MASARYK UNIVERSITY

FACULTY OF SCIENCE

Genomics of chromosomal dysploidy in plants

Ph.D. Thesis

Yile Huang

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

National Centre for Biomolecular Research
Mendel Centre for Plant Genomics and Proteomics

Brno 2025

Bibliographic Entry

Author:	Yile Huang
	Faculty of Science, Masaryk University National Centre for Biomolecular Research Central European Institute of Technology

Title of Thesis:	Genomics of chromosomal dysploidy in plants
Degree programme:	Genomics and Proteomics
Field of Study:	Genomics and Proteomics
Supervisor:	Prof. Mgr. Martin Lysák, Ph.D., DSc.
Academic Year:	2025/2026
Number of Pages:	160
Keywords:	whole-genome duplication, subgenomes, diploidization, descending dysploidy, chromosome number variation, karyotype evolution, Brassicaceae, Heliophileae, Biscutelleae

Bibliografický záznam

Autor:	Yile Huang Přírodovědecká fakulta, Masarykova univerzita Národní centrum pro výzkum biomolekul Středoevropský technologický institut

Název práce:	Genomika chromozomální dysploidie u rostlin
Studijní program:	Genomika a proteomika
Studijní obor:	Genomika a proteomika
Vedoucí práce:	Prof. Mgr. Martin Lysák, Ph.D., DSc.
Akademický rok:	2025/2026
Počet stran	160
Klíčová slova:	celogenomové duplikace, subgenomy, diploidizace, descendentní dysploidie, evoluce karyotypu, Brassicaceae, Heliophileae, Biscutelleae

Abstract

Genomic redundancy caused by whole-genome duplication (WGD) creates opportunities for errors in double-strand break misrepair, which can result in chromosomal rearrangements (CRs) and a reduction in chromosome number, a phenomenon known as descending dysploidy. While recurrent cycles of polyploidization and post-polyploid diploidization (PPD) are common in flowering plants, the evolutionary pathways and consequences of descending dysploidy remain poorly understood.

The chromosome-scale genome assembly of *Heliophila variabilis* (~334 Mb, 2*n* = 22) represents the first sequenced octoploid genome in the Brassicaceae. This genome shows evidence of an allo-octoploid origin ~12 million years ago (Mya). Rediploidization of the ancestral genome was marked by extensive reorganization of parental subgenomes (39 CRs), a 2.7-fold reduction in chromosome number (*n* = ~30 → *n* = 11), and speciation events in the genus *Heliophila*. These findings highlight the roles of WGD and PPD to the ecomorphological diversification and adaptation in the Cape flora region.

The eight *Biscutella* genomes (Brassicaceae; 0.6 – 1.1 Gb; *n* = 6, 8 and 9) were traced to a shared allotetraploid ancestor (*n* = 14) ~11 to 13 Mya, followed by independent descending dysploid trajectories that produced at least two evolutionary clades: early-diverging species (*n* = 6/8) and late-diverging species (*n* = 9). Both clades showed comparable degrees of subgenome fractionation; however, early-diverging species demonstrated an accelerated removal of LTR retrotransposons, frequent transitions between A and B chromatin compartments, and considerable variability in the sizes of topologically associated domains (TADs). Moreover, fourteen chromosomal breakage hotspots rich in repeats—often overlapping with TAD boundaries—highlight the influence of chromatin topology on genome restructuring. Collectively, these findings show that despite a shared ancestry, diploidization proceeds along clade-specific trajectories that differ both in rates and extent.

Collectively, these findings provide new insights into how WGD, PPD, and descending dysploidy have shaped chromosome numbers, genome architecture, and evolutionary innovation in seed plants. They highlight that while polyploidy is a near-universal

feature of plant genomes, the subsequent diploidization and chromosomal remodeling processes can follow lineage-dependent trajectories with distinct mechanisms, rates and extent. This contributes to the remarkable diversity and adaptability of seed plants.

Abstrakt

Genomová redundance způsobená celogenomovou duplikací (WGD) vytváří podmínky pro chybné opravy dvouřetězcových zlomů DNA, což může vést k chromozomovým přestavbám (CRs) a redukci počtu chromozomů, označované jako descendentní dysploidie. Ačkoli jsou opakované cykly polyploidizace a postpolyploidní diploidizace (PPD) u krytosemenných rostlin časté, evoluční mechanismy a důsledky descendentní dysploidie zůstávají nedostatečně objasněny.

Genomová sestava *Heliophila variabilis* v chromozomovém měřítku (~334 Mb, 2*n* = 22) představuje první sekvenovaný oktoploidní genom v čeledi brukvovitých. Analýzy ukazují na allo-oktoploidní původ tohoto genomu před přibližně 12 miliony let. Rediploidizace vedla k rozsáhlé reorganizaci rodičovských subgenomů (39 CRs), 2,7násobnému snížení počtu chromozomů (z ~30 na 11) a k speciacím v rámci rodu *Heliophila*. Tyto výsledky zdůrazňují význam WGD a PPD pro ekomorfologickou diverzifikaci a adaptaci flóry Kapské oblasti.

Osm genomů rodu *Biscutella* (Brassicaceae; 0,6–1,1 Gb; *n* = 6, 8 a 9) bylo vystopováno k společnému allotetraploidnímu předku (*n* = 14) přibližně před 11–13 miliony let, po čemž následovaly nezávislé trajektorie sestupné dysploidie, které vedly ke vzniku nejméně dvou evolučních kladů: raně divergujících druhů (*n* = 6/8) a později divergujících druhů (*n* = 9). Oba klady vykazovaly srovnatelný stupeň frakcionace subgenomů, nicméně časně divergované druhy měly zrychlenou eliminaci LTR retrotranspozonů, častější přechody mezi A a B chromatinovými kompartmenty a výraznější variabilitu ve velikosti topologicky asociovaných domén (TADs). Dále čtrnáct hotspotů chromozomálních zlomů bohatých na repetitivní sekvence – často překrývajících se s hranicemi TAD – zdůrazňuje vliv chromatinové topologie na přestavbu genomu. Souhrnně tato zjištění ukazují, že navzdory společnému původu probíhá diploidizace podél kladově specifických trajektorií, které se liší jak rychlostí, tak rozsahem.

Tyto výsledky poskytují nové poznatky o tom, jak WGD, PPD a descendentní dysploidie formují počty chromozomů, strukturu genomu a evoluční inovace u semenných rostlin. Zdůrazňují, že polyploidie je sice téměř univerzálním rysem rostlinných genomů, avšak

následná diploidizace a remodelace chromozomů probíhají podle liniově specifických trajektorií, které se liší mechanismy, rychlostí i rozsahem. To přispívá k mimořádné diverzitě a adaptabilitě semenných rostlin.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. Martin Lysák, for his invaluable guidance, patience and support throughout my PhD journey. His knowledge and experience were instrumental in the completion of this thesis.

I am also grateful to all my colleagues in the Lysak Research Group for fostering a collaborative and supportive working environment. In particular, I would like to thank Dr. Terezie Mandáková, Dr. Xinyi Guo, and Milan Pouch for their generous help.

Special appreciation goes to Prof. Christian Parisod (University of Fribourg), Prof. Feng Cheng (Chinese Academy of Agricultural Sciences), and Dr. Alexandros Bousios (University of Sussex), for their valuable insights and contribution to our collaborative projects.

Finally, I am deeply grateful to my family and friends for their constant love, support, and encouragement throughout this journey.

Plant Sciences core facility of CEITEC Masaryk University is acknowledged for the technical support. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. This work was supported by the Czech Science Foundation (grants nos. 21-07748L and 19-07487S), the project TowArdsNextGENeration Crops (no. 17 CZ.02.01.01/00/22_008/0004581), the Masaryk University Grant Agency (MUNI/R/1268/2022), and the National Geographic Society (grant no. 9345-13).

Declaration

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

Date: 04/09/2025 Signed: 

Author's publications

Publication 1 (APPENDIX 1):

Huang, Y., Guo, X., Zhang, K., Mandáková, T., Cheng, F., & Lysak, M. A. (2023). The meso-octoploid *Heliophila variabilis* genome sheds a new light on the impact of polyploidization and diploidization on the diversity of the Cape flora. *The Plant Journal*, 116(2), 446-466. https://doi.org/10.1111/tpj.16383

Publication 2 (APPENDIX 2):

Huang, Y., Poretti, M., Mandáková, T., Pouch, M., Guo, X., Perez-Roman, E., Crespo, M. B., Grob, S., Bousios, A., Parisod C., & Lysak, M. A. (2025). Post-polyploid chromosomal diploidization in plants is affected by clade divergence and constrained by shared genomic features. *In revision, Nature Communications,* https://doi.org/10.21203/rs.3.rs-6440714/v1

For both publications, H. Y. performed bioinformatics analyses of the sequence data, including genome assembly, gene and repeatome annotation, subgenome phasing, phylogenetic analysis, inference of whole-genome duplications and ancestral genomes, reconstruction of chromosomal rearrangements, and 3D genome analysis. H. Y. interpreted the results and wrote the manuscripts with contributions of other authors.

Abbreviations

ACBK	Ancestral Karyotype of Camelinodae and Brassicodae
ACK	Ancestral Crucifer Karyotype
AK	Ancestral Karyotype
ancPCK	ancestral Proto-Calepineae Karyotype
BrassiToL	Brassicaceae Tree of Life
CCP	Comparative Chromosome Painting
CFR	Cape Floristic Region
c-NHEJ	canonical Non-Homologous End-Joining
CR	Chromosomal Rearrangement
CTW	Context-Tree Weighting
DSB	Double-Strand Break
EET	End-to-End Translocation
EPL	Expected Ploidy Level
FISH	Fluorescence *In Situ* Hybridization
FoSTeS	Fork Stalling and Template Switching
GB	Genomic Block
GO	Gene Ontology
Ipa	Paracentric Inversion
Ipe	Pericentric Inversion
ITS	Internal Transcribed Spacer
*K*a	Non-synonymous substitution rate
*K*s	Synonymous substitution rate
LCR	Low-Copy Repeat
LF	Less Fractionated
LTR-RT	Long Terminal Repeat Retrotransposon
MF	More Fractionated
ML	Maximum Likelihood
MMBIR	Microhomology-Mediated Break-Induced Replication
MMEJ	Microhomology-Mediated End-Joining
Mya	Million years ago
NAHR	Non-Allelic Homologous Recombination

NCI	Nested Chromosome Insertion
NGS	Next-Generation Sequencing
NHEJ	Non-Homologous End-Joining
ONT	Oxford Nanopore Technologies
PCK	Proto-Calepineae Karyotype
PPD	Post-Polyploid Diploidization
rDNA	Ribosomal DNA
RT	Reciprocal Translocation
satDNA	satellite DNA
TAD	Topologically Associated Domain
TE	Transposable Element
tPCK	Translocation Proto-Calepineae Karyotype
TPM	Transcripts Per Million
unbalanced-RT	unbalanced Reciprocal Translocation
WGD	Whole-genome Duplication
WGT	Whole-genome Triplication
3D	Three-dimensional

Table of Contents

1. Introduction 16

1.1 Polyploidization and post-polyploid diploidization in plants 16

1.1.1 Polyploidization 16

1.1.2 Post-polyploid diploidization 19

1.2 Chromosomal rearrangements during post-polyploid diploidization 23

1.2.1 Mechanisms of chromosomal rearrangements 23

1.2.2 Dysploid chromosomal rearrangements 27

1.2.3 Impact of dysploid changes on diversification of the Brassicaceae family 32

1.3 Origins and mechanisms of biased subgenome fractionation 39

1.4 Taxa of interest 40

1.4.1 Heliophileae 40

1.4.2 Biscutelleae 42

2. Aims 47

3. Materials and Methods 48

3.1 Plant material, library preparation and sequencing 48

3.2 Genome size measurement by flow cytometry and *K*-mer frequency 49

3.3 Genome assembly, scaffolding, and quality assessment 49

3.4 Gene prediction and functional annotation 50

3.5 Identification of syntenic genes and fragments 51

3.6 Repetitive element annotation 52

3.7 FISH experiment and comparative chromosome painting 52

3.8 Phylogenetic inference and divergence time calibration 53

3.9 Subgenome phasing and validation 54

3.10 *K*a/*K*s analysis 54

3.11 Reconstruction of the ancestral karyotypes 54

3.12 Gene family classification and functional enrichment 55

3.13 3D genome analysis in *Biscutella* genomes....56

4. Results....58

4.1 The meso-octoploid *Heliophila variabilis* genome sheds a new light on the impact of polyploidization and diploidization on the diversity of the Cape flora....58

4.2 Post-polyploid chromosomal diploidization in plants is affected by clade divergence and constrained by shared genomic features....65

5. Discussion....76

5.1 Extensive descending dysploidy shaped the meso-octoploid *Heliophila* genome over millions of years....76

5.2 Chromosomal diploidization in *Biscutella* proceeded through independent descending dysploidy, LTR deletion and chromatin reorganization....77

6. Conclusions and Final Remarks....79

7. References....81

APPENDIX 1....95

APPENDIX 2....116

Curriculum vitae....157

List of Figures

Figure 1 Genomic features of different polyploid forms and post-polyploid genome diploidization. 22
Figure 2 Molecular mechanisms underlying chromosomal rearrangements (CRs). 27
Figure 3 Proposed dysploid chromosomal rearrangement types in angiosperms, adopted from Mandáková and Lysak (2018). 31
Figure 4 Time-calibrated genus-level Brassicaceae Tree of Life (BrassiToL) inferred from a maximum likelihood analysis of a 297 nuclear gene supermatrix, adopted from Hendriks et al. (2023) 35
Figure 5 Two proposed karyotype evolution pathways in the Brassicaceae. 37
Figure 6 Four representative trajectories of karyotype (genome) evolution in Brassicaceae. 38
Figure 7 Schematic relationships among 15 *Heliophila* species and *Chamira circaeoides* based on the ITS phylogeny (Dogan et al., 2021) 42
Figure 8 Proposed origin and evolution of the Biscutelleae diploid–polyploid genome complex by Guo et al. (2021). 45
Figure 9 Morphological characteristics of *Heliophila variabilis* and its genome structure. 60
Figure 10 Evolution of four subgenomes in the octoploid *H. variabilis* genome. 61
Figure 11 Four reconstructed ancestral subgenomes of *H. variabilis* and their structure within the extant *H. variabilis* genome. 62
Figure 12 TE element content and distribution of TE insertion times of *H. variabilis* 63
Figure 13 Heat map of 21-mer enrichment shared by four subgenomes of *H. variabilis* in 22 GBs (A to X). 64
Figure 14 Genome size, repeat composition, chromosome collinearity, and subgenome phylogenomics in eight *Biscutella* species. 70
Figure 15 Karyotype evolution of *Bicutella* genomes. 71
Figure 16 Evolution of LTR retrotransposons and 3D chromatin organization. 72
Figure 17 Position and characteristics of chromosome breakage hotspots. 74

1. Introduction

1.1 Polyploidization and post-polyploid diploidization in plants

Polyploidy, defined as the presence of more than two complete sets of chromosomes in an organism, results from polyploidization—also known as whole-genome duplication (WGD). Polyploidization occurs frequently across land plants, particularly in angiosperms, and provides raw material for evolutionary innovations that underpin diversification of plants (Cui et al., 2006; Wood et al., 2009; Jiao et al., 2011). Using probabilistic models, Carta et al. (2020) inferred that the ancestral haploid chromosome number of angiosperms was $n = 7$, with an estimated ancestral 1C genome size of 1,692 Mb. Theoretically, successive polyploidization events should result in exponential increases in both chromosome number and genome size. However, empirical observations deviate strikingly from this expectation: extant angiosperms exhibit a wide range of chromosome numbers, from $2n = 4$ in some species (e.g., *Colpodium biebersteinianum* in the Poaceae and *Rhynchospora tenuis* in the Cyperaceae) to over 600 in others (e.g., *Echeveria bakery* in the Crassulaceae) (Sokolovskaya and Probatova, 1977; Vanzela et al., 1996), and genome sizes spanning over 2,500-fold—from only ~63 Mb in *Genlisea aurea* (Lentibulariaceae) to nearly 150 Gb in *Paris japonica* (Melanthiaceae) (Greilhuber et al., 2006; Pellicer and Leitch, 2020). The extreme variation indicates that polyploidization alone cannot fully explain the remarkable diversification observed in angiosperms. Instead, newly formed polyploids typically undergo a gradual and multifaceted process termed post-polyploid diploidization (PPD), which encompasses large-scale chromosomal rearrangements (CRs), genome reorganization, and various small-scale modifications such as point mutations, insertions and deletions (indels), as well as transposable element (TE) activation and epigenetic reprogramming. These genomic changes collectively contribute to the reestablishment of a functionally diploid-like genome. Among them, CRs frequently result in reduction of chromosome number, i.e., descending dysploidy, a process that can promote reproductive isolation and thereby lead to speciation and cladogenesis.

1.1.1 Polyploidization

Polyploidization is considered to be a trigger for speciation and biodiversity in plants (e.g., Van de Peer et al., 2021; Heslop-Harrison et al., 2023). To date, more than 240

putative ancient WGDs have been identified in the green plants (Li and Barker, 2020). One ancient WGD event, known as the epsilon (ε) WGD, is hypothesized to have occurred ~300 million years ago (Mya) during the Carboniferous period, preceding the divergence of angiosperms from other seed plants, subsequently enabling the rapid diversification of angiosperms into more than 350,000 species (Jiao et al., 2011; AG Project, 2013). Despite the shared ε-WGD, the number and timing of subsequent polyploidizations differ among lineages. Some species have undergone only one or two WGDs, e.g., *Amborella trichopoda* and *Aristolochia fimbriata* show evidence of only the ε event, grape (*Vitis vinifera*), sugar beet (*Beta vulgaris*), coffee (*Coffea canephora*), rose (*Rosa chinensis*), and cacao (*Theobroma cacao*) experienced both ε and the later gamma (γ) whole genome triplication (WGT) (Jaillon et al., 2007; Argout et al., 2011; AG Project, 2013; Denoeud et al., 2014; Dohm et al., 2014; Hibrand Saint-Oyant et al., 2018; Cui et al., 2022). However, other species, such as *Arabidopsis thaliana*, rapeseed (*Brassica napus*), sugarcane (*Saccharum officinarum*), and maize (*Zea mays*), have experienced multiple rounds of WGDs (AG Initiative, 2000; Haberer et al., 2005; Song et al., 2020; Healey et al., 2024).

Polyploids are generally classified as autopolyploids and allopolyploids based on their genomic origin and chromosomal composition (Soltis et al., 2007; Parisod et al., 2010; Lv et al., 2024). Autopolyploids typically arise within a single species through doubling of structurally similar, homologous genomes (e.g., AAAA, where identical chromosome sets are derived from genome A), whereas allopolyploids originate from interspecific hybridization, leading to the coexistence of divergent but homeologous chromosome sets (e.g., AABB, where A and B represent distinct genomes) (**Figure 1a**). A principal distinction between these two types lies in their chromosomal pairing behavior during meiosis. Allopolyploids generally exhibit predominantly bivalent pairing between homologous chromosomes, resulting in disomic inheritance, where two alleles segregate per locus. Autopolyploids are expected to have multisomic inheritance, where more than two alleles per locus can segregate due to multivalent pairing (Jackson R. C. and Jackson J. W., 1996; Hauber et al., 1999; Landergott et al., 2006; Li et al., 2021). It is important to point out that even though strictly bivalent pairing can occur in some autopolyploids (e.g., Qu et al., 1998; Stift et al., 2008), but random segregation of homologous chromosomes during meiosis may still result in multisomic inheritance. Therefore, the

occurrence of multisomic inheritance is a unique feature that can define autopolyploids (Parisod et al., 2010; Tayalé and Parisod, 2013).

Autopolyploids and allopolyploids represent opposite ends of a spectrum in the genetic divergence of two or more polyploid genomes. When parental genomes are only partially divergent, intermediate polyploid forms may exhibit both homologous and homeologous chromosome pairing, thereby blurring the classical distinction between auto- and allopolyploidy. For example, segmental allopolyploids harbor partially homologous genomes (e.g., $A_1A_1A_2A_2$, where A_1 and A_2 represent partially divergent genomes). Another example is autoallopolyploids, which contain both autopolyploid and allopolyploid segments or chromosome sets (e.g., AAAABB or AABBBB) (**Figure 1a**; Stebbins, 1947; Mason and Wendel, 2020). The genetic behavior of such intermediate polyploids is described as mixosomic (Soltis et al., 2016), reflecting the presence of both disomic and multisomic segregation within the same organism.

Compared to diploids, polyploids often possess larger somatic cells and increased vacuolar content, which contribute to enhanced vegetative growth (Doyle and Coate, 2019). In allopolyploids, the merging of divergent genomes can give rise to heterosis (Baranwal et al., 2012), enhancing overall vigor and environmental adaptability. Their high genetic plasticity also imparts greater tolerance to both biotic and abiotic stressors (Rapp et al., 2009; Cheng et al., 2018). On the other hand, autopolyploids may suffer from meiotic irregularities due to multivalent formation, resulting in unbalanced gametes and inbreeding depression (Abel and Becker, 2007). These potential disadvantages partly explain why allopolyploidy is often considered more prevalent and evolutionarily favored in plants (Rapp et al., 2009; Ainouche and Wendel, 2014; Behling et al., 2020). However, based on the incidence of polyploid taxa within genera, Barker et al. (2016) reported that 13% of the surveyed 4,003 vascular plant species across 47 genera were autopolyploids and 11% were allopolyploids. Building on this dataset, Li et al. (2021) conducted a detailed examination of meiotic chromosome behavior in 208 polyploid species representing 40 genera. Among the 208 species, 118 had previously been classified by Barker et al. as allopolyploids and 90 as autopolyploids. Li et al. found that 92 species exhibited strictly bivalent pairing, whereas 116 displayed mixed or multivalent pairing. These findings collectively highlight the complexity and diversity of

plant polyploid forms and suggest that the number of polyploids with mixed or multivalent chromosome pairing may be largely underestimated.

Based on the age of a WGD and the extent and speed of diploidization, polyploids can also be broadly categorized as neopolyploids (e.g., *Arabidopsis suecica*), mesopolyploids (e.g., *Brassica rapa*) and paleopolyploids (e.g., *V. vinifera*) (Mandáková et al., 2010; Mandáková and Lysak, 2018; Zhang et al., 2019). Neopolyploids arise recently and typically retain an additive number of parental chromosomes and genome size being a sum of parental C-values. Their subgenomes remain largely intact, structurally distinguishable, and often traceable to extant progenitor species. Through progressive diploidization over evolutionary timescales, neopolyploids turn into mesopolyploids and eventually into paleopolyploids. Mesopolyploid genomes generally exhibit diploid-like chromosome numbers and meiotic behavior, while the reshuffled parental subgenomes can still be identified by comparative cytogenetic and bioinformatic approaches. Paleopolyploids represent the most diploidized polyploids, with highly restructured subgenomes and few detectable homeologous regions (**Figure 1b**; Mandáková and Lysak, 2018). This classification of polyploids is inherently relative, typically applied within a specific clade or lineage. For example, functionally diploid species of the *Glycine* genus (2*n* = 38, 40; e.g., *G. max* and *G. soja*) have experienced two rounds of WGD: a shared paleo-polyploidization event ~65 Mya as all papilionoid legumes, and a *Glycine*-specific mesopolyploid WGD within the last ~13 million years (Doyle and Egan, 2010; Koenen et al., 2020). Remarkably, within the past 350,000 years, several *Glycine* species have undergone independent allopolyploidization events, giving rise to a number of neo-allopolyploid species (2*n* = 78, 80; e.g., *G. tomentella* and *G. dolichocarpa*) (Zhuang et al., 2022).

1.1.2 Post-polyploid diploidization

Despite the frequent occurrence of WGD events and the potential advantages conferred by polyploids, mature neopolyploids remain far less prevalent than paleo- and mesopolyploids (Stebbins, 1950; Wagner, 1970). Two factors are thought to underlie this pattern: (i) pervasive polyploid extinction, supported by evidence that newly formed polyploids had lower diversification rates and higher extinction rates compared to their diploid relatives (Mayrose et al., 2011, 2015); and (ii) frequent reversion to

diploidy and disomic inheritance following an initial polyploid phase (Wendel, 2015; Soltis et al., 2016; Li et al., 2021).

Diploidization encompasses multiple mechanisms that can be broadly categorized into two primary processes: cytological diploidization (see Section 1.2) and genic diploidization/fractionation (see Section 1.3). Cytological diploidization frequently entails substantial chromosomal changes—such as fissions and fusions—that progressively restructure the genome and restore diploid-like meiotic pairing behavior (Ma and Gustafson, 2005). Genic diploidization results in the loss of gene duplicates generated by WGD, with only a subset retained as paralogs over time (Freeling et al., 2012; Garsmeur et al., 2014). Both processes may occur soon after WGD (sometimes within just a few dozen generations; e.g., Xiong et al., 2011; Buggs et al., 2012), and can occur independently (e.g., Ma et al., 2024) as well as at different rates (e.g., Mandáková et al., 2017b), yielding a diversity of genomes with different patterns of diploidization following polyploidy across lineages (Otto and Whitton, 2000; Wolfe, 2001).

Chromosome number itself imposes evolutionary constraints on genome organization and stability. Extremely low counts can limit meiotic segregation and reduce genetic variability, whereas extremely high counts increase the risk of mis-segregation during nuclear division (Schubert and Vu, 2016). Therefore, genomes need a balance between the stability required for faithful inheritance and the variability necessary for adaptive potential. Among the two possible directions of chromosome number change, descending dysploidy is generally more common, as it usually proceeds through chromosomal fusions accompanied by centromere loss or inactivation (see Section 1.2.2). By contrast, ascending dysploidy often requires centromere fission or the formation of new centromeres, making it structurally and energetically less feasible. As anticipated, ascending dysploidy appears to be rare and has only been documented in a few cases. One such example is *Capsicum rhomboideum* (*n* = 13; Solanaceae), which possesses one more chromosome than its presumed ancestral karyotype (*n* = 12). This increase is thought to result from the fission of ancestral chromosome A12 into two distinct chromosomes, thus generating a higher chromosome number compared to other *Capsicum* species (Zhang et al., 2025).

The rate of PPD varies considerably—not only across different land plant lineages, but also among the descendants of a single WGD event. Some allopolyploids undergo chromosomal restructuring within only a few generations. For example, the synthetic allotetraploid *Tragopogon miscellus* (Asteraceae) exhibits substantial chromosomal variation after approximately 40 generations (Xiong et al., 2011). In contrast, others, such as allopolyploid cotton (*Gossypium*), which arose ~1–2 Mya, have retained relatively stable genomes over millions of years (Chen et al., 2020). Mandáková et al. (2017b) further demonstrated that even lineages descending from the same ancestral allopolyploid can exhibit markedly different diploidization trajectories. Within the tribe Microlepidieae (Brassicaceae), which originated via intertribal hybridization ~11 Mya, three major subclades display distinct degrees of descending dysploidy: the crown-group genera possess highly reshuffled genomes and low chromosome numbers ($n = 15 \rightarrow n = 4 – 7$), *Pachycladon* experienced slower diploidization ($n = 15 \rightarrow n = 10$) and some *Arabidella* species have the least diploidized genomes ($n = 15 \rightarrow n = 12$). These findings lead to fundamental questions: why does the rate of diploidization differ among polyploid species/lineages? and what are the primary factors driving these differences? Answering these questions remains highly challenging due to the deep evolutionary timescales over which diploidization occurs, the difficulty of estimating diploidization rates, and the extensive genomic data required for robust comparative analyses.

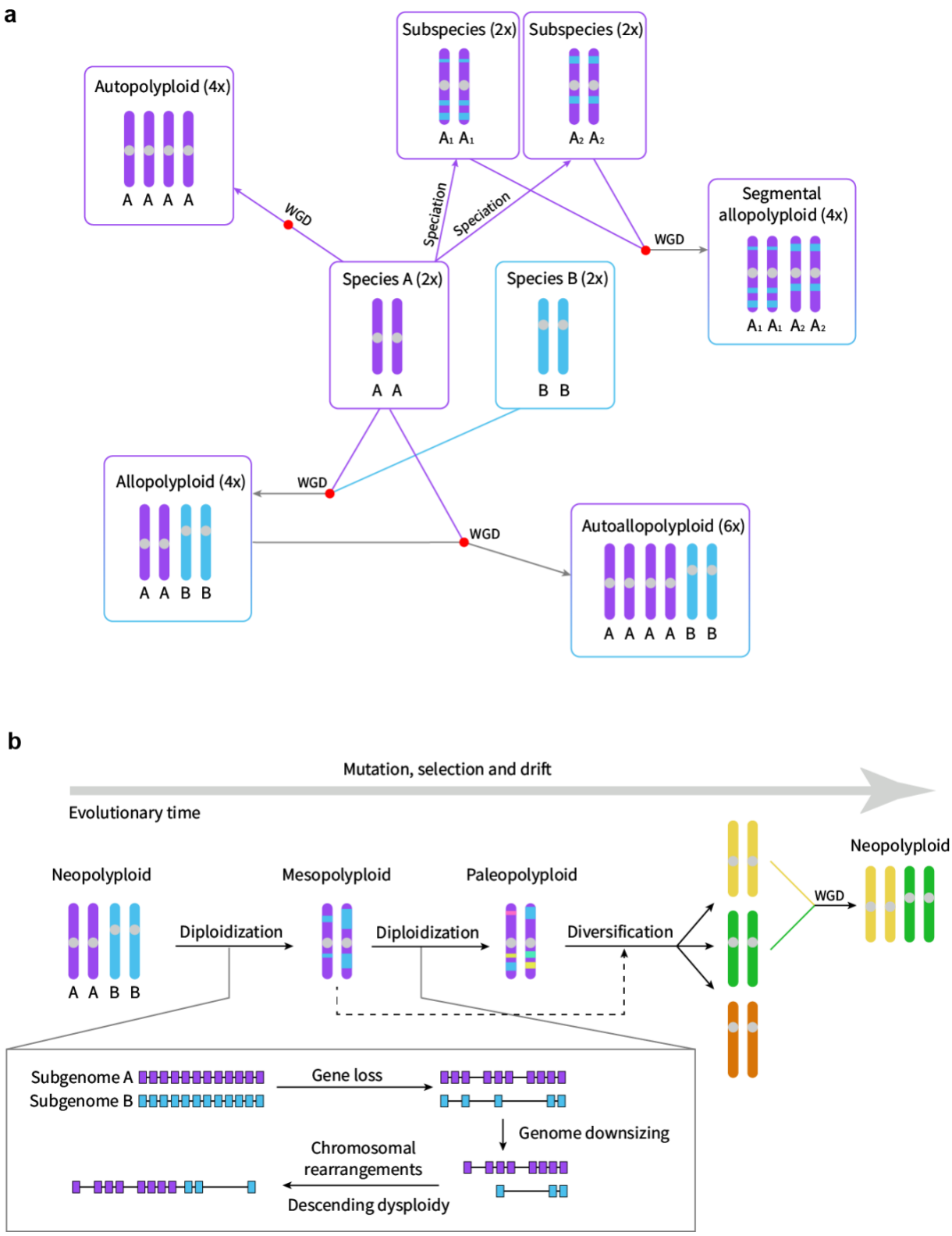


Figure 1 Genomic features of different polyploid forms and post-polyploid genome diploidization.

(**a**) Schematic representations of four typical polyploid forms. An autopolyploid (AAAA) arises from genome doubling within a single diploid species, resulting in multiple copies of

homologous chromosome sets. A segmental allopolyploid ($A_1A_1A_2A_2$) results from polyploidization between closely related genomes (A_1A_1 and A_2A_2) that are only partially genetically differentiated. An allopolyploid (AABB) originates from hybridization between two distinct diploid genomes (AA and BB). A subsequent hybridization involving one of the parental genomes (AA) can result in an autoallopolyploid (AAAABB), containing both homologous and homeologous chromosome sets.

(**b**) A polyploid can be classified as neopolyploid, mesopolyploid or paleopolyploid depending on the time elapsed since the WGD and the degree of PPD. Neopolyploid represents the most recently formed polyploid that retains most of the parental genome structures and chromosome numbers. Over time, as diploidization proceeds, the neopolyploid genome transits into mesopolyploid and eventually into paleopolyploid. Diploidization is a progressive and complex process that involves gene loss, genome downsizing, chromosomal rearrangements including descending dysploidy. Diploidized genome(s) can enter a new round of polyploidization, forming a new neopolyploid subjected to next-round diploidization.

1.2 Chromosomal rearrangements during post-polyploid diploidization

At a given ploidy level, karyotype evolution is primarily mediated by CRs, which can alter chromosome size, structure, morphology and number. This section summarizes the key mechanisms underlying CR formation, the major types of rearrangements contributing to descending dysploidy, and the impact of dysploid changes on plant genome evolution and diversification.

1.2.1 Mechanisms of chromosomal rearrangements

Certain structural features can predispose genomic regions to instability and rearrangements. Recurrent CRs often share similar breakpoints across unrelated individuals, whereas non-recurrent CRs are typically unique, with distinct breakpoints in each case (Liu et al., 2012; Carvalho and Lupski, 2016; Burssed et al., 2022). Regardless of recurrence, most CRs originate from DNA double-strand breaks (DSBs), which can arise from exogenous factors (e.g., radiation and chemical mutagens) or endogenous sources (e.g., free radicals, mechanical forces, DNA replication and transcription errors) (Cannan and Pederson, 2016). Here, four major mechanisms underlying CR formation are summarized.

(1) Non-allelic homologous recombination (NAHR)

When DSBs occur in DNA, homologous recombination typically ensures high-fidelity repair by using an allelic sequence on the sister chromatid as a template. However, when

DSBs occur within or near repetitive sequences, the homology search may misidentify a homologous region elsewhere in the genome, leading to non-allelic repair (Hastings et al., 2009). NAHR occurs via ectopic crossover between low-copy repeats (LCRs) (**Figure 2a**), primarily during meiotic prophase I (especially at the pachytene stage), but can also take place during mitotic divisions (Sasaki et al., 2010; Startek et al., 2015). LCRs act as recombination substrates due to their sequence similarity, with their size, degree of homology, orientation, and genomic position influencing local genome architecture and instability (Stankiewicz and Lupski, 2002). Therefore, NAHR is regarded as a primary driver of recurrent CRs (Liu et al., 2012). The spatial orientation of the recombining LCRs dictates rearrangement topology: repeats in direct orientation mediate intra- or inter-chromosomal exchanges that produce deletions or duplications, whereas repeats in opposite orientation led to intra-chromosomal inversions. Pairing between LCRs on different chromosomes can additionally result in reciprocal translocations (RTs) (**Figure 2a**; Burssed et al., 2022; Vervoort and Vermeesch, 2023).

(2) Non-homologous end-joining (NHEJ)

The predominant DSB repair pathway throughout the cell cycle comprises two distinct modalities: canonical NHEJ (c-NHEJ) and microhomology-mediated end-joining (MMEJ). The c-NHEJ pathway directly ligates broken DNA ends through Ku70/80-mediated synapsis, often introducing small insertions/deletions (indels) at the repair junctions (**Figure 2b**; Barnes, 2001; Burssed et al., 2022). MMEJ relies on short microhomologies near the DSB site and is mediated by polymerase POLQ (Deriano and Roth, 2013; Kent et al., 2015; Schimmel et al., 2017). POLQ facilitates annealing of the resected ends and completes repair through low-fidelity DNA synthesis, frequently generating deletions or templated insertions of short sequence tracts at the break site (**Figure 2b**; Sfeir and Symington, 2015; Ahrabi et al., 2016; Merker et al., 2024).

(3) Replication-based mechanisms

The replication-based mechanisms, such as microhomology-mediated break-induced replication (MMBIR) and fork stalling and template switching (FoSTeS), have been proposed to explain the origin of complex rearrangements showing discontinuous duplications mixed with deletions, inverted duplications and triplications (Smith et al., 2007; Malkova and Ira, 2013). These events are initiated by replication fork stalling or

collapse, which triggers aberrant template switching. In forward template switching, the lagging strand invades a downstream replication fork, resulting in deletions, whereas backward invasion into an upstream fork generates duplications. Template switching between different chromosomes can yield translocations, and depending on the orientation of strand invasion, inversions may also arise (**Figure 2c**).

(4) Chromoanagenesis (chromosome rebirth)

The term chromoanagenesis, introduced by Holland and Cleveland (2012), encompasses three distinct but related mechanisms—chromothripsis, chromoplexy and chromoanasynthesis—which describe extensive, complex CRs arising from a single catastrophic event (Holland and Cleveland, 2012; Pellestor, 2019). Chromothripsis (chromosome shattering) occurs when a chromosome or chromosomal arm is fragmented into dozens to hundreds of pieces, often due to clustered DSBs (Shoshani et al., 2021; Simovic and Ernst, 2022). These fragments are then reassembled in a disordered manner, with random order and orientation relative to the original chromosome (**Figure 2d**). Chromoplexy (chromosome restructuring) involves a series of rearrangements that break and rejoin DNA segments from different chromosomes through c-NHEJ or MMEJ, thus leading to the formation of translocations (Baca et al., 2013; Zepeda-Mendoza and Morton, 2019). Unlike chromothripsis, chromoplexy typically affects fewer genomic segments (tens rather than hundreds), but the rearrangements are dispersed across multiple chromosomes (**Figure 2d**). Chromoanasynthesis (chromosome reconstitution) arises from defects in DNA replication. When replication forks stall or collapse, the lagging strand of a defective fork may disengage and undergo successive MMBIR and FoSTeS events with other nearby replication forks before completing synthesis on the original template (**Figure 2d**). This process can lead to complex structural variants such as deletions, duplications, triplications, and inversions, forming highly rearranged genomic regions (Pellestor and Gatinois, 2018).

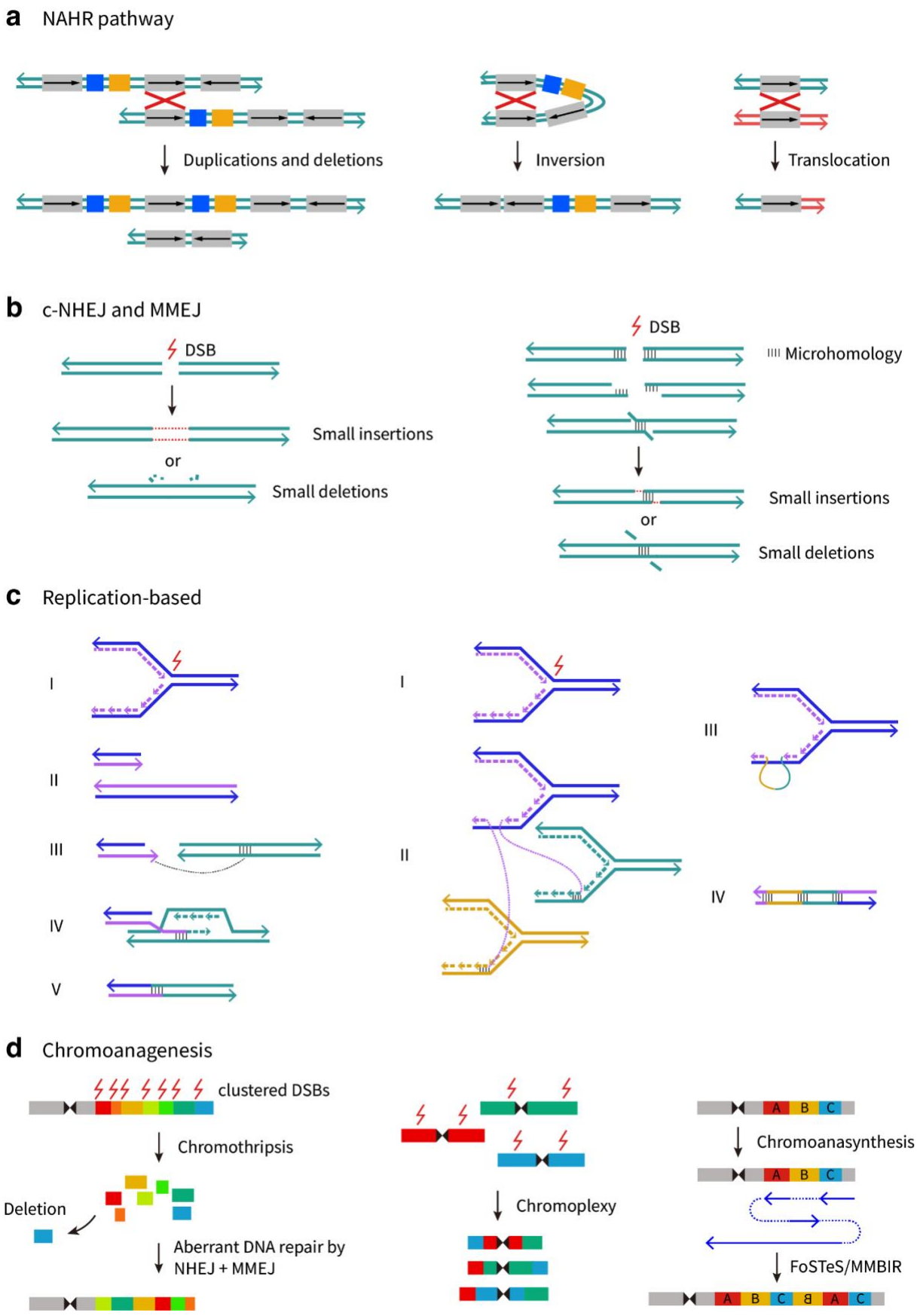

a NAHR pathway
Duplications and deletions
Inversion
Translocation
b c-NHEJ and MMEJ
DSB
Small insertions
or
Small deletions
DSB
Microhomology
Small insertions
or
Small deletions
c Replication-based
I
II
III
IV
V
I
II
III
IV
d Chromoanagenesis
clustered DSBs
Chromothripsis
Deletion
Aberrant DNA repair by
NHEJ + MMEJ
Chromoplexy
A B C
Chromoanasynthesis
A B C
FoSTeS/MMBIR
A B C A C

Figure 2 Molecular mechanisms underlying chromosomal rearrangements (CRs).
(**a**) Non-allelic homologous recombination (NAHR) leading to the formation of duplications/ deletions, inversions or reciprocal translocations. Low-copy repeats (LCRs) are indicated by grey duplicons with arrows indicating their direct or inverted orientation.
(**b**) Mechanisms of double-strand break (DSB) repair. The left panel illustrates canonical non-homologous end-joining (c-NHEJ), which repairs DSB by directly bridging and ligating the broken ends. This can happen through the editing of the broken ends with small insertions or deletions. The right panel illustrates microhomology-mediated end-joining (MMEJ). After a DSB, 5' to 3' resection generates two 3' single-stranded overhangs, which are aligned through regions of microhomology. Subsequent gap filling and ligation may result in deletions or insertions.
(**c**) Replication-based mechanisms. Microhomology-mediated break-induced replication (MMBIR) is shown in the left: (I) the replication fork encounters a DNA lesion (red) and stalls; (II) fork stalling leads to fork collapse and cleavage by an endonuclease, generating a single-ended DSB, depicted as the shorter DNA strand; (III) 5' to 3' resection exposes a region of microhomology and produces a 3' single-stranded overhang; (IV) a displacement loop (D-loop) forms as the overhang invades a homologous template strand at the microhomologous site, initiating DNA synthesis; and (V) synthesis is continued to the end of the chromosome. The resulting product contains rearranged genomic segments joined via microhomology. Fork stalling and template switching (FoSTeS) is shown on the right. (I) when a replication fork stalls, (II) the lagging strand disengages from its original template and through microhomology-mediated annealing, invades alternative replication forks (green and yellow) to resume synthesis. (III) eventually, the strand can return to its original template and restart synthesis. (IV) the final product contains segments from multiple genomic loci, brought together through microhomology-dependent template switching.
(**d**) Chromoanagenesis-derived CRs. Chromothripsis can result from shattering of one chromosome or chromosome arm followed by incomplete DNA repair through NHEJ and MMEJ, generating complex rearrangements. Chromoplexy involves simultaneous DSBs on multiple chromosomes, which are repaired and rearranged into derivative configurations. Chromoanasynthesis occurs through replication-based processes (FoSTeS and MMBIR), generating templated insertions that may have increased copy number and variable orientation.

1.2.2 Dysploid chromosomal rearrangements

Non-allelic recombination between two or more non-homologous chromosomes forms the mechanistic foundation of descending dysploidy. At least four major types of whole-chromosome or chromosome-arm translocation events have been implicated in the reduction of chromosome numbers: nested chromosome insertion (NCI), end-to-end

translocation (EET), Robertsonian translocation and repeated unbalanced reciprocal translocations (unbalanced-RTs) (Zhuang et al., 2014; Sun et al., 2024; Jiang et al., 2025).

NCI refers to "insertion" of an entire chromosome into or near the pericentromeric region of a recipient chromosome. This process typically requires at least three DSBs: two at both ends of the inserted chromosome to render the telomeric ends "sticky", and one within the pericentromere of the recipient chromosome. The inserted chromosome is then integrated between the arms of the recipient chromosome, resulting in the formation of a single, fused chromosome (**Figure 3**) (Luo et al., 2009; Wang et al., 2015; Lusinska et al., 2018; Schubert, 2021; Lysak, 2022). If the pericentromeric breakage in the recipient chromosome and/or subsequent resection of the resulting centromeric fragments do not remove the CENH3-containing nucleosomes, a NCI would generate a dicentric or tricentric fusion chromosome with one "strong" centromere from the inserted chromosome and one or two "weaker" centromere(s) from the recipient chromosome (**Figure 3**). Thus, inactivation or deletion of the recipient centromere or its fragments is required to stabilize the resulting fusion chromosome.

Several chromosome-level genome assemblies in grasses should, in principle, provide insights into the fate of (sub)telomeric repeats from the inserted chromosome and pericentromeric sequences of the recipient chromosome following NCI events. Breakage within or near the centromeric region of the recipient chromosome can displace pericentromeric sequences toward internal regions on both arms of the fusion chromosome (**Figure 3**) (Wang et al., 2015; Schubert, 2021; Lysak, 2022). As a result, the two arms of the fusion chromosome are expected to retain adjacent clusters of pericentromeric and telomeric repeat sequences (**Figure 3**). However, given the limited reports of pericentromeric repeat peaks at fusion junctions, it is likely that these sequences were deleted either during the rejoining process or shortly thereafter. For example, the extant chromosomes of Pooidea grasses with $n = 7$ (e.g., barley and *Aegilops tauschii*) and *Brachypodium* ($n = 5$) are thought to have formed independently through sequential NCIs in the last tens of millions of years, following their divergence from *Oryza* (rice; $n = 12$) (Wang et al., 2019). In *B. distachyon*, footprints of the recipient centromere in the form of 156-bp satellite repeat and retrotransposons have been detected on all five chromosomes (Li et al., 2018). However, no reports of centromeric

or (sub)telomeric repeat sequence peaks have been found at other chromosomal locations, especially at the presumed fusion junctions, reflecting that such repeats are likely to have been erased.

EETs mediate tandem fusions of two chromosomes after DSB misrepair in their (sub)telomeric regions (**Figure 3**). The resulting (dicentric) chromosome harbors two centromeres separated by a considerable physical distance. To ensure proper segregation, one of the centromeres must become inactive and/or be eliminated, likely via epigenetic silencing and/or recombinational loss of CENH3-containing nucleosomes (Zhang et al., 2010; Stimpson et al., 2012; Lysak, 2014). EETs can occur between any types of chromosomes, including metacentric, acrocentric and telocentrics ones.

A specific form of terminal fusion—when two acrocentric or telocentric chromosomes join at their centromeric ends (i.e., the short arms)—is referred to as a Robertsonion translocation (Robinson et al., 1994; Bandyopadhyay et al., 2002; Keymolen et al., 2011). The product of Robertsonian translocation is a large fused chromosome alongside a small, centromere-bearing minichromosome, which is often unstable and prone to loss (**Figure 3**). While Robertsonian translocations are well documented in human genomes (Jarmuz-Szymczak et al., 2014; Poot and Hochstenbach, 2021), they are less frequently reported in plants. This underrepresentation may stem from limited resolution in comparative genome maps, structural changes that obscure fusion points, and lack of knowledge about ancestral karyotypes. These challenges make it difficult to clearly differentiate Robertsonian translocations from other types of chromosomal fusions such as EETs (Sun et al., 2024; Jiang et al., 2025).

In contrast to reciprocal translocations (RTs), where two chromosomes mutually exchange segments, unbalanced-RTs involve the unidirectional transfer of one or more segments from a donor chromosome to the recipient(s) (**Figure 3**; Schubert and Lysak, 2011; Chang et al., 2013; Parisod and Badaeva, 2020). While both RTs and single unbalanced-RT do not alter chromosome number, two or more unbalanced-RTs can redistribute segments from a single donor to multiple recipient chromosomes, ultimately resulting in complete relocation of the donor chromosome (**Figure 3**). Dysploid unbalanced-RTs generally involve at least three DSBs: one on the donor

chromosome and two on different recipient chromosomes. In principle, these DSBs can occur at any location on the involved chromosomes (Schubert and Lysak, 2011), but typical situation is that one DSB occurs near/at the pericentromeric region of the donor chromosome, and two DSBs near/at the (sub)telomeric regions of different recipient chromosomes (Artandi et al., 2000; Ali et al., 2013). The donor chromosome fragments can be separately inserted into these DSB sites on the recipient chromosomes, resulting in two fused chromosomes (**Figure 3**).

Regardless of the type of chromosome fusion, cells need to cope with the survival challenges posed by dicentric or multicentric chromosomes and the substantial increase in chromosome length. In eukaryotes with localized (monocentric) centromeres, a single functional centromere per chromosome is the evolutionarily preferred configuration, whereas dicentric chromosomes are often associated with genomic instability and deleterious consequences. The classical view of dicentric behavior, first described by McClintock (1939; 1941) in maize, shows that dicentrics are inherently unstable and typically undergoing repeated cycles of anaphase bridge formation and breakage. Similar behaviors have been observed in Arabidopsis (Yokota et al., 2011) and wheat (Fu et al., 2012). To stabilize newly formed fused chromosomes arising from EETs, some NCIs or unbalanced-RTs, a common cellular strategy is to suppress and/or epigenetically eliminate the additional, often weaker centromeres. These processes may involve loss of key centromere and kinetochore proteins (e.g., CENP-A, CENP-C, and CENP-E; Earnshaw et al., 1989; Sullivan and Schwartz, 1995), epigenetically mediated chromatin remodeling (MacKinnon and Campbell, 2011; Sato et al., 2012; Fu et al., 2012; Stimpson et al., 2012), and removal of centromeric satellite DNA (Chiatante et al., 2017).

Another challenge posed by chromosome fusion is the alteration of chromosome morphology and arm length, particularly when one arm becomes significantly extended by incorporating an entire additional chromosome. Experimental evidence using engineered chromosomes in field bean lines demonstrated that the longest chromosome arm must not exceed half of the average length of the spindle axis at telophase (Schubert and Oud, 1997). If a sister chromatid arm exceeds this threshold, it may fail to separate properly during mitosis, and its distal region can become physically broken during new cell wall formation (Li et al., 2011). This mechanical constraint provides a plausible

explanation for why sequential fusions involving three or more chromosomes are rarely observed in nature, and may also contribute to the selective advantage of metacentric chromosomes with more balanced arm lengths (Wyttenbach and Hausser, 1996; Fedyk and Chętnicki, 2007).

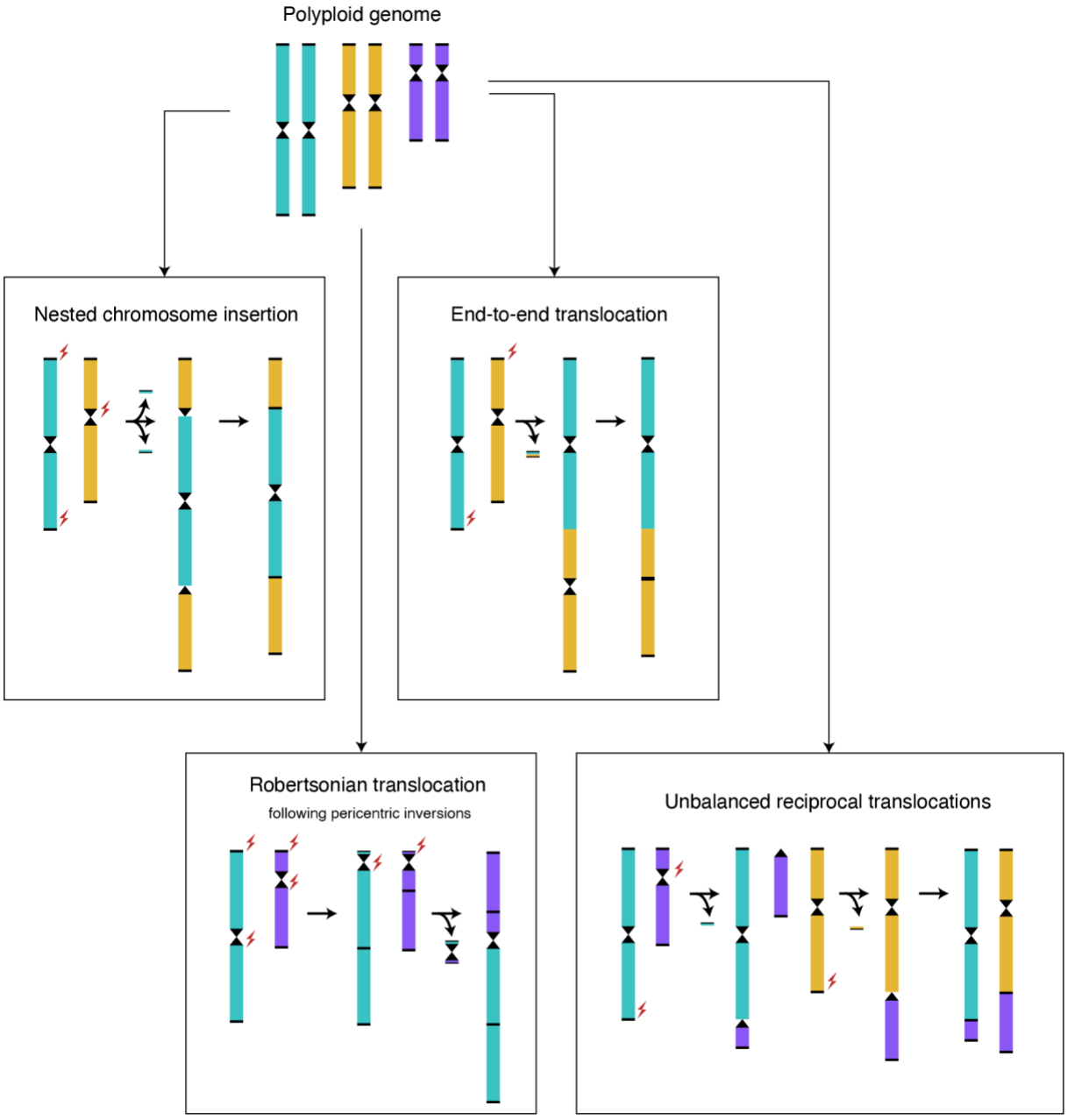


Figure 3 Proposed dysploid chromosomal rearrangement types in angiosperms, adopted from Mandáková and Lysak (2018).
Nested chromosome insertions (NCI) are translocation events in which an "insertion" chromosome becomes integrated into the (peri)centromeric regions of a "recipient" chromosome. End-to-end translocations (EET) involve the fusion of two non-homologous

chromosomes via double-strand breaks (DSBs) at chromosome ends, resulting in the formation of a dicentric chromosome. To ensure chromosome stability, one of the two centromeres must be inactivated and/or eliminated. Descending dysploidy through Robertsonian translocation occurs via recombination between two telo/acrocentric chromosomes. The outcome is a large translocation chromosome—comprising substantial portions of the original chromosome arms—and a small minichromosome primarily derived from the centromeric region of the telo/acrocentric chromosome. Due to its meiotic instability, the minichromosome is typically eliminated. In the case of descending dysploidy mediated by unbalanced reciprocal translocations (unbalanced-RTs), repeated unidirectional recombination events occur between a donor chromosome and multiple recipient chromosomes. This process results in the complete redistribution of the donor chromosome's segments to other chromosomes.

1.2.3 Impact of dysploid changes on diversification of the Brassicaceae family

The Brassicaceae (Cruciferae) belongs to one of the 15 largest angiosperm families, comprising ~4,158 species in 357 genera (German et al., 2023; Al-Shehbaz, 2025). Phylogenetic studies have proposed five major lineages (supertribes), each consisting of monophyletic clades classified as tribes (**Figure 4**; Al-Shehbaz, 2012; Hendriks et al., 2023). All extant tribes are descended from a paleotetraploid ancestor that originated from the At-α WGD event, dated to ~35 Mya (Bowers et al., 2003; Hohmann et al., 2015; Walden et al., 2020; Hendriks et al., 2023). Following the At-α event, several tribe-specific mesopolyploidizations have been identified, and a combined analysis of cytogenomic and transcriptomic data found that at least 11 out of 52 Brassicaceae tribes had independent mesopolyploidizations followed by different degrees of diploidization (Mandáková et al., 2017a). Collectively, these additional WGD events potentially impacted around 130 genera, averaging 1.45 WGDs per genus over the past 35 million years—a rate notably higher than the angiosperm-wide average (Landis et al., 2018). Moreover, more than half (~43.3 %) of Brassicaceae taxa are classified as neopolyploids (Hohmann et al., 2015). Given that paleo- and mesopolyploidy are widely recognized as important driving forces of past genome evolution in angiosperms, and that neopolyploidy may fuel ongoing and future diversification (Hohmann et al., 2015), the Brassicaceae represents an excellent model system to study diversification driven by WGD and PPD.

The Ancestral Crucifer Karyotype (ACK), proposed by Schranz et al. (2006), consists of eight chromosomes and initially represented the putative ancestral genome of the Camelinodae supertribe (Lineage I). The boundaries of 22 conserved genomic blocks (GBs) in ACK (n = 8) were defined based on comparative genetic mapping and cross-species fluorescence *in situ* hybridization (FISH) using chromosome-specific BAC clones from *A. thaliana* and related species (**Figure 5a**; Schranz et al., 2006; Lysak et al., 2016). Another ancestral genome structure, the ancestral Proto-Calepineae Karyotype (ancPCK; n = 8), was proposed for the Biscutelleae clade (Geiser et al., 2016; Mandáková et al., 2018; Guo et al., 2021) and differs from ACK by a single RT (**Figure 5a**). Based on shared homeologous GBs, the Proto-Calepineae Karyotype (PCK; n = 7; Mandáková and Lysak, 2008) is considered the ancestral karyotype of Brassicodae (Lineage II) and has been retained in tribes such as Calepineae, Conringieae, and Noccaeeae, whereas karyotypes of Eutremeae, Isatideae, Schrenkielleae, Thlasipideae, and Sisymbrieae are characterized by an additional translocation (translocation Proto-Calepineae Karyotype; tPCK; n = 7; Mandáková and Lysak, 2008).

For a long time, the ACK genome has served as the foundational reference for studying karyotype evolution in the Brassicaceae, with ancPCK, PCK, and tPCK proposed to have evolved from ACK in a stepwise manner (**Figure 5a**; Mandáková and Lysak, 2008; Mandáková et al., 2018; Kang et al., 2020; Bayat et al., 2021). Nevertheless, a recent study introduced a revised ancestral karyotype shared by Camelinodae and Brassicodae, termed the Ancestral Camelinodae-Brassicodae Karyotype (ACBK), consisting of eight protochromosomes (Jiang et al., 2025). Notably, this study challenges the traditional view by proposing that the alternative ancestor of Camelinodae (Lineage I) may have a Cardamineae-like genome, designated as Ancestral Karyotype I (AKI), which differs from ACK by a single RT (**Figure 5b**; Jiang et al., 2025). This hypothesis is primarily based on phylogenetic evidence positioning Cardamineae as the basal group to most other Camelinodae tribes (Hendriks et al., 2023). However, the proposed scenario gives rise to a "chicken-and-egg" dilemma: does the genome structure of Cardamineae, as an outgroup, truly reflect the ancestral state of Camelinodae, or is the translocation observed in AKI a lineage-specific feature of Cardamineae, with ACK representing the shared ancestral state of other Camelinodae tribes? (**Figure 5c** and **d**). Addressing this question remains challenging due to several unresolved issues: (1) it remains unknown

whether additional extant or extinct taxa are positioned at the base of Camelinodae (Lineage I), and if so, what their genome structures might reveal. (2) the ancestral karyotypes of the supertribes Hesperodae (Lineage III) and Arabodae (Lineage IV)—both located outside Camelinodae—remain unresolved, but could be informative if either contains the sister lineage to Camelinodae.

Therefore, comparative genomic studies of Brassicaceae have uncovered four major trajectories of chromosome evolution (**Figure 6**): (1) Karyotypic stability: some lineages have preserved highly conserved genome structures over millions of years (e.g., *A. lyrata* and *Capsella* spp.) (**Figure 6a**). (2) Dysploid taxa: several taxa have undergone descending dysploidy from the ancestral $n = 8$ to $n = 7$, 6, or 5 (e.g., *Camelina hispida*, *C. neglecta*, and *A. thaliana*) (**Figure 6b**). (3) Post-WGD karyotype stability: certain lineages experienced additional WGD event(s), yet retained stable karyotypes after polyploidization. For example, *Streptanthus diversifolius* ($n = 14$; Thelypodieae) arose through hybridization between two PCK-like genomes ~7.4 Mya and has retained a largely intact ancestral genome structure (Davis et al., 2025; **Figure 6c**). (4) Post-WGD genome reshuffling with descending dysploidy: other lineages underwent CRs and descending dysploidy following WGD event(s). For example, in the Brassiceae, a mesohexaploid ancestor formed from hybridization among three tPCK-like genomes (~11–13 Mya) gave rise to species with varying chromosome numbers ($n = 10$, 9, or 8), reflecting substantial karyotypic diversification (Lysak et al. 2005, 2007; Cheng et al., 2014; Huang et al., 2020; Cai et al., 2021; Yang et al., 2023; **Figure 6d**).

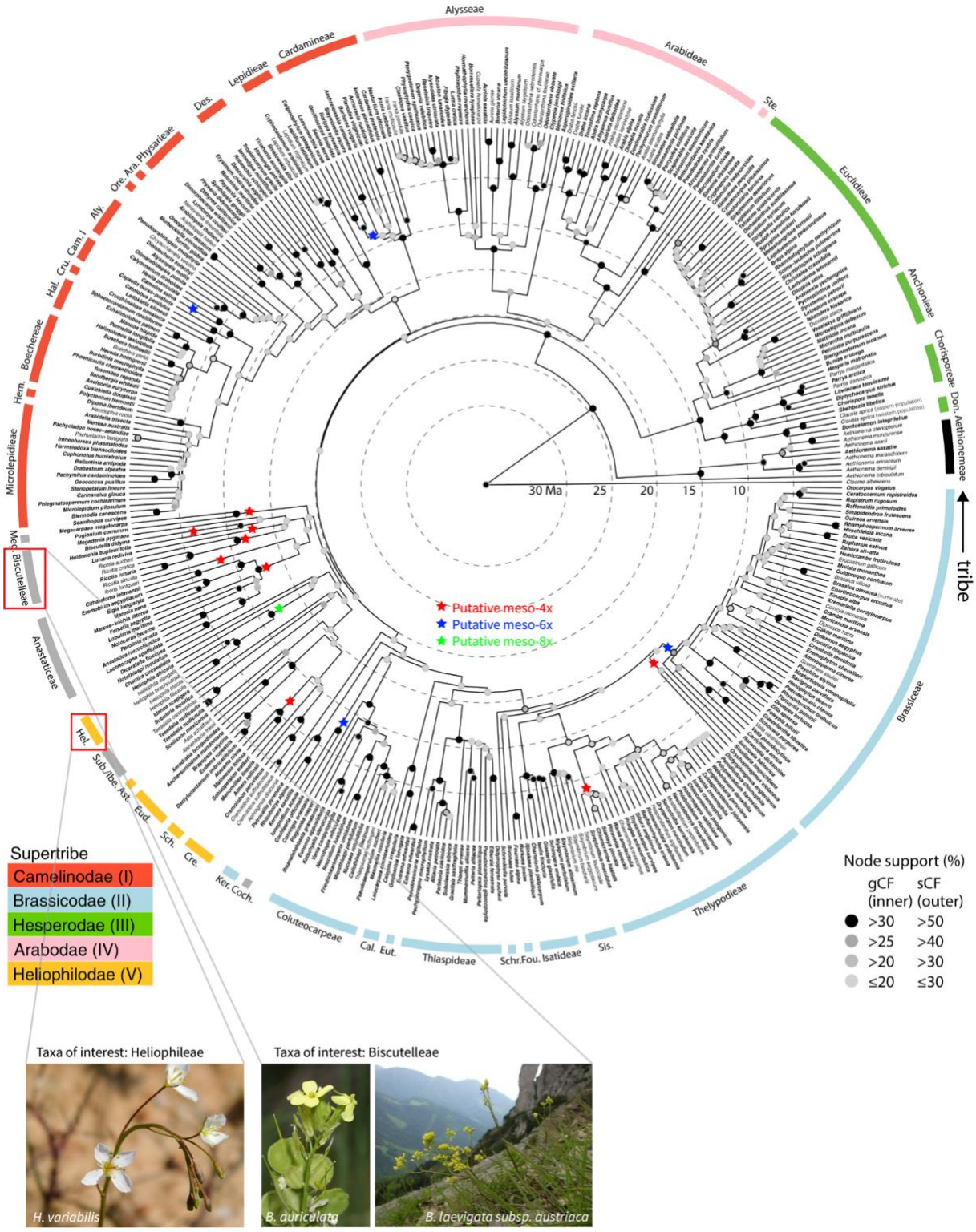


Figure 4 Time-calibrated genus-level Brassicaceae Tree of Life (BrassiToL) inferred from a maximum likelihood analysis of a 297 nuclear gene supermatrix, adopted from Hendriks et al. (2023).

Colored groups are the main core Brassicaceae lineages (supertribes I–V). Genus type species are highlighted in bold. All tribes with more than a single representative are listed. Putative mesotetraploidization (red stars), mesohexaploidization (blue stars), and

mesooctoploidization (green star) events mapped onto the BrassiToL based on data by Mandáková et al. (2017a) and Huang et al. (2023). These symbols indicate tribes where polyploidization may have occurred, without implying the exact timing of these events. The taxa of interest in this thesis—Heliophileae and Biscutelleae (see Section 1.4)—are highlighted with red boxes. Tribe abbreviations are as follows: Aly.: Alyssopsideae; Ara.: Arabidopsideae; Ast.: Asteae; Cal.: Calepineae; Cam. I: Camelineae I; Coch.: Cochlearieae; Cre.: Cremolobeae; Cru.: Crucihimalayeae; Des.: Descurainieae; Don.: Dontostemoneae; Eud.: Eudemeae; Eut.: Eutremeae; Fou.: Fourraeeae; Hal.: Halimolobeae; Hel.: Heliophileae; Hem.: Hemilophieae; Ibe.: Iberideae; Ker.: Kernereae; Meg.: Megacarpaeeae; Ore.: Oreophytoneae; Sch.: Schizopetaleae; Schr.: Schrenkielleae; Sis.: Sisymbrieae; Ste.: Stevenieae; Sub.: Subularieae.

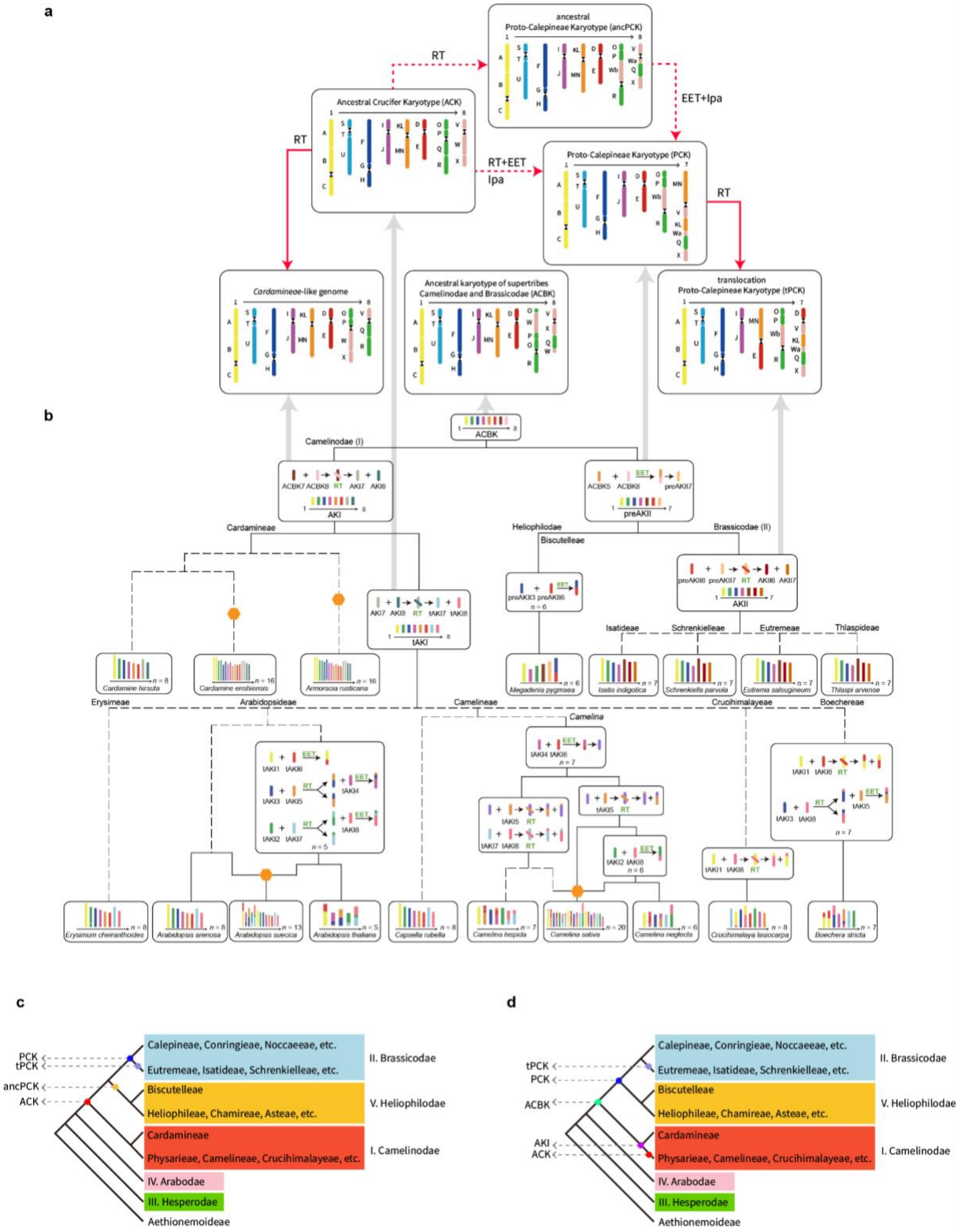


Figure 5 Two proposed karyotype evolution pathways in the Brassicaceae.

(**a**) Traditional pathway (red arrows) originally proposed by Mandáková and Lysak (2008) and refined in later studies (Lysak et al., 2016; Mandáková et al., 2018; Kang et al., 2020; Bayat et al., 2021). In this pathway, the Ancestral Crucifer Karyotype (ACK) comprises eight chromosomes and 22 conserved genomic blocks (A–X). The ancestral Proto-Calepineae

Karyotype (ancPCK) arose from ACK chromosomes via a reciprocal translocation (RT) event. The Proto-Calepineae Karyotype (PCK) evolved either from ancPCK through an end-to-end translocation (EET) and a paracentric inversion (Ipa) (Mandáková et al., 2018; Bayat et al., 2021), or directly from ACK (Mandáková and Lysak, 2008). The translocation PCK (tPCK) subsequently was formed from PCK via a RT event.

(**b**) Alternative karyotype evolution scenario proposed by Jiang et al. (2025). In this scenario, ACBK represents the ancestral karyotype shared by Camelinodae and Brassicodae. Within this framework, the Ancestral Karyotype of Camelinodae (AKI) is derived from ACBK and differs from it by a single RT.

(**c**) and (**d**) Differences between the traditional and Jiang et al.'s scenario regarding the phylogenetic placement of ancestral karyotypes (ACK, ancPCK, PCK, and tPCK) within a simplified BrassiToL topology (Hendriks et al., 2023). (**c**) In the traditional pathway, ACK is considered as the ancestral genome of Camelinodae, Heliophilodae and Brassicodae (Schranz et al., 2006), with Cardamineae exhibiting a tribe-specific RT. Then the ancPCK, PCK and tPCK were evolved subsequently (ACK→ ancPCK; ACK → PCK → tPCK). (**d**) In the scenario proposed by Jiang et al., ACBK is hypothesized to represent the ancestral genome of Camelinodae, Heliophilodae and Brassicodae. AKI is considered ancestral to Camelinodae, and PCK is regarded as the ancestral genome of Brassicodae and Heliophilodae, from which tPCK later evolved (ACBK → AKI → ACK; ACBK → PCK → tPCK).

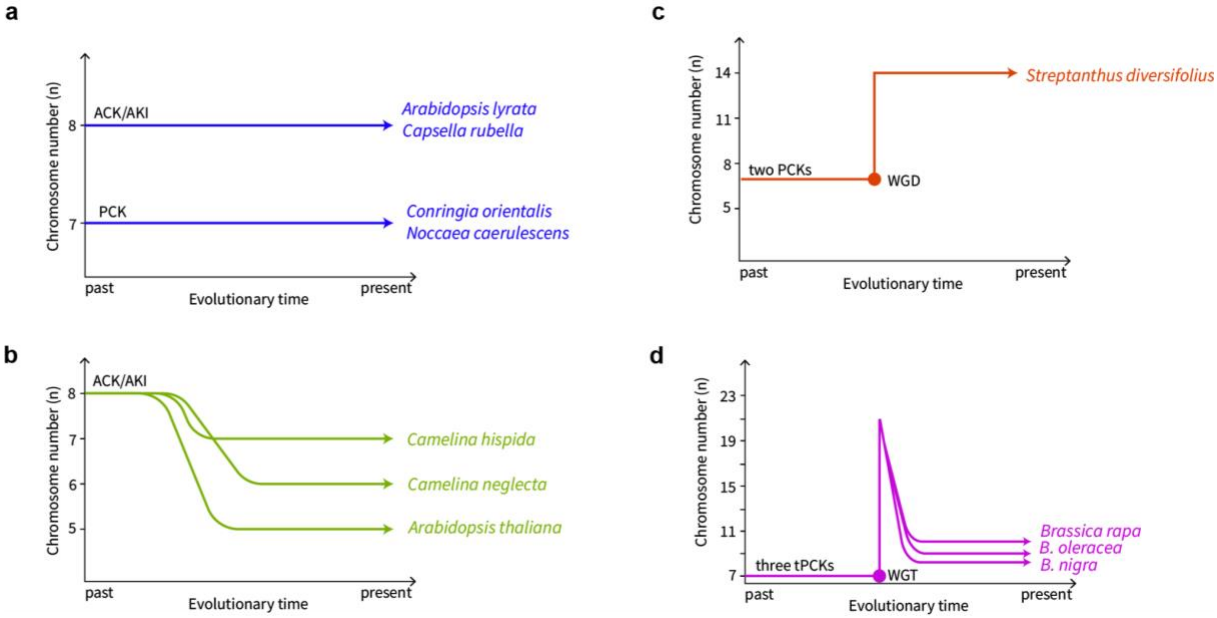


Figure 6 Four representative trajectories of karyotype (genome) evolution in Brassicaceae. Each panel illustrates cases of chromosome evolution dynamics, with the y-axis representing chromosome number and the x-axis representing evolutionary time (from past to present). Panels (**b**) and (**d**) illustrate the extent of chromosome number reduction, and do not reflect the actual speeds of diploidization among species.

1.3 Origins and mechanisms of biased subgenome fractionation

Regardless of karyotypic configuration, when a post-polyploid genome contains two or more recognizable and largely intact parental genomes, these constituents are referred to as subgenomes (Schnable et al., 2011). A phenomenon commonly observed in allopolyploids—but rarely in autopolyploids—is biased subgenome fractionation, characterized by asymmetrical gene retention and expression (Thomas et al., 2006; Bird et al., 2018; Cheng et al., 2018). Typically, one subgenome becomes dominant, retaining more protein-coding genes and generally exhibiting higher expression levels relative to its counterpart(s).

Understanding the mechanisms underlying subgenome dominance and biased fractionation is a key focus of polyploidy research. Firstly, the abundance and distribution of TEs are thought to be associated with biased gene expression in many allopolyploid genomes (Alger and Edger, 2020). TE activation can compromise genome stability and often represses the expression of nearby genes. Consequently, TE-rich subgenomes often exhibit lower gene expression levels compared to TE-poor subgenomes (e.g., Woodhouse et al., 2010; Schnable et al., 2011; Cheng et al., 2016; Edger et al., 2017; Xu et al., 2020). However, an increasing number of counterexamples—such as the synthesized *Brassica* and *Nicotiana* allotetraploids—indicate that the TE density alone does not fully explain subgenome expression bias (Mhiri et al., 2019; Zhang et al., 2023). Secondly, subgenome dominance may be shaped by genetic incompatibilities. Although hybridization is the primary route to allopolyploid formation, it can also disrupt pre-existing regulatory networks. The merger of divergent diploid progenitor genomes into a single nucleus may interfere with the coordination of essential metabolic, signaling, and transcriptional processes. As a result, distinct biological pathways may become preferentially regulated by different subgenomes, potentially leading to a division of phenotypic traits between subgenomes (Flagel et al., 2008; Bird et al., 2018). A third factor is chromatin dynamics, which are increasingly recognized as key contributors to subgenome expression asymmetry. Recent studies in soybean have demonstrated that the gain or loss of DNA sequences, along with mutations in *cis*-regulatory elements located within accessible chromatin regions, can alter the expression levels and tissue specificity of duplicated genes (Fang et al., 2023; Huang et al., 2024). Similarly, in the octoploid strawberry

(*Fragaria* × *ananassa*), dominant subgenome expression is strongly correlated with chromatin accessibility (Fang et al., 2024), underscoring the role of chromatin architecture and epigenetic regulation in establishing and maintaining subgenome dominance.

Many questions in subgenome fractionation research remain unresolved. For instance, is subgenome dominance predetermined by inherent differences between the parental genomes? In some studies, synthetic polyploids have been generated and used as surrogates for "original" natural polyploids. This approach assumes that, like their natural counterparts, synthetic polyploids have experienced minimal evolution and selection. Many studies have reported that both types of polyploids consistently share the same "dominant" subgenome with comparable levels of subgenome bias. For example, genome-wide expression comparisons between synthetic and natural allohexaploid wheat revealed that fractionation consistently favored the same subgenome (Chagué et al., 2010). Similarly, in six independent resynthesized lines of *Brassica napus*, the same parental genome was repeatedly dominant in expression, matching patterns observed in the ~7500-year-old natural *B. napus* (Bird et al., 2021). Another key question concerns whether the degree of subgenome fractionation increases over time. In *Mimulus peregrinus*, Edger et al. (2017) compared gene expression differences among a di-haploid hybrid, a synthetic allopolyploid and a natural genotype, and found that gene expression dominance strengthened over successive generations. This, in turn, raises the question whether such increased fractionation ultimately stabilizes. While natural neopolyploids offer valuable insight into early post-polyploid dynamics, their relatively short evolutionary timescales may limit the detection of long-term trends, making it challenging to fully assess the evolutionary trajectories of paleo- and mesopolyploid lineages.

1.4 Taxa of interest

1.4.1 Heliophileae

Members of the family Brassicaceae occur on every continent except Antarctica. While many are globally widespread, certain tribes are restricted to specific (sub)continents or smaller biogeographic regions, providing valuable models for investigating genome evolution under long-term geographic isolation (Al-Shehbaz, 2012; Raza et al., 2020).

One such lineage is the tribe Heliophileae, which is endemic to southern Africa (eSwatini, Lesotho, Namibia, and South Africa), and represents the most morphologically diverse crucifer clade (Mummenhoff et al., 2005; Mandáková et al., 2012; Al-Shehbaz, 2025). The genus *Heliophila*, comprising 106 species, encompasses diverse life forms, from small, short-lived annual herbs to perennial herbs, subshrubs, and tall woody shrubs (e.g., *H. brachycarpa*). Species differ markedly in a range of traits, including leaf dissection (entire to variously divided), petal length (1.2–30 mm) and color (white, pink, lavender, purple, blue, or yellow), ovule number (1–80), fruit length (2–120 mm) and shape (linear, lanceolate, oblong, ovate, elliptic, or orbicular) (Mummenhoff et al., 2005; Mandáková et al., 2012; Dogan et al., 2021).

Phylogenetic analyses based on internal transcribed spacer (ITS) sequences (Mummenhoff et al., 2005) and low-coverage whole-genome sequencing (Dogan et al., 2021) support the monophyly of *Heliophila* and its sister-group relationship with *Chamira*. Within *Heliophila*, four major subclades (designated A–D; **Figure 7**) have been resolved, and chromosome number variation largely mirrors these clades. Most species in clade A possess chromosome numbers $2n = 20$, while $2n = 22$ is prevalent in clade C. The chromosome number of $2n = 44$ occurs in two morphologically related species in clade D but is rare in other clades. The known chromosome numbers of clade B species exhibit the greatest variation ($2n = 16$, 22, 26, 32, and 64; Mandáková et al., 2012; Dogan et al., 2021). This broad variation in chromosome number likely reflects a complex interplay of an ancient polyploidization, subsequent PPD, and more recent WGD events (e.g., $2n = 64$ species in clade B and $2n = 44$ species in clade D). Supporting this view, Dogan et al. (2021) identified similarly positioned synonymous substitution (*Ks*) peaks in four *Heliophila* species and one *Chamira* species, indicating that both genera likely share a mesopolyploid WGD, dated to ~26–29 Mya. Earlier cytogenomic analyses also proposed that the tribe Heliophileae originated likely from a mesohexaploid ancestral genome (Mandáková et al., 2012; Mandáková et al., 2017a), but the identification of more than three homeologous copies of some GBs via comparative chromosome painting (CCP) implied that the tribe might have a more complex evolutionary history (Mandáková et al., 2012).

Before the work presented here (Huang et al., 2023), no reference genome had been available for any *Heliophila* species, which substantially hindered genomic research not only within the genus itself but also across the Heliophilodae supertribe (Lineage V; **Figure 4**).

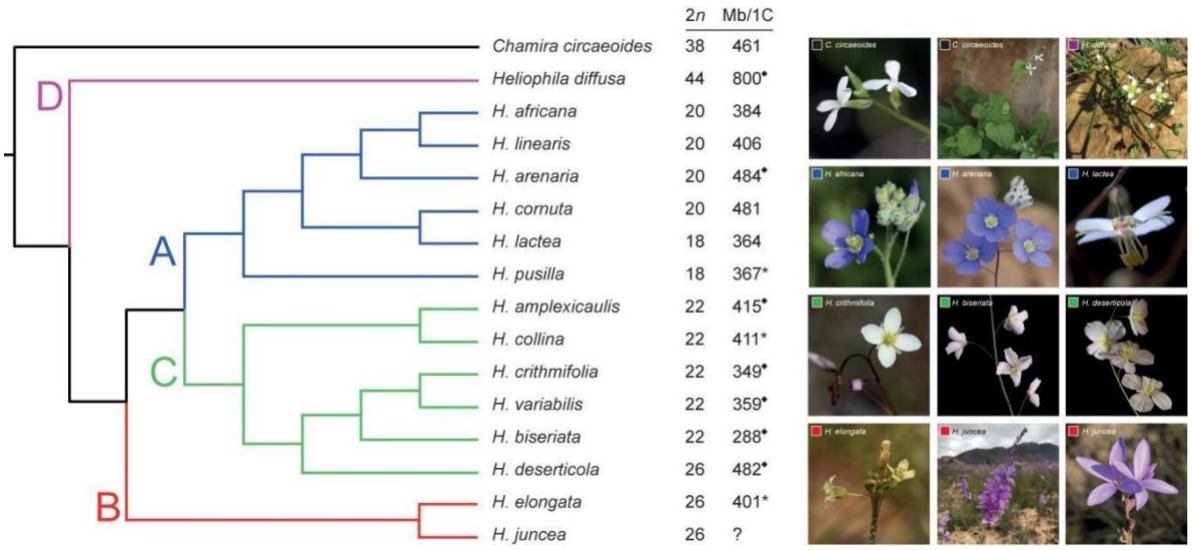


Figure 7 Schematic relationships among 15 *Heliophila* species and *Chamira circaeoides* based on the ITS phylogeny (Dogan et al., 2021).
Capital letters refer to four *Heliophila* clades (A-D). Diamond symbols indicate average C-values based on estimates for two or more populations; asterisks indicate C-values obtained by flow-cytometric analysis of dried leaf material.

1.4.2 Biscutelleae

The Biscutelleae tribe (~74 species in five genera; POWO, 2025) was re-established by German and Al-Shehbaz (2008), including *Biscutella* and *Megadenia* due to their phylogenetic and morphological affinities (German et al., 2009). Three additional genera—*Heldreichia*, *Lunaria* and *Ricotia*—were later included based on molecular phylogenetic analyses (Özüdoğru et al., 2015, 2016, 2017). Species belonging to Biscutelleae exhibit a broad and ecologically diverse geographic distribution. The genus *Biscutella* L. comprising ~45–53 species of annual herbs and dwarf shrubs, is widely distributed through the Mediterranean basin, Central Europe, and Southwest Asia (Al-Shehbaz, 2012; POWO, 2025). The genus *Ricotia*, containing 10 recognized species, is narrowly restricted to the eastern Mediterranean region (Özüdoğru et al., 2022). The genus *Heldreichia* has its diversity center in Anatolia and is represented by a single,

morphologically polymorphic species (*H. bupleurifolia*; Parolly et al., 2010). *Lunaria*, native to Europe and comprises three species (*L. annua*, *L. rediviva*, and *L. telekiana*) (Khapugin and Chugunov, 2023; POWO, 2025). *Megadenia* is an alpine genus, primarily distributed on the Qinghai–Tibetan Plateau, with disjunct populations also found in northern China and Russia (Yang et al., 2021; 2023).

The phylogenetic placement of the tribe Biscutelleae remains contentious. In the plastome-based phylogenies of the Brassicaceae, Biscutelleae occupies a basal position within the Brassicodae (Lineage II) (Mandáková et al., 2018; Nikolov et al., 2019; Walden et al., 2020). However, nuclear phylogenomic analyses have yielded conflicting topologies. In an ASTRAL species tree derived from transcriptomic data (Guo et al., 2021), as well as a phylogeny based on 1,421 exon trees (Nikolov et al., 2019), Biscutelleae was recognized as the sister group to Camelinodae (Lineage I). The latter view was confirmed by Hendriks et al., (2023) using 297 nuclear genes constructed in BrassiToL, although the Biscutelleae was not assigned to any of the five supertribes (**Figure 4**). The incongruence between plastid and nuclear phylogenies likely reflects a complex evolutionary history, possibly shaped by reticulate evolution and hybridization events. The first genomic insight, from *Biscutella laevigata*, showed a WGD followed by biased fractionation, extensive structural diploidization, and descending dysploidy, with the mesotetraploid ancestral genome ($n = 16$) inferred to have arisen from the merger of two ancPCK-like genomes ($n = 8$) (Geiser et al., 2016). A subsequent study demonstrated that another Biscutelleae genus, *Ricotia*, had also a polyploid origin, but via a more distant hybridization between a maternal ancPCK-like genome ($n = 8$) and a paternal PCK-like genome ($n = 7$) (Mandáková et al. 2018). Nevertheless, Guo et al., (2021) proposed that the Biscutelleae tribe represents an assemblage of paleotetraploid, mesotetraploid, and neotetraploid genera. The ancient ancPCK-like genome diversified into several $n = 8$ genomes, some further altered by dysploid rearrangements to $n = 6$ (*Megadenia*) and $n = 5$ (ancestral diploid genome of *Heldreichia*). Allotetraploids originated independently—either from hybridization between closely related ancPCK-like genomes (8 + 8; *Biscutella*) or from more distant crosses between ancPCK ($n = 8$) and PCK ($n = 7$) genomes (8 + 7; *Lunaria* and *Ricotia*) (**Figure 8**).

Before the work presented here (Huang et al., 2025), several important questions regarding *Biscutella*—the largest genus within the tribe Biscutelleae—remained unresolved. Although variation in genome size (ranging from 0.69 to 1.83 pg; Pellicer and Leitch, 2020) and chromosome number (n = 6, 8, or 9; Rice et al., 2015) has been documented within the genus (excluding neotetraploids), it was unclear how PPD had shaped these genomes. Specifically, it was not clear whether diploidization processes in *Biscutella* species had occurred through a single shared trajectory or through multiple and independent processes. If independent, did these parallel diploidization trajectories have led to convergent or divergent genomic consequences?

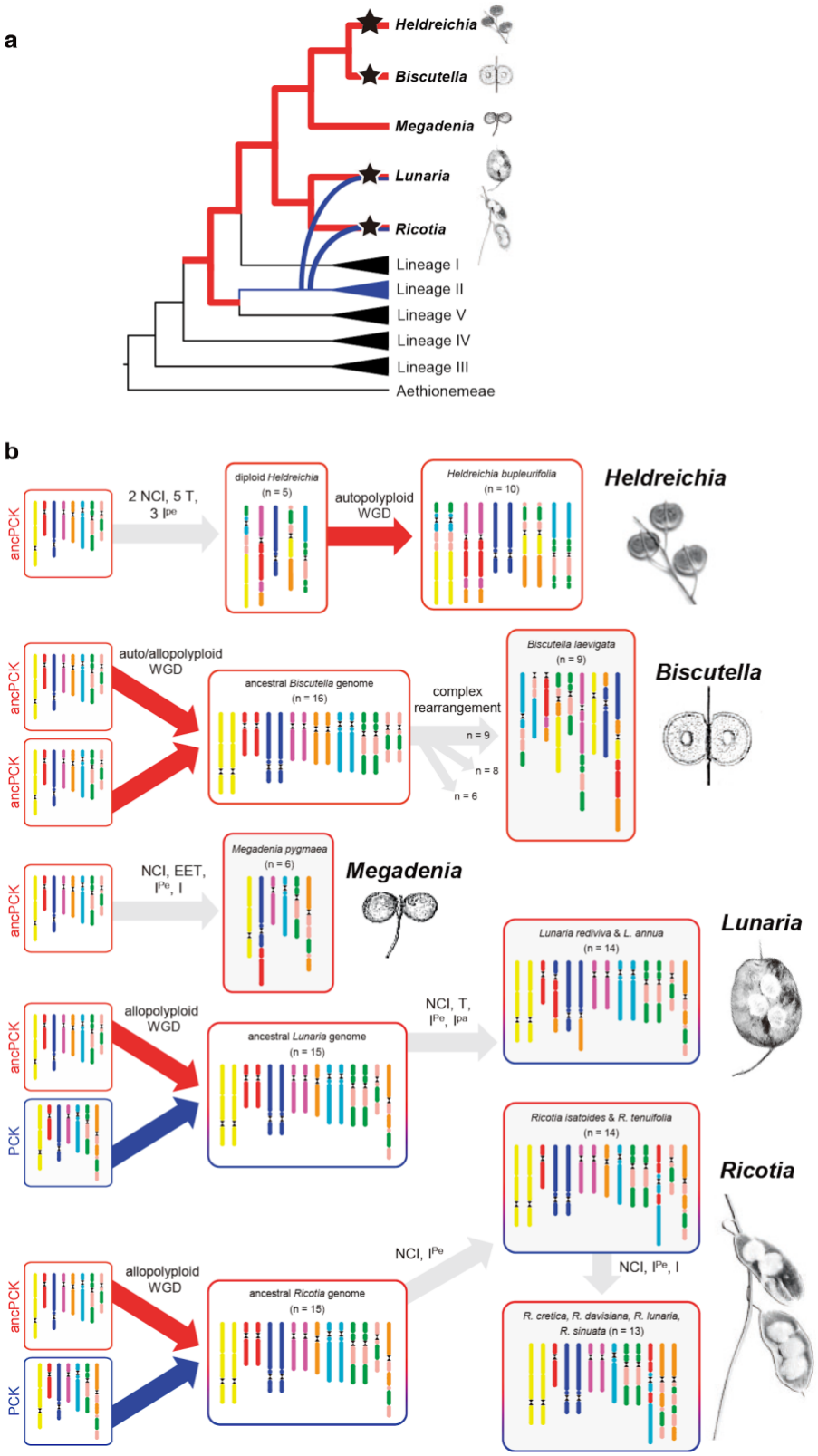


Figure 8 Proposed origin and evolution of the Biscutelleae diploid–polyploid genome complex by Guo et al. (2021).

(**a**) A simplified phylogenetic scheme showing the position of the Biscutelleae in the Brassicaceae family (based on Nikolov et al. 2019). Red- and blue-labeled branches indicate

evolutionary trajectory of ancestral ancPCK (*n* = 8) and PCK (*n* = 7) genomes, respectively. Star symbols indicate the genus-specific WGDs.

(**b**) Reconstructed origin and genome evolution for individual genera of Biscutelleae. All extant Biscutelleae genomes have descended from the ancestral ancPCK-like genome (red contours and arrows). Its divergence led to the origin of ancestral diploid genomes of *Heldreichia* (*n* = 5) and *Megadenia* (*n* = 6). The extant *Heldreichia* genome originated via autopolyploidization. An ancestral *Biscutella* genome (*n* = 16) was formed by hybridization between two ancPCK-like genomes. Allotetraploid genomes of *Lunaria* and *Ricotia* (*n* = 15) originated through recurrent hybridizations between ancPCK (*n* = 8) and PCK genome (*n* = 7; blue contours and arrows). WGDs in *Biscutella*, *Lunaria*, and *Ricotia* were followed by genus- and species-specific descending dysploidies mediated by nested chromosome insertions (NCI) and end-to-end translocations (EET). Nondysploid rearrangements included translocations (T), as well as paracentric (Ipa) and pericentric (Ipe) inversions.

2. Aims

The main goal of this work was to improve our understanding of the evolutionary mechanisms and constraints underlying chromosomal dysploidy in angiosperms by integrating next-generation sequencing (NGS), cytogenomics and in-depth bioinformatic analyses.

The specific objectives were as follows:

1. Generate high-quality, chromosome-scale genome assemblies and annotations for nine crucifer species, including *Heliophila variabilis* (*n* = 11; Heliophileae) and eight *Biscutella* species (Biscutelleae): *B. lyrata* (*n* = 6), *B. auriculata* (*n* = 8), *B. baetica* (*n* = 8), *B. didyma* (*n* = 8), *B. laevigata* subsp. *austriaca* (*n* = 9), *B. laevigata* subsp. *varia* (*n* = 9), *B. frutescens* (*n* = 9), and *B. prealpina* (*n* = 9).

2. Characterize the proposed WGD events and reconstruct the genome structure of the mesopolyploid genomes to elucidate the timing, mode and consequences of WGDs.

3. Infer chromosome rearrangement (CR) pathways leading to the present-day genomes, quantify both dysploid and non-dysploid CRs in the studied species, and assess whether the type and number of CRs reflect convergent or divergent evolutionary patterns across lineages.

4. Identify chromosome breakage hotspots and evaluate their potential association with repetitive sequences.

5. Analyze patterns of gene retention and fractionation between subgenomes and explore their functional implications.

6. Investigate the impact of post-polyploid diploidization on three-dimensional (3D) genome architecture, with a particular focus on potential changes in chromatin compartments and the organization of topologically associated domains (TADs).

3. Materials and Methods

3.1 Plant material, library preparation and sequencing

The following accessions were used: *H. variabilis* (Namaqualand, Springbok, Goegap Nature Reserve, 29°40′17″S, 17°59′55″ E), *B. laevigata* subsp. *varia* (V12-4; Germany, Beuron), *B. laevigata* subsp. *austriaca* (Jord.) Mach.-Laur. (A2Schnee 3B; Austria, Schneealpe Altenberg), *B. prealpina* Raffaelli & Baldoin (RCBO_NC17; Italy, Recoaro Terme), *B. frutescens* Coss. (PI 650129, USDA collection; Spain), *B. auriculata* L. (PI 650127, USDA collection; Spain, Ames), *B. didyma* L. (LBN-00579, LBN seed bank; Lebanon), *B. baetica* Boiss. & Reut. (Gaucín; Spain, Gaucín), and *B. lyrata* L. (Cádiz; Spain, Cádiz). Plants were grown from seed and cultivated under standard conditions in growth chambers (21/18°C, 16/8 h of light/dark) or a greenhouse (22/19°C, 16/8 h of light/dark).

Illumina: Genomic DNA was isolated from young leaf tissue using the NucleoSpin Plant II kit (Macherey-Nagel). Illumina sequencing libraries were prepared using the TruSeq Nano DNA HT Sample preparation kit following the manufacturer's recommendations. Genome sequencing was performed using the Illumina HiSeq Xten platform (Illumina; San Diego, CA, USA).

Nanopore and PacBio HiFi: Long-read sequencing processes followed standard protocols of Oxford Nanopore Technologies (ONT; Oxford, UK) and PacBio HiFi technology (PacBio; San Diego, CA, USA). DNA libraries with an average fragment length of >20 kb were constructed for ONT sequencing and sequenced on the PromethION platform. The PacBio circular consensus sequencing library for *B. lyrata* was additionally produced and sequenced on one SMRT cell of the PacBio Sequel II system.

Hi-C and Omni-C: Hi-C libraries were prepared from leaf samples using the Proximo Hi-C Kit according to the manufacturer's protocol at Phase Genomics (Seattle, USA). The Hi-C libraries were sequenced on Illumina HiSeq X-Ten instruments to generate 150-base paired-end reads. The Hi-C library for *B. lyrata* was generated using the Omni-C Proximity Ligation Assay (Dovetail Genomics, Scotts Valley, CA) following the manufacturer's "Non-Mammalian Samples Protocol version 1.2B". The library was then sequenced on the Illumina HiSeq X Ten platform.

IsoSeq and RNAseq: Total RNA was extracted from leaf, flower, and root tissues using the Quick-RNA Miniprep Kit (Zymo Research). The quality and quantity of the extracted RNA were assessed using NanoDrop 2000c Spectrophotometer (Thermo Scientific), Qubit 4 Fluorometer (Thermo Scientific), and Fragment Analyzer (Agilent). Total RNAs from the tissues of leaves and flowers were mixed equally, along with individual RNA from root tissues, for long-read (PacBio Iso-Seq) sequencing technology to generate transcriptome data. The Iso-Seq cDNA libraries were constructed according to the PacBio standard protocol and sequenced on the PacBio sequel II platform. The short-reads RNA-seq data for *B. baetica* and *B. lyrata* were downloaded from (Guo et al., 2021).

3.2 Genome size measurement by flow cytometry and *K*-mer frequency

Holoploid genome size was estimated by flow cytometry. The young intact leaf, ~1 cm in length, were prepared according to Doležel et al. (2007). The samples were stained using a solution containing propidium iodide + RNAase IIA, both at final concentrations of 50 µg/ml, for 5 min at room temperature and analysed using a CyFlow cytometer Partec equipped with a 532 nm diodepumped solid-state laser Cobolt Samba. A fluorescence intensity of 5,000 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). Three different samples for each species measured on three consecutive days was used for genome size estimation. Genome sizes were further confirmed by *K*-mer frequency (*K* = 21) analysis with the findGSE (v1.94; Sun et al., 2018), after counting 21-mers with Jellyfish (Marçais and Kingsford, 2011).

3.3 Genome assembly, scaffolding, and quality assessment

We employed two strategies for de novo genome assembly: an ONT-based strategy for *H. variabilis*, *B. auriculata*, *B. baetica*, *B. didyma*, and all *n* = 9 *Biscutella* species (Parisod et al., 2025), and a PacBio-based strategy for *B. lyrata*. Initial assemblies were generated from Nanopore long reads using NextDenovo (v2.2; https://github.com/Nextomics/NextDenovo) with its standard pipeline, and the resulting contigs were polished in two iterative rounds with both long and short reads

using NextPolish (v1.1.0; Hu et al., 2020). For *B. lyrata*, haplotype-resolved assemblies were generated using hifiasm (v0.19.5; Cheng et al., 2021) in hybrid mode by integrating PacBio HiFi reads, ONT reads (>= 25kb), and Hi-C reads. The program Khaper (Zhang et al., 2021) was used to select primary contigs and filter redundant sequences from the initial assemblies of the highly heterozygous genomes *B. auriculata* and *B. baetica*. Completeness of genome assemblies was evaluated using BUSCO (v3.0.1; Simão et al., 2015) with embryophyta_odb10 database (1,614 total BUSCOs). For assemblies generated using the ONT-based strategy, Hi-C reads for each genome were aligned to the corresponding contigs using Juicer (v1.5.7; Durand et al., 2016). The 3D-DNA pipeline (v180922; Dudchenko et al., 2017) was used to correct potential mistakes and to order, orient and scaffold the sequences. For *B. lyrata*, the Omni-C data was used to scaffold the genome assembly by YaHS (v1.1; Zhou et al., 2023) with default parameters. The linkage results were manually curated to correct misjoins and misassemblies based on visualization using JuiceBox (v1.1.08; Robinson et al., 2018). The chloroplast genome was assembled based on Illumina short reads using the GetOrganelle toolkit (v1.7.7; Jin et al., 2020).

3.4 Gene prediction and functional annotation

Protein-coding genes were predicted using an evidence-based annotation workflow that integrated multiple sources of evidence. For transcriptome-based predictions, RNAseq data were mapped using Hisat2 (v2.1.0; Kim et al., 2015) and subsequently assembled into transcripts by StringTie (v2.1.4; Pertea et al., 2015). IsoSeq datasets for each species were aligned to the genome assemblies using GMAP (v2018-07-04; Wu and Watanabe, 2005). All transcripts from RNAseq and IsoSeq were merged with StringTie into a pool of candidate transcripts. TransDecoder (v5.5.0; http://transdecoder.github.io) identified potential coding regions in the resulting transcripts. Additionally, two rounds of PASA (v.2.3.3; Haas et al., 2003) were conducted to refine gene models by identifying untranslated regions and isoforms, using transcripts generated by genome-guided Trinity (v2.11.0; Haas et al., 2013) assemblies. *De novo* gene predictions were generated using AUGUSTUS (v3.4.0; Stanke et al., 2006), with species-specific AUGUSTUS gene models trained using GeneMark-ET (v4.0; Lomsadze et al., 2014). This model leveraged RNA-seq and IsoSeq evidence in two iterative rounds of predictions to refine parameters. For annotation of homologs, protein sequences from *A. thaliana*, *Cadarmine*

hirsuta, *Eutrema salsugineum*, *Thlaspi arvense*, and *B. rapa* were aligned to each target genomes to identify the homologous genes using GenomeThreader (v1.7.1; Gremme et al., 2005). Finally, all gene predictions were integrated into a final gene model set using EVidenceModeler (v1.1.1; Haas et al., 2008) after removing pseudogenes and non-coding genes using a custom Python script. FeatureCounts (Liao et al., 2014) was used to extract the mapped reads for each gene, allowing for the calculation of transcripts per million (TPM) values as the expression level of the genes. Functional assignments for the predicted protein-coding genes were performed with BLAST (v2.12.0+; Altschul et al., 1990) to align coding sequences against public protein databases, including NCBI non-redundant protein (Pruitt et al., 2007), SwissProt (Bairoch and Boeckmann, 1992), and InterProScan (Hunter et al., 2009). Gene Ontology (GO) terms for each gene were provided by InterProScan.

3.5 Identification of syntenic genes and fragments

Syntenic gene pairs between *Heliophila*, *Biscutella* and the 22 conserved GBs (Schranz et al., 2006; Lysak et al., 2016) were identified using SynOrths (Cheng et al., 2012), which integrates sequence similarity and the syntenic context of flanking genes. Homologous gene pairs were first detected by BLASTP embedded in SynOrths, with only best hits or gene pairs showing an E-value < 1×10^{-20} retained. A gene pair was considered as syntenic if at least 20% of the 20 flanking genes on either side of the query gene were homologous to those within a 100-gene window surrounding the subject gene. Syntenic gene pairs that were continuously distributed along the target genomes and 22 GBs were considered as ancestral fragments inherited from the progenitors. Due to local structural variations and potential genome assembly errors, local syntenic gene pairs may not be distributed immediately adjacent to other syntenic genes. Thus, adjacent syntenic gene pairs were merged into a single syntenic fragment if they were separated by fewer than 50 intervening genes or a physical distance of <300 kb. In addition, tandem repeat arrays were also identified using SynOrths. Each tandem repeat array consisted of continuously distributed homologous genes and was not allowed to be interrupted by more than six non-homologous genes.

3.6 Repetitive element annotation

The Extensive de novo TE Annotator (EDTA; v1.8.3) (Ou et al., 2019) was utilized to annotate TEs in each species with the following parameters: "--species others --step all --anno 1". Within EDTA, LTRharvest (Ellinghaus et al., 2008), LTR_FINDER_parallel (Ou and Jiang, 2019) and LTR_retriever (Ou and Jiang, 2018) were used for identification of long terminal repeat retrotransposons (LTR-RTs). Tandem repeats were identified using TRASH (Wlodzimierz et al., 2023). The ribosomal DNA (rDNA) sequences were predicted with Barrnap (v0.9; https://github.com/tseemann/barrnap) using the Eukaryota database. For precise classification of LTR-RTs at the lineage level, TEsorter (v1.4.6; Zhang et al., 2022) was employed. SoloLTRs were identified using the soloLTRseeker pipeline (https://github.com/estpr/soloLTRseeker).

The coordinates of syntenic genes at the ends of GBs flanking gene-poor regions were used to delineate the boundaries of pericentromeric regions. DNA compression was generated using the context-tree weighting (CTW) function of the BCT package in R (https://www.rdocumentation.org/packages/BCT/versions/1.2) to further refine centromere localization. Tandem repeats that were significantly enriched in pericentromeric regions and overlapped with the troughs of the CTW values were identified as centromeric satellite DNA (satDNA) candidates, and those located at the ends of chromosomes were identified as (sub)telomeric satDNA candidates. The screened candidates were further validated through FISH experiments.

Insertion ages of LTR-RTs were estimated using the method proposed by SanMiguel et al. (1998), which compared the 5'- and 3'-LTRs of each full-length element. Nucleotide substitutions per site (K) between LTR pairs were calculated using Kimura's two-parameter model (Kimura, 1980). Following Koch et al. assumptions (Koch et al., 2000), we employed the mutation rate (r) of 1.5×10^{-8} substitutions per year per synonymous site and calculated the insertion times (T) of LTR-RTs with the formula $T = K/2r$.

3.7 FISH experiment and comparative chromosome painting

The mitotic and meiotic (pachytene) chromosome spreads were prepared from young anthers. Oligoprobes (60-bp in length) were used to visualize the identified tandem repeats on chromosomes. The most conserved regions within the consensus sequences

of monomers were selected, with a preference for regions with low GC content (30–50%) and minimal self-annealing. For chromosomal localization of conserved GBs, *A. thaliana* BAC clones were assembled to represent the 22 GBs of the ancestral crucifer karyotype (Lysak et al., 2016). DNA probes were labeled with biotin-dUTP, digoxigenin-dUTP, or Cy3-dUTP by nick translation as described by Mandáková and Lysak (2016). Labeled BAC DNAs were pooled, precipitated, and resuspended in 20 µl of hybridization mixture (50% formamide and 10% dextran sulfate in 2×SSC) per slide. Labeled 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 for 16 to 72 hours. Post-hybridization washing was performed in 20% formamide in 2×SSC at 42°C. Fluorescence signals were analyzed with an Axioimager Z2 epifluorescence microscope (Zeiss) and CoolCube CCD camera (MetaSystems).

3.8 Phylogenetic inference and divergence time calibration

To infer the ancestral origin of the identified syntenic fragments, we constructed subgenome-aware gene and species trees using maximum likelihood (ML) and coalescent-based approaches. Orthologous gene groups were identified among *H. variabili*s, eight *Biscutella* species, and 15 additional Brassicaceae species (*Aethionema arabicum*, *Euclidium syriacum*, *Arabis alpina*, *Pugionium cornutum*, *E. salsugineum*, *Schrenkiella parvula*, *B. rapa*, *C. hirsuta*, *Boechera stricta*, *Arabidopsis halleri*, *A. thaliana, Lunaria rediviva*, *Ricotia lunaria*, *Megadenia pygmaea*, and *Heldreichia bupleurifolia*). For each gene group, coding sequences were aligned using MAFFT (v7.427; Katoh et al., 2002) and trimmed using TrimAL (v1.4; Capella-Gutiérrez et al., 2009). ML phylogenies were infered using IQ-TREE (v1.6.11; Nguyen et al., 2015) with default parameters. Coalescent-based species trees were constructed using ASTRAL-Pro (v1.1.3), which allows species-tree inference in the presence of paralogy (Zhang et al., 2020). To further estimates divergence times, single-copy ortholog groups identified using OrthoFinder (v2.5.4, Emms and Kelly, 2019), were used to construct a ML phylogenetic tree with IQ-TREE, followed by time calibration using three methods: r8s (v1.81; Sanderson, 2003), PATHd8 (v1.0; Britton et al., 2007) and RelTime method in MEGA X (Kumar et al., 2018). The divergence at 20.6 Mya between crucifer Lineage I and Lineage II (http://www.timetree.org/) was used as the calibration node.

3.9 Subgenome phasing and validation

Using the 22 GBs as the reference, four and two genomic copies of each diploid chromosome in the octoploid ancestor of *Heliophila* and the tetraploid ancestor of *Biscutella* were identified, respectively. To reconstruct the subgenomes, three key rules were applied: (i) no overlapping or redundant regions were allowed within each reconstructed chromosome of the subgenomes; (ii) each ancestral chromosome adhered to the gene density distribution pattern between the subgenomes; and (iii) the phylogenetic relationship of the syntenic gene fragments in the 22 GBs was consistent with the subgenomic branches. Based on these rules, syntenic fragments were categorized into four groups in *Heliophila* (the least fractionated subgenome – sub #1, and the more fractionated subgenomes – sub #2, sub #3, sub #4) and two groups in *Biscutella* (the less fractionated subgenome – LF, and the more fractionated subgenomes – MF). To validate the subgenome phasing, hierarchical clustering of subgenome-specific repetitive DNA sequences and principal component analysis were performed using SubPhaser (v1.2; Jia et al., 2022).

3.10 *K*a/*K*s analysis

We performed self-to-self BLASTP alignment for *B. rapa*, *P. cornutum*, *T. arvense*, *E. salsugineum*, *C. hirsuta*, *A. alpina*, *L. rediviva*, *R. lunaria*, *M. pygmaea*, *H. bupleurifolia*, and *A. thaliana*, respectively, and selected the best hits among the homologous gene pairs with identity > = 90%. Paralogues from the *Heliophila* and eight *Biscutella* species, along with the homologues from other species were used for the non-synonymous (*K*a) and synonymous (*K*s) rate calculations. Each pair of protein sequences was aligned by MAFFT and pairwise nucleotide sequence alignments were generated by transforming protein alignments into codon alignments with ParaAT (v2.0; Zhang et al., 2012). The *K*s values and *K*a/*K*s ratios were calculated based on the Nei–Gojobori method implemented in KaKs_Calculator (v2.0; Wang et al., 2010).

3.11 Reconstruction of the ancestral karyotypes

We utilized both top-down and bottom-up strategies to reconstruct the ancestral karyotypes. The top-down strategy involved comparing GB associations in target species with established ancestral karyotypes of the Brassicaceae (ACK, ancPCK, and PCK)

(Schranz et al., 2006; Mandáková and Lysak, 2008; Geiser et al., 2016; Lysak et al., 2016). Conserved GB associations shared among multiple species were considered to be inherited from their ancestral genome. For *Biscutella* genomes, we also utilized the WGDI tool (Sun et al., 2022) following a bottom-up strategy, which is applicable to species with unknown ancestral karyotypes, as applied in Lamiales and Buxales (Wang et al., 2022; Hou et al., 2025). Synteny maps generated by WGDI among target species enabled the inference of ancestral karyotypes at various evolutionary nodes based on phylogeny. Intact chromosomes with continuous synteny were initially identified as ancestral chromosomes. Chromosomal breaks or fusions shared by multiple species, as well as those unique to particular species, were systematically characterized. We traced the origins of these breaks or fusions for each species hierarchically, which allowed us to reconstruct the karyotype for each node sequentially, progressing from the youngest to the oldest. To classify CR types, we employed a graph-based approach in which ancestral protochromosomes were color-coded and mapped onto the chromosomes of extant genomes (target chromosomes): NCIs were characterized by the presence of the same ancestral protochromosome at both ends of a given target chromosome, with one or more different protochromosomes inserted between them. EETs were identified when two ancestral protochromosomes were found to be joined at their terminal ends.
RTs were inferred when adjacent segments from two different ancestral protochromosomes were exchanged and detected on at least two distinct target chromosomes. Unbalanced RTs were characterized by the unidirectional fragmentation of an ancestral protochromosome, with its segments redistributed across multiple target chromosomes without reciprocal exchange. Inversions were recognized when one or more fragments from an ancestral protochromosome were oriented in the opposite direction relative to adjacent collinear fragments on the same target chromosome. Shifts were defined by noncontiguous arrangements of fragments from the same protochromosome along a single target chromosome, regardless of their orientation.
The CR rate was calculated as the total number of CRs divided by the elapsed time between any two nodes.

3.12 Gene family classification and functional enrichment

The genes that had two or more gene copies in four subgenomes of *H. variabilis* were considered as multiple-copy genes. The remaining genes were classified as singletons.

We also classified orthologous gene families across eight *Biscutella* species into core, softcore, and private categories. Genes with two copies in both subgenomes of *Biscutella* species were considered as duplicates, and the remaining genes were classified as singletons. Thus, five types of gene sets, i.e., core duplicates shared in all eight *Biscutella* species, core singletons shared in all eight *Biscutella* species, softcore duplicates and singletons shared in late-diverging clade comprised by all $n = 9$ species or in early-diverging species comprised by *B. auriculata*, *B. didyma*, *B. baetica*, and *B. lyrata*, and private singletons specific to a certain species, were further subjected to GO functional enrichment analysis. The GO terms and pathways of gene enrichment were identified by the clusterProfiler package (v3.14.3; Yu et al., 2012). GO enrichments were estimated using one-sided Fisher's exact tests, and an adjusted P-value < 0.05 was set as the cutoff criterion for the significance of the gene enrichment.

The Pfam protein domains identified by InterproScan were used to examine the conserved domains of multiple-copy gene families in *Heliophila* and core duplicate gene families in *Biscutella*. We assumed that gene duplicates could either split functions (subfunctionalization) or generate a new function (neofunctionalization; Birchler and Yang, 2022). For each gene family, if both copies retained identical protein domain(s), we considered that the domain(s) inherited from the ancestral gene, and if one copy lost the conserved domain or diverged into a new domain distinct from the other, we considered that the gene undergone sub/neofunctionalization.

3.13 3D genome analysis in *Biscutella* genomes

Hi-C sequencing reads were mapped to their corresponding reference genomes using BWA-MEM (v0.7.17; Li & Durbin, 2009) with the parameters "-A1 -B4 -E50 -L0". Hi-C contact matrices were generated at multiple resolutions (10, 25, 50, and 100 kb) using the hicBuildMatrix option in HiCExplorer (v3.7.2; Ramírez et al., 2018) with the parameters "--restrictionSequence GATC --danglingSequence GATC --binSize 10000 20000 50000 100000". Matrix correction was performed using the hicCorrectMatrix option, with filtering thresholds ("--filterThreshold -1.5 5") determined as described by Ramírez et al. (2018). The contact matrix plots for each chromosome were generated using hicPlotMatrix, and the 50-kb resolution matrices were selected for visualization and downstream analyses. A/B compartments were identified using the hicPCA option

in HiCExplorer, which computes Pearson correlation matrices from the raw contact matrices (50-kb resolution), followed by calculation of covariance matrices. The eigenvectors of the covariance matrices were used to assign genomic regions to compartments. Manual examination of compartment assignments for each chromosome was performed to determine the final eigenvector orientation, where negative values corresponded to B compartments (gene-poor regions) and positive values corresponded to A compartments (gene-rich regions). Based on genome-wide compartment assignments, we examined whether orthologous gene pairs between different species and paralogous gene pairs within species shared identical or distinct compartment assignments. TAD-like structures were identified from Hi-C data using a separation score–based approach implemented in the hicFindTADs option of HiCExplorer, with the parameters "--thresholdComparisons 0.01 --delta 0.01 --correctForMultipleTesting fdr". The overlap between TADs, A/B compartments, and subgenomes was calculated, and TADs were able to be categorized as entirely within the A compartment (or LF subgenome), entirely within the B compartment (or MF subgenome), or spanning both compartments (or subgenomes). Conserved TADs between species were identified following the method of Shen et al. (2023), in which two TADs from different species were considered conserved if syntenic gene pairs were found within 100 kb on both sides of their boundaries.

4. Results

4.1 The meso-octoploid *Heliophila variabilis* genome sheds a new light on the impact of polyploidization and diploidization on the diversity of the Cape flora

Huang, Y., Guo, X., Zhang, K., Mandáková, T., Cheng, F., & Lysak, M. A., 2023. *The Plant Journal*, 116(2), 446-466. DOI: https://doi.org/10.1111/tpj.16383

The genome of *H. variabilis* (*n* = 11; 1C = 334 Mb; **Figure 9a**) was sequenced using a combination of Illumina short reads, Oxford Nanopore long reads, and Hi-C technologies. We obtained a total of 77 contigs with a total length of 300.5 Mb and a contig N50 size of 11.32 Mb. These contigs were assigned to 11 pseudochromosomes ranging in size from 23.85 to 32.26 Mb based on the contact information of the Hi-C data. The resulting chromosome-level assembly achieved 98.7% completeness of BUSCO score. A total of 32,351 protein-coding genes were annotated, and transposable elements accounted for 131.83 Mb (43.87%) of the genome.

Synteny analysis combined with CCP experiment revealed quadruplicated GBs, supporting an allooctoploid origin of the *H. variabilis* genome (**Figure 9b**). To further test whether hybridization contributed to the origin of the octoploid genome, we performed ML and coalescent-based analyses separately for each of the syntenic fragments, enabling inference of the putative progenitor lineages. At least two phylogenetic origins of the ancestors of *H. variabilis* were retrieved: one that formed a clade with *Megadenia*, which was sister to the present-day crucifer supertribe Camelinodae (Lineage I), and the other being sister to the Brassicodae (Lineage II) (**Figures 10a** and **b**). These results suggest that the ancestral *Heliophila* genome had a hybrid origin involving progenitor genomes from divergent crucifer lineages. Subgenome-specific gene fractionation was evident, allowing us to classify the genome into four distinct subgenomes. Subgenomes #1 and #2 retained the most genes (10,230 and 7,839, respectively), whereas subgenomes #3 and #4 were more fractionated (3,965 and 2,518 genes, respectively) (**Figure 10c**).

Comparative analyses of GB associations with known ancestral karyotypes (ACK, ancPCK, and PCK) revealed that the octoploid *Heliophila* genome originated from four

distinct progenitor genomes: two ancPCK-like genomes (n = 8; sub #2 and sub #4), which were successively sister to the tribe Biscutelleae and Lineage I, a PCK-like genome (n = 7; sub #3), and an unknown genome (n = 7; sub #1), which both were closely related to the ancestral 7-chromosome genomes of Lineage II. The inferred ancestral octoploid genome had at most 30 chromosome pairs (n = 7 + 8 + 7 + 8; i.e., 2n = 8x = ~60), which were reduced to n = 11 in *H. variabilis* by extensive descending dysploidy (**Figure 11**). The allo-octoploid genome originated through hybridization between genomes with lower ploidies (2x, 4x, or 6x), either by hybridization between two tetraploid genomes (4x × 4x → 8x) or via a hexaploid bridge (4x → 6x → 8x).

TE dynamics in *H. variabilis* also provided evidence for its hybrid origin. While the total length of TEs was associated with the level of gene fractionation (i.e., sub #1 > sub #2 > sub #3 > sub #4; **Figure 12a**), the proportions of different repeat categories were comparable among subgenomes, except for the most fractionated sub #4 (**Figures 12b** and **c**). Analysis of TE insertion times revealed distinct evolutionary histories for two major TE classes: DNA transposons showed a large peak indicating transpositional bursts at ~18 Mya (**Figure 12d**), whereas the LTR-RTs exhibited two apparent peaks that were mainly caused by the activities of Gypsy retroelements at ~8 Mya and ~22 Mya (**Figure 12e**). Genomic repeats contain abundant phylogenetic signals for subgenome phasing. Specifically, the numbers of 21-mer sequences shared between subgenomes were much lower than that of subgenome-specific ones. Nevertheless, in 13 of the 22 GBs, clustering of shared 21-mer sequences grouped the less fractionated subgenomes (#1 and #2) separately from the more fractionated subgenomes (#3 and #4) (**Figure 13**). Although the observed pattern could be due to subgenome-biased sequence loss, it could alternatively support the two-step origin of the allo-octoploid *Heliophila* genome by hybridization between two allotetraploid progenitor genomes, one comprising subgenomes #1 and #2 and the second combining subgenomes #3 and #4.

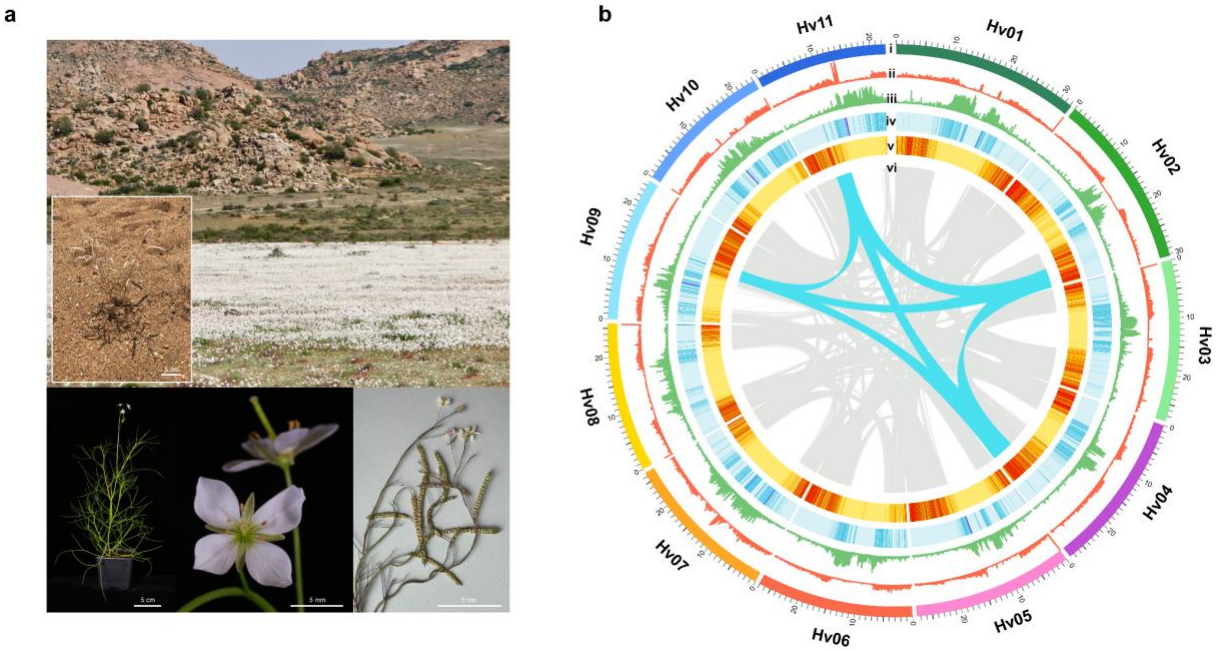


Figure 9 Morphological characteristics of *Heliophila variabilis* and its genome structure.
(**a**) Spring flower carpet in Goegap Nature Reserve (Namaqualand, South Africa) dominated by *H. variabilis*, and a close-up of a fruiting plant at the same site. Morphological characters of *H. variabilis* are shown below.
(**b**) Collinearity within the *H. variabilis* genome. The circles from outside to inside show (i) 11 chromosomes (named Hv01-Hv11); (ii) densities of DNA transposons; (iii) densities of LTR retrotransposons; (iv) densities of all TEs; (v) densities of genes; (vi) gene synteny between 11 chromosomes of *H. variabilis* is shown by the grey lines, the blue lines represent a set of four duplicated genomic fragments, corresponding to genomic block A in the ancestral genome.

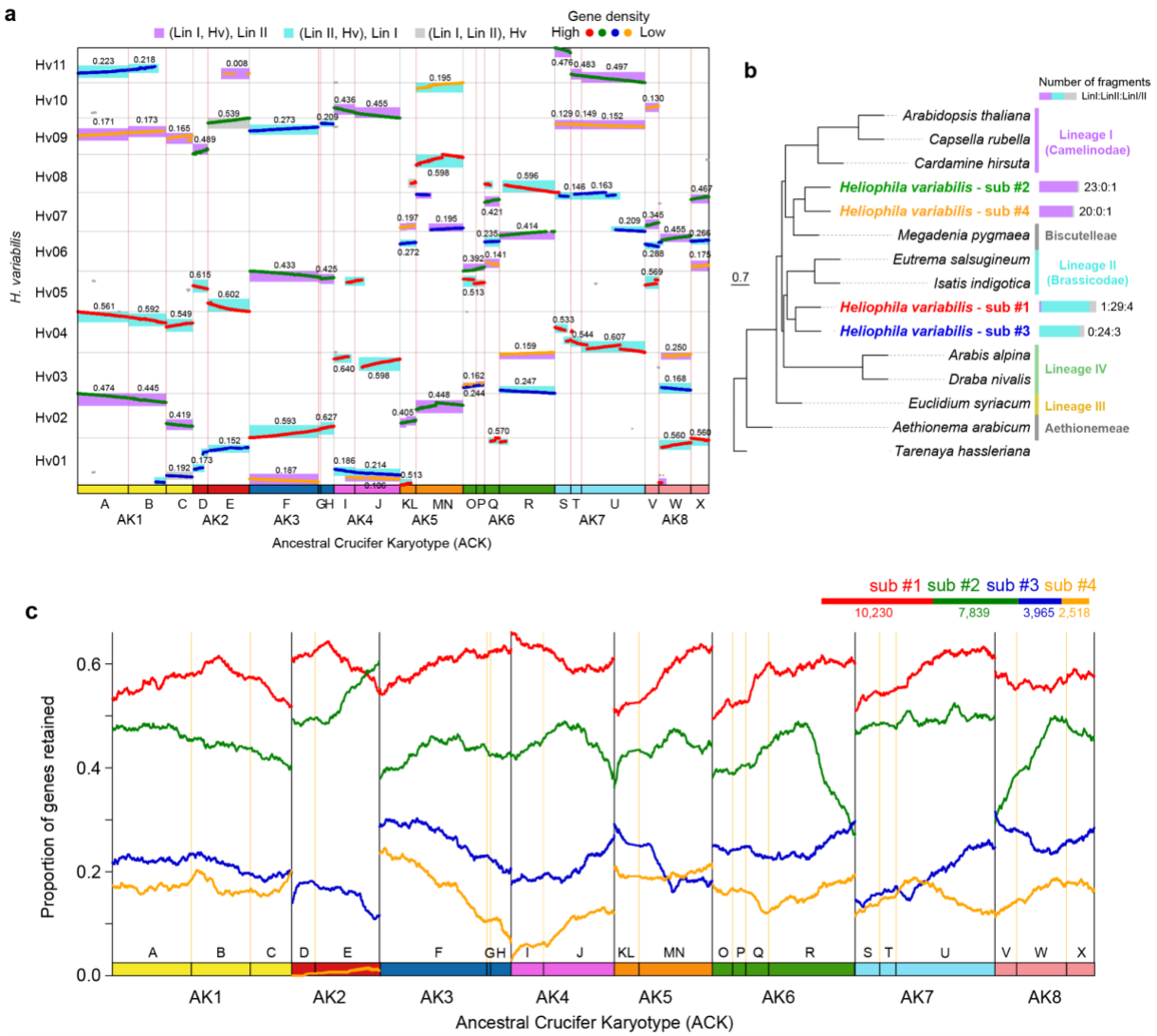


Figure 10 Evolution of four subgenomes in the octoploid *H. variabilis* genome.

(**a**) The synteny map showing syntenic relationships between 11 chromosomes of *H. variabilis* and 22 genomic blocks of the Ancestral Crucifer Karyotype (ACK). Syntenic fragments are labeled according to their average gene density, numerically expressed above each fragment. The background coloring of the syntenic segments corresponds to the phylogenetic placements in (**b**).

(**b**) Coalescent-based phylogenetic tree including the four subgenomes of *H. variabilis* and 11 other crucifer genomes representing the major Brassicaceae lineages. Statistics on the phylogenetic topology of the 106 syntenic fragments shared between *H. variabilis* and ACK (see (**a**)) are shown to the right of the stacked bar charts.

(**c**) The density of syntenic genes in four subgenomes of *H. variabilis* compared to the ACK genome. The total number of genes contained in each subgenome is labeled on the upper right.

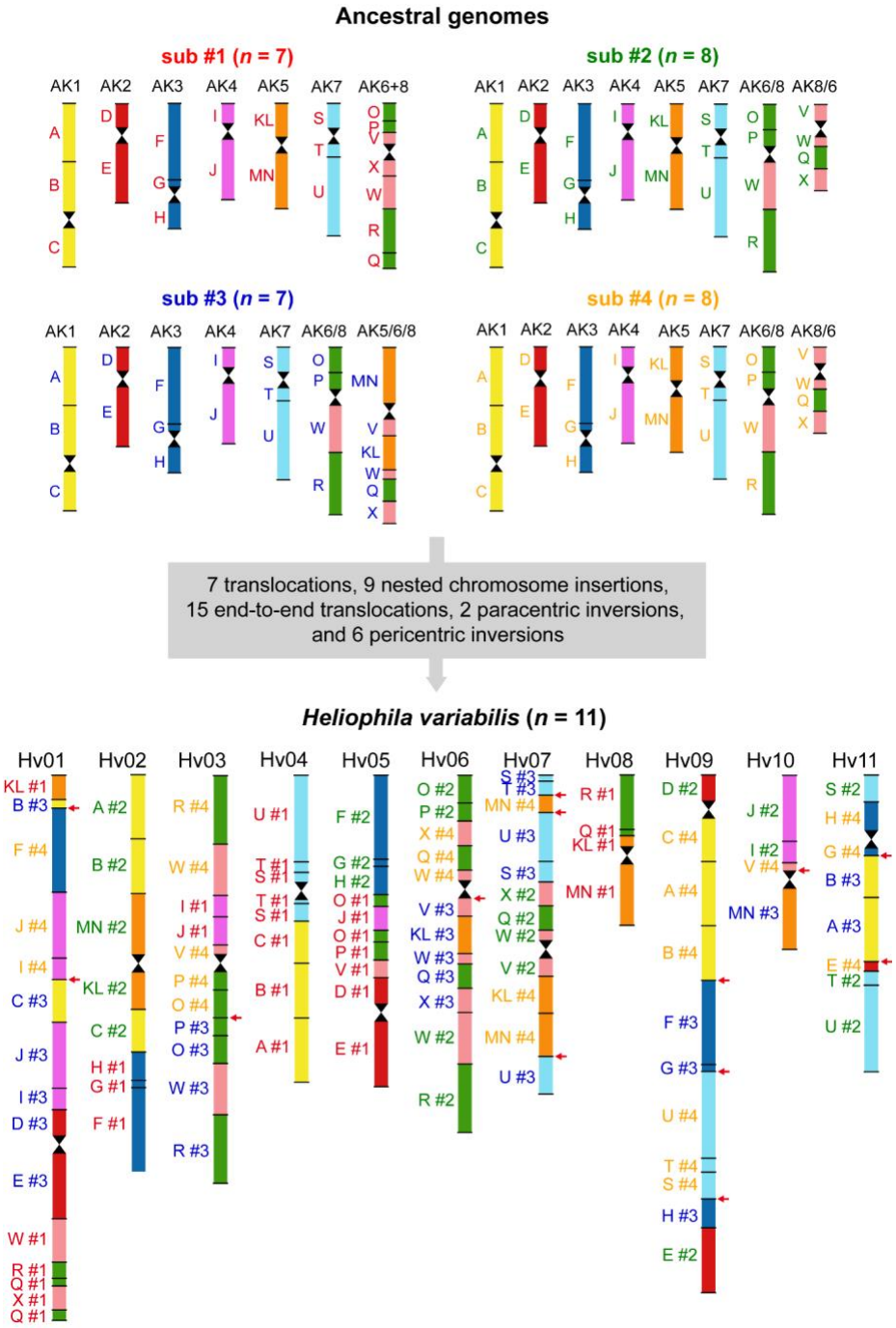


Figure 11 Four reconstructed ancestral subgenomes of *H. variabilis* and their structure within the extant *H. variabilis* genome.

Three subgenomes resemble previously inferred ancestral Brassicaceae genomes, namely ancPCK (sub #2 and #4, both *n* = 8) and PCK (sub #3, *n* = 7), whereas sub #1 (*n* = 7) could not be attributed to any of the previously inferred ancestral genomes. The color coding and capital letters (A to X) correspond to 22 genomic blocks shared among the four parental genomes and the 11 chromosomes of *H. variabilis*. Sub #3 and sub #4 formed the highest number of GB associations (red arrows).

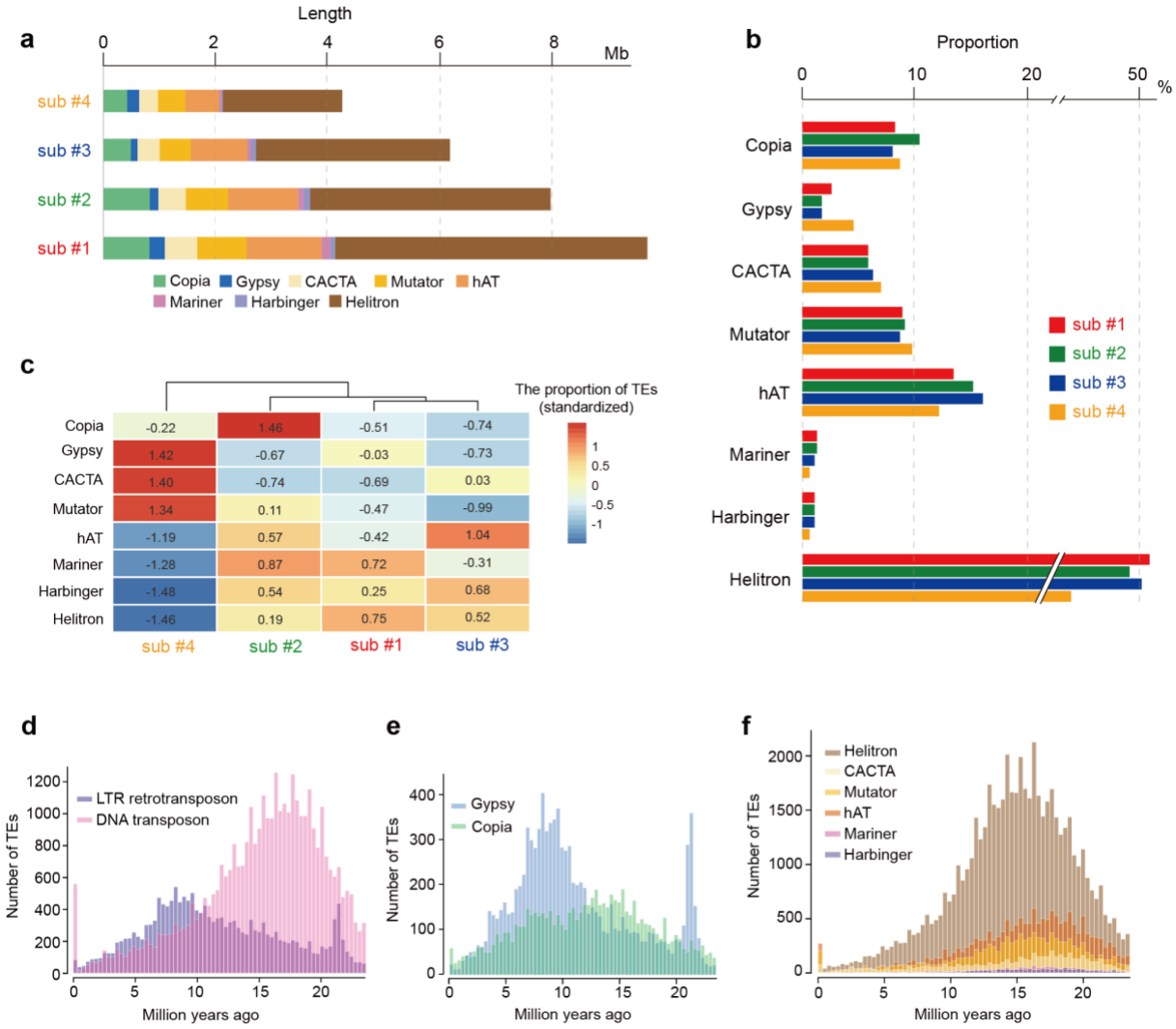


Figure 12 TE element content and distribution of TE insertion times of *H. variabilis.*

(**a**) Length of TEs in the four subgenomes.

(**b**) Proportion of eight TE types in the four subgenomes.

(**c**) Standardized proportions of eight TE types in the four subgenomes.

(**d**) Insertion time distribution of LTR retrotransposons and DNA transposons.

(**e**) Insertion time distribution of two classes of LTR retrotransposons: Gypsy and Copia.

(**f**) Insertion time distribution of six DNA transposon elements.

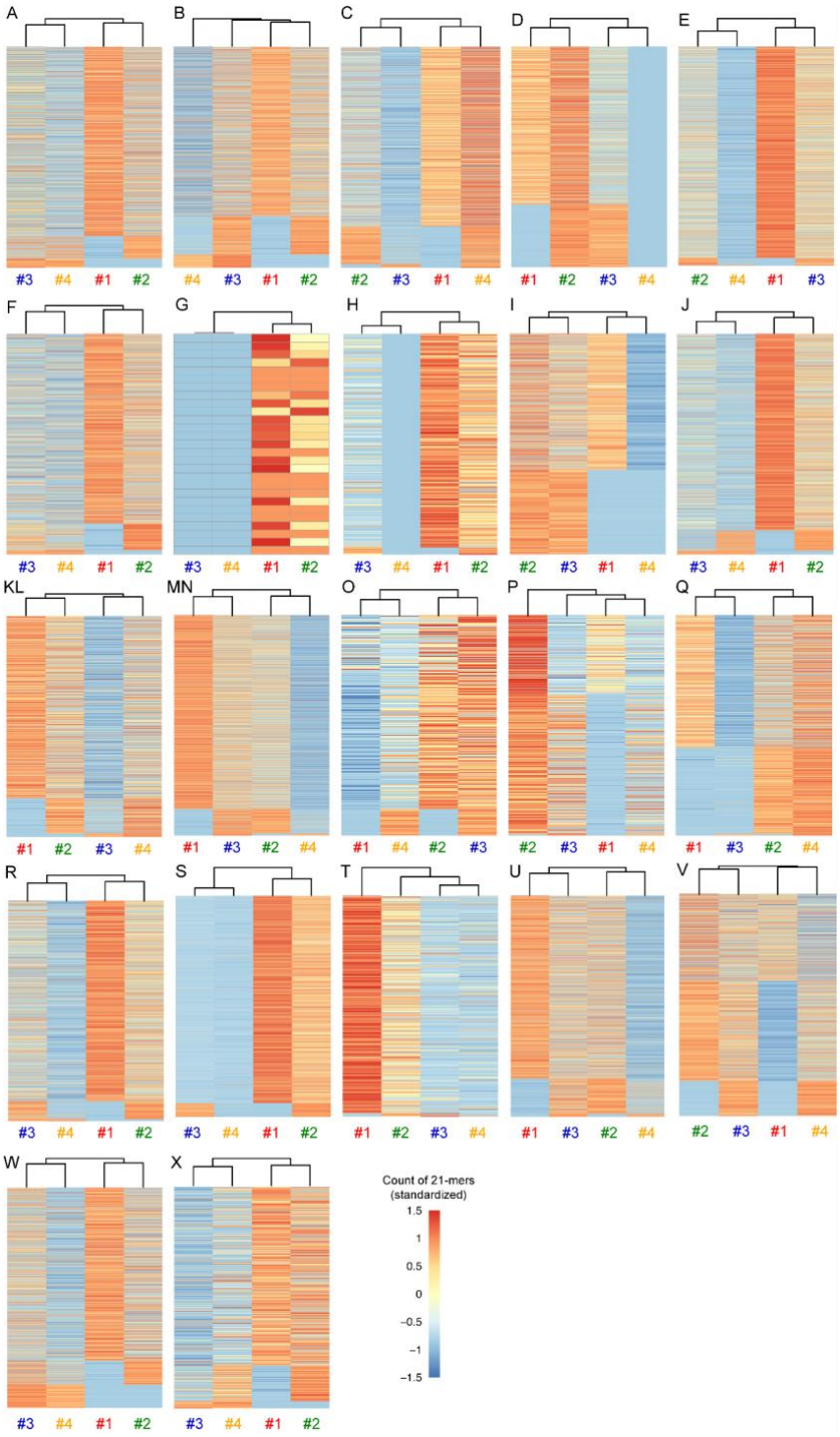


Figure 13 Heat map of 21-mer enrichment shared by four subgenomes of *H. variabilis* in 22 GBs (A to X).

4.2 Post-polyploid chromosomal diploidization in plants is affected by clade divergence and constrained by shared genomic features

Huang, Y., Poretti, M., Mandáková, T., Pouch, M., Guo, X., Perez-Roman, E., Crespo, M. B., Grob, S., Bousios, A., Parisod C., & Lysak, M. A., 2025. *In revision, Nature Communications*. DOI: https://doi.org/10.21203/rs.3.rs-6440714/v1

We generated chromosome-scale genome assemblies using Illumina short-read (c. 83× genome coverage), ONT long-read (c. 62×), and Hi-C sequencing for eight *Biscutella* species, including *B. laevigata* subsp. *varia* (*n* = 9; 1C = 814 Mb; referred to as *B. varia*), *B. laevigata* subsp. *austriaca* (*n* = 9; 904 Mb; referred to as *B. austriac*a), *B. prealpina* (*n* = 9; 916 Mb), *B. frutescens* (*n* = 9; 936 Mb), *B. auriculata* (*n* = 8; 686 Mb), *B. didyma* (*n* = 8; 980 Mb), *B. baetica* (*n* = 8; 1.1 Gb), and *B. lyrata* (*n* = 6; 806 Mb). The PacBio HiFi sequencing (c. 28×) was additionally performed for *B. lyrata*. Eight genome assemblies with total lengths of c. 515 to 1,168 Mb were obtained following either ONT- or PacBio-based assembly strategy (**Figure 14a**). The 32,292 to 50,236 high-confidence protein-coding genes and 268 to 764 Mb of TEs were annotated (**Figure 14b**). Synteny analyses further identified 22,492 to 31,397 gene pairs in the eight *Biscutella* genomes, corresponding to 22 ancestral GBs, which confirms a meso-tetraploid WGD predating diversification of the genus (**Figure 14c**).

Subgenomes were identified based on unequal gene density of two genomic copies (**Figure 14d**), designating as the less fractionated (LF) and more fractionated (MF) subgenomes, and further validated by phylogenetic and *K*-mer sequence similarity analyses. Phylogenetic analyses across 22 GBs revealed topological discordance among homeologous gene trees and suggested a reticulate origin of the clade. Despite this complexity, subgenome-aware coalescent and ML trees consistently placed the LF subgenome with the neotetraploid *Heldreichia*, while the MF subgenome formed a monophyletic clade outside this sister group (**Figure 14e**). Molecular clock analyses and *Ks* distribution dated the hybridization event forming *Biscutella* to ~11–13 Mya (**Figure 14f** and **g**). Lineage splitting within the genus revealed a temporal gap between early-diverging species (*n* = 8/6) and a recent radiation of the *n* = 9 species (~3.3 Mya) (**Figure 14f**).

Ancestral genome reconstruction using both top-down (based on GB associations) and bottom-up (phylogeny-guided WGDI inference) approaches revealed that the mesotetraploid *Biscutella* genome was derived from hybridization between two progenitors with distinct karyotypes. The LF subgenome originated from an ancPCK-like genome (n = 8) with a paracentric inversion event, and that the MF subgenome underwent additional NCI and EET events, which reduced its chromosome number to six (n = 6), while preserving more ancestral GB associations during subsequent speciation events (**Figure 15a**). The post-polyploid evolution of the allotetraploid ancestor (n = 14) to the modern species (n = 9, 8 and 6) was reconstructed by integrating multiple CR events (**Figure 15b**). Only one NCI event is shared by all eight species, whereas the remaining CRs are clade- or species-specific. The early-diverging species experienced 2 to 14 independent, private CR events, while the late-diverging species had only one to three (**Figure 15b**). Unbalanced reciprocal translocations were the predominant CR type across all *Biscutella* genomes (5 to 9 events per species) and, when accompanied by stable centromere inactivation in dicentric chromosomes, contributed to the reduction of the ancestral chromosome number (n = 14) to present-day chromosome numbers.

LTR-RTs constitute the major repetitive fraction in *Biscutella* genomes (34.3% – 54.3% of TEs), yet only ~18.7% of them remain intact, reflecting extensive removal following recent bursts of activity. Most LTR-RTs are also subject to rapid deletion, as we identified a substantial number (4,869 – 8,401) of recombination remnants defined as soloLTRs. Analysis of different LTR-RT lineages revealed that the Athila and CRM dominated the landscape and collectively accounted for 65% of intact LTRs and 55.9% of soloLTRs (**Figure 16a**). Early-diverging *Biscutella* species exhibited significantly higher solo:intact LTR-RT ratios for Gypsy-derived lineages (e.g., Athila, CRM, Retand, Tekay) than late-diverging species (**Figure 16b**). Nevertheless, no specific pattern of soloLTR loss was observed between subgenomes (**Figure 16c**), indicating that the deletion of LTR-RTs is not linked to gene fractionation.

Hi-C analysis revealed that *Biscutella* genomes are partitioned into distinct A/B compartments, with A compartments enriched near telomeres and B compartments occupying gene-poor, TE-rich pericentromeric regions (**Figures 16d** and **16e**). Except

for the smallest *B. auriculata* genome, which contains more A compartments (50.6% of the genome), the remaining genomes had a slightly higher proportion of B compartments (~53%; **Figure 16f**). The examination of orthologous gene pair composition between any two species demonstrated the variability in the A/B compartment assignments. The proportion of orthologous gene pairs assigned to different chromatin compartments was significantly higher in the early-diverging species (14.63% – 30.42%) than the late-diverging species (3.51% – 7.26%; **Figure 16g**). Furthermore, the organization of TADs also underwent significant restructuring during post-polyploid diploidization. The late-diverging species had more and relatively shorter TADs (942 – 1,050 TADs, 0.65 – 0.72 Mb), whereas early-diverging species showed considerable variability in TAD number and length (502 – 1,039 TADs; 0.81 – 1.35 Mb), with the fewest but longest TADs in *B. lyrata* (502 TADs; 1.35 Mb average length; **Figures 16h** and **16i**). TAD conservation, assessed by comparing orthologous gene pairs at TAD boundaries, indicated that only ~9.49% of TADs were conserved between any two species, with a higher proportion of conserved TADs retained between late-diverging species (11.09% – 13.44%) than between early-diverging species (5.57% – 10.43%, **Figure 16j**).

Synteny comparisons of eight *Biscutella* genomes with their ancestral (sub)genomes revealed 14 breakpoints shared by multiple species, including eight inherited from ancestors at different phylogenetic nodes and six that arose recurrently across clades, designated as breakage hotspots (HOT regions; **Figure 17a**). In the ancestral *Biscutella* genome, nine HOT regions belonged to the LF subgenome (n = 8) and five to the MF subgenome (n = 6) (**Figure 17a**). Based on the inferred ancestral centromere positions, which we define as paleocentromeres (Yang et al., 2021; Guo et al., 2021), eight HOTs were associated with breaks at/near these regions, while six were located within ancestral chromosome arms (**Figure 17a**). In the extant *Biscutella* genomes, more than half of the rejoined junctions aligned with pericentromeric regions (**Figure 17b**). Analysis of the LTR-RT content in 100-kb regions on each side of rejoined junctions showed that more than 97% of these regions had higher LTR-RT content than the genome-wide average (**Figure 17b**), suggesting that interspersed stretches of similar TEs may serve as be substrates for NAHR and permit the continuous accumulation of TE insertions.

To assess the relationship between 3D genome organization and chromosome breakage, we examined A/B compartment assignments and TAD distribution within the HOT regions. We found that up to 68% of the breakpoint junctions showed inconsistent A/B compartment assignments and 64% matched TAD boundaries across *Biscutella* genomes (**Figure 17b**). For example, HOT9, a hotspot at/near the paleocentromere of ancestral chromosome AK8, resulted in the rearrangement of blocks MN and KL on arms of chromosome 1 and 4 in *B. lyrata*. The distal end of the MN block likely contains paleocentromeric remnants, characterized by gene depletion and CRM accumulation (**Figure 17c**). Despite being located within a chromosome arm, the rejoined junction at MN block remains in the B compartment, whereas the KL junction occupies the A compartment (**Figure 17c**), indicating an A/B compartment shift. More interestingly, the rejoined junction at the MN block colocalized with TAD boundaries (**Figure 17c**), suggesting a relationship between chromosome break and TAD architecture.

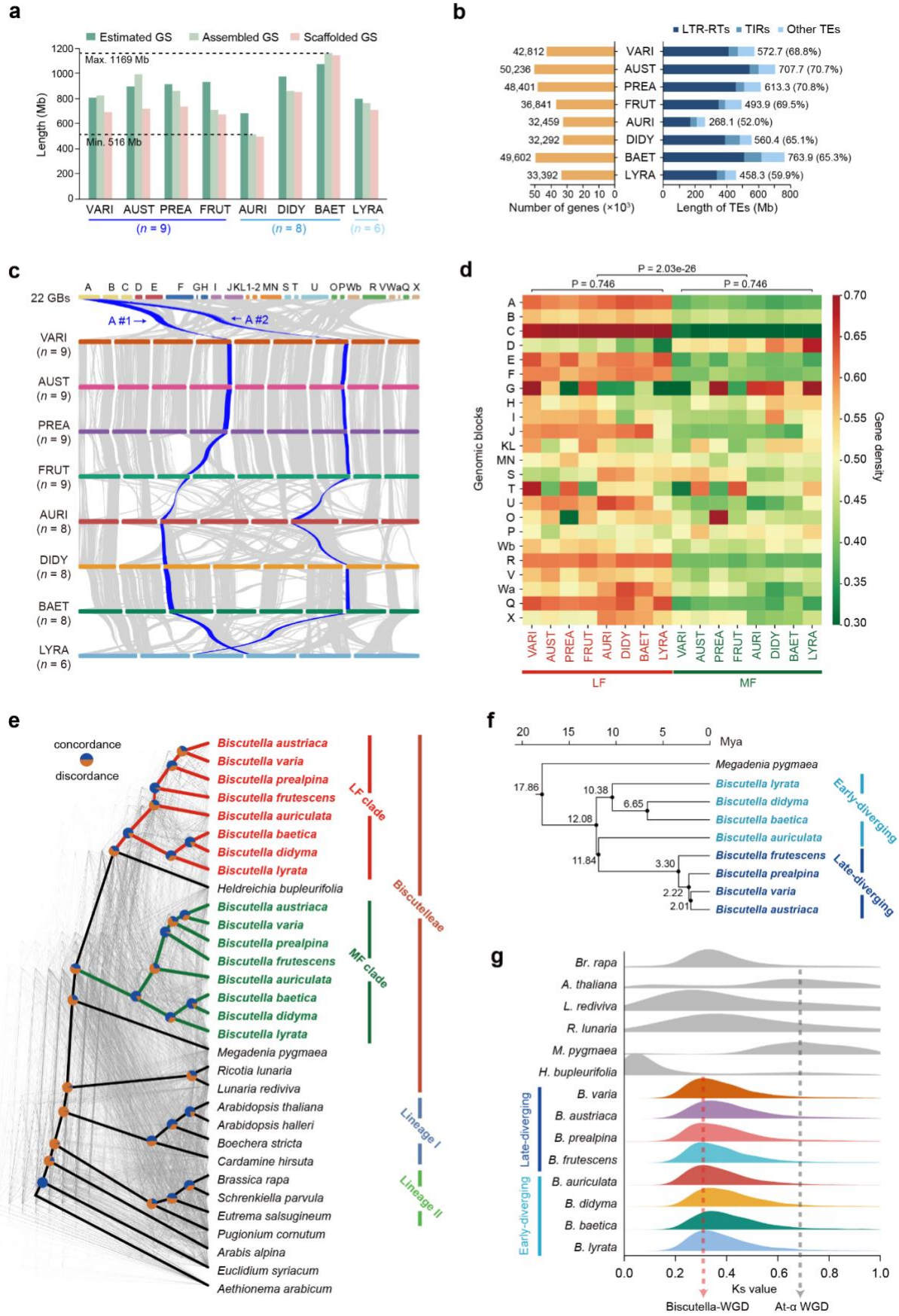

a
Estimated GS
Assembled GS
Scaffolded GS
Length (Mb)
Max. 1169 Mb
Min. 516 Mb
VARI AUST PREA FRUT AURI DIDY BAET LYRA
(n = 9)
(n = 8)
(n = 6)
b
LTR-RTs
TIRs
Other TEs
42,812
50,236
48,401
36,841
32,459
32,292
49,602
33,392
VARI
AUST
PREA
FRUT
AURI
DIDY
BAET
LYRA
572.7 (68.8%)
707.7 (70.7%)
613.3 (70.8%)
493.9 (69.5%)
268.1 (52.0%)
560.4 (65.1%)
763.9 (65.3%)
458.3 (59.9%)
Number of genes (×10³)
Length of TEs (Mb)
c
22 GBs
A #1
A #2
VARI (n = 9)
AUST (n = 9)
PREA (n = 9)
FRUT (n = 9)
AURI (n = 8)
DIDY (n = 8)
BAET (n = 8)
LYRA (n = 6)
d
P = 2.03e-26
P = 0.746
P = 0.746
Genomic blocks
Gene density
LF
MF
e
concordance
discordance
Biscutella austriaca
Biscutella varia
Biscutella prealpina
Biscutella frutescens
Biscutella auriculata
Biscutella baetica
Biscutella didyma
Biscutella lyrata
Heldreichia bupleurifolia
LF clade
MF clade
Biscutelleae
Lineage I
Lineage II
Megadenia pygmaea
Ricotia lunaria
Lunaria rediviva
Arabidopsis thaliana
Arabidopsis halleri
Boechera stricta
Cardamine hirsuta
Brassica rapa
Schrenkiella parvula
Eutrema salsugineum
Pugionium cornutum
Arabis alpina
Euclidium syriacum
Aethionema arabicum
f
Mya
Megadenia pygmaea
Biscutella lyrata
Biscutella didyma
Biscutella baetica
Biscutella auriculata
Biscutella frutescens
Biscutella prealpina
Biscutella varia
Biscutella austriaca
17.86
10.38
6.65
12.08
11.84
3.30
2.22
2.01
Early-diverging
Late-diverging
g
Br. rapa
A. thaliana
L. rediviva
R. lunaria
M. pygmaea
H. bupleurifolia
B. varia
B. austriaca
B. prealpina
B. frutescens
B. auriculata
B. didyma
B. baetica
B. lyrata
Late-diverging
Early-diverging
Ks value
Biscutella-WGD
At-α WGD

Figure 14 Genome size, repeat composition, chromosome collinearity, and subgenome phylogenomics in eight *Biscutella* species.

(**a**) Genome size (GS) variation. The estimated (flow cytometry-based; green), assembled (light green), and scaffolded (pink) genome size is shown for the eight *Biscutella* species. The assembled genome size ranged from c.516 to 1,169 Mb.

(**b**) Comparison of gene number and TE length. The number of annotated protein-coding genes (left), total TE length (right, Mb), and proportion of TEs relative to the entire genome (%) are shown.

(**c**) Syntenic relationships between the eight *Biscutella* genomes. A riparian plot illustrates macrosynteny across 22 ancestral genomic blocks (GBs A – X) and the eight assemblies, with the dark blue lines highlighting the two genomic copies of block A. Species acronyms are as follows: *B. laevigata* subsp. *varia* (VARI), *B. laevigata* subsp. *austriaca* (AUST), *B. prealpina* (PREA), *B. frutescens* (FRUT), *B. auriculata* (AURI), *B. didyma* (DIDY), *B. baetica* (BAET), and *B. lyrata* (LYRA).

(**d**) Syntenic gene density in two subgenomes. The heatmap shows the density of syntenic genes in the LF and MF subgenomes in eight *Biscutella* species compared to the 22 ancestral genomic blocks. The color gradient represents the average percentage of retained homeologous genes in the *Biscutella* subgenomes surrounding each ancestral gene. Here, 300 genes flanking each side of a given gene locus were analyzed, giving a total window size of 601 genes. A significant difference in gene density was found between the LF and MF subgenomes (Wilcoxon rank-sum test), whereas interspecific differences were not significant (Kruskal-Wallis test).

(**e**) Subgenome-aware phylogenetic analysis. A cloud tree shows concordance and discordance among 1,234 nuclear gene trees. The coalescent-based tree, which includes the two *Biscutella* subgenomes and 15 other crucifer genomes representing the major Brassicaceae lineages, is shown with thick black lines. Pie charts indicate the proportions of gene trees concordant or discordant with the species tree topology. The *Aethionema arabicum* genome was used as an outgroup.

(**f**) A simplified time-calibrated tree constructed using r8s. The inferred divergence times for each clade are labelled on the phylogenetic nodes. The *Biscutella* species can be divided into two groups based on their divergence times: late-diverging species (*B. varia*, *B. austriaca*, *B. prealpina*, and *B. frutescens*) and early-diverging species (*B. auriculata*, *B. didyma*, *B. baetica*, and *B. lyrata*).

(**g**) *K*s-value distribution. The distribution of *K*s-value was analyzed for homeologous gene pairs in eight *Biscutella* genomes and six other Brassicaceae species, including *Br. rapa*, *A. thaliana*, *L. rediviva*, *R. lunaria*, *M. pygmaea*, and *H. bupleurifolia*.

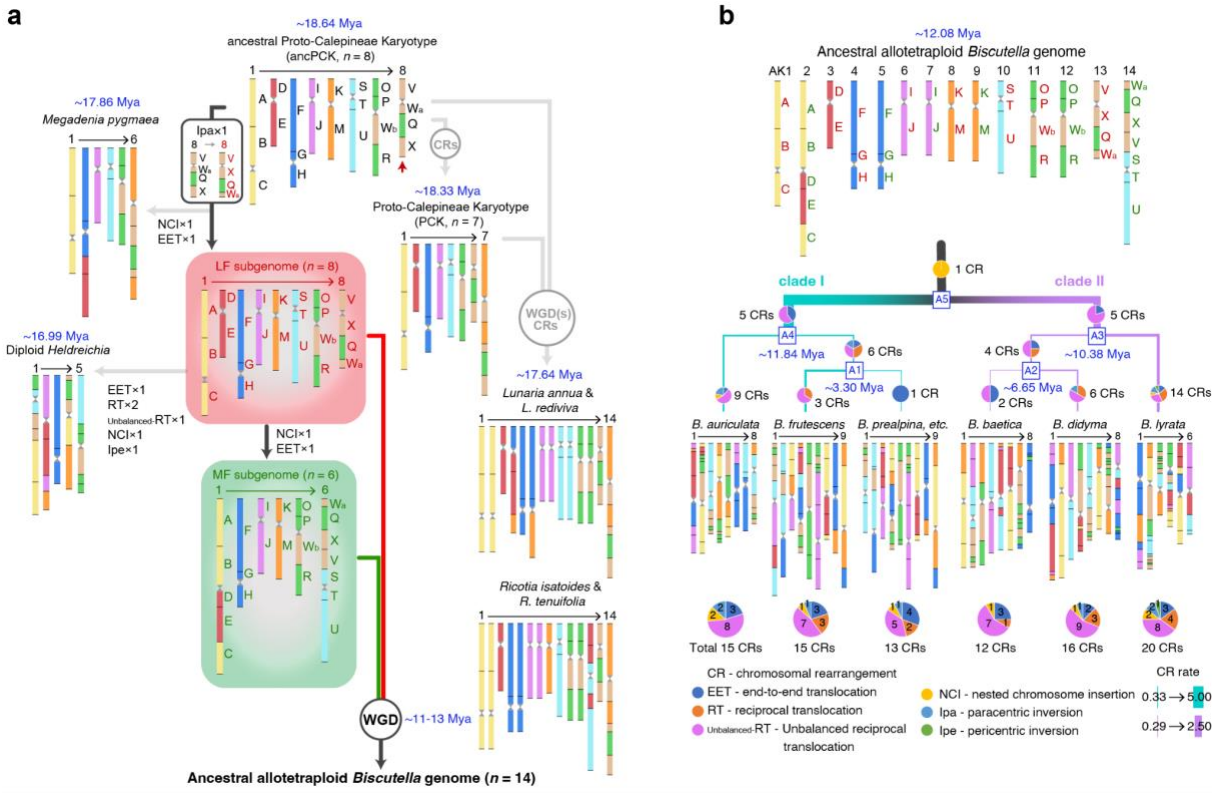


Figure 15 Karyotype evolution of *Bicutella* genomes.

(**a**) Reconstructed genome evolution in the tribe Biscutelleae. The earliest ancestor of Biscutelleae structurally resembled the ancPCK-like genome (*n* = 8). The ancPCK-like genome diversified into an 8-chromosome genome with a paracentric inversion on chromosome AK8/6 (V+Wa+Q+X → V+X+Q+Wa) and by descending dysploidy into PCK genome (*n* = 8 → *n* = 7). One or more likely two hybridization events between the ancPCK-like genome (*n* = 8) and PCK-like genome (*n* = 7) resulted in the ancestral tetraploid genome(s) of *Lunaria* and *Ricotia*, which were subsequently diploidized to 14 chromosomes (*n* = 8+7 → *n* = 14; Guo et al., 2021). The inversion ancPCK-like genome, structurally resembling the LF subgenome of *Biscutella* (*n* = 8), underwent three independent descending dysploidies forming (i) the *Megadenia pygmaea* genome (*n* = 8 → *n* = 6), (ii) the *Heldreichia bupleurifolia* genome (*n* = 8 → *n* = 5), and (iii) the MF subgenome of the tetraploid *Biscutella* ancestor (*n* = 8 → *n* = 6). The KL block is abbreviated to “K” and the MN block to “M”. EET: end-to-end translocation, RT: reciprocal translocation, unbalanced-RT: unbalanced reciprocal translocation, NCI: nested chromosome insertion, Ipa: paracentric inversion, Ipe: pericentric inversion, CR: chromosomal rearrangement.

(**b**) Post-polyploid genome evolution in *Biscutella*. The ancestral allotetraploid genome (*n* = 14) underwent three CRs and diverged into two clades: clade I (cyan lines) includes *B. auriculata* (*n* = 8) and the *n* = 9 species, clade II (purple lines) contains *B. baetica*, *B. didyma* (both *n* = 8) and *B. lyrata* (*n* = 6). The conserved genomes of the *n* = 9 species (*B. austriaca*, *B. prealpina* and *B. varia*) are only represented by *B. prealpina*. The ancestral karyotype at each evolutionary node (A1 to A5) was inferred using WGDI. The different types of CRs are shown as pie charts, with the total number of CRs per species indicated below. CR rate was calculated

as the total number of CRs divided by the elapsed time and indicated by line thickness (thin: low rate, thick: high rate).

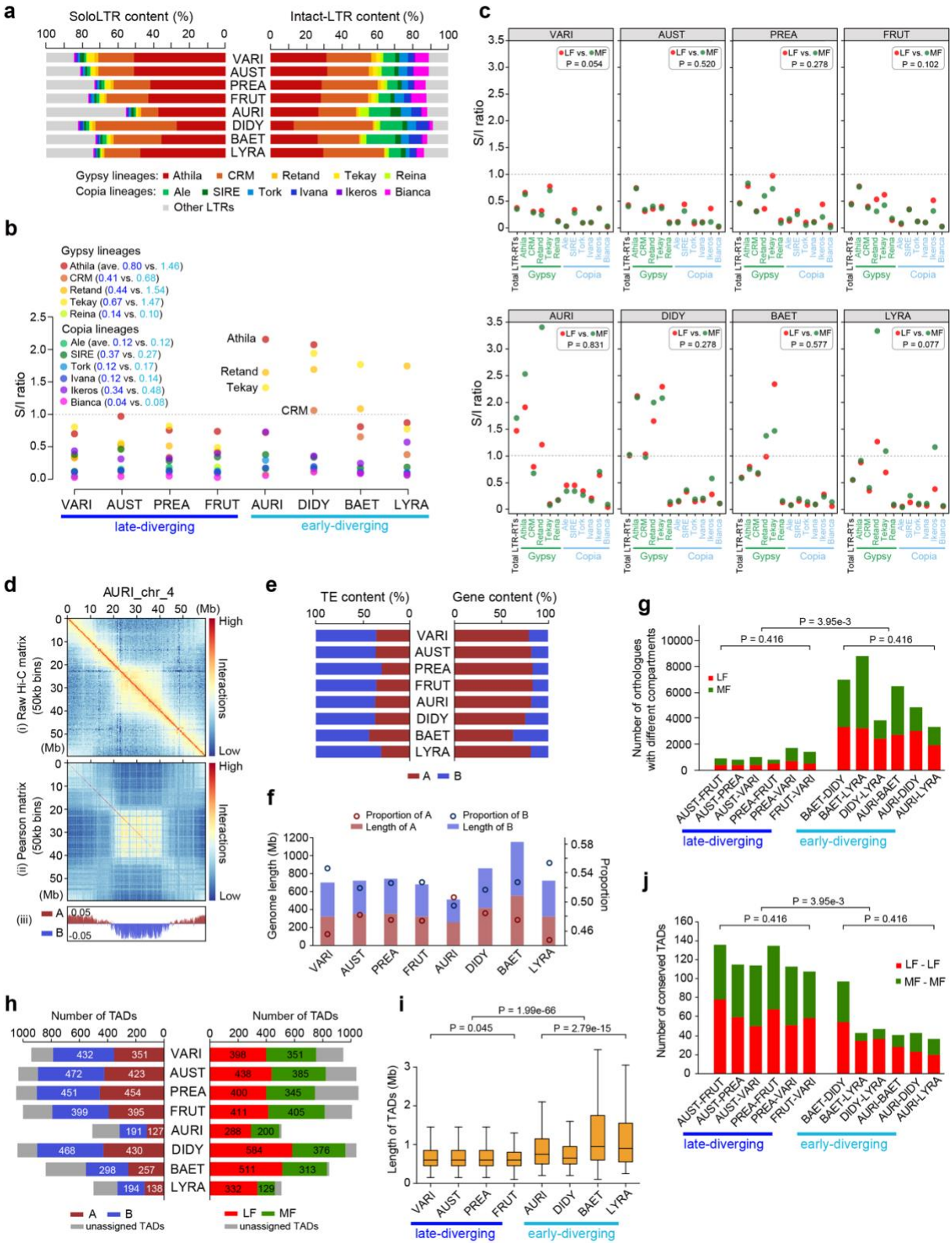


Figure 16 Evolution of LTR retrotransposons and 3D chromatin organization.

(**a**) Content of solo and intact LTRs. Shown are the relative proportions of soloLTRs (left) and intact copies of LTR-RTs (right) in five major Gypsy lineages and six major Copia lineages
(**b**) Solo:intact (S/I) LTR ratios. S/I ratios for 11 Gypsy or Copia lineages were compared in eight *Biscutella* species to assess the relative abundance of soloLTRs versus intact LTR-RTs. The average values of S/I LTR ratio for late-diverging (dark blue) and early-diverging (light blue) species are labelled.
(**c**) S/I LTR ratios in subgenomes. S/I ratios of total LTR-RTs (first column of Y-axis) and 11 Gypsy or Copia lineages (2nd to 12th column of Y-axis) were analyzed for the LF and MF subgenomes, respectively. No significant differences were observed between the two subgenomes of the eight species (Wilcoxon signed-rank test).
(**d**) Hi-C map of chromosome 4 in *B. auriculata*. The tracks from top to bottom represent: (i) the raw Hi-C interaction matrix at 50-kb resolution, (ii) the Pearson matrix generated from raw Hi-C interaction matrix, and (iii) the first principal component (PC1) derived from the PCA analysis of this matrix was used to define the A and B compartments. Positive PC1 values are shown in red, representing A compartments, and negative PC1 values are shown in blue and designated as B compartments.
(**e**) TE and gene content in the nuclear A/B compartments. The proportions of TEs (left) and genes (right) were measured in the A/B compartments as a percentage of the total number of TEs and genes assigned to each compartment.
(**f**) Length and proportion of nuclear A/B compartments. The total length (bars; Y-axis left) and proportion (lines; Y-axis right) of A/B compartments were calculated for each genome. The proportion represents the percentage of the genome assigned to one of the two compartments.
(**g**) Orthologous gene pairs with different compartments. The number of orthologous gene pairs between any two species with different compartments was analyzed, with statistical comparisons among species (Kruskal-Wallis test) and between late- and early-diverging groups (Wilcoxon rank-sum test).
(**h**) The total number of TADs in A/B compartments (left) and LF/MF subgenomes (right). The unassigned TADs include TADs with mixed compartments or subgenome regions, such that they cannot be fully assigned to either the A/B compartment or one of the two subgenomes.
(**i**) Length of the TADs. The distribution of TAD lengths in the eight *Biscutella* genomes was analyzed and compared using statistical tests as in (**g**).
(**j**) Conserved TADs between species. The number of conserved TADs between any two species was quantified and analyzed using statistical tests as in (**g**). Two TADs are considered conserved between species if an orthologous gene pair in two species were located within 100-kb on either side (200-kb in total) of the TAD boundaries.

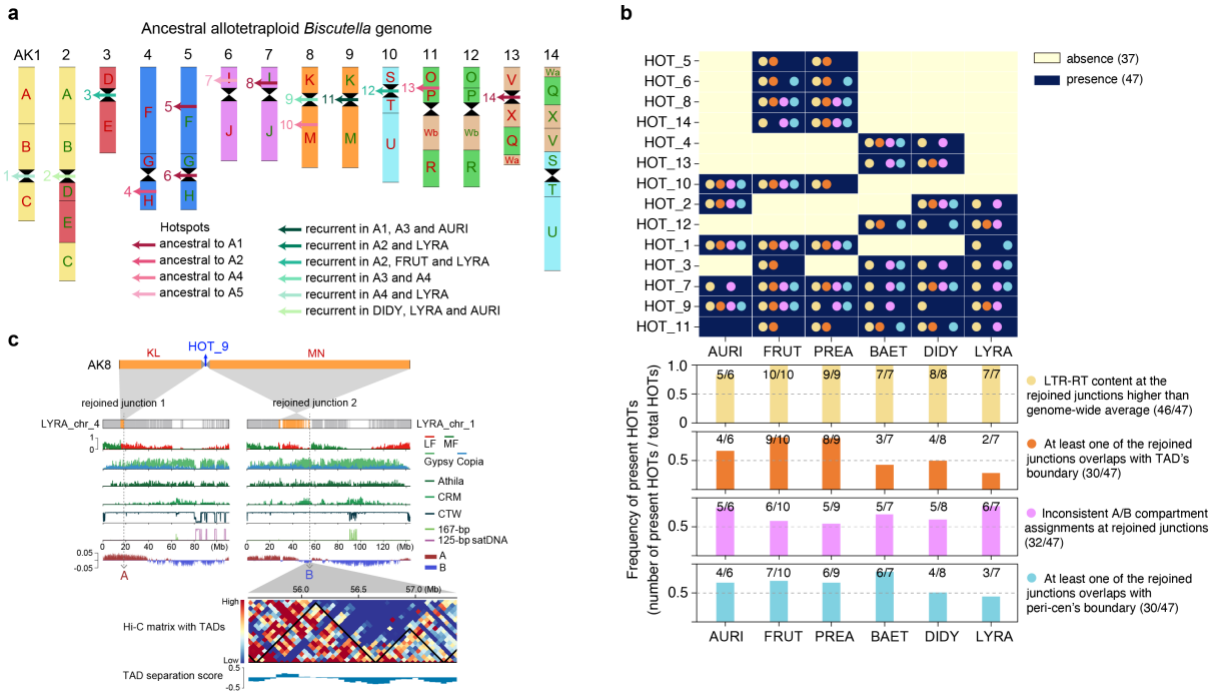


Figure 17 Position and characteristics of chromosome breakage hotspots.

(**a**) Fourteen chromosome breakage hotspots (HOT1 to HOT14) were identified in the ancestral allotetraploid *Biscutella* genome. The HOT region is defined as the flanking 100-kb region (200-kb in total) on each side of the neighbouring genes of two joined GBs. Eight hotspots are inherited from ancestors at different phylogenetic nodes (see **Figure 15b**; gradient pink arrows) and six that arose recurrently across clades (gradient green arrows).

(**b**) Presence and absence of the 14 HOTs in *Biscutella* genomes. The conserved genomes of the *n* = 9 species (*B. austriaca*, *B. prealpina* and *B. varia*) are only represented by *B. prealpina*. Three HOTs are shared by all species, and the remaining are shared by at least two species. A total of 47 rejoined junction pairs were detected across all analyzed genomes, with the number determined by counting each detected breakage event across species. The absence (37) and presence (47) of breaks are indicated by light yellow and dark blue squares. Of the 47 rejoined junctions, 46 had higher LTR-RT content than genome-wide average (yellow dots), 30 overlapped with TAD boundaries (orange dots), and 32 were associated with shifts in A/B compartment assignments (magenta dots), and 30 overlapped with pericentromeric region boundaries (blue dots).

(**c**) Origin and evolution of HOT9. HOT9, a breakage hotspot near the paleocentromere of ancestral chromosome AK8, underwent CRs that relocated blocks KL and MN to chromosomes 1 and 4, respectively, in *B. lyrata*. The tracks from top to bottom show: the ancestral chromosome AK8; syntenic relationship between AK8 and chromosomes 1 and 4 in *B. lyrata*; distributions of genes on chromosomes 1 and 4; gene density in LF (red) and MF (green) subgenomes; density of Ty3/Gypsy (light green) and Ty1/Copia (light blue) retrotransposons; density of Athila retroelements; density of CRM retroelements; CTW values; density of 125-bp (pink) and 167-bp (green) centromeric tandem repeats.; A/B compartment assignment; a

Hi-C matrix for the 55.0 – 57.5 Mb region, along with TADs and corresponding TAD separation scores.

5. Discussion

5.1 Extensive descending dysploidy shaped the meso-octoploid *Heliophila* genome over millions of years

While polyploidy is widespread in angiosperms, octoploid genomes remain exceptional, and among those, few have been sequenced and assembled. The ~300 Mb *H. variabilis* genome is the first octoploid crucifer genome to be sequenced at the chromosome level. Unlike the stability of chromosome number observed in other sequenced octoploid genomes, such as the allo-octoploid strawberries (*Fragaria* spp., $2n = 8x = 56$; Edger et al., 2019), octoploid sugarcane (*Saccharum officinarum*, $2n = 8x = 80$; *S. spontaneum*, $2n = 8x = 64$; Zhang et al., 2018), holocentric beak-sedge *Rhynchospora pubera* ($2n = 8x = 10$; Hofstatter et al., 2022), and the recent sequenced *Phyllanthus emblica* (Phyllanthaceae; $2n = 8x = 104$) and *Sauropus spatulifolius* (Phyllanthaceae; $2n = 8x = 103$) (Li et al., 2024), the 30 chromosome pairs in *Heliophila* have undergone a 2.7-fold reduction to only 11 chromosome pairs ($n = \sim30 \rightarrow n = 11$). The extensive descending dysploidy is likely a consequence of the older origin of the *Heliophila* octoploid genome. While strawberry, sugarcane, *R. pubera* and the octoploid Phyllanthaceae species originated within the past 5 million years (Zhang et al., 2018; Edger et al., 2019; Hofstatter et al., 2022; Li et al., 2024), the ancestral octoploid *Heliophila* genome originated before the divergence of the genus in the Miocene (at least 12 Mya; Walden et al., 2020; Dogan et al., 2021; Hendriks et al., 2022). Given the at least partially shared allopolyploid origin with the sister genus *Chamira* (Dogan et al., 2021), the hybridization-facilitated genome merger events could be dated earlier than the *Chamira*/*Heliophila* split (c. 18 Mya; Walden et al., 2020; Dogan et al., 2021; Hendriks et al., 2022), which is associated with the remnants of ancient TE expansion in the *H. variabilis* genome. Consequently, rediploidization of the meso-octoploid genome may have been a protracted process spanning more than 12 million years and accompanied by adaptive radiation and infrageneric cladogenesis, broadly associated with the establishment of a summer-dry climate in the Cape Floristic Region (CFR) in the mid-to-late Miocene (Verboom et al., 2009; Van Santen and Linder, 2020).

Patterns of gene fractionation in *H. variabilis* further support its allopolyploid origin. Two subgenomes (sub #1 and #2) exhibit less fractionation, while the remaining two

(sub #3 and #4) are more eroded, consistent with an allopolyploid origin involving hybridization between two tetraploid progenitors (*n* = 15 each). Both tetraploid parental genomes originated through distant intertribal hybridization, merging an unknown *n* = 7 genome (sub #1), two ancPCK-like genomes (*n* = 8; sub #2 and sub #4), and a PCK-like genome (*n* = 7; sub #3). While more ancestral ancPCK-like genomes were phylogenetically close to Lineage I (Camelinodae) and the tribe Biscutelleae (Heliophilodae), both 7-chromosomal diploid genomes belonged to Lineage II (Brassicodae). The inferred origin of the meso-octoploid *Heliophila* genome highlights the importance of distant hybridizations and WGDs for cladogenesis, adaptive radiation, and colonization of new habitats, including long-distance dispersals.

The geographical distribution of extant *Heliophila* genus and putative ancestral genomes supports a scenario in which the ancestral *Heliophila* genome originated in northern Africa, the Mediterranean or southwestern Asia and subsequently reached southern Africa—either by long-distance dispersal or stepwise migration through Miocene habitats such as wooded grasslands (Peppe et al., 2023). Once established in the CFR, diploidization processes may have facilitated the diversification of *Heliophila* species in newly invaded habitats in southern Africa. The CFR is one of the world's biodiversity hotspots with a high proportion of endemic taxa, particularly in plants (9000 species, 70% endemic; Mittermeier et al., 1998; Myers et al., 2000; Goldblatt and Manning, 2002). The assembled mesopolyploid genome of *H. variabilis* has shown that polyploidization–diploidization cycles may be more important for the diversity of the Cape flora than previously thought (Oberlander et al., 2016).

5.2 Chromosomal diploidization in *Biscutella* proceeded through independent descending dysploidy, LTR deletion and chromatin reorganization

Although the presence of multiple base chromosome numbers in *Biscutella* is not unusual among monophyletic angiosperm genera, the existence of several base numbers within the genus reflects repeated post-polyploid chromosomal reductions rather than a direct consequence of mesopolyploidy alone. In crucifers, polybasic clades often result from independent descending dysploidy events following polyploidization (Mandáková et al., 2017a, 2017b). The extant species of *Biscutella* descend from an allotetraploid

ancestor that arose 11–13 million years ago with *n* = 14, followed by independent chromosome number reductions to *n* = 9, 8, and 6 (Geiser et al., 2016; Guo et al., 2021; Beringer et al., 2024).

Despite differences in chromosome numbers, levels of diploidization, and divergence times, post-polyploid descending dysploidy in species with *n* = 8 and 9 involved 12 to 16 CRs, while *B. lyrata* (*n* = 6) experienced at least 20 CRs. The inferred CRs include comparable numbers of EETs, NCIs, RTs and unbalanced-RTs. Such structural changes resemble those mediating descending dysploidy in other eudicots (Feng et al., 2024; Sun et al., 2024). The early stages of cytological diploidization in *Biscutella* were marked by elevated TE activities and higher CR rates, resulting in extensive karyotypic restructuring, compared to the later stages characterized by lower CR rates. Thus, initial inter-subgenome homogenization through CRs and descending dysploidy associated with gene fractionation may progressively decelerate cytological diploidization.

Fourteen chromosomal breakage hotspots were identified across several *Biscutella* species. These HOTs showed enrichment in LTR-RTs, particularly Athila and CRM elements, and frequently colocalized with paleocentromeric regions, supporting regions with frequent DSBs and NAHR. In addition, many HOTs are also colocalized with boundaries of TADs, indicating that CRs are not randomly distributed with respect to higher-order chromatin (A/B) compartments and local TADs (Li et al., 2023). In contrast to the younger polyploids such as cotton and horseradish, which retain largely conserved TADs and A/B compartments (Wang et al., 2018; Shen et al., 2023), *Biscutella* shows substantial alteration of such chromatin features. Only a limited number of conserved domains were detected between orthologous chromosomes, reflecting the progressive erosion of 3D genome organization over extended evolutionary timescales.

6. Conclusions and Final Remarks

In my PhD research, I generated chromosomal-scale, high-quality genome assemblies and comprehensive annotations for selected crucifer species using a combination of ONT long reads, PacBio HiFi reads, Illumina short reads, Hi-C, and RNA-seq data. These genomic resources enabled a suite of downstream analyses, including subgenome phasing, phylogenetic inference, ancestral karyotype reconstruction, identification of CRs, repeatome profiling, and 3D genome architecture analysis. By integrating these genomic datasets, my research resolved the complex evolutionary histories of two crucifer clades: Heliophileae and Biscutelleae.

The chromosome-scale assembly of the *H. variabilis* genome reveals its complex allooctoploid origin from distant, intertribal hybridization among four distinct crucifer progenitors. Subsequent PPD involved extensive chromosome restructuring, resulting in a drastic reduction from an inferred ancestral karyotype of ~30 chromosomes to the present-day $n = 11$. TE dynamics in *H. variabilis* support its hybrid, allo-octoploid origin. Differences in gene number retention and subgenome-specific repeat content, along with clustering of shared *K*-mer sequences, distinguish the less fractionated subgenomes (#1 and #2) from the more fractionated ones (#3 and #4). These patterns are consistent with a two-step formation of the genome via hybridization between two allotetraploid progenitors.

Comparative genomic analyses of eight *Biscutella* species demonstrate that post-polyploid chromosomal diploidization proceeds through independent descending dysploid trajectories, with unbalanced-RTs predominating. Structural breakage hotspots—biased toward paleocentromeric regions of the less fractionated subgenome—frequently coincide with TAD boundaries and are enriched in LTR-RTs, implicating TE-rich, 3D structural features in recurrent chromosome breakage. Contrasting structural dynamics between early- and late-diverging clades underscore the influence of independent diploidization histories and phylogenetic divergence on the tempo and mode of chromosomal evolution.

Given the vast diversity and chromosomal variability among land plants, the broader comparative analysis of representative seed plant genomes spanning multiple families

revealed divergent, lineage-specific patterns of descending dysploidy. While further investigations across a wider range of seed plant lineages are needed to validate chromosomal evolutionary trajectories, the collective findings from these three studies provide new insights into the genomic and mechanistic diversity of PPD, offering a refined conceptual model for how polyploid genomes undergo structural reorganization over evolutionary time.

7. References

Abel S., Becker H. C. (2007) The effect of autopolyploidy on biomass production in homozygous lines of *Brassica rapa* and *Brassica oleracea*. *Plant Breeding* 126:642-643.

Ahrabi S., Sarkar S., Pfister S. X., Pirovano G., Higgins G. S., Porter A. C., et al. (2016) A role for human homologous recombination factors in suppressing microhomology-mediated end joining. *Nucleic Acids Research* 44:5743-5757.

Ainouche M. L., Wendel J. F. (2014) Polyploid speciation and genome evolution: lessons from recent allopolyploids. In Evolutionary biology: genome evolution, speciation, coevolution and origin of life. Cham: *Springer International Publishing* 87-113.

Amborella Genome Project (2013) The *Amborella* genome and the evolution of flowering plants. *Science* 342:1241089.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796-815.

Alger E. I., Edger P. P. (2020) One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Current Opinion in Plant Biology* 54:108-113.

Ali H., Daser A., Dear P., Wood H., Rabbitts P., Rabbitts T. (2013) Nonreciprocal chromosomal translocations in renal cancer involve multiple DSBs and NHEJ associated with breakpoint inversion but not necessarily with transcription. *Genes, Chromosomes and Cancer* 52:402-409.

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

Al-Shehbaz I. A. (2025) The Brassicaceae then and now: Advancements in the past three decades, a review. *Annals of Botany* mcaf055.

Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.

Argout X., Salse J., Aury J. M., Guiltinan M. J., Droc G., Gouzy J., et al. (2011) The genome of *Theobroma cacao*. *Nature Genetics* 43:101-108.

Artandi S. E., Chang S., Lee S. L., Alson S., Gottlieb G. J., Chin L., et al. (2000) Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* 406:641-645.

Baca S. C., Prandi D., Lawrence M. S., Mosquera J. M., Romanel A., Drier Y., et al. (2013) Punctuated evolution of prostate cancer genomes. *Cell* 153:666-677.

Bairoch A., Boeckmann B. (1992) The SWISS-PROT protein sequence data bank. *Nucleic Acids Research* 20:2019.

Bandyopadhyay R., Heller A., Knox-DuBois C., McCaskill C., Berend S. A., Page S. L., et al. (2002) Parental origin and timing of *de novo* Robertsonian translocation formation. *The American Journal of Human Genetics* 71:1456-1462.

Baranwal V. K., Mikkilineni V., Zehr U. B., Tyagi A. K., Kapoor S. (2012) Heterosis: emerging ideas about hybrid vigour. *Journal of Experimental Botany* 63:6309-6314.

Barker M. S., Arrigo N., Baniaga A. E., Li Z., Levin D. A. (2016) On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* 210:391-398.

Barnes D. E. (2001) Non-homologous end joining as a mechanism of DNA repair. *Current Biology* 11:R455-R457.

Bayat S., Lysak M. A., Mandáková T. (2021) Genome structure and evolution in the cruciferous tribe. Thlaspideae (Brassicaceae). *The Plant Journal* 108:1768-1785.

Behling A. H., Shepherd L. D., Cox M. P. (2020) The importance and prevalence of allopolyploidy in Aotearoa New Zealand. *Journal of the Royal Society of New Zealand* 50:189-210.

Beringer M., Choudhury R. R., Mandáková T., Grünig S., Poretti M., Leitch I. J., et al. (2024) Biased retention of environment-responsive genes following genome fractionation. *Molecular Biology and Evolution* 41:msae155.

Birchler J. A., Veitia R. A. (2012) Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences* 109:14746–14753.

Bird K. A., Niederhuth C. E., Ou S., Gehan M., Pires J. C., Xiong Z., et al. (2021) Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid *Brassica napus*. *New Phytologist* 230:354-371.

Bird K. A., VanBuren R., Puzey J. R., Edger P. P. (2018) The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytologist* 220:87-93.

Britton T., Anderson C. L., Jacquet D., Lundqvist S., Bremer K. (2007) Estimating divergence times in large phylogenetic trees. *Systematic Biology* 56:741–752

Bowers J. E., Chapman B. A., Rong J., Paterson A. H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433-438.

Bredeson J. V., Lyons J. B., Oniyinde I. O., Okereke N. R., Kolade O., Nnabue I., et al. (2022) Chromosome evolution and the genetic basis of agronomically important traits in greater yam. *Nature Communications* 13:2001.

Buggs R. J., Chamala S., Wu W., Tate J. A., Schnable P. S., Soltis D. E., et al. (2012) Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Current Biology* 22:248-252.

Burssed B., Zamariolli M., Bellucco F. T., Melaragno M. I. (2022) Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics* 15:23.

Cai X., Chang L., Zhang T., Chen H., Zhang L., Lin R., et al. (2021) Impacts of allopolyploidization and structural variation on intraspecific diversification in *Brassica rapa*. *Genome Biology* 22:166.

Cannan W. J., Pederson D. S. (2016) Mechanisms and consequences of double-strand DNA break formation in chromatin. *Journal of Cellular Physiology* 231:3-14.

Capella-Gutiérrez S., Silla-Martínez J. M., Gabaldón T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972-1973.

Carta A., Bedini G., Peruzzi L. (2020) A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytologist* 228:1097-1106.

Carvalho C. M., Lupski, J. R. (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics* 17:224-238.

Chagué V., Just J., Mestiri I., Balzergue S., Tanguy A. M., Huneau C., et al. (2010) Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. *New Phytologist* 187:1181-1194.

Chang Y. W., Wang P. H., Li W. H., Chen L. C., Chang C. M., Sung P. L., et al. (2013) Balanced and unbalanced reciprocal translocation: an overview of a 30-year experience in a single tertiary medical center in Taiwan. *Journal of the Chinese Medical Association* 76:153-157.

Chen X., Tong C., Zhang X., Song A., Hu M., Dong W., et al. (2021) A high-quality *Brassica napus* genome reveals expansion of transposable elements, subgenome evolution and disease resistance. *Plant Biotechnology Journal* 19:615-630.

Chen Z. J., Sreedasyam A., Ando A., Song Q., De Santiago L. M., Hulse-Kemp A. M., et al. (2020) Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nature Genetics* 52:525-533.

Cheng F., Sun C., Wu J., Schnable J., Woodhouse M. R., Liang J., et al. (2016) Epigenetic regulation of subgenome dominance following whole genome triplication in *Brassica rapa*. *New Phytologist* 211:288-299.

Cheng F., Wu J., Cai X., Liang J., Freeling M., Wang X. (2018) Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants* 4:258-268.

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

Cheng F., Wu J., Wang X. (2014) Genome triplication drove the diversification of *Brassica* plants. *Horticulture Research* 1:1-8.

Cheng H., Concepcion G. T., Feng X., Zhang H., Li H. (2021) Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* 18:170-175.

Chiatante G., Giannuzzi G., Calabrese F. M., Eichler E. E., Ventura M. (2017) Centromere destiny in dicentric chromosomes: new insights from the evolution of human chromosome 2 ancestral centromeric region. *Molecular Biology and Evolution* 34:1669-1681.

Cui L., Wall P. K., Leebens-Mack J. H., Lindsay B. G., Soltis D. E., Doyle J. J., et al. (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16:738-749.

Cui X., Meng F., Pan X., Qiu X., Zhang S., Li C., et al. (2022) Chromosome-level genome assembly of *Aristolochia contorta* provides insights into the biosynthesis of benzylisoquinoline alkaloids and aristolochic acids. *Horticulture Research* 9:uhac005.

Davis J. T., Li Q., Grassa C. J., Davis M. W., Strauss S. Y., Gremer J. R., et al. (2025) A chromosome-level genome assembly of the varied leaved jewelflower, *Streptanthus diversifolius*, reveals a recent whole genome duplication. *G3: Genes, Genomes, Genetics* 15:jkaf022.

Denoeud F., Carretero-Paulet L., Dereeper A., Droc G., Guyot R., Pietrella M., et al. (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345:1181-1184.

Deriano L., Roth D. B. (2013) Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage. *Annual Review of Genetics* 47:433-455.

Dogan M., Pouch M., Mandáková T., Hloušková P., Guo X., Winter P., et al. (2021) Evolution of tandem repeats is mirroring post-polyploid cladogenesis in *Heliophila* (Brassicaceae). *Frontiers in Plant Science* 11:607893.

Dohm J. C., Minoche A. E., Holtgräwe D., Capella-Gutiérrez S., Zakrzewski F., Tafer H., et al. (2014) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* 505:546-549.

Doležel J., Greilhuber J., Suda J. (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* 2:2233-2244.

Doyle J. J., Coate J. E. (2019) Polyploidy, the nucleotype, and novelty: the impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences* 180:1-52.

Doyle J. J., Egan A. N. (2010) Dating the origins of polyploidy events. *New Phytologist* 186:73-85.

Dudchenko O., Batra S. S., Omer A. D., Nyquist S. K., Hoeger M., Durand N. C., et al. (2017) *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356:92-95.

Durand N. C., Shamim M. S., Machol I., Rao S. S., Huntley M. H., Lander E. S., et al. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* 3:95-98.

Earnshaw W. C., Ratrie III H., Stetten, G. (1989) Visualization of centromere proteins CENP-B and CENP-C on a stable dicentric chromosome in cytological spreads. *Chromosoma* 98:1-12.

Edger P. P., Smith R., McKain M. R., Cooley A. M., Vallejo-Marin M., Yuan Y., et al. (2017) Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *The Plant Cell* 29:2150-2167.

Edger P. P., Poorten T. J., VanBuren R., Hardigan M. A., Colle M., McKain M. R. et al. (2019) Origin and evolution of the octoploid strawberry genome. *Nature Genetics* 51:541–547.

Ellinghaus D., Kurtz S., Willhoeft U. (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* 9:1-14.

Emms D. M., Kelly S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* 20:1–14.

Fang C., Jiang N., Teresi S. J., Platts A. E., Agarwal G., Niederhuth C., et al. (2024) Dynamics of accessible chromatin regions and subgenome dominance in octoploid strawberry. *Nature Communications* 15:2491.

Fang C., Yang M., Tang Y., Zhang L., Zhao H., Ni H., et al. (2023) Dynamics of *cis*-regulatory sequences and transcriptional divergence of duplicated genes in soybean. *Proceedings of the National Academy of Sciences* 120:e2303836120.

Fedyk S., Chętnicki W. (2007) Preferential segregation of metacentric chromosomes in simple Robertsonian heterozygotes of *Sorex araneus*. *Heredity* 99:545-552.

Feng X., Chen Q., Wu W., Wang J., Li G., Xu S., He Z. (2024) Genomic evidence for rediploidization and adaptive evolution following the whole-genome triplication. *Nature Communications* 15:1635

Flagel L., Udall J., Nettleton D., Wendel J. (2008) Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biology* 6:16.

Fonsêca A., Ferraz M. E., Pedrosa-Harand A. (2016) Speeding up chromosome evolution in *Phaseolus*: multiple rearrangements associated with a one-step descending dysploidy. *Chromosoma* 125:413-421.

Freeling M., Woodhouse M. R., Subramaniam S., Turco G., Lisch D., Schnable, J. C. (2012) Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology* 15:131-139.

Fu S., Gao Z., Birchler J., Han F. (2012) Dicentric chromosome formation and epigenetics of centromere formation in plants. *Journal of Genetics and Genomics* 39:125-130.

Garsmeur O., Schnable J. C., Almeida A., Jourda C., D'Hont A., Freeling M. (2014) Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution* 31:448-454.

Geiser C., Mandáková T., Arrigo N., Lysak M. A., Parisod C. (2016) Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in buckler mustard. *The Plant Cell* 28:17-27.

German D. A., Al-Shehbaz I. A. (2008) Five additional tribes (Aphragmeae, Biscutelleae, Calepineae, Conringieae, and Erysimeae) in the Brassicaceae (Cruciferae). *Harvard Papers in Botany* 13:165-170.

German D. A., Friesen N., Neuffer B., Al-Shehbaz I. A., Hurka H. (2009) Contribution to ITS phylogeny of the Brassicaceae, with special reference to some Asian taxa. *Plant Systematics and Evolution* 283:33-56.

German D. A., Hendriks K. P., Koch M. A., Lens F., Lysak M. A., Bailey C. D., et al. (2023) An updated classification of the Brassicaceae (Cruciferae). *PhytoKeys* 220:127.

Goldblatt P., Manning J. C. (2002) Plant diversity of the Cape region of southern Africa. *Annals of the Missouri Botanical Garden* 89:281–302.

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.

Gremme G., Brendel V., Sparks M. E., Kurtz S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* 47:965-978.

Guo X., Mandáková T., Trachtová K., Özüdoğru B., Liu J., Lysak M. A. (2021) Linked by ancestral bonds: multiple whole-genome duplications and reticulate evolution in a Brassicaceae tribe. *Molecular Biology and Evolution* 38:1695-1714.

Haas B. J., Delcher A. L., Mount S. M., Wortman J. R., Smith Jr R. K., Hannick L. I., et al. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* 31:5654-5666.

Haas B. J., Papanicolaou A., Yassour M., Grabherr M., Blood P. D., Bowden J., et al. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8:1494-1512.

Haas B. J., Salzberg S. L., Zhu W., Pertea M., Allen J. E., Orvis J., et al. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* 9:1-22.

Haberer G., Young S., Bharti A. K., Gundlach H., Raymond C., Fuks G., et al. (2005) Structure and architecture of the maize genome. *Plant Physiology* 139:1612-1624.

Hastings P. J., Lupski J. R., Rosenberg S. M., Ira, G. (2009) Mechanisms of change in gene copy number. *Nature Reviews Genetics* 10:551-564.

Hauber D. P., Reeves A., Stack S. M. (1999) Synapsis in a natural autotetraploid. *Genome* 42: 936-949.

Healey A., Garsmeur O., Lovell J., Shengquiang S., Sreedasyam A., Jenkins J., et al. (2024) The complex polyploid genome architecture of sugarcane. *Nature* 628:804-810.

Hendriks K. P., Kiefer C., Al-Shehbaz I. A., Bailey C. D., van Huysduynen A. H., Nikolov L. A., et al. (2023) Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. *Current Biology* 33:4052-4068.

Heslop-Harrison J., Schwarzacher T., Liu Q. (2023) Polyploidy: its consequences and enabling role in plant diversification and evolution. *Annals of Botany* 131:1-10.

Hibrand Saint-Oyant L., Ruttink T., Hamama L., Kirov I., Lakhwani D., Zhou N., et al. (2018) A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nature Plants* 4:473-484.

Hofstatter P. G., Thangavel G., Lux T., Neumann P., Vondrak T., Novak P., et al. (2022) Repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell* 185:3153-3168.

Hohmann N., Wolf E. M., Lysak M. A., Koch M. A. (2015) A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *The Plant Cell* 27:2770-2784.

Holland A. J., Cleveland D. W. (2012) Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nature Medicine* 18:1630-1638.

Hou L., Niu Z., Zheng Z., Zhang J., Luo C., Wang X., et al. (2025) The *Isodon serra* genome sheds light on tanshinone biosynthesis and reveals the recursive karyotype evolutionary histories within Lamiales. *The Plant Journal* 121:e17170.

Hu J., Fan J., Sun Z., Liu S. (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36:2253-2255.

Huang X., Wang Y., Zhang S., Pei L., You J., Long Y., et al. (2024) Epigenomic and 3D genomic mapping reveals developmental dynamics and subgenomic asymmetry of transcriptional regulatory architecture in allotetraploid cotton. *Nature Communications* 15:10721.

Huang X. C., German D. A., Koch M. A. (2020) Temporal patterns of diversification in Brassicaceae demonstrate decoupling of rate shifts and mesopolyploidization events. *Annals of Botany* 125:29-47.

Huang Y., Guo X., Zhang K., Mandáková T., Cheng F., Lysak M. A. (2023) The meso-octoploid *Heliophila variabilis* genome sheds a new light on the impact of polyploidization and diploidization on the diversity of the Cape flora. *The Plant Journal* 116:446-466.

Huang Y., Poretti M., Mandáková T., Pouch M., Guo X., Perez-Roman E., et al. (2025) Post-polyploid chromosomal diploidization in plants is affected by clade divergence and constrained by shared genomic features. https://doi.org/10.21203/rs.3.rs-6440714/v1

Hunter S., Apweiler R., Attwood T. K., Bairoch A., Bateman A., Binns D., et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Research* 37:D211-D215.

Jackson R. C., Jackson J. W. (1996) Gene segregation in autotetraploids: prediction from meiotic configurations. *American Journal of Botany* 83:673-678.

Jaillon O., Aury J. M., Noel B., Policriti A., Clepet C., Casagrande A., et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463-467.

Jarmuz-Szymczak M., Janiszewska J., Szyfter K., Shaffer L. G. (2014) Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. *Chromosome Research* 22:517-532.

Jia K., Wang Z., Wang L., Li G., Zhang W., Wang X., et al. (2022) SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *New Phytologist* 235:801-809.

Jiang X., Hu Q., Mei D., Li X., Xiang L., Al-Shehbaz I. A., et al. (2025) Chromosome fusions shaped karyotype evolution and evolutionary relationships in the model family Brassicaceae. *Nature Communications* 16:1-10.

Jiao Y., Li J., Tang H., Paterson A. H. (2014) Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *The Plant Cell* 26:2792-2802.

Jiao Y., Wickett N. J., Ayyampalayam S., Chanderbali A. S., Landherr L., Ralph P. E., et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:97-100.
Jin J., Yu W., Yang J., Song Y., DePamphilis C. W., Yi T., et al. (2020) GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biology* 21:1-31.
Kang M., Wu H., Yang Q., Huang L., Hu Q., Ma T., et al. (2020) A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine: An *Isatis* genome. *Horticulture Research* 7:18.
Katoh K., Misawa K., Kuma K. I., Miyata T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30:3059-3066.
Kent T., Chandramouly G., McDevitt S. M., Ozdemir A. Y., Pomerantz R. T. (2015) Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase θ. *Nature Structural & Molecular Biology* 22:230-237.
Keymolen K., Van Berkel K., Vorsselmans A., Staessen C., Liebaers I. (2011) Pregnancy outcome in carriers of Robertsonian translocations. *American Journal of Medical Genetics Part A* 155:2381-2385.
Khapugin A. A., Chugunov G. G. (2023) Population status of a regionally endangered plant, *Lunaria rediviva* (Brassicaceae), near the eastern border of its range. *Biology* 12:761.
Kim D., Langmead B., Salzberg S. L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 12:357-360.
Kimura M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111–120.
Koch M. A., Haubold B., Mitchell-Olds T. (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution* 17:1483–1498
Koenen E. J., Ojeda D. I., Steeves R., Migliore J., Bakker F. T., Wieringa J. J., et al. (2020) Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytologist* 225:1355-1369.
Kumar S., Stecher G., Li M., Knyaz C., Tamura K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* 35:1547–1549
Landergott U., Naciri Y., Schneller J. J., Holderegger R. (2006) Allelic configuration and polysomic inheritance of highly variable microsatellites in tetraploid gynodioecious *Thymus praecox* agg. *Theoretical and Applied Genetics* 113:453-465.
Landis J. B., Soltis D. E., Li Z., Marx H. E., Barker M. S., Tank D. C., et al. (2018) Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany* 105:348-363.
Li F., Hou Z., Xu S., Han D., Li B., Hu H., et al. (2024) Haplotype-resolved genomes of octoploid species in Phyllanthaceae family reveal a critical role for polyploidization and hybridization in speciation. *The Plant Journal* 119:348-363.
Li H., Durbin R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754-1760.
Li X., Wang J., Yu Y., Li G., Wang J., Li C., et al. (2023) Genomic rearrangements and evolutionary changes in 3D chromatin topologies in the cotton tribe (Gossypieae). *BMC Biology* 21:56.
Li X., Zhu C., Lin Z., Wu Y., Zhang D., Bai G., et al. (2011) Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. *Molecular Biology and Evolution* 28:1901-1911.
Li Y., Zuo S., Zhang Z., Li Z., Han J., Chu Z., et al. (2018) Centromeric DNA characterization in the model grass *Brachypodium distachyon* provides insights on the evolution of the genus. *The Plant Journal* 93:1088-1101.
Li Z., Barker M. S. (2020) Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. *GigaScience* 9:giaa004.

Li Z., McKibben M. T., Finch G. S., Blischak P. D., Sutherland B. L., Barker M. S. (2021) Patterns and processes of diploidization in land plants. *Annual Review of Plant Biology* 72:387-410.
Liao Y., Smyth G. K., Shi W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923-930.
Liu P., Carvalho C. M., Hastings P., Lupski J. R. (2012) Mechanisms for recurrent and complex human genomic rearrangements. *Current Opinion in Genetics & Development* 22:211-220.
Lomsadze A., Burns P. D., Borodovsky M. (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* 42:e119-e119.
Luo M. C., Deal K. R., Akhunov E. D., Akhunova A. R., Anderson O. D., Anderson J. A., et al. (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proceedings of the National Academy of Sciences* 106:15780-15785.
Lusinska J., Majka J., Betekhtin A., Susek K., Wolny E., Hasterok R. (2018) Chromosome identification and reconstruction of evolutionary rearrangements in *Brachypodium distachyon*, *B. stacei* and *B. hybridum*. *Annals of Botany* 122:445-459.
Lv Z., Addo Nyarko C., Ramtekey V., Behn H., Mason A. S. (2024) Defining autopolyploidy: cytology, genetics, and taxonomy. *American Journal of Botany* 111:e16292.
Lysak M. A. (2014) Live and let die: centromere loss during evolution of plant chromosomes. *New Phytologist* 203: 1082-1089.
Lysak M. A. (2022) Celebrating Mendel, McClintock, and Darlington: on end-to-end chromosome fusions and nested chromosome fusions. *The Plant Cell* 34:2475-2491.
Lysak M. A., Berr A., Pecinka A., Schmidt R., McBreen K., Schubert I. (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proceedings of the National Academy of Sciences* 103:5224-5229.
Lysak M. A., Cheung K., Kitschke M., Bures P. (2007) Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. *Plant Physiology* 145:402-410.
Lysak M. A., Koch M. A., Pecinka A., Schubert I. (2005) Chromosome triplication found across the tribe Brassiceae. *Genome Research* 15:516-525.
Lysak M. A., Mandáková T., Schranz M. E. (2016) Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology* 30:108-115.
Ma X. F., Gustafson J. P. (2005) Genome evolution of allopolyploids: a process of cytological and genetic diploidization. *Cytogenetic and Genome Research* 109:236-249.
MacKinnon R. N., Campbell L. J. (2011) The role of dicentric chromosome formation and secondary centromere deletion in the evolution of myeloid malignancy. *Genetics Research International* 2011:643628.
Malkova A., Ira G. (2013) Break-induced replication: functions and molecular mechanism. *Current Opinion in Genetics & Development* 23:271-279.
Mandáková T., Guo X., Özüdoğru B., Mummenhoff K., Lysak M. A. (2018) Hybridization-facilitated genome merger and repeated chromosome fusion after 8 million years. *The Plant Journal* 96:748-760.
Mandáková T., Joly S., Krzywinski M., Mummenhoff K., Lysak M. A. (2010) Fast diploidization in close mesopolyploid relatives of Arabidopsis. *The Plant Cell* 22:2277-2290.
Mandáková T., Li Z., Barker M. S., Lysak M. A. (2017a) Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *The Plant Journal* 91:3-21.
Mandáková T., Lysak M. A. (2008) Chromosomal phylogeny and karyotype evolution in x = 7 crucifer species (Brassicaceae). *The Plant Cell* 20:2559-2570.
Mandáková T., Lysak M. A. (2016) Chromosome preparation for cytogenetic analyses in *Arabidopsis*. *Current Protocols in Plant Biology* 1:43-51.

Mandáková T., Lysak M. A. (2018) Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology* 42:55-65.

Mandáková T., Mummenhoff K., Al-Shehbaz I. A., Mucina L., Mühlhausen A., Lysak M. A. (2012) Whole genome triplication and species radiation in the southern African tribe Heliophileae (Brassicaceae). *Taxon* 61:989-1000.

Mandáková T., Pouch M., Harmanová K., Zhan S., Mayrose I., Lysak M. A. (2017b) Multispeed genome diploidization and diversification after an ancient allopolyploidization. *Molecular Ecology* 26:6445-6462.

Marçais G., Kingsford C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764-770.

Mason A. S., Wendel J. F. (2020) Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Frontiers in Genetics* 11:1014.

Mayrose I., Zhan S. H., Rothfels C. J., Arrigo N., Barker M. S., Rieseberg L. H., Otto S. P. (2015) Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al.(2014). *New Phytologist* 206:27-35.

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

McClintock B. (1939) The behavior in successive nuclear divisions of a chromosome broken at meiosis. *Proceedings of the National Academy of Sciences* 25:405-416.

McClintock B. (1941) The stability of broken ends of chromosomes in *Zea mays*. *Genetics* 26:234.

Merker L., Feller L., Dorn A., Puchta H. (2024) Deficiency of both classical and alternative end-joining pathways leads to a synergistic defect in double-strand break repair but not to an increase in homology-dependent gene targeting in *Arabidopsis*. *The Plant Journal* 118:242-254.

Mhiri C., Parisod C., Daniel J., Petit M., Lim K. Y., Dorlhac de Borne F., et al. (2019) Parental transposable element loads influence their dynamics in young *Nicotiana* hybrids and allotetraploids. *New Phytologist* 221:1619-1633.

Mittermeier R. A., Myers N., Thomsen J. B., Da Fonseca G. A., Olivieri, S. (1998) Biodiversity hotspots and major tropical wilderness areas: approaches to setting conservation priorities. *Conservation Biology* 12:516–520.

Mummenhoff K., Al-Shehbaz I. A., Bakker F. T., Linder H. P., Mühlhausen A. (2005) Phylogeny, morphological evolution, and speciation of endemic Brassicaceae genera in the Cape flora of southern Africa. *Annals of the Missouri Botanical Garden* 400-424.

Murat F., Xu J., Tannier E., Abrouk M., Guilhot N., Pont C., et al. (2010) Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Research* 20:1545-1557.

Myers N., Mittermeier R. A., Mittermeier C. G., Da Fonseca G. A., Kent J. (2000) Biodiversity hotspots for conservation priorities. *Nature* 403:853-858.

Nguyen L. T., Schmidt H. A., Von Haeseler A., Minh B. Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32:268-274.

Nikolov L. A., Shushkov P., Nevado B., Gan X., Al-Shehbaz I. A., Filatov D., et al. (2019) Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist* 222:1638-1651.

Oberlander K. C., Dreyer L. L., Goldblatt P., Suda J., Linder H. P. (2016) Species-rich and polyploid-poor: Insights into the evolutionary role of whole-genome duplication from the Cape flora biodiversity hotspot. *American Journal of Botany* 103:1336-1347.

Ondov B. D., Treangen T. J., Melsted P., Mallonee A. B., Bergman N. H., Koren S., et al. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* 17:132.

Otto S. P., Whitton, J. (2000) Polyploid incidence and evolution. *Annual Review of Genetics* 34:401-437.

Ou S., Jiang N. (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* 10:48.

Ou S., Jiang N. (2018) LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* 176:1410-1422.

Ou S., Su W., Liao Y., Chougule K., Agda J. R., Hellinga A. J., et al. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* 20:1-18.

Özüdoğru B., Akaydın G., Erik S., Al-Shehbaz I. A., Mummenhoff K. (2015) Phylogeny, diversification and biogeographic implications of the eastern Mediterranean endemic genus *Ricotia* (Brassicaceae). *Taxon* 64:727-740.

Özüdoğru B., Akaydın G., Erik S., Mummenhoff K. (2016) Seed morphology of *Ricotia* (Brassicaceae) and its phylogenetic and systematic implication. *Flora-Morphology, Distribution, Functional Ecology of Plants* 222:60-67.

Özüdoğru B., Al-Shehbaz I. A., Mummenhoff K. (2017) Tribal assignment of *Heldreichia* Boiss.(Brassicaceae): evidence from nuclear ITS and plastidic *ndh*F markers. *Plant Systematics and Evolution* 303:329-335.

Özüdoğru B., Karacaoğlu Ç., Akaydın G., Erik S., Mummenhoff K., Sağlam İ. K. (2022) Ecological specialization promotes diversity and diversification in the Eastern Mediterranean genus *Ricotia* (Brassicaceae). *Journal of Systematics and Evolution* 60:331-343.

Parisod C., Badaeva E. D. (2020) Chromosome restructuring among hybridizing wild wheats. *New Phytologist* 226:1263-1273.

Parisod C., Holderegger R., Brochmann C. (2010) Evolutionary consequences of autopolyploidy. *New Phytologist* 186:5-17.

Parisod C., Poretti M., Mandáková T., Choudhury R., Lysak M. A. (2025) The role of centromeric transposable elements in shaping chromosome evolution. https://doi.org/10.21203/rs.3.rs-5461468/v1

Paritosh K., Yadava S. K., Singh P., Bhayana L., Mukhopadhyay A., Gupta V., et al. (2021) A chromosome-scale assembly of allotetraploid *Brassica juncea* (AABB) elucidates comparative architecture of the A and B genomes. *Plant Biotechnology Journal* 19:602-614.

Parolly G., Nordt B., Bleeker W., Mummenhoff K. (2010) *Heldreichia* Boiss.(Brassicaceae) revisited: a morphological and molecular study. *Taxon* 59:187-202.

Paterson A., Bowers J., Chapman B. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences* 101:9903-9908.

Pellestor F. (2019) Chromoanagenesis: cataclysms behind complex chromosomal rearrangements. *Molecular Cytogenetics* 12:6.

Pellestor F., Gatinois V. (2018) Chromoanasynthesis: another way for the formation of complex chromosomal abnormalities in human reproduction. *Human Reproduction* 33:1381-1387.

Pellicer J., Leitch I. J. (2020) The Plant DNA C-values database (release 7.1). *New Phytologist* 226:301-305.

Peppe D. J., Cote S. M., Deino A. L., Fox D. L., Kingston J. D., Kinyanjui R. N., et al. (2023) Oldest evidence of abundant C4 grasses and habitat heterogeneity in eastern Africa. *Science* 380:173-177.

Pertea M., Pertea G. M., Antonescu C. M., Chang T. C., Mendell J. T., Salzberg S. L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33:290-295.

POWO (2025) Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet. Available online: http://www.plantsoftheworldonline.org (accessed on 16 January 2025)

Poot M., Hochstenbach R. (2021) Prevalence and phenotypic impact of Robertsonian translocations. *Molecular Syndromology* 12:1-11.

Pruitt K. D., Tatusova T., Maglott D. R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35:D61-D65.

Qin L., Hu Y., Wang J., Wang X., Zhao R., Shan H., et al. (2021) Insights into angiosperm evolution, floral development and chemical biosynthesis from the *Aristolochia fimbriata* genome. *Nature Plants* 7:1239-1253.

Qu L., Hancock J. F., Whallon J. H. (1998) Evolution in an autopolyploid group displaying predominantly bivalent pairing at meiosis: genomic similarity of diploid *Vaccinium darrowi* and autotetraploid *V. corymbosum* (Ericaceae). *American Journal of Botany* 85:698-703.

Ramírez F., Bhardwaj V., Arrigoni L., Lam K. C., Grüning B. A., Villaveces J., et al. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications* 9:189.

Rapp R. A., Udall J. A., Wendel J. F. (2009) Genomic expression dominance in allopolyploids. *BMC Biology* 7:1-10.

Raza A., Hafeez M. B., Zahra N., Shaukat K., Umbreen S., Tabassum J., et al. (2020) The plant family. Brassicaceae: Introduction, biology, and importance. *The Plant Family Brassicaceae: Biology and Physiological Responses to Environmental Stresses*:1-43.

Rice A., Glick L., Abadi S., Einhorn M., Kopelman N. M., Salman-Minkov A., et al. (2015) The Chromosome Counts Database (CCDB)–a community resource of plant chromosome numbers. *New phytologist* 206:19-26.

Robinson J. T., Turner D., Durand N. C., Thorvaldsdóttir H., Mesirov J. P., Aiden E. L. (2018) Juicebox. js provides a cloud-based visualization system for Hi-C data. *Cell Systems* 6:256-258.

Robinson W. P., Bernasconi F., Basaran S., Yüksel-Apak M., Neri G., Serville F., et al. (1994) A somatic origin of homologous Robertsonian translocations and isochromosomes. *American Journal of Human Genetics* 54:290.

Salse J., Bolot S., Throude M., Jouffe V., Piegu B., Quraishi U. M., et al. (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *The Plant Cell* 20:11-24.

Sanderson M. J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302

SanMiguel P., Gaut B. S., Tikhonov A., Nakajima Y., Bennetzen J. L. (1998) The paleontology of intergene retrotransposons of maize. *Nature Genetics* 20:43–45

Sasaki M., Lange J., Keeney S. (2010) Genome destabilization by homologous recombination in the germ line. *Nature Reviews Molecular Cell Biology* 11:182–195.

Sato H., Masuda F., Takayama Y., Takahashi K., Saitoh S. (2012) Epigenetic inactivation and subsequent heterochromatinization of a centromere stabilize dicentric chromosomes. *Current Biology* 22:658-667.

Schnable J. C., Springer N. M., Freeling M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences* 108:4069-4074.

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

Schubert I. (2021) Boon and bane of DNA double-strand breaks. *International Journal of Molecular Sciences* 22:5171.

Schubert I., Lysak M. A. (2011) Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends in Genetics* 27:207-216.

Schubert I., Oud J. L. (1997) There is an upper limit of chromosome size for normal development of an organism. *Cell* 88:515-520.

Schubert I., Vu G. T. (2016) Genome stability and evolution: attempting a holistic view. *Trends in Plant Science* 21:749-757.

Sfeir A., Symington L. S. (2015) Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends in Biochemical Sciences* 40:701-714.
Shen F., Xu S., Shen Q., Bi C., Lysak M. A. (2023) The allotetraploid horseradish genome provides insights into subgenome diversification and formation of critical traits. *Nature Communications* 14:4102.
Shoshani O., Brunner S. F., Yaeger R., Ly P., Nechemia-Arbely Y., Kim D. H., et al. (2021) Chromothripsis drives the evolution of gene amplification in cancer. *Nature* 591:137-141.
Simão F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V., Zdobnov E. M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210-3212.
Simovic M., Ernst A. (2022) Chromothripsis, DNA repair and checkpoints defects. *Seminars in Cell & Developmental Biology* 123:110-114.
Slotte T., Hazzouri K. M., Ågren J. A., Koenig D., Maumus F., Guo Y. L., et al. (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics* 45:831-835.
Smith C. E., Llorente B., Symington L. S. (2007) Template switching during break-induced replication. *Nature* 447:102-105.
Soltis D. E., Soltis P. S., Schemske D. W., Hancock J. F., Thompson J. N., Husband B. C., et al. (2007) Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* 56:13-30.
Soltis D. E., Visger C. J., Marchant D. B., Soltis P. S. (2016) Polyploidy: pitfalls and paths to a paradigm. *American Journal of Botany* 103:1146-1166.
Sokolovskaya A. P., Probatova N. S. (1977) O naimenshem chisle khromosom (2*n* = 4) u *Colpodium versicolor* (Stev.) Woronow. Bot Zhurn (Moscow & Leningrad) 62:241–245
Song J., Guan Z., Hu J., Guo C., Yang Z., Wang S., et al. (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nature Plants* 6:34-45.
Song X., Wei Y., Xiao D., Gong K., Sun P., Ren Y., et al. (2021) *Brassica carinata* genome characterization clarifies U's triangle model of evolution and polyploidy in *Brassica*. *Plant Physiology* 186:388-406.
Stanke M., Keller O., Gunduz I., Hayes A., Waack S., Morgenstern B. (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research* 34:W435-W439.
Stankiewicz P., Lupski J. R. (2002) Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* 18:74-82.
Startek M., Szafranski P., Gambin T., Campbell I. M., Hixson P., Shaw C. A., et al. (2015) Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination. *Nucleic Acids Research* 43:2188-2198.
Stebbins G. L. (1947) Types of polyploids: their classification and significance. *Advances in Genetics* 1:403-429.
Stebbins G. L. (1950) Variation and evolution in plants. *Oxford University Press*.
Stimpson K. M., Matheny J. E., Sullivan B. A. (2012) Dicentric chromosomes: unique models to study centromere function and inactivation. *Chromosome Research* 20:595-605.
Stift M., Berenos C., Kuperus P., van Tienderen P. H. (2008) Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *Rorippa* (yellow cress) microsatellite data. *Genetics* 179:2113-2123.
Sullivan B. A., Schwartz S. (1995) Identification of centromeric antigens in dicentric Robertsonian translocations: CENP-C and CENP-E are necessary components of functional centromeres. *Human Molecular Genetics* 4:2189-2197.
Sun H., Ding J., Piednoël M., Schneeberger K. (2018) findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* 34:550-557.
Sun P., Jiao B., Yang Y., Shan L., Li T., Li X., et al. (2022) WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Molecular Plant* 15:1841-1851.

Sun P., Lu Z., Wang Z., Wang S., Zhao K., Mei D., et al. (2024) Subgenome-aware analyses reveal the genomic consequences of ancient allopolyploid hybridizations throughout the cotton family. *Proceedings of the National Academy of Sciences* 121:e2313921121.

Tang H., Bowers J. E., Wang X., Paterson A. H. (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences* 107:472-477.

Tayalé A., Parisod C. (2013) Natural pathways to polyploidy in plants and consequences for genome reorganization. *Cytogenetic and Genome Research* 140:79-96.

Thomas B. C., Pedersen B., Freeling, M. (2006) Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* 16:934-946.

Trávníček P., Ponert J., Urfus T., Jersáková J., Vrána J., Hřibová E., 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 Part A* 87:958-966.

Van de Peer Y., Ashman T. L., Soltis P. S., Soltis D. E. (2021) Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell* 33:11-26.

Van Santen M., Linder H. P. (2020) The assembly of the Cape flora is consistent with an edaphic rather than climatic filter. *Molecular Phylogenetics and Evolution* 142:106645.

Vanzela A. L. L., Guerra M., Luceño M. (1996) *Rhynchospora tenuis* Link (Cyperaceae), a species with the lowest number of holocentric chromosomes. *Cytobios* 88:219-228.

Verboom G. A., Archibald J. K., Bakker F. T., Bellstedt D. U., Conrad F., Dreyer L. L., et al. (2009) Origin and diversification of the Greater Cape flora: ancient species repository, hot-bed of recent radiation, or both? *Molecular Phylogenetics and Evolution* 51:44-53.

Vervoort L., Vermeesch J. R. (2023) Low copy repeats in the genome: From neglected to respected. *Exploration of Medicine* 4:166-175.

Walden N., German D. A., Wolf E. M., Kiefer M., Rigault P., Huang X., et al. (2020) Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. *Nature Communications* 11:1-12.

Wang D., Zhang Y., Zhang Z., Zhu J., Yu J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics & Bioinformatics* 8:77-80.

Wang M., Wang P., Lin M., Ye Z., Li G., Tu L., et al. (2018) Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nature Plants* 4:90-97.

Wang X., Jin D., Wang Z., Guo H., Zhang L., Wang L., et al. (2015) Telomere-centric genome repatterning determines recurring chromosome number reductions during the evolution of eukaryotes. *New Phytologist* 205:378-389.

Wang Z., Li Y., Sun P., Zhu M., Wang D., Lu Z., et al. (2022) A high-quality *Buxus austro-yunnanensis* (Buxales) genome provides new insights into karyotype evolution in early eudicots. *BMC Biology* 20:216.

Wang Z., Wang J., Pan Y., Lei T., Ge W., Wang L., et al. (2019) Reconstruction of evolutionary trajectories of chromosomes unraveled independent genomic repatterning between Triticeae and *Brachypodium*. *BMC Genomics* 20:180.

Wagner Jr W. H. (1970) Biosystematics and evolutionary noise. *Taxon* 19:146-51.

Wendel J. F. (2015) The wondrous cycles of polyploidy in plants. *American Journal of Botany* 102:1753-1756

Wlodzimierz P., Hong M., Henderson I. R. (2023) TRASH: tandem repeat annotation and structural hierarchy. *Bioinformatics* 39:btad308.

Wolfe K. H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2:333-341.

Wood T. E., Takebayashi N., Barker M. S., Mayrose I., Greenspoon P. B., Rieseberg L. H. (2009) The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* 106:13875-13879.

Woodhouse M. R., Schnable J. C., Pedersen B. S., Lyons E., Lisch D., Subramaniam S., et al. (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biology* 8:e1000409.

Wu T. D., Watanabe C. K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21:1859-1875.

Wyttenbach A., Hausser J. (1996) The fixation of metacentric chromosomes during the chromosomal evolution in the common shrew (*Sorex araneus*, Insectivora). *Hereditas* 125:209-217.

Xie D., Xu Y., Wang J., Liu W., Zhou Q., Luo S., et al. (2019) The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. *Nature Communications* 10:5158.

Xiong Z., Gaeta R. T., Pires J. C. (2011) Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proceedings of the National Academy of Sciences* 108:7908-7913.

Xu W., Zhang Q., Yuan W., Xu F., Muhammad Aslam M., Miao R., et al. (2020) The genome evolution and low-phosphorus adaptation in white lupin. *Nature Communications* 11:1069.

Yang T., Cai B., Jia Z., Wang Y., Wang J., King G. J., et al. (2023) *Sinapis* genomes provide insights into whole-genome triplication and divergence patterns within tribe Brassiceae. *The Plant Journal* 113:246-261.

Yang W., Feng L., Jiao P., Xiang L., Yang L., Olonova M. V., et al. (2023) Out of the Qinghai-Tibet plateau: Genomic biogeography of the alpine monospecific genus *Megadenia* (Biscutelleae, Brassicaceae). *Molecular Ecology* 32:492-503.

Yang W., Zhang L., Mandáková T., Huang L., Li T., Jiang J., et al. (2021) The chromosome-level genome sequence and karyotypic evolution of *Megadenia pygmaea* (Brassicaceae). *Molecular Ecology Resources* 21:871-879.

Yokota E., Shibata F., Nagaki K., Murata M. (2011) Stability of monocentric and dicentric ring minichromosomes in *Arabidopsis*. *Chromosome Research* 19:999-1012.

Yu G., Wang L., Han Y., He Q. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* 16:284-287.

Zepeda-Mendoza C. J., Morton C. C. (2019) The iceberg under water: unexplored complexity of chromoanagenesis in congenital disorders. *The American Journal of Human Genetics* 104:565-577.

Zhang C., Scornavacca C., Molloy E. K., Mirarab S. (2020) ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution* 37:3292-3307.

Zhang J., Zhang X., Tang H., Zhang Q., Hua X., Ma X. et al. (2018) Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nature Genetics*, 50:1565–1573.

Zhang K., Wang X., Cheng F. (2019) Plant polyploidy: origin, evolution, and its influence on crop domestication. *Horticultural Plant Journal* 5:231-239.

Zhang K., Yu H., Zhang L., Cao Y., Li X., Mei Y., et al. (2025) Transposon proliferation drives genome architecture and regulatory evolution in wild and domesticated peppers. *Nature Plants* 11:359-375.

Zhang K., Zhang L., Cui Y., Yang Y., Wu J., Liang J., et al. (2023) The lack of negative association between TE load and subgenome dominance in synthesized *Brassica* allotetraploids. *Proceedings of the National Academy of Sciences* 120:e2305208120.

Zhang R. G., Li G. Y., Wang X. L., Dainat J., Wang Z., Ou S., et al. (2022) TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research* 9:uhac017.

Zhang R. G., Liu H., Shang H. Y., Shu H., Liu D. T., Yang H., et al. (2025) Convergent patterns of karyotype evolution underlying karyotype uniformity in conifers. *Advanced Science* 12:2411098.

Zhang W., Friebe B., Gill B. S., Jiang J. (2010) Centromere inactivation and epigenetic modifications of a plant chromosome with three functional centromeres. *Chromosoma* 119:553-563.

Zhang X., Chen S., Shi L., Gong D., Zhang S., Zhao Q., et al. (2021) Haplotype-resolved genome assembly. provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nature Genetics* 53:1250-1259.

Zhang Z., Xiao J., Wu J., Zhang H., Liu G., Wang X., et al. (2012) ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications* 419:779-781.

Zhou C., McCarthy S. A., Durbin R. (2023) YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* 39:btac808.

Zhuang X., Huang J., Jin X., Yu Y., Li J., Qiao J., et al. (2014) Chromosome aberrations and spermatogenic disorders in mice with Robertsonian translocation (11; 13). *International Journal of Clinical and Experimental Pathology* 7:7735.

Zhuang Y., Wang X., Li X., Hu J., Fan L., Landis J. B., et al. (2022) Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. *Nature Plants* 8:233-244.

APPENDIX 1

the plant journal

The Plant Journal (2023) **116,** 446–466 doi: 10.1111/tpj.16383

The meso-octoploid *Heliophila variabilis* genome sheds a new light on the impact of polyploidization and diploidization on the diversity of the Cape flora

Yile Huang[1,2,†], Xinyi Guo[1,†], Kang Zhang[3], Terezie Mandáková[1,4,*], Feng Cheng[3,*] and Martin A. Lysak[1,2,*]

[1]*Central European Institute of Technology (CEITEC), Masaryk University, Kamenice 5, Brno 625 00, Czech Republic,*
[2]*National Centre for Biomolecular Research (NCBR), Masaryk University, Kamenice 5, Brno 625 00, Czech Republic,*
[3]*State Key Laboratory of Vegetable Biobreeding, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture and Rural Affairs, Sino-Dutch Joint Laboratory of Horticultural Genomics, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China, and*
[4]*Department of Experimental Biology, Masaryk University, Kamenice 5, Brno 625 00, Czech Republic*

Received 17 February 2023; revised 5 June 2023; accepted 3 July 2023; published online 10 July 2023.
*For correspondence (e-mail martin.lysak@ceitec.muni.cz; terezie.mandakova@ceitec.muni.cz; chengfeng@caas.cn).
†These authors contributed equally to this work.

SUMMARY

Although the South African Cape flora is one of the most remarkable biodiversity hotspots, its high diversity has not been associated with polyploidy. Here, we report the chromosome-scale genome assembly of an ephemeral cruciferous species *Heliophila variabilis* (~334 Mb, *n* = 11) adapted to South African semiarid biomes. Two pairs of differently fractionated subgenomes suggest an allo-octoploid origin of the genome at least 12 million years ago. The ancestral octoploid *Heliophila* genome (2*n* = 8*x* = ~60) has probably originated through hybridization between two allotetraploids (2*n* = 4*x* = ~30) formed by distant, intertribal, hybridization. Rediploidization of the ancestral genome was marked by extensive reorganization of parental subgenomes, genome downsizing, and speciation events in the genus *Heliophila*. We found evidence for loss-of-function changes in genes associated with leaf development and early flowering, and over-retention and sub/neofunctionalization of genes involved in pathogen response and chemical defense. The genomic resources of *H. variabilis* will help elucidate the role of polyploidization and genome diploidization in plant adaptation to hot arid environments and origin of the Cape flora. The sequenced *H. variabilis* represents the first chromosome-scale genome assembly of a meso-octoploid representative of the mustard family.

Keywords: genome assembly, octoploidy, polyploidy, whole-genome duplication, genome diploidization, chromosomal rearrangements, adaptive evolution, southern Africa, Brassicaceae.

INTRODUCTION

Polyploidization, whole-genome duplications (WGDs), occurred frequently during land plant evolution. Dozens of WGDs have been identified at the base of major, as well as minor, plant lineages (Jaillon et al., 2007; Jiao et al., 2011; The Arabidopsis Genome Initiative, 2000; Van de Peer et al., 2017). More than 240 putative ancient WGDs have been inferred to date across the Viridiplantae (Li & Barker, 2020). Since the distribution of WGDs along the reconstructed phylogenies follows a non-random pattern, ancient polyploidization events most likely increased the chance of polyploid plants to adapt to new ecological niches and survive under stressful conditions (Van de Peer et al., 2017, 2021). Whole-genome duplication events are usually followed by rediploidization of the genome, which gradually erases the signature of polyploidization by merging the parental subgenome into a pseudodiploid genome. Genomic changes along with biased gene retention and functional divergence following WGDs may have played an important role in plant trait innovation, speciation, and cladogenesis as WGDs per se (Clark & Donoghue, 2017; Mandáková & Lysak, 2018; Robertson et al., 2017; Schranz et al., 2012; Vekemans et al., 2012).

In many clades of angiosperms, the process of genome rediploidization was followed by one or more new rounds of hybridization (allopolyploidization) or autopolyploidization, so that some clades have genomes formed by multiple WGD events of different ages. In this case, ancient

paleopolyploid WGDs were overwritten by more recent mesopolyploid and even more recent neopolyploid events, as shown by the example of the canola (*Brassica napus*) genome, which could be reconstructed as containing 72 monoploid chromosome sets (Chalhoub et al., 2014). Mechanistically, tetraploidization is the simplest pathway to polyploidy in plants, and indeed, genome duplications are more common than whole-genome triplications (WGTs; Mandáková et al., 2017; Qiao et al., 2022). Nevertheless, paleohexaploidization events have frequently occurred during angiosperm diversification, such as the gamma triplication that preceded the diversification of the core eudicots (Jaillon et al., 2007; Vekemans et al., 2012), two WGTs in the order Solanales (Zhang et al., 2022), the mesopolyploid WGT at the base of the tribe Brassiceae (Lysak et al., 2007; Wang et al., 2011), or the allohexaploid origin of the wheat (International Wheat Genome Sequencing Consortium, 2018) and oat (Kamal et al., 2022) genomes. In contrast, ancient octoploidizations occurred much less frequently. Only three octoploid angiosperm genomes have been analyzed with chromosome-level precision until now: allo-octoploid strawberry genomes (*Fragaria* spp., $2n = 8x = 56$; Edger et al., 2019), the octoploid sugarcane species (*Saccharum officinarum*, $2n = 8x = 80$; *S. spontaneum*, $2n = 8x = 64$; Zhang et al., 2018), and the recently sequenced auto-octoploid genome of beak-sedge *Rhynchospora pubera* ($2n = 10$; Hofstatter et al., 2022).

The evolution of crucifer genomes (Brassicaceae, the mustard family) was marked by the occurrence of more than 13 genus- and tribe-specific mesopolyploid genome duplications (Guo et al., 2021; Mandáková et al., 2017), post-dating the family-specific At-α paleotetraploid WGD (The Arabidopsis Genome Initiative, 2000) and pre-dating numerous neopolyploid events (e.g., *Arabidopsis suecica*, *B. napus*, *Capsella bursa-pastoris*). One of the largest crucifer tribes, the monogeneric tribe Heliophileae, has been hypothesized to diversify following a mesopolyploid WGD (Dogan et al., 2021; Mandáková et al., 2012). The genus *Heliophila* is restricted to southern Africa (eSwatini, Lesotho, Namibia, and South Africa) and is considered the most morphologically diverse crucifer clade (Mandáková et al., 2012; Mummenhoff et al., 2005). More than a 100 *Heliophila* species include both short-lived annuals and perennial herbs (including the lianelle *H. scandens*), subshrubs, and tall shrubs (*H. juncea*), and they also vary widely in the type and size of their fruits (Dogan et al., 2021; Mummenhoff et al., 2005). Flower colors are equally variable, ranging from white and yellow to pink, purple, or blue, with blue flowers known only in *Heliophila* and Himalayan *Solms-laubachia* in the Brassicaceae (Dogan et al., 2021). The greatest diversity of *Heliophila* species is restricted to southwestern South Africa, home to two global biodiversity hotspots—the Cape Floristic Region (CFR) and the Succulent Karoo—sometimes united as the Greater Cape Floristic Region (Hopper et al., 2016).

Previous cytogenomic analyses indicated that the Heliophileae most likely descended from a mesohexaploid genome (Mandáková et al., 2012, 2017), but some genomic blocks identified by comparative chromosome painting as more than three homeologous copies (Mandáková et al., 2012) indicated that the origin of the *Heliophila* genomes may be more complex. To elucidate the polyploid history of *Heliophila* genomes, we sequenced and analyzed the genome of *Heliophila variabilis* ($2n = 22$, 1C = 334 Mb). White-flowered *H. variabilis* is an ephemeral annual species that occurs primarily in the southwesternmost corner of South Africa, throughout the Succulent Karoo (Richtersveld and Namaqualand), and less frequently in the CFR. The species is adapted to extreme summer aridity by escaping drought through rapid germination and growth after winter rains, completing its life cycle before the onset of drought. Locally, *H. variabilis* contributes significantly to the massive spring flower display that many ephemeral plant species produce after winter rains (Figure 1a; Van Rooyen et al., 1992).

Here, we report a chromosome-level assembly of the 334-Mb genome of *H. variabilis* ($2n = 22$) and show that this species and probably the entire tribe Heliophileae (Mandáková et al., 2012) descend from an allo-octoploid genome ($2n = \sim 60$). The last WGD in *Heliophila* occurred at least 12 million years ago (Mya) and was followed by genome rediploidization, as revealed by biased gene fractionation, dominant gene expression, and ancient bursts of transposable elements. The retained genes after WGD showed evidence of functional compatibility between subgenomes and further divergence during evolution. In addition, we showed that genes with distinct fate after duplication might have jointly contributed to the adaptation and survival of *Heliophila* species in semiarid environments in southern Africa. The elucidated evolution of the *H. variabilis* genome challenges the traditional picture of the limited contribution of WGDs to the diversity of the South African Cape flora (Oberlander et al., 2016).

RESULTS

Assembly and gene annotation of the *Heliophila variabilis* genome

We sequenced the *H. variabilis* genome by combining Illumina short-read, Oxford Nanopore long-read, and high-throughput chromosome conformation capture (Hi-C) sequencing technologies. We generated 48.41 Gb Illumina paired-end reads (representing a 161× of the 334-Mb genome) to perform genome survey before assembly (Table **S1**). The genome size of *H. variabilis* was estimated to be 265–310 Mb based on the distribution of 21-mers (Figure **S1**). The genome heterozygosity and the proportion

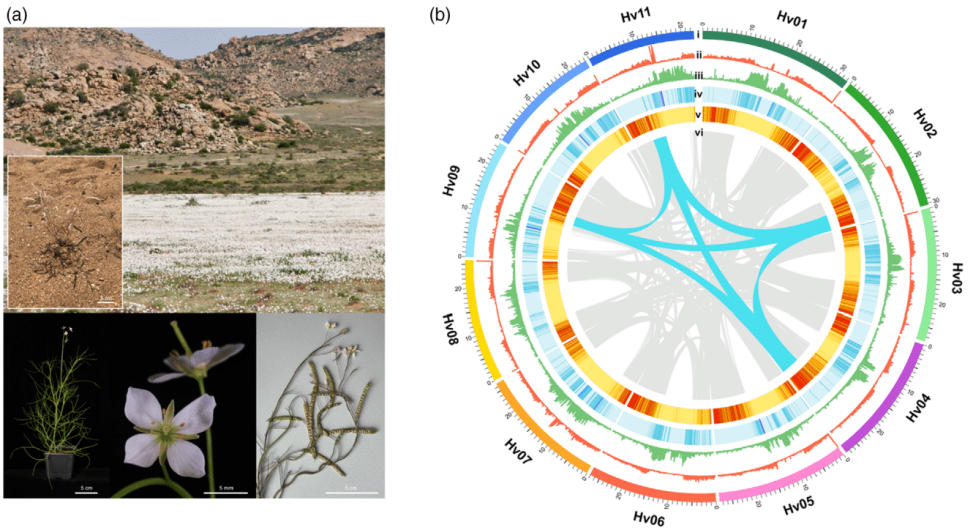


Figure 1. Morphological characteristics of *Heliophila variabilis* and its genome structure.
(a) Spring flower carpet in Goegap Nature Reserve (Namaqualand, South Africa) dominated by *H. variabilis*, and a close-up of a fruiting plant at the same site. Morphological characters of *H. variabilis* are shown below.
(b) Collinearity within the *H. variabilis* genome. The circles from outside to inside show (i) 11 chromosomes (named Hv01-Hv11); (ii) densities of DNA transposons; (iii) densities of LTR retrotransposons; (iv) densities of all TEs; (v) densities of genes; (vi) gene synteny between 11 chromosomes of *H. variabilis* is shown by the gray lines, the blue lines represent a set of four duplicated genomic fragments, corresponding to genomic block A in the ancestral genome.

of repeats were estimated to be 0.0859% and 33.44%, respectively (Figure **S1**). We generated 51.63 Gb (172×) Oxford Nanopore long reads (N50 length, 35.16 kb) (Table **S1**) and performed genome assembly, followed by sequence polishing and filtering with both long and short reads (see Experimental Procedures). We obtained a total of 77 contigs with a total length of 300.50 Mb and a contig N50 size of 11.32 Mb (Table S2). These contigs were assigned to 11 pseudochromosomes ranging in size from 23.85 to 32.26 Mb based on the contact information of the Hi-C data (Table S3), corresponding to ~99.41% of all assembled sequences (Figure S2). Benchmarking Universal Single-Copy Ortholog (BUSCO) analysis of the assembled sequences showed the high completeness of genome assembly (98.7%; Figure S3a).

A total of 32 351 protein-coding genes in the *H. variabilis* genome were annotated by a combination of *de novo* prediction, homology search, and transcriptome-based prediction with mRNA-seq data from floral, leaf, silique, and stem tissues (Table S4). The homology search revealed that 96.25% of the predicted genes had their best hits to at least one of the plant sequences from the databases of National Center for Biotechnology Information (NCBI), Interpro, Swissprot, Gene Ontology (GO), or Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (Table S5). Most (96.8%) of the 1614 core genes in the Embryophyta database were identified in the predicted gene set of *H. variabilis*, confirming the reliability of gene prediction (Figure S3b).

Repeatome composition

A total of 131.83 Mb (43.87%) of the assembled sequences were predicted as transposable elements (TEs), including class I (retrotransposons: 24.15 Mb), class II (DNA transposons: 34.58 Mb), and other TE groups (73.10 Mb; Figure 1b and Table S6). Helitron (20.68 Mb) and *Gypsy* (19.72 Mb) were the most abundant subfamilies of TEs, accounting for 59.80% and 81.65% of DNA transposons and retrotransposons, respectively (Table S6). Previous studies have shown that long terminal repeat (LTR) retrotransposons are the most abundant mobile element in almost all plants (Feschotte et al., 2002; Kidwell, 2002). Although retrotransposons also prevailed in *H. variabilis*, we also identified a remarkable number of DNA transposons (26.23% of all TEs). Therefore, we compared the distribution of the different TE elements in the pericentromere and chromosome arm regions of the 11 chromosomes. The total length of TEs in the pericentromere and arm regions was approximately 91.61 Mb (69.49%) and 38.37 Mb (29.11%), respectively. Long terminal repeat retrotransposons (LTR-RTs)

and simple tandem repeats were predominantly distributed in the pericentromeric regions (85.47% and 96.87%, respectively), whereas DNA transposons were mainly distributed along the chromosome arms (75.11%; Figures S4a,b).

We identified a substantial number of tandem repeats in the *H. variabilis* genome, including rDNAs (5S and 35S, 2.49 Mb, 0.82%) and satellites (37.58 Mb, 12.51%). However, the proportion of rDNA sequences in the genome assembly was much lower than that identified by Dogan et al. based on short reads (c. 5%; Dogan et al., 2021), suggesting an incomplete representation of these sequences in the genome assembly. Nearly half of the sequences (1.2 Mb) were placed in pseudomolecules Hv02, Hv03, and Hv04 (Figure S5a). Consistently, our FISH analyses identified putative nucleolus organizer regions (NORs) on three different chromosomes (Figure S5b). We identified two major satellite repeats, including a 177-bp putative centromere-specific repeat in all chromosomes and a 168-bp repeat in pericentromeric regions of eight chromosomes (Figure S5a), which accounted for 10.55 Mb (3.5%) and 24.37 Mb (8.1%) of the genome assembly, respectively.

An octoploid origin of the *H. variabilis* genome

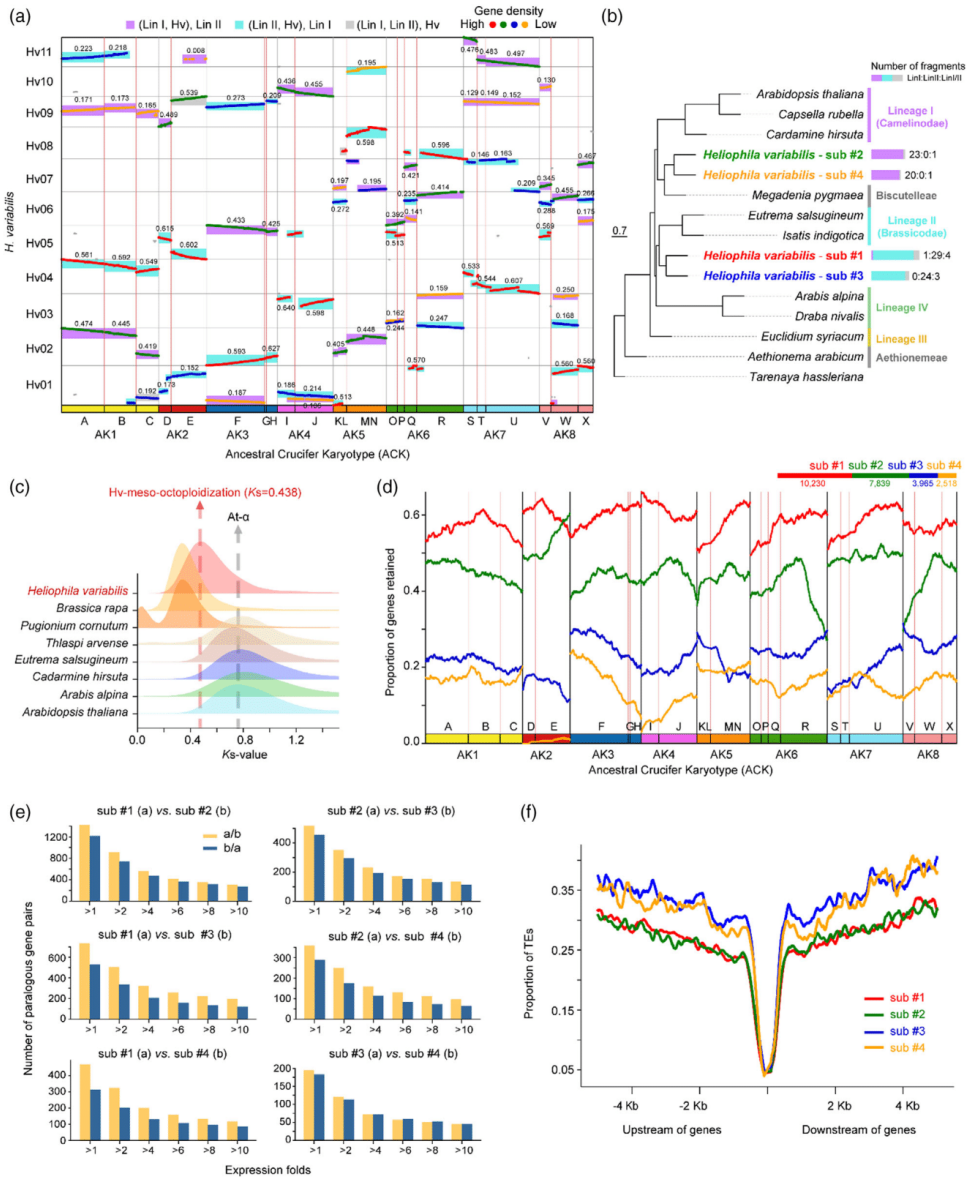
Genomic synteny analyses revealed large quadruple chromosomal regions in the *H. variabilis* genome (Figure 1b) and indicated a 4:1 correspondence between the *H. variabilis* and *Arabidopsis thaliana* genomes (Figure S6). To facilitate integrative analyses using both genomic and cytogenetic approaches, we identified 106 syntenic fragments between the *H. variabilis* genome and the Ancestral Crucifer Karyotype (ACK; $n = 8$). Four copies were detected for 19 of 22 ancestral genomic blocks (GBs; labeled A through X; Lysak et al., 2016; Schranz et al., 2006), while only three genomic copies were identified for Blocks D and H and two copies for Block G (Figure 2a). However, we were able to identify homeologous genes at the presumed locations of the missing blocks, indicating a loss of genomic synteny due to extensive gene loss in these regions. Therefore, each of the four subgenomes is a complete set of seven or eight ancestral chromosomes comprising 22 GBs. The association of the GBs was verified by a series of comparative chromosome painting experiments, which also allowed the detection of the four copies for most of the 22 GBs (Figure S7). Taken together, these results support an octoploid origin of the *H. variabilis* genome.

To test whether hybridization (allopolyploidy) contributed to the origin of the octoploid genome, we performed maximum likelihood (ML) and coalescent-based analyses separately for each of the 106 chromosomal fragments based on 23 503 low-copy genes, which allowed us to infer the phylogenetic placement of the putative progenitor species (Figure 2b). Despite the lack of major gene tree topologies, at least two phylogenetic origins of the ancestors of *H. variabilis* were supported: one that formed a clade with *Megadenia pygmaea* (tribe Biscutelleae), which was sister to the present-day crucifer Lineage I (supertribe Camelinodae), supported by 44 fragments (9331 genes), and the other being sister to the Lineage II (Brassicodae), supported by 53 fragments (13 069 genes). In nine genomic regions (1103 genes), we observed the placement of *H. variabilis* outside Lineage I + Lineage II (Figure 2a,b and Table S7), indicating a strong incongruence among gene tree topologies. These results suggest that the ancestral *Heliophila* genome had a hybrid origin involving progenitor genomes from divergent crucifer lineages. This is supported by the major peak in the K_S distribution of 7482 pairs of paralogs in *H. variabilis* at approximately 0.44 (Figure 2c), which is similar to the previously estimated synonymous divergence between Lineages I and II (Kagale et al., 2014).

Inference of at least three different ancestral genomes involved in the octoploidization

Comparison with the ACK genome revealed distinct levels of synteny loss (or gene fractionation) among the different copies of GBs, with syntenic gene density ranging from 0.008 to 0.640 (Figure 2a). To infer the genome structure of the four diploid ancestors of *H. variabilis*, we combined the evidence for synteny loss and phylogenetic affinity of the 106 GB fragments with ancestral genomes of either

Figure 2. Evolution of four subgenomes in the octoploid *H. variabilis* genome.
(a) The synteny map showing syntenic relationships between 11 chromosomes of *H. variabilis* and eight chromosomes and 22 genomic blocks of the Ancestral Crucifer Karyotype (ACK). Syntenic fragments are labeled according to their average gene density, numerically expressed above each fragment. The background coloring of the syntenic segments corresponds to the phylogenetic placements in (b).
(b) Coalescent-based phylogenetic tree including the four subgenomes of *H. variabilis* and 11 other crucifer genomes representing the major Brassicaceae lineages. Statistics on the phylogenetic topology of the 106 syntenic fragments shared between *H. variabilis* and ACK (see (a)) are shown to the right of the stacked bar charts.
(c) The K_S-value distribution of paralogous gene pairs in *H. variabilis* and other seven Brassicaceae species. At-α refers to the paleotetraploidization shared by the entire Brassicaceae family.
(d) The density of syntenic genes in four subgenomes of *H. variabilis* compared with ACK genome. The total number of genes contained in each subgenome is labeled on the upper right.
(e) The number of dominantly expressed genes in flower tissues between any two of the four subgenomes. See Table S9 for expression differences of other tissues analyzed.
(f) The average TE density at the vicinity of 24 552 genes distributed in the four subgenomes.



Lineage I or II (Table S7). We were able to sort the 106 genomic regions into four GB groups corresponding to four subgenomes (sub #1, sub #2, sub #3, and sub #4). The number of protein-coding genes was much higher in sub #1 (10230) and sub #2 (7839) than in sub #3 (3965) and sub #4 (2518) (Figure 2d and Table S8). Interestingly,

subgenome-specific K_S analyses based on pairs of homeologues revealed comparable distribution profiles (Figure S8) that were consistent with the overall K_S distribution (Figure 2c). Given the inferred phylogenetic placement (Figure 2b), it is highly likely that the more closely related progenitor genomes, that is, sub #1 and sub #3 *vs.* sub #2 and sub #4, diverged rapidly after the split between crucifer Lineage I and Lineage II approximately 25 million years ago (Mya; Kagale et al., 2014).

Comparison of gene expression of 7330 retained homeologue pairs between any two of the four subgenomes further corroborated the dominance of the less fractionated subgenomes, that is, sub #1 > sub #2 > sub #3 > sub #4 (Figure 2e and Table S9). There were also subgenomic differences in expression levels between 97 gene families with four complete copies (Figure S9). The density of TEs surrounding genes (2 kb upstream or downstream) was comparable between sub #1 and sub #2, both of which were lower than those in sub #3 and sub #4 (Figure 2f). Comparative chromosome painting analysis showed that GBs of the four homeologues consistently differ in their physical length and overall fluorescence intensity (Figure S10). Genomic blocks within sub #1 were on average 1.4, 2.6, and 3.3 times longer and 1.6, 4.7, and 8 times more fluorescent than homeologous GBs in sub #2, #3, and #4, confirming the biased fractionation of the four subgenomes.

We then compared the associations of GBs in the *H. variabilis* genome with those of three known ancestral karyotypes (ACK, ancPCK, and PCK; Geiser et al., 2016; Lysak et al., 2016; Mandáková & Lysak, 2008; Schranz et al., 2006). We observed 25 GB associations that were specific to one or the other ancestral karyotype, with the highest number of associations within sub #3 and sub #4 (13 in total; Figure 3 and Table S10). Since the main difference between the known ancestral karyotypes (ACK, ancPCK, and PCK) lies in the association of seven GBs (O, P, W, R, V, Q, and X), we were able to identify the involvement of the different ancestral genomes based on the position of these GBs on chromosomes Hv01, Hv03, Hv06, and Hv07 (Figure 3, see Supporting Information and Figure S11 for detailed reconstruction of chromosome origins).

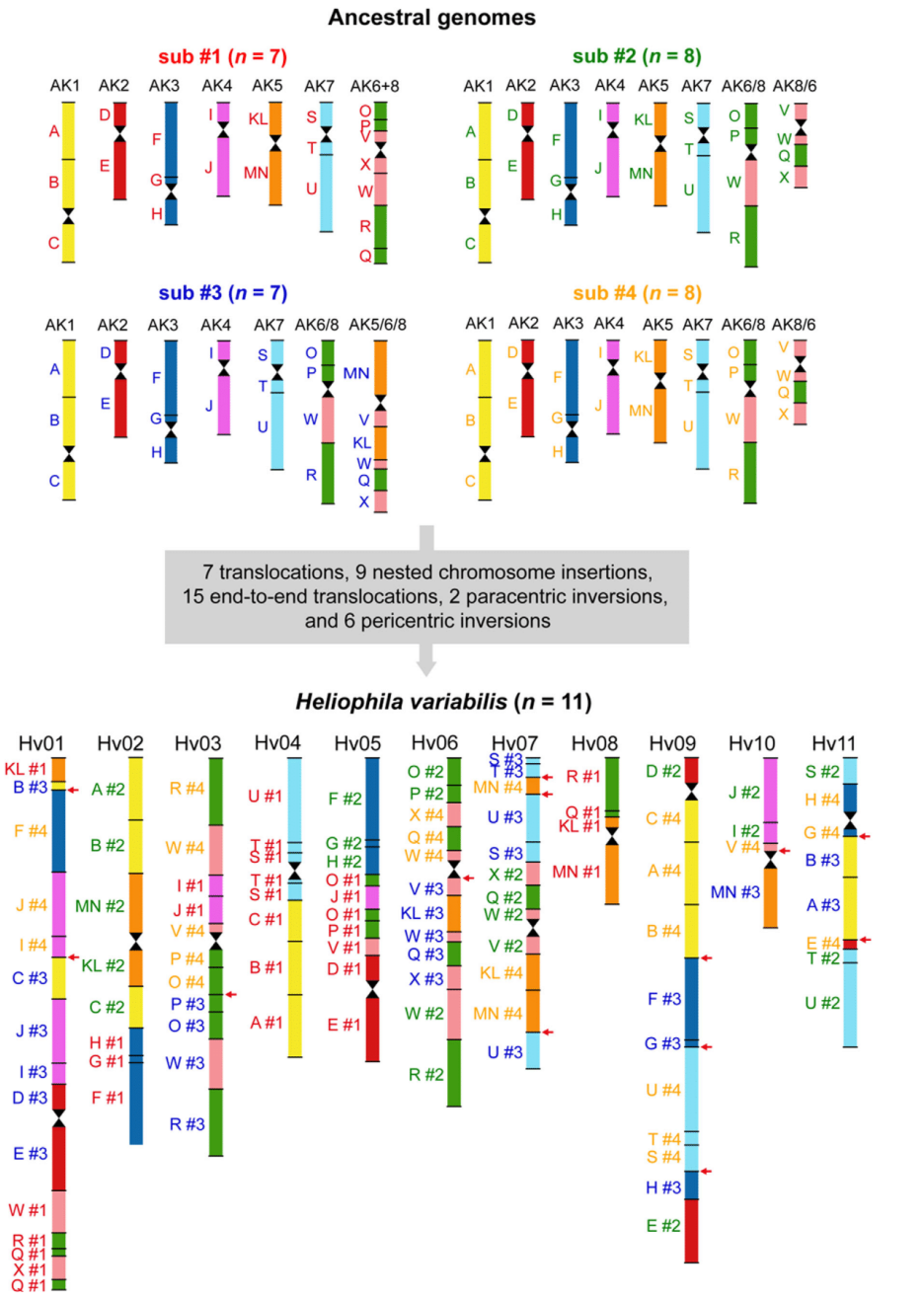
Despite an extensive restructuring after polyploidization (39 chromosomal rearrangements; Table S11), our analyses showed that the origin of the octoploid *Heliophila* genome included four progenitor genomes, that is, two ancPCK-like genomes ($n = 8$; sub #2 and sub #4), which were successively sister to the tribe Biscutelleae and Lineage I, a PCK-like genome ($n = 7$; sub #3), and an unknown genome ($n = 7$; sub #1), which both were closely related to the ancestral 7-chromosome genomes of Lineage II (Figure 2b). The inferred ancestral octoploid genome had at most 30 chromosome pairs ($n = 7 + 8 + 7 + 8$; i.e., $2n = 8x = \sim 60$), which were reduced to $n = 11$ in *H. variabilis* by extensive descending dysploidy. The allo-octoploid genome originated through hybridization between genomes with lower ploidies ($2x$, $4x$, or $6x$), either by hybridization between two tetraploid genomes ($4x \times 4x \rightarrow 8x$) or via a hexaploid bridge ($4x \rightarrow 6x \rightarrow 8x$). The first hypothesis is supported by the analysis of exonic deletions in each subgenome, with genes with such deletions in sub #1 and sub #2 having a comparable proportion (9.67% and 9.54%, respectively) in contrast to the slightly higher proportion of genes in sub #3 (11.51%) and sub #4 (11.46%; Table S12).

Relics of TE activities corroborate ancient hybridization events

To further characterize the four subgenomes as well as the origin of the octoploid *Heliophila* genome, we examined the abundance of genomic repeats and the insertion time of TEs. While the total length of repetitive sequences was associated with the level of gene fractionation (i.e., sub #1 > sub #2 > sub #3 > sub #4; Figure S12a), the proportions of different repeat categories were comparable among subgenomes, except for the most fractionated sub #4 (Figure S12b,c). The distribution of insertion times revealed distinct histories between two TE classes: The DNA transposons showed a large peak indicating transpositional bursts at ~18 Mya, whereas the LTR-RTs exhibited two apparent peaks that were mainly caused by the activities of *Gypsy* retroelements at ~8 Mya and ~ 22 Mya (Figure S13). In addition, pairwise comparisons between subgenomes showed a moderate-to-strong correlation (Pearson's $r > 0.5$) in the insertion times of all TE superfamilies except for *Gypsy* (Figure S14). Collectively, the TE landscape of *H. variabilis* suggested largely concerted subgenome evolution after large-scale TE insertion events that predated the divergence time (~12 Mya) of all *Heliophila* species or even the split between *Heliophila* and its sister genus *Chamira* (~18 Mya; Dogan et al., 2021; Hendriks et al., 2022; Walden et al., 2020), which may have experienced shared WGD(s) (Dogan et al., 2021).

Genomic repeats contain abundant phylogenetic signals for subgenome phasing (Gordon et al., 2019), but their effectiveness in mesopolyploid genomes has not been

Figure 3. Four reconstructed ancestral subgenomes of *H. variabilis* and their structure within the extant *H. variabilis* genome. Three subgenomes resemble previously inferred ancestral Brassicaceae genomes, namely ancPCK (sub #2 and #4, both $n = 8$) and PCK (sub #3, $n = 7$), whereas sub #1 ($n = 7$) could not be attributed to any of the previously inferred ancestral genomes. The color coding and capital letters (A to X) correspond to 22 genomic blocks shared among the four parental genomes and the 11 chromosomes of *H. variabilis*. Sub #3 and sub #4 formed the highest number of GB associations (red arrows).



The Plant Journal, (2023), **116**, 446–466

adequately explored. The numbers of 21-mer sequences shared between subgenomes were much lower than that of subgenome-specific ones (Table S13). Nevertheless, in 13 of the 22 GBs, clustering of shared 21-mer sequences resulted in the grouping of the less fractionated (sub #1 and sub #2) and the more fractionated subgenomes (sub #3 and sub #4), respectively (Figure S15). Although the observed pattern could be due to subgenome-biased sequence loss, it could alternatively support the two-step origin of the allo-octoploid *Heliophila* genome by hybridization between two allotetraploid progenitor genomes, one comprising subgenomes #1 and #2 and the second combining subgenomes #3 and #4.

Erosion of LTR retrotransposons contributes to post-octoploid genome shrinkage

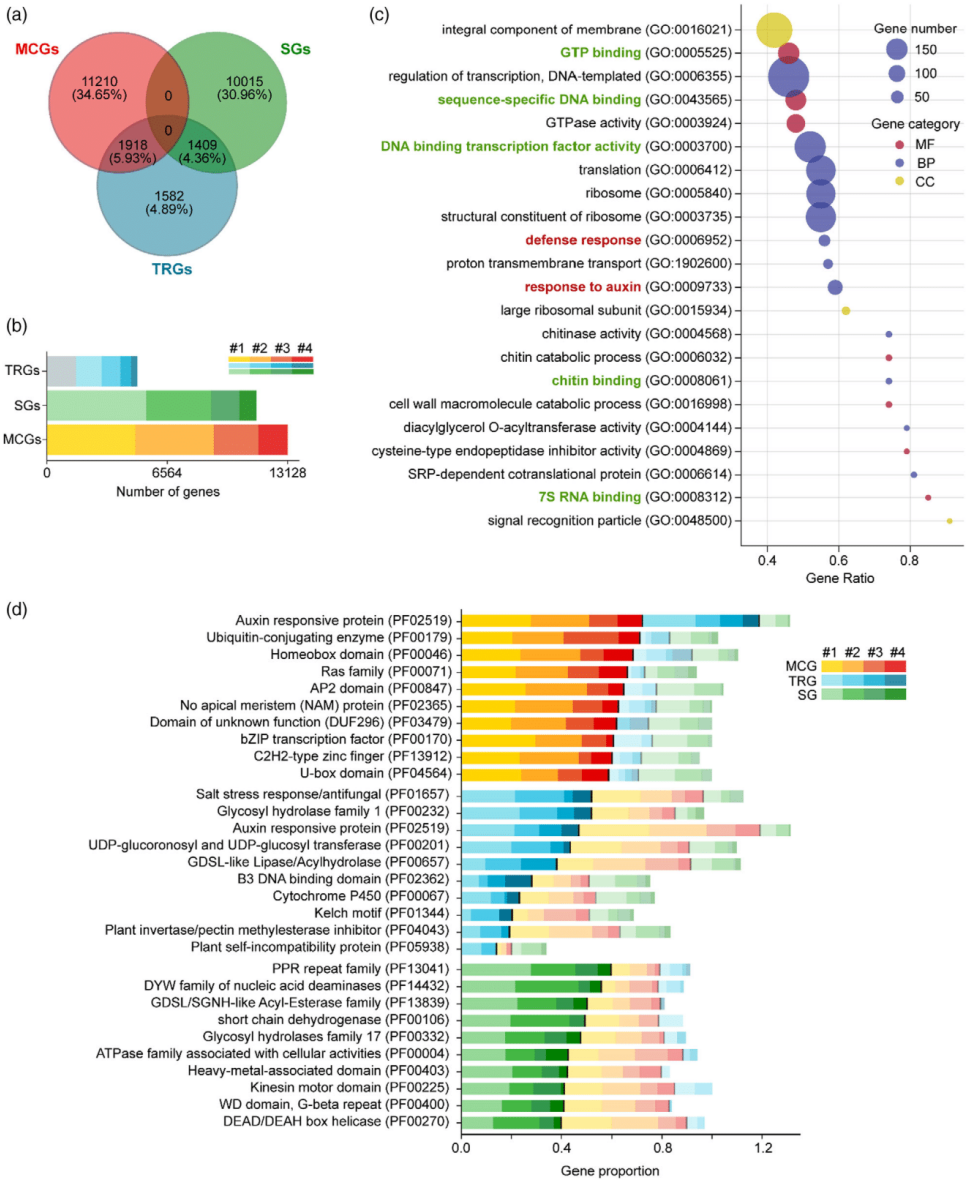
In contrast to the octoploid genome history, most *Heliophila* species have small genome sizes (GS, 288–484 Mb; Dogan et al., 2021), corresponding to the overall tendency of genome downsizing after WGD in Brassicaceae (Hohmann et al., 2015; Lysak et al., 2009). Assuming that the ancestral GS (^{anc}1C) for Brassicaceae was 0.5 pg (~490 Mb; Lysak et al., 2009), we could infer a huge GS decrease from ~1960 Mb in the octoploid ancestor to the size of 334 Mb in *H. variabilis*. The remarkably 5.87-fold change represents the largest extent of GS reduction among six crucifer genomes of different ploidy levels (Figure S16), followed by the mesohexaploid *B. rapa* (from ~1470 to ~356 Mb, 3.23-fold) and the diploid *A. thaliana* (from ~490 Mb to ~160 Mb, 3.06-fold). The numbers of both full-length LTR-RTs and solo LTRs in the *H. variabilis* genome were much lower than those of the remaining genomes, except for *A. thaliana* (Figure S17a). In addition, the ratio between solo LTRs and full-length LTR-RTs (S/F) was high in the *H. variabilis* genome (S/F = 8.38) despite a lack of correlation (Pearson's $r = 0.58$ with P-value = 0.23) between the S/F ratio and GS variation among the analyzed species. These results highlighted effective purging of LTR-RTs, possibly via ectopic recombination (Bennetzen, 2002; Devos et al., 2002; Grover & Wendel, 2010), as a mechanism for genome contraction in *Heliophila*. We also observed a discrepancy between subgenome-specific TE bursts (Figure S14b) and the turnover of LTR in the *H. variabilis* genome (Figure S17b), suggesting that there was a time lag between the ancient WGDs and subsequent radiation with genome downsizing (Hohmann et al., 2015; Schranz et al., 2012) and/or that selection may have affected genome composition (Cheng et al., 2018).

Biased gene retention, functional divergence, and compatibility in the four subgenomes

To investigate the preferential retention of genes after WGDs, we identified 13 128 (40.58%) multiple-copy genes (MCGs, i.e., with homeologues retained in at least two subgenomes) and 11 424 (35.32%) singleton genes (SGs, i.e., genes specific to one subgenome) in the 106 syntenic fragments (Figure 4a,b), which could be classified into 16 060 syntenic gene families. Of these genes, 3327 (13.55%) were also present as tandem repeat genes (TRGs; Figure 4a). The function of MCGs was associated with protein binding and defense responses, as reflected by the over-representation of GO terms such as '7S RNA binding', 'chitin binding', 'defense response', and 'response to auxin' (Figure 4c), and SGs were associated with DNA activities such as 'binding', 'repair', and 'replication' (Figure S18). In contrast, TRGs were enriched in GO terms related to stress responses, including 'response to osmotic stress' and 'defense response' (Figure S19). Analyses of PFAM domains further corroborated the different functional compositions between gene categories. For example, MCGs and TRGs retained the highest proportions of 'auxin-responsive protein' (PF02519, 72%) and 'salt stress response/antifungal' (PF01657, 68%) domains, respectively (Figure 4d and Table S14). However, the protein domains associated with response to environmental changes were not identified among SGs, which preferred to retain domains associated with organelle biosynthesis and regulation (Figure 4d). The distinct functions of the retained genes were largely consistent with the duplicability of genes in angiosperms (Li et al., 2016) and suggested that some of them may be dosage-sensitive genes (Birchler & Veitia, 2012; Papp et al., 2003).

To explore the extent of functional divergence among retained genes, we compared the presence and absence of 2248 protein domains identified in 13 128 MCGs that comprised 5504 syntenic gene families. Whereas the same domains were found among syntelogs of 4682 (85.07%) gene families, different domain composition was detected

Figure 4. Retention characteristics of genes with different amplified forms.
(a) A Venn diagram showing the number and proportion of multiple-copy genes (MCGs), singleton genes (SGs), and tandem repeat genes (TRGs).
(b) The numbers of retained MCGs, SGs, and TRGs in the four subgenomes. The gray bar for TRGs represents retained genes not attributed to any subgenome.
(c) GO enrichment for MCGs (Figures S18 and S19 for SGs and TRGs, respectively). The x-axis represents the ratio between the number of enriched genes to the number of background genes. Circles of different sizes correspond to the number of genes enriched, whereas the colors represent molecular function (MF), biological process (BP), and cellular component (CC) gene categories, respectively. The GO terms related to substance binding and biotic or abiotic stress are highlighted in green and red, respectively.
(d) Over-represented conserved protein domains in MCGs (red), TRGs (blue), and SGs (green). Different transparencies represent the four subgenomes of *H. variabilis*, which is consistent with the color legend in (b).



between gene counterparts of the remaining 822 (14.93%) families, likely reflecting sub/neofunctionalization of genes. A total of 571 domains were identified that were variable among homeologues, of which 13 domains from 45 families were not shared between genes. Interestingly, close examination of domain functions revealed that genes that showed evidence of sub/neofunctionalization were frequently associated with 'response to stress'. For example,

27 (45%) of the 60 domains that exhibited high variability (i.e., were variable in more than 50% of gene families carrying the same domain) were involved in 'resistance' (Table S15).

The highly diploidized (fractionated) genome of *H. variabilis* suggests that the proteins encoded by the four subgenomes may be functionally incompatible. To test this hypothesis, we analyzed the protein–protein interaction (PPI) network of *H. variabilis* proteins based on a map of experimentally validated interactions of their syntenic orthologs in the *A. thaliana* genome (McWhite et al., 2020). Based on the high-confidence PPI network (CF-MS score > = 0.70), we identified a total of 18 153 edges (interactions) between 1099 nodes (genes) in *H. variabilis* that formed 90 independent interacting clusters (Figure S20a and Table S16). Multiple-to-multiple interactions (14 133 edges between 768 MCGs) accounted for the majority of edges, which was significantly more than the other categories (*t*-test, *P*-values = 0; Figure S20b). Interestingly, we observed mixed node associations between subgenomes, favoring interactions with proteins from multiple subgenomes. For example, we detected 973 nodes involving at least three subgenomes, whereas only 126 nodes were connected proteins from one or two subgenomes (Figure S20c). Thus, protein interactions in allopolyploid genomes appeared to be compatible in general regardless of ploidy levels (Hao et al., 2021) and the phylogenetic affinity between subgenomes. In addition, interactions between more distantly related subgenomes tended to be more favored than those between more closely related subgenomes. For example, we observed more interactions between sub #1 and sub #2, and between sub #3 and sub #4 (70 nodes) than between sub #1 and sub #3, and between sub #2 and sub #4 (24 nodes, Figure S20c).

Post-polyploid diploidization shaped the stress-response regimen of *H. variabilis*

Given the extensive genome diploidization, we next examined the extent to which this evolutionary process contributed to the adaptation of the short-lived *Heliophila* to the winter-rainfall desert. To this end, we focused on eight categories of candidate genes, including 123 genes in 47 gene families, that could be responsible for adaptive morphological, physiological, and biochemical traits (Figure 5). In contrast to the excessive retention of genes associated with seed dormancy and response to stress (proportion of genes retained: 71.6%), most genes related to early flowering and leaf development underwent extensive deletion (proportion of genes retained: 32.3%; Figure 5 and Table S17).

Notably, genes encoding the transcription factor *GROWTH-REGULATING FACTOR5* (*GRF5*) were completely lost in *H. variabilis* (Table S18), whereas orthologs of *GRF5* were identified in other mesopolyploid genomes, including *B. rapa* (three orthologs) and the psammophyte *P. cornutum* (one ortholog; Table S18). The *GRF5* genes promote leaf growth and play a major role among members of the *GRF* family in Arabidopsis (Horiguchi et al., 2005; Kim et al., 2003; Kim & Lee, 2006). Overexpression of *GRF5* increases leaf area, whereas downregulation of *GRF5* results in the formation of narrower leaves with fewer cells in the *grf5* mutant (Gonzalez et al., 2012). The complete loss of *GRF5* could contribute to the unique pinnately divided leaves of *H. variabilis* (Figure S21), reducing leaf area and reducing water loss. In addition, the remaining *GRF* genes in *H. variabilis* were characterized by the presence of non-canonical motifs (Figure S22), increased mutation in their canonical motifs compared with other species (Figure S23), and low expression levels (TPM-values <20) across different tissues (Figure S24). Taken together, these results suggest that post-polyploid diploidization had a multifaceted effect on leaf morphogenesis in *H. variabilis* (see Supporting Information for detailed analyses of the *GRF* gene family).

Previous studies have identified strongly antifungal peptides (AFPs) in *Heliophila coronopifolia* (De Beer & Vivier, 2011; Weiller et al., 2017), suggesting that *Heliophila* species may have evolved enhanced chemical defenses during plant growth. Here, we identified eight *AFP* genes in *H. variabilis*, including two arrays of tandem duplicates in sub #1 (Hv05) and sub #3 (Hv01) that are both syntenic with *AT1G75830* in Arabidopsis (Figure S26). Comparative phylogenetic analyses that included sequences from other crucifer species confirmed the distinct origins of the duplicated genes in *H. variabilis* (Figure S27). The number of *AFP* genes varied between genomes regardless of ploidy level (Table S19). The AFP proteins share a highly conserved *β2-β3* loop consisting of four cystine bridges that are essential for maintaining the secondary structure of the protein (Schaaper et al., 2001). Interestingly, this structure appeared to be well-conserved in *H. variabilis* AFPs, as only four amino acid mutations were found in one of the eight peptides (Figure S27 and Table S20), suggesting that most of the *AFP* genes retained in the *H. variabilis* genome were functional (Supporting Information). The boosted chemical defense in *H. variabilis* was also supported by the retention/amplification of several duplicated genes related to the formation of endoplasmic reticulum (ER)-derived structures (ER bodies) that accumulate β-glucosidases/myrosinases (Stefanik et al., 2020), including *NAI2* and its paralog *TONSOKU-ASSOCIATING PROTEIN 1* (*TSA1*), which is derived from the At-α event (De Beer & Vivier, 2011), and *BGLU18*, a neighboring gene of *TSA1* (Figure S28). Again, the evolution of these genes was accelerated by the process of post-polyploid rediploidization, with the number of exons and total length of CDSs varying strongly with the presence or absence of core motifs (Figure S29 and Table S21, and Supporting Information).

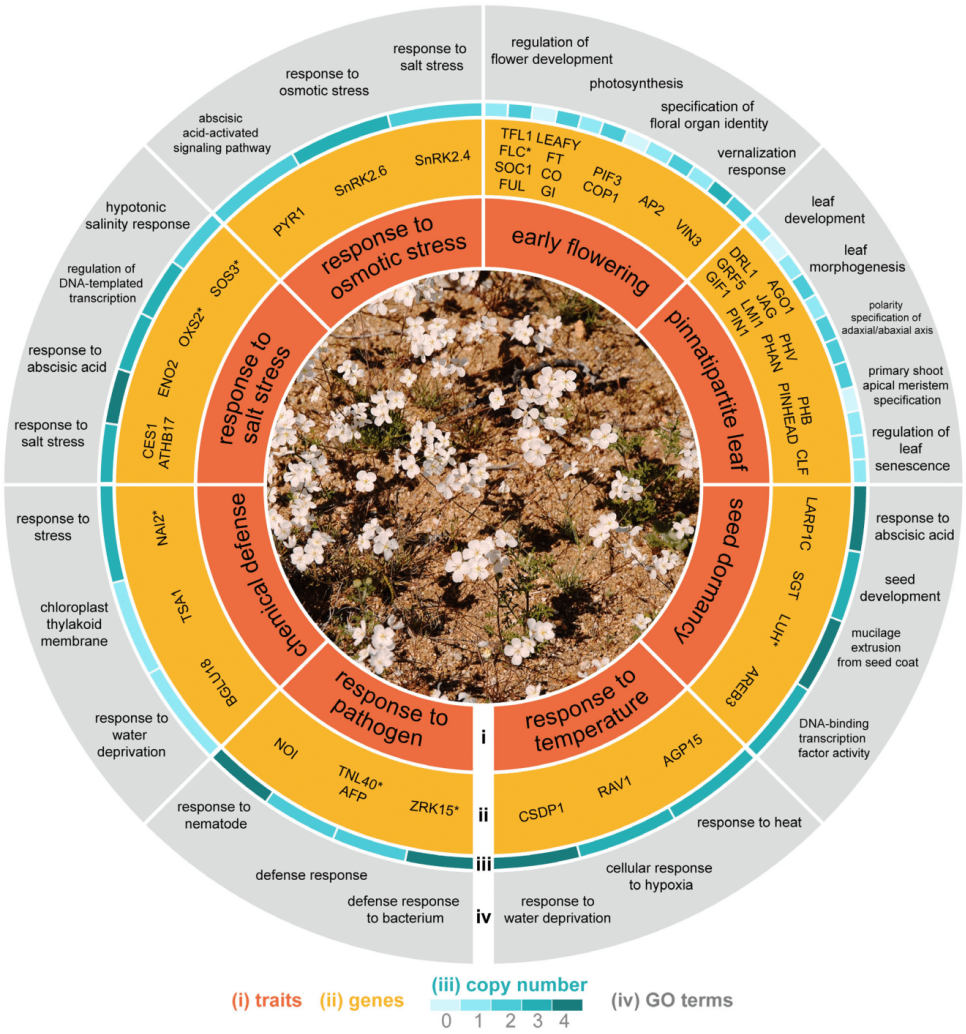


Figure 5. Impact of post-polyploid diploidization on different types of genes accounting for the adaptive diversification in *H. variabilis*. (i) The eight typical traits of *H. variabilis*. (ii) Examples of candidate genes. The asterisk to the right of the gene name (*) indicates that the candidate gene in *H. variabilis* has undergone sub/neofunctionalization after the octoploidization. (iii) The copy number (from 0 to 4) of candidate genes after the octoploidization. (iv) GO terms of the candidate genes.

DISCUSSION

Whereas auto- and allotetraploids are common in angiosperms, hexaploids are less common and polyploids of higher ploidy levels are even rarer. Accordingly, the number of octoploid plant genomes sequenced and annotated to date is small. The 334-Mb *H. variabilis* genome is the first octoploid crucifer genome to be sequenced at the chromosome level. Only three other octoploid angiosperm genomes have been sequenced and assembled to date. The allo-octoploid strawberries (*Fragaria* spp., $2n = 8x = 56$; Edger et al., 2019) are the most analyzed octoploid plant genomes, while octoploid origins have also

been detected for sugarcane species (*Saccharum officinarum*, 2*n* = 8*x* = 80; *S. spontaneum*, 2*n* = 8*x* = 64; Zhang et al., 2018) and the beak-sedge *R. pubera* with holokinetic chromosomes (2*n* = 10; Hofstatter et al., 2022).

While the eight chromosome sets have remained stable in strawberry and sugarcane species, the 30 chromosome pairs in *Heliophila* have undergone a 2.7-fold reduction to only 11 chromosome pairs (*n* = ~30 → *n* = 11). The different degree of diploidization and descending dysploidy could be largely due to the different ages of these octoploid genomes. In *R. pubera*, the second WGD (4*x* → 8*x*) occurred 2.1 Mya, and the 1.6-fold descending dysploidy (*n* = 8 → *n* = 5) might have been facilitated by holocentricity (Hofstatter et al., 2022). The origin of the octoploid strawberry and sugarcane genomes has been dated to approximately 1 Mya and <1 Mya, respectively (Pompidor et al., 2021). In contrast, the ancestral octoploid *Heliophila* genome originated before the divergence of the genus in the Miocene (at least 12 Mya; Dogan et al., 2021; Hendriks et al., 2022; Walden et al., 2020). Given the at least partially shared allopolyploid origin with the sister genus *Chamira* (Dogan et al., 2021), the hybridization-facilitated genome merger events could be dated earlier than the *Chamira/Heliophila* split (c. 18 Mya; Dogan et al., 2021; Hendriks et al., 2022; Walden et al., 2020), which is associated with the remnants of ancient TE expansion in the *H. variabilis* genome. Consequently, rediploidization of the meso-octoploid genome may have been a protracted process spanning more than 12 million years and accompanied by adaptive radiation and infrageneric cladogenesis, broadly associated with the establishment of a summer-dry climate in the Greater Cape Floristic Region in the mid-to-late Miocene (Van Santen & Linder, 2020; Verboom et al., 2009).

With few exceptions, allo- and autopolyploid plant genomes show distinct patterns of gene fractionation (Burns et al., 2021; Garsmeur et al., 2014; Sun et al., 2017). In *Brassica* species, biased gene fractionation provides evidence for a two-step origin of their mesohexaploid ancestral genome (Cheng et al., 2018; Tang et al., 2012). Also in *H. variabilis*, two less fractionated subgenomes (sub #1 and sub #2) and two more fractionated subgenomes (sub #3 and sub #4) suggest an allopolyploid origin of the ancestral *Heliophila* genome (*n* = ~30), probably by hybridization between two (*n* = 15) tetraploids (Figure 6). Analyses of several mesopolyploid crucifer genomes have revealed their allopolyploid origin including distant hybridizations between species from different, early-diverged lineages/supertribes (Dogan et al., 2021; Guo et al., 2021; Hu et al., 2021; Mandáková & Lysak, 2018). The monogeneric Heliophileae and nine other tribes most likely of mesopolyploid origin (Dogan et al., 2021; Hendriks et al., 2022; Mandáková et al., 2017) form the evolutionary lineage V (Nikolov et al., 2019) or the supertribe Heliophilodae

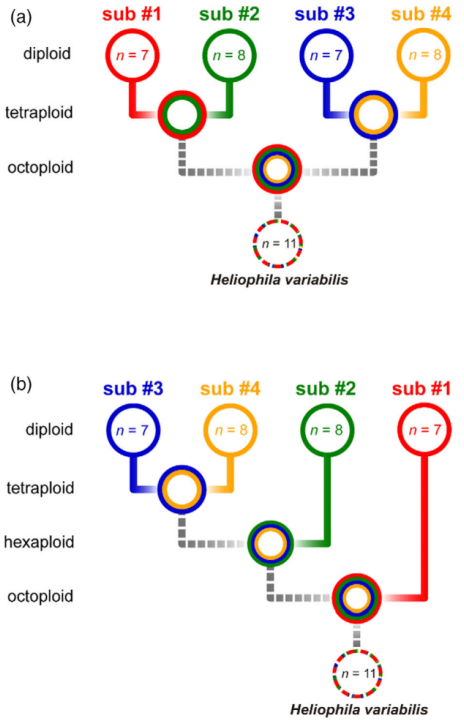


Figure 6. Two models of the origin of the octoploid *H. variabilis* genome. The allo-octoploid ancestral genome (*n* = ~30) was either formed by hybridization between two tetraploid (*n* = ~15) genomes (a) or via a hexaploid (*n* = ~23) (b). Dashed lines indicate post-polyploid genome diploidization and the diploidized genome of *H. variabilis*.

(Hendriks et al., 2022). Similar to tribe Biscutelleae (Guo et al., 2021), both tetraploid parental genomes of the ancestral octoploid *Heliophila* genome originated through distant intertribal hybridization, merging an unknown *n* = 7 genome (sub #1) and an ancPCK-like genome (*n* = 8; sub #2), as well as a PCK-like genome (*n* = 7; sub #3) and an ancPCK-like genome (*n* = 8; sub #4) (Figure 3). While more ancestral ancPCK-like genomes were phylogenetically close to Lineage I (Camelinodae) and the tribe Biscutelleae (Heliophilodae), both 7-chromosomal diploid genomes belonged to Lineage II (Brassicodae) (Figure 2b). The inferred origin of the meso-octoploid *Heliophila* genome highlights the importance of distant hybridizations and WGDs for cladogenesis, adaptive radiation, and colonization of new habitats, including long-distance dispersals. Since both the genus *Heliophila* and its monotypic sister genus *Chamira* are restricted to southern Africa, and their

putative ancestral genomes are distributed in northern Africa, the Mediterranean, and southwestern Asia, it is likely that the ancestral octoploid genome reached southern Africa via long-distance dispersal or by migration through heterogeneous Miocene habitats (e.g., wooded grasslands) of eastern Africa (Peppe et al., 2023). Subsequent diploidization has facilitated the diversification of *Heliophila* species in newly invaded habitats in southern Africa. The CFR is one of the world's biodiversity hotspots with a high proportion of endemic taxa, particularly in plants (9000 species, 70% endemic; Chang et al., 2023; Goldblatt & Manning, 2002; Mittermeier et al., 1998; Myers et al., 2000). The assembled mesopolyploid genome of *H. variabilis* has shown that polyploidization and post-polyploid diploidization have played an important role in the ecomorphological divergence and adaptation (e.g., Mandáková et al., 2012; Mummenhoff et al., 2005) of *Heliophila* species and that polyploidization–diploidization cycles may be more important for the diversity of the Cape flora than previously thought (Oberlander et al., 2016).

Most *Heliophila* species are short-lived herbaceous plants adapted to semidesert conditions, completing their life cycle in a short period after seasonal rainfalls. Typical traits of these ephemeral plants include early flowering, extremely small seeds that can survive a long dormant period, and narrow leaves that correspond to the drought escape strategy (Shavrukov et al., 2017). Interestingly, candidate genes associated with these adaptations showed different fates after duplication. On the one hand, we discovered loss-of-function changes in genes associated with leaf development and early flowering in *H. variabilis* (Figure 5), including the complete loss of *GRF5*, a regulator of cell proliferation and leaf size (Horiguchi et al., 2005), as well as the accelerated accumulation of mutations on the core motifs of other *GRF* genes (Figure S23), which could affect leaf cell size and quantity (Debernardi et al., 2014). On the other hand, genes associated with biotic/abiotic stress are over-represented and sequentially conserved in the *H. variabilis* genome, including *AFP* genes involved in pathogen response and *NAI2/TSA1-BGLU18* involved in chemical defense (Figure 5 and Supporting Information). More importantly, we found evidence for frequent sub/neofunctionalization of these resistance genes during post-polyploid diploidization (Table S15), which may allow for an enhanced defense system adapted to challenging environmental conditions.

EXPERIMENTAL PROCEDURES

Genome sequencing

Seeds and specimens of *H. variabilis* were collected in the Northern Cape, South Africa (Namaqualand, Springbok, Goegap Nature Reserve, 29°40′17″S, 17°59′55″ E). Herbarium specimens were deposited in the Compton Herbarium (South African National Biodiversity Institute; Kirstenbosch, South Africa, NGS224) and the herbarium of Masaryk University (BRNU 681243). Plants were grown from seed and cultivated under standard conditions in growth chambers (21/18°C, 16/8 h of light/dark) or a greenhouse (22/19°C, 16/8 h of light/dark).

High-molecular-weight genomic DNA was extracted from the isolated nuclei for sequencing library construction. The Illumina HiSeq Xten (Illumina, San Diego, CA, USA) was used for genome sequencing. Illumina sequencing libraries were prepared using the TruSeq Nano DNA HT Sample preparation kit (Illumina, USA) following the manufacturer's recommendations. A total of 161.58 million reads (~48.41 Gb, ~161× coverage of the assembled genome) were obtained (Table **S1**), of which 96.8% had base quality values above 20 and 91.5% above 30.

The long-read sequencing process followed the standard protocol of Oxford Nanopore Technologies (ONT; Deamer et al., 2016). DNA libraries with a mean fragment length of >20 kb were constructed for ONT sequencing and sequenced on the PromethION platform (Oxford Nanopore Technologies, Oxford, UK). Nanopore sequencing yielded a total of 55.92 GB of raw data and 51.63 GB of clean data (~172×) after quality control (Table **S1**).

Hi-C libraries were constructed according to the suggested procedure (Lieberman-Aiden et al., 2009). Briefly, leaf samples were fixed with formaldehyde solution before chromatin extraction, and chromatin was digested with 100 units of the restriction enzyme HindIII at 37°C. The DNA ends were labeled with biotin, and DNA ligation was performed with T4 DNA ligase (NEB, Ipswich, MA, USA). After ligation, proteinase K was added for reverse cross-linking. The DNA fragments were then purified and dissolved. The purified DNA was fragmented to 300–600 bp, and the DNA ends were repaired. After a quality check, the Hi-C libraries were sequenced on Illumina HiSeq X-Ten instruments to generate 150-base paired-end reads. A total of 50.06 GB of clean data were obtained, of which 94.3% had base quality values above 30 (Table **S1**).

Four different types of organs and tissues of *H. variabilis*, including flowers, leaves, stems, and siliques, were collected for transcriptome sequencing. The eukaryotic mRNA was enriched with Oligo (dT) by magnetic beads, and the mRNA was randomly interrupted by adding fragmentation buffer. The effective library concentration was accurately quantified by Q-PCR to ensure library quality. High-throughput sequencing was then performed using NovaSeq 6000 (Illumina, USA) with a sequencing read length of PE150. Approximately 10 Gb of clean data were generated for each sample, resulting in a total of ~40 Gb of clean data (Table S4).

Genome assembly and quality assessment

Genome size, heterozygosity ratio, and repeat sequence ratio were assessed by *K*-mer distribution analysis ($K = 21$) using Illumina short reads. *De novo* genome assembly was performed based on Nanopore long reads using NextDenovo (v2.2-beta.0; https://github.com/Nextomics/NextDenovo) with the standard pipeline. The following parameters were used in the assembly: read_cutoff = 5 k, seed_cutoff = 43 000, genome_size = 300 m, sort_options = −m 5 g -t 8, minimap2_options_raw = −x ava-ont -t 8. Finally, 71 contigs were obtained. After the completion of the pre-assembly (297 Mb), three iterative rounds of polishing were performed with the cleaned Illumina reads using Pilon (v1.23; Walker et al., 2014). The sizes, contig numbers, and contig N50 values of the draft genome assemblies are summarized in Table S2. Completeness of genome assembly was assessed using the

embryophyta_odb10 database of BUSCO (v3.0.1; Simão et al., 2015) software (1614 total BUSCOs).

Genome scaffolding based on Hi-C data

To further improve the contiguity of the assembly, the Hi-C reads were mapped to the assembled contigs using BWA (v0.7.17; Li & Durbin, 2009). These mapped datasets were submitted to Juicer (v1.5.7; Durand et al., 2016) and 3ddna (v180922; Dudchenko et al., 2017) software for grouping and ordering. Two of the 71 raw contigs were manually separated, yielding 77 contigs. These contigs were linked into scaffolds using ALLHIC (v0.9.8; Zhang et al., 2019). The linkage results were also manually curated to correct misjoins and misassemblies based on visualization using JuiceBox (v1.1.08; Robinson et al., 2018).

Gene annotations

Protein-coding genes were predicted using an evidence-based annotation workflow by integrating different sources of evidence. *De novo* gene predictions were generated using AUGUSTUS (v3.4.0; Stanke et al., 2006). For this purpose, a *Heliophila*-specific AUGUSTUS gene model was trained using GeneMark-ET (v4.0; Lomsadze et al., 2014) with the following parameters: --et_score 5 --min_contig 10 000 --max_intron 30 000 --max_gap 2000 --max_intergenic 1 000 000. GeneMark-ET uses RNA-Seq evidence as training data and performs two rounds of iterative gene predictions to train the model parameters. The 2000 gene models with the highest scores were used as training data for AUGUSTUS. The resulting gene models were then used to predict the coding genes using AUGUSTUS (–gff3 = on -hintsfile = hints.gff -extrinsicCfgFile = extrinsic.cfg -allow_hinted_splicesites = gcag,atac -min_intron_len = 30 -softmasking = 1). For mRNA-seq-based prediction, we used mRNA-seq datasets from flower, leaf, silique, and stem tissues of *H. variabilis* to assist gene prediction. Hisat2 (v2.1.0; Kim et al., 2015) was used to map RNA reads to the *H. variabilis* genome. Then, the primary transcripts were assembled using StringTie (v2.1.4; Pertea et al., 2015) without guide reference annotation for each tissue individually. The GTF files from the transcript assembly of each tissue were then merged using StringTie. TransDecode (v5.5.0; https://github.com/TransDecoder/TransDecoder) was used to identify the potential coding regions in the resulting transcripts. Meanwhile, RNA-Seq reads were *de novo* assembled into transcripts with Trinity (v2.11.0; Haas et al., 2013) using the genome-guided mode, and PASApipeline (v.2.3.3; Haas et al., 2003) was used for gene prediction from these transcripts. For annotation of homologs, protein sequences from *A. thaliana* and *B. rapa* were aligned to the *H. variabilis* genome to identify the homologous genes using Exonerate (v2.4.0; http://www.ebi.ac.uk/~guy/exonerate/). Finally, all gene predictions were integrated into a final gene annotation set using EVidenceModeler (v1.1.1; Haas et al., 2008; parameters: -segmentSize 1 000 000 -overlapSize 100 000) after removing pseudogenes and non-coding genes using a custom Perl script. A total of 32 351 genes were identified with a total CDS length of 34.45 Mb.

Identification of syntenic genes and fragments

The syntenic orthologs of *H. variabilis* genes in ACK genome (Lysak et al., 2016; Schranz et al., 2006) and sequenced Brassicaceae genomes were determined based on both sequence similarity and sequence homozygosity of their flanking genes using SynOrths (Cheng et al., 2012) with *H. variabilis* as the query genome and each of the other Brassicaceae genomes as the subject genome. An algorithm similar to that implemented in SynOrths was used to identify syntenic gene pairs (paralogs) within the genome. For each of the genes in the *H. variabilis* genome, up to three syntenic paralogs (corresponding to four copies of genes generated by octoploidization) were chosen from the sorted candidate pairs of syntenic paralogs, which were sorted from high to low based on their sequence homology and the support ratio of their flanking genes. A total of 23 520 syntenic genes were identified by comaparing *H. variabilis* and ACK genomes. Tandem repeat arrays were also identified using SynOrths. Each tandem repeat array consisted of continuously distributed homologous genes (E-value $<1.0 \times 10^{-5}$) and was not allowed to be interrupted by more than six non-homologous genes. Syntenic gene pairs that were continuously distributed along the *H. variabilis* genome and 22 genomic blocks (A to XLysak et al., 2016; Schranz et al., 2006) were considered as ancestral fragments inherited from the progenitors. Due to factors of local structural variation and genome assembly errors in the *H. variabilis* genome, local syntenic gene pairs may not be distributed immediately adjacent to other syntenic genes. Thus, when two syntenic gene pairs were interrupted by fewer than 50 genes or had a distance of less than 300 kb, they were combined into a pair of syntenic fragments. Finally, we identified 106 syntenic fragments in the *H. variabilis* genome corresponding to the 22 conserved macrosynteny GBs. In addition, we identified pericentromeres on each chromosome based on the regions where syntenic gene density dramatically decreased. The coordinates of the two ends of each pericentromere were defined by the start and end of adjacent syntenic gene fragments.

Phylogenetic analysis

To infer the ancestral origin of the 106 genomic fragments, gene trees were inferred using maximum likelihood (ML) and then subjected to coalescent-based analysis for each fragment. Homologous gene groups were called using the Broccoli pipeline (v1.2; Derelle et al., 2020) for amino acid (AA) sequences of *H. variabilis* and 10 crucifer species with publicly available genomes (*Aethionema arabicum*, *Arabidopsis thaliana*, *Arabis alpina*, *Capsella rubella*, *Cardamine hirsuta*, *Draba nivalis*, *Euclidium syriacum*, *Eutrema salsugineum*, *Isatis indigotica*, and *Megadenia pygmaea*), including *Tarenaya hassleriana* (also known as *Cleome hassleriana*) as an outgroup. Next, we filtered for low-copy genes groups that contained (1) at least one *H. variabilis* gene, (2) no more than eight genes in other species, and (3) no more than 60 genes in a single group. For each gene group, AA sequences were aligned using MAFFT (v7.427; Katoh et al., 2002) with parameters '--genafpair --maxiterate 1000'. Alignments were trimmed and back-translated into coding sequence alignments using TrimAL (v1.4; Capella-Gutiérrez et al., 2009) with parameters '-automated1 –split-bystopcodon'. Maximum likelihood analyses were performed using IQ-TREE (v1.6.11; Nguyen et al., 2015) with parameters '-st CODON -bb 1000 -nt 2 –bnni', which used a codon-based substitution model with 1000 replicates of ultrafast bootstrap to obtain branch support (Hoang et al., 2018). For gene groups with multiple *H. variabilis* genes, we performed an individual ML analysis for each gene, while keeping the genes of the remaining species the same. Coalescent-based inference of the species tree was performed using ASTRAL-Pro (v1.1.3), which allows quartet-based species-tree inference in the presence of paralogy (Zhang et al., 2020).

Inference of genome structure and four subgenomes

Using the inferred eight chromosomes of the ACK genome (Lysak et al., 2016; Schranz et al., 2006) as the reference, the four copies of each diploid chromosome in the octoploid ancestor of

H. variabilis were reconstructed. Then, three rules were followed to reconstruct the four *H. variabilis* subgenomes: (i) Each reconstructed chromosome in each of the four subgenomes has no overlapping or redundant genomic regions; (ii) the phylogenetic relationship of the syntenic gene fragments in the 22 GBs of *H. variabilis* was consistent with each two fragments corresponding to the present-day crucifer Lineage I and Lineage II (Hendriks et al., 2022); and (iii) for each of the ancestral chromosomes, the reconstructed chromosome always follows a specific distribution of gene density, corresponding to the distribution pattern of the four subgenome gene densities (sub #1 > sub #2 > sub #3 > sub #4). Based on the above rules, the 106 genomic fragments can be divided into four groups. We have designated the least fractionated group as sub #1 and the remaining three groups in order of the extent of gene fractionation as sub #2, sub #3, and sub #4.

Exonic deletion analysis

Based on the abovementioned alignments of coding sequences (see part of Phylogenetic analysis), we counted the number of exonic deletions and deleted bases within the *H. variabilis* genes by searching for legacies of short direct repeats. Following the methods of Tang *et al.*, we focused on deletions that are >30 bases (Tang et al., 2012). In both 25-bp sequences upstream (Pool_A) and downstream (Pool_B), the deletions in the alignments were extracted using a custom Perl script. For each gene, the Pool_B sequences were aligned against the Pool_A sequences by Blastn search (Boratyn et al., 2013) under the 'blastn-short' mode with a word size of 7 bp, whereby a resulting hit with 'plus/plus' strand orientation was treated as evidence for the presence of direct repeats. Accordingly, the deletions flanked by direct repeats were recorded as exonic deletions and summarized for each subgenome.

K-mer distribution

Each subgenome exhibits a characteristic pattern of past activity of TEs recorded by a particular combination of enriched *K*-mers. We expect that each subgenome will share repetitive content inherited from their respective common progenitors (Figure S15; Gordon et al., 2019). Operationally, we determined the distribution of 21-mers that are specific to each of the subgenomes and shared by any two of the four subgenomes comprising 22 GBs. The 21-mer frequencies on each of subgenomes were counted using Jellyfish (v1.1.10; Marçais & Kingsford, 2011). An in-house python script was then used to select repeated 21-mers (frequency >1 on each subgenome and shared between at least two subgenomes). *K*-mer distribution mapping and enrichment was performed using the pheatmap package (v1.0.12; http://rpackages.ianhowson.com/cran/pheatmap/).

Genome size variation analyses

We compared the genomic components of *H. variabilis* and other five sequenced crucifer species with different ploidy levels (*Arabidopsis thaliana*, diploid; *Arabis alpina*, diploid; *Thlaspi arvense*, diploid; *Brassica rapa*, mesohexaploid; and *Pugionium cornutum*, mesotetraploid). The download information of these crucifer genomic data is listed in Table S22. We extrapolated GS values into genome sizes in megabase pairs following with 1 pg DNA = 980 Mb (Doležel et al., 2007). The inferred diploid ancestral genome size is 0.5 pg, or 490 Mb (Lysak et al., 2009).

K_a/K_s analysis

We performed self-to-self BLASTP (v2.12.0+; Altschul et al., 1990) alignment for *H. variabilis*, *B. rapa*, *P. cornutum*, *T. arvense*, *E. salsugineum*, *Cardamine hirsuta*, *Arabis alpina*, and *A. thaliana*, respectively, and selected the best hits among the homologous gene pairs with identity > = 80 for the non-synonymous (K_a) and synonymous (K_s) rate calculations. Protein sequences of homologous gene pairs were aligned by ParaAT (v2.0; Zhang et al., 2012). Protein alignments were then converted into coding sequence alignments based on the indexed records of coding information using a custom Perl script. K_s values were further calculated based on the coding sequence alignments using the Nei–Gojobori method implemented in KaKs_Calculator (v2.0; Wang et al., 2010).

Comparison of dominant expression between paralogs

The mRNA-seq datasets from four tissues (flower, leaf, silique, and stem) were used for this analysis. The paired-end Illumina reads from the RNA samples described above were mapped to gene models of *H. variabilis* using Hisat2 (v2.1.0; Kim et al., 2015), and FeatureCounts (Liao et al., 2014) was used to extract the mapped reads for each gene to calculate transcripts per million (TPM) values as the expression level of the genes. Paralogous gene pairs (gene doublets) were extracted from the homologous genes of the four *H. variabilis* subgenomes. To avoid the effects of mismapped reads, all genes were sorted from high to low based on their expression levels. The lowest 1% of genes were considered as not expressed. When comparing the dominant expression status in each gene doublet, at least one of the two genes had an expression value >5, and the difference in expression values between the two genes was >2-fold.

TE distribution in neighboring regions of *H. variabilis* genes

We used a 50-bp sliding window with a 10-bp step moving across the 5′ and 3′ flanking regions of genes to estimate the TE density around each gene. In each 50-bp window, we calculated the ratio of TE nucleotides and then averaged the ratio across subsets of the *H. variabilis* genes. The averaged values were plotted as TE density in the flanking region of these subsets of *H. variabilis* genes.

Analysis of protein interaction network in *H. variabilis*

The protein complex map of *A. thaliana* was downloaded from the plant. MAP database (McWhite et al., 2020). Interaction genes with CF-MS score >0.7 were considered as high confidence. Arabidopsis genes represented the ancestral locus after the α-WGD event, and we selected loci in each subgenome of *H. variabilis* and mapped their *A. thaliana* orthologs onto network nodes. As ubiquitination regulates complex interactions between various proteins and can lead to overconnectivity in the protein interaction network (Huang & Dixit, 2016), we manually removed nodes associated with ubiquitination regulation. Finally, we identified 18 153 edges between 1099 nodes in *H. variabilis* that formed 90 independent interacting clusters (Table S16). Cytoscape (v3.9.0; Shannon et al., 2003) was used for visualization. We further asked about the retention status of each *H. variabilis* gene, labeling interactions among multiple-to-multiple copy genes, multiple-to-single copy genes, and single-to-single copy genes with different colors. Similarly, interactions between different subgenomes were also labeled with different colors (Figure S20).

Gene function enrichment

The genes that had two or more gene copies in four subgenomes of *H. variabilis* were considered as multiple-copy genes. The remaining genes were classified as singletons. Tandem

repeat arrays were identified using the SynOrths tool (Cheng et al., 2012). If a singleton existed as an array of tandem repeats, we used the longest sequence as the putative functional singleton. The three types of gene sets, that is, multiple-copy genes, singleton genes, and tandem repeat genes, were further subjected to GO function enrichment analysis. The GO terms and pathways of gene enrichment were identified by the clusterProfiler package (v3.14.3; Yu et al., 2012). GO enrichments were estimated using one-sided Fisher's exact tests, and an adjusted *P*-value <0.05 was set as the cutoff criterion for the significance of the gene enrichment.

Identification of conserved protein domains

We used the CD-search tool (Marchler-Bauer & Bryant, 2004; https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) to query the conserved domain of genes. The search was based on the CDD (v3.19; Marchler-Bauer et al., 2010) database with an expected value threshold of 0.01. Composition-based statistics is a new 'Composition-aware' measure implemented in RPS-blast, and it largely eliminates the necessity of using a low-complexity filter. We investigated the conserved protein domains of all 13 128 multiple-copy genes and expect to observe the fate of gene duplication in *H. variabilis* after the octoploidization. We assume that two copies of a gene after duplication can split functions (subfunctionalization) or can diverge to generate a new function (neofunctionalization; Birchler & Yang, 2022). For a multiple-copy gene family, if all copies in this family retain an identical conserved domain, we consider that the domain may be inherited from the ancestral gene, and if a gene in this family loses a conserved domain or diverges into a new conserved domain that is different from the others, we consider that the gene may be undergone sub/neofunctionalization. If all copies in this family had different conserved domains, they were all included in the count. Based on the above rules, we finally identified 822 gene families that underwent potential sub/neofunctionalization, as 571 conserved protein domains were newly differentiated. Sub/neofunctionalized genes were further submitted to PRGdb4.0 (Calle García et al., 2022) for testing whether they were resistance (R) genes.

Identification protein motifs

The classic model of MEME (v5.4.1; Bailey & Elkan, 1994; https://meme-suite.org/meme/tools/meme) tool has been used to identify the protein motifs of *AFP, NAI2/TSA1-BGLU18* and *GRF* gene family members, which structures were visualized by GSDS (v2.0; Hu et al., 2015).

ACKNOWLEDGMENTS

We thank Dr. P. Trávníček for estimation of genome size by flow cytometry and Maxie Jonk for providing the photograph of the *H. variabilis* population in Figure 1a. Plant Sciences core facility of CEITEC Masaryk University is acknowledged for the technical support. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. This work was supported by the Czech Science Foundation (grant no. 19-07487S), the Masaryk University Grant Agency (MUNI/R/1268/2022), and by the National Geographic Society (grant no. 9345-13).

AUTHOR CONTRIBUTIONS

TM and MAL conceived and designed the project. YH assembled and annotated the genome. YH, XG, TM, and KZ conducted experimental work and data analysis. Interpretation of data and results was led by TM, FC, and MAL All authors wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AND MATERIAL AVAILABILITY

The genome assembly and annotations of *H. variabilis* have been deposited to http://www.bioinformaticslab.cn/files/HV_data/. All other data needed to evaluate the conclusions in the manuscript are present in the paper and/or the Supporting Information.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Data S1. Supplementary Methods.

Figure S1. Distribution of *K*-mers ($K = 21$) in short-read sequence data of *H. variabilis* analyzed by findGSE (a) and GenomeScope (b) tools.

Figure S2. Genome-wide all-by-all Hi-C interactions of 11 pseudo-chromosomes in *H. variabilis*. Heat map shows the density of Hi-C interactions (calculated by pair-end reads) between contigs. The intensity of red color is proportional to the interaction frequency. The blue boxes indicate the 11 pseudo-chromosomes and other scaffolds, whereas the green boxes indicate the contigs that connect these scaffolds.

Figure S3. The BUSCO analysis of *H. variabilis* and other seven sequenced Brassicaceae genomes using the genome model (a) and protein model (b).

Figure S4. The length and proportion of transposable elements (TEs) in the *H. variabilis* genome. (a) TE lengths in the pericentromeres (upper bar) and chromosomal arm regions (lower bar) for 11 chromosomes. (b) The proportion of TE lengths in the pericentromeres and chromosomal arm regions of the *H. variabilis* genome.

Figure S5. Tandem repeats in the *H. variabilis* genome. (a) Circos plot of satellite repeats and rDNA sequences within the *H. variabilis* genome assembly. From outer to inner circles: 1st ring, distribution of all tandem repeats (dark grey bars), as well as the 832-bp satellite repeat specific to Hv01 (green bars); 2nd ring, distribution of the most abundant 168-bp satellite repeat (blue bars); 3rd ring, distribution of the putative 177-bp centromeric satellite repeat (red bars) found in all chromosomes; 4th ring, distribution of 5S (yellow) and 35S (cyan) rDNA sequences. (b) FISH localization of the selected satellite repeats (168-bp, 177-bp and 832-bp) and nuclear rDNA sequences (5S and 35S) on pachytene chromosomes. Chromosomes were counterstained by DAPI; FISH signals are shown in colors as indicated. Scale bars, 10 μm.

Figure S6. Synteny relationships between 11 chromosomes of *H. variabilis* and five chromosomes of *A. thaliana*.

Figure S7. Comparative chromosome painting in *H. variabilis*. *A. thaliana* BAC contigs corresponding to 22 genomic blocks (GBs, A to X) and eight chromosomes (AK1 to AK8) of the Ancestral Crucifer Karyotype (Lysak et al., 2016) were differentially labeled and polarized, and used as painting probes on pachytene bivalents. Quadruplicated GBs were tentatively assigned as #1 to #4. Split of a GB is indicated as "a" and "b". Scale bars, 10 μm.

Figure S8. Distribution of *Ks*-values between homeologous gene pairs. #1 to #4 refer to the four *H. variabilis* subgenomes. The numbers in parentheses indicate the numbers of identified homeologous gene pairs.

Figure S9. Expression profiles of gene families with four subgenomic copies retained in the *H. variabilis* genome. Four tissues were investigated: S (stems), L (leaves), F (flowers), and Sq (siliques). Clustering of the expression levels of gene families (TPM-values) is denoted on the left, whereas clustering of the expression levels of the four subgenomes is displayed on the top.

Figure S10. Quantification of quadruplicated genomic block A in *H. variabilis* revealed by comparative chromosome painting. The four genomic copies differ consistently in their fluorescence intensity and physical length. The histograms show ratios of fluorescence intensity and physical length of genomic block A between individual four subgenomes (#1 to #4) of *H. variabilis*. Scale bar, 10 μm.

Figure S11. The reconstructed origin of 11 chromosomes of *H. variabilis* (Hv) from 30 ancestral (AK) chromosomes and 88 genomic blocks. The capital letters (A to X) correspond to 22 genomic blocks of the four ancestral karyotypes/subgenomes color-coded as red (#1), green (#2), blue (#3), and orange (#4). The chromosomal rearrangements include: T, translocation; $T^{unequal}$, unequal translocation; EET, end-to-end translocation; I^{pa}, paracentric inversion; I^{pe}, pericentric inversion; NCI, nested chromosome insertion. Perientromeres are indicated by black hourglass symbols; lost ancestral pericentromeres are indicated by grey hourglass symbols. Black arrows show breakpoints. "Block + number" (e.g., Q1, Q2, and Q3) denotes that the GB was broken into multiple segments.

Figure S12. The investigation of TE element content in the four subgenomes of *H. variabilis*. (a) The length of TEs in the four subgenomes. (b) The proportion of eight TE types in the four subgenomes. The standardized proportions of these elements were further clustered and are shown in (c).

Figure S13. The distribution of TE insertion times in the *H. variabilis* genome. (a) The insertion time distribution of LTR retrotransposons and DNA transposons. (b) The insertion time distribution of two classes of LTR retrotransposons: Gypsy and Copia. (c) The insertion time distribution of six DNA transposon elements.

Figure S14. The distribution of TE insertion times in subgenomes and pericentromere regions of *H. variabilis*. (a) The distribution of insertion times of eight TE types in four subgenomes and pericentromere regions. Burst peaks are indicated by dark color bars and the corresponding insertion times are displayed. (b) The Pearson correlation of TE insertion times between any two subgenomes and between subgenomes and pericentromeres.

Figure S15. Heat map of 21-mer enrichment shared by four subgenomes of *H. variabilis* in 22 GBs (A to X).

Figure S16. Genomic sequence composition and length in six Brassicaceae species, including three diploids (2*x*), one tetraploid (4*x*), one hexaploid (6*x*), and the octoploid *H. variabilis* (8*x*). Different genomic components are indicated by different colors. The putative ancestral genome size (anc1C) for Brassicaceae is 0.5 pg (490 Mb; Lysak et al., 2009). The extent of genome size reduction (x-fold; the expected genome size (490 Mb × ploidy level) / (2 × actual genome size)) is labeled at the end of each species bar. For genome size values of the five crucifer species see Table S22.

Figure S17. The numbers of full-length LTR-RTs (F) and solo LTRs (S) and the S/F ratios in six Brassicaceae species (a) and within the four subgenomes of *H. variabilis* (b).

Figure S18. GO enrichment for singleton genes in *H. variabilis*. The x-axis represents the ratio between the number of enriched genes to the number of background genes. Circles of different sizes correspond to the number of genes enriched, whereas the colors represent molecular function (MF) and biological process (BP) gene categories, respectively. The GO terms that are involved in DNA activities are marked in yellow.

Figure S19. GO enrichment for tandem repeat genes (TRGs) in *H. variabilis*. The x-axis represents the ratio between the number of enriched genes to the number of background genes. Circles of different sizes correspond to the number of genes enriched, whereas the colors represent molecular function (MF), biological process (BP), and cellular component (CC) gene categories, respectively. The GO terms involved in biotic or abiotic stresses are marked in red.

Figure S20. The investigation of protein interactions in the *H. variabilis* genome. (a) Protein interaction network of *H. variabilis* genes. The dots represent the nodes (genes) and the connecting lines show the edges (interactions). The left and right panels are the same network, distinguished only by the meaning represented by the color of the gene: the left panel shows the copy number of genes, the right panel the subgenomic assignment. (b) The number of edges between interacting genes with 34 different copy numbers. (c) The number of nodes that form interactions between different subgenomes. The asterisks in (b) and (c) represent the level of significant difference. Two stars indicate a statistically significant level (*P*-value <= 0.05), three stars indicate the *P*-value close to 0. (d) GO functional enrichment of genes that formed interaction relationships between different subgenomes. Based on the phylogenetic relationships between subgenomes (see Figure 2d), we classified protein interactions as closely related (combinations between single subgenomes, sub #1-to-sub #3, and sub #2-to-sub #4), distantly related (sub #1-to-sub #2; sub #3-to-sub #4), and mixed (combinations of more than two subgenomes).

Figure S21. The pinnatipartite leaves of *H. variabilis*.

Figure S22. The 48 conserved protein motifs of the GRF gene family identified in six Brassicaceae species (*A. arabicum, A. thaliana, B. rapa, P. cornutum, S. parvula*, and *H. variabilis*). The phylogenetic relationship of each gene family member is marked to the left of the protein motif.

Figure S23. Similarity of *GRF* gene family protein motifs in five Brassicaceae species compared to the MEME conserved protein motif database (Bailey et al., 2006). Each *GRF* gene family was represented individually (except for *GRF5*). The x-axis represents the protein motifs distributed in the *GRF* gene family, and the y-axis represents the similarity level ($-\log_{10}$ (*P*-value)). The higher the value is, the higher the similarity is, where the $-\log_{10}$ (*P*-value) = 0 means that the motif is completely lost in the gene. The gene family members in each species are represented by dots of different colors.

Figure S24. The expression levels of GRF family members in *H. variabilis*.

Figure S25. Syntenic relationship between 10 GRF genes in *H. variabilis* (orange chromosome bars) and 7 *GRF* genes in *A. thaliana* (grey chromosome bars). Genes with syntenic relationships are labeled, and the colors correspond to the four *H. variabilis* subgenomes.

Figure S26. Syntenic relationship of the *AFP* genes localized on chromosome At1 in *Arabidopsis thaliana* and chromosomes Hv01 and Hv05 in *H. variabilis*.

Figure S27. Neighbour-joining tree and protein sequence alignment of conserved domains in *AFP* gene family. The phylogenetic tree consist of three groups of 40 *AFP* gene family members, indicated by green, yellow, and blue colors, respectively. The four *AFP* genes previously identified in *Heliophila coronopifolia*

(*HcAFP1-4*, De Beer & Vivier, 2011) are marked in red and can be divided into two groups. The protein sequence alignment of the conserved domain of the *AFP* gene family is shown on the right of the phylogenetic tree. The conserved $\beta 2$-$\beta 3$ loops are marked in yellow backgrounds. The 19-mer conserved amino acid sequences encoding active functions are marked in blue backgrounds.

Figure S28. Outline of genome structure changes at the *BGLU18-NAI2/TSA1* locus in seven Brassicaceae species, including four diploids, one tetraploid, one hexaploid, and the octoploid *H. variabilis*. Gene is indicated by square and the lost genes are marked as dashed lines. *NAI2* and *TSA1* is homeologous gene pair and *BGLU18* is a neighbouring gene of *TSA1*. "-TA" after the gene name indicates a tandem array of the gene.

Figure S29. Gene structure and protein conserved motifs of the 45 members of the *BGLU18- NAI2/TSA1* gene family. The gene names are arranged in the order of *NAI2, TSA1* and *BGLU18* on the left, the distribution of CDS and UTR regions on the genes is shown in the middle, and the conserved protein motifs in the CDS region are shown on the right. Six genes in the *NAI2* family and three genes in the *BGLU18* family lost some motifs as marked by small circles of different colors.

Table S1. Sequencing reads used for assembly of the *H. variabilis* genome.

Table S2. Statistics of the final assembly of the *H. variabilis* genome.

Table S3. Lengths of all pseudo-chromosomes and scaffolds assembled.

Table S4. RNA-seq data obtained from four tissues of *H. variabilis*.

Table S5. The function annotation of predicted genes of *H. variabilis*.

Table S6. Classification of repetitive elements in the *H. variabilis* genome.

Table S7. The chromosome-level coordinates of syntenic gene segments shared between the *H. variabilis* and ACK genomes.

Table S8. The difference in retained gene numbers among the four copies of 22 GBs in *H. variabilis*.

Table S9. The number of dominantly expressed genes between paralogous gene pairs in *H. variabilis*.

Table S10. The GB associations whthin subgenomes and between any two subgenomes.

Table S11. Quantification of chromosomal rearrangement types inferred in the origin of the 11 chromosomes of *H. variabilis*.

Table S12. Summary of exonic deletions on the basis of subgenome assignment.

Table S13. The number of K-mer repeated in the four *H. variabilis* subgenomes.

Table S14. Potential functions of conserved protein domains over-represented in the *H. variabilis* genome.

Table S15. Conserved protein domains involved in sub/neo-functionalization of multiple-copy (MCG) families in *H. variabilis*.

Table S16. Number of edges and nodes in the *H. variabilis* protein interaction network.

Table S17. Eight categories of candidate genes that involved in adaptive evolution.

Table S18. Gene members of the *GRF* family in six Brassicaceae species.

Table S19. Gene members of the *AFP* family in seven Brassicaceae species.

Table S20. Statistics on the conserved structure of the *AFP* genes in seven Brassicaceae species.

Table S21. The CDS number and length of *BGLU18-NAI2/TSA1* gene family members.

Table S22. Genome sequence datasets used for comparative genomic analysis.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Bailey, T.L. & Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proceedings - International Conference on Intelligent Systems for Molecular Biology*, **2**, 28–36.

Bennetzen, J.L. (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, **115**, 29–36.

Birchler, J.A. & Veitia, R.A. (2012) Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, **109**, 14746–14753.

Birchler, J.A. & Yang, H. (2022) The multiple fates of gene duplications: deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *The Plant Cell*, **34**, 2466–2474.

Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Research*, **41**, W29–W33.

Burns, R., Mandáková, T., Gunis, J., Soto-Jiménez, L.M., Liu, C., Lysak, M.A. *et al.* (2021) Gradual evolution of allopolyploidy in *Arabidopsis suecica*. *Nature Ecology & Evolution*, **5**, 1367–1381.

Calle García, J., Guadagno, A., Paytuvi-Gallart, A., Saera-Vila, A., Amoroso, C.G., D'Esposito, D. *et al.* (2022) PRGdb 4.0: an updated database dedicated to genes involved in plant disease resistance process. *Nucleic Acids Research*, **50**, D1483–D1490.

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A., Tang, H., Wang, X. *et al.* (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*, **345**, 950–953.

Chang, J., Duong, T.A., Schoeman, C., Ma, X., Roodt, D., Barker, N. *et al.* (2023) The genome of the king protea, *Protea cynaroides*. *The Plant Journal*, **113**, 262–276.

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M. & Wang, X. (2018) Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants*, **4**, 258–268.

Cheng, F., Wu, J., Fang, L. & Wang, X. (2012) Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. *Frontiers in Plant Science*, **3**, 198.

Clark, J.W. & Donoghue, P.C. (2017) Constraining the timing of whole genome duplication in plant evolutionary history. *Proceedings of the Royal Society B: Biological Sciences*, **284**, 20170912.

De Beer, A. & Vivier, M.A. (2011) Four plant defensins from an indigenous South African Brassicaceae species display divergent activities against two test pathogens despite high sequence similarity in the encoding genes. *BMC Research Notes*, **4**, 1–19.

Deamer, D., Akeson, M. & Branton, D. (2016) Three decades of nanopore sequencing. *Nature Biotechnology*, **34**, 518–524.

Debernardi, J.M., Mecchia, M.A., Vercruyssen, L., Smaczniak, C., Kaufmann, K., Inze, D. *et al.* (2014) Post-transcriptional control of *GRF* transcription factors by microRNA miR396 and GIF co-activator affects leaf size and longevity. *The Plant Journal*, **79**, 413–426.

Derelle, R., Philippe, H. & Colbourne, J.K. (2020) Broccoli: combining phylogenetic and network analyses for orthology assignment. *Molecular Biology and Evolution*, **37**, 3389–3396.

Devos, K.M., Brown, J.K. & Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Research*, **12**, 1075–1079.

Dogan, M., Pouch, M., Mandáková, T., Hloušková, P., Guo, X., Winter, P. *et al.* (2021) Evolution of tandem repeats is mirroring post-polyploid cladogenesis in *Heliophila* (Brassicaceae). *Frontiers in Plant Science*, **11**, 607893.

Doležel, J., **Greilhuber, J.** & **Suda, J.** (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, **2**, 2233–2244.
Dudchenko, O., **Batra, S.S.**, **Omer, A.D.**, **Nyquist, S.K.**, **Hoeger, M.**, **Durand, N.C.** ***et al.*** (2017) De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–95.
Durand, N.C., **Shamim, M.S.**, **Machol, I.**, **Rao, S.S.**, **Huntley, M.H.**, **Lander, E.S.** ***et al.*** (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, **3**, 95–98.
Edger, P.P., **Poorten, T.J.**, **VanBuren, R.**, **Hardigan, M.A.**, **Colle, M.**, **McKain, M.R.** ***et al.*** (2019) Origin and evolution of the octoploid strawberry genome. *Nature Genetics*, **51**, 541–547.
Feschotte, C., **Jiang, N.** & **Wessler, S.R.** (2002) Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, **3**, 329–341.
Garsmeur, O., **Schnable, J.C.**, **Almeida, A.**, **Jourda, C.**, **D'Hont, A.** & **Freeling, M.** (2014) Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution*, **31**, 448–454.
Geiser, C., **Mandáková, T.**, **Arrigo, N.**, **Lysak, M.A.** & **Parisod, C.** (2016) Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in buckler mustard. *The Plant Cell*, **28**, 17–27.
Goldblatt, P. & **Manning, J.C.** (2002) Plant diversity of the Cape region of southern Africa. *Annals of the Missouri Botanical Garden*, **89**, 281–302.
Gonzalez, N., **Vanhaeren, H.** & **Inzé, D.** (2012) Leaf size control: complex coordination of cell division and expansion. *Trends in Plant Science*, **17**, 332–340.
Gordon, S.P., **Levy, J.J.** & **Vogel, J.P.** (2019) Polycracker, a robust method for the unsupervised partitioning of polyploid subgenomes by signatures of repetitive DNA evolution. *BMC Genomics*, **20**, 1–14.
Grover, C.E. & **Wendel, J.F.** (2010) Recent insights into mechanisms of genome size change in plants. *Journal of Botany*, **2010**, 1–8.
Guo, X., **Mandáková, T.**, **Trachtová, K.**, **Özüdoğru, B.**, **Liu, J.** & **Lysak, M.A.** (2021) Linked by ancestral bonds: multiple whole-genome duplications and reticulate evolution in a Brassicaceae tribe. *Molecular Biology and Evolution*, **38**, 1695–1714.
Haas, B.J., **Delcher, A.L.**, **Mount, S.M.**, **Wortman, J.R.**, **Smith, R.K., Jr.**, **Hannick, L.I.** ***et al.*** (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**, 5654–5666.
Haas, B.J., **Papanicolaou, A.**, **Yassour, M.**, **Grabherr, M.**, **Blood, P.D.**, **Bowden, J.** ***et al.*** (2013) De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.
Haas, B.J., **Salzberg, S.L.**, **Zhu, W.**, **Pertea, M.**, **Allen, J.E.**, **Orvis, J.** ***et al.*** (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, **9**, 1–22.
Hao, Y., **Mabry, M.E.**, **Edger, P.P.**, **Freeling, M.**, **Zheng, C.**, **Jin, L.** ***et al.*** (2021) The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. *Genome Research*, **31**, 799–810.
Hendriks, K.P., **Kiefer, C.**, **Al-Shehbaz, I.A.**, **Bailey, C.D.**, **Hooft Van Huysduynen, A.**, **Nikolov, L.A.** ***et al.*** (2022) Global phylogeny of the Brassicaceae provides important insights into gene discordance. *bioRxiv* 2022-09. https://doi.org/10.1101/2022.09.01.506188
Hoang, D.T., **Chernomor, O.**, **Von Haeseler, A.**, **Minh, B.Q.** & **Vinh, L.S.** (2018) UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, **35**, 518–522.
Hofstatter, P.G., **Thangavel, G.**, **Lux, T.**, **Neumann, P.**, **Vondrak, T.**, **Novak, P.** ***et al.*** (2022) Repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell*, **185**, 3153–3168.
Hohmann, N., **Wolf, E.M.**, **Lysak, M.A.** & **Koch, M.A.** (2015) A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *The Plant Cell*, **27**, 2770–2784.
Hopper, S.D., **Silveira, F.A.** & **Fiedler, P.L.** (2016) Biodiversity hotspots and Ocbil theory. *Plant and Soil*, **403**, 167–216.
Horiguchi, G., **Kim, G.T.** & **Tsukaya, H.** (2005) The transcription factor *AtGRF5* and the transcription coactivator *AN3* regulate cell proliferation in leaf primordia of *Arabidopsis thaliana*. *The Plant Journal*, **43**, 68–78.
Hu, B., **Jin, J.**, **Guo, A.Y.**, **Zhang, H.**, **Luo, J.** & **Gao, G.** (2015) GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics*, **31**, 1296–1297.
Hu, Q., **Ma, Y.**, **Mandáková, T.**, **Shi, S.**, **Chen, C.**, **Sun, P.** ***et al.*** (2021) Genome evolution of the psammophyte *Pugionium* for desert adaptation and further speciation. *Proceedings of the National Academy of Sciences*, **118**, e2025711118.
Huang, X. & **Dixit, V.M.** (2016) Drugging the undruggables: exploring the ubiquitin system for drug development. *Cell Research*, **26**, 484–498.
International Wheat Genome Sequencing Consortium (IWGSC), **Appels, R.**, **Eversole, K.**, **Stein, N.**, **Feuillet, C.**, **Keller, B.** ***et al.*** (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, eaar7191.
Jaillon, O., **Aury, J.**, **Noel, B.**, **Policriti, A.**, **Clepet, C.**, **Casagrande, A.** ***et al.*** (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
Jiao, Y., **Wickett, N.J.**, **Ayyampalayam, S.**, **Chanderbali, A.S.**, **Landherr, L.**, **Ralph, P.E.** ***et al.*** (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.
Kagale, S., **Robinson, S.J.**, **Nixon, J.**, **Xiao, R.**, **Huebert, T.**, **Condie, J.** ***et al.*** (2014) Polyploid evolution of the Brassicaceae during the Cenozoic era. *The Plant Cell*, **26**, 2777–2791.
Kamal, N., **Tsardakas Renhuldt, N.**, **Bentzer, J.**, **Gundlach, H.**, **Haberer, G.**, **Juhász, A.** ***et al.*** (2022) The mosaic oat genome gives insights into a uniquely healthy cereal crop. *Nature*, **606**, 113–119.
Katoh, K., **Misawa, K.**, **Kuma, K.I.** & **Miyata, T.** (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.
Kidwell, M.G. (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, **115**, 49–63.
Kim, D., **Langmead, B.** & **Salzberg, S.L.** (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357–360.
Kim, J.H., **Choi, D.** & **Kende, H.** (2003) The *AtGRF* family of putative transcription factors is involved in leaf and cotyledon growth in *Arabidopsis*. *The Plant Journal*, **36**, 94–104.
Kim, J.H. & **Lee, B.H.** (2006) GROWTH-REGULATING FACTOR4 of *Arabidopsis thaliana* is required for development of leaves, cotyledons, and shoot apical meristem. *Journal of Plant Biology*, **49**, 463–468.
Li, H. & **Durbin, R.** (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
Li, Z. & **Barker, M.S.** (2020) Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. *GigaScience*, **9**, giaa004.
Li, Z., **Defoort, J.**, **Tasdighian, S.**, **Maere, S.**, **Van de Peer, Y.** & **De Smet, R.** (2016) Gene duplicability of core genes is highly consistent across all angiosperms. *The Plant Cell*, **28**, 326–344.
Liao, Y., **Smyth, G.K.** & **Shi, W.** (2014) FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
Lieberman-Aiden, E., **Van Berkum, N.L.**, **Williams, L.**, **Imakaev, M.**, **Ragoczy, T.**, **Telling, A.** ***et al.*** (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
Lomsadze, A., **Burns, P.D.** & **Borodovsky, M.** (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, **42**, e119.
Lysak, M.A., **Cheung, K.**, **Kitschke, M.** & **Bures, P.** (2007) Ancestral chromosomal blocks are triplicated in Brassiceae species with varying chromosome number and genome size. *Plant Physiology*, **145**, 402–410.
Lysak, M.A., **Koch, M.A.**, **Beaulieu, J.M.**, **Meister, A.** & **Leitch, I.J.** (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. *Molecular Biology and Evolution*, **26**, 85–98.
Lysak, M.A., **Mandáková, T.** & **Schranz, M.E.** (2016) Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. *Current Opinion in Plant Biology*, **30**, 108–115.
Mandáková, T., **Li, Z.**, **Barker, M.S.** & **Lysak, M.A.** (2017) Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *The Plant Journal*, **91**(1), 3–21.
Mandáková, T. & **Lysak, M.A.** (2008) Chromosomal phylogeny and karyotype evolution in x= 7 crucifer species (Brassicaceae). *The Plant Cell*, **20**, 2559–2570.

Mandáková, T. & **Lysak, M.A.** (2018) Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*, **42**, 55–65.
Mandáková, T., Mummenhoff, K., Al-Shehbaz, I.A., Mucina, L., Mühlhausen, A. & **Lysak, M.A.** (2012) Wholegenome triplication and species radiation in the southern African tribe Heliophileae (Brassicaceae). *Taxon*, **61**, 989–1000.
Marçais, G. & **Kingsford, C.** (2011) A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, **27**, 764–770.
Marchler-Bauer, A. & **Bryant, S.H.** (2004) CD-search: protein domain annotations on the fly. *Nucleic Acids Research*, **32**, W327–W331.
Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C. *et al.* (2010) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, **39**, D225–D229.
McWhite, C.D., Papoulas, O., Drew, K., Cox, R.M., June, V., Dong, O.X. *et al.* (2020) A pan-plant protein complex map reveals deep conservation and novel assemblies. *Cell*, **181**, 460–474.
Mittermeier, R.A., Myers, N., Thomsen, J.B., Da Fonseca, G.A. & **Olivieri, S.** (1998) Biodiversity hotspots and major tropical wilderness areas: approaches to setting conservation priorities. *Conservation Biology*, **12**, 516–520.
Mummenhoff, K., Al-Shehbaz, I.A., Bakker, F.T., Linder, H.P. & **Mühlhausen, A.** (2005) Phylogeny, morphological evolution, and speciation of endemic Brassicaceae genera in the Cape flora of southern Africa. *Annals of the Missouri Botanical Garden*, **92**, 400–424.
Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A. & **Kent, J.** (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.
Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. & **Minh, B.Q.** (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.
Nikolov, L.A., Shushkov, P., Nevado, B., Gan, X., Al-Shehbaz, I.A., Filatov, D. *et al.* (2019) Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist*, **222**, 1638–1651.
Oberlander, K.C., Dreyer, L.L., Goldblatt, P., Suda, J. & **Linder, H.P.** (2016) Species-rich and polyploid-poor: insights into the evolutionary role of whole-genome duplication from the Cape flora biodiversity hotspot. *American Journal of Botany*, **103**, 1336–1347.
Papp, B., Pal, C. & **Hurst, L.D.** (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**, 194–197.
Peppe, D.J., Cote, S.M., Deino, A.L., Fox, D.L., Kingston, J.D., Kinyanjui, R.N. *et al.* (2023) Oldest evidence of abundant C4 grasses and habitat heterogeneity in eastern Africa. *Science*, **380**, 173–177.
Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. & **Salzberg, S.L.** (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295.
Pompidor, N., Charron, C., Hervouet, C., Bocs, S., Droc, G., Rivallan, R. *et al.* (2021) Three founding ancestral genomes involved in the origin of sugarcane. *Annals of Botany*, **127**, 827–840.
Qiao, X., Zhang, S. & **Paterson, A.H.** (2022) Pervasive genome duplications across the plant tree of life and their links to major evolutionary innovations and transitions. *Computational and Structural Biotechnology Journal*, **20**, 3248–3256.
Robertson, F.M., Gundappa, M.K., Grammes, F., Hvidsten, T.R., Redmond, A.K., Lien, S. *et al.* (2017) Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biology*, **18**, 1–14.
Robinson, J.T., Turner, D., Durand, N.C., Thorvaldsdóttir, H., Mesirov, J.P. & **Aiden, E.L.** (2018) Juicebox. js provides a cloud-based visualization system for Hi-C data. *Cell Systems*, **6**, 256–258.
Schaaper, W., Posthuma, G., Meloen, R., Plasman, H., Sijtsma, L., Van Amerongen, A. *et al.* (2001) Synthetic peptides derived from the β2−β3 loop of *Raphanus sativus* antifungal protein 2 that mimic the active site. *The Journal of Peptide Research*, **57**, 409–418.
Schranz, M.E., Lysak, M.A. & **Mitchell-Olds, T.** (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends in Plant Science*, **11**, 535–542.
Schranz, M.E., Mohammadin, S. & **Edger, P.P.** (2012) Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Current Opinion in Plant Biology*, **15**, 147–153.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.
Shavrukov, Y., Kurishbayev, A., Jatayev, S., Shvidchenko, V., Zotova, L., Koekemoer, F. *et al.* (2017) Early flowering as a drought escape mechanism in plants: how can it aid wheat production? *Frontiers in Plant Science*, **8**, 1950.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & **Zdobnov, E.M.** (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
Stanke, M., Tzvetkova, A. & **Morgenstern, B.** (2006) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology*, **7**, 1–8.
Stefanik, N., Bizan, J., Wilkens, A., Tarnawska-Glatt, K., Goto-Yamada, S., Strzałka, K. *et al.* (2020) NAI2 and TSA1 drive differentiation of constitutive and inducible ER body formation in Brassicaceae. *Plant and Cell Physiology*, **61**, 722–734.
Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y. *et al.* (2017) Karyotype stability and unbiased fractionation in the paleo-allotetraploid *Cucurbita* genomes. *Molecular Plant*, **10**, 1293–1306.
Tang, H., Woodhouse, M.R., Cheng, F., Schnable, J.C., Pedersen, B.S., Conant, G. *et al.* (2012) Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics*, **190**, 1563–1574.
Van de Peer, Y., Ashman, T.L., Soltis, P.S. & **Soltis, D.E.** (2021) Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell*, **33**, 11–26.
Van de Peer, Y., Mizrachi, E. & **Marchal, K.** (2017) The evolutionary significance of polyploidy. *Nature Reviews Genetics*, **18**, 411–424.
Van Rooyen, M., Grobbelaar, N., Theron, G. & **Van Rooyen, N.** (1992) The ephemerals of Namaqualand: effect of germination date on development of three species. *Journal of Arid Environments*, **22**, 51–66.
Van Santen, M. & **Linder, H.P.** (2020) The assembly of the Cape flora is consistent with an edaphic rather than climatic filter. *Molecular Phylogenetics and Evolution*, **142**, 106645.
Vekemans, D., Proost, S., Vanneste, K., Coenen, H., Viaene, T., Ruelens, P. *et al.* (2012) Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Molecular Biology and Evolution*, **29**, 3793–3806.
Verboom, G.A., Archibald, J.K., Bakker, F.T., Bellstedt, D.U., Conrad, F., Dreyer, L.L. *et al.* (2009) Origin and diversification of the greater Cape flora: ancient species repository, hot-bed of recent radiation, or both? *Molecular Phylogenetics and Evolution*, **51**, 44–53.
Walden, N., German, D.A., Wolf, E.M., Kiefer, M., Rigault, P., Huang, X.C. *et al.* (2020) Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. *Nature Communications*, **11**, 3795.
Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.
Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & **Yu, J.** (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics & Bioinformatics*, **8**, 77–80.
Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S. *et al.* (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*, **43**, 1035–1039.
Weiller, F., Moore, J.P., Young, P., Driouich, A. & **Vivier, M.A.** (2017) The Brassicaceae species *Heliophila coronopifolia* produces root border-like cells that protect the root tip and secrete defensin peptides. *Annals of Botany*, **119**, 803–813.
Yu, G., Wang, L.G., Han, Y. & **He, Q.Y.** (2012) ClusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology*, **16**, 284–287.
Zhang, C., Scornavacca, C., Molloy, E.K. & **Mirarab, S.** (2020) ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution*, **37**, 3292–3307.
Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X. *et al.* (2018) Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nature Genetics*, **50**, 1565–1573.

Zhang, X., **Zhang, S.**, **Zhao, Q.**, **Ming, R.** & **Tang, H.** (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants*, **5**, 833–845.

Zhang, Y., **Zhang, L.**, **Xiao, Q.**, **Wu, C.**, **Zhang, J.**, **Xu, Q.** ***et al.*** (2022) Two independent allohexaploidizations and genomic fractionation in Solanales. *Frontiers in Plant Science*, **13**, 1001402.

Zhang, Z., **Xiao, J.**, **Wu, J.**, **Zhang, H.**, **Liu, G.**, **Wang, X.** ***et al.*** (2012) ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications*, **419**, 779–781.

APPENDIX 2

Post-polyploid chromosomal diploidization in plants is affected by clade divergence and constrained by shared genomic features

Martin Lysak

martin.lysak@ceitec.muni.cz

Central European Institute of Technology, Masaryk University https://orcid.org/0000-0003-0318-4194

Yile Huang

Central European Institute of Technology – Masaryk University

Manuel Poretti

University of Fribourg https://orcid.org/0000-0001-6915-2238

Terezie Mandáková

CEITEC - Central European Institute of Technology, Masaryk University

Milan Pouch

CEITEC, Masaryk University

Xinyi Guo

Central European Institute of Technology – Masaryk University https://orcid.org/0000-0001-5416-7787

Estela Perez-Roman

School of Life Sciences, University of Sussex

Manuel B Crespo

Department of Environmental Sciences and Natural Resources (dCARN), University of Alicante https://orcid.org/0000-0002-3294-5637

stefan Grob

Strasbourg IBMP

Alexandros Bousios

School of Life Sciences, University of Sussex

Christian Parisod

University of Fribourg https://orcid.org/0000-0001-8798-0897

Article

Keywords: whole-genome duplication, diploidization, subgenomes, recurrent chromosome breakpoint, LTR retrotransposons, TAD, descending dysploidy, karyotype evolution, angiosperms, Brassicaceae

Posted Date: May 12th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6440714/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

Genomic redundancy resulting from whole-genome duplication creates opportunities for double-strand misrepair that can lead to chromosomal rearrangements and a reduction in chromosome number, known as descending dysploidy. Although flowering plants often undergo post-polyploid rediploidization, the pathways and consequences of descending dysploidy are still poorly understood. In this study, we sequenced and assembled the genomes of eight *Biscutella* species varying in size from 0.6 to 1.1 Gb and exhibiting chromosome numbers of *n* = 6, 8 and 9. Our analysis revealed an estimated 12 million years of diploidization of an allotetraploid ancestral genome (*n* = 14) characterized by independent descending dysploidy, resulting in chromosome numbers of *n* = 9, 8, and 6. We identified clades of early-diverging (*n* = 8/6) and late-diverging (*n*= 9) genomes that exhibited both convergent and divergent features. While both clades showed similar levels of subgenome fractionation and preferential retention of polyploidy-derived genes, the early-diverging genomes exhibited a higher removal ratio of LTR retrotransposons and greater variability in the size of topologically associated domains (TADs). In addition, we identified 12 chromosome breakage hotspots enriched in LTR retrotransposons and frequently located at TAD boundaries. In addition, we identified 12 chromosome breakpoint hotspots enriched for LTR retrotransposons and frequently located at TAD boundaries. This suggests that although post-polyploid descendig dysploid appears to be an independent and superficially random process, some shared genomic features may favor the occurrence of recurrent chromosome breakpoints in different species.

Introduction

Land plants are known for their remarkable karyological variation, with chromosome numbers varying from two pairs (*n* = 2) to several hundred pairs. This extraordinary variation is due to two antagonistic evolutionary processes: polyploidy (whole-genome duplication, WGD) and diploidization. Polyploidy contributes to an increase in the number of chromosomes, whereas diploidization is typically associated with the restoration of bivalent pairing and a decrease in the number of chromosomes, a process known as descending dysploidy. The cyclic nature of the independent emergence of genomic redundancy followed by its gradual reduction (Wendel et al., 2016, Qiao et al., 2019) is widespread in most clades of land plants. Although significant progress has been made in understanding WGDs, our knowledge of the triggers, mechanisms, and implications of post-polyploid diploidization remains relatively sparse (Li et al., 2021).

In plants, a common way of descending dysploidy is non-allelic homologous recombination (NAHR) between two or more different chromosomes, reducing the number of chromosomes typically by one (e.g., Schubert & Lysak, 2011; Mayrose & Lysak, 2021). Diploidizing descending dysploidy is particularly common after WGD events, as polyploidization multiplies homologous sequences as substrates for illegitimate recombination, leading first to simple fusion chromosomes and later, as diploidization progresses, to complex fusion chromosomes (CFCs), in which segments of multiple non-homologous chromosomes are combined. CFCs frequently occur in diploidized tetraploid or hexaploid plant

genomes, such as cotton (*Gossypium*, Sun et al., 2024), soybean (Zhao et al., 2017), *Brassica* (Cheng et al., 2013) and camelina (*Camelina*, Mandáková et al., 2019). The frequency and stability of CFCs depends, among other factors, on the degree of DNA homology and repeat content, the frequency of DSBs and NAHR as well as structural (e.g., the size of CFCs) and functional (e.g., altered gene expression) constraints.

WGD events often occur before the diversification of a clade, leading to the assumption that all clade members are descended from a single polyploid genome or metapopulation. The polyploid clade diversifies as a result of adaptive radiation and is typically associated with some level of genetic, epigenetic, or cytological diploidization. Assuming that the polyploid populations are genetically identical or nearly identical, the results of independent diploidization may be convergent, or alternatively, variable abiotic and biotic conditions may determine or modulate the outcomes of the diploidization process. Nevertheless, the question of the likelihood of convergent vs. divergent post-polyploid diploidization is still largely unresolved. For example, if the genetic makeup of spatially separated polyploid populations or descendant species is the same, are there chromosomal regions or sequences that are more prone to DSBs and chromosomal rearrangements? Furthermore, what is the likelihood that DSB misrepair at recombination hotspots in different diploidizing populations or species will lead to the same chromosomal rearrangements and eventually to the same or very similar fusion chromosomes?

The cruciferous genus *Biscutella* comprises about 60 species distributed throughout the Mediterranean basin, and as far as Iran and the Arabian Peninsula (POWO 2025). *Biscutella* is a polybasic genus with three different base chromosome numbers: x = 6 (*B. lyrata* L.), 8 (a dozen) and 9 (most species) (Olowokudejo & Heywood, 1984). Intrageneric karyological variation attracted scientific interest and was first recognized as an “interesting evolutionary problem” by Manton (1932), who then introduced the *B. laevigata* L. species complex as one of the first models for studying the role of autopolyploidy in the evolution of plant genomes (Manton 1934). The identification of x = 9 in most *Biscutella* species led to the conclusion that nine chromosomes represent the ancestral condition, with lower chromosome numbers (*n* = 8 and 6) interpreted as resulting from chromosome loss (aneuploidy; Olowokudejo & Heywood, 1984). Much later, however, transcriptomic and cytogenomic studies of the diploid species *B. laevigata* subsp. *varia* (Dumort.) Rouy & Foucaud (2*n* = 18; Geiser et al., 2016), *B. baetica* Boiss. & Reuter (2*n* = 16) and *B. lyrata* (2*n* = 12; Mandáková et al., 2017a) indicated that these species underwent a common WGD event. Consequently, the current chromosome numbers are thought to have evolved from a mesotetraploid progenitor genome through a process of descending dysploidy that accompanied post-polyploid diploidization (Geiser et al., 2016; Mandáková et al., 2017a; Guo et al., 2021; Beringer et al., 2024).

To investigate the pathways and constraints of cytological diploidization, we generated chromosome-level genome assemblies for eight *Biscutella* species from morphologically and phylogenetically distinct infrageneric clades, with three different base numbers (*n* = 6, 8 and 9). We show (1) post-polyploid cytological diploidization mediated by independent descending dysploidy in early- and late-diverging species, (2) convergence in the extent of subgenome fractionation and retention of polyploidy-derived

genes, (3) divergence in the removal of LTR retrotransposons and conservation of topologically associated domains (TADs) between early- and late-diverging species, and (4) the role of LTR retrotransposons and 3D genome architecture in establishing chromosome breakage hotspots.

Results

Chromosome-scale genome assembly and annotation of eight *Biscutella* species with varying chromosome numbers

We generated chromosome-scale genome assemblies using Illumina short-read (c. 83× genome coverage), Oxford Nanopore long-read (ONT; c. 62×), and high-throughput chromosome conformation capture (Hi-C) sequencing for eight *Biscutella* species, including *B. laevigata* subsp. *varia* (*n* = 9; haploid nuclear genome size [1C] = 814 Mb; referrred to as *B. varia*), *B. laevigata* subsp. *austriaca* (*n* = 9; 904 Mb; referred to as *B. austriaca*), *B. prealpina* (*n* = 9; 916 Mb), *B. frutescens* (*n* = 9; 936 Mb), *B. auriculata* (*n* = 8; 686 Mb), *B. didyma* (*n* = 8; 980 Mb), *B. baetica* (*n* = 8; 1.1 Gb), and *B. lyrata* (*n* = 6; 806 Mb). The PacBio high-fidelity (HiFi; c. 28×) sequencing was additionally performed for *B. lyrata* (**Table S1**). Eight genome assemblies with total lengths of c. 515 to 1,168 Mb were obtained following either ONT- or PacBio-based assembly strategy (**Figure 1a**; **Methods**). Compared to *Biscutella* genomes assembled via the ONT-based strategy, which yielded an average of 1,258 contigs with an average contig N50 length of ~2.9 Mb, the *B. lyrata* genome, assembled using the PacBio-based strategy, consists of 119 contigs and a contig N50 length of 22.89 Mb (**Table S2**), representing the most continuous *Biscutella* genome assembly (**Figure 1b**). The assembled contigs were subsequently organized into 6, 8, or 9 pseudochromosomes ranging in size from 56 to 159 Mb based on Hi-C contact maps (**Figure S1**). Benchmarking Universal Single-Copy Ortholog (BUSCO) analysis of the assembled sequences showed high completeness of all eight assemblies (92.0% – 99.1%; **Table S3**).

Using a combination of *ab initio*, homology-based and evidence-based approaches (**Methods**), we annotated 32,292 to 50,236 high-confidence protein-coding genes (**Figure 1c**), with comparable average gene lengths (~2,160 bp) and exon numbers (average ~5 per gene; **Table S4**). Synteny analysis further identified 22,492 to 31,397 gene pairs in the eight *Biscutella* genomes, corresponding to 22 ancestral genomic blocks (GBs A – X; Schranz et al., 2006; Lysak et al., 2016; **Figure S2**). Syntenic genes are predominantly distributed along chromosome arms, whereas gene-poor regions are typically associated with (peri)centromeres. The position of syntenic genes delimiting the GBs flanking the gene-poor regions were used to define the boundaries of pericentromeric regions (**Table S5**). Utilizing these syntenic gene pairs as pillars, we identified two genomic copies for nearly all 22 GBs (**Figure S2**), supporting the reliability of the gene annotations. An exception was block G in *B. prealpina* and *B. varia*, where only one copy was detected, likely due to extensive gene loss or misassembly in these regions. The overall 2:1 syntenic relationship between the *Biscutella* genomes and the 22 GBs confirms a meso-tetraploid WGD predating diversification of the genus (Geiser et al., 2016; Guo et al., 2021; Beringer et al., 2024). However, only the closely-related *B. austriaca*, *B. varia* and *B. prealpina* (Grünig et al., 2024) retained

relatively stable syntenic relationships, while the remaining genomes contain extensive chromosomal rearrangements (CRs; **Figure 1d**).

A total of 268 to 764 Mb of transposable elements (TEs) were identified, constituting 52% to 71% of the assembled genomes (**Figure 1c**). The most abundant TEs were long terminal repeat retrotransposons (LTR-RTs; 34.3% – 54.3%), followed by helitrons (8.8% – 14.1%) and terminal inverted repeats (TIRs; 5.9% – 10.4%; **Table S6**). Tandem repeat sequences ranged from 19.5 to 88.7 Mb (**Table S7**), with centromeric satellite DNA (satDNA) identified in all eight species and (sub)telomeric satDNA detected in four species (**Methods**). In n = 9 species, two types of centromeric satDNAs were identified (monomer length of 213 bp and 234 bp), while in the n = 8/6 species, the monomer lengths of centromeric satDNAs ranged from 124 to 218 bp, comprising seven types (**Figures S3 – S5** and **Table S8**). Phylogenetic analysis and pairwise comparisons further revealed that the 234-bp satDNAs form a distinct conserved clade with high sequence similarities (**Figures 1e** and **S6**), suggesting a conserved centromere structure in recently diverging n = 9 species. In contrast, other centromeric monomers displayed lower sequence similarity, reflecting their higher variability.

Subgenome divergence revealed the allotetraploid origin of *Biscutella*

Subgenomes of *Biscutella* were identified based on unequal gene density of two genomic copies and verified by phylogeny and *K*-mer sequence similarity. Biased gene fractionation was observed between duplicated GBs, with one copy generally exhibiting higher gene density than the other (**Figure 2a**), which was confirmed by comparative chromosome painting experiments (**Figure S7**). This discrepancy in gene density permitted the classification of the genomic regions into two distinct groups corresponding to the two subgenomes: the less fractionated subgenome (LF) and the more fractionated subgenome (MF). The LF subgenome harbored more protein-coding genes and generally a larger total length of TEs (12,700 – 14,899 genes; 142.5 – 433.2 Mb TEs) compared to the MF subgenome (9,656 – 11,025 genes; 95.0 – 288.2 Mb TEs; **Figures 2b** and **2c**), while the proportion of different TE categories was comparable between subgenomes in all species (**Figure 2c**), suggesting similarity of the two progenitor genomes in TE content prior to the hybridization.

Subgenome-aware maximum likelihood (ML) and coalescent analyses were performed on orthologous gene groups from eight *Biscutella* species and 15 other Brassicaceae genomes representing major crucifer lineages to infer the phylogenetic placement of the putative progenitor species. Considerable topological discordance was observed among 1,234 syntenic gene trees and the coalescent-based tree (**Figure 2d**). The internodes with conflicting topologies were particularly evident in the placements of *B. auriculata*, the neotetraploid *Heldreichia bupleurifolia* (Biscutelleae), and the basal node of the Biscutelleae clade (the diploid *Megadenia pygmaea*, and the mesotetraploid *Lunaria* and *Ricotia* species), suggesting a complex, non-bifurcating phylogeny. The ten distinct topologies observed among the phylogenetic trees constructed separately for the 22 GBs (**Figure S8**) consistently identified

two coherent clades: one comprising all *n* = 9 species (*B. varia*, *B. austriaca*, *B. prealpina*, and *B. frutescens*) and another comprising *B. baetica*, *B. didyma* and *B. lyrata*. Despite different placement of *B. auriculata* in the nuclear gene dataset, the coalescent-based tree and the chloroplast tree (**Figure S9**) positioned *B. auriculata* as a sister to the *n* = 9 clade, despite its chromosome number being the same as in *B. baetica* and *B. didyma* (*n* = 8). The LF subgenome formed a sister clade with the neotetraploid *H. bupleurifolia*, whereas the MF subgenome formed a monophyletic clade outside this sister group, confirming the hierarchical clustering of *K*-mer abundance and principal component analysis (**Figure S10**).

The ancestral *Biscutella* genome originated from hybridization between two divergent progenitor species around 11 – 13 million years ago (Mya), based on the time-calibrated trees (**Figures 2e** and **S11**) and the distribution of synonymous substitution rate for homeologue pairs (*K*s-value peak = ~0.33; **Figure 2f**). Time-calibrated phylogenies revealed that the divergence of *B. lyrata*, *B. didyma*, and *B. baetica* (~10.4 Mya; hereafter referred to as "early-diverging species") significantly predates that of the *n* = 9 species (~3.3 Mya; hereafter referred to as "late-diverging species"; **Figure 2e**). Although the coalescent-based tree retrieved *B. auriculata* as sister to the *n* = 9 clade (**Figure 2d**), it diverged from the *n* = 9 species as early as ~11.8 Mya (and is therefore also categorized asan "early-diverging species").

Reconstruction of the ancestral mesotetraploid genome of *Biscutella*

The ancestral karyotype of *Biscutella* genomes was reconstructed using both top-down and bottom-up strategies. The top-down approach included comparing frequent GB associations in *Biscutella* species with those in the known ancestral karyotypes of Brassicaceae (e.g., ACK, ancPCK, and PCK; Schranz et al., 2006; Mandáková & Lysak, 2008; Geiser et al., 2016; Lysak et al., 2016). The ancPCK karyotype (*n* = 8) best explained the observed GB associations in *Biscutella*, with 73.3% and 66.7% of GB associations in ancPCK also found in the LF and MF subgenomes, respectively (**Figures S12**). A shared paracentric inversion event (Ipa), which restructured the ancestral chromosome AK8/6 by rearranging the V+Wa+Q+X association into V+X+Q+Wa, was identified in both subgenomes in all *Biscutella* species, as well as in two other Biscutelleae species – *M. pygmaea* and *H. bupleurifolia* (**Figures 3a** and **S12a**). Furthermore, additional conserved GB associations identified in the MF subgenome, e.g., D+E+C and S+V+X+Q+Wa, but absent in the LF subgenome (**Figure S12a**), indicate that the two diploid *Biscutella* ancestors had distinct karyotypes. Based on these observations, we infer that the LF subgenome originated from an ancPCK-like genome (*n* = 8) with the Ipa on chromosome AK8/6, and that the MF subgenome underwent additional nested chromosome insertion (NCI) and end-to-end translocation (EET) events, which reduced its chromosome number to six (*n* = 6), while preserving more ancestral GB associations during subsequent speciation events (**Figure S12a**). To further validate the accuracy of the top-down inference, we used a bottom-up strategy with the WGDI tool (Sun et al., 2022) to hierarchically reconstruct ancestral karyotypes at each phylogenetic node, from youngest to oldest (**Figure S13**). The most recent common allotetraploid ancestral karyotype of *Biscutella* was eventually reconstructed and matched the results of the top-down strategy (**Figures 3a** and **3b**). Our integrative analysis revealed that the ancestral mesotetraploid *Biscutella* genome comprised 14 chromosome pairs (*n* = 6 + 8; 2*n* = 4x = 28)

formed through hybridization between two distinct parental genomes: LF (*n* = 8) and MF (*n* = 6) (**Figure 3a**).

Independent descending dysploidy

The post-polyploid evolution of the allotetraploid ancestor (*n* = 14) to the modern species (*n* = 9, 8 and 6) was reconstructed by integrating multiple CR events, including EET, reciprocal translocation (RT), unequal reciprocal translocation (T_{uneq}), NCI, Ipa and pericentric inversion (Ipe) (**Figure 3b**, see **Figures S14 – S17** fordetailed evolutionary pathways). Only one NCI event is shared by all eight species (**Figure 3b**), whereas the remaining CRs are clade- or species-specific. The early-diverging species experienced 2 to 14 independent, private CR events, while the late-diverging species had only one to three. Unequal reciprocal translocations were the predominant CR type across all *Biscutella* genomes (5 to 9 events per species) and, when accompanied by stable centromere inactivation in dicentric chromosomes (**Figure S14**), contributed to the reduction of the ancestral chromosome number (*n* = 14) to present-day chromosome numbers. Despite the identical chromosome number (*n* = 8), *B. auriculata* and the *B. baetica*–*B. didyma* clade followed distinct evolutionary trajectories. The descending dysploidy from *n* = 14 to *n* = 8 in *B. auriculata* involved 15 CRs, primarily driven by Ts_{uneq} (8), EETs (3), and NCIs (2). However, *B. baetica* and *B. didyma* underwent 12 and 16 CRs, respectively, with major contributions from Ts_{uneq} (7 and 9), EETs (3 and 2), and RTs (1 and 3) (**Figure 3b**). A comparable number of CRs, 15 and 13, was observed in the *n* = 9 genomes, also dominated by Ts_{uneq}, EETs, and RTs. The highest number of CRs (20) was identified in the highly diploidized *B. lyrata* (*n* = 6), although the level of descending dysploidy and T_{uneq} frequency were not correlated, e.g., *B. didyma* (*n* = 8) exhibited 9 Ts_{uneq}, whereas *B. lyrata* (*n* = 6) had 8 Ts_{uneq}. Furthermore, the CR rate varied along the phylogeny, peaking during early divergence (between A5 and A4/A3) with an average rate of 3.75 chromosomal rearrangements per million years (CRs/Mya) and decreasing to ~0.72 CRs/Mya towards the tips (between A4/A3 and extent species; **Figure 3b** and **Table S9**).

Ty3/Gypsy retrotransposons show a higher removal rate in early diverging *Biscutella* species

Although LTR-RT sequences represent the predominant component of repeats in the *Biscutella* genomes, intact LTR-RT elements accounted for only about 18.7% of the total LTR-RTs (**Table S6**). Most of these intact LTR-RTs are relatively young, with the average insertion times of the two major superfamilies – Ty3/Gypsy and Ty1/Copia – both within 1 million years (**Figure 4a**). Alongside their recent amplification, most LTR-RTs are also subject to rapid deletion, as we identified a substantial number (4,869 – 8,401; **Table S10**) of recombination remnants defined as soloLTRs, that is, LTR-RT sequences where the internal domain and one of the two LTRs have been deleted (**Methods**). Analysis of different LTR-RT lineages revealed that the Athila and CRM dominated the landscape and collectively accounted for 65% of intact LTRs and 55.9% of soloLTRs (**Figure 4b**). CRM elements were primarily concentrated around pericentromeric regions, while Athila elements were more randomly distributed throughout the chromosomes (**Figures 1b**, **S3,** and **S4**). Moreover, the Ty3/Gypsy lineages exhibited higher removal rates

in early-diverging species than in late-diverging species, as shown by the increased solo:intact LTR ratios (S/I ratios) for Athila, CRM, Retand, and Tekay (average S/I ratio of ~0.577 in late-diverging species vs. ~1.290 in early-diverging species; **Figure 4c**). Nevertheless, no specific pattern of soloLTR loss was observed between subgenomes (**Figure 4d**), indicating that the deletion of LTR-RTs (at least through soloLTR formation) is not linked to gene fractionation.

Two major *Biscutella* species clades show contrasting chromatin organization

Hi-C data provided a higher-order perspective of chromatin organization in the *Biscutella* genomes. The Hi-C maps revealed that contact densities were predominantly concentrated along the main diagonals, with weak *trans*-interactions (**Figures 4e** and **S18**). PCA-based analysis of Hi-C contact data at 50-kb resolution partitioned the *Biscutella* chromosomes into distinct A and B compartments (**Figures 4e** and **S18**). Consistent with observations in other plant genomes (e.g., pepper and soybean; Liao et al., 2022; Ni et al., 2023), the A compartments were located near the telomeres whereas B compartments occupied the central, repeat-rich regions of the chromosomes, consistent with the global distribution of gene and TE sequences (**Figures 4f** and **S18**). With the exception of the smallest *B. auriculata* genome, which contains more A compartments (50.6% of the genome), the remaining genomes had a slightly higher proportion of B compartments (~53%; **Figure 4g**). The examination of orthologous gene pair composition between any two species demonstrated the variability in the A/B compartment assignments. The proportion of orthologous gene pairs assigned to different chromatin compartments was significantly higher in the early-diverging species (14.63% – 30.42%) than the late-diverging species (3.51% – 7.26%; **Figure 4h**). Furthermore, the organization of topologically associated domains (TADs) also underwent significant restructuring during post-polyploid diploidization. The late-diverging species had more and relatively shorter TADs (942 – 1,050 TADs, 0.65 – 0.72 Mb), whereas early-diverging species showed considerable variability in TAD number and length (502 – 1,039 TADs; 0.81 – 1.35 Mb), with the fewest but longest TADs in *B. lyrata* (502 TADs; 1.35 Mb average length; **Figures 4i** and **4j**). TAD conservation, assessed by comparing orthologous gene pairs at TAD boundaries (**Methods**), indicated that only ~9.49% of TADs were conserved between any two species, with a higher proportion of conserved TADs retained between late-diverging species (11.09% – 13.44%) than between early-diverging species (5.57% – 10.43%, **Figure 4k**).

Chromosome breakage hotspots are associated with LTR-RT enrichment and changes in 3D genome organization

Synteny comparisons of eight *Biscutella* genomes with their ancestral (sub)genomes revealed 12 breakpoints shared by multiple species and designated as breakage hotspots (HOT regions; **Figures 5a** and **S19**; **Methods**). In the ancestral *Biscutella* genome, eight HOT regions belonged to the LF subgenome (n = 8) and four to the MF subgenome (n = 6) (**Figure 5a**). Based on the inferred ancestral centromere positions, which we define as paleocentromeres (Yang et al., 2021; Guo et al., 2021), seven HOTs were associated with breaks at/near these regions, while five were located within ancestral

chromosome arms (**Figure 5a**). In the extant *Biscutella* genomes, more than half of the rejoined junctions aligned with pericentromeric regions (**Figure 5b**). Analysis of the LTR-RT content in 100-kb regions on each side of rejoined junctions showed that 67.5% of these regions had higher LTR-RT content than the genome-wide average, mainly contributed by Athila and CRM retroelements (**Figure 5c**). This is consistent with previous findings in Arabidopsis and *Orychophragmus* (Jiao & Schneeberger, 2020; Zhang et al., 2023a), suggesting that interspersed stretches of similar TEs may be substrate for NAHR that facilitates the recurrent occurrence of some CRs.

To assess the relationship between 3D genome organization and chromosome breakage, we examined A/B compartment assignments and TAD distribution within the HOT regions. We found that up to 62.5% of the breakpoint junctions showed inconsistent A/B compartment assignments and 75% matched TAD boundaries across *Biscutella* genomes (**Figure 5b**). For example, HOT7, a hotspot at/near the paleocentromere of ancestral chromosome AK8, resulted in the rearrangement of blocks MN and KL on arms of chromosome 1 and 4 in *B. lyrata.* The distal end of the MN block likely contains paleocentromeric remnants, characterized by gene depletion and CRM accumulation (**Figure 5d**). Despite being located within a chromosome arm, the rejoined junction at MN block remains in the B compartment, whereas the KL junction occupies the A compartment (**Figure 5d**), indicating an A/B compartment shift. More interestingly, the rejoined junction at the MN block colocalized with TAD boundaries (**Figure 5d**), suggesting a relationship between chromosome break and TAD architecture. This overlap implies that breakpoints preferentially occurred at conserved TAD boundaries (Okhovat et al., 2023; Li et al., 2023), highlighting the presumed role of TADs as functional entities during chromosome evolution.

Gene duplication and deletion, and gene sub/neofunctionalization contributed to the emergence of unique traits in *Biscutella*

To explore the preferential retention of WGD-derived genes following infrageneric cladogenesis and speciation, we identified 4,572 syntenic core gene families shared by all eight *Biscutella* genomes. Of these, 1,049 families (22.94%) retained duplicates (core duplicates, CD, i.e., paralogues retained in both subgenomes) and 3,523 families (77.06%) retained singletons (core singletons, CS, i.e., genes specific to one subgenome, **Figure 6a**). In addition, we identified 246 syntenic softcore gene families exclusive to either late- or early-diverging species (softcore duplicates and singletons, SD and SS) and 466 private singletons (PS) specific to individual species (**Figure 6a**). No species-specific duplicates were detected. Core gene families, especially CDs, displayed lower *K*a/*K*s ratio (**Figure 6b**), were flanked by fewer TEs (**Figure 6c**), and showed significantly higher expression levels (**Figure 6d**), suggesting they have undergone more stringent selective pressures. Moreover, TE accumulation flanking retained genes was comparable between late- and early-diverging species, yet the late-diverging species had significantly higher *K*a/*K*s ratios (**Figures 6e** and **6f**), indicating more relaxed purifying selection.

Functional differences between the retained duplicates and the singletons were assessed through Gene Ontology (GO) functional enrichment. Only three GO terms – “plasma membrane”, “regulation of transcription” and “endoplasmic reticulum” – were shared between the CD and CS gene families (**Figure 6g**), while 84 and 44 unique GO terms were identified for each gene family (**Tables S11** and **S12**), revealing functional biases for gene retention. A substantial number of GO terms involved in environmental adaptation were specific to the CD family, particularly those related to ionic and salinity stress, e.g., “response to cadmium ion”, “response to salt stress”, and “hyperosmotic salinity response” (**Figure 6g**). The over-representation of ion/salt response-related functions align with the unique ecological niche of *Biscutella*, i.e., more than 50 species are native to rocky, dry and warm habitats throughout the Mediterranean (Jalas et al., 1996). Notably, domain examination between paralogues revealed 45.28% of CD duplicates (475/1,049 pairs) exhibited divergent protein domain compositions, with 220 pairs showing species-specific alterations (**Figure S20**). GO terms related to chloroplast function and DNA damage repair were exclusively enriched in the CS family, whereas they were underrepresented in the CD family (**Figure 6g**). This is consistent with results in 20 flowering plants (De Smet et al., 2013), as these functions likely involve dosage-sensitive gene networks that constrain gene loss (Edger & Pires, 2009; Makino & McLysaght, 2010). Although duplicates and singletons showed significant functional divergence, no GO terms were enriched among genes specifically retained in early- and late-diverging species.

In contrast to over-retention of genes related to environmental adaptation, the development of lens-like silicles in *Biscutella* might be associated with excessive gene deletions and relaxed selection. Compared to the slender, cylindrical siliques of *Arabidopsis* or *Brassica*, *Biscutella* species produce characteristic didymous flattened silicles with a prominent and accumbent radicle (**Figure 6h,** Murley, 1951; Vaughan & Whitehouse 1971). We examined copy number variation in homeologues involved in the ABCDE flowering model (Theißen, 2001; Theißen et al., 2016; **Figure 6h**) and found that, unlike A-, B-, C-, or E-class genes, which regulate sepal, petal, stamen, and carpel development, respectively, all D-class genes in *Biscutella*, including SEEDSTICK (STK), SHATTERPROOF1 (SHP1), and SHP2, specifically regulate ovule development, transitioned from duplicates to singletons (**Figure 6i**). In addition, D-class genes in *Biscutella* species experienced relaxed selection, as indicated by their generally higher *K*a/*K*s ratios compared to other seven Brassicaceae species (**Figure S21**). Specifically, SHP1, a MADS-box gene member specifying ovule integument identity in *A. thaliana* (Ehlers et al., 2016), remains intact in the mesohexaploid *Br. rapa* (three copies) and the mesotetraploid *Pugionium cornutum* (two copies), but has reverted to a singleton or even lack syntenic SHP1 copies in *Biscutella* species (**Figure 6i** and **Table S13**). The retained SHP1 singletons exhibit lower motif conservation in the NCBI-CDD database compared to other Brassicaceae species (**Figure S22**). Given the key roles of D-class genes in ovule primordial development, we propose that their fractionation and relaxed selection contributed to the evolution of lens-like silicles in *Biscutella*. This finding aligns with observations in *Arabidopsis*, where loss-of-function mutants of D-class genes produce shorter siliques with rounder, smaller seeds (Pinyopich et al., 2003; Di Marzo et al., 2020).

Discussion

The three base chromosome numbers in *Biscutella* are not exceptional among angiosperm monophyletic genera. In crucifers, multiple base numbers need not be directly associated with mesopolyploidy and the diploidization process, but more commonly post-polyploid diploidization is associated with independent reductions in chromosome number, rendering genera and clades polybasic (e.g., Mandáková et al., 2017a, 2017b). Modern *Biscutella* species are descendants of an allotetraploid ancestral genome that originated 13 to 11 Mya and whose chromosome number (n = 14) was reduced 1.6 to 2.3-fold to n = 9, 8 and 6 (Geiser et al., 2016; Guo et al., 2021; Beringer et al., 2024).

We show that although modern *Biscutella* species differ in chromosome number, degree of chromosomal diploidization and divergence times, the independent post-polyploid descending dysploidy of all *Biscutella* species with n = 8 and 9 involved 12 to 16 CRs (20 CRs in *B. lyrata*, n = 6), with a comparable number of end-to-end translocations, nested chromosome insertions and chromosome translocations. This is consistent with inferred CR types mediating descending dysploidy in other eudicots (Feng et al., 2024; Sun et al., 2024) and presumably consistent with genome instability after WGD (Raeside et al., 2014; Tong et al., 2025). The early stages of cytological diploidization in *Biscutella* were indeed accompanied by high TE dynamics (Beringer et al., 2024) and higher CR rates, resulting in extensive chromosomal restructuring, compared to the later stages characterized by lower CR rates. Thus, initial inter-subgenome homogenization through CRs and descending dysploidy associated with gene fractionation may gradually slow down the process of cytological diploidization.

Interestingly, we identified 12 chromosome breakage hotspots (HOTs) shared by several *Biscutella* species. HOTs have higher LTR-RT content, supporting regions with frequent DSBs and NAHR, as evidenced here by HOTs localized within paleocentromeric regions enriched in Athila and CRM being still detectable at chromosomal junctions. HOTs also frequently colocalize with boundaries of TADs, suggesting that CRs are not randomly distributed with respect relative to higher-order chromatin (A/B) compartments and local TADs (Li et al., 2023). In contrast to the highly conserved TADs observed in the diploid subgenomes of more recent polyploids such as cotton (origin ~ 1–2 Mya; Wang et al., 2018) and horseradish (~ 5 Mya; Shen et al., 2023), *Biscutella* species retained only a limited number of conserved A/B compartments and TADs between orthologs, underscoring the erosion of chromatin topologies over deep evolutionary time scales.

Online Methods

Plant material, library preparation and sequencing

Plant material: The following *Biscutella* accessions were used: *B. laevigata* subsp. *varia* (V12-4; Germany, Beuron), *B. laevigata* subsp. *austriaca* (Jord.) Mach.-Laur. (A2Schnee 3B; Austria, Schneealpe Altenberg), *B. prealpina* Raffaelli & Baldoin (RCBO_NC17; Italy, Recoaro Terme), *B. frutescens* Coss. (PI 650129, USDA collection; Spain), *B. auriculata* L. (PI 650127, USDA collection; Spain, Ames), *B. didyma* L.

(LBN-00579, LBN seed bank; Lebanon), *B. baetica* Boiss. & Reut. (Gaucín; Spain, Gaucín), and *B. lyrata* L. (Cádiz; Spain, Cádiz).

B. auriculata is the type species of section *Iondraba* (Medik.) DC., whereas the remaining 7 species belong to section *Biscutella* that includes two different lineages: series *Biscutellae* with *B. baetica*, *B. didyma* and *B. lyrata*, and series *Laevigatae* Malin. with *B. frutescens*, *B. laevigata* subsp. *austriaca*, *B. laevigata* subsp. *varia,* and *B. prealpina* (Olowokudejo, 1986; Vicente et al., 2020). The plants were grown from seed and cultivated under standard conditions (21/18°C, 16/8 h of light/dark cycle) in growth chambers.

***Illumina*:** Genomic DNA was isolated from young leaf tissue using the NucleoSpin Plant II kit (Macherey-Nagel). Illumina sequencing libraries were prepared using the TruSeq Nano DNA HT Sample preparation kit following the manufacturer's recommendations. Genome sequencing was performed using the Illumina HiSeq Xten platform (Illumina; San Diego, CA, USA).

***Nanopore and PacBio HiFi*:** High-molecular-weight (HMW) DNA for long-read sequencing was extracted from approximately 1 g of young leaves (following a 2–3 day dark treatment) using the "Arabidopsis (*Arabidopsis thaliana* LEr) leaf DNA" protocol (https://nanoporetech.com/document/extraction-method/arabidopsis-leaf-dna) with minor modifications. To remove DNA fragments shorter than 10 kb, the Short Read Eliminator (SRE) XS kit (PacBio) was used. The quality and quantity of the extracted genomic DNA were assessed using NanoDrop 2000c Spectrophotometer (Thermo Scientific), Qubit 4 Fluorometer (Thermo Scientific), and Genomic DNA ScreenTape assay (Agilent).

Long-read sequencing processes followed standard protocols of Oxford Nanopore Technologies (ONT; Oxford, UK) and PacBio HiFi technology (PacBio; San Diego, CA, USA). DNA libraries of *Biscutella* species with an average fragment length of >20 kb were constructed for ONT sequencing and sequenced on the PromethION platform. The PacBio circular consensus sequencing (CCS) library for *B. lyrata* was additionally produced and sequenced on one SMRT cell of the PacBio Sequel II system.

***Hi-C and Omni-C*:** Hi-C libraries were prepared from leaf samples using the Proximo Hi-C Kit according to the manufacturer's protocol at Phase Genomics (Seattle, USA). The Hi-C libraries were sequenced on Illumina HiSeq X-Ten instruments to generate 150-base paired-end reads. The Hi-C library for *B. lyrata* was generated using the Omni-C Proximity Ligation Assay (Dovetail Genomics, Scotts Valley, CA) following the manufacturer's "Non-Mammalian Samples Protocol version 1.2B". The library was then sequenced on the Illumina HiSeq X Ten platform.

***IsoSeq and RNAseq*:** Total RNA was extracted from leaf, flower, and root tissues using the Quick-RNA Miniprep Kit (Zymo Research). The quality and quantity of the extracted RNA were assessed using NanoDrop 2000c Spectrophotometer (Thermo Scientific), Qubit 4 Fluorometer (Thermo Scientific), and Fragment Analyzer (Agilent). Total RNAs from the tissues of leaves and flowers were mixed equally, along with individual RNA from root tissues, for long-read (PacBio Iso-Seq) sequencing technology to generate transcriptome data. The Iso-Seq cDNA libraries were constructed according to the PacBio

standard protocol and sequenced on the PacBio sequel II platform. The short-reads RNA-seq data for *B. baetica* and *B. lyrata* were downloaded from Guo et al., 2021. The outputs of all sequencing reads are summarized in **Table S1**.

Genome size measurement by flow cytometry and *K*-mer frequency

Holoploid genome size was estimated by flow cytometry. The young intact leaf, ~1 cm in length, were prepared according to Doležel et al. (2007). The samples were stained using a solution containing propidium iodide + RNAase IIA, both at final concentrations of 50 µg/ml, for 5 min at room temperature and analysed using a CyFlow cytometer Partec equipped with a 532 nm diodepumped solid-state laser Cobolt Samba. A fluorescence intensity of 5,000 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). Three different samples for each species measured on three consecutive days was used for genome size estimation. Genome sizes of *Biscutella* species were further confirmed by *K*-mer frequency (*K* = 21) analysis with the findGSE (v1.94; Sun et al., 2018), after counting 21-mers with Jellyfish (Marcais & Kingsford, 2011).

Genome assembly, scaffolding, and quality assessment

We employed two strategies for *de novo* genome assembly: an ONT-based strategy for *B. auriculata*, *B. baetica*, *B. didyma*, and all *n* = 9 species (Parisod et al., 2025), and a PacBio-based strategy for *B. lyrata*. Initial assemblies were generated from Nanopore long reads using NextDenovo (v2.2; https://github.com/Nextomics/NextDenovo) with its standard pipeline, and the resulting contigs were polished in two iterative rounds with both long and short reads using NextPolish (v1.1.0; Hu et al., 2020). For *B. lyrata*, haplotype-resolved assemblies were generated using hifiasm (v0.19.5; Cheng et al., 2021) in hybrid mode by integrating PacBio HiFi reads, ONT reads (>= 25kb), and Hi-C reads. The program Khaper (Zhang et al., 2021) was used to select primary contigs and filter redundant sequences from the initial assemblies of the highly heterozygous genomes *B. auriculata* and *B. baetica*. The sizes, contig numbers, and contig N50 values of the draft genome assemblies are summarized in **Table S2**. Completeness of genome assemblies was evaluated using BUSCO (v3.0.1; Simão et al., 2015) with embryophyta_odb10 database (1,614 total BUSCOs). For assemblies generated using the ONT-based strategy, Hi-C reads for each genome were aligned to the corresponding contigs using Juicer (v1.5.7; Durand et al., 2016). The 3D-DNA pipeline (v180922; Dudchenko et al., 2017) was used to correct potential mistakes and to order, orient and scaffold the sequences. For *B. lyrata*, the Omni-C data was used to scaffold the genome assembly by YaHS (v1.1; Zhou et al., 2023) with default parameters. The linkage results were manually curated to correct misjoins and misassemblies based on visualization using JuiceBox (v1.1.08; Robinson et al., 2018). The chloroplast genome was assembled based on Illumina short reads using the GetOrganelle toolkit (v1.7.7; Jin et al., 2020).

Gene prediction and functional annotation

Protein-coding genes were predicted for *n* = 8/6 species using an evidence-based annotation workflow that integrated multiple sources of evidence. The gene annotations for *n* = 9 species were sourced from Parisod et al. (2025). For transcriptome-based predictions, RNAseq data from *B. baetica* and *B. lyrata* were mapped using Hisat2 (v2.1.0; Kim et al., 2015) and subsequently assembled into transcripts by StringTie (v2.1.4; Pertea et al., 2015). IsoSeq datasets for each species were aligned to the genome assemblies using GMAP (v2018-07-04; Wu & Watanabe 2005). All transcripts from RNAseq and IsoSeq were merged with StringTie into a pool of candidate transcripts. TransDecoder (v5.5.0; http://transdecoder.github.io) identified potential coding regions in the resulting transcripts. Additionally, two rounds of PASA (v.2.3.3; Haas et al., 2003) were conducted to refine gene models by identifying untranslated regions and isoforms, using transcripts generated by genome-guided Trinity (v2.11.0; Haas et al., 2013) assemblies. *De novo* gene predictions were generated using AUGUSTUS (v3.4.0; Stanke et al., 2006), with *Biscutella*-specific AUGUSTUS gene models trained using GeneMark-ET (v4.0; Lomsadze et al., 2014). This model leveraged RNA-seq and IsoSeq evidence in two iterative rounds of predictions to refine parameters. For annotation of homologs, protein sequences from *Arabidopsis thaliana*, *Cadarmine hirsuta*, *Eutrema salsugineum*, *Thlaspi arvense*, and *Brassica rapa* were aligned to each *Biscutella* genomes to identify the homologous genes using GenomeThreader (v1.7.1; Gremme et al., 2005). Finally, all gene predictions were integrated into a final gene model set using EVidenceModeler (v1.1.1; Haas et al., 2008) after removing pseudogenes and non-coding genes using a custom Python script. FeatureCounts (Liao et al., 2014) was used to extract the mapped reads for each gene in *B. baetica* and *B. lyrata*, allowing for the calculation of transcripts per million (TPM) values as the expression level of the genes.

Functional assignments for the predicted protein-coding genes were performed with BLAST (v2.12.0+; Altschul et al., 1990) to align coding sequences against public protein databases, including NCBI non-redundant (NR) protein (Pruitt et al., 2007), SwissProt (Bairoch & Boeckmann, 1992), and InterProScan (Hunter et al., 2009). Gene Ontology (GO) terms for each gene were provided by InterProScan.

Identification of syntenic genes and fragments

Syntenic orthologs of *Biscutella* genes within 22 GBs (Schranz et al., 2006; Lysak et al., 2016) were identified using SynOrths (Cheng et al., 2012), based on both sequence similarity and the sequence homozygosity of their flanking genes. Syntenic gene pairs that were continuously distributed along the *Biscutella* genomes and 22 GBs were considered as ancestral fragments inherited from the progenitors. Due to local structural variations and potential genome assembly errors, local syntenic gene pairs may not be distributed immediately adjacent to other syntenic genes. Thus, when two syntenic gene pairs were interrupted by fewer than 50 genes or had a distance of less than 300 kb, they were combined into a pair of syntenic fragments. A total of 50 to 55 syntenic fragments were identified in eight genomes,

including 11 GB breaks (**Table S5**). Macrosynteny relationships across multiple species were visualized as a riparian plot (**Figure 1d**) using NGenomeSyn (He et al., 2023).

Repetitive element annotation

The Extensive *de novo* TE Annotator (EDTA v1.8.3; Ou et al., 2019) was utilized to annotate TEs in each *Biscutella* species with the following parameters: "--species others --step all --anno 1". Within EDTA, LTRharvest (Ellinghaus et al., 2008), LTR_FINDER_parallel (Ou & Jiang, 2019) and LTR_retriever (Ou & Jiang, 2018) were used for LTR-RTs identification. Tandem repeats were identified using TRASH (Wlodzimierz et al., 2023). The ribosomal DNA (rDNA) sequences were predicted with Barrnap (v0.9; https://github.com/tseemann/barrnap) using the Eukaryota database. For precise classification of LTR-RTs at the lineage level, TEsorter (v1.4.6; Zhang et al., 2022) was employed. SoloLTRs were identified using the new soloLTRseeker pipeline (https://github.com/estpr/soloLTRseeker). In brief, soloLTRseeker first generates a high-quality non-redundant LTR library using LTRs of intact elements classified in advance to LTR lineages by TEsorter. It then runs BLASTn on the genome with 99% and 80% thresholds for query coverage and sequence identity respectively. 200-bp sequences flanking each putative soloLTR locus are compared with sequences of the same length from both termini of the internal domain of intact elements of the same lineage to filter out cases of partially deleted TEs (that are not soloLTRs). Finally, for each candidate, upstream and downstream sequences of 6-bp length are locally aligned to annotate target site duplications.

The coordinates of syntenic genes at the ends of GBs flanking gene-poor regions were used to delineate the boundaries of pericentromeric regions. DNA compression was generated using the context-tree weighting (CTW) function of the BCT package in R (https://www.rdocumentation.org/packages/BCT/versions/1.2) to further refine centromere localization. Tandem repeats that were significantly enriched in pericentromeric regions and overlapped with the troughs of the CTW values were identified as centromeric satDNA candidates, and those located at the ends of chromosomes were identified as (sub)telomeric satDNA candidates. The screened candidates were further validated through FISH experiments.

Insertion ages of LTR-RTs were estimated using the method proposed by SanMiguel et al. (SanMiguel et al., 1998), which compared the 5'- and 3'-LTRs of each full-length element. Nucleotide substitutions per site (K) between LTR pairs were calculated using Kimura's two-parameter model (Kimura, 1980). Following Koch et al. assumptions (Koch et al., 2000), we employed the mutation rate (r) of 1.5×10^{-8} substitutions per year per synonymous site and calculated the insertion times (T) of LTR-RTs with the formula T = K/2r.

FISH experiment and comparative chromosome painting

The mitotic and meiotic (pachytene) chromosome spreads were prepared from young anthers. Oligoprobes (60-bp in length) were used to visualize the identified tandem repeats on chromosomes (**Table S8**). The most conserved regions within the consensus sequences of monomers were selected, with a preference for regions with low GC content (30–50%) and minimal self-annealing. For chromosomal localization of conserved genomic blocks (Lysak et al., 2016), *A. thaliana* BAC clones were assembled to represent the 22 GBs of the ancestral crucifer karyotype (Lysak et al., 2016). DNA probes were labeled with biotin-dUTP, digoxigenin-dUTP, or Cy3-dUTP by nick translation as described by Mandáková & Lysak (2016). Labeled BAC DNAs were pooled, precipitated, and resuspended in 20µl of hybridization mixture (50% formamide and 10% dextran sulfate in 2×SSC) per slide. Labeled 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 for 16 to 72 hours. Post-hybridization washing was performed in 20% formamide in 2×SSC at 42°C. Fluorescence signals were analyzed with an Axioimager Z2 epifluorescence microscope (Zeiss) and CoolCube CCD camera (MetaSystems).

Phylogenetic analysis

To infer the ancestral origin of the identified syntenic fragments, gene trees were inferred using maximum likelihood (ML) and coalescent-based analyses for each fragment. The protein-coding genes of eight *Biscutella* species and 15 other Brassicaceae species (*Aethionema arabicum*, *Euclidium syriacum*, *Arabis alpina*, *Pugionium cornutum*, *E. salsugineum*, *Schrenkiella parvula*, *Br. rapa*, *C. hirsuta*, *Boechera stricta*, *Arabidopsis halleri*, *A. thaliana*, *Lunaria rediviva*, *Ricotia lunaria*, *Megadenia pygmaea*, and *Heldreichia bupleurifolia*) were used to generated syntenic gene groups. For each gene group, coding sequences were aligned using MAFFT (v7.427; Katoh et al., 2002) and trimmed using TrimAL (v1.4; Capella-Gutiérrez et al., 2009). ML analyses were performed using IQ-TREE (v1.6.11; Nguyen et al., 2015) with default parameters. Coalescent-based inference of the species tree was performed using ASTRAL-Pro (v1.1.3), which allows species-tree inference in the presence of paralogy (Zhang et al., 2020). To further visualize single-gene tree conflicts, the Python package Toytree (v.3.0.5; Eaton, 2020) was utilized in the construction of cloud tree plots.

Subgenome phasing and validation

Using the 22 GBs as the reference, two genomic copies of each diploid chromosome in the tetraploid ancestor of *Biscutella* were identified. To reconstruct the two subgenomes, three key rules were applied: (i) no overlapping or redundant regions were allowed within each reconstructed chromosome of the subgenomes; (ii) each ancestral chromosome adhered to the gene density distribution pattern (LF > MF) between the subgenomes; and (iii) the phylogenetic relationship of the syntenic gene fragments in the 22 GBs was consistent with the two subgenomic branches. Based on these rules, syntenic fragments were categorized into two groups: the less fractionated (LF) and more fractionated (MF) subgenomes, following the nomenclature for *Brassica* subgenomes (Cheng et al., 2012). To validate the subgenome

phasing, hierarchical clustering of subgenome-specific repetitive DNA sequences and principal component analysis were performed using SubPhaser (v1.2; Jia et al., 2022).

Calibration of the WGD timing

We performed self-to-self BLASTP alignments for *Br. rapa*, *A. thaliana*, *L. rediviva*, *R. lunaria*, *M. pygmaea*, and *H. bupleurifolia*, respectively, and selected the best hits among the homologous gene pairs with identity > = 90%. Paralogues from the eight *Biscutella* species and homologues from other species were used for the non-synonymous (*K*a) and synonymous (*K*s) rate calculations. Each pair of protein sequences was aligned by MAFFT and pairwise nucleotide sequence alignments were generated by transforming protein alignments into codon alignments with ParaAT (v2.0; Zhang et al., 2012). The *K*s values and *K*a/*K*s ratios were calculated based on the Nei–Gojobori method implemented in KaKs_Calculator (v2.0; Wang et al., 2010). Timing of subgenome divergence was calculated according to the formula:

$$T = Ks/2r$$

where r is the mutation rate of 1.5×10^{-8} substitutions per year per synonymous site (Koch et al., 2000).

To further validate divergence times for *Biscutella*, two time-calibrated trees were constructed using r8s (v1.81; Sanderson et al., 2003), PATHd8 (v1.0; Britton et al., 2007) and RelTime method in MEGA X (Kumar et al., 2018). A total of 94 single-copy ortholog groups, identified using OrthoFinder (v2.5.4, Emms & Kelly, 2019), were used to construct a ML phylogenetic tree with IQ-TREE. The divergence at 20.6 Mya between crucifer Lineage I and Lineage II (http://www.timetree.org/) was used as the calibration node.

Reconstruction of the ancestral karyotype and identification of chromosome breakage hotspots

We utilized both top-down and bottom-up strategies to reconstruct the ancestral *Biscutella* karyotype. The top-down strategy involved comparing GB associations in *Biscutella* species with established ancestral karyotypes of the Brassicaceae (ACK, ancPCK, and PCK; Schranz et al., 2006; Mandáková & Lysak, 2008; Geiser et al., 2016; Lysak et al., 2016). Conserved GB associations shared among multiple species were considered to be inherited from their ancestral genome. For validation, we utilized the WGDI tool (Sun et al., 2022) following a bottom-up strategy, which is applicable to species with unknown ancestral karyotypes, as applied in Lamiales and Buxales (Wang et al., 2022; Wang et al., 2024). Synteny maps generated by WGDI among *Biscutella* species enabled the inference of ancestral karyotypes at various evolutionary nodes based on phylogeny (**Figure 3b**). Intact chromosomes with continuous synteny were initially identified as ancestral chromosomes. Chromosomal breaks or fusions shared by multiple species, as well as those unique to particular species, were systematically characterized. We

traced the origins of these breaks or fusions for each species hierarchically and considered six typical chromosome rearrangement (CR) types, i.e., end-to-end translocation (EET), reciprocal translocation (RT), unequal reciprocal translocation (T_{uneq}), nested chromosome insertion (NCI), paracentric inversion (Ipa), and pericentric inversion (Ipe), which allowed us to reconstruct the karyotype for each node sequentially, progressing from the youngest to the oldest (nodes A1 to A5; **Figures S14 – S17**). The CR rate was calculated as the total number of CRs divided by the elapsed time between any two nodes, The elapsed time is rounded to get integers, and those less than 1 Mya are calculated as 1 Mya (**Table S9**).

WGDI was further utilized to generate homologous gene dot plots between the eight species and the reconstructed ancestral *Biscutella* genome (using the gene order of *B. baetica* for karyotype projection). Species from the two phylogenetic clades, i.e., *n* = 9 species-*B. auriculata* clade (clade I) and *B. baetica*-*B. didyma*-*B. lyrata* clade (clade II), were sequentially compared to the ancestral genome. The number of occurrences of each breakpoint shared by at least two species was counted, and further categorized as shared among all species, shared within clade I or II, and shared within the *n* = 9 clade or the *B. baetica* and *B. didyma* clade (**Figure S19**). To mitigate the influence of local gene rearrangements or potential assembly errors, the region between the adjacent border genes of two rejoined GBs, along with 100-kb of flanking sequences on each side, was analyzed. The HOT regions in each genome were extracted and subsequently analyzed for their LTR-RT content, A/B compartments, and aligned with TAD boundaries.

Gene family classification and functional enrichment

We classified orthologous gene families across eight *Biscutella* species into core, softcore, and private categories. Genes with two copies in both subgenomes of *Biscutella* species were considered as duplicates, and the remaining genes were classified as singletons. Thus, five types of gene sets, i.e., core duplicates shared in all eight *Biscutella* species (CD), core singletons shared in all eight *Biscutella* species (CS), softcore duplicates and singletons shared in late-diverging clade comprised by all *n* = 9 species or in early-diverging species comprised by *B. auriculata*, *B. didyma*, *B. baetica*, and *B. lyrata* (SD and SS), and private singletons specific to a certain species (PS), were further subjected to GO functional enrichment analysis. The GO terms and pathways of gene enrichment were identified by the clusterProfiler package (v3.14.3; Yu et al., 2012). GO enrichments were estimated using one-sided Fisher's exact tests, and an adjusted P-value < 0.05 was set as the cutoff criterion for the significance of the gene enrichment.

The Pfam protein domains identified by InterproScan were used to examine the conserved domains of 1,049 CD gene families. We assumed that gene duplicates could either split functions (subfunctionalization) or generate a new function (neofunctionalization; Birchler & Yang, 2022). For each CD gene family, if both copies retained identical protein domain(s), we considered that the domain(s) inherited from the ancestral gene, and if one copy lost the conserved domain or diverged into a new domain distinct from the other, we considered that the gene undergone sub/neofunctionalization. Based

on these criteria, we identified 220 species-specific (protein domains in a certain species differ from others) gene families that underwent potential sub/neofunctionalization.

Calculation of the loss rate for syntelogs

We investigated 12 syntenic gene families (AP1 – 3, PI, AG, STK, SHP1 – 2, SEP 1 – 4) belonging to the ABCDE model (Theißen, 2001; Theißen et al., 2016) using the *A. thaliana* genome as a reference, and the diploid genomes *Bo. stricta*, *Capsella rubella*, *Car. hirsuta*, *M. pygmaea*, and *S. parvula*, the meso-tetraploid genomes *P. cornutum*, *Biscutella*, and the meso-hexaploid genome *Br. rapa* as queries. The expected number of copies retained per gene family was predetermined based on the ploidy level of each species: one copy in diploid species, two copies in meso-tetraploid species, and three copies in meso-hexaploid species. The loss rate of syntelogs (r) was then calculated using the formula:

$$r = 1 - (Actual\ copy\ number\ /\ Expected\ copy\ number)$$

TE distribution in neighboring regions of genes

The TE density around genes was calculated following the method described by Zhang et al. (2023b). Specifically, the 5-kb upstream and downstream regions of each gene were defined as flanking sequences (totaling 10-kb). Flanking sequences overlapping with the coding regions of adjacent genes were hard-masked as “N”. A 50-bp sliding window with a 10-bp step size was used to scan these flanking regions and gene bodies. The TE density for each gene was then determined as the proportion of TE-derived bases across all sliding windows within gene bodies and its flanking sequences.

3D genome analysis

Hi-C reads were mapped against the corresponding reference genomes with BWA-MEM (v0.7.17; Li & Durbin, 2009). Hi-C contact matrices were generated using HiCExplorer (v3.7.2; Ramírez et al., 2018) at different resolutions (10, 25, 50, 100 kb). The hicPCA program embedded in HiCExplorer was used to delineate A/B compartments at a 50-kb resolution. TAD-like structures were identified using the hicFindTADs program embedded in HiCExplorer using the following parameters: “--thresholdComparisons 0.01 --delta 0.01 --correctForMultipleTesting fdr”. Conserved TADs between species were identified following the method proposed by Shen et al. (Shen et al., 2023), i.e., two TADs were considered conserved if syntelogs were within 100 kb on both sides of the TAD boundaries in both species.

Declarations

Acknowledgements

We thank dr. Pavel Trávníček for genome size estimation by flow cytometry. We thank dr. Rimjhim Roy Choudhury, dr. Nicolas Blavet, and dr. Hussein Anani for insightful discussions during this work. Plant Sciences core facility of CEITEC Masaryk University is acknowledged for the technical support. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. This work was supported by the Czech Science Foundation (grant no. 21-07748L to M.A.L.) and Swiss National Science Foundation (nos. 31003A_178938 and 310030L_197839 to C.P.). Additional funding was provided by the project TowArdsNextGENeration Crops (no. 17 CZ.02.01.01/00/22_008/0004581) of the ERDF Programme Johannes Amos Comenius and through Royal Society awards UF160222, RF/ERE/221032, URF/R/221024, RGF/R1/180006, RGF/EA/201030, and RF/ERE/210069 to A.B.

Data availability

The genome assembly and annotation data for *Biscutella* have been deposited in Figshare (https://doi.org/10.6084/m9.figshare.28451828.v1). All other data needed to evaluate the conclusions in the manuscript are present in the paper and/or the Supplementary Information.

Author contributions

M.A.L. and C.P. conceived and designed the project. Y.H. and M. Poretti assembled and annotated the genomes. Y.H. and X.G. performed evolutionary analyses. T.M. and M. Pouch conducted experimental work. A.B. and E.P.R. contributed to the repeatome analyses. S.G. contributed to the 3D genome analyses. M.B.C. contributed *Biscutella* materials. Interpretation of data and results was led by M.A.L. and C.P. All authors wrote and revised the manuscript.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
2. Bairoch A, Boeckmann B (1992) The SWISS-PROT protein sequence data bank. *Nucleic Acids Research*, 20, 2019
3. Beringer M, Choudhury RR, Mandáková T, Grünig S, Poretti M, Leitch IJ, Parisod C (2024) Biased retention of environment-responsive genes following genome fractionation. Mol Biol Evol 41:msae155
4. Birchler JA, Yang H (2022) The multiple fates of gene duplications: deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. Plant Cell 34:2466–2474
5. Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K (2007) Estimating divergence times in large phylogenetic trees. Syst Biol 56:741–752
6. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973

7. 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
8. Cheng F, Wu J, Fang L, Wang X (2012) Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. Front Plant Sci 3:198
9. Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat Methods 18:170–175
10. De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110, 2898–2903
11. Di Marzo M, Herrera-Ubaldo H, Caporali E, Novák O, Strnad M, Balanzà V, Colombo L (2020) SEEDSTICK controls *Arabidopsis* fruit size by regulating cytokinin levels and FRUITFULL. Cell Rep 30:2846–2857
12. Doležel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. Nat Protoc 2:2233–2244
13. Dornelas MC, Dornelas O (2005) From leaf to flower: revisiting Goethe's concepts on the metamorphosis of plants. Braz J Plant Physiol 17:335–344
14. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Aiden EL (2017) *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356:92–95
15. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 3:95–98
16. Eaton DA (2020) Toytree: A minimalist tree visualization and manipulation library for Python. Methods Ecol Evol 11:187–191
17. Edger PP, Pires JC (2009) Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res 17:699–717
18. Ehlers K, Bhide AS, Tekleyohans DG, Wittkop B, Snowdon RJ, Becker A (2016) The MADS box genes ABS, SHP1, and SHP2 are essential for the coordination of cell divisions in ovule and seed coat development and for endosperm formation in *Arabidopsis thaliana*. PLoS ONE 11:e0165075
19. Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics 9:1–14
20. Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:1–14
21. Feng X, Chen Q, Wu W, Wang J, Li G, Xu S, He Z (2024) Genomic evidence for rediploidization and adaptive evolution following the whole-genome triplication. Nat Commun 15:1635
22. 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 mustard. Plant Cell 28:17–27

23. Gremme G, Brendel V, Sparks ME, Kurtz S (2005) Engineering a software tool for gene structure prediction in higher organisms. Inf Softw Technol 47:965–978
24. Grünig S, Patsiou T, Parisod C (2024) Ice age-driven range shifts of diploids and expanding autotetraploids of *Biscutella laevigata* within a conserved niche. New Phytol 244:1616–1628
25. Guo X, Mandáková T, Trachtová K, Özüdoğru B, Liu J, Lysak MA (2021) Linked by ancestral bonds: multiple whole-genome duplications and reticulate evolution in a Brassicaceae tribe. Mol Biol Evol 38:1695–1714
26. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, White O (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654–5666
27. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Regev A (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512
28. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9:1–22
29. He W, Yang J, Jing Y, Xu L, Yu K, Fang X (2023) NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. Bioinformatics 39:btad121
30. Hu J, Fan J, Sun Z, Liu S (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 36:2253–2255
31. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37:D211–D215
32. Jalas J, Suominen J, Lampinen R (1996) Atlas Florae Europaeae. 11. Cruciferae (*Ricotia* to *Raphanus*). Helsinki University Printing House, Helsinki
33. Jia KH, Wang ZX, Wang L, Li GY, Zhang W, Wang XL, Mao JF (2022) SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. New Phytol 235:801–809
34. Jiao WB, Schneeberger K (2020) Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun 11:989
35. Jin JJ, Yu WB, Yang JB, Song Y, DePamphilis CW, Yi TS, Li DZ (2020) GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. Genome Biol 21:1–31
36. Katoh K, Misawa K, Kuma KI, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066
37. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360

38. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120
39. Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol Biol Evol 17:1483–1498
40. Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549
41. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760
42. Li X, Wang J, Yu Y, Li G, Wang J, Li C, Gong L (2023) Genomic rearrangements and evolutionary changes in 3D chromatin topologies in the cotton tribe (Gossypieae). BMC Biol 21:56
43. Li Z, McKibben MT, Finch GS, Blischak PD, Sutherland BL, Barker MS (2021) Patterns and processes of diploidization in land plants. Annu Rev Plant Biol 72:387–410
44. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930
45. Liao Y, Wang J, Zhu Z, Liu Y, Chen J, Zhou Y, Chen C (2022) The 3D architecture of the pepper genome and its relationship to function and evolution. Nat Commun 13:3479
46. Lomsadze A, Burns PD, Borodovsky M (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res 42:e119–e119
47. Lysak MA, Mandáková T, Schranz ME (2016) Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. Curr Opin Plant Biol 30:108–115
48. Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences*, 107, 9270–9274
49. Mandáková T, Lysak MA (2008) Chromosomal phylogeny and karyotype evolution in x = 7 crucifer species (Brassicaceae). Plant Cell 20:2559–2570
50. Mandáková T, Lysak MA (2016) Chromosome preparation for cytogenetic analyses in *Arabidopsis*. Curr Protocols Plant Biology 1:43–51
51. Mandáková T, Li Z, Barker MS, Lysak MA (2017a) Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J 91:3–21
52. Mandáková T, Pouch M, Brock JR, Al-Shehbaz IA, Lysak MA (2019) Origin and evolution of diploid and allopolyploid *Camelina* genomes were accompanied by chromosome shattering. Plant Cell 31:2596–2612
53. Mandáková T, Pouch M, Harmanová K, Zhan SH, Mayrose I, Lysak MA (2017b) Multispeed genome diploidization and diversification after an ancient allopolyploidization. Mol Ecol 26:6445–6462
54. Manton I (1932) Introduction to the general cytology of the Cruciferae. Ann Botany 46:509–556

55. Manton I (1934) The problem of *Biscutella laevigata* L. Z für induktive Abstammungs-und Vererbungslehre 67:41–57
56. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770
57. Mayrose I, Lysak MA (2021) The evolution of chromosome numbers: mechanistic models and experimental approaches. Genome Biol Evol 13:evaa220
58. Murley MR (1951) Seeds of the Cruciferae of northeastern North America. Am Midl Nat, 1–81
59. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274
60. Ni L, Liu Y, Ma X, Liu T, Yang X, Wang Z, Tian Z (2023) Pan-3D genome analysis reveals structural and functional differentiation of soybean genomes. Genome Biol 24:12
61. Okhovat M, VanCampen J, Nevonen KA, Harshman L, Li W, Layman CE, Carbone L (2023) TAD evolutionary and functional characterization reveals diversity in mammalian TAD boundary properties and function. Nat Commun 14:8111
62. Olowokudejo JD (1986) The infrageneric classification of *Biscutella* (Cruciferae). Brittonia, 86–88
63. Olowokudejo JD, Heywood VH (1984) Cytotaxonomy and breeding system of the genus *Biscutella* (Cruciferae). Plant Syst Evol 145:291–309
64. Ou S, Jiang N (2018) LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol 176:1410–1422
65. Ou S, Jiang N (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA 10:48
66. Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, Hufford MB (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 20:1–18
67. Parisod C, Poretti M, Mandáková T, Choudhury R, Lysak M (2025) The role of centromeric transposable elements in shaping chromosome evolution. https://doi.org/10.21203/rs.3.rs-5461468/v1
68. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295
69. Pinyopich A, Ditta GS, Savidge B, Liljegren SJ, Baumann E, Wisman E, Yanofsky MF (2003) Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature 424:85–88
70. POWO (2025) Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet. Available online: http://www.plantsoftheworldonline.org/ (accessed on 16 January 2025)
71. Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490

72. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65
73. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Paterson AH (2019) Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biol 20:1–23
74. Raeside C, Gaffé J, Deatherage DE, Tenaillon O, Briska AM, Ptashkin RN, Schneider D (2014) Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. MBio 5:10–1128
75. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Manke T (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun 9:189
76. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL (2018) Juicebox. js provides a cloud-based visualization system for Hi-C data. Cell Syst 6:256–258
77. Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301–302
78. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45
79. 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
80. Schubert I, Lysak MA (2011) Interpretation of karyotype evolution should consider chromosome structural constraints. Trends Genet 27:207–216
81. Shen F, Xu S, Shen Q, Bi C, Lysak MA (2023) The allotetraploid horseradish genome provides insights into subgenome diversification and formation of critical traits. Nat Commun 14:4102
82. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212
83. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res 34:W435–W439
84. Sun H, Ding J, Piednoël M, Schneeberger K (2018) findGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. Bioinformatics 34:550–557
85. Sun P, Jiao B, Yang Y, Shan L, Li T, Li X, Liu J (2022) WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol Plant 15:1841–1851
86. Sun P, Lu Z, Wang Z, Wang S, Zhao K, Mei D, Liu J (2024) Subgenome-aware analyses reveal the genomic consequences of ancient allopolyploid hybridizations throughout the cotton family. *Proceedings of the National Academy of Sciences*, 121, e2313921121
87. Theißen G (2001) Development of floral organ identity: stories from the MADS house. Curr Opin Plant Biol 4:75–85
88. Theißen G, Melzer R, Rümpler F (2016) MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. Development 143:3259–

3271

89. Tong K, Datta S, Cheng V et al (2025) Genome duplication in a long-term multicellularity evolution experiment. Nature. https://doi.org/10.1038/s41586-025-08689-6
90. Trávníček P, Ponert J, Urfus T, Jersáková J, Vrána J, Hřibová E, Suda J (2015) Challenges of flow-cytometric estimation of nuclear genome size in orchids, a plant group with both whole-genome and progressively partial endoreplication. Cytometry Part A 87:958–966
91. Vaughan JG, Whitehouse JM (1971) Seed structure and the taxonomy of the Cruciferae. Bot J Linn Soc 64:383–409
92. Vicente A, Alonso MÁ, Crespo MB (2020) Born in the Mediterranean: Comprehensive taxonomic revision of *Biscutella* ser. *Biscutella* (Brassicaceae) based on morphological and phylogenetic data. Ann Mo Bot Gard 105:195–231
93. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteom Bioinf 8:77–80
94. Wang J, Song B, Yang M, Hu F, Qi H, Zhang H, Wang X (2024) Deciphering recursive polyploidization in Lamiales and reconstructing their chromosome evolutionary trajectories. Plant Physiol, kiae151
95. Wang M, Wang P, Lin M, Ye Z, Li G, Tu L, Zhang X (2018) Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. Nat Plants 4:90–97
96. Wang Z, Li Y, Sun P, Zhu M, Wang D, Lu Z, Yang Y (2022) A high-quality *Buxus austro-yunnanensis* (Buxales) genome provides new insights into karyotype evolution in early eudicots. BMC Biol 20:216
97. Wendel JF, Jackson SA, Meyers BC, Wing RA (2016) Evolution of plant genome architecture. Genome Biol 17:1–14
98. Wlodzimierz P, Hong M, Henderson IR (2023) TRASH: tandem repeat annotation and structural hierarchy. Bioinformatics 39:btad308
99. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875
100. Yang W, Zhang L, Mandáková T, Huang L, Li T, Jiang J, Hu Q (2021) The chromosome-level genome sequence and karyotypic evolution of *Megadenia pygmaea* (Brassicaceae). Mol Ecol Resour 21:871–879
101. Yu G, Wang LG, Han Y, He QY (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16:284–287
102. Zhang C, Scornavacca C, Molloy EK, Mirarab S (2020) ASTRAL-Pro: quartet-based species-tree inference despite paralogy. Mol Biol Evol 37:3292–3307
103. Zhang K, Yang Y, Zhang X, Zhang L, Fu Y, Guo Z, Cheng F (2023a) The genome of *Orychophragmus violaceus* provides genomic insights into the evolution of Brassicaceae polyploidization and its distinct traits. Plant Commun, 4
104. Zhang K, Zhang L, Cui Y, Yang Y, Wu J, Liang J, Cheng F (2023b) The lack of negative association between TE load and subgenome dominance in synthesized *Brassica* allotetraploids. *Proceedings*

of the National Academy of Sciences, 120, e2305208120

105. Zhang RG, Li GY, Wang XL, Dainat J, Wang ZX, Ou S, Ma Y (2022) TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic Res 9:uhac017
106. Zhang X, Chen S, Shi L, Gong D, Zhang S, Zhao Q, You M (2021) Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. Nat Genet 53:1250–1259
107. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L (2012) ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun 419:779–781
108. Zhao M, Zhang B, Lisch D, Ma J (2017) Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. Plant Cell 29:2974–2994
109. Zhou C, McCarthy SA, Durbin R (2023) YaHS: yet another Hi-C scaffolding tool. Bioinformatics 39:btac808

Figures

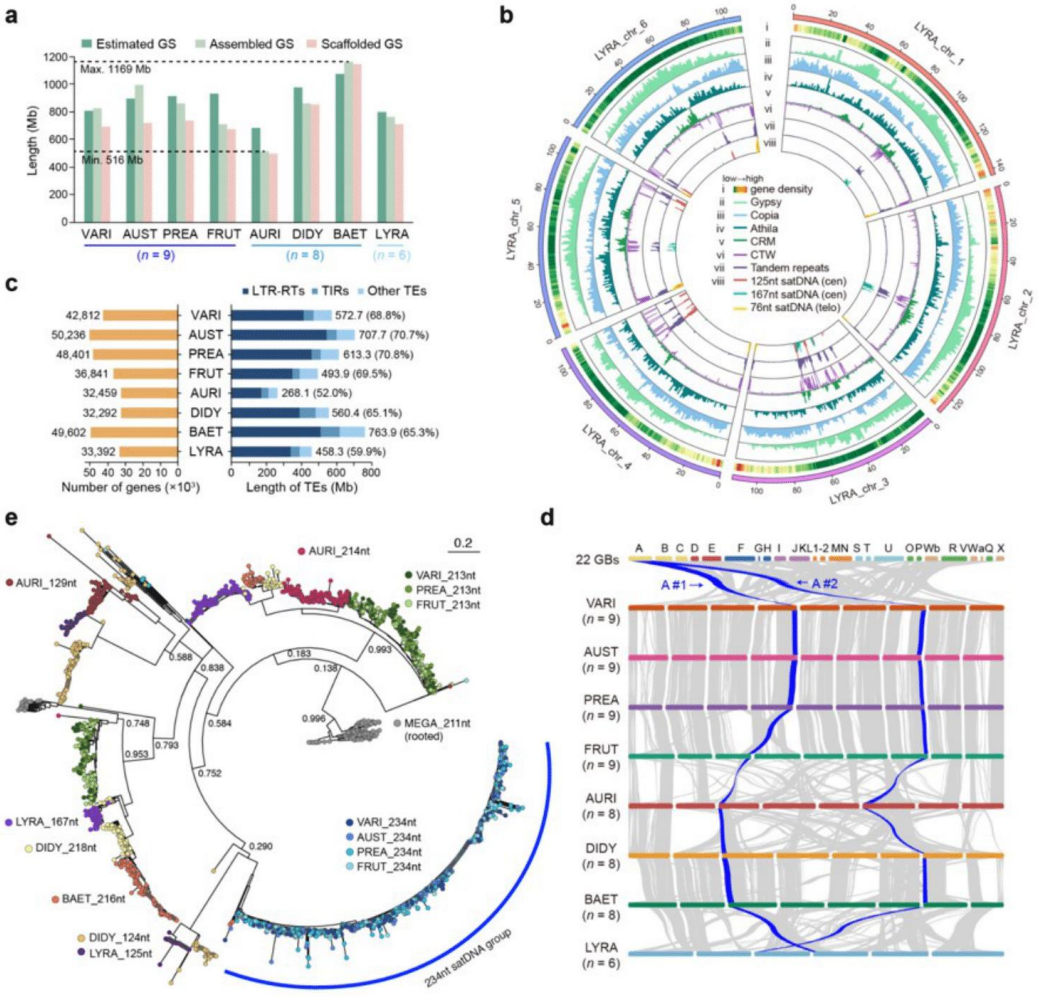


Figure 1

Genome size variation, repeatome, and inter-species chromosome collinearity of eight *Biscutella* species.

(**a**) Genome size (GS) variation. The estimated (flow cytometry-based; green), assembled (light green), and scaffolded (pink) genome size is shown for the eight *Biscutella* species. The assembled genome size ranged from c.516 to 1,169 Mb.

(**b**) Chromosomal organization in *B. lyrata*. The six chromosomes of *B. lyrata* are depicted as circular tracks (from outer to inner) representing: (i) gene density, (ii) Gypsy density, (iii) Copia density, (iv) Athila density, (v) CRM density, (vi) DNA compression (CTW), (vii) tandem repeat density, and (viii) densities of

125-bp (pink) and 167-bp (green) centromeric (cen) and 76-bp (yellow) subtelomeric (telo) satellite DNAs. Densities were calculated using the sliding window method, counting each element within a non-overlapping 1-Mb window across the genome. Chromosomal organizations for the other seven *Biscutella* species are shown in **Figures S3** and **S4**.

(**c**) Comparison of gene number and TE length. The number of annotated protein-coding genes (left), total TE length (right, Mb), and proportion of TEs relative to the entire genome (%) are shown.

(**d**) Syntenic relationships between the eight *Biscutella* genomes. A riparian plot illustrates macrosynteny across 22 ancestral genomic blocks (GBs A – X) and the eight assemblies, with the dark blue lines highlighting the two genomic copies of block A. Species acronyms are as follows: *B. laevigata* subsp. *varia* (VARI), *B. laevigata* subsp. *austriaca* (AUST), *B. prealpina* (PREA), *B. frutescens* (FRUT), *B. auriculata* (AURI), *B. didyma* (DIDY), *B. baetica* (BAET), and *B. lyrata* (LYRA).

(**e**) Phylogenetic analysis of centromeric satellite DNAs. A maximum-likelihood tree was constructed using 2,358 centromeric satDNA monomers from eight *Biscutella* species and 168 monomers of 211-bp satDNA from *Megadenia pygmaea* (MEGA) using FastTree (Price et al., 2010). Local support values are labelled on nodes of the main branches. Each dot represents an individual monomer. Different satDNA types are color-coded at the tips of the tree to indicate their classification.

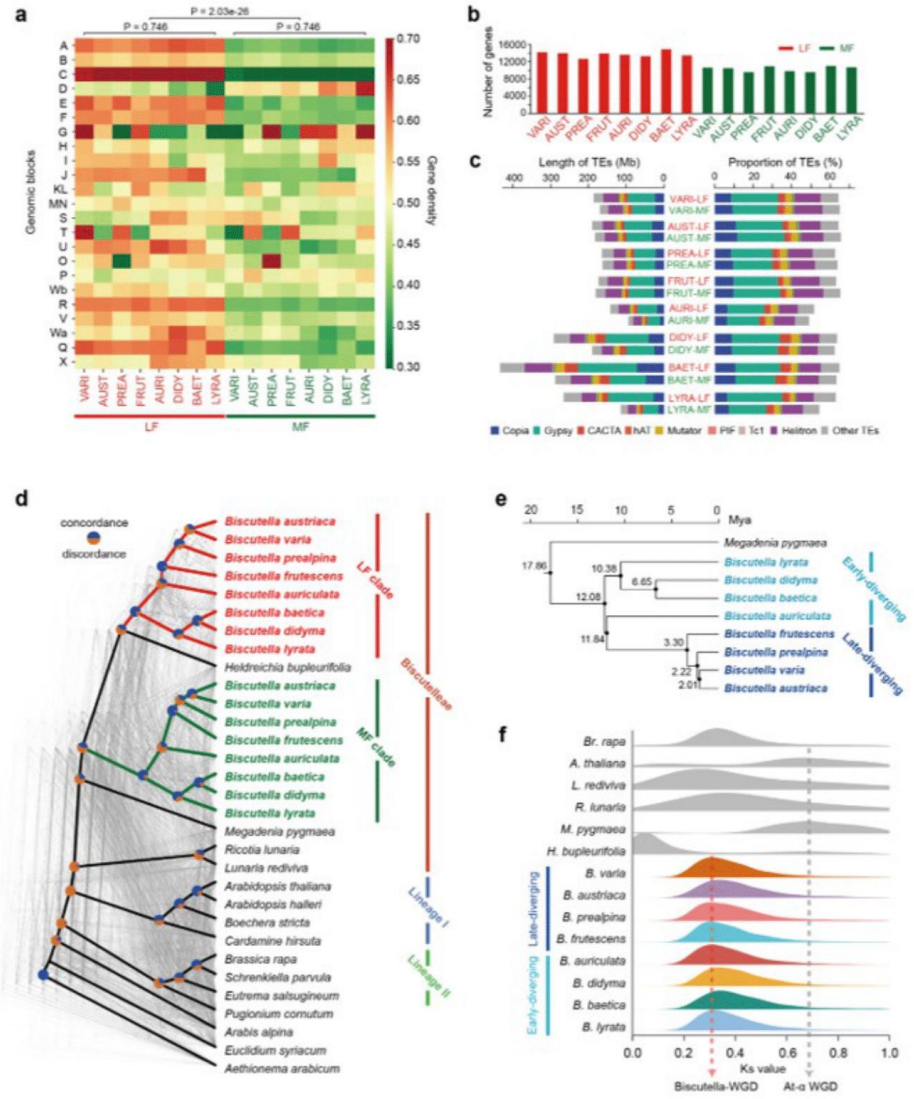


Figure 2

Structural and phylogenomic characteristics of the two *Biscutella* subgenomes.

(**a**) Syntenic gene density in two subgenomes. The heatmap shows the density of syntenic genes in the LF and MF subgenomes in eight *Biscutella* species compared to the 22 ancestral genomic blocks. The color gradient represents the average percentage of retained homeologous genes in the *Biscutella*

subgenomes surrounding each ancestral gene. Here, 300 genes flanking each side of a given gene locus were analyzed, giving a total window size of 601 genes. A significant difference in gene density was found between the LF and MF subgenomes (Wilcoxon rank-sum test), whereas interspecific differences were not significant (Kruskal-Wallis test).

(**b**) The total number of genes assigned to the LF and MF subgenomes is shown.

(**c**) TE content in two subgenomes. The length (left) and proportion (right) of eight major TE types in LF and MF subgenomes are displayed. The size of the subgenomes (Mb) was based on the total length of syntenic gene fragments in each subgenome. The proportion of TEs (%) was calculated as the ratio of the total TE length in a subgenome to its size.

(**d**) Subgenome-aware phylogenetic analysis. A cloud tree shows concordance and discordance among 1,234 nuclear gene trees. The coalescent-based tree, which includes the two *Biscutella* subgenomes and 15 other crucifer genomes representing the major Brassicaceae lineages, is shown with thick black lines. Pie charts indicate the proportions of gene trees concordant or discordant with the species tree topology. The *Aethionema arabicum* genome was used as an outgroup.

(**e**) A simplified time-calibrated tree constructed using r8s. The inferred divergence times for each clade are labelled on the phylogenetic nodes. The *Biscutella* species can be divided into two groups based on their divergence times: late-diverging species (*B. varia*, *B. austriaca*, *B. prealpina*, and *B. frutescens*) and early-diverging species (*B. auriculata*, *B. didyma*, *B. baetica*, and *B. lyrata*). Three additional time-calibrated trees that include more species are shown in **Figure S11**.

(**f**) *Ks*-value distribution. The distribution of *Ks*-value was analyzed for homeologous gene pairs in eight *Biscutella* genomes and six other Brassicaceae species, including *Br. rapa*, *A. thaliana*, *L. rediviva*, *R. lunaria*, *M. pygmaea*, and *H. bupleurifolia*. The *Ks*-value peak for *Biscutella* is ~0.33 (red dashed line) and, using a mutation rate of 1.5×10^{-8} substitutions per site per year (Koch et al., 2000), the age of the shared WGD is ~11 Mya.

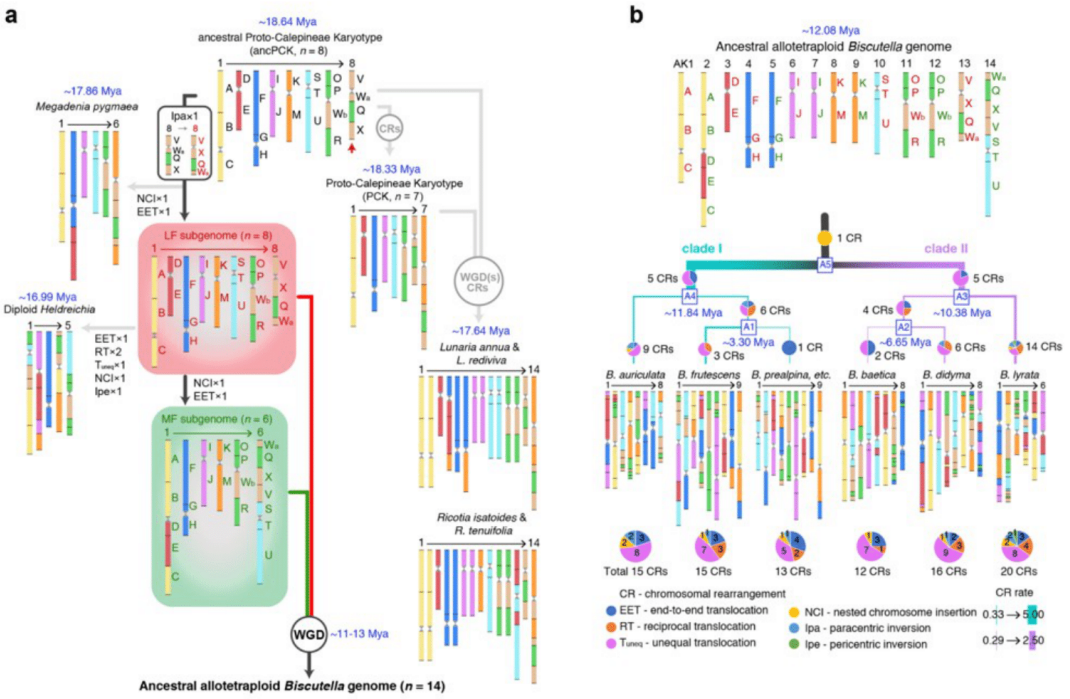


Figure 3

Karyotype evolution of *Bicutella* genomes.

(**a**) Reconstructed genome evolution in the tribe Biscutelleae. The earliest ancestor of Biscutelleae structurally resembled the ancPCK-like genome (n = 8). The ancPCK-like genome diversified into an 8-chromosome genome with a paracentric inversion on chromosome AK8/6 (V+Wa+Q+X → V+X+Q+Wa) and by descending dysploidy into PCK genome (n = 8 → n = 7). One or more likely two hybridization events between the ancPCK-like genome (n = 8) and PCK-like genome (n = 7) resulted in the ancestral tetraploid genome(s) of *Lunaria* and *Ricotia*, which were subsequently diploidized to 14 chromosomes (n = 8+7 → n = 14; Guo et al., 2021). The inversion ancPCK-like genome, structurally resembling the LF subgenome of *Biscutella* (n = 8), underwent three independent descending dysploidies forming (i) the *Megadenia pygmaea* genome (n = 8 → n = 6), (ii) the *Heldreichia bupleurifolia* genome (n = 8 → n = 5), and (iii) the MF subgenome of the tetraploid *Biscutella* ancestor (n = 8 → n = 6). The KL block is abbreviated to “K” and the MN block to “M”. The time estimates (Mya) are adopted from the r8s-calibrated phylogeny (**Figure S11**) and the WGD date was inferred from time-calibrated trees and *Ks*-based analyses (**Figures 2f** and **S11**). EET: end-to-end translocation, RT: reciprocal translocation, T_{uneq}: unequal reciprocal translocation, NCI: nested chromosome insertion, Ipa: paracentric inversion, Ipe: pericentric inversion, CR: chromosomal rearrangement.

(**b**) Post-polyploid genome evolution in *Biscutella*. The ancestral allotetraploid genome (*n* = 14) underwent three CRs and diverged into two clades: clade I (cyan lines) includes *B. auriculata* (*n* = 8) and the *n* = 9 species, clade II (purple lines) contains *B. baetica*, *B. didyma* (both *n* = 8) and *B. lyrata* (*n* = 6). The conserved genomes of the *n* = 9 species (*B. austriaca*, *B. prealpina* and *B. varia*) are only represented by *B. prealpina*. The ancestral karyotype at each evolutionary node (A1 to A5) was inferred using WGDI based on a bottom-up strategy (the detailed evolutionary pathway is shown in **Figures S15 – S17**). The time estimates (Mya) are adopted from the r8s-calibrated phylogeny (**Figure S11**). The different types of CRs are shown as pie charts, with the total number of CRs per species indicated below. CR rate was calculated as the total number of CRs divided by the elapsed time and indicated by line thickness (thin: low rate, thick: high rate). The detailed CR rates are listed in **Table S9**.

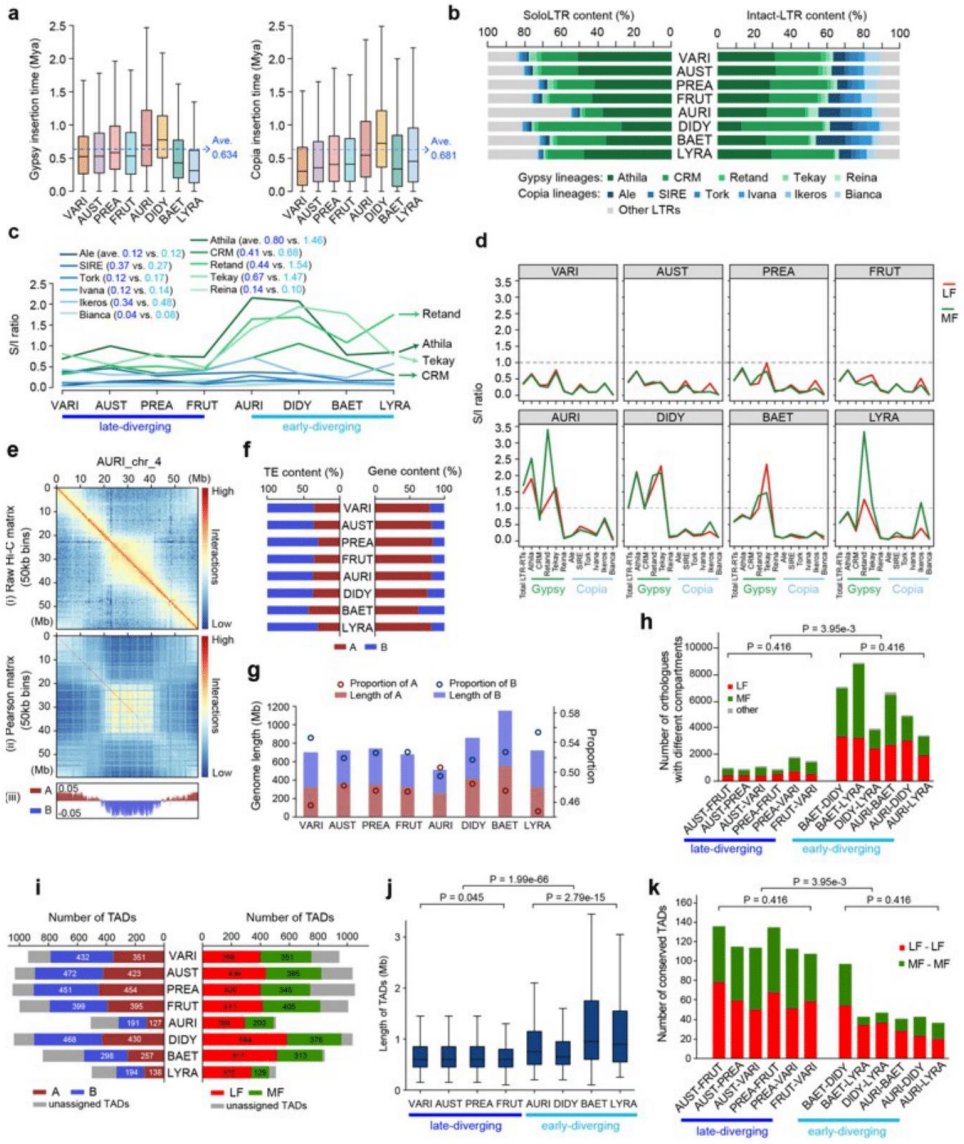


Figure 4

The evolution of LTR retrotransposons and 3D chromatin organization.

(**a**) Insertion time distribution of Gypsy (left) and Copia (right) retrotransposons inferred from the divergence of LTRs among intact copies.

(**b**) Content of solo and intact LTRs. Shown are the relative proportions of soloLTRs (left) and intact copies of LTR-RTs (right) in five major Gypsy lineages (shaded green) and six major Copia lineages (shaded blue).

(**c**) Solo:intact (S/I) LTR ratios. S/I ratios for 11 Gypsy or Copia lineages were compared in eight *Biscutella* species to assess the relative abundance of soloLTRs versus intact LTR-RTs. The average values of S/I LTR ratio for late-diverging (dark blue) and early-diverging (light blue) species are labelled.

(**d**) S/I LTR ratios in subgenomes. S/I ratios of total LTR-RTs (first column of Y-axis) and 11 Gypsy or Copia lineages (2nd to 12th column of Y-axis) were analyzed for the LF and MF subgenomes, respectively.

(**e**) Hi-C map of chromosome 4 in *B. auriculata*. The tracks from top to bottom represent: (i) the raw Hi-C interaction matrix at 50-kb resolution, (ii) the Pearson matrix generated from raw Hi-C interaction matrix, and (iii) the first principal component (PC1) derived from the PCA analysis of this matrix was used to define the A and B compartments. Positive PC1 values are shown in red, representing A compartments, and negative PC1 values are shown in blue and designated as B compartments. The Hi-C maps of other chromosomes of the analyzed *Biscutella* species are shown in **Figure S18**.

(**f**) TE and gene content in the nuclear A/B compartments. The proportions of TEs (left) and genes (right) were measured in the A/B compartments as a percentage of the total number of TEs and genes assigned to each compartment.

(**g**) Length and proportion of nuclear A/B compartments. The total length (bars; Y-axis left) and proportion (lines; Y-axis right) of A/B compartments were calculated for each genome. The proportion represents the percentage of the genome assigned to one of the two compartments.

(**h**) Orthologous gene pairs with different compartments. The number of orthologous gene pairs between any two species with different compartments was analyzed, with statistical comparisons among species (Kruskal-Wallis test) and between late- and early-diverging groups (Wilcoxon rank-sum test).

(**i**) The total number of TADs in A/B compartments (left) and LF/MF subgenomes (right). The unassigned TADs include TADs with mixed compartments or subgenome regions, such that they cannot be fully assigned to either the A/B compartment or one of the two subgenomes.

(**j**) Length of the TADs. The distribution of TAD lengths in the eight *Biscutella* genomes was analyzed and compared using statistical tests as in (**h**).

(**k**) Conserved TADs between species. The number of conserved TADs between any two species was quantified and analyzed using statistical tests as in (**h**). Two TADs are considered conserved between species if an orthologous gene pair in two species were located within 100-kb on either side (200-kb in total) of the TAD boundaries.

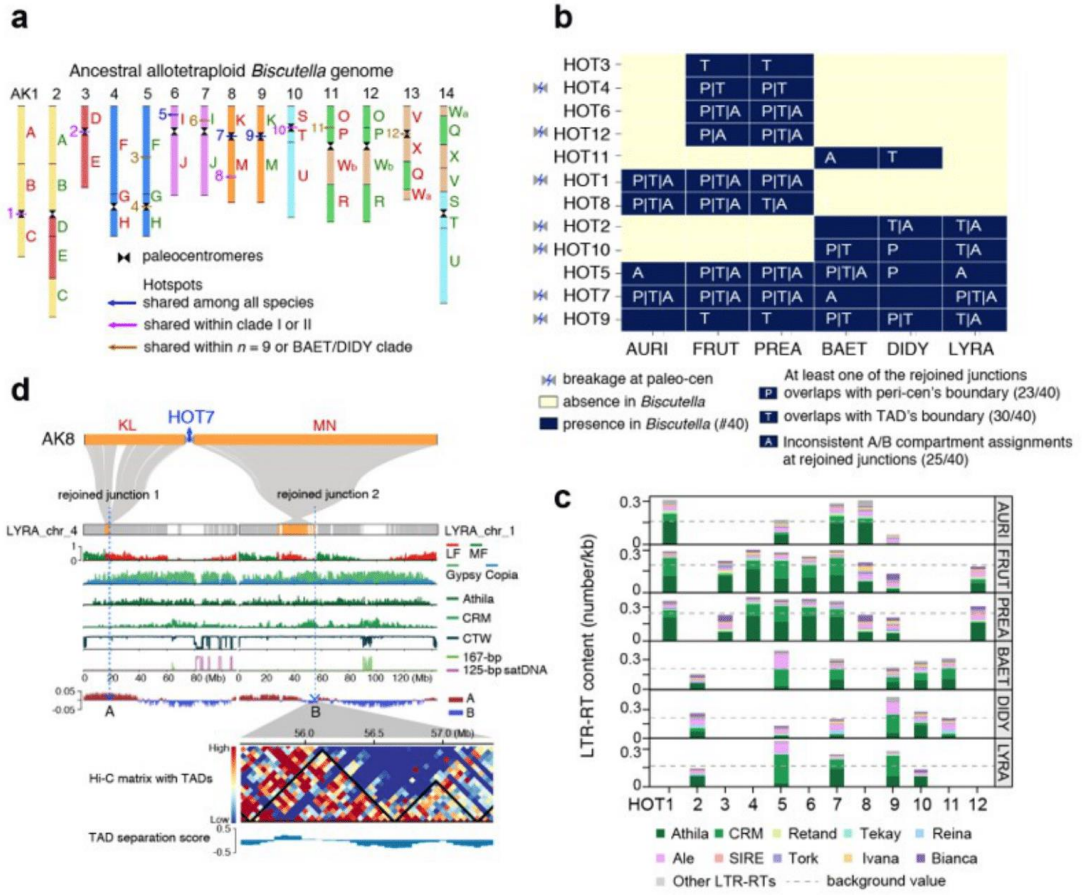


Figure 5

Position and characteristics of chromosome breakage hotspots.

(**a**) Twelve chromosome breakage hotspots (HOT1 to HOT12) were identified in the ancestral allotetraploid *Biscutella*genome. The HOT region is defined as the flanking 100-kb region (200-kb in total) on each side of the neighbouring genes of two joined GBs. Three hotspots are shared by all species (HOT5, 7, and 9), four hotspots are shared by either clade I species (HOT1 and 8; *B. auriculata* and the *n* = 9 species) or clade II species (HOT2 and 10; *B. baetica*, *B. didyma*, and *B. lyrata*), and five hotspots are shared by either *n* = 9 species (HOT3, 4, 6, and 12) or *B. baetica* and *B. didyma* (HOT11). Based on cytogenomically determined centromere positions in the closely related diploid *M. pygmaea* and autopolyploid *H. bupleurifolia* (Yang et al., 2021; Guo et al., 2021), we inferred the most likely centromere

positions in the ancestral allotetraploid *Biscutella* genome and defined them as paleocentromeres (opposite triangles).

(**b**) Presence and absence of the twelve HOTs in *Biscutella* genomes. The conserved genomes of the *n* = 9 species (*B. austriaca*, *B. prealpina*and *B. varia*) are only represented by *B. prealpina*. A total of 40 rejoined junction pairs were detected across all analyzed genomes, with the number determined by counting each detected breakage event across species. For example, if three breakages were shared by six genomes, then 18 junctions contribute to the total. The absence (32) and presence (40) of breaks are indicated by light yellow and dark blue squares; triangular symbols indicate chromosomal breaks at/near paleocentromeres (HOT1, 2, 4, 7, 9, 10, and 12). Of the 40 rejoined junctions, 23 overlapped with pericentromeric region boundaries (letter “P”), 30 overlapped with TAD boundaries (“T”), and 25 were associated with shifts in A/B compartment assignments (“A”).

(**c**) LTR-RT content within the 12 HOTs. The LTR-RT content was calculated as the total number of LTR-RTs divided by the size of a HOT region (kb). The genome-wide average LTR-RT content is marked as a grey dashed line.

(**d**) Origin and evolution of HOT7. HOT7, a breakage hotspot near the paleocentromere of ancestral chromosome AK8, underwent CRs that relocated blocks KL and MN to chromosomes 1 and 4, respectively, in *B. lyrata*. The tracks from top to bottom show: the ancestral chromosome AK8; syntenic relationship between AK8 and chromosomes 1 and 4 in *B. lyrata*; distributions of genes on chromosomes 1 and 4; gene density in LF (red) and MF (green) subgenomes; density of Ty3/Gypsy (light green) and Ty1/Copia (light blue) retrotransposons; density of Athila retroelements; density of CRM retroelements; CTW values; density of 125-bp (pink) and 167-bp (green) centromeric tandem repeats.; A/B compartment assignment; a Hi-C matrix for the 55.0 – 57.5 Mb region, along with TADs and corresponding TAD separation scores. Density calculations for each element follow the methods shown in **Figure 1b**.

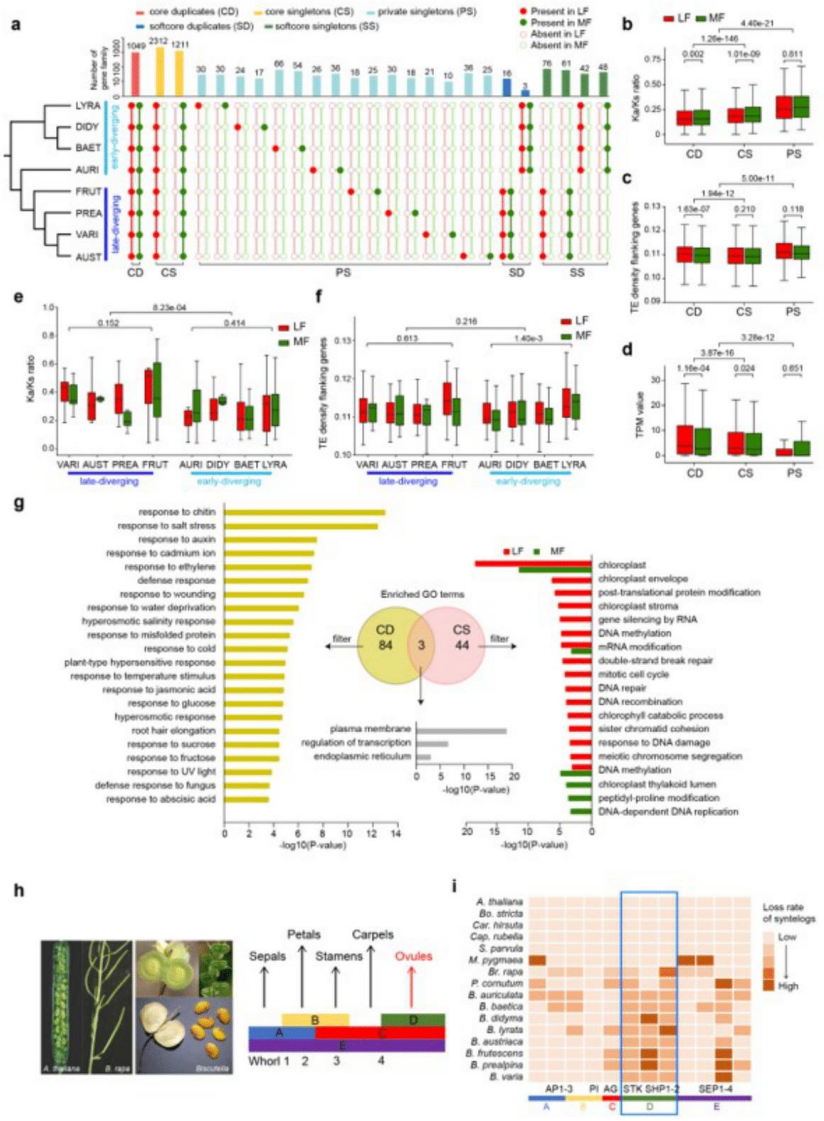


Figure 6

Analysis of WGD-drived gene families in *Biscutella* species.

(**a**) Presence and absence of gene families in eight *Biscutella* genomes. The left panel shows the species tree (see **Figure 2c**), whereas the upper panel shows the number of the five gene family categories: core duplicates and singletons shared by all eight species (CD and CS), softcore duplicates

and singletons shared by late- or early-diverging species (SD and SS), and private singletons specific to a particular species (PS). The presence or absence of these gene families in LF and MF subgenomes is represented by filled and open circles.

(**b**) *K*a/*K*s ratios of genes in CD, CS and PS family categories. *K*a/*K*s ratios of genes in each category were calculated using syntelogs in *M. pygmaea* as reference (Wilcoxon rank-sum test).

(**c**) TE density at the gene bodies and their 5-kb flanking regions in the CD, CS and PS family categories is shown (Wilcoxon rank-sum test; see **Methods** for a detailed calculation strategy).

(**d**) Gene expression levels. The TPM values of genes in the CD, CS and PS family categories for *B. baetica*and *B. didyma* (Wilcoxon rank-sum test; see **Methods** for a detailed calculation strategy).

(**e**) *K*a/*K*s ratio of genes in the PS category. The *K*a/*K*s ratios were calculated using syntelogs in *M. pygmaea* as reference. Significant differences in *K*a/*K*s ratios were found between late- and early-diverging species (Wilcoxon rank-sum test), while no significant differences were found between the four late-diverging species (Kruskal-Wallis test).

(**f**) TE density at the gene bodies and their 5-kb flanking regions in the PS category. Test methods as in (**e**).

(**g**) GO functional enrichments for CD (left) and CS (right) family categories. The x-axis shows the -log10 (P-value) of GO enrichment. A Venn diagram shows the shared and specific GO terms enriched for both CD and CS family categories; only three GO terms are shared between the two categories, with grey bars showing the GO enrichment in the CD category.

(**h**) Fruit and seed morphology and the ABCDE model. The left panel shows the differences in fruit and seed morphology between *Biscutella* species, *A. thaliana* and *Br. rapa*. The refined ABCDE model was adapted from Dornelas M.C. & Dornelas O. (2005).

(**i**) Loss rate of syntelogs for ABCDE model gene families. The loss rates of syntelogs for 12 gene families of the ABCDE model, including class A genes APETALA1 and 2 (AP1 and 2), class B genes APETALA3 (AP3) and PISTILLATA (PI), class C gene AGAMOUS (AG), class D genes SEEDSTICK (STK), SHATTERPROOF1 and 2 (SHP1 – 2), and class E genes SEPALLATA1, 2, 3, and 4 (SEP1 – 4), are shown for eight *Biscutella* species and seven other Brassicaceae species (*A. thaliana*, *Boechera stricta*, *Br. rapa*, *Capsella rubella*, *Cardamine hirsuta*, *M. pygmaea*, *Pugionium cornutum*, and *Schrenkiella parvula*). The loss rates of syntelogs are indicated by gradient colors (light = slight loss, dark = severe loss; see **Methods** for a detailed calculation strategy).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplTable113.xlsx
- BiscutellaSupplementaryFigures.pdf
- NCOMMS2528001rs.pdf

Curriculum vitae

Yile Huang

SUMMARY

• Seven years of experience in bioinformatics, genomics, and molecular biology
• Proficient in processing, analyzing, and mining next-generation sequencing (NGS) data
• Strong reading, writing, and presentation skills, with several publications in high-impact journals
• Passionate about exploring and learning cutting-edge advancements in biology and bioinformatics

EDUCATION

• Ph.D. in Genomics and Proteomics — Sept 2021 – Oct 2025 (expected)
Central European Institute of Technology (CEITEC), Masaryk University — Brno, Czechia
• Master of Science in Agriculture — Sept 2018 - July 2021
Chinese Academy of Agricultural Sciences — Beijing, China
& Henan Agricultural University (joint program) — Zhengzhou, China
• Bachelor of Science in Agriculture — Sept 2014 - July 2018
Henan Agricultural University — Zhengzhou, China

RESEARCH EXPERIENCE

• Research specialist - PhD student — Sept 2021 – present
CEITEC, Masaryk University (Supervisor: Prof. Martin A. Lysak) — Brno, Czechia
➢ Led three major projects focused on assembling and annotating several complex polyploid genomes using NGS data
➢ Explored the mechanisms of chromosome number reduction during diploidization after polyploidization in Brassicaceae species based on comparative genomics
➢ Collected plant genome assemblies to develop a chromosomal rearrangement database using Shiny (ongoing project)

➢ These projects led to the submission of two publications (one under review and one published in *The Plant Journal*)

• Academic visitor Sept 2024 - Oct 2024

University of Sussex (Collaborator: Dr. Alexandros Bousios) Brighton, United Kingdom

➢ Advanced an international collaborative project and be responsible for annotation and analysis of transposable elements in *Biscutella* (Brassicaceae) species

• Master student Sept 2018 - July 2021

Chinese Academy of Agricultural Sciences (Supervisor: Prof. Feng Cheng) Beijing, China

➢ Identified tandem duplicated genes (TDG) in several Solanaceae species and inferred the general patterns of TDG amplification based on comparative genomics (published in *Journal of Integrative Agriculture*)

➢ Resolved ancestral evolutionary history of paleo-hexaploidized Solanaceae genome based on published genomic data (published in *Plant Communications*)

➢ Processed resequencing data of over 1,000 tomato accessions, called SNP/Indel variants and performed GWAS analysis

• Student intern July 2017 - Dec 2017

Henan Academy of Agricultural Sciences Zhengzhou, China

➢ Acquired wet lab techniques, including isolated anther and microspore culture methods to obtain homozygous plants

PUBLICATIONS

• **Huang, Y.**, Poretti, M., Mandáková, T., Pouch, M., Guo, X., Perez-Roman, E., … & Lysak, M. A. (2025). Post-polyploid chromosomal diploidization in plants is affected by clade divergence and constrained by shared genomic features. *In revision*, *Nature Communications*. https://doi.org/10.21203/rs.3.rs-6440714/v1

• Zhang, L.†, Liu, Y.†, **Huang, Y.†**, Zhang, Y., Fu, Y., Xiao, Y., ... & Cheng, F. (2025). Solanaceae pan-genomes reveal extensive fractionation and functional innovation of duplicated genes. *Plant Communications*. 6(3).

• **Huang, Y.†**, Guo, X.†, Zhang, K., Mandáková, T., Cheng, F., & Lysak, M. A. (2023). The meso-octoploid *Heliophila variabilis* genome sheds a new light on the impact of polyploidization and diploidization on the diversity of the Cape flora. *The Plant Journal*, 116(2), 446-466.

• **Huang, Y.**, Zhang, L., Zhang, K., Chen, S., Hu, J., & Cheng, F. (2022). The impact of tandem duplication on gene evolution in Solanaceae species. *Journal of Integrative Agriculture*, 21(4), 1004-1014.

• Yang, Y.†, Zhang, K.†, Xiao, Y., Zhang, L., **Huang, Y.**, Li, X., ... & Cheng, F. (2022). Genome assembly and population resequencing reveal the geographical divergence of Shanmei (*Rubus corchorifolius*). *Genomics, Proteomics & Bioinformatics*, 20(6), 1106-1118.

• Cui, Y.†, Zhuang, M.†, Wu, J., Liu, J., Zhang, Y., Zhang, L., **Huang, Y.**, ... & Cheng, F. (2020). Segmental translocation contributed to the origin of the *Brassica* S-locus. *Horticultural Plant Journal*, 6(3), 167-178.

† co-first authorship

CONFERENCE PRESENTATIONS

Oral presentation

• Sip of Science – CEITEC Brno, Czechia, April 16 2025

The long road to stability: Chromosome evolution in *Biscutella* over 12 million years

• XX International Botanical Congress 2024 Madrid, Spain, July 21-27 2024

The pathways of post-polyploid diploidization and descending dysploidy in *Biscutella* (Brassicaceae)

• The Czech Plant Nucleus Workshop Brno, Czechia, Jun 20-21 2023

The meso-octoploid *Heliophila variabilis* genome sheds a new light on the impact of polyploidization and diploidization on the diversity of the Cape flora

• ELIXIR CZ Annual Conference 2022 Trest, Czechia, Sept 19-21 2022

Structure and evolution of the meso-octoploid genome of *Heliophila variabilis* (Brassicaceae)

Poster presentation

• Plant Chromosome Biology: Cytogenetics meeting 2023 Brno, Czechia, Sept 11-13 2023

The pathways of post-polyploid diploidization and descending dysploidy in *Biscutella* (Brassicaceae)

• Polyploidy 2023 Palm Coast, Florida, United States, May 9-12 2023

The pathways of post-polyploid diploidization and descending dysploidy in *Biscutella* (Brassicaceae)

• Congress of the European Society for Evolutionary Biology Prague, Czechia, Aug 14-19 2022

Structure and evolution of the meso-octoploid genome of *Heliophila variabilis* (Brassicaceae)

OTHER SCIENTIFIC ACTIVITIES AND AWARDS

• Awarded for the MUNI Scientist 2023

• Awarded for the JCMM (South Moravian Centre for International Mobility) Scholarship 2021

• Awarded for the National First Prize Scholarship by the Chinese Ministry of Education 2018-2021

• Volunteer at Chinese Genomics Meet-up online (CGM), responsible for inviting speakers, 2023

• Participated as a guest in the National Congress of Plant Biology 2019 (Chengdu, China, October 11-14, 2019)

• Participated as a guest in the Mathematics, Computers and Life Sciences Workshop (Beijing, China, May 18-19, 2019)

SKILLS

• Sequencing data processing: familiar with PacBio HiFi, Nanopore, Illumina, Hi-C, Omni-C, RNA-seq, Iso-seq

• *De novo* genome assembly: genome survey, genome assembly, Hi-C assisted assembly, gene/TE annotation

• Comparative genomics: synteny analysis, phylogenetic analysis, gene family analysis

• Population genetics: SNPs/InDels detection and annotation, genome-wide association studies

• Programming: Python (Advanced), Bash (Advanced), R (Basic)

• Personal skills: scientific writing, presentation, problem solving, collaboration, continuous learning

• Languages: Mandarin (Native), English (Professional working), German (Learning)