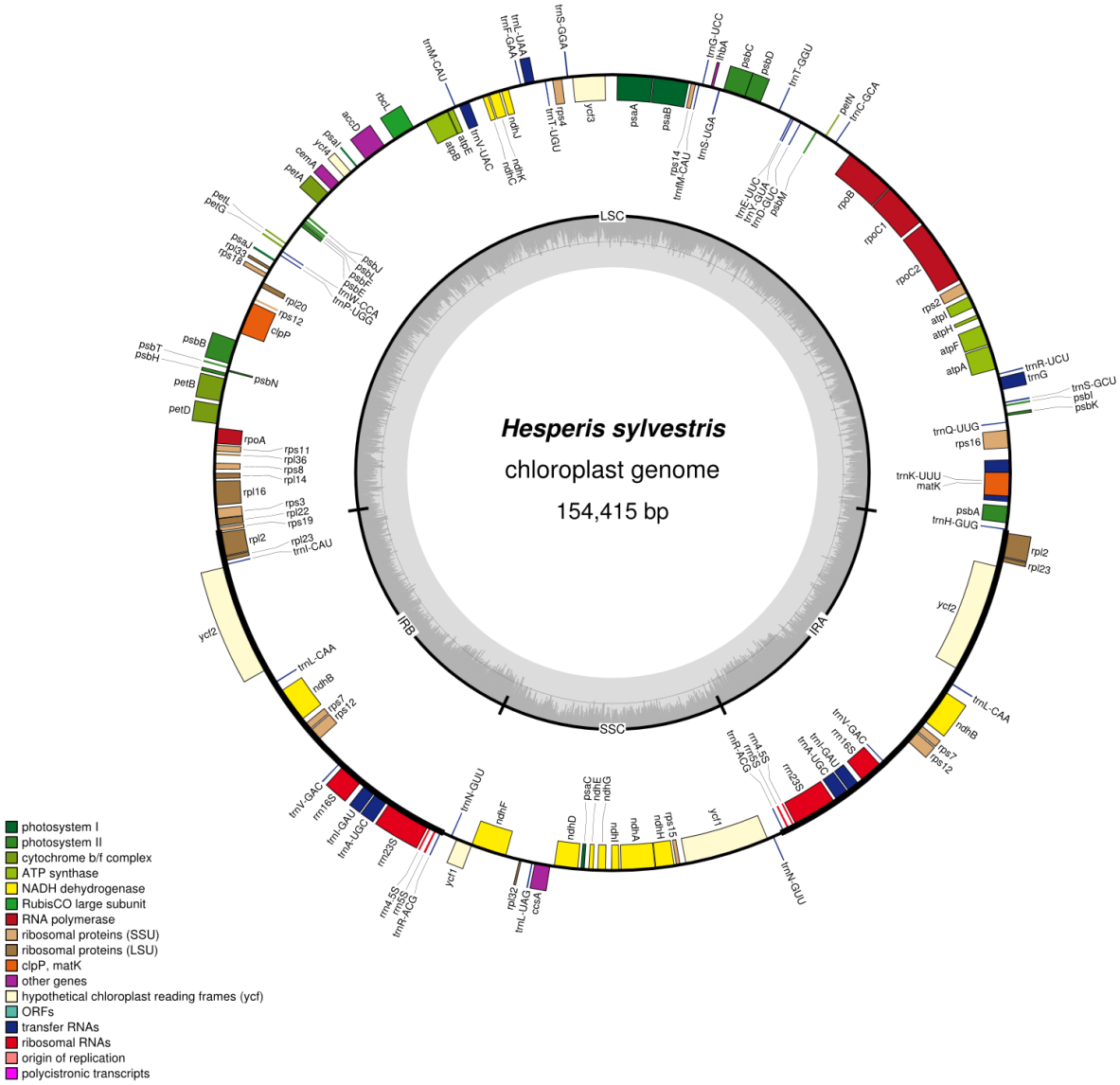


Figure 5 Chloroplast genome of *Hesperis sylvestris* (NC_035512.1; Mandáková *et al.*, 2017). The annotation of genome was done by GeSeq tool (Tillich *et al.*, 2017) and visualization by OGDRAW v1.3.1 (Greiner *et al.*, 2019). Genes inside the circle are transcribed clockwise, genes outside the circle counter clockwise.

IRs vary in length (usually ranging from 20,000 to 30,000 bp in angiosperms) and are very similar but not completely identical. The IRs are highly conserved among land plants but loss or partial loss of genes (e.g. genes *ycf1* and *rps16*) can occur during evolution (Xu *et al.*, 2015). In some cases these lost genes have been transferred to the nucleus (e.g. Wakasugi *et al.*, 1994; Timmis *et al.*, 2004; Sheppard *et al.*, 2008) or to mitochondria

(Smith, 2011). There is also well documented evidence of a mitochondrial DNA to chloroplast genome transfer (Iorizzo *et al.*, 2012, Straub *et al.*, 2013; Ma *et al.*, 2015).

For a long time it has been thought that chloroplast genome is a strictly circular DNA molecule but it has been shown that the majority of cpDNA is arranged in concatemers of two or more molecules in either circularized or linear form (e.g. Lilly *et al.*, 2001; Bendich, 2004).

In most angiosperms cpDNAs are usually inherited from a single parent. Paternal inheritance is mostly observed in gymnosperms, while flowering plants often inherit chloroplasts maternally. Biparental (cpDNA inherited from both parents) chloroplast inheritance occurs in some cases, e.g. in alfalfa (Lee *et al.*, 1988) or *Passiflora* (Hansen *et al.*, 2007).

The number of sequenced and known chloroplast genome sequences is growing rapidly with advances in NGS. Low coverage genome sequencing data contain beside nuclear DNA also chloroplast DNA if DNA is extracted from whole tissues. Thus, NGS data are useful for phylogenetic studies based on selected cpDNA markers or the complete assembly of chloroplast genomes.

The first complete chloroplast genomes were reported for tobacco, *Nicotiana tabacum* (Shinozaki *et al.*, 1986) and a liverwort, *Marchantia polymorpha* (Ohyama *et al.*, 1986). The complete structure of the chloroplast genome of the model plant *Arabidopsis thaliana* has been determined already 20 years ago (Sato *et al.*, 1999). The 154,478-bp genome was found to possess the typical quadripartite structure of angiosperms, consisting of two IRs (26,264 bp) separated by a LSC region (84,170 bp) and SSC region (17,780 bp), and 87 potential protein-coding genes (eight genes duplicated in the IR regions), four ribosomal RNA genes and 37 tRNA genes (Sato *et al.*, 1999).

So far (January 2020), there are more than 3,000 complete chloroplast genomes of land plants available in the National Center for Biotechnology Information (NCBI) organelle genome database (https://www.ncbi.nlm.nih.gov/genome/organelle/). According to this database, the smallest assembled chloroplast genome is that of a parasite angiosperm plant *Cytinus hypocistis* (Cytinaceae). The genome is highly reduced (19,400

bp; NC_031150), with only 23 genes and no inverted repeat regions (Roquet *et al.*, 2016). The largest assembled chloroplast genomes were identified in *Pelargonium* species, *P. transvaalense* (242,575 bp; NC_031206; Geraniaceae) and *P.* × *hortorum* (217,942 bp; NC_008454), with greatly expanded IRs (75,741 bp each; Chumley *et al.*, 2006).

1.1.2.1 Chloroplast-based phylogeny

Generally, the phylogenetic tree represents hypotheses about the history of evolutionary relationships among a given group of species. Different methods can be employed to generate phylogenetic trees: (i) distance-based methods, (ii) maximum parsimony, (iii) maximum likelihood, and (iv) Bayesian inference.

Distance-based methods seek to reconstruct the tree topology based on the matrix of distances between pairs of taxonomic units (Brocchieri 2001), i.e., between each pair of sequences; the more similar ones should be evolutionary more related. Commonly used distance-based methods are Neighbour Joining and UPGMA. Maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference methods, together called character-based methods, search for the most probable tree(s) for a specific sequence set, based on characters at each position of the sequence alignment (Sleator, 2013).

Due to the uniparental inheritance in most species and absence of recombination, the cpDNA is phylogenetically linear over generations, changing only by occasional mutations (substitution rates are about 10 times lower than those of nuclear genomes; Xu *et al.*, 2015), and thus, cpDNA sequences have been used extensively for inferring relationships among plants at all taxonomic levels.

Frequently sequenced cpDNA regions for resolving relationships at family or generic level were the *rbc*L gene (gene for large subunit of Rubisco; e.g. Savolainen *et al.*, 2000; Soltis *et al.*, 2000), *atp*B (gene coding beta subunit of ATP synthase; Soltis *et al.*, 2000), *ndh*F (coding a subunit of chloroplast NADH-dehydrogenase; Clark *et al.*, 1995; Alverson *et al.*, 1999), *mat*K (gene coding maturase; Wojciechowski *et al.*, 2004), and the *trnL-trnF* region. The *trnL-trnF* region is used for resolving relationship among closely related

species as accumulation of indels were observed in this region of cpDNA (other non-coding cpDNA sequences used for phylogenetic studies were reviewed in Shaw *et al.*, 2005; 2007).

Important phylogenetic trees have been resolved using chloroplast genome sequences. Jansen *et al.* (2007) used 81 plastid genes from 64 sequenced genomes to estimate relationships among the major angiosperm clades. Moore *et al.* (2007) performed phylogenetic analyses of 61 plastid genes for 45 taxa to resolve basal angiosperm relationships. Early diversification of eudicots has been resolved by Moore *et al.* (2010) by using 83 protein-coding and rRNA genes from the plastid genome of 86 species. Hohmann *et al.* (2015) resolved the phylogenetic relationships among taxa from the whole Superrosidae clade based on a set of 73 genes, including 51 protein-coding genes, 19 tRNAs, and three ribosomal RNA genes.

1.2 SEQUENCING TECHNOLOGIES

Since 1953, when the structure of double stranded DNA has been discovered and described by Watson and Crick (Watson and Crick , 1953), deciphering of the primary structure of DNA has become one of the major focuses of molecular biology. DNA sequencing is the process which allows us to determine the order of nucleotides in DNA primary structure.

Current sequencing technologies are generally classified into generations. These methods differ in basic principles, read length and number of reads. In the late 1970s, a method of chemical sequencing was invented by Maxam and Gilbert (1977). This was the first widely adopted method of DNA sequencing. In the same year, Sanger and his colleagues (Sanger *et al.*, 1977) developed a new method known as an enzymatic sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication. These two methods are now referred as the first generation of DNA sequencing and Sanger's

sequencing method was the first commercially available automated sequencing. It produces relatively long (500–1000 bp) high quality DNA sequences.

Improvements in biotechnology and computer technology led to development of next-generation sequencing methods. The methods of the second generation (454, Illumina, Ion Torrent and SOLiD) have enabled producing more sequencing in parallel at low cost. There was no longer needed to clone DNA molecules in a microbial host (*in vivo*) as the methods instead involve ligation of DNA fragments to special adapters for clonal amplification *in vitro*. The analyzed DNA molecules must be fragmented prior to sequencing, these fragments (reads) can be read from one end (single-end reads) or both ends (paired-end or mate-pair reads).

In recent years, third-generation technologies emerged (PacBio, Oxford Nanopore) that aim to sequence individual molecules and to produce significantly longer reads at least of several thousand base pairs in length.

1.2.1 SECOND GENERATION

The DNA sequencing technologies of the second generation, or massively parallel sequencing, are capable of processing millions or billions of DNA reads in parallel at high speed.

In 2005 (Margulies *et al.*, 2005), Roche 454 sequencing (also known as pyrosequencing) appeared. It is a sequencing by synthesis technology that is based on the detection of released pyrophosphate during DNA synthesis (Ronaghi, 2001). Sequencing by ligation was introduced by SOLiD (sequencing by oligonucleotide ligation and detection). Non-optical method (Rothberg *et al.*, 2011), which does not use fluorescently labelled nucleotides, was developed by IonTorrent and was based on detection of the release of H+ during a polymerization reaction through a solid-state sensor (Levy and Myers, 2016).

The most used NGS technology is Illumina sequencing based on detection of fluorescently modified nucleotides. The Illumina paired-end DNA sequencing process

consists of four steps (**Figure 6**), preceded by DNA extraction (Goodwin *et al.*, 2016; https://www.illumina.com):

a) DNA library preparation.

A DNA library is a processed sample material that serves as the input material for NGS applications. The preparation of a DNA library includes random fragmentation (by shearing, sonication, or nebulization) of genomic DNA to obtain fragments of required length and attaching oligonucleotide sequencing adapters to both ends of the DNA fragments.

b) Cluster amplification

The prepared library is loaded into a flow cell and DNA fragments are bound to primers on the flow cell surface. Unlabelled nucleotide bases and DNA polymerase are added to initiate the bridge amplification using primers complementary to the sequencing adapters.

c) Sequencing by synthesis and image analysis.

Primers, DNA polymerase, and fluorescently tagged nucleotides are added to the flow cell. Every nucleotide has a reversible terminator that stops DNA synthesis to prevent multiple additions in one round. Only one base is added per one round. Each of the four nucleotides has a unique label that can be excited by lasers to emit a characteristic wavelength, and after each round, fluorescence is detected by a camera, and a computer records which base was incorporated (base calling).

d) Post-sequencing data processing and analysis.

In the last step, raw sequencing reads are obtained in form of text file formats (usually fastq format). This data is then quality checked and pre-processed (e.g. filtering by quality, adapter trimming) before bioinformatic analyses.

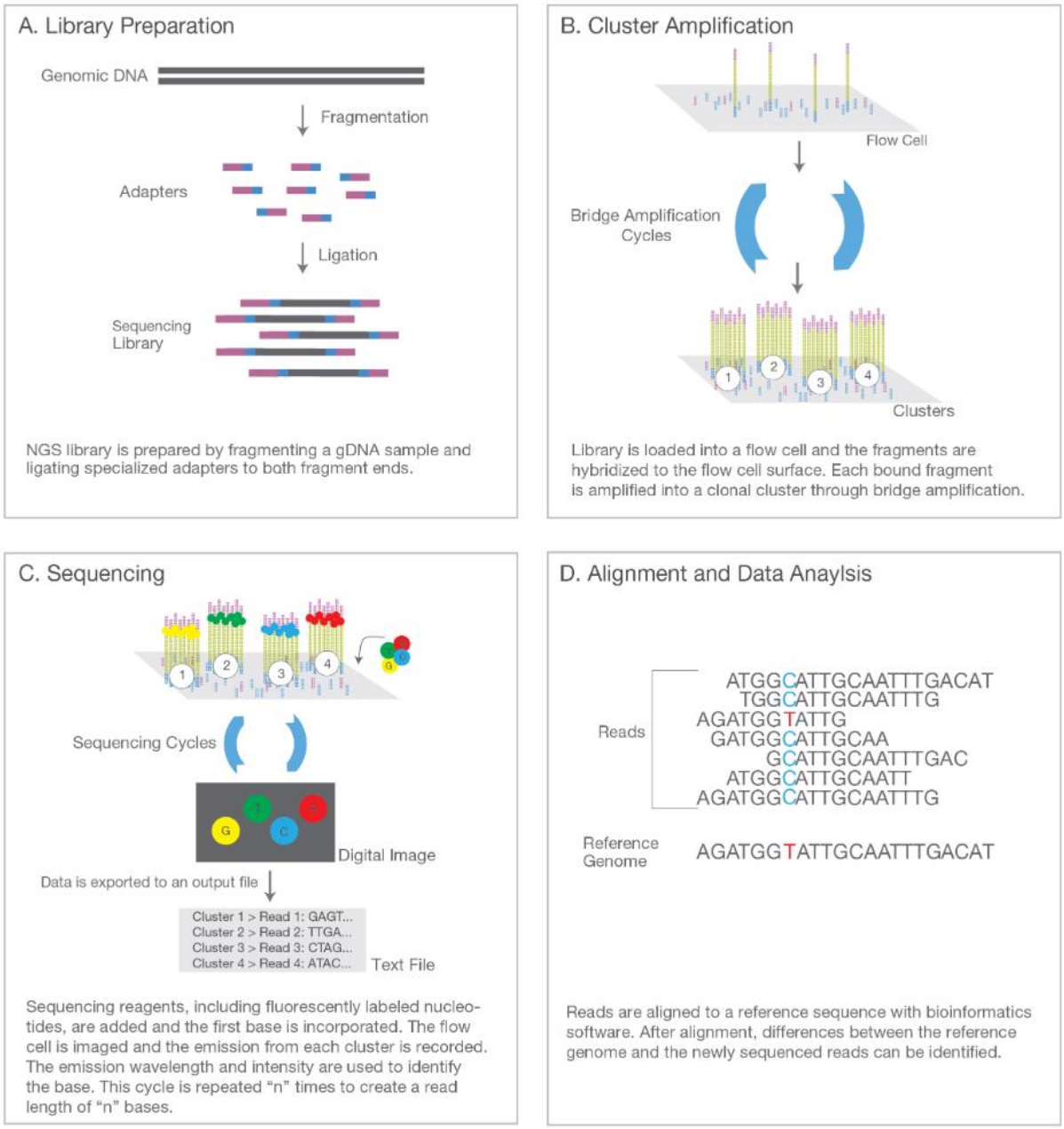


Figure 6 Overview of Illumina sequencing technology. (Adopted from https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf).

1.2.2 THIRD GENERATION

Third generation methods are currently under active development. Nucleotide sequences are sequenced at the single molecule level, in contrast to existing methods that require breaking long strands of DNA into small fragments.

The available third generation DNA sequencing technologies are Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing, and the Oxford Nanopore Technologies sequencing platform. Genome assemblies of *A. thaliana* (Michael *et al.*, 2018) and maize (Jiao *et al.*, 2017) have already been improved by single-molecule sequencing technologies.

SMRT sequencing overcomes the short length limitations of the second generation sequencing technologies, without the need for PCR amplification. The template (SMRTbell template; Travers *et al.*, 2010) is a double-stranded DNA molecule flanked by two ligated hairpin adapters which create closed circle, enabling a repeated sequencing of artificially created circular DNA (Travers *et al.*, 2010). The basis of the sequencing method is an uninterrupted template-directed synthesis performed by DNA polymerase using four distinguishable fluorescently labelled nucleotides. The simultaneous detection of the growing DNA strand is done by nanopore zero-mode waveguides, which enable single-fluorophore detection (Eid *et al.*, 2009). SMRT sequencing produces read lengths up to 100,000 bp but with an error rate of 10% to 15% (Lee *et al.*, 2016). To correct for the error rate, long reads are often combined with short accurate PacBio or Illumina reads.

The Oxford Nanopore sequencing device MinION is the smallest sequencing device currently available. It has the size of a USB flash drive and it is powered by USB so it can be plugged directly to a computer (Lu *et al.*, 2016). The device passes an ionic current through protein nanopores and measures the changes in current to identify the sequenced molecule as the bases go through the nanopore channels or near them in flow cell in different combinations (Lu *et al.*, 2016; https://nanoporetech.com/how-it-works#).

However, short reads of the second generation sequencing platforms are still less erroneous (0.01-1%) than long reads of new third generation sequencing platforms (4-15%).

1.2.3 LOW-PASS SEQUENCING

Method of genome skimming (Straub *et al.*, 2012) using low-coverage (low-pass) shotgun genome sequencing approach allows to retrieve sequences presented in genome in high abundances, such as mentioned repetitive DNA, rDNA, and organelle genomes (cpDNA, mtDNA) and high-copy genes.

1.2.4 OUTPUT OF SEQUENCING

The standard format for storing NGS data from sequencing platforms is fastq file (Cock *et al.*, 2009). Fastq is a text-based sequencing data file format that stores both raw sequence data and quality values encoded in ASCII code. A fastq file typically consists of four lines: (i) a line starting with @ and containing the sequence identifier, (ii) the actual sequence, (iii) a separator, a line starting with +, which can be followed by an optional sequence identifier, and (iv) a line with quality scores encoded as an ASCII character (e.g. the quality score for encoding Illumina 1.8+ is represented as the character with an ASCII code equal to its value + 33; https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html).

This format can be further directly used for data analysis or can be converted to fasta format which is also text-based sequencing data file format that does not include information of data quality.

A quality score assigned to each base of read is called Phred score (Ewing and Green, 1998) and it determines the probability of error for that base. The Phred quality score (Q) is defined as the negative logarithm with base 10 of the base-calling error probability (P):

$$Q = -10 \log_{10} P$$

That means if the Phred score is 30 then the chance that this base is called incorrectly is 1 in 1000 (base call accuracy 99.9%). Information about quality encoding is important for determination of quality cut-off, quality threshold for quality filtering. Phred scores can be used to compare different sequencing methods efficiency.

1.2.5 QUALITY CONTROL

The crucial part of bioinformatic analyses based on NGS data is the preprocessing of NGS raw reads. The preprocessing comprises a series of steps that involve low-quality read filtering, trimming of adapters, and filtering of unwanted sequences.

Before analyzing NGS data, initial quality control check of data should be always performed, to ensure that the raw data are fulfilling expected quality and quantity (e.g. read length and number of reads) and that there are no problems or biases in data which could affect further analyses. A quality control check of NGS data is usually made by FastQC tool (Andrews, 2010).

FastQC is a java application designed to provide an overview of basic quality control metrics for raw NGS data, to spot and highlight potential problems in datasets. It can be run from command line or as a GUI (graphical user interface). It runs a set of analyses (modules) on raw sequence files in fastq format and produces a report which summarizes the results. The output is an html file that can be viewed in web browser. It includes basic information about input file and generates graphs showing overall data quality statistics.

A warning is raised if some of the FastQC statistics fails. If the raw data fail the quality check it is usually related to poor sequencing quality or base calling errors, small insertions/deletions, presence of adapter/primer sequences in data or to a library contamination. In contrast to classical Sanger sequencing, associated error rates of NGS platforms are about 0.1% to 15% (depending on a method and read lengths).

The per base sequence quality, which measures quality score statistics at each position along all reads in the file, is probably the most important information on how reads are good or bad overall. It provides an overview of the overall quality and it helps to decide how many bases to trim from the 5'- or 3'-end, and to set a quality trimming threshold (Andrews, 2010; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/; https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/).

For Illumina reads, usually the base quality decreases towards the end of the read (error rates increase). This phenomenon is caused by the sequencing by synthesis process. Bases are added one at a time and the consensus is determined in a cluster of identical sequences but if molecule fails to elongate properly or advances too fast, not all sequences in a cluster will then grow at the same rate (phasing error; Dohm *et al.*, 2008). The cluster signal can fade with increasing read length and it can lead to the error accumulation, resulting in higher error rates towards the end of the read (Schirmer *et al.*, 2015). Differences in per base sequence quality of 100- and 300-bp long Illumina reads are shown in **Figure 7**.

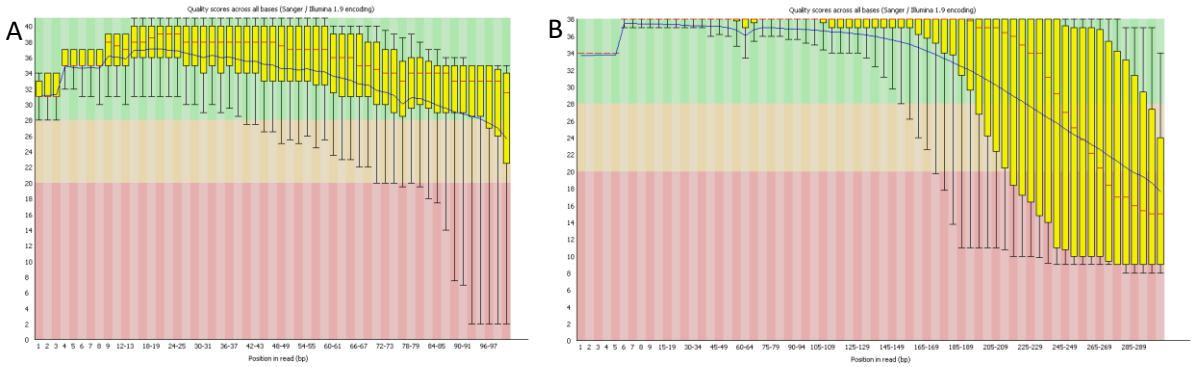


Figure 7 Differences in per base sequence quality of (A) HiSeq 100-bp *Hesperis matronalis* reads, and (B) MiSeq 300-bp long reads of *Hesperis sylvestris*.

Another useful FastQC output plot is the per base sequence content which indicates proportion of each base position in a file and can reveal if primers or adapters are still present in reads based on the non-random nucleotide content (**Figure 8A and 8B**). The over-represented sequences table informs whether particular trimming is needed. Reads with presence of adapter/primer sequences are either trimmed or completely removed from the analyzed dataset.

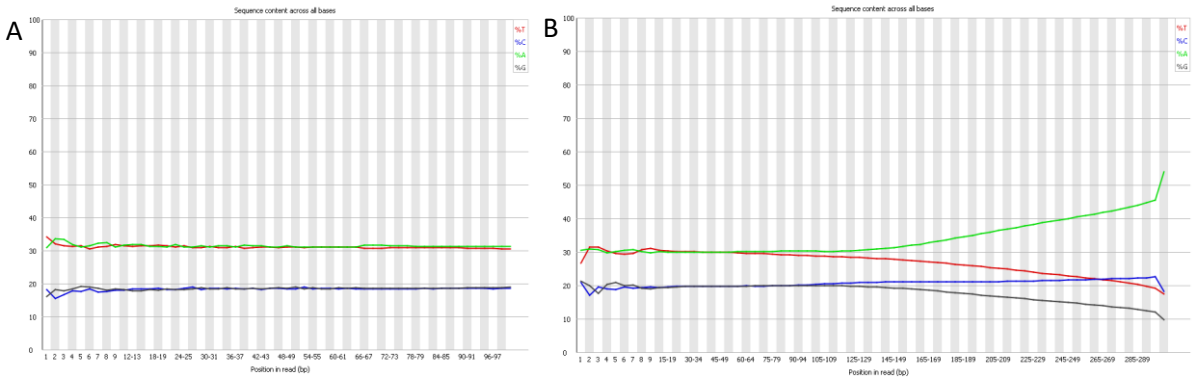


Figure 8 Per base sequence content plot obtained by FastQC program. Non-random nucleotide content displayed as zigzag structure at the beginning of reads is observed in both datasets (A – *Anastatica hierochuntica* reads, B – *Aubrieta canescens* reads) indicating presence of primers or adapters. The non-random nucleotide content towards the end of *A. canescens* reads in (B) indicates over-represented adapter sequences.

Per sequence GC content measures the GC content of each base position in a file and compares it to a modelled normal distribution of GC content (**Figure 9A**). Central peak corresponds to the overall GC content mean of the underlying genome. It can indicate library contamination if unusually shaped distribution (e.g. two peaks) is observed (**Figure 9B**). When known adapter or primer sequences are detected, a simple removal or trimming of these contaminating sequences can make a huge difference in the quality of reads. The FastQC is rerun afterwards to see whether that resolves the problem. If the GC still does not follow the normal distribution, library contamination is indicated. Often sequencing reads may contain the Phi X 174 bacteriophage reads, as Phi X is regularly used as a positive control in DNA sequencing (https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html).

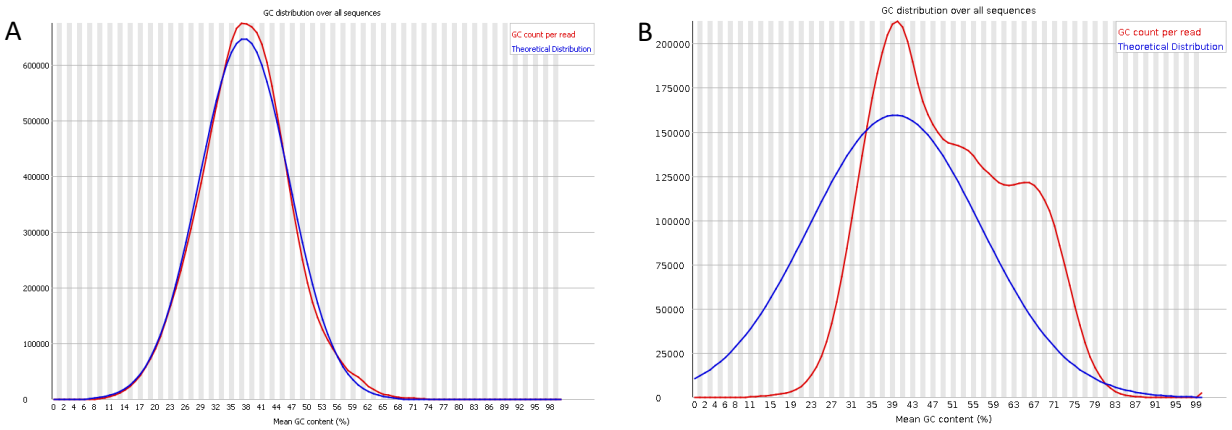


Figure 9 Per sequence GC content. Normal distribution of GC content throughout (A) *Hesperis matronalis* reads and biased GC content (B) in *Azolla filiculoides* reads caused by the presence of contamination by cyanobacterial symbiont *Nostoc azollae* reads are shown (unpublished data of P. Hloušková).

After the raw reads are adapter and quality trimmed and filtered, paired-end reads can be interlaced into a single fastq file; complete pairs are kept, single reads are discarded. Pre-processed reads can be then converted from fastq to fasta file format. Commonly used softwares for NGS data preprocessing are FASTX-toolkit (Gordon and Hannon, 2010), cutadapt (Martin, 2011), and Trimmomatic (Bolger *et al.*, 2014).

1.3 ASSEMBLY

Genome assembly is a process of putative reconstructing an original genome sequence (target) from thousands of small reads produced by sequencing instruments. The whole genome sequence (WGS) assembly is defined as a reconstruction of the target sequence up to the chromosome level. The assembly is possible when the target is oversampled in a way that reads overlap (Miller *et al.*, 2010) and contigs (contiguous sets of overlapping DNA fragments) can be built. As contig length is a function of genome coverage and read length, short reads require a much higher coverage to reach the same expected contig length (Schatz *et al.*, 2010). Desai *et al.* (2013) have showed that 35-50× read depth of the second generation sequencing techniques is sufficient for a *de novo* assembly of small genomes such as bacteria and yeast. The recommended coverage for a *de novo* assembly of plant genome is 60-80× and in some cases even 100× (Schatz *et*

al., 2010; Li and Harkess, 2018) using Illumina short reads and then additional scaffolding approach is needed to achieve a reasonably good assembly. A 40× coverage is recommended for a PacBio-only assembly or a small-to-moderate-sized genome (<1 Gbp; Li and Harkess, 2018). Reference-based assemblies, or resequencing projects, can be done at a lower coverage. Assembling of any genome requires the proper combination of coverage, read length and good quality reads with low error rates (Schatz *et al.*, 2010). An assembly quality depends also on repeat composition as high repeat contents are challenging. Long reads and paired-end reads help to resolve repeated regions and improve the genome assembly (Berglund *et al.*, 2011).

There are two different types of genome assembly, a reference-based and a *de novo* assembly. The reference-based method consists of mapping (aligning) sequencing reads to a known reference genome of the same (resequencing project) or closely related species, and reconstructing a consensus based on that genome. The aim is to align each read to the reference genome, allowing mismatches, indels and clipping of some short fragments on the two ends of the reads (Sung, 2017). Popular mapping programs are BWA (Li and Durbin, 2009), Bowtie and Bowtie2 (Langmead *et al.*, 2009; Langmead and Salzberg, 2012), whereas a *de novo* assembly method reconstructs the template sequence without the guidance of a reference genome. This approach is used for sequencing of non-model organisms and it is computationally much more challenging and demanding than the reference-based approach (Berglund *et al.*, 2011). There are two standard *de novo* assembly approaches, based either on Overlap-Layout-Consensus (OLC) or de Bruijn graph (DBG).

The OLC assemblers use an overlap graph which is built by finding the overlaps between each pair of sequencing reads by all-against-all pair-wise read comparison (Miller *et al.*, 2010). Multiple sequence alignment determines the precise layout and then the consensus sequence is reconstructed (Miller *et al.*, 2010). Many assembler programs adopted this algorithm, such as ARACHNE (Batzoglou *et al.*, 2002), Newbler (designed specifically for assembling sequence data generated by the 454; Margulies *et al.*, 2005), Celera Assembler (Myers *et al.*, 2000), Euler (Pevzner *et al.*, 2001), or CAP3 (Huang and

Madan, 1999), which is used for read assembly within the repeat clusters in RepeatExplorer pipeline (Novák *et al.*, 2013).

DBG is an anti-intuition algorithm as it first chops reads into much shorter k-mers and then uses all the k-mers to form a DBG and finally infers the genome sequence based on the DBG (Li *et al.*, 2012). The k-mer graph does not require all-against-all overlap discovery as the OLC method and it compresses redundant sequences (Miller *et al.*, 2010). Many programs employing DBG algorithm have been developed, such as Velvet (Zerbino *et al.*, 2008), ABySS (Simpson *et al.*, 2009), SOAPdenovo (Li *et al.*, 2010), Ray (Boisvert *et al.*, 2010), and AllPath-LG (Gnerre *et al.*, 2011).

1.3.1 METRICS FOR MEASURING ASSEMBLY OUTPUT

Assemblies are measured by the size and number of contigs and scaffolds. Assembly size is usually given by statistics including maximum length, average length, combined total length, N50, and L50 (Miller *et al.*, 2010). Contig or scaffold N50 is a weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value. L50 is the rank of the contig that corresponds to the N50 length, thus N50 describes a sequence length, whereas L50 describes a number of sequences.

The tool QUAST (Gurevich *et al.*, 2013) was developed for evaluating and comparing different assemblers or different assembler's settings (e.g. different k-mers values) and it helps to choose the best pipeline for a given dataset.

1.3.2 ASSEMBLY OF CHLOROPLAST GENOME

Whole genome shotgun sequences and even low-coverage shotgun genome sequences of plant genomes often contain an abundance of reads that originate from the chloroplast genome. Assembling of chloroplast genomes is still challenging because of the inverted repeat regions. It is usually required and recommended to obtain at least 20× coverage of cpDNA. Using a combination of long and short reads, hybrid assemblies

seem to be more accurate than long-read-only or short-read-only assemblies (Wang *et al.*, 2018).

Using the reference-based approach, reads are identified and assembled into chloroplast genomes based on homology to chloroplasts from related species with already known sequence. Disadvantages of this approach is that it ignores structural differences between the genomes and it is not suitable for non-model species without an available reference genome. An alternative and preferred approach is a *de novo* assembly from total genomic DNA sequences or filtered chloroplast sequences. There are newly developed programs specifically designed for *de novo* chloroplast assembly, e.g. Novoplasty (Dierckxsens *et al.*, 2016) or pipeline GetOrganelle (Jin *et al.*, 2018).

MITObim v1.7 (Hahn *et al.* 2013) was designed for a mitochondrial genome assembly but was successfully employed to assemble whole chloroplast genomes (e.g. Silva *et al.*, 2017; Xie *et al.*, 2017)

1.3.2.1 Annotation of the chloroplast genome

The next step is the annotation of assembled chloroplast genome. For the last 15 years, software DOGMA (Dual Organellar GenoMe Annotator; Wyman *et al.*, 2004) was often used for automated annotation of chloroplast genomes. It was replaced by newly developed and more user-friendly packages, e.g. GeSeq (Tillich *et al.*, 2017), CpGAVAS (Liu *et al.* 2012), CpGAVAS2 (Shi *et al.*, 2019), Plann (Huang and Cronk, 2015) and PGA (Qu *et al.*, 2019), which allow and facilitate visualization and preparation of GenBank-formatted reference of assembled genomes. The tRNA genes can be detected and predicted using ARAGORN (Laslett and Canback, 2004) or tRNAscan-SE (Lowe and Chan, 2016).

1.3.3 SEQUENCING AND ASSEMBLY OF PLANT GENOMES

The first complete plant genome assembly was that of the *Arabidopsis thaliana*. It was finished in 2000 (Arabidopsis Genome Initiative, 2000) and took long ten years to finish.

This was only the third published genome assembly of a multicellular eukaryotic genome (after *Caenorhabditis elegans* and *Drosophila melanogaster*) but the first plant genome assembly.

Since then, many other plant genomes have been sequenced and assembled. To integrate genomic data, number of plant genome databases have been established. Phytozome (Goodstein *et al.*, 2012; http://www.phytozome.net) is one of a such database which collects genome and gene family data of plants which were sequenced, assembled and annotated at the Joint Genome Institute (https://jgi.doe.gov/), including major crop species (rice, maize, wheat) and many other plant.

But not all the plant genomes are easy or even feasible to assemble due to their genome size, repeat structure, and heterozygosity (Li and Harkess, 2018). Hybrid assemblies, combination of long third generation reads and shorter less erroneous reads, are currently the way how to make sequencing projects feasible.

2 AIMS AND OBJECTIVES

The goal of this thesis is to complement cytogenomic laboratory methods with bioinformatics based on NGS data, to investigate structure and evolution of plant genomes in the mustard family (Brassicaceae).

Main objectives of the NGS data analysis in the frame of the thesis:

1. to mine phylogenetically informative chloroplast markers for deciphering phylogenetic relationships among plant species (phylogenetic analysis),
2. to better understand plant genome composition through *in silico* identification of repetitive sequences, and thus, to elucidate sources of genome size variation in plants,
3. to *de novo* identify centromere-associated tandem repeats that can be further used as DNA probes and markers in cytogenetic studies,
4. to design oligoprobes for identified repeat sequences, to describe the genomic distribution of different repetitive sequences on chromosomes.

This thesis is based on the following three publications. The motivation for each publication as well as the bioinformatics aims (my contribution to the study) are summarized as follows:

Publication 1

Mandáková T, **Hloušková P**, German DA, Lysak MA. (2017). Monophyletic origin and evolution of the largest crucifer genomes. *Plant Physiology*. 174(4), 2062-2071.

Motivation: The *Hesperis* clade (a.k.a. lineage III or clade E), is one of the major Brassicaceae clades. The clade is noteworthy as it harbors species with the largest nuclear genomes among the crucifers but relatively low chromosome numbers (n = 5–7). By applying comparative cytogenetic analysis and whole-chloroplast phylogenetics,

we wanted to construct cytogenomic maps for selected representatives of clade E tribes, reconstruct their putative ancestral genome, date their diversification, and investigated their relationships in a family-wide context.

Bioinformatics aims:

1. To assemble the chloroplast genome from Illumina 100-bp, 156-bp, and 300-bp paired-end reads in selected species.
2. To annotate the assembled chloroplast genomes.
3. To infer phylogenetic relationships based on whole-chloroplast sequences and to employ methods of molecular dating to estimate divergence times within the *Hesperis* clade and across the family Brassicaceae.
4. To confirm that the tribe Anastaticeae, previously classified as a member of the *Hesperis* clade, does not belong to this lineage.

Publication 2

Hloušková P, Mandáková T, Pouch M, Trávnícek P, Lysak MA. (2019). The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing. *Annals of Botany*. 124(1), 103–120.

Motivation: Most Brassicaceae species have small nuclear genomes but the *Hesperis* clade is known for its large genome size variation, ranging from relatively small genomes (*Euclidium syriacum*, 254 Mb) to large genomes (*Hesperis sylvestris*, 4,264 Mb). We aimed to identify, quantify and localize *in situ* repetitive sequences from which these genomes are built and so to uncover sequences which are responsible for genome size variation within the *Hesperis* clade. We analyzed Illumina low-coverage genome sequencing data in seven diploid species, covering the phylogenetic and genome size breadth of the clade.

Bioinformatics aims:

1. To identify and estimate the abundance of different repeat families in selected species representing six out of the seven *Hesperis* clade tribes.

2. To uncover whether the genome size variation is caused by the accumulation of either a single/a few repeat types or many different repeat families.
3. To discover whether the genome size variation correlates with the repeat content and life forms.
4. To reconstruct the ancestral genome size of the *Hesperis* clade and the evolution of the genome size in the clade.
5. To design probes specific to the most abundant repetitive sequences further used for fluorescent *in situ* localization of the investigated repeat families on chromosomes.

Publication 3

Mandáková T, **Hloušková P**, Koch MA, Lysak MA. (2020). Genome evolution in Arabideae was marked by frequent centromere repositioning. *Plant Cell*. Doi: 10.1105/tpc.19.00557

Motivation: Centromere repositioning and evolutionary new centromeres (ENCs) were frequently encountered during vertebrate genome evolution, but only rarely observed in plants. We aimed to analyze genome structure of 10 species from the tribe Arabideae (Brassicaceae) by BAC-based comparative chromosome painting to uncover frequency of ENCs. As Arabideae genomes show a remarkable stasis of chromosome numbers and genome structure we intended to prove that centromere repositioning is a primary process underlying structural differentiation of Arabideae chromosomes and whole genomes.

Bioinformatics aims:

1. To perform repeatome analysis of selected Arabideae species.
2. To identify centromere-associated tandem repeats which can be further used as cytogenetic markers to delimit centromeres.

3. To map the distribution of repeats, genes, DNA and histone methylation along pseudo-chromosome sequences to identify potential distinct peaks corresponding to ENCs or ancestral centromeric regions.
4. To perform comparative genomic analysis between genomes of *Arabidopsis lyrata* and *Arabis alpina* to prove chromosomal collinearity and absence of chromosomal rearrangements (mainly hemi- and pericentric inversions).

3 MATERIAL AND METHODS

3.1 MATERIAL

The genome sequencing of ten species belonging to lineage III and expanded lineage II *(Anastatica hierochuntica*, *Braya humilis*, *Bunias orientalis*, *Chorispora tenella*, *Dontostemon micranthus*, *Euclidium syriacum*, *Farsetia stylosa*, *Hesperis matronalis*, *Lobularia libyca* and *Morettia canescens*), generating 100 bp paired-end reads, was performed on an Illumina HiSeq 2000 platform at GATC Biotech (Konstanz, Germany), whereas genomes of two species (*Hesperis sylvestris* and *Matthiola incana*) were sequenced using an Illumina MiSeq, paired 300-bp reads, and MiSeq v3 reagents, at the sequencing core facility of the Oklahoma Medical Research Foundation (Oklahoma City, USA). Genome of *Alyssum gmelinii* was sequenced using an Illumina MiSeq, paired 156-bp reads at CEITEC Genomics Core Facility (Brno, Czechia). Number of reads produced, genome size and additional information for each species are shown in **Table 2**.

Table 2 Number of sequenced reads in 13 selected species belonging to Lineage III and expanded lineage II (EII) with genome size for each species and estimated genome coverage of performed sequencing.

Species	Tribe	Lineage	Genome size (pg/Mb)		Number of PE reads	Genome coverage
Alyssum gmelinii	Alysseae	EII	-	-	14,382,900	-
Anastatica hierochntica	Anastaticeae	EII	0.87	850.9	36,164,256	4.3
Braya humilis	Euclidieae	III	1.63	1,594.1	20,445,958	1.3
Bunias orientalis	Buniadeae	III	2.67	2,611.3	49,513,064	1.9
Chorispora tenella	Chorisporeae	III	0.35	342.3	31,642,304	9.2
Dontostemon micranthus	Dontostemoneae	III	1.66	1,623.5	102,904,072	6.3
Euclidium syriacum	Euclidieae	III	0.26	254.3	29,007,668	11.4
Farsetia stylosa	Anastaticeae	EII	0.64	625.9	51,936,504	8.3
*Hesperis matronalis**	Hesperideae	III	8.11	7,931.6	23,189,130	0.3
Hesperis sylvestris	Hesperideae	III	4.36	4,264.1	6,269,600	0.4
Lobularia libyca	Anastaticeae	EII	0.52	508.6	25,181,904	5.0
Mathiola incana	Anchonieae	III	2.2	2,151.6	5,715,646	0.8
Morettia canescens	Anastaticeae	EII	0.85	831.3	29,712,804	3.6

* Tetraploid species.

The genome sequencing of Arabideae species was performed using an Illumina MiSeq, paired 300-bp reads, and MiSeq v3 reagents, at the sequencing core facility of the

Oklahoma Medical Research Foundation (Oklahoma City, USA). Number of reads produced, genome size and additional information for each species are shown in **Table 3**. Publicly available data were used for species *A. alpina* (BioProject: PRJNA241291) and *A. montbretiana* (BioProject: PRJNA258048).

Table 3 Number of sequenced reads in selected Arabideae species with genome size for each species and estimated genome coverage (or depth) of performed sequencing. (EII – Expanded lineage II)

Species	Tribe	Lineage	Genome size (pg/Mb)		Number of PE reads	Genome coverage
Arabis auriculata	Arabideae	EII	0.21	205.4	5,907,124	8.6
Arabis cypria	Arabideae	EII	0.39	381.4	5,610,682	4.4
Arabis planisiliqua	Arabideae	EII	-	-	5,815,994	-
Aubrieta canescens	Arabideae	EII	0.40	391.2	4,536,844	3.5
Draba hispida	Arabideae	EII	-	-	4,961,206	-
Draba muralis	Arabideae	EII	0.47	454.8	5,133,548	3.4
Draba nemorosa	Arabideae	EII	0.32/0.24	313.0/234.7	6,449,486	6.2/8.2
Pseudoturritis turrita	Arabideae	EII	0.38	371.6	5,257,476	4 .2

In both read datasets, the majority of reads originated from nuclear genome but chloroplast, and mitochondrial DNA reads were presented as well because we performed sequencing of genomic DNA extracted from whole leaf tissues. Chloroplast reads were extracted and/or removed from genome sequencing data depending on further purpose of bioinformatics analysis. Up to 14% of reads in analyzed species were identified as chloroplast reads.

3.2 WORKFLOWS

Three workflows were designed and used in this work:

3.2.1 REPEAT IDENTIFICATION AND CHARACTERIZATION

1. Preprocessing of raw sequence data

    - Quality filtering and trimming
    - Adapter and contamination removing (e.g. removing phiX reads)
    - Sequence filtering (cpDNA reads can be filtered out before next steps)
    - Discarding single (orphan) reads
    - Interlacing
    - Sampling
    - Modification of sequence names (for comparative study)
    - Converting fastq to fasta

2. Run RepeatExplorer pipeline

3. Run TAREAN with option merge clusters (can identify longer tandem repeat monomers)

5. Run TRF on merged reads for identification of short tandem repeats (monomer shorter than read length)

6. Annotation of unknown clusters with other bioinformatics tools – Censor, blastn and blastx using GenBank databases

7. The oligoprobe and primer design for FISH-based *in situ* localization of selected repetitive sequences

3.2.2 CHLOROPLAST GENOME ASSEMBLY

1. Preprocessing of raw sequence data

    - Quality filtering and trimming

- Adapter and contamination removing (e.g. removing phiX reads)
- Interlacing
- Sampling
- Converting fastq to fasta

2. Use all reads and/or only extracted cpDNA reads

3. Assembly

    - *De novo* assembly
    - Assembly quality comparison between outputs from different assembly settings

4. Closing gaps and scaffolding

5. Building consensus sequence

8. Annotation of genes, tRNA, repeats

9. extracting genes/markers for further analyses

3.2.3 PHYLOGENETIC ANALYSIS

1. Data selection

2. Multiple sequence alignment

3. Alignment edits

4. Evaluation of the best substitution/evolutionary model

5. Construction of a phylogenetic tree

6. Evaluation of the phylogenetic tree

7. Visualization of the phylogenetic tree

3.3 METHODS

3.3.1 PREPROCESSING OF RAW SEQUENCE DATA

3.3.1.1 Quality filtering and trimming

A quality check of paired-end reads (**Table 2** and **3**) was carried out using FastQC (Andrews, 2010). Raw sequencing data preprocessing was done before assembly and clustering analysis. Presence of adapters and their sequences were discovered by FastQC. Read-quality filtering (Phred score >20 and cutoff value 80 %), adapter trimming (removal of adapter-containing reads) and conversion of fastq to fasta were performed using the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) either through its implementation within the Galaxy environment of Repeatexplorer pipeline (Novák *et al.*, 2013) or by locally installed instance, and cutadapt (Martin, 2011). MiSeq reads were trimmed to 100 bp for comparative analysis of *Hepseris* clade species (**Table 2**) and to 200 bp for Arabideae species (**Table 3**).

FASTX-toolkit (Gordon and Hannon, 2010) is a collection of command line tools for preprocessing of fastq and fasta files.

FASTQ-Quality-Filter removes low-quality sequences from FASTQ files.

Example usage:

```
fastq_quality_filter -v -q 20 -p 80 -i input.fastq -o output.fastq
```

Example command will filter reads of input file input.fastq based on quality settings that 80% of bases in read must have quality equal to or higher than 20 (phred cut-off value). Reads which fulfill these given criteria will be saved in outputfile output.fastq, the parameter -v will report number of sequences after filtering.

FASTX-Clipper removes adapters, sequences shorter than given length, leaves only reads with or without adapter sequences.

Example usage:

```
fastx_clipper -a AACCGGTT -l 100 -C -v -i input.fastq -o output.fastq
```

Example command will discard (-C) reads which contained the given adapter sequence (-a) and also reads shorter than 100 bp will be discarded (-l).

FASTQ/A-Trimmer trims sequences in a fasta or fastq file, e.g. removes barcodes or noise.

Example usage:

```
fastx_trimmer -f 11 -l 111 -v -i input.fastq -o output.fastq
```

Example command will shorten reads; first base to keep (-f) will be 11th and the last base to keep (-l) will be 111th base.

Other useful tools from this package are e.g. FASTQ-to-FASTA (`fastq_to_fasta`, which converts a FASTQ file to FASTA file, FASTA Formatter (`fasta_formatter`), which changes the width of sequences line in a fasta file, or FASTQ/A Reverse-Complement (`fastx_reverse_complement`), which produces the reverse complement of each sequence in a fastq/fasta file. More details can be found at http://hannonlab.cshl.edu/fastx_toolkit/.

Cutadapt (Martin, 2011) is another tool for removing contamination, trimming and quality filtering. It finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from NGS reads.

Example usage:

```
cutadapt -q 15,10 -a AACCGGTT -o output.fastq input.fastq
```

Example command will trim low quality ends from reads (-q; the 5’ end will then be trimmed with a cutoff of 15, and the 3’ will be trimmed with a cutoff of 10) and after that a 3'-end adapter with known given sequence will be trimmed from reads (-a; -g for 5'-end or -b option for both possible, 5’ or 3’). There is also option -u, which can be used

to remove a fixed number of bases from the beginning or end of each read and can be combined with the other options (adapter trimming is done after -u step).

Cutadapt can be performed also on paired-end reads (use uppercase options, -A, -B, -G):

```
cutadapt -a AACCGGTT -A AACCGGTT -o output_forward.fastq -p output_reverse.fastq input_forward.fastq intput_reverse.fastq
```

3.3.1.2 Additional filtering, extracting sequences

To remove phi X 174 bacteriophage reads (used during sequencing as positive control), blast search against local built database in combination with basic bash commands was used.

Steps:

- Build nucleotide database (makeblastdb) of known contaminant fasta sequence(s).
- Blast search reads (in fasta format) against built local database (blastn).
- Get IDs (headers) of contaminant reads and of all reads using AWK and grep commands, sort the outputs.
- Compare IDs of all reads and contaminant reads (comm command).
- Get filtered reads or get reads of contaminant (seqtk tool).

This step was also used to remove cpDNA reads from Arabideae NGS data, nucleotide database was built from known cpDNA sequences of related Brassicaceae species.

Another way how to do filtering of reads is to map reads to the known reference sequence (PhiX genome sequence or cpDNA sequence of related species) using, for instance, software Bowtie2 (Langmead and Salzberg 2012), BWA (Li and Durbin, 2009), or BBMap (Bushnell 2014). Unmapped reads should be only nuclear genome reads.

Run Bowtie2:

```
bowtie2-build -f contamination.fasta dbname
bowtie2 -x dbname -1 input_forward.fastq -2 input_reverse.fastq --un-conc <path>
```

Run BWA:

```
bwa index contamination.fasta
bwa mem -M -t 16 contamination.fasta input_forward.fastq input_reverse.fastq > aln.sam
```

3.3.2 REPEAT IDENTIFICATION AND CHARACTERIZATION

3.3.2.1 Repeat identification

For repeat identification and characterization we used RepeatExplorer pipeline (Novák *et al.*, 2013). RepeatExplorer can be used through Galaxy based web interface on the public Galaxy server accessed at https://galaxy-elixir.cerit-sc.cz (RepeatExplorer2) or it can be installed locally using its freely available source code.

Principles of this computational pipeline were described by Novák *et al.* (2010, 2013). Pipeline is based on graph-based clustering. Sequence overlaps between the reads are transformed to a graph where the reads are represented as nodes (vertices) and their similarities as edges connecting the nodes (**Figure 10A**). Graph structure is divided into separate subgraphs, called clusters (**Figure 10B**). Cluster is a set of frequently overlapping reads (frequently connected nodes) which creates a repeat family.

Clusters have dense connections between the nodes within the clusters but sparse connections between nodes in different clusters (**Figure 10B**). Different repeat classes produce different shapes of graph (**Figure 11**). In the presence of paired-end reads clusters can be connected to higher structures, superclusters, depending on how much read pairs they share.

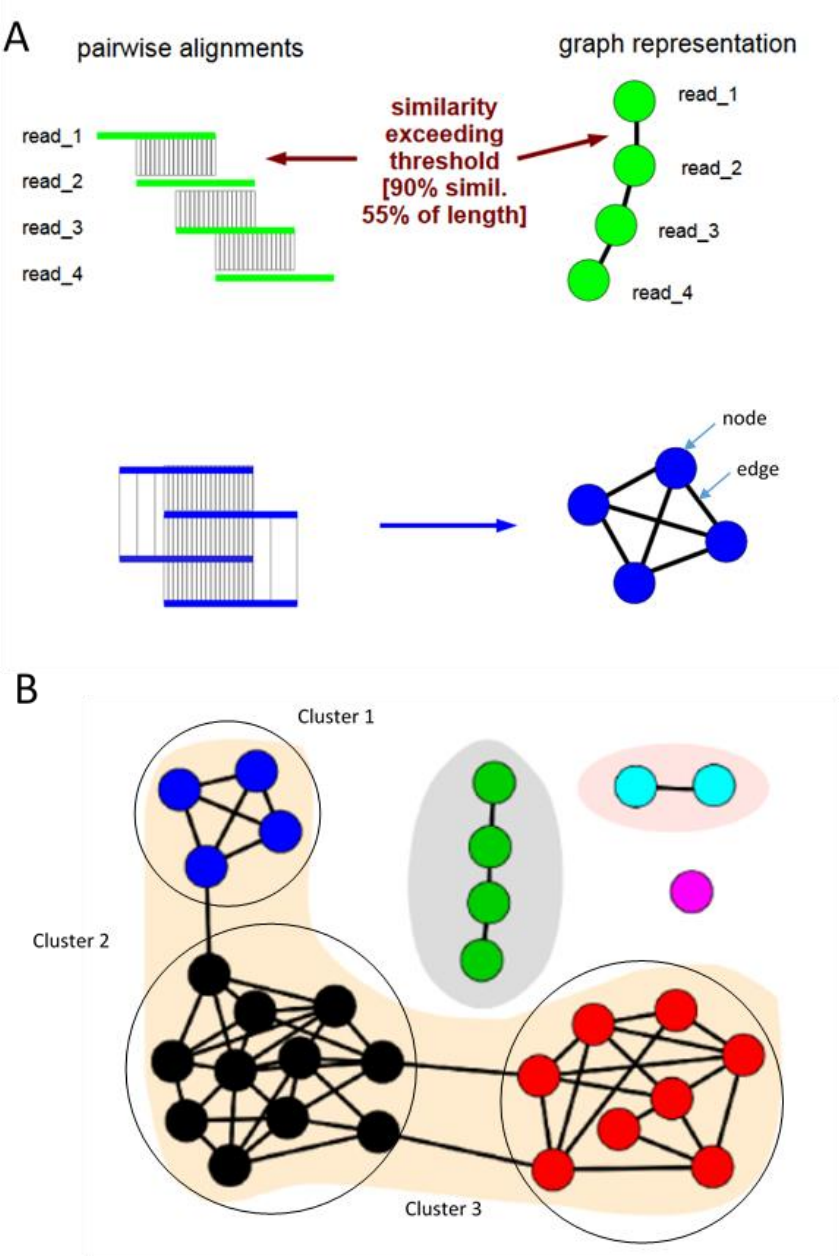


Figure 10 Principles of graph-based clustering method (Adapted from http://repeatexplorer.org/).

Clustering can be run in two different modes: (i) full repeat analysis done by RepeatExplorer clustering which focuses on all types of repeats but is less sensitive to satellite detection, or (ii) tandem repeat analysis done by TAREAN tool which focuses on more sensitive tandem repeat detection (http://repeatexplorer.org/).

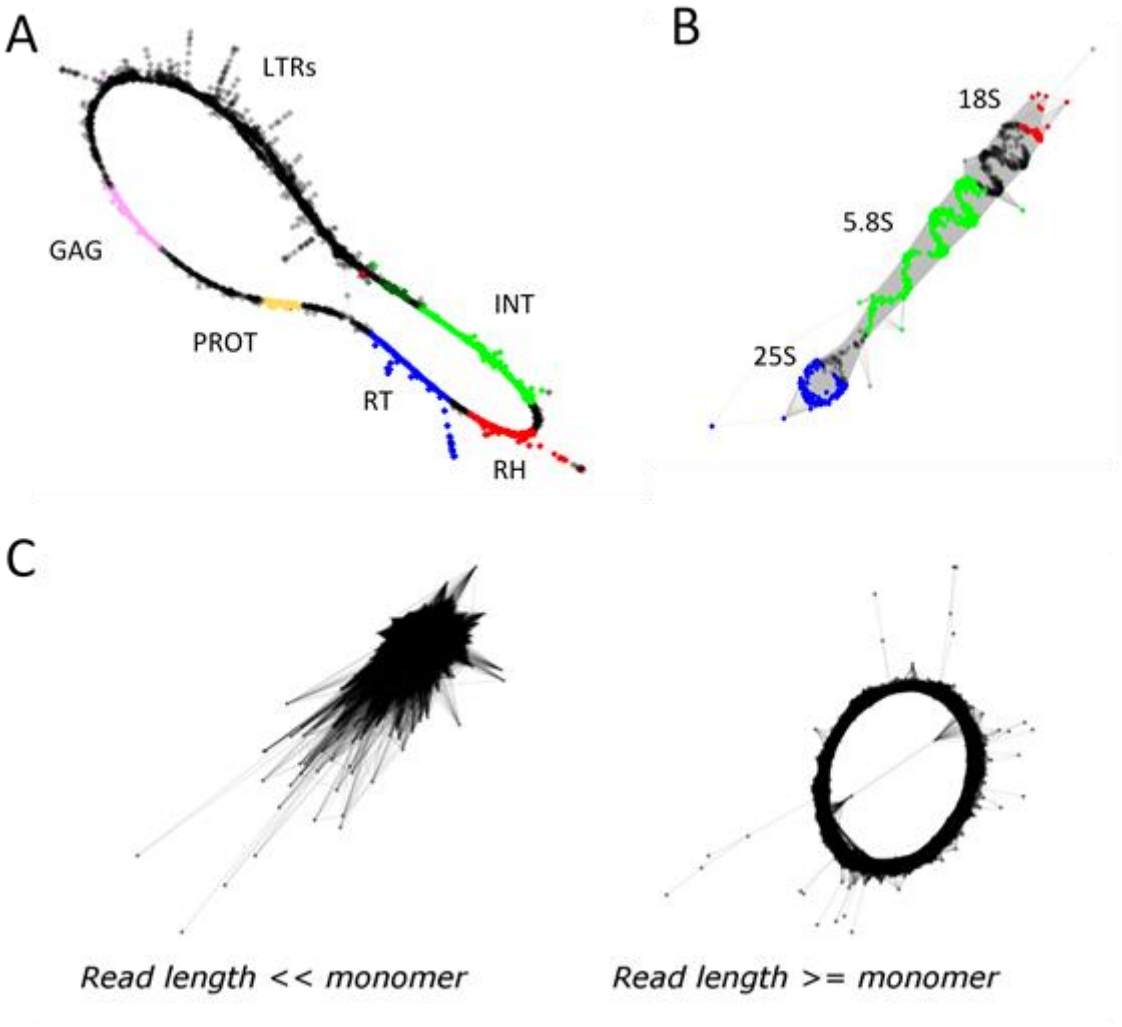


Figure 11 Different repeat classes produce different shapes of graph: (A) Ty1/*copia* LTR retroelement, (B) rDNA, (C) tandem repeats with different monomer size.

Characterization and identification of repetitive sequences by RepeatExplorer workflow used in our studies consist of several steps (all tools used, except for contamination/cpDNA read filtering, are implemented under Galaxy web-based environment; https://galaxy-elixir.cerit-sc.cz/):

0. Checking the sequencing data quality, contamination removing (incl. removing of cpDNA reads)
1. Data upload (Get data)
2. Preprocessing of raw fastq reads (Pre-processing and QC utilities) - only complete pairs are kept, single reads are discarded

a. Trimming (optional)

   All reads should have uniform length. This step is mandatory for 454 reads with variable reads length.

b. Quality filtering (80 % of bases in a read must have quality Phred score >20)

c. Adapter filtering (sequences of adapters specified by FastQC tool are used)

d. Converting fastq to fasta

e. Interlacing two fasta files

f. Sampling (optional; genome coverage of 0.01 - 0.50× is recommended)

g. Modification of sequence names

   For comparative analysis sequence names must contain code (prefix) for each group.

3. RepeatExplorer2 clustering tool works in sequential steps:

   a. Prerun – the pipeline estimates genome repetitiveness and data size limit (number of reads which can be processed with assigned RAM)

   b. All-to-all comparison is done by mgblast (default parameters are set to sequence similarities > 90% over at least 55% of the read length for reads longer than 100 bp)

   c. Cluster analysis

      i. Clustering - creating clusters

      ii. Analyzing pairs

      iii. Assembly

      Reads are assembled to contigs by CAP3 program, each cluster separately. Putative satellite clusters are not assembled by CAP3, instead TAREAN generates k-mer based consensus.

      iv. Tandem repeat annotation

      v. Comparison of all reads with known sequences – similarity-based annotation

         - Database of protein domains (REXDb) - blastx or diamond aligner

- DNA database (rDNA, tRNA, cpDNA, mtDNA, potential contaminant sequences)
- Custom database (optional) can be used

d. Result synthesis – report is created

Output includes an HTML summary with a table listing all analyzed clusters. More detailed information about clusters is provided in additional files and directories.

Together with RepeatExplorer analysis, TAREAN tool (Novák *et al.*, 2017) was used for a more precise detection of tandem repeats. TAREAN calculates graph layout and provides automatic analysis of graph topology with the aim to identify tandem repeats. Tandem repeats with longer monomers tend to split onto multiple clusters in full repeat analysis done by RepeatExplorer clustering. Thus, advanced TAREAN's option 'perform cluster merging' (clusters connected through paired-end reads are merged) was used for all our analyzed species. TAREAN reports cluster analysis which includes a list of monomer tandem repeat sequence variants reconstructed from the most frequent k-mers. The reconstructed consensus sequences are sorted based on their significance and genome proportion, and for each sequence graph image and sequence logo are produced. DNA logo represents nucleotide sequence conservation of reconstructed consensus sequences. The most conserved part of the sequence can be then selected to design oligonucleotide probes for further cytogenomic testing.

Although the clusters are automatically annotated if similarity is found between cluster reads and known sequences from database, checking and correction of automatic repeat classification is recommended. RepeatExplorer cannot identify some low-complexity repeats or simple repeats, such as telomeric motifs or microsatellites which can be underestimated or not detected at all. Also non-autonomous TEs, possessing truncated protein coding domains, and mutated TEs are difficult or impossible to classify using protein domain sequences.

We manually checked all clusters in analyzed genomes. If the cluster was not annotated, we used tool Censor to compare contigs of unknown cluster to Viridiplantae repeat

database (Repbase) and blastn and blastx search using GenBank databases. Unknown sequences without any similarities to known sequences and without any structural characteristics were left as unclassified repeats.

Software Dotter (Sonnhammer and Durbin, 1995) was used for plotting self-dotplot of putative satellite sequences identified by the RepeatExplorer pipeline to check whether contigs contain characteristic repeating pattern. Default setting was used.

Clustering analysis for each species was run on maximum number of reads and on different sampling size (genome coverage of 0.01 - 0.50×). Each genome coverages were analyzed with more replicates to check consistency of results.

3.3.2.2 Tandem repeat finder

Tandem repeat finder (TRF, Benson, 1999) was used to detect tandem repeats with short monomers (shorter than read length). Interlaced reads were concatenated to obtain one long fasta file instead of multiple fasta file. Fasta headers were replaced by string of Ns. The modified file was used as input for TRF, default parameters were used and then adjusted for a more accurate search:

```
trf concatenated_reads.fasta 2 5 7 80 10 25 25 -d
```

TRF output files were parsed using TRAP (Sobreira *et al.*, 2006) to generate sorted (by number of loci or monomer size) output summary. By this approach we verified increased abundance of telomeric repeats in some *Hesperis* clade species.

TRF was also used to screen clusters with putative satellites which were not identified by the TAREAN pipeline.

3.3.2.3 Design of FISH oligonucleotide probes

Oligo-probes for tandem repeats with monomer size below 500 bp were design from reconstructed monomer sequences. Most conserved regions with appropriate GC content (30-50%) within the consensus sequences were manually selected for ~60-bp

oligoprobes. The Geneious software platform (https://www.geneious.com, Kearse *et al.* 2012) was used to check designed sequences to minimize self-annealing and formation of hairpins.

For shared tandem repeats and tandem repeat families, design of oligonucleotide probes was targeted to conserved DNA regions of multiple alignments generated by MAFFT (Katoh and Standley 2013).

The specificity of designed probes was tested by blastn search against the database built from all clusters retrieved from individual RepeatExplorer clustering analysis.

3.3.2.4 PCR primer design

For LTR retrotransposon probes, PCR primers were designed to the *gag* gene sequence of various retrotransposon families. PCR primers for amplification from genomic DNA were also designed for tandem repeats with long monomers (>500 bp; Ávila Robledillo *et al.* 2018) identified by TAREAN pipeline.

Protein domain finder tool (Domain based ANnotation of Transposable Elements - DANTE Tool) embedded in RepeatExplorer Galaxy interface was used to find and classify all TE protein domains in individual species contigs from clustering RepeatExplorer analysis. This tool uses database of Viridiplantae protein domains derived from transposable elements (Neumann at al., 2019). The protein domain searching is done by aligning program LAST (Kiełbasa *et al.*, 2011). Domains are subsequently annotated and classified using phylogenetic approach. To retrieve only contigs with *gag* domains, the Protein domain filter tool was used with default settings (minimum identity 35 %, minimum similarity 45 % and minimum alignment length 80 %). Software Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi; Untergasser *et al.*, 2007) was then used to design primers (~20 bp) from sequences.

3.3.3 CHLOROPLAST GENOME ASSEMBLY

In Mandáková *et al.* (2017) we performed *de novo* assembly (using Ray assembler; Boisvert *et al.*, 2010) in combination with reference-guided assembly (using software package Geneious; Kearse et a., 2012) of 12 chloroplast genomes from *Hesperis* clade species and species from Expanded lineage II (**Table 2**).

For each species, chloroplast genome reads were identified by BLAST software (Altschul *et al.*, 1990). All raw reads were aligned (using BLASTn) to the reference genomes of *Arabidopsis thaliana* (AP000423) and *Lobularia maritima* (NC_009274), and only the reads with positive hits were used for *de novo* assembly. Before *de novo* assembly cp reads were downsampled to 100× coverage.

3.3.3.1 Assembly

Ray (Boisvert *et al.*, 2010) is an open source de Bruijn graph-based assembler written in C++ that can assemble genomes in parallel using the message-passing interface. It supports Illumina, 454, and SOLiD NGS data.

Example usage:

```
mpiexec -n 16 Ray -k 31 -p input_forward.fastq
input_reverse.fastq -o new_assembly
```

Paired reads can be provided as two files (-p) or as one file containing interleaved sequences (-i). Parameter -n specifies number of cores to use; -o specifies output directory. Files generated by the Ray include files Contigs.fasta and Scaffolds.fasta.

Different k-mers were tested for our data, k-mer 29 was chosen as the assembly result provided the best assembly quality metrics.

Tool QUAST (Gurevich *et al.*, 2013) was used for genome assembly evaluation and comparison to select the best assembly results that were further used for mapping to the reference genomes.

The increasing number of studies employing multigene chloroplast phylogenies has led to the development of new assembly software designed specifically for organellar assembly.

NOVOPlasty (Dierckxsens *et al.*, 2016) is a *de novo* seed-extend based assembler written in perl that provides a fast and straightforward extraction of the extranuclear genomes from NGS data in one high quality circular contig. The input reads have to be uncompressed raw Illumina reads (fastq/fasta files) or gz/bz2 zipped files. It is recommended not to filter or quality trim the reads, only adapters should be removed. Assembly output contains two versions of the assembly that vary only in the orientation of the region between the two inverted repeat regions. The correct orientation needs to be resolved manually. The software is open source and can be downloaded at https://github.com/ndierckx/NOVOPlasty.

To run NOVOPlasty a configuration file is used to specify input files and all parameters needed (e.g. k-mer size):

```
perl NOVOPlasty3.2.pl -c config.txt
```

3.3.3.2 Closing gaps and scaffolding

With short Illumina sequences (100 bp) we were not able to assembly complete chloroplast sequences using the Ray assembler. Gapfiller (Boetzer and Pirovano, 2012) was used to attempt to fill sequence gaps in scaffolds. GapFiller is a stand-alone perl program for closing gaps within preassembled scaffolds (in fasta file) using the distance information of paired-read data (fasta or fastq files; a library file has to be generated). The algorithm starts with removing low quality nucleotides from the sequence edges. Pair-end reads are mapped to the scaffolds and kept if one pair aligns to a scaffold sequence and one pair to a gapped region (Boetzer and Pirovano, 2012). The final gap-filled scaffolds are provided in output fasta file.

Example usage:

```
perl GapFiller.pl -l library.txt -s scaffolds.fasta -m 30 -
o 2 -r 0.7 -n 10 -d 50 -t 10 -T 1 -i 1 -b output
```

3.3.3.3 Building consensus sequence

We built consensus sequences in few iterative steps:

- Mapping the scaffolds to reference genome (cpDNA of *Arabidopsis thaliana* or *Lobularia maritima*) to anchor contigs and determine their orientation, and creating a preliminary pseudo-reference.
- Mapping all cp reads to build consensus (filling gaps by non-homologous sequences), manual inspection and editing.
- Mapping of all reads to consensus sequence to correct misassemblies and ambiguities which were introduced by reference-based approach.

All steps were done in software package Geneious v8.1.7 (Kearse et a., 2012). Geneious is a desktop software application framework for the organization and analysis of biological and bioinformatics data using numerous analytical tools, with interactive visualizations. The basic usage can be extended with plugins for assembly (e.g. Velvet and Bowtie), alignment (e.g. MAFFT and Clustal), and phylogenetics tools (e.g. RAxML, MrBayes).

3.3.3.4 Annotation of cp genes and tRNA

Annotation of all assembled cp genomes was performed on the Dual Organellar GenoMe Annotator (DOGMA; Wyman *et al.*, 2004). Each DOGMA annotation was manually corrected for the start and stop codons or intron/exon junctions by comparison to known homologous chloroplast genes. tRNA genes were checked by ARAGORN (Laslett and Canback, 2004). Sequences were then submitted to GenBank via a web-based sequence submission tool BankIt (https://www.ncbi.nlm.nih.gov/WebSub/).

3.3.3.5 Extracting genes/markers for further analyses

Selected genes or markers for further analyses were extracted by custom-made bash script, software BEDTools (Quinlan and Hall, 2010) and/or seqtk tool (available at https://github.com/lh3/seqtk). BEDTools is a tool package for a wide range of genomics analytical tasks. seqtk is a toolkit of programs for working with sequence data in fasta and fastq formats.

Example usage of BEDTools:

```
bedtools getfasta -fi input.fasta -bed genes.bed -fo extracted_genes.fasta
```

(genes.bed file contains gene names or identifiers and their coordinates in genome.)

Example usage of seqtk tool:

```
seqtk subseq input.fasta genes.ids > output.fasta
```

Command extracts sequences with names in a given file (genes.ids, one sequence name per line).

```
seqtk subseq input.fasta gene_regions.bed > output.fasta
```

Command extracts sequences in regions listed in a given file (gene_regions.bed).

3.3.4 PHYLOGENETIC ANALYSIS

3.3.4.1 Data selection

Both protein and nucleotide data can be used for generating phylogenetic trees. Protein sequences are more conserved than nucleotide sequences and thus it is possible to recover more conserved sites in a multiple protein sequence alignment. Non-coding DNA regions tend to be more variable than coding regions as they are not under functional constraint. Two most commonly used non-coding markers are nuclear ITS (Internal transcribed spacer) and chloroplast region *trnL-trnF* (Calonje *et al.*, 2009). It is also important to choose an appropriate taxon sampling and markers with regard to

what relationships we want to solve, and on which taxonomic level (e.g. species, clade, family).

For purpose of our study (Mandáková et al., 2017) to elucidate position of *Hesperis* clade within Brassicaceae family, we used chloroplast sequences of our *de novo* assembled species in combination with data from study by Hohmann *et al.* (2015) and chloroplast sequences available from GenBank to have better taxon sampling. We chose the same set of markers as in Hohmann et al. (2015), i.e., a set of 73 genes, including 51 protein-coding genes (introns were excluded), 19 tRNAs, and three, giving 38,622 bp for each species.

To reconstruct the evolution of genome size in the *Hesperis* clade, ITS and *ndhF* phylogenies were constructed using sequences retrieved from GenBank and BrassiBase (Kiefer et al., 2014; https://brassibase.cos.uni-heidelberg.de) for species with known genome size.

3.3.4.2 Multiple sequence alignment

Multiple sequence alignment compares three or more sequences. It arranges DNA or protein sequences to identify regions of similarity, shifts the data by inserting gaps to line up all the conserved (homologous) sites into vertical columns. There are several methods of alignments with different strategies (e.g. progressive alignment method, iterative methods or Hidden Markov Models). Most common alignment programs are MUSCLE (MUltiple Sequence Comparison by Log-Expectation; Edgar, 2004), MAFFT (Multiple Alignment using Fast Fourier Transform; Katoh *et al.*, 2002) and ClustalW (Thompson, 2003).

We conducted multiple sequence alignments of our analyzed sequences in MAFFT v7.017 (Katoh and Standley 2013) as a plugin in the Geneious platform (v8.1.7; Kearse et a., 2012). MAFFT offers various multiple alignment strategies (progressive and iterative refinement methods).

3.3.4.3 Alignment edit

After multiple sequence alignment, it is necessary to remove the most variable parts (non-conserved positions). This editing step was done manually and by using program Gblocks (Castresana, 2000) which eliminates poorly aligned positions and variable regions of a DNA alignment and makes the alignment more suitable for phylogenetic analyses.

3.3.4.4 Evaluation of the best substitution/evolutionary model

The crucial step is to use the correct model of evolution to obtain an accurate phylogenetic tree. Programs, such as jModelTest (Posada, 2008) for nucleotide sequences or ProtTest (Abascal *et al.*, 2005) for protein sequences, can predict which model algorithm would best capture evolution of the given dataset. The best model is usually the one with the lowest AIC (Akaike Information Criterion) and/or BIC (Bayesian Information Criterion). This model is later used to build the phylogenetic tree.

PartitionFinder (Lanfear *et al.*, 2016) is a program to select not only models of molecular evolution but also best-fit partitioning schemes by estimating independent models of molecular evolution for subsets of sites that most likely have evolved in similar ways.

Using PartitionFinder, three subsets were found and the GTR+Γ+I substitution model was chosen for chloroplast sequences multiple sequence alignment.

3.3.4.5 Construction of phylogenetic tree

Pre-run of phylogenetic analysis was conducted by ML method using RAxML software (Stamatakis, 2014). Final trees were conducted by softwares MrBayes (Huelsenbeck and Ronquist, 2001) and BEAST (Bouckaert et al., 2014).

RAxML (Randomized Axelerated Maximum Likelihood; Stamatakis, 2014) is a tool for maximum-likelihood based phylogenetic inference which produces bootstrapped Maximum likelihood (ML) phylogenies.

Example:

```
raxmlHPC -m GTRGAMMAI -p 12345 -q partition_file.txt -s multialignment.phy -n output -o outgoup_name -f a -x 123456 -N1000
```

-p and –x provide seed numbers, so that the program can generate random numbers for the bootstrapping process; -m specifies model; -s gives the name of input phylip file; -n gives the name of output file, which will be made into a tree.

MrBayes (Huelsenbeck and Ronquist, 2001) is a software for the Bayesian estimation of phylogeny. MrBayes uses Markov chain Monte Carlo (MCMC) methods to approximate the posterior distribution. The posterior probability of phylogenetic trees can be calculated using Bayes theorem.

Example of a nexus input file:

```
#nexus
begin data;
dimensions ntax=3 nchar=11;
format datatype=dna gap=-;
matrix
species1 ACCATTGGCT
species2 AC-GTTGGCT
outgroup_species AC-GTCAGCC
;
end;
begin mrbayes;
log start filename=Bayes.log replace;
set autoclose=yes nowarn=yes;
outgroup outgroup_species;
lset nst=6 rates=gamma;
mcmcp ngen=1000000 samplefreq=1000
printfreq=500 nchains=4 nruns=2;
mcmc;
sump burnin=300;
sumt burnin=300;
log stop;
end;
```

Run MrBayes:

```
execute Bayes.log
```

BEAST (Bouckaert *et al.*, 2014) is another widely used Bayesian phylogenetic software based on MCMC methods. It estimates rooted, time-measured phylogenies using strict or relaxed molecular clock models (https://www.beast2.org/). BEAST can be run with a graphical user interface (GUI) and as the input takes an XML command file and returns output log files. BEAST XML files are generated by a GUI application BEAUti which is a part of BEAST package.

Divergence time estimation in our study (Mandáková et al., 2017) was conducted in BEAST v2.4.4 (Bouckaert *et al.*, 2014) using independent site and clock models. *Vitis vinifera* was defined as an outgroup. Fossil constraints were adopted from Hohmann *et al.* (2015), normal distribution was used. Two independent MCMC runs were generated with 300,000,000 generations each, sampled every 30,000 generations. LogCombiner v1.8.3 (https://www.beast2.org/programs/) was used to combine trees from the two runs, and 10% of trees were discarded as burn in. TreeAnnotator v1.8.3 (https://www.beast2.org/treeannotator/) was used to generate a maximum clade credibility tree. Parameter values of each run were checked using Tracer v.1.6 (Rambaut and Drummond, 2009). Tracer is a graphical tool for visualization and diagnosing the convergence of chains of MCMC output generated by Bayesian runs. It can be used to analyze runs of BEAST and MrBayes programs.

Unrooted phylogenetic trees for ITS and *ndhF* datasets in Hloušková *et al.* (2019) were reconstructed using MrBayes v3.2.6. In all Bayesian analyses, starting trees were random, four simultaneous Markov chains were run for 5,000,000 generations, burnin values were set at 500,000 and trees were sampled every 5,000 generations. Bayesian posterior probabilities were calculated using a MCMC sampling approach. The 50% majority rule was used for constructing consensus trees. All parameters were checked by Tracer v1.6 (Rambaut and Drummond, 2009) for convergence (ESS >200).

3.3.4.6 Evaluation of the phylogenetic tree

The last but important step is the evaluation of the phylogenetic tree. The most common method to test the amount of tree support is to evaluate the statistical support for each node on the tree. This is achieved e.g. by bootstrapping, jackknifing or posterior probabilities.

The value that indicates a statistically well-supported grouping is considered to be 70% or above for ML bootstrap values (Hillis and Bull, 1993). For Bayesian analysis, a good statistical support is for posterior probabilities of 0.95 or higher (Hillis and Bull, 1993).

3.3.4.7 Visualization of the phylogenetic tree

Visualization of the final phylogenetic trees was done in FigTree software (Rambaut, 2014).

3 RESULTS

3.1 MONOPHYLETIC ORIGIN AND EVOLUTION OF THE LARGEST CRUCIFER GENOMES

Mandáková T, **Hloušková P**, German DA, Lysak MA. (2017). Monophyletic origin and evolution of the largest crucifer genomes. *Plant Physiology*. 174(4), 2062-2071.

PH performed *de novo* assembly of 13 chloroplast genomes, phylogenetic analysis based on chloroplast sequences, and the divergence time estimates. PH wrote the respective parts of Materials and Methods and Results.

Summary

Clade E, or the *Hesperis* clade, is one of the major Brassicaceae clades, classified into seven tribes (Anchonieae, Buniadeae, Chorisporeae, Dontostemoneae, Euclidieae, Hesperideae, and Shehbazieae). The clade is known for large genome size variation (more than a 30-fold variation) but low numbers of chromosomes (n = 5–7).

Methods of comparative cytogenetic analysis and phylogenetics based on chloroplast sequences were used to construct the cytogenetic maps in selected representatives of the clade E tribes and to investigate their relationships in a family-wide context.

Comparative cytogenetic maps were constructed by chromosome painting for three species: *Chorispora tenella*, *Euclidium syriacum* and *Strigosella africana*. By comparing the karyotype structure of the analyzed species, a putative structure of an ancestral karyotype of clade E (CEK; n = 7) was inferred.

The low-coverage whole-genome Illumina sequencing of eight genomes of clade E species (representing six out of the seven tribes), four genomes from tribe Anastaticeae tribe, and one genome from tribe Alysseae were performed. Chloroplast reads were filtered out from sequencing data and chloroplast genome sequences were assembled

and used for phylogenetic tree reconstruction together with the already published whole-chloroplast data available for Brassicaceae species.

Our cytogenetic analyses, along with whole-chloroplast phylogeny, support the monophyletic origin of the *Hesperis* clade. The divergence time estimates, based on chloroplast genes, dated the origin of the *Hesperis* clade to the Oligocene, followed by subsequent Miocene tribal diversifications.

The tribe Anastaticeae was formerly treated as belonging to the crucifer clade E. However, the absence of the clade E-specific chromosomal rearrangements on pachytene chromosomes of three investigated Anastaticeae species and their phylogenetic position outside clade E supported our hypothesis that Anastaticeae does not belong to clade E.

Check for updates

Research Report

Monophyletic Origin and Evolution of the Largest Crucifer Genomes[1]

Terezie Mandáková,[a] Petra Hloušková,[a] Dmitry A. German,[b,c] and Martin A. Lysak[a,2]

[a]Plant Cytogenomics Research Group, Central European Institute of Technology, and Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic
[b]Department of Biodiversity and Plant Systematics, Centre for Organismal Studies, Heidelberg University, 69120 Heidelberg, Germany
[c]South-Siberian Botanical Garden, Altai State University, 656049 Barnaul, Russia

ORCID IDs: 0000-0001-6485-0563 (T.M.); 0000-0001-7951-1644 (D.A.G.); 0000-0003-0318-4194 (M.A.L.).

Clade E, or the *Hesperis* clade, is one of the major Brassicaceae (Crucifereae) clades, comprising some 48 genera and 351 species classified into seven tribes and is distributed predominantly across arid and montane regions of Asia. Several taxa have socioeconomic significance, being important ornamental but also weedy and invasive species. From the comparative genomic perspective, the clade is noteworthy as it harbors species with the largest crucifer genomes but low numbers of chromosomes (n = 5–7). By applying comparative cytogenetic analysis and whole-chloroplast phylogenetics, we constructed, to our knowledge, the first partial and complete cytogenetic maps for selected representatives of clade E tribes and investigated their relationships in a family-wide context. The *Hesperis* clade is a well-supported monophyletic lineage comprising seven tribes: Anchonieae, Buniadeae, Chorisporeae, Dontostemoneae, Euclidieae, Hesperideae, and Shehbazieae. The clade diverged from other Brassicaceae crown-group clades during the Oligocene, followed by subsequent Miocene tribal diversifications in central/southwestern Asia. The inferred ancestral karyotype of clade E (CEK; n = 7) originated from an older n = 8 genome, which also was the purported progenitor of tribe Arabideae (KAA genome). In most taxa of clade E, the seven linkage groups of CEK either remained conserved (Chorisporeae) or were reshuffled by chromosomal translocations (Euclidieae). In 50% of Anchonieae and Hesperideae species, the CEK genome has undergone descending dysploidy toward n = 6 (−5). These genomic data elucidate early genome evolution in Brassicaceae and pave the way for future whole-genome sequencing and assembly efforts in this as yet genomically neglected group of crucifer plants.

Already, the Romans prized the dame's rocket (*Hesperis matronalis*) and stocks (*Matthiola incana* and *Matthiola longipetala*) for their delightful fragrances, which develop in the late afternoon and persist long through the evening and night. However, these plants and their close relatives, classified today as members of clade E, are not only attractive for their scent but also for their large, diversely colored flowers, decorating our gardens today (*Matthiola* spp.) as well as mainly Asian steppes, grasslands, rocky outcrops, and sparsely vegetated screes of high mountains (e.g. *Chorispora*, *Clausia*, *Hesperis*, *Matthiola*, *Parrya*, *Solms-laubachia*, and *Tchihatchewia* spp.; Fig. 1). On the less attractive side, several clade E species also are regarded as noxious weeds (*Chorispora tenella* and *Strigosella africana*) and invasive elements entering naturally occurring plant communities (*Bunias orientalis* and *H. matronalis*; Francis et al., 2009, CABI, 2012). According to the Global Naturalized Alien Flora database covering 843 regions worldwide (van Kleunen et al., 2015), the two most invasive clade E species are *H. matronalis*, reported to be naturalized in 97 regions, and *B. orientalis* in 53 regions, followed by *M. incana* (44 regions), *S. africana* (28 regions), and *Euclidium syriacum* (19 regions).

According to the most recent tribal treatment of Brassicaceae (Al-Shehbaz, 2012), lineage III (Beilstein et al., 2006) or clade E (Huang et al., 2016) includes seven tribes, namely Anastaticeae (ANAS; 13 genera/65 species), Anchonieae (ANCH; 10/75), Buniadeae (BUNI; one/two), Chorisporeae (CHOR; four/55), Dontostemoneae (DONT; two/17), Euclidieae (EUCL; 28/149), and Hesperideae (HESP; two/52), plus the recently described monotypic Shehbazieae (SHEH; one/one; German and Friesen, 2014). In congruence with some previous studies (for review, see German et al., 2011), this circumscription of lineage III was not fully supported by the multigene analysis of Huang et al. (2016), due to ANAS (*Lobularia maritima*) being positioned outside of the monophyletic clade E or *Hesperis* clade of six tribes (ANCH, BUNI, CHOR,

[1] This work was supported by a research grant from the Czech Science Foundation (grant no. P501/12/G090) and the CEITEC 2020 (grant no. LQ1601) project. D.A.G. was supported by a research grant from the DFG (grant no. KO2302-13/1,2).

[2] Address correspondence to martin.lysak@ceitec.muni.cz.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Martin A. Lysak (martin.lysak@ceitec.muni.cz).

M.A.L. and T.M. conceived the project and research plans; T.M. and P.H. performed most of the experiments; M.A.L., D.A.G., and T.M. wrote the article.

www.plantphysiol.org/cgi/doi/10.1104/pp.17.00457

 © 2017 American Society of Plant Biologists. All Rights Reserved.

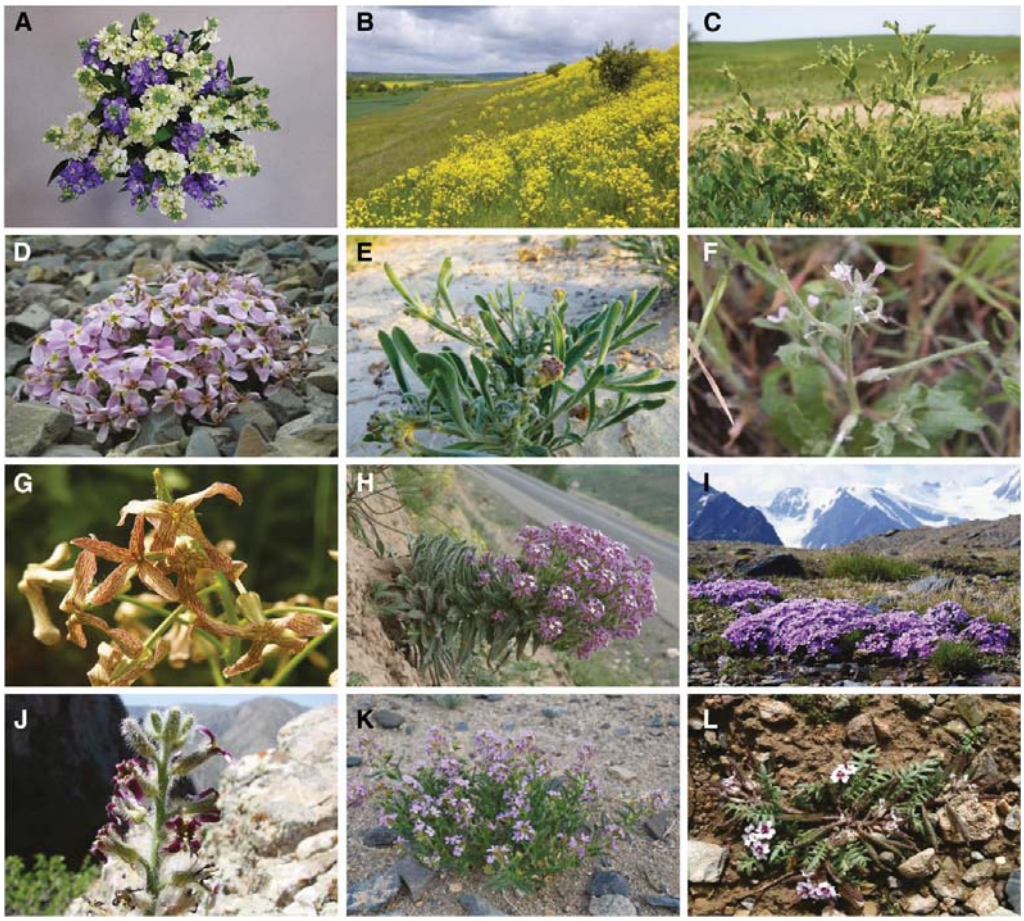


Figure 1. Representatives of the seven tribes of clade E. A, A bouquet of *Matthiola incana* (ANCH). B, *Bunias orientalis* (BUNI). C, *Euclidium syriacum* (EUCL). D, *Leiospora exscapa* (EUCL). E, *Tetracme quadricornis* (EUCL). F, *Strigosella africana* (EUCL). G, *Hesperis tristis* (HESP). H, *Tchihatchewia isatidea* (HESP). I, *Chorispora bungeana* (CHOR). J, *Parrya olgae* (CHOR). K, *Dontostemon elegans* (DONT). L, *Shehbazia tibetica* (SHEH). Photographs are by T. Mandáková (A), K. Schneider (B), P.E. Yevseyenkov (C), I.E. Smelyansky (D), A.L. Ebel (E), S.V. Smirnov (F), L. Hoskovec (G), E. Rencová (H), P.A. Kosachev (I and K), N. Yu Beshko (J), and Q. Lin (L). All photographs are reproduced with the permission of their authors.

DONT, EUCL, and HESP; SHEH was not studied but should be assigned here because it represents an ancient hybrid between CHOR and DONT). ANAS consistently clustered with representatives of Biscutelleae, Cochlearieae, and Iberideae, as a newly recognized clade C (Huang et al., 2016; Guo et al., 2017). Without ANAS, the seven tribes of the *Hesperis* clade include 48 genera and 351 species and represent 9% of the total species diversity of the family (BrassiBase; Kiefer et al., 2014).

The *Hesperis* clade has a special position among all Brassicaceae lineages and clades due to its unusual, more than 30-fold variation in genome size. Whereas most Brassicaceae species possess very small genomes with a mean size of 0.62 Gb (Lysak et al., 2009), the largest genomes have been found among clade E species (Lysak et al., 2009; Kiefer et al., 2014). Crucifer genomes larger than 2 Gb are represented by species of *Bunias* (BUNI), *Clausia* (DONT), *Hesperis* (HESP), and *Matthiola* (ANCH). The largest known genome of clade E and the whole family was estimated for *H. matronalis* (8 Gb; $2n = 24$ and 28), whereas the smallest genomes in clade E (0.26 Gb) were reported for *Diptychocarpus strictus* (CHOR; $2n = 14$) and *E. syriacum* (EUCL; $2n = 14$).

The two smallest genomes were chosen to be sequenced within the framework of the BMAP initiative (JGI Genome Portal; accessed January 31, 2017).

Despite its genomic, phylogenetic, and ecogeographical distinctiveness within the Brassicaceae, as well as its socioeconomic importance, virtually nothing is known about the origin and genome evolution of the *Hesperis* clade. Therefore, to our knowledge for the first time, we investigated genome evolution in tribes assigned to clade E by comparative chromosome painting, with the aim to reconstruct its ancestral genome and elucidate the genomic processes that have shaped the origin of this lineage. Our cytogenetic analyses, along with whole-chloroplast phylogeny, support the monophyly of the *Hesperis* clade, allowing us to construct, to our knowledge, the first cytogenomic maps and propose an ancestral genome for the lineage. This phylogenomic analysis is an important step toward achieving a better understanding of early genome evolution in the Brassicaceae.

RESULTS

Karyotypes of Clade E Species

Comparative cytogenetic maps were constructed by chromosome painting for the following species: *C. tenella* ($2n = 14$; CHOR), *E. syriacum* ($2n = 14$; EUCL), and *S. africana* ($2n = 28$; EUCL; Fig. 2). The karyotypes were then compared with the reference ACK genome comprising eight chromosomes and 22 genomic blocks (GBs; Schranz et al., 2006; Lysak et al., 2016). In *C. tenella*, only chromosome Ct3 structurally resembled the ancestral chromosome AK3, whereas the remaining GB associations (except for D-E) were reshuffled by chromosomal rearrangements (Fig. 2A).

In *E. syriacum* (Fig. 2B), none of its seven chromosomes retained the ACK structure; however, chromosomes Es4 and Es5 were structurally identical to Ct4 and Ct5 of *C. tenella*. Among the other five linkage groups, GBs on the upper arms of chromosomes Es3 and Es6 resembled the structures of Ct3 and Ct6 in *C. tenella*. Similarly, the upper arms of chromosomes Es2 and Es7 had the same GB composition as the bottom arms of Ct2 and Ct7 in *C. tenella*. Chromosome Es1 differed from its Ct1 homolog by a paracentric inversion on the upper arm (Fig. 2C).

The tetraploid genome of *S. africana* resembled that of *Euclidium* spp., with all but one homolog pair having the same structure. Chromosomes Sa3 and Sa3′ differed from the Es3 homolog by a paracentric inversion on the bottom arm (Fig. 2B).

As large-scale comparative chromosome painting (CCP) on pachytene chromosomes in ANCH, CHOR, DONT, and HESP genomes with a high repeat content was challenging (for details, see "Materials and Methods"), only the unique GB associations shared among karyotypes of *Chorispora*, *Euclidium*, and *Strigosella* (i.e. Ct1/Es1/Sa1, Ct4/Es4/Sa4, and Ct5/Es5/Sa5) were identified successfully on mitotic chromosomes of *B. orientalis* ($2n = 14$; BUNI), *Dontostemon micranthus* ($2n = 14$; DONT), *Hesperis sylvestris* ($2n = 12$; HESP), and *M. incana* ($2n = 14$; ANCH). CCP localization of linkage group 1 (GBs A and B) in the four species is shown in Figure 2C. Chromosome 1 in *M. incana*, *D. micranthus*, and *B. orientalis* resembled Ct1 in *C. tenella*, with the upper arm bearing GBs A and Ba and the bottom arm formed by Bb. In the two latter species, the terminal part of the upper arm (A-Ba) remained unlabeled after applying the painting probe for chromosome Ct1. In *H. sylvestris*, chromosome 1 was structurally similar to its homolog in EUCL species (Fig. 2B); however, its terminal parts were not painted by the probe corresponding to chromosome Es1. These findings suggest that chromosome 1 in BUNI, DONT, and HESP species participated in a taxon-specific translocation event(s). In *H. sylvestris*, the structure of chromosome 1 may indicate that this chromosome was formed via an insertion-like translocation event (nested chromosome insertion) responsible for the descending dysploidy from $n = 7$ to $n = 6$.

CCP with probes corresponding to homologs Ct4/Es4/Sa4 and Ct5/Es5/Sa5 did not uncover any specific chromosomal rearrangements in any of the ANCH, BUNI, DONT, and HESP species analyzed. Thus, these two chromosomes are shared by all the analyzed clade E species.

Ancestral Karyotype of Clade E

By comparing the karyotype structure of the analyzed species, we inferred a putative structure of the ancestral genome shared by all clade E tribes (Fig. 2D). The CEK genome had seven linkage groups and was structurally closest to the analyzed genome of *C. tenella* (Fig. 2A), whereby only a single reciprocal translocation differentiates the two genomes. Three reciprocal translocations transformed the CEK genome into the *Euclidium*/*Strigosella* karyotype (Fig. 2B; Supplemental Fig. S1).

ANAS Did Not Descend from CEK

As tribe ANAS was formerly treated as belonging to clade E, we attempted to identify CEK-specific GB associations on pachytene chromosomes of three ANAS species, namely *Farsetia stylosa* ($2n = 20$), *Lobularia libyca* ($2n = 22$), and *Morettia canescens* ($2n = 22$). However, we failed to identify three unique GB associations (i.e. A-B, M-Jb, and U) in the ANAS genomes analyzed. Instead, the three tested chromosomes of *F. stylosa*, *L. libyca*, and *M. canescens* exhibited ACK-derived associations of GBs (data not shown). As two genomic copies of each GB were consistently observed in haploid complements of ANAS species with the lowest known chromosome numbers for the tribe, these genomes probably

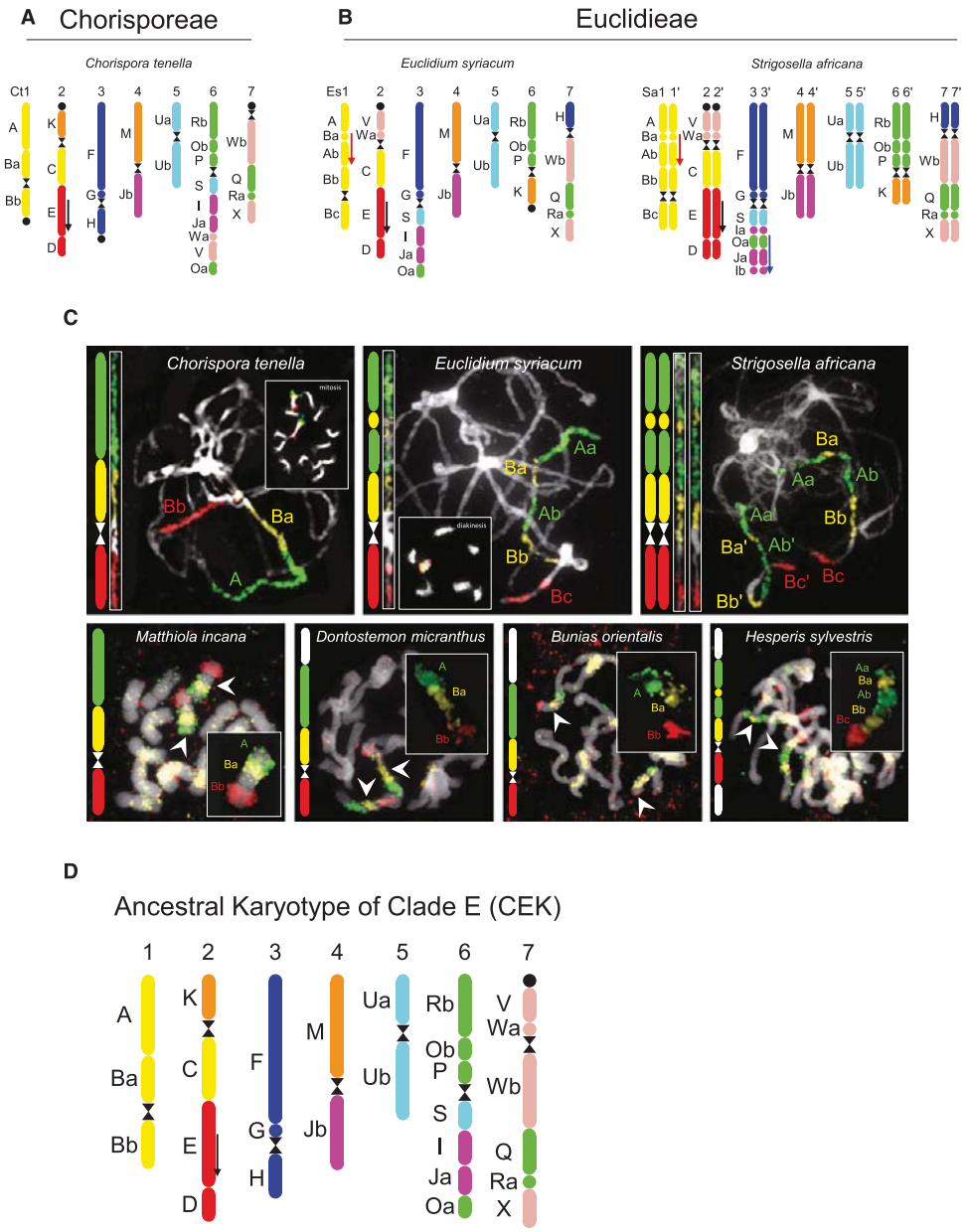


Figure 2. Ideograms of the extant and ancestral genomes of clade E tribes and examples of cytogenetic analyses. A, Genome structure of *C. tenella* (CHOR). B, Genome structures of *E. syriacum* and the neotetraploid *S. africana* (both EUCL). Black arrows refer to the inverted collinearity of block E in relation to the ancestral crucifer karyotype (ACK); red arrows show a EUCL-specific paracentric inversion on chromosome 1; the blue arrow indicates a paracentric inversion differentiating chromosome 3 in *E. syriacum* and *S. africana*. C, Identification of genomic blocks A and B (linkage group 1) by comparative chromosome painting analysis on pachytene chromosomes (top three species) and mitotic chromosomes of seven clade E species. D, Ancestral karyotype of clade E (CEK). The color code and capital letters correspond to the eight chromosomes and 22 genomic blocks of ACK, respectively. The black circle marks the position of the 35S rDNA locus. Colors in C correspond to epifluorescence of biotin-, digoxigenin- and Cy3-labeled painting contigs. Chromosomes were counterstained by 4′,6-diamidino-2-phenylindole (DAPI).

originated by a whole-genome duplication event(s) not detected in clade E genomes.

Comparison of CEK with Other Ancestral Genomes

After inferring CEK, we aimed to elucidate its closest relatives among the yet proposed crucifer ancestral genomes. The seven linkage groups of CEK hinted at descending dysploidy from an older $n = 8$ genome. We realized that chromosomes CEK_1 (GBs A-B), CEK_3 (F-G-H), and CEK_5 (Ua-Ub) resembled chromosomes KAA_1, KAA_3, and KAA_7 in the KAA genome of *Arabis alpina* (Willing et al., 2015). CEK and KAA share the structure of the bottom arm of chromosomes CEK_4 and KAA_4 (GB Jb), and the GB compositions of chromosomes CEK_7 and KAA_8 are notably similar. Chromosome CEK_3 has the same structure as its homologs in ACK, the proto-Calepineae karyotype (PCK; Lysak et al., 2016), and KAA (except for the different centromere position in KAA; Willing et al., 2015). GB association D-E can be identified as either an entire chromosome in ACK, PCK, and KAA or as a part of chromosome CEK_2. Altogether, extant as well as reconstructed chromosomal structures link the inferred CEK and KAA genomes of *A. alpina* (Willing et al., 2015). We propose that the two lineages (i.e. clade E and tribe Arabideae) descended from a genome with eight linkage groups (Fig. 3). This $n = 8$ ancestral genome presumably shared a common ancestor with ACK ($n = 8$), which was retained up to the current time in tribes of lineage I and reshuffled to form the pre-PCK genome ($n = 8$) of clade C (Geiser et al., 2016) and the PCK genome ($n = 7$) of clade B/lineage II (Mandáková and Lysak, 2008).

Clade E Is a Monophyletic Lineage with Miocene Tribal Diversification

To corroborate the monophyly of clade E retrieved by cytogenetic analyses, we sequenced whole-chloroplast genomes of eight clade E species (representing six out of seven tribes) and four ANAS representatives. Our sequence data were analyzed together with all the whole-chloroplast data hitherto available for Brassicaceae species (Hohmann et al., 2015; Guo et al., 2017; GenBank accessions). In the phylogenetic tree (Fig. 4), the core Brassicaceae taxa were divided into two clades: clade A (lineage I) and all other crown-group clades. Within the latter group, clade E was retrieved as sister to the three remaining clades (clades B, C, and D) with high statistical support (Bayesian posterior probability of 100%). The ANAS genomes clustered together with other clade C genera outside of clade E.

Using four divergence time estimates (Magallón et al., 2015), we inferred the Aethionemeae-core Brassicaceae clade split to have occurred 40.07 million years ago (mya), with 95% high posterior density of 29.44 to 54.66 mya. The origin of clade E was dated to 29.27 mya (Oligocene), and the diversification of clade E tribes commenced at 24.60 mya in the Late Oligocene and continued throughout the Miocene (Fig. 4).

DISCUSSION

The *Hesperis* Clade Is a Well-Supported Monophyletic Lineage

Clade E, or the *Hesperis* clade, is an evolutionary unit defined by multiple parameters. (1) Morphologically, clade E species share multicellular glands (a unique character in the family), simple, nonauriculate leaves with blades usually gradually narrowing to a petiole, and often lobed stigmas with connivent lobes and/or filaments of median stamens united in pairs (Fig. 1). (2) The majority of species are native to Asia, with fewer taxa occurring in Europe and Africa and very few in North America. (3) The group includes the largest nuclear genomes in Brassicaceae, where increases in genome size usually are not associated with neopolyploidy. (4) The vast majority of clade E species has seven chromosome pairs (diploids) or chromosome complements based on $x = 7$ (polyploids). (5) The genome structures described here and the inferred ancestral genome (CEK) point to a monophyletic origin of the clade. (6) Phylogenetic analyses based on nuclear and chloroplast gene markers repeatedly retrieved the *Hesperis* clade as being a monophyletic lineage (Beilstein et al., 2006, 2008, 2010; German et al., 2009, 2011; Couvreur et al., 2010; Warwick et al., 2010; Huang et al., 2016; Guo et al., 2017; this study).

Our chloroplast tree, congruent with Beilstein et al. (2006, 2008, 2010), German et al. (2009), Couvreur et al. (2010), Huang et al. (2016), and Guo et al. (2017), showed that ANAS does not belong to clade E. The distinct phylogenetic history of ANAS also is supported by its base chromosome numbers equal to eight to 13 but not six or seven (BrassiBase; Kiefer et al., 2014) and by the absence of clade E-specific chromosomal rearrangements. Furthermore, this study and Mandáková et al. (2017) revealed that the extant diploid ANAS species represent diploidized mesotetraploid genomes. In contrast, no evidence for a mesopolyploid event in the ancestry of clade E was obtained.

Phylogenomic Evidence of Two Major Intraclade Branches

Within clade E, Huang et al. (2016) retrieved two subclades: the first one containing CHOR and DONT and the second one harboring ANCH, BUNI, EUCL, and HESP. SHEH, formed via an intertribal hybridization between CHOR and DONT (German and Friesen, 2014), should belong to the CHOR/DONT subclade. Species from the CHOR/DONT subclade have simple trichomes and often winged or margined seeds, whereas the larger subclade is characterized predominantly by branched trichomes and wingless seeds. The same dichotomy was retrieved in our chloroplast phylogeny, based however on only 10 chloroplast

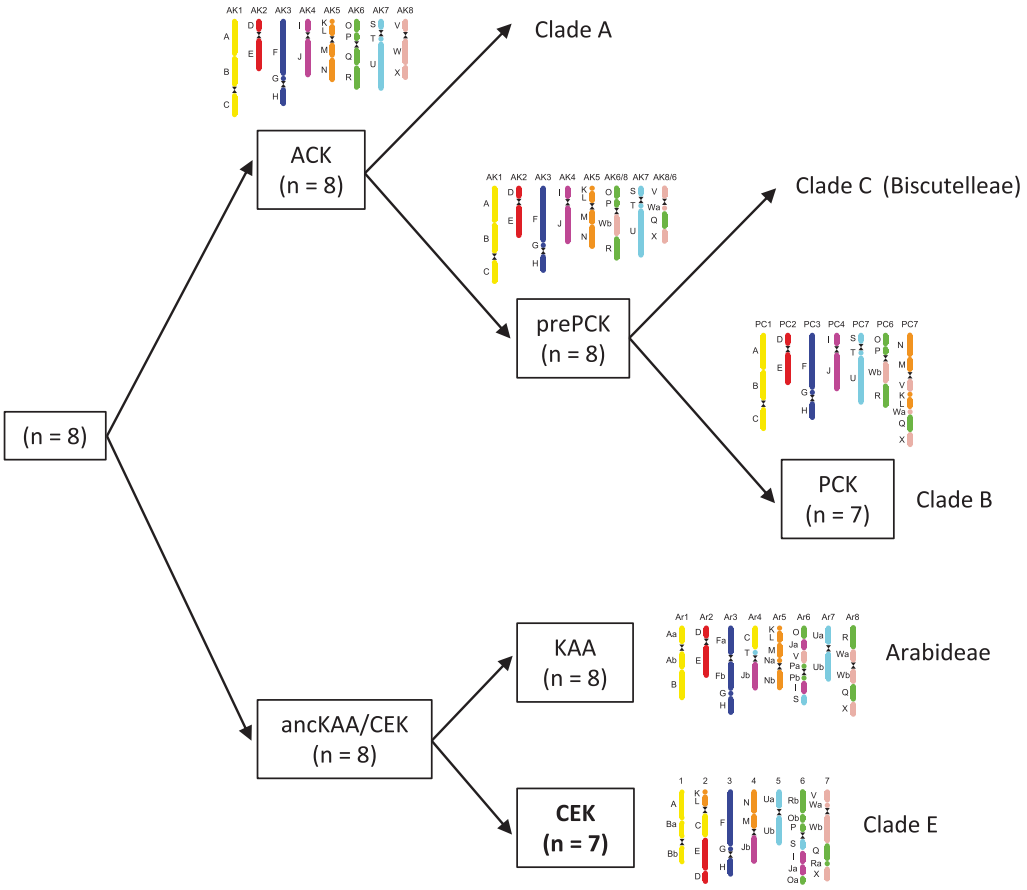


Figure 3. Parsimonious evolutionary scenario of the origin of the CEK ($n = 7$) and KAA ($n = 8$) genomes from evolutionarily older $n = 8$ genome(s). The presumed relationships of these genomes to other inferred ancestral Brassicaceae genomes are outlined.

sequences. Although within clade E, topology differs considerably among authors, this split was often observed, as by Beilstein et al. (2006, 2008, 2010), Franzke et al. (2009), Couvreur et al. (2010), and partly by German et al. (2011). It could be assumed that the tribal dichotomy also is reflected by differences between reconstructed genome structures of CHOR and EUCL (Fig. 2, A and B), whereby the CHOR/DONT subclade would represent more ancestral, CEK-like genomes with a slower rate of karyotype evolution.

Oligocene Origin and Miocene Diversification of the *Hesperis* Clade

Our divergence time estimates based on chloroplast genes dated the origin of the *Hesperis* clade to the Oligocene, and its later diversification occurred throughout the Miocene. These time estimates are largely congruent with the purported Oligocene divergence of major Brassicaceae clades (Huang et al., 2016) as well as with other inferred emergence dates for clade E of 21.4 mya (Couvreur et al., 2010) and 21 mya (maximum stem age; Hohmann et al., 2015). Hohmann et al. (2015) estimated the emergence of clade E tribes at 17 mya, and the same estimate (17.2 mya) for the most recent common ancestor of clade E was reported by Huang et al. (2016). A Middle Miocene divergence (15 mya) also was proposed for the basal split within DONT (between *Clausia* and *Dontostemon*; Friesen et al., 2016). Because the vast majority of Aethionemeae, a sister clade to all other Brassicaceae clades, occurs in the Irano-Turanian region (predominantly in Turkey) and one of the main diversity hotspots of the family is located there as well, this region is often referred to as the cradle of the family (Hedge, 1976; Franzke et al., 2009, 2011; Couvreur

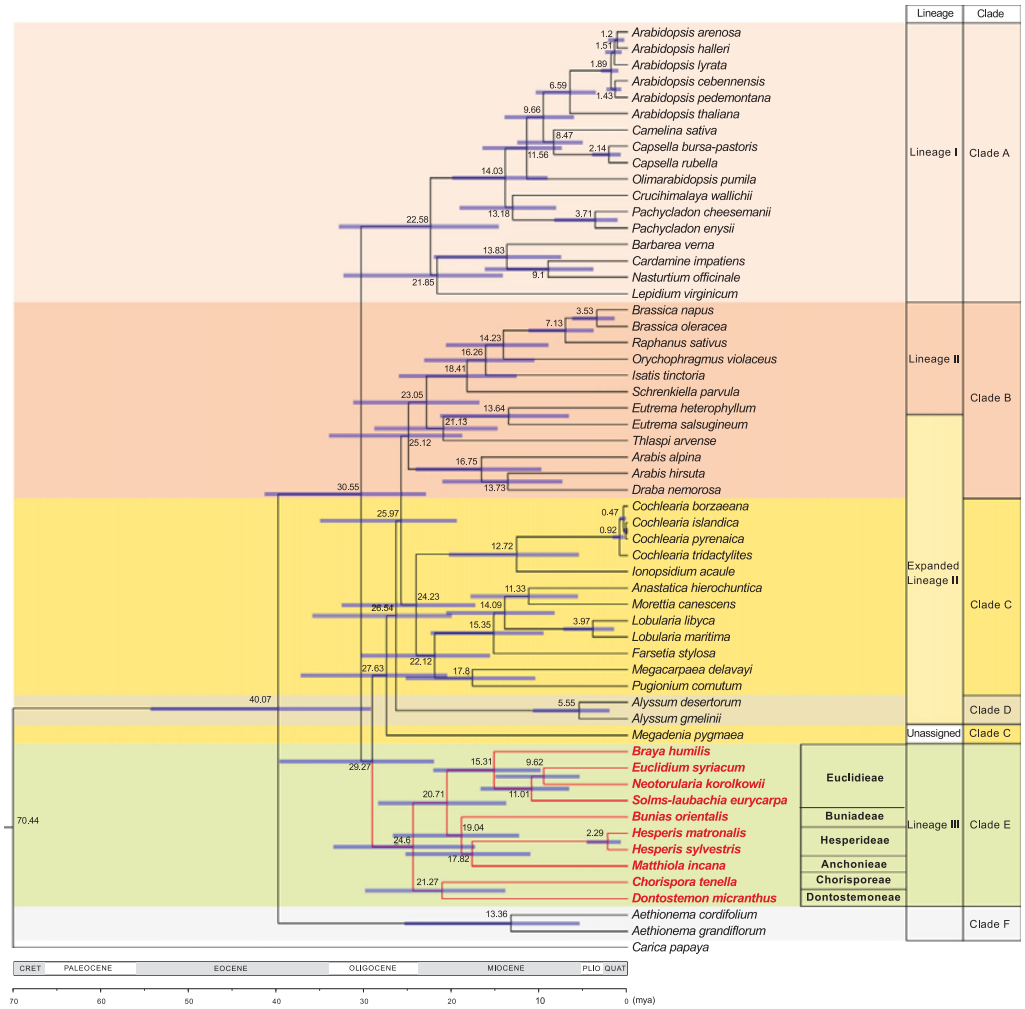


Figure 4. Brassicaceae family tree/chronogram showing the phylogenetic positions and divergence times of clade E tribes. A maximum clade credibility tree was produced by BEAST analysis based on whole-chloroplast sequence data of Brassicaceae taxa. Divergence times based on a relaxed clock log normal model are shown, with blue lines representing 95% high posterior density intervals. Classification to lineages and clades follows Franzke et al. (2011) and Huang et al. (2016), respectively.

et al., 2010). Extant taxa of *Hesperis* clade tribes also occur predominantly in the Ancient Mediterranean floristic subkingdom, especially in the western Asiatic subregion of the Irano-Turanian region (Takhtajan, 1986) or the Irano-Turanian region sensu (Hedge, 1976), which could mean that the emergence of the clade is close to that of the whole family. On the other hand, DONT and some EUCL genera demonstrate diversification in the eastern part of the Irano-Turanian floristic region (central Asiatic subregion sensu [Takhtajan, 1986] or outside the Irano-Turanian floristic region sensu [Hedge, 1976]) and even in mountainous areas of the eastern Asian region, assuming that the already early branching of clade E might be somewhat more eastern than the origin of the whole family. Generally, the increase in open habitats during the Late Oligocene/Early Miocene could have facilitated the diversification of clade E (Franzke et al., 2009). This might be particularly true for the evolution of DONT, apparently driven by the development of the Eurasian steppe belt (Friesen et al., 2016),

enabling the genetic diversification of DONT, associated with a more eastward (mainly central Asian) distribution of the tribe. The diversification of several high-mountain and alpine genera of EUCL (e.g. *Braya*, *Lepidostemon*, *Sisymbriopsis*, and *Solms-laubachia*), also characterized by more eastern centers of present-day diversity, was probably triggered by uplifts of the Qinghai-Tibetan Plateau and mountains of the Hengduan-Himalayan region (Wen et al., 2014, and refs. therein). However, in all cases, current distribution patterns should be interpreted with caution. For example, even for the relatively young (Late Pliocene/Early Pleistocene) *Solms-laubachia*, an ancestral, more western distribution compared with its current center of diversity was detected by Yue et al. (2009).

Clade E and Early Genome Evolution in Brassicaceae

Unlike previous authors, Huang et al. (2016) claimed that clade E branched out early after the split of the Aethionemeae (clade F) from the crown group and that the clade is sister to all the remaining crown-group clades (ABCD). Within our chloroplast tree, interclade relationships are more ambiguous, with clade E being sister to clades B, C, and D and, again, this superclade being sister to clade A. This topology is congruent with the plastome phylogeny of Guo et al. (2017) and to a large extent is supported by the scenario proposed here of an ancient genome evolution (Fig. 3). In accord with Guo et al. (2017), our phylogenetic analysis retrieved clade C paraphyletic due to *Megadenia* being sister to clades B, C, and D.

All the inferred ancestral genomes in Brassicaceae have descended from a common post-At-α genome, which later diversified into an ancestral clade F genome and an ancestral genome (n = 8) shared by all crown-group clades (A–E). The evolution of the latter genome is still rather elusive due to the lack of genomic data on clades C (except for Biscutelleae; Geiser et al., 2016) and D. Comparisons of structurally characterized modern genomes of clades A, B, C, and E plus Arabideae suggest that the ancestral crown-group genome further evolved into ACK (n = 8; Schranz et al., 2006) and another n = 8 genome shared by Arabideae and clade E. ACK either remained conserved in clade A (Lysak et al., 2006, 2016; Mandáková et al., 2013), was altered by a reciprocal translocation in clade C (pre-PCK of Biscutelleae; Geiser et al., 2016), or underwent descending dysploidy toward the PCK genome of clade B (n = 7; Mandáková and Lysak, 2008; Cheng et al., 2013; Mandáková et al., 2015). Although CEK of clade E and KAA of Arabideae (Willing et al., 2015) share some unique genomic features (Fig. 3), this genomic affinity is not corroborated by plastome phylogeny (Guo et al., 2017; Fig. 4), and more work is needed to settle this discrepancy.

Trends of Genome Evolution in Clade E

Based on chromosome counts collated by BrassiBase (Kiefer et al., 2014), 96% to 100% of BUNI, CHOR, DONT, and EUCL species possessed genomes based on x = 7 (usually $2n$ = 14 or 28). In HESP, 57% of available counts corresponded to x = 7, 36% to x = 6, and 1% to x = 5. A very similar pattern was observed in ANCH, with x = 7 in 50% of chromosome counts, 44% corresponding to x = 6, and 4% corresponding to x = 5 (note that 5.3% and 2.5% of $2n$ = 16 counts in HESP and ANCH, respectively, are most likely miscounts of $2n$ = 14). The prevalence of x = 7 across all tribes further justifies CEK as an ancestral genome of the *Hesperis* clade and points to its apparent stasis. This is demonstrated by almost identical genomes of *Euclidium* and *Strigosella*, also suggesting that intratribal diversification was not associated with major chromosomal reshuffling. It remains to be seen whether the strong tendency for descending dysploidy from n = 7 to n = 6 (−5) in *Hesperis* and *Matthiola* could be associated with speciation events in these genera. A comparable karyotype and chromosome number stasis was reported previously for clade B (expanded lineage II; Mandáková and Lysak, 2008), despite containing some 25 tribes (Al-Shehbaz, 2012). Such genomes represent well-tuned genetic systems that have not been affected by major genomic alterations for the last 20 million years. In both clades, the lack of genome repatterning and extensive descending dysploidies also can be attributed to the absence of mesopolyploid whole-genome duplications. Independent polyploidizations frequently triggered major genomic rearrangements and descending dysploidies across the Brassicaceae (Mandáková et al., 2017).

MATERIALS AND METHODS

Plant Material

Plants used for cytogenetic and/or phylogenetic analyses were grown from seeds or collected in the field (for origins, see Supplemental Table S1).

Chromosome Preparation

Inflorescences containing young flower buds were collected into fixative (3:1, 96% ethanol:glacial acetic acid) and kept at −20°C until needed. Mitotic and meiotic chromosome preparations were prepared from anthers as described by Mandáková and Lysak (2016a). Preparations were staged using a phase-contrast microscope, and suitable slides containing tapetal mitoses and/or meiosis I chromosomes were postfixed in 4% formaldehyde in distilled water for 10 min and air dried. Chromosome preparations were treated with 100 μg mL^{-1} RNase in 2× SSC (20× SSC = 3 M sodium chloride and 300 mM trisodium citrate, pH 7) for 60 min and with 0.1 mg mL^{-1} pepsin in 0.01 M HCl at 37°C for 3 to 15 min, then postfixed in 4% formaldehyde in 2× SSC for 10 min, washed in 2× SSC twice for 5 min, and dehydrated in an ethanol series (70%, 90%, and 100%, 2 min each).

CCP

For CCP in CHOR and EUCL species, chromosome-specific BAC clones of Arabidopsis (*Arabidopsis thaliana*) were grouped into contigs according to 22 GBs of ACK (Lysak et al., 2016). To determine and characterize paracentric inversions (see chromosomes Ct2, Es1, Es2, Sa1, Sa1′, Sa2, Sa2′, Sa3, and Sa3′ in Fig. 2, A and B) and splits of block I (see chromosomes Sa3 and Sa3′ in Fig. 2B; Supplemental Fig. S1B), BAC contigs corresponding to GBs A, B, E, and I were subdivided after initial CCP experiments into smaller, differentially labeled contigs. We were not able to detect block T, probably due to its close proximity to (peri)centromeric heterochromatin. Arabidopsis BAC clone T15P10 (AF167571), bearing 35S rRNA gene repeats, was used for in situ localization of nucleolar organizer regions and Arabidopsis clone pCT 4.2 (M65137). All DNA

probes were labeled with biotin-, digoxigenin-, or Cy3-dUTP by nick translation as described by Mandáková and Lysak (2016b). A total of 100 ng of labeled DNA of each selected BAC clone was pooled together, ethanol precipitated, dissolved in 20 μL of hybridization mixture containing 50% formamide and 10% dextran sulfate in 2× SSC, and pipetted onto each chromosome preparation. The slides were heated at 80°C for 2 min and incubated at 37°C overnight. Hybridized probes were visualized either as direct fluorescence of Cy3-dUTP (yellow) or through fluorescently labeled antibodies against biotin-dUTP (red) and digoxigenin-dUTP (green), as described by Mandáková and Lysak (2016b). Chromosomes were counterstained with DAPI (2 μg mL^{-1}) in Vectashield antifade. Fluorescent signals were analyzed and photographed using a Zeiss AxioImager epifluorescence microscope equipped with a CoolCube camera (MetaSystems). Individual images were merged and processed using Photoshop CS software (Adobe Systems). Painted chromosomes in Figure 2C were straightened using the plugin Straighten Curved Objects in ImageJ (Kocsis et al., 1991).

CCP on pachytene chromosomes in ANCH, CHOR, DONT, and HESP species with large genomes resulted in nonspecific hybridization signals covering multiple chromosomes. This was probably caused by a high repeat content of these genomes and/or high levels of chromosome heterochromatinization. However, a modified CCP protocol enabled us to identify common GB associations of CHOR and EUCL on mitotic chromosomes of *Bunias orientalis*, *Dontostemon micranthus*, *Hesperis sylvestris*, and *Matthiola incana*. The combinations of BAC contigs building up chromosomes Ct1/Es1/Sa1/Sa1′, Ct4/Es4/Sa4/Sa4′, and Ct5/Es5/Sa5/Sa5′ in *Chorispora*, *Euclidium*, and *Strigosella*, respectively, were used as painting probes. The following modifications were applied: (1) the concentration of each selected labeled BAC clone was increased 5 times (500 ng per slide); (2) denaturation time was increased (4 min); (3) hybridization times at 37°C were prolonged (68–72 h); and (4) stringent posthybridization washing was prolonged (three washes in 20% formamide in 2× SSC, 10 min each time). After CCP, chromosomes were counterstained with half-concentrated DAPI (1 μg mL^{-1}) in Vectashield antifade.

Chloroplast Genome de Novo Assembly

Leaf material of 12 species, presumably belonging to clade E, and that of *Alyssum gmelinii* (Supplemental Table S1) was harvested and dried using silica gel. For the de novo assembly of chloroplast genomes, reads from low-coverage whole-genome sequencing (Illumina; 2× 100 bp, 2× 350 bp) were used. Chloroplast reads make up to 6% of all reads. Quality filtering (Phred score > 20 and cutoff value of 80%), adaptor trimming, and converting fastq to fasta were performed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

For each species, chloroplast genome reads were identified by BLAST software (Altschul et al., 1990). All raw reads were aligned (using BLASTn) on the reference genome of Arabidopsis (AP000423) and *Lobularia maritima* (NC_009274), and only reads with positive hits were used for de novo assembly. Before de novo assembly, chloroplast reads were down sampled to 100× coverage (i.e. 150,000 100-bp paired-end reads and 45,000 350-bp paired-end reads). De novo assembly was performed by Ray assembler (Boisvert et al., 2010), and sequence gaps in scaffolds were attempted to be filled by Gapfiller (Boetzer and Pirovano, 2012). All contigs were mapped to the reference chloroplast genome of Arabidopsis by Geneious 8.1.7 (Kearse et al., 2012), and all chloroplast reads were then mapped back to the consensus, with sequences being checked manually to remove misalignments and mismatches between the newly assembled and reference genomes.

Genome Annotation

Annotations of all 13 chloroplast genomes were performed on the Dual Organellar GenoMe Annotator (DOGMA; Wyman et al., 2004). Each DOGMA annotation was manually corrected for the start and stop codons or intron/exon junctions by comparison with known homologous chloroplast genes, and tRNA genes were checked by ARAGORN (Laslett and Canback, 2004).

Phylogenetic Analysis

For phylogenetic analysis, we used the alignment published by Hohmann et al. (2015), whole-chloroplast sequences of 14 Brassicaceae species from GenBank, and our 13 newly assembled chloroplast genomes. From assembled and annotated genomes, genes used by Hohmann et al. (2015) were extracted, aligned to the Arabidopsis reference genome, and start/stop codons, introns, and insertions/deletions were removed. Extracted genes were then ordered and added to the nexus of 72 species of Hohmann et al. (2015). Sequence alignments have been deposited in the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.54df2). Divergence time estimation was conducted in BEAST version 2.4.4 (Bouckaert et al., 2014) using independent site and clock models. *Vitis vinifera* was defined as an outgroup. We used four fossil constraints, as used by Magallón et al. (2015) and Hohmann et al. (2015), and a normal distribution was used for these. We ran two independent MCMC runs with 300,000,000 generations each, sampled every 30,000 generations. LogCombiner version 1.8.3 was used to combine trees from the two runs, and 10% of trees were discarded as burn in. TreeAnnotator version 1.8.3 was used to generate a maximum clade credibility tree. All phylogenetic analyses were computed through the Cyberinfrastructure for Phylogenetic Research portal (http://www.phylo.org/; Miller et al., 2010). Visualization was performed in FigTree version 1.4.2 (Rambaut, 2014).

Accession Numbers

Newly assembled chloroplast sequences from this article can be found in GenBank under accession numbers KY912021 to KY912032 and MF169880 (*A. gmelinii*).

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Parsimoniously reconstructed origins of chromosomes in clade E species.

Supplemental Table S1. Collection data for the species used in this study.

ACKNOWLEDGMENTS

We thank the core Global Naturalized Alien Flora project members for providing data on naturalized species of clade E and Kateřina Beková for help with analyzing sequence data. Several of our colleagues (see Fig. 1 legend) are acknowledged for giving us permission to reproduce their photographs of clade E species.

Received April 3, 2017; accepted June 26, 2017; published June 30, 2017.

LITERATURE CITED

Al-Shehbaz IA (2012) A generic and tribal synopsis of the Brassicaceae (Cruciferae). Taxon **61:** 931–954

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

Beilstein MA, Al-Shehbaz IA, Kellogg EA (2006) Brassicaceae phylogeny and trichome evolution. Am J Bot **93:** 607–619

Beilstein MA, Al-Shehbaz IA, Mathews S, Kellogg EA (2008) Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: tribes and trichomes revisited. Am J Bot **95:** 1307–1327

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc Natl Acad Sci USA **107:** 18724–18728

Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. Genome Biol **13:** R56

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol **17:** 1519–1533

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLOS Comput Biol **10:** e1003537

CABI (2012) *Bunias orientalis* (Turkish warty-cabbage). Datasheet Invasive Species Compendium. http://www.cabi.org/isc/datasheet/109130

Cheng F, Mandáková T, Wu J, Xie Q, Lysak MA, Wang X (2013) Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. Plant Cell **25:** 1541–1554

Couvreur TLP, Franzke A, Al-Shehbaz IA, Bakker FT, Koch MA, Mummenhoff K (2010) Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol Biol Evol **27:** 55–71

Francis A, Cavers PB, Warwick SI (2009) The biology of Canadian weeds, 140: *Hesperis matronalis* L. Can J Plant Sci **89:** 189–204
Franzke A, German D, Al-Shehbaz IA, Mummenhoff K (2009) Arabidopsis family ties: molecular phylogeny and age estimates in Brassicaceae. Taxon **58:** 425–437
Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K (2011) Cabbage family affairs: the evolutionary history of Brassicaceae. Trends Plant Sci **16:** 108–116
Friesen N, German DA, Hurka H, Herden T, Oyuntsetseg B, Neuffer B (2016) Dated phylogenies and historical biogeography of Dontostemon and Clausia (Brassicaceae) mirror the palaeogeographic history of the Eurasian steppe. J Biogeogr **43:** 738–749
Geiser C, Mandáková T, Arrigo N, Lysak MA, Parisod C (2016) Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in Buckler mustards. Plant Cell **28:** 17–27
German DA, Friesen N, Neuffer B, Al-Shehbaz IA, Hurka H (2009) Contribution to ITS phylogeny of the Brassicaceae, with a special reference to some Asian taxa. Plant Syst Evol **283:** 33–56
German DA, Friesen NW (2014) *Shehbazia* (Shehbazieae, Cruciferae), a new monotypic genus and tribe of hybrid origin from Tibet. Turczaninowia **17:** 17–23
German DA, Grant JR, Lysak MA, Al-Shehbaz IA (2011) Molecular phylogeny and systematics of the tribe Chorisporeae (Brassicaceae). Plant Syst Evol **294:** 65–86
Guo X, Liu J, Hao G, Zhang L, Mao K, Wang X, Zhang D, Ma T, Hu Q, Al-Shehbaz IA, et al (2017) Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics **18:** 176
Hedge IC (1976) A systematic and geographical survey of the Old World Cruciferae. *In* JG Vaughan, AJ Macleod, BMG Jones, eds, The Biology and Chemistry of the Cruciferae. Academic Press, London, pp 1–45
Hohmann N, Wolf EM, Lysak MA, Koch MA (2015) A time-calibrated road map of Brassicaceae species radiation and evolutionary history. Plant Cell **27:** 2770–2784
Huang CH, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-Shehbaz I, Edger PP, et al (2016) Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol Biol Evol **33:** 394–412
Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28:** 1647–1649
Kiefer M, Schmickl R, German DA, Mandáková T, Lysak MA, Al-Shehbaz IA, Franzke A, Mummenhoff K, Stamatakis A, Koch MA (2014) BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. Plant Cell Physiol **55:** e3
Kocsis E, Trus BL, Steer CJ, Bisher ME, Steven AC (1991) Image averaging of flexible fibrous macromolecules: the clathrin triskelion has an elastic proximal segment. J Struct Biol **107:** 6–14
Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res **32:** 11–16
Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, Schubert I (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. Proc Natl Acad Sci USA **103:** 5224–5229
Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. Mol Biol Evol **26:** 85–98
Lysak MA, Mandáková T, Schranz ME (2016) Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. Curr Opin Plant Biol **30:** 108–115
Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T (2015) A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol **207:** 437–453
Mandáková T, Kovarík A, Zozomová-Lihová J, Shimizu-Inatsugi R, Shimizu KK, Mummenhoff K, Marhold K, Lysak MA (2013) The more the merrier: recent hybridization and polyploidy in *Cardamine*. Plant Cell **25:** 3280–3295
Mandáková T, Li Z, Barker MS, Lysak MA (2017) Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J **91:** 3–21
Mandáková T, Lysak MA (2008) Chromosomal phylogeny and karyotype evolution in x=7 crucifer species (Brassicaceae). Plant Cell **20:** 2559–2570
Mandáková T, Lysak MA (2016a) Chromosome preparation for cytogenetic analyses in *Arabidopsis*. Curr Protoc Plant Biol **1:** 1–9
Mandáková T, Lysak MA (2016b) Painting of *Arabidopsis* chromosomes with chromosome-specific BAC clones. Curr Protoc Plant Biol **1:** 359–371
Mandáková T, Singh V, Krämer U, Lysak MA (2015) Genome structure of the heavy metal hyperaccumulator *Noccaea caerulescens* and its stability on metalliferous and nonmetalliferous soils. Plant Physiol **169:** 674–689
Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *In* Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, LA, November 14, 2010, 1–8
Rambaut A (2014) FigTree version 1.4.2. http://tree.bio.ed.ac.uk/software/figtree/
Schranz ME, Lysak MA, Mitchell-Olds T (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci **11:** 535–542
Takhtajan AL (1986) Floristic Regions of the World. University of California Press, Berkeley
van Kleunen M, Dawson W, Essl F, Pergl J, Winter M, Weber E, Kreft H, Weigelt P, Kartesz J, Nishino M, et al (2015) Global exchange and accumulation of non-native plants. Nature **525:** 100–103
Warwick SI, Mummenhoff K, Sauder CA, Koch MA, Al-Shehbaz IA (2010) Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. Plant Syst Evol **285:** 209–232
Wen J, Zhang JQ, Nie ZL, Zhong Y, Sun H (2014) Evolutionary diversification of plants on the Qinghai-Tibetan Plateau. Front Genet **5:** 4
Willing E-M, Rawat V, Mandáková T, James GV, Nordström KJ, Maumus F, Becker C, Warthmann N, Chica C, Szarzynska B, et al (2015) Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. Nat Plants **1:** 14023
Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20:** 3252–3255
Yue JP, Sun H, Baum DA, Li JH, Al-Shehbaz IA, Ree R (2009) Molecular phylogeny of *Solms-laubachia* (Brassicaceae) s.l., based on multiple nuclear and plastid DNA sequences, and its biogeographic implications. J Syst Evol **47:** 402–415

Supplemental Material online

Mandáková et al.:
Monophyletic origin and evolution of the largest crucifer genomes

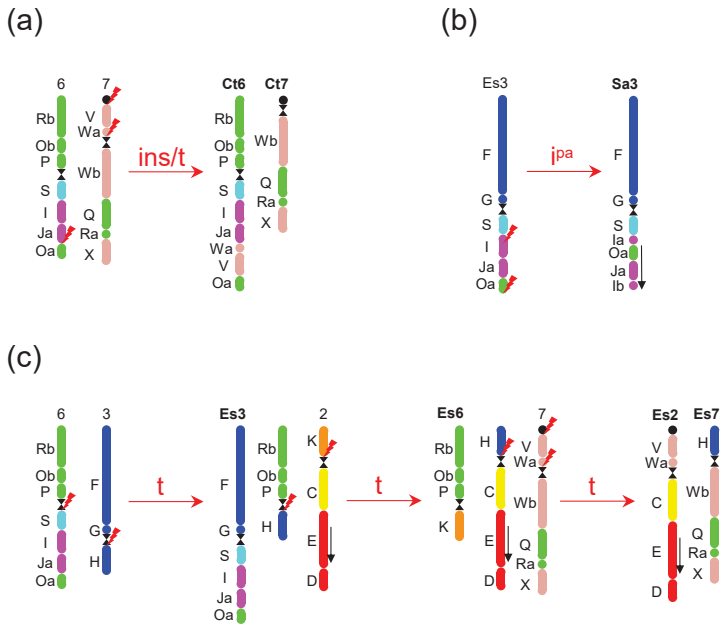


Figure S1. Parsimoniously reconstructed origins of chromosomes in Clade E species. (a) Origin of chromosomes Ct6 and Ct7 in *Chorispora tenella* (CHOR) from two chromosomes of CEK. (b) Reshuffling of homeologue 3 in EUCL. Chromosome Sa3 in *Strigosella africana* was altered by a paracentric inversion as compared to an evolutionary older chromosome Es3 in *Euclidium syriacum*. (c) Origin of chromosomes Es2, Es3, Es6 and Es7 in *E. syriacum* from ancestral chromosomes 2, 3, 6 and 7 of CEK. Capital letters and color coding correspond to seven chromosomes and 21 genomic blocks of CEK (Fig. 2d; NB. the position of the 22nd block, T, was not determined). ins: insertion, t: translocation, ipa: paracentric inversion. Black solid circles mark the position of 35S rDNA loci.

Table S1. Collection data for the species used in the present study.

Species	Tribe	Chloroplast phylogeny	Comparative chromosome painting	Collector/accession no.	Origin
Anastatica hierochuntica L.	Anastaticeae	+		MSB 0210409	Egypt, Western Desert
Farsetia stylosa R. Br.	Anastaticeae	+	+	MSB 0085243	Mauritania
Lobularia libyca (Viv.) Webb & Berthel.	Anastaticeae	+	+	Botanical Garden Denmark	
Morettia canescens Boiss.	Anastaticeae	+	+	MSB 0089942	Marocco, Ouarzazate
Matthiola incana W. T. Aiton	Anchonieae	+	+	AEV-VS company	
Bunias orientalis L.	Buniadeae	+	+	T. Mandáková	Czech Republic, Brno, Dvorakova str.
Dontostemon micranthus C. A. Mey.	Dontostemoneae	+	+	D. A. German et al.	Russia, Altai Republic, Ongudai district, right bank of Chuya between mouth and Iodro, 50°24' N, 86°50' E, 31 July 2004
Euclidium syriacum (L.) W. T. Aiton	Euclidieae	+	+	A. Pečinka	Ladakh, in the Mulbekh village
Braya humilis (C. A. Mey.) B. L. Rob.	Euclidieae	+		Rocky Mountain Rare Plants (2521)	
Strigosella africana (L.) Botsch.	Euclidieae		+	D. A. German et al.	China, Xinjiang, Altai, right bank of Kran river, 47°50' N, 88°12' E, 900 m a.s.l, rocky steppe slopes, 22 May 2004
Hesperis matronalis L.	Hesperideae	+		Botanical Garden Brno	
Hesperis sylvestris Crantz	Hesperideae	+	+	T. Mandáková	Czech Republic, South Moravia, Palava Hills
Chorispora tenella (Pall.) DC.	Chorisporeae	+	+	Botanical Garden Berlin-Dahlem (1382)	Macedonia, Pangaion Hills
Alyssum gmelinii Jord.	Alysseae	+		T. Mandáková	Czech Republic, South Moravia, Palava Hills

MSB - Millenium Seed Bank, Royal Botanic Gardens, Kew

3.2 THE LARGE GENOME SIZE VARIATION IN THE *HESPERIS* CLADE WAS SHAPED BY THE PREVALENT PROLIFERATION OF DNA REPEATS AND RARER GENOME DOWNSIZING

Hloušková P, Mandáková T, Pouch M, Trávnícek P, Lysak MA. (2019). The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing. Annals of Botany. 124(1), 103–120.

PH performed ancestral genome size reconstruction, all presented statistical analyses, repetitive DNA analysis including the qualitative and quantitative characterization, participated in designing oligoprobes and primers for cytogenetic experiments. PH interpreted data, made the story and wrote, reviewed and edited the manuscript.

Summary

The crucifer species with the largest genomes occur within the monophyletic *Hesperis* clade (Clade E or Lineage III), monoploid genome sizes vary 16-fold (254–4264 Mb) in analyzed species. Whereas most chromosome numbers in the clade are n = 6 or 7.

Genome sizes of analyzed species were estimated by flow cytometry. The evolution of genome size in the *Hesperis* clade was simulated on ITS and *ndhF* phylogenies for all species with known genome size. In the ITS phylogeny, ancestral genome size of the *Hesperis* clade was estimated as 1,790 Mb and a similar value (1,524 Mb) was inferred based on the *ndhF* tree. The most recent common ancestor of the *Hesperis* clade has experienced genome upsizing due to TE amplification. Genome size variation was found to be correlated with life histories. The *Hesperis*-clade species show statistically significant tendency of annual species to have smaller genomes, than biennial or perennial ones. The median genome size was significantly lower in annuals than in perennials.

Low-coverage genome sequencing data of seven species (*Braya humilis*, *Bunias orientalis*, *Chorispora tenella*, *Dontostemon micranthus*, *Euclidium syriacum*, *Hesperis sylvestris* and *Matthiola incana*) were used for identification and quantification of repeat sequences employing RepeatExplorer pipeline. We identified main types of repetitive sequences and estimated their genome proportions in all the seven species analyzed. In all genomes, LTR retroelements made up the majority of repeatomes, mainly represented by the Athila clade of the Ty3/*gypsy* superfamily. Tandem repeats were found in different abundances, from a very low (in *Br. humilis* and *D. micranthus*) to a high genome proportion in *H. sylvestris*. The identified tandem repeat monomers were variable in length among the seven species, ranging from 20 to 825 bp. Overall, a strong positive correlation was found between the repeat content and genome size. The most abundant repeats were localized on chromosomes by fluorescence *in situ* hybridization. We compared chromosomal organization in small versus large crucifer genomes, to challenge the stereotype of the arabidopsis-type chromosomal organization being universal for all crucifer taxa. The amplification of TEs and tandem repeats impacted the chromosomal architecture of the *Hesperis* clade species. The arabidopsis-like chromosomal architecture is characteristic for species with smallest genome (*C. tenella* and *E. syriacum*). With increasing nuclear genome size, the TE repeat content increases and the arabidopsis-like chromosomal architecture disappears. In larger genomes (>1,500 Mb) TEs are equally distributed along the entire chromosome length, except for distinct subtelomeric and pericentromeric loci occupied by tandem repeats.

Annals of Botany **XX**: 1–17, 2019
doi: 10.1093/aob/mcz036, available online at www.academic.oup.com/aob

ANNALS OF BOTANY Founded 1887

The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing

Petra Hloušková[1], Terezie Mandáková[1], Milan Pouch[1], Pavel Trávníček[2] and Martin A. Lysak[1,*]

[1]CEITEC - Central European Institute of Technology, and Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic and [2]Institute of Botany, Czech Academy of Sciences, Zámek 1, 252 43 Průhonice, Czech Republic
**For correspondence. E-mail martin.lysak@ceitec.muni.cz*

Received: 13 December 2018 Returned for revision: 10 January 2018 Editorial decision: 22 February 2019 Accepted: 28 February 2019

• **Background and Aims** Most crucifer species (Brassicaceae) have small nuclear genomes (mean 1C-value 617 Mb). The species with the largest genomes occur within the monophyletic *Hesperis* clade (Mandáková *et al.*, *Plant Physiology* **174**: 2062–2071; also known as Clade E or Lineage III). Whereas most chromosome numbers in the clade are 6 or 7, monoploid genome sizes vary 16-fold (256–4264 Mb). To get an insight into genome size evolution in the *Hesperis* clade (~350 species in ~48 genera), we aimed to identify, quantify and localize *in situ* the repeats from which these genomes are built. We analysed nuclear repeatomes in seven species, covering the phylogenetic and genome size breadth of the clade, by low-pass whole-genome sequencing.
• **Methods** Genome size was estimated by flow cytometry. Genomic DNA was sequenced on an Illumina sequencer and DNA repeats were identified and quantified using RepeatExplorer; the most abundant repeats were localized on chromosomes by fluorescence *in situ* hybridization. To evaluate the feasibility of bacterial artificial chromosome (BAC)-based comparative chromosome painting in *Hesperis*-clade species, BACs of arabidopsis were used as painting probes.
• **Key Results** Most biennial and perennial species of the *Hesperis* clade possess unusually large nuclear genomes due to the proliferation of long terminal repeat retrotransposons. The prevalent genome expansion was rarely, but repeatedly, counteracted by purging of transposable elements in ephemeral and annual species.
• **Conclusions** The most common ancestor of the *Hesperis* clade has experienced genome upsizing due to transposable element amplification. Further genome size increases, dominating diversification of all *Hesperis*-clade tribes, contrast with the overall stability of chromosome numbers. In some subclades and species genome downsizing occurred, presumably as an adaptive transition to an annual life cycle. The amplification versus purging of transposable elements and tandem repeats impacted the chromosomal architecture of the *Hesperis*-clade species.

Key words: Genome size evolution, repetitive DNA, tandem repeats, retrotransposons, interstitial telomeric repeats (ITRs), chromosome organization, *Bunias*, *Hesperis*, *Matthiola*, Lineage III, Brassicaceae.

INTRODUCTION

Angiosperms, flowering plants, exhibit 2440-fold variation in nuclear genome size. The smallest genome has only ~60 Mb, whereas the size of the largest angiosperm genome is almost 150 000 Mb and the mean and modal genome size equals 5020 and 587 Mb, respectively (Pellicer *et al.*, 2018). Nuclear genomes expand as the consequence of whole-genome duplications (polyploidy) and due to the accumulation of transposable elements (TEs) and tandem repeats (e.g. Kubis *et al.*, 1998; Macas *et al.*, 2015; Willing *et al.*, 2015; Gaiero *et al.*, 2018; Pellicer *et al.*, 2018). Genome expansion is counterbalanced by deletion-biased double-strand break repair, including transposon excision and homologous and illegitimate recombination (e.g. Devos *et al.*, 2002; Hawkins *et al.*, 2009; Waterworth *et al.*, 2011; Vu *et al.*, 2017). Large chromosome regions can be lost as the consequence of chromosomal rearrangements, such as deletions and translocations (Schubert and Lysak, 2011), and inversions moving inverted regions to more proximal chromosomal positions can increase the elimination of repetitive sequences due to higher illegitimate recombination rates in these regions (Ren *et al.*, 2018). As genome expansion and downsizing mechanisms can be (in)active to strikingly different extents, huge genome and chromosome size variation can be encountered even in plant groups with overall constant chromosome numbers, such as grasses and the Pinaceae (Heslop-Harrison and Schwarzacher, 2011).

In comparison with the 2440-fold variation across all angiosperms, genome sizes of crucifer species (the mustard family or Brassicaceae) vary only by 52-fold (from 157 Mb in *Arabidopsis thaliana* to 8117 Mb in the tetraploid *Hesperis matronalis*; Bennett *et al.*, 2003; Kiefer *et al.*, 2014; https://brassibase.cos.uni-heidelberg.de), with most species having a small genome size (mean and modal genome size is 617 and 392 Mb, respectively; Lysak *et al.*, 2009). In fact, a crucifer species, namely arabidopsis (*A. thaliana*) was considered to have the smallest genome (157 Mb; Bennett *et al.*, 2003) among flowering plants until its special position was replaced by the extremely small genomes (~60 Mb) of the bladderwort family (Lentibulariaceae; Greilhuber *et al.*, 2006).

© The Author(s) 2019. Published by Oxford University Press on behalf of the Annals of Botany Company.
All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

When analysing genome size variation across 3977 crucifer species classified in 341 genera and 52 tribes (Kiefer *et al.*, 2014; https://brassibase.cos.uni-heidelberg.de), it becomes evident that the variation is not equally distributed among the tribes and six or so super-tribes, i.e. lineages or clades (Beilstein *et al.*, 2006; Huang *et al.*, 2016). With some rare exceptions, crucifer species with very large as well as the largest genome sizes (Lysak *et al.*, 2009) belong to the *Hesperis* clade (Mandáková *et al.*, 2017), also known as Lineage III (Beilstein *et al.*, 2006) or Clade E (Huang *et al.*, 2016). The monophyletic *Hesperis* clade comprises seven tribes harbouring ~350 species classified in ~48 genera (Mandáková *et al.*, 2017; but see Chen *et al.*, 2018; German and Al-Shehbaz, 2017, 2018 for recent taxonomic reappraisals in the clade). Among the several crucifer super-tribes, the *Hesperis* clade not only contains the largest genomes, but also exhibits the broadest range of genome sizes. Holoploid genome size varies by >30-fold, ranging from 265 Mb in *Diptychocarpus strictus* and *Euclidium syriacum* (Kiefer *et al.*, 2014; https://brassibase.cos.uni-heidelberg.de; this study) to 8117 Mb in the tetraploid *Hesperis matronalis* (Kiefer *et al.*, 2014; https://brassibase.cos.uni-heidelberg.de). Monoploid genome size varies 16.8-fold, ranging from 265 to 4273 Mb in *H. sylvestris* (this study). Interestingly, the extensive genome size variation contrasts with the evolutionary stability of chromosome numbers, with most species having rather low chromosome numbers ($n = 6$ or $n = 7$) (Mandáková *et al.*, 2017). As noted by early scholars (Jaretzky, 1928; Manton, 1932), few chromosomes accommodating a large nuclear genome make the chromosomes of the *Hesperis*-clade species some of the largest chromosomes in the Brassicaceae.

In the present study, we aimed to analyse repeatomes of selected *Hesperis*-clade species to get a deeper insight into processes underlying genome size variation across the clade. To this end, we carried out low-pass Illumina sequencing of genomic DNA in seven diploid species representing six tribes as well as the 16-fold genome size variation within the *Hesperis* clade. In the context of gene-based phylogenetic hypotheses, our objective was to elucidate the directionality of repeatome evolution in the clade; in particular we aimed to analyse why genome obesity is not a universal feature of all species belonging to the apparently monophyletic super-tribe (Mandáková *et al.*, 2017). In the case of large-genome species, we asked whether these genomes were inflated by only a few abundant repeats amplified to high copy numbers or due to the proliferation of many repeat types with fewer genomic copies. Finally, yet importantly, we aimed to compare chromosomal organization in small versus large crucifer genomes, to challenge the stereotype of the arabidopsis-type chromosomal organization being universal for all crucifer taxa.

MATERIALS AND METHODS

Plant material

Plants used in this study were grown from seeds or collected in the field (for the origins see Mandáková *et al.*, 2017). Genomic DNA was extracted from fresh or silica-dried leaves using the NucleoSpin Plant II kit (Macherey-Nagel). Young inflorescences from several plants of the analysed species were collected and fixed in freshly prepared fixative (ethanol:acetic acid, 3:1) overnight, transferred to 70 % ethanol and stored at −20 °C until further use.

Genome size measurements

Holoploid genome sizes were estimated by flow cytometry. For each species, preferentially two intact petals or one young, intact leaf, ~1 cm in length, was prepared according to the two-step procedure of Otto (1990) in a simplified version (Doležel *et al.*, 2007). The samples were stained (solution containing propidium iodide + RNAase IIA, both at final concentrations of 50 µg mL^{-1}) for 5 min at room temperature and analysed using a CyFlow cytometer (Partec) equipped with a 532 nm diode-pumped solid-state laser (Cobolt Samba; Cobolt). A fluorescence intensity of 5000 particles was recorded. *Pisum sativum* 'Ctirad' (1C = 4.38 pg; Trávníček *et al.*, 2015) served as the primary reference standard and *Solanum pseudocapsicum* as the secondary standard (1C = 1.29 pg recalculated against the primary reference). One individual of each species measured on three consecutive days was used for genome size estimation.

Low-pass genome sequencing

Genome sequencing of five species (*Braya humilis*, *Bunias orientalis*, *Chorispora tenella*, *Dontostemon micranthus*, *Euclidium syriacum*), generating 100-bp paired-end reads, was performed on an Illumina HiSeq 2000 platform at GATC Biotech (Konstanz, Germany), and genomes of two species (*Hesperis sylvestris* and *Matthiola incana*) were sequenced using an Illumina MiSeq, paired 300-bp reads, and MiSeq v3 reagents, at the sequencing core facility of the Oklahoma Medical Research Foundation (Oklahoma City, USA).

Phylogenetic analysis and ancestral genome size reconstruction

Internal transcribed spacer (ITS) sequences were obtained from BrassiBase (Kiefer *et al.*, 2014; https://brassibase.cos.uni-heidelberg.de/) and *ndh*F sequences from NCBI GenBank (www.ncbi.nlm.nih.gov). Nucleotide sequences were aligned and manually checked using Geneious v11.1.5 (https://www.geneious.com; Kearse *et al.*, 2012). Only sequences of *Hesperis*-clade species with known genome sizes were used for further phylogenetic analyses and reconstruction of genome size evolution. C-values of *Hesperis*-clade species were either estimated in the present study or adopted from Greilhuber and Obermayer (1999), Suda *et al.* (2005), Lysak *et al.* (2009), Kubešová *et al.* (2010) and BrassiBase (Kiefer *et al.*, 2014; https://brassibase.cos.uni-heidelberg.de/) (Supplementary Data Table S1).

Phylogenetic unrooted trees for ITS and *ndh*F datasets were reconstructed using MrBayes v3.2.6 (Ronquist *et al.*, 2012). In all Bayesian analyses, starting trees were random, four simultaneous Markov chains were run for 5 000 000 generations, burn-in values were set at 500 000 and trees were sampled every 5000 generations. Bayesian posterior probabilities were calculated using a Markov chain Monte Carlo sampling approach. The

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

50 % majority rule was used for constructing consensus trees. All parameters were inspected with Tracer v.1.6 (Rambaut and Drummond, 2009).

The R package GEIGER (Harmon *et al.*, 2007) was used to estimate Pagel's λ, measuring phylogenetic dependence of the observed trait, i.e. genome size. A λ value equal or close to 1 suggests trait evolution according a Brownian motion model. As λ values were close to 1 (0.96) for both datasets we used a Brownian motion model for further analyses.

Ancestral genome sizes were reconstructed for each node using the function ace in the R package APE (Paradis *et al.*, 2004) using the Brownian motion-based maximum likelihood. The reconstructions were subsequently mapped onto the Bayesian phylograms using the function contMap in the package phytools (Revell, 2012).

Genome size and life forms

Information on life forms was obtained from Hohmann *et al.* (2015). *Hesperis*-clade species with known genome sizes were divided into two categories based on their life forms (annuals versus biennials and perennials). The Shapiro–Wilk normality test showed that the genome size values did not have a normal distribution. Thus, we used an unpaired two-sided Mann–Whitney test to find whether genome size differs significantly in annuals versus biennials and perennials. To test the correlation between genome size and life form we performed Spearman's rank correlation test.

Data pre-processing and de novo *identification of repetitive sequences*

A quality check of paired-end reads was carried out using FastQC (Andrews, 2010). Raw sequencing data pre-processing was done before clustering analysis. Removal of reads with similarity to the PhiX was done using our custom-made script. Read-quality filtering (Phred score >20 and cutoff value 80 %), adapter trimming (removal of adapter-containing reads) and conversion of fastq to fasta were performed using the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) implemented within the Galaxy environment (Afgan *et al.*, 2018). MiSeq reads were trimmed to 100 bp.

Repeat identification by similarity-based clustering of reads was performed using local installation of the RepeatExplorer pipeline (Novák *et al.*, 2013) using (1) the maximum number of reads possible, and (2) the number of reads representing 0.05× genome coverage. Each species was analysed separately. The settings for each analysis were left at the default with the minimum overlap length for clustering set as 55 %, and the minimal overlap for assembly set as 40 %. Repeat clusters with genome proportions >0.01 % were annotated in detail. Both genome coverages were analysed with two or three replicates.

The detailed repeatome analysis was based on clustering with maximum reads as we aimed to capture all repetitive sequences responsible for genome size variation; a higher genome coverage (at least 0.01×) has to be used to estimate abundance of repeats with low(er) genome proportions (http://repeatexplorer.org/?page_id=179).

Clusters with known protein domains were classified by the RepeatExplorer pipeline directly. Other clusters were further analysed using similarity search tool BLAST (Altschul *et al.*, 1990) against GenBank nucleotide and protein databases, and the software tool CENSOR (Kohany *et al.*, 2006), which screens query sequences against a Viridiplantae reference database of repeats. Contigs of clusters classified as putative satellites were manually inspected and analysed using Tandem Repeat Finder (TRF; Benson, 1999) and Dotter (Sonnhammer and Durbin, 1995). Reconstruction of consensus monomer sequences of satellites was performed using the tandem repeat analyser TAREAN (Novák *et al.*, 2017) pipeline; interlaced paired-end reads of individual species were used as inputs. TAREAN's advanced option Perform cluster merging was used to merge clusters connected through paired-end reads. TAREAN is available as part of the RepeatExplorer2 pipeline (https://repeatexplorer-elixir.cerit-sc.cz/galaxy/).

Up to 14 % chloroplast DNA (cpDNA) was found in cluster analysis. It has been reported that cpDNA could be found incorporated in the nuclear genome (Roark *et al.*, 2010). However, the significantly high proportion of cpDNA and high similarity to cpDNA of other crucifer species, verified by BLASTN to the NCBI nucleotide database, suggested that it might have come from the DNA extraction process, and thus we excluded cpDNA clusters from our analyses.

Cluster analysis of *H. sylvestris* data using the maximum number of reads resulted in an error due to high computation demands. Therefore, we used automatic filtering of abundant satellite repeats option as this automatic filtering tries to identify the most abundant tandem repeats and removes such sequences partially (10 % left) from analysis. Removal of abundant tandem repeats enabled us to analyse less abundant repeats and a higher number of reads in total. The modified clustering parameters helped to identify additional copies of TEs, particularly LTR retrotransposons (*Ty3-gypsy*/Athila and *Ty1-copia*/Ale elements).

Additionally, *E. syriacum* sequence data from the study by Jiao *et al.* (2017) were downloaded from the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra). SRA archives (ERR1773712 to ERR1773714) were converted into fastq files with fastq-dump from SRA Toolkit v2.4.2. These data were submitted to the TAREAN pipeline (Novák *et al.*, 2017). The assembled *E. syriacum* genome (Jiao *et al.*, 2017) was downloaded from the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena), BioProject ID PRJEB16743. Sequence contigs were analysed using TRF (Benson, 1999). Satellite monomers obtained from the two *E. syriacum* datasets were compared and mapped to *E. syriacum* contigs by BLASTN (e-value $1e^{-3}$, identity >70 %). The Integrative Genomics Viewer (Robinson *et al.*, 2011) was used to visualize satellite localization on assembled scaffolds (data not shown).

A comparative analysis of repetitive sequences of *Hesperis*-clade species was done on pooled reads of all species sampled to 0.01× genome coverage. The settings for the comparative analysis were the same as those for the individual species cluster analyses.

Correlation between genome size and repeat content

To test whether there were correlations between the amounts of different types of repeats with genome size variation in the *Hesperis* clade, we used the function lm for linear regression

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

in the package stats in R software (R Development Core Team, 2013) using absolute amounts of repeats estimated for individual species.

Construction of phylogenetic tree for TE reverse transcriptase domains

Protein domain finder tool embedded in RepeatExplorer Galaxy platform (https://repeatexplorer-elixir.cerit-sc.cz/galaxy?tool_id=domains_finder&version=1.0.0) was used to find and classify all TE protein domains in concatenated contigs from the individual cluster analyses (contigs from individual species were distinguished by sample code). This tool uses the external aligning program LAST (Kiełbasa *et al.*, 2011) and the RepeatExplorer database of TE protein domains (Viridiplantae). The Protein domain filter tool was then applied to filter out only contigs with reverse transcriptase (RT) domains. Default alignment quality criteria were used: minimum identity 35 %, minimum similarity 45 % and minimum alignment length 80 %. To extract protein sequences of RT domains, the Protein domain search tool was used. A database of protein domains derived from plant mobile elements is used in this tool for a similarity search using the fasty36 program (Pearson *et al.*, 1997). Raw fasty36 output was filtered for minimal quality of alignment. The output consisted of protein sequences translated from query DNA and best matching sequences from the protein database. Two output datasets were created according to LTR retroelement superfamilies, for *Ty1-copia*-related sequences and for *Ty3-gypsy* sequences. Multialignment of protein domains was done in MAFFT v7.017 (Katoh and Standley, 2013) in Geneious v11.1.5 (https://www.geneious.com; Kearse *et al.*, 2012) and manually checked. The phylogenetic trees were built using a Bayesian methods algorithm by MrBayes 3.2.6 (Ronquist *et al.*, 2012); the number of generations was set to 5 000 000 and burn-in values were set at 500 000. Parameter values of each run were checked using Tracer v.1.6 (Rambaut and Drummond, 2009).

Identification of shared tandem repeats

Putative satellite sequences from all species were compared with each other by BLASTN (*e*-value $1e^{-3}$, identity >70 %) to assess their sequence similarity. BLAST searching against the GenBank nucleotide database of each satellite was done to investigate whether they showed similarity hits to already known satellite sequences from Brassicaceae species.

Chromosome preparations

Chromosome spreads from fixed young flower buds containing immature anthers were prepared according to published protocols (Mandáková and Lysak, 2016*a*). Briefly, selected flower buds were rinsed in distilled water and citrate buffer, and digested in 0.3 % cellulase, cytohelicase and pectolyase (all from Sigma–Aldrich) in citrate buffer at 37 °C for 3 h. After digestion, individual anthers were dissected and spread in 20 µL of 60 % acetic acid on a microscope slide placed on a metal hot plate (50 °C) for ~30 s. The preparation was then fixed in freshly prepared fixative (ethanol:acetic acid, 3:1) by dropping the fixative around the remaining drop of acetic acid and into it. Chromosome spreads were dried using a hair dryer, post-fixed in freshly prepared 4 % formaldehyde in distilled water and air-dried. Preparations were kept in a dust-free box at room temperature until used.

Fluorescence in situ *hybridization probes*

Oligonucleotide probes were designed from consensus DNA sequences of tandem repeat sequences (Supplementary Data Table S2). Target sequences (59–82 nt) were manually selected to obtain a high level of sequence complexity to maximize probe specificity and ensure a GC content between 30 and 50 %. The sequences were checked to minimize self-annealing and formation of hairpin structures in Geneious 11.1.5 (https://www.geneious.com, Kearse *et al.*, 2012). The double-stranded DNA probes were generated and labelled with biotin-dUTP, digoxigenin-dUTP or Cy3-dUTP by nick translation as described by Mandáková and Lysak (2016*b*).

For retrotransposon probes, PCR primers were designed to the *gag* gene of various retrotransposon families (Supplementary Data Table S3). PCR products were sequenced at Macrogen Ltd. to validate them and then labelled by nick translation according to Mandáková and Lysak (2016*b*).

For comparative chromosome painting (CCP), chromosome-specific bacterial artificial chromosome (BAC) clones of *A. thaliana* grouped into contigs according to genomic blocks Jb and M of the Ancestral Crucifer Karyotype (Lysak *et al.*, 2016) were used and labelled with biotin-dUTP and digoxigenin-dUTP, respectively (Mandáková and Lysak, 2016*b*).

Fluorescence in situ *hybridization and microscopy*

Labelled probes were pooled, ethanol-precipitated, dried and dissolved in 20 µL of 50 % formamide and 10 % dextran sulphate in 2× saline–sodium citrate (SSC) per slide. Then 20 µL of the labelled probe was pipetted onto a suitable slide and denatured on a hotplate at 80 °C for 2 min. Hybridization was carried out in a moist chamber at 37 °C overnight. Post-hybridization washing was performed in 20 % formamide in 2× SSC at 42 °C. The immunodetection of hapten-labelled probes was performed as described by Mandáková and Lysak (2016*b*). Chromosomes were counterstained with 2 µg mL^{-1} 4′,6-diamidino-2-phenylindole (DAPI) in Vectashield. The preparations were photographed using a Zeiss Axioimager Z2 epifluorescence microscope with a CoolCube camera (MetaSystems). Images were acquired separately for all four fluorochromes using appropriate excitation and emission filters (AHF Analysentechnik). At least ten chromosome figures were photographed for each probe localized; however, due to combining different probes, almost all probes were localized on several slides repeatedly. The four monochromatic images were pseudocoloured, merged, processed and cropped using Photoshop CS (Adobe Systems). The images were processed only using the software functions applying to all pixels of the image.

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

Quantification of selected repeats using dot-blot analysis

Four repeats were quantified using a dot-blot analysis in *C. tenella* and *H. sylvestris*. We chose one satellite (ChTe2 and HeSy1) and one LTR retrotransposon (*gag* domain) from the Athila lineage (ChTe_Athila and HeSy_Athila) for each species. The radioactively labelled probes (synthesized oligonucleotides for satellites as described above, and purified and cloned PCR products for retroelements) were hybridized to diluted standards of unlabelled probes (0.125, 0.25, 0.5, 1 and 2 ng) and genomic DNA (1, 5, 50, 100 and 200 ng) of the two species onto Hybond-XL membrane (GE Healthcare). The dot-blot signals were quantified using a Typhoon FLA 9500 (GE Healthcare).

RESULTS

Extensive genome size variation versus chromosome number stasis

Our initial analysis confirmed that all the seven species analysed were diploid, with either $2n = 12$ (*H. sylvestris*) or $2n = 14$ (*Br. humilis*, *Bu. orientalis*, *C. tenella*, *D. micranthus*, *E. syriacum* and *M. incana*). Flow-cytometric analysis of nuclear DNA content revealed and confirmed extensive genome size variation among the seven species (Table 1). The smallest genome sizes were estimated for *E. syriacum* (254 Mb) and *C. tenella* (342 Mb), whereas *H. sylvestris* had the largest genome (4264 Mb). The four remaining species had medium to large genomes ranging from 1594 to 2611 Mb. Thus, the analysed species have comparable numbers of chromosomes, while their genome sizes differ by 16-fold and average chromosome size (genome size/haploid chromosome number) varies 20-fold (Table 1).

Genome size evolution

To reconstruct the evolution of genome size in the *Hesperis* clade, ITS and *ndh*F phylogenies were constructed using sequences retrieved from GenBank and BrassiBase (Kiefer *et al.*, 2014; https://brassibase.cos.uni-heidelberg.de) for species with known C-values (Fig. 1). Although the two trees showed similar basal dichotomy, splitting the six tribes into two groups, the position of Hesperideae (HESP) was not consistent among the ITS and *ndh*F trees. Due to the conflicting position of HESP, both trees were used to model genome size evolution and infer ancestral genome size (ancCS) for the *Hesperis* clade (Fig. 1, Supplementary Data Table S4).

For both phylogenies, Pagel's λ was estimated to determine the phylogenetic signal of genome size variation. As the λ values (0.96) were close to 1.0 for both trees, genome size evolution should be correlated with the tree structure. In the ITS phylogeny, ancGS was estimated as 1790 Mb (Fig. 1A, Supplementary Data Table S4) and a similar value was inferred based on the *ndh*F tree: 1524 Mb (Fig. 1B, Supplementary Data Table S4).

While the topology of the *ndh*F tree (Fig. 1B) supports morphological differences between the two tribal subclades (Mandáková *et al.*, 2017), namely between Chorisporeae (CHOR) and Dontostemoneae (DONT) on the one hand and Anchonieae (ANCH), Buniadeae (BUNI), Euclidieae (EUCL) and HESP on the other hand, the ~40 % more species used in the ITS tree provide a more realistic picture of genome size variation within the clade. In the context of the ITS tree (Fig. 1A), the inferred ancGS value points to independent genome size increases in ANCH, BUNI, DONT and HESP (note that only two C-values for DONT do not reflect the real extent of variation), accompanied by decreases in CHOR and EUCL. The maternal phylogeny (Fig. 1B) congruently suggests independent genome size increases in ANCH, BUNI and HESP, and downsizing in CHOR and EUCL (and DONT). As both inferred ancGS values are substantially bigger than the family's mean (617 Mb) and modal (392 Mb) genome sizes (Lysak *et al.*, 2009), the early diversification of the *Hesperis* clade was most likely marked by a genome size increase. The elevated ancestral genome size was subjected to stasis or further increase in ANCH, BUNI, DONT and HESP, while ~6-fold genome reductions occurred in CHOR and EUCL.

Genome size variation is correlated with life histories

In species with known genome sizes we tested whether the inter-species genome size differences are related to life-history strategies (Supplementary Data Table S1). The median and mean genome size of annual species ($n = 9$) was 697 and 1003 Mb, respectively. The species with prevalent perennial or biennial life history ($n = 20$) had median and mean genome size of 2054 and 2381 Mb, respectively. The median genome size was significantly lower in annuals than in perennials (Mann–Whitney test, $P = 0.0024$). We found a weak but significant positive

TABLE 1. *Chromosome numbers and genome sizes of the analysed* Hesperis*-clade plants*

Species	Tribe	2*n*	Genome size (1C)		Average chromosome size	
			(pg)	(Mb)	(pg)	(Mb)
E. syriacum	Euclidieae	14	0.26	254.28	0.04	36.33
C. tenella	Chorisporeae	14	0.35	342.30	0.05	48.90
Br. humilis	Euclidieae	14	1.63	1594.14	0.23	227.73
D. micranthus	Dontostemoneae	14	1.66	1623.48	0.24	231.93
M. incana	Anchonieae	14	2.20	2151.60	0.31	307.37
Bu. orientalis	Buniadeae	14	2.67	2611.26	0.38	373.04
H. sylvestris	Hesperideae	12	4.36	4264.08	0.73	710.68

1 pg = 978 Mb (Doležel *et al.*, 2003).

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

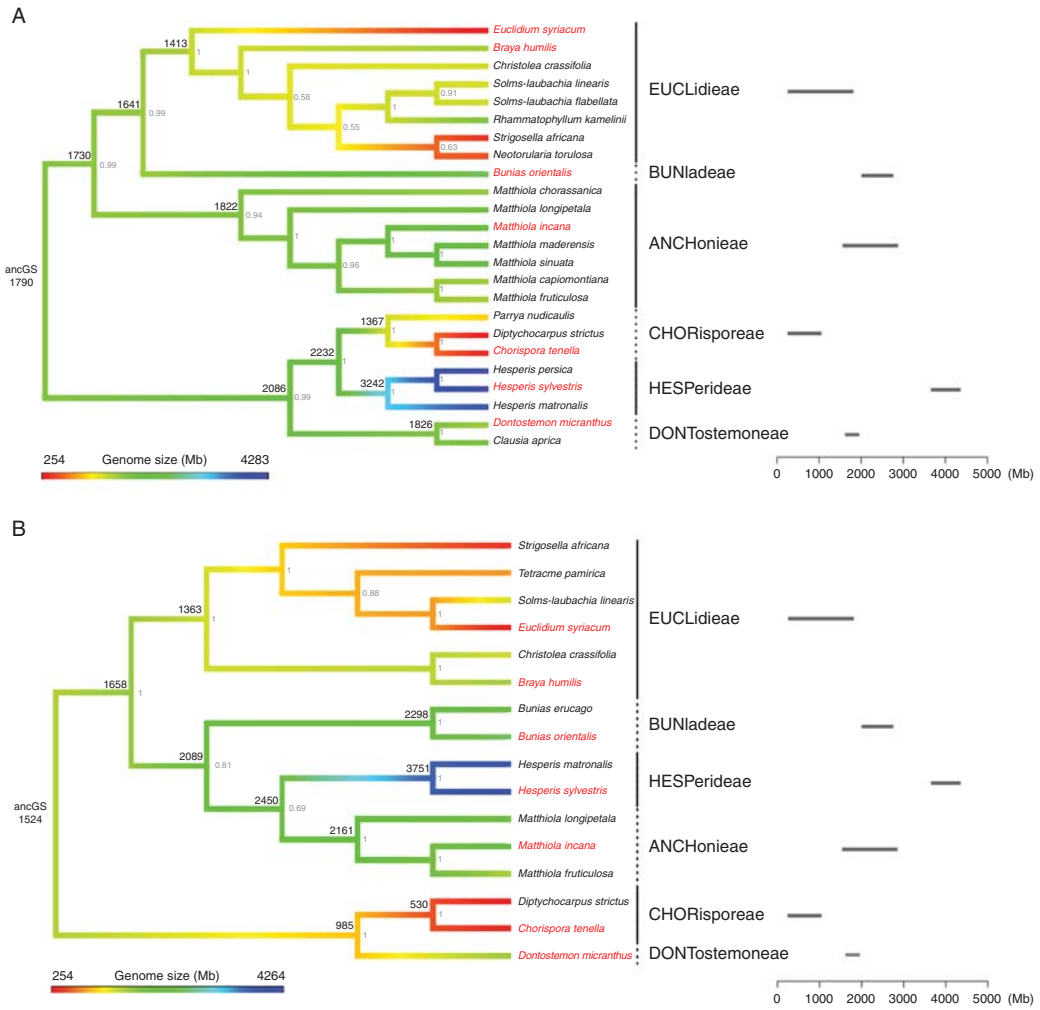


FIG. 1. Bayesian phylogenetic trees of the *Hesperis* clade with results of ancestral genome size reconstruction. (A) ITS tree. (B) *ndh*F tree. See Supplementary Data Table 1 for GenBank accession numbers. The reconstructed genome sizes (Mb) are shown at the nodes; posterior probability values are shown in grey. Horizontal bars represent the range of Cx-values for each tribe (C-values are from Table 1 and Kiefer *et al.*, 2014, https://brassibase.cos.uni-heidelberg.de). Species with an analysed repeatome are labelled in red.

correlation between increased genome size and perenniality (Spearman's rank correlation test, $\rho = 0.5549$, $P = 0.0018$).

Repeatome analysis

To identify and analyse the underlying sequences responsible for genome size variation in the *Hesperis* clade, whole-genome shotgun sequencing was performed in the seven species (Table 1) using an Illumina platform, generating 100- (*Br. humilis*, *Bu. orientalis*, *C. tenella*, *D. micranthus* and *E. syriacum*) or 300-nucleotide (*H. sylvestris* and *M. incana*) paired-end reads. All the longer reads were trimmed to 100 nucleotides prior conducting analyses embedded in the RepeatExplorer pipeline (Novák *et al.*, 2013). The cluster analysis permitting identification of reads derived from repetitive sequences was carried out for each species separately with two different samplings: (1) the maximum number of reads (genome coverage from 0.02× to 1.85×), and (2) at 0.05× genome coverage (Table 2). The detailed repeatome analysis was based on clustering with maximum reads.

We identified main types of repetitive sequences and their genome proportions in all the seven species analysed (Table 3, Fig. 2). Small genomes, i.e. those of *E. syriacum* and *C. tenella*, exhibited the lowest proportion of repeats: 24.31 and 33.33 %, respectively. In medium-sized genomes, repetitive sequences represented at least 40 % of their genomes (*Br. humilis*, 42.4

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

%; *D. micranthus*, 60.3 %; *M. incana*, 62.4 %; *Bu. orientalis*, 65.5 %). Within the largest genome of *H. sylvestris* at first only 52.82 % of repetitive DNA (sampled at 0.01× genome coverage; data not shown) was identified. However, after filtering out the most abundant tandem repeats (see Materials and methods section for details), a new round of cluster analysis retrieved 10.96 % additional repetitive sequences, increasing the total repeat content in *H. sylvestris* to 63.78 % (62.39 % when excluding cpDNA reads; Table 2). Among all seven species, low- or single-copy sequences constituted 35 % (900 Mb, *Bu. orientalis*) to 76 % (192 Mb, *E. syriacum*) of the sequence data and ~4–14 % of repeats remained unclassified (Table 3).

To determine how reliable our *in silico* estimates of repeat abundances were, we quantified the number of genomic copies for one tandem repeat and Athila retrotransposon (*gag* domain) by a dot-blot analysis in two species with contrasting genome sizes. The dot-blot and *in silico* estimates were largely congruent for *C. tenella* (1C = 342 Mb) and *H. sylvestris* (1C = 4264 Mb), except for the ChTe2 tandem repeat in *C. tenella*, being 1.75-fold more abundant in the dot-blot analysis (Supplementary Data Table S5). This discrepancy suggests that *in silico*-estimated abundances of tandem repeats can be somewhat underestimated compared with dot-blot or Southern blot analyses due to G/C bias in Illumina reads (Chen *et al.*, 2013) and tandem repeats usually being A/T-rich.

Retrotransposon diversity and abundances

In all seven genomes, LTR retrotransposons made up the majority of repeatomes, ranging from 11.11 % in *C. tenella* to nearly 48.11 % in *M. incana* (Table 3). Although *H. sylvestris* has the largest genome among the species analysed (Table 1), only 40.56 % (1 729.51 Mb) of its genome was identified to be built from LTR retrotransposons. The identified *Ty1-copia* elements belonged to seven lineages (Ale, Angela, Bianca, Ivana/Oryco, Maximus/SIRE, TAR and Tork; Table 3) out of the 16 known lineages (Neumann *et al.*, 2019). The identified *Ty3-gypsy* elements belonged to two major lineages (Neumann *et al.*, 2019): Chromovirus (represented by CRM and Tekay clades) and non-Chromovirus (Athila and Ogre/Tat clades; Table 3).

In all genomes, LTR retroelements of the *Ty3-gypsy* superfamily prevailed and were mainly represented by the Athila clade, followed by Ogre/Tat (Table 3). The abundance of Athila elements ranged from 2.19 % in *E. syriacum* to 22.62 % in *D. micranthus*. In the smallest genome, that of *E. syriacum*, the Ogre/Tat element was the most abundant *Ty3-gypsy* element (2.42 %), followed by Athila (2.19 %) and Chromovirus (1.27 %, mainly CRM lineage). However, the Ogre/Tat clade was most amplified in genomes of *M. incana* (6.57 %) and *D. micranthus* (8.05 %). Some *Ty3-gypsy* elements remained unclassified, as we were not able to assign them clearly to any lineage; the highest proportion of unclassified *Ty3-gypsy* elements was identified in the larger genomes of *H. sylvestris* (~10 %) and *Bu. orientalis* (~13 %). In all but one species, the Chromovirus lineage was represented by the CRM and Tekay clades; in *H. sylvestris*, the Tekay clade was more abundant than CRM.

Ty1-copia retroelements, represented mainly by the Angela lineage, were 2- to 5-fold less abundant than *Ty3-gypsy* elements (Table 3). Angela retroelements occupied 1.70 % (*E. syriacum*) to 9.76 % (*M. incana*) of the genome. Other common lineages in medium- and large-sized genomes were Ale (from 1.36 % in *Br. humilis* to 2.75 % in *M. incana*), Bianca (from 0.80 % in *Br. humilis* to 1.94 % in *Bu. orientalis*) and Maximus (from 0.07 % in *M. incana* to 1.15 % in *D. micranthus*). The representation of *Ty1-copia* retroelements in *D. micranthus* was significantly lower than in other medium-sized genomes (e.g. Ale was not identified and Angela elements occupied only 3.38 % of the genome). In small-sized genomes, after Angela, the second most abundant *Ty1-copia* element was Maximus in *E. syriacum* (0.55 %) and Bianca in *C. tenella* (0.80 %). Other *Ty1-copia* lineages, such as Ivana/Oryco, TAR and Tork, were found only in low amounts or were absent in repeat clusters constituting at least 0.01 % of a genome (Table 3).

From non-LTR retrotransposons, LINE elements were identified only at very low genome proportions in the analysed species: 0.08 % in *H. sylvestris* to 0.51 % in *D. micranthus* (Table 3). MITE and SINE elements were not detected in clusters

TABLE 2. *Numbers of high-throughput sequencing reads used in the RepeatExplorer bioinformatic pipeline and clustering statistics*

Species	Maximum no. of reads					Genome coverage 0.05×		
	No. of reads	Genome coverage	Total repeats* (%)	Total repeats excluding cpDNA† (%)	No. of clusters	No. of reads	Total repeats* (%)	No. of clusters
E. syriacum	4 711 370	1.85	38.05	24.31	385	128 516	22.96	333
C. tenella	1 898 952	0.55	41.10	33.33	401	171 150	25.98	354
Br. humilis	4 235 224	0.27	53.46	42.40	454	796 316	43.28	444
D. micranthus	4 258 534	0.26	62.41	60.30	475	809 780	54.67	380
M. incana	2 589 598	0.12	65.98	62.40	467	1 075 800	61.43	430
Bu. orientalis	2 969 920	0.11	67.19	65.50	440	1 305 300	59.58	471
H. sylvestris	1 221 831	0.03	63.78‡	62.39‡	323	2 133 750	No data¶	

*Percentage of repeats in clusters constituting at least 0.01 % of the genome.

†Percentage of repeats in clusters constituting at least 0.01 % of the genome; clusters annotated as cpDNA were excluded.

‡RepeatExplorer analysis was performed with the advanced option of automatic filtering out the most abundant tandem repeats.

¶Not possible to compute due to computational resources restriction.

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

TABLE 3. *Genome proportions and classification of repetitive sequences from the individual RepeatExplorer analyses performed using the maximum number of reads for clustering*

		E. syriacum	*C. tenella*	*Br. humilis*	*D. micranthus*	*M. incana*	*Bu. orientalis*	*H. sylvestris*
	Genome size (Mb)	254.28	342.30	1 594.14	1 623.48	2 151.60	2 611.26	4 264.08
Repeat family	Lineage	% of the genome/Mb						
LTR retrotransposons		11.65/29.63	11.11/38.03	31.02/494.51	43.74/710.12	48.11/1035.14	44.48/1161.41	40.56/1729.51
Ty1-copia	Ale	0.09/0.23	0.10/0.35	1.36/21.69	0/0	2.75/59.17	2.20/57.45	2.00/85.28
	Angela	1.70/4.33	1.74/5.96	9.39/149.69	3.38/54.88	9.76/210.00	6.19/161.64	7.30/311.28
	Bianca	0.30/0.77	0.90/3.09	0.80/12.76	1.05/17.05	1.13/24.32	1.94/50.66	1.79/76.33
	Ivana/Oryco	0.14/0.36	0.08/0.28	0.20/3.19	0.24/3.90	0.06/1.30	0.21/5.48	0.05/2.13
	Maximus/SIRE	0.55/1.4	0.08/0.28	0.85/13.56	1.15/18.68	0.07/1.51	0.90/23.50	1.14/48.61
	TAR	0/0	0.17/0.59	0.06/0.96	0.11/1.79	0.35/7.54	0.38/9.92	0.22/9.38
	Tork	0.12/0.31	0.12/0.42	0.26/4.15	0.35/5.69	0.20/4.31	0.51/13.32	0.38/16.20
	Unclassified	0.30/0.77	0.88/3.02	0.79/12.60	1.03/16.73	0.84/18.08	5.02/131.09	0/0
	Total	3.20/8.14	4.06/13.90	13.71/218.56	7.31/118.68	15.16/326.19	17.36/453.31	12.88/ 549.21
Ty3-gypsy	Chromovirus	1.27/3.23	0.64/2.20	1.79/28.54	1.69/27.44	1.63/35.08	1.68/43.87	1.75/74.62
	CRM	1.23/3.13	0.62/2.12	1.53/24.39	1.16/18.83	0.98/21.09	1.04/27.16	0.75/31.98
	Tekay	0.04/3.10	0.02/0.08	0.26/4.15	0.53/8.61	0.65/13.99	0.64/16.71	1.00/42.64
	Athila	2.19/5.57	3.25/11.13	11.29/179.98	22.62/367.24	22.56/485.41	10.73/280.19	20.54/875.84
	Ogre/Tat	2.42/6.16	0.16/0.55	0.35/5.58	8.05/130.7	6.57/141.37	1.33/34.73	0.75/31.98
	Unclassified	2.57/6.54	3.00/10.27	3.88/61.86	4.07/66.08	2.20/47.34	13.38/349.39	4.64/197.85
	Total	8.45/21.49	7.06/24.17	17.31/275.95	36.43/591.44	32.95/708.96	27.12/708.17	27.68/1180.30
DNA transposons	CACTA	0.70/1.78	0.59/2.02	1.62/25.83	1.81/29.39	1.71/36.80	2.09/54.58	0.63/26.86
	Helitron	0.62/1.58	0.36/1.24	0.22/3.51	0.03/0.49	0.22/4.74	0.02/0.52	0.05/2.13
	Mutator	0.37/0.95	0.87/2.98	1.27/20.25	0.40/6.50	0.38/8.18	1.38/36.04	0.36/15.35
	Unclassified	1.56/3.97	0.45/1.55	1.61/25.67	0.08/1.30	0.46/9.90	0.50/13.06	0/0
	Total	3.25/8.27	2.27/7.78	4.72/75.25	2.32/37.67	2.77/59.60	3.99/104.19	1.04/44.35
LINE		0.14/0.36	0.31/1.07	0.26/4.15	0.51/8.28	0.32/6.89	0.23/6.01	0.08/3.41
rDNA		1.38/3.51	1.64/5.62	1.69/26.95	0.68/11.04	0.63/13.56	1.20/31.36	0.25/10.66
Satellites		2.69/6.85	7.88/26.98	0.26/4.15	0.46/7.47	0.87/18.72	0.80/20.89	8.77/373.96
Unclassified repeats		5.20/13.23	9.98/34.17	3.96/63.13	12.53/203.43	9.70/208.71	13.90/362.97	11.68/498.04
Low/single-copy sequences		75.69/192.47	66.67/228.22	57.60/918.23	39.70/644.53	37.60/809.01	34.50/900.88	37.61/1603.72
All repeats total		24.31/61.82	33.33/114.09	42.40/675.92	60.30/978.96	62.40/1342.60	65.50/1710.38	62.39/2660.36

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

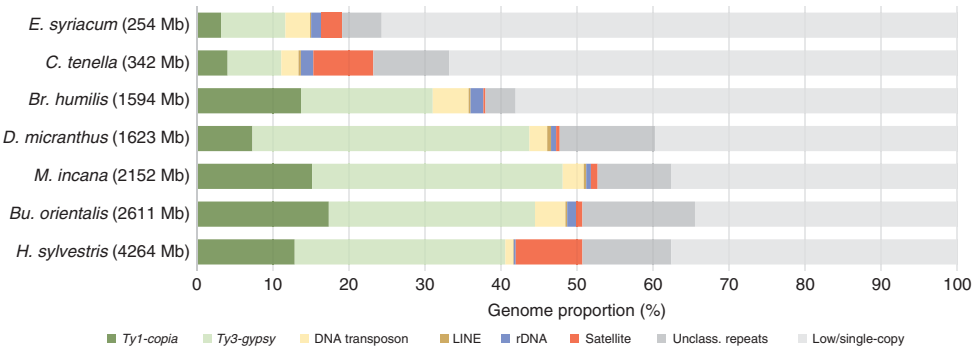


FIG. 2. Relative abundances of repeat families and low/single-copy sequences identified in genomes of the seven *Hesperis*-clade species analysed.

constituting at least 0.01 % of a genome. DNA transposons were represented by abundances ranging from 1.04 % in *H. sylvestris* up to 4.72 % in *Br. humilis*; the most abundant of these were CACTA and Mutator elements (Table 3).

With the exception of H. sylvestris*, tandem repeats did not contribute significantly to genome expansion in the* Hesperis*-clade species*

Tandem repeats were found in different abundances, from a very low (0.26 % in *Br. humilis* and 0.46 % in *D. micranthus*) to a high genome proportion in *H. sylvestris* (8.77 %) (Table 3 and Fig. 2). Results of tandem repeat analysis are summarized in Table 4. The identified monomer sizes were variable among the seven species, ranging from 20 to ~350 bp. The 825-bp ChTe2 satellite identified in *C. tenella* had an exceptional monomer length.

In the small-sized genomes of *E. syriacum* and *C. tenella*, tandem repeats occupied 2.78 % (four different repeats) and 7.61 % (seven different repeats), respectively. In the *E. syriacum* genome, while only 0.08 % of the genome was found to consist of typical tandemly repeated DNA, a satellite family of non-homogeneous monomers containing a 60-bp repetitive motif occupied ~ 2.70 % of the genome. All contigs from the RepeatExplorer cluster analysis whose graph shapes indicated putative tandem repeats were further analysed using Dotter and TRF to create self-dot plots and to identify satellite monomer lengths, respectively. The 60-bp motif was identified by TRF using all reads (average 75 % matches) and by the TAREAN pipeline using sampled reads, which additionally identified satellites with monomer lengths of 519, 179, 60 and 40 bp. To further investigate these sequences, we analysed the sequenced *E. syriacum* genome (Jiao *et al.*, 2017) by TRF and TAREAN. Whereas TAREAN identified two satellites with a monomer length of 717 and 377 bp, the TRF analysis revealed two more monomer lengths: 357 and 397 bp. All the identified monomers contained the 60-bp motif (Supplementary Data Fig. S1). In *C. tenella*, approximately one-third (2.72 %) of the tandem repeats identified were represented by ITRs derived from the arabidopsis-type telomeric repeat (TTTAGGG).

In species with medium-sized genomes, tandem repeats represented <0.9 % of their genomes. In *M. incana* and *Bu. orientalis*, tandem repeats constituted only 0.87 % (five different satellites) and 0.77 % (nine different satellites) of the genome, respectively. Among the five identified tandem repeats in *H. sylvestris*, the 91-bp HeSy1 satellite repeat occupied 7.38 % of the genome.

Chromosomal localization of the identified repeats

Chromosomal localization of the identified repeats was determined by fluorescence *in situ* hybrization (FISH) of fluorochrome- or hapten-labelled DNA probes to mitotic chromosomes. To localize retrotransposons, probes designed to the *gag* domain of Angela, Athila and Chromovirus were used.

In species with a small genome size (*C. tenella* and *E. syriacum*), tandem repeats as well as retrotransposons clustered within heterochromatic pericentromere regions. In *E. syriacum*, FISH of DNA probes corresponding to consensus monomer sizes of 357 bp (EuSy1A) and 377 bp (EuSy1B), containing the 60-bp repetitive motif, showed that these repeats occurred on two and one chromosome pair(s), respectively (Fig. 3A). In *C. tenella*, three major tandem repeats formed pericentromere chromatin (Fig. 3B). The 39-bp ChTe1 satellite (2.27 % of the genome) localized to four chromosome pairs, the 825-bp ChTe2 repeat (1.60 %) provided weak hybridization signals on three chromosome pairs, and the 139-bp ChTe3 repeat (0.84 %) gave a stronger hybridization signal at the heterochromatin/euchromatin boundary of four chromosome pairs. The large blocks of ITRs (~2.7 %) were located at all pericentromeres in *C. tenella* (Fig. 3B), whereas telomeric repeats were localized only at chromosome ends in *E. syriacum* (Fig. 3A). In both species, LTR retrotransposons were present in all pericentromere regions (Fig. 4A–C), largely co-localizing with the identified tandem repeats (Fig. 4M). Apart from the pericentromeric heterochromatin, Chromovirus and Athila retroelements co-localized with four terminal nucleolar organizing regions (NORs) in *E. syriacum* (Fig. 4A) and eight NORs in *C. tenella* (Fig. 4B). The DNA probe for the Angela retrotransposon hybridized to all pericentromeres and the eight NORs in *C. tenella* (Fig. 4C).

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

TABLE 4. *Tandem repeats identified by RepeatExplorer and TAREAN analyses. Only repeats with genome proportion >0.01 % were analysed and are listed*

Species	Tandem repeat	Monomer length (bp)	Genome proportion (%)
E. syriacum	EuSy1 (EuSy1A, EuSy1B)	60 (motif)	2.70
	EuSy2	20	0.05
	EuSy3	354	0.03
	EuSy4	132	0.01
C. tenella	ChTe1	39	2.27
	ChTe2	825	1.60
	ChTe3	139	0.84
	ChTe4	102	0.12
	ChTe5	28	0.04
	ChTe6	52	0.02
	ITR and telomeric repeat	7	2.72
Br. humilis	BrHu1	161	0.16
	BrHu2	295	0.04
	BrHu3	87	0.02
	BrHu4	338	0.02
	BrHu5	345	0.02
	telomeric repeat	7	0.24
D. micranthus	DoMi1	36	0.30
	DoMi2	143	0.06
	DoMi3*	350	0.05
	DoMi4	26	0.03
	DoMi5	354	0.02
	DoMi6	182	0.01
M. incana	MaIn1*	352	0.58
	MaIn2	355	0.10
	MaIn3	69	0.08
	MaIn4	88	0.06
	MaIn5	590	0.05
Bu. orientalis	BuOr1*	352	0.36
	BuOr2	192	0.18
	BuOr3	179	0.10
	BuOr4	20	0.09
	BuOr6	171	0.01
	BuOr7	77	0.01
	BuOr8	177	0.01
	BuOr9	490	0.01
	telomeric repeat	7	0.40
H. sylvestris	HeSy1	91	7.38
	HeSy2	161	0.69
	HeSy3	91	0.08
	HeSy4	200	0.07
	HeSy5	174	0.06

*Shared repeats.

In medium-sized genomes (*Br. humilis*, *D. micranthus*, *Bu. orientalis* and *M. incana*), tandem repeats predominantly constituted pericentromere and subtelomere heterochromatic regions. In *D. micranthus*, the 36-bp DoMi1 satellite (0.30 % of genome) localized to subtelomeric regions of six chromosome pairs and, together with the 143-bp DoMi2 satellite (0.06 %), to the pericentromere of an additional chromosome pair (Fig. 3C). The 350-bp DoMi3 satellite (0.05 %) localized to subtelomeric regions of four chromosome pairs, whereas the 26-bp DoMi4 repeat (0.03 %) occurred on three chromosomes (Fig. 3C). In *Bu. orientalis*, the 352-bp tandem repeat BuOr1 (0.36 % of genome) showed localization at chromosome termini of six out of the seven chromosome pairs (Fig. 3D). FISH with the BuOr1 satellite and the arabidopsis-like telomeric repeat showed that the newly identified repeat occupied the most distal chromosome regions immediately adjacent to the telomeric repeats at 11 chromosome ends (Fig. 3D). The 192-bp satellite BuOr2 (0.18 %), together with the 179-bp BuOr3 (0.10 %), was localized on the same arm of a single chromosome pair (Fig. 3D).

In *Br. humilis* and *M. incana,* LTR retrotransposons co-localized with pericentromeric heterochromatin and both adjacent chromosome arms, except for the most proximal regions (Fig. 4D–F). In *Bu. orientalis* and *D. micranthus* the Athila and Angela retrotransposons showed dispersed distribution along the entire length of all chromosomes (Fig. 4G–J).

In *H. sylvestris*, the most abundant 91-bp tandem repeat, HeSy1 (7.38 % of the genome), was localized at pericentromeres of only three chromosome pairs (Fig. 3E). The 91-bp HeSy3 (0.08 %), showing 80 % sequence identity to HeSy1, co-localized with HeSy1 on three chromosome pairs, in addition to a solo localization on a fourth chromosome. The 161-bp satellite HeSy2 (0.69 %) localized to ten subtelomeric regions; the 174-bp HeSy5 (0.06 %) had a similar localization, with signals on four chromosome pairs, and the 200-bp HeSy4 (0.07

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

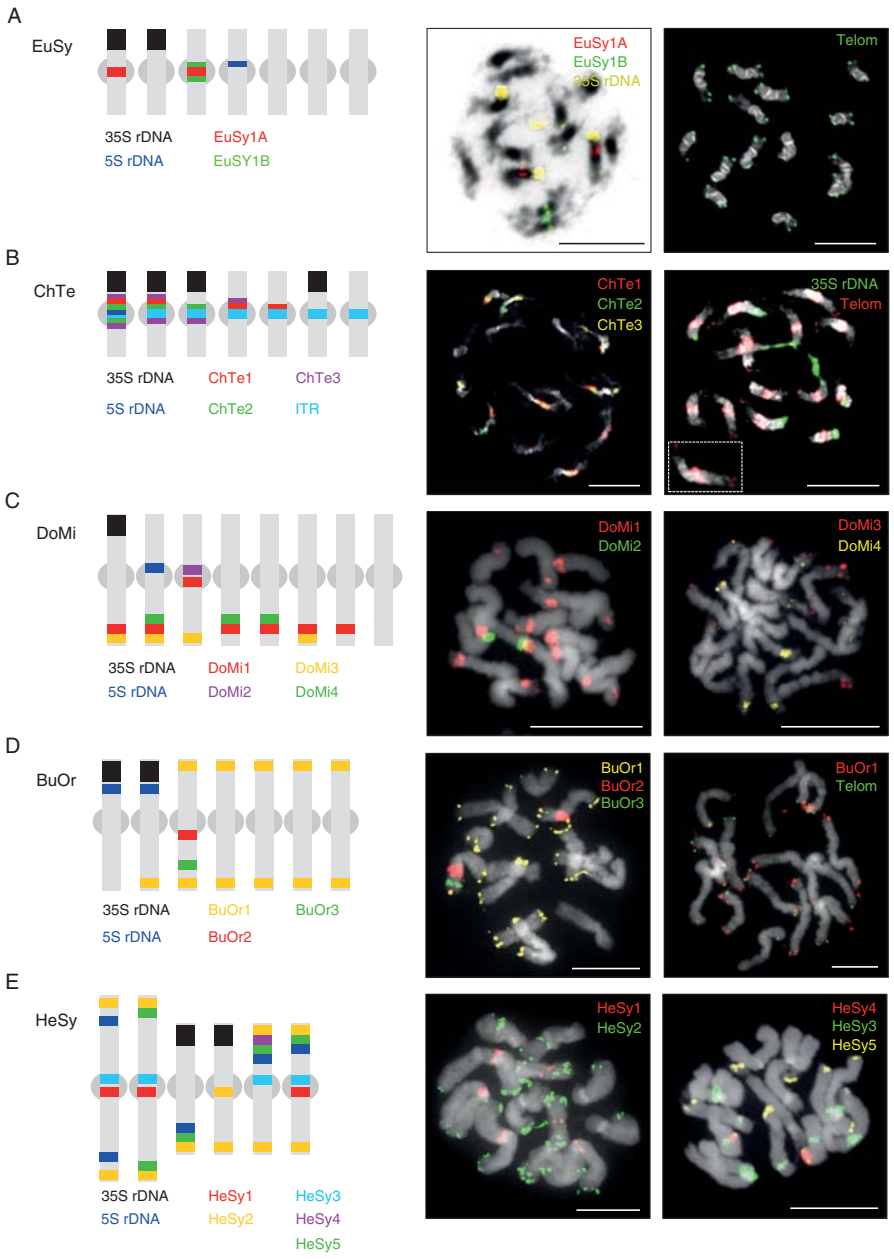


FIG. 3. Chromosomal localization of the most abundant tandem repeats and rDNA loci on mitotic metaphase chromosomes in five *Hesperis*-clade species. Telomeres and ITRs (B) were localized using a FISH probe for the arabidopsis-type telomeric repeat. Chromosomes were counterstained by DAPI (displayed in black and white); FISH signals are shown in colour as indicated. Grey spheroids in the schematic ideograms represent (peri)centromeric regions. EuSy, *E. syriacum*, ChTe, *C. tenella*; DoMi, *D. micranthus*; BuOr, *Bu. orientalis*; HeSy, *H. sylvestris*. All scale bars = 10 µm.

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

%) localized to one arm of a single chromosome pair (Fig. 3E). In *H. sylvestris*, retrotransposon probes hybridized along the entire length of all chromosome pairs, except for (peri)centromeric regions (Fig. 4K, L). FISH with DNA probes for HeSy1 and Athila and Angela (not shown) retrotransposons confirmed that the (peri)centromeric regions with a low abundance of dispersed repeats were occupied by tandem repeats (Fig. 4N).

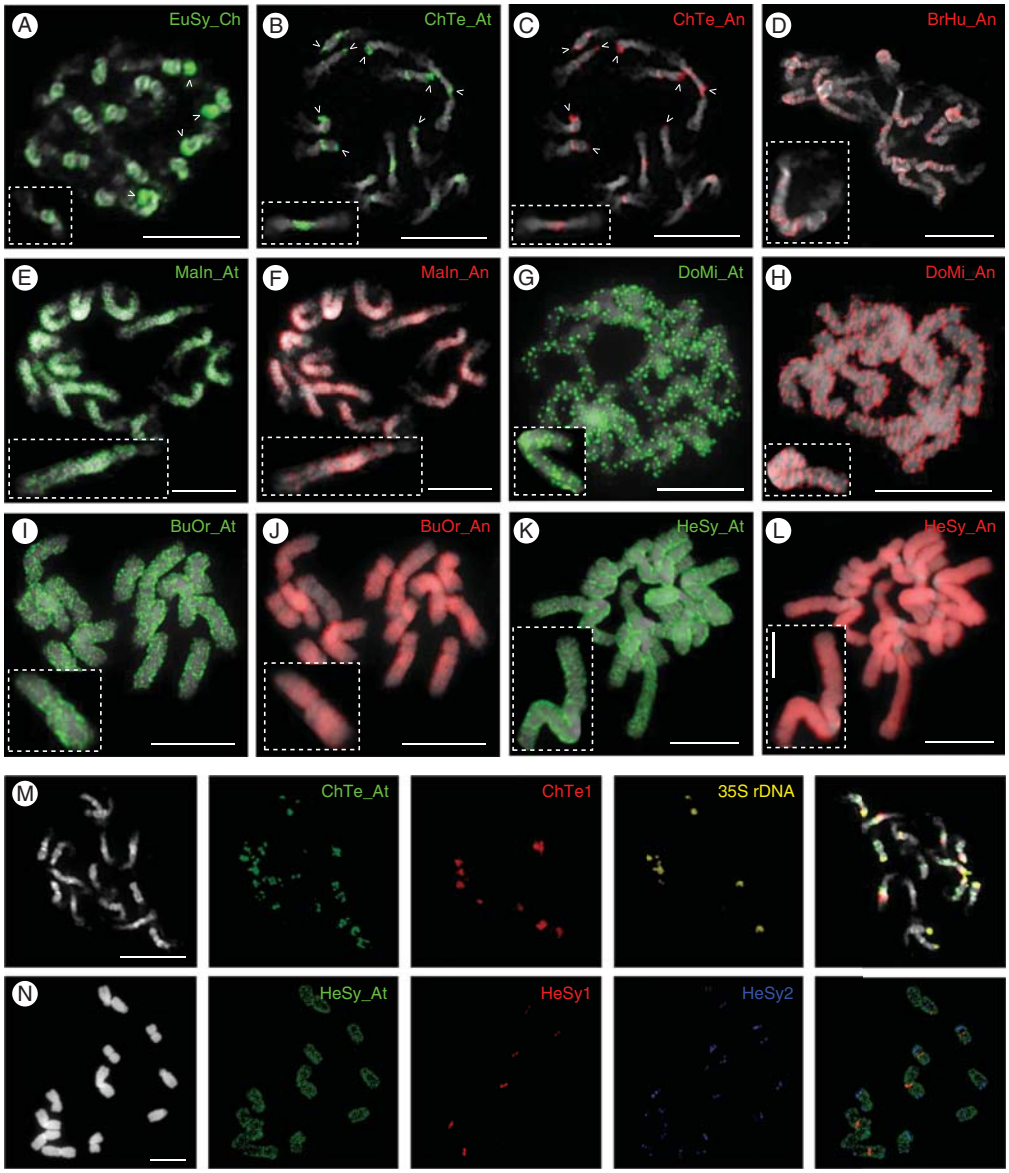


FIG. 4. Chromosomal localization of dispersed repeats on mitotic metaphase chromosomes in seven *Hesperis*-clade species. (A) *E. syriacum* (EuSy), (B, C) *C. tenella* (ChTe), (D) *Br. humilis* (BrHu), (E, F) *M. incana* (MaIn) (G, H) *D. micranthus* (DoMi), (I, J) *Bu. orientalis* (BuOr), (K, L) *H. sylvestris* (HeSy) (M, N) Co-localization of tandem repeats (Fig. 3) and Athila retrotransposons in *C. tenella* (M) and *H. sylvestris* (N). Arrowheads in (A–C) point to 35S rDNA (NOR) loci. Chromosomes were counterstained by DAPI (displayed in black and white); FISH signals are shown in colour as indicated. Lineage abbreviations: An, Angela; At, Athila; Ch, Chromovirus. Scale bars (complete chromosome spreads) = 10 µm; (insets) = 5 µm (L).

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

Repeat content is positively correlated with genome size

We tested whether the estimated abundances of the identified repeats reflect the ~16-fold genome size variation among the analysed species (Supplementary Data Table S6). A strong positive correlation was found between the repeat content and genome size (R^2 = 0.984, P = 1.041e^{-5}), but also between abundances of both LTR retrotransposon superfamilies and increasing genome size (*Ty1-copia* plus *Ty3-gypsy*, R^2 = 0.967, P = 4.168e^{-5}; *Ty3-gypsy*, R^2 = 0.940, P = 0.0003; *Ty1-copia*, R^2 = 0.918, P = 0.0007), and in particular for both families, Athila and Angela (Athila, R^2 = 0.874, P = 0.0020; Angela, R^2 = 0.885, P = 0.0016). Despite a significant positive correlation between tandem repeat contents and genome size variation, the lower R^2 value (0.604) indicates that tandem repeat amplification influenced genome size expansions across the *Hesperis* clade to a lesser extent than proliferation of LTR retrotransposons.

Phylogenetic relationships among the identified LTR retroelement lineages

Phylogenetic analyses based on the reverse transcriptase (RT) sequence of the identified LTR retrotransposons were carried out to assess whether their relationships reflect tribal relationships within the *Hesperis* clade. As expected, *Ty1-copia* and *Ty3-gypsy* elements clustered into major clades, namely Angela, Ale, Bianca, Ivana/Oryco, Maximus/SIRE, TAR and Tork (Supplementary Data Fig. S2), and Athila, Chromovirus and Ogre/Tat (Supplementary Data Fig. S3), respectively. No species-specific LTR retroelements were found. These analyses further supported the antiquity and ubiquity of LTR retrotransposon lineages shared among the six tribes.

Identification of shared repeats within the Hesperis *clade and across the Brassicaceae*

A comparative RepeatExplorer cluster analysis was performed by pooling single dataset reads used in individual analyses as random samples corresponding to 0.01× genome coverage; the total of 1 284 124 reads were analysed, the number of reads per species ranging from 25 704 (*E. syriacum*) to 409 540 (*H. sylvestris*). The detailed analysis of the first 300 clusters showed variation in the proportion of reads contributed by individual species due to a positive correlation between abundances of species-specific reads and genome size of the analysed species. The majority of the first 300 repeat clusters contained sequences of all or most species analysed and were annotated as LTR retrotransposons (Supplementary Data Fig. S4).

In contrast to LTR retrotransposons, most of the tandem repeat clusters were made up of reads originating from a single species and no tandem repeat was shared among all the species analysed. To compare the sequence identity of all the identified tandem repeats, we BLASTed these sequences against each other and against known sequences in the NCBI GenBank database. A tandem repeat with an average monomer size of 352 bp was found to be shared among *Bu. orientalis* (BuOr1), *M. incana* (MaIn1) and partly *D. micranthus* (DoMi3). These three satellites showed hits to *Brassica oleracea* (pBoSTRb) and *Brassica rapa* (pBrSTRb) subtelomeric satellites (Koo *et al.*, 2011) with identities up to 79 % for BuOr1 and MaIn1, 67 % for DoMi3 (Supplementary Data Fig. S5).

The CRAMBO tandem repeat (CRAMBO7 and CRAMBO.6 both 338 bp, and CRAMBO.11 309 bp), previously found only in *Cardamine* species (Mandáková *et al.*, 2013), was found in the *Br. humilis* genome as the 338-bp BrHu4 tandem repeat (0.02 % of the genome, Table 4). In a pairwise BLASTN comparison, the BrHu4 repeat and the three CRAMBO variants (accession numbers JQ412178, JQ412179 and JQ412180) exhibited 96–99 % sequence identity with 81–95 % query coverage. The remaining identified tandem repeats did not show a significant sequence similarity to already known repeats.

Feasibility of comparative BAC-based painting decreases with increasing genome size and repeat content

While reconstructing genome evolution in *Hesperis*-clade tribes by CCP based on arabidopsis BAC clones (Mandáková *et al.*, 2017), we noticed that the method was less efficient or even not applicable to species with large(r) genomes. Here we used the repeatome data and chromosomal localization of the identified repeats to reassess the feasibility of BAC painting in crucifer species with large genomes.

In six *Hesperis*-clade species and under identical experimental conditions, CCP with BAC contigs spanning genomic blocks Jb and M and forming chromosome 4 of CEK (ancestral karyotype of Clade E; Mandáková *et al.*, 2017) demonstrated that chromosome specificity of painting probes and overall efficacy of CCP gradually decreased with increasing repeat content and genome size (Supplementary Data Fig. S6). Whereas in *C. tenella* and *E. syriacum* both painting probes provided highly specific and strong hybridization signals, weaker, less specific and homogeneous signals were observed in *D. micranthus* and *M. incana*. In *Bu. orientalis* and *H. sylvestris*, painting was even more compromised, with fluorescent signals hardly specific and distinguishable.

DISCUSSION

Due to the prominent role of the minute arabidopsis genome in plant research, crucifers are traditionally viewed as an angiosperm lineage harbouring species with comparably small genomes. Here we showed that species and genera of the *Hesperis* clade represent an exception to the rule and that these genomes followed evolutionary trajectories different from most crucifer taxa.

Genome size evolution in the Hesperis *clade: genome obesity, with rare genome downsizing*

Although based on a very limited dataset, our reconstruction of ancestral genome size suggested that the common ancestor of the *Hesperis* clade (called CEK; Mandáková *et al.*,

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

2017) most likely had a genome larger (~1600 Mb) than the modal (392 Mb) and mean (617 Mb) C-values for Brassicaceae species, and that the expansion of the ancestral genome has preceded the tribal diversification within the clade. As the ancestral genome upsizing was followed by further genome size increases in all six tribes (Fig. 1), the *Hesperis* clade genomes must have intrinsic propensities to tolerate or benefit from further genome expansion. When plotting the available C-values on phylogenetic trees of two tribes harbouring species with small and large genomes, namely Chorisporeae and Euclidieae, the prevailing tendency for genome expansion is further supported. In Chorisporeae (~63 species in four genera), the small *Chorispora*/*Diptychocarpus* subclade (12 species), containing species with small genomes, is sister to or younger than (German *et al.*, 2011; BrassiBase, https://brassibase.cos.uni-heidelberg.de) the species-rich *Parrya*/*Litwinowia* subclade of 43 perennial species with genomes presumably as large as that of *Parrya nudicaulis*. Thus, these phylogenies point to a more recent origin of *Chorispora*/*Diptychocarpus* genomes followed by genome downsizing. In the diverse and species-rich Euclidieae (28 genera and 152 species; Chen *et al.*, 2018), large genomes of perennial species prevail (Supplementary Data Table S1), whereas small genomes have been identified so far only in the annual species *E. syriacum* (one species in the genus), *Neotorularia torulosa* (~14 species) and *Strigosella africana* (24 species). The dominance of large genomes and the phylogenetic position of the three genera in the tribe (Chen *et al.*, 2018) point to genome downsizing specific for *Euclidium* and (some) species of *Neotorularia* and *Strigosella*.

As the genome obesity of *Hesperis*-clade species was caused mainly by the activity of LTR retrotransposons, particularly *Ty3-gypsy* elements, whole-genome duplication(s) (WGD) as a possible mechanism underlying the genome size increases can be ruled out. This was corroborated by earlier CCP analyses which failed to detect duplicated genomic regions in all *Hesperis*-clade species analysed (Mandáková *et al.*, 2017). Interestingly, when comparing genome sizes in species from the 13 crucifer clades (Lysak *et al.*, 2009; Kiefer *et al.*, 2014; https://brassibase.cos.uni-heidelberg.de/; Hohmann *et al.*, 2015) that have undergone a mesopolyploid WGD (Mandáková *et al.*, 2017), it turns out that these species have usually substantially smaller genome sizes than many *Hesperis*-clade species. This is due to long-lasting and genome-wide post-polyploid diploidization effectively downsizing the inflated mesopolyploid genomes. The peculiar exception to this trend is the 2300-Mb genome of *Physaria bellii* (n = 4, Physarieae; Lysak and Lexer, 2006; Lysak *et al.*, 2009). In this species, and potentially in some of its congeners, the tribe-specific whole-genome triplication (Mandáková *et al.*, 2017) was followed by extensive diploidization, including descending dysploidy to only four chromosome pairs, and amplification of repetitive sequences increasing the average chromosome size in *P. bellii* (575 Mb) to values comparable with *Hesperis*-clade species (Fig. 1).

Genome expansion through amplification of TEs

Genome size variation among crucifer species with the arabidopsis-like chromosomal architecture was associated with the expansion (or contraction) of repeat-rich pericentromeres (Hall *et al.*, 2006), as the insertion of amplified retrotransposon copies and other repeats into pericentromeres is potentially less harmful than targeting gene-rich chromosome arms. Although this has certainly occurred in some species, as evidenced by the ITR arrays at all pericentromeres in *C. tenella*, here we showed that the *Hesperis*-clade genomes expanded due to the chromosome-wide amplification of LTR retrotransposons and, to a lesser extent, the origin and amplification of tandem repeats. Whereas the diversity of TEs was comparable among all the sequenced genomes, the abundances of individual TE families differed substantially among the genomes and were positively correlated with increasing genome sizes. In all the large-genome species, *Ty3-gypsy* elements were identified as the key repeatome components driving the observed genome expansions. The frequently dominating role of *Ty3-gypsy* elements in genome size upsizing was documented in species from diverse plant families (e.g. Park *et al.*, 2012; Macas *et al.*, 2015; Willing *et al.*, 2015; Dodsworth *et al.*, 2017). Based on our partial repeatome analysis, tandem repeats represented only 0.26–0.8 % of repeatomes in four genomes >1500 Mb (Table 1). The high genome abundance (~7 %) of the HeSy1 repeat in *H. sylvestris* makes one notable exception. It remains unclear whether the accumulation of this repeat at three pericentromeres in *H. sylvestris* could have a functional role and whether this or similar high-copy tandem repeats can be found in genomes of other Hesperideae species.

The small genomes characterizing annual species of Chorisporeae (*Chorispora* and *Diptychocarpus*) and Euclidieae (*Euclidium*, *Neotorularia* and *Strigosella*) presumably represent independent subclade- (Chorisporeae) or species-specific (Euclidieae) genome downsizing events. Although our repeatome analysis, together with the phylogenetic position of these taxa, points to genome purging, it is difficult to pinpoint the underlying mechanism(s) using short read sequences (Macas *et al.*, 2015). Repetitive sequences can be removed by recombination within or between repeat copies (Devos *et al.*, 2002; Hawkins *et al.*, 2009) or during double-strand break repair (e.g. Vu *et al.*, 2017). However, a first prerequisite of deeper understanding of DNA purging in these tribes is more supported phylogenetic relationships with the aim of identifying species and genus pairs with and without genome contraction.

Chromosomal architecture in Brassicaceae species

Repetitive sequences in plant genomes usually show specific chromosomal organization, with tandem repeats localized in spatially separated domains, while TEs have more ubiquitous chromosomal distribution (e.g. Schmidt and Heslop-Harrison, 1998; Heslop-Harrison and Schwarzacher, 2011). Tandemly repeated sequences usually constitute chromosomal heterochromatic arrays, whereas TEs, despite their frequent co-localization with tandem repeats, can intersperse throughout gene-rich euchromatic regions. The angiosperm plants with small nuclear genomes, exemplified by the arabidopsis genome, show non-uniform distribution of repetitive sequences, which are preferentially localized in heterochromatic pericentromeric regions and knobs, and mostly absent on chromosome arms (Fransz *et al.*, 1998, 2002; Lim *et al.*, 1998; Cheng *et al.*, 2001; Grob *et al.*, 2013; Simon *et al.*, 2015; Underwood *et al.*, 2017; Morata *et al.*, 2018). This distribution

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

of repetitive sequences is widespread across Brassicaceae (as indirectly evidenced by dozens of CCP analyses carried out in our laboratory), as most crucifer species possess small genomes (modal 1C-value 392 Mb; Lysak *et al.*, 2009; Hohmann *et al.*, 2015). In the *Hesperis* clade, the arabidopsis-like chromosomal architecture is characteristic of species with smallest genome sizes, i.e. *C. tenella*, *E. syriacum* and *Strigosella africana* (390 Mb; Lysak *et al.*, 2009). As nuclear genome size, average chromosome size and TE content increase in most *Hesperis* clade species (Tables 1 and 3), the longitudinal chromosomal compartmentalization disappears. In genomes larger than 1500 Mb and average chromosome sizes above 200 Mb, TEs are evenly distributed along the entire chromosome length, except for distinct subtelomeric and pericentromeric loci occupied by tandem repeats. The increasing chromosome arm lengths pose a serious challenge to centromeres to ensure a correct segregation of 'obese' chromosomes during cell division. It should be interesting to analyse whether the increasing chromosome arm length was reflected by a corresponding increase in centromere size and copy number of centromeric tandem repeats (Zhang and Dawe, 2012).

Genome size and life-form transition

There is a substantial body of evidence linking genome size, ecophysiological parameters and life-history strategies in plant species. Whereas species with small genomes can grow in more diverse habitats and can adopt any life form, species with larger genomes are confined to narrower ecological amplitudes and perenniality (Bennett, 1987; Knight *et al.*, 2005; Suda *et al.*, 2015; Pellicer *et al.*, 2018). The *Hesperis*-clade species show the statistically significant tendency of ephemeral or annual species to have small genomes, whereas species with large(er) genomes are more likely to adopt a biennial or perennial life history. Scarce C-value data are not sufficient to rigorously test this causal relationship in closely related or sister species of different life forms. The inferred correlation is found, for example, in *Bunias* and *Chorispora*. Genome size of the annual *Bunias erucago* (2083 Mb) is 0.8-fold the C-value (2585 Mb) of the perennial *Bu. orientalis* (Greilhuber and Obermayer, 1999). Whereas the annual *C. tenella* has a 342-Mb genome, the genome size of the perennial *C. bungeana*, confined to high alpine environment (2000–4200 m; Song *et al.*, 2015), is 817–830 Mb (Liu, 2017). Altogether, our data suggest that the smaller *Hesperis*-clade genomes could have been selected for, as genome downsizing enables short-lived ephemerals and annuals to adapt to time-limited habitats. For example, in the Asian cold deserts ephemeral crucifer species are an important component of the flora. In the Junggar Desert of northwest China, of the 24 ephemeral Brassicaceae species, ten belong to the *Hesperis* clade and nine are annual herbs with indehiscent or dehiscent fruits (Liu and Tan, 2007; Lu *et al.*, 2017). Among the nine taxa, small C-values are known for four species and a comparably small genome can be predicted for the remainder of the annuals. However, most *Hesperis*-clade species are biennials and perennials with large genomes. Longer life cycles of perennials, associated with genome inflation, were important in the adaptation of *Hesperis*-clade species to extreme mountain and alpine conditions, with frequent fluctuations of temperature and precipitation, long-lasting snow cover and high solar radiation (e.g. Hughes and Atchison, 2015).

The feasibility of chromosome painting

Chromosome painting based on BAC in plants is based on hybridization and subsequent visualization of non-repetitive sequences on chromosomes (Lysak *et al.*, 2003; Betekhtin *et al.*, 2014). Large-scale CCP in plants takes advantage of small genomes, such as that of arabidopsis, with euchromatic gene-rich chromosome arms and most repetitive sequences clustered within heterochromatic pericentromeres. The amplification and mobility of repeats, underlying genome upsizing, transform arabidopsis-type chromosomes into the less compartmentalized chromosomes characterizing most plant genomes (Kejnovsky *et al.*, 2009). In *Hesperis*-clade genomes with >40 % of repetitive sequences, CCP is significantly compromised or unfeasible (Mandáková *et al.*, 2017; this study) due to the changed chromosomal architecture. As genome sequences of these species are not available, we may only hypothesize that painting probes, based on single-copy coding sequences, render weaker hybridization signals as the target sequences are interspersed with abundant dispersed repeats. Moreover, heterochromatinization, including DNA methylation and histone modifications, may further hinder the accessibility of target sequences for the DNA probe.

Conclusions

The *Hesperis* clade represents a unique crucifer lineage grouping taxa with unusually large nuclear genomes and low chromosome numbers. We demonstrated that the phylogenetically shared genome obesity has not been caused by a clade-specific WGD, but by proliferation of LTR retrotransposons, initially in the *Hesperis*-clade ancestor and subsequently in taxa of the six tribes. It is assumed that the predominance of genome obesity was associated with the selection for biennial or perennial life histories. Rarely, but repeatedly, genome expansion was counteracted by purging of TEs, enabling in some species an adaptive transition to the annual life strategy. Genome downsizing versus expansion significantly impacted chromosome size and architecture of the *Hesperis*-clade species towards small and highly compartmentalized chromosomes (e.g. *C. tenella*, *E. syriacum*) versus large and less structured chromosomes (e.g. *Matthiola* and *Hesperis* spp.).

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/aob and consist of the following. Table S1: summary of chromosome number, monoploid genome size, phylogenetic and life form data available for *Hesperis*-clade species. Table S2: nucleotide sequences of consensus satellites monomers. Table S3: sequences of PCR primers designed to amplify the *gag* domain in selected LTR retrotransposons. Table S4: reconstruction of ancestral genome sizes based on ITS and *ndh*F phylogenies. Table S5: comparison of *in silico* and dot-blot estimates of repeat abundances in two *Hesperis*-clade species with contrasting genome size. Table S6: correlation between repeat amounts and genome size variation in the analysed *Hesperis*-clade species. Figure S1: self dot-plot comparison of non-homogenized monomers of

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

a satellite family with a 60-bp repetitive motif in the *E. syriacum* genome. Figure S2: phylogenetic analysis of *Ty1-copia* LTR retrotransposons based on multiple alignment of their RT domains. Figure S3: phylogenetic analysis of *Ty3-gypsy* LTR retrotransposons based on multiple alignment of their reverse transcriptase domains. Figure S4: repeat sequence proportions in the 50 largest clusters based on the comparative clustering analysis. Figure S5: dot plot showing sequence similarities between satellites identified in *Bu. orientalis*, *M. incana*, *D. micranthus*, *Brassica rapa* and *Brassica oleracea*. Figure S6: comparative chromosome painting in *Hesperis*-clade species.

ACKNOWLEDGEMENTS

We thank Dr Dmitry German (Heidelberg University) and Dr Jiří Macas (Institute of Plant Molecular Biology, České Budějovice) for their advice on the manuscript. This work was supported by research grants from the Czech Science Foundation (grants P501/12/G090 and 18-20134S), the CEITEC 2020 (grant LQ1601) project and by the Czech Academy of Sciences (long-term research development project RVO 67985939).

LITERATURE CITED

Afgan E, Baker D, Batut B, *et al.* 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* **46**: W537–W544.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

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

Beilstein MA, Al-Shehbaz IA, Kellogg EA. 2006. Brassicaceae phylogeny and trichome evolution. *American Journal of Botany* **93**: 607–619.

Bennett MD. 1987. Variation in genomic form in plants and its ecological implications. *New Phytologist* **106**: 177–200.

Bennett MD, Leitch IJ, Price HJ, Johnston JS. 2003. Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25 % larger than the *Arabidopsis* genome initiative estimate of ~125 Mb. *Annals of Botany* **91**: 547–557.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573–580.

Betekhtin A, Jenkins G, Hasterok R. 2014. Reconstructing the evolution of *Brachypodium* genomes using comparative chromosome painting. *PLoS ONE* **9**: e115108.

Chen H, Al-Shehbaz IA, Yue J, Sun H. 2018. New insights into the taxonomy of tribe Euclidieae (Brassicaceae), evidence from nrITS sequence data. *PhytoKeys* **100**: 125–139.

Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. 2013. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE* **8**: e62856.

Cheng Z, Buell CR, Wing RA, Gu M, Jiang J. 2001. Toward a cytological characterization of the rice genome. *Genome Research* **11**: 2133–2141.

Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Research* **12**: 1075–1079.

Dodsworth S, Jang TS, Struebig M, Chase MW, Weiss-Schneeweiss H, Leitch AR. 2017. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). *Plant Systematics and Evolution* **303**: 1013–1020.

Doležel J, Bartoš J, Voglmayr H, Greilhuber J. 2003. Nuclear DNA content and genome size of trout and human. *Cytometry* **51**: 127–128.

Doležel J, Greillhuber J, Suda J. 2007. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* **2**: 2233–2244.

Fransz P, Armstrong S, Alonso-Blanco C, Fischer TC, Torres-Ruiz RA, Jones G. 1998. Cytogenetics for the model system *Arabidopsis thaliana*. *Plant Journal* **13**: 867–876.

Fransz P, De Jong JH, Lysak M, Castiglione MR, Schubert I. 2002. Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. *Proceedings of the National Academy of Sciences of the USA* **99**: 14584–14589.

Gaiero P, Vaio M, Peters SA, Schranz ME, de Jong H, Speranza PR. 2018. Comparative analysis of repetitive sequences among species from the potato and the tomato clades. *Annals of Botany* **123**: 521–532.

German DA, Al-Shehbaz IA. 2017. A taxonomic note on *Sterigmostemum* and related genera (Anchonieae, Cruciferae). *Novosti Sistematicheski Vysshikh Rasteniĭ* **48**: 78–83.

German DA, Al-Shehbaz IA. 2018. A reconsideration of *Pseudofortuynia* and *Tchihatchewia* as synonyms of *Sisymbrium* and *Hesperis*, respectively (Brassicaceae). *Phytotaxa* **334**: 95–98.

German DA, Grant JR, Lysak MA, Al-Shehbaz IA. 2011. Molecular phylogeny and systematics of the tribe Chorisporeae (Brassicaceae). *Plant Systematics and Evolution* **294**: 65–86.

Greilhuber J, Obermayer R. 1999. Cryptopolyploidy in *Bunias* (Brassicaceae) revisited—a flow-cytometric and densitometric study. *Plant Systematics and Evolution* **218**: 1–4.

Greilhuber J, Borsch T, Müller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biology* **8**: 770–777.

Grob S, Schmid MW, Luedtke NW, Wicker T, Grossniklaus U. 2013. Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. *Genome Biology* **14**: R129.

Hall AE, Kettler GC, Preuss D. 2006. Dynamic evolution at pericentromeres. *Genome Research* **16**: 355–364.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2007. GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**: 129–131.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proceedings of the National Academy of Sciences of the USA* **106**: 17811–17816.

Heslop-Harrison JS, Schwarzacher T. 2011. Organisation of the plant genome in chromosomes. *Plant Journal* **66**: 18–33.

Hohmann N, Wolf EM, Lysak MA, Koch MA. 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* **27**: 2770–2784.

Huang CH, Sun R, Hu Y, *et al.* 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution* **33**: 394–412.

Hughes CE, Atchison GW. 2015. The ubiquity of alpine plant radiations: from the Andes to the Hengduan Mountains. *New Phytologist* **207**: 275–282.

Jaretzky R. 1928. Untersuchungen über Chromosomen und Phylogenie bei einigen Cruciferen. *Jahrbücher für Wissenschaftliche Botanik* **68**: 1–45.

Jiao WB, Accinelli GG, Hartwig B, *et al.* 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Research* **27**: 778–786.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

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

Kejnovsky E, Leitch IJ, Leitch AR. 2009. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends in Ecology & Evolution* **24**: 572–582.

Kiefer M, Schmickl R, German DA, *et al.* 2014. BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. *Plant and Cell Physiology* **55**: e3.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Research* **21**: 487–493.

Knight CA, Molinari NA, Petrov DA. 2005. The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany* **95**: 177–190.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**: 474.

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

Koo DH, Hong CP, Batley J, *et al.* 2011. Rapid divergence of repetitive DNAs in *Brassica* relatives. *Genomics* **97**: 173–185.
Kubešová M, Moravcova L, Suda J, Jarošík V, Pyšek P. 2010. Naturalized plants have smaller genomes than their non-invading relatives: a flow cytometric analysis of the Czech alien flora. *Preslia* **82**: 81–96.
Kubis S, Schmidt T, Heslop-Harrison JS. 1998. Repetitive DNA elements as a major component of plant genomes. *Annals of Botany* **82**: 45–55.
Lim KY, Leitch IJ, Leitch AR. 1998. Genomic characterisation and the detection of raspberry chromatin in polyploid *Rubus*. *Theoretical and Applied Genetics* **97**: 1027–1033.
Liu L. 2017. *The epigenetic modifications of* Chorispora bungeana *and the function of ADH1 in cold response*. Thesis retrieved from China Integrated Knowledge Resources Database. http://cdmd.cnki.com.cn/Article/CDMD-10730-1018803968.htm.
Liu XF, Tan DY. 2007. Diaspore characteristics and dispersal strategies of 24 ephemeral species of Brassicaceae in the Junggar Desert of China. *Journal of Plant Ecology* **31**: 1019–1027.
Lu JJ, Tan DY, Baskin CC, Baskin JM. 2017. Role of indehiscent pericarp in formation of soil seed bank in five cold desert Brassicaceae species. *Plant Ecology* **218**: 1187–1200.
Lysak MA, Lexer C. 2006. Towards the era of comparative evolutionary genomics in Brassicaceae. *Plant Systematics and Evolution* **259**: 175–198.
Lysak MA, Mandáková T. 2013. Analysis of plant meiotic chromosomes by chromosome painting. *Methods in Molecular Biology* **990**: 13–24.
Lysak MA, Pecinka A, Schubert I. 2003. Recent progress in chromosome painting of *Arabidopsis* and related species. *Chromosome Research* **11**: 195–204.
Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. 2009. The dynamic ups and downs of genome size evolution in Brassicaceae. *Molecular Biology and Evolution* **26**: 85–98.
Lysak MA, Mandáková T, Schranz ME. 2016. Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology* **30**: 108–115.
Macas J, Novak P, Pellicer J, *et al.* 2015. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe *Fabeae*. *PLoS ONE* **10**: e0143424.
Mandáaková T, Kovařík A, Zozomová-Lihová J, Shimizu-Inatsugi R, Shimizu KK, Mummenhoff K, Marhold K, Lysak MA. 2013. The more the merrier: recent hybridization and polyploidy in *Cardamine*. *Plant Cell* **25**: 3280–3295.
Mandáková T, Lysak MA. 2016*a*. Chromosome preparation for cytogenetic analyses in *Arabidopsis*. *Current Protocols in Plant Biology* **1**: 43–51.
Mandáková T, Lysak MA. 2016*b*. Painting of *Arabidopsis* chromosomes with chromosome-specific BAC clones. *Current Protocols in Plant Biology* **1**: 359–371.
Mandáková T, Hloušková P, German D, Lysak MA. 2017. Monophyletic origin and evolution of the largest crucifer genomes. *Plant Physiology* **174**: 2062–2071.
Manton I. 1932. Introduction to the general cytology of the Cruciferae. *Annals of Botany* **46**: 509–556.
Morata J, Tormo M, Alexiou KG, *et al.* 2018. The evolutionary consequences of transposon-related pericentromer expansion in melon. *Genome Biology and Evolution* **10**: 1584–1595.
Neumann P, Novák P, Hoštáková N, 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.
Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**: 792–793.
Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. 2017. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research* **45**: e111.
Otto F. 1990. DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. In: Darzynkiewicz Z, Crissman HA, eds. *Methods in Cell Biology,* Vol. **33**. New York: Academic Press, 105–110.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.
Park M, Park J, Kim S, *et al.* 2012. Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. *Plant Journal* **69**: 1018–1029.
Pearson WR, Wood T, Zhang Z, Miller W. 1997. Comparison of DNA sequences with protein sequences. *Genomics* **46**: 24–36.
Pellicer J, Hidalgo O, Dodsworth S, Leitch I. 2018. Genome size diversity and its impact on the evolution of land plants. *Genes* **9**: 88.
R Development Core Team. 2013. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.
Rambaut A, Drummond A. 2009. *Tracer v1.6*. http://tree.bio.ed.ac.uk/software/tracer.
Ren L, Huang W, Cannon EK, Bertioli DJ, Cannon SB. 2018. A mechanism for genome size reduction following genomic rearrangements. *Frontiers in Genetics* **9**: 454.
Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217–223.
Roark LM, Hui AY, Donnelly L, Birchler JA, Newton KJ. 2010. Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. *Cytogenetic and Genome Research* **129**: 17–23.
Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.* 2011. Integrative genomics viewer. *Nature Biotechnology* **29**: 24.
Ronquist F, Teslenko M, Van Der Mark P, *et al.* 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**: 539–542.
Schmidt T, Heslop-Harrison JS. 1998. Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends in Plant Science* **3**: 195–199.
Schubert I, Lysak MA. 2011. Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends in Genetics* **27**: 207–216.
Simon L, Voisin M, Tatout C, Probst AV. 2015. Structure and function of centromeric and pericentromeric heterochromatin in *Arabidopsis thaliana*. *Frontiers in Plant Science* **6**: 1049.
Song Y, Liu L, Feng Y, *et al.* 2015 Chilling- and freezing-induced alterations in cytosine methylation and its association with the cold tolerance of an alpine subnival plant, *Chorispora bungeana*. *PLoS ONE* **10**: e0135485.
Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**: GC1–GC10.
Suda J, Kyncl T, Jarolímová V. 2005. Genome size variation in Macaronesian angiosperms: forty percent of the Canarian endemic flora completed. *Plant Systematics and Evolution* **252**: 215–238.
Suda J, Meyerson LA, Leitch IJ, Pyšek P. 2015. The hidden side of plant invasions: the role of genome size. *New Phytologist* **205**: 994–1007.
Trávníček P, Ponert J, Urfus T, *et al.* 2015. Challenges of flow-cytometric estimation of nuclear genome size in orchids, a plant group with both whole-genome and progressively partial endoreplication. *Cytometry* **87A**: 958–966.
Underwood CJ, Henderson IR, Martienssen RA. 2017. Genetic and epigenetic variation of transposable elements in *Arabidopsis*. *Current Opinion in Plant Biology* **36**: 135–141.
Vu GT, Cao HX, Reiss B, Schubert I. 2017. Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytologist* **214**: 1712–1721.
Waterworth WM, Drury GE, Bray CM, West CE. 2011. Repairing breaks in the plant genome: the importance of keeping it together. *New Phytologist* **192**: 805–822.
Willing E-M, Rawat V, Mandáková T, *et al.* 2015. Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants* **1**: 14023.
Zhang H, Dawe RK. 2012. Total centromere size and genome size are strongly correlated in ten grass species. *Chromosome Research* **20**: 403–412.

Downloaded from https://academic.oup.com/aob/advance-article-abstract/doi/10.1093/aob/mcz036/5445375 by OUP Restricted Live University Test, Johnny McAdmin on 23 April 2019

SUPPLEMENTARY INFORMATION

Hloušková *et al.* **The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing**

Supplementary Data Table S1. Summary of chromosome number, monoploid genome size, phylogenetic and life form data available for *Hesperis*-clade species. C-values were estimated in this study or adopted from published resources; average values are given in case of multiple records. In case of multiple life strategies, the life form in bold was considered in correlation analyses.

Species	Tribe	ITS*	*ndh*F*	1C$_x$ (pg)	n	Life form	Reference (C$_x$-values)
Matthiola capiomontiana	Anchonieae	DQ357564	-	1.66	5	annual	BrassiBase (Kiefer *et al.*, 2014)
Matthiola fragrans	Anchonieae	-	-	2.74	?	perennial	BrassiBase (Kiefer *et al.*, 2014)
Matthiola fruticosa	Anchonieae	-	-	1.67	?	?	BrassiBase (Kiefer *et al.*, 2014)
Matthiola fruticulosa	Anchonieae	DQ357566	KF023008	1.60	6	perennial	Lysak *et al.* (2009)
Matthiola chorassanica	Anchonieae	DQ518396	-	1.72	6	perennial	Lysak *et al.* (2009)
Matthiola incana	Anchonieae	DQ249848	KY912030	2.20	7	perennial	Lysak *et al.* (2009)/this study
Matthiola longipetala	Anchonieae	FN821619	LC090138	2.03	7	**annual**/biennial	BrassiBase (Kiefer *et al.*, 2014)
Matthiola maderensis	Anchonieae	DQ249849	-	2.35	7	perennial	Suda *et al.* (2005)/Lysak *et al.* (2009)
Matthiola ovatifolia	Anchonieae	FJ687327†	-	2.93	?	perennial	BrassiBase (Kiefer *et al.*, 2014)
Matthiola sinuata	Anchonieae	KX165786	-	2.25	7	annual/**biennial**/perennial	Lysak *et al.* (2009)
Bunias orientalis	Buniadeae	FM958515	DQ288744	2.67	7	biennial/**perennial**	Greilhuber and Obermayer (1999)/ Lysak et al. (2009)/ BrassiBase (Kiefer *et al.*, 2014)/this study
Bunias erucago	Buniadeae	GQ497885‡	NC_036110	2.10		**annual**/biennial	Greilhuber and Obermayer (1999)/ Lysak et al. (2009)/
Clausia aprica	Dontostemoneae	DQ357529	-	1.99	?	perennial	BrassiBase (Kiefer *et al.*, 2014)
Dontostemon micranthus	Dontostemoneae	LK021230	KY912023	1.66	7	annual	this study
Braya humilis	Euclidieae	AY237318	NC_035515	1.63	7	perennial	Lysak *et al.* (2009)/this study

Species	Tribe	ITS*	*ndh*F*	1C$_x$ (pg)	n	Life form	Reference (C$_x$-values)
Euclidium syriacum	Euclidieae	DQ357543	DQ288767	0.26	7	annual	Lysak *et al.* (2009)/this study
Christolea crassifolia	Euclidieae	DQ523423	DQ288754	1.41	7	perennial	Lysak *et al.* (2009)
Neotorularia torulosa	Euclidieae	AY353164	-	0.58	?	annual	BrassiBase (Kiefer *et al.*, 2014)
Rhammatophyllum kamelinii	Euclidieae	DQ357587	-	1.85	7	perennial	BrassiBase (Kiefer *et al.*, 2014)
Solms-laubachia flabellata	Euclidieae	GQ424562	-	1.40	7	perennial	BrassiBase (Kiefer *et al.*, 2014)
Solms-laubachia linearis	Euclidieae	DQ523417	DQ288760	1.38	7	perennial	Lysak *et al.* (2009)
Strigosella africana	Euclidieae	AY237307	DQ288793	0.40	14	annual	Lysak *et al.* (2009)
Tetracme pamirica	Euclidieae	-	DQ288837	0.85	?	annual	Lysak *et al.* (2009)
Hesperis persica	Hesperideae	HG426449	-	4.38	8	perennial	BrassiBase (Kiefer *et al.*, 2014)
Hesperis matronalis	Hesperideae	DQ357547	DQ28877	3.98	?	biennal	BrassiBase (Kiefer *et al.*, 2014)/ Kubešová *et al.* (2010)
Hesperis sylvestris	Hesperideae	this study	NC_035512	4.36	6	biennial	this study
Hesperis tristis	Hesperideae	-	-	4.45	7	**biennial**/perennial	BrassiBase (Kiefer *et al.*, 2014)
Diptychocarpus strictus	Chorisporeae	DQ357534	DQ288762	0.27	7	annual	BrassiBase (Kiefer *et al.*, 2014)
Chorispora tenella	Chorisporeae	KJ623505	DQ288753	0.35	7	Annual	Lysak *et al.* (2009)/this study
Parrya nudicaulis	Chorisporeae	FN821534	-	1.08	7	Perennial	Lysak *et al.* (2009)

* GenBank accession numbers.

† Partial ITS sequence, not used in phylogenetic analyses.

‡ Probably a misplaced sequence, not used in phylogenetic analyses.

REFERENCES

Greilhuber J, Obermayer R. 1999. Cryptopolyploidy in *Bunias* (*Brassicaceae*) revisited—A flow-cytometric and densitometric study. *Plant Systematics and Evolution* **218:** 1-4.

Kiefer M, Schmickl R, German DA, *et al*. 2014. BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. *Plant and Cell Physiology* **55:** e3-e3.

Kubešová M, Moravcova L, Suda J, Jarošík V, Pyšek P. 2010. Naturalized plants have smaller genomes than their non-invading relatives: a flow cytometric analysis of the Czech alien flora. *Preslia* **82:** 81-96.

Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. 2009. The dynamic ups and downs of genome size evolution in Brassicaceae. *Molecular Biology and Evolution* **26:** 85-98.

Suda J, Kyncl T, Jarolímová V. 2005. Genome size variation in Macaronesian angiosperms: forty percent of the Canarian endemic flora completed. *Plant Systematics and Evolution* **252:** 215-238.

Supplementary Data Table S2. Reconstruction of ancestral genome size (ancGS) based on ITS and *ndh*F phylogenies (**Fig. 1**). AncGS values (inc. 95% confidence intervals, CI) for each tribe/node and the most recent common ancestors (MRCA) are given.

	Estimated ancGS (95% CI) (Mb)
ITS tree (Fig. 1a)	
Hesperideae	3 242.15 (2 882.08 - 3 602.22)
Chorisporeae	1 367.02 (1 006.95 - 1 727.09)
(Hesperideae – Chorisporeae)	2 231.64 (1 839.96 - 2 623.32)
(Hesperideae – Chorisporeae - Dontostemoneae)	2085.75 (1 610.99 - 2 560.52)
Dontostemoneae	1 825.74 (1 518.10 - 2 133.38)
Anchonieae	1 821.68 (1 315.02 - 2 328.34)
(Buniadeae - Euclidieae)	1 640.80 (1 115.05 - 2 166.55)
Euclidieae	1 412.53 (912.08 - 1 912.97)
MRCA	1 789.65 (1 150.84 - 2 428.47)
***ndh*F tree (Fig. 1b)**	
(Chorisporeae – Dontostemoneae)	984.54 (535.25 - 1 433.83)
Chorisporeae	530.30 (195.85 - 864.75)
(Buniadeae – Hesperideae – Anchonieae)	2 088.99 (1 604.87 - 2 573.09)
(Hesperideae – Anchonieae)	2 450.00 (2 001.91 - 2 898.06)
Hesperideae	3 750.65 (3 411.02 - 4 090.27)
Anchonieae	2 160.68 (1 752.51 - 2 568.83)
Buniadeae	2 297.74 (1 951.67 - 2 643.80)
Euclidieae	1 362.58 (875.03 - 1 850.11)
MRCA	1 523.63 (890.55 - 2 156.69)

Supplementary Data Table S3. Comparison of *in silico* and dot-blot estimates of repeat abundances in two *Hesperis*-clade species with contrasting genome sizes.

Species	Repeat	*In silico* estimate (%)	Dot-blot estimate (%)
C. tenella	ChTe2	1.60	2.80
	Athila	0.18*	0.20
H. sylvestris	HeSy1	7.38	8.00
	Athila	1.12*	1.35

* *In silico* estimates correspond to repeat clusters representing the *gag* domain for which the PCR primers were designed.

Supplementary Data Table S4. Correlation between repeat contents and genome size variation in the analyzed *Hesperis*-clade species.

Repeat family	R^2	*p*-value
LTR retroelements	0.967	$4.168E^{-5}$*
Ty1-copia	0.918	0.0007*
Ale	0.831	0.0042*
Angela	0.885	0.0016*
Bianca	0.929	0.0005*
Ivana/Oryco	0.207	0.3047
Maximus/SIRE	0.784	0.0080*
TAR	0.733	0.0139*
Tork	0.880	0.0018*
Ty3-gypsy	0.940	0.0003*
Athila	0.874	0.0020*
Chromovirus	0.990	$3.333E^{-6}$*
Ogre/Tat	0.049	0.6345
LINE	0.175	0.3500
DNA transposons	0.293	0.2097
Satellites	0.604	0.0397*
rDNA	0.133	0.4213
All repeats	0.984	$1.041E^{-5}$*

* *p*-value < 0.05

Supplementary Data Table S5. Nucleotide sequences of consensus satellites monomers.

Species	Satellite	Monomer length (bp)	Genome proportion (%)	Consensus monomer
E. syriacum	EuSy1: EuSy1A	357		GCACCATAGCTAAAACTTGTCAAACTTAGCTAACATTTGTTCACAATAGCTAAAACTTTGGCACTAGAGGTAA AAGTTATCGAACTTACCTTAAACTTCTTTACCATAACTAAATGTTTGGCACCAGAGCCAAAACTTGTAAAACT TAGCTAAACTTGTGAAACTTAGCTTAAACTTATGCACCATAGCTAAATATTTCATACCATAGTGTAAAACTTG TCAAACTTAGCTAAAACTTGTGCACCATAGCTATAACTTTGGCACTAGAGAAAAAACTTGTCAAACTTACCAA AACTTGTGCAAAACAGCTTAATGTCGCCACTAGAGCTAAACCTTGTAAAAATTAGCTAACACTTGT
	EuSy1: EuSy1B	377		CTAGCTAAGTTTGATTAGTTTTCACTTTGATGCCAATGTTTTAGTTTTGTAGCACAAGTTTTAGCTAAGTTTG ACAAGTTTTAGTTATGGTGCTCAAGTTTTACTTATTTTGACTAGGTTTAGCTATGGTGCTCAAGTTCTAGCTA TTAGCAGAAGTCTTAGCTAAGTTCTCCTAGTTCTAGATTGTGGGCCGAACTATTACCTATGGTGCACAATTGT AGGTAAGTTTGACAAGTTTTAGCCCTGGTACTAAAATCTTAGCTATTATGCACAAGTTCTAGCATAGTTTGAT AAGTCTTGGCTCTCGTGCAAGACTTTTAGCTATATAGCACAAGTTTTCGATAAGTTGGACTAGTTATAACACT GGTGCAAAAGTTTTGGCTATGGTGCATAACT
	EuSy2	81	0.32	AGGAATCTATACCCTAGACCCTCCTTTGGGAACCCAAACACATTCATATACAGTCCACGATATCTACGATACGT TTCTCTG
	EuSy3	20	0.05	CATAGCTAAAACTTGTCCAA
	EuSy4	354	0.03	CAAACAAGCGTTTGGATGATAGATCATTTTCTGTTAACTGTTAGCCAACAATTTTAGGAAAAACTCCCAATAA CCCTTGAATACCTTGTCTAACCCTTAAGAGATTTCACATTTAGTGAGAACTGTGGTCATTTATGCACTCAATTC TCATTGATATTATCATAAATAGTCCATGTGGACGATTTCAGGCCATTTTACATGAAAAATAAAAAGTTGAAAA TTGGCCCCTATATCAAGAGGTTTCTCGAGAGGATATCTTAAAATTTGGGAAAAATGGTGAAAATAACATATAT TTTCTCTGCGGATCTGTCTCCGAAACACCTAGCCACGTCATACCTACCTAGTTTTATGATT
	EuSy5	132	0.01	CAAAACCACATGTTAGTCCGGAGGAAGTCGAAACGGAAGTCACTGAGGGAGTTGCAAAGGTACCCGAAGATCC ACCTACTGCTCCACACGAGGTTCCAGGGAAAGCTGAAGGGGATGATGAAGACGGCTCTA
C. tenella	ChTe1	39	2.27	ACCGAAAAAGGTTTCTGTTGAGCTTGATACCAAAAAAAG

Species	Satellite	Monomer length (bp)	Genome proportion (%)	Consensus monomer
	ChTe2	825	1.6	CGCACGAATTTTGCACAAACATGATTATAAACTTGAATTTTGCAAAATTTTGTTATAACATGATTTTACTAGG TTTTTTCACATGTATTTTGAACATAATTCGTTTTGTACAATTTTTCTTATAAATTATCTCTACATTTTCTATTC GCACAAGTTTCTTACAATTTCTTTTTTACCCAATTGTTTTTGCACAAATCCAGATTTAACCCCAAATTTTTCAT GGTTTCGTTTGTCATGTGTTTTTAATTGATATAGTTTTACCTAAACTTTTCACTATTCCATTTTTCATGTTTTA ATAAGATCTTGTTTTCTTCGAATTTTGTATTCACACGAGTTTTGCACAGTTATAACTCTTCTGTGCAAATTCTA TAATCATGTTCTTACAAAATCGTGCGAATATAAAAGGTTTGGTAACACTCTATGTGAAAACCATATGAATCCA TGTCTAATACTTTGTTTGGAAATCCAGATTTTAGAAAGACTCGGAAAAATACACTCATGCTAATTTTGGATAA AATATTTTGTGTATTTTCGAACACATTTCAGGTATGCAGAAGCTATACAAAATAAAAATATGCAATTTTCTCG CAAAACCATATCCTTTCAGAACATGACGAAGACATAATCATGGAAAATTTGAGTTAAAATTGTATCAATTAAA AGCACTGAACAAACACATGCAAAATTTGGCTATATCAAAATTGCAAACCTGAATTTTTCATGACTATATTTTA TCTCGTCTTTCAAAATTCTGGATTTCTACTCAAAGTATTAGAAATGCTTAGATTATACGGTGTTCACAGATTT TTACCAATTTTTTATATC
	ChTe3	139	0.84	CTTTCCCGAATTTTTAAAACTTTCCTTAAAAAGAATTTTGATCAATACAAGAACATTCAAAAAAAAAAATTCAA GAAAGCCTTTGAAAATAAAAATTAATTAAAATTCAGGAAGGCCTTTCCCAAACCTTTGTTATTCGA
	ChTe4	102	0.12	CGAAGAATTTTTCCCGCATCCTAAGGCCAAGGTCCGAGACTTTTTCGATTCTCCCGGACCGAGGACCGACGAAT CACGCCAGGATGCGGTTTTTCCCGGACC
	ChTe5	28	0.04	TCTGTTGAGCTTGAGACCGAAAAAGGAT
	ChTe6	52	0.02	ACGTTTCGGTTTGACGTTTGGTTTTGTGAAGAAGAAGATGATTTAGGTTTTC
Br. humilis	BrHu1	161	0.16	CTGATTATTTGGAATCAAAATGCTCATTTAGAAAGCCCTTACCAAATGGTACAATTACTTCGTGAATCGACTT CAAATTGGCTGATATGGTACTATTACACTTCATTCCAAGTCTGGGGATAGATTAGTTCAAAAAAACACATTCA AAATCGGCTGAAAAA

Species	Satellite	Monomer length (bp)	Genome proportion (%)	Consensus monomer
	BrHu2	295	0.04	ACCATTTTCAAGCTTAGAATCCCAAAAACGGAGTTATGACCCTTCGTGATAGAAATGAGAGTGAATAGAAGTG TTGAAGTCTCAAAGTAGTTGAAGCATGCTACCAAGGTGGAGAAGCAAGTCAAAGTAAGAAACTAGCGTTATAG CTTATCCCACAGTTCTGTTAGTTCTTTGAATCTCATAGGAGTTTCTAATGGTGAATTTGGGAGTTGTAGAAAG AAGACTCATATGGAGGCTGAAACAACACTTTTTGATGTGAATATGAAGAATCAATGCTTTCATTTCCAAGGAT TTG
	BrHu3	87	0.02	CTTTCAGGGACATTTGGAGAGGTAATTAAGATTACCTCAACCGAGCCGAGGCATTTGCTTTAGATGACTTTTT ACGACTCTCGTTTTA
	BrHu5	338	0.02	TCAAAATACGGGTTATGTTACAATTTACCAAGAGAAACATTGGTCATGGTCAAACTTTCGAAAATACCCCGTT TTACAGCTAATATTTCTCGACTAAAACTTCTCAAATACATAAAGGAAGAGTGTAGGATATCTAATTGGAACTA AAATTCCCACTAAACGGTCCCTTAGTGTCACTAAACGGCACTAAACGGGTCACTACTCGTGTCACTAAAGGTGA AAAATAAAAGTTGCACAAATGATGATTTTTCTTGTCCAAATGCCCTTATATAAGTAATAAACACGATCCATCG AAAAAAATCAGGTCACTATTGGGTCACTAACCGATCAACACACGG
	BrHu6	355	0.02	GTTCTGATAACCAAAACCGTCTAGAATTCAGGCCATTTCCCTATGAGATTTCTTATTTTGAAATATGACTGTA TAAGTTGGTTTAGCACTACTAGGGCATACATTTAGGTGTATACTAACACATTTAGTAACATATTTATGTATTC TTTGAGTATGCTCAATACTTTGTTGAGAACCAAAATCGTCCATAATCGTCCATAATCGTCCACAATCGTCCAGG ATTCAAGCCCTTTCCCTATAAAATTTCAAATTTTGAAATATGACTCTATAAGTTGTTTCGCCACTATTAGGGC ATACTATTAGATGTATACAAACATATTTTGTAATATATTCGACTATTCAATACCGTTTGATC
D. micranthus	DoMi1	36	0.3	TTTGTAGTCTGTTTTTATAGGGGTAGCATCACAGAA
	DoMi2	143	0.06	TGTTCTCGCAGTGACGGAAACTCGGCCCTGGAGCTTCCGGTTTTGAAGTCTTTGGGTGTCCGGAGGAGTACAGA ACTAGTAGAACAAACCGAGCATGGTATCCTGTCCAGTCTCCTGGTCCGTTAGTGGTGCCATGTTCAACC
	DoMi3	350	0.05	TGTAATTTTGAAAATTTTCACGTCATTGTTTTAATGCATAGCTGTATCCACCTTGGATGTGGGAACTTGACGA AAGAGGTTGCAGTCTCTGATTGTCCTCGATTATACAAGAAATGGAGGTTTAACTGCTGTAAAACCACCTAGTT CGATGGTATCTAGAGATTTGTCAAATTACTTGTTAAAAAATTAAGAAACATGCCTCATGACCATCAAAACTGG TATAGTGAACATTAATATAGACTTAGGATGTGTTTAGAAAATCTTAGAATGGACTAAAGGTGATAAAAATAC

Species	Satellite	Monomer length (bp)	Genome proportion (%)	Consensus monomer
				TTATTTACCTGATTTATCGAGGAAACCGAGTTTTTTGTAGGGGATATAGCTCACAGAGCG
	DoMi4	26	0.03	AGAATTTTCTGTGATGCACCCCCTAG
	DoMi5	354	0.02	ATAAAATCACCCGGTTAACACTTAGGGTATCTAAAAATCTGGAGAAAATGCTTTAAATAATCAATATTTTACA TTAGGATCTATCTCCGGAACACCTAGCCACGTTAAACCTACCTAGTATTATGATTCAAAGTAGTTTTTGGAGCT TAGGTCCTTTAGGTTAGCTATTTAGCCAACAATTATAGGAAAAACACCTAATAACCCTTTAATACCTCGCCTA ACTCCTATGAGATTTCACTTTTACTGAGGAGTGGGGACATTTATGCACCCAATTCTCGTTGATATCATCATTTA TAATCGTTGTGGACGGTTTCACGTTGTTTTATCTAAAAAATTACACATTTCAAAATTCTC
	DoMi6	182	0.01	AGTCTTTATGGGTTCTTACACTCGCGGTTCTCACACCCCATACCCTCGTGTCCAGGTTTTAAGTCTTCATGAGT TCTTACATCCACACATCTCCAACCTCTCACGTCCAAGCTTTATGTTTTCATCGATTTTCACACCCTCATGATTC AAGACTTCATAGGTTCTCACATCTCCAGGATACA
M. incana	MaIn1	352	0.58	CTTTCTACATTTGTATTAATGTATGCTAAACTTTATTTAGTGGTAAATGGTAATTATTTCTAAGTTTTTAACC ATTAATTTCATAAAACTTCGTATACACCCTAAATACGTGTAGTAACCGCTATGAAATCACTTTTATCCATAGA AACGGAGACAGCGAAGGTCCCAAACCACTTTGACCATCATCCTATGTCATTGCACGATACAAATATGCATTAA AACAAGATTTTCATTTTTTTGCAGTTTTGGTCAAAATCAGTGGGTTAACCCCTACGAAAATATCGAGTTTTTC GGTAAATGCTCTATATACTGAATTATATAGTCTTTAGCGTTGTTTAAAAAATTCTAACCA
	MaIn2	355	0.1	TTTGAATCATAAAACTAGGTAGGTTTAGCGTGGCTACGTGTTCCGGAGGAAGGTCTATGTGGAAAATATTGAT TATTTTAAGCATTTTCGCCAGATTTCGAGATATCATTAGGCCTAAAGAGGTGGTTTAAGGAGATTTTTCAAAA CTTCAATTTTTTGGATAAAACAATGCGAAACTGTCCAAAATGACTAGAAACGATGATATCAACGAGAATTGGG TGCATAAATGTCCCTACTCCTCCTTAAAAGCATAATCACATAGGGGTTAGGGAAGGTGTTTAAGGGTTATTCG ATGTTTTTCCTATAATTGTTGGCTAAACAGGTAAGAGAAAGGAACTAATGGCCCAAAAACTGG
	MaIn3	69	0.08	ATCTCCGAATCGCGCCATACTTAGTCAAATAAGCCGAGTTTCGCCTCGCAAATCGTATGGAGTAAGCCG

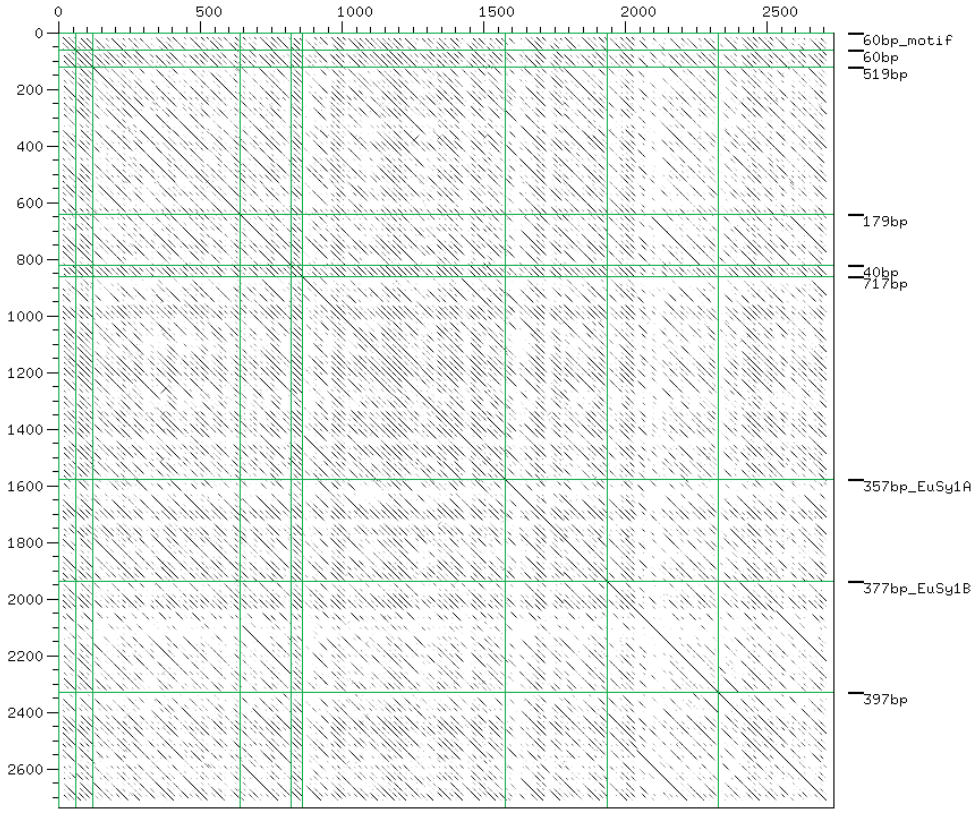
Species	Satellite	Monomer length (bp)	Genome proportion (%)	Consensus monomer
	MaIn4	88	0.06	TGTAAAAATAACAATCGCAGTTGGCTACGTGCACCAGTTAGGGTGTATGAATAATGCTTGGGTTAAGGCATTA CCCTCTAGGGTATAG
	MaIn5	590	0.05	AAGTAACCAAAACCCACTAAACAACGAGTTTCAAATCAATTTAACAAACGCTAGACTTACGATTGTTTCTTAG GATATCCTAATTACAATAGAAATCTAAGGTGAGTTTGTTTTAGCACCGCTCTCGAACACAAACAATGGTATAG TGAACCCGCTTCCCTTGTGAACCCTAACTATCCTTAGATAATAGAAAATCAAATTCCTTTGTGAATTCTTAATA TCTAAGCATGCATTACCTACAAGTTTGCTTTAAACAAATCCAAAATAGAAGGTTTGGGGTTTTGGAATGAAAT GGTTGATAAAGAGTAGCTTGATGAGTTGTTTCAAGAAGTAAAAGGGTATTGTGATGAAAAAGTCGACGAGGA TGAAAAAAAGAAAGAAAATATTATAACGTCATGATTTCGAGGAATGGTAAATAGGTAAAGAGCTAAGGTTTG AGTAAGTACACATCAAAACTAGTTTAAAAGGTAACGATTAAGGTTAGGAAAGAAGTTGGCATATCCAATAGT AATTAGTACGATCAAGAGAGTGAAAATTCTAGATTCAACTCAAACTAAATGTAAAAGATTTAAAAGTAAAAG CAAATAACT
Bu. orientalis	BuOr1	352	0.36	ACGAAAATATCGAGTTTTTCGATAAATGCTCTATATAATGAAATTTATCCTCTTTAGTGTTGTTTTAAAATGC ATAACCACATTCTACACTTATATTAATGTACCCTAAGCTTGTTTTCATGATAAATGGGCATCGTTTCTAGTTT ATTAACAAGTAATTCAATAAAACTTCATATACTGCGTAAATCAGAGGAGTAACCCCTATGAAATCTCAATTTT CTATAGAAACGGGGACAATTAAGGTCCCAAACCTCTTTGATCTTCATCCTATGACAGTGCTGGATGCTACTAT GCATTAAAACAAAAATATCAAATTTCTCAAAATTATCTCAAAATCCGTGGGCTAACCCCT
	BuOr2	192	0.18	TGAAAATCATTCCCGAACGGTTAGTTATCCGGTGATAGAGCTTCTAAAATATTTGAAAGCACGTTTTTCAATG TCTGATGCTCTTTCCTAACTCGAAACATCTTTCGAAAGAGTACTAAGCGGAACAAATGAAAGAGCTTTAAGCG ATTTAGGTAGGTTTTTTAAGAATTCGGGAAAGGGTTCCCAAATTTT
	BuOr3	179	0.1	CGAACCGAAAACCTTGAAAAACGGACAAATCACACAAACTCAAGAGACGAAGAGACTAATTAATCACTAAGAG ACGATTTCTCCAGTAGCAGGACGATCGTATTTGATTCTGAGAAATTCGACGCGCGTACCTTGTTCAGCTTAGGA GAAATCGCACGAAGCAGAGAGAAATACGCGAA
	BuOr4	20	0.09	ACAACTTTTATCTCCCGTGG

Species	Satellite	Monomer length (bp)	Genome proportion (%)	Consensus monomer
	BuOr6	171	0.01	GTCAAAACTACAATAAATAGGATTATGGGCTTTATGTGAATTTAATGTTCAACAGGTTTTCTTCTCTTTTAAG ACAAAGAGAAGTACTTGTCTTCTTTTTATCCTAAATGATGAAATGAAAAATATAAAATCTCGTGGACATATTT TTAAAAGTTTATTGTGTTTTGTCTG
	BuOr7	77	0.01	GGTCGCCCGAAAACCCTAAATTATTTTCGGGTTCAGACGCACCCAGCGGTCGCTGGCAAACTTGCCCAGCGCTC GCT
	BuOr8	177	0.01	AGATTCAAGAGACTAGGTTGTTTTACGAAGAAATGGTTGTCAGAGAAGTGTGGGCTGGACCGGAGCTCGCTGG TGAGGATGATTCGTATGTAGAAAGTTACATAAGCACTATTGGAGTCGATGAACACCAAACTGTGGACTTTGTT GAACCTGCCTTCGAACCTCTCAATATGAATC
	BuOr9	490	0.01	TATCCTTCTAAGTGTGGTAAAAAGTAATATAATGGAACACATTGGTATCTCCACCAAGTTCAAATTATTGGGT AAAGAGAAATCATTGTACTATTGCAGAGCAATAGCAACGTCATGTTTCAGCAACTCAAAGTTGTACAAATCAG AGCAAATTGAGTGGCCTATAATGGAACACATTGGTATGTTTTAGTCCTTGTGTTTTAGGGTTTCCATCAATTT CCTTGTCTATTGAAGCATTTGTGTTGTTAAGAGAAGACGTGTGGAACATATAATCAGTAAAAAAAATTATTAA GTATTTTTCTGGTAGATCTGAAGACTGAACAAAGGTAAGAAGATTCTTCAACTCTTGAATATTTTCCTGGTAG ATCCGACATTTTTTTAATAAATTTATCATTAAAAAGTTTTATTTTCTCTATTGTTAAATTCATTTGTCTTTTT TAATAATTGTTTTTGAGTATTGCTCAATTTGATCATATAATTCAACTCTTGA
H. sylvestris	HeSy1	91	7.38	AGAGCCATAATCTACATTATGTTTCAAGTAAAGCATGTGTAAACTACATGAAACTAATGAAAACCTAATTATA CGTAAACATGTTATTTAT
	HeSy2	161	0.69	AGACAAACAAACATCAGATGTCCCGTGCAAAACTCGGGTATCCCAAATCTGCACTTAATGATCCGTGCAAATC TCGTGCATGATCATATTGTGCAGAATAACAAACAAACAGCAGAAAATCTTGCTAATTTATAACATACTCATGA AACAACAGTATGTAA
	HeSy3	91	0.08	GTTATTTATAAAGCCATAACCTAGATTATATTTAAAAACCACTATGTATAAACTACTTTAAACTAATCAAACC CTAATTATACGTAAATAT

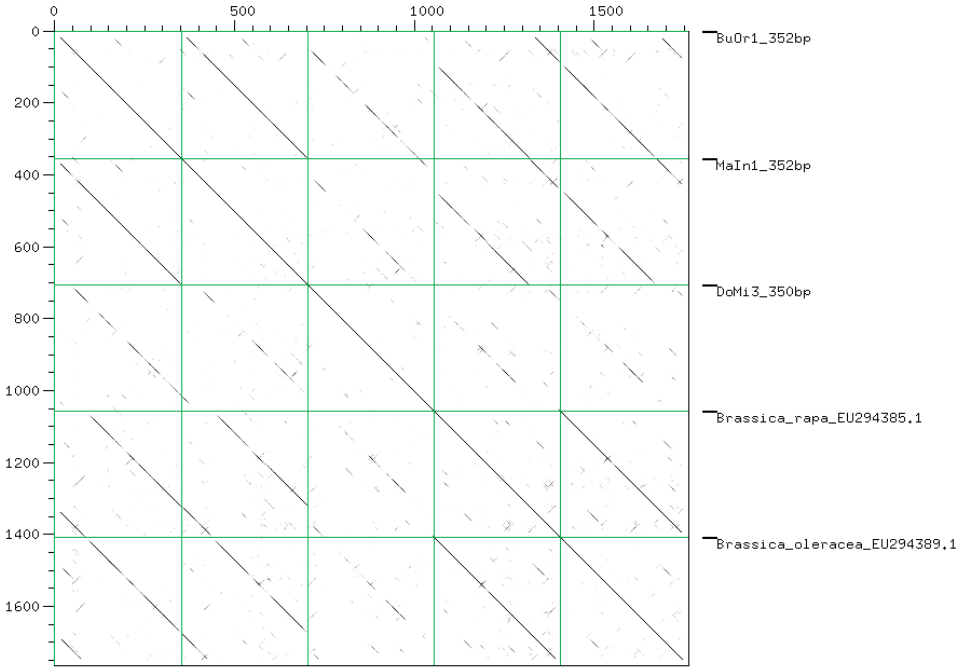
Species	Satellite	Monomer length (bp)	Genome proportion (%)	Consensus monomer
	HeSy4	200	0.07	AGCATCTTACAACTACTCGAACCGAGATGTTTCAACTCGGTGAAAATACACTACGTAGATTTTTCAGGAAGGC TTTCCCGAATTAGTGTTTACCCTTCCCGAATTTTCAATTACTCATTATTTTATCCATGAAACTTTCACTAAGGC GCATACAAAGTGTTTTTAAATACCAGTTTCGGCTTTTGCAACCGAAAATAAAA
	HeSy5	174	0.06	AGAAACTCATAGGAGTTTGTCCACGACCTCTGAAAATTTAGGGCGTTTTGACGTAATTTCCTGATAATGCACG GAACATGTCAGAATGACAATTAAGGTGTTTCAACACTTTAACTTGCTAAAAGAGACTCTATAAAATACCACTA ACACAAGTTCAAGGTTTCAAAAGTTTTC

Supplementary Data Table S6. Sequences of PCR primers designed to amplified the *gag* domain of the selected LTR-retrotransposons.

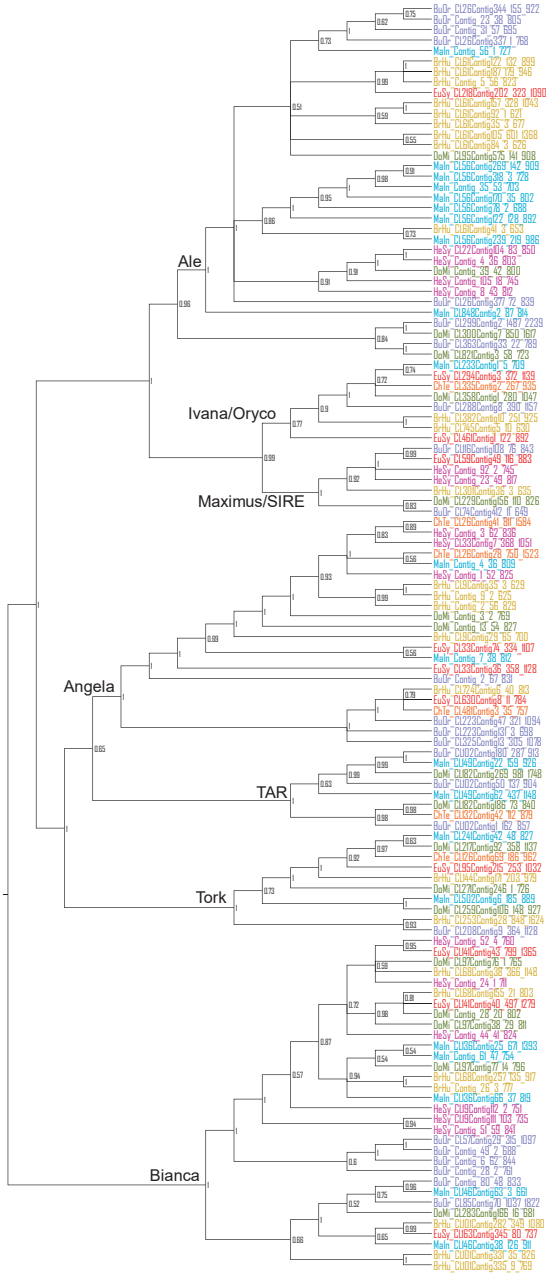
Species		PCR primer
E. syriacum	Angela	TCGACACCATCACTGTGTCG
		TGCCTTATGCTGAGTCGAGA
	Athila	ATGATTCTGGGTCGGATCGG
		GACTTAGACCGCTGCATGGT
	Chromovirus/CRM	GGTGGTGAAAGTTGCGGTTT
		TTCGCACCCCTTGGAAGTTT
	Ogre/Tat	TTCCAAGCCCGAGGAACATC
		CATGCCCTACTTCGCATCCA
C. tenella	Angela	TGACATGACGACTCACTGGA
		ACTTGCGGGAGCAATATGAG
	Athila	CGAAGAAGGCATTTTTGGTG
		GGTGCTCAGCAATTGAGGTT
Br. humilis	Angela	GAGCATTTCATGCAGCTCAG
		GCCAGACTTTGAGGCATCTC
	Athila	GCCAAGCTCTTCCATCAAAA
		ACCATGATGAGGGCATTGAT
D. micranthus	Angela	CATGAACTCGGACTTGCAGA
		CAATCCAGTTTTGCCAGGTT
	Athila	GCTCCTAAATCGGTCCCAAG
		AGTATCCGAGGACGCTCTCA
M. incana	Angela	ATCATTGCGTCCCTGAAGAG
		ACTAGGAGGCAGCGACTGAA
	Athila	TACTTGGGAGGAGACGAGGA
		GTCGGTGAATCCATGATGTG
Bu. orientalis	Angela	TTGATGTTAGGCTCAGCTGTCT
		GGACTGTCCCATTAGCCAAG
	Athila	GACTGGCTGAGGTCACTTCC
		TTGAACCTGGTGTACGCATC
H. sylvestris	Angela	GATCATAACACGCAGGCAGA
		TCCAGGAACAGGCTCGTACT
	Bianca	ATGCTGCCACAGGCTAAGTT
		GCACGGTACTGTTGCTGAAG
	Athila	CGCTAGCGCACATCAATAGT
		GAAGCTCGAGATGTCGTTCC
	Chromovirus/Tekay	CAACGAAGTCTCGCCTTAGC
		CCTTGCTGCAATTGACGATA



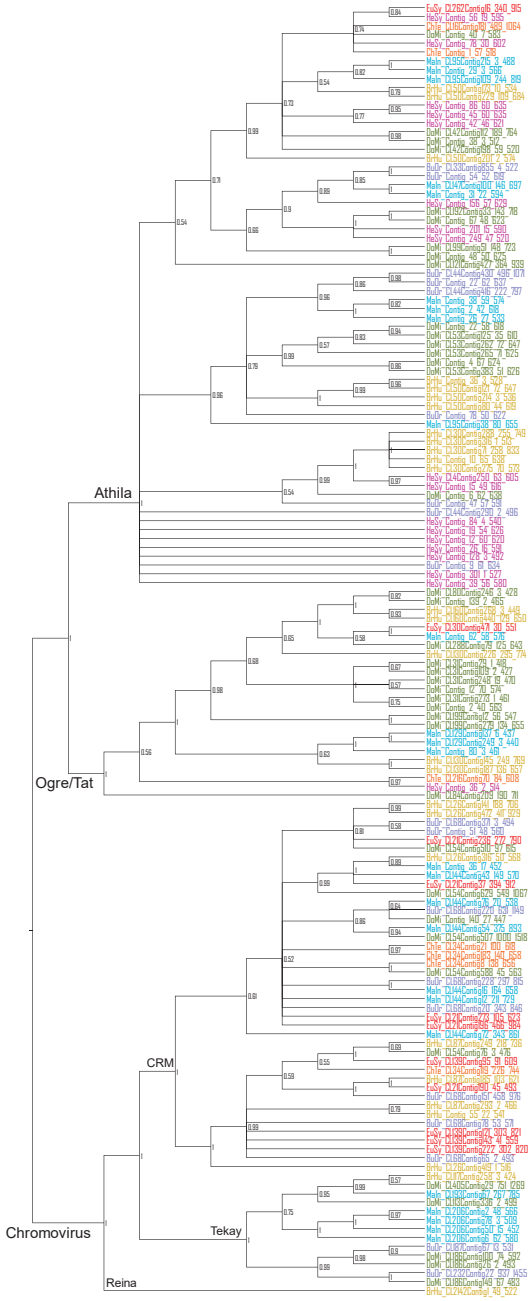
Supplementary Data Figure S1. Self dot-plot comparison of non-homogenized monomers of satellite family (EuSy1) with a 60-bp repetitive motif in the *E. syriacum*genome.



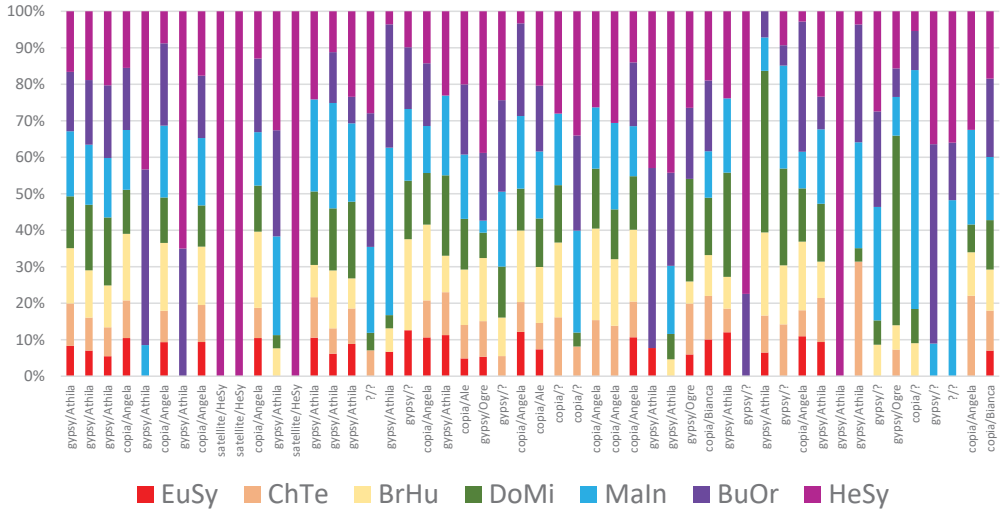
Supplementary Data Figure S2. Dot-plot showing sequence similarities between satellites identified in *Bu. orientalis* (BuOr1), *M. incana* (MaIn1), *D. micranthus* (DoMi3), *Brassica rapa* (EU294385.1) and *B. oleracea* (EU294389.1).



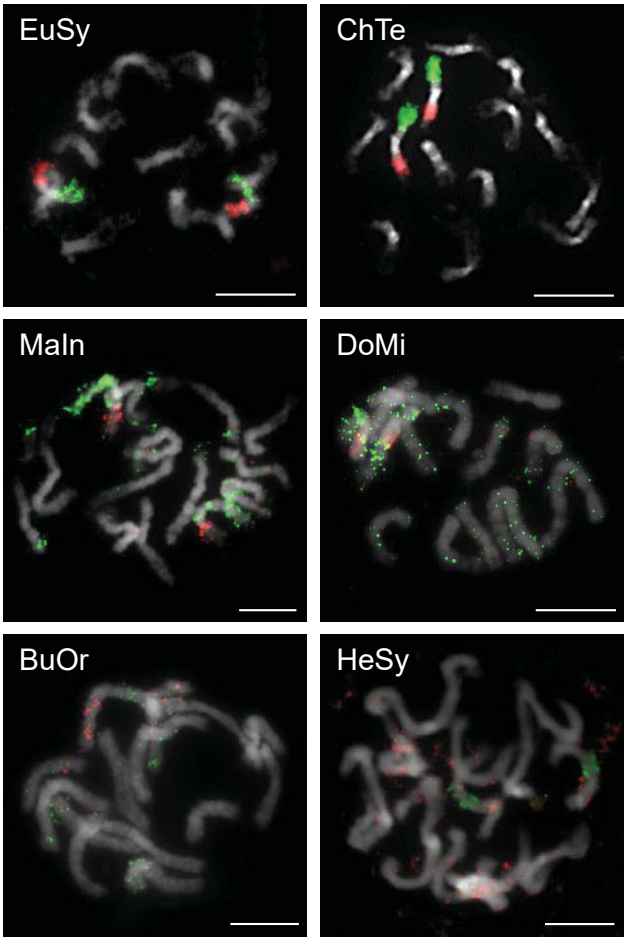
Supplementary Data Figure S3. Bayesian phylogenetic analysis of *Ty1-copia* LTR-retrotransposons based on multiple alignment of their reverse transcriptase domains. Ale, Angela, Bianca, Ivana/Oryco, Maximus/SIRE, TAR and Tork lineages formed well-distinguished and supported monophyletic clades. Analyzed species were distinguished by colors.



Supplementary Data Figure S4. Bayesian phylogenetic analysis of *Ty3-gypsy* LTR-retrotransposons based on multiple alignment of their reverse transcriptase domains. Athila, Ogre/Tat and Chromovirus lineages formed well-distinguished and supported monophyletic clades. Analyzed species are distinguished by colors.



Supplementary Data Figure S5. Repeat sequence proportions in the 50 largest clusters based on the comparative clustering analysis. The seven analyzed species are represented by different colors. "?" = an unknown repeat sequence.



Supplementary Data Figure S6. Comparative chromosome painting (CCP) in *Hesperis*-clade species. Identification of genomic blocks Jb (red fluorescence) and M (green fluorescence) by CCP using arabidopsis BAC contigs on mitotic chromosomes of *E. syriacum* (EuSy), *C. tenella* (ChTe), *M. incana* (MaIn), *D. micranthus* (DoMi), *Bu. orientalis* (BuOr) and *H. sylvestris* (HeSy). Scale bars, 10 μm.

3.3 GENOME EVOLUTION IN ARABIDEAE WAS MARKED BY FREQUENT CENTROMERE REPOSITIONING

Mandáková T, **Hloušková P**, Koch MA, Lysak MA. (2020). Genome evolution in Arabideae was marked by frequent centromere repositioning. *Plant Cell*. 32(3), 650-665.

PH performed the bioinformatics analyses of the NGS data, identified repetitive sequences, design the oligoprobes and primers further used as cytogenetic probes, performed comparative genome analysis, and participated on writing of the respective parts of the manuscript.

Summary

Centromere repositioning refers to a *de novo* centromere formation in an alternative position on the same chromosome with conserved chromosomal collinearity. Compared to animals, there are less reports on centromere repositioning in plant species.

Tribe Arabideae exhibits notable uniformity in chromosome numbers and genome structure. However, in the absence of gross chromosomal rearrangements, chromosomes may still be differentiated through centromere repositioning.

Comparative cytogenomic maps of the investigated 10 Arabideae species were constructed by multicolor comparative chromosome painting using chromosome specific repeat-free BAC contigs of *Arabidopsis thaliana*. Low-pass whole-genome Illumina sequence data of eight Arabideae species and publicly available genomic data of *A. alpina* and *A. montbretiana* were analyzed by the RepeatExplorer and TAREAN pipelines to identify the most abundant tandem repeats. Only tandem repeats constituting at least 0.01 % of a genome were further tested by FISH as potential chromosomal landmarks identifying (peri)centromeric regions in the analyzed species.

Positional shift was observed on five out of six chromosomes with conserved chromosomal collinearity in the crown-group Arabideae species. Precise

characterization of the repositioned centromeres was carried out by chromosome painting using a BAC by BAC approach along with FISH localization of newly identified centromere-associated tandem repeats. To further corroborate centromere repositioning events at sequence level in Arabideae genomes, we verified intergenome conserved collinearity and the absence of inversions between *A. alpina* and *A. lyrata* assembled genomes.

The Plant Cell, Vol. 32: 650–665, March 2020, www.plantcell.org © 2020 ASPB.

Check for updates

Genome Evolution in Arabideae Was Marked by Frequent Centromere Repositioning

Terezie Mandáková,[a] Petra Hloušková,[a] Marcus A. Koch,[b] and Martin A. Lysak[a,1]

[a] Central European Institute of Technology (CEITEC) and Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic
[b] Centre for Organismal Studies (COS) Heidelberg, Biodiversity and Plant Systematics/Botanical Garden and Herbarium (HEID), Heidelberg University, Heidelberg, Germany

ORCID IDs: 0000-0001-6485-0563 (T.M.); 0000-0003-4121-2547 (P.H.); 0000-0002-1693-6829 (M.A.K.); 0000-0003-0318-4194 (M.A.L.)

Centromere position may change despite conserved chromosomal collinearity. Centromere repositioning and evolutionary new centromeres (ENCs) were frequently encountered during vertebrate genome evolution but only rarely observed in plants. The largest crucifer tribe, Arabideae (~550 species; Brassicaceae, the mustard family), diversified into several well-defined subclades in the virtual absence of chromosome number variation. Bacterial artificial chromosome–based comparative chromosome painting uncovered a constancy of genome structures among 10 analyzed genomes representing seven Arabideae subclades classified as four genera: *Arabis*, *Aubrieta*, *Draba*, and *Pseudoturritis*. Interestingly, the intra-tribal diversification was marked by a high frequency of ENCs on five of the eight homoeologous chromosomes in the crown-group genera, but not in the most ancestral *Pseudoturritis* genome. From the 32 documented ENCs, at least 26 originated independently, including 4 ENCs recurrently formed at the same position in not closely related species. While chromosomal localization of ENCs does not reflect the phylogenetic position of the Arabideae subclades, centromere seeding was usually confined to long chromosome arms, transforming acrocentric chromosomes to (sub)metacentric chromosomes. Centromere repositioning is proposed as the key mechanism differentiating overall conserved homoeologous chromosomes across the crown-group Arabideae subclades. The evolutionary significance of centromere repositioning is discussed in the context of possible adaptive effects on recombination and epigenetic regulation of gene expression.

INTRODUCTION

Although the role of chromosomal rearrangements as a reproductive barrier promoting speciation is frequently discussed (White, 1978; Grant, 1981; Rieseberg, 2001; Levin, 2002), the significance and patterns of chromosomal speciation might differ between various, even closely related, groups or clades. Comparative genomic studies indicate that besides groups with genomes differentiated by distinct chromosome numbers and gross chromosomal rearrangements, there are also clades exhibiting striking uniformity in chromosome numbers and genome structure. Apparently, clades may differ in the frequency of chromosomal rearrangements and probability of their fixation, with a higher or lower intensity of purifying natural selection acting on carriers of chromosomal rearrangements. The scale of chromosomal rearrangements can vary, too. The lack of large-scale rearrangements, such as kilobase- to megabase-sized inversions or translocations, does not mean that the genomes cannot be differentiated through innumerable small inversions, insertions, and deletions (indels). Moreover, in the absence of gross chromosomal rearrangements, chromosomes may still be differentiated through centromere repositioning, observed for the first time in primates some 20 years ago (Montefalcone et al., 1999).

[1] Address correspondence to martin.lysak@ceitec.muni.cz.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Martin A. Lysak (martin.lysak@ceitec.muni.cz).
www.plantcell.org/cgi/doi/10.1105/tpc.19.00557

Centromere repositioning refers to de novo centromere formation in a different position on the same chromosome. The currently accepted model of centromere repositioning assumes formation of a new functional centromere (centromere seeding), containing nucleosomes with the centromere-specific histone H3 variant CenH3 and a concurrent or very fast decay of the old centromere (Rocchi et al., 2012; Schneider et al., 2016; Chiatante et al., 2017; Comai et al., 2017; Tolomeo et al., 2017; Schubert, 2018). The process usually includes the depletion of centromere-specific or -associated repeats in the inactive centromere, whereas the newly formed centromeres tend to be repeat poor initially but are colonized by repeats over the long term. Although centromere repositioning was observed repeatedly and the overall picture seems to be convergent across different eukaryotic lineages, the mechanism of de novo centromere formation and decay remains elusive. The most baffling puzzle is the timing of these events, namely, how the cell avoids potential problematic chromosome division in the presence of an acentric or dicentric chromosome (for a recent review, see Schubert, 2018).

Centromere repositioning, as one of the mechanisms of chromosomal evolution, has been deduced from complete chromosome collinearity between two individuals or species, but differently positioned centromeres (Montefalcone et al., 1999; Rocchi et al., 2012). In phylogenetic contexts, a repositioned centromere is known as an evolutionary new centromere (ENC; Ventura et al., 2004). Several reports on ENCs in primates (Montefalcone et al., 1999; Ventura et al., 2007; Chiatante et al., 2017; Tolomeo et al., 2017) and other vertebrates (Carbone et al., 2006; Piras et al., 2010; Rocchi et al., 2012) suggest centromere

IN A NUTSHELL

Background: The centromere is a chromosomal structure that ensures sister-chromatid cohesion and regular chromosome segregation during mitosis and meiosis. Despite their crucial importance, the physical position of centromeres may change without apparent chromosomal rearrangements corrupting chromosomal collinearity. These evolutionary new centromeres (ENCs) have repeatedly been documented in yeast, flies (*Drosophila*) and vertebrates (e.g., in equids and primates), and less frequently in plants. Triggers, mechanism and evolutionary consequences of ENC formation in diverse eukaryotic phyla are far from being understood.

Question: The unexpected finding of centromere relocation despite the conserved chromosomal collinearity, previously identified in the sequenced Alpine rockcress (*Arabis alpina*) genome, made us wonder whether the phenomenon is a common feature underlying genome evolution in the whole Arabideae clade from the mustard family (Brassicaceae). We analyzed this by comparative chromosome painting and using the identified centromere-associated tandem repeats.

Findings: We found that the intra-tribal cladogenesis in Arabideae (c. 550 species in 18 genera) was marked by a high frequency of ENCs on five out of the eight homoeologous chromosomes in the crown-group genera, but not in the most ancestral *Pseudoturritis* genome. The high centromere mobility was contrasted by the absence of gross chromosomal rearrangements. While chromosomal localization of ENCs does not reflect the phylogenetic relationship of the Arabideae subclades, centromere repositioning is proposed as the key mechanism differentiating overall conserved homoeologous chromosomes across the crown-group subclades. Our study documented frequent centromere repositioning in the largest species set within a single monophyletic plant clade.

Next steps: Researchers aim to understand what triggers centromere repositioning and whether the formation of ENCs may confer adaptive advantage to their carriers. The role of ENCs as a post-zygotic reproductive barrier in Arabideae is testable by analyzing genomes of closely related (sub)species pairs differentiated by centromere repositioning, and their natural or artificial hybrids. Our work shows Arabideae as a suitable model group to analyze the evolutionary significance of centromere repositioning in plant genome evolution.

repositioning as an important mechanism of chromosomal evolution. A particularly high frequency of ENCs has occurred during equine genome evolution (genus *Equus*: donkeys, horses, and zebras; Carbone et al., 2006; Piras et al., 2010; Huang et al., 2015) and in recent human and primate evolution (Ventura et al., 2007; Rocchi et al., 2012; Chiatante et al., 2017). More recently, inter-species comparison of several high-quality reference genomes in the *Drosophila obscura* group revealed that in these flies telocentric chromosomes were transformed to metacentric chromosomes most likely due to centromere repositioning, and not through inversions as traditionally believed (Bracewell et al., 2019).

Compared to animals, there are fewer reports on centromere repositioning in plant species. By fluorescence in situ hybridization (FISH) mapping, Han et al. (2009) observed different centromere positions between two pairs of chromosomes in cucumber (*Cucumis sativus*) and melon (*Cucumis melo*), although a subsequent comparative analysis, based on sequence data, concluded that one of the shifts most likely resulted from multiple chromosomal rearrangements (Yang et al., 2014). An inter-species comparison of pericentromere sequences among two *Oryza* (rice) genomes and the outgroup genome of *Leersia perrieri* revealed a centromere repositioning event on chromosome 12 in *Oryza brachyantha*, moving the centromere ~400 kb away (Liao et al., 2018). De novo centromere formation and centromere inactivation have been studied most comprehensively in maize (*Zea mays*). In oat (*Avena sativa*) × maize hybrid lines, centromere repositioning (~16 Mb away from its original location) and centromere size expansion were observed on one of the eight maize chromosomes after transfer to the oat background (Wang et al., 2014). Liu et al. (2015) showed that a newly formed centromere on an engineered maize chromosome can become inactive again, and de novo centromeres are formed elsewhere on that chromosome. However, some regions can be (more) prone to de novo centromere formation, as documented by at least three independent centromere formation events within the latent region 2 Mb away from the original centromere on maize chromosome 3 (Zhao et al., 2017). In a large-scale study, Schneider et al. (2016) uncovered the high centromere mobility after maize domestication. In more than 20 lines of domesticated maize, they documented 57 independent centromere shifts associated with decay of the original centromeres. The new centromeres originated by repositioning of CenH3 either by expansion to nearby regions or through hemicentric inversions. A hemicentric inversion, that is, an inversion with one breakpoint in the centromere, was shown to mediate positional shift of the kinetochore-forming region on maize chromosome 8 (Lamb et al., 2007). Later studies revealed that centromere repositioning caused by hemicentric inversions was frequent in maize, during the origin and early evolution of the maize genome (Wang and Bennetzen, 2012; Wolfgruber et al., 2016) and after maize domestication (Schneider et al., 2016), and probably less frequent in rice genomes (Liao et al., 2018).

Whole-genome sequencing and comparative painting analysis in the Alpine rockcress (*Arabis alpina*, Brassicaceae) genome identified centromere repositioning on three of eight chromosomes in comparison with a parsimoniously inferred ancestral genome (Willing et al., 2015). The unexpected finding of ENCs in *A. alpina* made us wonder whether the phenomenon is limited only to this species or whether it is a more common feature underlying genome evolution in the whole Arabideae clade. The tribe

Arabideae is the largest tribe of the family Brassicaceae or Cruciferae (mustard family), containing at least 550 species (13.7% of the family) in 18 genera (BrassiBase: https://brassibase.cos.uni-heidelberg.de/), including *Pseudoturritis turrita* (Kiefer et al., 2019; M.A. Koch, unpublished data). While *Arabis* within tribe Arabideae has ~100 species, including several paraphyletic species assemblages, the genus *Draba* (~400 species) is the largest monophyletic group in Arabideae and the entire family (Jordon-Thaden et al., 2010). The remaining 16 genera are smaller, and some are mono- or oligotypic, with only one or a few species (Al-Shehbaz, 2012; Karl and Koch, 2013; Huang et al., 2019). All major genera and clades are well defined (Jordon-Thaden et al., 2010; Karl et al., 2012; Koch et al., 2012, 2017; Karl and Koch, 2014), and there is a backbone phylogeny available for tribal-wide evolutionary inferences (Karl and Koch, 2013; Kiefer et al., 2017).

Here, we aimed to characterize chromosome structure and centromere positions across the Arabideae by chromosome-specific painting probes and centromere-associated tandem repeats. To this end, we have identified the most ancestral Arabideae genome and subsequently constructed comparative genomic maps for species representing the main Arabideae subclades. We asked how frequently ENCs formed during the diversification of the Arabideae and to what extent is the centromere mobility associated with the intra-tribal relationships and other characteristics of these plant genomes. In the absence of gross chromosomal rearrangements, centromere repositioning is proposed as the key mechanism differentiating otherwise collinear chromosomes of Arabideae species.

RESULTS

Structural Stasis of Arabideae Genomes

Comparative cytogenomic maps of the investigated Arabideae species (Supplemental Table 1) were constructed by multicolor comparative chromosome painting using chromosome-specific bacterial artificial chromosome (BAC) contigs of Arabidopsis (*Arabidopsis thaliana*). The differentially labeled BAC contigs were successfully used as painting probes to identify all 22 ancestral genomic blocks of crucifer genomes (Lysak et al., 2016) in the Arabideae genomes analyzed (Figures 1 and 2).

As *P. turrita* has the most ancestral position in the tribe (Figure 1), its genome structure was inferred prior to analyzing genomes from younger Arabideae clades. Next, the eight chromosomes of *P. turrita* (Pt1 to Pt8; Figures 1 and 2) were used as the reference for reconstructing chromosomal evolution in the remaining Arabideae genomes.

Comparisons of the *P. turrita* genome with genome structure of the remaining Arabideae species allowed us to infer an ancestral crown-group Arabideae genome with eight chromosomes (Ar1 to Ar8; Figures 1 and 2). Whereas chromosomes Ar4 and Ar6 differ from their ancestral homoeologues (Pt4 and Pt6) by two reciprocal translocations, six chromosomes of *P. turrita* (Pt1 to Pt3, Pt5, Pt7, and Pt8) are collinear with the six Ar chromosomes (Ar1 to Ar3, Ar5, Ar7, and Ar8). Despite the perfect collinearity of genomic blocks, five homoeologues (Pt1/Ar1, Pt2/Ar2, Pt3/Ar3, Pt5/Ar5, and Pt7/Ar7) differ by centromere positions, whereas chromosome Ar8 retained its ancestral organization (Figure 2). Species-specific rearrangements were encountered only in two species, namely, an ~1.9-Mb paracentric inversion within one Ar1 homoeolog in the tetraploid *Arabis blepharophylla* (breakpoints within GBs A, between BAC clones F17F16 and F20D23, and B, between F16L1 and T22J18), and a whole-arm translocation between chromosomes Ar5 and Ar8 in *A. montbretiana* (Figure 1; Madrid et al., unpublished data).

With the exception of *A. auriculata* ($n = 7$), the remaining species retained the ancestral chromosome number ($n = 8$). Descending dysploidy from $n = 8$ to $n = 7$ in *A. auriculata* was mediated by a nested chromosome insertion, placing chromosome Ar2 into (peri)centromere of Ar1, later followed by an ~3.27-Mb paracentric inversion (breakpoints within GBs A, between BAC clones F20D23 and F2864, and B, between T1K7 and T24P13). A reciprocal translocation between chromosomes Ar6 and Ar8 was identified in this species (Figure 1).

Newly Identified Centromere-Associated Repeats

Low-pass whole-genome Illumina sequence data of eight Arabideae species (Supplemental Table 3; Supplemental File 1) and publicly available data of *A. alpina* and *A. montbretiana* were analyzed by the RepeatExplorer (Novák et al., 2013) and TAREAN (Novák et al., 2017) pipelines to identify the most abundant tandem repeats. The repeat content varied between 11% (*A. auriculata* and *A. montbretiana*) and 38% (*Aubrieta canescens*) and was broadly related to genome size differences (Supplemental Table 3; Supplemental File 1).

The tandem repeat content varied from 0.05% in *A. montbretiana* to 5.81% in *Arabis cypria* (Supplemental Table 3). Only tandem repeats constituting at least 0.01% of a genome were further tested by FISH as potential chromosomal landmarks identifying (peri)centromeric regions in the analyzed species. The satellites specifically hybridizing to (peri)centromeres (Figure 3A) are listed in Supplemental Table 4, and their sequences are given in Supplemental File 2. Dot-plot comparison of monomer consensus sequences revealed similarities among some repeats in *A. alpina*, *A. auriculata*, and *A. cypria* and among satellites identified in *Draba hispida* and *D. nemorosa* (Supplemental Figure 2).

As the most abundant tandem repeats in *A. alpina* (ArAl1, 496 bp; 3.3% of the genome) and *A. cypria* (ArCy1, 495 bp; 2.7% of the genome) showed more than 92% sequence identity (Supplemental Figure 3A), a universal oligo probe (Ar_univ4) was synthesized to identify centromeres in the two *Arabis* species (Figure 3A). The 136-bp repeat ArMo3, occupying only 0.014% of the genome, specifically localized to centromeres in *A. montbretiana*. This sequence was highly similar (97.8%) to both satellites ArAl2 (136 bp; 1%) and ArCy2 (136 bp; 1.2%) in *A. alpina* and *A. cypria*, respectively (Supplemental Figure 3B). A 136-bp consensus monomer (Ar_univ8 oligo) was used to identify centromeres in *A. montbretiana*. In *A. auriculata*, three tandem repeats with monomer sizes of 621 bp (0.13%), 494 bp (0.08%), and 875 bp (0.08%) were identified and used as oligo probes (ArAu1, ArAu2, and ArAu3) in this species (Figure 3A). The 494-bp tandem repeat showed 62.1 and 62.5% sequence

identity to ArAl1 and ArCy1 repeats (Supplemental Figure 3A). Centromeres in *A. blepharophylla* were targeted using the Ar_univ3 probe designed from a consensus 170-bp satellite sequence shared by *A. auriculata* and *A. planisiliqua* (pairwise identity 97.6%; Supplemental Figure 3C). In *Au. canescens* and *Au. parviflora*, an oligo probe based on the 170-bp AuCa repeat (3.2% of the genome) localized to centromeres in these species (Figure 3A).

In the *D. hispida* genome, the most abundant tandem repeat was represented by a family of four sequence variants (DrHi1 to DrHi4) occupying ~2% of the genome. The variants of different monomer length (144, 145, and 148 bp in two variants) showed sequence similarity ranging between 78.4 and 93.2% (Supplemental Figure 3D). DrHi oligo probe, designed as universal for all four repeat variants (Supplemental Figure 3D), was used to identify centromeric regions in *D. hispida*. Five sequence variants (DrNe1 to DrNe5; 4.15% of the genome) were found in the repeatome of *D. nemorosa*. The five monomer variants, ranging from 135 to 143 bp, showed pairwise identity between 81.8 and 95.6% (Supplemental Figure 3D). The DrNe oligo, designed as a universal

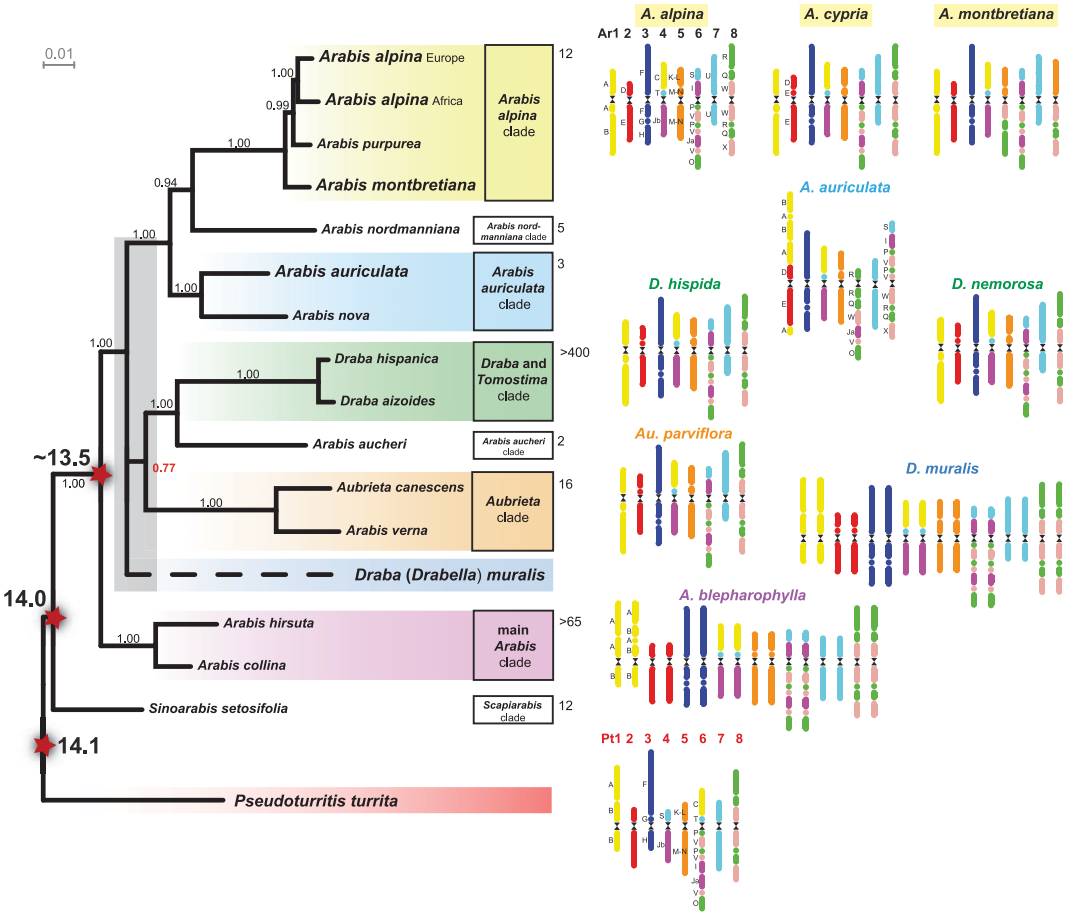


Figure 1. Comparative Genome Structures of Arabideae Species in a Phylogenetic Context.

The phylogenetic scheme shows the assignment of analyzed genomes to 10 Arabideae clades. The core phylogeny was adopted from Karl and Koch (2013) and Kiefer et al. (2017). Gray box, not fully resolved phylogenetic signal as revealed from a tribal-wide RAxML analysis (Supplemental Figure 1). Divergence time estimates for three nodes are taken from Karl and Koch (2013) [14 mya]; Karl and Koch (2013), Hohmann et al. (2015), and Guo et al. (2017) [14.1 mya]; and Karl et al. (2012) [13 to 14 mya, mean 13.5 mya]. Stem group ages and time of diversification for the intra-tribal clades are provided in Supplemental Table 2. Numbers denote the number of species within a given clade. The color coding of the constructed genomes corresponds to eight chromosomes and 22 genomic blocks of the ACK (Lysak et al., 2016).

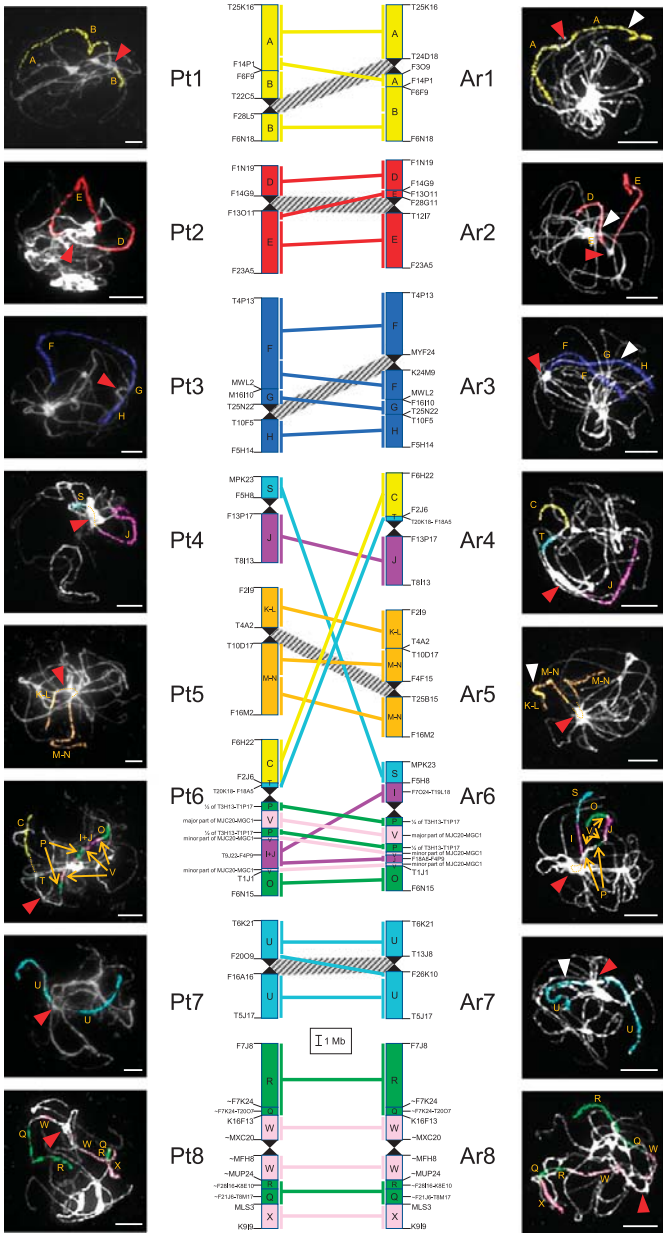


Figure 2. Chromosome-Level Comparison of *Pseudoturritis turrita* and *Arabis cypria* Genomes.

Collinearity relationship of the eight ancestral chromosomes of *P. turrita* (Pt1 to Pt8) to the eight chromosomes of *A. cypria* (Ac1 to Ac8) as inferred from comparative painting experiments. Centromeres are depicted as sandglass-like symbols in the central cartoon and labeled by arrowheads in FISH images (red arrowhead, functional centromere; white arrowhead, inactive ancestral centromere). Hatched ribbons connect the ancestral and repositioned

probe for all five repeat variants (Supplemental Figure 3D), was synthesized to identify centromeric positions in *D. nemorosa* (Figure 3A). Both *Draba* repeat families show considerable sequence similarity (68.9 to 80.4%), whereby the highest similarity (76.4 to 80.4%) was found between the four DrHi satellites and DrNe1 (Supplemental Figure 3D).

In *P. turrita*, in silico analysis identified a major satellite family represented by two sequence variants, PsTu1 (146 bp; 0.41% of the genome) and PsTu2 (175 bp; 0.14% of the genome), with the pairwise homology of 83.6% (Supplemental Figure 3E). The oligo PsTu, designed to cover both repeat variants, hybridized to centromeric regions in *P. turrita* (Figure 3A).

Centromere Repositioning as the Key Mechanism of Karyotype Differentiation in Arabideae

As a positional shift was observed on five of six chromosomes with conserved chromosomal collinearity in the crown-group Arabideae species (Figure 1), precise characterization of the repositioned centromeres was performed by chromosome painting using a BAC-by-BAC approach (Figure 3B), along with FISH localization of newly identified centromere-associated tandem repeats (Supplemental Table 4). The fine-scale chromosome painting experiments using repeat-free Arabidopsis BAC clones gave compelling evidence that all five ENCs were formed outside repeat-rich pericentromeres of the respective ancestral genomes. Most ENCs are smaller and less hetorochromatic than larger heterochromatic pericentromeres in the ancestral genome of *P. turrita* (Figure 4; Supplemental Figure 4). Centromere positions on chromosomes of *P. turrita* have been considered as the ancestral ones, and length of chromosome arms (i.e., centromere position) was approximated using physical length of Arabidopsis BAC contigs (https://www.arabidopsis.org/) in all the analyzed species. The BAC-by-BAC analysis of ENCs (exemplified in Figure 3B) is detailed below and graphically summarized in Figure 4; for more experimental data on centromere localization, see Supplemental Figure 3.

ENC1 (Chromosome Ar1)

The centromere on *Pseudoturritis* chromosome Pt1 is located at position 9.62 Mb, within genomic block B. In the crown-group Arabideae species, the centromeres moved along the longer upper arm 3.17 to 4.19 Mb away from the ancestral position. Three *Arabis* (*A. alpina*, *A. cypria*, and *A. montbretiana*) and two *Draba* species (*D. hispida* and *D. nemorosa*) show the identical centromere repositioning to position 5.5 and 6.07 Mb, respectively. ENCs in *Au. parviflora* and *Draba muralis* are located at yet different chromosomal positions—6.45 and 5.43 Mb. Centromere repositioning transformed the acrocentric Pt1-like chromosome to metacentric Ar1 homoeologues, as also reflected by increased centromeric indexes (CIs; length of the short arm to the total chromosome length × 100). Whereas CI of the ancestral chromosome equalled 18%, CI increased to 45% in *Au. parviflora,* and 48% in *D. hispida* and *D. nemorosa*.

ENC2 (Chromosome Ar2)

The centromere on Pt2 is located between blocks D and E, at position 3.05 Mb. Whereas the ancestral centromere position remained conserved in *A. alpina*, *A. blepharophylla*, and *A. montbretiana*, homoeologous centromeres in *A. cypria*, *D. muralis*, and *Au. parviflora* repositioned along the longer bottom arm: 0.76, 2.08, and 2.71 Mb away from the original location. Repositioning increased the CI from 33% in *P. turrita* to 38% in *Au. parviflora*, 41% in *A. cypria*, and 45% in *D. muralis*, transforming the ancestral (Pt2) submetacentric to metacentric chromosomes.

ENC3 (Chromosome Ar3)

The Pt3 centromere is located between genomic blocks G and H, at position 10.77 Mb. Centromeres on Ar3 homoeologues were repositioned in all seven crown-group Arabideae species analyzed. The shift relocated the centromere along the longer upper arm, 3.54 Mb (*A. blepharophylla*) to 6.54 Mb (*Aubrieta*) away from its ancestral position. In two *Arabis* (*A. alpina* and *A. cypria*) and two *Draba* species (*D. muralis* and *D. nemorosa*), respectively, the ENCs have identical positions. At least five independent centromere repositioning events transformed the ancestral acrocentric chromosome (CI 23%) into modern submetacentric (*D. muralis*, *D. nemorosa*: CI 37%; *Au. parviflora*: CI 30%) and metacentric chromosomes (*A. blepharophylla*: CI 48%; *A. montbretiana*: CI 49%), respectively.

ENC5 (Chromosome Ar5)

The centromere on chromosome Pt5 is located between blocks K-L and M-N, at position 4.07 Mb. The ancestral centromere position has changed in all eight Arabideae species analyzed. At least six independent centromere shifts relocated the centromere along the longer bottom arm 1.53 Mb (*A. blepharophylla*) to 3.87 Mb (*D. muralis*) away from its original position. In *A. alpina*, *A. cypria*, and *D. hispida*, Ar5 homoeologues have the same centromere position. As the ancestral chromosome was acrocentric (CI 35%), centromere repositioning up to 2.5 Mb rendered the Ar5 homoeologues (sub)metacentric (CI 42 and 49% in *A. auriculata* and *A. blepharophylla*, respectively). However, larger shifts moved the centromere closer to opposite chromosome end and made the chromosomes acrocentric again (CIs from 31% in *D. muralis* to 36% in *D. nemorosa*).

Figure 2. (continued).

centromeres. The 22 genomic blocks (A to X) of ACK (Lysak et al., 2016) are colored according to their position on chromosomes AK1 to AK8 and their size equals to the size of these regions in the Arabidopsis genome (The Arabidopsis Information Resource [TAIR]; http://www.arabidopsis.org); boundaries of the genomic (sub)blocks are given as corresponding Arabidopsis BAC clones. Fluorescence of painting probes was captured as black-and-white photographs and pseudocolored to match the 22 blocks and eight AK chromosomes. Bars in FISH images = 10 μm.

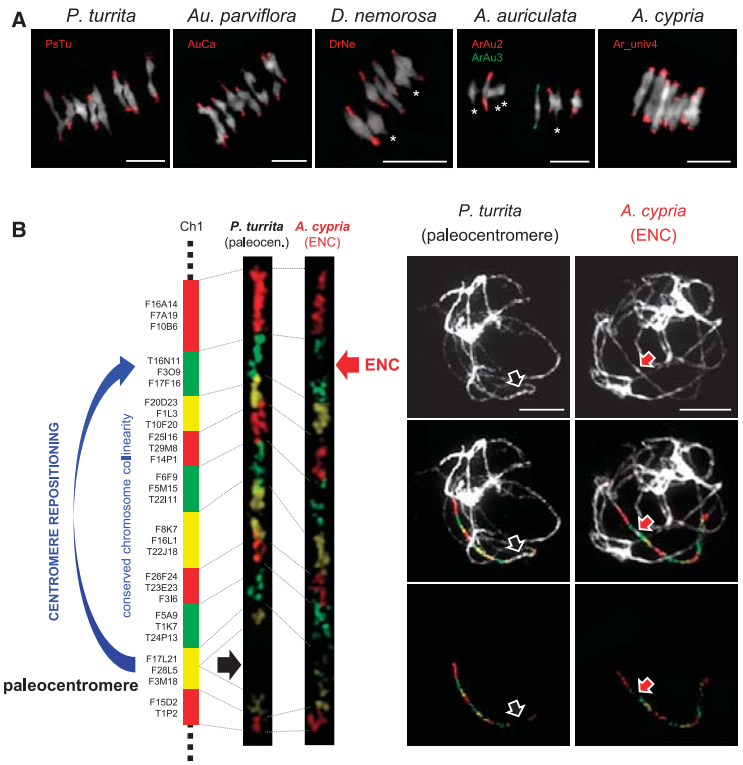


Figure 3. Chromosomal Localization of Centromere-Associated Tandem Repeats and Experimental Proof-of-Concept of Centromere Repositioning in Arabideae Species.

(A) Localization of the identified tandem repeats (Supplemental Table 3) to centromeres of metaphase I bivalents in five Arabideae species analyzed. Stars indicate chromosomes without hybridization signals. Bars = 10 μm.
(B) Comparative chromosome painting using 29 Arabidopsis BAC clones (TAIR; http://www.arabidopsis.org) on pachytene bivalents in *P. turrita* and *A. cypria*. The differentially labeled BAC contig from Arabidopsis chromosome 1 hybridized to homoeologous chromosome regions on chromosome Pt1 and Ar1. Conserved inter-species collinearity reveals the position of the paleocentromere (black arrows) and ENC1 (red arrows) in *P. turrita* and *A. cypria*, respectively. Bars = 10 μm.

ENC7 (Chromosome Ar7)

The Pt7 centromere is located within genomic block U, at position 4.17 Mb (CI 50%). Whereas centromere on the Ar7 homoeologue in *A. cypria* remained at the ancestral position, centromeres in other species moved along both chromosome arms (five of six ENCs on the upper arm). Along the upper arm, the largest shift of 1.28 Mb was identified in *D. muralis* and a comparable shift of 1.32 Mb on the bottom arm was observed in *Aubrieta*. The repositioning events on opposite chromosome arms impacted the chromosome symmetry in the same fashion, rendering Ar7 homoeologues acrocentric in *D. muralis* (CIs 35%) and *Au. parviflora* (34%). Chromosomes in the remaining species are submetacentric.

Collinearity between Homoeologous Chromosomes in *A. alpina* and *Arabidopsis lyrata* Corroborates Centromere Repositioning in Arabideae Genomes

The sequenced genomes of *A. alpina* and *Arabidopsis lyrata* have descended from a common ancestor (Willing et al., 2015; Lysak et al., 2016). To further corroborate centromere repositioning events in Arabideae genomes, we re-analyzed the level of inter-genome collinearity between chromosome-scale assemblies of *A. alpina* (Willing et al., 2015) and *A. lyrata* ($n = 8$; Hu et al., 2011). Five chromosomes in *A. alpina* (Ar1 to Ar3, Ar5, and Ar7) are partly or entirely collinear with their homoeologous counterparts in the *A. lyrata* genome (Al1 to Al3, Al5, and Al7). Pairwise sequence alignments of the five homoeologues confirmed their perfect

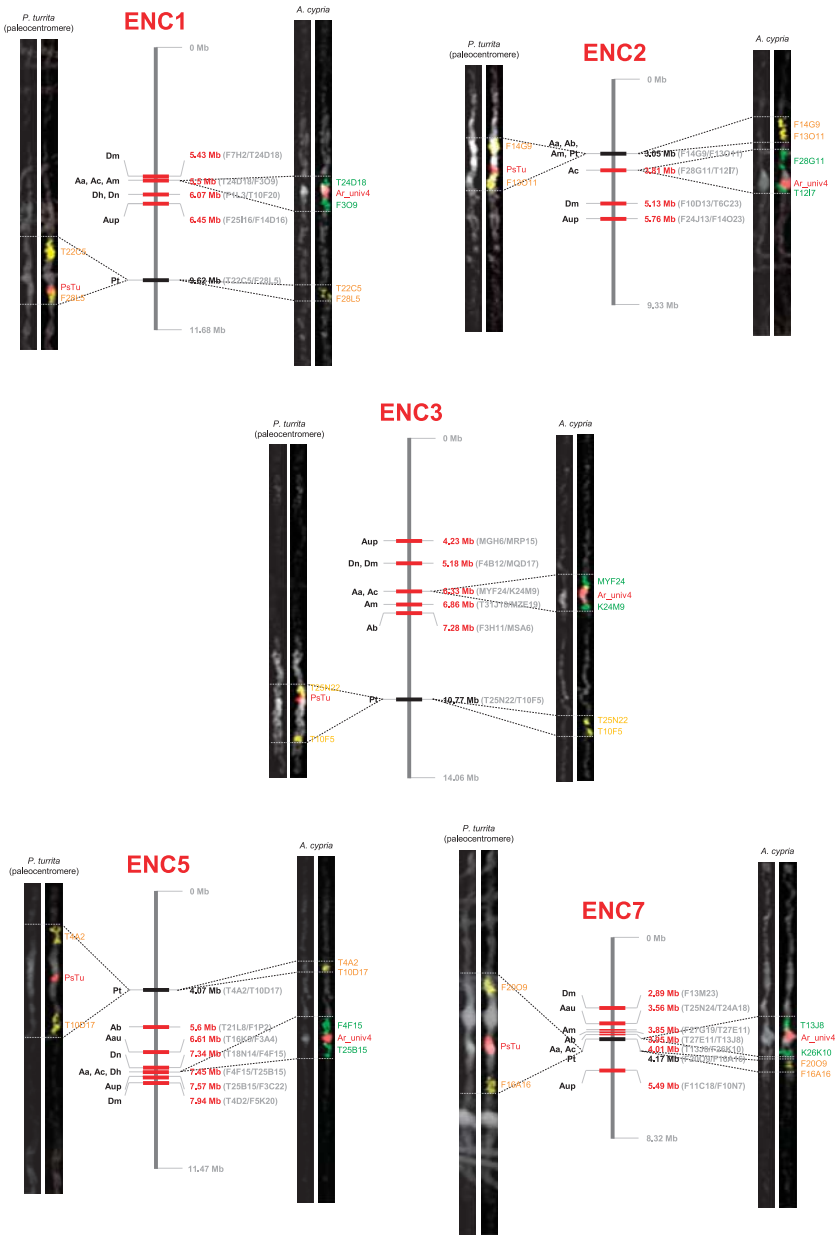


Figure 4. Centromere Repositioning Events on Five Homoeologous Chromosomes in Arabideae Species.

Black bars indicate the ancestral centromeres in *P. turrita*, whereas red bars show ENCs (1 to 3, 5, and 7). Centromere positions in megabases were inferred from FISH localization of centromere-facing Arabidopsis BAC clones on chromosomes of Arabideae species and approximated as the position of these BACs on Arabidopsis pseudomolecules (TAIR; http://www.arabidopsis.org). FISH of centromere-facing BACs along with centromere-associated tandem

collinearity despite differently positioned centromeres (Supplemental Figure 5). Thus, conserved inter-genome chromosomal collinearity and the absence of inversions further corroborated repositioning as the mechanism driving centromere mobility in Arabideae genomes.

Chromatin Features of Ancestral Centromeres and ENCs Do Not Differ in *A. alpina*

Using the available pseudomolecule assemblies in *A. alpina* (Willing et al., 2015), we questioned whether the sequence and chromatin context of paleocentromeres, inactive centromeres, and ENCs differ. Chromosomal profiling does not identify distinct peaks of repeat density, DNA methylation, and H3K27 monomethylation (a hallmark of pericentromeric heterochromatin; Willing et al., 2015) around any of the eight functional centromeres (Figure 5). Instead, the eight chromosomes exhibit rather broad pericentromere regions with indistinct transitions to chromatin of both chromosome arms. Taken together, in the *A. alpina* genome, pericentromeres of ancestral and ENCs, defined as H3K27me1 enrichment, do not strikingly differ in size, repeat content and DNA methylation levels (Figure 5).

DISCUSSION

The Ancestral Arabideae Genome in the Context of Crucifer Genome Evolution

Comparisons of the ancestral Arabideae genome, that is, the *Pseudoturritis* genome, with other inferred ancestral genomes in the family Brassicaceae corroborate the antiquity of the Arabideae clade apparent from some phylogenetic analyses (Nikolov et al., 2019). Some ancestral Arabideae chromosomes are shared by other inferred ancestral genomes, such as chromosome 3 (Pt3), which is structurally stable across all ancestral karyotypes, chromosome 2 (Pt2), structurally mirroring chromosome AK2 in the ancestral crucifer karyotype (ACK; Schranz et al., 2006; Lysak et al., 2016), and homoeologues in the ancestral proto-Calepineae karyotype (ancPCK; Geiser et al., 2016) and PCK (Mandáková and Lysak, 2008). Chromosome 5 (Pt5) was retained in ACK and ancPCK, whereas chromosomes 1 (Pt1) and 7 (Pt7) are shared by the ancestral (CEK) karyotype of the *Hesperis* clade (Lineage III/Clade E; Mandáková et al., 2017). Altogether, the chromosomes shared with other crucifer lineages and clades corroborate the ancestral position of centromeres in the *P. turrita* genome, contrasted by centromere repositioning events in evolutionary younger Arabideae clades.

Based on the early divergence of *Hesperis* (Lineage III) and Arabideae clades from the remaining main crucifer lineages (Mandáková et al., 2017; Nikolov et al., 2019) and three chromosomes shared between these two clades, an ancestral $n = 8$ genome predating the divergence of main crucifer lineages most likely contained five paleo-chromosomes (homoeologues of Pt1, Pt2, Pt3, Pt5, and Pt7), in part conserved in younger lineage-specific ancestral genomes. The paleostructure of the remaining three chromosomes (homoeologues of Pt4, Pt6, and Pt8) cannot be unambiguously inferred due to lineage-specific chromosomal rearrangements.

Centromere Repositioning in the Arabideae Is Exceptional among Crucifer Lineages

Despite being diversified into several subclades and occupying diverse habitats mostly across the Northern Hemisphere (Jordon-Thaden et al., 2013; Karl and Koch, 2013, 2014), members of the Arabideae show a remarkable stasis of chromosome numbers and genome structure. Here, we proved that the high frequency of centromere repositioning is a primary process underlying structural differentiation of Arabideae chromosomes and whole genomes.

While ENCs originated frequently during diversification of Arabideae, the phenomenon was documented in only three other crucifer genera. In *Neslia paniculata* (Lysak et al., 2006), *Cardamine rivularis* (Mandáková et al., 2013), and *Camelina sativa* (Mandáková et al. 2019b), (sub)metacentric chromosomes have become telocentric or the centromere of an acrocentric chromosome was repositioned more distally (*Camelina*) without apparently disrupting chromosomal collinearity. At least five ENCs were inferred on different chromosomes in three tetraploid and octoploid *Cardamine* species (Mandáková et al., 2019a). Collectively, these instances illustrate that although centromere repositioning has occurred in some crucifer taxa, such events are very rare in diploid genomes showing overall karyotype stasis (Camelineae, Cardamineae, Lineage II tribes, *Hesperis*/Lineage III tribes; Lysak et al., 2006, 2016; Mandáková and Lysak, 2008; Mandáková et al., 2013, 2017, 2019a, 2019b). The high number of identified ENCs in the Arabideae is exceptional among Brassicaceae genera and tribes. The drivers underlying the unusually frequent incidence of ENCs in the tribe cannot be identified easily. Although Arabideae is the largest tribe of Brassicaceae (Al-Shehbaz, 2012), particularly due to frequent hybridization, polyploidization, and speciation events in the genus *Draba* (>390 species; Jordon-Thaden et al., 2010, 2013), its global distribution, life histories, and ecological requirements do not differ from many other crucifer tribes primarily confined to the Northern Hemisphere. Future analysis of more Arabideae species, including high-quality genome assemblies, should provide more insight into causes of the accelerated centromere mobility in this tribe.

Mechanism of Centromere Repositioning in Arabideae

The high incidence of ENCs in Arabideae is comparable with frequent neocentromere formation during the evolution of grass

Figure 4. (continued).

repeats (PsTu and Ar_univ4) to pachytene bivalents in *P. turrita* and *A. cypria* mark the position of paleocentromeres and ENCs on homoeologous chromosomes of the two species. Aa, *A. alpina*; Aau, *A. auriculata*; Ab, *A. blepharophylla*; Ac, *A. cypria*; Am, *A. montbretiana*; Aup, *Au. parviflora*; Dh, *D. hispida*; Dm, *D. muralis*; Dn, *D. nemorosa*; Pt, *Pseudoturritis turrita*.

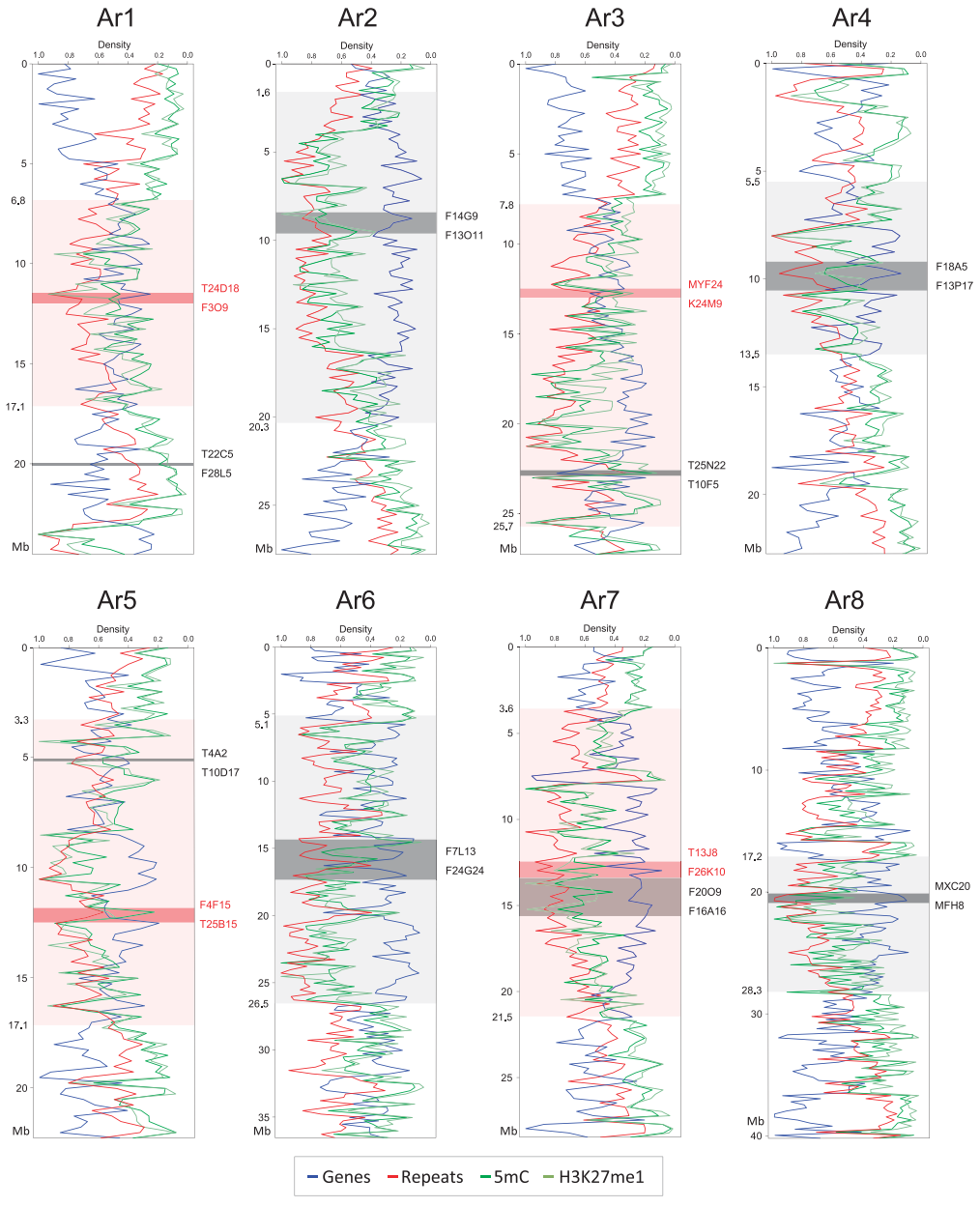


Figure 5. Position of Functional and Inactive Centromeres on Chromosomes of *A. alpina*.

The location of repositioned centromeres (ENCs, red bars), and ancestral and inactive centromeres (dark gray bars) are shown in the context of gene, repeat, DNA methylation, and H3K27 monomethylation chromosomal density profiles (Willing et al., 2015). The positions of all centromeres are delimited by Arabidopsis BAC clones based on comparative chromosome painting in *A. alpina* and *P. turrita*. The elevated H3K27me1 density (marked as light red and light gray ranges) approximately marks pericentromeres in *A. alpina* (Willing et al., 2015).

genomes. Several studies in maize and rice (Lamb et al., 2007; Wang and Bennetzen, 2012; Liu et al., 2015; Wolfgruber et al., 2016; Zhao et al., 2017; Liao et al., 2018) have shown that centromeres can change their position by two different mechanisms. The entire or partial centromeric region can be repositioned to a new chromosomal position by pericentric (Liao et al., 2018) or hemicentric (Lamb et al., 2007) inversions, disrupting the original collinearity within the inverted chromosomal segment. The alternative mode of centromere shift includes centromere repositioning or seeding via spreading or transposition of CenH3 to a new chromosomal position and concurrent or fast inactivation of the original centromere. Centromere seeding or sliding does not disrupt the collinearity in the region between the old and new centromere. In Arabideae, the conserved collinearity of neocentromere-bearing chromosomes, revealed by BAC-by-BAC chromosome painting, strongly suggests that these centromeres were repositioned. Although hemi- and/or pericentric inversions followed by other inversions, restoring the original collinearity, cannot be fully dismissed, traces of such inversions have not been found for any of the 32 documented ENCs in the analyzed genomes.

A de novo centromere is more likely to become fixed when formed in a gene-poor (gene desert) region, as suggested by several published studies (Cardone et al., 2006; Ventura et al., 2007; Lomiento et al., 2008; Zhao et al., 2017; Lu and He, 2019). However, random seeding in anonymous, gene-containing regions was observed, too (Tolomeo et al., 2017; Bracewell et al., 2019), as well as repositioning to multiple chromosomal regions (Ketel et al., 2009). In Arabideae, localization of BAC clones, containing single- and low-copy (coding) sequences, between the original and new centromeres on all five ENC-bearing chromosomes (Figure 4; Supplemental Figure 3) indicates that new centromeres have formed in initially anonymous euchromatic regions, and not in repeat-rich pericentromeres. The conserved chromosomal collinearity within regions spanning the old and new centromeres, along with the absence of heterochromatic islands at the original centromere sites, suggest that the new centromeres formed through spreading or loading of CenH3 into new chromosomal positions.

Based on the analysis of 57 independent CenH3 relocation events in more than 20 lines of domesticated maize, Schneider et al. (2016) proposed an attractive model of centromere repositioning driven by selection for key centromere-linked genes. Long-term inbreeding for favorable centromere-linked alleles results in a loss of centromere-specific tandem repeats, and this decay triggers spreading or transposition (i.e., repositioning) of CenH3 to a new position. The new, initially repeat-free, centromere is invaded by transposable elements, some of which may evolve into centromere-specific satellite repeats (Gong et al., 2012; Sharma et al., 2013). The selection for centromere-linked genes might have resulted in the frequent emergence of ENCs in Arabideae genomes.

Recurrent Centromere Repositioning

Our BAC-by-BAC approach allowed us to compare precise positions of ENCs on five homoeologues in the eight analyzed species. Some ENCs are apparently shared among closely related species (subclades) by descent, such as ENC1 in *A. alpina*, *A. cypria*, and *A. montbretiana*, ENC1 in *D. hispida* and *D. nemorosa*, and ENC3 in *A. alpina* and *A. cypria*. Two interesting instances are ENC3, shared by *D. muralis* and *D. nemorosa*, and ENC5 having the same position in two *Arabis* species (*A. alpina* and *A. cypria*) and *D. hispida* (Figure 4; Supplemental Figure 3). The same position of two ENCs in two different, well-defined clades (Figure 1) suggests their independent emergence at the same chromosomal region (Ventura et al., 2004; Cardone et al., 2006).

Directionality of Centromere Repositioning

In Arabideae and *Cardamine* species (Mandáková et al., 2013, 2019a), centromere repositioning frequently transformed acrocentric chromosomes into (sub)metacentric chromosomes. The same trend of centromere repositioning, reverting ancestral telocentrics to metacentrics, was observed in species of the *D. obscura* group (Bracewell et al., 2019). ENCs in Arabideae species show clear clustering due to their preferential formation on one chromosome arm (with the single exception of ENC7 in *Au. parviflora*). Taking the ancestral *P. turrita* genome as a reference, ENCs occurred on a longer arm of all five chromosomes. While the evolutionary significance of this pattern is not immediately apparent, it can be assumed that ENCs formed on short arms of acrocentric chromosomes are not fixed as reduced recombination around new (peri)centromeres could compromise the frequently observed minimum of two crossovers (COs) per chromosome pair (one on each arm). The absence of the obligate number of COs can be deleterious and lead to production of aneuploid gametes (Ritz et al., 2017). In two acrocentric Arabidopsis chromosomes (chromosomes 2 and 4), the mean chiasma frequency is higher in the long than in the short arm (López et al., 2012). This difference was also observed for chromosome 4 by analyzing the CO rate variation after single-nucleotide polymorphism genotyping. Despite the lower number of COs in the short arm, a significant recombination hot spot was detected close to the nucleolar organizer region on the short arm (Drouaud et al., 2006). Thus, de novo centromere formation on a short arm could stop recombination in this region and negatively impact chromosome segregation due the decreased chiasma frequency. On the contrary, increasing chromosome symmetry of acrocentrics by centromere seeding in long arms may increase CO frequency along the former short arm and increase the number of chiasmata. Whole-genome sequencing and analysis of recombination frequencies in species with differently positioned ENCs can help to elucidate these assumptions.

How Old Are ENCs in Arabideae?

Pseudoturritis and the remaining Arabideae clades diverged probably in the Middle Miocene ~14 million years ago (mya; Figure 1; Karl and Koch, 2013), and this divergence time estimate marks the earliest possible emergence of a neocentromere(s) in the core Arabideae clades. As the main *Arabis* clade originated ~13.5 mya (Figure 1), three ENCs in *A. blepharophylla* might have emerged more than 10 mya. The shared ENC1 among three and ENC3 among two species of the *A. alpina* clade (Figure 4) suggest

that these centromeres emerged close to the origin of the clade 6 to 10 mya (Supplemental Table 2). ENC1 shared by two species of the *Draba*/*Tomostima* clade can be up to 10 to 12 mya (Supplemental Table 2). With the caveat that ENCs in the individual clades might have emerged with a time lag after the divergence events, it is safe to argue that the age of at least some ENCs is a few million years. This is in accord with centromere repositioning events on homoeologues of cucumber chromosome 7 (Han et al., 2009), which probably occurred several million years ago (Yang et al., 2014). In *Drosophila*, the oldest ENCs emerged 16 to 20 mya, while the youngest ENCs emerged 3 to 9 mya (Bracewell et al., 2019). Since the 25-million-year divergence between macaque and humans, 14 ENCs emerged in either the macaque or the human lineage (Ventura et al., 2007), whereby the nine ENCs in macaques are at least 16 million years old (Tolomeo et al., 2017). By contrast, five ENCs have arisen in the donkey genome after the donkey–zebra divergence only 1 mya (Carbone et al., 2006) and the repeat-free ENC12 in orangutans (*Pongo*) probably emerged less than 1 mya (>400,000 years ago; Tolomeo et al., 2017). Schneider et al. (2016) showed that the frequent centromere shifts in maize inbred lines postdated domestication ~10,000 years ago.

The Evolutionary Significance of Centromere Repositioning in Arabideae

Interestingly, centromere repositioning in Arabideae genomes has been observed exclusively on five paleo-chromosomes (Figure 4). Whereas these five chromosomes have been altered by the emergence of ENCs, three chromosomes (4, 6, and 8), liable to repatterning across all the crucifer clades, were reshuffled by inversions and translocations (Figure 1). Thus, centromere repositioning and collinearity-disrupting chromosomal rearrangements seem to be mutually exclusive in Arabideae. It should be interesting to analyze an even broader spectrum of Arabideae genomes to bolster this observation.

While chromosome number and karyotype stasis seem to be universal for all Arabideae clades (except the four homoeologues in *A. auriculata*), paleocentromeres followed two opposing evolutionary scenarios. Whereas centromeres on chromosomes in *P. turrita* have retained their ancestral positions probably for the past ~13 million years, the same homoeologous centromeres were repositioned frequently in later-diverging clades. The available data do not permit to infer whether the emergence of ENCs was concurrent with or even triggering the intra-tribal diversification or whether it was merely incidental to the divergence of crown-group Arabideae clades. While we assume that centromere repositioning events may modify recombination frequencies (see above), they certainly have the potential to change expression levels of active genes. Centromere repositioning alters or regulates transcription of genes originally residing in euchromatin, now being embedded in pericentromeric heterochromatin. Reciprocally, expression gene levels may increase within shrinking original pericentromeres, devoid of repeats (Schneider et al., 2016; Bracewell et al., 2019). Future comparative sequence and transcriptome analysis of multiple Arabideae genomes may interrogate genes near the inactive paleocentromeres and those adjacent to ENCs. Such analysis should clarify how the altered epigenetic marking and repeat loss versus invasion impacted gene expression near old and new centromeres.

Induced kinetochore mutations in fission yeast showed that the kinetochore impairment may easily initiate centromere repositioning. Crosses between yeast cells with the original and repositioned centromere resulted in defective meiotic chromosome segregation due to a centromere mismatch (Lu and He, 2019). These results suggest that if an individual or population becomes homozygous for a neocentromere-bearing chromosome, it may become partially or fully reproductively isolated from individuals with the original centromere. It is reasonable to predict that the degree of centromere mismatch, chromosome segregation distortion, and thus the strength of potential reproductive barrier will be positively correlated with the physical distance between the old and new centromere. Heterozygous plants with the original and new centromere at a close distance may still exhibit normal homologous pairing and full fertility as shown, for instance, in maize hybrids heterozygous for a hemicentric inversion (Lamb et al., 2007). Similarly, the original centromere and ENC on chromosome 12 in Bornean (*P. pygmaeus*) and Sumatran (*P. abelii*) orangutans occur in the heterozygous condition for 400,000 years without being fixed in either species (Locke et al., 2011; Tolomeo et al., 2017). The role of ENCs as a postzygotic reproductive barrier in Arabideae is testable by analyzing genomes of closely related (sub)species pairs differentiated by centromere repositioning, and their natural or artificial hybrids.

METHODS

Plant Material

Origins of analyzed species are listed in Supplemental Table 1. Plants were grown from seeds and cultivated under standard conditions in growth chambers (150 μmol m^{-2} s^{-1}; 21/18°C, day/night; 16/8 h of light/dark) or in a greenhouse (150 μmol m^{-2} s^{-1}; 22/19°C, day/night; 16/8 h of light/dark). Leaves and inflorescences of *Arabis auriculata* (Pálava Mts.), *Draba muralis*, *D. nemorosa*, and *Pseudoturrits turrita* were harvested from plants localized in the field.

Low-Pass Whole-Genome Sequencing

Isolated genomic DNA of *Aubrieta canescens*, *A. auriculata*, *A. cypria*, *A. planisiliqua*, *Draba hispida*, *D. muralis*, *D. nemorosa*, and *P. turrita* was sequenced using an Illumina MiSeq, paired 300-bp reads, and MiSeq v3 reagents, at the sequencing core facility of the Oklahoma Medical Research Foundation (Oklahoma City).

Sequence Data of *A. alpina* and *A. montbretiana*

A. alpina raw paired-end reads (SRR1652423) were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA; Leinonen et al., 2011) under BioProject PRJNA241291 (Willing et al., 2015) using SRA Toolkit (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/). Pseudo paired-end reads for *A. montbretiana* were created from assembled contigs (Bioproject PRJNA258048), deposited at GenBank under accession number LNCH00000000, by the tool "Get pseudo short paired end reads from long reads" (Galaxy version 0.1.0; available at https://repeatexplorer-elixir.cerit-sc.cz) with the following parameters: read length 200 bp, insert length 500 bp, and a 0.5× coverage; 490,928 reads were created.

Sequence Data Preprocessing

The quality of raw sequence reads was checked by FastQC (Andrews, 2010). The 300-bp reads were trimmed to 200 bp, and then all reads were filtered by quality (quality cutoff value was set to 20 and at least 80% of bases had to fulfill this quality); reads containing adaptors were removed. Reads with similarity to PhiX bacteriophage and chloroplast DNA were removed from the data based on a similarity search by our custom-made script.

De Novo Identification of Repetitive Sequences

Repetitive sequences in *Au. canescens*, *Arabis alpina*, *A. auriculata*, *A. cypria*, *A. montbretiana*, *D. hispida*, *D. muralis*, *D. nemorosa*, and *P. turrita* were identified and characterized by similarity-based clustering of Illumina reads using RepeatExplorer pipeline (Novák et al., 2013) through Galaxy (https://repeatexplorer-elixir.cerit-sc.cz/galaxy/). The reads were randomly subsampled to a 0.25× coverage in each species (256,750 to 568,500 reads), except 450,000 reads for *D. hispida* with unknown genome size. Clustering analysis was done individually for each genome, using default clustering parameter settings. The clusters building up at least 0. 01% of the analyzed genomes were annotated directly through RepeatExplorer where classification of transposable elements is based on the similarity to the reference database of transposable element protein domains (REXdb; Neumann et al., 2019) or additionally annotated by CENSOR (Kohany et al., 2006) using Repbase library (Bao et al., 2015).

Identification of Tandem Repeats

Tandem repeats were identified using the Tandem Repeat Analyzer (TAREAN) pipeline (https://repeatexplorer-elixir.cerit-sc.cz/galaxy/; Novák et al., 2017) that performed unsupervised identification of tandem repeats from unassembled sequence reads. The advanced option "Perform cluster merging" was used to merge clusters connected through paired-end reads. Consensus monomer sequences of the identified tandem repeats were reconstructed, and all the repeats were compared with each other by blastn to discover potential shared repeats. Reconstructed monomer sequences of shared tandem repeats were aligned by MAFFT v7.017 (Katoh and Standley, 2013) in Geneious software 11.1.5 (https://www.geneious.com; Kearse et al., 2012) to calculate pairwise percentage identity. The identified tandem repeats are listed in Supplemental Table 4.

Oligo-Probe Design

Oligonucleotide probes (60 to 68 bp) with GC content 30 to 50% were designed for the identified tandem repeats (Supplemental File 1). Probes for shared tandem repeats and tandem repeat families were targeted to conserved DNA regions of multiple alignments generated by MAFFT v7.017 (Katoh and Standley, 2013). The specificity of probes was tested by blastn search against the database built from TAREAN clusters retrieved from all sequenced genomes. The Geneious package v11.1.5 (https://www.geneious.com; Kearse et al., 2012) was used to check sequences to minimize self-annealing and formation of hairpins. The synthetic double-stranded DNA probes were labeled with biotin-dUTP, digoxigenin-dUTP, or Cy3-dUTP by nick translation as described by Mandáková and Lysak (2016a). PCR primers for amplification from genomic DNA were designed for tandem repeats with long monomers (>500 bp; Ávila Robledillo et al., 2018).

Assembly Data and Genomic Resources of *A. alpina*

The *A. alpina* genome assembly, version V4 (http://www.arabis-alpina.org; Willing et al., 2015), and sequences of Arabidopsis (*Arabidopsis thaliana*) BAC clones were downloaded from NCBI database. DNA methylation and histone modification data (Willing et al., 2015) were downloaded from NCBI database (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA237036), whereas gene and repeat annotations were downloaded from www.arabis-alpina.org. Each Arabidopsis BAC was mapped onto the reference assembled *A. alpina* chromosomes using nucmer (Delcher et al., 2002) to determine their position of chromosomes. The genomes of *Arabidopsis lyrata* (Hu et al., 2011) and *A. alpina* (Willing et al., 2015) were aligned and compared (Supplemental Figure 5) to identify regions of synteny using the nucmer software (Delcher et al., 2002) and default settings. Beside this comparison, the tool SynOrths (Cheng et al., 2012) was used to identify pairwise syntenic genes between *A. lyrata* and *A. alpina* genomes and to inspect the conserved order of orthologous genes based on protein sequences (data not shown).

Chromosome Preparation, Probe Labeling, and Comparative Chromosome Painting

Inflorescences of the investigated Arabideae species were fixed in freshly prepared ethanol:acetic acid fixative (3:1) overnight, transferred into 70% (v/v) ethanol, and stored at −20°C until use. Selected inflorescences were rinsed in distilled water and citrate buffer (10 mM sodium citrate, pH 4.8) and digested by a 0.3% (w/v) mix of pectolytic enzymes (cellulase, cytohelicase, and pectolyase; all from Sigma-Aldrich) in citrate buffer at 37°C for ~3 h. Mitotic and meiotic (pachytene) chromosome spreads were prepared from pistils and anthers, respectively, as described by Mandáková and Lysak (2016b). Suitable slides were pretreated by RNase (100 μg/mL; AppliChem) and pepsin (0.1 mg/mL; Sigma-Aldrich). In total, 674 chromosome-specific BAC clones of Arabidopsis grouped into contigs according to 22 genomic blocks (Lysak et al., 2016) were used. Based on the known genome structure of *A. alpina* (Willing et al., 2015), Arabidopsis BAC contigs were designed according to the structure of the eight chromosomes and used as painting probes in other Arabideae species. To determine fine-scale chromosome structures, uncover species-specific chromosomal rearrangements, and precisely characterize chromosome breakpoints and centromere positions, the initial painting experiments were followed by painting using shorter BAC (sub)contigs. In BAC-by-BAC characterization of centromere positions, individual differentially labeled BAC clones were applied. For all painting probes, individual BAC clones were labeled with biotin-dUTP, digoxigenin-dUTP, and Cy3-dUTP by nick translation and then pooled, precipitated, and resuspended in 20 μL of hybridization mixture (50% [v/v] formamide and 10% [w/v] dextran sulfate in 2× SSC) per slide as described by Mandáková and Lysak (2016a). Probes and chromosomes were denatured together on a hot plate at 80°C for 2 min and incubated in a moist chamber at 37°C overnight. Post-hybridization washing was performed in 20% formamide in 2× SSC at 42°C. Fluorescent detection was 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; digoxigenin-dUTP was detected by mouse anti-digoxigenin (Jackson ImmunoResearch) and goat anti-mouse Alexa Fluor 488 (Molecular Probes). Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI; 2 μg/mL) in Vectashield (Vector Laboratories). Fluorescent signals were analyzed and photographed using a Zeiss Axioimager epifluorescence microscope and a CoolCube camera (MetaSystems). Images were acquired separately for the four fluorochromes using appropriate excitation and emission filters (AHF Analysentechnik). The monochromatic images were pseudocolored and merged using Photoshop CS6 software (Adobe Systems). Pachytene chromosomes were straightened using the "straighten-curved-objects" plugin in ImageJ (Kocsis et al., 1991).

Assessing Phylogenetic Structure

A first tribe-wide analysis, characterizing all major Arabideae clades and species assemblages, has been introduced by Karl and Koch (2013) and

Kiefer et al. (2017). These studies also presented a robust backbone phylogeny based on four nuclear genes and two regions from the plastid genome. This backbone phylogeny has been used to illustrate genome evolution in Arabideae (Figure 1). In respective large-scale analysis using the ITS (internal transcribed spacer) marker system (ITSs 1 and 2 of nuclear rDNA), *P. turrita* has been considered as an outgroup of the remaining Arabideae clades. In order to account for this new finding (Kiefer et al., 2019), we re-analyzed the ITS alignment used previously by Karl and Koch (2013) and used *Macropodium* and *Stevenia* (tribe Stevenieae) as an outgroup, while *P. turrita* was included in the tribe Arabideae (Supplemental Figure 1). In total, the alignment consists 312 taxa (Supplemental File 3). A maximum likelihood analysis was performed using RAxML-NG (Kozlov et al., 2019). Analyses were run under the most optimal GTR+FO+I+G4m model, and bootstrap support was calculated from 1000 replicates. We further constrained the input file with the backbone phylogeny of Arabideae as shown in Figure 1. This approach has been chosen to highlight potential phylogenetic uncertainties of single taxa. In order to provide a temporal perspective on the evolution of major clades and phylogenetic branching patterns in Arabideae, we evaluated the most recent literature providing relevant divergence time estimates and allowing to compare across studies. A comprehensive overviews and discussions on this topic are provided for Arabideae and the entire Brassicaceae by Karl and Koch (2013) and Huang et al. (2019), respectively.

Supplemental Data

Supplemental Figure 1. Phylogenetic tree as revealed by constrained RAxML NG analysis.

Supplemental Figure 2. Dot-plot pairwise comparison of monomer consensus sequences of the identified centromere-associated tandem repeats.

Supplemental Figure 3. Sequence comparison of shared or similar (peri-)centromeric tandem repeats identified.

Supplemental Figure 4. Centromere repositioning events (ENCs) on five homoeologous chromosomes of the analyzed Arabideae species.

Supplemental Figure 5. Comparison of homoeologous chromosomes between *Arabis alpina* and *A. lyrata*.

Supplemental Table 1. The origin of Arabideae species analyzed.

Supplemental Table 2. Divergence time estimates.

Supplemental Table 3. Spectrum and genome proportions of repeat families identified in the sequenced Arabideae genomes.

Supplemental Table 4. List of (peri-)centromeric tandem repeats identified in the sequenced repeatomes of Arabideae species and corresponding FISH probes.

Supplemental File 1. Results of repeatome analysis in eight Arabideae species.

Supplemental File 2. Consensus monomer sequences of identified tandem repeats.

Supplemental File 3. Nexus format alignments used for ITS analyses (Figure 1).

ACKNOWLEDGMENTS

This article is dedicated to Ihsan Al-Shehbaz (Missouri Botanical Garden, St. Louis) on the occasion of his 80th birthday. We greatly appreciate Ihsan's expertise, help, and encouragement of our research on evolution of crucifer genomes, so close to his heart. We thank Milan Pouch and Christiane Kiefer for technical help. We thank Gernot Presting for discussing centromere repositioning in maize. This work was supported by the Czech Science Foundation (grant 15-18545S awarded to M.A.L.) and by the CEITEC 2020 project (grant LQ1601). Computational resources were provided by the ELIXIR-CZ project (grant LM2015047), part of the international ELIXIR infrastructure.

AUTHOR CONTRIBUTIONS

M.A.L. and T.M. conceived and designed the study; T.M. and P.H. performed the experiments; M.A.L. wrote the article with contribution of T.M., P.H., and M.A.K. All authors read and approved the final article.

Received July 24, 2019; revised December 2, 2019; accepted January 9, 2020; published January 9, 2020.

REFERENCES

Al-Shehbaz, I.A. (2012). A generic and tribal synopsis of the Brassicaceae (Cruciferae). Taxon **61:** 931–954.

Ávila Robledillo, L., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., Schubert, I., and Macas, J. (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.

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

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. Mob. DNA **6:** 11.

Bracewell, R., Chatla, K., Nalley, M.J., and Bachtrog, D. (2019). Dynamic turnover of centromeres drives karyotype evolution in *Drosophila*. eLife **8:** e49002.

Carbone, L., et al. (2006). Evolutionary movement of centromeres in horse, donkey, and zebra. Genomics **87:** 777–782.

Cardone, M.F., et al. (2006). Independent centromere formation in a capricious, gene-free domain of chromosome 13q21 in Old World monkeys and pigs. Genome Biol. **7:** R91.

Cheng, F., Wu, J., Fang, L., and Wang, X. (2012). Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. Front. Plant Sci. **3:** 198.

Chiatante, G., Capozzi, O., Svartman, M., Perelman, P., Centrone, L., Romanenko, S.S., Ishida, T., Valeri, M., Roelke-Parker, M.E., and Stanyon, R. (2017). Centromere repositioning explains fundamental number variability in the New World monkey genus *Saimiri*. Chromosoma **126:** 519–529.

Comai, L., Maheshwari, S., and Marimuthu, M.P.A. (2017). Plant centromeres. Curr. Opin. Plant Biol. **36:** 158–167.

Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. **30:** 2478–2483.

Drouaud, J., Camilleri, C., Bourguignon, P.Y., Canaguier, A., Bérard, A., Vezon, D., Giancola, S., Brunel, D., Colot, V., Prum, B., Quesneville, H., and Mézard, C. (2006). Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". Genome Res. **16:** 106–114.

Geiser, C., Mandáková, T., Arrigo, N., Lysak, M.A., and Parisod, C. (2016). Repeated whole-genome duplication, karyotype reshuffling,

and biased retention of stress-responding genes in Buckler mustard. Plant Cell **28:** 17–27.

Gong, Z., Wu, Y., Koblízková, A., Torres, G.A., Wang, K., Iovene, M., Neumann, P., Zhang, W., Novák, P., Buell, C.R., Macas, J., and Jiang, J. (2012). Repeatless and repeat-based centromeres in potato: Implications for centromere evolution. Plant Cell **24:** 3559–3574.

Grant, V. (1981). Plant Speciation. (New York: Columbia University Press).

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Wang, X., Zhang, D., Ma, T., Hu, Q., Al-Shehbaz, I.A., and Koch, M.A. (2017). Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics **18:** 176.

Han, Y., Zhang, Z., Liu, C., Liu, J., Huang, S., Jiang, J., and Jin, W. (2009). Centromere repositioning in cucurbit species: Implication of the genomic impact from centromere activation and inactivation. Proc. Natl. Acad. Sci. USA **106:** 14937–14941.

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.

Hu, T.T., et al. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. **43:** 476–481.

Huang, J., Zhao, Y., Bai, D., Shiraigol, W., Li, B., Yang, L., Wu, J., Bao, W., Ren, X., Jin, B., Zhao, Q., and Li, A., et al. (2015). Donkey genome and insight into the imprinting of fast karyotype evolution. Sci. Rep. **5:** 14106.

Huang, X.-C., German, D.A., and Koch, M.A. (2019). Temporal patterns of diversification in Brassicaceae demonstrate decoupling of rate shifts and mesopolyploidization events. Ann. Bot. **125:** 29–47.

Jordon-Thaden, I., Hase, I., Al-Shehbaz, I., and Koch, M.A. (2010). Molecular phylogeny and systematics of the genus *Draba* (Brassicaceae) and identification of its most closely related genera. Mol. Phylogenet. Evol. **55:** 524–540.

Jordon-Thaden, I.E., Al-Shehbaz, I.A., and Koch, M.A. (2013). Species richness of the globally distributed, arctic–alpine genus *Draba* L. (Brassicaceae). Alp. Bot. **123:** 97–106.

Karl, R., and Koch, M.A. (2013). A world-wide perspective on crucifer speciation and evolution: Phylogenetics, biogeography and trait evolution in tribe Arabideae. Ann. Bot. **112:** 983–1001.

Karl, R., and Koch, M.A. (2014). Phylogenetic signatures of adaptation: The *Arabis hirsuta* species aggregate (Brassicaceae) revisited. Perspect. Plant Ecol. Evol. Syst. **16:** 247–264.

Karl, R., Kiefer, C., Ansell, S.W., and Koch, M.A. (2012). Systematics and evolution of Arctic-Alpine *Arabis alpina* (Brassicaceae) and its closest relatives in the eastern Mediterranean. Am. J. Bot. **99:** 778–794.

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.

Kearse, M., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28:** 1647–1649.

Ketel, C., Wang, H.S., McClellan, M., Bouchonville, K., Selmecki, A., Lahav, T., Gerami-Nejad, M., and Berman, J. (2009). Neocentromeres form efficiently at multiple possible loci in *Candida albicans*. PLoS Genet. **5:** e1000400.

Kiefer, C., Severing, E., Karl, R., Bergonzi, S., Koch, M., Tresch, A., and Coupland, G. (2017). Divergence of annual and perennial species in the Brassicaceae and the contribution of cis-acting variation at *FLC* orthologues. Mol. Ecol. **26:** 3437–3457.

Kiefer, C., Willing, E.-M., Jiao, W.-B., Sun, H., Piednoël, M., Hümann, U., Hartwig, B., Koch, M.A., and Schneeberger, K. (2019). Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. Nat. Plants **5:** 846–855.

Koch, M.A., Karl, R., German, D.A., and Al-Shehbaz, I.A. (2012). Systematics, taxonomy and biogeography of three new Asian genera of Brassicaceae tribe Arabideae: An ancient distribution circle around the Asian high mountains. Taxon **61:** 955–969.

Koch, M.A., Karl, R., and German, D.A. (2017). Underexplored biodiversity of eastern Mediterranean biota: Systematics and evolutionary history of the genus *Aubrieta* (Brassicaceae). Ann. Bot. **119:** 39–57.

Kocsis, E., Trus, B.L., Steer, C.J., Bisher, M.E., and Steven, A.C. (1991). Image averaging of flexible fibrous macromolecules: The clathrin triskelion has an elastic proximal segment. J. Struct. Biol. **107:** 6–14.

Kohany, O., Gentles, A.J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics **7:** 474.

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.

Lamb, J.C., Meyer, J.M., and Birchler, J.A. (2007). A hemicentric inversion in the maize line knobless Tama flint created two sites of centromeric elements and moved the kinetochore-forming region. Chromosoma **116:** 237–247.

Leinonen, R., Sugawara, H., and Shumway, M. (2011) International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. Nucleic Acids Res. 39: D19–D21.

Levin, D.A. (2002). The Role of Chromosomal Change in Plant Evolution. (New York: Oxford University Press).

Liao, Y., Zhang, X., Li, B., Liu, T., Chen, J., Bai, Z., Wang, M., Shi, J., Walling, J.G., Wing, R.A., Jiang, J., and Chen, M. (2018). Comparison of *Oryza sativa* and *Oryza brachyantha* genomes reveals selection-driven gene escape from the centromeric regions. Plant Cell **30:** 1729–1744.

Liu, Y., Su, H., Pang, J., Gao, Z., Wang, X.J., Birchler, J.A., and Han, F. (2015). Sequential de novo centromere formation and inactivation on a chromosomal fragment in maize. Proc. Natl. Acad. Sci. USA **112:** E1263–E1271.

Locke, D.P., et al. (2011). Comparative and demographic analysis of orang-utan genomes. Nature **469:** 529–533.

Lomiento, M., Jiang, Z., D'Addabbo, P., Eichler, E.E., and Rocchi, M. (2008). Evolutionary-new centromeres preferentially emerge within gene deserts. Genome Biol. **9:** R173.

López, E., Pradillo, M., Oliver, C., Romero, C., Cuñado, N., and Santos, J.L. (2012). Looking for natural variation in chiasma frequency in *Arabidopsis thaliana*. J. Exp. Bot. **63:** 887–894.

Lu, M., and He, X. (2019). Centromere repositioning causes inversion of meiosis and generates a reproductive barrier. Proc. Natl. Acad. Sci. USA **116:** 21580–21591.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K., and Schubert, I. (2006). Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. Proc. Natl. Acad. Sci. USA **103:** 5224–5229.

Lysak, M.A., Mandáková, T., and Schranz, M.E. (2016). Comparative paleogenomics of crucifers: Ancestral genomic blocks revisited. Curr. Opin. Plant Biol. **30:** 108–115.

Mandáková, T., and Lysak, M.A. (2008). Chromosomal phylogeny and karyotype evolution in x=7 crucifer species (Brassicaceae). Plant Cell **20:** 2559–2570.

Mandáková, T., and Lysak, M.A. (2016a). Painting of *Arabidopsis* chromosomes with chromosome-specific BAC clones. Curr. Protoc. Plant Biol. **1:** 359–371.

Mandáková, T., and Lysak, M.A. (2016b). Chromosome preparation for cytogenetic analyses in *Arabidopsis*. Curr. Protoc. Plant Biol. **1:** 43–51.

Mandáková, T., Kovarík, A., Zozomová-Lihová, J., Shimizu-Inatsugi, R., Shimizu, K.K., Mummenhoff, K., Marhold, K., and Lysak, M.A. (2013). The more the merrier: Recent hybridization and polyploidy in *cardamine*. Plant Cell **25:** 3280–3295.

Mandáková, T., Hloušková, P., German, D.A., and Lysak, M.A. (2017). Monophyletic origin and evolution of the largest crucifer genomes. Plant Physiol. **174:** 2062–2071.

Mandáková, T., Zozomová-Lihová, J., Kudoh, H., Zhao, Y., Lysak, M.A., and Marhold, K. (2019a). The story of promiscuous crucifers: Origin and genome evolution of an invasive species, *Cardamine occulta* (Brassicaceae), and its relatives. Ann. Bot. **124:** 209–220.

Mandáková, T., Pouch, M., Brock, J.R., Al-Shehbaz, I.A., and Lysak, M.A. (2019b). Origin and evolution of diploid and allopolyploid *Camelina* genomes were accompanied by chromosome shattering. Plant Cell **31:** 2596–2612.

Montefalcone, G., Tempesta, S., Rocchi, M., and Archidiacono, N. (1999). Centromere repositioning. Genome Res. **9:** 1184–1188.

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. Mob. DNA **10:** 1.

Nikolov, L.A., Shushkov, P., Nevado, B., Gan, X., Al-Shehbaz, I.A., Filatov, D., Bailey, C.D., and Tsiantis, M. (2019). Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. New Phytol. **222:** 1638–1651.

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.

Novák, P., Ávila Robledillo, L., 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.

Piras, F.M., Nergadze, S.G., Magnani, E., Bertoni, L., Attolini, C., Khoriauli, L., Raimondi, E., and Giulotto, E. (2010). Uncoupling of satellite DNA and centromeric function in the genus *Equus*. PLoS Genet. **6:** e1000845.

Rieseberg, L.H. (2001). Chromosomal rearrangements and speciation. Trends Ecol. Evol. (Amst.) **16:** 351–358.

Ritz, K.R., Noor, M.A.F., and Singh, N.D. (2017). Variation in recombination rate: Adaptive or not? Trends Genet. **33:** 364–374.

Rocchi, M., Archidiacono, N., Schempp, W., Capozzi, O., and Stanyon, R. (2012). Centromere repositioning in mammals. Heredity **108:** 59–67.

Schneider, K.L., Xie, Z., Wolfgruber, T.K., and Presting, G.G. (2016). Inbreeding drives maize centromere evolution. Proc. Natl. Acad. Sci. USA **113:** E987–E996.

Schranz, M.E., Lysak, M.A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: Building blocks of crucifer genomes. Trends Plant Sci. **11:** 535–542.

Schubert, I. (2018). What is behind "centromere repositioning"? Chromosoma **127:** 229–234.

Sharma, A., Wolfgruber, T.K., and Presting, G.G. (2013). Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14: 142.

Tolomeo, D., et al. (2017). Epigenetic origin of evolutionary novel centromeres. Sci. Rep. **7:** 41980.

Ventura, M., et al. (2004). Recurrent sites for new centromere seeding. Genome Res. **14:** 1696–1703.

Ventura, M., Antonacci, F., Cardone, M.F., Stanyon, R., D'Addabbo, P., Cellamare, A., Sprague, L.J., Eichler, E.E., Archidiacono, N., and Rocchi, M. (2007). Evolutionary formation of new centromeres in macaque. Science **316:** 243–246.

Wang, H., and Bennetzen, J.L. (2012). Centromere retention and loss during the descent of maize from a tetraploid ancestor. Proc. Natl. Acad. Sci. USA **109:** 21004–21009.

Wang, K., Wu, Y., Zhang, W., Dawe, R.K., and Jiang, J. (2014). Maize centromeres expand and adopt a uniform size in the genetic background of oat. Genome Res. **24:** 107–116.

White, M.J.D. (1978). Modes of Speciation. (San Francisco: W.H. Freeman & Co.).

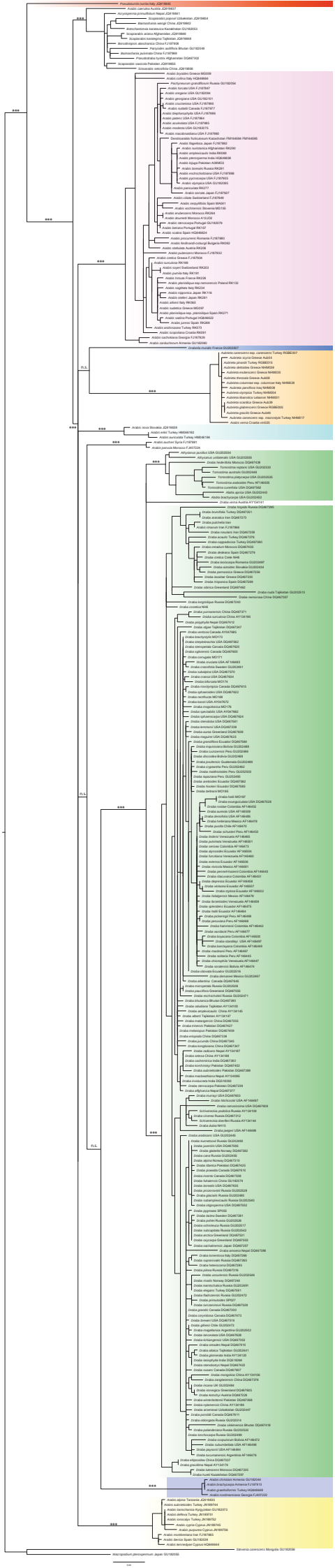
Willing, E.-M., et al. (2015). Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. Nat. Plants **1:** 14023.

Wolfgruber, T.K., Nakashima, M.M., Schneider, K.L., Sharma, A., Xie, Z., Albert, P.S., Xu, R., Bilinski, P., Dawe, R.K., Ross-Ibarra, J., Birchler, J.A., and Presting, G.G. (2016). High quality maize centromere 10 sequence reveals evidence of frequent recombination events. Front. Plant Sci. **7:** 308.

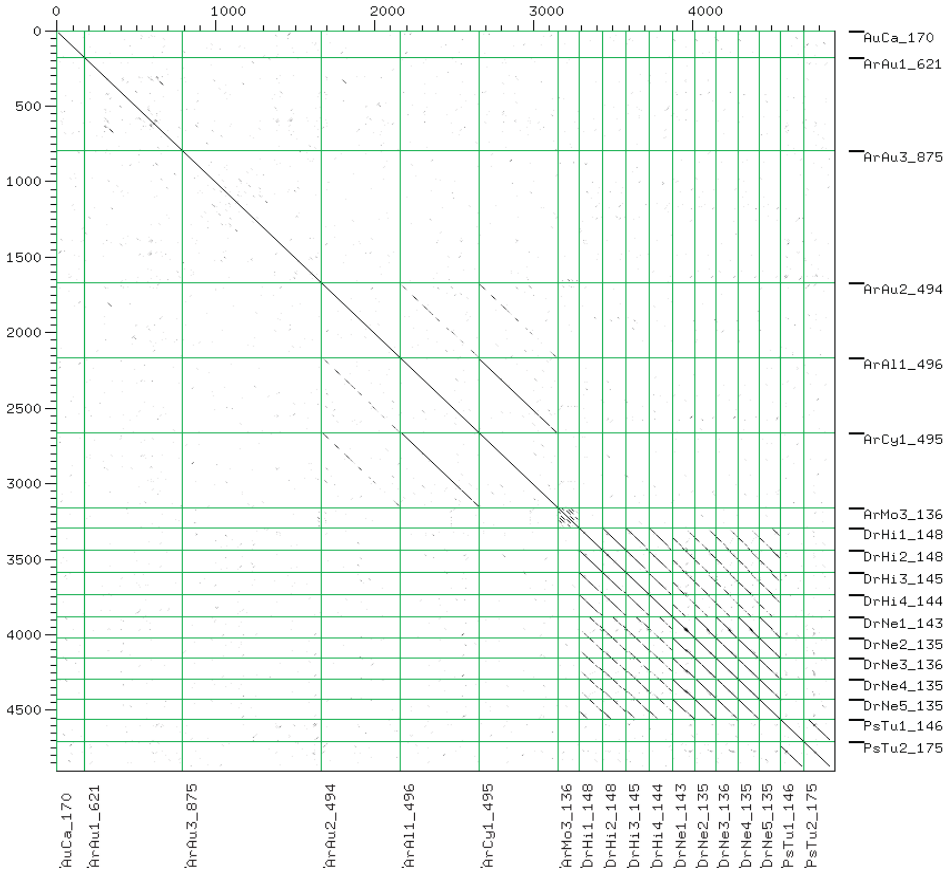
Yang, L., et al. (2014). Next-generation sequencing, FISH mapping and synteny-based modeling reveal mechanisms of decreasing dysploidy in *Cucumis*. Plant J. **77:** 16–30.

Zhao, H., Zeng, Z., Koo, D.H., Gill, B.S., Birchler, J.A., and Jiang, J. (2017). Recurrent establishment of de novo centromeres in the pericentromeric region of maize chromosome 3. Chromosome Res. **25:** 299–311.

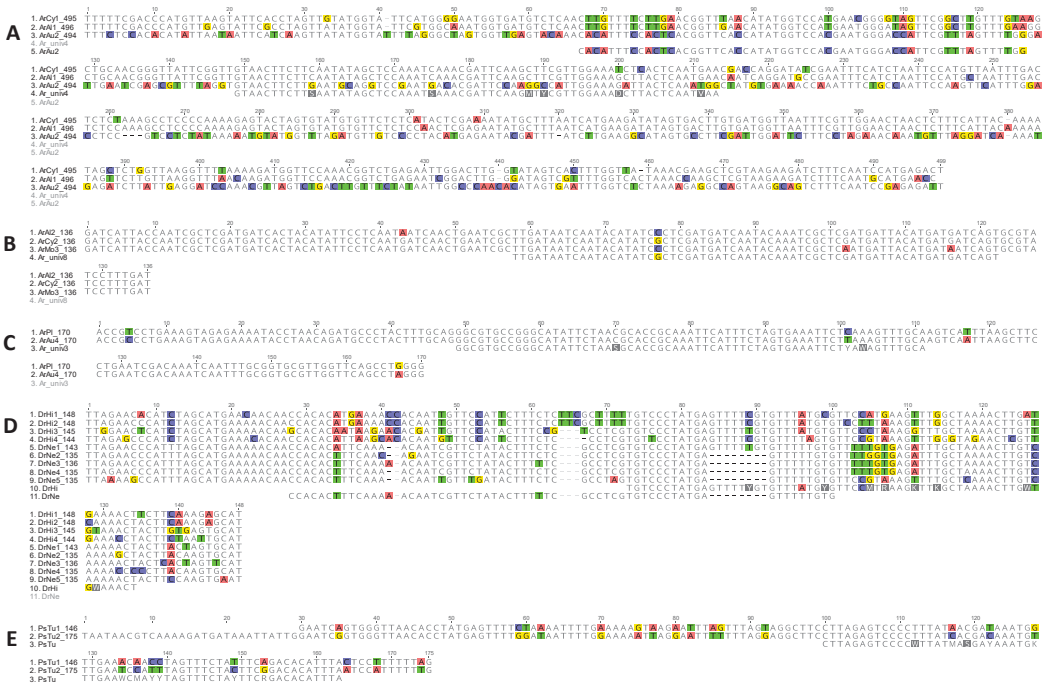
Supplemental Figure 1. Phylogenetic tree as revealed by constrained RAxML NG analysis. Result from RAxML analysis of tribal wide analysis of ITS sequence data (Karl and Koch 2013). Supports Figure 1. Tree topology is constrained by the backbone phylogeny as shown in Figure 1 (Kiefer et al. 2019). All major clades (underlaid boxes) are supported by bootstrap >95% (***) [bootstrap support within clades is not shown]. *Draba* (*Drabella*) *muralis* has been placed separately from any of the major clades. n.s.: not significant. See Supplemental Data 3 for sequence alingments. References: Karl R, Koch MA (2013) *Ann Bot* 112: 983-1001. Kiefer C et al. (2019) *Nat Plants* 5: 846-855.



Supplemental Figure 2. Dot-plot pairwise comparison of monomer consensus sequences of the identified centromere-associated tandem repeats (Supplemental Table 4). **Supports Figure 3A and Supplemental Table 4.** Default settings were used in Dotter (Sonnhammer and Durbin 1995).

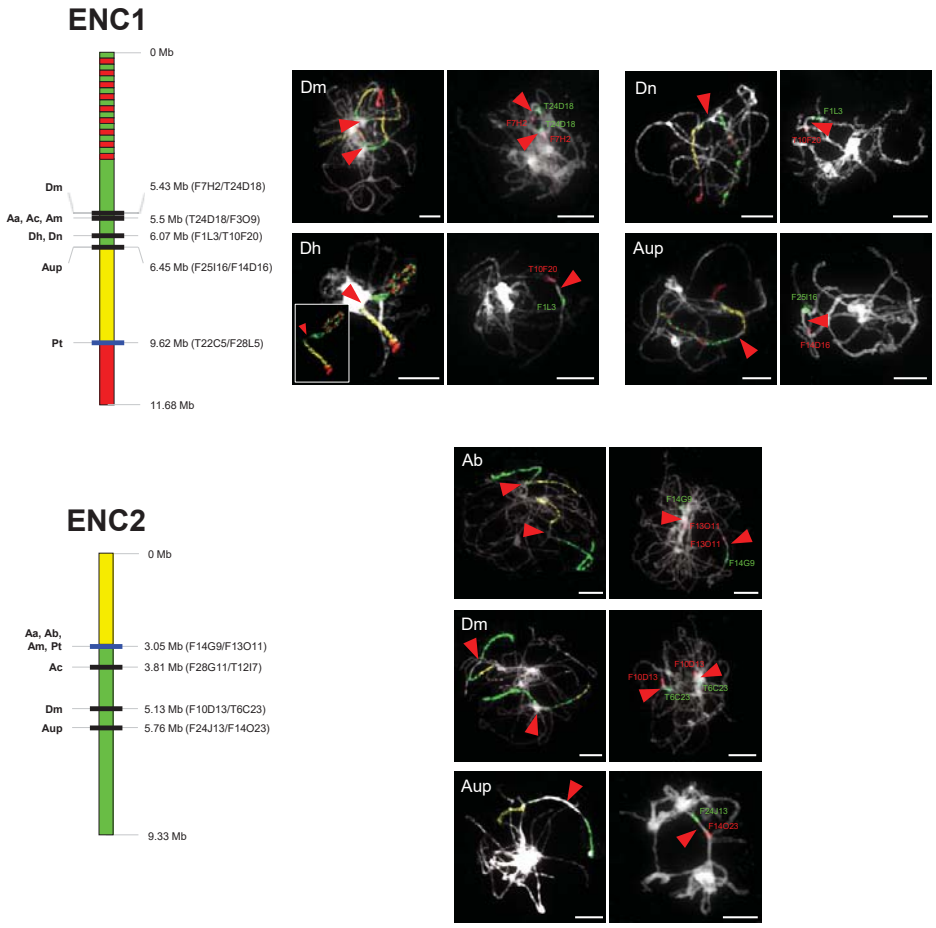


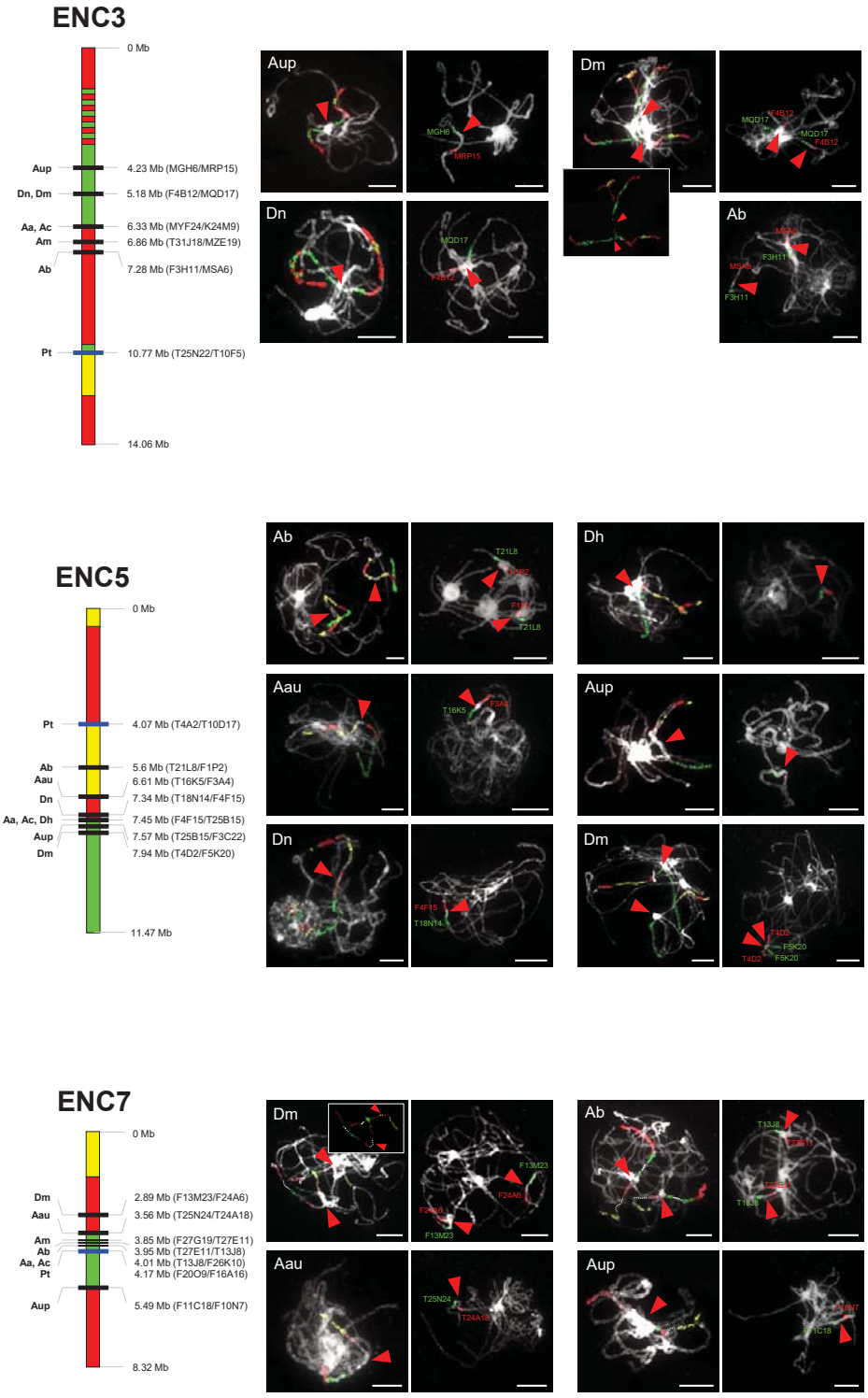
Supplemental Figure 3. Sequence comparison of shared or similar (peri-)centromeric tandem repeats identified. Supports Figure 3A and Supplemental Table 4. Each multiple aligment includes consensus monomer sequences and designed oligo probes (see Supplemental Table 2). **(A)** Multiple alignment of repats ArCy1 (495 bp; 2.70% of genome), ArAl1 (496 bp, 3.30%), ArAu2 (494 bp, 0.08%), and oligo probes Ar_univ4 and ArAu2; pairwise homology - ArCy1 and ArAl1: 92.6%, ArAl1 and ArAu2: 62.1%, ArCy1 and ArAu2 62.5%. **(B)** Multiple alignment of shared (identity 97.8 - 98.5%) 136-bp satellites ArAl2 (1.00%), ArCy2 (1.20%) and ArMo3 (0.01%), and the Ar_univ8 oligo probe. **(C)** Multiple alignment of satellite ArPl, ArAu4 (pairwise homology 97.6%) and the Ar_univ3 oligo probe used as a FISH probe in *Ar. blepharophylla*. **(D)** Multiple alignment of tandem repeat families in *D. hispida* (variants DrHi1 - DrHi4; 2.03%; monomer lengths: 144, 145 and two variants of 148 bp; pairwise homologies among the variants range between 78.4 and 93.2%) and in *D. nemorosa* (DrNe1 - DrNe5; 4.15%; monomer lenghts: 135 to 143 bp; pairwise homologies among the variants range between 81.8% and 95.6%), and DrHi and DrNe oligo probes. Pairwise homology between DrNe1 and DrHi1 - DrHi4 ranged between 76.4 and 80.4%, and sequence homology between variants DrNe2 -DrNe5 and DrHi1 - DrHi4 was varied between 68.9 and 75.7%. **(E)** Multiple alignment of satellite family variants PsTu1 (146 bp, 0.41%) and PsTu2 (175 bp, 0.14%), and the PsTu oligo probe; pairwise homology between the two variants was 83.6%.



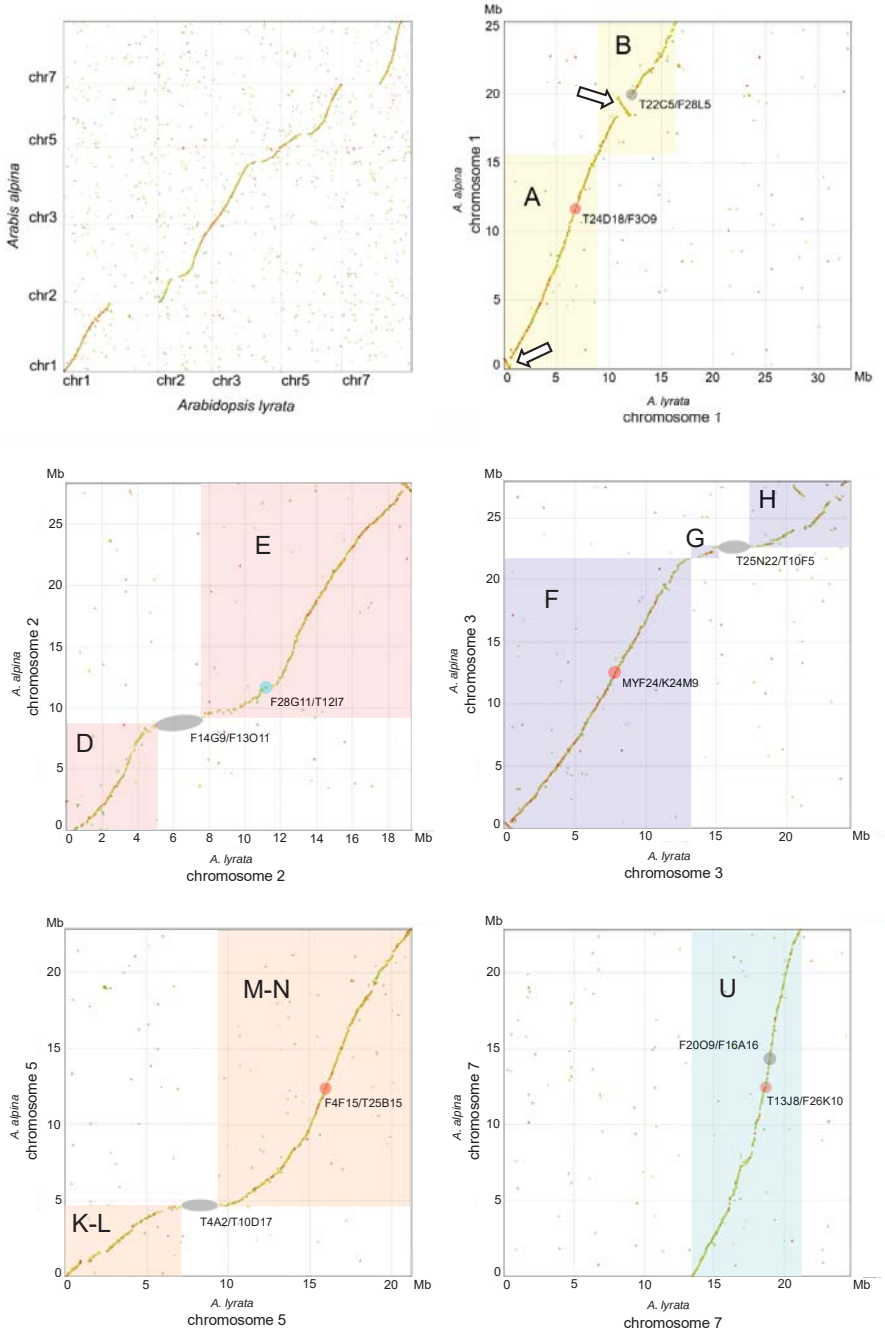
Supplemental Figure 4. Centromere repositioning events (ENCs) on five homeologous chromosomes of the analyzed Arabideae species. Supports Figure 4. Blue bars indicate the paleocentromeres in *Pseudoturritis turrita* (Pt), whereas black bars show the position of ENCs in crown-group Arabideae species (ENCs 1-3, 5 and 7). Centromere positions in Mb were inferred from FISH localization of centromere-facing arabidopsis BAC clones and approximated as the position of these BACs on the *Arabidopsis thaliana* pseudomolecules (www.arabidopsis.org). Examples of whole-chromosome painting experiments with differently labeled BAC clone contigs arranged as shown in the color schemes, and FISH of centromere-facing BACs, on pachytene DAPI-stained bivalents in *A. auriculata* (Aau), *A. blepharophylla* (Ab), *Aubrieta parviflora* (Aup), *Draba hispida* (Dh), *D. muralis* (Dm) and *D. nemorosa* (Dn). Red arrowheads point to the ENCs. Scale bars, 10 µm.

Detailed experimental evidence of centromere positions in *Arabis cypria* (Ac) and *P. turita* is shown in Figures 2, 3B and 4. Chromosome structure and centromere positions in *A. alpina* (Aa) and *A. montbretiana* (Am) are presented by Willing et al. (2015, *Nat Plants*) and Madrid et al. (2020, submitted), respectively.




ENC3
0 Mb
Aup
4.23 Mb (MGH6/MRP15)
Dn, Dm
5.18 Mb (F4B12/MQD17)
Aa, Ac
6.33 Mb (MYF24/K24M9)
Am
6.86 Mb (T31J18/MZE19)
Ab
7.28 Mb (F3H11/MSA6)
Pt
10.77 Mb (T25N22/T10F5)
14.06 Mb
Aup
MGH6
MRP15
Dm
MQD17
F4B12
Dn
MQD17
F4B12
Ab
MSA6
F3H11
ENC5
0 Mb
Pt
4.07 Mb (T4A2/T10D17)
Ab
5.6 Mb (T21L8/F1P2)
Aau
6.61 Mb (T16K5/F3A4)
Dn
7.34 Mb (T18N14/F4F15)
Aa, Ac, Dh
7.45 Mb (F4F15/T25B15)
Aup
7.57 Mb (T25B15/F3C22)
Dm
7.94 Mb (T4D2/F5K20)
11.47 Mb
Ab
T21L8
F1P2
F1P2
T21L8
Dh
Aau
T16K5
F3A4
Aup
Dn
F4F15
T18N14
Dm
T4D2
F5K20
T4D2
F5K20
ENC7
0 Mb
Dm
2.89 Mb (F13M23/F24A6)
Aau
3.56 Mb (T25N24/T24A18)
Am
3.85 Mb (F27G19/T27E11)
Ab
3.95 Mb (T27E11/T13J8)
Aa, Ac
4.01 Mb (T13J8/F26K10)
Pt
4.17 Mb (F20O9/F16A16)
Aup
5.49 Mb (F11C18/F10N7)
8.32 Mb
Dm
F13M23
F24A6
F24A6
F13M23
Ab
T13J8
T27E11
T13J8
Aau
T25N24
T24A18
Aup
F10N7
F11C18

Supplemental Figure 5. Comparison of homeologous chromosomes between *Arabis alpina* and *Arabidopsis lyrata*. Supports Figure 1. An inter-genome comparison of five homeologous chromosomes (1, 2, 3, 5 and 7) with ENCs in *A. alpina*. Arabidopsis BAC clones mark the position of ancestral centromeres in *A. lyrata* (grey ovals) and ENCs in *A. alpina* (red spheres, NB. – chromosome 2 centromere in *A. alpina* has the ancestral position, hence the ENC2 position in *A. cypria* is indicated instead – the cyan sphere). Capital letters refer to genomic blocks in the ACK genome. Two small inversions differentiating chromosome 1 homeologues were identified to be specific for *A. thaliana* chromosome 1, and absent in *A. alpina* and *A. lyrata* (data not shown). These assembly errors originated due to synteny-based scaffolding using the *A. thaliana* genome as a reference (Willing et al. 2015). Alignment of the two genomes was done using nucmer and visualized by mummerplot; data from Hu et al. (2011, *Nat Genet* 43) and Willing et al. (2015, *Nat Plant* 1).



Supplemental Table 1. The origin of Arabideae species analyzed.

Species	Chromosome number	Origin/Seed number/Herbarium voucher
Arabis alpina L.	16	Spain, the Cordillera Cantábrica Mts., Pajares, 42.5932N, 5.4532W; 1400 m a. s. l.; analyzed in Willing et al. (2015)
Arabis auriculata Lam.	14	ex cult. BG Heidelberg; originally: Italy: S. Pio delle Camere (AQ), 830 m a. s. l. B-2006-0556; HEID914439
	14	Czech Republic: Pálava Mts., Sirotčí hrádek (castle hill)
Arabis blepharophylla Hook. & Arn.	32	commercially available variety (grown as an ornamental plant)
Arabis cypria Holmboe	16	Cyprus: Pentadaktylos Peak, Kyrenia range. B-2010-0201; HEID930016
Arabis montbretiana Boiss.	16	Turkey: Hakkari, between Şemdinli-Şapatan passageway, 2 km from Şemdinli. BM7968; HEID809801, RK043 in Karl and Koch (2013); analyzed in Madrid et al. (2019),
Arabis planisiliqua (Pers.) Rchb.	16	National Botanical Garden of Georgia
Aubrieta canescens subs. *macrostyla* Cullen & Hub.	16	Turkey: north-east Turkey, Silvas Province, B-2015-0013-2; HEID923307
Aubrieta parviflora Boiss.	16	Iran: Montes de Fashand, B-2015-0038-6; HEID923332
Draba hispida Willd.	16	Georgia: Guria region, sourroundings of Bakhmaro. B-2010-0292-17; HEID920526,
Draba muralis L.	32	Slovakia: Vinosady, c. 500 m NW from the village, approx. 48.3166122N, 17.2840417E
Draba nemorosa L.	16	Slovakia: Heľpa, the Heľpa railway station, along the railway tracks
Pseudoturritis turrita (L.) Al-Shehbaz	16	Slovakia: Vinosady, forest c. 1 km NW from the village, approx. 48.3226406N, 17.2799728E

References

Karl R, Koch MA (2013) A world-wide perspective on crucifer speciation and evolution: phylogenetics, biogeography and trait evolution in tribe Arabideae. *Ann Bot* 112: 983-1001.

Madrid E, Kiefer C, Jiao W-B, Severing E, Mandáková T, de Ansorena E, Walkemeier B, Lysak MA, Schneeberger K, Coupland G (2019) Analysis of divergence of annual and perennial crucifers by genome comparisons and interspecific genetics reveals contribution of reproductive assurance (submitted).

Willing E-M, Rawat V, Mandáková T, Maumus F, Velikkakam James G, Nordström KJV, Becker C, Warthmann N, Chica C, Szarzynska B, Zytnicki M, Albani MC, Kiefer C, Bergonzi S, Castaings L, Mateos JL, Berns MC, Bujdoso N, Piofczyk T, de Lorenzo L, Barrero-Sicilia C, Mateos I, Piednoël M, Hagmann J, Chen-Min-Tao R, Iglesias-Fernández R, Schuster SC, Alonso-Blanco C, Roudier F, Carbonero P, Paz-Ares J, Davis SJ, Pecinka A, Quesneville H, Colot V, Lysak MA, Weigel D, Coupland G, Schneeberger K (2015) Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nat Plants* 1: 14023.

Supplemental Table 2. Divergence time estimates. Divergence time estimates are based on the tribal-level phylogeny and divergence time estimation (internal transcribed spacer of nuclear ribosomal DNA, ITS) constraint by a nuclear-gene-based core-phylogeny (Karl and Koch 2013).

	Species	Time of origin (stem group age, mya)	Time of diversification (crown group age, mya)
Arabis alpina clade	12	6 - 10	2.7 - 3.3
Arabis auriculata clade	3	11 - 12	3.0 - 10.2
Arabis nordmanniana clade	5	6 - 10	1.1 - 1.4
Draba & *Tomostima* clade	>400	10 - 12	6.7 - 10.9
Arabis aucheri clade	2	10 - 12	0.8 - 0.9
Aubrieta clade	16	11 - 12.5	2.7 - 5.2
main *Arabis* clade	>65	10 - 13	6.6 - 7.1
Scapiarabis clade	12	11.5 - 14	4.2 - 10.6

Reference

Karl R, Koch MA (2013) A world-wide perspective on crucifer speciation and evolution: phylogenetics, biogeography and trait evolution in tribe Arabideae. *Ann Bot* 112: 983-1001.

Supplemental Table 3. Spectrum and genome proportions of repeat families identified in the sequenced Arabideae genomes. Genome proportions and classification of repeat families based on graph-based clustering using RepeatExplorer pipeline for each analyzed genome (0.25× genome coverage). ArAl: *Arabis alpina*, ArAu: *Arabis auriculata*, ArCy: *Arabis cypria*, ArMo: *Arabis montbretiana*, AuCa: *Aubrieta canescens*, DrHi: *Draba hispida*, DrNe: *D. nemorosa*, PsTu: *Pseudoturritis turrita*.

Repeat family (%)	ArAl	ArAu	ArCy	ArMo	AuCa	DrHi	DrNe	PsTu
LTR retrotransposon	18.36	3.71	14.88	6.70	23.13	13.38	5.85	10.51
Ty3/gypsy	14.92	2.68	11.61	6.56	20.93	10.87	4.36	9.05
Athila	4.37	0.70	4.31	3.00	4.79	3.82	1.25	2.64
Chromovirus	6.20	0.21	5.12	1.63	14.50	5.57	2.11	1.23
CRM	0.11	0.06	0.12	0.00	0.01	0.49	0.17	0.83
Galadriel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
Reina	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.02
Tekay	6.09	0.06	5.00	1.63	14.48	5.08	1.94	0.30
Tat/Ogre	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
Tat/Retand	4.09	0.40	1.62	1.93	1.44	1.32	1.00	3.49
Ty1/copia	3.44	1.03	3.27	0.14	2.20	2.51	1.49	1.46
Ale	1.30	0.53	0.40	0.00	0.57	0.46	0.48	0.04
Angela	0.00	0.00	0.00	0.00	0.00	0.27	0.00	0.09
Bianca	0.91	0.17	0.79	0.06	0.55	0.65	0.45	0.54
Ikeros	0.05	0.03	0.14	0.00	0.05	0.12	0.00	0.00
Ivana	0.02	0.16	0.07	0.00	0.11	0.05	0.00	0.02
SIRE	0.09	0.00	0.08	0.01	0.02	0.03	0.16	0.00
TAR	0.95	0.00	1.62	0.06	0.58	0.51	0.18	0.63
Tork	0.13	0.04	0.05	0.00	0.00	0.39	0.16	0.15
LINE	0.13	0.09	0.31	0.02	0.08	0.09	0.01	0.03
DNA transposon	3.14	1.37	3.37	0.25	4.30	2.29	1.26	2.13
CACTA	0.54	0.08	0.79	0.05	0.24	0.14	0.47	0.36
hAT	0.00	0.02	0.18	0.00	0.02	0.00	0.00	0.01
Harbinger	0.11	0.04	0.02	0.01	0.13	0.12	0.00	0.13
Helitron	0.34	0.45	0.17	0.04	0.03	0.00	0.00	0.55
Mariner	0.25	0.21	0.54	0.13	0.73	0.22	0.00	0.00
MITE	0.00	0.13	0.23	0.00	0.00	0.00	0.33	0.00
MuDR	1.67	0.04	0.09	0.01	1.90	0.83	0.46	0.60
Pararetrovirus	0.00	0.00	0.00	0.00	0.12	0.03	0.00	0.00
rDNA	0.12	1.11	0.70	0.04	2.02	2.65	1.35	1.99
Satellite	2.01	1.17	5.81	0.05	2.85	2.41	4.36	1.78
Unclassified repeats	7.16	3.75	7.54	4.10	5.61	8.31	7.61	10.31
Low/single copy sequences	69.07	88.79	67.39	88.85	61.89	70.83	79.57	73.24
All repeats (%)	**30.93**	**11.21**	**32.61**	**11.15**	**38.11**	**29.17**	**20.43**	**26.76**
Genome size (Mb)[a]	**372**	**205**	**382**	**275**	**392**	?	**313**	**372**

[a] see Supplemental Table 4 for data sources.

Supplemental Table 4. List of centromeric tandem repeats identified in the sequenced repeatomes of Arabideae species and corresponding FISH probes.

Species	Genome size (pg and Mb/1C)[a]		Source (genome size)	Tandem repeat	Monomer length (bp)	Genome proportion (%)[b]	Oligo-probe for FISH
Arabis alpina	0.38	372	Lysak *et al.* (2009)	ArAl1	496	3.30	Ar_univ4[c]
A. auriculata	0.21	205	our estimate	ArAu1	621	0.13	ArAu1[d]
				ArAu2	494	0.08	ArAu2
				ArAu3	875	0.08	ArAu3[d]
A. cypria	0.39	382	our estimate	ArCy1	495	2.70	Ar_univ4[c]
A. blepharophylla	-		-	-	-	-	Ar_univ3[e]
A. montbretiana	0.28	275	Hoffmann *et al.* (2010)	ArMo3	136	0.01	Ar_univ8[f]
Aubrieta canescens	0.40	392	BrassiBase	AuCa	170	3.20	AuCa
Au. parviflora	0.41	401	BrassiBase	-	-	-	AuCa
Draba hispida	-		-	DrHi1 - DrHi4	144, 145, 148	2.03	DrHi
D. nemorosa	0.32/ 0.24	313/ 235	our estimate /Johnston *et al.* (2005)	DrNe1 - DrNe5	135, 136, 143	4.15	DrNe
Pseudoturritis turrita	0.38	372	Lysak *et al.* (2009)	PsTu1, PsTu2	146, 175	0.55	PsTu

[a] 1 pg = 978 Mb, Doležel J, Bartoš J, Voglmayr H, Greilhuber J (2003) Nuclear DNA content and genome size of trout and human. *Cytometry* 51: 127-128.

[b] genome proportion calculated by the RepeatExplorer pipeline.

[c] Ar_univ4 probe was designed from a consensus 495-bp satellite sequence shared by *A. alpina* (ArAl1) and *A. cypria* (ArCy1) (pairwise identity 92.6%, 460 identical sites).

[d] PCR amplification of selected tandem repeats with monomer length >500 bp.

[e] Ar_univ3 probe was designed from a consensus 170-bp satellite sequence shared by *Arabis auriculata* (ArAu4) and *A. planisiliqua* (ArPl) (pairwise identity 97.6%, 166 identical sites).

[f] Ar_univ8 probe was designed from a shared 136-bp satellite identified in repeatomes of *A. alpina* (ArAl2), *A. cypria* (ArCy2) and *A. montbretiana* (ArMo3) (pairwise identity between 97.8 and 98.5%).

References

BrassiBase: https://brassibase.cos.uni-heidelberg.de/, visited on June 18, 2019.

Hoffmann MH, Schmuths H, Koch C, Meister A, Fritsch RM (2010) Comparative analysis of growth, genome size, chromosome numbers and phylogeny of *Arabidopsis thaliana* and three cooccurring species of the Brassicaceae from Uzbekistan. *J. Bot.* 2010: 504613.

Johnston JS1, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, Lopez R, Price HJ (2005) Evolution of genome size in Brassicaceae. *Ann. Bot.* 95:229-235.

Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol. Biol. Evol.* 26:85-98.

4 FINAL REMARKS

In my thesis, I have focused on the utilization of low-pass genome sequencing data using modern bioinformatic approaches to explore and elucidate the structure of plant genomes from the Brassicaceae family. NGS data were used to analyze genomic repeat abundances and to assemble of high-copy organellar chloroplast DNA.

By phylogenetic analysis based on whole chloroplast sequences we resolved tribal relationships within the *Hesperis* clade and dated the diversification in this clade. This study allows more insight into mechanisms responsible for genome size growth. The increase in genome size in *Hesperis* clade was due to an increase in repetitive DNA content, specifically Ty3/*gypsy* transposable elements. *In silico* identified satellite sequences were used as FISH markers to delimit (peri)centromeric regions of chromosomes in Arabideae species.

5 BIBLIOGRAPHY

Abascal, F., Zardoya, R. and Posada, D. (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104-2105.

Abrusán, G., Grundmann, N., DeMester, L. and Makalowski, W. (2009) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, **25**, 1329-1330.

Ågren, J. A. and Clark, A. G. (2018) Selfish genetic elements. *PLoS Genetics*, **14**, e1007700.

Alverson, W. S., Whitlock, B. A., Nyffeler, R., Bayer, C. and Baum, D. A. (1999) Phylogeny of the core Malvales: evidence from *ndhF* sequence data. *American Journal of Botany*, **86**, 1474-1486.

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I. J., Koblížková, A., Macas, J. and Lysak, M. A. (2011) Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. *Annals of Botany*, **107**, 255-268.

Ananiev, E. V., Phillips, R. L. and Rines, H. W. (1998) Chromosome-specific molecular organization of maize (Zea mays L.) centromeric regions. *Proceedings of the National Academy of Sciences*, **95**, 13073-13078.

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

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796-815.

Arumuganathan, K. and Earle, E. D. (1991) Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter*, **9**, 208-218.

Ávila Robledillo, L., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., Schubert, I., and Macas, J. (2018). Satellite DNA in Vicia faba is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Scientific reports*, **8**, 1-11.

Bao, W., Kojima, K. K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.

Bao, Z. and Eddy, S. R. (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research*, **12**, 1269-1276.

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P. and Lander, E. S. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Research*, **12**, 177-189.

Bendich, A. J. (2004) Circular chloroplast chromosomes: the grand illusion. *Plant Cell*, **16**, 1661-1666.

Bennett, M. D. and Smith, J. B. (1976) Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, **274**, 227-274.

Bennett, M. D. and Smith, J. B. (1991) Nuclear DNA amounts in angiosperms. *Philosophical Transactions: Biological Sciences*, **334**, 309-345.

Bennett, M. D., Leitch, I. J., Price, H. J. and Johnston, J. S. (2003) Comparisons with *Caenorhabditis* (∼100 Mb) and Drosophila (∼175 Mb) using flow cytometry show genome size in Arabidopsis to be ∼157 Mb and thus ∼25% larger than the Arabidopsis genome initiative estimate of ∼125 Mb. *Annals of Botany*, **91**, 547-557.

Bennetzen, J. L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology*, **42**, 251-269.

Bennetzen, J. L. (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, **115**, 29-36.

Bennetzen, J. L. and Kellogg, E. A. (1997) Do plants have a one-way ticket to genomic obesity?. *Plant Cell*, **9**, 1509.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573-580.

Berglund, E. C., Kiialainen, A. and Syvänen, A. C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, **2**, 23.

Bergman, C. M. and Quesneville, H. (2007) Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, **8**, 382-392.

Biscotti, M. A., Canapa, A., Forconi, M., Olmo, E. and Barucca, M. (2015b) Transcription of tandemly repetitive DNA: functional roles. *Chromosome Research*, **23**, 463-477.

Biscotti, M. A., Olmo, E. and Heslop-Harrison, J. P. (2015a) Repetitive DNA in eukaryotic genomes. *Chromosome Research*, **23**, 415–420.

Boetzer, M. and Pirovano, W. (2012) Toward almost closed genomes with GapFiller. *Genome Biology*, **13**, R56.

Boisvert, S., Laviolette, F. and Corbeil, J. (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, **17**, 1519-1533.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M.A., Rambaut, A. and Drummond, A. J. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, **10**, e1003537.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L. and Feschotte, C. (2018) Ten things you should know about transposable elements. *Genome Biology*, **19**, 1-12.

Brocchieri, L. (2001) Phylogenetic inferences from molecular sequences: review and critique. *Theoretical population biology*, **59**, 27-40.

Bureau, T. E. and Wessler, S. R. (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell*, **4**, 1283-1294.

Bushnell, B. (2014) BBMap: a fast, accurate, splice-aware aligner (No. LBNL-7065E). *Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).*

Calonje, M., Martín-Bravo, S., Dobeš, C., Gong, W., Jordon-Thaden, I., Kiefer, C., Kiefer, M., Paule, J., Schmickl, R. and Koch, M. A. (2009) Non-coding nuclear DNA markers in phylogenetic reconstruction. *Plant Systematics and Evolution*, **282**, 257-280.

Cappello, J., Handelsman, K. and Lodish, H. F. (1985) Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. *Cell*, **43**, 105-115.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540-552.

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C. R., Gu, M., Blattner, F. R. and Jiang, J. (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell*, **14**, 1691-1704.

Chu, C., Nielsen, R. and Wu, Y. (2016) REPdenovo: inferring de novo repeat motifs from short sequence reads. *PloS ONE*, **11**, e0150719.

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L. and Jansen, R. K. (2006) The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molecular Biology and Evolution*, **23**, 2175-2190.

Čížková, J., Hřibová, E., Humplíková, L., Christelová, P., Suchánková, P. and Doležel, J. (2013) Molecular analysis and genomic organization of major DNA satellites in banana (Musa spp.). *PloS ONE*, **8**, e54808.

Clark, L. G., Zhang, W. and Wendel, J. F. (1995) A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. *Systematic Botany*, **20**, 436-460.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, **38**, 1767-1771.

de Vries, J., Habicht, J., Woehle, C., Huang, C., Christa, G., Wägele, H., Nickelsen, J., Martin, W. F. and Gould, S. B. (2013) Is *ftsH* the key to plastid longevity in sacoglossan slugs?. *Genome Biology and Evolution*, **5**, 2540-2548.

Desai, A., Marwah, V. S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V. and Jere, A. (2013) Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PloS ONE*, **8**, e60204.

Devos, K. M., Brown, J. K. and Bennetzen, J. L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Research*, **12**, 1075-1079.

Dierckxsens, N., Mardulyn, P. and Smits, G. (2016) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, **45**, e18-e18.

Díez, C. M., Gaut, B. S., Meca, E., Scheinvar, E., Montes-Hernandez, S., Eguiarte, L. E. and Tenaillon, M. I. (2013) Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytologist*, **199**, 264-276.

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 Systematics and Evolution*, **303**, 1013-1020.

Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, **36**, e105.

Doolittle, W. F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601.

Du, C., Caronna, J., He, L. and Dooner, H. K. (2008) Computational prediction and molecular confirmation of *Helitron* transposons in the maize genome. *BMC Genomics*, **9**, 51.

Eddy, S. R. (2009) A new generation of homology search tools based on probabilistic inference. *In Genome Informatics 2009: Genome Informatics Series Vol. 23*, 205-211.

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792-1797.

Edgar, R. C. and Myers, E. W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics*, **21**, i152-i158.

Eichler, E. E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *TRENDS in Genetics*, **17**, 661-669.

Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

Elliott, T. A. and Gregory, T. R. (2015) Do larger genomes contain more diverse transposable elements?. *BMC Evolutionary Biology*, **15**, 69.

Estep, M. C., DeBarry, J. D. and Bennetzen, J. L. (2013) The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity*, **110**, 194.

Evgen'ev, M. B. and Arkhipova, I. R. (2005) *Penelope*-like elements – a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenetic and Genome Research*, **110**, 510-521.

Evgen'ev, M. B., Zelentsova, H., Shostak, N., Kozitsina, M., Barskyi, V., Lankenau, D. H. and Corces, V. G. (1997) *Penelope*, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis*. *Proceedings of the National Academy of Sciences*, **94**, 196-201.

Ewing, A. D. (2015) Transposable element detection from whole genome sequence data. *Mobile DNA*, **6**, 24.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186-194.

Fajkus, P., Peška, V., Sitová, Z., Fulnečková, J., Dvořáčková, M., Gogela, R., Sýkorová, E., Hapala, J. and Fajkus, J. (2016) *Allium* telomeres unmasked: the unusual telomeric sequence $(CTCGGTTATGGG)_n$ is synthesized by telomerase. *Plant Journal*, **85**, 337-347.

Fertin, G., Jean, G., Radulescu, A. and Rusu, I. (2015) Hybrid de novo tandem repeat detection using short and long reads. *BMC Medical Genomics*, **8**, S5.

Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M. L. and Levine, D. (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biology and Evolution*, **1**, 205-220.

Finn, R. D., Clements, J. and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, **39**, W29-W37.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. and Bateman A. (2010) The Pfam protein families database. *Nucleic Acids Research*, **38**, D211-D222.

Finnegan, D. J. (1989) Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, **5**, 103-107.

Fleischmann, A., Michael, T. P., Rivadavia, F., Sousa, A., Wang, W., Temsch, E. M., Greilhuber, J., Müller, K. F. and Heubl, G. (2014) Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany*, **114**, 1651-1663.

Fultz, D., Choudury, S. G. and Slotkin, R. K. (2015) Silencing of active transposable elements in plants. *Current Opinion in Plant Biology*, **27**, 67-76.

Gao, C., Xiao, M., Ren, X., Hayward, A., Yin, J., Wu, L., Fu, D. and Li, J. (2012) Characterization and functional annotation of nested transposable elements in eukaryotic genomes. *Genomics*, **100**, 222-230.

Garrido-Ramos, M. (2017) Satellite DNA: An evolving topic. *Genes* **8**, 230.

Garrido-Ramos, M. A. (2015) Satellite DNA in plants: more than just rubbish. *Cytogenetic and Genome Research*, **146**, 153-170.

Gill, N., Findley, S., Walling, J. G., Hans, C., Ma, J., Doyle, J., Stacey, G. and Jackson, S. A. (2009) Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiology*, **151**, 1167-1174.

Gnerre, S., MacCallum, I., Przybylski, *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, **108**, 1513-1518.

Goerner-Potvin, P. and Bourque, G. (2018) Computational tools to unmask transposable elements. *Nature Reviews Genetics*, **19**, 688-704.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D. S. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, **40**, D1178-D1186.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**, 333.

Goodwin, T. J. and Poulter, R. T. (2004) A new group of tyrosine recombinase-encoding retrotransposons. *Molecular Biology and Evolution*, **21**, 746-759.

Gordon, A. and Hannon, G. J. (2010) Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab.cshl.edu/fastx_toolkit, 5.

Grandbastien, M. A., Audeon, C., Bonnivard, E., Casacuberta, J. M., Chalhoub, B., Costa, A. P., Le, Q. H., Melayah, D., Petit, M., Poncet, C., Tam, S. M., van Sluys M. A. and Mhiri C. (2005) Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenetic and Genome Research*, **110**, 229-241.

Green, B. R. (2011) Chloroplast genomes of photosynthetic eukaryotes. *Plant journal*, **66**, 34-44.

Gregory, T. R. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological reviews*, **76**, 65-101.

Greiner, S., Lehwark, P. and Bock, R. (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research*, **47**, W59-W64.

Guo, R., Li, Y. R., He, S., Ou-Yang, L., Sun, Y. and Zhu, Z. (2018) RepLong: *de novo* repeat identification using long read sequencing data. *Bioinformatics*, **34**, 1099-1107.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* **29**, 1072-1075.

Hahn, C., Bachmann, L. and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, **41**, e129-e129.

Han, Y. and Wessler, S. R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, **38**, e199-e199.

Hansen, A. K., Escobar, L. K., Gilbert, L. E. and Jansen, R. K. (2007) Paternal, maternal and biparental inheritance of the chloroplast genome in *Passiflora* (Passifloraceae): implications for phylogenetic studies. *American Journal of Botany*, **94**, 42-46.

Hawkins, J. S., Kim, H., Nason, J. D., Wing, R. A. and Wendel, J. F. (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Research*, **16**, 1252-1261.

He, L., Liu, J., Torres, G. A., Zhang, H., Jiang, J. and Xie, C. (2013) Interstitial telomeric repeats are enriched in the centromeres of chromosomes in *Solanum* species. *Chromosome Research*, **21**, 5-13.

Henikoff, S., Ahmad, K. and Malik, H. S. (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293**, 1098-1102.

Heslop-Harrison, J. S., Brandes, A. and Schwarzacher, T. (2003) Tandemly repeated DNA sequences and centromeric chromosomal regions of *Arabidopsis* species. *Chromosome Research*, **11**, 241-253.

Heslop-Harrison, J. S., Murata, M., Ogura, Y., Schwarzacher, T. and Motoyoshi, F. (1999) Polymorphisms and genomic organization of repetitive DNA from centromeric regions of Arabidopsis chromosomes. *Plant Cell*, **11**, 31-42.

Hillis, D. M. and Bull, J. J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic biology*, **42**, 182-192.

Hloušková, P., Mandáková, T., Pouch, M., Trávníček, P. and Lysak, M. A. (2019) The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing. *Annals of Botany*, **124**, 103-120.

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.

Hollister, J. D. and Gaut, B. S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research*, **19**, 1419-1428.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H. and Haberer, G. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, **43**, 476-481.

Huang, D. I. and Cronk, Q. C. (2015) Plann: A command-line application for annotating plastome sequences. *Applications in Plant Sciences*, **3**, 1500026.

Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Research*, **9**, 868-877.

Huelsenbeck, J. P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754-755.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, et al.*et al.* (2008) InterPro: the integrative protein signature database. *Nucleic Acids Research*, **37**, D211-D215.

International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711-716.

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Iorizzo, M., Senalik, D., Szklarczyk, M., Grzebelus, D., Spooner, D. and Simon, P. (2012) *De novo* assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. *BMC Plant Biology*, **12**, 61.

Jaillon, O., Aury, J. M., Noel, B., *Policriti, A., Clepet, C., Casagrande, A., et al.et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463-467.

Janicki, M., Rooke, R. and Yang, G. (2011) Bioinformatics, and genomic analysis of transposable elements in eukaryotic genomes. *Chromosome Research*, **19**, 787.

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., DePamphilis, C. W., Leebens-Mack, J., Müller, K. F., Guisinger-Bellian, M., Haberle, R. C., Hansen, A. K., Chumley, T. W., Lee, S. B., Peery, R., McNeal, J. R., Kuehl, J.V. and Boore, J. L. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences*, **104**, 19369-19374.

Jiang, J., Birchler, J. A., Parrott, W. A. and Dawe, R. K. (2003) A molecular view of plant centromeres. *Trends in Plant Science*, **8**, 570-575.

Jiang, N. and Wessler, S. R. (2001) Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. *Plant Cell*, **13**, 2553-2564.

Jiao, Y., Peluso, P., Shi, J., *Liang, T., Stitzer, M. C., Wang, B., et al.et al. (* 2017) Improved maize reference genome with single-molecule technologies. *Nature*, **546**, 524.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S. and Soltis, D. E. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97-100.

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., Yi, T. S. and Li, D. Z. (2018) GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular chloroplast genome using genome skimming data. *BioRxiv*, 256479.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research*, **110**, 462-467

Kalendar, R., Tanskanen, J., Chang, W., Antonius, K., Sela, H., Peleg, O. and Schulman, A. H. (2008) *Cassandra* retrotransposons carry independently transcribed 5S RNA. *Proceedings of the National Academy of Sciences*, **105**, 5833-5838.

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A. and Schulman, A. H. (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics*, **166**, 1437-1450.

Kalyanaraman, A. and Aluru, S. (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. *Journal of Bioinformatics and Computational Biology*, **4**, 197-216.

Kapitonov, V. V. and Jurka, J. (2001) Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, **98**, 8714-8719.

Kapitonov, V. V. and Jurka, J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, **9**, 411-412.

Katoh, K. and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772-780.

Katoh, K., Misawa, K., Kuma, K. I. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059-3066.

Kearse, M., Moir, R., Wilson, A., *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647-1649.

Kejnovský, E., Michalovova, M., Steflova, P., Kejnovska, I., Manzano, S., Hobza, R., Kubat, Z., Kovarik, J., Jamilena, M. and Vyskot, B. (2013) Expansion of microsatellites on evolutionary young Y chromosome. *PloS ONE*, **8**, e45519.

Kelly, L. J. and Leitch, I. J. (2011) Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Research*, **19**, 939-953.

Kelly, L. J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., Lysak, M. A., Day, P. D., Berger, M., Fay, M. F., Nichols, R. A., Leitch, A. R. and Leitch I. J. (2015) Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytologist*, **208**, 596-607.

Kiefer, M., Schmickl, R., German, D. A., Mandáková, T., Lysak, M. A., Al-Shehbaz, I. A., Franzke, A., Mummenhoff, K., Stamatakis, A. and Koch, M. A. (2014) BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. *Plant and Cell Physiology*, **55**, e3-e3.

Koch, P., Platzer, M. and Downie, B. R. (2014) RepARK—de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research*, **42**, e80-e80.

Kohany, O., Gentles, A. J., Hankus, L. and Jurka, J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, **7**, 474.

Koo, D. H., Hong, C. P., Batley, J., Chung, Y. S., Edwards, D., Bang, J. W., Hur, Y. and Lim, Y. P. (2011) Rapid divergence of repetitive DNAs in Brassica relatives. *Genomics*, **97**, 173-185.

Kurtz, S. and Schleiermacher, C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426–427.

Kurtz, S., Narechania, A., Stein, J. C. and Ware, D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, **9**, 517.

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. *Molecular Biology and Evolution*, **34**, 772-773.

Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357.

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 Biology*, **10**, R25.

Laslett, D. and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, **32**, 11-16.

Laten, H. M., Majumdar, A. and Gaucher, E. A. (1998) *SIRE-1*, a *copia/Ty1*-like retroelement from soybean, encodes a retroviral envelope-like protein. *Proceedings of the National Academy of Sciences*, **95**, 6897-6902.

Lee, D. J., Blake, T. K. and Smith, S. E. (1988) Biparental inheritance of chloroplast DNA and the existence of heteroplasmic cells in alfalfa. *Theoretical and Applied Genetics*, **76**, 545-549.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W. R. and Schatz, M. (2016) Third-generation sequencing and the future of genomics. *BioRxiv*, 048603.

Leeton, P. R. and Smyth, D. R. (1993) An abundant LINE-like element amplified in the genome of *Lilium speciosum*. *Molecular and General Genetics MGG*, **237**, 97-104.

Leitch, I. J., Beaulieu, J. M., Chase, M. W., Leitch, A. R. and Fay, M. F. (2010) Genome size dynamics and evolution in monocots. *Journal of Botany*, 2010.

Lerat, E. (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, **104**, 520-533.

Levy, S. E. and Myers, R. M. (2016) Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, **17**, 95-115.

Li, F. W. and Harkess, A. (2018) A guide to sequence your favorite plant genomes. *Applications in Plant Sciences*, **6**, e1030.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G. K. and Wang, J. (2005) ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Computational Biology*, **1**, e43.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J. and Wang, J. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, **20**, 265-272.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B. and Fan, W. (2012) Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, **11**, 25-37.

Lilly, J. W., Havey, M. J., Jackson, S. A. and Jiang, J. (2001) Cytogenomic analyses reveal the structural plasticity of the chloroplast genome in higher plants. *Plant Cell*, **13**, 245-254.

Lisch, D. (2013) How important are transposons for plant evolution?. *Nature Reviews Genetics*, **14**, 49-61.

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X. and Guan, X. (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, **13**, 715.

Llorens, C., Futami, R., Bezemer, D. and Moya, A. (2007) The *Gypsy* Database (GyDB) of mobile genetic elements. *Nucleic Acids Research*, **36**, D38-D46.

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G. P. , Maumus, F., Munoz-Pomer, A., Sempere, J. M., Latorre, A. and Moya, A. (2010) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Research*, **39**, D70-D74.

Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H. and Moya, A. (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biology direct*, **4**, 41.

Lowe, T. M. and Chan, P. P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research*, **44**, W54-W57.

Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics & Bioinformatics*, **14**, 265-279.

Lucier, J. F., Perreault, J., Noel, J. F., Boire, G. and Perreault, J. P. (2007). RTAnalyzer: a web application for finding new retrotransposons and detecting L1 retrotransposition signatures. *Nucleic Acids Research*, **35**, W269-W274.

Ma, P. F., Zhang, Y. X., Guo, Z. H. and Li, D. Z. (2015) Evidence for horizontal transfer of mitochondrial DNA to the plastid genome in a bamboo genus. *Scientific Reports*, **5**, 11608.

Macas, J. and Neumann, P. (2007) Ogre elements—a distinct group of plant Ty3/gypsy-like retrotransposons. *Gene*, **390**, 108-116.

Macas, J., Novak, P., Pellicer, J., Čížková, J., Koblížková, A., Neumann, P., Fuková, I., Doležel, J., Kelly, L. J. and Leitch, I. J. (2015) In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. *PloS ONE*, **10**, e0143424.

Macas, J., Požárková, D., Navrátilová, A., Nouzová, M. and Neumann, P. (2000) Two new families of tandem repeats isolated from genus Vicia using genomic self-priming PCR. *Molecular and General Genetics MGG*, **263**, 741-751.

Mader, M., Pakull, B., Blanc-Jolivet, C., Paulini-Drewes, M., Bouda, Z., Degen, B., Small, I. and Kersten, B. (2018) Complete chloroplast genome sequences of four Meliaceae species and comparative analyses. *International Journal of Molecular Sciences*, **19**, 701.

Malé, P. J. G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., Lhuillier, E., Scotti-Saintagne, C., Tinaut, A. and Chave, J. (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources*, **14**, 966-975.

Mandáková, T. and Lysak, M. A. (2018) Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, **42**, 55-65.

Mandáková, T., Hloušková, P., German, D. A. and Lysak, M. A. (2017) Monophyletic origin and evolution of the largest crucifer genomes. *Plant Physiology*, **174**, 2062-2071.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al.*et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, **17**, 10-12.

Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, **74**, 560-564.

McCarthy, E. M. and McDonald, J. F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, **19**, 362-367.

McClintock, B. (1950) The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Science*, **36**, 344-355.

McKinley, K. L. and Cheeseman, I. M. (2016) The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology*, **17**, 16.

Mehrotra, S. and Goyal, V. (2014) Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics, proteomics & Bioinformatics*, **12**, 164-171.

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J. F., DeRisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., Chan, S. W. L. (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, **14**, R10.

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., Loudet, O., Weigel, D. and Ecker, J. R. (2018) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nature Communications*, **9**, 1-8.

Miller, J. R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315-327.

Moore, M. J., Bell, C. D., Soltis, P. S. and Soltis, D. E. (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences*, **104**, 19363-19368.

Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. and Soltis, D. E. (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences*, **107**, 4623-4628.

Myers, E. W., Sutton, G. G., Delcher *et al.* (2000) A whole-genome assembly of *Drosophila*. *Science*, 287, 2196-2204.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J. and Macas, J. (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA*, **2**, 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.

Novák, P., Ávila Robledillo, L., 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 Research*, **45**, e111-e111.

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.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., et al.*et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579-584.

O'Dushlaine, C.T. and Shields, D. C. (2006) Tools for the identification of variable and potentially variable tandem repeats. *BMC Genomics*, **7**, 290.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H., Ozeki, H. (1986) Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*, **322**, 572-574.

Okamoto, H. and Hirochika, H. (2001) Silencing of transposable elements in plants. *Trends in Plant Science*, **6**, 527-534.

Oliver, K. R. and Greene, W. K. (2009) Transposable elements: powerful facilitators of evolution. *Bioessays*, **31**, 703-714.

Orgel, L. E. and Crick, F. H. (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604-607.

Palmer, J. D. (1985) Comparative organization of chloroplast genomes. *Annual Review of Genetics*, **19**, 325-354.

Pélissier, T., Tutois, S., Deragon, J. M., Tourmente, S., Genestier, S. and Picard, G. (1995) *Athila*, a new retroelement from *Arabidopsis thaliana*. *Plant Molecular Biology*, **29**, 441-452.

Pellicer, J., Fay, M. F. and Leitch, I. J. (2010) The largest eukaryotic genome of them all?. *Botanical Journal of the Linnean Society*, **164**, 10-15.

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.

Peška, V., Fajkus, P., Fojtová, M., Dvořáčková, M., Hapala, J., Dvořáček, V., Polanská, P., Leitch, A.R., Sýkorová, E. and Fajkus, J. (2015) Characterisation of an unusual telomere motif

(TTTTTTAGGG)$_n$ in the plant *Cestrum elegans* (Solanaceae), a species with a large genome. *Plant Journal*, **82**, 644-654.

Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, **98**, 9748-9753.

Piednoël, M., Aberer, A. J., Schneeweiss, G. M., Macas, J., Novak, P., Gundlach, H., Temsch, E.M. and Renner, S. S. (2012) Next-generation sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. *Molecular Biology and Evolution*, **29**, 3601-3611.

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.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D. S., Jackson, S., Wing, R. A. and Panaud, O. (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research*, **16**, 1262-1269.

Posada, D. (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253-1256.

Pray, L. (2008) Eukaryotic genome complexity. *Nature Education*, **1**:96.

Price, A. L., Jones, N. C. and Pevzner, P. A. (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics*, **21**, i351-i358.

Pritham, E. J., Putliwala, T. and Feschotte, C. (2007) *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene*, **390**, 3-17.

Qu, X. J., Moore, M. J., Li, D. Z. and Yi, T. S. (2019) PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods*, **15**, 50.

Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

Rambaut, A. (2014) FigTree version 1.4.2. http://tree.bio.ed.ac.uk/software/figtree/

Rambaut, A., Drummond, A. (2009) Tracer v1.6. http://tree.bio.ed.ac.uk/software/tracer.

Renny-Byfield, S., Kovarik, A., Kelly, L. J., Macas, J., Novak, P., Chase, M. W., Nichols, R. A., Pancholi, M. R., Grandbastien, M. A. and Leitch, A. R. (2013) Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *Plant Journal*, **74**, 829-839.

Richard, G. F., Kerrest, A. and Dujon, B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Rev*iews, **72**, 686-727.

Roessler, K., Bousios, A., Meca, E. and Gaut, B. S. (2018) Modeling interactions between transposable elements and the plant epigenetic response: a surprising reliance on element retention. *Genome Biology and Evolution*, **10**, 803-815.

Ronaghi, M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Research*, **11**, 3-11.

Roquet, C., Coissac, É., Cruaud, C., Boleda, M., Boyer, F., Alberti, A., Gielly, L., Taberlet, P., Thuiller, W., Van Es, J. and Lavergne, S. (2016) Understanding the evolution of holoparasitic plants: the complete plastid genome of the holoparasite *Cytinus hypocistis* (Cytinaceae). *Annals of Botany*, **118**, 885-896.

Rothberg, J. M., Hinz, W., Rearick, T. M., *Schultz, J., Mileski, W., Davey, M., et al.et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature, **475**, 348-352.

Saha, S., Bridges, S., Magbanua, Z. V. and Peterson, D. G. (2008a) Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Research*, **36**, 2284-2294.

Saha, S., Bridges, S., Magbanua, Z. V. and Peterson, D. G. (2008b) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Tropical Plant Biology*, **1**, 85-96.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**, 5463-5467.

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z. and Bennetzen, J. L. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765-768.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E. and Tabata, S. (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Research*, **6**, 283-290.

Savolainen, V., Fay, M. F., Albach, D. C., Backlund, A., Van der Bank, M., Cameron, K. M., Johnson, S. A., LLedo, M. D., Pintaud, J.C., Powell, M., Sheahan, M. C., Soltis, D.E., Soltis, P. S., Weston, P., Whitten, W. M., Wurdack, K. J. and Chase, M. W. (2000) Phylogeny of the eudicots: a nearly complete familial analysis based on *rbcL* gene sequences. *Kew Bulletin*, **55**, 257-309.

Schatz, M. C., Delcher, A. L. and Salzberg, S. L. (2010) Assembly of large genomes using second-generation sequencing. *Genome Research*, **20**, 1165-1173.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. and Quince, C. (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, **43**, e37-e37.

Schmidt, T. (1999) LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. *Plant Molecular Biology*, **40**, 903-910.

Schnable, P. S., Ware, D., Fulton, *R. S., Stein, J. C., Wei, F., Pasternak, S., et al.et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112-1115.

Schubert, I. and Vu, G. T. (2016) Genome stability and evolution: attempting a holistic view. *Trends in Plant Science*, **21**, 749-757.

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., Siripun, K. C., Winder, C. T., Schilling E. E. and Small, R. L. (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American journal of botany*, **92**, 142-166.

Shaw, J., Lickey, E. B., Schilling, E. E. and Small, R. L. (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany*, **94**, 275-288.

Sheppard, A. E., Ayliffe, M. A., Blatch, L., Day, A., Delaney, S. K., Khairul-Fahmy, N., Li, Y., Madesis, P., Pryor, A.J. and Timmis, J. N. (2008) Transfer of plastid DNA to the nucleus is elevated during male gametogenesis in tobacco. *Plant Physiology*, **148**, 328-336.

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L. and Liu, C. (2019) CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Research*, **47**, W65-W73.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., *et al.* (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO Journal*, **5**, 2043-2049.

Shirasu, K., Schulman, A. H., Lahaye, T. and Schulze-Lefert, P. (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Research*, **10**, 908-915.

Shivakumar, V. S., Appelhans, M. S., Johnson, G., Carlsen, M. and Zimmer, E. A. (2017) Analysis of whole chloroplast genomes from the genera of the Clauseneae, the curry tribe (Rutaceae, *Citrus* family). *Molecular Phylogenetics and Evolution*, **117**, 135-140.

Silva, S. R., Pinheiro, D. G., Meer, E. J., Michael, T. P., Varani, A. M. and Miranda, V. F. (2017) The complete chloroplast genome sequence of the leafy bladderwort, *Utricularia foliosa* L. (Lentibulariaceae). *Conservation Genetics Resources*, **9**, 213-216.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117-1123.

Sleator, R. D. (2013) A beginner's guide to phylogenetics. *Microbial ecology*, **66**, 1-4.

Smit, A.F.A., Hubley, R. and Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

Smit, A.F.A., Hubley, R. RepeatModeler Open-1.0. 2008-2015 <http://www.repeatmasker.org>.

Smith, D. R. (2011) Extending the limited transfer window hypothesis to inter-organelle DNA migration. *Genome Biology and Evolution*, **3**, 743-748.

Sobreira, T.J.; Durham, A.M. and Gruber, A. (2006) TRAP: automated classification, quantification and annotation of tandemly repeated sequences. *Bioinformatics*, **22**, 361-362.

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., Sankoff, D., de Pamphilis, C. W., Wall, P. K. and Soltis, P. S. (2009) Polyploidy and angiosperm diversification. *American Journal of Botany*, **96**, 336-348.

Soltis, D. E., Soltis, P. S., Chase, M. W., Mort, M. E., Albach, D. C., Zanis, M., Savolainen, V., Hahn, W. H., Hoot, S. B., Fay, M. F., Axtell, M., Swensen, S. M., Prince, L. M., Kress, W. J., Nixon, K. C. and Farris, J. S. (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbcL,* and *atpB* sequences. *Botanical Journal of the Linnean Society*, **133**, 381-461.

Sonnhammer, E. 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.

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

Straub, S. C., Cronn, R. C., Edwards, C., Fishbein, M. and Liston, A. (2013) Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae) *Genome Biology and Evolution*, **5**, 1872-1885.

Straub, S. C., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C. and Liston, A. (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349-364.

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**, 1435-1444.

Sung, W. K. (2017) Algorithms for next-generation sequencing. *Chapman and Hall/CRC*.

Sykorova, E., Lim, K. Y., Chase, M. W., Knapp, S., Leitch, I. J., Leitch, A. R. and Fajkus, J. (2003) The absence of *Arabidopsis*-type telomeres in *Cestrum* and closely related genera *Vestia* and *Sessea* (Solanaceae): first evidence from eudicots. *Plant Journal*, **34**, 283-291.

Szak, S. T., Pickeral, O. K., Makalowski, W., Boguski, M. S., Landsman, D. and Boeke, J. D. (2002) Molecular archeology of L1 insertions in the human genome. *Genome Biology*, **3**, research0052-1.

Tek, A. L., Song, J., Macas, J. and Jiang, J. (2005) Sobo, a recently amplified satellite repeat of potato and its implications for the origin of tandemly repeated sequences. *Genetics*, **170**, 1231-1238.

Tetreault, H. M. and Ungerer, M. C. (2016) Long terminal repeat retrotransposon content in eight diploid sunflower species inferred from next-generation sequence data. *G3: Genes, Genomes, Genetics*, **6**, 2299-2308.

Thomas Jr, C. A. (1971) The genetic organization of chromosomes. *Annual Review of Genetics*, **5**, 237-256.

Thompson, J. D., Gibson, T. J. and Higgins, D. G. (2003) Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, **1**, 2-3.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R. and Greiner, S. (2017) GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, **45**, W6-W11.

Timmis, J. N., Ayliffe, M. A., Huang, C. Y. and Martin, W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, **5**, 123-135.

Torres, G. A., Gong, Z., Iovene, M., Hirsch, C. D., Buell, C. R., Bryan, G. J., Novák, P., Macas, J. and Jiang, J. (2011) Organization and evolution of subtelomeric satellite repeats in the potato genome. *G3: Genes, Genomes, Genetics*, **1**, 85-92.

Tran, T. D., Cao, H. X., Jovtchev, G., Neumann, P., Novák, P., Fojtová, M., Vu, G. T. H., Macas, J., Fajkus, J., Schubert, I. and Fuchs, J. (2015) Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus *Genlisea*. *Plant Journal*, **84**, 1087-1099.

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S. and Turner, S. W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, **38**, e159-e159.

Tu, Z. (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proceedings of the National Academy of Sciences*, **98**, 1699-1704.

Tu, Z., Li, S. and Mao, C. (2004) The changing tails of a novel short interspersed element in Aedes aegypti: genomic evidence for slippage retrotransposition and the relationship between 3' tandem repeats and the poly (dA) tail. *Genetics*, **168**, 2037-2047.

Tymms, M. J. and Schweiger, H. G. (1985) Tandemly repeated nonribosomal DNA sequences in the chloroplast genome of an *Acetabularia mediterranea* strain. *Proceedings of the National Academy of Sciences*, **82**, 1706-1710.

Uchida, W., Matsunaga, S., Sugiyama, R. and Kawano, S. (2002) Interstitial telomere-like repeats in the *Arabidopsis thaliana* genome. *Genes & Genetic Systems*, **77**, 63-67.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. and Leunissen, J. A. M. (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*, **35**, W71-W74.

Van Straalen, N. M. and Roelofs, D. (2011) Introduction to Ecological Genomics. *Oxford University Press, Oxford, UK*.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., *et al.* (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PloS ONE*, **2**, e1326.

Vitte, C. and Panaud, O. (2003) Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Molecular Biology and Evolution*, **20**, 528-540.

Vu, G. T., Schmutzer, T., Bull, F., Cao, H. X., Fuchs, J., Tran, T. D., Jovtchev, G., Pistrick, K., Stein, N., Pecinka, A., Neumann, P., Novak, P., Macas, J., Dear, P. H., Blattner, F. R., Scholz, U. and Schubert, I. (2015) Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. *Plant Genome*, **8**.

Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T. and Sugiura, M. (1994) Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proceedings of the National Academy of Sciences*, **91**, 9794-9798.

Wang, W., Schalamun, M., Morales-Suarez, A., Kainer, D., Schwessinger, B. and Lanfear, R. (2018) Assembly of chloroplast genomes with long-and short-read data: a comparison of approaches using *Eucalyptus pauciflora* as a test case. *BMC Genomics*, **19**, 977.

Watson, J. D. and Crick, F. H. (1953) Molecular structure of nucleic acids. *Nature*, **171**, 737-738.

Wei, L., Xiao, M., An, Z., Ma, B., Mason, A. S., Qian, W., Li, J. and Fu, D. (2013) New insights into nested long terminal repeat retrotransposons in Brassica species. *Molecular Plant*, **6**, 470-482.

Weilguny, L. and Kofler, R. (2019) DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition. *Molecular Ecology Resources*, **19**, 1346-1354.

Weiss-Schneeweiss, H., Leitch, A. R., McCann, J., Jang, T. S. and Macas, J. (2015) Employing next generation sequencing to explore the repeat landscape of the plant genome. *Next Generation Sequencing in Plant Systematics. Regnum Vegetabile*, **157**, 155-179.

Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A. and Liston, A. (2014) Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, **2**, 1400042.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. and Schulman, A. H. (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, **8**, 973-982.

Witte, C. P., Le, Q. H., Bureau, T. and Kumar, A. (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences*, **98**, 13778-13783.

Wojciechowski, M. F., Lavin, M. and Sanderson, M. J. (2004) A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany*, **91**, 1846-1862.

Wright, D. A. and Voytas, D. F. (1998) Potential retroviruses in plants: Tat*1* is related to a group of *Arabidopsis thaliana* Ty3/*gypsy* retrotransposons that encode envelope-like proteins. *Genetics*, **149**, 703-715.

Wyman, S. K., Jansen, R. K. and Boore, J. L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252-3255.

Xia, X., Selvaraj, G. and Bertrand, H. (1993) Structure and evolution of a highly repetitive DNA sequence from *Brassica napus*. *Plant Molecular Biology*, **21**, 213-224.

Xie, H., Jiao, J., Fan, X., Zhang, Y., Jiang, J., Sun, H. and Liu, C. (2017) The complete chloroplast genome sequence of Chinese wild grape *Vitis amurensis* (Vitaceae*: Vitis* L.). *Conservation Genetics Resources*, **9**, 43-46.

Xiong, W., He, L., Lai, J., Dooner, H. K. and Du, C. (2014) HelitronScanner uncovers a large overlooked cache of *Helitron* transposons in many plant genomes. *Proceedings of the National Academy of Sciences*, **111**, 10263-10268.

Xu, J. H., Liu, Q., Hu, W., Wang, T., Xue, Q. and Messing, J. (2015) Dynamics of chloroplast genomes in green plants. *Genomics*, **106**, 221-231.

Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**, W265-W268.

Xuan, Y. H., Zhang, J., Peterson, T. and Han, C. D. (2012) *Ac/Ds*-induced chromosomal rearrangements in rice genomes. *Mobile Genetic Elements*, **2**, 67-71.

Zerbino, D. R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821-829.

Zhang, J. and Peterson, T. (2004) Transposition of reversed Ac element ends generates chromosome rearrangements in maize. *Genetics*, **167**, 1929-1937.

Zhang, N., Wen, J. and Zimmer, E. A. (2015) Congruent deep relationships in the grape family (Vitaceae) based on sequences of chloroplast genomes and mitochondrial genes via genome skimming. *PloS ONE*, **10**, e0144701.

Zuccolo, A., Sebastian, A., Talag, J., Yu, Y., Kim, H., Collura, K., Kudrna, D. and Wing, R. A. (2007) Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. *BMC Evolutionary Biology*, **7**, 152.

Zytnicki, M., Akhunov, E. and Quesneville, H. (2014) Tedna: a transposable element de novo assembler. *Bioinformatics*, **30**, 2656-2658.