MASARYK UNIVERSITY

FACULTY OF SCIENCE AND CEITEC

Doctoral Thesis

Sheng Zuo

Brno 2022

MASARYK
UNIVERSITY

FACULTY OF SCIENCE AND CEITEC

Plastome, repeatome and kinetochore protein evolution in land plants

Doctoral Thesis

Sheng Zuo

Supervisors: Prof. Mgr. Martin A. Lysak, Ph.D., DSc.
Dr. rer. nat. Inna Lermontova

Brno 2022

Bibliografický záznam

Autor:	Sheng Zuo Přírodovědecká fakulta, Masarykova univerzita CEITEC, Masarykova univerzita
Název práce:	Evoluce plastomů, repetitivních sekvencí a kinetochorových proteinů u vyšších rostlin
Studijní program:	Vědy o živé přírodě
Specializace:	Bio-omika
Vedoucí práce:	Prof. Mgr. Martin A. Lysak, Ph.D., DSc. Dr. rer. nat. Inna Lermontova
Akademický rok:	2022/2023
Počet stran:	171
Klíčová slova:	Chloroplastová DNA; Evoluce; Kinetochor; KNL2; Fylogenomika; Rostlinný genom; Polyploidie; Repetitivní DNA

Bibliographic Entry

Author:	Sheng Zuo Faculty of Science, Masaryk University CEITEC, Masaryk University
Title of the Thesis:	Plastome, repeatome and kinetochore protein evolution in land plants
Degree Programme:	Life Sciences
Field of Study:	Bio-omics
Supervisor:	Prof. Mgr. Martin A. Lysak, Ph.D., DSc. Dr. rer. nat. Inna Lermontova
Academic Year:	2022/2023
Number of Pages:	171
Keywords:	Chloroplast DNA; Evolution; Kinetochore; KNL2; Phylogenomics; Plant genome; Polyploidy; Repeatome

Abstrakt

Předmetem této práce je studium struktury a evoluce rostlinného genomu pomocí celogenomových sekvenačních dat a fylogenomických přístupů. Autor se během svého doktorandského studia podílel na pěti odborných publikacích, které tvoří jádro dizertační práce.

První část práce je zaměřena na malou rostlinnou čeleď mokřadkovité (Limnanthaceae) z řádu brukvotvaré (Brassicales), která má pouze dva rody a osm druhů. Pomocí sekvenačních dat s nízkým pokrytím jsme zrekonstruovali fylogenetické vztahy a charakterizovali repetitivní sekvence těchto genomů. Trojrozměrná fluorescenční *in situ* hybridizace prokázala, že pět chromozomových párů v interfázních jádrech druhů rodu *Limnanthes* zaujímá Rablovu polarizovanou konfiguraci. Genomy čeledi Limnanthaceae byly zkoumány jako potenciální modelové systémy.

Druhá část práce se zabývá diploidizací genomu v tribu Microlepidieae (Brassicaceae), který obsahuje přibližně 17 rodů a 60 druhů vyskytující se v Austrálii a na Novém Zélandu. Tribus Microlepidieae vykazuje rozdílné tempo diploidizace genomu a rozsáhlou morfologickou konvergenci. Analýzou fylogenetických vztahů a morfologických znaků v tomto tribu jsme poskytli fylogenomický důkaz, že rozdílné tempo post-polyploidní diploidizace je spojeno s intratribovou kladogenezí, morfologickou disparitou a změnou životních forem. Také jsme ukázali, že rychlejší diploidizace genomu je pozitivně korelovaná s evolucí chloroplastových genů. Na základě nových fylogenomických poznatků, byla revidována taxonomie rodů *Arabidella*, *Cuphonotus* a *Lemphoria*.

Třetí část je věnována evoluční historii proteinu KINETOCHORE NULL2 (KNL2) a jeho funkci při depozici CENH3 (centromerická varianta histonu H3). Ukázali jsme, že gen KNL2 prošel třemi nezávislými dávnými duplikacemi, a to u kapradin, trav a dvouděložných rostlin. Neklasifikované geny KNL2 mohou být rozděleny do dvou kladů: αKNL2 a βKNL2 u dvouděložných rostlin a γKNL2 a δKNL2 u trav. Potvrzená centromerická lokalizace βKNL2 a mutační analýza naznačují, že se protein účastní depozice nového CENH3 do centromery, podobně jako αKNL2. Navíc jsme zjistili, že mutant KNL2 by mohl být využit k indukci haploidních rostlin. Nově identifikovaný βKNL2 se tak může stát nástrojem pro získání haploidů u huseníčku i zemědělských plodin.

Abstract

The subject of this thesis is the study of the plant genome structure and evolution using whole genome sequencing data and phylogenomic approaches. The author's doctoral studies resulted in five publications that form the thesis framework.

The first part addresses the knowledge gap in the meadowfoam family (Limnanthaceae), one of the small families in the order Brassicales, which harbors only two genera and eight species. Using low coverage sequencing data, we reconstructed phylogenetic relationships and characterized the repeatomes of Limnanthaceae genomes. A three-dimensional fluorescence *in situ* hybridization analysis demonstrated that the five chromosome pairs in interphase nuclei of *Limnanthes* species adopt the Rabl-like configuration, a special interphase chromosome arrangement. We examined the Limnanthaceae genomes as a potential model system for 3D genome organization.

The second part focuses on the genome diploidization in the crucifer tribe Microlepidieae (Brassicaceae), which contains c. 17 genera and 60 species endemic to Australia and New Zealand. The tribe Microlepidieae exhibits differently paced genome diploidization and extensive morphological convergence. By analyzing phylogenetic relationships and morphological characters in this tribe, we provided clear phylogenomic evidence that differently paced post-polyploid diploidization was associated with intra-tribal cladogenesis, morphological disparity, and life-form transitions. We also showed that faster genome diploidization is positively correlated with the evolution of chloroplast genes. The taxonomic limits of *Arabidella*, *Cuphonotus*, and *Lemphoria* were revisited based on phylogenomic findings.

The third part focuses on the evolutionary history of KINETOCHORE NULL2 (KNL2) protein and its function in CENH3 (centromeric histone H3 variant) loading. We showed that the *KNL2* gene underwent three independent ancient duplications in ferns, grasses, and eudicots. The unclassified *KNL2* genes could be divided into two clades: *αKNL2* and *βKNL2* in eudicots, and *γKNL2* and *δKNL2* in grasses. The confirmed centromeric localization of βKNL2 and mutant analysis suggested that the protein participates in the loading of new CENH3 into the centromere, similarly to αKNL2. Moreover, we reported that a *KNL2* mutant could be used as a haploid inducer. Thus, the newly identified *βKNL2* may become the subject of manipulations to obtain haploids in *Arabidopsis thaliana* and crop species.

Acknowledgement

First of all, I would like to thank my supervisor Prof. Mgr. Martin A. Lysak sincerely for his support and guidance throughout my studies. I also would like to thank my co-supervisor Dr. rer. nat. Inna Lermontova for scientific and mental support.

I would like to thank Dr. Terezie Mandáková for her valuable contribution. Her hardworking and motivation inspired me to complete the project. I would like to thank Dr. Xinyi Guo for his fruitful discussion and bioinformatic support. I would also like to thank Michaela Kubová, Ramakrishna Yadala, Ulkar Ahmadli for their valuable contribution. I am also grateful for the support from Mert Dogan, Milan Pouch, Dr. Petra Hloušková, and Wenbo Shan.

Additionally, I would like to thank all members of the Lysak Lab at CEITEC for their assistance and the excellent working atmosphere. I would like to thank all the people I met in the CEITEC during my Ph.D. study.

Finally, I would like to thank my parents. Without their encouragement and support, I could not pursue my dreams.

Original publications and definition of the author’s contribution

The thesis is based on five manuscripts written by the author of the thesis (Sheng Zuo, S.Z.).

Article 1 (Annex I to this doctoral thesis)

Zuo, S., Mandáková, T., Kubová, M., Lysak, M. A. (2022), Genomes, repeatomes and interphase chromosome organization in the meadowfoam family (Limnanthaceae, Brassicales). Plant Journal, 110: 1462-1475. (IF = 6.417)

Research article. **Z.S.** participated in the design of the study, independently carried out the bioinformatic analyses described in the manuscript, prepared the draft manuscript, and finalized the text after receiving comments from the co-authors.

Article 2 (Annex II to this doctoral thesis)

Zuo, S., Guo, X., Mandáková, T., Edginton, M., Al-Shehbaz, I. A., Lysak, M. A. (2022). Genome diploidization associates with cladogenesis, trait disparity, and plastid gene evolution. Plant Physiology, 190: 403-420. (IF = 8.34)

Research article. **Z.S.** participated in the design of the study, independently carried out the phylogenomic analyses described in the manuscript, prepared the draft manuscript, and finalized the text after receiving comments from the co-authors.

Article 3 (Annex III to this doctoral thesis)

Lysak, M. A., Edginton, M., **Zuo, S.**, Guo, X., Mandáková, T., Al-Shehbaz, I. A. (2022). Transfer of two *Arabidella* and two *Cuphonotus* species to the genus *Lemphoria* (Brassicaceae) and a description of the new species *L. queenslandica*. Phytotaxa, 549: 235-240. (IF = 1.17)

Research article. **Z.S.** performed the phylogenetic analysis.

Article 4 (Annex IV to this doctoral thesis)

Zuo, S., Yadala, R., Yang, F., Talbert, P., Fuchs, J., Schubert, V., Ahmadli, U., Rutten, T., Pecinka, A., Lysak, M. A., Lermontova, I. (2022). Recurrent plant-specific duplications of KNL2 and its conserved function as a kinetochore assembly factor. Molecular Biology and Evolution, 39: msac123. (IF = 16.24)

Research article. **Z.S.** participated in the design of the study, independently carried out the bioinformatic analyses described in the manuscript, prepared the draft manuscript, and finalized the text after receiving comments from the co-authors.

Article 5 (Annex V to this doctoral thesis)

Ahmadli, U., Kalidass, M., Khaitova, L. C., Fuchs, J., Cuacos, M., Demidov, D., **Zuo, S.**, Pecinkova, J., Mascher, M., Ingouff, M., Heckmann, S., Houben, A., Riha, K., Lermontova, I. (2022). High temperature increases centromere-mediated genome elimination frequency in *Arabidopsis* deficient in cenH3 or its assembly factor KNL2. BioRxiv, 10.1101/2022.03.24.485459

Research article. **Z.S.** performed the RNA-seq data analysis and participated in the writing of the respective parts of the manuscript.

I hereby declare on my honor that the work presented in this thesis is original and independent and that I have used only the literature listed in the bibliography.

29.09.2022 左 胜

TABLE OF CONTENTS

1 INTRODUCTION 11

2 LITERATURE REVIEW 13

- 2.1 Plant genome structure and organization 13
  - 2.1.1 Repetitive sequences 14
  - 2.1.2 Centromere 20
  - 2.1.3 Telomere 24
  - 2.1.4 Nucleolar organizer region and rDNA 27
  - 2.1.5 Organellar genomes 28
- 2.2 Polyploidy and post-polyploid diploidization in plants 30
  - 2.2.1 Polyploidization 30
  - 2.2.2 Genome diploidization 32
- 2.3 Sequencing technologies and applications 34
  - 2.3.1 First-generation DNA sequencing 35
  - 2.3.2 Second-generation DNA sequencing 36
  - 2.3.3 Third-generation DNA sequencing 37
  - 2.3.4 Genome assembly 40
  - 2.3.5 Phylogenomic applications 44
  - 2.3.6 Identification of repetitive sequences 47

3 AIMS OF THE THESIS 50

4 RESULTS AND DISCUSSION 51

- 4.1 Genome analysis and nuclear organization in Limnanthaceae 51
- 4.2 Genome diploidization in the tribe Microlepidieae (Brassicaceae) 54
- 4.3 Evolutionary history and function of the *KNL2* in plants 56

5 CONCLUSIONS 58

6 REFERENCES 60

7 LIST OF PUBLICATIONS (I-V) 83

8 LIST OF ABBREVIATIONS 167

9 CURRICULUM VITAE 169

1 INTRODUCTION

The structure and organization of genomes are the fundamental characteristics of every living organism. With technical advances in next-generation sequencing (NGS) and long read sequencing approaches, we have witnessed an enormous change in the understanding of the evolution and structure of plant genomes in recent years. These sequencing approaches have revealed the genomic diversity in exquisite detail and led to many insights into plant genome function and evolution.

This thesis summarizes the author's contributions to the fields of the plant genome structure and evolution using whole genome sequencing data and phylogenomic approaches. The thesis is divided into introduction, literature review, aims of the thesis, brief results of work conducted along with the published articles, and conclusions. The literature review focuses on plant genome structure and organization, polyploidy and post-polyploid diploidization, sequencing technologies and their applications.

The thesis presents three different but interrelated phylogenomic projects. The first project focused on genomes and repeatomes of the meadowfoam family (Limnanthaceae), one of the genomically underexplored families in the order Brassicales. The Limnanthaceae harbors only two genera, *Limnanthes* and *Floerkea*. The genus *Limnanthes* (meadowfoams) has seven species, while the genus *Floerkea* contains only one species (*F. proserpinacoides*, false mermaidweed), all native to North America. Limnanthaceae has a rather basal position within the Brassicales, being placed between the Setchellanthaceae (*Setchellanthus caeruleus*) and the large clade consisting of the core Brassicales and four small families (Edger *et al.*, 2018). Given the knowledge gap extending from the Caricaceae (the papaya genome, *Carica papaya*) to the Brassicaceae, the phylogenetic position within the Brassicales, low chromosome numbers, annual herbaceous life history and seed availability make the Limnanthaceae potentially attractive for gaining more insights into genome evolution of the Brassicales. Here, we have applied phylogenomic approaches to reconstruct phylogenetic relationships and characterize the repeatomes of Limnanthaceae genomes using low-coverage whole genome sequencing data. We have also used the *de novo* identified repeats to analyze interphase chromosome organization in this family for the first time.

In the second project, we focused on the post-polyploid evolution in the tribe Microlepidieae from the mustard family (Brassicaceae). Angiosperm genome evolution was marked by many clade-specific whole genome duplication (WGD) events. The Microlepidieae is one of several tribe-level clades in Brassicaceae formed after an ancestral allotetraploidization. The ancestral allotetraploidization was followed by speciation events and different levels of genome diploidization resulting in the extant diversity of about 17 genera and 60 species endemic to Australia and New Zealand. Here, we reconstructed phylogenetic relationships in this tribe using complete chloroplast sequences, entire 35S rDNA units, and abundant repetitive sequences. The four recovered intra-tribal clades mirror the varied diploidization of Microlepidieae genomes, suggesting that the intrinsic genomic features underlying the extent of diploidization are shared among genera and species within one clade. In addition, we showed that faster genome diploidization is positively correlated with mean morphological disparity and evolution of chloroplast genes (plastid–nuclear genome coevolution). Our results along with the close phylogenetic relatedness to *Arabidopsis thaliana* (hereafter Arabidopsis) make Microlepidieae an excellent model system to investigate the evolutionary consequences of post-polyploid genome evolution.

In the third project, we identified a new KINETOCHORE NULL2 gene (KNL2, also termed M18BP1) in Arabidopsis and reconstructed the evolutionary history of the *KNL2* gene in plants using phylogenomic approaches. Centromeres are specific chromosomal regions where kinetochore protein complexes assemble in mitosis and meiosis to attach chromosomes to the spindle microtubules. Centromere identity is specified epigenetically by the presence of the histone H3 variant termed CENH3 (also named CENP-A in mammals) which triggers the assembly of a functional kinetochore. KNL2 plays a crucial role in new CENH3 deposition after replication. In most metazoan genomes, only one *KNL2* gene was identified containing the characteristic SANTA (SANT-associated) domain. Here, we reconstructed the evolutionary history of the *KNL2* gene in the plant kingdom. Our results indicated that the *KNL2* gene in plants underwent three independent ancient duplications, namely in ferns, grasses and eudicots. Additionally, we demonstrated that previously unclassified *KNL2* genes could be divided into two clades *αKNL2* and *βKNL2* in eudicots and *γKNL2* and *δKNL2* in grasses, respectively. *KNL2s* of all clades encode the conserved SANTA domain, but only the *αKNL2* and *γKNL2* groups additionally encode the CENPC-k motif. The confirmed centromeric localization of βKNL2 and mutant analysis suggested that the protein participates in the loading of new CENH3, similarly to αKNL2. Taken together, our results provide new insights into the evolutionary diversification of the plant kinetochore assembly gene *KNL2* and suggest that the plant-specific duplicated *KNL2* genes are involved in centromere and kinetochore assembly.

2 LITERATURE REVIEW

2.1 Plant genome structure and organization

The plant nuclear genome is organized into discrete chromosomes, consisting of DNA, histone, and other associated proteins. Each non-replicated chromosome and metaphase chromatid consists of a linear and unbroken DNA molecule (Heslop-Harrison & Schwarzacher, 2011), and structural features of chromosomes, such as centromeres, telomeres, and nucleolar organizer region (NOR), are conserved (**Figure 1**). In contrast, genome size and chromosome numbers in plants have tremendous diversity, with approximately a 2400-fold range from 65 Mbp/1C to 150 Gbp/1C and a 300-fold range from $n = 2$ to $n = 600$ chromosomes, respectively (Bennett & Leitch, 2005; Zonneveld *et al.*, 2005; Bennett & Leitch, 2011; Fleischmann *et al.*, 2014). The genome size variation is primarily caused by the proliferation of repetitive DNA sequences and whole genome duplication (WGD) events. Along with the nuclear genome, the plant cell also contains mitochondrial, chloroplast or plastid genomes, and these organellar genomes may influence the organization and evolution of the nuclear genome.

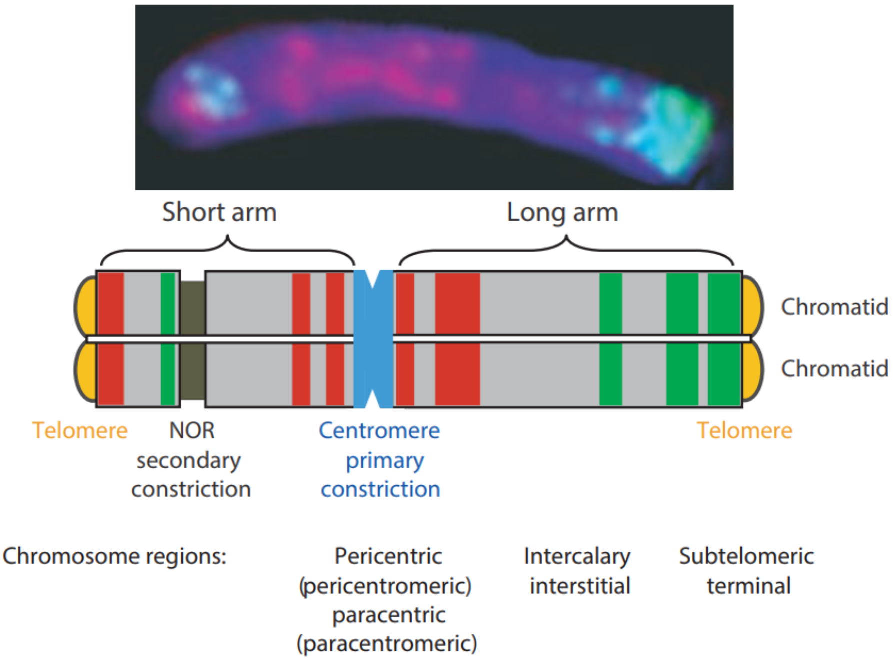


Figure 1. The organization and features of a plant chromosome. Top: A fluorescent light micrograph of a metaphase chromosome stained blue with the DNA-binding fluorochrome 4′,6-diamidino-2-phenylindole (DAPI). *In situ* hybridization shows the location of two tandemly repeated DNA sequences detected as red and green fluorescence. Bottom: Diagram of the structure of a metaphase chromosome with two chromatids. Adapted from Heslop-Harrison and Schwarzacher (2011).

2.1.1 Repetitive sequences

Genome sizes across flowering plants vary from 65 Mbp/1C in Lentibulariaceae (Fleischmann *et al.*, 2014) to 150 Gbp/1C in *Paris japonica* (Bennett & Leitch, 2011). Whereas the number of coding genes is relatively similar in plant genomes, the variation in the number of non-coding sequences and repetitive elements largely influenced the size and evolution of plant genomes. Complex plant genomes are heavily occupied by various types of repetitive sequences (**Figure 2**), including mobile elements (transposable elements, TEs) dispersed throughout the chromosome, and tandem repeats (satellite repeats) that comprise most of the heterochromatic chromosomal regions. Although TEs and tandem repeats are the two main groups of repetitive elements, low copy repeats (LCRs) and other types of repetitive sequences also exist in plant genomes (Bailey *et al.*, 2001).

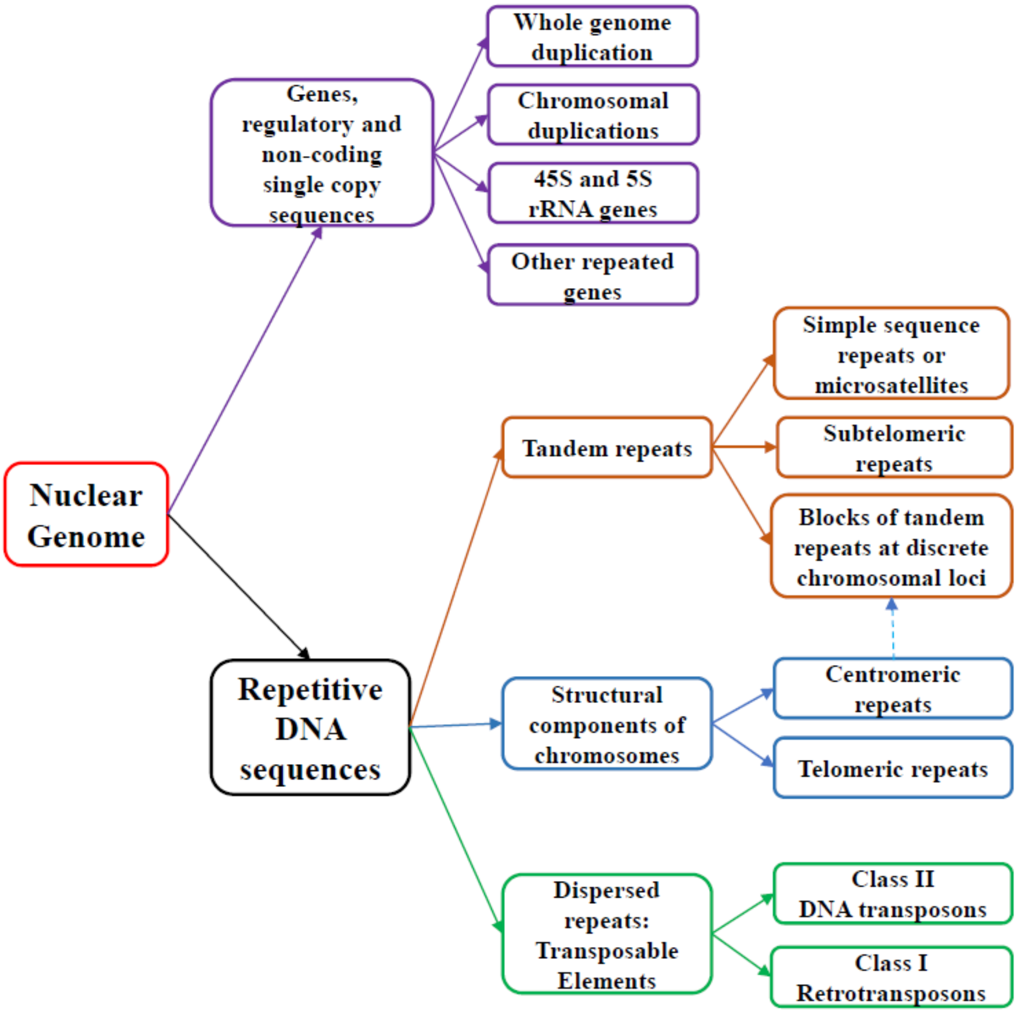


Figure 2. Nuclear genome composition and repetitive sequences classification. Adapted from Biscotti *et al.* (2015).

Transposable elements

Transposable elements (TEs) are fragments of DNA that can insert into new chromosomal locations. TEs were discovered in maize by Barbara McClintock more than 70 years ago as these elements are responsible for the sector of altered pigmentation on mutant kernels. Later, TEs had been identified in the genomes of *Drosophila melanogaster*, yeast (*Saccharomyces cerevisiae*), *Escherichia coli*, *Caenorhabditis elegans* and humans (Feschotte *et al.*, 2002). Although TEs are the largest component of repetitive sequences of most eukaryotes, active elements comprise only a tiny fraction of the TE complement of plant genomes and most other genomes. Epigenetic mechanisms, such as RNA interference or DNA methylation, are possible pathways to silence TE activity.

The plant genome may be viewed as an ecosystem occupied by diverse communities of TEs, and TEs are not randomly distributed in the genome. TEs may exhibit various levels of preference on insertion within certain compartments of the genome, following a balancing act of facilitating propagation while mitigating deleterious effects on host cell function (Sultana *et al.*, 2017). The distribution and accumulation of TEs may also be influenced by natural selection and genetic drift (Bourque *et al.*, 2018). Therefore, some TEs are more likely to be retained in certain genomic locations than others. Thus, the diversity of TEs in a genome is modified by properties intrinsic to the elements and evolutionary forces acting at the level of the host genomes.

The first TE classification system was proposed by David J. Finnegan based on the DNA or RNA intermediate replication mechanism (Finnegan, 1989). Based on the different transposition mechanisms, TEs can be divided into Class I TEs (retrotransposons) and Class II TEs (DNA transposons). Retrotransposons insert into a new genomic location via RNA intermediate, called the "copy-paste" replication mechanism (**Figure 3**), which can result in an increased copy number of a retrotransposon (Wicker *et al.*, 2007). Retrotransposons are generally the major contributor to the repetitive sequence content in plant genomes because of their replication mechanism. Two high-copy Class I TE superfamilies, *Copia* and *Gypsy*, are generally prevalent in plant genomes (Macas *et al.*, 2015; Wicker *et al.*, 2018). DNA transposons move to a new genomic location via DNA intermediate, termed the "cut-paste" mechanism (**Figure 3**). Terminal inverted repeats (TIRs) and a transposase enzyme are two unique features of most DNA transposons (Wicker *et al.*, 2007).

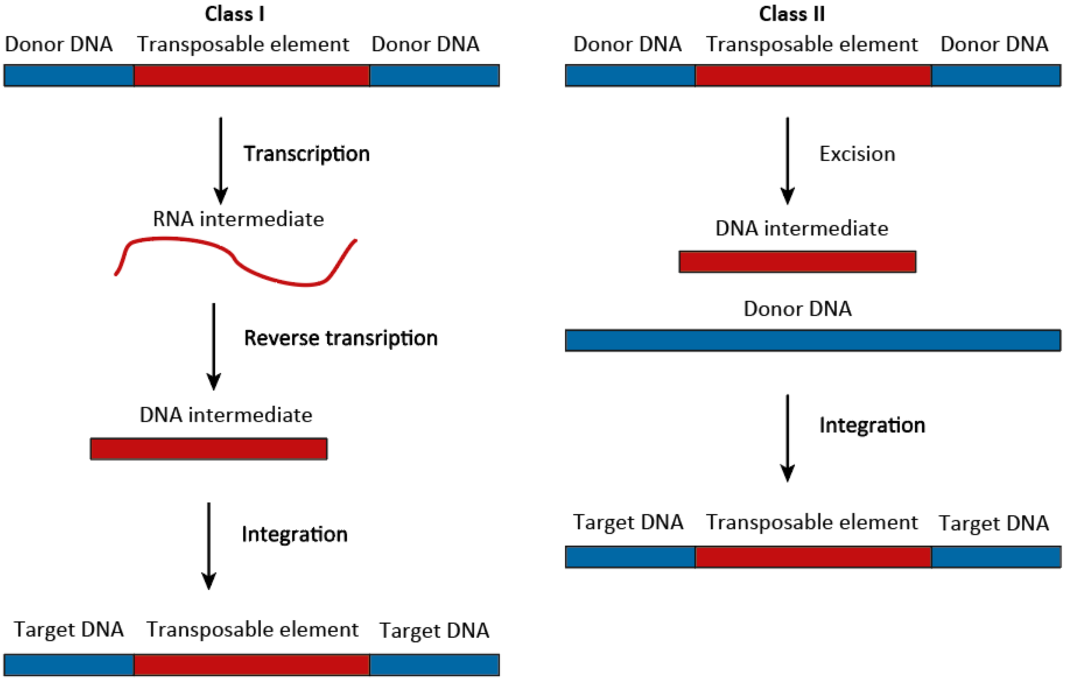
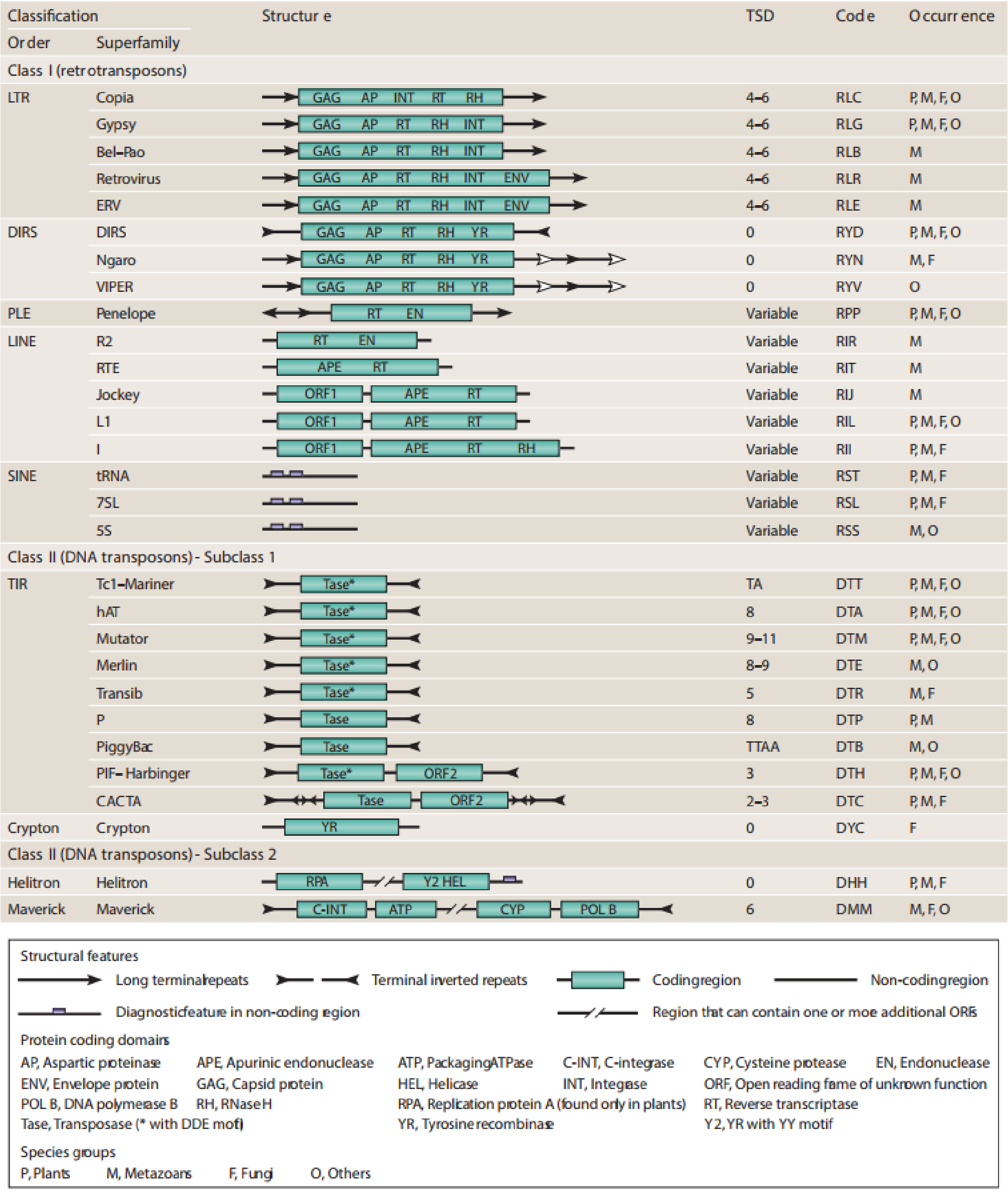


Figure 3. Class I and II transposons and mechanisms of their amplification integration. Adapted from Agren (2014).

Later, Wicker *et al.* (2007) proposed the first unified hierarchical classification system for TE by considering the transposition mechanisms, sequence similarities, and structural relationships. Therefore, TE can be further divided into subclasses, orders and superfamilies (**Figure 4**). Following this classification system, retrotransposons can be divided into five orders, including long terminal repeat (LTR) retrotransposons, long interspersed nuclear elements (LINEs), *DIRS*-like elements, *Penelope*-like elements (PLEs), and short interspersed nuclear elements (SINEs). The DNA transposons can be divided into two subclasses: subclass I is spread through classic conservative transposition (cut-and-paste), and subclass II elements spread through a rolling-circle replicative transposition mechanism using a rolling circle replication protein and a helicase (Kapitonov & Jurka, 2001).

LTR retrotransposons usually comprise the largest portion of the TEs in plant genomes (Macas *et al.*, 2015; Zhang *et al.*, 2017; Neumann *et al.*, 2019). Moreover, two superfamilies, Ty1/*copia* and Ty3/*gypsy*, occupied the major proportion of LTR retrotransposons (Neumann *et al.*, 2019). Although the LTR retrotransposons are diverse, their structure is highly conserved. The presence of long terminal repeats (LTRs) at both ends is a common feature of LTR retrotransposons. Most LTR retrotransposons have a primer binding site (PBS) downstream of the 5' LTR and a polypurine tract (PPT) upstream of the 3' LTR. The central part contains two open reading frames (ORFs) for the *gag* gene and polyprotein *pol*. The *gag* gene encodes a structural protein essential for the assembly of viral-like particles, while the *pol* gene encodes four proteins, including a protease (PR), a ribonuclease H (RH), a reverse transcriptase (RT), and an integrase (INT). *Copia* and *gypsy* LTR retrotransposons differ in the arrangement of the protein domains encoded within the *pol* gene. Recently, based on phylogenetic analyses of the three most conserved polyprotein domains (RT, RH, and INT), a comprehensive LTR retrotransposons classification system was proposed (Neumann *et al.*, 2019), dividing Ty1/*copia* into 16 lineages and Ty3/*gypsy* into two major lineages (chromovirus and non-chromovirus). In addition, a comprehensive database of retrotransposon protein domains (REXdb) was established for repetitive sequence analysis provided a unified annotation of LTR retrotransposons in plant genomes (Neumann *et al.*, 2019; Novák *et al.*, 2020).



Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	Copia	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	DIRS	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	Penelope	RT EN	Variable	RPP	P, M, F, O
LINE	R2	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	tRNA		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	Tc1–Mariner	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	Crypton	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	Helitron	RPA Y2 HEL	0	DHH	P, M, F
Maverick	Maverick	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features

Long terminal repeats
Terminal inverted repeats
Coding region
Non-coding region
Diagnostic feature in non-coding region
Region that can contain one or more additional ORFs

Protein coding domains

AP, Aspartic proteinase
APE, Apurinic endonuclease
ATP, Packaging ATPase
C-INT, C-integrase
CYP, Cysteine protease
EN, Endonuclease
ENV, Envelope protein
GAG, Capsid protein
HEL, Helicase
INT, Integrase
ORF, Open reading frame of unknown function
POL B, DNA polymerase B
RH, RNase H
RPA, Replication protein A (found only in plants)
RT, Reverse transcriptase
Tase, Transposase (* with DDE motif)
YR, Tyrosine recombinase
Y2, YR with YY motif

Species groups

P, Plants
M, Metazoans
F, Fungi
O, Others

Figure 4. Classification system for transposable elements (TEs). Adapted from Wicker *et al.* (2007).

Role of TEs in plant genomes

Transposition is an important mechanism of genome expansion that is counteracted by the removal of DNA via deletion over time. The balance of the two mechanisms is a major driving force of plant genome evolution (Schubert & Vu, 2016; Bourque *et al.*, 2018). As the insertion and removal of TEs are usually imprecise, these processes may influence adjacent sequences. If these events occur at a high frequency, the host genome can accumulate a vast amount of duplication and reshuffling, including genes and regulatory sequences. For example, Pack-MULE transposable elements in rice contain fragments derived from more than 1 000 cellular genes (Jiang *et al.*, 2004). TEs also induce genomic structural variation even without mobile activity, as recombination events can occur between the highly homologous TEs sequences at distant positions within the genome and result in large-scale inversions, deletions and duplications (Bennetzen & Wang, 2014).

Although TEs have been considered junk DNA for a long time, there is growing evidence that TE insertion can provide the raw material for the emergence of protein-coding genes and non-coding RNAs (Naville *et al.*, 2016; Joly-Lopez & Bureau, 2018). For example, TEs can donate their genes to the host genome by adding exons to the existing host genes. In line with Barbara McClintock's predictions, TEs could be a rich source of material for the modulation of eukaryotic gene expression. Indeed, TEs can insert into promoters and enhancers, transcription factor binding sites, insulator sequences, and repressive elements (Chuong *et al.*, 2017). For example, the methylation level of a LINE retrotransposon, in the intron of the homeotic gene *DEFICIENS*, controls whether or not the plants bear oil-rich fruit (Ong-Abdullah *et al.*, 2015).

Tandem repeats

Tandem repeats (TRs) are DNA sequence motifs that contain adjacent repeating units. TRs are usually ubiquitous in plant genomes. A single TR can make up to 36% of a nuclear genome (Ambrozová *et al.*, 2011). Based on the monomer length, TRs are typically classified as microsatellites (simple sequence repeats, SSRs), minisatellites, and satellite DNA (satDNA). SSRs are widely detected in plant species (Gemayel *et al.*, 2010). For example, the dimer motifs are more frequent in green algae, bryophytes, and ferns, whereas the trimer motifs are more frequent in flowering plants (Victoria *et al.*, 2011). Satellite DNAs are long arrays of tandemly, head to tail, arranged highly conserved motifs (repeat units, monomers). The monomer length of satellite DNAs can be as short as simple sequence repeats (<10 bp) or reach over 5 kb (Gong *et al.*, 2012; Heckmann *et al.*, 2013), but they are usually hundreds of nucleotides long (Vondrak *et al.*, 2020). The satellite DNAs located preferentially in heterochromatic regions, mainly in (peri)centromeric or subtelomeric regions (Mehrotra & Goyal, 2014; Zhang *et al.*, 2017; Li *et al.*, 2018), while the micro- and minisatellites located both in euchromatin and heterochromatin regions (Garrido-Ramos, 2015; Garrido-Ramos, 2017).

TRs are extremely unstable and mutation rates of TRs are usually much higher than those in other parts of the genome (Gemayel *et al.*, 2010). Most mutations in TRs are repeat polymorphisms that occur when the number of the repeated unit changes, not by point mutations (Gemayel *et al.*, 2010). In other words, most of these changes consist of the addition or deletion of complete repeat units, while additions or deletions of part of one unit are very rare. In plants, the famous Bur-0 IIL1 defect in Arabidopsis that generates a detrimental phenotype is caused by the expansion of triplet TTC/GAA in the intron of *IIL1* gene (Sureshkumar *et al.*, 2009).

There are two models for explaining the mechanisms of TRs expansions or contractions: strand-slippage replication and recombination (**Figure 5**) (Paques *et al.*, 1998; Gemayel *et al.*, 2010). Briefly, strand-slippage replication occurs during the replication of the TRs when there is mispairing between the template and nascent DNA strands. If the template strand is looped out, then contraction of the TR occurs, whereas if the nascent strand loops out, then an expansion will result. Recombination events, including unequal crossing over and gene conversion, can also lead to contraction and expansion of TR sequences.

Although satellite DNAs are a general component of plant genomes, their sequence composition is highly variable even within one species (Gong *et al.*, 2012). Centromeric satellite DNAs, localized in centromeric regions, are rapidly evolving DNA sequences in plant species (Henikoff *et al.*, 2001; Gong *et al.*, 2012; Melters *et al.*, 2013). A comprehensive comparative analysis of several hundred species including plants and animals showed no sequence conservation in centromeric TRs (Melters *et al.*, 2013).

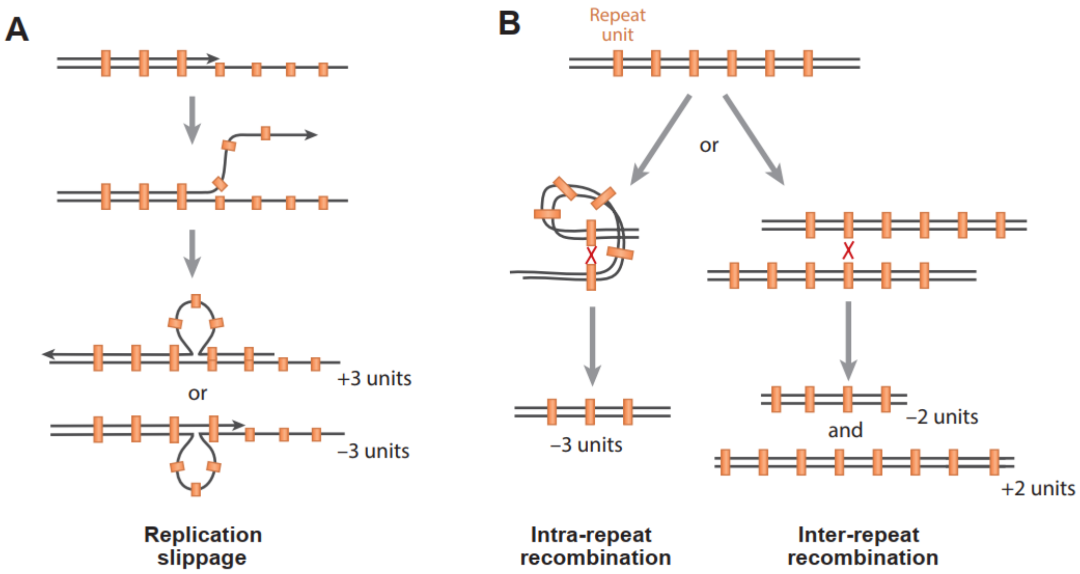


Figure 5. The simplified illustrations of two major mechanisms of tandem repeat (TR) expansions and contractions. (A) Replication slippage. (B) Recombination. Adapted from Gemayel *et al.* (2010).

2.1.2 Centromere

The centromere is the specialized chromosomal region that provides the site of assembly for the kinetochore, which interacts with spindle microtubules. Centromeres are responsible for accurate chromosome segregation and stabilization and ensure equal division of genetic material between daughter cells during mitosis and meiosis. Despite the fundamental role of centromeres two different types, monocentromere and holocentrmere, are observed across plants (**Figure 6**). Most plant species possess monocentric chromosomes, forming primary contraction on chromosomes in metaphase. However, the centromere sizes are remarkably diverse among eukaryotes. The budding yeast have point centromeres, and the length of their DNA sequences is 125 base pairs (bp), while in many plants, centromeres contain several megabase pairs (Mbp) of repeats (Henikoff *et al.*, 2001; Cheng *et al.*, 2002; Talbert *et al.*, 2002; Cleveland *et al.*, 2003; Nagaki *et al.*, 2003; Jin *et al.*, 2004; Wu *et al.*, 2004). In contrast, holocentric chromosomes lack primary constriction and are attributed to a kinetochore activity along almost the entire chromosome length during mitosis and meiosis (Steiner & Henikoff, 2014; Neumann *et al.*, 2015; Marques & Pedrosa-Harand, 2016; Schubert *et al.*, 2020).

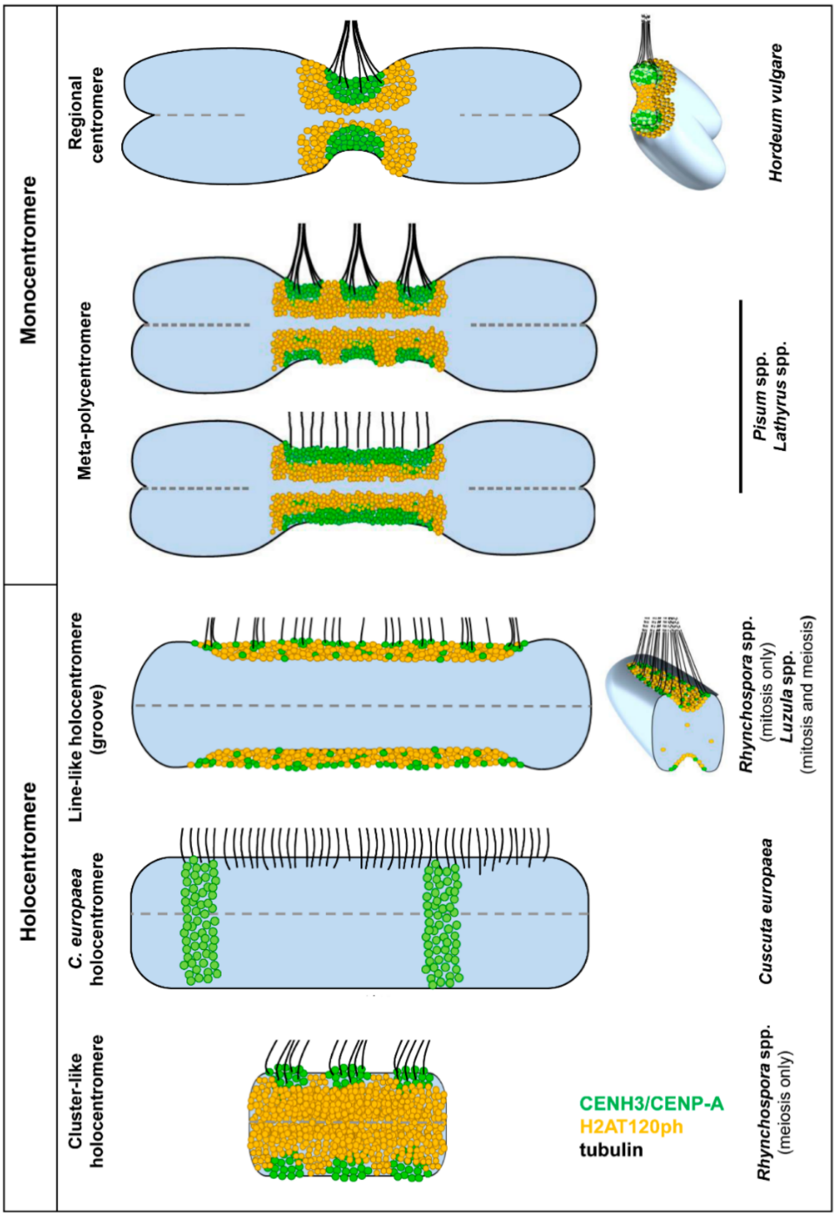

Monocentromere
Holocentromere
Regional centromere
Meta-polycentromere
Line-like holocentromere (groove)
C. europaea holocentromere
Cluster-like holocentromere
Hordeum vulgare
Pisum spp.
Lathyrus spp.
Rhynchospora spp. (mitosis only)
Luzula spp. (mitosis and meiosis)
Cuscuta europaea
Rhynchospora spp. (meiosis only)
CENH3/CENP-A
H2AT120ph
tubulin

Figure 6. Summary of the different mono- and holocentromere types in plant species. Adapted from Schubert *et al.* (2020).

A specialized histone H3 variant, CENH3, is a hallmark of active centromeres in plants (Talbert *et al.*, 2002; Zhong *et al.*, 2002; Naish *et al.*, 2021), which is called CID in *Drosophila* (Malik *et al.*, 2002), Cse4p in budding yeast (Meluh *et al.*, 1998), and CENP-A in humans (Palmer *et al.*, 1991). The function of the CENH3 is conserved among different species, but the CENH3 protein and centromeric DNA sequences differ even between closely related species (Henikoff *et al.*, 2001; Malik & Henikoff, 2002; Jiang *et al.*, 2003; Lamb *et al.*, 2004; Melters *et al.*, 2013; Lermontova *et al.*, 2014; Lermontova *et al.*, 2015). In general, the plant centromere contains ubiquitous and abundant repetitive DNA sequences, including satellite repeats, transposons, and retrotransposons (Ananiev *et al.*, 1998; Heslop-Harrison *et al.*, 1999; Cheng *et al.*, 2002; Jin *et al.*, 2004; Nagaki *et al.*, 2004; Wu *et al.*, 2004; Ma *et al.*, 2007). Centromeric satellite repeats usually comprise megabase-sized arrays of simple tandem repeats, and some of them exhibit higher-order repeat (HOR) structures. These satellite DNAs are often intermingled by LTR retrotransposons. As CENH3 presents exclusively in active or functional centromeres, in recent years, chromatin immunoprecipitation following high-throughput sequencing (ChIP-seq) technology has been widely used to characterize centromeric DNA sequences (Gong *et al.*, 2012; Bloom, 2014; Zhang *et al.*, 2014; Zhang *et al.*, 2017; Li *et al.*, 2018; Yang *et al.*, 2018; Robledillo *et al.*, 2020; Huang *et al.*, 2021), which significantly facilitated the understanding of the centromere and kinetochore function.

Kinetochores are multi-protein complexes that connect chromosomes to microtubules of the mitotic and meiotic spindles. Kinetochore complexes assemble on the centromeric chromatin, in which CENH3/CENP-A specifies the position of the kinetochore. The kinetochore can be divided into several sub-complexes (**Figure 7**), including the constitutive centromere-associated network (CCAN) and the KMN-network (KNL1, MIS12, and NDC80 complexes) (McKinley & Cheeseman, 2016; Pesenti *et al.*, 2016). The KMN-network is recruited to the kinetochore via the CCAN complex, and the NDC80 complex of the KMN-network directly interacts with microtubules (McKinley & Cheeseman, 2016; Hara & Fukagawa, 2018). The KMN network also has spindle assembly checkpoint (SAC) functions that complete the assembly of the entire kinetochore complex (Varma & Salmon, 2012). In plants, CENH3 nucleosomes directly bind to two inner kinetochore proteins, CENP-C and KNL2 (M18BP1) (Dawe *et al.*, 1999; Lermontova *et al.*, 2013; Sandmann *et al.*, 2017). While CENH3-containing nucleosomes bind to CENP-C and KNL2, CENP-C interacts with the MIS12 complex, which associates with the NDC80 complex. The major kinetochore components, including KNL1, KNL2, CENP-C, MIS12, and NDC80, are conserved in plants (Dawe *et al.*, 1999; Talbert *et al.*, 2004; Sato *et al.*, 2005; Du & Dawe, 2007; Lermontova *et al.*, 2013).

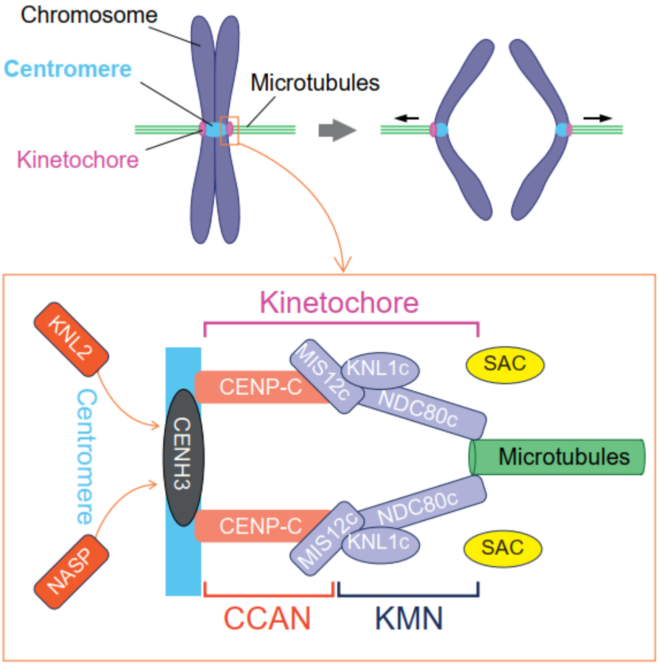


Figure 7. A model of basic kinetochore structure in plants. The main structure of the kinetochore is formed of constitutive centromere-associated network (CCAN) and the KMN (KNL1, MIS12, and NDC80 complexes. Adapted from Hara and Fukagawa (2020).

KNL2 function

The correct assembly of the kinetochore complex requires the deposition of CENH3 at the centromeric region, depending on CENH3 assembly factors and chaperones (Silva & Jansen, 2009), centromeric repeats transcripts (Bobkov *et al.*, 2018; Talbert & Henikoff, 2018), and the epigenetic modification of the centromeric chromatin (Bergmann *et al.*, 2011; Kim *et al.*, 2012). KNL2 (M18BP1) plays a crucial role in new CENH3 deposition after replication (**Figure 7**). M18BP1 in vertebrates is part of the Mis18 complex, including Mis18α and Mis18β. However, Mis18α and Mis18β have not yet been identified in plants. The KNL2 proteins identified so far contain the SANTA (SANT-associated) domain (Zhang *et al.*, 2006), a protein module of ~90 amino acids. The function of the SANTA domain has remained obscure for a long time. For instance, deleting the SANTA domain in Arabidopsis KNL2 has not impaired its targeting to centromeres (Lermontova *et al.*, 2013) nor disrupted its interaction with DNA (Sandmann *et al.*, 2017). A conserved CENPC-k motif, which is highly similar to the CENPC motif of the CENP-C protein (Sugimoto *et al.*, 1994; Talbert *et al.*, 2004; Kato *et al.*, 2013), was identified on the C-terminal part of the KNL2 homologs in a broad spectrum of eukaryotes (Kral, 2015). The importance of this domain for the centromeric targeting of KNL2 was demonstrated in Arabidopsis (Sandmann *et al.*, 2017), *Xenopus* (French *et al.*, 2017) and chicken (Hori *et al.*, 2017). Moreover, direct binding of CENPC-k to CENH3 nucleosomes was shown (French *et al.*, 2017; Hori *et al.*, 2017). KNL2 in eutherian mammals lacks a CENPC-k motif (Kral, 2015), and centromeric localization of human KNL2 may be achieved by directly binding the SANTA domain to CENP-C (French & Straight, 2019). Depletion of KNL2 in different organisms causes defects in CENH3 assembly (Fujita *et al.*, 2007; Lermontova *et al.*, 2013). For instance, knockout of M18BP1 and other components of the Mis18 complex in human HeLa cells with RNAi abolished centromeric recruitment of newly synthesized CENP-A, leading to chromosome missegregation and interphase micronuclei (Fujita *et al.*, 2007). The homozygous *knl2* mutant of Arabidopsis is viable despite reduced CENH3 levels and mitotic and meiotic abnormalities resulting in reduced growth rate and fertility (Lermontova *et al.*, 2013). The fact that the *knl2* mutant CENH3 is still localized at the centromeres suggests that this is not the only mechanism responsible for the centromeric loading of CENH3 in plants. Taken together, although KNL2 protein homologs have been identified in different organisms as components of the CENH3 loading machinery, they differ considerably in the composition of their functional domains, interacting partners, and localization timing in the mitotic cell cycle.

2.1.3 Telomere

Telomeres are the key components of the chromosome, which solve the "end-protection" problem by distinguishing the ends of chromosomes from DNA damage and the "end-replication" problem by facilitating the complete replication of chromosomal ends via DNA replication machinery and telomerase. In addition, telomeres may also involve in a process called interstitial telomere loops (ITLs) or telomere position effects over long distances (TPE-OLD) (Robin *et al.*, 2014; Kim *et al.*, 2016), which could influence gene expression over much larger distances. Telomere length in plants is maintained by telomerase, a specialized reverse transcriptase. Telomerase is a ribonucleoprotein (RNP) enzyme that minimally contains a telomerase reverse transcriptase (TERT) protein subunit, which provides catalytic activity, and a long noncoding telomerase RNA subunit, a small part of which serves as a template for synthesis of short sequence motifs of telomeric DNA (Greider & Blackburn, 1987; Shakirov *et al.*, 2022). Telomere function relies on the proper length of the telomeric DNAs and telomere-binding protein complexes.

The telomeric repeat sequence is relatively conserved across kingdoms, represented by the TTAGGG motif (human-type) in animals (Moyzis *et al.*, 1988) and TTTAGGG motif（Arabidopsis-type）in plants (Richards & Ausubel, 1988). However, there are several switch-points identified in the divergence of telomeric motifs during the evolution of land plants (**Figure 8**) (Schrumpfová *et al.*, 2016; Peška & Garcia, 2020), including the carnivorous plant *Genlisea hispidula* (TTCAGG/TTTCAGG) (Tran *et al.*, 2015), the genus *Cestrum* (Solanaceae;

TTTTTTAGGG) (Peška *et al.*, 2015), and plants from the Asparagales order with either a vertebrate-type telomere repeat TTAGGG (Sýkorová *et al.*, 2003) or the genus-specific CTCGGTTATGGG repeat in *Allium* [onion, garlic species; (Fajkus *et al.*, 2016)]. In addition, telomeric repeats also vary in red algae, green algae and Glaucophytes (**Figure 8**). For example, in addition to the Arabidopsis-type of telomeric motif, the *Chlamydomonas*-type (TTTTAGGG), human-type (TTAGGG), and a novel TTTTAGG repeat have been described in algae (Schrumpfová *et al.*, 2016).

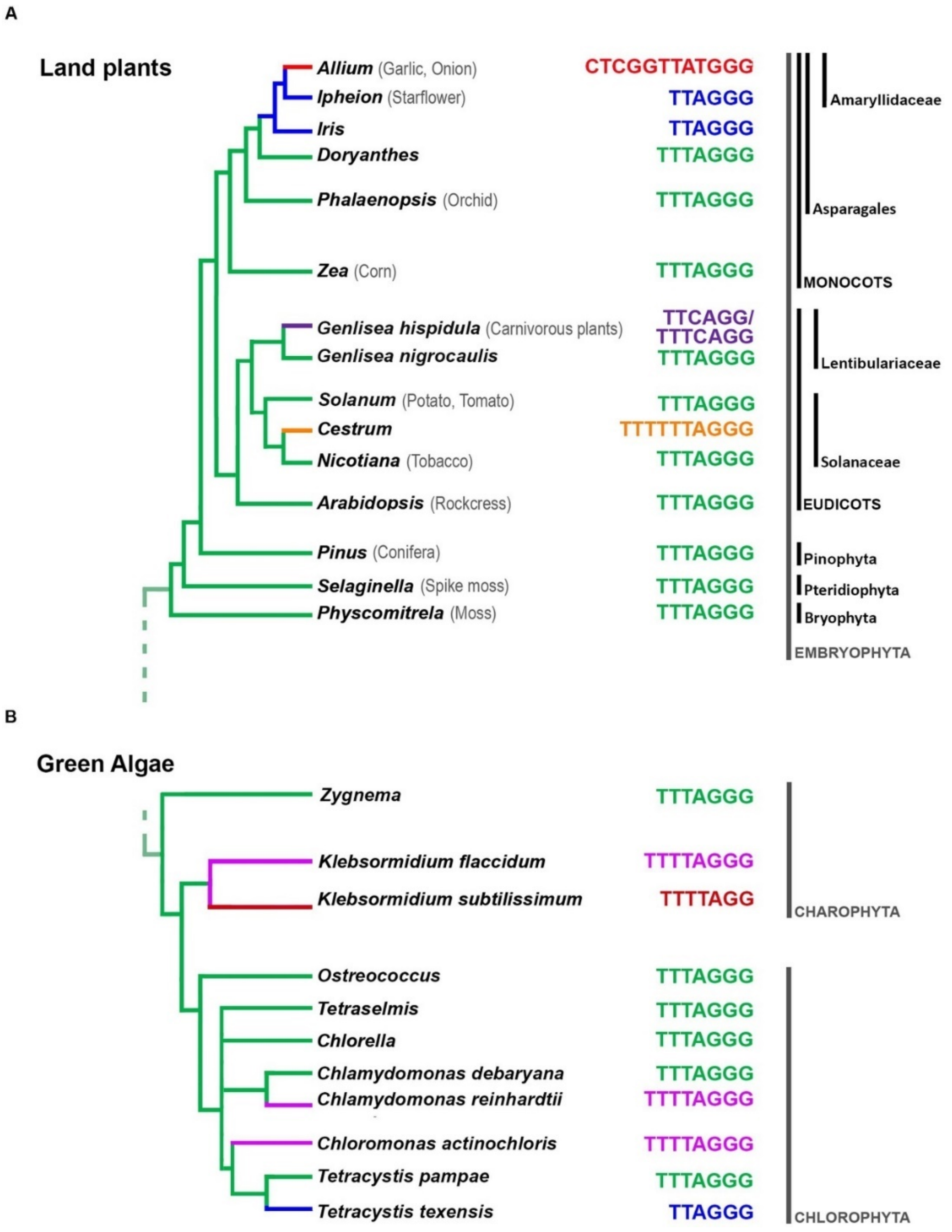


Figure 8. Summary of the telomeric motifs in land plants (A) and green algae (B). Adapted from Schrumpfová *et al.* (2016).

Telomeric repeats are not observed exclusively at the end of plant chromosomes. In fact, telomeric repeats are present in multiple internal sites of chromosomes in many plant species (**Figure 9**) (Fuchs *et al.*, 1995; Majerová *et al.*, 2014; Aksenova & Mirkin, 2019). Such sequences are named Interstitial Telomeric Repeats (ITRs) and can be divided into two major groups: heterochromatic ITRs and short ITRs (Aksenova & Mirkin, 2019). Heterochromatic ITRs are large blocks of telomeric repeats that mainly occupy centromeric or pericentromeric regions, while short ITRs are usually distributed at various positions in chromosomes.

Although we observed ubiquitous ITRs across plants, the mechanisms for the ITRs at intrachromosomal sites are not fully understood. One possible mechanism is that most short ITRs resulted from the insertion of telomeric repeats when a double-stranded break in DNA was repaired by non-homologous end joining with possible telomerase recruitment (Jia & Chai, 2018). Other possible mechanisms of heterochromatic ITRs formation could be the chromosomal rearrangements involving telomeric regions, transposition of telomeric repeats by mobile elements or heterologous recombination (Fuchs *et al.*, 1995; Souza *et al.*, 2016). In addition, it may be hypothesized that the ITRs evolved by similar mechanisms to the dynamic satellite DNA sequences due to the telomeric motif as a particular kind of minisatellites.

ITRs have long been regarded as junk DNA associated with chromosomal rearrangements and aberrations. However, during the last decade, accumulated data changed our understanding that ITRs may involve telomere maintenance, genome-wide regulation of gene expression, and 3D genome structure (Wood *et al.*, 2014; Wood *et al.*, 2015; Shay, 2018). For example, large blocks of ITRs such as heterochromatic ITRs are supposed to confer even more fragility and contribute to genome evolution (Bolzan, 2012).

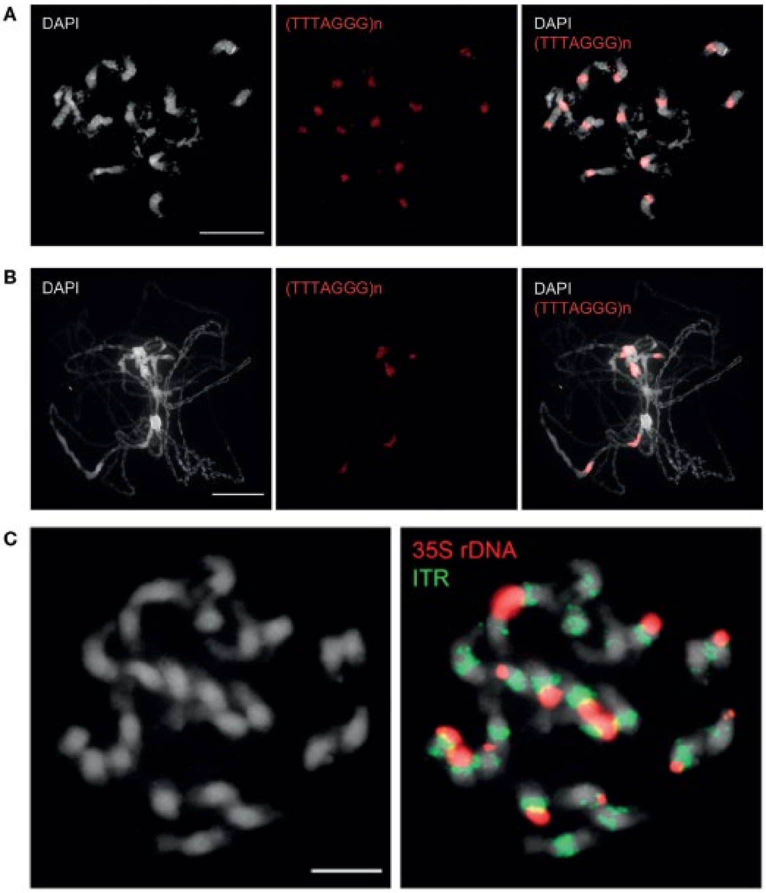


Figure 9. Fluorescence *in situ* localization of the telomeric repeats in Brassicaceae species. In *Ballantinia antipoda*, the telomere repeats (red) hybridize preferentially to centromeres, whereas minor signals at chromosome termini are less prominent on mitotic (A) and pachytene (B) chromosomes. (C) Localization of 35S rDNA (red signals) and interstitial telomeric repeats (green signals) in *Cardamine cordifolia*. Adapted from Majerová *et al.* (2014) and Mandáková *et al.* (2016).

2.1.4 Nucleolar organizer region and rDNA

Nucleolar organizer regions (NORs) are chromosomal landmarks that consist of tandemly repeated sequences of ribosomal RNA genes. Only loci with active rRNA transcription and processing during the interphase can form a nucleolus. The nucleolus is a prominent nuclear condensate that plays a central role in ribosome biogenesis. In eukaryotes, ribosomal RNA genes are transcribed by RNA polymerase I into a large primary precursor, which is then processed into the 18S, 5.8S, and 26S rRNAs (Turowski & Tollervey, 2015). The amount of active rRNAs varies with cellular demand for ribosome production and protein synthesis. In interspecific hybrids or allopolyploid species, the NORs of one (sub)genome can be dominant over the NORs of another (sub)genome, which is referred to as "nucleolar dominance" (Jiang & Gill, 1994; Pikaard, 2000; McStay, 2006; Tucker *et al.*, 2010; Borowska-Zuchowska *et al.*, 2020). Nucleolar dominance is an epigenetic phenomenon that describes the expression of 35S rRNA genes inherited from one progenitor due to the silencing of the other progenitor's rRNA genes. For example, in *A. suecica*, the allotetraploid hybrid of Arabidopsis and *A. arenosa*, the Arabidopsis-derived rRNA genes are silenced (Preuss *et al.*, 2008). Nucleolar dominance is primarily regulated by epigenetic conditions, such as DNA methylations and histone modifications (Pikaard, 2000; Tucker *et al.*, 2010).

In eukaryotes, two types of rDNA are present, including 35S (in plants) / 45S (in animals) rDNA encoding 18S-5.8S-26S rRNA genes, and 5S rDNA encoding 5S rRNA (Garcia *et al.*, 2017). The 35S rDNA unit usually has 8 - 14 kb containing coding regions, internal transcribed spacers (ITS), and intergenic spacers (IGS). In most plants, the chromosomal loci of 18S-5.8S-26S rRNA genes are separated from the 5S rRNA genes (Separated or S-type arrangement). In rare cases, they are linked in the same unit (Linked or L-type arrangement) (Sone *et al.*, 1999; Garcia *et al.*, 2009; Garcia & Kovarik, 2013). As there are several thousand rDNA units in plant genomes, the concerted evolution process may maintain the integrity and homogeneity of 35S and 5S rDNA units (Eickbush & Eickbush, 2007). Frequent whole genome duplication (WGD) or hybridization events have been attributed to the variability of rDNA and the presence of multiple rDNA loci. However, after polyploidization events, concerted evolution has been observed in allopolyploids, and a single type of rDNA is commonly found (Wendel *et al.*, 1995; Volkov *et al.*, 1999; Kotseruba *et al.*, 2003; Bao *et al.*, 2010; Weiss-Schneeweiss *et al.*, 2012). Moreover, even in recently formed polyploids, the paternally inherited rDNA genes and loci were eliminated, and the pattern of concerted evolution was observed, such as in the trigenomic allopolyploid *Cardamine* × *schulzii* (Zozomová-Lihová *et al.*, 2014). To explain the concerted evolution process, there are two scenarios, namely stochasticity and driven by selection (Zozomová-Lihová *et al.*, 2014). For instance, whether the maternal or paternal parents donated them, *Melampodium* polyploids homogenized the same parental rDNA repeats (Weiss-Schneeweiss *et al.*, 2012). In contrast, cotton (Wendel *et al.*, 1995) and rice (Bao *et al.*, 2010) polyploids homogenized to alternative progenitor diploids in different allopolyploid derivatives.

2.1.5 Organellar genomes

Many pieces of evidence support the idea that the chloroplast and mitochondrial genomes are remnants of their prokaryotic endosymbiont genomes (Archibald, 2009; Green, 2011). Indeed, many cyanobacterial genes have been transferred to the host cell nucleus, and their products are targeted back to the chloroplast. In addition, some of them were lost because their products were no longer needed, and some nuclear genes were recruited to the chloroplast service by adding the appropriate target sequences. Thus, a steady stream of organelle DNA appears to be bombarding the nucleus and integrating into the nuclear genome. Large amounts of chloroplast

DNA were found in plant nuclear genomes, and even the rice genome had a complete mitochondrial genome integrated into one of its chromosomes (Huang *et al.*, 2005; Kleine *et al.*, 2009).

The mitochondrial DNA (mtDNA) has important role in plants, which is to encode essential components of the mitochondrial electron transfer chain (Gualberto & Newton, 2017). In addition, the mtDNA can also encode a few proteins involved in the assembly of functional respiratory complexes. Although the number of mitochondrial genes is generally conserved, the size of the mtDNA varies over more than a 100-fold range in plants. For instance, the angiosperm mtDNA size varies significantly, between 200 and 700 kb, and can be as large as 11 Mb in *Silene conica* (Sloan *et al.*, 2012). The non-coding sequences that are not conserved across species contribute significantly to the large size or the size variation of plant mtDNA (Gualberto & Newton, 2017). Interestingly, plant mtDNA evolves more slowly in sequence than animal mtDNA, and gene sequences have very low base substitution rates (Wolfe *et al.*, 1987). But high homologous recombination activity and rearrangements were observed in plant mtDNA. Thus, plant mtDNA evolves rapidly in structure by recombination. For example, sequence duplications, inversions, deletions, and insertions were identified in comparing mtDNA from different accessions of maize and Arabidopsis (Allen *et al.*, 2007; Arrieta-Montiel *et al.*, 2009).

The standard picture of a plastid genome is a circular DNA molecule, 100-200 kbp in size, with a "tetrad" structure consisting of two inverted repeats (IRs) dividing the circle into large and small single-copy regions. Much of the difference in genome size is due to the repetitive sequences contained in the IR, while the number of protein-coding genes and tRNAs was very similar in plants. The plastid genome generally has 16S, 23S, and 5S rRNA genes and 27-31 tRNA genes, which sufficient translate all amino acids, and at least three of the four subunits of prokaryotic RNA polymerases (rpoB, C1, C2). It also has most of the genes for photosystem I, photosystem II, cytochrome b6f complex, and ATP synthase polypeptides.

Although photosynthesis is generally considered a vital function of the plastid, they also play essential roles in other aspects of plant physiology and development, including the synthesis of amino acids, nucleotides, fatty acids, plant hormones, vitamins, and various metabolites, as well as assimilation of sulfur and nitrogen (Daniell *et al.*, 2016). The entire plastid genome sequence of vascular plants was first reported in tobacco (Shinozaki *et al.*, 1986). With the development of next-generation sequencing technologies, we have rapidly acquired complete chloroplast genomes at a low cost. Currently, the National Center for Biotechnology Information (NCBI) archives have thousands of chloroplast genomes, including all major lineages of the plant kingdom. Insights from complete chloroplast genome sequences have improved our understanding of plant biology and diversity. In addition, chloroplast genomes have contributed significantly to phylogenetic studies of multiple plant families and the resolution of evolutionary relationships within phylogenetic clades (Daniell *et al.*, 2016). Furthermore, chloroplast genome sequences show considerable variation within and between plant species in sequence and structural variations. This information is precious for understanding the process of photosynthesis and the climate adaptation of crops.

2.2 Polyploidy and post-polyploid diploidization in plants

2.2.1 Polyploidization

Polyploids [whole genome duplication (WGD)] have three or more sets of chromosomes. Plant polyploidy has been studied for more than a century. In the early 1900s, researchers demonstrated that polyploids might be formed and highlighted the frequency of polyploids in nature (Lutz, 1907; Gates, 1909; Winge, 1917). Two general types of polyploids have long been proposed: those involving the multiplication of one chromosome set and those resulting from the merger of structurally different chromosome sets. Kihara and Ono (1926) used the terms autopolyploidy (auto = "same") and allopolyploidy (allo = "different") to define the different polyploids (**Figure 10**). In addition, genomic research demonstrated that allopolyploids have genome dominance and biased fractionation, whereas autopolyploids do not have these features (Garsmeur *et al.*, 2014). Allopolyploidy has been considered much more common than autopolyploidy in plants. Yet, polyploidy remains underexplored, and its roles and impact in biological processes and across phylogeny are unclear.

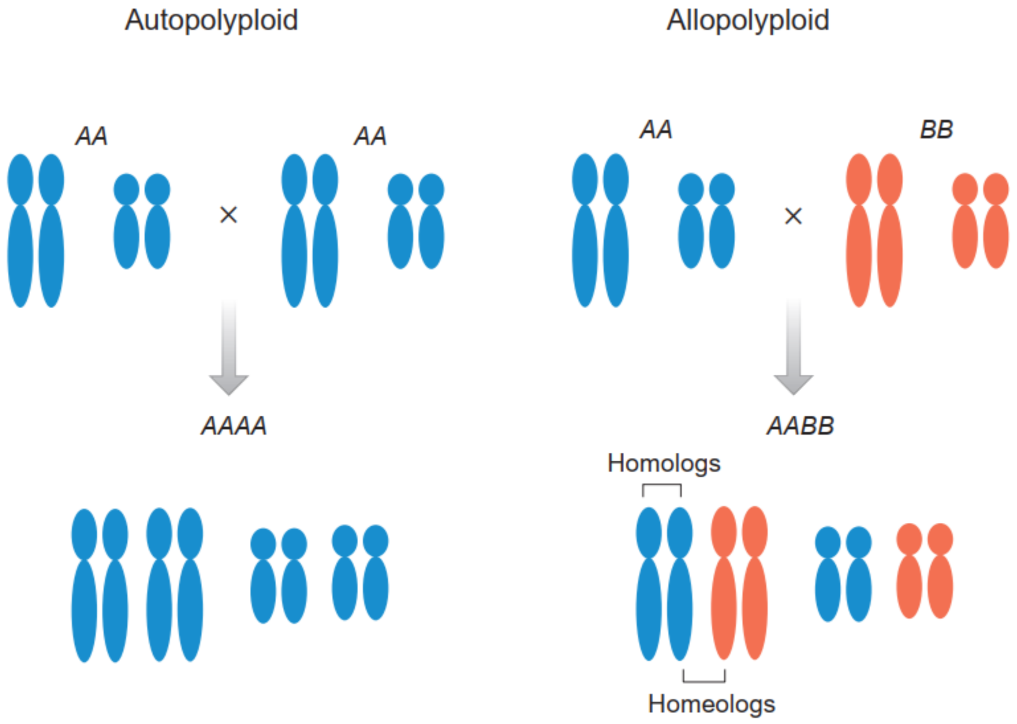


Figure 10. Two types of polyploids: autopolyploid and allopolyploid. Adapted from Yoo *et al.* (2014).

It has long been recognized that polyploidy is a major evolutionary factor in plants. The angiosperms (flowering plants) have received much attention regarding the occurrence of polyploidy. Masterson (1994) estimated that 70% of all angiosperms had experienced one or more episodes of polyploidy in their ancestry. With the genomic technology advance over the last two decades, research in flowering plants demonstrated that the evolution of the angiosperms is influenced by pervasive whole genome duplication events (Jiao *et al.*, 2011; Vekemans *et al.*, 2012; Albert *et al.*, 2013; Ruprecht *et al.*, 2017), such as the ancient *Epsilon* WGD event shared by angiosperm species (**Figure 11**). In addition, many plant lineages have experienced extra ancient or recent WGD events (Bowers *et al.*, 2003; Jiao *et al.*, 2011; Vanneste *et al.*, 2014; Mandáková *et al.*, 2017; Guo *et al.*, 2021). For instance, Brassicaceae, a cosmopolitan plant family comprising almost 4,000 species in 351 genera, have descended from a paleotetraploid ancestor formed by the At-α WGD (Bowers *et al.*, 2003; Haudry *et al.*, 2013; Hohmann *et al.*, 2015). Moreover, more than a dozen genus- and clade-specific mesopolyploid WGDs, post-dating the family-specific paleotetraploid (At-α) WGD, were also identified in different Brassicaceae tribes (Mandáková *et al.*, 2017; Guo *et al.*, 2021).

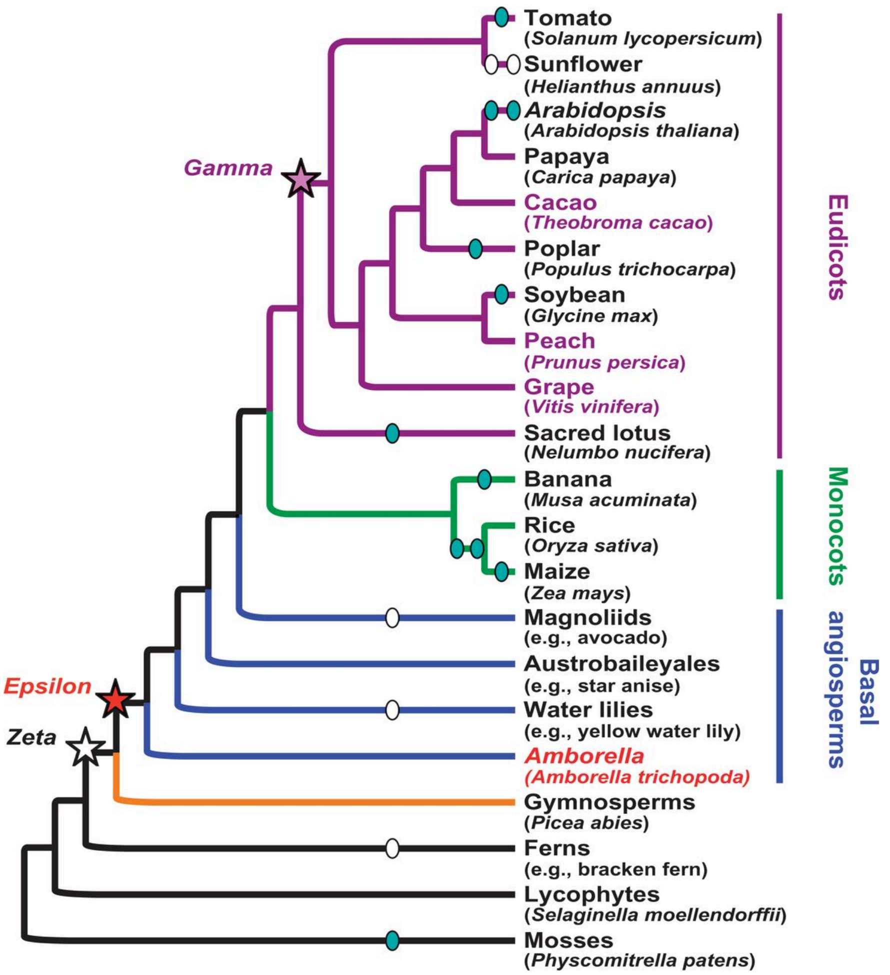


Figure 11. An overview of whole genome duplication (WGD) events during the evolution of land plants. Different color symbols mark the hypothesized polyploidy events. Adapted from Albert *et al.* (2013).

2.2.2 Genome diploidization

In the last two decades, polyploid genome evolution has become a prevalent research topic in evolutionary biology. Polyploidization or WGD is frequently followed by post-polyploid diploidization (PPD) (Soltis *et al.*, 2009; Vanneste *et al.*, 2014; Hohmann *et al.*, 2015; Walden *et al.*, 2020). The resulting polyploid genomes generally have not remained static, but returned to pseudo-diploid genomes through the process collectively named diploidization (Thomas *et al.*, 2006), gradually erasing and concealing the signatures of ancient WGD events. In plants, PPD plays an important evolutionary force in promoting diversification and speciation. For instance, polyploid genomes may undergo diploidization potentially resulting in a continuum of more or less reproductively isolated populations, and eventually species and clades.

Based on the time elapsed since a WGD and the diploidization rate, WGD events can be broadly classified as paleopolyploid, mesopolyploid, and neopolyploid (Carman, 1997; Mandáková *et* *al.*, 2010). Due to progressive genome diploidization, neopolyploids may turn into mesopolyploids and paleopolyploids over time (Mandáková & Lysak, 2018). PPD is associated with a wide range of processes, such as genome downsizing, chromosomal rearrangements, modulation of gene expression, and epigenetic reprogramming (Conant *et al.*, 2014; Vicient & Casacuberta, 2017; Mandáková & Lysak, 2018). Descending dysploidy is one of the most crucial diploidization routes, resulting in decreased base chromosome number (x). In addition, biased gene fractionation has been widely observed across angiosperm lineages, which means paralogous genes of one subgenome in a diploidized polyploid genome are preferentially retained and exhibit higher gene expression levels relative to the other (more fractionated) subgenome(s) (Freeling *et al.*, 2012; Geiser *et al.*, 2016; Mandáková *et al.*, 2017).

Genome reshuffling and decreases do not proceed with the same speed and intensity along all clades descending from a single WGD event (**Figure 12**) (Mandáková *et al.*, 2017; Guo *et al.*, 2021). Life histories, mating systems, and other factors may influence the different rates of descending dysploidy. For example, descending dysploidies are thought to proceed faster in annuals than in perennial and woody plants (Luo *et al.*, 2015; Miguel *et al.*, 2015). The higher number of generations in annuals is associated with a higher probability of DSB misrepair, potentially generating chromosome rearrangements.

Chromosome number variation coupled with PPD can be correlated with speciation events, adaptive radiations, and cladogenesis (**Figure 12**) (Mandáková & Lysak, 2018). The WGD radiation lag-time model explains that the post-polyploid diversification of the crown group frequently commenced millions of years after the corresponding WGD (Schranz *et al.*, 2012). The lag between a WGD and subsequent diversification demonstrated that the ancestor polyploid genome or its populations must undergo some "adjustment" — the process of genome diploidization. PPD acting with differing intensities on the primary polyploid may generate genetically variable descendants with reproductive barriers, eventually resulting in speciation and cladogenetic events (**Figure 12**) (Mandáková & Lysak, 2018). The evolutionary role of post-polyploid diploidization will finally be elucidated with new paleogenomics and phylogenomics data in genome evolution and speciation.

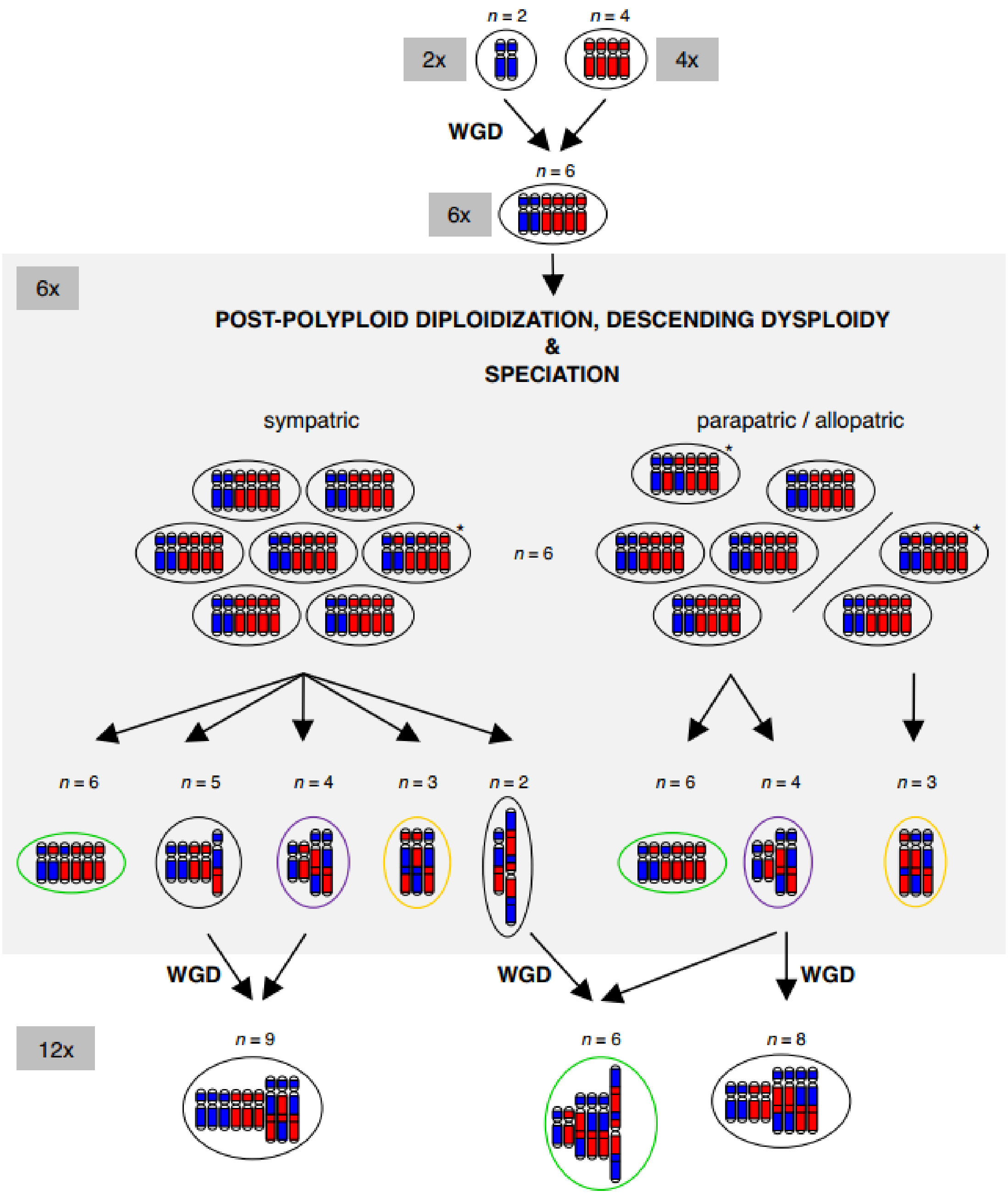


Figure 12. Speciation and diversification driven by post-polyploid diploidization (PPD). Adapted from Mandáková and Lysak (2018).

2.3 Sequencing technologies and applications

Deoxyribonucleic acid (DNA) is the genetic material in plants and animals. DNA sequences carry the genetic information of life and guide biological development. Since Watson and Crick (1953) described the structure of double-stranded DNA, deciphering the nucleic acid sequence has become one of the major focuses of molecular biology. DNA sequencing is the method or technology to determine the order of the four bases: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA sequencing can be used to decipher the sequence of individual gene sequences, whole chromosomes, or complete genomes and has accelerated biological and medical research. In the past four decades, sequencing technology has experienced three stages of development (**Figure 13**). Based on these sequencing technologies, we have obtained valuable knowledge from struggling towards the deduction of the coding sequence of a single gene to whole genome sequencing. During the same time, bioinformatics was born as an interdisciplinary subject with broad application prospects in improving data processing capabilities and generating valuable biological information.

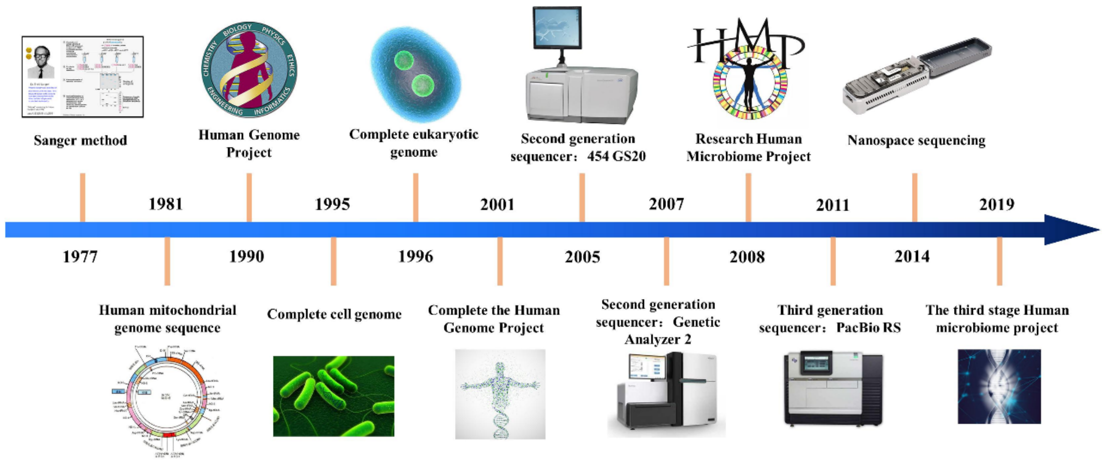

Sanger method
Human Genome Project
Complete eukaryotic genome
Second generation sequencer：454 GS20
HMP
Research Human Microbiome Project
Nanospace sequencing
1981
1995
2001
2007
2011
2019
1977
1990
1996
2005
2008
2014
Human mitochondrial genome sequence
Complete cell genome
Complete the Human Genome Project
Second generation sequencer：Genetic Analyzer 2
Third generation sequencer：PacBio RS
The third stage Human microbiome project

Figure 13. History of sequencing technology. Adapted from Yang *et al.* (2020).

2.3.1 First-generation DNA sequencing

Two types of DNA sequencing technologies were developed independently in the 1970s. In the late 1970s, Sanger and colleagues (Sanger & Coulson, 1975; Sanger *et al.*, 1977) described a method using DNA polymerase, which makes use of inhibitors that terminate the newly synthesized chains at specific residues. This technology developed by Sanger and colleagues, commonly referred to as Sanger sequencing, is still widely used in conventional sequencing applications. On the other hand, Maxam and Gilbert proposed the chain degradation method to sequence DNA (Maxam & Gilbert, 1977). These methods are collectively called first-generation sequencing technology. However, the sequencing cost and throughput seriously affect its large-scale application.

In 1990, the human genome project (HGP) (Lander *et al.*, 2001), an international project to decipher the complete sequence of nucleotide base pairs of a human, was launched, funded mainly by the National Institutes of Health (NIH). A parallel project (Venter *et al.*, 2001) was initiated in 1998 and launched by Celera Corporation (Alameda, CA, USA); both initiatives released a draft genome in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001). Thanks to the Sanger method applied in the human genome sequencing project, a complete genome was released in 2004 (Collins *et al.*, 2004), but assembling a gap-less genome is still a big challenge. Even after two decades, the human reference genome is being improved and corrected (Miga *et al.*, 2020; Logsdon *et al.*, 2021; Nurk *et al.*, 2022).

2.3.2 Second-generation DNA sequencing

The development and commercialization of various next-generation sequencing (NGS) technologies were stimulated by the human genome project, which required the use of cheaper and higher throughput DNA sequencing methods to supplement the time- and resources-consuming Sanger sequencing. The newly developed methods included 454 Life Sciences (now Roche) (Margulies *et al.*, 2005), Solexa/Illumina, SOLiD, Ion Torrent platform, Complete Genomics (Beijing Genomics Institute, BGI) (Drmanac *et al.*, 2010), and Polonator, which have enabled producing more sequencing in parallel at low cost. The pyrosequencing method by 454 Life Sciences was the first NGS technology released in 2005 (Margulies *et al.*, 2005). One year later, the Solexa/Illumina sequencing platform was commercialized. Applied Biosystems (now Life Technologies) released sequencing technology by Oligo Ligation Detection (SOLiD) in 2007 (Valouev *et al.*, 2008). In 2010, Ion Torrent (now Life Technologies) released the Personal Genome Machine (PGM), resembling the 454 system. Complete Genomics platform used the combinatorial probe–anchor ligation (cPAL) or combinatorial probe–anchor synthesis (cPAS) to perform DNA sequencing (Drmanac *et al.*, 2010).

Nowadays, the Illumina platform is the most widely used NGS technology based on a synthesis approach and detection of fluorescently modified nucleotides. The Illumina sequencing process consists of three major steps: (i) DNA library preparation, (ii) Cluster amplification, and (iii) Sequencing by synthesis and image analysis. The Illumina technology allows the sequencing of fragments up to 300 bp and provides the ultra-high-throughput NovaSeq 6000 system (6,000 gigabases per sequencing run). The enormous numbers of reads generated by NGS enabled the sequencing of entire genomes at an unprecedented speed. However, a drawback of NGS technologies was their relatively short reads. This made genome assembly more complex and required the development of novel alignment algorithms. A comparison of Sanger sequencing and Illumina next-generation sequencing is depicted in **Figure 14**.

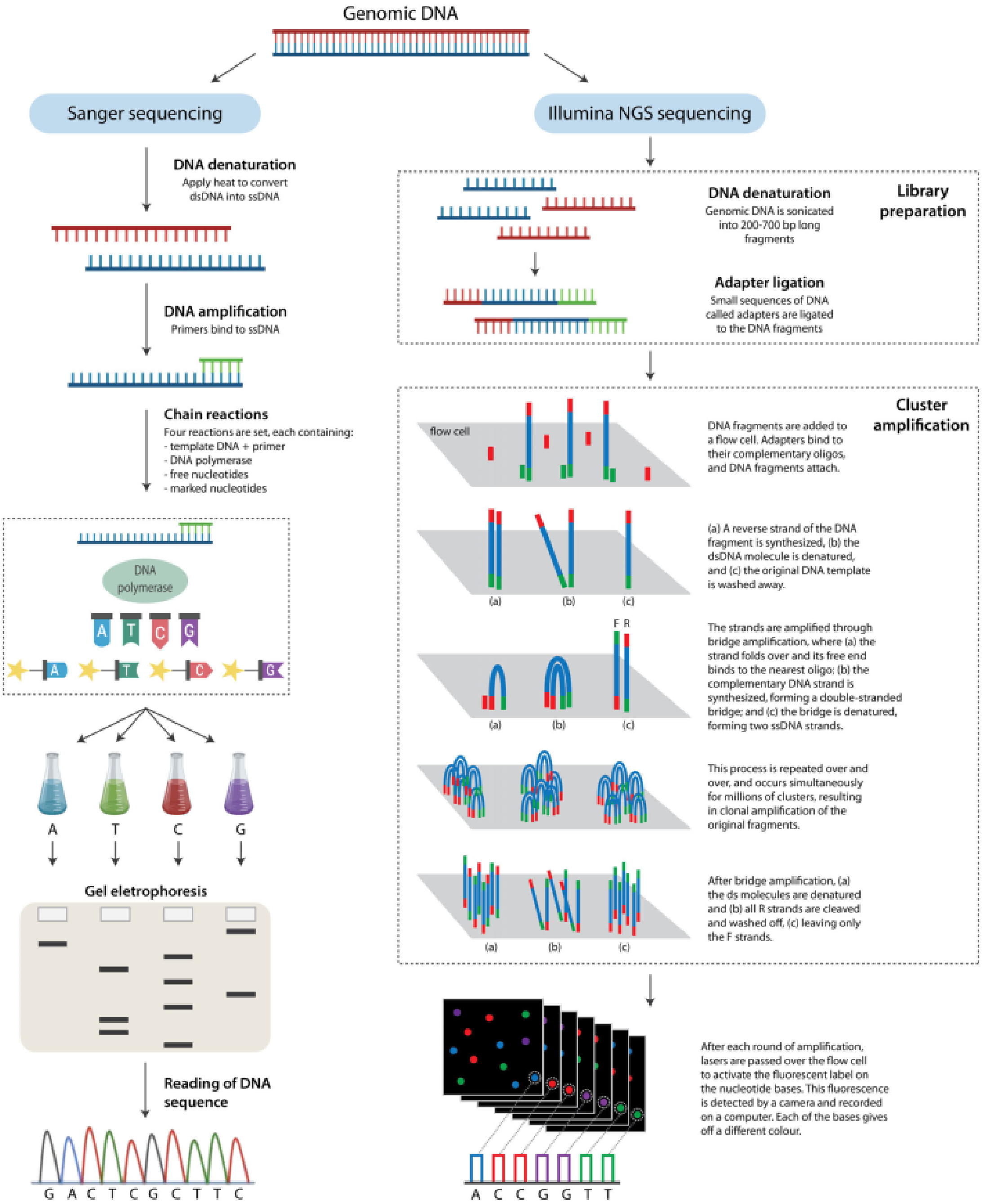

Genomic DNA
Sanger sequencing
Illumina NGS sequencing
DNA denaturation
Apply heat to convert dsDNA into ssDNA
DNA amplification
Primers bind to ssDNA
Chain reactions
Four reactions are set, each containing:
- template DNA + primer
- DNA polymerase
- free nucleotides
- marked nucleotides
DNA polymerase
A T C G
A T C G
A T C G
Gel eletrophoresis
Reading of DNA sequence
G A C T C G C T T C
DNA denaturation
Genomic DNA is sonicated into 200-700 bp long fragments
Library preparation
Adapter ligation
Small sequences of DNA called adapters are ligated to the DNA fragments
Cluster amplification
flow cell
DNA fragments are added to a flow cell. Adapters bind to their complementary oligos, and DNA fragments attach.
(a) (b) (c)
(a) A reverse strand of the DNA fragment is synthesized, (b) the dsDNA molecule is denatured, and (c) the original DNA template is washed away.
F R
(a) (b) (c)
The strands are amplified through bridge amplification, where (a) the strand folds over and its free end binds to the nearest oligo; (b) the complementary DNA strand is synthesized, forming a double-stranded bridge; and (c) the bridge is denatured, forming two ssDNA strands.
This process is repeated over and over, and occurs simultaneously for millions of clusters, resulting in clonal amplification of the original fragments.
(a) (b) (c)
After bridge amplification, (a) the ds molecules are denatured and (b) all R strands are cleaved and washed off, (c) leaving only the F strands.
After each round of amplification, lasers are passed over the flow cell to activate the fluorescent label on the nucleotide bases. This fluorescence is detected by a camera and recorded on a computer. Each of the bases gives off a different colour.
A C C G G T T

Figure 14. Comparison of Sanger sequencing (left) and Illumina next-generation sequencing (right). Adapted from Young and Gillung (2020).

2.3.3 Third-generation DNA sequencing

Both plant and animal genomes are highly complex with many repetitive elements and satellite sequences. Short-read technologies are generally insufficient to assemble them because of the repetitive fragments longer than the read length. In addition, short-read technologies are not able to sequence full length of transcriptome. To overcome these issues, long-read or third-generation sequencing was introduced, as they can deliver reads over ten kilobases (kb). Long reads can span complex or repetitive regions with a single continuous read. There are two main types of long-read technologies: real single-molecule long-read sequencing and synthetic long-read sequencing (Reuter *et al.*, 2015; Jiao & Schneeberger, 2017). The first technology sequences the full-length DNA/RNA molecules, while the latter allows for the assembly of long reads from short-read sequencing data.

Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing (www.pacb.com) and Oxford Nanopore Technologies (ONT; nanoporetech.com) are two of the major platforms offering real long-read sequencing. SMRT sequencing overcomes the short-length limitations of the NGS technologies, generating reads with an average size of nearly 20 kb. Genomes of Arabidopsis (Berlin *et al.*, 2015) and *Oropetium thomaeum* (VanBuren *et al.*, 2015) are among the first sequenced plant genomes using PacBio data. In ONT sequencing, single DNA molecules are guided to pass through nanopores, which directly detect the sequences of the nucleotides (Clarke *et al.*, 2009). Thus, the read length is theoretically only limited by the size of the DNA molecules, wherein the most extended reads are up to several hundred kb. In 2014, the first consumer prototype of the nanopore sequencer MinION became available, the smallest sequencing device from ONT. With the unprecedented read length, ONT sequencing allowed assembling complex genomic regions of the Arabidopsis genome, including sequences of the ribosomal DNA (rDNA) and centromeric regions (Michael *et al.*, 2018). In addition to their long read length, both SMRT and ONT technologies can directly identify regular and modified bases such as inosine, pseudouridine, or methylated adenosine (Simpson *et al.*, 2017). A schematic representation of both platforms is shown in **Figure 15**. As long-read sequencing reads have up to 15% sequencing error rates, correction with short sequencing reads or self-correction with sufficient sequencing data is needed (Koren *et al.*, 2012; Chin *et al.*, 2013). Due to the new sequencing strategy, PacBio recently offered a solution to overcome the high error rate in long-read sequencing, which was achieved by generating the circular consensus sequence (CCS; also known as HiFi read) of the same DNA molecules.

Unlike real-time long-read sequencing, synthetic long-read sequencing is based on existing short-read sequencers and is achieved by a barcoding system (Voskoboynik *et al.*, 2013; McCoy *et al.*, 2014). Unique barcoding helps to identify sequencing reads originating from the same molecule, therefore, the long DNA fragment can be computationally re-assembled. Currently, two technologies are available for generating synthetic long-reads, including the Illumina synthetic long-read sequencing platform and the 10X Genomics emulsion-based system (**Figure 15**).

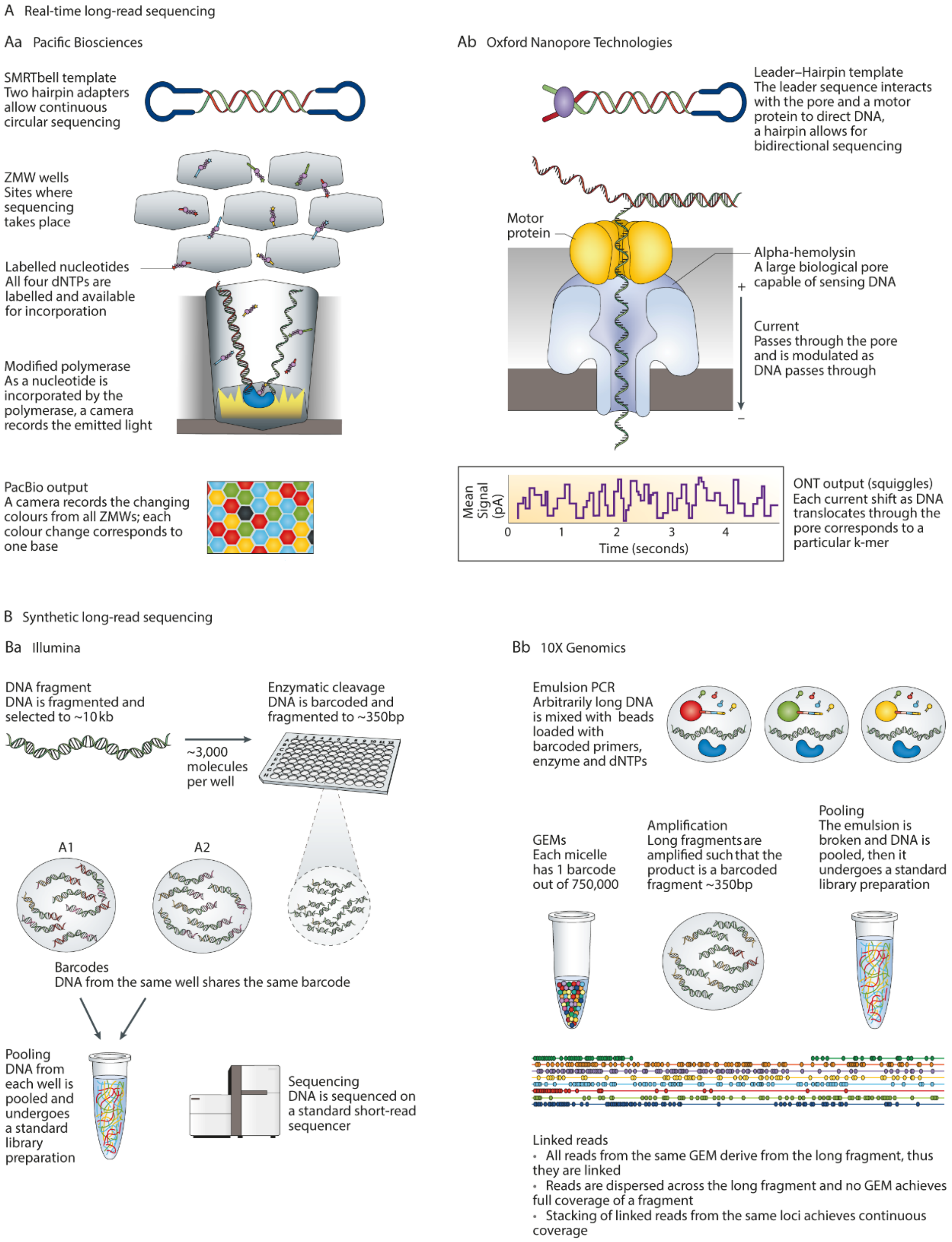

A Real-time long-read sequencing
Aa Pacific Biosciences
SMRTbell template
Two hairpin adapters allow continuous circular sequencing
ZMW wells
Sites where sequencing takes place
Labelled nucleotides
All four dNTPs are labelled and available for incorporation
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light
PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base
Ab Oxford Nanopore Technologies
Leader–Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing
Motor protein
Alpha-hemolysin
A large biological pore capable of sensing DNA
+
Current
Passes through the pore and is modulated as DNA passes through
–
Mean Signal (pA)
0
1
2
3
4
Time (seconds)
ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer
B Synthetic long-read sequencing
Ba Illumina
DNA fragment
DNA is fragmented and selected to ~10kb
~3,000 molecules per well
Enzymatic cleavage
DNA is barcoded and fragmented to ~350bp
A1
A2
Barcodes
DNA from the same well shares the same barcode
Pooling
DNA from each well is pooled and undergoes a standard library preparation
Sequencing
DNA is sequenced on a standard short-read sequencer
Bb 10X Genomics
Emulsion PCR
Arbitrarily long DNA is mixed with beads loaded with barcoded primers, enzyme and dNTPs
GEMs
Each micelle has 1 barcode out of 750,000
Amplification
Long fragments are amplified such that the product is a barcoded fragment ~350bp
Pooling
The emulsion is broken and DNA is pooled, then it undergoes a standard library preparation
Linked reads
• All reads from the same GEM derive from the long fragment, thus they are linked
• Reads are dispersed across the long fragment and no GEM achieves full coverage of a fragment
• Stacking of linked reads from the same loci achieves continuous coverage

Figure 15. Real-time long-read sequencing and synthetic long-read sequencing platforms. Adapted from Goodwin *et al.* (2016).

2.3.4 Genome assembly

Thanks to the advances in sequencing technologies, the number of published plant genomes has increased dramatically in the past 20 years. The growing wealth of genomic data has enabled the development of bioinformatic and genomic approaches to address many interesting questions in genome biology and evolution. The first sequenced plant genome was the model plant Arabidopsis in 2001 (Kaul *et al.*, 2000). The Arabidopsis genome, with small genome size (~140 Mb), represents a gold standard for plant genome sequencing, wherein the quality of the assembly for its five chromosomes has been dramatically improved while a few gaps remain (Naish *et al.*, 2021; Wang *et al.*, 2021; Hou *et al.*, 2022). Since 2000, hundreds of plant genomes have been sequenced, assembled, and updated (**Figure 16**). Although DNA sequencing technologies have dramatically developed, getting a gap-less genome assembly remains challenging, particularly for polyploid plant species.

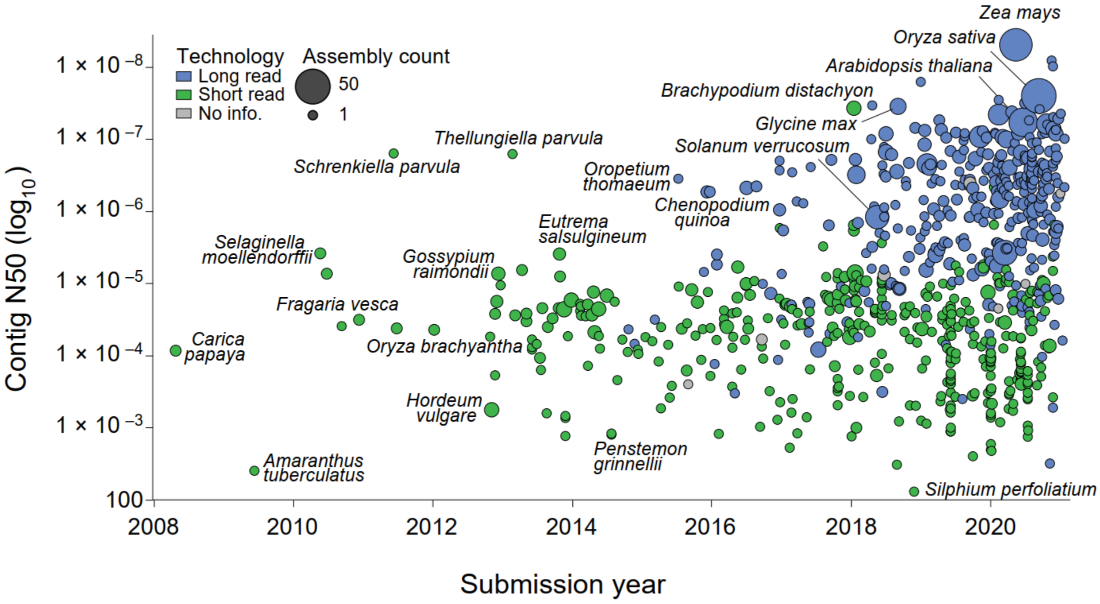


Figure 16. Changes in genome assembly quality and availability over time in land plants. Adapted from Marks *et al.* (2021).

Although hundreds of plant genome resources are available, only a small fraction of the extant land plants have had their genomes, and these efforts have not been evenly distributed across clades. Some orders of land plants are over-represented in genome assembly databases based on species richness. For instance, high-quality genome assemblies are available, and thousands of accessions or ecotypes have been resequenced in some model plants and crop species (Bayer *et al.*, 2020). Brassicaceae is the most heavily sequenced plant family, with genome assemblies for dozens of species, including Arabidopsis and numerous vegetables. In contrast, for most other groups, none or only a single species has a genome assembly.

De novo assembly strategies

As innovative sequencing technologies were introduced, genome assembly approaches were rapidly developed. *De novo* assembly strategy for short NGS reads has three main steps (**Figure 17**): contig assembly, scaffolding, and gap-filling (Paszkiewicz & Studholme, 2010; Compeau *et al.*, 2011; El-Metwally *et al.*, 2013; Nagarajan & Pop, 2013; Simpson & Pop, 2015). In the first step, the short reads are assembled as contigs without gaps. Then, the contigs are connected by large-insert (pair-end/mate-pair) reads, and an ordered set of connected contigs is defined as a scaffold. The gaps between the contigs can be filled using other independent reads (gap-filling step) to complete the assembly. *De novo* genome assembly using short reads still need to overcome many computational challenges, including the correction of sequencing errors, uneven read depth, the topological complexity of repetitive elements, and high computation cost (Sohn & Nam, 2018).

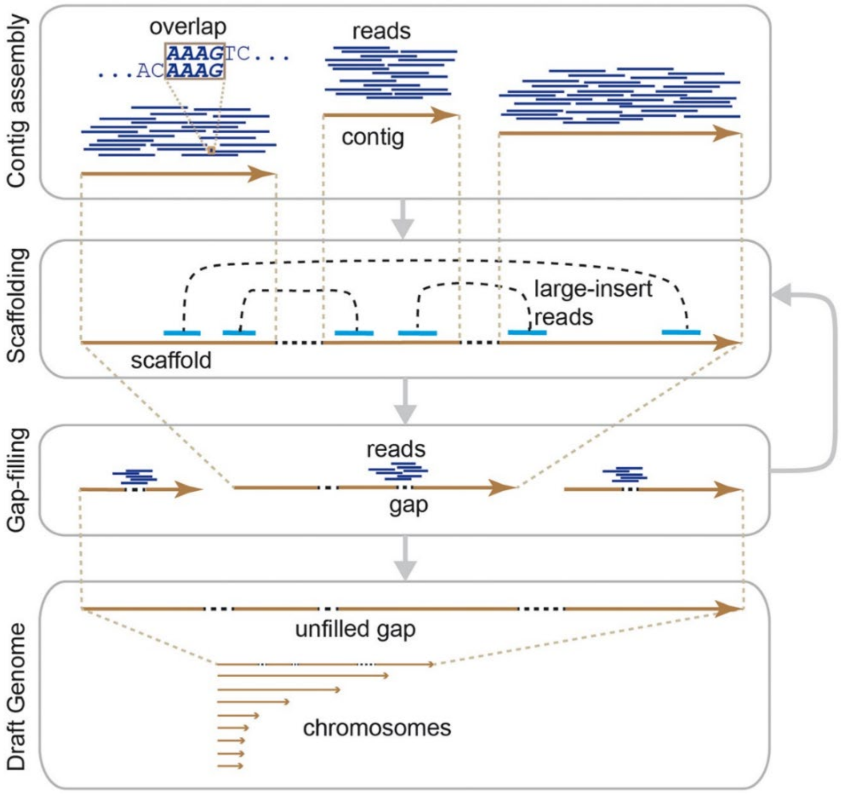


Figure 17. Workflow of the *de novo* assembly of a whole genome using short NGS reads. Adapted from Sohn and Nam (2018).

A lot of tools were developed for *de novo* genome assembly using NGS reads. The de Bruijn graph-based algorithm has been applied to many assemblers as an efficient genome assembly approach for short-read data (Nagarajan & Pop, 2013). For example, ALLPATHS-LG, based on the Eulerian de Bruijn graph, requires relatively large memory for large genomes (Gnerre *et al.*, 2011). Velvet (Zerbino & Birney, 2008) and SPAdes (Bankevich *et al.*, 2012) are the Eulerian de Bruijn graph assemblers. SparseAssembler (Ye *et al.*, 2012) and SOAPdenovo2 (Luo *et al.*, 2012) were developed based on the sparse k-mer and reduced required memory. There are other assemblers for short reads, including ABySS (Simpson *et al.*, 2009), SGA (Simpson & Durbin, 2012), MaSuRCA (Zimin *et al.*, 2013), Meraculous (Chapman *et al.*, 2011), and JR-Assembler (Chu *et al.*, 2013).

High-throughput chromosomal conformation capture (Hi-C) sequencing technology can determine how a genome is folded by measuring the frequency of contact between pairs of loci (Lieberman-Aiden *et al.*, 2009). Contact frequency mostly depends on the one-dimensional (1D) distance between a pair of loci. For instance, loci separated by 10 kb in the Arabidopsis genome form contacts more often than at a distance of 100 kb. Thus, Hi-C data can provide links across various length scales, spanning even whole chromosomes. Indeed, Hi-C has been used to improve draft genome assemblies and to produce chromosome-length scaffolds (Lieberman-Aiden *et al.*, 2009; Marie-Nelly *et al.*, 2014; Naish *et al.*, 2021). In this process, Hi-C data are used to assign draft scaffolds to chromosomes and order and orient the draft scaffolds within each chromosome.

As discussed earlier, there are substantial challenges in *de novo* genome assembly, particularly for resolving the gap-less assembly in genomic regions with repeats. Fortunately, long sequencing reads could address these problems, as long sequencing reads can cover the repeats and are less biased to regions with high GC or AT contents. Whether or not to combine short-read data, the assembly software for long reads can be classified into hybrid or long-read-only methods. The hybrid techniques took advantage of the accuracy of short reads to increase the assembly quality and reduce sequencing costs. In contrast, long-read-only methods use only long reads to generate the genome assembly. Long sequencing reads usually have high error rates, and therefore two strategies are used to correct errors in the *de novo* genome assembly, including "correction then assembly" and "assembly then correction". Many genome assembly tools that correct long sequencing reads and then assemble the genome using corrected reads, such as Falcon (Chin *et al.*, 2016), Canu (Koren *et al.*, 2017), MECAT (Xiao *et al.*, 2017), and NECAT (Chen *et al.*, 2021). Whereas other assemblers, such as miniasm (Li, 2016), Flye (Kolmogorov *et al.*, 2019), wtdbg2 (Ruan & Li, 2020), Shasta (Shafin *et al.*, 2020), and Raven (Vaser & Šikić, 2021), assemble the genome using error-prone reads following by correction. Moreover, other methods were also introduced to increase assembly contiguity and accuracy, such as optical mapping (Kronenberg *et al.*, 2018; Udall & Dawe, 2018; Miga *et al.*, 2020).

Haplotype-resolved *de novo* assembly

De novo genome assemblies have traditionally been pseudo-haploid in nature. Haplotype-phased genome assembly provides a complete picture of genomes and their complex genetic variations. However, haplotype-resolved *de novo* assembly is still a challenge, even introduced with long-read sequencing technologies. Most genome assembly tools collapse the different homologous haplotypes into a mosaic consensus. To address this challenge, FALCON and Falcon-Unzip assemble haplotype contigs (haplotigs) using phasing information from heterozygous positions (Chin *et al.*, 2016). It can produce one primary assembly representing a mosaic of homologous haplotypes and one alternate assembly composed of short haplotype-specific contigs for alleles. Another tool, trio binning (Koren *et al.*, 2018), simplifies haplotype assembly by resolving allelic variation before assembly, which uses short reads from two parental genomes. HiCanu tries to keep the contiguity of one parental haplotype and produces Falcon-Unzip-style primary/alternate assemblies (Nurk *et al.*, 2020). Another assembler for PacBio's long high-fidelity (HiFi) reads, hifiasm (Cheng *et al.*, 2021), was developed recently and can generate a well-connected assembly graph and produce better-phased assemblies in practice. Hifiasm performs all-versus-all read overlap alignment and then corrects sequencing errors. After completing three rounds of error correction, it does overlap alignment again and builds a string graph (Myers, 2005), in which a pair of heterozygous alleles will be represented by a "bubble" in the string graph. Like Falcon-Unzip and HiCanu, hifiasm arbitrarily selects one side of each bubble and outputs a primary assembly without additional data. Using hifiasm, the high-quality haplotype-resolved potato genome was assembled based on high-quality long reads and Hi-C data (Sun *et al.*, 2022). By combining advances in long-read assembly and Hi-C-based phasing, DipAsm (Garg *et al.*, 2021) can accurately reconstruct the two haplotypes in a diploid individual using only HiFi reads and Hi-C data, both at ~30-fold coverage (**Figure 18**).

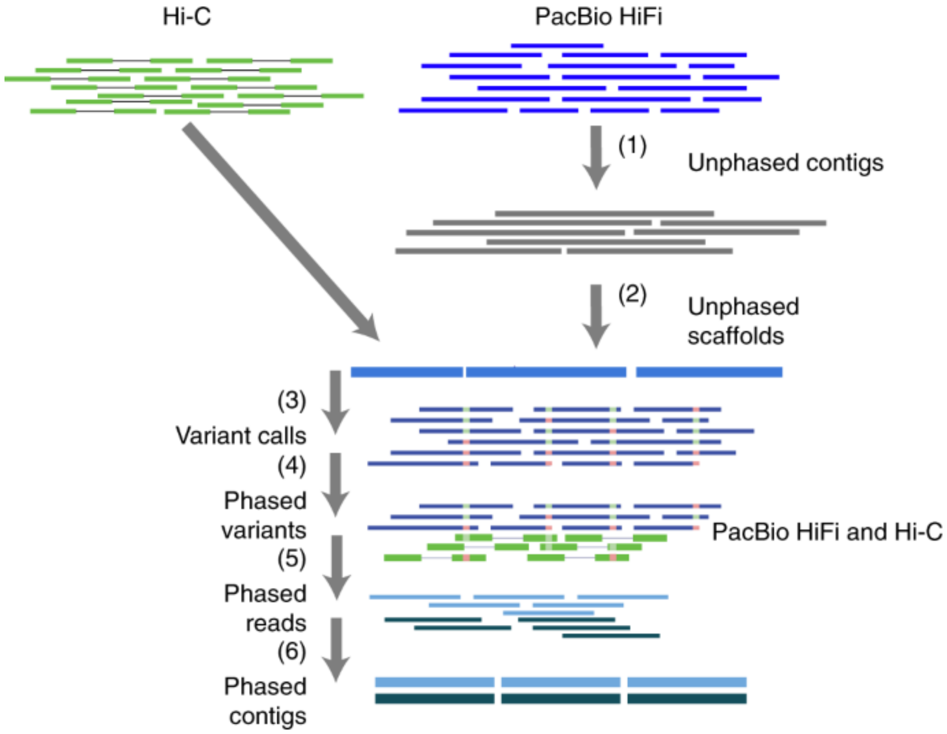


Figure 18. Outline of the phased assembly algorithm DipAsm. Adapted from Garg *et al.* (2021).

2.3.5 Phylogenomic applications

The phylogenetic analysis aims to elucidate the evolutionary history and relationship among a group of organisms. Different methods were proposed to construct phylogenetic trees, including distance-based, maximum parsimony, maximum likelihood, and Bayesian methods. Generally, these methods can be classified into distance-based and character-based. UPGMA (Unweighted Pair Group Method with Arithmetic, Sokal and Michener, 1958) and NJ (Neighbor-joining) (Saitou & Nei, 1987) methods are the representative distance-based methods, which use evolutionary distance matrix. The advantage of the distance-based method is its short calculation time, and thus this method can handle a large amount of data. MEGA (Tamura *et al.*, 2021) is the representative software and is now widely used for inferring phylogenetic trees. Maximum parsimony, maximum likelihood, and Bayesian method are the representative character-based methods that use aligned sequences directly during the tree inference. Maximum parsimony, assuming a common character was derived from a common ancestor, is the origin of character-based methods. In contrast, maximum likelihood uses statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees. Therefore, the calculation time is longer than those of other methods. The Bayesian method is based on posterior probabilities under the estimated best model for inferring a phylogenetic tree. Posterior probabilities are obtained by exploring tree space using Markov chain Monte Carlo (MCMC) algorithms. Maximum likelihood and

Bayesian methods are now widely used because of their implementation of elaborated evolutionary models based on statistical methods. PhyML (Guindon *et al.*, 2010), RAxML/ExaML (Stamatakis *et al.*, 2005; Kozlov *et al.*, 2015), MrBayes (Huelsenbeck & Ronquist, 2001), TOPALi v2(Milne *et al.*, 2009), FastTree (Price *et al.*, 2009) and IQ-TREE 2 (Minh *et al.*, 2020) are the most widely used programs for inferring phylogenetic tree by these methods.

Generally, statistical methods, such as maximum likelihood, can generate more reliable results than distance and parsimony methods (Yang & Rannala, 2012; Whelan & Morrison, 2017). However, they are also computationally more expensive. PhyML, RAxML/ExaML, FastTree, and IQ-TREE are popular fast maximum likelihood-based phylogenetic programs. These tools offer different tradeoffs between the extent of tree space searched and speed in fast phylogenetic inference. Zhou *et al.* (2018) conducted a systematic examination and evaluation of the fast maximum likelihood-based phylogenetic programs on diverse sets of empirical phylogenomic data. The results showed that IQ-TREE has a very appealing performance, as IQ-TREE represents the latest development in fast phylogenetic programs and has implemented a novel data structure to facilitate concatenation analysis (Chernomor *et al.*, 2016). After a phylogenetic tree was inferred, the tree should be evaluated, for example, the validity of the tree shape, evolutionary distance, and the validation of each internal branch. Concerns about reproducibility in phylogenetics has historically been discussed. An investigation of reproducibility in maximum likelihood phylogenetic inference showed that 18.11% IQ-TREE-inferred and 9.34% RAxML-NG-inferred maximum likelihood gene trees are topologically irreproducible when executing two replicates (Shen *et al.*, 2020). Model selection was considered an essential step in the phylogenetic reconstruction process. However, a recent study (Abadi *et al.*, 2019) reported that using the most complex nucleotide substitution model GTR+I+G for all datasets, rather than performing a model selection step, resulted in phylogenies and ancestral sequences as accurate as those obtained when the model selection was performed.

During the early stages of molecular phylogenetics, plant phylogenetic studies relied on a few universal molecular markers, primarily sequences of the chloroplast and nuclear ribosomal DNA. For example, the *rbcL*, *atpB*, *ndhF*, and *matK* genes from plastid genome and internal transcribed spacer (ITS) sequences from nuclear genome are frequently used for inferring the phylogenetic relationships (Alverson *et al.*, 1999; Soltis *et al.*, 2000; Wojciechowski *et al.*, 2004; Shaw *et al.*, 2007). With the advances of sequencing technology, a number of NGS-based methods are developed (Paula, 2021), for example, whole-genome sequencing (WGS), restriction site-associated DNA sequencing (RAD-seq), genotyping-by-sequencing (GBS), multiplexed inter-simple sequence repeats (ISSR) genotyping-by-sequencing (MIG-seq), target sequence capture/hybrid enrichment, amplicon sequencing, RNA-Seq, metabarcoding, metagenomics, and direct DNA shotgun sequencing.

With the available of genomic sequences, phylogenomic approaches are now widely used to resolve species relationships based on hundred low-copy genes or complete plastid gene datasets. The standard method for estimating the phylogeny of species is to calculate alignments for each gene, join these alignments into a super-alignment, and then estimate a tree from the super-alignment. The second strategy is a coalescent-based species tree method by providing a statistically consistent estimation of the actual species tree from unrooted gene trees. However, it is hard to estimate a reliable species tree due to plant genome evolution, such as polyploidy, periods of rapid speciation, extinction, horizontal gene transfer, incomplete lineage sorting, and gene duplication and loss (Yang & Warnow, 2011). For example, the one thousand plant transcriptomes initiative provided a robust phylogenomic framework for examining the evolution of green plants (**Figure 19**). Most inferred species relationships are well supported, but discordance occurred among plastid and nuclear gene trees at a few important nodes.

Due to the uniparental inheritance and absence of recombination, the chloroplast DNA is phylogenetically linear over generations with significantly lower mutation rates than the nuclear DNA. Thus, chloroplast DNA sequences have been used extensively for inferring relationships in plants. During the past decades, phylogenetic analyses based on complete plastomes have achieved significant progress in clarifying the backbone relationships of angiosperms (Davis *et al.*, 2014; Li *et al.*, 2019; Walden *et al.*, 2020; Li *et al.*, 2021; Zhao *et al.*, 2021). For example, a comprehensive plastid phylogenomic study (Li *et al.*, 2019) generated and assembled a large DNA dataset comprising 80 genes from 2,881 plastomes and constructed a phylogenetic tree across angiosperm.

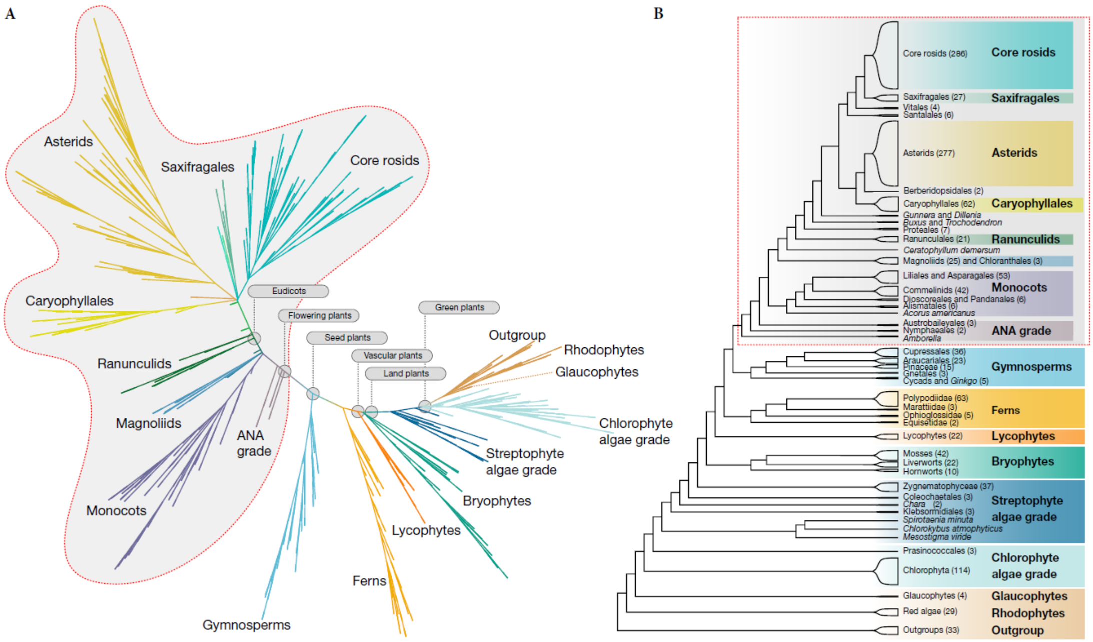


Figure 19. Phylogenetic inferences were based on ASTRAL analysis of 410 single-copy nuclear gene families in green plant (Viridiplantae) species. (A) Phylogram showing internal branch lengths proportional to coalescent units between branching events, (B) Relationships among major clades with red box outlining flowering plant clade. Adapted from Leebens-Mack *et al.* (2019).

2.3.6 Identification of repetitive sequences

Transposable elements (TEs), tandem repeats (TRs) and other repetitive sequences are essential contributors to genome composition. TEs even make up approximately 85% of the genomes of wheat (*Triticum aestivum*) and maize (Schnable *et al.*, 2009; Consortium, 2018). Even with the advance in NGS and long-read sequencing, the identification and annotation of repetitive sequences are still challenging. More practically, the repetitive sequences pose fundamental challenges to genome sequencing, assembly, annotation, and alignment. Several methods were proposed to analyze these repeats, such as similarity-based, signature-based, or *de novo* methods (**Figure 20**) (Bergman & Quesneville, 2007; Goerner-Potvin & Bourque, 2018).

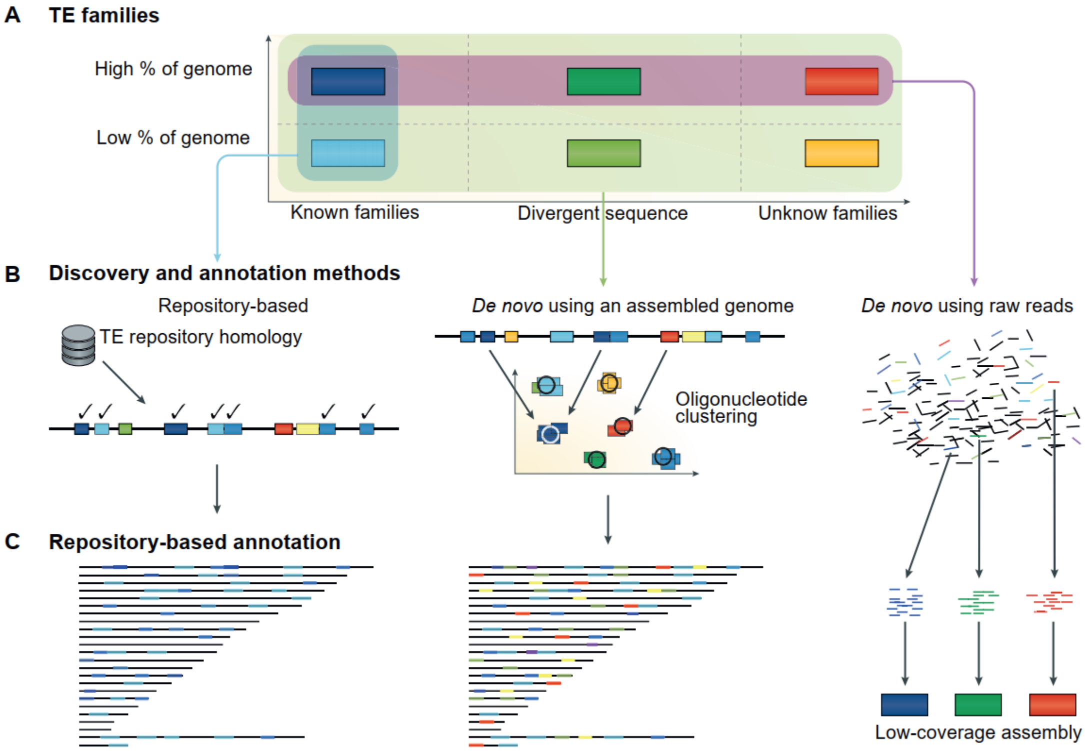


Figure 20. Comparison of TE identification and annotation approaches. Adapted from Goerner-Potvin and Bourque (2018).

Repeats identification and annotation can be performed with or without a genome assembly. Repository-based annotation and *de novo* annotation are two main strategies based on assembled genomes. The idea of repository-based annotation tools is to perform genome-wide searches of repetitive consensus sequences, and their performance is related to the quality and specificity of the sequence databases. The most widely used query tool for the TE annotation is RepeatMasker, which queries against the RepBase and Dfam databases (Bao *et al.*, 2015; Hubley *et al.*, 2016). The censor tool (https://www.girinst.org/censor/index.php) was developed to identify repetitive elements compared to known repeats from RepBase (Kohany *et al.*, 2006). Other tools in this category search for known motifs or genomic structures in plant genomes, including MASiVE, HelitronScanner and LTR Annotator, which can detect Sirevirus, Helitrons and LTRs, respectively (Darzentas *et al.*, 2010; Xiong *et al.*, 2014; Goerner-Potvin & Bourque, 2018). In addition, MGEScan (Lee *et al.*, 2016) can detect non-LTRs and LTRs, and LTRdigest (Steinbiss *et al.*, 2009) can classify LTRs according to internal sequence structure. Moreover, LTRclassifier (Monat *et al.*, 2016), a web server, can classify a set of LTR retrotransposons in their respective superfamily and provide automatically functional annotation of these elements. TransposonUltimate (Riehl et al., 2022), a powerful bundle of three modules for transposon classification, annotation, and detection, was recently developed.

Although the repository-based method is widely used for TE annotation, *de novo* approaches offer the potential to identify novel TE families. Some of the most popular *de novo* annotation tools for assembled genomes have been developed, including RECON (Bao & Eddy, 2002), RepeatScout (Price *et al.*, 2005), RepeatModeler (Flynn *et al.*, 2020), phRAIDER (Schaeffer *et al.*, 2016) and Red (Girgis, 2015). Moreover, TEdenovo of the REPET suite uses more classical multiple alignments and clustering methods to provide comprehensive TE annotation (Flutre *et al.*, 2011). Recently, a comprehensive pipeline, Extensive *de-novo* TE annotator (EDTA), was proposed to generate *de novo* TE libraries that can be used for repeats annotation (Ou *et al.*, 2019). DeepTE classifies TEs using machine learning with good performance in terms of accuracy and sensitivity against other similar programs (Yan *et al.*, 2020).

De novo identification and annotation using raw reads is an alternative method that does not require genome assembly. Tools in this category use low-coverage sequencing data to assemble TEs and other repetitive sequences directly from raw reads. The widely used tool RepeatExplorer (Novák *et al.*, 2010), a graph-based clustering algorithm to identify TEs from sequencing reads, was published in 2010, and RepeatExplorer was updated with extended features and deployed on a Galaxy server later (Novák *et al.*, 2013; Novák *et al.*, 2020). Following the same principle, dnaPipeTE (Goubert *et al.*, 2015) generates quantitative annotation using the Trinity assembler from NGS reads. Tedna (Zytnicki *et al.*, 2014), RepARK (Koch *et al.*, 2014), REPdenovo (Chu *et al.*, 2016), and RepAHR (Liao *et al.*, 2020) are similar tools, which assemble TEs from raw reads. Replong (Guo *et al.*, 2018) and LongRepMarker (Liao *et al.*, 2019) were recently published as the long-read *de novo* TE annotation tools. Using LongRepMarker, the most comprehensive multi-species repeats database, msRepDB, was constructed, covering >80 000 species, containing more complete repeat families than RepBase and Dfam databases (Liao *et al.*, 2022). Together, these tools provide an excellent opportunity to annotate repetitive elements in newly sequenced species without a reference genome.

De novo identification of tandem repeats can be carried out by the program Tandem Repeat Finder (TRF) without specifying either the pattern or pattern size (Benson, 1999). Recently, a novel computational pipeline, tandem repeat analyzer (TAREAN), was developed to detect satellite repeats and reconstruct repeat monomers directly from short raw reads (Novák *et al.*, 2017).

3 AIMS OF THE THESIS

I Genome analysis and nuclear organization in the meadowfoam family (Limnanthaceae, Brassicales)

The first aim of the thesis was: (i) to reconstruct phylogenetic relationships of Limnanthaceae based on whole-chloroplast, rDNA and repetitive sequences, (ii) to characterize and to compare repeatomes of Limnanthaceae species and (iii) using the *de novo* identified repeats to analyze interphase chromosome organization in Limnanthaceae species by three-dimensional fluorescence *in situ* hybridization (3D FISH).

II Genome diploidization associates with cladogenesis, trait disparity and plastid gene evolution in the tribe Microlepidieae (Brassicaceae)

The second aim of the thesis was: (i) to understand the reticulate phylogenomic patterns and differently phased genome diploidization within the tribe Microlepidieae (Brassicaceae) and (ii) to evaluate the extent of morphological convergence and disparity, and plastid–nuclear coevolution during post-polyploid genome diploidization and cladogenesis.

III Evolutionary history and function of the KNL2 in plants

The third aim of the thesis was: (i) to reconstruct the evolutionary history of the *KNL2* gene in the plant kingdom; and (ii) to clarify the KNL2 function and its role in CENH3 deposition and kinetochore assembly.

4 RESULTS AND DISCUSSION

The results and discussion are presented as short commentaries of the published peer-reviewed articles.

4.1 Genome analysis and nuclear organization in Limnanthaceae

The meadowfoam family (Limnanthaceae, Brassicales) harbors two genera and eight species. Whereas *Limnanthes* species have five pairs of chromosomes, similar to the number of Arabidopsis chromosomes, chromosome sizes in *Limnanthes* are much larger. While it is unclear whether the At-β WGD occurred before or after the divergence of Setchellanthaceae, the origin of Limnanthaceae likely post-dated the At-β WGD and the five-chromosome genomes of today's Limnanthaceae species might arise by descending dysploidy that occurred during post-At-β genome diploidization.

In this project, we performed low-coverage whole genome sequencing, including short and long reads, in four *Limnanthes* (sub)species and *F. proserpinacoides* (Zuo *et al.*, 2022b). We assembled the complete chloroplast (cp) genome and nuclear 35S rDNA sequence of the five Limnanthaceae accessions. All Limnanthaceae cp genomes exhibited the typical quadripartite structure, consisting of a pair of inverted repeat (IR) regions separated by a large single-copy region (LSC) and a small single-copy (SSC) region. The Limnanthaceae cp genome contains 112 unique genes, including 78 protein-coding genes (PCGs), 30 tRNAs, and four rRNAs. The cp genome of most land plants consists of two structural haplotypes that differ only in the orientation of their SSC sequences. Using ONT long reads, we estimated that the two structural haplotypes occurred with approximately equal frequencies in the Limnanthaceae species.

Based on 72 protein-coding genes (PCGs) in 22 plastomes of Limnanthaceae and other Brassicales species, we compiled a gap-free alignment matrix with 43 263 columns, of which 4 773 were parsimony informative. Both the maximum likelihood (ML) and the Bayesian inference (BI) trees (BS = 100 and PP = 1) confirmed the within-family split corresponding to the genera *Floerkea* and *Limnanthes*. In the genus *Limnanthes*, there were two well-supported clades (BS = 100 and PP = 1): (i) *L. alba* and *L. floccosa* and (ii) *L. douglasii*. The two clades corresponded to the infrageneric sections *Inflexae* and *Limnanthes* (formerly *Reflexae*). The rDNA-based ML and BI phylogenetic trees (BS = 100; PP = 1) had congruent topology with plastomes. The fully resolved backbone phylogeny of Limnanthaceae provides the basis for future comparative studies.

Flow cytometric analysis revealed that genome size varied 1.4-fold, from 1 516 Mb in *F. proserpinacoides* to 2 102 Mb in *L. douglasii*. Chromosome counting in somatic tissues of young anthers confirmed five pairs of (sub)metacentric chromosomes ($2n = 10$) in all analyzed Limnanthaceae taxa. Although Limnanthaceae and Arabidopsis (135 Mb) have the same chromosome numbers, their size and structure differ significantly. The average chromosome size (genome size/haploid chromosome number) is over 300 Mb (340–420 Mb) in the Limnanthaceae species, while it is only 32 Mb in Arabidopsis. While in Arabidopsis, most of the repetitive sequences are located in the heterochromatic pericentromeric regions, in Limnanthaceae the repeats are distributed almost evenly over the >300-Mb-long chromosomes. We observed that there is no strong eu-/heterochromatin boundary and heterochromatin is equally distributed throughout chromosomes of Limnanthaceae species.

The identified repetitive sequences accounted for 58.12–66.22% of the Limnanthaceae genomes. In all repeatomes, long terminal repeat (LTR) retrotransposons accounted for the majority of repeats, ranging from 21.04% in *F. proserpinacoides* to 24.59% in *L. douglasii*. The genome proportion of Ty3/*gypsy* elements ranged from 14.39% in *F. proserpinacoides* to 20.10% in *L. floccosa* subsp. *grandiflora*, whereas Ty1/*copia* retrotransposons were three to five times less abundant than Ty3/*gypsy* elements. The identified tandem repeats constituted less than 0.5% of Limnanthaceae nuclear genomes. Seven satellite repeats were identified in Limnanthaceae taxa with very low abundances ranging from 0.03% in *L. alba* to 0.27% in *L. floccosa* subsp. *bellingeriana*. In Brassicales, genome sizes range from 135 Mb to 4.6 Gb (Lysak, 2018) and are primarily determined by the proportion of non-coding and repetitive sequences (Elliott & Gregory, 2015; Hloušková *et al.*, 2019). In the absence of evidence for a family-specific WGD, the non-coding DNAs of Limnanthaceae genomes must have originated either from the At-β WGD or from the proliferation of transposable elements after the genome duplication but most likely before the *Floerkea–Limnanthes* divergence. In either case, selective purging of TEs and/or suppression of their activity may have been less effective in Limnanthaceae. Since no significant shifts in TE abundance were associated with WGDs in Brassicales (Beric *et al.*, 2021), genome obesity in Limnanthaceae is likely due to the proliferation of TEs rather than the At-β genome duplication.

The LiFlo-TR34 repeat occupied all centromeres and two subtelomeric regions of one chromosome pair in subsp. *bellingeriana* and subsp. *grandiflora* of *L. floccosa*. Carnoy's fixed interphase nuclei isolated from young anthers were hybridized with telomeric, centromeric LiFlo-TR34 and 35S rDNA probes. Polarized (Rabl-like) positioning of centromeres and telomeres at the opposite nuclear poles was observed in 83% and 86% of nuclei in subsp. *bellingeriana* and subsp. *grandiflora* of *L. floccosa*, respectively. Telomeric signals were frequently clustered with 35S rDNA loci (nucleolus), whereas centromeres were positioned within the more heterochromatic opposite pole. To further analyze the spatial arrangement of centromeres and telomeres in *L. floccosa* subsp. *bellingeriana*, their distribution was further investigated by 3D FISH using paraformaldehyde-fixed interphase nuclei isolated from three different tissues (root tips, stem leaves and petals). In most leaf nuclei, 3D FISH showed that centromeres were located at one nuclear pole, whereas telomeres and 35S rDNA (nucleolus) were found at the opposite pole (Rabl-like configuration, 81%).

Among the Brassicales, chromatin organization in interphase nuclei has been analyzed only in the Brassicaceae family, with Arabidopsis being the most extensively studied species (Fransz *et al.*, 2002; Pecinka *et al.*, 2004; Pontvianne & Grob, 2020; Shan *et al.*, 2021). In the small Arabidopsis genome, telomeres within interphase nuclei generally associate with the nucleolus, while centromeres are positioned peripherally at the nuclear membrane. In Brassicaceae species with large nuclear genomes (2 600–4 300 Mb) and a small number of chromosomes (n = 6, 7), the spatial arrangement of centromeres and telomeres resembles the Rabl model, or they are scattered in the nuclear interior. In *Limnanthes*, the predominant nuclear phenotype resembles the polarized Rabl configuration, in which the centromeres are usually located at one nuclear pole and the telomeres, together with the nucleolus (or nucleoli), at the opposite pole. Because Rabl organization resembles chromosome configuration in the mitotic anaphase, centromere–telomere polarization in Limnanthaceae species nuclei could be mechanistically interpreted as an effective arrangement of long metacentric (V-shaped) chromosomes within the limited nuclear space, possibly reducing topological entanglement of chromatin fibers (Pouokam *et al.*, 2019). Indeed, chromosome length, not just genome size (Dong & Jiang, 1998), maybe a more important factor in determining the Rabl configuration of interphase chromosomes (Saunders & Houben, 2001; Shan *et al.*, 2021).

4.2 Genome diploidization in the tribe Microlepidieae (Brassicaceae)

Hybridization and polyploidization (or WGD) frequently accompanied the diversification and speciation in plants. In Brassicaceae, more than a dozen genus- and clade-specific mesopolyploid WGDs, postdating the family-specific paleotetraploid (At-α) WGD, were identified. The monophyletic tribe Microlepidieae has descended from a common allotetraploid genome (n = 15) formed by an intertribal cross between parental species closely related to the extant tribes Crucihimalayeae (n = 8) and Smelowskieae (n = 7) during the Late Miocene. The post-polyploid diploidization and diversification in the Microlepidieae did not proceed with equal intensity. In addition, the widespread convergent evolution of morphological characters used for the delimitation of genera and species in the Microlepidieae.

In this project (Zuo *et al.*, 2022a), we retrieved the Microlepidieae as a monophyletic group sister to the tribe Crucihimalayeae and inferred four strongly supported intra-tribal clades based on 76 PCGs in 60 plastomes. According to the plastome phylogeny and two secondary calibration points, the split between Microlepidieae and Crucihimalayeae was dated 10.46 million years ago during the Late Miocene (Tortonian). Clade A represents the previously defined crown-group genera, including two *Arabidella* species (*A. eremigena* and *A. procumbens*) and *Menkea crassa*. Clade B, consisting of four *Arabidella* species (*A. filifolia*, *A. glaucescens*, *A. nasturtium*, and *A. trisecta*) and *Irenepharsus magicus* are sisters to the crown-group clade. Clades C and D appear as successive sisters to clades A + B, whereby clade C harbors only *Pachycladon* species, and clade D includes *A. chrysodema* and two *Menkea* species (*M. sphaerocarpa* and *M. villosula*). The taxonomic limits of *Arabidella*, *Cuphonotus*, and *Lemphoria* were revised based on recent cytogenomic and molecular phylogenetic findings (Lysak *et al.*, 2022). As a result, *Lemphoria* was re-established to include two species previously placed in *Cuphonotus* and two in *Arabidella* (*A. eremigena* and *A. procumbens*). *Lemphoria queenslandica* was described as a new species, and the new combinations *L. andraeana*, *L. eremigena*, *L. humistrata*, and *L. procumbens* were proposed. Keys to distinguish *Arabidella* and *Lemphoria* species and an expanded generic description of *Lemphoria* were provided (Lysak *et al.*, 2022).

With expanded taxon sampling, we have obtained robust phylogenies of the tribe Microlepidieae, allowing for phylogenetically informed analysis of post-polyploid genome diploidization and cladogenesis. Whereas mesotetraploid genomes of the early branching *A. chrysodema*/*Menkea* clade and the crown group have been extensively diploidized, *Arabidella* and *Pachycladon* genomes are slowly diploidizing. We observed significantly higher synonymous substitution rates in plastomes of the fast-diploidizing Microlepidieae clades than in less diploidized genomes of *Arabidella*, *Irenepharsus*, and *Pachycladon*. Our results demonstrated that plastid genes may co-evolve with the nuclear genomes undergoing slow or fast post-polyploid diploidization. The variation in morphological characters among the diploidizing genomes and species is largely controlled by gene expression changes. These processes may have several possible outcomes, such as morphological disparity despite the shared ancestry or morphological convergence despite independent diploidization of polyploid genomes. Morphological convergence or disparity may hamper retrieving true phylogenetic relationships among species of diploidizing polyploid lineages. Morphological convergence was frequently observed across Brassicaceae tribes. In Microlepidieae, highly supported phylogenetic analyses uncovered several instances of convergent evolution of some morphological characters and, conversely, considerable intra-tribal phenotypic disparity. We detected higher disparity in the crown-group genera, especially in *Lemphoria* and *Stenopetalum*. In contrast, *Arabidella* and *Pachycladon* displayed lower mean disparity than genera with the same or even smaller number of species.

Our Bayesian analysis of macroevolutionary mixtures (BAMM) analyses revealed a continuous decrease in diversification rates after the initial divergence of Microlepidieae c. 10 Mya. Our BAMM analyses failed to detect any rate shifts during the diversification of Microlepidieae. The lack of shifts in speciation rate across the Microlepidieae phylogeny supported the notion that diversification was largely decoupled from WGDs and/or diploidization. In addition, our binary state speciation and extinction (BiSSE) analyses pinpointed higher speciation rates in perennials (0.468 species/Myr) than annuals (0.107 species/Myr), with a stronger tendency of transition from perenniality to annuality than in the opposite direction. In Microlepidieae, higher speciation rates in *Arabidella* and *Pachycladon* could be tentatively linked to their stable genome structures, which may allow for frequent homoploid hybridization. Ancestral state reconstruction inferred annuality, with a likelihood of 78.4%, to be the most likely ancestral life form in Microlepidieae. To detect probable hybridization events, a tribe-wide analysis of 5S rDNA clustering graphs was applied to the genus *Arabidella* (clade B). Our analysis showed that all three populations of *A. nasturtium* represented presumably homoploid interspecies hybrids and suggested that the entire species could have a hybridogenous origin. These findings supported the view that the genus *Arabidella* is a polyploid complex of closely related mesotetraploid ($2n = 24$) and neomesotetraploid ($2n = 48$) genomes.

Altogether, we provided clear phylogenomic evidence that differently paced postpolyploid diploidization was associated with (1) intratribal cladogenesis, (2) morphological disparity, (3) selection pressure on genes involved in cytonuclear interaction, and (4) life-form transitions.

4.3 Evolutionary history and function of the *KNL2* in plants

The KNL2 plays a crucial role in new CENH3 deposition. The KNL2 protein contains a conserved module designated as SANTA due to its association with the SANT domain. Most metazoan genomes have only one *KNL2* gene with the SANTA domain, while in Arabidopsis two *KNL2* copies (At5g02520 and At1g58210) were identified. The KNL2 protein (At5g02520) contains the SANTA domain and a conserved CENPC-like motif (CENPC-k) at its C terminus that is required for the centromeric localization of KNL2. The At1g58210 gene encodes a protein of 281 amino acids, including only the SANTA domain. We designated it as *βKNL2* and previously characterized *KNL2* as *αKNL2*.

In this project (Zuo *et al.*, 2022c), we retrieved two *KNL2* copies from water ferns, eudicots, and grasses, whereas only one *KNL2* copy was found in bryophytes and gymnosperms. To understand the evolutionary history of the *KNL2* gene across the plant kingdom, we reconstructed their phylogenetic relationships. The topology of the maximum likelihood (ML) tree showed that KNL2 proteins cluster into two branches in three plant clades—heterosporous water ferns (Salviniaceae), eudicots, and grasses (Poaceae) - indicating ancient gene duplications. These KNL2 proteins present conserved features: the N-terminus contains the conserved SANTA domain in all KNL2 proteins, whereas only the αKNL2-type C-terminus possesses the CENPC-k motif. We identified positive selection sites in and near the SANTA domain of KNL2 in the analyzed Brassicaceae species, similar to what has been previously reported for CENH3 (Talbert *et al.*, 2002) and CENP-C (Talbert *et al.*, 2004). However, the mechanisms of adaptively evolving regions remain to be elucidated.

The CENPC-k motif was found in KNL2 of diverse eukaryotes, including non-mammalian vertebrates and plants. In eudicots, the conserved CENPC-k motif was present in the αKNL2 clade, but was absent from βKNL2. Similarly, in most grass species the CENPC-k motif was conserved in the γKNL2 clade, while the δKNL2 clade did not have the motif. However, we found a RRLRSGKV/I motif in the δKNL2 clade possibly related to the beginning of the CENPC-k motif (KRSRSGRV/LLVSPLEFW). It remains to be elucidated whether KNL2 variants with the truncated CENPC-k motif can target CENH3 nucleosomes. Among all grass species with sequenced genomes, maize represents an exception since it has only one *KNL2* gene, which belongs to the δKNL2 clade with the truncated CENPC-k and has no γKNL2 protein variant. Interestingly, in sorghum, closely related to maize, the γKNL2 protein can be identified. This suggested that maize may have evolved a different mechanism for CENH3 deposition compared with other grasses.

We demonstrated that βKNL2 colocalizes with CENH3 at centromeres, despite lacking a CENPC-k motif (Zuo *et al.*, 2022c). Due to the lack of the CENPC-k motif in βKNL2, we proposed that in Arabidopsis βKNL2 might localize to centromeres by binding to CENP-C through the SANTA domain as it was shown for *Xenopus* (French & Straight, 2019), or through the conserved N-terminal motif located upstream of the SANTA domain similar to what was previously described in human (Stellfox *et al.*, 2016) or through both of these regions. In general, both variants of Arabidopsis KNL2 showed a similar localization pattern during interphase. However, in contrast to αKNL2, βKNL2 can be detected on chromosomes during metaphase and early anaphase. The centromeric location of βKNL2 suggests that *βKNL2* may partially compensate for the loss of *αKNL2* in the corresponding Arabidopsis mutant, which showed only reduced, but not completely abolished CENH3 loading, which would be lethal (Lermontova *et al.*, 2013). Homozygous T-DNA insertions for βKNL2 resulted in plant death at the seedling stage and probably because of reduced root development. As reciprocal crosses of *βknl2* mutants with the wild-type (WT) resulted in normal seed development in both directions, we hypothesized that the *βKNL2* null mutations do not affect gametes or fertilization processes, but rather postzygotic cell divisions. In support of this hypothesis, ploidy analysis of young seedlings revealed that in contrast to the WT with distinct 2C and 4C peaks, homozygous mutants showed a shift toward endopolyploidization, potentially a consequence of disrupted cell divisions. Thus, our data strongly suggested the involvement of βKNL2 in CENH3 loading.

Double haploid production is the most effective way of creating true-breeding lines in a single generation. In Arabidopsis, haploid induction via mutation of the CENH3 has been shown when outcrossed to wild-type. Here we reported that a mutant of the CENH3 assembly factor KNL2 could be used as a haploid inducer (Ahmadli *et al.*, 2022). We elucidated that the short temperature stress of the *knl2* mutant increased the efficiency of haploid induction from 1 to 10%. Moreover, we demonstrated that a point mutation in the CENPC-k motif of KNL2 is sufficient to generate haploid inducing lines, suggesting that haploid inducing lines in crops can be identified in a naturally occurring or chemically induced mutant population, avoiding the GMO approach at any stage.

Taken together, our results suggested that the *KNL2* gene underwent ancient duplication events with the core function of CENH3 deposition to define the centromere region. We demonstrated that *KNL2* genes exist in two copies in eudicots (*α*, *βKNL2*) and monocots (*γ*, *δKNL2*). The conserved gene structure and expression patterns of *α/γKNL2* in both eudicots and monocots suggest that *α/γKNL2* mutations could be used to develop in vivo haploid induction systems in different crop plants. Similarly, the newly identified *βKNL2* may become the subject of manipulations to obtain haploids both in Arabidopsis and crop species.

5 CONCLUSIONS

The results of this thesis were summarized in three main parts together with five publications. The first part addressed the knowledge gap in the genome evolution of the meadowfoam family (Limnanthaceae) (Zuo *et al.*, 2022b). Using low-coverage whole genome sequencing data, we re-examined phylogenetic relationships and characterized the repeatomes of Limnanthaceae genomes. Phylogenies based on complete chloroplast and 35S rDNA sequences corroborated the sister relationship between *Floerkea* and *Limnanthes* and two major clades in the latter genus. The genome size of Limnanthaceae species ranges from 1.5 to 2.1 Gb, apparently due to the large increase in DNA repeats, which constitute 60–70% of their genomes. Repeatomes are dominated by long terminal repeat retrotransposons, while tandem repeats represent less than 0.5% of the genomes. A three-dimensional fluorescence *in situ* hybridization analysis demonstrated that the five chromosome pairs in interphase nuclei of *Limnanthes* species adopt the Rabl-like configuration. Taking together, we examined the Limnanthaceae genomes as a potential model system for 3D genome organization.

The second part focused on patterns of genome diploidization in the tribe Microlepidieae, Brassicaceae (Zuo *et al.*, 2022a). We analyzed phylogenetic relationships in this tribe using complete chloroplast sequences, entire 35S rDNA units, and abundant repetitive sequences. The four recovered intra-tribal clades mirror the varied diploidization of Microlepidieae genomes, suggesting that the intrinsic genomic features underlying the extent of diploidization are shared among genera and species within one clade. Nevertheless, even congeneric species may exert considerable morphological disparity (e.g. in fruit shape), whereas some species within different clades experience extensive morphological convergence despite the different pace of their genome diploidization. We showed that faster genome diploidization is positively associated with mean morphological disparity and evolution of chloroplast genes (plastid–nuclear genome coevolution). Higher speciation rates in perennials than in annual species were observed. The taxonomic limits of *Arabidella*, *Cuphonotus*, and *Lemphoria* (Microlepidieae, Brassicaceae) are revised based on morphology and molecular phylogenetic findings (Lysak *et al.*, 2022). Altogether, our results confirmed the potential of Microlepidieae as a promising subject for the analysis of post-polyploid genome diploidization in Brassicaceae.

The third part of this thesis focused on the evolutionary history of *KNL2* and its function in kinetochore assembly (Zuo *et al.*, 2022c). Our results demonstrated that the *KNL2* gene in plants underwent three independent ancient duplications in ferns, grasses, and eudicots. Additionally, we showed that previously unclassified *KNL2* genes could be divided into two clades *αKNL2* and *βKNL2* in eudicots and *γKNL2* and *δKNL2* in grasses, respectively. KNL2s of all clades encode the conserved SANTA domain, but only the αKNL2 and γKNL2 groups additionally encode the CENPC-k motif. The confirmed centromeric localization of βKNL2 and mutant analysis suggest that it participates in the loading of new CENH3, similarly to αKNL2. A high rate of seed abortion was found in heterozygous *βknl2* plants, and the germinated homozygous mutants did not develop beyond the seedling stage. Moreover, we reported that a mutant of the CENH3 assembly factor KNL2 could be used as a haploid inducer, and thus the newly identified βKNL2 may become the subject of manipulations to obtain haploids both in Arabidopsis and crop species (Ahmadli *et al.*, 2022). Taken together, our study provided a new understanding of the evolutionary diversification of the *KNL2*, and suggested that the duplicated *KNL2* genes are involved in centromere and/or kinetochore assembly.

6 REFERENCES

Abadi S, Azouri D, Pupko T, Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications* **10**(1).

Agren JA. 2014. Evolutionary transitions in individuality: insights from transposable elements. *Trends in Ecology & Evolution* **29**(2): 90-96.

Ahmadli U, Kalidass M, Khaitova LC, Fuchs J, Cuacos M, Demidov D, Zuo S, Pecinkova J, Mascher M, Ingouff M, et al. 2022. High temperature increases centromere-mediated genome elimination frequency in *Arabidopsis* deficient in cenH3 or its assembly factor KNL2. *bioRxiv*.

Aksenova AY, Mirkin SM. 2019. At the beginning of the end and in the middle of the beginning: structure and maintenance of telomeric DNA repeats and interstitial telomeric sequences. *Genes* **10**(2).

Albert VA, Barbazuk WB, Depamphilis CW, Der JP, Leebens-Mack J, Ma H, Palmer JD, Rounsley S, Sankoff D, Schuster SC, et al. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* **342**(6165): 1467-1469.

Allen JO, Fauron CM, Minx P, Roark L, Oddiraju S, Lin GN, Meyer L, Sun H, Kim K, Wang CY, et al. 2007. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. *Genetics* **177**(2): 1173-1192.

Alverson WS, Whitlock BA, Nyffeler R, Bayer C, Baum DA. 1999. Phylogeny of the core Malvales: evidence from *ndhF* sequence data. *American Journal of Botany* **86**(10): 1474-1486.

Ambrozová K, Mandáková T, Bures P, Neumann P, Leitch IJ, Koblízková A, Macas J, Lysak MA. 2011. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria lilies*. *Annals of Botany* **107**(2): 255-268.

Ananiev EV, Phillips RL, Rines HW. 1998. Chromosome-specific molecular organization of maize (*Zea mays L.*) centromeric regions. *Proceedings of the National Academy of Sciences of the United States of America* **95**(22): 13073-13078.

Archibald JM. 2009. The puzzle of plastid evolution. *Current Biology* **19**(2): 81-88.

Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA. 2009. Diversity of the *Arabidopsis* mitochondrial genome occurs via nuclear-controlled recombination activity. *Genetics* **183**(4): 1261-1268.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Research* **11**(6): 1005-1017.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**(5): 455-477.

Bao WD, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**(1): 1-6.

Bao Y, Wendel JF, Ge S. 2010. Multiple patterns of rDNA evolution following polyploidy in *Oryza*. *Molecular Phylogenetics and Evolution* **55**(1): 136-142.

Bao ZR, Eddy SR. 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research* **12**(8): 1269-1276.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference. *Nature Plants* **6**(11): 1389-1389.

Bennett MD, Leitch IJ. 2005. Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Annals of Botany* **95**(1): 45-90.

Bennett MD, Leitch IJ. 2011. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Annals of Botany* **107**(3): 467-590.

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology* **65**: 505-530.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**(2): 573-580.

Bergman CM, Quesneville H. 2007. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics* **8**(6): 382-392.

Bergmann JH, Rodriguez MG, Martins NMC, Kimura H, Kelly DA, Masumoto H, Larionov V, Jansen LET, Earnshaw WC. 2011. Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore. *Embo Journal* **30**(2): 328-340.

Beric A, Mabry ME, Harkess AE, Brose J, Schranz ME, Conant GC, Edger PP, Meyers BC, Pires JC. 2021. Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales). *G3-Genes Genomes Genetics* **11**(7).

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology* **33**(6): 623-628.

Biscotti MA, Olmo E, Heslop-Harrison JS. 2015. Repetitive DNA in eukaryotic genomes. *Chromosome Research* **23**(3): 415-420.

Bloom KS. 2014. Centromeric heterochromatin: the primordial segregation machine. *Annual review of genetics* **48**: 457-484.

Bobkov GOM, Gilbert N, Heun P. 2018. Centromere transcription allows CENP-A to transit from chromatin association to stable incorporation. *Journal of Cell Biology* **217**(6): 1957-1972.

Bolzan AD. 2012. Chromosomal aberrations involving telomeres and interstitial telomeric sequences. *Mutagenesis* **27**(1): 1-15.

Borowska-Zuchowska N, Kovarik A, Robaszkiewicz E, Tuna M, Tuna GS, Gordon S, Vogel JP, Hasterok R. 2020. The fate of 35S rRNA genes in the allotetraploid grass *Brachypodium hybridum*. *Plant Journal* **103**(5): 1810-1825.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome Biology* **19**(3).

Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**(6930): 433-438.

Carman JG. 1997. Asynchronous expression of duplicate genes in angiosperms may cause apomixis, bispory, tetraspory, and polyembryony. *Biological Journal of the Linnean Society* **61**(1): 51-94.

Chapman JA, Ho I, Sunkara S, Luo SJ, Schroth GP, Rokhsar DS. 2011. Meraculous: *de novo* genome assembly with short paired-end reads. *PLoS One* **6**(8).

Chen Y, Nie F, Xie SQ, Zheng YF, Dai Q, Bray T, Wang YX, Xing JF, Huang ZJ, Wang DP, et al. 2021. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications* **12**(1).

Cheng HY, Concepcion GT, Feng XW, Zhang HW, Li H. 2021. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**(2): 170-174.

Cheng ZK, Dong FG, Langdon T, Shu OY, Buell CR, Gu MH, Blattner FR, Jiang JM. 2002. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**(8): 1691-1704.

Chernomor O, Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology* **65**(6): 997-1008.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**(6): 563-567.

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**(12): 1050-1054.

Chu C, Nielsen R, Wu YF. 2016. REPdenovo: inferring *de novo* repeat motifs from short sequence reads. *PLoS One* **11**(3).

Chu TC, Lu CH, Liu TL, Lee GC, Li WH, Shih ACC. 2013. Assembler for *de novo* assembly of large genomes. *Proceedings of the National Academy of Sciences of the United States of America* **110**(36): 3417-3424.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* **18**(2): 71-86.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4**(4): 265-270.

Cleveland DW, Mao YH, Sullivan KF. 2003. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**(4): 407-421.

Collins FS, Lander ES, Rogers J, Waterston RH, Conso IHGS. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.

Compeau PEC, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* **29**(11): 987-991.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* **19**(1): 91-98.

Consortium IWGS. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**(6403): eaar7191.

Daniell H, Lin CS, Yu M, Chang WJ. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* **17**(2).

Darzentas N, Bousios A, Apostolidou V, Tsaftaris AS. 2010. MASiVE: mapping and analysis of SireVirus elements in plant genome sequences. *Bioinformatics* **26**(19): 2452-2454.

Davis CC, Xi ZX, Mathews S. 2014. Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. *Bmc Biology* **12**(1).

Dawe RK, Reed LM, Yu H-G, Muszynski MG, Hiatt EN. 1999. A maize homolog of mammalian CENPC is a constitutive component of the inner kinetochore. *Plant Cell* **11**(7): 1227-1238.

Dong FG, Jiang JM. 1998. Non-Rabl patterns of centromere and telomere distribution in the interphase nuclei of plant cells. *Chromosome Research* **6**(7): 551-558.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**(5961): 78-81.

Du Y, Dawe RK. 2007. Maize NDC80 is a constitutive feature of the central kinetochore. *Chromosome Research* **15**(6): 767-775.

Edger PP, Hall JC, Harkess A, Tang M, Coombs J, Mohammadin S, Schranz ME, Xiong ZY, Leebens-Mack J, Meyers BC, et al. 2018. Brassicales phylogeny inferred from 72 plastid genes: a reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses. *American Journal of Botany* **105**(3): 463-469.

Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175**(2): 477-485.

El-Metwally S, Hamza T, Zakaria M, Helmy M. 2013. Next-generation sequence assembly: four stages of data processing and computational challenges. *Plos Computational Biology* **9**(12).

Elliott TA, Gregory TR. 2015. What's in a genome? the C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B-Biological Sciences* **370**(1678).

Fajkus P, Peška V, Sitová Z, Fulnečková J, Dvořáčková M, Gogela R, Sýkorová E, Hapala J, Fajkus J. 2016. Allium telomeres unmasked: the unusual telomeric sequence (CTCGGTTATGGG)(n) is synthesized by telomerase. *Plant Journal* **85**(3): 337-347.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics* **3**(5): 329-341.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends in Genetics* **5**(4): 103-107.

Fleischmann A, Michael TP, Rivadavia F, Sousa A, Wang WQ, Temsch E, Greilhuber J, Muller KF, Heubl G. 2014. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany* **114**(8): 1651-1663.

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* **6**(1).

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**(17): 9451-9457.

Fransz P, Jong JH, Lysak M, Castiglione MR, Schubert I. 2002. Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. *Proceedings of the National Academy of Sciences of the United States of America* **99**(22): 14584-14589.

Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology* **15**(2): 131-139.

French BT, Straight AF. 2019. CDK phosphorylation of *Xenopus laevis* M18BP1 promotes its metaphase centromere localization. *Embo Journal* **38**(4): e100093.

French BT, Westhorpe FG, Limouse C, Straight AF. 2017. *Xenopus laevis* M18BP1 directly binds existing CENP-A nucleosomes to promote centromeric chromatin assembly. *Developmental Cell* **42**(2): 190-199.

Fuchs J, Brandes A, Schubert I. 1995. Telomere sequence localization and karyotype evolution in higher plants. *Plant Systematics and Evolution* **196**(3-4): 227-241.

Fujita Y, Hayashi T, Kiyomitsu T, Toyoda Y, Kokubu A, Obuse C, Yanagida M. 2007. Priming of centromere for CENP-A recruitment by human hMis18 alpha, hMis18 beta, and M18BP1. *Developmental Cell* **12**(1): 17-30.

Garcia S, Kovarik A. 2013. Dancing together and separate again: gymnosperms exhibit frequent changes of fundamental 5S and 35S rRNA gene (rDNA) organisation. *Heredity* **111**(1): 23-33.

Garcia S, Kovarik A, Leitch AR, Garnatje T. 2017. Cytogenetic features of rRNA genes across land plants: analysis of the plant rDNA database. *Plant Journal* **89**(5): 1020-1030.

Garcia S, Lim KY, Chester M, Garnatje T, Pellicer J, Valles J, Leitch AR, Kovarik A. 2009. Linkage of 35S and 5S rRNA genes in *Artemisia* (family Asteraceae): first evidence from angiosperms. *Chromosoma* **118**(1): 85-97.

Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nature Biotechnology* **39**(3): 309-312.

Garrido-Ramos MA. 2015. Satellite DNA in plants: more than just rubbish. *Cytogenetic and Genome Research* **146**(2): 153-170.

Garrido-Ramos MA. 2017. Satellite DNA: an evolving topic. *Genes* **8**(9).

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. 2014. Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution* **31**(2): 448-454.

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**(1): 17-27.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics* **44**: 445-477.

Girgis HZ. 2015. Red: an intelligent, rapid, accurate tool for detecting repeats *de-novo* on the genomic scale. *Bmc Bioinformatics* **16**(1).

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **108**(4): 1513-1518.

Goerner-Potvin P, Bourque G. 2018. Computational tools to unmask transposable elements. *Nature Reviews Genetics* **19**(11): 688-704.

Gong ZY, Wu YF, Koblizkova A, Torres GA, Wang K, Iovene M, Neumann P, Zhang WL, Novak P, Buell CR, et al. 2012. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell* **24**(9): 3559-3574.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**(6): 333-351.

Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. 2015. *De novo* assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biology and Evolution* **7**(4): 1192-1205.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. *Plant Journal* **66**(1): 34-44.

Greider CW, Blackburn EH. 1987. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell* **51**(6): 887-898.

Gualberto JM, Newton KJ. 2017. Plant mitochondrial genomes: dynamics and mechanisms of mutation. *Annual Review of Plant Biology* **68**: 225-252.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**(3): 307-321.

Guo R, Li YR, He S, Le OY, Sun YW, Zhu ZX. 2018. RepLong: *de novo* repeat identification using long read sequencing data. *Bioinformatics* **34**(7): 1099-1107.

Guo XY, Mandáková T, Trachtová K, Özüdoğru B, Liu JQ, Lysak MA. 2021. Linked by ancestral bonds: multiple whole-genome duplications and reticulate evolution in a Brassicaceae tribe. *Molecular Biology and Evolution* **38**(5): 1695-1714.

Hara M, Fukagawa T. 2018. Kinetochore assembly and disassembly during mitotic entry and exit. *Current Opinion in Cell Biology* **52**(1): 73-81.

Hara M, Fukagawa T. 2020. Dynamics of kinetochore structure and its regulations during mitotic progression. *Cellular and Molecular Life Sciences* **77**(15): 2981-2995.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics* **45**(8): 891-898.

Heckmann S, Macas J, Kumke K, Fuchs J, Schubert V, Ma L, Novak P, Neumann P, Taudien S, Platzer M, et al. 2013. The holocentric species *Luzula elegans* shows interplay between centromere and large-scale genome organization. *Plant Journal* **73**(4): 555-565.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**(5532): 1098-1102.

Heslop-Harrison JS, Murata M, Ogura Y, Schwarzacher T, Motoyoshi F. 1999. Polymorphisms and genomic organization of repetitive DNA from centromeric regions of *Arabidopsis* chromosomes. *Plant Cell* **11**(1): 31-42.

Heslop-Harrison JS, Schwarzacher T. 2011. Organisation of the plant genome in chromosomes. *Plant Journal* **66**(1): 18-33.

Hloušková P, Mandáková T, Pouch M, Trávníček P, Lysak MA. 2019. The large genome size variation in the *Hesperis* clade was shaped by the prevalent proliferation of DNA repeats and rarer genome downsizing. *Annals of Botany* **124**(1): 103-119.

Hohmann N, Wolf EM, Lysak MA, Koch MA. 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* **27**(10): 2770-2784.

Hori T, Shang WH, Hara M, Ariyoshi M, Arimura Y, Fujita R, Kurumizaka H, Fukagawa T. 2017. Association of M18BP1/KNL2 with CENP-A nucleosome is essential for centromere formation in non-mammalian vertebrates. *Developmental Cell* **42**(2): 181-189.

Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. 2022. A near-complete assembly of an *Arabidopsis thaliana* genome. *Molecular Plant* **15**(8): 1247-1250.

Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiology* **138**(3): 1723-1733.

Huang YJ, Ding WJ, Zhang MQ, Han JL, Jing YNF, Yao W, Hasterok R, Wang ZH, Wang K. 2021. The formation and evolution of centromeric satellite repeats in *Saccharum* species. *Plant Journal* **106**(3): 616-629.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao WD, Smit AFA, Wheelers TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44**(D1): D81-D89.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8): 754-755.

Jia PP, Chai WH. 2018. The MLH1 ATPase domain is needed for suppressing aberrant formation of interstitial telomeric sequences. *DNA Repair* **65**: 20-25.

Jiang J, Gill BS. 1994. New 18S.26S ribosomal RNA gene loci: chromosomal landmarks for the evolution of polyploid wheats. *Chromosoma* **103**(3): 179-185.

Jiang JM, Birchler JA, Parrott WA, Dawe RK. 2003. A molecular view of plant centromeres. *Trends in Plant Science* **8**(12): 570-575.

Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**(7008): 569-573.

Jiao WB, Schneeberger K. 2017. The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology* **36**(1): 64-70.

Jiao YN, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang HY, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**(7345): 97-113.

Jin WW, Melo JR, Nagaki K, Talbert PB, Henikoff S, Dawe RK, Jiang JM. 2004. Maize centromeres: organization and functional adaptation in the genetic background of oat. *Plant Cell* **16**(3): 571-581.

Joly-Lopez Z, Bureau TE. 2018. Exaptation of transposable element coding sequences. *Current Opinion in Genetics & Development* **49**(1): 34-42.

Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **98**(15): 8714-8719.

Kato H, Jiang JS, Zhou BR, Rozendaal M, Feng HQ, Ghirlando R, Xiao TS, Straight AF, Bai YW. 2013. A conserved mechanism for centromeric nucleosome recognition by centromere protein CENP-C. *Science* **340**(6136): 1110-1113.

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY, et al. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**(6814): 796-815.

Kim IS, Lee M, Park KC, Jeon Y, Park JH, Hwang EJ, Jeon TI, Ko S, Lee H, Baek SH, et al. 2012. Roles of Mis18α in epigenetic regulation of centromeric chromatin and CENP-A loading. *Molecular Cell* **46**(3): 260-273.

Kim W, Ludlow AT, Min J, Robin JD, Stadler G, Mender I, Lai TP, Zhang N, Wright WE, Shay JW. 2016. Regulation of the human telomerase gene *TERT* by telomere position effect—over long distances (TPE-OLD): implications for aging and cancer. *Plos Biology* **14**(12).

Kleine T, Maier UG, Leister D. 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual Review of Plant Biology* **60**: 115-138.

Koch P, Platzer M, Downie BR. 2014. RepARK-*de novo* creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research* **42**(9): e80.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *Bmc Bioinformatics* **7**(1).

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**(5): 540-549.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. *De novo* assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology* **36**(12): 1174-1179.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, et al. 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology* **30**(7): 692-699.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**(5): 722-736.

Kotseruba V, Gernand D, Meister A, Houben A. 2003. Uniparental loss of ribosomal DNA in the allotetraploid grass *Zingeria trichopoda* (2*n*=8). *Genome* **46**(1): 156-163.

Kozlov AM, Aberer AJ, Stamatakis A. 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**(15): 2577-2579.

Kral L. 2015. Possible identification of CENP-C in fish and the presence of the CENP-C motif in M18BP1 of vertebrates. *F1000Research* **4**: 474.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* **360**(6393): 1085-1092.

Lamb JC, Theuri J, Birchler JA. 2004. What's in a centromere? *Genome Biology* **5**(2).

Lander ES, Consortium IHGS, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lee H, Lee M, Ismail WM, Rho M, Fox GC, Oh S, Tang HX. 2016. MGEScan: a Galaxy-based system for identifying retrotransposons in genomes. *Bioinformatics* **32**(16): 2502-2504.

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**(7780): 679-685.

Lermontova I, Kuhlmann M, Friedel S, Rutten T, Heckmann S, Sandmann M, Demidov D, Schubert V, Schubert I. 2013. *Arabidopsis* KINETOCHORE NULL2 is an upstream component for centromeric histone H3 variant cenH3 deposition at centromeres. *Plant Cell* **25**(9): 3389-3404.

Lermontova I, Sandmann M, Demidov D. 2014. Centromeres and kinetochores of Brassicaceae. *Chromosome Research* **22**(2): 135-152.

Lermontova I, Sandmann M, Mascher M, Schmit AC, Chaboute ME. 2015. Centromeric chromatin and its dynamics in plants. *Plant Journal* **83**(1): 4-17.

Li H. 2016. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* **32**(14): 2103-2110.

Li HT, Luo Y, Gan L, Ma PF, Gao LM, Yang JB, Cai J, Gitzendanner MA, Fritsch PW, Zhang T, et al. 2021. Plastid phylogenomic insights into relationships of all flowering plant families. *Bmc Biology* **19**(1).

Li HT, Yi TS, Gao LM, Ma PF, Zhang T, Yang JB, Gitzendanner MA, Fritsch PW, Cai J, Luo Y, et al. 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants* **5**(5): 461-470.

Li YJ, Zuo S, Zhang ZL, Li ZJ, Han JL, Chu ZQ, Hasterok R, Wang K. 2018. Centromeric DNA characterization in the model grass *Brachypodium distachyon* provides insights on the evolution of the genus. *Plant Journal* **93**(6): 1088-1101.

Liao XY, Gao X, Zhang XK, Wu FX, Wang JX. 2020. RepAHR: an improved approach for *de novo* repeat identification by assembly of the high-frequency reads. *Bmc Bioinformatics* **21**(1).

Liao XY, Hu K, Salhi A, Zou Y, Wang JX, Gao X. 2022. msRepDB: a comprehensive repetitive sequence database of over 80 000 species. *Nucleic Acids Research* **50**(D1): D236-D245.

Liao XY, Zhang XK, Wu FX, Wang JX. 2019. *De novo* repeat detection based on the third generation sequencing reads. *IEEE International Conference on Bioinformatics and Biomedicine (Bibm)*: 431-436.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950): 289-293.

Logsdon GA, Vollger MR, Hsieh P, Mao YF, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, et al. 2021. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**(7857): 101-108.

Luo MC, You FM, Li PC, Wang JR, Zhu TT, Dandekar AM, Leslie CA, Aradhya M, McGuire PE, Dvorak J. 2015. Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. *Bmc Genomics* **16**(1).

Luo RB, Liu BH, Xie YL, Li ZY, Huang WH, Yuan JY, He GZ, Chen YX, Pan Q, Liu YJ, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**.

Lysak MA. 2018. Brassicales: an update on chromosomal evolution and ancient polyploidy. *Plant Systematics and Evolution* **304**(6): 757-762.

Lysak MA, Edginton M, Zuo S, Guo XY, Mandakova T, Al-Shehbaz IA. 2022. Transfer of two *Arabidella* and two *Cuphonotus* species to the genus *Lemphoria* (Brassicaceae) and a description of the new species *L. queenslandica*. *Phytotaxa* **549**(2): 235-240.

Ma JX, Wing RA, Bennetzen JL, Jackson SA. 2007. Plant centromere organization: a dynamic structure with conserved functions. *Trends in Genetics* **23**(3): 134-139.

Macas J, Novák P, Pellicer J, Čížková J, Koblížková A, Neumann P, Fuková I, Doležel J, Kelly LJ, Leitch IJ. 2015. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. *PLoS One* **10**(11).

Majerová E, Mandáková T, Vu GTH, Fajkus J, Lysak MA, Fojtová M. 2014. Chromatin features of plant telomeric sequences at terminal vs. internal positions. *Frontiers in Plant Science* **5**.

Malik HS, Henikoff S. 2002. Conflict begets complexity: the evolution of centromeres. *Current Opinion in Genetics & Development* **12**(6): 711-718.

Malik HS, Vermaak D, Henikoff S. 2002. Recurrent evolution of DNA-binding motifs in the *Drosophila* centromeric histone. *Proceedings of the National Academy of Sciences of the United States of America* **99**(3): 1449-1454.

Mandáková T, Gloss AD, Whiteman NK, Lysak MA. 2016. How diploidization turned a tetraploid into a pseudotriploid. *American Journal of Botany* **103**(7): 1187-1196.

Mandáková T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA. 2010. Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *Plant Cell* **22**(7): 2277-2290.

Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *Plant Journal* **91**(1): 3-21.

Mandáková T, Lysak MA. 2018. Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology* **42**(1): 55-65.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Marie-Nelly H, Marbouty M, Cournac A, Flot JF, Liti G, Parodi DP, Syan S, Guillen N, Margeot A, Zimmer C, et al. 2014. High-quality genome (re) assembly using chromosomal contact data. *Nature Communications* **5**(1).

Marks RA, Hotaling S, Frandsen PB, VanBuren R. 2021. Representation and participation across 20 years of plant genome sequencing. *Nature Plants* **7**(12): 1571-1579.

Marques A, Pedrosa-Harand A. 2016. Holocentromere identity: from the typical mitotic linear structure to the great plasticity of meiotic holocentromeres. *Chromosoma* **125**(4): 669-681.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**(2): 560-564.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS. 2014. Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9**(9).

McKinley KL, Cheeseman IM. 2016. The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology* **17**(1): 16-29.

McStay B. 2006. Nucleolar dominance: a model for rRNA gene silencing. *Genes & Development* **20**(10): 1207-1214.

Mehrotra S, Goyal V. 2014. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics Proteomics Bioinformatics* **12**(4): 164-171.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology* **14**(1).

Meluh PB, Yang PR, Glowczewski L, Koshland D, Smith MM. 1998. Cse4p is a component of the core centromere of *Saccharomyces cerevisiae*. *Cell* **94**(5): 607-613.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. 2018. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nature Communications* **9**(1).

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**(7823): 79-86.

Miguel M, Bartholome J, Ehrenmann F, Murat F, Moriguchi Y, Uchiyama K, Ueno S, Tsumura Y, Lagraulet H, De-Maria N, et al. 2015. Evidence of intense chromosomal shuffling during conifer evolution. *Genome Biology and Evolution* **7**(10): 2799-2809.

Milne I, Lindner D, Bayer M, Husmeier D, McGuire G, Marshall DF, Wright F. 2009. TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics* **25**(1): 126-127.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37**(5): 1530-1534.

Monat C, Tando N, Tranchant-Dubreuil C, Sabot F. 2016. LTRclassifier: a website for fast structural LTR retrotransposons classification in plants. *Mobile Genetic Elements* **6**(6): e1241050.

Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, Meyne J, Ratliff RL, Wu JR. 1988. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* **85**(18): 6622-6626.

Myers EW. 2005. The fragment assembly string graph. *Bioinformatics* **21**: 79-85.

Nagaki K, Cheng ZK, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang JM. 2004. Sequencing of a rice centromere uncovers active genes. *Nature Genetics* **36**(2): 138-145.

Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang JM. 2003. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics* **163**(3): 1221-1225.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nature Reviews Genetics* **14**(3): 157-167.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A, Mandakova T, Jamge B, Lambing C, Kuo P, et al. 2021. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**(6569): 840-845.

Naville M, Warren A, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, Galiana D, Volff JN. 2016. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clinical Microbiology and Infection* **22**(4): 312-323.

Neumann P, Novák P, Hoštáková N, Macas J. 2019. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA* **10**(1): 1-17.

Neumann P, Pavlíková Z, Koblížková A, Fuková I, Jedličková V, Novák P, Macas J. 2015. Centromeres off the hook: massive changes in centromere size and structure following duplication of CenH3 gene in Fabeae species. *Molecular Biology and Evolution* **32**(7): 1862-1879.

Novák P, Ávila-Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. 2017. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research* **45**(12): e111.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *Bmc Bioinformatics* **11**(1).

Novák P, Neumann P, Macas J. 2020. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols* **15**(11): 3745-3776.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**(6): 792-793.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**(6588): 44-49.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research* **30**(9): 1291-1305.

Ong-Abdullah M, Ordway JM, Jiang N, Ooi SE, Kok SY, Sarpan N, Azimi N, Hashim AT, Ishak Z, Rosli SK, et al. 2015. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* **525**(7570): 533-539.

Ou SJ, Su WJ, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* **20**(1).

Palmer DK, Oday K, Trong HL, Charbonneau H, Margolis RL. 1991. Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proceedings of the National Academy of Sciences of the United States of America* **88**(9): 3734-3738.

Paques F, Leung WY, Haber JE. 1998. Expansions and contractions in a tandem repeat induced by double-strand break repair. *Molecular and Cellular Biology* **18**(4): 2045-2054.

Paszkiewicz K, Studholme DJ. 2010. *De novo* assembly of short sequence reads. *Briefings in Bioinformatics* **11**(5): 457-472.

Paula DP. 2021. Next-generation sequencing and its impacts on entomological research in ecology and evolution. *Neotropical Entomology* **50**(5): 679-696.

Pecinka A, Schubert V, Meister A, Kreth G, Klatte M, Lysak MA, Fuchs J, Schubert I. 2004. Chromosome territory arrangement and homologous pairing in nuclei of *Arabidopsis thaliana* are predominantly random except for NOR-bearing chromosomes. *Chromosoma* **113**(5): 258-269.

Pesenti ME, Weir JR, Musacchio A. 2016. Progress in the structural and functional characterization of kinetochores. *Current Opinion in Structural Biology* **37**(1): 152-163.

Peška V, Fajkus P, Fojtová M, Dvořáčková M, Hapala J, Dvořáček V, Polanská P, Leitch AR, Sýkorová E, Fajkus J. 2015. Characterisation of an unusual telomere motif (TTTTTTAGGG)(n) in the plant *Cestrum elegans* (Solanaceae), a species with a large genome. *Plant Journal* **82**(4): 644-654.

Peška V, Garcia S. 2020. Origin, diversity, and evolution of telomere sequences in plants. *Frontiers in Plant Science* **11**.

Pikaard CS. 2000. The epigenetics of nucleolar dominance. *Trends in Genetics* **16**(11): 495-500.

Pontvianne F, Grob S. 2020. Three-dimensional nuclear organization in *Arabidopsis thaliana*. *Journal of Plant Research* **133**(4): 479-488.

Pouokam M, Cruz B, Burgess S, Segal MR, Vazquez M, Arsuaga J. 2019. The Rabl configuration limits topological entanglement of chromosomes in budding yeast. *Scientific Reports* **9**.

Preuss SB, Costa-Nunes P, Tucker S, Pontes O, Lawrence RJ, Mosher R, Kasschau KD, Carrington JC, Baulcombe DC, Viegas W, et al. 2008. Multimegabase silencing in nucleolar dominance involves siRNA-directed DNA methylation and specific methylcytosine-binding proteins. *Molecular Cell* **32**(5): 673-684.

Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**(1): I351-I358.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26**(7): 1641-1650.

Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. *Molecular Cell* **58**(4): 586-597.

Richards EJ, Ausubel FM. 1988. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* **53**(1): 127-136.

Robin JD, Ludlow AT, Batten K, Magdinier F, Stadler G, Wagner KR, Shay JW, Wright WE. 2014. Telomere position effect: regulation of gene expression with progressive telomere shortening over long distances. *Genes & Development* **28**(22): 2464-2476.

Robledillo LA, Neumann P, Koblížková A, Novák P, Vrbová I, Macas J. 2020. Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe Fabeae. *Molecular Biology and Evolution* **37**(8): 2341-2356.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nature Methods* **17**(2): 155-159.

Ruprecht C, Lohaus R, Vanneste K, Mutwil M, Nikoloski Z, Van de Peer Y, Persson S. 2017. Revisiting ancestral polyploidy in plants. *Science Advances* **3**(7): e1603195.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**(4): 406-425.

Sandmann M, Talbert P, Demidov D, Kuhlmann M, Rutten T, Conrad U, Lermontova I. 2017. Targeting of *Arabidopsis* KNL2 to centromeres depends on the conserved CENPC-k motif in its C terminus. *Plant Cell* **29**(1): 144-155.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94**(3): 441-448.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**(12): 5463-5467.

Sato H, Shibata F, Murata M. 2005. Characterization of a Mis12 homologue in *Arabidopsis thaliana*. *Chromosome Research* **13**(8): 827-834.

Saunders VA, Houben A. 2001. The pericentromeric heterochromatin of the grass *Zingeria biebersteiniana* (2*n*=4) is composed of Zbcen1-type tandem repeats that are intermingled with accumulated dispersedly organized sequences. *Genome* **44**(6): 955-961.

Schaeffer CE, Figueroa ND, Liu XL, Karro JE. 2016. phraider: pattern-hunter based rapid ab initio detection of elementary repeats. *Bioinformatics* **32**(12): 209-215.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**(5956): 1112-1115.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Current Opinion in Plant Biology* **15**(2): 147-153.

Schrumpfová PP, Schořová Š, Fajkus J. 2016. Telomere-and telomerase-associated proteins and their functions in the plant cell. *Frontiers in Plant Science* **7**.

Schubert I, Vu GTH. 2016. Genome stability and evolution: attempting a holistic view. *Trends in Plant Science* **21**(9): 749-757.

Schubert V, Neumann P, Marques A, Heckmann S, Macas J, Pedrosa-Harand A, Schubert I, Jang TS, Houben A. 2020. Super-resolution microscopy reveals diversity of plant centromere architecture. *International Journal of Molecular Sciences* **21**(10).

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. *Nature Biotechnology* **38**(9): 1044-1051.

Shakirov EV, Chen JJL, Shippen DE. 2022. Plant telomere biology: the green solution to the end-replication problem. *Plant Cell* **34**(7): 2492–2504.

Shan W, Kubová M, Mandáková T, Lysak MA. 2021. Nuclear organization in crucifer genomes: nucleolus-associated telomere clustering is not a universal interphase configuration in Brassicaceae. *Plant Journal* **108**(2): 528-540.

Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94**(3): 275-288.

Shay JW. 2018. Telomeres and aging. *Current Opinion in Cell Biology* **52**(1): 1-7.

Shen XX, Li YN, Hittinger CT, Chen XX, Rokas A. 2020. An investigation of irreproducibility in maximum likelihood phylogenetic inference. *Nature Communications* **11**(1).

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchishinozaki K, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *Embo Journal* **5**(9): 2043-2049.

Silva MCC, Jansen LET. 2009. At the right place at the right time: novel CENP-A binding proteins shed light on centromere assembly. *Chromosoma* **118**(5): 567-574.

Simpson JT, Durbin R. 2012. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Research* **22**(3): 549-556.

Simpson JT, Pop M. 2015. The theory and practice of genome sequence assembly. *Annual Review of Genomics and Human Genetics* **16**: 153-172.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Research* **19**(6): 1117-1123.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods* **14**(4): 407-412.

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *Plos Biology* **10**(1).

Sohn JI, Nam JW. 2018. The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics* **19**(1): 23-40.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96**(1): 336-348.

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF, et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* **133**(4): 381-461.

Sone T, Fujisawa M, Takenaka M, Nakagawa S, Yamaoka S, Sakaida M, Nishiyama R, Yamato KT, Ohmido N, Fukui K, et al. 1999. Bryophyte 5S rDNA was inserted into 45S rDNA repeat units after the divergence from higher land plants. *Plant Molecular Biology* **41**(5): 679-685.

Souza G, Vanzela ALL, Crosa O, Guerra M. 2016. Interstitial telomeric sites and Robertsonian translocations in species of *Ipheion* and *Nothoscordum* (Amaryllidaceae). *Genetica* **144**(2): 157-166.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**(4): 456-463.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Research* **37**(21): 7002-7013.

Steiner FA, Henikoff S. 2014. Holocentromeres are dispersed point centromeres localized at transcription factor hotspots. *Elife* **3**.

Stellfox ME, Nardi IK, Knippler CM, Foltz DR. 2016. Differential binding partners of the Mis18α/β YIPPEE domains regulate Mis18 complex recruitment to centromeres. *Cell Reports* **15**(10): 2127-2135.

Sugimoto K, Yata H, Muro Y, Himeno M. 1994. Human centromere protein C (CENP-C) is a DNA-binding protein which possesses a novel DNA-binding motif. *Journal of Biochemistry* **116**(4): 877-881.

Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics* **18**(5): 292-308.

Sun HQ, Jiao WB, Campoy JA, Krause K, Goel M, Folz-Donahue K, Kukat C, Huettel B, Schneeberger K. 2022. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature Genetics* **54**(3): 342-348.

Sureshkumar S, Todesco M, Schneeberger K, Harilal R, Balasubramanian S, Weigel D. 2009. A genetic defect caused by a triplet repeat expansion in *Arabidopsis thaliana*. *Science* **323**(5917): 1060-1063.

Sýkorová E, Lim KY, Kunická Z, Chase MW, Bennett MD, Fajkus J, Leitch AR. 2003. Telomere variability in the monocotyledonous plant order Asparagales. *Proceedings of the Royal Society B-Biological Sciences* **270**(1527): 1893-1904.

Talbert PB, Bryson TD, Henikoff S. 2004. Adaptive evolution of centromere proteins in plants and animals. *Journal of biology* **3**(4): 18-27.

Talbert PB, Henikoff S. 2018. Transcribing centromeres: noncoding RNAs and kinetochore assembly. *Trends in Genetics* **34**(8): 587-599.

Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S. 2002. Centromeric localization and adaptive evolution of an *Arabidopsis* histone H3 variant. *Plant Cell* **14**(5): 1053-1066.

Tamura K, Stecher G, Kumar S. 2021. MEGA11: molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution* **38**(7): 3022-3027.

Thomas BC, 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**(7): 934-946.

Tran TD, Cao HX, Jovtchev G, Neumann P, Novák P, Fojtová M, Vu GTH, Macas J, Fajkus J, Schubert I, et al. 2015. Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus *Genlisea*. *Plant Journal* **84**(6): 1087-1099.

Tucker S, Vitins A, Pikaard CS. 2010. Nucleolar dominance and ribosomal RNA gene silencing. *Current Opinion in Cell Biology* **22**(3): 351-356.

Turowski TW, Tollervey D. 2015. Cotranscriptional events in eukaryotic ribosome synthesis. *Wiley Interdisciplinary Reviews: RNA* **6**(1): 129-139.

Udall JA, Dawe RK. 2018. Is it ordered correctly? Validating genome assemblies by optical mapping. *Plant Cell* **30**(1): 7-14.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research* **18**(7): 1051-1063.

VanBuren R, Bryant D, Edger PP, Tang HB, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, et al. 2015. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**(7579): 508-U209.

Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Research* **24**(8): 1334-1347.

Varma D, Salmon ED. 2012. The KMN protein network - chief conductors of the kinetochore orchestra. *Journal of Cell Science* **125**(24): 5927-5936.

Vaser R, Šikić M. 2021. Time- and memory-efficient genome assembly with Raven. *Nature Computational Science* **1**(5): 332-336.

Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, Ruelens P, Maere S, Van de Peer Y, Geuten K. 2012. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Molecular Biology and Evolution* **29**(12): 3793-3806.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1309.

Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. *Annals of Botany* **120**(2): 195-207.

Victoria FC, da Maia LC, de Oliveira AC. 2011. In *silico* comparative analysis of SSR markers in plants. *Bmc Plant Biology* **11**(1).

Volkov RA, Borisjuk NV, Panchuk II, Schweizer D, Hemleben V. 1999. Elimination and rearrangement of parental rDNA in the allotetraploid *Nicotiana tabacum*. *Molecular Biology and Evolution* **16**(3): 311-320.

Vondrak T, Robledillo LA, Novák P, Koblížková A, Neumann P, Macas J. 2020. Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *Plant Journal* **101**(2): 484-500.

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, et al. 2013. The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife* **2**.

Walden N, German DA, Wolf EM, Kiefer M, Rigault P, Huang XC, Kiefer C, Schmickl R, Franzke A, Neuffer B, et al. 2020. Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. *Nature Communications* **11**(1).

Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, et al. 2021. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *Genomics Proteomics Bioinformatics* **20**(1): 4-13.

Weiss-Schneeweiss H, Bloch C, Turner B, Villasenor JL, Stuessy TF, Schneeweiss GM. 2012. The promiscuous and the chaste: frequent allopolyploid speciation and its genomic consequences in American daisies (*Melampodium* sect. *Melampodium*; Asteraceae). *Evolution* **66**(1): 211-228.

Wendel JF, Schnabel A, Seelanan T. 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences of the United States of America* **92**(1): 280-284.

Whelan S, Morrison DA. 2017. Inferring trees. *Methods in Molecular Biology* **1525**: 349-377.

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramirez-Gonzalez RH, De Oliveira R, Mayer KFX, Paux E, Choulet F, et al. 2018. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology* **19**(1).

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**(12): 973-982.

Wojciechowski MF, Lavin M, Sanderson MJ. 2004. A phylogeny of legumes (Leguminosae) based on analyses of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany* **91**(11): 1846-1862.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America* **84**(24): 9054-9058.

Wood AM, Danielsen JMR, Lucas CA, Rice EL, Scalzo D, Shimi T, Goldman RD, Smith ED, Le Beau MM, Kosak ST. 2014. TRF2 and lamin A/C interact to facilitate the functional organization of chromosome ends. *Nature Communications* **5**(1).

Wood AM, Laster K, Rice EL, Kosak ST. 2015. A beginning of the end: new insights into the functional organization of telomeres. *Nucleus* **6**(3): 172-178.

Wu JZ, Yamagata H, Hayashi-Tsugane M, Hijishita S, Fujisawa M, Shibata M, Ito Y, Nakamura M, Sakaguchi M, Yoshihara R, et al. 2004. Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell* **16**(4): 967-976.

Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, Xie Z. 2017. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nature Methods* **14**(11): 1072-1079.

Xiong WW, He LM, Lai JS, Dooner HK, Du CG. 2014. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences of the United States of America* **111**(28): 10263-10268.

Yan HD, Bombarely A, Li S. 2020. DeepTE: a computational method for *de novo* classification of transposons with convolutional neural network. *Bioinformatics* **36**(15): 4269-4275.

Yang AM, Zhang W, Wang JH, Yang K, Han Y, Zhang LM. 2020. Review on the application of machine learning algorithms in the sequence data mining of DNA. *Frontiers in Bioengineering and Biotechnology* **8**.

Yang XM, Zhao HN, Zhang T, Zeng ZX, Zhang PD, Zhu B, Han YH, Braz GT, Casler MD, Schmutz J, et al. 2018. Amplification and adaptation of centromeric repeats in polyploid switchgrass species. *New Phytologist* **218**(4): 1645-1657.

Yang ZH, Rannala B. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**(5): 303-314.

Ye CX, Ma ZSS, Cannon CH, Pop M, Yu DW. 2012. Exploiting sparseness in *de novo* genome assembly. *Bmc Bioinformatics* **13**(1).

Yoo MJ, Liu XX, Pires JC, Soltis PS, Soltis DE. 2014. Nonadditive gene expression in polyploids. *Annual review of genetics* **48**: 485-517.

Young AD, Gillung JP. 2020. Phylogenomics: principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology* **45**(2): 225-247.

Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **18**(5): 821-829.

Zhang D, Martyniuk CJ, Trudeau VL. 2006. SANTA domain: a novel conserved protein module in Eukaryota with potential involvement in chromatin regulation. *Bioinformatics* **22**(20): 2459-2462.

Zhang HQ, Koblizkova A, Wang K, Gong ZY, Oliveira L, Torres GA, Wu YF, Zhang WL, Novak P, Buell CR, et al. 2014. Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* species: rapid evolution of DNA sequences associated with centromeres. *Plant Cell* **26**(4): 1436-1447.

Zhang WP, Zuo S, Li ZJ, Meng Z, Han JL, Song JQ, Pan YB, Wang K. 2017. Isolation and characterization of centromeric repetitive DNA sequences in *Saccharum spontaneum*. *Scientific Reports* **7**.

Zhao F, Chen YP, Salmaki Y, Drew BT, Wilson TC, Scheen AC, Celep F, Brauchler C, Bendiksby M, Wang Q, et al. 2021. An updated tribal classification of Lamiaceae based on plastome phylogenomics. *Bmc Biology* **19**(1).

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang JM, Dawe RK. 2002. Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell* **14**(11): 2825-2836.

Zhou XF, Shen XX, Hittinger CT, Rokas A. 2018. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Molecular Biology and Evolution* **35**(2): 486-503.

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**(21): 2669-2677.

Zonneveld BJ, Leitch IJ, Bennett MD. 2005. First nuclear DNA amounts in more than 300 angiosperms. *Annals of Botany* **96**(2): 229-244.

Zozomová-Lihová J, Mandáková T, Kovaříková A, Muhlhausen A, Mummenhoff K, Lysak MA, Kovařík A. 2014. When fathers are instant losers: homogenization of rDNA loci in recently formed *Cardamine* X *schulzii* trigenomic allopolyploid. *New Phytologist* **203**(4): 1096-1108.

Zuo S, Guo XY, Mandáková T, Edginton M, Al-Shehbaz IA, Lysak MA. 2022a. Genome diploidization associates with cladogenesis, trait disparity, and plastid gene evolution. *Plant Physiology* **190**(1): 403-420.

Zuo S, Mandáková T, Kubová M, Lysak MA. 2022b. Genomes, repeatomes and interphase chromosome organization in the meadowfoam family (Limnanthaceae, Brassicales). *Plant Journal* **110**(5): 1462-1475.

Zuo S, Yadala R, Yang F, Talbert P, Fuchs J, Schubert V, Ahmadli U, Rutten T, Pecinka A, Lysak MA, et al. 2022c. Recurrent plant-specific duplications of KNL2 and its conserved function as a kinetochore assembly factor. *Molecular Biology and Evolution* **39**(6): msac123.

Zytnicki M, Akhunov E, Quesneville H. 2014. Tedna: a transposable element *de novo* assembler. *Bioinformatics* **30**(18): 2656-2658.

7 LIST OF PUBLICATIONS (I-V)

the plant journal

SEB
SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **110,** 1462–1475 doi: 10.1111/tpj.15750

Genomes, repeatomes and interphase chromosome organization in the meadowfoam family (Limnanthaceae, Brassicales)

Sheng Zuo[1,2], Terezie Mandáková[1,3], Michaela Kubová[1,3] and Martin A. Lysak[1,2,*]

[1]*CEITEC – Central European Institute of Technology, Masaryk University, Brno CZ-625 00, Czech Republic,*
[2]*National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno CZ-625 00, Czech Republic, and*
[3]*Department of Experimental Biology, Faculty of Science, Masaryk University, Brno CZ-625 00, Czech Republic*

Received 21 January 2022; revised 17 March 2022; accepted 28 March 2022; published online 30 March 2022.

*For correspondence (e-mail martin.lysak@ceitec.muni.cz).

SUMMARY

The meadowfoam family (Limnanthaceae) is one of the smallest and genomically underexplored families of the Brassicales. The Limnanthaceae harbor about seven species in the genus *Limnanthes* (meadowfoam) and *Floerkea proserpinacoides* (false mermaidweed), all native to North America. Because all *Limnanthes* and *Floerkea* species have only five chromosome pairs, i.e., a chromosome number rare in Brassicales and shared with *Arabidopsis thaliana* (Arabidopsis), we examined the Limnanthaceae genomes as a potential model system. Using low-coverage whole-genome sequencing data, we reexamined phylogenetic relationships and characterized the repeatomes of Limnanthaceae genomes. Phylogenies based on complete chloroplast and 35S rDNA sequences corroborated the sister relationship between *Floerkea* and *Limnanthes* and two major clades in the latter genus. The genome size of Limnanthaceae species ranges from 1.5 to 2.1 Gb, apparently due to the large increase in DNA repeats, which constitute 60–70% of their genomes. Repeatomes are dominated by long terminal repeat retrotransposons, while tandem repeats represent only less than 0.5% of the genomes. The average chromosome size in Limnanthaceae species (340–420 Mb) is more than 10 times larger than in Arabidopsis (32 Mb). A three-dimensional fluorescence *in situ* hybridization analysis demonstrated that the five chromosome pairs in interphase nuclei of *Limnanthes* species adopt the Rabl-like configuration.

Keywords: Brassicales, chromosomes, DNA repeats, interphase, *Limnanthes*, meadowfoam, Rabl, repeatome.

INTRODUCTION

In addition to iconic and taxon-rich families, such as Brassicaceae, Capparaceae and Cleomaceae, the order Brassicales contains several monotypic and oligotypic families (Cardinal-McTeague et al., 2016; Edger et al., 2018; Swanepoel et al., 2020; The Angiosperm Phylogeny Group IV, 2016). Limnanthaceae, the meadowfoam family, harbors only two genera and eight annual species (Tucker, 1993). While the genus *Limnanthes* (meadowfoams) has seven species, *Floerkea* contains only one species (*F. proserpinacoides*, false mermaidweed). The family occurs disjunctly throughout temperate North America, with the greatest species diversity of *Limnanthes* restricted to California. *Floerkea* is a spring ephemeral annual of deciduous or coniferous forests, with a life cycle completed in only 60–70 days (Baskin et al., 1988). *Limnanthes* species are mostly spring flowering annuals of Californian and southern Oregon seasonal wetland habitats (vernal pools). Due to human-driven habitat destruction, many of the wild *Limnanthes* species are considered endangered (Meyers et al., 2010). White meadowfoam (*Limnanthes alba*) is an oilseed crop providing seed oil attractive for the cosmetic industry (Agerbirk et al., 2022; Jenderek & Hannan, 2009), while the poached egg flower (*Limnanthes douglasii*) is widely grown as a showy ornamental plant.

Helge Stenar was probably the first to note that *Limnanthes* species have only five pairs of chromosomes by analyzing the course of female and male meiosis in *L. douglasii* (Stenar, 1925). This low chromosome number corresponded to that of Arabidopsis (then known as *Stenogramma thalianum*; Laibach, 1907), but the chromosomes of *Limnanthes* were much larger. The small number of large (5–10 μm; Fries, 1936) and morphologically distinct chromosomes attracted the interest of some researchers

© 2022 Society for Experimental Biology and John Wiley & Sons Ltd.

(e.g., Arroyo, 1973; Fries, 1936; Propach, 1934), but the meadowfoam species never established themselves as important research models.

The Limnanthaceae have a rather basal position within the Brassicales, being placed between the Setchellanthaceae (*Setchellanthus caeruleus*) and the large clade consisting of the core Brassicales and four small families (Bataceae, Koeberliniaceae, Salvadoraceae and Tiganophytaceae) (Edger et al., 2018; Swanepoel et al., 2020). The evolution of Brassicales genomes has been influenced by several whole-genome duplications (WGD), of which the At-β paleotetraploid duplication is shared by most Brassicales families (e.g., Barker et al., 2009). While it is not clear whether the At-β WGD occurred before or after the divergence of Setchellanthaceae, the origin of Limnanthaceae likely post-dated the At-β WGD (Edger et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019). Although the chromosome number of the At-β tetraploid has not yet been reliably inferred (Lysak, 2018), it should be assumed that the five-chromosome genomes of today's Limnanthaceae species arose by descending dysploidy (i.e., reduction of chromosome number) that occurred during post-At-β genome diploidization.

Given the large knowledge gap extending from the Caricaceae (the papaya genome, *Carica papaya*) to the Brassicaceae, the phylogenetic position within the Brassicales, low chromosome number, annual herbaceous life history and seed availability make the Limnanthaceae potentially attractive for gaining more insight into genome evolution in the Brassicales. Here, we performed low-coverage whole-genome sequencing, including short and long reads, in four *Limnanthes* (sub)species and *Floerkea* with a threefold goal: (i) to reconstruct phylogenetic relationships based on whole-chloroplast and rDNA sequences, (ii) to characterize and compare repeatomes of Limnanthaceae species and (iii) to use the *de novo* identified repeats to analyze interphase chromosome organization in Limnanthaceae by three-dimensional fluorescence *in situ* hybridization (3D FISH).

RESULTS

Characterization of plastomes and nuclear 35S rDNA

Using the low-coverage whole-genome sequencing data, we assembled the complete chloroplast genome (cp genome) and nuclear 35S rDNA sequence of the five Limnanthaceae accessions. The length of the complete cp genomes ranged from 151 411 bp (*Floerkea proserpinacoides*) to 152 711 bp (*L. alba*). All cp genomes exhibited the typical quadripartite structure of angiosperm plastomes, consisting of a pair of inverted repeat (IR) regions separated by a large single-copy region and a small single-copy (SSC) region. The total guanine-cytosine (GC) content ranged from 35.9% to 36.0%. The Limnanthaceae cp genome contains a total of 112 unique genes, including 78 protein-coding genes (PCGs), 30 tRNAs, and four rRNAs. Nineteen genes were duplicated in the IR region, including eight PCGs, seven tRNAs and four rRNAs. The assembled length of the nuclear 35S rDNA sequences varied from 6747 bp in *Limnanthes floccosa* subsp. *bellingeriana* to 10 837 bp in *L. floccosa* subsp. *grandiflora*. Due to incomplete assembly of the highly variable intergenic spacer region, we used only the conservative 18S-ITS1-5.8S-ITS2-26S region for phylogenetic analysis.

The cp genome of most land plants consists of two structural haplotypes that differ only in the orientation of their SSC sequences (Palmer, 1983; Wang & Lanfear, 2019). To analyze this phenomenon in Limnanthaceae, we developed a BLAST-dependent approach to detect and quantify chloroplast structural heteroplasmy using Oxford Nanopore (ONT) long reads (see Experimenal Procedures section). By mapping selected ONT reads (longer than the IR length), we observed two types of cp genomes, distinguished by the relative orientation of the SSC sequences (Figure S1). Based on the frequency of mapped ONT reads, we estimated that the two structural haplotypes occurred with approximately equal frequencies.

Phylogenomic analysis retrieved two genus-level clades in Limnanthaceae

To clarify phylogenetic relationships within Limnanthaceae, we extracted PCG sequences from the assembled cp genomes. Based on 72 PCGs in 22 plastomes of Limnanthaceae and other Brassicales species, we compiled a gap-free alignment matrix with 43 263 columns, of which 4773 were parsimony informative. Both the maximum likelihood (ML) tree and the Bayesian inference (BI) tree confirmed ($BS = 100$ and $PP = 1$) the within-family split corresponding to the genera *Floerkea* and *Limnanthes* (Figure 1a; Figure S2). In the genus *Limnanthes*, there were two well-supported clades ($BS = 100$ and $PP = 1$): (i) *L. alba* and *L. floccosa* and (ii) *L. douglasii*. The two clades correspond to the infrageneric sections *Inflexae* and *Limnanthes* (formerly *Reflexae*), as defined previously (Mason, 1952; Meyers et al., 2010).

Both the non-partitioned and partitioned ML phylogenetic trees based on the 5928-bp alignment matrix of the nuclear 35S rDNA sequences strongly supported ($BS = 100$) the two-genus topology of the family tree (Figure S3a,b). In addition, the rDNA-based BI phylogeny strongly supported ($PP = 1$) the sister relationship between *Floerkea* and *Limnanthes*, as well as the *Inflexae* and *Limnanthes* sections in *Limnanthes* (Figure S3c).

Genome size and chromosome number variation in Limnanthaceae

Flow cytometric analysis revealed that genome size varied 1.4-fold, from 1516 Mb in *F. proserpinacoides* to 2102 Mb in *L. douglasii* (Table 1). Chromosome counting in somatic

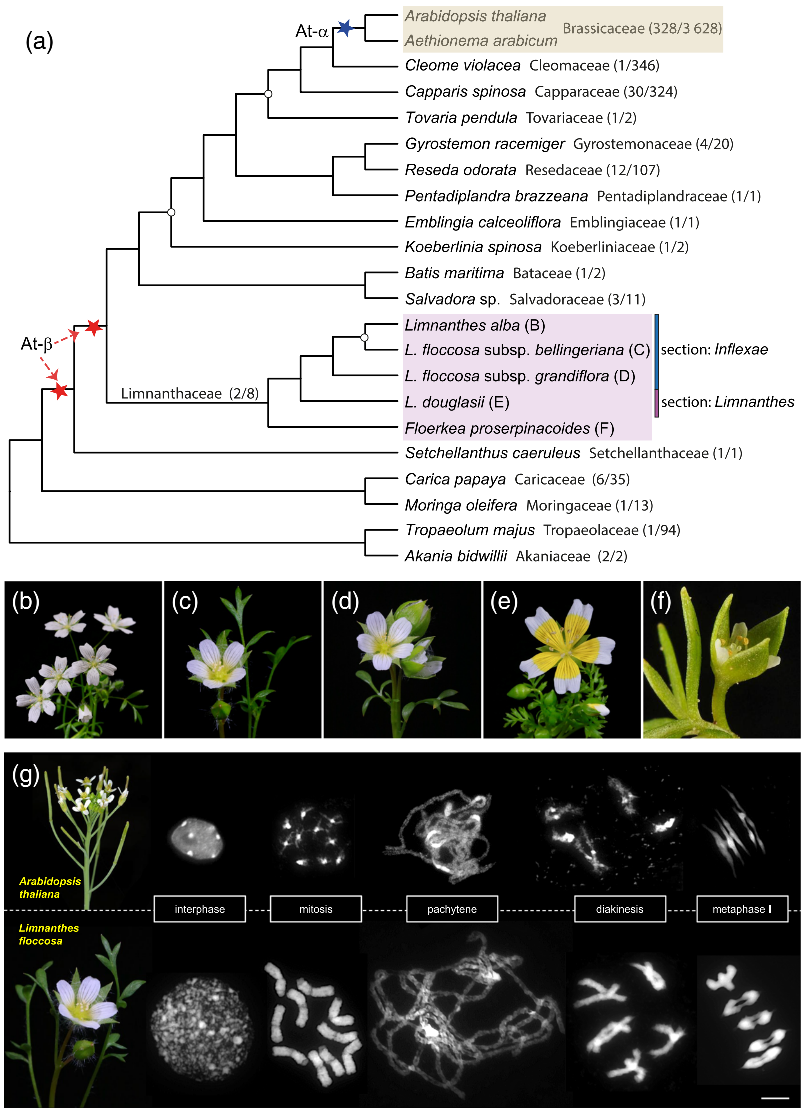

(a)
At-α
Arabidopsis thaliana
Aethionema arabicum
Brassicaceae (328/3 628)
Cleome violacea Cleomaceae (1/346)
Capparis spinosa Capparaceae (30/324)
Tovaria pendula Tovariaceae (1/2)
Gyrostemon racemiger Gyrostemonaceae (4/20)
Reseda odorata Resedaceae (12/107)
Pentadiplandra brazzeana Pentadiplandraceae (1/1)
Emblingia calceoliflora Emblingiaceae (1/1)
Koeberlinia spinosa Koeberliniaceae (1/2)
Batis maritima Bataceae (1/2)
Salvadora sp. Salvadoraceae (3/11)
At-β
Limnanthes alba (B)
L. floccosa subsp. bellingeriana (C)
section: Inflexae
L. floccosa subsp. grandiflora (D)
Limnanthaceae (2/8)
L. douglasii (E)
section: Limnanthes
Floerkea proserpinacoides (F)
Setchellanthus caeruleus Setchellanthaceae (1/1)
Carica papaya Caricaceae (6/35)
Moringa oleifera Moringaceae (1/13)
Tropaeolum majus Tropaeolaceae (1/94)
Akania bidwillii Akaniaceae (2/2)
(b)
(c)
(d)
(e)
(f)
(g)
Arabidopsis thaliana
interphase
mitosis
pachytene
diakinesis
metaphase I
Limnanthes floccosa

Figure 1. Phylogenetic position and chromosomes of Limnanthaceae species. (a) The maximum likelihood tree based on 72 chloroplast genes shows the relationship between 17 Brassicalaes families (Tiganophytaceae not included) and five Limnanthaceae taxa analyzed herein. Two red stars indicate the uncertain phylogenetic placement of the At-β whole-genome duplication (WGD) (Edger et al., 2018); the At-α WGD (blue star) occurred prior to the divergence of Brassicaceae. Numbers of genera/species are given for each family (Lysak, 2018). Bootstrap values of <100 are marked as white circles. Capital letters after species names correspond to images of Limnanthaceae taxa (b–f). (g) Comparison of interphase nuclei, mitotic ($2n = 10$) and meiotic ($n = 5$) chromosomes in *Arabidopsis thaliana* (Brassicaceae) and *Limnanthes floccosa* subsp. *bellingeriana*. In both species, chromosomes and nuclei were isolated from young anthers and counterstained by DAPI. Scale bar, 10 μm.

tissues of young anthers confirmed five pairs of (sub)metacentric chromosomes ($2n = 10$) in all Limnanthaceae taxa analyzed. The average size of the highly condensed mitotic metaphase chromosomes ranged from 8 to 12 μm (Figures 1g and 3j,k). Although Limnanthaceae and *Arabidopsis thaliana* (135 Mb, Brassicaceae) have the same number of chromosomes (Lysak, 2018), their size and structure differ significantly (Figure 1g). The average chromosome size (genome size/haploid chromosome number) is over 300 Mb (340–420 Mb) in Limnanthaceae species, while it is only 32 Mb in Arabidopsis. We observed that there is no strong eu–/heterochromatin boundary and heterochromatin is rather equally distributed throughout the chromosomes (Figures 1g and 3f–k; Figure S4). While in Arabidopsis most of the repetitive sequences are located at the heterochromatic pericentromeres, in Limnanthaceae the repeats are distributed almost evenly over the >300-Mb-long chromosomes (Figure S4).

Repeatome composition: transposable elements

To identify and analyze the sequences constituting genomes of Limnanthaceae taxa (Table 1), the RepeatExplorer2

© 2022 Society for Experimental Biology and John Wiley & Sons Ltd.,

Table 1 Genome size estimation in Limnanthaceae

Species	2*n*	Genome size (pg/1C)	Genome size (Mb/1C)
Floerkea proserpinacoides	10	1.55	1515.90
Limnanthes douglasii	10	2.15	2102.41
Limnanthes floccosa subsp. *bellingeriana*	10	1.81	1770.58
Limnanthes floccosa subsp. *grandiflora*	10	1.91	1865.57

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

pipeline was used to identify the major types of repetitive sequences and their genome representation. The identified repetitive sequences accounted for an estimated 58.12–66.22% of the analyzed genomes (Figure 2a; Table S1). In all repeatomes, long terminal repeat (LTR) retrotransposons accounted for the majority of repeats, ranging from 21.04% in *F. proserpinacoides* to 24.59% in *L. douglasii*. In Limnanthaceae genomes, the identified Ty1-*copia* elements belonged to six lineages (Ale, Alesia, Ikeros, Ivana, TAR and Tork; Neumann et al., 2019), while Ty3-*gypsy* elements belonged to two major lineages, chromovirus (CRM, Galadriel and Reina clades) and non-chromovirus (Athila, Ogre/Tat and Retand clades). The Ty1-*copia* elements were mainly represented by the Tork and Ale lineages, while the Ty3-*gypsy* superfamily was mostly represented by the Retand elements. The genome proportion of Ty3-*gypsy* elements ranged from 14.39% in *F. proserpinacoides* to 20.10% in *L. floccosa* subsp. *grandiflora*, whereas Ty1-*copia* retroelements were three to five times less abundant than Ty3-*gypsy* elements (Table S1).

Among non-LTR retrotransposons, long interspersed nuclear elements were identified with genome proportions ranging from 0.21% in *L. floccosa* subsp. *bellingeriana* to 1.55% in *L. douglasii* (Table S1). Short interspersed nuclear elements were not detected in clusters that accounted for at least 0.01% of the nuclear genome. DNA transposons were represented at frequencies ranging from 2.19% in *L. floccosa* subsp. *bellingeriana* to 4.76% in *L. alba*; mutator elements were the most abundant DNA transposons in Limnanthaceae genomes (Table S1).

The chromosomal distribution of selected retrotransposons (Table S2) was determined by FISH in *L. douglasii*. All five retroelement-based probes tested (Li-Dou1, Li-Dou32, Li-Dou38, Li-Dou41 and Li-Dou49) yielded

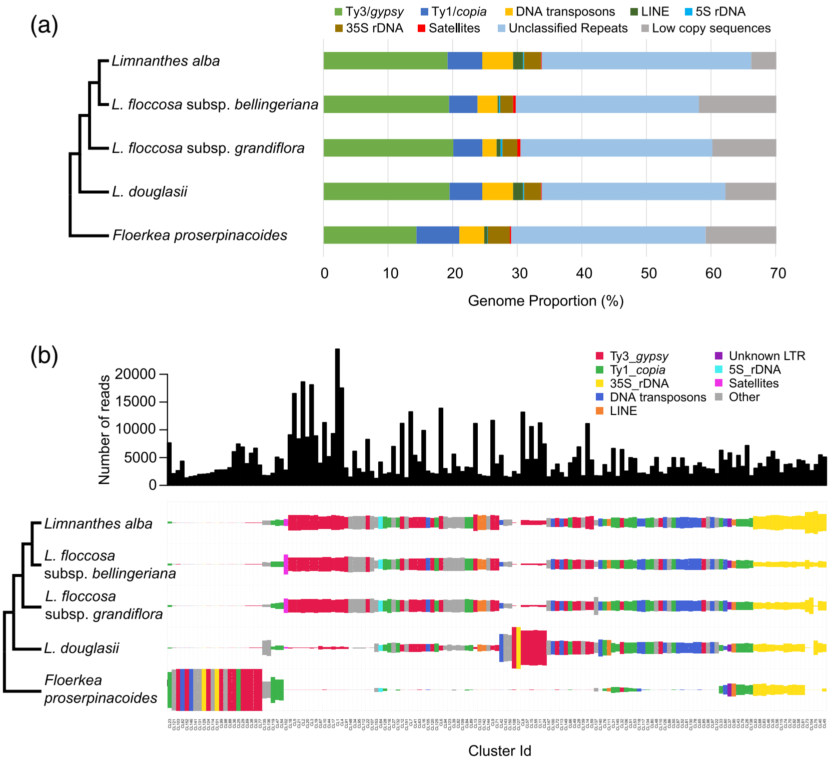


Figure 2. Repeatome composition and comparative clustering analysis in Limnanthaceae taxa. (a) Relative abundances of repeat and low-copy sequences in Limnanthaceae genomes. Low-copy sequences above 70% are not shown. The simplified plastome-based tree was adapted from Figure 1a. (b) Comparative repeat profiles of Limnanthaceae taxa. Comparative analysis of the five Limnanthaceae genomes was performed using the graphic clustering method: (i) 500 000 reads per species were sampled as input data for the RepeatExplorer2 pipeline, (ii) most abundant repeat clusters (>0.05% of the total input reads) were annotated. A bar plot on the top of the graph depicts the number of reads per top clusters. Differently colored rectangles represent different repeat types and their sizes are proportional to the number of reads in a given cluster. Hierarchical clustering was used to sort the read clusters. *Floerkea* contains abundant species-specific retrotransposons.

© 2022 Society for Experimental Biology and John Wiley & Sons Ltd.,

similarly strong signals that were evenly distributed along all chromosomes (Figure S4).

Repeatome composition: tandem repeats

The identified tandem repeats constituted only less than 0.5% of Limnanthaceae nuclear genomes. Seven satellite repeats were identified in Limnanthaceae taxa with very low abundances ranging from 0.03% in *L. alba* to 0.27% in *L. floccosa* subsp. *bellingeriana* (Figure 2a; Table S2). The 173-bp LiFlo-TR34 satellite was shared by two subspecies of *L. floccosa*.

rDNA loci and tandem repeats as chromosomal landmarks in *Limnanthes*

Terminal nucleolar organizer regions (NORs, 35S rDNA) were identified on two chromosome pairs (chromosomes 1 and 2) in *Floerkea*, while NORs were detected on three chromosome pairs in *Limnanthes* taxa (Figure 3a–o). The 35S rDNA loci were often fragile and broken off from the chromosomes (Figure 3g,i). In all taxa, 5S rDNA loci were identified at interstitial positions on one (*Floerkea*, *L. alba*; Figure 3a,b) or two chromosome pairs (remaining *Limnanthes* taxa; Figure 3c–e).

The chromosomal distribution of selected tandem repeats (Table S2) was determined by FISH in *L. alba*, *L. douglasii*, *L. floccosa* subsp. *bellingeriana* and *L. floccosa* subsp. *grandiflora* (Figure 3). The 173-bp LiFlo-TR34 repeat decorated all centromeres and two subtelomeric regions of one chromosome pair in *L. floccosa* subsp. *bellingeriana* and *L. floccosa* subsp. *grandiflora* (Figure 3o–s). An additional 96-bp tandem repeat (LiFlo-TR143) was identified as a single subtelomeric locus on chromosome 5 in all *Limnanthes* accessions except *L. alba* (Figure 3k,l). Species-specific satellites were identified in *L. alba* (122-bp LiAlb-TR120 in the subtelomeric region of chromosome 5 and 172-bp LiAlb-TR94 in centromeres of chromosomes 3 and 5; Figure 3b,g,h), *L. douglasii* (515-bp LiDou-TR200 in the centromere of chromosome 1 and 1887-bp LiDou-TR92 in the subtelomeric region of chromosome 2; Figure 3c,i,j) and *L. floccosa* subsp. *bellingeriana* (92-bp LiFlo-TR169 located interstitially on chromosome 5; Figure 3d,l).

Infrageneric abundance and divergence of centromeric repeat LiFlo-TR34

The centromeric repeat LiFlo-TR34 has a 173-bp consensus sequence inferred using TAREAN (Novák et al., 2017). Its monomer size was shorter than the centromeric repeat (approximately 200 bp) in the closely related Caricaceae (*C. papaya*), but similar in size to many other centromeric satellites (approximately 180 bp) in Brassicaceae (Melters et al., 2013). To analyze the occurrence of LiFlo-TR34, we searched for the repeat in the Limnanthaceae genomes using multiple approaches. Graph clustering with RepeatExplorer2 generated sphere-shaped graphs only in *L. floccosa* subsp. *bellingeriana* and *L. floccosa* subsp. *grandiflora* (Figure S5). Using BLASTn, this repeat could only be identified in reads of *L. floccosa* subsp. *bellingeriana* and *L. floccosa* subsp. *grandiflora*, while it was absent in reads of *L. alba*, which belongs to the same section (*Inflexae*). To expand our search, we performed BLAST searches of the LiFlo-TR34 consensus sequence against the NCBI nucleotide collection and found that LiFlo-TR34 had only four hits with approximately 80% similarity to the chromosome sequences of *Solanum tuberosum* (query cover: approximately 32%). Altogether these data indicate that LiFlo-TR34 is specific for *L. floccosa*.

To examine the LiFlo-TR34 profile in the two *L. floccosa* subspecies, RepeatProfiler (Negm et al., 2021) was used to estimate the level of variation. A total of 15 million paired-end reads were sampled, and 26 992 and 22 136 reads were retrieved as LiFlo-TR34 reads, respectively. LiFlo-TR34 copy number variation (CNV) profiles showed that LiFlo-TR34 is more abundant in *L. floccosa* subsp. *bellingeriana* than in *L. floccosa* subsp. *grandiflora* (Figure S6), consistently with the results of RepeatExplorer2. LiFlo-TR34 reads have approximately 82.2% identity in subsp. *bellingeriana* and 82.1% identity in subsp. *grandiflora*, indicating a substantial amount of sequence variation within the subspecies. However, variant profile graphs are similar among both *L. floccosa* accessions (Figure S6). Additionally, some recurrent variants were observed that appeared to be prone to variation in the two *L. floccosa* accessions (Figure S6).

Repeatome variation within Limnanthaceae and repeat-based phylogeny

A total of 2.5 million reads from five Limnanthaceae accessions, accounting for 3.3–4.6% of their genomes, were sampled as input data submitted to the RepeatExplorer platform to perform comparative clustering analysis (Table 1). Approximately 0.75 million reads were grouped into 153 clusters representing moderately or highly abundant repeat families for further annotation (Figure 2b). Only 45 of these 153 clusters were shared among all Limnanthaceae accessions, whereas most repeats (85/153) were shared only among *Limnanthes* species. Five clusters absent (5/85) and two species-specific clusters in *L. douglasii* reflected the two infrageneric sections retrieved in plastome and rDNA trees (Figure 1a; Figure S3). The remaining 21 clusters were exclusively present in *F. proserpinacoides* (Figure 2b), congruently with the bigeneric phylogeny of Limnanthaceae (Figure 1a).

To support the plastome- and rDNA-based phylogenetic hypotheses, we reconstructed the phylogenetic relationships among five Limnanthaceae taxa based on the sequence similarities of all repeat types. Neighbor-joining (NJ) trees were constructed based on 29 of the top 100

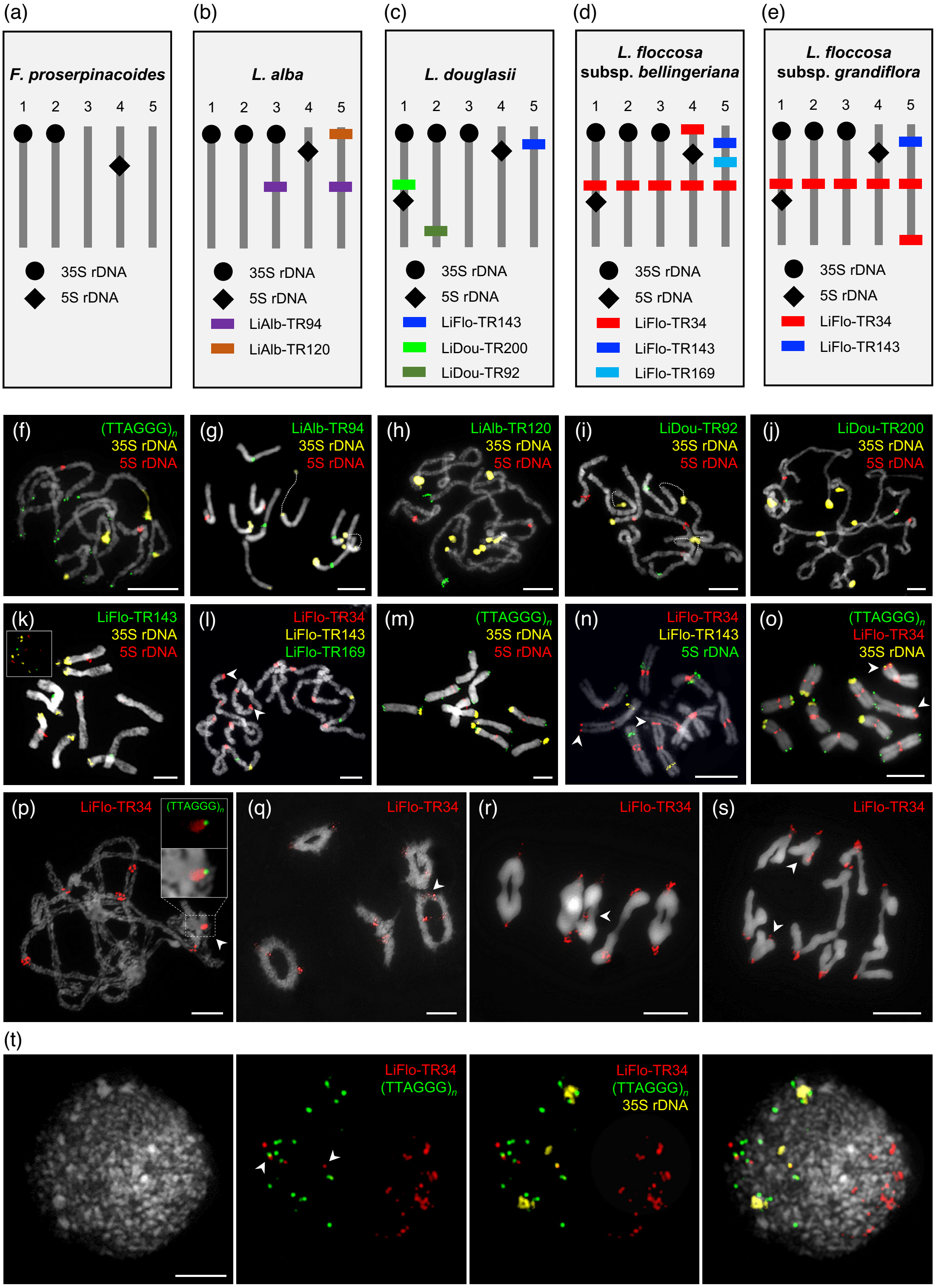

(a)
F. proserpinacoides
1 2 3 4 5
● 35S rDNA
◆ 5S rDNA
(b)
L. alba
1 2 3 4 5
● 35S rDNA
◆ 5S rDNA
LiAlb-TR94
LiAlb-TR120
(c)
L. douglasii
1 2 3 4 5
● 35S rDNA
◆ 5S rDNA
LiFlo-TR143
LiDou-TR200
LiDou-TR92
(d)
L. floccosa subsp. bellingeriana
1 2 3 4 5
● 35S rDNA
◆ 5S rDNA
LiFlo-TR34
LiFlo-TR143
LiFlo-TR169
(e)
L. floccosa subsp. grandiflora
1 2 3 4 5
● 35S rDNA
◆ 5S rDNA
LiFlo-TR34
LiFlo-TR143
(f) (TTAGGG)$_n$ 35S rDNA 5S rDNA
(g) LiAlb-TR94 35S rDNA 5S rDNA
(h) LiAlb-TR120 35S rDNA 5S rDNA
(i) LiDou-TR92 35S rDNA 5S rDNA
(j) LiDou-TR200 35S rDNA 5S rDNA
(k) LiFlo-TR143 35S rDNA 5S rDNA
(l) LiFlo-TR34 LiFlo-TR143 LiFlo-TR169
(m) (TTAGGG)$_n$ 35S rDNA 5S rDNA
(n) LiFlo-TR34 LiFlo-TR143 5S rDNA
(o) (TTAGGG)$_n$ LiFlo-TR34 35S rDNA
(p) LiFlo-TR34 (TTAGGG)$_n$
(q) LiFlo-TR34
(r) LiFlo-TR34
(s) LiFlo-TR34
(t)
LiFlo-TR34 (TTAGGG)$_n$
LiFlo-TR34 (TTAGGG)$_n$ 35S rDNA

Figure 3. Schematic repeat-based karyotypes and chromosomal localization of tandem repeats in *Limnanthes* and *Floerkea* species. (a–n) Mitotic chromosomes of *Floerkea proserpinacoides* (a, f), *Limnanthes alba* (b, g, h), *Limnanthes douglasii* (c, i–k), *Limnanthes floccosa* subsp. *bellingeriana* (d, l, m) and *Limnanthes floccosa* subsp. *grandiflora* (e, n) probed with identified tandem repeats and rDNA probes. (o–t) Mitotic chromosomes, the first meiotic division and an interphase nucleus of *L. floccosa* subsp. *bellingeriana* probed with the centromeric 173-bp repeat LiFlo-TR34 (red), Arabidopsis-type telomeric repeat (green) and 35S rDNA (yellow). (o) Mitotic chromosomes. (p) Pachytene. (q) Diakinesis. (r) Metaphase I. (s) Anaphase I. (t) Interphase. Arrowheads indicate the terminal locus of LiFlo-TR34 on chromosome 4. Chromosomes and nuclei were isolated from young anthers and counterstained by DAPI. Detailed information on the localized repeats is provided in Table S2. Scale bars, 10 μm.

repeat clusters from the comparative clustering analysis. A filtered supernetwork based on 29 NJ trees separated *Floerkea* and *Limnanthes*, and supported the two sections within *Limnanthes* (Figure S7).

Repeatome variation across Brassicales

We performed a comparative clustering analysis using reads from the Limnanthaceae and closely related Brassicales families (i.e., Bataceae, Caricaceae, Moringaceae and Setchellanthaceae) and Brassicaceae. The results showed that most clusters represent family-specific repeats. Although Ty3-*gypsy* elements are present in all the Brassicales genomes, they proliferated extremely in Limnanthaceae (Figure S8).

Interphase nuclear organization in *Limnanthes*

In our previous study, we demonstrated that the 35S rDNA probe can be used for *in situ* detection of nucleoli in the Brassicaceae (Shan et al., 2021). The 35S rDNA probe also proved to be a reliable indicator of nucleoli in Limnanthaceae (Figure S9). In the two subspecies of *L. floccosa*, Carnoy's fixed interphase nuclei isolated from young anthers were hybridized with telomeric, centromeric LiFlo-TR34 and 35S rDNA probes. Polarized (Rabl-like) positioning of centromeres and telomeres at the opposite nuclear poles was observed in 83% and 86% of nuclei in subsp. *bellingeriana* and subsp. *grandiflora*, respectively (Figure 3t). Telomeric signals were frequently clustered with 35S rDNA loci (nucleolus), whereas centromeres were positioned within the more heterochromatic opposite pole. The more heterochromatic ('centromeric') pole was also clearly visible as a nuclear region more densely stained with 4′,6-diamidino-2-phenylindole (DAPI) (Figures 1g and 3t). In some nuclei (17% and 14% in subsp. *bellingeriana* and subsp. *grandiflora*, respectively), the centromeres were localized at one pole, whereas the telomeres were scattered in the nuclear interior, with no obvious connection to the nucleolus.

To further analyze the spatial arrangement of centromeres and telomeres in *L. floccosa* subsp. *bellingeriana*, their distribution was further investigated by 3D FISH using paraformaldehyde-fixed interphase nuclei isolated from three different tissues (root tips, stem leaves and petals) and embedded in polyacrylamide pads. In the majority of leaf nuclei, 3D FISH showed that centromeres were located at one nuclear pole, whereas telomeres and 35S rDNA (nucleolus) were located at the opposite pole (Rabl-like configuration, 81%; Figure 4a; Table S3; Movie S1). In the minority of leaf nuclei, centromeres were located at one pole and telomeres were scattered throughout the nuclear interior ('centromeric polarization', 16%; Figure S10a) or centromeres were located at the nuclear periphery and telomeres were associated with the nucleolus ('rosette-like organization', 3%; Figure S10b). Of nuclei isolated from petals, 77% exhibited the Rabl-like pattern (Figure 4b, Movie S2), whereas 23% of nuclei had both centromeres and telomeres scattered within the nucleus ('dispersed distribution'; Figure S10c). Most of the nuclei isolated from root-tip tissue exhibited the Rabl-like pattern (93%; Figure 4c; Movie S3). In addition to the dominant spherical nuclei, the much rarer spindle-shaped nuclei also predominantly exhibited the Rabl-like configuration (data not shown). In very few nuclei, the centromeres were scattered throughout the nuclear interior and the telomeres were located together with the nucleolus at one nuclear pole ('telomeric polarization', 3%; Figure S10d), or both the centromeric and telomeric probes were scattered within the nuclear interior ('dispersed distribution', 3%; Figure S10e; Table S3).

In all tissues analyzed by 3D FISH the number of centromeric and telomeric signals was, on average, lower than theoretically expected. Instead of the expected 10 centromeric and 20 telomeric signals, typically four centromeres (2–6) and 10 telomeres (4–17) were observed in leaf tissues, three centromeres (2–6) and 12 telomeres (6–19) were observed in petal tissues and five centromeres

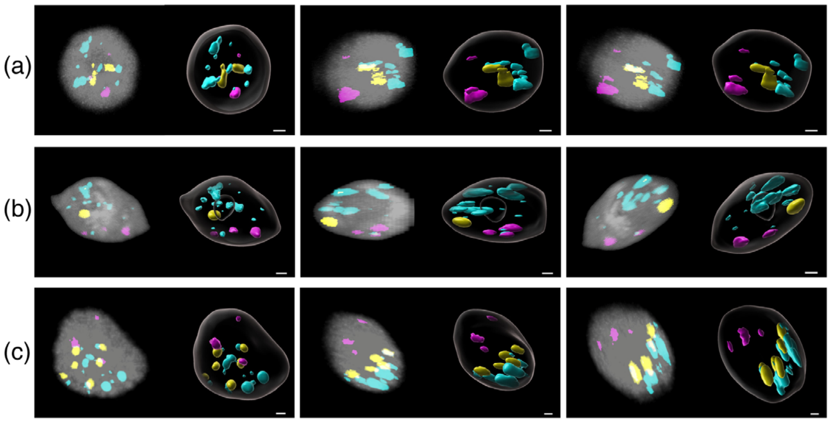


Figure 4. Three-dimensional fluorescence *in situ* hybridization (3D FISH) in *Limnanthes floccosa* subsp. *bellingeriana*. *In situ* localization and corresponding Imaris 3D projection of centromeric (LiFlo-TR34, magenta), telomeric ($(TTTAGGG)_n$, cyan blue) and 35S rDNA (yellow) repeats in paraformaldehyde-fixed nuclei isolated from leaves (a), petals (b) and root tips (c). Nuclei were counterstained with DAPI (gray). Scale bars, 1 μm.

© 2022 Society for Experimental Biology and John Wiley & Sons Ltd.,

(2–9) and 16 telomeres (7–24) were observed in root-tip tissues (Figure 4).

DISCUSSION

The backbone phylogeny of Limnanthaceae

Mason (1952) divided *Limnanthes* into two species sections – sections *Inflexae* and *Reflexae* (section *Limnanthes*; Meyers et al., 2010) – based on inflexing and reflexing petals after fertilization. The two infrageneric clades were confirmed by phylogenetic analysis using one nuclear gene and two chloroplast loci and comprehensive taxon sampling (Meyers et al., 2010). Here, using complete chloroplast sequences, nuclear rDNA genes and identified repeats, we confirmed the sister relationship of *Floerkea* and *Limnanthes* and retrieved two strongly supported infrageneric clades in the latter genus. Although not fully supported, section *Limnanthes* (*L. douglasii*) consistently had a more ancestral position than section *Inflexae* (*L. alba*, *L. floccosa*) (Figure 1a; Figure S3). The fully resolved backbone phylogeny of Limnanthaceae provides the basis for future comparative studies.

Genome evolution in Limnanthaceae

Among all Brassicales families, five chromosome pairs ($n = 5$) have been identified so far only in all Limnanthaceae species and a handful of Brassicaceae species including *A. thaliana* (Lysak, 2018). In the Brassicaceae, the five chromosomes arose during genome-wide diploidization after the family-specific At-α WGD (*A. thaliana*) or after younger clade-specific mesopolyploid WGDs (e.g., in *Stenopetalum lineare*; Mandáková et al., 2010). Post-polyploid diploidization in the lineage leading to *A. thaliana* was accompanied by chromosomal rearrangements and genome downsizing resulting in the small 135-Mb Arabidopsis genome with an average chromosome size of 32 Mb. In contrast, the genomes (1500–2100 Mb) and chromosomes (340–420 Mb) of Limnanthaceae species are at least 10 times larger. In the absence of evidence for a family-specific WGD (Edger et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019), the non-coding DNA of Limnanthaceae genomes must have originated either from the At-β WGD or from the proliferation of transposable elements (TEs) after the genome duplication but most likely before the *Floerkea*–*Limnanthes* divergence. In either case, selective purging of TEs and/or suppression of their activity may have been less effective in Limnanthaceae. The relatively large nuclear genomes might be tentatively associated with autumn or winter germination of these hardy annuals and their general tolerance to low (spring) temperatures (e.g., Baskin et al., 1988; Houle, 2002).

The chromosome numbers and genome sizes of Limnanthaceae species are particularly interesting when compared with other closely related Brassicales families (Figure 1a). Although the monospecific Setchellanthaceae (*S. caeruleus*, native to Mexico) appears to share the At-β WGD with Limnanthaceae (Edger et al., 2018) and both families are closely related (Edger et al., 2018; Li et al., 2021), the unknown genome size and chromosome number in *S. caeruleus* preclude a comparison of the post-At-β genome evolution in both New World families. Divergence of two other Brassicales families, Caricaceae and Moringaceae, occurred prior to the At-β genome duplication (Edger et al., 2018), so comparisons with Limnanthaceae genomes can provide only limited insights. The ancestral chromosome number of the Caricaceae (six genera/35 species) has been inferred to be $2n = 18$ (Rockinger et al., 2016), whereby this number has been maintained in papaya (*C. papaya*) but reduced to $2n = 16$ or $2n = 14$ in other genera due to descending dysploidy. Interestingly, Rockinger et al. (2016) and Zerpa-Catanho et al. (2021) found more than two-fold genome size variation (401–1022 Mb) within the Caricaceae that was not related to polyploidization. Among the Moringaceae (one genus/13 species), the drumstick tree (*Moringa oleifera*) has 14 chromosome pairs ($2n = 28$) and a relatively small genome size (217–315 Mb; Chang et al., 2019; Tian et al., 2015). Both Caricaceae and Moringaceae did not undergo an additional genome duplication after the At-γ whole-genome triplication (Tian et al., 2015).

In summary, the currently available data do not allow us to reconstruct the chromosome number of the most recent common ancestor of Limnanthaceae. We can safely assume that the ancestral chromosome number was higher than $2n = 10$ and that the five chromosome pairs arose due to post-polyploid descending dysploidy associated with TE amplification. A chromosome-level genome assembly of a *Limnanthes* species should provide information on the course of post-polyploid diploidization including the structure of the At-β paleotetraploid genome.

Repeatomes and genome size variation in Limnanthaceae and Brassicales

In Brassicales, genome sizes range from 135 Mb to 4.6 Gb (Lysak, 2018) and are largely determined by the proportion of non-coding and repetitive sequences (Elliott & Gregory, 2015; Hloušková et al., 2019). Using low-coverage sequencing data from 71 Brassicales taxa, Beric et al. (2021) confirmed that repeat content, along with the gene content and tandem repeats, is an important contributor to genome size variation in the order. In Brassicaceae and Cleomaceae, TEs account for 21% of the Arabidopsis genome (Quesneville, 2020) and 43% of Spider Flower (*Cleome hassleriana*) genome (Cheng et al., 2013) (Figure 5). In contrast, repeatomes account for 58–66% of the Limnanthaceae genomes (Figure 2a).

Here we have shown that Limnanthaceae genomes expanded through chromosome-wide amplification of LTR

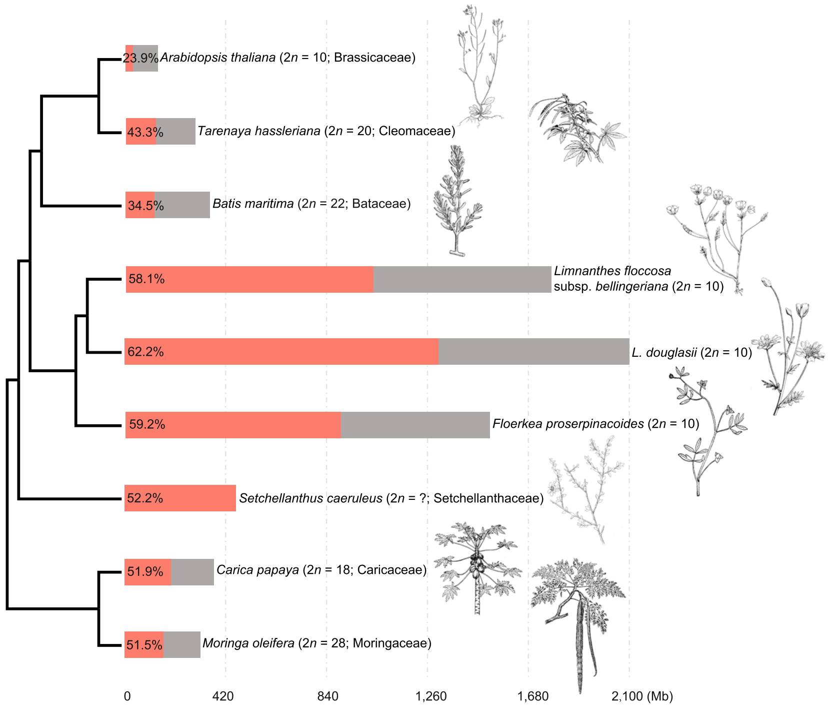


Figure 5. Nuclear genome and repeatome size variation in Limnanthaceae and Brassicales. The simplified plastome-based tree showing phylogenetic relationships among selected Brassicales genomes was adapted from Figure 1a. Red bars correspond to repetitive fractions (%) of the nuclear genomes (data source: *Arabidopsis thaliana*, Wang et al., 2021*; Carica papaya*, Ming et al., 2008; *Moringa oleifera*, Tian et al., 2015; *Tarenaya hassleriana*, Cheng et al., 2013; repeatome proportions in the remaining genomes were estimated by RepeatExplorer using low-coverage sequence data from Li et al., 2021, Beric et al., 2021 and this study). Genome size and chromosome number of *S. caeruleus* are unknown.

retrotransposons, particularly Ty3-*gypsy* elements. In contrast, tandem repeats did not amplify and/or were purged (Figure 2a). Since no significant shifts in TE abundance were associated with WGDs in Brassicales (Beric et al., 2021), genome obesity in Limnanthaceae is likely due to the proliferation of TEs rather than the At-β genome duplication. Recombination-based processes are known to remove repeats from the genome (Novák et al., 2020). For example, recent studies have shown that the lowest rates of unequal recombination between the long terminal repeats of LTR retrotransposons were found in the largest genomes analyzed (Charlesworth et al., 1994; Cossu et al., 2017; Jedlicka et al., 2020). DNA repair, particularly non-homologous end joining, is thought to play a role in controlling the rate of DNA removal, with larger deletions occurring in small plant genomes (Novák et al., 2020; Vu et al., 2017). Therefore, in species with large genomes, the slow degradation of repeats could lead to ever-increasing accumulation of repeat sequences (Figure 5; Kelly et al., 2015; Maumus & Quesneville, 2014). Future whole-genome sequencing and assemblies will elucidate the origin of the five large chromosome pairs in Limnanthaceae.

Different nuclear organization patterns in Brassicales

Among the Brassicales, chromatin organization in interphase nuclei has been analyzed only in the Brassicaceae family, with *A. thaliana* being the most extensively analyzed species (e.g., Fransz et al., 2002; Pecinka et al., 2004; Pontvianne & Grob, 2020; Shan et al., 2021). In the small Arabidopsis genome (135 Mb), telomeres within interphase nuclei generally associate with the nucleolus, while centromeres are positioned peripherally, at the nuclear membrane (Armstrong et al., 2001). In Brassicaceae species with large nuclear genomes (2600–4300 Mb) and a small number of chromosomes ($n = 6$, 7), the spatial arrangement of centromeres and telomeres resembles the Rabl pattern or they are scattered in the nuclear interior. It has been suggested that these chromatin configurations in interphase may be due to the small number of large chromosomes that lack the distinct longitudinal compartmentalization typical of small Brassicaceae genomes (Shan et al., 2021). In *Limnanthes*, the predominant nuclear phenotype resembles the polarized Rabl configuration, in which the centromeres are usually located at one nuclear pole and the telomeres, together with the nucleolus (or nucleoli), at the opposite pole. The position of the nucleolus at the telomeric nuclear pole reflects the fact that three of the five *Limnanthes* chromosomes carry terminal NORs. The Rabl configuration in Limnanthaceae genomes is congruent with their genome sizes ranging between 1700 and 2100 Mb, and only five predominantly (sub)metacentric, large (340–420-Mb) chromosomes. Because Rabl organization resembles chromosome configuration in the mitotic anaphase, centromere–telomere polarization in Limnanthaceae species nuclei could be mechanistically interpreted as an effective arrangement of long metacentric (V-shaped) chromosomes within the limited nuclear space, possibly reducing topological entanglement of chromatin fibers

© 2022 Society for Experimental Biology and John Wiley & Sons Ltd.,

(Pouokam et al., 2019). Indeed, chromosome length, not just genome size (Dong & Jiang, 1998), may be a more important factor determining the Rabl configuration of interphase chromosomes (e.g., Saunders & Houben, 2001; Shan et al., 2021).

EXPERIMENTAL PROCEDURES

Plant material

A list of all analyzed Limnanthaceae accessions and GenBank accessions is provided in Table S4.

DNA sequencing

Genomic DNA was extracted from young leaves of five Limnanthaceae accessions using the NucleoSpin Plant II kit (Macherey-Nagel, Dueren, Germany). DNA sequencing libraries were prepared and subsequently sequenced using the Illumina NextSeq platform (150-bp paired-end reads). ONT long-read sequencing was performed for *L. douglasii*. High-molecular-weight genomic DNA was extracted using the CTAB-based protocol adapted from Healey et al. (2014) and then treated using the Short Read Eliminator depletion kit (Circulomics, Menlo Park, CA, USA). The Ligation Sequencing Kit (Sqk Lsk109) was used to prepare the sequencing library following the manufacturer's protocol, which was sequenced in a MinION device. Both Illumina and Nanopore DNA libraries were sequenced at CEITEC's core facility Genomics.

Genome size estimation

Holoploid genome size was estimated by flow cytometry in *F. proserpinacoides*, *L. douglasii*, *L. floccosa* subsp. *bellingeriana* and *L. floccosa* subsp. *grandiflora*. Nuclear suspension was prepared from fully developed intact leaf tissue according to Doležel et al. (2007), and isolated nuclei were stained using propidium iodide and RNAase IIA (both 50 μg/ml) for 5 min at room temperature and analyzed using a Partec CyFlow cytometer; the fluorescence intensity of 5000 particles was recorded. *Solanum pseudocapsicum* (1C = 1.30 pg) served as the primary reference standard. One individual of each accession measured on three consecutive days was analyzed.

Chloroplast genome and 35S rDNA assembly, and phylogenetic analysis

The Illumina raw reads were filtered and trimmed using fastp-v0.20.1 software (Chen et al., 2018) with the following parameters: -z 4 -q 20 -u 30 -n 0 -f 4 -t 6 --length_required 140 -b 140. The complete cp genomes of five Limnanthaceae taxa were generated using GetOrganelle (Jin et al., 2020). The Limnanthaceae plastomes were annotated using Plann software (Huang & Cronk, 2015) and manually curated by Sequin software. To search for 35S rDNA sequences, the reads were assembled using GetOrganelle with the following parameters: -R 10 -t 15 -k 21, 35, 45, 65, 85, 115 -F embplant_nr. The transcription unit (18S-ITS1-5.8S-ITS2-26S) was selected for phylogenetic analysis.

Chloroplast protein-coding sequences of an additional 17 species representing major Brassicales lineages were retrieved from Edger et al. (2018). We utilized BLASTn combined with published data from Brassicales to parse target protein-coding sequences in Limnanthaceae species. Multiple sequence alignments were generated using MAFFT v7.450 (Katoh & Standley, 2013) and columns were removed (i.e., across all taxa) if a base-pair position was missing in one species using Gblocks v0.91b (Talavera & Castresana, 2007). We obtained a complete 43 263-bp data matrix derived from 72 different protein-coding regions of the plastid genome. For 35S rDNA, seven sequences were aligned, including those of *A. thaliana*, *S. caeruleus* and five Limnanthaceae accessions. Plastome protein-coding gene matrixes and 35S rDNA alignment were subsequently used to construct ML trees using IQ-TREE v1.6.10 (Nguyen et al., 2015). In addition, BI trees were constructed using the NGPhylogeny.fr portal.

Clustering analysis of repetitive DNA

Raw sequencing reads were pre-processed as described above. A quality check of paired-end reads was carried out using FastQC software. All reads were trimmed to 140 nucleotides for clustering analysis. Conversion of reads format from fastq to fasta was performed using the 'sed' command with parameters sed -n '1~4 s/^@/>/p;2~4p', and paired reads were interlaced using seqtk with the mergepe option. Repeatome analysis was performed through similarity-based clustering analysis using the RepeatExplorer platform (Novák et al., 2013). The number of reads representing 0.1× genome coverage were sampled and analyzed for each species. Default settings were used for each clustering analysis. Repeat clusters with a genome proportion of >0.01% were further annotated in detail. Tandem repeat analyzer TAREAN (Novák et al., 2017) was used to identify consensus monomer sequences of satellites.

For comparative clustering analysis, 500 000 reads from each Limnanthaceae accession were sampled. Additionally, 250 000 reads from each Brassicales species were sampled. The settings for the comparative analysis were the same as those for the individual clustering analysis. However, only repeat clusters with a genome proportion of >0.05% were annotated in detail for further analysis. Custom R scripts (Novák et al., 2020) were used to construct a graphical representation of repeat distribution between the species as proportionally scaled rectangles representing the number of reads in a given cluster.

Repetitive sequence similarity-based phylogeny

Comparative clustering analysis of five Limnanthaceae accessions was performed by RepeatExplorer2 with default parameters. The abundant repeat clusters (genome proportion > 0.05%) were employed for phylogenetic analysis. The repetitive sequence similarity matrices obtained from the comparative clustering analysis were employed to infer phylogenetic relationships (Vitales et al., 2020). Briefly, the more similar repeats of two species have a higher number of edges between the reads of those species; these similarity matrices were transformed into distance matrices. Then, the pairwise distance matrices were used to construct an NJ tree for each cluster by using the NJ function in the ape package. Finally, a filtered consensus network was reconstructed in Newick format from all NJ trees using SplitsTree5 (Bagci et al., 2021). Custom R scripts were used to process RepeatExplorer2 output results and phylogenetic analyses.

Processing of ONT reads

The raw ONT reads were basecalled using Guppy (ver. 2.3.1) with the following parameters: --flowcell FLO-MIN106 --kit SQK-LSK109 -r --num_callers 10 --cpu_threads_per_caller 12. NanoPlot software (Coster et al., 2018) was used for quality checking to show a bivariate plot comparing log transformed read length with average basecall Phred quality score. NanoFilt software (Coster et al., 2018) was used for read filtering and trimming with the following parameters: -l 1000 --headcrop 10 -q 7. To quantify chloroplast structural heteroplasmy in *L. douglasii*, ONT reads (longer than the IR length) were mapped to the two different structural haplotypes (A and B) of the *L. douglasii* plastome using BLASTn. Further statistics analyses were implemented using custom Python scripts.

DNA probes

A list of all designed oligo probes and primers specific for repetitive elements is provided in Table S2. Synthetic oligonucleotide probes were used for tandem repeats with shorter monomers (<500 bp). Target sequences (60 nt) with GC content 30–50% were selected from DNA alignments using the Geneious v11.1.5 software package (https://www.geneious.com) to minimize self-annealing and the formation of hairpin structures. For satellites with longer monomers and retrotransposons, PCR primers were designed to face outwards from the monomer; therefore, PCR amplification was performed only between tandemly arrayed monomers.

PCR products were purified using NucleoSpin Gel and PCR Clean-up kits (Macherey-Nagel). The BAC clone T15P10 (AF167571) of *A. thaliana* (Arabidopsis) bearing 35S rRNA gene repeats was used for *in situ* localization of NORs, and the Arabidopsis clone pCT4.2 (M65137), corresponding to a 500-bp 5S rDNA repeat, was used for localization of 5S rDNA loci. Telomeric sequences were detected using the Arabidopsis-type telomere repeat $(TTTAGGG)_n$ amplified following Ijdo et al. (1991).

Two-dimensional fluorescence in situ hybridization

For preparations of mitotic and meiotic chromosomes, as well as interphase nuclei, young inflorescences were collected from plants in the field (*Floerkea*) or cultivated in the greenhouse (*Limnanthes*). The inflorescences were fixed in freshly prepared fixative (ethanol:acetic acid, 3:1) overnight, transferred into 70% ethanol and stored at −20°C until use. Chromosome and/or nuclear spreads were prepared from immature anthers following the published protocol (Mandáková & Lysak, 2016a). All DNA probes were labeled with biotin-dUTP, digoxigenin-dUTP or Cy3-dUTP by nick translation as described previously (Mandáková & Lysak, 2016b). Briefly, 20 μl of the hybridization mix containing labeled DNA probes (100 ng each) dissolved in 50% formamide and 10% dextran sulfate in 2× SSC was pipetted on a suitable chromosome-containing slide and immediately denatured on a hot plate at 80°C for 2 min. FISH was carried out in a moist chamber at 37°C for 24 h. Post-hybridization washing was performed in 20% formamide in 2× SSC at 42°C. The immunodetection of hapten-labeled probes was performed as described by Mandáková and Lysak (2016b): biotin-dUTP was detected by avidin–Texas Red (Vector Laboratories, Burlingame, CA, USA) and the signal was amplified by biotin-conjugated goat anti-avidin (Vector Laboratories); digoxigenin-dUTP was detected by mouse anti-digoxigenin (Jackson ImmunoResearch, West Grove, PA, USA) and Alexa Fluor 488-conjugated goat anti-mouse (Invitrogen, now Thermo-Fisher Scientific, Waltham, MA, USA). The estimated stringency of FISH was 80–85%. Chromosomes were counterstained with DAPI (2 μg/ml) in Vectashield. The preparations were photographed using a Zeiss Axioimager Z2 epifluorescence microscope with a CoolCube camera (MetaSystems, Altlussheim, Germany). Images were acquired separately for each fluorochrome using appropriate excitation and emission filters (AHF Analysentechnik, Tübingen, Germany). Individual monochromatic images were pseudocolored, merged and cropped using Photoshop CS (Adobe Systems, San Jose, CA, USA).

Three-dimensional fluorescence *in situ* hybridization

Freshly harvested tissues (stem leaves, root tips and petals) were used to prepare paraformaldehyde (PFA)-fixed suspension nuclei following Shan et al. (2021). Polyacrylamide gel mix (80 mM KCl, 20 mM NaCl, 15 mM PIPES-NaOH, 0.5 mM EGTA, 2 mM EDTA, 80 mM sorbitol, 1 mM DTT, 0.15 mM spermine, 0.5 mM spermidine and 15% acrylamide/bis solution 29:1) was prepared in an Eppendorf tube. Freshly prepared 20% ammonium persulfate and 20% sodium sulfite solution were added to the mixture, which was quickly vortexed. After that, 13 μl of the PFA-fixed nuclear suspension was pipetted on a silane-coated slide (Sigma-Aldrich, St. Louis, MO, USA), followed by adding 6.5 μl of the polyacrylamide gel mix and mixing. The mixture was covered with a 24 × 24 mm coverslip and polymerized at room temperature for 1 h. The coverslip was removed using a razor blade. To get rid of unpolymerized gel, 200 μl of buffer A salts (80 mM KCl, 20 mM NaCl, 15 mM PIPES-NaOH, 0.5 mM EGTA, 2 mM EDTA, 80 mM sorbitol, 1 mM DTT, 0.15 mM spermine and 0.5 mM spermidine) was pipetted onto the polyacrylamide pad (Howe et al., 2013; Hurel et al., 2018). Then, 20 μl of the labeled probe was pipetted on the polyacrylamide pad and immediately denatured on a hot plate at 96°C for 6 min. Hybridization was carried out in a moist chamber at 37°C for approximately 48 h. Post-hybridization washing was performed in 0.1× SSC at 42°C, 2× SSC at 42°C, 2× SSC at room temperature and 4× SSC at room temperature by shaking on an orbital shaker (5 min each step). The immunodetection of hapten-labeled probes was carried out as described above for 2D FISH. After immunodetection, the preparations were stained with DAPI (2 μg/ml) in Vectashield, covered with a precision coverslip (22 × 22 mm) and sealed with nail polish. Fluorescence signals were photographed using a Zeiss Axio Observer.Z1 laser scanning microscope with LSM 780 laser scanning unit. Nuclei of comparable size and shape (spherical or oval) were analyzed preferentially. Scanning and deconvolution were performed using ZEN BLUE (Carl Zeiss, Oberkochen, Germany). IMARIS (Oxford Instruments, Abingdon, UK) was used for channel contrast adjustment (function 'Channel Adjustment'), projection of the centromere–telomere arrangement (function 'Surface') and creation of videos (function 'Animation').

Immunodetection of fibrillarin

Polyacrylamide pads containing PFA-fixed suspension of leaf nuclei were prepared as described above and incubated first with 5% bovine serum albumin (BSA) solution in 4× SSC with 0.2% Tween-20 at 37°C for 30 min and subsequently with 100 μl of anti-fibrillarin (Abcam; 1:100 in BSA) at 37°C overnight. After washing twice in 2× SSC (5 min each step), samples were incubated with Alexa Fluor 488-conjugated goat anti-mouse at 37°C for 30 min, followed by washing twice in 2× SSC (5 min each step). Immediately after the washing steps, nuclei were counterstained with DAPI (2 μg/ml) and photographed as described above.

ACKNOWLEDGMENTS

We acknowledge the core facility CELLIM and Genomics of CEITEC, Masaryk University, supported by the MEYS CR (project no. LM2018129), and CEITEC's core facility Plant Sciences. Computational resources were supplied by the project 'e-Infrastruktura CZ' (project no. LM2018140) supported by the MEYS CR. The authors acknowledge networking support from the 'INDEPTH' COST Action CA16212. We thank Anthony Reznicek for providing the leaf and flower material of *Floerkea*, Pavel Trávníček for genome size measurements and Milan Pouch and Petra Hloušková for DNA extraction and their assistance with probe design. This work was supported by the Czech Science Foundation (project no. 18-20134S) and Operational Programme Research, Development and Education – 'Project Internal Grant Agency of Masaryk University' (No. CZ.02.2.69/0.0/0.0/19_073/0016943).

AUTHOR CONTRIBUTIONS

MAL designed the research; SZ, MK and TM performed the experiments and analyzed the data; MAL, SZ, TM and MK wrote the manuscript.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All raw reads generated in this study are available from the NCBI database under BioProject PRJNA783938. The assembled cp genomes and 35S rDNA sequences are available from GenBank under accession numbers ON088447–ON088451 and OM970789–OM970794, respectively. All data included in this study are available within the paper and its supporting information published online.

SUPPORTING INFORMATION

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

Figure S1. Characterization of two natural haplotypes of the chloroplast genome in *L. douglasii*.

Figure S2. Bayesian analysis phylogeny of 72 plastome protein-coding genes in Limnanthaceae.

Figure S3. Maximum likelihood and Bayesian inference phylogenies based on 35S rDNA sequences of Limnanthaceae taxa.

Figure S4. Chromosomal localization of retrotransposons in *L. douglasii*.

Figure S5. Sphere-shaped graphs of LiFlo-TR34 satellite in *L. floccosa* generated by RepeatExplorer.

Figure S6. A head-to-tail organization and repeat profiles of the LiFlo-TR34 satellite in *L. floccosa* accessions.

Figure S7. Phylogeny based on repeat sequence similarities.

Figure S8. Comparative repeat profiles of representative Brassicales genomes.

Figure S9. Localization and corresponding Imaris 3D projection of the anti-fibrillarin antibody and 35S rDNA probe in paraformaldehyde-fixed leaf nuclei of *L. floccosa* subsp. *bellingeriana*.

Figure S10. Minor nuclear configurations in paraformaldehyde-fixed tissues of *L. floccosa* subsp. *bellingeriana*.

Table S1. Individual repeatome composition of the five Limnanthaceae taxa.

Table S2. Primers and oligos used in this study.

Table S3. Different patterns of interphase nuclear organization in *L. floccosa* subsp. *bellingeriana*.

Table S4. The origin and GenBank accession numbers of the analyzed Limnanthaceae taxa.

Movie S1. 3D visualization of dominating Rabl nuclear organization pattern in leaf tissue of *L. floccosa* subsp. *bellingeriana*.

Movie S2. 3D visualization of dominating Rabl nuclear organization pattern in petal tissue of *L. floccosa* subsp. *bellingeriana*.

Movie S3. 3D visualization of dominating Rabl nuclear organization pattern in root-tip tissue of *L. floccosa* subsp. *bellingeriana*.

REFERENCES

Agerbirk, N., Pattison, D., Mandáková, T., Lysak, M.A., Montaut, S. & Staerk, D. (2022) Ancient biosyntheses in an oil crop: glucosinolate profiles in *Limnanthes alba* and its relatives (Limnanthaceae, Brassicales). *Journal of Agricultural and Food Chemistry*, **70**, 1134–1147.

Armstrong, S.J., Franklin, F.C. & Jones, G.H. (2001) Nucleolus-associated telomere clustering and pairing precede meiotic chromosome synapsis in *Arabidopsis thaliana*. *Journal of Cell Science*, **114**, 4207–4217.

Arroyo, M.T.K. (1973) Chiasma frequency evidence on the evolution of autogamy in *Limnanthes floccosa* (Limnanthaceae). *Evolution*, **27**, 679–688.

Bagci, C., Bryant, D., Cetinkaya, B. & Huson, D.H. (2021) Microbial phylogenetic context using phylogenetic outlines. *Genome Biology and Evolution*, **1**, evab213.

Barker, M.S., Vogel, H. & Schranz, M.E. (2009) Paleopolyploidy in the Brassicales: analyses of the cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution*, **1**, 391–399.

Baskin, J.M., Baskin, C.C. & McCann, M.T. (1988) A contribution to the germination ecology of *Floerkea proserpinacoides* (Limnanthaceae). *Botanical Gazette*, **149**, 427–431.

Beric, A., Mabry, M.E., Harkess, A.E., Brose, J., Schranz, M.E., Conant, G.C. *et al.* (2021) Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales). *Genes, Genomes, Genetics*, **11**, jkab140.

Cardinal-McTeague, W.M., Sytsma, K.J. & Hall, J.C. (2016) Biogeography and diversification of Brassicales: a 103 million year tale. *Molecular Phylogenetics and Evolution*, **99**, 204–224.

Chang, Y., Liu, H., Liu, M. *et al.* (2019) The draft genomes of five agriculturally important African orphan crops. *GigaScience*, **8**, giy152.

Charlesworth, B., Sniegowski, P. & Stephan, W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, **371**, 215–220.

Chen, S., Zhou, Y., Chen, Y. & Gu, J. (2018) Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Cheng, S., Bergh, V.D.E., Zeng, P. *et al.* (2013) The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. *The Plant Cell*, **25**, 2813–2830.

Cossu, R.M., Casola, C., Giacomello, S., Vidalis, A., Scofield, D.G. & Zuccolo, A. (2017) LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. *Genome Biology and Evolution*, **9**, 3449–3462.

Coster, W.D., D'Hert, S., Schultz, D.T., Cruts, M. & Broeckhoven, C.V. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**, 2666–2669.

Doležel, J., Bartos, J., Voglmayr, H. & Greilhuber, J. (2003) Nuclear DNA content and genome size of trout and human. *Cytometry*, **51**, 127–128.

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

Dong, F. & Jiang, J. (1998) Non-Rabl patterns of centromere and telomere distribution in the interphase nuclei of plant cells. *Chromosome Research*, **6**, 551–558.

Edger, P.P., Hall, J.C., Harkess, A., Tang, M., Coombs, J., Mohammadin, S. *et al.* (2018) Brassicales phylogeny inferred from 72 plastid genes: a reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses. *American Journal of Botany*, **105**, 463–469.

Elliott, T.A. & Gregory, T.R. (2015) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **370**, 20140331.

Fransz, P., De Jong, J.H., Lysak, M., Castiglione, M.R. & Schubert, I. (2002) Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 14584–14589.

Fries, M. (1936) Über die Chromosomenzahl bei zwei Limnanthes-Arten. *Svensk Botanisk Tidskrift*, **30**, 440–442.

Healey, A., Furtado, A., Cooper, T. & Henry, R.J. (2014) Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods*, **10**, 1–8.

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

Houle, G. (2002) The advantage of early flowering in the spring ephemeral annual plant *Floerkea proserpinacoides*. *New Phytologist*, **154**, 689–694.

Howe, E.S., Murphy, S.P. & Bass, H.W. (2013) Three-dimensional acrylamide fluorescence in situ hybridization for plant cells. *Methods in Molecular Biology*, **990**, 53–66.

Huang, D.I. & Cronk, Q.C. (2015) Plann: a command-line application for annotating plastome sequences. *Applications in Plant Sciences*, **3**, 1500026.

Hurel, A., Phillips, D., Vrielynck, N., Mézard, C., Grelon, M. & Christophorou, N. (2018) A cytological approach to studying meiotic recombination and chromosome dynamics in *Arabidopsis thaliana* male meiocytes in three dimensions. *The Plant Journal*, **95**, 386–396.

Ijdo, J.W., Wells, R.A., Baldini, A. & Reeders, S.T. (1991) Improved telomere detection using a telomere repeat probe (TTAGGG)n generated by PCR. *Nucleic Acids Research*, **19**, 4780.

Jedlicka, P., Lexa, M. & Kejnovsky, E. (2020) What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Frontiers in Plant Science*, **11**, 644.

Jenderek, M.M. & Hannan, R.M. (2009) Diversity in seed production characteristics within the USDA-ARS *Limnanthes alba* germplasm collection. *Crop Science*, **49**, 1387–1394.

Jin, J.J., Yu, W.B., Yang, J.B., Song, Y., DePamphilis, C.W., Yi, T.S. *et al.* (2020) GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biology*, **21**, 1–31.

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

Kelly, L.J., Renny-Byfield, S., Pellicer, J. *et al.* (2015) Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytologist*, **208**, 596–607.

Laibach, F. (1907) Zur Frage nach der Individualität der Chromosomen im Pflanzenreich. *Botanisches Centralblatt*, **22**, 191–210.

Li, H.-T., Luo, Y., Gan, L., Ma, P.-F., Gao, L.-M., Yang, J.-B. *et al.* (2021) Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biology*, **19**, 1–13.

Lysak, M.A. (2018) Brassicales: an update on chromosomal evolution and ancient polyploidy. *Plant Systematics and Evolution*, **304**, 757–762.

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. & Lysak, M.A. (2016a) Chromosome preparation for cytogenetic analyses in *Arabidopsis*. *Current Protocols in Plant Biology*, **1**, 43–51.

Mandáková, T. & Lysak, M.A. (2016b) Painting of *Arabidopsis* chromosomes with chromosome-specific BAC clones. *Current Protocols in Plant Biology*, **1**, 359–371.

Mason, C.T. (1952) A systematic study of the genus *Limnanthes* R. Br. *Br. University of California Publications in Botany*, **25**, 455–512.

Maumus, F. & Quesneville, H. (2014) Deep investigation of *Arabidopsis thaliana* junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PLoS One*, **9**, e94101.

Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, J.G. et al. (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, **14**, 1–20.

Meyers, S.C., Liston, A. & Meinke, R. (2010) A molecular phylogeny of *Limnanthes* (Limnanthaceae) and investigation of an anomalous *Limnanthes* population from California, USA. *Systematic Botany*, **35**, 552–558.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H. et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, **452**, 991–996.

Negm, S., Greenberg, A., Larracuente, A.M. & Sproul, J.S. (2021) RepeatProfiler: a pipeline for visualization and comparative analysis of repetitive DNA profiles. *Molecular Ecology Resources*, **21**, 969–981.

Neumann, P., Novák, P., Hoštáková, N. & Macas, J. (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA*, **10**, 1–17.

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.

Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P. & Macas, J. (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*, **45**, e111.

Novák, P., Guignard, M.S., Neumann, P. *et al.* (2020) Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nature Plants*, **6**, 1325–1329.

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

One Thousand Plant Transcriptomes Initiative. (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, **574**, 679–685.

Palmer, J.D. (1983) Chloroplast DNA exists in two orientations. *Nature*, **301**, 92–93.

Pecinka, A., Schubert, V., Meister, A., Kreth, G., Klatte, M., Lysak, M.A. *et al.* (2004) Chromosome territory arrangement and homologous pairing in nuclei of *Arabidopsis thaliana* are predominantly random except for NOR-bearing chromosomes. *Chromosoma*, **113**, 258–269.

Pontvianne, F. & Grob, S. (2020) Three-dimensional nuclear organization in *Arabidopsis thaliana*. *Journal of Plant Research*, **133**, 479–488.

Pouokam, M., Cruz, B., Burgess, S., Segal, M.R., Vazquez, M. & Arsuaga, J. (2019) The Rabl configuration limits topological entanglement of chromosomes in budding yeast. *Scientific Reports*, **9**, 1–10.

Propach, H. (1934) Cytological investigations on *Limnanthes douglasii* R. Br. *Zeitschrift Zellforschung*, **21**, 357–375.

Quesneville, H. (2020) Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. *Mobile DNA*, **11**, 1–13.

Rockinger, A., Sousa, A., Carvalho, F.A. & Renner, S.S. (2016) Chromosome number reduction in the sister clade of *Carica papaya* with concomitant genome size doubling. *American Journal of Botany*, **103**, 1082–1088.

Saunders, V.A. & Houben, A. (2001) The pericentromeric heterochromatin of the grass *Zingeria biebersteiniana* ($2n = 4$) is composed of Zbcen1-type tandem repeats that are intermingled with accumulated dispersedly organized sequences. *Genome*, **44**, 955–961.

Shan, W., Kubová, M., Mandáková, T. & Lysak, M.A. (2021) Nuclear organization in crucifer genomes: nucleolus-associated telomere clustering is not a universal interphase configuration in Brassicaceae. *The Plant Journal*, **108**, 528–540.

Stenar, H. (1925) Embryologische und zytologische Studien über *Limnanthes douglasii* R.Br. *Svensk Botanisk Tidskrift*, **19**, 133–152.

Swanepoel, W., Chase, M.W., Christenhusz, M.J., Maurin, O., Forest, F. & Vanwyk, A.E. (2020) From the frying pan: an unusual dwarf shrub from Namibia turns out to be a new brassicalean family. *Phytotaxa*, **439**, 171–185.

Talavera, G. & Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564–577.

The Angiosperm Phylogeny Group IV. (2016) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, **181**, 1–20.

Tian, Y., Zeng, Y., Zhang, J., Yang, C., Yan, L., Wang, X. et al. (2015) High quality reference genome of drumstick tree (*Moringa oleifera* Lam.), a potential perennial crop. *Science China Life Sciences*, **58**, 627–638.

Tucker, G.C. (1993) Limnanthaceae. *Flora of North America Editorial Committee*, **7**, 172–183.

Vitales, D., Garcia, S. & Dodsworth, S. (2020) Reconstructing phylogenetic relationships based on repeat sequence similarities. *Molecular Phylogenetics and Evolution*, **147**, 106766.

Vu, G.T.H., Cao, H.X., Reiss, B. & Schubert, I. (2017) Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytologist*, **214**, 1712–1721.

Wang, B., Jia, Y., Jia, P., Dong, Q., Yang, X. & Ye, K. (2021) High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2021.08.003.

Wang, W. & Lanfear, R. (2019) Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants. *Genome Biology and Evolution*, **11**, 3372–3381.

Zerpa-Catanho, D.P., Jatt, T. & Ming, R. (2021) Karyotype and genome size determination of *Jarilla chocola*, an additional sister clade of *Carica papaya*. *Plant Omics*, **14**, 50–56.

https://doi.org/10.1093/plphys/kiac268
PLANT PHYSIOLOGY 2022: 190: 403–420

Plant Physiology®

Research Article

Genome diploidization associates with cladogenesis, trait disparity, and plastid gene evolution

Sheng Zuo (左胜),[1,2] Xinyi Guo (郭新异),[1] Terezie Mandáková,[1,3] Mark Edginton,[4] Ihsan A. Al-Shehbaz[5] and Martin A. Lysak[1,2,*]

1 CEITEC – Central European Institute of Technology, Masaryk University, Brno, CZ-625 00, Czech Republic
2 National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, CZ-625 00, Czech Republic
3 Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, CZ-625 00, Czech Republic
4 Queensland Herbarium, Department of Environment and Science, Brisbane Botanic Gardens, Mt Coot-tha Road, Toowong, QLD 4066, Australia
5 Missouri Botanical Garden, St. Louis, Missouri, 63110, USA

*Author for correspondence: martin.lysak@ceitec.muni.cz

These authors contributed equally (S.Z. and X.G.).

M.A.L. conceived the project. S.Z. and X.G. performed the analyses based on DNA sequence and morphological data. T.M. performed the cytogenomic experiments. M.E. and I.A.A. performed taxonomic revision of the taxa analyzed and contributed to the analyses of morphological disparity. M.A.L., S.Z., and X.G. wrote the article with contributions from the other authors.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/general-instructions) is Martin A. Lysak (martin.lysak@ceitec.muni.cz).

Abstract

Angiosperm genome evolution was marked by many clade-specific whole-genome duplication events. The Microlepidieae is one of the monophyletic clades in the mustard family (Brassicaceae) formed after an ancient allotetraploidization. Postpolyploid cladogenesis has resulted in the extant c. 17 genera and 60 species endemic to Australia and New Zealand (10 species). As postpolyploid genome diploidization is a trial-and-error process under natural selection, it may proceed with different intensity and be associated with speciation events. In Microlepidieae, different extents of homoeologous recombination between the two parental subgenomes generated clades marked by slow ("cold") versus fast ("hot") genome diploidization. To gain a deeper understanding of postpolyploid genome evolution in Microlepidieae, we analyzed phylogenetic relationships in this tribe using complete chloroplast sequences, entire 35S rDNA units, and abundant repetitive sequences. The four recovered intra-tribal clades mirror the varied diploidization of Microlepidieae genomes, suggesting that the intrinsic genomic features underlying the extent of diploidization are shared among genera and species within one clade. Nevertheless, even congeneric species may exert considerable morphological disparity (e.g. in fruit shape), whereas some species within different clades experience extensive morphological convergence despite the different pace of their genome diploidization. We showed that faster genome diploidization is positively associated with mean morphological disparity and evolution of chloroplast genes (plastid–nuclear genome coevolution). Higher speciation rates in perennials than in annual species were observed. Altogether, our results confirm the potential of Microlepidieae as a promising subject for the analysis of postpolyploid genome diploidization in Brassicaceae.

Received November 24, 2021. Accepted May 09, 2022. Advance access publication June 7, 2022
© American Society of Plant Biologists 2022. All rights reserved. For permissions, please email: journals.permissions@oup.com

Introduction

Brassicaceae (crucifers) is a cosmpolitan plant family occurring on all continents, except for Antarctica. Hybridization and polyploidization (or whole-genome duplication [WGD]) frequently accompanied the diversification of the Brassicaceae (Kagale et al., 2014; Hohmann et al., 2015; Mandáková et al., 2017a; Walden et al., 2020). The resulting polyploid genomes have not remained static, but returned to pseudo-diploid genomes through the process collectively named diploidization (Thomas et al., 2006), gradually erasing and concealing the signatures of ancient WGD events. In Brassicaceae, more than a dozen of genus- and clade-specific mesopolyploid WGDs, postdating the family-specific paleotetraploid (At-α) WGD (Bowers et al., 2003), were identified (Mandáková et al., 2017a); and even more genome duplications remain to be uncovered (e.g. Guo et al., 2021). While both the mesopolyploidization events and subsequent diploidizations have the potential to increase phenotypic diversity, it was suggested that no key morphological characters or innovations evolved after clade-specific WGDs in Brassicaceae (Walden et al., 2020). Still, fruits, trichomes, leaves, and embryos might be quite diverse even among closely related species, and thus, these characters have been used extensively for the past two centuries in the generic delimitations and tribal classifications in Brassicaceae. Tremendous diversity in fruit morphology, which is the most utilized organ in the classification of Brassicaceae, can be found even in small genera such as *Tropidocarpum* Hook (4 species; Al-Shehbaz, 2003), or a unigeneric tribe like the Eutremeae (44 spp.), harboring species with the shortest (2 mm) and longest (35 cm) fruits in the family. Indeed, two very different fruit shapes, such as a heart-shaped silicle (*Capsella* Medik.) and a cylindrical siliqua [*Arabidopsis* (DC.) Heynh.], may originate through different patterns of anisotropic growth, despite both closely related genera possess a cylindrically shaped gynoecium in the early phase of fruit development (Eldridge et al., 2016). On the other hand, cruciferous taxa are known for virtually every conceivable feature being subject to considerable convergence and reversals (Al-Shehbaz, 2012; Huang et al., 2016; Dong and Ostergaard, 2019; Nikolov et al., 2019). For instance, flat-shaped fruits evolved independently several times in Brassicaceae (Dong and Ostergaard, 2019), floral convergence between distantly related crucifer species may allow for exploitation of the same pollinators (Gómez et al., 2021) and two independently emerged *Capsella* species (*C. orientalis* Klokov and *C. rubella* Reut.) have undergone the convergent reduction of flower size (selfing syndrome) due to similar modulation of gene expression (Wozniak et al., 2020). Due to extensive family-wide morphological parallelism, lacking or incongruent molecular phylogenies, inferring phylogenetic relationships, especially at the tribal level, continue to be problematic.

Australia and New Zealand are home to many endemic crucifer species which, however, belong mostly to only four phylogenetic groups, namely *Barbarea* W.T.Aiton, *Cardamine* L., *Lepidium* L., and the tribe Microlepidieae (Mummenhoff et al., 2001; Heenan, 2017). Tribe Microlepidieae was expanded based on phylogenetic analyses to contain 16 genera and 56 species (Heenan et al., 2012). Only *Pachycladon* Hook.f. (11 species) is predominantly endemic to New Zealand (one species in Tasmania), whereas the other genera are indigenous to the Australian mainland and adjacent islands (e.g. Kangaroo Island, Tasmania). Among the 15 genera on the Australian mainland, 11 are mono- or oligospecific (i.e. with two or three species), whereas *Arabidella* (F.Muell.) O.E.Schulz (7 spp.), *Menkea* Lehm. (6 spp.), *Phlegmatospermum* O.E.Schulz (4 spp.), and *Stenopetalum* R.Br. ex DC. (10 spp.) harbor most species (Hewson, 1982; Heenan et al., 2012).

Attracted by chromosome numbers lower than in Arabidopsis [*Arabidopsis thaliana* (L.) Heynh., $2n = 10$], Mandáková et al. (2010a) analyzed chromosome complements of three Microlepidieae species by comparative chromosome painting (CCP) to find out that the bona fide diploid genomes ($2n = 8$, 10, and 12) originated through an unexpected WGD followed by genome-wide diploidization including descending dysploidy (DD), that is, reduction of chromosome number. A follow-up, more comprehensive phylogenomic study (Mandáková et al., 2017b) showed that the entire monophyletic tribe has descended from a common allotetraploid genome ($n = 15$) formed by an intertribal cross between parental species closely related to the extant tribes Crucihimalayeae (♀, $n = 8$) and Smelowskieae (♂, $n = 7$) during the Late Miocene. Following a long-distance dispersal from northeastern Asia or western North America, the mesotetraploid genome diversified into several clades on the Australian mainland (Mandáková et al., 2017b).

The postpolyploid diversification and diploidization in the Microlepidieae did not proceed with equal intensity throughout the tribe—three major clades distinguished by the level of diploidization were detected (Mandáková et al., 2017b). Whereas several genera possess highly reshuffled genomes and low chromosome numbers ($n = 4$–7; 2.1- to 3.75-fold DD from $n = 15$), *Pachycladon* experienced slower diploidization ($n = 10$; 1.5-fold DD) and some *Arabidella* species have the least diploidized genomes ($n = 12$; 1.25-fold DD). Remarkably, a two-fold difference in the level of diploidization was revealed among *Arabidella* species (Mandáková et al., 2017b). Whereas *Arabidella eremigena* (F. Muell.) E.A.Shaw has undergone major ("hot") diploidization ($n = 15 \rightarrow n = 6$), *Arabidella trisecta* (F.Muell.) O.E.Schulz showed the slowest ("cold") postpolyploid diploidization ($n = 15 \rightarrow n = 12$).

The available phylogenetic studies in Microlepidieae (Heenan et al., 2012; Mandáková et al., 2017b) clearly demonstrated the widespread convergent evolution of morphological characters used for the delimitation of genera and species in the tribe. A case in point, branched trichomes evolved independently at least 5 times, in *Harmsiodoxa* O.E.Schulz (3 spp.), *Microlepidium* F.Muell. (2 spp.), *Pachycladon*, *Pachymitus* O.E. Schulz (1 sp.), and *Stenopetalum*. Other features widely used taxonomically, particularly fruit shape, also exhibit tremendous convergence across the tribe. For instance, *A. eremigena* and *A. trisecta* share similar cylindrically shaped fruits, despite the fact that they differ markedly by their genome structure and phylogenetic position (Mandáková et al., 2017b).

Among the 13 + Brassicaceae clades of mesopolyploid origin (Mandáková et al., 2017a), Microlepidieae has become the tribe with the highest number of comparative genomic maps, and for which most extensive knowledge of postpolyploid genome evolution was acquired (Mandáková et al., 2010a, 2010b, 2017b), surpassing even the well-researched tribe Brassiceae (i.e. *Brassica* crops and their closest relatives). Therefore, the tribe has potential to be the subject for analysis of the course and impacts of postpolyploid genome diploidization including evolution of morphological traits across the diverse geography and climates of Australia and New Zealand.

The aim of this study is to further knowledge on the reticulate phylogenomic patterns and differently phased genome diploidization within Microlepidieae through robust phylogenetic hypotheses. We conducted low-coverage whole-genome sequencing in 39 Microlepidieae genomes, with a focus on *Arabidella* species differing by the extent of their genome diploidization (Mandáková et al., 2017b). In the absence of a robust nuclear genome phylogeny, phylogenetic relationships within Microlepidieae were resolved using complete chloroplast (cp) sequences, entire 35S ribosomal DNA (rDNA) units and nuclear DNA repeats. These phylogenetic frameworks were used to evaluate the extent of morphological convergence and disparity, and plastid–nuclear coevolution during postpolyploid genome diploidization and cladogenesis. Also, 5S rDNA sequence reads were analyzed to detect potential hybridization events.

Results

Characterizing the plastomes, nuclear rDNAs, and repeatomes

Using the low-coverage whole-genome sequencing data, we assembled the cp genomes, retrieved the sequences of the 35S rDNA, and characterized repeatomes of 39 Microlepidieae genomes (for accession data, see Supplemental Table S1). The length of the plastome sequences ranged from 153,821 bp in *Stenopetalum nutans* F.Muell. to 155,476 bp in *Arabidella chrysodema* Lepschi & Wege (Supplemental Table S2). We annotated a total of 132 genes (113 unique genes), including 87 protein coding, 37 tRNA, and 8 rRNA genes. The assembled length of nuclear 35S rDNA sequences varied from 5,939 bp in *S. decipiens* E.A. Shaw to 8,236 bp in *A. filifolia* (F.Muell.) E.A.Shaw (Supplemental Table S1). Due to incomplete assemblies of the highly variable intergenic spacer (IGS) region, we only utilized the conservative 18S-ITS1-5.8S-ITS2-26S region in downstream analyses.

By performing de novo repeatome identification using RepeatExplorer2 (Novak et al., 2013, 2020), the major repeat content in 39 Microlepidieae genomes was estimated to range from ~23% to 63% (Supplemental Table S3), taking into account the limitations of RepeatExplorer applied to low-coverage sequence data, for example, that less abundant repeats may be missed (Novak et al., 2020). In all Microlepidieae genomes, the predominant repeat type was the long terminal repeat (LTR) retrotransposons. The abundance of satellite repeats was highly variable, with the most remarkable expansion of the 174-bp BaSAT1 satellite repeat (Finke et al., 2019) accounting for > 10% of the *Ballantinia antipoda* (F.Muell.) E.A.Shaw genome. BaSAT1-like satellite sequences (c. 70% sequence similarity) were also identified in *Blennodia pterosperma* R.Br., *Menkea villosula* (F.Muell. & Tate) J.M.Black and *Phlegmatospermum richardsii* (F.Muell.) E.A.Shaw.

Phylogenomic analyses retrieved four intra-tribal clades in Microlepidieae

Based on 76 PCGs in 60 plastomes of Microlepidieae and outgroup species, we compiled a gap-free alignment matrix with 60,987 columns, of which 3,911 were parsimony informative. The same topology was inferred through maximum likelihood (ML) and Bayesian inference (BI) approaches, retrieving the Microlepidieae as a monophyletic group sister to the tribe Crucihimalayeae and resolving four strongly supported intra-tribal clades (bootstrap support [BS] > 90 and posterior probability = 1; Figure 1; Supplemental Figure S1). Clade A represents the previously defined crown-group genera (Mandáková et al., 2017b) including two *Arabidella* species (*A. eremigena*, *Arabidella procumbens* [Tate] E.A.Shaw) and *Menkea crassa* E.A.Shaw. Clade B, consisting of four *Arabidella* species (*A. filifolia*, *Arabidella glaucescens* E.A. Shaw, *Arabidella nasturtium* [F.Muell.] E.A.Shaw, and *A. trisecta*), and *Irenepharsus magicus* Hewson are sisters to the crown-group clade. Clades C and D appear as successive sisters to clades A + B, whereby clade C harbors only *Pachycladon* species, and clade D includes *A. chrysodema* and two *Menkea* species (*M. sphaerocarpa* F.Mull. and *M. villosula*). Of the six genera represented by at least two species, three genera (*Blennodia*, *Pachycladon*, and *Stenopetalum*) were retrieved as being monophyletic, *Arabidella* and *Menkea* as polyphyletic, and both *Cuphonotus* O.E.Schulz species clustered with the crown-group *Arabidella* species. Hence, seven *Arabidella* species were placed in clades A, B, and D, whereas three *Menkea* species were split between clades A and D.

The 5,875-bp alignment matrix of nuclear 35S rDNA sequences was based on 45 accessions analyzed (Supplemental Table S1). Whereas our ML and BI analyses (Supplemental Figures S2 and S3) corroborated the four intra-tribal clades within the plastome-based tree, the relationships between the clades differed. The rDNA-based phylogeny strongly supported (BS = 100) a sister relationship between the crown group (clade A) and *I. magicus*, and moderately supported (BS = 89) grouping of this clade and *Pachycladon*. Clade B (4 *Arabidella* spp.) was weakly supported (BS = 67) as sister to clade D. Congruently with the

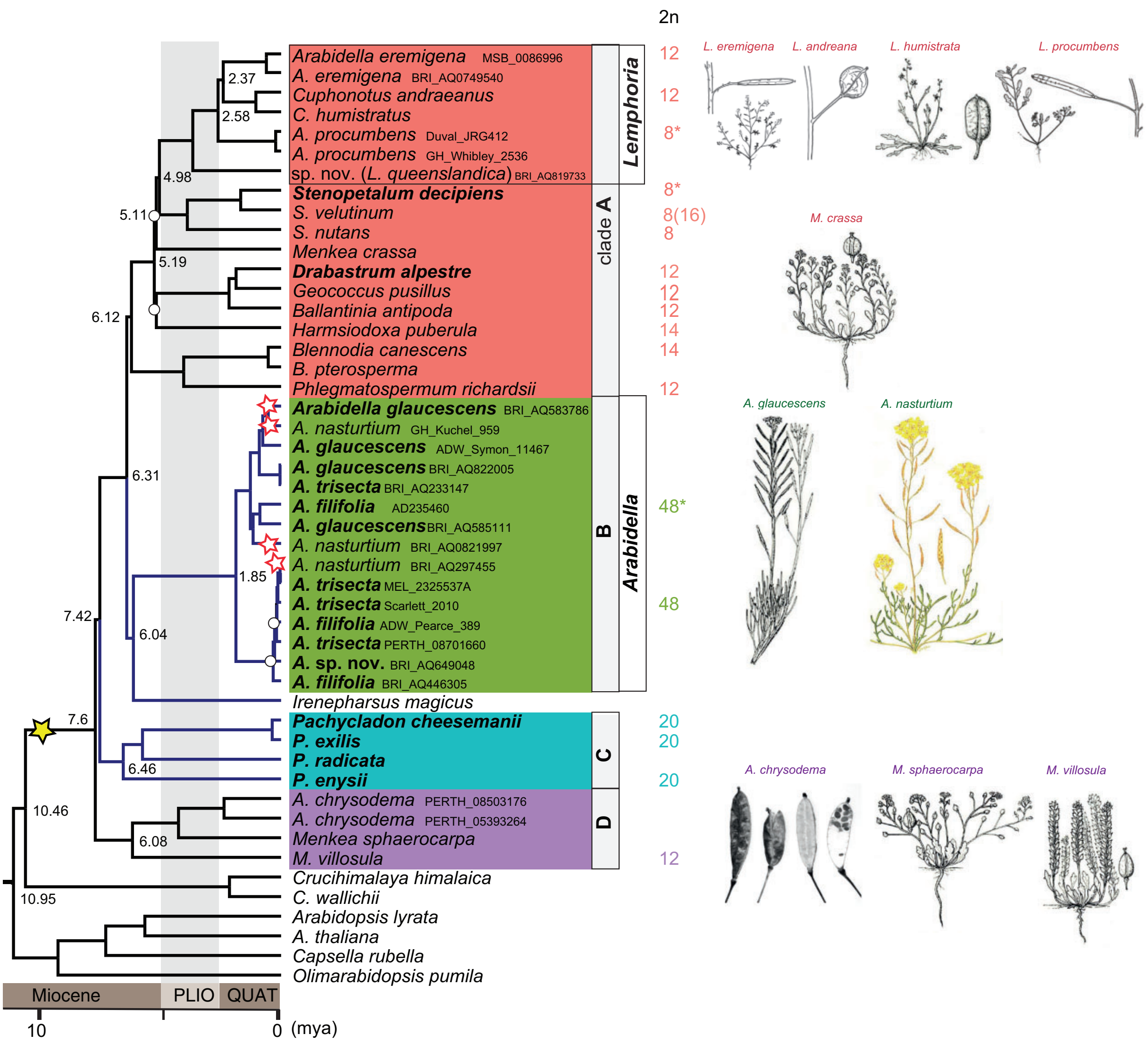


Figure 1 Plastome phylogeny of Microlepidieae. A simplified ML phylogeny based on concatenated 76 plastid PCGs (see Supplemental Figure S8 for a full version of the tree). Four intra-tribal clades were retrieved as clade A (the crown group), B (*Arabidella*), C (*Pachycladon*), and D (*A. chrysodema* and two *Menkea* spp.); blue-colored branches highlight slowly diploidizing clades and perennial species are in bold. Chromosome counts (2*n*) for the analyzed accessions are given, asterisked chromosome counts were established in this study. The yellow star indicates the tribe-specific allotetraploidization, while red stars represent the minimal number of neo-mesotetraploid WGDs in *Arabidella* based on chromosome counting and cytogenomic analyses (Figure 2). Numbers within the tree represent divergence-time estimates based on MCMC tree analysis (Supplemental Figure S8). White circles at tree nodes represent bootstraps <90. PLIO: Pliocene, QUAT: Quaternary.

plastome trees, the rDNA phylogenies corroborated *Arabidella* and *Menkea* being polyphyletic, and *Cuphonotus* clustered with crown-group *Arabidella* species.

To support the above inferred phylogenetic hypotheses, we analyzed relationships among Microlepidieae taxa based on similarities between shared repeat clusters. First, we have comparatively analyzed abundances of major repeat types in the 19 Microlepidieae genomes representing four intra-tribal clades using the RepeatExplorer2 platform. We identified 260 clusters of repetitive sequences that showed moderate to high abundances representing 18 repeat classes (Supplemental Figure S4). Topologies of the repeatome-based tree (Supplemental Figure S5) overall resembled the rDNA-based reconstruction. A consensus network (Supplemental Figure S5A), summarized from neighbor-joining (NJ) trees based on 33 out of the top 100 repeat clusters, corroborated the grouping of 12 Microlepidieae genomes into the four intra-tribal clades and retrieved a fifth clade formed by *I. magicus*. When *Pachycladon* (clade C) was omitted, *I. magicus* clustered with the crown group (Supplemental Figure S5B). Subsequently, we tested the performance of the repeatome-based phylogenies with respect to the diversity of identified repeated sequences (Supplemental Figure S6). Ty3/*Gypsy* retrotransposons, the most abundant repeat type in Microlepidieae genomes, produced phylogeny most congruent with the network based on all repeats. Consensus networks based on less abundant repeat types (e.g. Ty1/*copia*, DNA transposons) provided less resolved relationships.

A repeat-based phylogenetic analysis of all sequenced *Arabidella* accessions retrieved three phylogenetic clusters (Supplemental Figure S5C), corresponding to clades A, B, and D in the plastome and rDNA phylogenies. Comparison of repeat graphs revealed three distinct repeatome profiles of the *Arabidella* acccessions (Supplemental Figure S7).

Taxonomic considerations

The largely congruent phylogenetic analyses of the well-sampled Microlepidieae clearly support their division into four intra-tribal clades (Figure 1). Four *Arabidella* species including the generic type (*A. trisecta*) always formed a monophyletic clade, whereas *A. eremigena*, *A. procumbens*, a recently recognized species and two *Cuphonotus* species clustered together as a sub-clade within the crown group. The latter five species ought to be recognized as members of the genus *Lemphoria* O.E.Schulz (Lysak et al., 2022) and are referred to as *Lemphoria* species from here on (*L. andraeana*, *L. eremigena*, *L. humistrata*, *L. procumbens* (Tate) O.E.Schulz and *L. queenslandica* Edginton, Al-Shehbaz & Lysak). As recently circumscribed (Lysak et al., 2022), the genus *Arabidella* harbors four species (*Arabidella filifolia*, *A. glaucescens*, *A. nasturtium*, and *A. trisecta*). The corresponding formal nomenclatoric treatments of *Arabidella* and *Lemphoria* will be published separately. To settle taxonomic assignment of taxa in Clade D (*A. chrysodema* and two *Menkea* species), further phylogenetic analysis including all *Menkea* species is required.

Dated plastome phylogeny revealed the late divergence of *Arabidella*

Based on the plastome phylogeny and two secondary calibration points (see “Materials and methods”), the split between Microlepidieae and its closest tribe, Crucihimalayeae, was dated to 10.46 million years ago (Mya; highest posterior density [HPD] interval = 7.78–13.29; Supplemental Figure S8) during the Late Miocene (Tortonian). The Microlepidieae tribe underwent episodes of rapid diversification between 6 and 8 Mya (Messinian), leading to the successive divergence of the four intra-tribal clades. Consequently, the cladogenesis within three of these clades occurred almost simultaneously: the crown-group clade at 6.12 Mya (95% HPD: 4.62–7.82), *Pachycladon* at 6.46 Mya (95% HPD: 4.68–8.44) and Clade D at 6.08 Mya (95% HPD: 4.38–7.81). The divergence of *Arabidella* species has occurred much later, at around 1.85 Mya (95% HPD: 1.16–2.59), in Pleistocene.

Mesotetraploid and neo-mesotetraploid *Arabidella* genomes

The 24 chromosome pairs in *A. trisecta* ($2n = 48$) were previously shown to result from a younger, most likely autopolyploid WGD postdating the tribe-specific mesotetraploid event (Mandáková et al., 2017b). Here, we aimed to further elucidate genome evolution in *Arabidella* through comparative cytogenomic analysis of more populations. Fluorescently labeled chromosome-specific BAC contigs of *A. thaliana* were used to paint chromosomes of *Arabidella* and *Lemphoria* species, and the genomically unknown *I. magicus*. As these painting probes identify unique chromosomal regions in the diploid Arabidopsis genome, two genomic copies are indicative of the tribe-specific WGD, whereas four copies reflect the tribal WGD plus an additional, more recent, genome duplication (Mandáková et al., 2017a, 2017b). Three painting probes labeling two homoeologous regions within the haploid genome of *L. eremigena* ($2n = 12$), *L. procumbens* ($2n = 8$), and *I. magicus* ($2n = 20$; Supplemental Figure S9) corroborated their mesotetraploid origin (Figure 2). In *A. trisecta*, two genomic copies were identified in the mesotetraploid ($2n = 24$) population, whereas four copies in the neo-mesotetraploid population ($2n = 48$). A younger WGD was also detected in the analyzed $2n = 48$ populations of *A. filifolia* and *A. nasturtium* (Figure 2). The neo-mesotetraploid ($2n = 48$) populations of *A. trisecta* originated from the 24-chromosome cytotype via autopolyploidy as evidenced by (1) morphological similarity of both cytotypes, (2) perfect collinearity of four chromosome sets in the neo-mesotetraploid cytotype (Mandáková et al., 2017b), and (3) duplicated number of 5S and 35S rDNA loci in the neo-mesotetraploid cytotype (Figure 2).

Evidence of frequent hybridization and introgression in *Arabidella*

Graph clustering of Illumina reads corresponding to the repetitive 5S rDNA could produce a simple circle in diploid species and more complex structures in allopolyploid/hybrid genomes (Garcia et al., 2020). Using this approach, we observed simple circular structures in most Microlepidieae species (Supplemental Figure S10). Complex structures, composed of two or more loops interconnected by a junction region, were only identified in three accessions of *A. nasturtium* (BRI_AQ0821997, BRI AQ297455, GH_Kuchel_959), one population of *A. glaucescens* (BRI_AQ583786) and *Harmsiodoxa puberula* E.A.Shaw. We next asked whether footprints of hybridization could be detected among assembled plastome and 35S rDNA sequences, as the phylogenies based on the two datasets show partly conflicting topologies, including the relationships among *Arabidella* species (Figure 1; Supplemental Figure S2). As expected, our HyDe analyses showed that *Arabidella* accessions, especially the above-mentioned ones, were frequently detected as potential hybrids before correction for multiple testing (Supplemental Table S4).

By testing for putative hybridizing triplets using comparative three-genomic analysis of both 5S and 35S rDNA sequences with RepeatExplorer, we further corroborated the hybridization events and identified putative progenitor genomes. For 5S rDNA, the accessions BRI_AQ822005 (*A. glaucescens*) and MEL_2325537A (*A. trisecta*) were identified as putative parental genomes of *A. nasturtium* (BRI AQ297455; Figure 3A). Similarly, the sequence of MEL_2325537A (*A. trisecta*) and ADW_Pearce_389 (*A. filifolia*) showed the highest affinity to

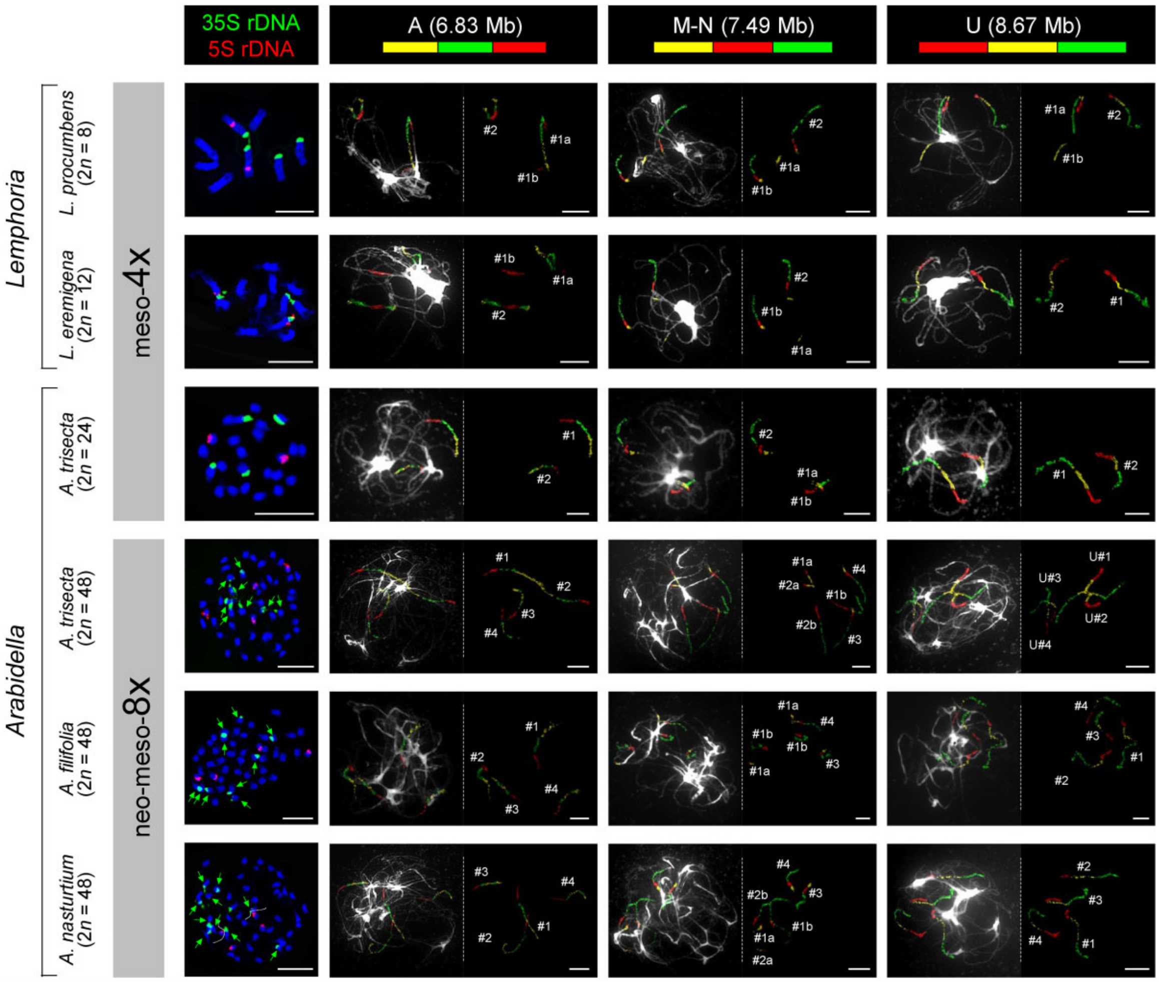


Figure 2 Cytogenetic analysis of *Arabidella* and *Lemphoria* genomes. The left side panel shows mitotic chromosome counts and FISH localization of 5S and 35S rDNA probes. The remaining panels display identification of three ancestral genomic blocks (GBs) by CCP on pachytene chromosomes of *Arabidella*/*Lemphoria* species. Two genomic copies of GBs (#1, #2) in *Lemphoria* spp. and the diploid cytotype of *A. trisecta* ($2n = 24$) reflect the tribe-specific mesotetraploid WGD. The four genomic copies (#1–#4) in the tetraploid cytotype of *A. trisecta* and two other *Arabidella* spp. (all $2n = 48$) correspond to the mesotetraploid WGD plus an additional genome duplication(s). DNA probes were detected as fluorescence of Cy3 (yellow), Alexa 488 (green) and Texas Red. Chromosomes at mitosis and pachytene were counterstained with DAPI. Scale bars, 10 μm.

A. glaucescens (BRI AQ583786; Figure 3B). The putative hybrid origin of the BRI AQ583786 accession was also confirmed by a three-genome comparison of the 18S-IGS region of 35S rDNA (Supplemental Figure S11). We failed to identify putative parental genomes contributing to the complex 5S rDNA structure in *H. puberula* (Supplemental Figure S12).

The evolution of chloroplast genomes associates with intra-tribal cladogenesis

To compare the evolutionary rates of the chloroplast genome among Microlepidieae clades, we estimated the substitution rates in PCGs in each species using *Crucihimalaya himalaica* (Edgew.) Al-Shehbaz et al. as a reference. The average rate across Microlepidieae species was 9.36×10^{-10} substitutions per site per year, with higher rates in the crown group (9.93×10^{-10}) and Clade D (1.02×10^{-9}), and lower rates in *Arabidella* (8.95×10^{-10}), *Pachycladon* (7.87×10^{-10}), and *I. magicus* (7.79×10^{-10}) (Figure 4A; Supplemental Table S5). We observed that plastome genes in the crown-group clade and Clade D species evolved significantly faster than those in *Arabidella* and *Pachycladon* ($P < 0.01$, two-tailed t test). In addition, genes in *Pachycladon* species evolved slower than those in *Arabidella* ($P < 0.01$, two-tailed t test).

To investigate the variation in selective pressure of chloroplast genes, we estimated the ratios of nonsynonymous (Ka) and synonymous (Ks) substitution rates between the sequences of 64 PCGs of Microlepidieae species and *C. himalaica*; the remaining 12 genes were excluded because of their extremely low variation (Ks = 0 in at least one species). The mean Ks values varied between 0.0285 and 0.0449, with significantly higher values in the crown-group clade and Clade D than in *Arabidella*, *Irenepharsus* Hewson, and *Pachycladon* ($P < 0.01$, two-tailed t test; Figure 4B; Supplemental Table S6). All PCGs of Microlepidieae plastomes showed signatures of purifying selection, that is, Ka/Ks values between 0 and 1, except for *matK* (maturase K) and *cemA* (chloroplast envelope membrane protein) in multiple

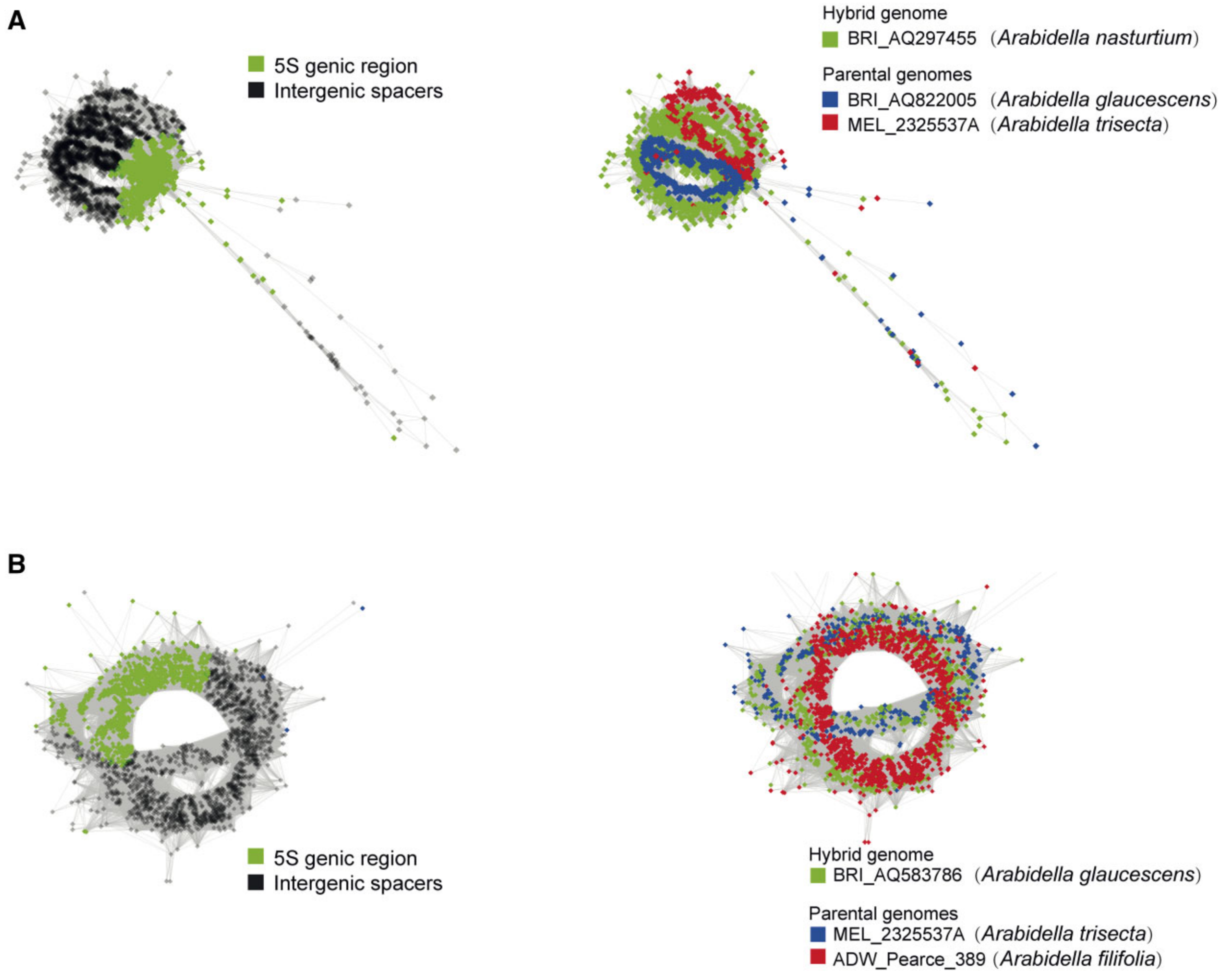


Figure 3 Graphical output of the three-genomic comparative 5S rDNA analyses including presumable hybrid and parental genomes. The results were visualized in two ways to show the clustering patterns of 5S genic region versus IGS domains, as well as the source of the input reads. A, Comparison between the hybrid accession BRI_AQ297455 (*Arabidella nasturtium*) and its putative parental accessions BRI_AQ822005 (*A. glaucescens*) and MEL_2325537A (*A. trisecta*). B, Comparison between the assumed hybrid accession BRI_AQ583786 (*A. glaucescens*) and its putative parental accessions MEL_2325537A (*A. trisecta*) and ADW_Pearce_389 (*A. filifolia*).

species (Supplemental Table S7). However, the clades with slow-evolving plastid genes (i.e. *Arabidella*, *Irenepharsus*, and *Pachycladon*) showed higher Ka/Ks values than those with fast-evolving genes (crown-group and D clades; Figure 4C). Interestingly, except for *matK* and *ycf2*, genes accounting for the difference between fast and slowly evolving clades encode subunits of four protein complexes including F-type ATP synthase, NADH dehydrogenase complex, cytochrome b_6f complex, and Photosystem II (Figure 4D), which are all involved in light-dependent reactions of photosynthesis (Wicke et al., 2011).

Species diversification and life-form transition

As a possible link between species diversification and life forms (annuality versus perenniality) was proposed in our earlier study (Mandáková et al., 2017b), we re-investigated this relationship based on the expanded phylogeny and taxon sampling. Our Bayesian Analysis of Macroevolutionary Mixtures (BAMM) analyses failed to detect any rate shifts during the diversification of Microlepidieae (Supplemental Figure S13A). The global speciation rate (λ), extinction rate (μ), and net diversification rate (γ) in Microlepidieae were estimated as 0.355 (95% quartile = 0.247–0.508), 0.112 (95% quartile = 0.007–0.327) and 0.244 (95% quartile = 0.140–0.340) species per million years (myr), respectively. Following the initial divergence of the tribe during late Miocene, the speciation rates showed a steady decline, while the extinction rates remained nearly constant (Supplemental Figure S13, B–D).

Ancestral state reconstruction inferred annuality, with a likelihood of 78.4%, to be the most likely ancestral life form in Microlepidieae (Supplemental Figure S14). The best-fitting model in BiSSE analyses suggested different speciation and transition rates but equal extinction rates between life forms (Supplemental Table S8). The model estimated a higher speciation rate in perennials than in annuals (0.468 versus 0.107 species/myr) with extremely low extinction rates ($<10^{-8}$) for both life forms. We observed multiple independent transitions from annuality to perenniality along the phylogeny, with *Drabastrum alpestre*

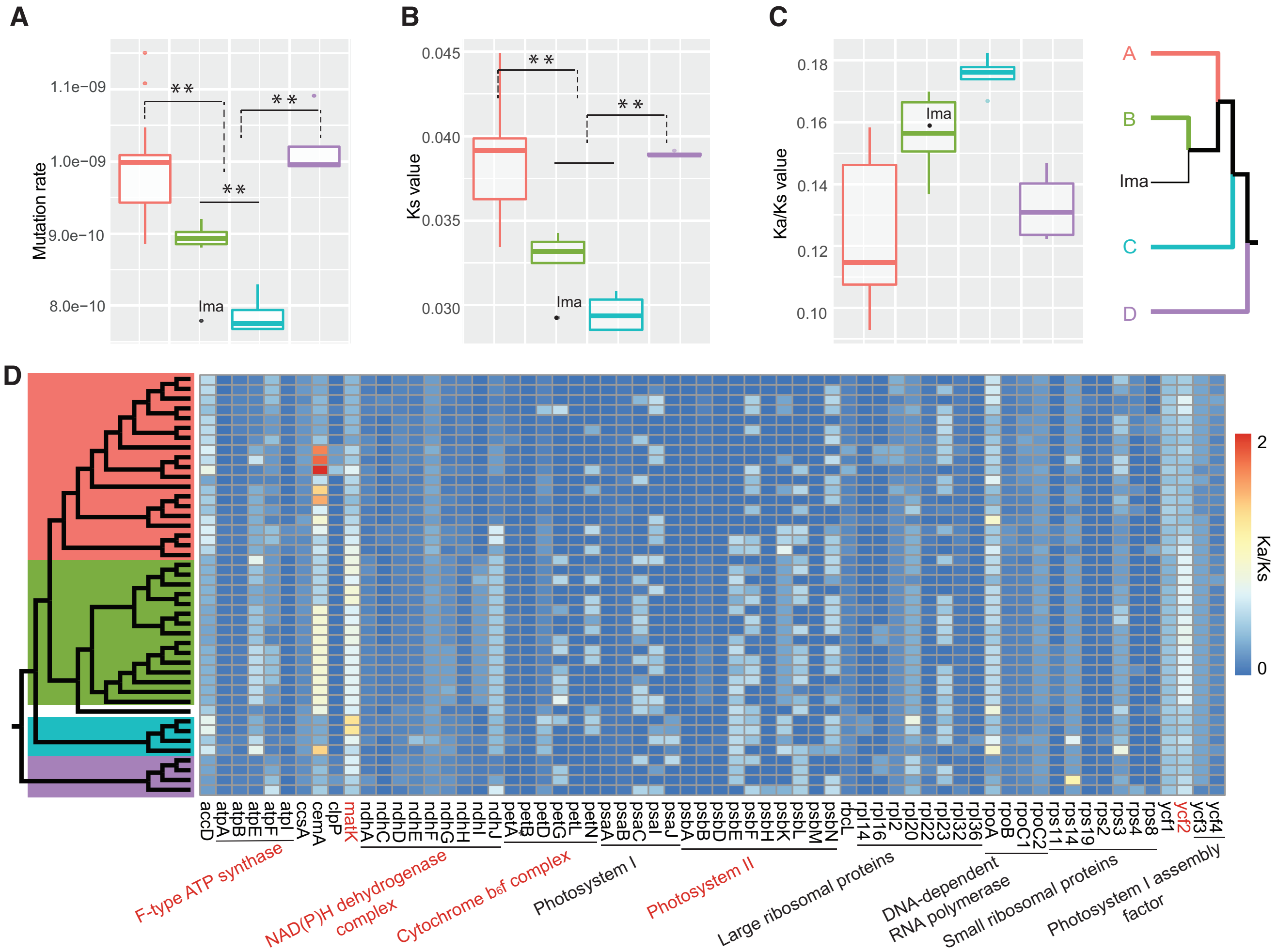


Figure 4 Mutation rate variation and selective pressure among the four intra-tribal Microlepidieae clades. A–C, Boxplots showing the difference in substitution rates (A), Ks (B), and Ka/Ks (C) of plastome genes between Microlepidieae species with fast (the crown group and Clade D, $n = 18, 4$) and slow diploidization rates (*Arabidella*, *Irenepharsus*, and *Pachycladon*, $n = 15$, 1, and 4, respectively). The plastid genes of *C. himalaica* were used as the reference. Boxes indicate the first and third quartiles, and horizontal bars represent the median values for each clade. Color coding in (A)–(C) corresponds to intra-tribal clades in simplified plastome phylogeny at the far right (clade A: the crown group, clade B: *Arabidella*; clade C: *Pachycladon*; clade D: *Arabidella chrysodema*/*Menkea* spp., Ima: *I. magicus*). D, The heatmap of Ka/Ks values sorted according to gene functions (*x*-axis) and the plastome phylogeny (*y*-axis). Red lettering indicates significant differences between fast and slowly diploidizing clades ($P < 0.01$; two-tails *t* test).

(F.Muell.) O.E.Schulz, some *Stenopetalum*, all *Pachycladon*, and three *Arabidella* species being perennials. Although *A. nasturtium* is usually recorded as having an annual life cycle, the species can be a short-living perennial in some circumstances (our observation). The perennial-to-annual transition rate was over 3 times higher than that of annual-to-perennial (0.474 versus 0.137 events/lineage/myr). In addition, our HiSSE analyses failed to detect the impact of any hidden effect on species diversification with life form transitions (Supplemental Table S9). Notably, our analyses were unlikely biased due to the oversampling of *Arabidella* accessions, as we recovered the difference between annuals and perennials using a pruned phylogeny with 36 (instead of 43) tips representing accessions/species sufficiently divergent from each other (see "Materials and methods").

Morphological trait evolution

To investigate the morphological evolution among Microlepidieae genera, we compiled a matrix for the presence/absence variation (PAV) in 27 states of eight phenotypic characters (coded as A–H; Figure 5A). The overall patterns of PAV among the selected traits showed high levels of homoplasy: 24 (out of 27) character states were present in more than 2 genera, and 21 character states were present in more than 2 clades (Figure 5A; Supplemental Figure S15). Whereas none of the characters had clade-specific PAV, 3 character states displayed genus-specific presence, including the fusiform fruit shape in *Scambopus* O.E.Schulz, the obcordate fruit shape in *Microlepidium*, and the aseriate seed arrangement in *Ballantinia* (Figure 5A; Supplemental Table S10).

The estimated means of morphological disparity (i.e. the proportion of states per morphological character within or

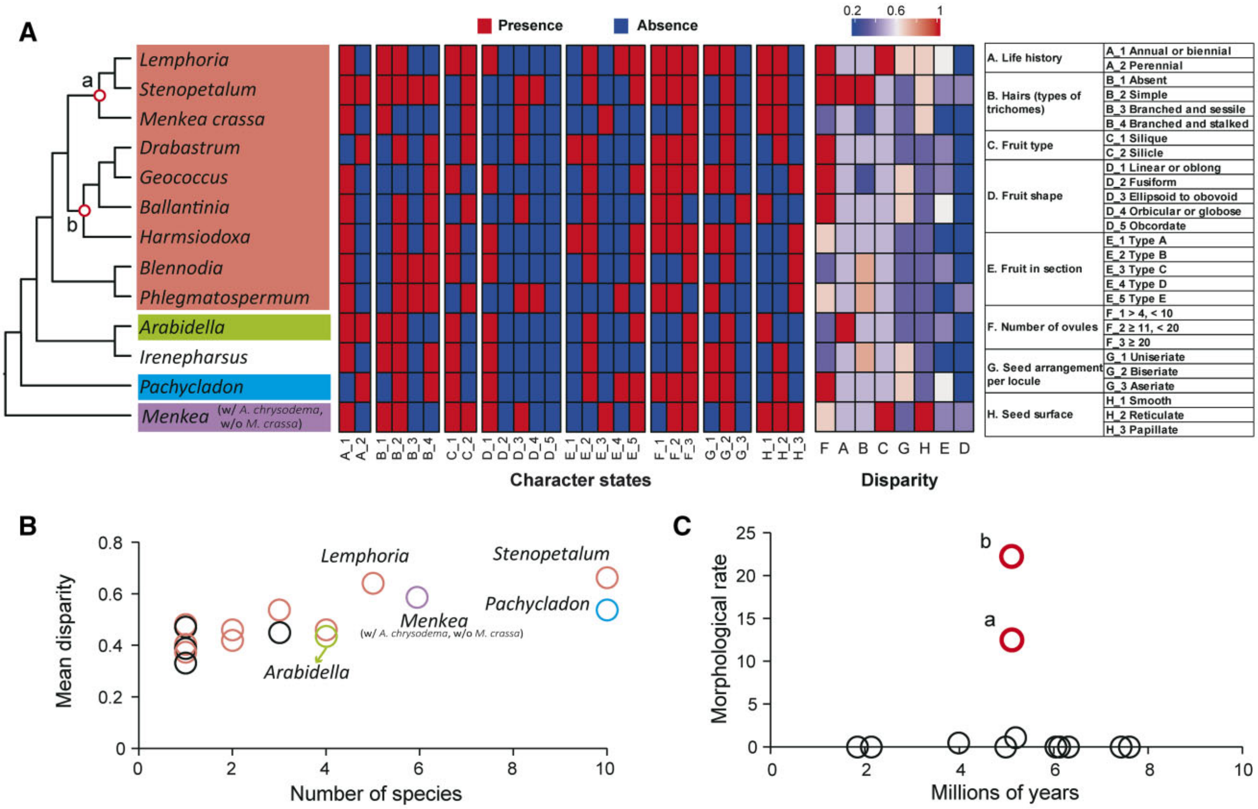


Figure 5 Morphological trait evolution across Microlepidieae. A, The simplified plastome phylogeny with heatmaps showing the PAV of 27 character states of 8 traits and the disparity for each trait realized at the genus level (see Supplemental Tables S10–S11). The circles (a and b) in the phylogeny indicate weakly supported nodes shown in Figure 1. The color scale bar indicates the range of disparity. The columns of disparity are sorted by the mean values for all genera, with the highest value on the left. The eight traits (A–H) and their states are listed in the rightmost box; states of trait E are detailed in Supplemental Figure S16. B, The relationship between mean disparity for all traits and species number across genera (see Supplemental Table S11 for more details). Colors correspond to four clades in (A) or otherwise indicate genera either not assigned to a clade or not included in the phylogeny in (A). C, Morphological evolutionary rate (i.e. the number of character changes myr) across Microlepidieae. Each circle represents an internal node in the phylogeny in (A); circles a and b represent outlier nodes with rapid morphological changes.

among species) showed a slight but significant positive correlation with the number of species (Pearson correlation: $n = 17$, $r = 0.753$, $P < 0.001$; Supplemental Table S11). We detected higher disparity in the crown-group genera, especially in *Lemphoria* and *Stenopetalum* (Figure 5, A and B; Supplemental Table S11). In contrast, *Arabidella* and *Pachycladon* displayed lower mean disparity than genera with the same or even smaller number of species (Figure 5B).

The inferred rates of character changes were generally low except for two early-diverging sub-clades of the crown-group clade (Figure 5C). Interestingly, these two nodes had the lowest BS in the plastome ML tree and displayed major conflicts between the plastome and rDNA phylogenies (Figure 1; Supplemental Figures S1–S3). Furthermore, δ-statistic analysis suggested that the presence or absence of character states was randomly distributed across the plastome phylogeny (Supplemental Table S12). Altogether, our analysis suggests that changes of most morphological traits occurred at a low rate and independently in each genus during the evolution of Microlepidieae.

Discussion

Four intra-tribal clades in Microlepidieae

With expanded taxon sampling and cp genome sequences, we have obtained a robust maternal phylogeny of the tribe Microlepidieae. The plastome-based phylogeny, largely congruent with phylogenetic analyses based on the nuclear 35S rDNA and repeatome sequences, (1) corroborated the repeatedly retrieved monophyly of the tribe (Warwick et al., 2010; Heenan et al., 2012; Mandáková et al., 2017b; Walden et al., 2020) resulting from an ancestral mesoallotetraploid event (Mandáková et al., 2017b) and (2) identified four highly supported intra-tribal clades (Figure 1). The four clades are formed by (A) the largest "crown group" of eight genera (c. 26 species), (B) genus *Arabidella* (4 spp.), (C) genus *Pachycladon* (11 spp.), and (D) clade comprising *A. chrysodema* and two *Menkea* spp. The position of *Irenepharsus* (3 spp.) remains unresolved due to the cytonuclear discordance placing the genus either as sister to *Arabidella* (Figure 1) or as an early branch within the crown-group clade (Supplemental Figure S2). Both phylogenies differ by the position of *Pachycladon*, which is sister to the crown group plus *Arabidella* in the plastome tree (Figure 1), but sister to the crown group and *Irenepharus* in the rDNA phylogeny (Supplemental Figure S2). Regardless the cytonuclear discordance, the four intra-tribal clades congruently recovered in all phylogenetic analyses, allow for phylogenetically informed analysis of postpolyploid genome diploidization and cladogenesis in Microlepidieae.

Intra-tribal cladogenesis mirrors different speed of postpolyploid diploidization

Whereas mesotetraploid genomes of the early branching *A. chrysodema*/*Menkea* clade and the crown group have been extensively diploidized, *Arabidella* and *Pachycladon* genomes are slowly diploidizing (Figure 1). Chromosome number of $2n = 20$ shared by *I. magicus* and *Pachycladon* species and the absence of an additional WGD, place *Irenepharsus* with the slowly diploidizing Microlepidieae genomes (Figure 1). The slowest diploidization occurred in *Arabidella* via the formation of four fusion chromosomes reducing the ancestral chromosome number from $n = 15$ to $n = 12$ (Mandáková et al., 2017b). Compared to *Arabidella*, the crown-group species experienced up to a three times more extensive chromosomal diploidization ($n = 15 \rightarrow n = 4$; Mandáková et al., 2010a, 2017b). The four recovered intra-tribal clades mirror the varied diploidization of Microlepidieae genomes, suggesting that the intrinsic genomic features underlying the extent of diploidization are shared among genera and species within one clade.

Plastome variation reveals cytonuclear interactions during postpolyploid diploidization

Establishing interactions between nuclear and plastome genomes is essential to the evolutionary success of hybrid lineages (Burton et al., 2013). In plant allopolyploids, challenges within the hybrid genomes might occur immediately or after several generations (Sehrish et al., 2015; Gyorfy et al., 2021) and can cause expression changes, homogenization, or loss of nucleus-encoded plastid-targeted genes (Gong et al., 2012, 2014; Sharbrough et al., 2021). In addition, accelerated chloroplast genome evolution could have resulted from exogenous selection (Muir et al., 2015). We observed significantly higher synonymous substitution rates in plastomes of the fast-diploidizing Microlepidieae clades than in less diploidized genomes of *Arabidella*, *Irenepharsus*, and *Pachycladon* (Figure 4). Interestingly, higher synonymous substitution rates were detected for duplicated nuclear genes in *S. nutans* (the crown-group, $2n = 8$) than those in *Pachycladon exilis* (Heenan) Heenan & A.D.Mitch. ($2n = 20$; Mandáková et al., 2017a). Moreover, we found that plastid genes participating in light-dependent reactions of photosynthesis showed different signatures of selection pressure between clades with fast versus slow genome diploidization. Thus, enzyme complexes comprising both nucleus- and plastid-encoded subunits might have different responses to a genome merger after hybridization. In diploidizing polyploids, cytonuclear interactions might be associated with the extent and tempo of diploidization of the nuclear genome (Sharbrough et al., 2017). While an increase in substitution rates of nuclear genes in diploidizing polyploids has occurred in multiple lineages (Kagale et al., 2014; Mandáková et al., 2017a; Guo et al., 2021), the evolutionary responses of plastomes were rarely addressed (Ferreira de Carvalho et al., 2019). Our results demonstrate, for a mesopolyploid Brassicaceae clade, that plastid genes may co-evolve with the nuclear genomes undergoing slower or faster postpolyploid diploidization (PPD). However, it remains unclear whether the cytonuclear interaction could be a direct (intrinsic) consequence of PPD or whether it can be linked to extrinsic factors such as climate or environmental changes (Muir et al., 2015; Hu et al. 2015). Similar studies in other mesopolyploid versus nonmesopolyploid Brassicaceae tribes should shed more light on the frequency and underlying factors of co-evolution of nuclear and plastid genes during diploidization.

Postpolyploid diploidization and evolution of morphological traits

After a WGD, polyploid genomes undergo diploidization potentially resulting in a continuum of more or less reproductively isolated populations, and eventually species and clades. Genome multiplication (autopolyploidy) or genome merger (allopolyploidy) may trigger diverse diploidization trajectories largely based on different fates of gene duplicates. The variation in morphological characters among the diploidizing genomes and species is to a large extent controlled by gene expression changes resulting from, for example, gene neofunctionalization, gene fractionation/loss, or biased gene expression (Zhu et al., 2017; Stitzer and Ross-Ibarra, 2018; Wu et al., 2018; Arya et al., 2021). These processes may have several possible outcomes, such as morphological disparity despite the shared ancestry, or morphological convergence despite independent diploidization of polyploid genomes (Tate and Simpson, 2003). In two *Brassica* L. species, *B. oleracea* L. and *B. rapa* L., similar morphotypes (e.g. cabbage versus Chinese cabbage) have evolved in parallel during independent diploidization associated with domestication of the mesohexaploid ancestral genome. In contrast, both *Brassica* species exhibit considerable intra-specific morphological variation, such as leaf heading cabbages and tuber-forming morphotypes (Cheng et al., 2016). Another mesohexaploid species, *Moricandia arvensis* (L.) DC., exhibits even within-individual seasonal disparity of floral characters (Gómez et al., 2020), whereby the seasonal plasticity may allow for exploitation of the same pollination niches due to phenotypic convergence between only distantly related crucifer species (Gómez et al., 2021). Some allopolyploid plant species may morphologically resemble more one of the two parental species due to biased gene expression after the genome merger (Alexander-Webber et al., 2016). Collectively, morphological convergence or disparity may hamper retrieving true phylogenetic relationship among species of diploidizing polyploid lineages.

Morphological convergence was frequently observed across Brassicaceae tribes (Hall et al., 2011; Huang et al., 2016; Hao et al., 2017; Walden et al., 2020). In Microlepidieae, highly supported phylogenetic analyses uncovered several instances of convergent evolution of some morphological characters and, conversely, considerable intra-tribal phenotypic disparity. For instance, *Arabidella* and *Lemphoria* show convergent fruit morphology which was interpreted by Shaw (1965) as a shared character justifying the merger of both genera. Superficially similar leaves of *L. procumbens* and *A. chrysodema* led to the inclusion of the latter species in *Arabidella* (Wege and Lepschi, 2007). On the contrary, two *Cuphonotus* species were recognized as a genus on its own and believed to be closely related to *Phlegmatospermu*m based on fruit morphology (Shaw, 1974). However, both *Cuphonotus* species form a monophyletic clade with *Lemphoria* species, and thus, both orbicular silicles (*L. andraeana*, *L. humistrata*) and cylindrical siliques (remaining three *Lemphoria* species) are found within a single genus. Comparable fruit-type disparity is encountered in a yet taxonomically unsettled clade harboring *A. chrysodema* (elliptic latiseptate siliques) and two *Menkea* species (obovoid silicles with a reduced or absent septum).

Our results indicate that morphological disparity does not necessarily correspond with species richness of Microlepidieae genera. The decoupling of morphological disparity and species diversity is quite common and could highlight the importance of other intrinsic or extrinsic factors that drive the morphological diversity (Oyston et al., 2016), for example, geographical distribution (Chartier et al., 2017, 2021) and additional WGDs (Walden et al., 2020). We noticed that clades/genera with higher mean disparity have experienced faster genome diploidization (Figure 5B). However, our analysis of morphological rate, congruently with recent studies in different organisms (Parins-Fukuchi et al., 2021; Stull et al., 2021), suggested overall low speed of character evolution with episodes of rapid changes that coincided with strong phylogenetic conflicts (Figure 5C). Therefore, the considerable morphological convergence across Microlepidieae could be best explained by gene-level diploidization that proceeded in parallel despite different pace of their genome-level (structural) diploidization. Although WGDs of different ages per se (without apparent diploidization) could alter phenotypic traits in polyploids (McCarthy et al., 2016), our study provides empirical evidence of the impact of PPD on morphological disparity in a mesopolyploie plant clade. As WGD-diploidization cycles occurred frequently during the angiosperm evolution, we propose that more polyploid models and rigorous tests (Clark and Donoghue, 2018) should be developed to investigate the impact of these genomic processes on evolution of morphological traits.

Although we have not examined the evolutionary trajectory of morphological evolution in Microlepidieae in the context of a robust nuclear phylogeny, our conclusions were based on evolutionary patterns at the level of major clades, whose monophyly has been consistently supported by multiple sources of evidence (plastome, rDNA, genomic repeats, and extent of chromosomal diploidization). In addition, the analysis of character PAV and disparity does not rely on statistical tests based on a fully resolved species-level phylogeny. Despite the apparent conflict with the plastome tree, we detected congruent patterns based on rDNA phylogeny, that is, low rates of morphological changes in general, a lack of phylogenetic signal in characters, and different diversification rates between annuals and perennials. Future studies combining comparative genomic and functional analyses should be able to identify the genetic basis underlying phenotypic changes during diploidization.

Life-form transitions during postpolyploid diversification

Our BAMM analyses revealed a continuous decrease in diversification rates after the initial divergence of Microlepidieae c. 10 Mya (Supplemental Figure S13A). The estimated mean values of both speciation (0.36 species/myr) and extintion rate (0.11 species/myr) were slightly higher than the recent estimates of 0.31 and 0.08 species/myr, respectively (Huang et al., 2020). The lack of shifts in speciation rate across the Microlepidieae phylogeny supports the notion that diversification was largely decoupled from WGDs and/or diploidization (Tank et al., 2015; Smith et al., 2018; Huang et al., 2020; Walden et al., 2020; but see Landis et al., 2018). In addition, our BiSSE analyses pinpointed higher speciation rates in perennials (0.468 species/myr) than annuals (0.107 species/myr), with a stronger tendency of transition from perenniality to annuality than in the opposite direction (Supplemental Table S8). Contrary to the expectation that rates of molecular evolution are higher in annuals (reviewed by Friedman, 2020), higher speciation rates in perennials were observed in several studies (e.g. Drummond et al., 2012; Azani et al., 2019) but the causes behind these observation remain unclear. In Microlepidieae, higher speciation rate in *Arabidella* and *Pachycladon* could be tentatively linked to their stable genome structures (Mandáková et al., 2010b; Mandáková et al., 2017b), which may allow for frequent homoploid hybridization (Becker et al., 2013; Joly et al., 2014). The directional bias in life-form transition appeared to be general in diverse herbaceous plant lineages (Bena et al., 1998; Andreasen and Baldwin, 2001; Datson et al., 2008; Lundgren and Marais, 2020), including Brassicaceae (Heidel et al., 2016). Regardless of its lower chance of formation, the perennial life style may have played an important role in adaptation to arid (e.g. *Arabidella*) and montane (e.g. *Pachycladon*, *Drabastrum alpestre*) habitats, as shown for Arabideae (Karl and Koch, 2013) and other plant lineages (Drummond et al., 2012; Jabbour and Renner, 2012; Ogburn and Edwards, 2015).

Hybridization and autopolyploidy drive genome evolution in *Arabidella*

The only known, but overlooked, chromosome number record for any *Arabidella* species was $2n = 24$ reported for *A. trisecta* from far northeastern South Australia (Rollins and Rüdenberg, 1971). Herein, 2*n* = 24 was identified in *A. trisecta* from northeastern South Australia (AD223503) and chromosome painting analysis has confirmed the mesotetraploid status of this plant (Figure 2). Other accessions of *A. trisecta*, *A. filifolia*, and *A. nasturtium* had 2*n* = 48 resulting from an additional WGD (s) (Figure 2; Mandáková et al., 2017b). Altogether, these findings support the view that the genus *Arabidella* is a polyploid complex of closely related mesotetraploid (2*n* = 24) and neomesotetraploid (2*n* = 48) genomes. Frequently reported intermediate morphotypes and difficult species assignment (e.g. Shaw, 1965; M. Edginton, unpublished observations) are reflecting most likely recurrent autopolyploidization and hybridization in *Arabidella*. Our analysis of homoeologous nuclear 5S rDNA sequences showed that all three populations of *A. nasturtium* represented presumably homoploid interspecies hybrids and suggest that the entire species could have a hybridogenous origin. Further population-level investigation is needed to assess the frequency of mesotetraploid and neopolyploid populations, as well as the level of inter-population gene flow between populations of the four sympatric *Arabidella* species.

Here, a tribe-wide analysis of 5S rDNA cluster graphs detecting probable hybridization events was applied to a Brassicaceae tribe. As hybridization in Brassicaceae is pervasive and genome-skimming data are accumulating, analysis of homologous 5S rDNA sequences in a wider spectrum of species may provide deeper insights into their origins and incongruent phylogenetic reconstructions.

Conclusions

Despite the prominence of genome duplication-diploidization cycles in evolution of angiosperm lineages, the role of postpolyploid diploidization in species divergence, including the phenotypic convergence and disparity, remains largely unknown. We addressed this question in the crucifer tribe Microlepidieae exhibiting differently paced genome diploidization and extensive morphological convergence. We provide clear phylogenomic evidence that differently paced postpolyploid diploidization was associated with (1) intratribal cladogenesis, (2) morphological disparity, (3) selection pressure on genes involved in cytonuclear interaction, and (4) life-form transitions. Our results along with the close phylogenetic relatedness to *A. thaliana* make Microlepidieae an excellent model system to investigate the evolutionary consequences of postpolyploid genome evolution.

Materials and methods

Taxon sampling

The list of the analyzed accessions and outgroup species, as well as the GenBank accessions of plastome and 35S rDNA sequences, are provided in Supplemental Table S1.

Low-coverage whole-genome sequencing

NucleoSpin Plant II kit (Macherey-Nagel, Düren, Germany) was used to extract genomic DNA from fresh or silica-dried leaves. DNA sequencing libraries were sequenced at the Core Facility Genomics, CEITEC, Masaryk University. The Illumina Miseq platform, generating 150-bp paired-end reads, was used for sequencing.

Sequence assembly and annotation

The raw reads were filtered using the fastp-version 0.20.1 software (Chen et al., 2018) with the following parameters: -z 4 -q 20 -u 30 -n 0 -f 5 -t 5. After quality control, cp genome assemblies were generated using NOVOPlasty (Dierckxsens et al., 2017) or GetOrganelle (Jin et al., 2020). When NOVOPlasty/GetOrganelle failed to return a complete assembly, plastid sequences were selected from contigs assembled with Velvet version 1.2.10 (Zerbino and Birney, 2008) through comparison with the Arabidopsis (*A. thaliana*) chloroplast genome (GenBank accession: NC_000932) and subsequently merged into a consensus linear sequence using Geneious software (Kearse et al., 2012). To demonstrate the accuracy of the assembled plastomes, we compared our assemblies with publicly available sequences in GenBank (Supplemental Table S13). Using publicly available cp genomes of Brassicaceae as the reference, the assembled plastomes were annotated using Plann software (Huang and Cronk, 2015) and manually curated with Sequin software. To search for 35S rDNA sequences, the reads were assembled by the Megahit software (Li et al., 2015) with the following parameters: -m 80,000,000,000 -t 12. The resulting contigs were mapped by the BWA software (Li and Durbin, 2009) to the *A. thaliana* 35S rDNA sequences (GenBank accession: X52322) and fully assembled 35S rDNA units (18S-ITS1-5.8S-ITS2-26S) were selected for downstream analyses.

Sequence alignment and phylogenetic analysis

For the cp data, we combined published cp genomes of 17 Brassicaceae species (Supplemental Table S1) with the generated sequences to build an alignment matrix of 76 protein-coding genes (PCGs) following (Guo et al., 2017). Each PCG was aligned by PRANK (Loytynoja and Goldman, 2008) and subjected to Gblock (Castresana, 2000) to trim ambiguously aligned regions. Then, the individual alignments were concatenated. For 35S rDNA data, due to the variation within the IGS sequences, only the unique 35S rDNA transcription units (18S-ITS1-5.8S-ITS2-26S) were used for the sequence alignment using the Mafft software (Katoh and Standley, 2013), and two ambiguous terminal regions were removed based on the *A. thaliana* 35S rDNA sequences. For both datasets, ML analyses were undertaken using IQ-tree program (Nguyen et al., 2015) by searching for the best substitution models for each of the partitions. Node supports were assessed with 1,000 rapid bootstrapping replicates. BI trees and divergence times were coestimated using BEAST version 2.5 (Bouckaert et al., 2019). The resulting trees were visualized and edited in FigTree version 1.4.1 (http://tree.bio.ed.ac.uk/software/figtree/). *Tarenaya hassleriana* (Chodat) Iltis (Cleomaceae) was used to root the phylogenetic trees.

Molecular dating based on plastome data

We used MCMCTree software implemented in the PAML4.9e package (Yang, 2007) to estimate divergent times with a codon-partitioned dataset from the concatenated 76 PCGs (Supplemental Figure S8). The independent rates clock model (Rannala and Yang, 2007) was applied with the gamma-Dirichlet prior (Dos Reis et al., 2014) for the overall substitution rate (rgene gamma) setting at G (4, 90, 1). The three parameters (birth rate λ, death rate μ, and sampling fraction ρ) for the birth–death process were specified as $\lambda = \mu = 1$ and $\rho = 0$. Due to the lack of reliable fossil calibration points in Brassicaceae (Franzke et al., 2016), we applied two secondary calibration points from (Walden et al., 2020): the crown age of Brassicaceae was set to 24.31–35.71 Mya, and the crown age of Camelineae was set to 5.56–9.78 Mya. The analyses were run for 5 million generations sampled every 500 generations after a burn in of 500,000 iterations. Two separate MCMC runs were compared for convergence with two different random seeds and similar results were observed.

Repeatome identification and phylogenetic analysis

The unassembled reads obtained from low-coverage genome sequencing were used for repetitive elements identification by applying the RepeatExplorer 2 (RE2) pipeline based on the graph-based clustering method (Novak et al., 2013, 2020). Reads were filtered as above described and then were sampled as input for the RE2 pipeline. Because genome size information was not available for most Microlepidieae species, an average of 200,000 reads per genome were sampled and analyzed using RE2 regardless of genome size. To verify that this sample size was sufficient to analyze the repeatomes of Microlepidieae species, we repeated our analysis with 400,000 reads per genome for selected species with different chromosome numbers and from different intra-tribal clades and found that these results were comparable to the 200,000 read samples. The most abundant repeat clusters (genome proportion $>0.01\%$ of the total input reads) were annotated and only the most abundant repeat types were summarized (Supplemental Table S3). Comparative clustering analyses were performed for Microlepidieae species by RE2 with default parameters. The repeat-sequence similarity matrices obtained from the comparative clustering analyses were employed to infer phylogenetic relationships using the most abundant clusters (Vitales et al., 2020). Briefly, the more similar repeats of two species have the higher number of edges between the reads of those species. These similarity matrices were transformed into distance matrices. Then, the pairwise genetic distance matrices were used to construct an NJ tree by using the NJ function in ape package for each abundant cluster (genome proportion $>0.01\%$). Finally, a consensus network was constructed by using SplitsTree5 (Bagci et al., 2021) in Newick format from all NJ trees (Supplemental Figure S5). Custom R scripts were used to process RE2 output data and phylogenetic analyses.

Chromosome preparation and cytogenetic analysis

For chromosome preparations, inflorescences with young flower buds were collected in fixative (3:1 ratio of ethanol 96% and glacial acetic acid, v/v) and kept cold until analysis. Mitotic and meiotic (pachytene) chromosome preparations were prepared from the fixed young flower buds containing immature anthers as described by Lysak and Mandáková (2013) and Mandáková and Lysak (2016). Chromosome preparations were treated with 100 μg mL^{-1} RNase in 2 × sodium saline citrate (SSC; 20 × SSC: 3-M sodium chloride, 300-mM trisodium citrate, pH 7.0) for 60 min and with 0.1 mg ml^{-1} pepsin in 0.01 M HCl at 37°C for 5 min; then postfixed in 4% formaldehyde in 2 × SSC (v/v) for 10 min, washed in 2 × SSC twice for 5 min, and dehydrated in an ethanol series (70%, 90%, and 100%, v/v, 2 min each). The BAC clone T15P10 (AF167571) of Arabidopsis bearing 35S rRNA gene repeats was used for in situ localization of nucleolar organizer regions (NORs), and the *Arabidopsis* clone pCT4.2 (M65137), corresponding to a 500-bp 5S rDNA repeat, was used for localization of 5S rDNA loci. Fluorescently labeled chromosome-specific Arabidopsis BAC contigs, representing three ancestral crucifer genomic blocks (Lysak et al., 2016), were used to paint pachytene chromosomes (block A: 32 BACs covering 6.85 Mb of the Arabidopsis chromosome At1; M-N: 45 BACs of Arabidopsis chromosome At3, 7.49 Mb; U: 48 BACs of At4, 8.67 Mb). All DNA probes were labeled with biotin-dUTP, digoxigenin-dUTP, or Cy3-dUTP by nick translation as described. Selected labeled DNA probes were pooled together, ethanol precipitated, dissolved in a 20 μL mixture containing 50% formamide (v/v), 10% dextran sulfate (w/v) and 2 × SSC, and pipetted onto each of the microscopic slides. The slides were heated at 80°C for 2 min and incubated at 37°C overnight. Hybridized probes were visualized either as the direct fluorescence of Cy3-dUTP (yellow) or through fluorescently labeled antibodies against biotin-dUTP (red) and digoxigenin-dUTP (green). Biotin-dUTP was detected by Avidin Texas Red and amplified by goat anti-avidin biotin and Avidin Texas Red; digoxigenin-dUTP was detected by mouse anti-digoxigenin and goat anti-Alexa Fluor 488. Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI; 2 μg mL^{-1}) in Vectashield antifade. Fluorescence signals were analyzed and photographed using a Zeiss Axioimager epifluorescence microscope and a CoolCube camera (MetaSystems, Heidelberg, Germany). Individual images were merged and processed using the Photoshop CS software (Adobe Systems, San Jose, CA, USA).

Hybridization events detection

The plant 5S rDNA includes c. 120-bp conserved coding region and variable IGSs. In graphic clustering analysis of 5S rDNA, the regular circular structures were observed in most diploid-like species, while the complex structures indicate potential hybrids (Garcia et al., 2020). We used graph clustering method implemented in the RepeatExplorer2 to analyze homoeologous 5S rDNA arrays at the genomic level searching for hybridogenic origin of the Microlepidieae species. Typically, 200,000 of pair-end reads were used as input for clustering. Then, in order to estimate the possible combination across our dataset, we used the HyDe program (Blischak et al., 2018) to predict the potential hybrids and parental species. By using 35S rDNA and plastome sequence matrixes, HyDe program outputs the possible combinations (Supplemental Table S4). We then summarized the HyDe results by recoding potential triplet combination to further test the intra- or inter-generic hybridization events. Finally, we carried out a comparative three-genomic analysis implemented in RepeatExplorer platform to test whether hybridization happened and to track the putative progenitor species.

Mutation rate and selection pressure analysis

Individual species-specific mutation rates from cp genome were calculated using the equation $R = m/(nT)$, where R is the rate of mutation per site per year, m is the number of observed mutation sites, n is the number of total nucleotide sites, and T is the divergence time of a node, as described by (Dong et al., 2020; Supplemental Table S5). We did not consider the multiple times of mutation on the same mutated site and nonobserved site. Codeml in the PAML package (Yang, 2007) was used to calculate the rates of nonsynonymous substitutions (Ka), synonymous substitutions (Ks), and their ratio (Ka/Ks) for Microlepidieae plastome PCGs (Supplemental Tables S6 and S7). The Crucihimalayeae species *C. himalaica* was used as a reference. As the accurate detection of genomic variations requires high read coverage (Sims et al., 2014), we calculated the sequencing depth for all plastome assemblies (Supplemental Table S2) following Negm et al. (2021).

Morphological trait matrix and analyses

We compiled a phenotypic dataset for Microlepidieae genera based on our proposed taxonomic treatment (Lysak et al., 2022). The matrix scored the absence (0) and presence (1) for 27 character states in eight discrete characters, including two general traits (life history and hairs/trichomes) and six reproductive ones in fruit and seed morphologies. To assess the PAV of these characters across genera, we calculated the morphological disparity (Walden et al., 2020), which indicated the fraction of character-state presence for each genus. Using the method developed by Parins-Fukuchi et al. (2021; https://figshare. com/ articles/ dtaset/ pf_ stull_ smith_ tgz/ 13190816/ 2, last accessed in July 20th, 2021), we estimated the number of character transitions along each branch using parsimony and further calculated the morphological evolutionary rates using the time estimates obtained from molecular dating analyses. To determine whether the presence/absence of character states occurred in more closely related taxa, that is, showed phylogenetic signal, we tested their correlation with plastome phylogeny using a recently developed approach that was specifically designed for categorical traits (Borges et al., 2019). For both analyses, we presented results based on a simplified version of the dated plastome phylogeny. Because the true species relationships remain unclear, we repeated the analyses using a simplified version of the dated ML tree of rDNA as the input phylogeny.

State-independent diversification analyses

First, we inferred the ancestral state of life form trait and the overall pattern of macroevolutionary rates in Microlepideae species. Based on the time-calibrated BEAST phylogeny inferred with the plastome data, we employed BAMM version 2.5.0 (Rabosky et al., 2014, 2017) to estimate rates of speciation (λ), extinction (μ), and net diversification (γ) for the Microlepidieae tribe at the genus level. After pruning outgroups, a total of 42 Microlepidieae accessions were used for BAMM analyses. According to a list of currently known Microlepidieae species, we accounted for nonrandom and incomplete taxon sampling by giving a percentage to each of the sampled and monophyletic genus; the overall sampling fraction was set to 0.88 assuming that the eight missing taxa were separate lineages across the phylogeny. The BAMM priors were generated with the function setBAMMpriors implemented in the R package BAMMtools (Rabosky et al., 2014) assuming one rate shift. We ran MCMC chains for 1 million generations with default settings, discarding the first 10% of samples as burn in. We checked the convergence of MCMC runs by plotting the trace of the log-likelihood as well as determining their effective sample sizes (>200) using the coda package in R (Plummer et al., 2006).

Ancestral state reconstruction for life forms and state-dependent diversification

We reconstructed the ancestral state of life forms for Microlepidieae species on the BEAST phylogeny of plastomes using the "ace" function in the ape R package. Trait data were taken from Mandáková et al. (2017b) and coded as 0 for annuals and 1 for perennials. Using the diversitree R package, we fitted two BiSSE models with unconstrained and constrained speciation rate (λ) to test whether the rate in perennials (λ1) was significantly increased compared with that in annuals (λ0). As described above for the state-independent analyses, we assumed a gross sampling fraction of 0.88 to account for incomplete sampling. As we included multiple accessions for *Arabidella* species, we further mitigated potential sampling bias by removing 7 tips from the dated plastome phylogeny to make pairwise species divergence time to be no less than 0.36 Mya, that is, the least divergence time between known species (*P. cheesemanii* and *P. exilis*). The best-fit model was selected based on the Akaike information criterion. Because we were estimating six parameters for a small tree with only 43 tips, we ran the MCMC analysis for 10,000 steps using an exponential prior with a rate of 1/(2r), where r is the diversification rate of the character. In addition, we employed the HiSSE framework implemented in the R package HiSSE version 1.9.19 (Beaulieu and O'Meara, 2016) to test the impact of any unmeasured factors (i.e. hidden states) on diversification rates in species with different life forms. We tested five models including: (1) the BiSSE model without any hidden effect; (2) three HiSSE models with two hidden states (A and B) vary independently

(HiSSE) or constrained with one of the life form states (HiSSE.0 and HiSSE.1); and (3) a null model with character-independent diversification (CID-2).

Accession numbers

All raw reads generated in this study are available from the NCBI database under the BioProject PRJNA752007. The assembled chloroplast genomes and 35S rDNA sequences are available from the GenBank under accession numbers OL364706–OL364744 and OL339480–OL339518, respectively.

Supplemental data

The following materials are available in the online version of this article.

Supplemental Figure S1. Plastome-based phylogeny of Microlepidieae based on Bayesian analysis of 76 PCGs.

Supplemental Figure S2. ML phylogenetic tree based on analysis of 35S rDNA sequences.

Supplemental Figure S3. Bayesian phylogenetic tree of Microlepidieae based on analysis of 35S rDNA sequences.

Supplemental Figure S4. A comparative repeat graph of Microlepidieae accessions.

Supplemental Figure S5. Phylogenetic relationships in Microlepidieae based on nuclear repeat sequence similarities.

Supplemental Figure S6. Performance of different repeat types employed to infer the phylogenetic relationships in Microlepidieae.

Supplemental Figure S7. A comparative repeat graph of the analyzed *Arabidella* genomes.

Supplemental Figure S8. Time-calibrated plastome phylogeny of Microlepidieae.

Supplemental Figure S9. Cytogenetic analysis of *I. magicus*.

Supplemental Figure S10. Graphic clustering analysis of 5S rDNA in Microlepidieae accessions.

Supplemental Figure S11. Graphical output of the three-genomic comparative analysis of 18S-IGS rDNA.

Supplemental Figure S12. Graphical output of the three-genomic comparative 5S rDNA analysis in *Harmsiodoxa puberula*.

Supplemental Figure S13. Macroevolutionary patterns of species diversification in Microlepidieae.

Supplemental Figure S14. Ancestral state reconstruction of life form trait in Microlepidieae.

Supplemental Figure S15. A histogram-based summary of character state presence across Microlepidieae genera and clades.

Supplemental Figure S16. Fruit types in cross section.

Supplemental Table S1. Origin and GenBank accession numbers of the analyzed Microlepidieae accessions.

Supplemental Table S2. Plastome read coverage in the analyzed Microlepidieae accessions.

Supplemental Table S3. Repeatome composition in 39 Microlepidieae accessions.

Supplemental Table S4. Detection of potential hybridization events and putative parental species in *Arabidella* using HyDe.

Supplemental Table S5. Plastome mutation rates of the analyzed Microlepidieae accessions.

Supplemental Table S6. Ks values of the analyzed Microlepidieae accessions.

Supplemental Table S7. Ka/Ks values of 64 plastome genes in the analyzed Microlepidieae accessions.

Supplemental Table S8. BiSSE analyses of species diversification in annuals and perennials.

Supplemental Table S9. HiSSE analyses of species diversification in annuals and perennials.

Supplemental Table S10. Matrix of morphological trait states in Microlepidieae genera.

Supplemental Table S11. Morphological disparity in selected characters within Microlepidieae genera.

Supplemental Table S12. Summary of δ-statistics test of phylogenetic signal in character states.

Supplemental Table S13. Comparison of de novo assembled and published plastome sequences for Microlepidieae taxa.

Acknowledgments

We thank Neville Walsh (Royal Botanic Gardens Victoria), Anthony Brach (Harvard University Herbaria), Gillian Brown (Queensland Herbarium), Daniel Duval (Botanic Gardens of South Australia), Neville Scarlett, Janet Terry (Millenium Seed Bank, Royal Botanic Gardens, Kew) and Kasper Hendriks (Naturalis Biodiversity Centre) for providing materials or sequence data. We thank Daniel Vitales for his suggestion on repeatome phylogeny and Roi Rodríguez González for sharing the R script. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic. Computational resources were also provided by the ELIXIR-CZ project (LM2018131), part of the international ELIXIR infrastructure.

Funding

This work was financially supported by a research grant from the Czech Science Foundation (20-03419Y).

Conflict of interest statement. None declared.

References

Al-Shehbaz IA (2003) A synopsis of *Tropidocarpum* (Brassicaceae). Novon **13**: 392–395

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

Alexander-Webber D, Abbott RJ, Chapman MA (2016) Morphological convergence between an allopolyploid and one of its parental species correlates with biased gene expression and DNA loss. J Heredity **107**: 445–454

Andreasen K, Baldwin BG (2001) Unequal evolutionary rates between annual and perennial lineages of checker mallows (Sidalcea, Malvaceae): evidence from 18S-26S rDNA internal and external transcribed spacers. Mol Biol Evol **18**: 936–944

Arya GC, Tiwari R, Bisht NC (2021) A complex interplay of G beta and G gamma proteins regulates plant growth and defence traits in the allotetraploid *Brassica juncea*. Plant Mol Biol **106**: 505–520

Azani N, Bruneau A, Wojciechowski MF, Zarre S (2019) Miocene climate change as a driving force for multiple origins of annual species in *Astragalus* (Fabaceae, Papilionoideae). Mol Phylogenet Evol **137**: 210–221

Bagci C, Bryant D, Cetinkaya B, Huson DH (2021) Microbial phylogenetic context using phylogenetic outlines. Genome Biol Evol **13**: evab213

Beaulieu JM, O'Meara BC (2016) Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst Biol **65**: 583–601

Becker M, Gruenheit N, Steel M, Voelckel C, Deusch O, Heenan PB, McLenachan PA, Kardailsky O, Leigh JW, Lockhart PJ (2013) Hybridization may facilitate in situ survival of endemic species through periods of climate change. Nat Climate Change **3**: 1039–1043

Bena G, Lejeune B, Prosperi JM, Olivieri I (1998) Molecular phylogenetic approach for studying life-history evolution: the ambiguous example of the genus *Medicago* L. Proc Royal Soc B-Biol Sci **265**: 1141–1151

Blischak PD, Chifman J, Wolfe AD, Kubatko LS (2018) HyDe: a Python package for genome-scale hybridization detection. Syst Biol **67**: 821–829

Borges R, Machado JP, Gomes C, Rocha AP, Antunes A (2019) Measuring phylogenetic signal between categorical traits and phylogenies. Bioinformatics **35**: 1862–1869

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchene S, Fourment M, Gavryushkina A, Heled J, Jones G, Kuhnert D, De Maio N, et al. (2019) BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLOS Comput Biol **15**: e1006650

Bowers JE, Chapman BA, Rong JK, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422**: 433–438

Burton RS, Pereira RJ, Barreto FS (2013) Cytonuclear genomic interactions and hybrid breakdown. Ann Rev Ecol Evol Syst **44**: 281–302

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol **17**: 540–552

Chartier M, Lofstrand S, von Balthazar M, Gerber S, Jabbour F, Sauquet H, Schonenberger J (2017) How (much) do flowers vary? Unbalanced disparity among flower functional modules and a mosaic pattern of morphospace occupation in the order Ericales. Proc Royal Soc B Biol Sci **284**: 20170066

Chartier M, von Balthazar M, Sontag S, Lofstrand S, Palme T, Jabbour F, Sauquet H, Schonenberger J (2021) Global patterns and a latitudinal gradient of flower disparity: perspectives from the angiosperm order Ericales. New Phytol **230**: 821–831

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**: i884–i890

Cheng F, Sun RF, Hou XL, Zheng HK, Zhang FL, Zhang YY, Liu B, Liang JL, Zhuang M, Liu YX, et al. (2016) Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. Nat Genet **48**: 1218–1224

Clark JW, Donoghue PCJ (2018) Whole-genome duplication and plant macroevolution. Trend Plant Sci **23**: 933–945

Datson PM, Murray BG, Steiner KE (2008) Climate and the evolution of annual/perennial life-histories in *Nemesia* (Scrophulariaceae). Plant Syst Evol **270**: 39–57

Dierckxsens N, Mardulyn P, Smits G (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res **45**: e18

Dong Y, Ostergaard L (2019) Fruit development and diversification. Curr Biol **29**: R781–R785

Dong W, Xu C, Wen J, Zhou S (2020) Evolutionary directions of single nucleotide substitutions and structural mutations in the chloroplast genomes of the family Calycanthaceae. BMC Evol Biol **20**: 1–12

Dos Reis M, Zhu TQ, Yang ZH (2014) The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. Syst Biol **63**: 555–565

Drummond CS, Eastwood RJ, Miotto STS, Hughes CE (2012) Multiple continental radiations and correlates of diversification in *Lupinus* (Leguminosae): testing for key innovation with incomplete taxon sampling. Syst Biol **61**: 443–460

Eldridge T, Langowski L, Stacey N, Jantzen F, Moubayidin L, Sicard A, Southam P, Kennaway R, Lenhard M, Coen ES, et al. (2016) Fruit shape diversity in the Brassicaceae is generated by varying patterns of anisotropy. Development **143**: 3394–3406

Ferreira de Carvalho J, Lucas J, Deniot G, Falentin C, Filangi O, Gilet M, Legeai F, Lode M, Morice J, Trotoux G, et al. (2019) Cytonuclear interactions remain stable during allopolyploid evolution despite repeated whole-genome duplications in *Brassica*. Plant J **98**: 434–447

Finke A, Mandáková T, Nawaz K, Vu GTH, Novak P, Macas J, Lysak MA, Pecinka A (2019) Genome invasion by a hypomethylated satellite repeat in Australian crucifer *Ballantinia antipoda*. Plant J **99**: 1066–1079

Franzke A, Koch MA, Mummenhoff K (2016) Turnip time travels: age estimates in Brassicaceae. Trend Plant Sci **21**: 554–561

Friedman J (2020) The evolution of annual and perennial plant life histories: ecological correlates and genetic mechanisms. Ann Rev Ecol Evol Syst **51**: 461–481

Garcia S, Wendel JF, Borowska-Zuchowska N, Ainouche M, Kuderova A, Kovarik A (2020) The utility of graph clustering of 5S ribosomal DNA homoeologs in plant allopolyploids, homoploid hybrids, and cryptic introgressants. Front Plant Sci **11**: 41

Gómez JM, Perfectti F, Armas C, Narbona E, Gonzalez-Megias A, Navarro L, DeSoto L, Torices R (2020) Within-individual phenotypic plasticity in flowers fosters pollination niche shift. Nat Commun **11**: 1–12

Gómez JM, Gonzalez-Megias A, Narbona E, Navarro L, Perfectti F, Armas C (2021) Phenotypic plasticity guides *Moricandia arvensis* divergence and convergence across the Brassicaceae floral morphospace. New Phytol **233**: 1479–1493

Gong L, Salmon A, Yoo MJ, Grupp KK, Wang ZN, Paterson AH, Wendel JF (2012) The cytonuclear dimension of allopolyploid evolution: an example from cotton using rubisco. Mol Biol Evol **29**: 3023–3036

Gong L, Olson M, Wendel JF (2014) Cytonuclear evolution of rubisco in four allopolyploid lineages. Mol Biol Evol **31**: 2624–2636

Guo X, Liu J, Hao G, Zhang L, Mao K, Wang X, Zhang D, Ma T, Hu Q, Al-Shehbaz IA, et al. (2017) Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics **18**: 1–9

Guo X, Mandáková T, Trachtova K, Ozudogru 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

Gyorfy MF, Miller ER, Conover JL, Grover CE, Wendel JF, Sloan DB, Sharbrough J (2021) Nuclear-cytoplasmic balance: whole genome duplications induce elevated organellar genome copy number. Plant J **108**: 219–230

Hall JC, Tisdale TE, Donohue K, Wheeler A, Al-Yahya MA, Kramer EM (2011) Convergent evolution of a complex fruit structure in the tribe Brassiceae (Brassicaceae). Am J Bot **98**: 1989–2003

Hao GQ, Al-Shehbaz IA, Ahani H, Liang QL, Mao KS, Wang Q, Liu JQ (2017) An integrative study of evolutionary diversification of *Eutrema* (Eutremeae, Brassicaceae). Bot J Linn Soc **184**: 204–223

Heenan PB (2017) A taxonomic revision of *Cardamine* L. (Brassicaceae) in New Zealand. Phytotaxa **330**: 1–154

Heenan PB, Goeke DF, Houliston GJ, Lysak MA (2012) Phylogenetic analyses of ITS and rbcL DNA sequences for sixteen genera of Australian and New Zealand Brassicaceae result in the expansion of the tribe Microlepidieae. Taxon **61**: 970–979

Heidel AJ, Kiefer C, Coupland G, Rose LE (2016) Pinpointing genes underlying annual/perennial transitions with comparative genomics. BMC Genomics **17**: 1–9

Hewson HJ (1982) Brassicaceae (Cruciferae). *In* AS George, ed, Flora of Australia, Vol. 8. Australian Government Publishing Service, Canberra, pp 231–357

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

Hu S, Sablok G, Wang B, Qu D, Barbaro E, Viola R, Li M, Varotto C (2015) Plastome organization and evolution of chloroplast genes in *Cardamine* species adapted to contrasting habitats. BMC Genomics **16**: 1–14

Huang CH, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-Shehbaz I, Edger PP, et al. (2016) Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol Biol Evol **33**: 394–412

Huang DI, Cronk QC (2015) Plann: a command-line application for annotating plastome sequences. Appl Plant Sci **3**: apps.1500026

Huang XC, German DA, Koch MA (2020) Temporal patterns of diversification in Brassicaceae demonstrate decoupling of rate shifts and mesopolyploidization events. Ann Bot **125**: 29–47

Jabbour F, Renner SS (2012) A phylogeny of Delphinieae (Ranunculaceae) shows that *Aconitum* is nested within *Delphinium* and that Late Miocene transitions to long life cycles in the Himalayas and Southwest China coincide with bursts in diversification. Mol Phylogenet Evol **62**: 928–942

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

Joly S, Heenan PB, Lockhart PJ (2014) Species radiation by niche shifts in New Zealand's rockcresses (*Pachycladon*, Brassicaceae). Syst Biol **63**: 192–202

Kagale S, Robinson SJ, Nixon J, Xiao R, Huebert T, Condie J, Kessler D, Clarke WE, Edger PP, Links MG, et al. (2014) Polyploid evolution of the Brassicaceae during the Cenozoic era. Plant Cell **26**: 2777–2791

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

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30**: 772–780

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**: 1647–1649

Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, Soltis PS (2018) Impact of whole-genome duplication events on diversification rates in angiosperms. Am J Bot **105**: 348–363

Li D, Liu CM, Luo R, Sadakane K, Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics **31**: 1674–1676

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754–1760

Loytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science **320**: 1632–1635

Lundgren MR, Marais DLD (2020) Life history variation as a model for understanding trade-offs in plant–environment interactions. Curr Biol **30**: R180–R189

Lysak MA, Mandáková T (2013) Analysis of plant meiotic chromosomes by chromosome painting. Method Mol Biol **990**: 13–24

Lysak MA, Mandáková T, Schranz ME (2016) Comparative paleogenomics of crucifers: ancestral genomic blocks revisited. Curr Opin Plant Biol **30**: 108–115

Lysak MA, Edginton M, Zuo S, Guo X, Mandáková T, Al-Shehbaz IA (2022) Transfer of two *Arabidella* and two *Cuphonotus* species to the genus *Lemphoria* (Brassicaceae) and a description of the new species *L. queenslandica*. Phytotaxa **549**: 235–240

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

Mandáková T, Heenan PB, Lysak MA (2010b) Island species radiation and karyotypic stasis in *Pachycladon* allopolyploids. BMC Evolut Biol **10**: 1–14

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

Mandáková T, Lysak MA (2016) Chromosome preparation for cytogenetic analyses in *Arabidopsis*. Curr Protocol Plant Biol **1**: 43–51

Mandáková T, Pouch M, Harmanova K, Zhan SH, Mayrose I, Lysak MA (2017b) Multispeed genome diploidization and diversification after an ancient allopolyploidization. Mol Ecol **26**: 6445–6462

McCarthy EW, Chase MW, Knapp S, Litt A, Leitch AR, Le Comber SC (2016) Transgressive phenotypes and generalist pollination in the floral evolution of *Nicotiana* polyploids. Nat Plants **2**: 1–9

Muir G, Ruiz-Duarte P, Hohmann N, Mable BK, Novikova P, Schmickl R, Guggisberg A, Koch MA. (2015) Exogenous selection rather than cytonuclear incompatibilities shapes asymmetrical fitness of reciprocal *A. rabidopsis* hybrids. Ecol Evol **5**: 1734–1745

Mummenhoff K, Bruggemann H, Bowman JL (2001) Chloroplast DNA phylogeny and biogeography of *Lepidium* (Brassicaceae). Am J Bot **88**: 2051–2063

Negm S, Greenberg A, Larracuente AM, Sproul JS (2021) RepeatProfiler: a pipeline for visualization and comparative analysis of repetitive DNA profiles. Mol Ecol Resource **21**: 969–981

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

Nikolov LA, Shushkov P, Nevado B, Gan X, Al-Shehbaz IA, Filatov D, Bailey CD, Tsiantis M (2019) Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. New Phytologist **222**: 1638–1651

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

Novak P, Neumann P, Macas J (2020) Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nat Protoc **15**: 3745–3776

Ogburn RM, Edwards EJ (2015) Life history lability underlies rapid climate niche evolution in the angiosperm clade Montiaceae. Mol Phylogenet Evol **92**: 181–192

Oyston JW, Hughes M, Gerber S, Wills MA (2016) Why should we investigate the morphological disparity of plant clades? Ann Bot **117**: 859–879

Parins-Fukuchi C, Stull GW, Smith SA (2021) Phylogenomic conflict coincides with rapid morphological innovation. Proc Natl Acad Sci USA **118**: e2023058118

Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. R News **6**: 7–11

Rabosky DL, Grundler M, Anderson C, Title P, Shi JJ, Brown JW, Huang H, Larson JG (2014) BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. Methods Ecol Evol **5**: 701–707

Rabosky DL, Mitchell JS, Chang J (2017) Is BAMM flawed? Theoretical and practical concerns in the analysis of multi-rate diversification models. Syst Biol **66**: 477–498

Rannala B, Yang Z (2007) Inferring speciation times under an episodic molecular clock. Syst Biol **56**: 453–466

Rollins R, Rüdenberg L (1971) Chromosome numbers of Cruciferae. II. Contribut Gray Herbarium Harvard Univ **201**: 117–133

Sehrish T, Symonds VV, Soltis DE, Soltis PS, Tate JA (2015) Cytonuclear coordination is not immediate upon allopolyploid formation in *Tragopogon miscellus* (Asteraceae) allopolyploids. PLoS One **10**: e0144339

Sharbrough J, Conover JL, Tate JA, Wendel JF, Sloan DB (2017) Cytonuclear responses to genome doubling. Am J Bot **104**: 1277–1280

Sharbrough J, Conover JL, Gyorfy MF, Grover CE, Miller ER, Wendel.JF, Sloan DB (2021) Global patterns of subgenome evolution in organelle-targeted genes of six allotetraploid angiosperms. Mol Biol Evol **39**: msac074

Shaw EA (1965) Taxonomic revision of some Australian endemic genera of Cruciferae. Trans Royal Soc South Australia **89**: 145–253

Shaw EA (1974) Revisions of some genera of Cruciferae native to Australia. Contribut Gray Herbarium Harvard Univ **205**: 147–162

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet **15**: 121–132

Smith SA, Brown JW, Yang Y, Bruenn R, Drummond CP, Brockington SF, Walker JF, Last N, Douglas NA, Moore MJ (2018) Disparity, diversity, and duplications in the Caryophyllales. New Phytologist **217**: 836–854

Stitzer MC, Ross-Ibarra J (2018) Maize domestication and gene interaction. New Phytologist **220**: 395–408

Stull GW, Qu XJ, Parins-Fukuchi C, Yang YY, Yang JB, Yang ZY, Hu Y, Ma H, Soltis PS, Soltis DE, et al. (2021) Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. Nat Plants **7**: 1015–1025

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ (2015) Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytologist **207**: 454–467

Tate JA, Simpson BB (2003) Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. Syst Bot **28**: 723–737

Thomas BC, 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 Res **16**: 934–946

Vitales D, Garcia S, Dodsworth S (2020) Reconstructing phylogenetic relationships based on repeat sequence similarities. Mol Phylogenet Evol **147**: 106766

Walden N, German DA, Wolf EM, Kiefer M, Rigault P, Huang XC, Kiefer C, Schmickl R, Franzke A, Neuffer B, et al. (2020) Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. Nat Commun **11**: 1–12

Warwick SI, Mummenhoff K, Sauder CA, Koch MA, Al-Shehbaz IA (2010) Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. Plant Syst Evol **285**: 209–232

Wege J, Lepschi B (2007) A new species of *Arabidella* (Brassicaceae) from Western Australia. Nuytsia **17**: 453–458

Wicke S, Schneeweiss GM, dePamphilis CW, Muller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol **76**: 273–297

Wozniak NJ, Kappel C, Marona C, Altschmied L, Neuffer B, Sicard A (2020) A similar genetic architecture underlies the convergent evolution of the Selfing Syndrome in *Capsella*. Plant Cell **32**: 935–949

Wu S, Zhang B, Keyhaninejad N, Rodriguez GR, Kim HJ, Chakrabarti M, Illa-Berenguer E, Taitano NK, Gonzalo MJ, Diaz A, et al. (2018) A common genetic mechanism underlies morphological diversity in fruits and other plant organs. Nat Commun **9**: 1–12

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**: 1586–1591

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**: 821–829

Zhu Z, Chen G, Guo X, Yin W, Yu X, Hu J, Hu Z (2017) Overexpression of SlPRE2, an atypical bHLH transcription factor, affects plant morphology and fruit pigment accumulation in tomato. Sci Rep **7**: 1–11

Phytotaxa 549 (2): 235–240
https://www.mapress.com/pt/
Copyright © 2022 Magnolia Press

Article

ISSN 1179-3155 (print edition)
PHYTOTAXA
ISSN 1179-3163 (online edition)

https://doi.org/10.11646/phytotaxa.549.2.9

Transfer of two *Arabidella* and two *Cuphonotus* species to the genus *Lemphoria* (Brassicaceae) and a description of the new species *L. queenslandica*

MARTIN A. LYSAK[1,2,6*], MARK EDGINTON[3,7], SHENG ZUO[1,2,8], XINYI GUO[1,9], TEREZIE MANDÁKOVÁ[1,4,10] & IHSAN A. AL-SHEHBAZ[5,11*]

[1] *CEITEC, Masaryk University, Brno, CZ-625 00, Czech Republic*
[2] *NCBR, Faculty of Science, Masaryk University, Brno, CZ-625 00, Czech Republic*
[3] *Queensland Herbarium, Department of Environment and Science, Brisbane Botanic Gardens, Mt Coot-tha Road, TOOWONG QLD 4066 Australia*
[4] *Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, CZ-625 00, Czech Republic*
[5] *Missouri Botanical Garden, St. Louis, Missouri, U.S.A. 63110*
[6] *martin.lysak@ceitec.muni.cz; https://orcid.org/0000-0003-0318-4194*
[7] *mark.edginton@des.qld.gov.au; https://orcid.org/0000-0002-6907-4310*
[8] *sheng.zuo@ceitec.muni.cz; https://orcid.org/0000-0002-2104-3726*
[9] *xinyi.guo@ceitec.muni.cz; https://orcid.org/0000-0001-5416-7787*
[10] *terezie.mandakova@ceitec.muni.cz; https://orcid.org/0000-0001-6485-0563*
[11] *ihsan.al-shehbaz@mobot.org; https://orcid.org/0000-0003-1822-4005*
**Authors for correspondence*

Abstract

The taxonomic limits of *Arabidella, Cuphonotus,* and *Lemphoria* (Brassicaceae, Microlepidieae) are revised based on a critical evaluation of morphology in light of recent cytogenomic and molecular phylogenetic findings. As a result, *Lemphoria* is re-established to include two species previously placed in *Cuphonotus* and two in *Arabidella. Lemphoria queenslandica* is described as a new species, and the new combinations *L. andraeana, L. eremigena, L. humistrata*, and *L. procumbens* are proposed. Keys to distinguish *Arabidella* and *Lemphoria* species and an expanded generic description of *Lemphoria* are provided.

Keywords: *Arabidella*, Australia, Brassicaceae, *Cuphonotus, Lemphoria,* New Zealand

Introduction

The Microlepidieae (Brassicaceae) was recognized as a monophyletic tribe comprising cruciferous genera endemic to Australia and New Zealand (Warwick *et al.*, 2010). Monophyly of the tribe has been confirmed by several phylogenetic analyses (e.g., Heenan *et al.*, 2012; Mandáková *et al.*, 2017; Zuo *et al.*, 2022). The Microlepidieae included 16 genera and 56 species (Heenan *et al.*, 2012), all of which except the New Zealand *Pachycladon* Hooker (1867: 724) (11 species) are endemic to the Australian mainland and adjacent islands (e.g., Kangaroo Island, Tasmania). Among the 15 Australian genera, *Stenopetalum* R.Br. ex Candolle (1821: 513) (10 spp.) and *Arabidella* (Mueller 1853: 368) Schulz (1924: 177) (7 spp.) were the largest, while *Cuphonotus* Schulz (1933: 92) (2 spp.) was among the smallest. The taxonomic accounts of the Australian taxa treated by Schulz (1924), Shaw (1965, 1974), and Hewson (1982) formed the basic background on which the molecular data are compared.

Cytogenomic and phylogenetic analyses have shown that a common ancestor of the tribe Microlepidieae underwent a shared whole-genome duplication. The allotetraploid genome(s) with presumably 30 chromosomes ($n = 15$) have diverged into new species, genera, and intra-tribal clades that exhibit varying degrees of genome (re)diploidization (Mandáková *et al.*, 2010a, b; 2017). The different pace of genome diploidization was also encountered among *Arabidella* species (Mandáková *et al.*, 2017). While *A. trisecta* (Mueller 1853: 368) Schulz (1924: 179) ($2n = 24$, 48) has 12 chromosomes in the base chromosome set, *A. eremigena* (Mueller 1861: 143) Shaw (1965: 197) has only six chromosome pairs ($2n = 12$), and the two species formed two different clades within the tribe. The most recent

Accepted by Karol Marhold: 23 Mar. 2022; published: 8 Jun. 2022

plastome-based phylogenetic tree of Microlepidieae (Zuo *et al.*, 2022) has confirmed the previous conclusion that the genus *Arabidella* is not monophyletic (Heenan *et al.*, 2012; Mandáková *et al.*, 2017), and the two major intra-generic clades differ by the level of post-polyploid genome diploidization.

The highly resolved plastome phylogeny (Figure 1; Zuo *et al.*, 2022) recovered the *Arabidella* species as three distinct clades, consistent with the different extent of their post-polyploid diploidization and previous analyses of two single-copy nuclear genes (Mandáková *et al.*, 2017), and grouped both *Cuphonotus* species with the rapidly diploidizing *Arabidella* species. Here we reflect these phylogenomic data by resurrecting the genus *Lemphoria* Schulz (1924: 267) to accommodate two former *Arabidella* species [*A. eremigena*, *A. procumbens* (Tate 1885: 67) Shaw (1965: 200)] and to embrace both *Cuphonotus* species [*C. andraeanus* (Mueller 1885: 49) Shaw (1974: 157) and *C. humistratus* (Mueller 1878: 25) Schulz (1933: 92]. *Lemphoria queenslandica* is described as new, and the newly circumscribed *Arabidella* contains four species, namely *A. nasturtium* (Mueller 1853: 368) Shaw (1965: 191), *A. filifolia* (Mueller 1853: 368) Shaw (1965: 188), *A. glaucescens* Shaw (1965: 184), and *A. trisecta* (Mueller 1853: 368) Schulz (1924: 179), and a potentially new species (Figure 1; Zuo *et al.*, 2022). The strongly supported phylogenetic affinity of *A. chrysodema* and some *Menkea* Lehm. (Figure 1; Zuo *et al.*, 2022) suggests further study and formal taxonomic treatment of this species.

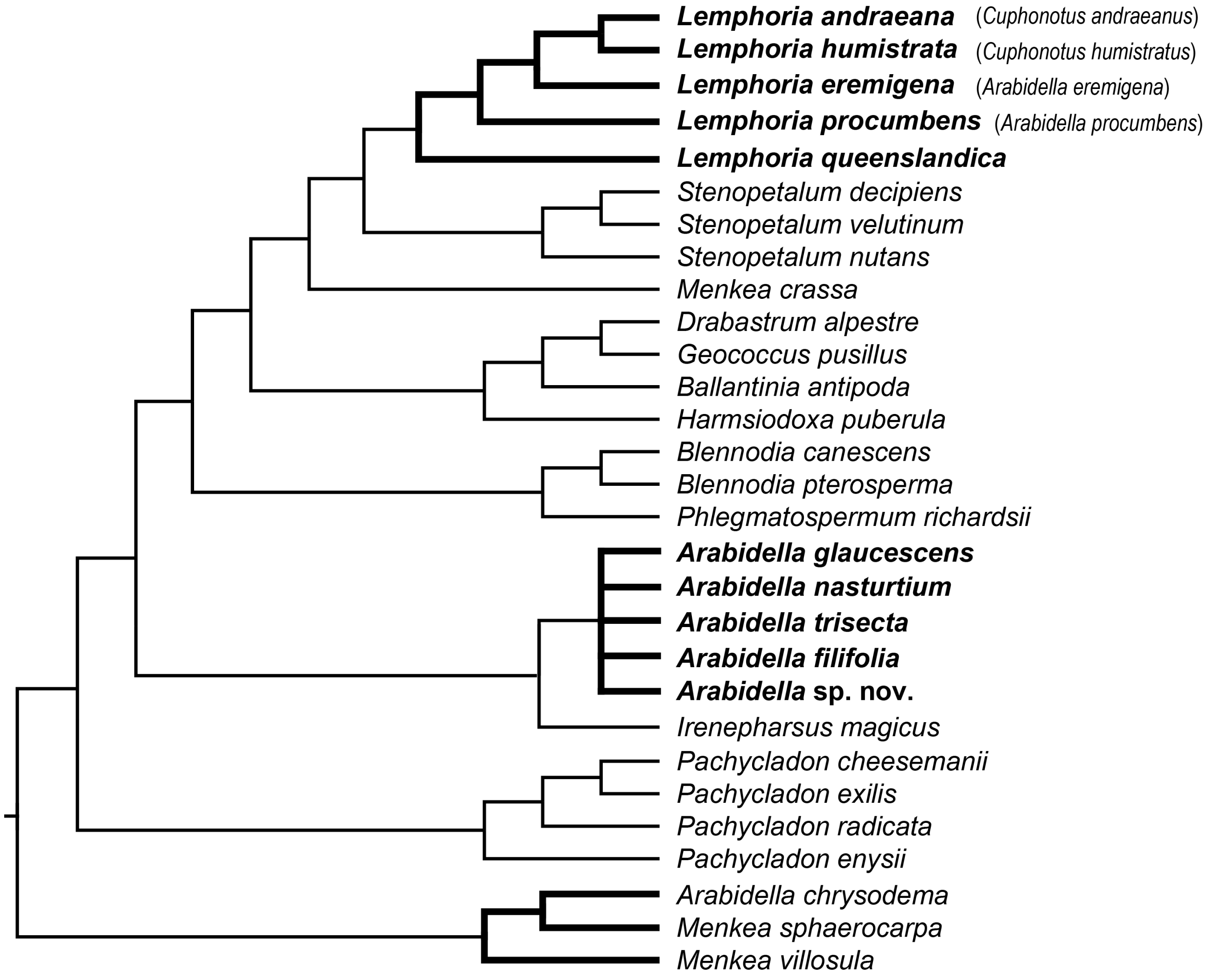


FIGURE 1. A simplified phylogenetic scheme of the tribe Microlepidieae based on whole chloroplast genome sequences (adapted from Zuo *et al.*, 2022). Three monophyletic clades (*Arabidella sensu stricto*, *Lemphoria*, and *Menkea*/*A. chrysodema*) within the former broadly circumscribed genus *Arabidella* are displayed in bold.

Arabidella and *Lemphoria* can be easily separated as follows:

1a. Shrubs or subshrubs, rarely annual herbs; lower leaves 2-, 3- or rarely multisect into linear to filiform lobes; nectar glands confluent, median glands present; ovules 20–90 per ovary; fruits linear, subterete or slightly latiseptate .... *Arabidella*
1b. Annual herbs; lower leaves pinnatifid to pinnatisect; nectar glands lateral, not confluent, median glands absent; ovules 6–70 per ovary; fruits oblong to linear siliques, or elliptic to suborbicular silicles, subterete or angustiseptate .... *Lemphoria*

Following the transfer of two of the six species of *Arabidella* sensu Shaw (1965) to *Lemphoria*, the remaining species can be distinguished by the following key.

1a. Annual herbs with basal leaf rosette at least during flowering; stems without bark .... *A. nasturtium*
1b. Woody shrubs or subshrubs without basal leaf rosette; stems usually with some bark.
2a. Fruiting pedicel divaricate to horizontal or slightly recurved; all leaves undivided, rarely some 2- or 3-sect .... *A. filifolia*
2b. Fruiting pedicels suberect to ascending, straight; almost all leaves 2-, 3-, or multisect.
3a. Stems glabrous; flowers in corymb; fruit 1–2.5 mm wide; gynophore (0.8–)1–2.2 mm long .... *A. glaucescens*
3b. Stems often papillate; flowers in lax raceme; fruit 0.8–1.5 mm wide; gynophore 0.1–0.5 mm long .... *A. trisecta*

Taxonomy

Lemphoria Schulz (1924: 267).

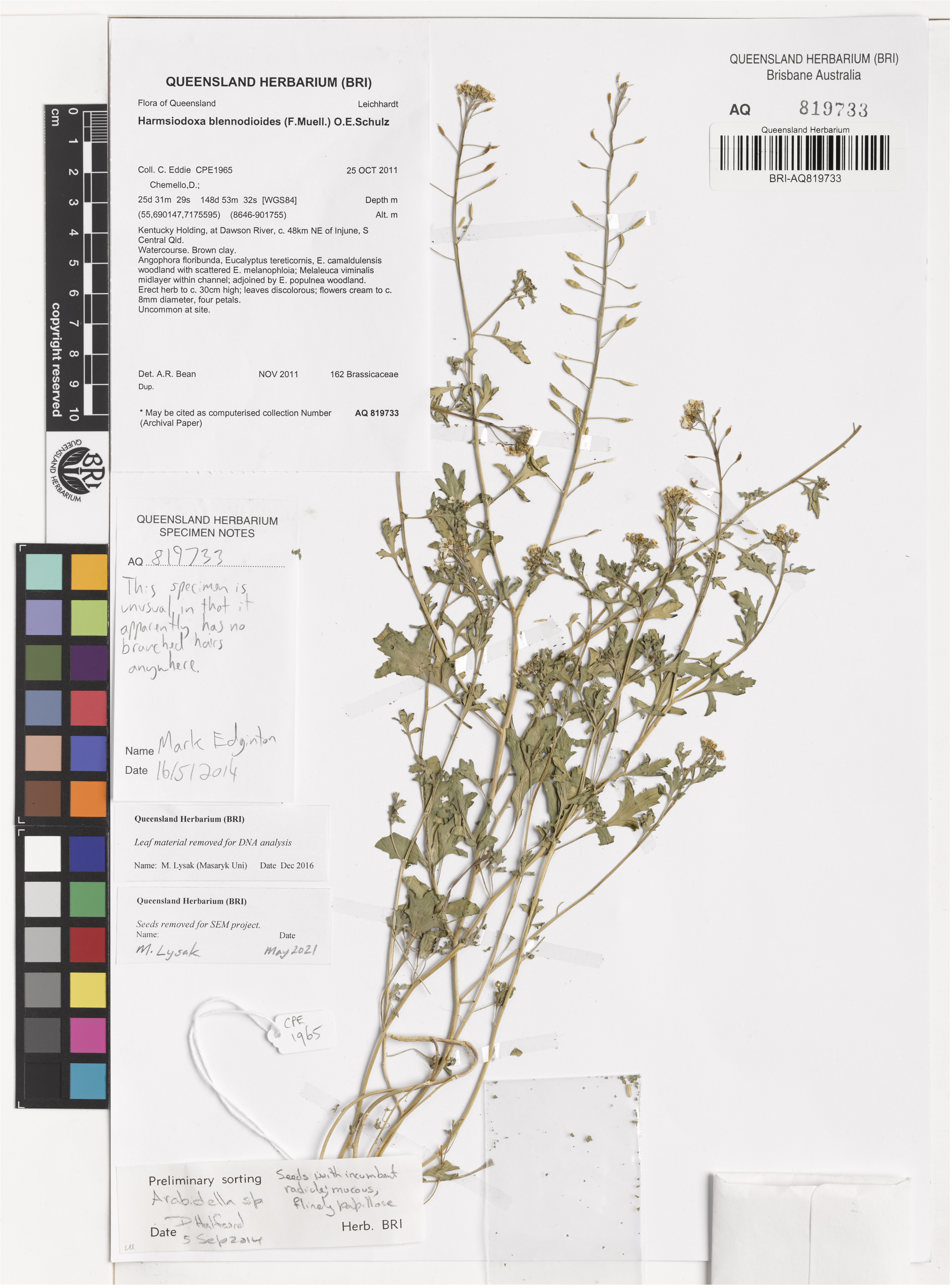
Type: *L. procumbens* (Tate) O.E.Schulz.
Cuphonotus Schulz (1933: 92), *syn. nov.* Lectotype (designated by Shaw 1974: 154): *C. humistratus* (F.Muell.) O.E.Schulz.

Herbs, annual. Trichomes absent or simple and slender, crisped or straight. Multicellular glands absent. Stems herbaceous, erect to ascending or decumbent, several from base, branched throughout, leafy, unarmed. Basal leaves rosulate or not, entire, pinnatifid to pinnatisect, subfleshy or not; cauline leaves more or less similar to basal leaves, short petiolate to subsessile, not auriculate at base, uppermost leaves entire. Racemes many flowered, ebracteate, corymbose, elongated considerably in fruit, rarely reduced to solitary flowers on pedicels originating from basal rosette; rachis straight; fruiting pedicels suberect to ascending or divaricate, persistent. Sepals ovate to oblong or elliptic, free, deciduous, spreading to reflexed, equal, base of lateral pair not saccate; petals white or yellow, spreading, subequaling to longer than sepals; blade oblanceolate to linear or obovate to suborbicular, apex obtuse; claw obscurely to strongly differentiated from blade, shorter than sepals, glabrous, unappendaged, entire; stamens 6, slightly exserted, erect to spreading, slightly tetradynamous; filaments wingless, unappendaged, glabrous, free, distinctly dilated at base; anthers ovate, obtuse at apex; nectar glands lateral, at both sides of lateral stamens, median glands absent; ovules 6−70 per ovary; placentation parietal. Fruit dehiscent, capsular, linear to oblong siliques or elliptic to suborbicular silicles, subterete or angustiseptate, not inflated, unsegmented; valves papery, prominently to obscurely veined, glabrous, not keeled, smooth, wingless, unappendaged; gynophore absent or obsolete; replum rounded, visible; septum complete, membranous, veinless or with an obscure midvein; style obsolete or 0.1−1 mm long, persistent; stigma capitate, entire or slightly 2-lobed, unappendaged. Seeds uniseriate or biseriate, wingless or margined, oblong to ellipsoid, plump or slightly flattened; seed coat minutely reticulate to nearly smooth, copiously mucilaginous when wetted; cotyledons incumbent.

Distribution: Endemic to Australia.

Key to *Lemphoria* species

1a. Fruit elliptic to suborbicular, strongly angustiseptate silicles; ovules 6−14 per ovary; seeds uniseriate.
2a. Plants pubescent; fruit valves smooth; petals 1−1.5 mm long; at least some leaves pinnatifid .... *L. andraeana*
2b. Plants glabrous to subglabrous; fruit valves conspicuously reticulate; petals 2−3 mm long; all leaves usually entire .... *L. humistrata*
1b. Fruit oblong or linear, subterete or slightly angustiseptate siliques; ovules 20−70 per ovary; seeds biseriate.
3a. Fruit oblong to ellipsoid, 2−2.5 mm wide; ovules 20−40 per ovary; cauline leaves broadly spatulate to obovate, 3−5-lobed .... *L. queenslandica*
3b. Fruit linear, 0.6−1.4 mm wide; ovules 50−70 per ovary; cauline leaves narrowly oblanceolate, (3−)5−15-lobed.
4a. Plants pilose throughout; petals 2.5–4 × 0.9–2 mm .... *L. eremigena*
4b. Plants glabrous; petals 1.4–2.2 × 0.4–0.7 mm .... *L. procumbens*


QUEENSLAND HERBARIUM (BRI)
Brisbane Australia
AQ 819733
Queensland Herbarium
BRI-AQ819733

QUEENSLAND HERBARIUM (BRI)
Flora of Queensland Leichhardt
Harmsiodoxa blennodioides (F.Muell.) O.E.Schulz

Coll. C. Eddie CPE1965 25 OCT 2011
Chemello,D.;
25d 31m 29s 148d 53m 32s [WGS84] Depth m
(55,690147,7175595) (8646-901755) Alt. m
Kentucky Holding, at Dawson River, c. 48km NE of Injune, S Central Qld.
Watercourse. Brown clay.
Angophora floribunda, Eucalyptus tereticornis, E. camaldulensis woodland with scattered E. melanophloia; Melaleuca viminalis midlayer within channel; adjoined by E. populnea woodland.
Erect herb to c. 30cm high; leaves discolorous; flowers cream to c. 8mm diameter, four petals.
Uncommon at site.

Det. A.R. Bean NOV 2011 162 Brassicaceae
Dup.
* May be cited as computerised collection Number AQ 819733
(Archival Paper)

QUEENSLAND HERBARIUM
SPECIMEN NOTES
AQ 819733
This specimen is unusual in that it apparently has no branched hairs anywhere.
Name Mark Edginton
Date 16/5/2014

Queensland Herbarium (BRI)
Leaf material removed for DNA analysis
Name: M. Lysak (Masaryk Uni) Date Dec 2016

Queensland Herbarium (BRI)
Seeds removed for SEM project.
Name: M. Lysak Date May 2021

CPE 1965

Preliminary sorting Arabidella sp D Halford
Date 5 Sep 2014
Seeds with incumbent radicle; mucous, finely papillose
Herb. BRI

FIGURE 2. Photo of the holotype of *Lemphoria queenslandica* at the Queensland Herbarium (BRI).

1. ***Lemphoria andraeana*** (F.Muell.) Al-Shehbaz & Lysak, *comb. nov.* Basionym: *Capsella andraeana* Mueller (1885: 49).
2. ***Lemphoria eremigena*** (F.Muell.) Al-Shehbaz & Lysak, *comb. nov.* Basionym: *Sisymbrium eremigenum* Mueller (1861: 143).
3. ***Lemphoria humistrata*** (F.Muell.) Al-Shehbaz & Lysak, *comb. nov.* Basionym: *Capsella humistrata* Mueller (1878: 25).
4. ***Lemphoria procumbens*** (Tate) Schulz (1924: 268). *Sisymbrium procumbens* Tate, (1885:67).
5. ***Lemphoria queenslandica*** Edginton, Al-Shehbaz & Lysak, *sp. nov.*

Diagnosis:—*Lemphoria queenslandica* is easily distinguished from the related *L. procumbens* and *L. eremigena* by having ellipsoid to oblong (vs. linear) fruits 2–2.5 (vs. 0.6–1.4) mm wide, 20–40 (vs. ca. 50–70) ovules per ovary, and broadly spatulate to obovate and 3–5-lobed (vs. narrowly oblanceolate and [3–]5–15-lobed) cauline leaves. It also differs from *L.procumbens* by being pubescent throughout (vs. glabrous).

Type:—AUSTRALIA, Queensland, Kentucky Holding, at Dawson River, ca. 48 km NE of Injune, S Central Qld., 25°31‘29“S, 148°53‘32“E, 25 Oct 2011, *C. Eddie CPE1965* (Holotype: BRI AQ-819733). Figure 2.

Description:—Herbs annual, usually pubescent with spreading trichomes 0.3–0.6 mm long. Stems erect, ca. 40 cm tall, branched. Basal leaves not seen. Cauline leaves obovate to broadly spatulate in outline, strongly 3- or pinnately 5-lobed, to 7 cm long, gradually smaller on the stem distally, sparsely to moderately pubescent; petiole 0.4–2 cm long; lobes dentate, subacute, laterals smaller than terminal one. Racemes corymbose, 20–40-flowered, elongated in fruit; rachis straight; fruiting pedicels ascending to divaricate, 8–16.5 mm long, sparsely to moderately pubescent. Sepals oblong-ovate, ca. 2.5 mm long, sparsely pubescent, hyaline margined; petals white or possibly yellow, ca. 4 × 2 mm, suborbicular, narrowed to claw ca. 2 mm long; stamens 6, tetradynamous, 2–3 mm long; anthers oblong; ovary sparsely pubescent, 20–40 ovuled. Fruits dehiscent, ellipsoid to oblong, 4–6.5 × 2–2.5 mm, sessile, slightly angustiseptate, reticulate veined, sparsely pubescent to subglabrous; septum complete, translucent; style 0.5–0.8 mm long; stigma entire. Seeds uniseriate, oblong to ovoid, plump, 0.7–1 × ca. 0.5 mm, brownish, finely papillate, mucilaginous when wetted; cotyledons incumbent.

Phenology:—As the novel species is only known from the type specimen collected on October 25, it is not possible to draw firm conclusions on the length and timing of the fruiting and flowering seasons. However, the inflorescence of the type specimen has mature fruits proximally and new flowers and buds distally. This implies that the species has quite a lengthy flowering and fruiting season.

Etymology:—The species epithet is named after the Australian state Queensland.

Distribution:—The species is known thus far only from the type collection on the Dawson River, 48 km NE of Injune, south-central Queensland.

Discussion:—*Lemphoria queenslandica* is most closely related to *L. eremigena* and *L. procumbens*, which it resembles by having siliques and biseriate seeds. It differs from both of them by the ellipsoid to oblong and slightly angustiseptate (vs. linear and subterete) fruits 2–2.5 (vs. 0.6−1.4) mm wide, fewer (20−40 vs. 50−70) ovules per ovary, and broadly spatulate to obovate (vs. narrowly oblanceolate) and 3−5-lobed [vs. (3−)5−15-lobed] cauline leaves. It differs from the other two species of the genus (*L. andraeana* and *L. humistrata*) by the oblong and slightly angustiseptate siliques (vs. strongly angustiseptate and elliptic to suborbicular silicles) and biseriate (vs. uniseriate) seeds in each fruit locule.

Acknowledgements

This work was financially supported by a research grant from the Czech Science Foundation (20-03419Y). We thank Neville Walsh (Royal Botanic Gardens Victoria), Anthony Brach (Harvard University Herbaria), Gillian Brown (Queensland Herbarium), Daniel Duval (Botanic Gardens of South Australia), Neville Scarlett, Brendan Lepschi (Centre for Australian National Biodiversity Research), Helen Vonow (State Herbarium of South Australia), and Janet Terry (Millennium Seed Bank, Royal Botanic Gardens, Kew).

References

Candolle, A.P. de (1821) *Regni vegetabilis systema naturalle, sive ordines, genera et species plantarum secundum methodi naturalis normus digestarum et descriptarum,* vol. 2.Treuttel and Würtz, Paris, 745 pp.

Heenan, P.B., Goeke, D.F., Houliston, G.J. & Lysak, M.A. (2012) Phylogenetic analyses ofITS and *rbc*L DNA sequences for sixteen genera of Australian and New Zealand Brassicaceae result in the expansion of the tribe Microlepidieae. *Taxon* 61: 970–979. https://doi.org/10.1002/tax.615004

Hewson, H.J. (1982) Brassicaceae. *In:* Briggs, B.G., Barlow, B.A., Eichler, H., Pedley, L., Ross, J., H., Symon, D.E. & Wilson, P.G. (Eds.) *Flora of Australia*, vol. 8. Australian Government Publishing Service, Canberra, 420 pp.

Hooker, J.D. (1867) *Handbook of the New Zealand flora: a systematic description of the native plants of New Zealand and the Chatham, Kermadec's, Lord Auckland's, Campbell'sand Macquarrie's Islands,* part 2. Reeve & Co, London, pp. 393–798. https://doi.org/10.5962/bhl.title.132966

Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K. & Lysak, M.A. (2010a) Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *The Plant Cell* 22: 2277–2290. https://doi.org/10.1105/tpc.110.074526

Mandáková, T., Heenan, P.B. & Lysak, M.A. (2010b) Island species radiation and karyotypic stasis in *Pachycladon* allopolyploids. *BMC Evolutionary Biology* 10: 1–14. https://doi.org/10.1186/1471-2148-10-367

Mandáková, T., Pouch, M., Harmanova, K., Zhan, S.H., Mayrose, I. & Lysak, M.A. (2017) Multispeed genome diploidization and diversification after an ancient allopolyploidization.*Molecular Ecology* 26: 6445–6462. https://doi.org/10.1111/mec.14379

Mueller, F.J.H.von (1853) Diagnoses et descriptions plantarum Novarum, quas in Nova Hollandia australi praecipua in regionibus interioribus. *Linnaea* 25: 367–445.

Mueller, F.J.H. von (1861) *Fragmenta Phytographiae Australiae*, vol. 2. Government Printer,John Ferres, Melbourne, 199 pp.

Mueller, F.J.H. von (1878) *Fragmenta Phytographiae Australiae*, vol. 11. GovernmentPrinter, John Ferres, Melbourne, 151 pp.

Mueller, F.J.H. von (1885) Definitions of some new Australian plants. *The Southern ScienceRecord and Magazine of Natural History, n. ser.* 1: 49–50.

Schulz, O.E. (1924) Cruciferae-Sisymbrieae. *In*: Engler, A. (Ed.) *Pflanzenreich* IV. 105 (Heft86). Verlag von Wilhelm Engelmann, Leipzig, 388 pp.

Schulz, O.E. (1933) Kurze Notizen über neue Gattungen, Sektionen und Arten der Cruciferen.*Botanische Jahrbücher fur Systematic Pflanzengeschichte und Pflanzengeographie* 66: 91–102.

Shaw, E.A. (1965) Taxonomic revision of some Australian endemic genera of Cruciferae. *Transactions of the Royal Society of South Australia* 89: 145–253.

Shaw, E.A. (1974) Revisions of some genera of Cruciferae native to Australia. *Contributionsfrom the Gray Herbarium of Harvard University* 205: 147–162.

Tate, R. (1885) Descriptions of new species of South Australian plants. *Transactions and Proceedings and Report of the Royal Society of South Australia* 7: 67–71.

Warwick, S.I., Mummenhoff, K., Sauder, C.A., Koch, M.A. & Al-Shehbaz, I.A. (2010) Closing the gaps: Phylogenetic relationships in the Brassicaceae based on DNA sequencedata of nuclear ribosomal ITS. *Plant Systematics and Evolution* 285: 209–232. https://doi.org/10.1007/s00606-010-0271-8

Zuo, S., Guo, X., Mandáková, T., Edginton, M., Al-Shehbaz, I.A. & Lysak, M.A. (2022) Genome diploidization associates with cladogenesis, trait disparity and plastid gene evolution in a crucifer tribe. *Plant Physiology.* [conditionally accepted] https://doi.org/10.1093/plphys/kiac268

Recurrent Plant-Specific Duplications of KNL2 and its Conserved Function as a Kinetochore Assembly Factor

Sheng Zuo,†[1,2] Ramakrishna Yadala,†[3] Fen Yang,[4,5] Paul Talbert,[6] Joerg Fuchs,[3] Veit Schubert,[3] Ulkar Ahmadli,[3] Twan Rutten,[3] Ales Pecinka,[4,5] Martin A. Lysak,[1,2] and Inna Lermontova*,[1,3]

[1]Central European Institute of Technology (CEITEC), Masaryk University, Kamenice 5, CZ-625 00 Brno, Czech Republic

[2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, CZ-625 00 Brno, Czech Republic

[3]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstrasse 3, D-06466 Seeland, Germany

[4]Institute of Experimental Botany, Czech Acad Sci, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, 779 00 Olomouc, Czech Republic

[5]Department of Cell Biology and Genetics, Faculty of Science, Palacký University, Šlechtitelů 27, 779 00 Olomouc, Czech Republic

[6]Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

***Corresponding author:** E-mail: lermonto@ipk-gatersleben.de.

Associate editor: Dr. Harmit Malik

†These authors contributed equally to this work.

Abstract

KINETOCHORE NULL2 (KNL2) plays key role in the recognition of centromeres and new CENH3 deposition. To gain insight into the origin and diversification of the *KNL2* gene, we reconstructed its evolutionary history in the plant kingdom. Our results indicate that the *KNL2* gene in plants underwent three independent ancient duplications in ferns, grasses, and eudicots. Additionally, we demonstrated that previously unclassified *KNL2* genes could be divided into two clades *αKNL2* and *βKNL2* in eudicots and *γKNL2* and *δKNL2* in grasses, respectively. *KNL2*s of all clades encode the conserved SANTA domain, but only the *αKNL2* and *γKNL2* groups additionally encode the CENPC-k motif. In the more numerous eudicot sequences, signatures of positive selection were found in both *αKNL2* and *βKNL2* clades, suggesting recent or ongoing adaptation. The confirmed centromeric localization of βKNL2 and mutant analysis suggests that it participates in loading of new CENH3, similarly to αKNL2. A high rate of seed abortion was found in heterozygous *βknl2* plants and the germinated homozygous mutants did not develop beyond the seedling stage. Taken together, our study provides a new understanding of the evolutionary diversification of the plant kinetochore assembly gene *KNL2*, and suggests that the plant-specific duplicated *KNL2* genes are involved in centromere and/or kinetochore assembly for preserving genome stability.

Key words: **adaptive evolution, CENH3, centromere, endopolyploidy, gene duplication, kinetochore, KNL2.**

Introduction

Centromeres are specific chromosomal regions where kinetochore protein complexes assemble in mitosis and meiosis to attach chromosomes to the spindle microtubules, and thus, are responsible for accurate segregation of chromosomes. Loss of centromere and kinetochore function causes chromosome missegregation, aneuploidy, and cell death (Fachinetti et al. 2013; McKinley and Cheeseman 2016; Barra and Fachinetti 2018). Centromere identity is specified epigenetically by the presence of the histone H3 variant termed CENH3 (also named CENP-A in mammals) which triggers the assembly of a functional kinetochore (Talbert et al. 2002). The kinetochore complexes are formed by dozens of proteins including the constitutive centromere-associated network complexes and outer kinetochore modules (Cheeseman and Desai 2008; Musacchio and Desai 2017; Hara and Fukagawa 2018).

KINETOCHORE NULL2 (KNL2, also termed M18BP1; Moree et al. 2011; Lermontova et al. 2013) plays a key role in new CENH3 deposition after replication. In vertebrates, M18BP1 (KNL2) is part of the Mis18 complex, including additionally Mis18α and Mis18β proteins. However, Mis18α and Mis18β in plants have not yet been identified. The human Mis18 complex is transiently present at centromeres prior to new CENH3 incorporation (Fujita et al. 2007); in chicken and *Xenopus*, the M18BP1 protein is present at centromeres throughout the cell cycle (French et al. 2017; Hori et al. 2017). In plants, KNL2 localizes at centromeres through the cell cycle, except from metaphase to late anaphase (Lermontova et al. 2013). The KNL2 proteins identified so far contain the characteristic SANTA (SANT-associated) domain (Zhang et al. 2006), a protein module of ∼90 amino acids which in some organisms is accompanied by a SANT/Myb-like putative DNA-binding domain. The functional role of

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Mol. Biol. Evol. 39(6):msac123 https://doi.org/10.1093/molbev/msac123 Advance Access publication June 7, 2022

SANTA and SANT domains has remained obscure for a long time. For instance, an interaction of KNL2 homologues containing the SANT/Myb domain with DNA has not yet been demonstrated, while *Arabidopsis thaliana* KNL2, which lacks this domain, showed DNA-binding capability *in vitro* and an association with the centromeric repeat *PAL1 in vivo* (Sandmann et al. 2017). Deletion of the SANTA domain in *Arabidopsis* KNL2 has not impaired its targeting to centromeres (Lermontova et al. 2013) nor disrupted its interaction with DNA (Sandmann et al. 2017). In *Xenopus*, a direct interaction of M18BP1 with CENH3 nucleosomes also did not require the SANTA domain (French et al. 2017). However, M18BP1 localizes at centromeres during metaphase—prior to CENH3 loading—by binding to CENP-C using the SANTA domain (French and Straight 2019).

A conserved CENPC-k motif, which is highly similar to the previously described CENPC motif of the CENP-C protein (Sugimoto et al. 1994; Talbert et al. 2004; Kato et al. 2013), was identified on the C-terminal part of the KNL2 homologues in a wide spectrum of eukaryotes (Kral 2016; Sandmann et al. 2017). The importance of this domain for the centromeric targeting of KNL2 was demonstrated in *Arabidopsis* (Sandmann et al. 2017), *Xenopus* (French et al. 2017), and chicken (Hori et al. 2017). Moreover, direct binding of CENPC-k to CENH3 nucleosomes was shown (French et al. 2017; Hori et al. 2017). In *Xenopus*, KNL2, similar to CENP-C, recruits the CENH3 chaperone HJURP to centromeres for new CENH3 assembly, and CENP-C competes with KNL2 for binding new CENH3 at centromeres (French et al. 2017). KNL2 in eutherian mammals lacks a CENPC-k motif (Kral 2016; Sandmann et al. 2017), and centromeric localization of human KNL2 may be achieved by direct binding of the SANTA domain to CENP-C (French and Straight 2019). Depletion of KNL2 in different organisms causes defects in CENH3 assembly (Fujita et al. 2007; Lermontova et al. 2013; French et al. 2017). For instance, knockout of M18BP1 as well as other components of the Mis18 complex in human HeLa cells with RNAi abolished centromeric recruitment of newly synthesized CENP-A, leading to chromosome missegregation and interphase micronuclei (Fujita et al. 2007). Embryos of homozygous mis18α mutant of mouse showed decreased DNA methylation, increased centromeric transcription, misaligned chromosomes, anaphase bridges, and lagging chromosomes, which was accompanied by embryo lethality (Kim et al. 2012). Unlike in mammals, the homozygous *knl2* mutant of *Arabidopsis* is viable despite reduced CENH3 levels and mitotic and meiotic abnormalities resulting in reduced growth rate and fertility (Lermontova et al. 2013). The fact that in the *knl2* mutant CENH3 is still localized at the centromeres suggests that this is not the only mechanism responsible for the centromeric loading of CENH3 in plants.

Although the functions of KNL2 are gradually being uncovered, research is still limited to a few model species, and in particular, the precise molecular mechanism of KNL2 interaction remains to be clarified. Up to now, robust phylogenetic analyses of the *KNL2* gene across large evolutionary time scales have not been reported. A better understanding of *KNL2* evolution may yield important insights into its role in CENH3 deposition and kinetochore assembly. To reconstruct the evolutionary history of the *KNL2* gene in plants, we compiled a data set of the proteins encoded by *KNL2* genes across major plant lineages from available genomic resources. Our phylogenetic analyses indicate that the *KNL2* gene in plants underwent three independent ancient duplications in ferns, grasses, and eudicots. We show that previously unclassified *KNL2* genes in eudicots could be divided into two clades (*αKNL2* and *βKNL2*). Both clades encode the conserved SANTA domain, but only the *αKNL2* group additionally encodes the conserved CENPC-k motif. Signatures of positive selection were found in both clades. Two additional *KNL2* clades (*γKNL2* and *δKNL2*) were identified in grasses. Similar to the divergence of αKNL2 and βKNL2 proteins, γKNL2 proteins retain the CENPC-k motif, while δKNL2 proteins have a shortened motif that resembles part of CENPC-k. In addition, analysis of RNA-seq data in *Arabidopsis* shows the *βKNL2* gene expression in nearly all tissues is considerably higher than the expression of *αKNL2*. Moreover, we provide the first evidence that βKNL2 localizes to centromeric regions in *Arabidopsis*. Mutant analysis of βKNL2 suggests that it participates in the loading of new CENH3 similarly to αKNL2. Taken together, our study provides a new understanding of the evolutionary origin and function of plant-specific duplicated KNL2 as a kinetochore assembly factor.

Results

Search for *KNL2* Genes in Plants Led to the Finding and Re-annotation of a New *KNL2* Variant in *Arabidopsis*

The KNL2 protein contains a conserved module designated as SANTA due to its association with the SANT domain. Although most metazoans have only one gene coding for a SANTA domain-containing protein, two genes (*At5g02520* and *At1g58210*) were identified in *Arabidopsis* (Zhang et al. 2006). Since the predicted protein encoded by the *At1g58210* gene contained in addition to the SANTA domain, a protein interaction kinase domain 1 (KIP1) and the C-terminal chromosome maintenance structural domain (SMC_Prok_B), completely atypical for previously described KNL2 proteins, we had previously excluded it from our research and focused on *At5g02520* (Lermontova et al. 2013).

However, based on the updated Araport-11 annotation (TAIR and Phytozome 13 database) and our in silico analysis, we found that the *At1g58210* gene encodes a protein of 281 amino acids including the SANTA domain but excluding KIP1 and SMC_Prok_B. We designated it as βKNL2 and the previously characterized KNL2 as αKNL2 (fig. 1A), in which full-length alpha and beta KNL2 have only 41.5% identity.

To investigate the origin and evolution of *KNL2* genes, we constructed a comprehensive proteome data set across major plant lineages including 90 representative species (fig. 1*B*,*C*). We performed a genome-wide search using the *Arabidopsis* αKNL2 (*At5g02520*) amino acid sequence and its conserved domains as the query for a local BLASTP search against the data set (supplementary fig. S1, Supplementary Material online). In total, 148 homologous conceptual protein sequences encoded by *KNL2* genes were identified in plant lineages including bryophytes (3 species:3 sequences), lycophytes (1:1), ferns (3:5), gymnosperms (7:7), and angiosperm species (67:132; fig. 1*B*,*D*; supplementary table S1and file S1, Supplementary Material online). For lycophytes, the *KNL2* gene was retrieved by TBLASTN search from *Selaginella moellendorffii* genome. Comparison with genomic and cDNA sequences in *S. moellendorffii* revealed that there is an intron right in the CENPC-k motif (supplementary file S2, Supplementary Material online). While the *KNL2* gene was detected in all investigated angiosperm species and ferns, it has not been identified in 4 out of 11 gymnosperm species investigated (*Cycas micholitzii*, *Ginkgo biloba*, *Gnetum montanum*, and *Taxus baccata*). The failure to find KNL2 in these species is likely because of incompletely assembled proteomes of gymnosperms at the time they were downloaded from the PLAZA genome database, not because of its absence in their genomes. Additionally, the *KNL2* gene also was not retrieved in any of the five algal species we examined. Based on the quality of the assembled algal proteomes (Merchant et al. 2007; Blanc et al. 2012; Collen et al. 2013), the *KNL2* gene may be absent in these genomes. However, we cannot exclude the possibility that *KNL2* has diverged beyond recognition by BLASTP and tBLASTN in algal genomes. In summary, the *KNL2* genes experienced recurrent ancient plant-specific duplication events.

KNL2 Gene in Plants Underwent Independent Duplications in Ferns, Grasses, and Eudicots

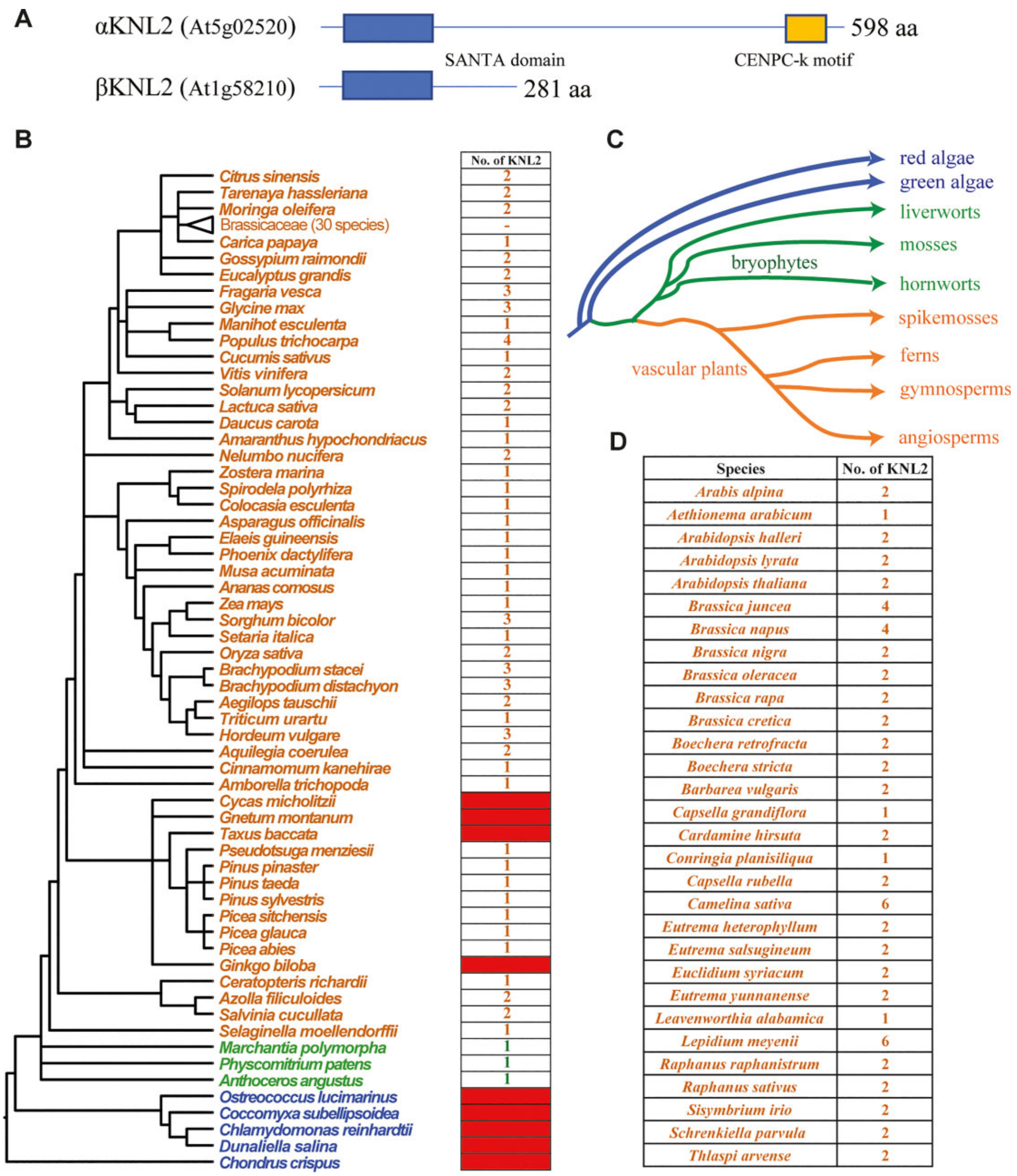
To better understand the *KNL2* gene diversification and evolution across the plant kingdom, we made a multiple sequence alignment of KNL2 proteins (supplementary file S3, Supplementary Material online) and constructed a phylogenetic tree. The topology of the Maximum Likelihood (ML) tree (fig. 2) shows that KNL2 proteins cluster into two branches in three plant clades—heterosporous water ferns (Salviniaceae), eudicots, and grasses (Poaceae)—indicating ancient gene duplications. Despite the deep divergence of the duplicated paralogs in ferns, their CENPC-k motifs are 83% identical. The grouping of a KNL2 protein of *Ceratopteris*, a member of the Polypodiales encompassing ∼80% of fern species, with one of the two KNL2 proteins of water ferns suggests that the duplication of *KNL2* in ferns occurred prior to the divergence of Salviniales and Polypodiales, more than 120 Ma (Qi et al. 2018). In angiosperms, gene duplication occurred after the divergence of *Amborella trichopoda* and monocots, but prior to the divergence of the basal eudicot *Nelumbo nucifera*, estimated at ∼100 Ma (Angiosperm Phylogeny website: http://www.mobot.org/MOBOT/research/APweb/; Friis et al. 2016). This duplication gave rise to the *αKNL2* and *βKNL2* genes of *Arabidopsis* and their orthologs in other eudicots. Monocots except for grasses (Poaceae) appear to have only one *KNL2* gene copy, while two paralogs in grasses indicate another gene duplication in the grass ancestor ∼100 Ma (Wu et al. 2018). In light of their separate origin from *αKNL2* and *βKNL2* in eudicots, these two paralogous copies in grasses were named γ*KNL2* and δ*KNL2*.

The αKNL2 and βKNL2 Paralogs Contain the SANTA Domain, but only αKNL2 is Characterized by the Presence of the C-terminal CENPC-k motif

Next, we focused on the *αKNL2* and *βKNL2* genes and their proteins mainly in Brassicales due to the extensive availability of genomic resources (supplementary fig. S2, supplementary file S4, Supplementary Material online). Except for a few neopolyploid species, the *αKNL2* and *βKNL2* gene numbers are conserved at one copy each across Brassicales species. These KNL2 proteins present several conserved features: the N-terminus contains the conserved SANTA domain in all KNL2 proteins, whereas only the αKNL2-type C-terminus possesses the CENPC-k motif. αKNL2 and βKNL2 sequences identified from Brassicales showed 41.0 and 57.2% pairwise identity, respectively.

We aligned all SANTA domains in KNL2 homologs from Brassicales species to show the conservation and variation and also made separate alignments for the SANTA domains in αKNL2 and βKNL2 paralogs (fig. 3*A*). The alignment results showed that SANTA domains from Brassicales species have 55.0% pairwise identity, while the similarity of these domains within αKNL2 paralogs is 71.0% and within βKNL2 paralogs is 72.3%, respectively. Many residues in the SANTA domains are conserved between both αKNL2 and βKNL2 paralogs. However, there are also amino acids specific to αKNL2 or βKNL2, suggesting that they might have different functions or interact with different proteins. For instance, one putative Aurora kinase phosphorylation consensus ($(R/K)X_{1\text{-}3}(S/T)$) can be detected in αKNL2 (fig. 3*A*, middle panel, aa 37–41) and three in βKNL2 (fig. 3*A*, lower panel, aa 37–41, 47–50, 69–72). In addition, we aligned SANTA domains from angiosperm species (minus Brassicales) and early diverging land plants (supplementary fig. S3, Supplementary Material online). As expected, SANTA domain variation increased with the phylogenetic divergence through evolutionary time. However, SANTA domains from nearly all paralogs maintain the previously identified conserved hydrophobic residues at the N- and C-termini, including the VxLxDW motif at the N-terminus of the SANTA domain and the GFxxxxxxxFxxGFPxxW motif at the C-terminus (Zhang et al. 2006).

In contrast to the SANTA domain, the CENPC-k motif is highly conserved throughout the plant kingdom where it



	No. of KNL2
Citrus sinensis	2
Tarenaya hassleriana	2
Moringa oleifera	2
Brassicaceae (30 species)	-
Carica papaya	1
Gossypium raimondii	2
Eucalyptus grandis	2
Fragaria vesca	3
Glycine max	3
Manihot esculenta	1
Populus trichocarpa	4
Cucumis sativus	1
Vitis vinifera	2
Solanum lycopersicum	2
Lactuca sativa	2
Daucus carota	1
Amaranthus hypochondriacus	1
Nelumbo nucifera	2
Zostera marina	1
Spirodela polyrhiza	1
Colocasia esculenta	1
Asparagus officinalis	1
Elaeis guineensis	1
Phoenix dactylifera	1
Musa acuminata	1
Ananas comosus	1
Zea mays	1
Sorghum bicolor	3
Setaria italica	1
Oryza sativa	2
Brachypodium stacei	3
Brachypodium distachyon	3
Aegilops tauschii	2
Triticum urartu	1
Hordeum vulgare	3
Aquilegia coerulea	2
Cinnamomum kanehirae	1
Amborella trichopoda	1
Cycas micholitzii	🟥
Gnetum montanum	🟥
Taxus baccata	🟥
Pseudotsuga menziesii	1
Pinus pinaster	1
Pinus taeda	1
Pinus sylvestris	1
Picea sitchensis	1
Picea glauca	1
Picea abies	1
Ginkgo biloba	🟥
Ceratopteris richardii	1
Azolla filiculoides	2
Salvinia cucullata	2
Selaginella moellendorffii	1
Marchantia polymorpha	1
Physcomitrium patens	1
Anthoceros angustus	1
Ostreococcus lucimarinus	🟥
Coccomyxa subellipsoidea	🟥
Chlamydomonas reinhardtii	🟥
Dunaliella salina	🟥
Chondrus crispus	🟥

Species	No. of KNL2
Arabis alpina	2
Aethionema arabicum	1
Arabidopsis halleri	2
Arabidopsis lyrata	2
Arabidopsis thaliana	2
Brassica juncea	4
Brassica napus	4
Brassica nigra	2
Brassica oleracea	2
Brassica rapa	2
Brassica cretica	2
Boechera retrofracta	2
Boechera stricta	2
Barbarea vulgaris	2
Capsella grandiflora	1
Cardamine hirsuta	2
Conringia planisiliqua	1
Capsella rubella	2
Camelina sativa	6
Eutrema heterophyllum	2
Eutrema salsugineum	2
Euclidium syriacum	2
Eutrema yunnanense	2
Leavenworthia alabamica	1
Lepidium meyenii	6
Raphanus raphanistrum	2
Raphanus sativus	2
Sisymbrium irio	2
Schrenkiella parvula	2
Thlaspi arvense	2

Fig. 1. Identification of the *KNL2* gene homologs across major plant lineages. (*A*) Protein structure of KNL2 in *Arabidopsis*. SANTA domain and CENPC-k motif are indicated by differently colored boxes. (*B*) The number of *KNL2* homologs in 90 representative plant species. The phylogenetic tree is adopted from the NCBI common tree. The blue-, green-, and orange-colored species names indicate alga, bryophytes, and vascular plants, respectively. The red filled boxes mean that we could not retrieved *KNL2* from these species. (*C*) Phylogenetic relationships of the analyzed species were adapted from Banks et al. (2011). (*D*) The number of *KNL2* homologs identified in analyzed crucifer (Brassicaceae) genomes.

is present (fig. 3*B*); however, the CENPC-k motif is missing from the βKNL2 and δKNL2 clades. Given that αKNL2 and βKNL2 paralogs may have been retained to perform distinct functions, we looked for additional conserved motifs in both variants from Brassicales species using the Multiple Em for Motif Elicitation (MEME) tool. Besides the motifs preserved in SANTA and CENPC-k regions (fig. 3), we also identified several additional conserved motifs that are unique to one or the other paralog (supplementary fig. S4, Supplementary Material online). For example, the N-termini of βKNL2 paralogs have a conserved motif 7 (21 aa), which is located upstream of the SANTA domain, but absent in αKNL2 paralogs (supplementary fig. S4, Supplementary Material online).

The KNL2 of Maize is Represented only by the δKNL2 Variant with a Truncated CENPC-k Motif

To observe the conserved features of KNL2, we also examined the γ*KNL2* and δ*KNL2* genes in grasses. γ*KNL2* encodes a SANTA domain and CENPC-k motif (supplementary file S5, Supplementary Material online), while δ*KNL2* encodes a SANTA domain and the motif RRLRSGKV/I, which resembles a truncated version of the

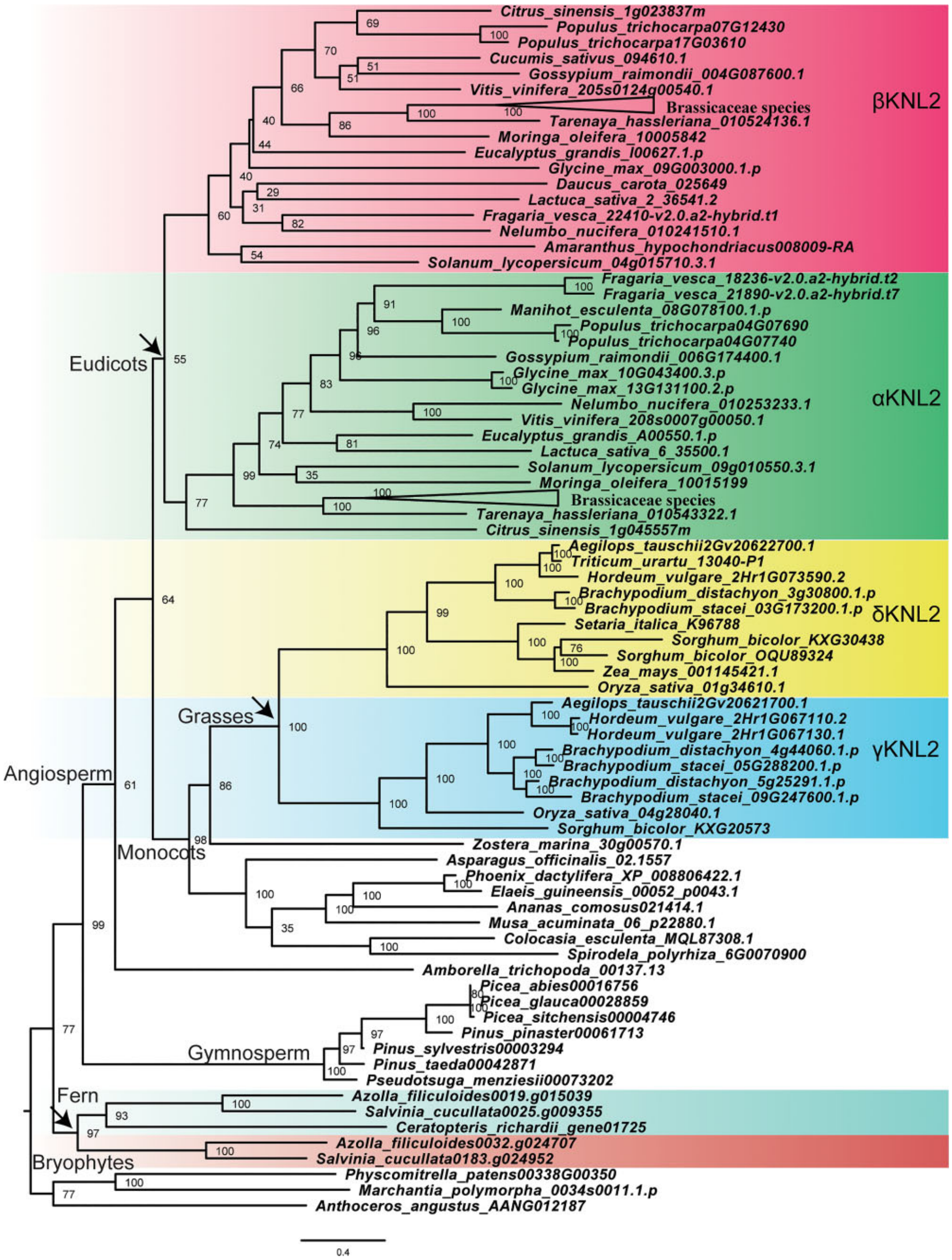

Citrus_sinensis_1g023837m
Populus_trichocarpa07G12430
Populus_trichocarpa17G03610
Cucumis_sativus_094610.1
Gossypium_raimondii_004G087600.1
Vitis_vinifera_205s0124g00540.1
Brassicaceae species
Tarenaya_hassleriana_010524136.1
Moringa_oleifera_10005842
Eucalyptus_grandis_I00627.1.p
Glycine_max_09G003000.1.p
Daucus_carota_025649
Lactuca_sativa_2_36541.2
Fragaria_vesca_22410-v2.0.a2-hybrid.t1
Nelumbo_nucifera_010241510.1
Amaranthus_hypochondriacus008009-RA
Solanum_lycopersicum_04g015710.3.1
βKNL2
Fragaria_vesca_18236-v2.0.a2-hybrid.t2
Fragaria_vesca_21890-v2.0.a2-hybrid.t7
Manihot_esculenta_08G078100.1.p
Populus_trichocarpa04G07690
Populus_trichocarpa04G07740
Gossypium_raimondii_006G174400.1
Glycine_max_10G043400.3.p
Glycine_max_13G131100.2.p
Nelumbo_nucifera_010253233.1
Vitis_vinifera_208s0007g00050.1
Eucalyptus_grandis_A00550.1.p
Lactuca_sativa_6_35500.1
Solanum_lycopersicum_09g010550.3.1
Moringa_oleifera_10015199
Brassicaceae species
Tarenaya_hassleriana_010543322.1
Citrus_sinensis_1g045557m
αKNL2
Eudicots
Aegilops_tauschii2Gv20622700.1
Triticum_urartu_13040-P1
Hordeum_vulgare_2Hr1G073590.2
Brachypodium_distachyon_3g30800.1.p
Brachypodium_stacei_03G173200.1.p
Setaria_italica_K96788
Sorghum_bicolor_KXG30438
Sorghum_bicolor_OQU89324
Zea_mays_001145421.1
Oryza_sativa_01g34610.1
δKNL2
Grasses
Aegilops_tauschii2Gv20621700.1
Hordeum_vulgare_2Hr1G067110.2
Hordeum_vulgare_2Hr1G067130.1
Brachypodium_distachyon_4g44060.1.p
Brachypodium_stacei_05G288200.1.p
Brachypodium_distachyon_5g25291.1.p
Brachypodium_stacei_09G247600.1.p
Oryza_sativa_04g28040.1
Sorghum_bicolor_KXG20573
γKNL2
Angiosperm
Monocots
Zostera_marina_30g00570.1
Asparagus_officinalis_02.1557
Phoenix_dactylifera_XP_008806422.1
Elaeis_guineensis_00052_p0043.1
Ananas_comosus021414.1
Musa_acuminata_06_p22880.1
Colocasia_esculenta_MQL87308.1
Spirodela_polyrhiza_6G0070900
Amborella_trichopoda_00137.13
Picea_abies00016756
Picea_glauca00028859
Picea_sitchensis00004746
Pinus_pinaster00061713
Pinus_sylvestris00003294
Pinus_taeda00042871
Pseudotsuga_menziesii00073202
Gymnosperm
Fern
Azolla_filiculoides0019.g015039
Salvinia_cucullata0025.g009355
Ceratopteris_richardii_gene01725
Azolla_filiculoides0032.g024707
Salvinia_cucullata0183.g024952
Bryophytes
Physcomitrella_patens00338G00350
Marchantia_polymorpha_0034s0011.1.p
Anthoceros_angustus_AANG012187
0.4

Fig. 2. Evolutionary relationship of KNL2 homologs in land plants. Maximum likelihood phylogenetic analysis was performed using IQ-tree with a protein alignment of KNL2 homologs in land plants. The *KNL2* genes cluster into two branches in three plant clades—heterosporous water ferns (Salviniaceae), eudicots, and grasses (Poaceae)—indicating ancient gene duplications (arrows). The KNL2 in eudicots and grasses can be classified into two major groups (αKNL2 and βKNL2, and γKNL2 and δKNL2, respectively). Bootstrap values obtained after 1,000 ultrafast bootstrap replicates (bb) are shown in the tree. The scale bar indicates the number of substitutions per site. The tree is arbitrarily rooted between bryophytes and tracheophytes.

CENPC-k motif (supplementary file S6, Supplementary Material online). γKNL2 and δKNL2 sequences from grasses showed 41.4 and 37.8% pairwise identity, respectively. Other non-grass monocot species only have one *KNL2* gene copy (fig. 2 and supplementary table S1, Supplementary Material online), and these single-copy *KNL2* genes more closely resemble the γ clade, encoding SANTA and CENPC-k motif, which is the ancestral state of *KNL2* before the grass-specific gene duplication. Interestingly, in eight reference proteomes of maize, we found only one copy of the *KNL2* gene, though with several splicing variants (supplementary fig. S5, Supplementary Material online). We also checked maize transcriptome data from different tissues and developmental stages; however, only *δKNL2* was identified (Maize RNA-seq Database: http://ipf.sustech.edu.cn/pub/zmrna/). We propose that unlike in other grass species, the maize genome contains only one copy of the *δKNL2* gene and has lost *γKNL2*.

Different Evolutionary Forces act on KNL2 Paralogs

We considered the possibility that selection may act differently on KNL2 paralogs. We used ML methods using the PAML suite (Yang 2007) to test for positive selection on each of the KNL2 paralogs in Brassicaceae species (supplementary file S7, Supplementary Material online). The branch-site model was used to test two KNL2 groups by using Codeml program (Yang 2007). Our PAML analyses revealed positive selection on both αKNL2 (fig. 4A, M1 vs. M2, $P = 2.104 \times 10^{-4}$ and M7 vs. M8, $P = 3.518 \times 10^{-5}$) and βKNL2 paralogs (M7 vs. M8, $P = 4.863 \times$

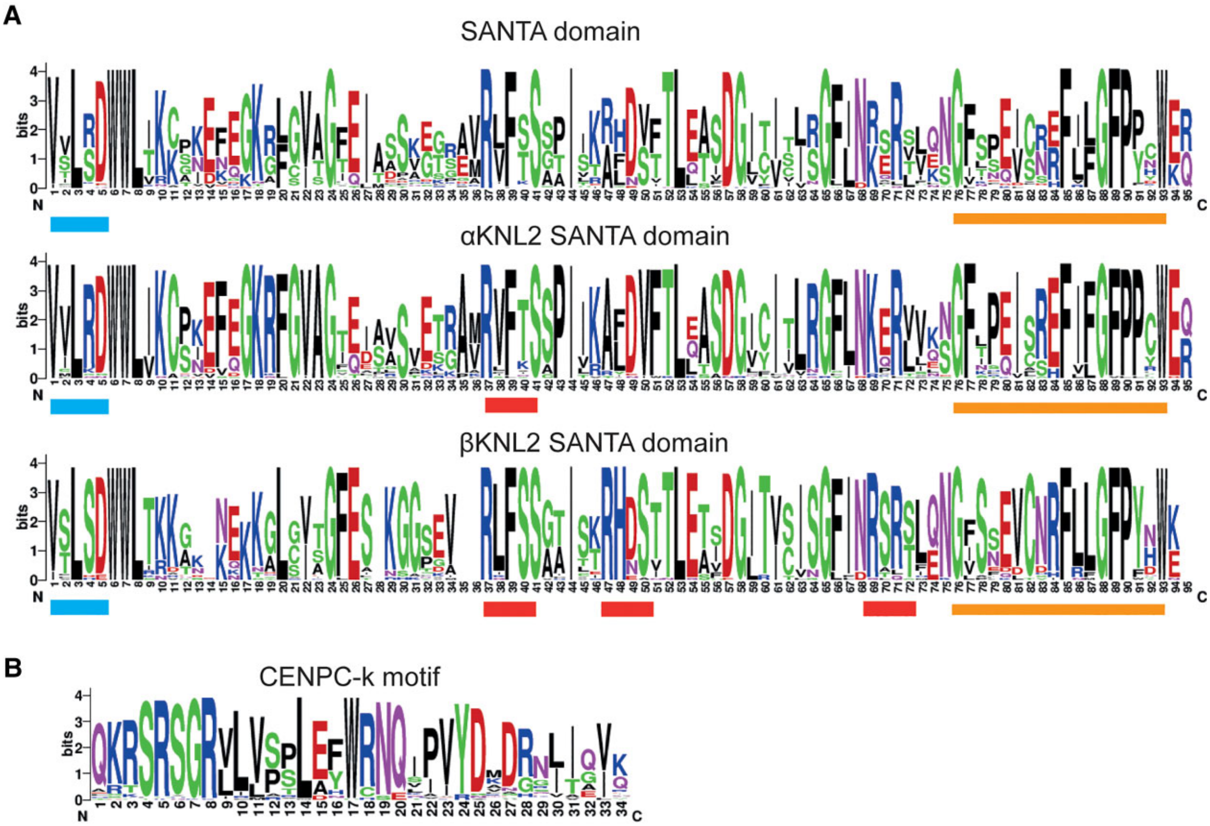
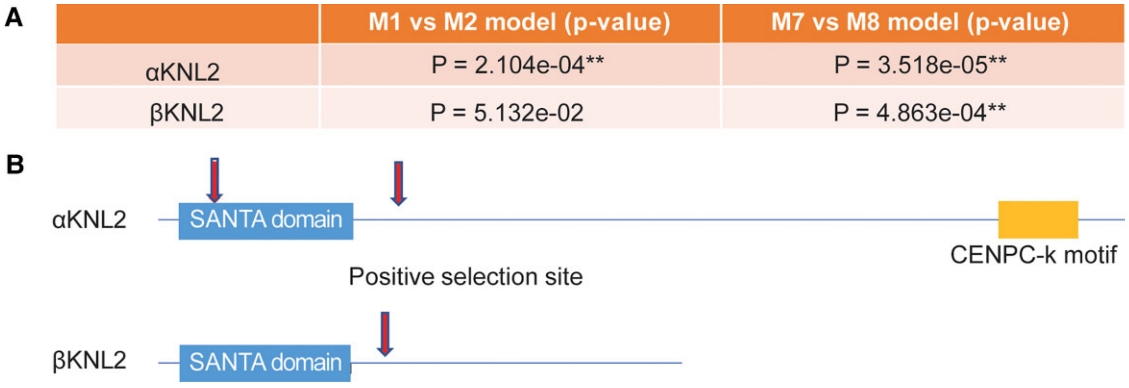


FIG. 3. Alignments of SANTA domain and CENPC-k motif in KNL2 homologs presented in LOGO format. (*A*) Variation map of the SANTA domain in the KNL2 homologs. The WebLogo program (http://weblogo.berkeley.edu/logo.cgi) was used to present SANTA domain alignments. The upper panel aligns SANTA domains of all KNL2 homologs from Brassicales, whereas the middle and bottom panels represent SANTA domain alignments of αKNL2 and βKNL2 homologs, respectively. The conserved N-terminal and C-terminal hydrophobic motifs are marked by blue and orange bars, respectively. Putative Aurora kinase phosphorylation consensus sites are underlined with red bars. (*B*) Alignment of CENPC-k motif of KNL2 homologs from land plants.



A

	M1 vs M2 model (p-value)	M7 vs M8 model (p-value)
αKNL2	P = 2.104e-04**	P = 3.518e-05**
βKNL2	P = 5.132e-02	P = 4.863e-04**

FIG. 4. Evolutionary pressures on the KNL2 paralogs. (*A*) Summary of tests for positive selection performed on KNL2 paralogs from Brassicaceae species. Statistically significant tests ($P < 0.05$) are indicated with asterisks. (*B*) A schematic of a representative KNL2 protein, showing sites evolving under positive selection identified by Bayes empirical Bayes analysis (posterior probability > 0.95).

10^{-4}). Bayes empirical Bayes analyses identified two amino acids in αKNL2 paralogs and one amino acid in βKNL2 paralogs as having evolved under positive selection with a high posterior probability ($>$0.95, fig. 4*B*). In αKNL2, the two positively selected sites are located in and slightly C-terminal to the SANTA domain (fig. 4*B*, supplementary fig. S6, Supplementary Material online). In βKNL2, the positively selected site also is located slightly C-terminal to the SANTA domain (fig. 4*B*, supplementary fig. S6, Supplementary Material online).

βKNL2 of *Arabidopsis* shows Centromeric Localization

We assessed the subcellular localization and putative biological function of the *Arabidopsis* βKNL2 variant *in vivo*.

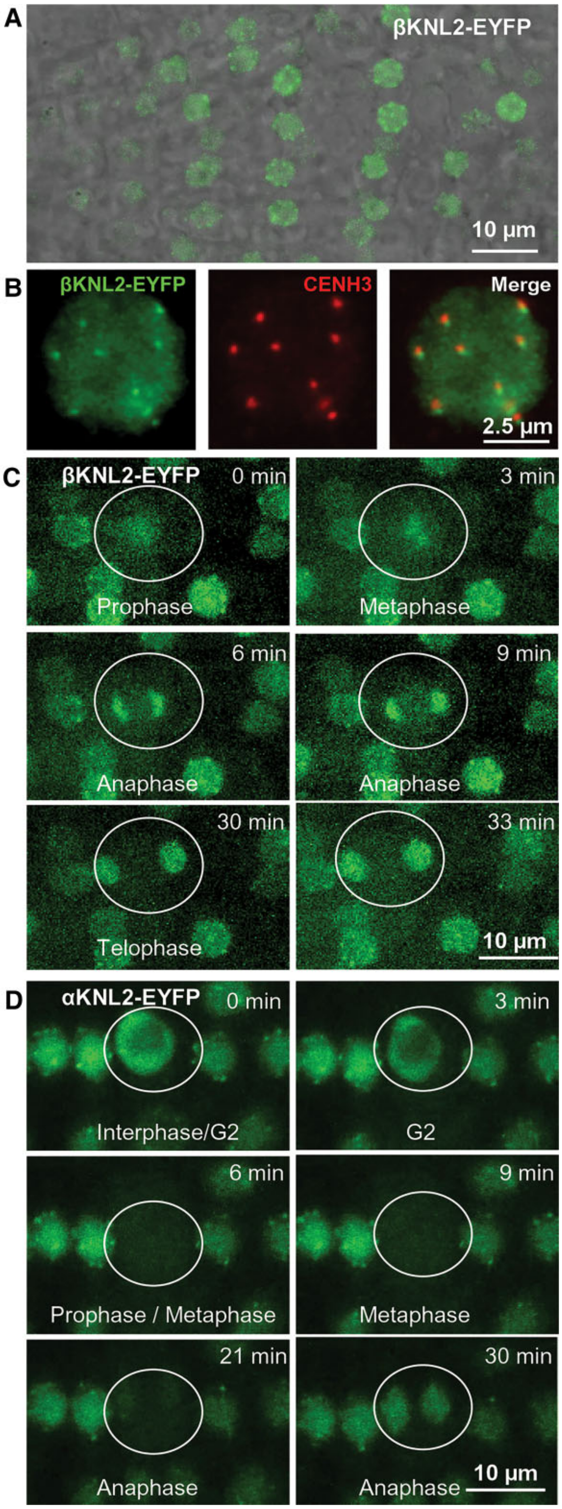


Fig. 5. Subcellular localization of βKNL2 in *Arabidopsis*. (*A*) Live imaging of root tip cells of *Arabidopsis* transformed with the βKNL2-EYFP and αKNL2-EYFP fusion constructs. Fluorescent signals showed distinct centromeric and diffused nucleoplasmic distribution. (*B*) Nucleus isolated from seedlings of the βKNL2-EYFP transformants after immunostaining with anti-GFP (left panel) and anti-CENH3 (middle panel) antibodies. Merge of both immunosignals (right panel). (*C*) Live imaging of root tip cells of *Arabidopsis* transformed with the βKNL2-EYFP fusion construct. (*D*) Live imaging of root tip cells of *Arabidopsis* transformed with the αKNL2-EYFP fusion construct. Cell undergoing mitosis is encircled.

To this end, the *βKNL2* cDNA was cloned into the pDONR221 vector and subcloned into pGWB641 (35Spro, C-EYFP) and pGWB642 (35Spro, N-EYFP) vector, respectively. In *Arabidopsis*, seedlings stably transformed with the βKNL2 fused to EYFP, fluorescent signals were detected at centromeres and in the nucleoplasm of the root tip nuclei (fig. 5*A*–*C*). An immunostaining experiment with anti-GFP and anti-CENH3 antibodies revealed the colocalization of βKNL2-EYFP with CENH3 at centromeres (fig. 5*B*). Live cell imaging of mitotic cells showed that βKNL2 is present at centromeres during interphase, almost not detectable shortly prior to mitosis, but appears again during the M phase (fig. 5*C*). In contrast, αKNL2 was not detectable during prophase, metaphase, and early anaphase in *Arabidopsis* root tip cells (fig. 5*D*; Lermontova et al. 2013).

In all Selected Meristematic Tissues, the Expression Level of *βKNL2* is Higher than that of *αKNL2*

To investigate the expression profiles of the *KNL2* genes in different tissues and developmental stages and to compare them with *CENH3* and *CENP-C*, we downloaded the available RNA-seq data in *Arabidopsis* from a public database (Klepikova et al. 2016) and additionally performed expression analysis using the eFP genome browser. In the eFP genome browser analysis, *βKNL2* was excluded from the analysis due to the mis-annotation and consequent lack of correct gene expression data, while we used the correct *βKNL2* annotation for our RNA-seq data analysis. The expression value of selected genes was normalized to the reference gene *MONENSIN SENSITIVITY1* (*MON1*; *At2g28390*) which shows stable transcription during plant development (Czechowski et al. 2005). The data showed that the *KNL2*, *CENH3*, and *CENP-C* genes have high transcriptional activity in tissues enriched for meristematically active cells (fig. 6, supplementary fig. S7, Supplementary Material online), indicating the involvement of these genes in cell division processes. In contrast, a low expression level of the selected genes was observed in the rosette and senescent leaves (supplementary fig. S7, Supplementary Material online). In general, the *CENP-C* and *CENH3* genes show higher expression than *KNL2*. Interestingly, the *βKNL2* has higher expression level than *αKNL2* in nearly all tissues.

βKNL2 Knockout Resulted in an Abnormal Seed Development and Semilethal Mutant Phenotype

To characterize and understand the *βKNL2* function, two T-DNA insertion lines SALK_135778 and SALK_091054 were identified and defined as *βknl2-1* and *βknl2-2*, respectively (fig. 7*A*). Both T-DNA insertions are present in the single exon of *βKNL2*, 270 and 335 nucleotides downstream from the transcription start. Thus, in *βknl2-1*, the T-DNA insertion is located upstream and in *βknl2-2* directly in the region encoding the SANTA domain (fig. 7*A*). Polymerase chain reaction (PCR)-based genotyping of soil-grown plants revealed no homozygous

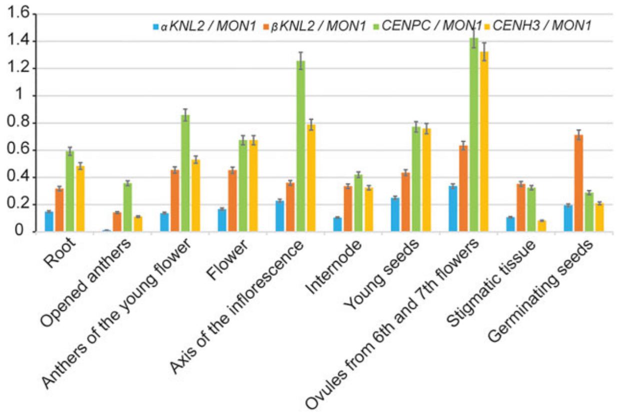


Fig. 6. The *CENH3*, *CENP-C*, and *KNL2* gene expression profiles in *Arabidopsis*. Column charts showing different expression levels of the *CENH3*, *CENP-C*, and *KNL2* genes in tissues enriched for dividing cells. The relative fragments per kilobase of exon per million mapped fragments (RPKM) values of *CENH3*, *CENP-C*, and *KNL2* were normalized to the reference gene *MON1* (*At2g28390*) in RNA-seq data sets. The corresponding gene id numbers are: *CENH3* (*At1g01370*), *CENP-C* (*At1g15660*), *αKNL2* (*At5g02520*), and *βKNL2* (*At1g58210*).

mutant lines in either mutant population obtained from the ABRC seed stock ($n = 26$ and $n = 38$, respectively) or in the next generation ($n = 195$ and $n = 220$, respectively). This suggested that the *βKNL2* knockout might be lethal.

Therefore, siliques of both mutants were tested for the seed phenotype. Heterozygous *βknl2* mutant lines show 11 ± 1% (supplementary fig. S8, Supplementary Material online) of abnormal seeds ($P \leq 0.01$), which look larger and whitish with glossy surface compared with normal green seeds (fig. 7*B*), whereas in the case of wild-type (WT) plants no such seeds were found. However, unlike *βknl2-2*, the *βknl2-1* mutant exhibited an ovule abortion phenotype (supplementary fig. S9, Supplementary Material online). The SALK_135778 (*βknl2-1*) line carries two additional T-DNA insertions in the *AT1G76850* and *AT3G13920* genes according to the ABRC database (https://abrc.osu.edu/stocks/618439). Furthermore, these two genes affect ovule development and pollen acceptance. The corresponding mutations cause an ovule lethal phenotype (Bush et al. 2015; Safavian et al. 2015). Therefore, we speculated that the ovule lethality found in *βknl2-1* might be due to these off-target mutations. Using primers specific to these additional T-DNA insertions, we selected clean *βknl2-1* plants carrying single T-DNA. Indeed, resulting *βknl2-1* lines did not show the aborted ovule phenotype and were selected for further analysis (fig. 7*B*). To assess whether the heterozygous or homozygous state of mutation causes the abnormal seed phenotype and maternal or paternal effects during embryogenesis, reciprocal crosses between WT and heterozygous *βknl2-1* and *βknl2-2* mutants were performed. All these crosses produced <3% of abnormal seeds (fig. 7*C,D* and supplementary table S2, Supplementary Material online) which is similar to the frequency observed in WT self-pollinated siliques. These findings indicate that the appearance of abnormal seeds in the siliques of heterozygous mutants is not the result of defective female gamete formation, but is rather due to defects during postzygotic development. The fact that the abnormal seeds were increased only in self-pollinated heterozygous mutants (fig. 7*C,D*, supplementary table S2, Supplementary Material online), suggests the recessive nature of this phenotype.

As mentioned above, homozygous *βknl2* mutants cannot be selected among the progeny population of heterozygous lines grown on soil. Therefore, we tested whether the abnormal seeds, possibly homozygous for *βknl2* mutations, could germinate and survive under *in vitro* conditions, where seeds and seedlings would be protected from the negative effects of environmental conditions and where the risk that homozygous seedlings would be overgrown by a population of heterozygous plants and WT plants would be minimized.

For both mutants, we found abnormal seedlings, with reduced growth rate and root development (fig. 7*E*). According to the genotyping results, abnormal seedlings represented homozygous mutants, which occur at a frequency of 2–6% of the total number of sown seeds. Unfortunately, our repeated attempts to transfer homozygous seedlings into the soil resulted in their death (fig. 7*F*). At the same time, heterozygous mutant seedlings were not distinguishable from the WT ones (fig. 7*E*). In heterozygous self- or manually pollinated mutants containing single T-DNA insertions, the siliques show <25% of abnormal seeds that does not correspond to the Mendelian monohybrid phenotypic ratio (fig. 7*C*). We hypothesized that this might be due to inaccuracy in the visual phenotyping of immature seeds. Therefore, as the next step, the dry-seed phenotype was analyzed in single siliques (fig. 7*G*–*J*). The heterozygous mutants in addition to normal seeds contain small, dark-colored, and shriveled ones (fig. 7*H*–*I*) in contrast to the WT (fig. 7*G*) with uniform seed size and color.

We observed that the abnormal dry-seed phenotype is significantly more frequent in the siliques of both heterozygous mutants compared with WT (fig. 7*J*, $P \leq 0.001$) and the frequency is similar to that of the whitish seeds in fresh siliques (supplementary fig. S8, Supplementary Material online). Thus, it can be assumed that a large part of the whitish seeds with a glossy surface became dark and small or shriveled on drying.

Additionally, we analyzed the germination rate of seeds obtained from single siliques of both heterozygous *βknl2* mutants and WT (fig. 8*A,B*). Compared with WT, mutants showed a significantly decreased germination rate (fig. 8*B*, $P \leq 0.01$) and increased number of abnormal seedlings per single silique (fig. 8*A,C*, $P < 0.01$). To test the Mendelian segregation of phenotype–genotype ratio, we also performed single silique genotyping. In the case of *βknl2-1*, the homozygous mutation represents ~16% per silique and *βknl2-2* ~25% (supplementary table S3, Supplementary Material online). The variation between the two mutants may be due to the different quality of the seeds harvested at two different time points and, as

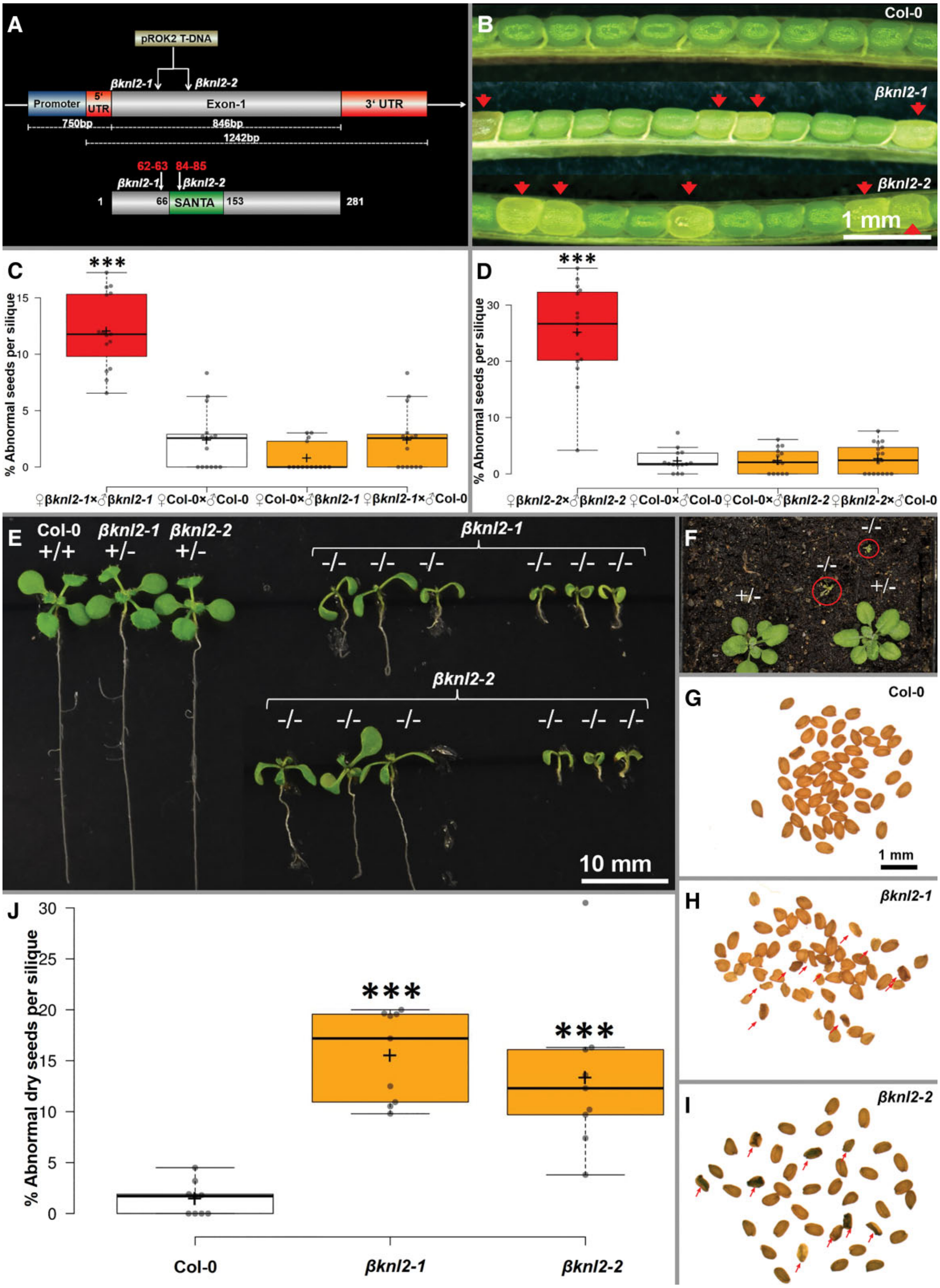


Fig. 7. Identification and primary analysis of *βknl2* mutant. (*A*) Schematic representation of the T-DNA insertion position in the genomic fragment and protein with the position of the SANTA domain. (*B*) Representative siliques with red arrowheads showing abnormal whitish glossy-seed phenotype from heterozygous *βknl2-1* and *βknl2-2* plants. (*C,D*) Boxplots depicting the number of abnormal seeds per silique data from the reciprocal crossing of WT and heterozygous *βknl2-1* and *βknl2-2* (***$P \leq 0.001$). (*E*) Two weeks old *in vitro* germinated seedlings from Col-0, *βknl2-1*, and *βknl2-2* heterozygous (+/−) and homozygous mutants (−/−). (*F*) *βknl2* homozygous (−/−) and heterozygous (+/−) mutants on soil, homozygous mutants turning yellow in the red circle. (*G–I*) Representative dry seeds of Col-0, *βknl2-1*, and *βknl2-2*. Red arrowheads indicate the abnormal seeds. (*J*) Boxplot depicting the significant increase of abnormal dry seeds per silique of heterozygous *βknl2-1* and *βknl2-2* compared with WT as control.

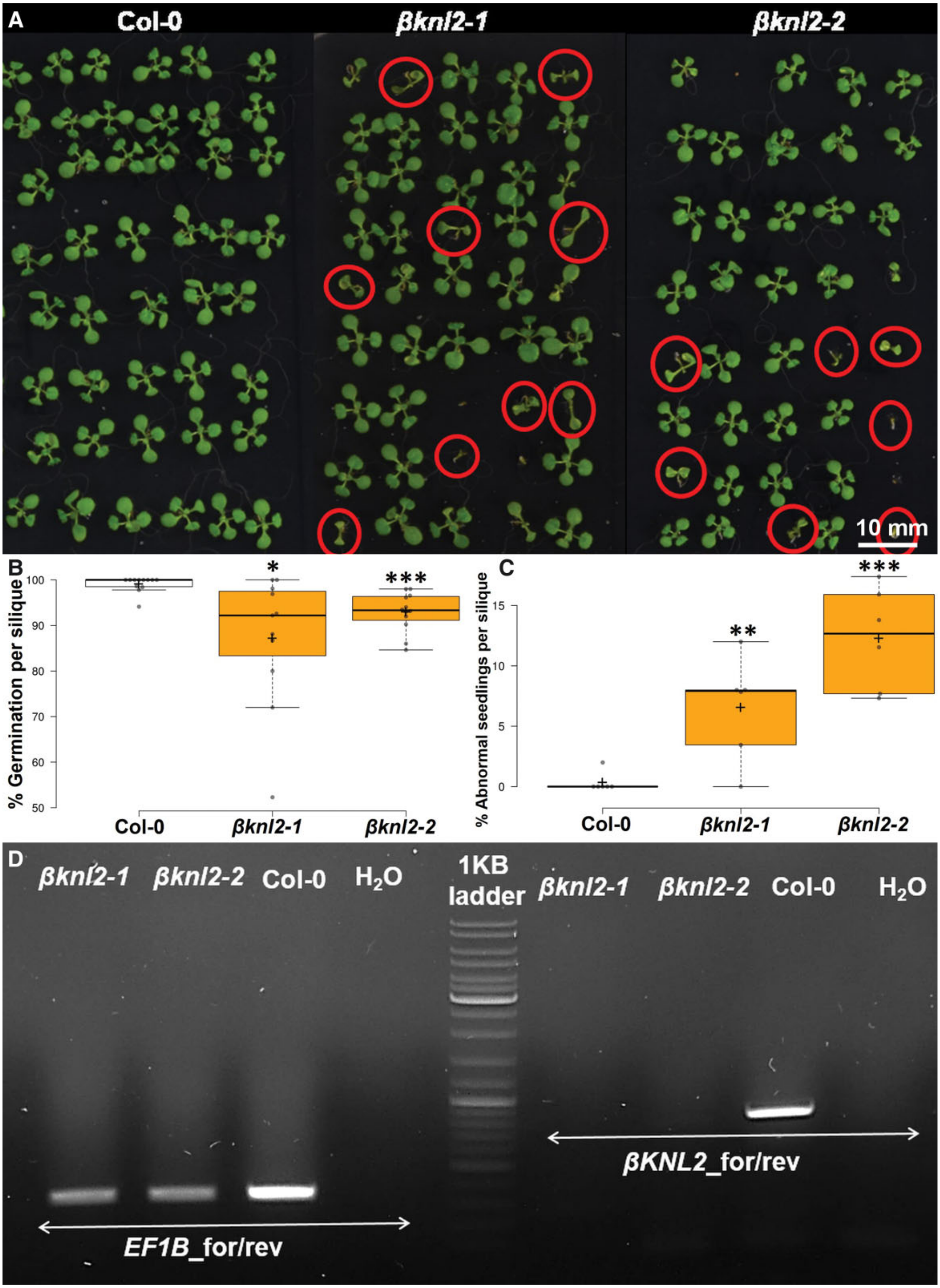


Fig. 8. Analysis of single siliques for seeds germination and presence of abnormal seedlings. (*A*) Two-week-old *in vitro* germinated seeds collected from single siliques of WT as control and heterozygous self-pollinated *βknl2-1* and *βknl2-2* plants. *βknl2* homozygous seedlings are indicated by red circles. Bars: 1 cm. (*B*) Boxplot depicting the significant decrease of germination percentage per silique of heterozygous *βknl2-1* and *βknl2-2* compared with WT as control ($*P \leq 0.05$, $***P \leq 0.001$). (*C*) Boxplot depicting the significant increase of abnormal seedlings (red color circled seedlings in (*A*) germinated from single silique seeds of heterozygous *βknl2-1* and *βknl2-2* compared with WT as control ($**P \leq 0.01$), $***P \leq 0.001$). (*D*) RT-PCR amplification of *βKNL2* from *βknl2-1* and *βknl2-2* homozygous null mutants and WT as the positive control with *βKNL2* (EMB1674) gene-specific primers and *EF1B* primers as housekeeping gene.

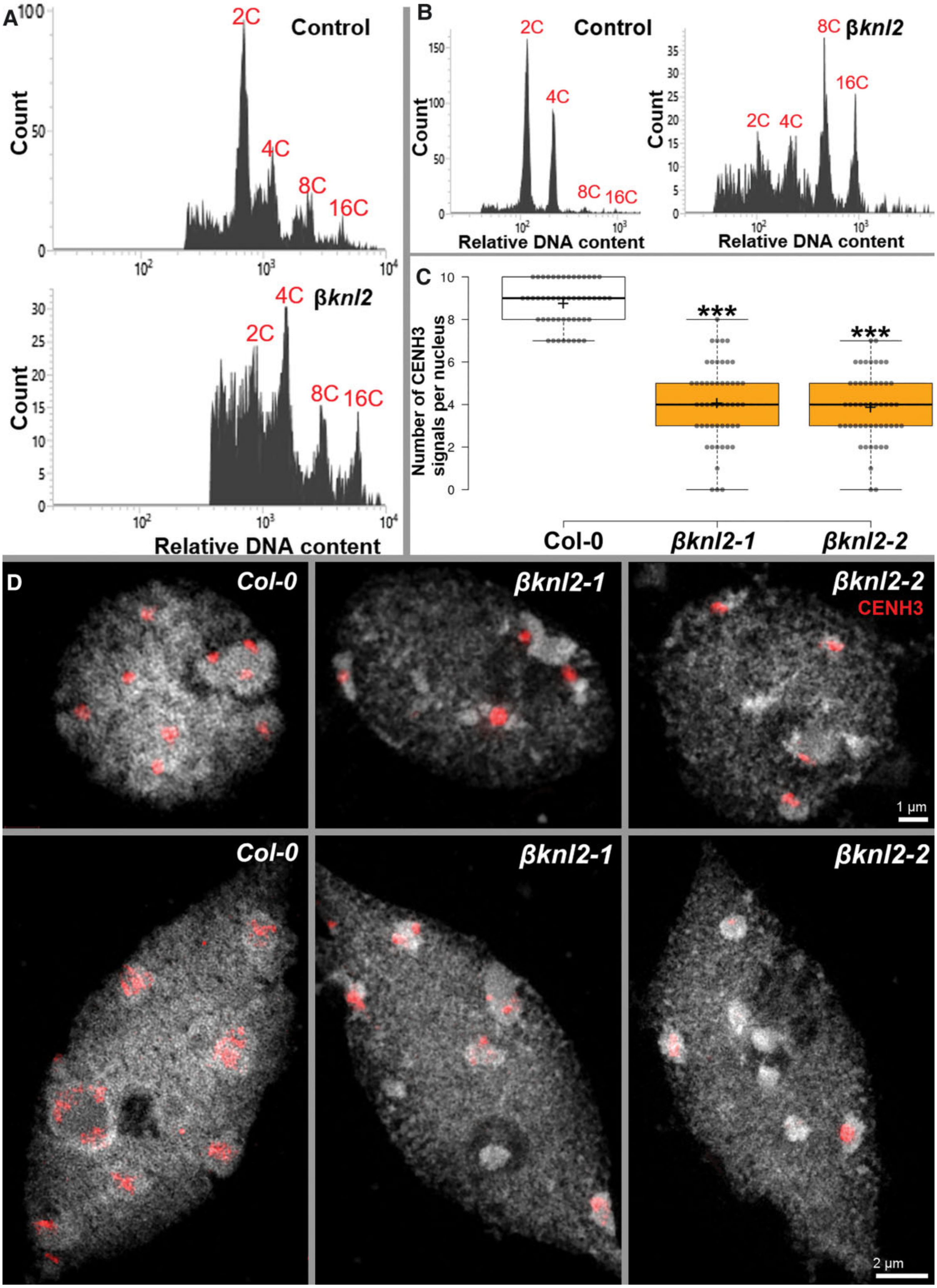


Fig. 9. Reduced CENH3 levels in *βknl2* null mutants leading to endoreduplication. (*A*) Representative ploidy analysis histogram of normal (green) seeds of heterozygous *βknl2* mutants and WT as control (upper panel) and white abnormal seeds from *βknl2* heterozygous mutants (lower panel). (*B*) Representative ploidy analysis histogram of WT seedlings as control (left panel) and abnormal seedlings of *βknl2* null mutants (right panel). (*C*) Boxplot showing a significant decrease in the number of centromeric CENH3 signals in *βknl2-1* and *βknl2-2* compared with WT as a control (***$P \leq 0.001$). (*D*) Super-resolution microscopy images showing nuclei of WT and *βknl2* null mutants immune-stained with anti-CENH3 antibodies in meristematic cells (top) and differentiated cells (bottom).

a result, the lower germination of the homozygous lines of one of the mutants.

To test whether abnormal seedlings (reduced seedling size and reduced root length) of both *βknl2* mutants possess the *βKNL2* transcripts, the reverse transcription-PCR (RT-PCR) analysis with gene-specific primers for *βKNL2* was performed on RNA isolated from three to five seedlings pooled together. The results showed an absence of full-length *βKNL2* transcript in both mutant lines *βknl2-1* and *βknl2-2*, suggesting that homozygous seedlings for further analysis can be selected based on their abnormal phenotype without additional genotyping (fig. 8*D*).

Arabidopsis βKNL2 is Required for Proper CENH3 Loading and Correct Somatic Cell Division

We showed that βKNL2 colocalizes at centromeres with CENH3 (fig. 5*B*) and has a localization pattern similar to that of αKNL2 (Lermontova et al. 2013). To analyze whether βKNL2, similar to αKNL2, is involved in the regulation of cell divisions and CENH3 loading, we used homozygous seedlings of both mutants for flow cytometry (FC) analysis and nuclei isolation for immunostaining. The seedlings were selected based on their abnormal phenotype. Thus, leaves of abnormal seedlings and additionally abnormal white seeds were checked by FC for ploidy levels. Comparison of the green seeds of heterozygous mutants with WT showed similar histogram profiles with a pronounced 2C embryo peak (fig. 9*A*, top), whereas the white seeds showed a clear shift toward increased endopolyploidy levels with the 4C nuclei being in most cases the dominant population (fig. 9*A*, bottom; supplementary fig. S10, Supplementary Material online). In addition, we noticed a reduced sharpness of the peaks probably due to the occurrence of aneuploid nuclei. In some cases, it was even impossible to identify nuclear peaks (supplementary fig. S10, Supplementary Material online). To analyze ploidy levels of seedlings we chopped a single leaf from six 14 days old seedlings of WT and homozygous *βknl2*. In contrast to WT leaves with distinct peaks of 2C and 4C nuclei, in mutant leaves high ploidy nuclei such as 8C and 16C were predominant (fig. 9*B*, supplementary fig. S11, Supplementary Material online).

To find whether the *βKNL2* knockout results in reduced loading of CENH3 at centromeres, similar to αKNL2 deregulation, we performed an immunostaining experiment with anti-CENH3 antibodies on nuclei isolated from 14-day-old seedlings of WT and *βknl2* mutants. In *A. thaliana* roots and leaves, there are predominantly two forms of nuclei (flattened sphere and spindle) occurring (Pecinka et al. 2004). Root meristems contain mainly spherical nuclei (fig. 5*A*), while in the elongated differentiated regions spindle-shaped nuclei appear. These differently shaped nuclei were included in the immunostaining experiment. We found that compared with WT, the mutant nuclei contain less CENH3 signals independent of nucleus shape. The CENH3 signals were counted in 50 round-shaped WT, *βknl2-1* and *βknl2-2* nuclei, respectively. In contrast to WT with eight to ten signals, both mutants showed on average only four signals (fig. 9*C* and supplementary fig. S12, Supplementary Material online). We performed the Student's *t*-test and found that the mutants have significantly lower number of CENH3 signals compared with WT (fig. 9*C*, $n \leq 6$, $P < 0.001$). Furthermore, Mean Fluorescence Intensities were calculated to quantify the centromeric CENH3 levels. Compared with WT, the signal intensities were reduced to 68.98% ($P < 0.001$) in *βknl2-1*, and to 79.47% ($P < 0.01$) in *βknl2-2*, respectively (supplementary fig. S13, Supplementary Material online). In spindle-shaped nuclei, the CENH3 immunosignals on chromocenters were mostly dispersed in the WT and both *βknl2* mutants, whereas in the mutants some chromocenters were completely free of signals. The observed dispersion of CENH3 signals in spindle-shaped nuclei with increased ploidy levels is in agreement with our previous observations (Lermontova et al. 2006). To analyze the chromatin ultrastructure in more detail, representative nuclei from the same slides were captured by spatial structured illumination super-resolution microscopy (3D-SIM; fig. 9*D*). We observed that in nuclei with reduced CENH3 levels the chromatin remains normal as in WT suggesting that intact non-degraded nuclei were selected for the analysis. In summary, our data suggest that the reduced CENH3 amount in the homozygous *βknl2-1&2* mutants lead to the inhibition of mitosis and switching of cells to endocycles.

Discussion

Duplication of *KNL2*

Most metazoan genomes have only one *KNL2* gene with the SANTA domain, except for the allotetraploid *Xenopus laevis*, where two *KNL2* genes were identified; both with identical CENPC-k motifs, nearly identical SANTA and Myb (SANT) domains, and 74% sequence similarity (Moree et al. 2011; French et al. 2017). In contrast, two genes containing the SANTA domain were identified in water ferns, eudicots, and grasses, whereas only one *KNL2* copy was found in bryophytes and gymnosperms (fig. 2). Though Brassicaceae species experienced multiple whole genome duplication (WGD) events such as the *At-α* and *At-β* WGDs (Edger et al. 2018), most species exhibit two *KNL2* gene copies, *αKNL2* and *βKNL2*, except for a few neopolyploid species which have experienced an extra recent WGD event(s).

We found strong conservation of the SANTA domain of KNL2, notably in the Vx**L**xD**W** motif at the N-terminus and the GFxxxxxxx**F**xxG**F**PxxW motif at the C-terminus (fig. 3*A*), where the bolded residues impaired CENP-C binding when mutated in *Xenopus* M18BP (French and Straight 2019), suggesting that plant KNL2s may also bind CENP-C through the SANTA domain. In addition, analysis of αKNL2 and βKNL2 protein sequences identified numerous paralog-specific motifs, suggesting that the paralogs might be subfunctionalized. A study in *Drosophila* has shown that Cid (CENH3) paralogs evolved new motifs following Cid duplication (Kursel and Malik 2017). Loss of ancestral motifs in *Drosophila* Cids was proposed as direct evidence of subfunctionalization (Kursel and Malik 2017; Kursel et al. 2020).

We identified positive selection sites in and near the SANTA domain of KNL2 in the analyzed Brassicaceae species, similar to what has been previously reported for CENH3 (Talbert et al. 2002) and CENP-C (Talbert et al. 2004). Thus, KNL2 might be responding to centromere drive through interaction with rapidly evolving CENH3 and CENH3 chaperone NASP$^{\mathrm{SIM3}}$, which recently was identified in *Arabidopsis* (Le Goff et al. 2020), or with CENP-C. However, the mechanisms of adaptively evolving regions remain to be elucidated.

Partial or Complete Loss of the CENPC-k Motif in KNL2 in Different Clades of Plants

The CENPC-k motif is found in KNL2 of diverse eukaryotes including non-mammalian vertebrates, many invertebrates, chytrid fungi, cryptomonads, and plants (Kral 2016; Sandmann et al. 2017). In eudicots the conserved CENPC-k motif is present in the αKNL2 clade, but is absent from βKNL2. Similarly, in most grass species the CENPC-k motif is conserved in γKNL2 clade, while δKNL2 clade does not have the motif. However, we found a RRLRSG**K**V/I motif in the δKNL2 clade possibly related to the beginning of the CENPC-k motif (KRSRSG**R**V/LLVSPLEFW; supplementary file S6, Supplementary Material online). We showed previously that the substitution of the bolded seventh Arg in the CENPC-k motif (above) by Ala abolishes centromere targeting of αKNL2 (Sandmann et al. 2017). In the truncated putative CENPC-k motif, Lys is present instead of Arg. Since these two amino acids have similar features, Lys might be required for the targeting of δKNL2 to centromeres. However, the truncated putative CENPC-k motif does not include the Trp which similar to Arg, is also needed for the targeting of αKNL2 to centromeres (Sandmann et al. 2017). Moreover, it remains to be elucidated whether KNL2 variants with the truncated CENPC-k motif can target CENH3 nucleosomes directly, without an additional interacting partner. Among all grass species with sequenced genomes, maize represents an exception, since it has only one *KNL2* gene which belongs to the δ*KNL2* clade with the truncated CENPC-k and has no γKNL2 protein variant with the complete CENPC-k motif. Interestingly, in sorghum, closely related to maize, the γKNL2 protein can be identified (supplementary file S5, Supplementary Material online). On the other hand, for other species, it may be postulated that centromeric targeting of βKNL2 and δKNL2 depends on αKNL2 and γKNL2, respectively, for maize this assumption cannot be applied. This suggests that maize may have evolved a different mechanism for CENH3 deposition compared with other grasses. Notably, δKNL2 retains the hydrophobic residues in the SANTA domain that are important for CENP-C binding in *Xenopus*. Perhaps the mechanism of localization and function of KNL2 in maize relies on CENP-C binding similar to *Xenopus*. Interestingly, two CENP-C proteins were identified in maize (Talbert et al. 2004), in contrast to other species.

The Function of βKNL2 in Plants

Although KNL2 protein homologues have been identified in different organisms as components of the CENH3 loading machinery, they differ considerably in the composition of their functional domains, interacting partners, and localization timing in the mitotic cell cycle. The mammalian M18BP1, composed of the conserved N-terminal (Mis18α-binding) region, SANTA domain, CENP-C-binding domain, SANT (Myb-like) domain and the C-terminus, is lacking the CENPC-k motif. The N-terminal (Mis18α-binding) region and the CENP-C-binding domain are required for centromere targeting (Stellfox et al. 2016). Deletion of the SANTA domain in mammalian and chicken M18BP1/KNL2 does not abolish its centromeric localization (Stellfox et al. 2016; Hori et al. 2017). In contrast, mutation of the SANTA domain in *Xenopus* reduced centromeric localization of M18BP1/KNL2 by 90% (French et al. 2017). Later, the same authors demonstrated that the SANTA domain is required for the interaction of M18BP1/KNL2 with CENP-C during metaphase (French and Straight 2019).

We showed previously that in *Arabidopsis* the centromeric localization of αKNL2 depends on the CENPC-k motif (Sandmann et al. 2017), while it was not abolished in the complete absence of the N-terminal part of KNL2 containing the SANTA domain (Lermontova et al. 2013). The C-terminal half of *Arabidopsis* KNL2 was not only sufficient for its targeting to centromeres, but also the interaction with DNA (Sandmann et al. 2017). In the present study, we demonstrated that βKNL2 colocalizes with CENH3 at centromeres, despite lacking a CENPC-k motif. In general, both variants of *Arabidopsis* KNL2 showed a similar localization pattern during interphase. However, in contrast to αKNL2, βKNL2 can be detected on chromosomes during metaphase and early anaphase (fig. 5*C*,*D*). The centromeric location of βKNL2 suggests that *βKNL2* may partially compensate for the loss of *αKNL2* in the corresponding *Arabidopsis* mutant which showed only reduced, but not completely abolished CENH3 loading which would be lethal (Lermontova et al. 2013). Homozygous T-DNA insertions for βKNL2 resulted in plant death at the seedling stage and probably because of reduced root development. However, it should be considered that in the analyzed *αknl2* mutant, the T-DNA was inserted after the SANTA domain coding region, whereas in the case of *βknl2* mutants, one T-DNA was inserted before and the other directly in the SANTA domain coding region. Therefore, it cannot be excluded that truncated αKNL2 with the full SANTA domain may retain some function in the mutant. As reciprocal crosses of *βknl2* mutants with the WT resulted in normal seed development in both directions, we hypothesized that the *βKNL2* null mutations do not affect gametes or fertilization processes, but rather postzygotic cell divisions. In support of this hypothesis, FC ploidy analysis of young seedlings revealed that in contrast to the WT with distinct 2C and 4C peaks, homozygous mutants showed a shift toward endopolyploidization (fig. 9*B*), potentially a consequence of disrupted cell divisions. Impaired mitotic divisions in mutant seedlings can be explained by the reduced levels of CENH3 on the centromeres of both mutants (supplementary figs. 9D and S13, Supplementary Material online). Thus, our data strongly suggest the involvement of βKNL2 protein in CENH3 loading. The ability of cells in homozygous seedlings to undergo some mitotic divisions can be explained by residual amounts of CENH3 from parental plants, and when CENH3 levels are highly diluted, cells switch from mitotic cycle to endocycles. We observed that the development of homozygous seedlings can be inhibited at different stages (fig. 7*E*).

Taken together, our results suggest that the *KNL2* gene in eudicots underwent an early duplication with the core function of CENH3 deposition to define the centromere region. Due to the lack of the CENPC-k motif in βKNL2, we propose that in *Arabidopsis* βKNL2 might localize to centromeres by binding to CENP-C through the SANTA domain as it was shown for *Xenopus* (French and Straight 2019), or through the conserved N-terminal motif located upstream of the SANTA domain similar to what was previously described in human (Stellfox et al. 2016), or through both of these regions.

Although in the SANTA domain of βKNL2, three putative Aurora kinase phosphorylation sites can be identified, there is only one in αKNL2 (fig. 4*A*). This fact might suggest that both KNL2 variants are involved in the formation of different protein complexes. We also could not rule out the possibility that βKNL2 assembles with a Mis18 complex to ensure centromeric localization and subsequent CENH3 deposition. So far, Mis18α and β proteins have not been identified and characterized in *Arabidopsis*. However, in silico analysis (https://bioinformatics.psb.ugent.be/plaza/) revealed a family of seven genes (*At2G40110*, *AT3G08990*, *AT3G11230*, *AT3G55890*, *AT4G27740*, *AT4G27745*, and *AT5G53940*) encoding proteins with the Yippee-Mis18 domain-specific to Mis18 proteins (Stellfox et al. 2016). Recently, it was demonstrated that the direct binding of *Schizosaccharomyces pombe* Mis18 to nucleosomal DNA is important for the recruitment of *sp*Mis18 and Cnp1 (CENH3) to the centromere in fission yeast (Zhang et al. 2020). In contrast to αKNL2, βKNL2 not only lacks the CENPC-k domain but also the part necessary for interaction with DNA. Thus, an association with Mis18 proteins, with the ability to bind to DNA, is plausible. We also cannot exclude that centromere targeting of βKNL2 depends on αKNL2.

We showed previously that manipulation of αKNL2 can be used for the production of haploids and subsequently of double haploids in *Arabidopsis* (Lermontova 2017; Ahmadli et al. 2022a). Double haploid production helps to accelerate plant breeding as it allows to generate true-breeding lines in one generation instead of the seven to nine generations required for conventional selection (Britt and Kuppu 2016; Kalinowska et al. 2019). Here we demonstrate that *KNL2* genes exist in two variants in eudicots (*α*, *βKNL2*) and monocots (*γ*, *δKNL2*). The conserved gene structure and expression patterns of *α*/*γKNL2* in both eudicots and monocots suggest that *α*/*γKNL2* mutations could be used to develop *in vivo* haploid induction systems in different crop plants. Similarly, the newly identified βKNL2 may become the subject of manipulations to obtain haploids both in *Arabidopsis* and in crops. As homozygous *βknl2* mutants are dying at the seedling stage, we can assume that the heterozygous mutant plants can also induce haploids similar to what was described for the heterozygous *cenh3* mutants of maize and wheat (Lv et al. 2020; Wang et al. 2021).

Materials and Methods

Data Sources and Sequences Retrieval

The KNL2 protein sequences of *A. thaliana* were identified by screening the *Arabidopsis* Information Resource (TAIR10) using the specific gene number. To obtain and annotate *KNL2* members in plants, we downloaded 88 representative species reference genomes or transcriptomes including red and green algae, bryophytes, lycophytes, ferns, gymnosperms, and angiosperms from the Phytozome database (Goodstein et al. 2012; https://phytozome.jgi.doe.gov/), NCBI genome database, Ensembl Plants database, PLAZA database, and other single genome website (supplementary table S1, Supplementary Material online). We used the homology search tool BLASTP to scan the reference proteome with a cutoff *e*-value of 0.01 using whole sequences and conserved domains from *Arabidopsis* αKNL2 as the query. TBLASTN was used as an additional method for failed identification case. Two KNL2 protein sequences from *Colocasia esculenta* and *Phoenix dactylifera* were retrieved from GenBank database. Then, we combined the BLAST results and deleted spliced variants in multiple sequence alignments. The protein data are summarized in supplementary table S1 and file S1, Supplementary Material online.

Alignments and Phylogenetic Analysis

To explore the phylogenetic relationships of the *KNL2* genes in plant lineages, KNL2 protein sequences were aligned using MAFFT software (Yamada et al. 2016) and potentially inaccurate regions of βKNL2 were excluded. Evolutionary relationships among *KNL2* gene family members were determined by using IQ-TREE software (Nguyen et al. 2015) and ML methods based on 1000 bootstrap alignments and single-branch tests. The phylogenetic trees were visualized and modified using the Fig-Tree v1.4.4 software (http://tree.bio.ed.ac.uk/software/figtree/). Sequence logos were generated using WebLogo3 (http://weblogo.berkeley.edu/; Crooks et al. 2004).

Sequence Motif Analysis

The unaligned amino acid sequences of KNL2 were collected to search for additional conserved motifs using MEME suite v5.1.0 (Bailey et al. 2009). Due to misleading annotation of the *βKNL2* gene (Lermontova et al. 2013), we manually removed the KIP1 domain regions in some species. The data set was submitted to the MEME server (http://meme-suite.org/) and the conserved domains and motifs were marked. We used the motif search algorithm MAST (Bailey and Gribskov 1998) to identify motifs.

Plasmid Construction, Plant Transformation, and Cultivation

The entire open reading frame of *βKNL2* (*At1g58210*) was amplified by RT-PCR with RNA isolated from flower buds of *Arabidopsis* WT and cloned into the pDONR221 vector (Invitrogen) via the Gateway BP reaction. From pDONR221 clones, the open reading frame was recombined via Gateway LR reaction (Invitrogen) into the two attR recombination sites of the Gateway-compatible vectors pGWB641and pGWB642 (http://shimane-u.org/nakagawa/gbv.htm), respectively, to study the localization of βKNL2 protein *in vivo*.

Plants of *Arabidopsis* accession Columbia-0 were transformed according to the flower dip method (Clough and Bent 1998). T1 transformants were selected on Murashige and Skoog (MS) medium (Murashige and Skoog 1962) containing 20 mg/l of phosphinotricine. Growth conditions in a cultivation room were 21 °C 8 h light/18 °C 16 h dark or 21 °C 16 h light/18 °C 8 h dark.

Analysis of T-DNA Insertion Mutants

Seeds of T-DNA insertion lines were obtained from the European *Arabidopsis* stock center (http://arabidopsis.info/). To confirm the presence of the T-DNA, and identify heterozygous versus homozygous T-DNA insertions, we performed PCR with pairs of gene-specific primers flanking the putative positions of T-DNA (supplementary table S4, Supplementary Material online) and with a pair of gene-specific and T-DNA end-specific primers (LBb3.1, supplementary table S4, Supplementary Material online). DNA isolation was performed as described in Edwards et al. (1991).

For the germination and segregation experiments, seeds from individual siliques were germinated *in vitro* on an MS medium as described above.

Flow Cytometry

For the analysis of (endopoly)ploidy of immature seeds, white and green seeds were selected from the same silique of the heterozygous mutant and compared with the green seeds of the WT. For the analysis of (endopoly)ploidy levels in seedlings, one leaf from 2-week-old heterozygous mutant and WT seedlings was used. Seeds and leaf tissue were chopped with a razor blade in 300 μl of nuclei extraction buffer (CyStain UV Ploidy; Sysmex-Partec). The resulting nuclei suspension was filtered through a 50 μm disposable CellTrics filter (Sysmex-Partec), incubated for 10 min on ice and measured on BD Influx cell sorter (BD Biosciences).

Immunostaining and Microscopy Analysis of Fluorescent Signals

For analysis of CENH3 loading in homozygous mutants and WT, 2-week-old seedlings were used. Slides were prepared using a cytospin and used for immunostaining as it was described by Ahmadli et al. (2022b). To determine the colocalization of βKNL2-EYFP protein with CENH3, immunostaining of nuclei/chromosomes with anti-CENH3 and anti-GFP antibodies and microscopic analysis of fluorescent signals were performed as previously described (Lermontova et al. 2013).

For time-lapse microscopy, seedlings of transformants were grown in cover slip chambers (Nalge Nunc International) for 7–10 days and analyzed with an LSM 510 META confocal laser scanning microscope (Carl Zeiss GmbH).

To investigate the interphase nucleus and centromeric chromatin ultrastructures at an optical lateral resolution of ∼100 nm (super-resolution achieved with a 405-nm laser excitation), we applied spatial structural illumination microscopy (3D-SIM) using a 63/1.40 objective of an Elyra PS.1 super-resolution microscope system (Carl Zeiss GmbH; Weisshart et al. 2016; Kubalova et al. 2021) DAPI (whole chromatin) and rhodamine (CENH3 signals) were excited by 405 and 561 nm lasers, respectively.

Expression Profile Analyses

The *Arabidopsis* genome assembly and gene annotation were downloaded from Araport11 (https://bar.utoronto.ca/thalemine/dataCategories.do) with integrative re-annotation (Cheng et al. 2017). The *KNL2* gene models were manually re-examined. The *Arabidopsis* RNA-seq data were downloaded from previous studies (Klepikova et al. 2016). RNA-seq data were selected from ten tissue types in *Arabidopsis*, including germinating seeds, stigmatic tissue, ovules from sixth and seventh flowers, young seeds, internode, the axis of the inflorescence, flower, anthers of the young flower, opened anthers, and root (NCBI SRA: SRR3581356, SRR3581684, SRR3581691, SRR3581693, SRR3581704, SRR3581705, SRR3581719, SRR3581727, SRR3581728, SRR3581732). Transcriptome analysis utilized a standard TopHat-Cufflinks pipeline with minor modification (Trapnell et al. 2012). Transcription levels were normalized to *MON1* and expressed in reads per kilobase of exon model per million mapped reads (RPKM). Expression levels of *CENH3*, *CENP-C*, and *KNL2* normalized to *MON1* in different tissues from microarray experiments were obtained from the *Arabidopsis* eFP Browser website (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi). The corresponding gene IDs are: *CENP-C* (*At1g15660*), *αKNL2* (*At5g02520*), *βKNL2* (*At1g58210*), and *CENH3* (*At1g01370*).

Positive Selection Analyses

PAML 4.8 software (Yang 2007) was used to test for positive selection on KNL2 homologs from Brassicaceae species. The *KNL2* gene alignments and gene trees were used as input into the CodeML of PAML. Alignments were manually refined as described in phylogenetic analysis. To determine whether αKNL2 and βKNL2 homologs evolve under positive selection, random-site models were selected. Random-site models allow ω to vary among sites but not across lineages. We compared two models that do not allow ω to exceed 1 (M1 and M7), and that allow $\omega > 1$ (M2 and M8). Positively selected sites were classified as those sites with a Bayes empirical Bayes posterior probability >95%.

Statistical Data Analysis

All statistical analyses were performed in Microsoft Excel using FTEST and two-tailed TTEST functions (supplementary file S8, Supplementary Material online). Box plots were generated using the online tool BoxPlotR (http://shiny.chemgrid.org/boxplotr/, Team RC, 2013).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank Xinyi Guo for help with the bioinformatic data analysis and Heike Kuhlmann and Pascal Jaroschinsky for technical assistance. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic. I.L. was supported by European Regional Development Fund-Project "MSCAfellowMUNI" (No. CZ.02.2.69/0.0/0.0/17_050/0008496), R.Y. by WIPANO Wissens und Technologietransfer durch Patente und Normen project grant (03THWST001) and breeding company Enza Zaden, UA by the Deutsche Forschungsgemeinschaft (LE2299/3-1). A.P. was supported by the Purkyně fellowship from the Czech Academy of Sciences, ERDF project "Plants as a tool for sustainable global development" (No. CZ.02.1.01/0.0/0.0/16_019/0000827) and with F.Y by the Czech Science Foundation grant 22-00871S. The authors acknowledge networking support from COST Action (STSM–CA16212-47992, http://www.cost.eu/COST_Actions/ca/CA16212). P.T. was supported by Howard Hughes Medical Institute. The publication of this article was funded by the German Research Foundation (DFG)—HE 9114/1-1.

Author Contributions

S.Z. and R.Y., contributed equally to this work. I.L., S.Z., R.Y., F.Y., P.T., A.P., and M.L. conceived the study and designed the experiments. S.Z, R.Y., F.Y., U.A., J.F., and V.S. performed the experiments. S.Z., R.Y., I.L., M.L., and P.T. wrote the manuscript. All authors read and approved the final manuscript.

Data Availability

All data used in this manuscript are available as supplementary files to this manuscript.

References

Ahmadli U, Kalidass M, Khaitova LC, Fuchs J, Cuacos M, Demidov D, Zuo S, Pecinkova J, Mascher M, Heckmann S, *et al.* 2022a. High temperature increases centromere-mediated genome elimination frequency in Arabidopsis deficient in cenH3 or its assembly factor KNL2. *BioRxive*.

Ahmadli U, Sandmann M, Fuchs J, Lermontova I. 2022b. Immunolabeling of nuclei/chromosomes in *Arabidopsis thaliana*. In: Caillaud MC, editor. *Plant cell division. Methods in molecular biology*, vol. 2382. New York (NY): Humana. p. 19–28.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren JY, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**(2):W202–W208.

Bailey TL, Gribskov M. 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics.* **14**(1):48–54.

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, *et al.* 2011. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science.* **332**(6032):960–963.

Barra V, Fachinetti D. 2018. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nat Commun.* **9**(1):4340.

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, *et al.* 2012. The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biol.* **13**(5):R39.

Britt AB, Kuppu S. 2016. Cenh3: an emerging player in haploid induction technology. *Front Plant Sci.* **7357**.

Bush MS, Crowe N, Zheng T, Doonan JH. 2015. The RNA helicase, eIF4A-1, is required for ovule development and cell size homeostasis in Arabidopsis. *Plant J.* **84**(5):989–1004.

Cheeseman IM, Desai A. 2008. Molecular architecture of the kinetochore-microtubule interface. *Nat Rev Mol Cell Biol.* **9**(1):33–46.

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**(4):789–804.

Clough SJ, Bent AF. 1998. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**(6):735–743.

Collen J, Porcel B, Carre W, Ball SG, Chaparro C, Tonon T, Barbeyron T, Michel G, Noel B, Valentin K, *et al.* 2013. Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida. *Proc Natl Acad Sci U S A.* **110**(13):5247–5252.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* **14**(6):1188–1190.

Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR. 2005. Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol.* **139**(1):5–17.

Edger PP, Hall JC, Harkess A, Tang M, Coombs J, Mohammadin S, Schranz ME, Xiong Z, Leebens-Mack J, Meyers BC, *et al.* 2018. Brassicales phylogeny inferred from 72 plastid genes: a reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses. *Am J Bot.* **105**(3):463–469.

Edwards K, Johnstone C, Thompson C. 1991. A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res.* **19**(6):1349.

Fachinetti D, Folco HD, Nechemia-Arbely Y, Valente LP, Nguyen K, Wong AJ, Zhu Q, Holland AJ, Desai A, Jansen LE, *et al.* 2013. A two-step mechanism for epigenetic specification of centromere identity and function. *Nat Cell Biol.* **15**(9):1056–1066.

French BT, Straight AF. 2019. CDK phosphorylation of *Xenopus laevis* M18BP1 promotes its metaphase centromere localization. *Embo J.* **38**(4):e100093.

French BT, Westhorpe FG, Limouse C, Straight AF. 2017. *Xenopus laevis* M18BP1 directly binds existing CENP-A nucleosomes to promote centromeric chromatin assembly. *Dev Cell.* **42**(2):190–199.

Friis EM, Pedersen KR, Crane PR. 2016. The emergence of core eudicots: new floral evidence from the earliest Late Cretaceous. *Proc R Soc B.* **283**(1845):20161325.

Fujita Y, Hayashi T, Kiyomitsu T, Toyoda Y, Kokubu A, Obuse C, Yanagida M. 2007. Priming of centromere for CENP-A recruitment by human hMis18 alpha, hMis18 beta, and M18BP1. *Dev Cell.* **12**(1):17–30.

Goodstein DM, Shu SQ, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, *et al.* 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**(D1):D1178–D1186.

Hara M, Fukagawa T. 2018. Kinetochore assembly and disassembly during mitotic entry and exit. *Curr Opin Cell Biol.* **52**:73–81.

Hori T, Shang WH, Hara M, Ariyoshi M, Arimura Y, Fujita R, Kurumizaka H, Fukagawa T. 2017. Association of M18BP1/KNL2 with CENP-A nucleosome is essential for centromere formation in non-mammalian vertebrates. *Dev Cell.* **42**(2):181–189.

Kalinowska K, Chamas S, Unkel K, Demidov D, Lermontova I, Dresselhaus T, Kumlehn J, Dunemann F, Houben A. 2019. State-of-the-art and novel developments of in vivo haploid technologies. *Theor Appl Genet.* **132**(3):593–605.

Kato H, Jiang JS, Zhou BR, Rozendaal M, Feng HQ, Ghirlando R, Xiao TS, Straight AF, Bai YW. 2013. A conserved mechanism for centromeric nucleosome recognition by centromere protein CENP-C. *Science.* **340**(6136):1110–1113.

Kim IS, Lee M, Park KC, Jeon Y, Park JH, Hwang EJ, Jeon TI, Ko S, Lee H, Baek SH, *et al.* 2012. Roles of Mis18alpha in epigenetic regulation of centromeric chromatin and CENP-A loading. *Mol Cell.* **46**(3):260–273.

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. 2016. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J.* **88**(6):1058–1070.

Kral L. 2016. Possible identification of CENP-C in fish and the presence of the CENP-C motif in M18BP1 of vertebrates. *F1000Res.* **4**:474.

Kubalova I, Nemeckova A, Weisshart K, Hribova E, Schubert V. 2021. Comparing super-resolution microscopy techniques to analyze chromosomes. *Int J Mol Sci.* **22**(4):1903.

Kursel LE, Malik HS. 2017. Recurrent gene duplication leads to diverse repertoires of centromeric histones in *Drosophila* species. *Mol Biol Evol.* **34**(6):1445–1462.

Kursel LE, Welsh FC, Malik HS. 2020. Ancient coretention of paralogs of *Cid* centromeric histones and *Cal1* chaperones in Mosquito species. *Mol Biol Evol.* **37**(7):1949–1963.

Le Goff S, Keceli BN, Jerabkova H, Heckmann S, Rutten T, Cotterell S, Schubert V, Roitinger E, Mechtler K, Franklin FCH, *et al.* 2020. The H3 histone chaperone NASP$^{\mathrm{SIM3}}$ escorts CenH3 in Arabidopsis. *Plant J.* **101**(1):71–86.

Lermontova I. 2017. Generation of haploid plants based on KNL 2. Available from: https://patents.google.com/patent/WO2017067714A1/en

Lermontova I, Kuhlmann M, Friedel S, Rutten T, Heckmann S, Sandmann M, Demidov D, Schubert V, Schubert I. 2013. *Arabidopsis* KINETOCHORE NULL2 is an upstream component for centromeric histone H3 variant cenH3 deposition at centromeres. *Plant Cell.* **25**(9):3389–3404.

Lermontova I, Schubert V, Fuchs J, Klatte S, Macas J, Schubert I. 2006. Loading of Arabidopsis centromeric histone CENH3 occurs mainly during G2 and requires the presence of the histone fold domain. *Plant Cell.* **18**(10):2443–2451.

Lv J, Yu K, Wei J, Gui H, Liu C, Liang D, Wang Y, Zhou H, Carlin R, Rich R, *et al.* 2020. Generation of paternal haploids in wheat by genome editing of the centromeric histone CENH3. *Nat Biotechnol.* **38**(12):1397–1401.

McKinley KL, Cheeseman IM. 2016. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol.* **17**(1):16–29.

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal- Drouard L, *et al.* 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**(5848):245–250.

Moree B, Meyer CB, Fuller CJ, Straight AF. 2011. CENP-C recruits M18BP1 to centromeres to promote CENP-A chromatin assembly. *J Cell Biol.* **194**(6):855–871.

Murashige T, Skoog F. 1962. A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol Plant.* **15**(3):473–497.

Musacchio A, Desai A. 2017. A molecular view of kinetochore assembly and function. *Biology* **6**(1):5.

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**(1):268–274.

Pecinka A, Schubert V, Meister A, Kreth G, Klatte M, Lysak MA, Fuchs J, Schubert I. 2004. Chromosome territory arrangement and homologous pairing in nuclei of *Arabidopsis thaliana* are predominantly random except for NOR-bearing chromosomes. *Chromosoma* **113**(5):258–269.

Qi XP, Kuo LY, Guo CC, Li H, Li ZY, Qi J, Wang LB, Hu Y, Xiang JY, Zhang CF, *et al.* 2018. A well-resolved fern nuclear phylogeny reveals the evolution history of numerous transcription factor families. *Mol Phylogenet Evol.* **127**:961–977.

Safavian D, Zayed Y, Indriolo E, Chapman L, Ahmed A, Goring DR. 2015. RNA silencing of exocyst genes in the stigma impairs the acceptance of compatible pollen in Arabidopsis. *Plant Physiol.* **169**(4):2526–2538.

Sandmann M, Talbert P, Demidov D, Kuhlmann M, Rutten T, Conrad U, Lermontova I. 2017. Targeting of Arabidopsis KNL2 to centromeres depends on the conserved CENPC-k motif in its C terminus. *Plant Cell.* **29**(1):144–155.

Stellfox ME, Nardi IK, Knippler CM, Foltz DR. 2016. Differential binding partners of the Mis18 α/β YIPPEE domains regulate Mis18 complex recruitment to centromeres. *Cell Rep.* **15**(10):2127–2135.

Sugimoto K, Yata H, Muro Y, Himeno M. 1994. Human centromere protein-C (CENP-C) is a DNA-binding protein which possesses a novel DNA-binding motif. *J Biochem.* **116**(4):877–881.

Talbert PB, Bryson TD, Henikoff S. 2004. Adaptive evolution of centromere proteins in plants and animals. *J Biol.* **3**(4):18.

Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S. 2002. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. *Plant Cell.* **14**(5):1053–1066.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* **7**(3):562–578.

Wang N, Gent JI, Dawe RK. 2021. Haploid induction by a maize cenh3 null mutant. *Sci Adv.* **7**(4):abe2299.

Weisshart K, Fuchs J, Schubert V. 2016. Structured Illumination Microscopy (SIM) and Photoactivated Localization Microscopy (PALM) to analyze the abundance and distribution of RNA polymerase II molecules on flow-sorted *Arabidopsis* nuclei. *Bio Protocol.* **6**(3):e1725.

Wu Y, You HL, Li XQ. 2018. Dinosaur-associated Poaceae epidermis and phytoliths from the early cretaceous of China. *Natl Sci Rev.* **5**(5):721–727.

Yamada KD, Tomii K, Katoh K. 2016. Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics* **32**(21): 3246–3251.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* **24**(8):1586–1591.

Zhang D, Martyniuk CJ, Trudeau VL. 2006. SANTA domain: a novel conserved protein module in Eukaryota with potential involvement in chromatin regulation. *Bioinformatics* **22**(20): 2459–2462.

Zhang M, Zheng F, Xiong YJ, Shao C, Wang CL, Wu MH, Niu XJ, Dong FF, Zhang X, Fu CH, *et al.* 2020. Centromere targeting of Mis18 requires the interaction with DNA and H2A-H2B in fission yeast. *Cell Mol Life Sci.* **78**(1):373–384.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.24.485459; this version posted April 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

High temperature increases centromere-mediated genome elimination frequency in Arabidopsis deficient in cenH3 or its assembly factor KNL2

Ulkar Ahmadli[1#], Manikandan Kalidass[1#], Lucie Crhak Khaitova[2], Joerg Fuchs[1], Maria Cuacos[1], Dmitri Demidov[1], Sheng Zuo[2], Jana Pecinkova[2], Martin Mascher[1], Mathieu Ingouff[3], Stefan Heckmann[1], Andreas Houben[1], Karel Riha[2] and Inna Lermontova[1]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstrasse 3, D-06466 Seeland, Germany

[2]Central European Institute of Technology (CEITEC) and National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic

[3]CIRAD, DIADE, IRD, University of Montpellier, 34393 Montpellier, France

***Author for correspondence:** lermonto@ipk-gatersleben.de

Phone: ++49/39482 5570

Fax: ++49/39482 5137

# - contributed equally to this work

Running title: High temperature increases haploid induction efficiency

Abstract

Double haploid production is the most effective way of creating true-breeding lines in a single generation. In *Arabidopsis*, haploid induction via mutation of the centromere-specific histone H3 (cenH3) has been shown when outcrossed to wild-type. Here we report that a mutant of the cenH3 assembly factor KNL2 can be used as a haploid inducer. We elucidated that short temperature stress of the *knl2* mutant increased the efficiency of haploid induction from 1 to 10%. Moreover, we have demonstrated that a point mutation in the CENPC-k motif of KNL2 is sufficient to generate haploid inducing lines, suggesting that haploid inducing lines in crops can be identified in a naturally occurring or chemically induced mutant population, avoiding the GMO approach at any stage. In addition, we have shown that the *cenh3-4* mutant, which does not induce haploids under standard growth conditions, functions as a haploid inducer after exposure to short temperature stress.

Keywords: Double haploid, centromere, cenH3, KNL2, temperature stress

Introduction

The haploid generation technology, followed by whole genome duplication, is an effective strategy for accelerating plant breeding, as it allows to obtain true-breeding lines with complete homozygosity in a single step. In the conventional breeding approach, these lines are obtained by inbreeding, and often 7 to 9 generations of inbreeding are performed over several years to achieve the desired level of homozygosity (Britt and Kuppu, 2016). To produce (double) haploids (DHs), two main approaches have been widely used such as the *in vitro* explantation of gametophytic tissues (mainly cultivation of anthers or microspores) and the selective loss of one parental chromosome set *in vivo* through interspecific or intraspecific hybridization (Kalinowska *et al.*, 2019). However, depending on the tissue culture or crossability of the species of interest both approaches can only be applied to a limited number of genotypes. Hence, alternative resource-efficient and reliable approaches to produce DHs are strongly required. One way to improve the DH effectiveness is to develop efficient inducer lines that guarantee a high haploid induction rate (HIR) combined with a high-throughput haploid selection system. One promising approach to induce haploids is through centromere-mediated genome elimination (Ravi and Chan, 2010, Kuppu *et al.*, 2015).

Centromeres are unique chromosomal regions that mediate the kinetochore protein complex formation and microtubule attachment during cell division (Verdaasdonk and Bloom, 2011, Schalch and Steiner, 2017). Most centromeres are epigenetically defined by nucleosomes containing the centromere-specific histone H3 variant, cenH3 (Allshire, 1997). The cenH3 protein contains two domains, the N-terminal tail, which is a target for post-translational modification, and the C-terminal histone fold domain, which interacts with DNA and other histones to form the nucleosome. The loading of cenH3 to centromeres initiates the assembly of the functional kinetochore complex. The cenH3 loading pathway can be divided into three steps: initiation (centromere licencing), deposition, and maintenance. The centromere licencing factor KNL2, identified in *A. thaliana* showed colocalization with cenH3 throughout the cell cycle except from metaphase to mid-anaphase (Lermontova *et al.*, 2013). Furthermore, Sandmann *et al.*, (2017) identified a cenH3 nucleosome binding CENPC-k motif of KNL2 at its C-terminal part. The complete deletion of this motif or mutating of its conserved amino acids abolished the localization of KNL2 at centromeres. Thus, it is evident that the CENPC-K motif is functionally required for centromeric localization of KNL2 in *A. thaliana* (Sandmann *et al.*, 2017).

Due to its essential function in chromosome segregation, inactivation of cenH3 has been shown to result in chromosome segregation errors and lethality (Ravi and Chan, 2010, Ravi *et al.*, 2011). RNAi-mediated knockdown of *cenH3* showed a reduction in its mRNA level (27-43%) and also resulted in a dwarf plant phenotype and meiotic defects in *Arabidopsis* (Lermontova *et al.*, 2011). Recently, a mutation in cenH3 named *cenh3-4* has been discovered from the genetic suppressor screen, which increased fertility and promoted meiotic exit in *smg7-6* plants (Capitao *et al.*, 2021). The *cenh3-4* is a point mutation (G→A) in the splicing donor site of the 3rd exon of cenH3 showed a reduced amount of cenH3 at centromeres and thus, forming small centromeres. Similar to the cenH3 RNAi transformants, a T-DNA insertion knockout mutant of KNL2 showed a reduced amount of cenH3 at centromeres, decreased growth rate, fertility, and meiotic defects (Lermontova *et al.*, 2013), supporting further the functional relationship of both proteins.

Ravi and Chan, (2010) discovered that haploid plants can be obtained by pollination of a *cenh3-1* mutant of *A. thaliana* complemented with a GFP-tail swap construct (fusion of N-terminus of conventional H3 to the C-terminus of cenH3) with different wild-type accessions. This process at the end has resulted in haploid progenies with the genome of the wild-type parent at frequencies as high as 25-45%. If a wild-type female was crossed to a GFP-tail swap male, the proportion of haploid plants was lower. In recent studies, haploids also were achieved by introducing point mutations or small deletions in *Arabidopsis* cenH3 (Karimi-Ashtiyani *et al.*, 2015, Kuppu *et al.*, 2015, Kuppu *et al.*, 2020). Marimuthu *et al.* (2021) showed that cenH3 variants complementing the *cenh3-1* mutant are selectively removed from centromeres during reproduction. Additionally, the authors have demonstrated that the null mutant of VIM1 (VARIANT IN METHYLATION 1) enhances haploid induction frequencies of the complemented *cenh3-1* mutant. The cenH3-based haploid induction approach was successfully extended from *Arabidopsis* to crop plants, but using of homozygous *cenh3* mutant complemented with an altered variant of cenH3 resulted in an average haploid induction frequency below 1% in maize (Kelliher *et al.*, 2016). However, recently it has been reported that the use of heteroallelic cenH3 mutation combinations, which are characterized by reduced transmission in female gametophytes, has increased the HIR in maize to 5% (Wang *et al.*, 2021). Application of a similar haploid induction approach to wheat resulted in HIR up to 8% (Lv *et al.*, 2020).

In accordance with previous studies, an altered cenH3 protein would be sufficient for haploid induction in *Arabidopsis,* but whether an alteration of cenH3 assembly factors such as KNL2 could also be used as a haploid inducer has not been studied yet. Therefore, in this study, we show that a T-DNA knockout mutant of KNL2 is an effective haploid inducer when crossed with *Arabidopsis* wild-type plants. We demonstrate that short-term exposure of *knl2* to heat stress leads to an increase of the haploid induction efficiency from 1% to 10%. Moreover, the stress treatment regime defined for the haploid induction process with the *knl2* mutant also appeared to be effective for the *cenh3-4* mutant. Additionally, we showed that the introduction of a point mutation in the CENPC-k motif of KNL2 is sufficient to create a haploid inducer line.

Results

A short temperature stress of the *knl2* mutant increases the efficiency of haploid induction

The T-DNA knockout mutation of the cenH3 loading factor KNL2 (*knl2* mutant) results in a decreased amount of cenH3 protein, suggesting an essential role of KNL2 in the loading of cenH3 at centromeres (Lermontova *et al.*, 2013). Therefore, we assumed that the crossing of *knl2* mutant with wild-type *Arabidopsis* might generate haploids similarly to the cenH3 based haploid induction process. To test this hypothesis, the *knl2* mutant was crossed reciprocally with *Arabidopsis* wild-type accession *Landsberg erecta* grown under standard conditions. Flow cytometric analysis (FC) of pools of up to six seeds revealed 1% haploid progeny when *knl2* was used as the female parent (Figure 1, Table 1).

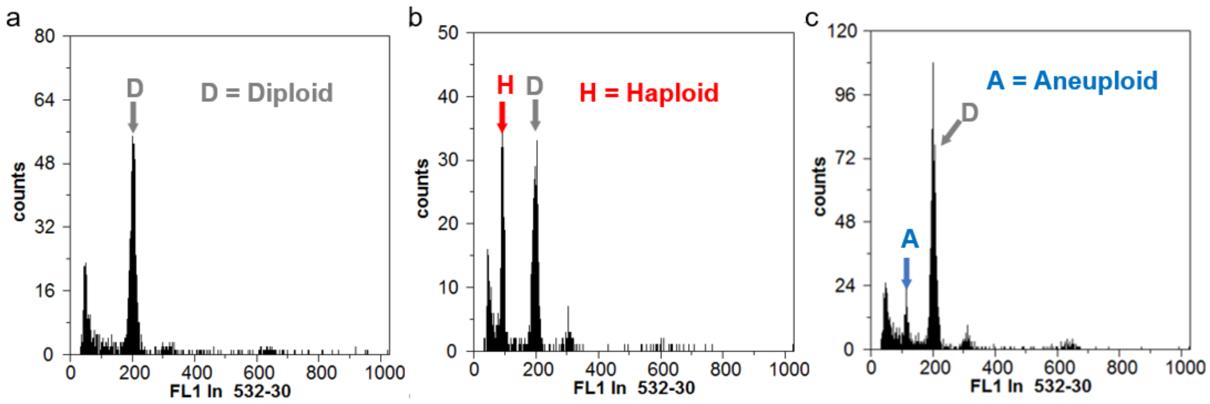


Figure 1 | Analysis of haploid seed pools from flow cytometry histograms.

a-c Flow cytometry histogram of 6-seed pools containing only diploid (a), haploid and diploid (b), aneuploid, and diploid (c) seeds. The presence of haploid/aneuploid seeds within the pools was determined by evaluating the PI fluorescence intensity on a linear scale. From the seed pools, we considered only one haploid seed per pool compared to the diploid population since the precise number of haploids/aneuploids per pool cannot be measured. Thus, the number of haploids/aneuploids is an underestimation rather than an overestimation.

Our previous RNAseq data analysis revealed a large number of stress-responsive genes that are differentially expressed in *knl2* seedlings and flower buds compared to wild-type (Boudichevskaia *et al.*, 2019). We therefore hypothesized that *knl2* mutant plants may be more sensitive to stress treatment than control plants and that exposure of *knl2* to the stress may increase HIR in its crosses with the wild-type. To support this assumption, the expression of cenH3 and cenH3 assembly factors KNL2, CENP-C, NASP under different stress conditions were analyzed. The gene expression data were retrieved from the *Arabidopsis* transcriptome data platform (http://ipf.sustech.edu.cn/pub/athrdb/). The results showed that these genes were significantly down-regulated in response to various stress treatments like heat, NPA (1-Naphthylphthalamic acid), and fluctuating light (Supplemental Table S1, Figure S1). Thus, increased temperature and light intensity has a strong effect on the expression of cenH3, KNL2 and other key kinetochore components.

To test the impact of stress on *knl2* growth and development and the induction of haploids, it was exposed to either high temperature or high light intensity before crossing. The usage of the *gl1-1* mutant as a crossing partner allows identifying haploids or double haploids based on its trichome-less phenotype (Kuppu *et al.*, 2015). First, all plants were cultivated for three weeks under long-day standard conditions (ST) at a temperature of 21/18°C day/night and light intensity at 100 µmol $m^{-2}sec^{-1}$. Then, one part of the plants remained under standard growth conditions while others were transferred either to a higher temperature (25/21°C day/night) or high light intensity (400 µmol $m^{-2}sec^{-1}$) (Figure 2a). For each growth condition, about 25 *knl2*, 15-20 wild-type, and 15-20 *gl1-1* plants were cultivated. At higher temperatures or light intensity, the phenotypic difference between the wild-type and the *knl2* mutant became more pronounced than under standard growth conditions (Figure S2). Reciprocal crosses were performed between *knl2* mutant plants and *gl1-1* cultivated under growth conditions as described above. The FC analysis of seed pools or trichome less *gl1-1* phenotype analysis of F1 plants revealed no increase in haploid induction efficiency in either type of continuous stress conditions (Table 1 and 2).

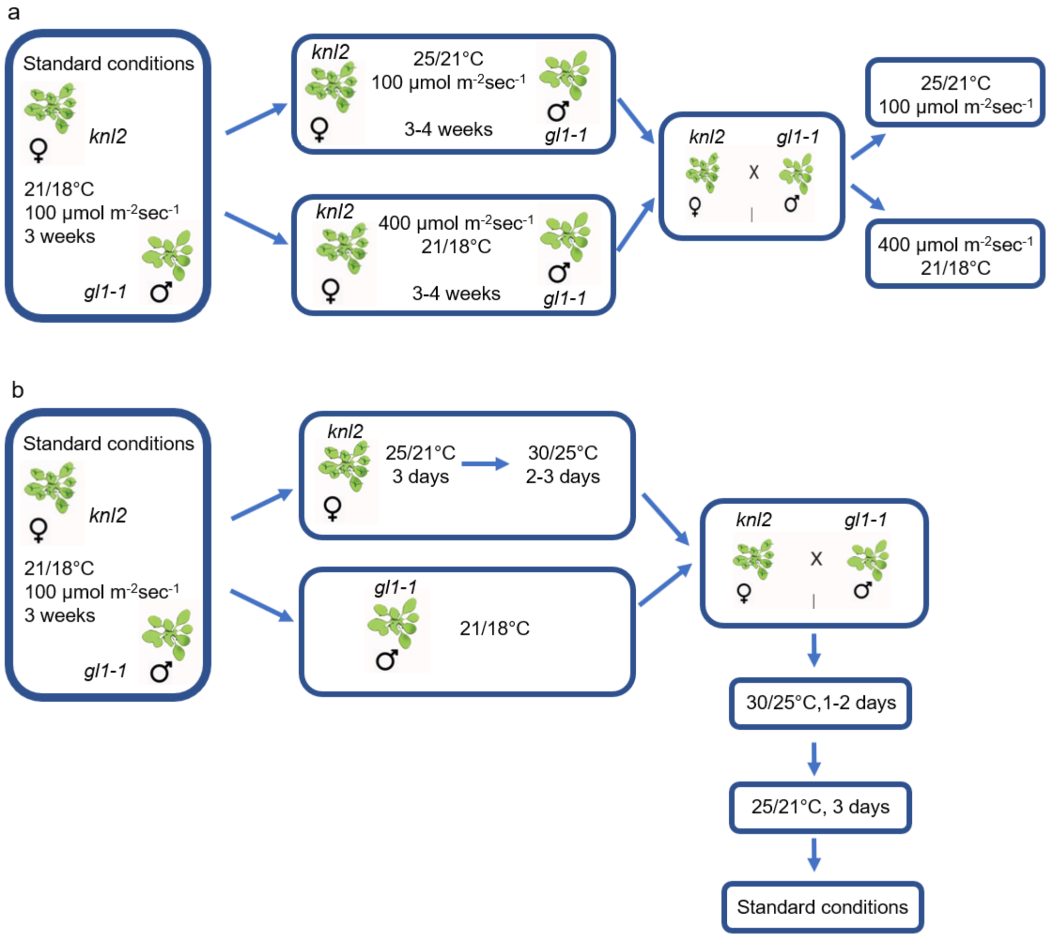
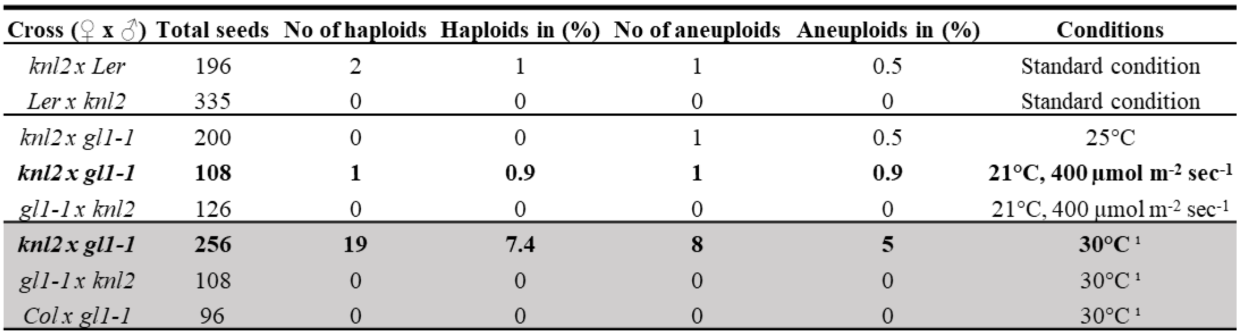


Figure 2 | Schematic representation of the crossing of the *knl2* mutant with the *gl1-1* marker line under temperature stress.

(a) The *knl2* and *gl1-1* plants were grown under standard growth conditions (21/18°C day/night and 100 µmol m^{-2}sec^{-1} light intensity) for three weeks, then the *knl2* mutant plants and *gl1-1* plants were transferred to a growth chamber with constantly increased temperature (25/21°C day/night) or light intensity (400 µmol m^{-2}sec^{-1}) for 3-4 weeks. The independent crossing of *knl2 and gl1-1* was carried out for increased temperature and light intensities and the plants have remained at the same conditions. (b) Similarly, *knl2* and *gl1-1* plants were grown under standard growth conditions until flowering. Afterward, *knl2* plants were moved to high temperatures (25-21°C day/night) for 3 days followed by short temperature stress (30/25°C day/night) for 2-3 days, while the *gl1-1 L. erecta* marker line was left under standard growth conditions. Then, the *knl2* and *gl1-1* plants were crossed and placed back to 30/25°C day/night for 1-2 days. The temperature was reduced stepwise: first to 25/21°C day/night for three days and then to the standard conditions.

Table 1 | Ploidy analysis of seeds derived from the reciprocal crosses of *knl2* with wild-type *A. thaliana* or *gl1-1*

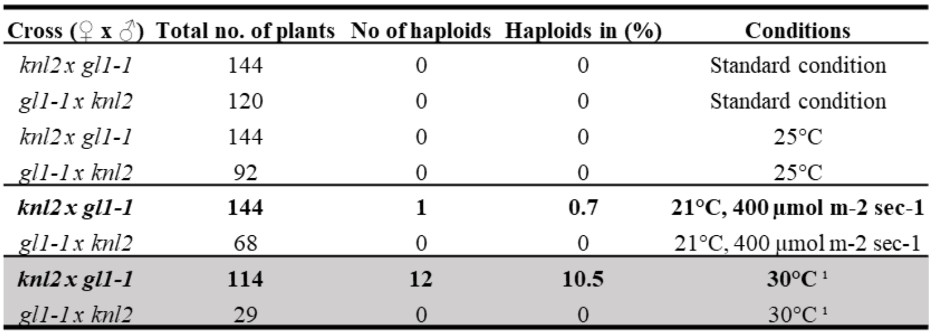


Cross (♀ x ♂)	Total seeds	No of haploids	Haploids in (%)	No of aneuploids	Aneuploids in (%)	Conditions
knl2 x Ler	196	2	1	1	0.5	Standard condition
Ler x knl2	335	0	0	0	0	Standard condition
knl2 x gl1-1	200	0	0	1	0.5	25°C
knl2 x gl1-1	**108**	**1**	**0.9**	**1**	**0.9**	**21°C, 400 µmol m^{-2} sec^{-1}**
gl1-1 x knl2	126	0	0	0	0	21°C, 400 µmol m^{-2} sec^{-1}
knl2 x gl1-1	**256**	**19**	**7.4**	**8**	**5**	**30°C [1]**
gl1-1 x knl2	108	0	0	0	0	30°C [1]
Col x gl1-1	96	0	0	0	0	30°C [1]

[1] The plants treated under temperature stress for short time before crossing

Assuming that we did not get an increase in HIR due to adaptation of the *knl2* mutant to continuous stress, the experimental setting was changed and *knl2* mutant plants were exposed to high temperature (30°C) for a short period (2-3 days) before crossing (Figure 2b), while the *gl1-1* crossing partner remained under standard conditions. The temperature was increased and decreased stepwise, as shown in Figure 2. The reciprocal crosses were repeated at least three times in two different growth chambers. FC analysis of seed pools and *gl1-1* mutant phenotype analysis of F1 plants (Figure 3), displayed a similar haploid induction efficiency of 7.4% and 10%, respectively, when heat-stressed *knl2* was used as the female. As the FC analysis of seeds was performed in pools of six seeds and counted only one haploid seed per pool, the number of haploids is compared to the diploid population, which resulted in an underestimation of the number of haploids. No haploids were detected when heat-stressed *knl2* was used as a pollen donor (Table 1 and 2).

Table 2 | Phenotype-based selection of plants derived from the reciprocal cross of *knl2* with ***gl1-1***



Cross (♀ x ♂)	Total no. of plants	No of haploids	Haploids in (%)	Conditions
knl2 x gl1-1	144	0	0	Standard condition
gl1-1 x knl2	120	0	0	Standard condition
knl2 x gl1-1	144	0	0	25°C
gl1-1 x knl2	92	0	0	25°C
knl2 x gl1-1	**144**	**1**	**0.7**	**21°C, 400 µmol m-2 sec-1**
gl1-1 x knl2	68	0	0	21°C, 400 µmol m-2 sec-1
knl2 x gl1-1	**114**	**12**	**10.5**	**30°C [1]**
gl1-1 x knl2	29	0	0	30°C [1]

[1] The plants treated under temperature stress for short time before crossing

The trichomeless plants were much smaller than the corresponding diploids (Figure 3a,b). The 1C nuclei of selected *gl1-1* plants revealed a maximum of 5 chromocenters (Figure 3c) and further, immunostaining of cenH3 also showed 5 chromocenter-localized signals, thus confirming haploidy (Figure 3d). Moreover, a sample flow histogram plot of haploids produced from *knl2* and *gl1-1* crosses and diploid control was shown (Figure 3e).

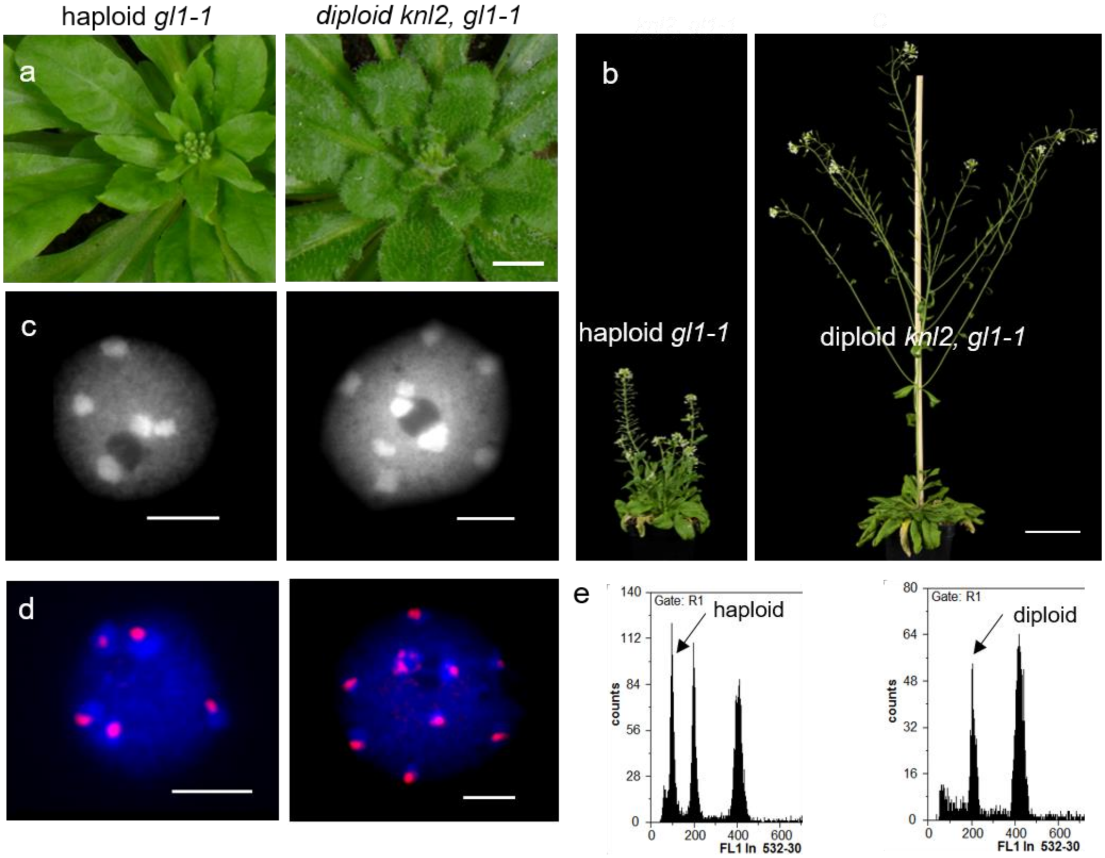


Figure 3 | Haploid progenies obtained by genome elimination in crosses of *knl2* with the trichome-less *gl1-1* marker.

(a) Comparison of the haploid plant without trichomes (left) and diploid hybrid phenotype with trichomes (right). Scale bar = 1 cm. (b) Phenotype of haploid *gl1-1* and diploid *knl2*, *gl1-1* hybrid plants during the generative development stage. Scale bar = 5 cm. (c) DAPI stained nuclei isolated from haploid and diploid plants showing a maximum of 5 and 10 heterochromatic chromocenters, respectively. (d) Anti-cenH3 labelled nuclei isolated from haploid and diploid plants showing a maximum of 5 and 10 immunosignals (in red), respectively. The scale bars represent 5 µm. (e) Histogram analysis of nuclei by flow cytometry for a *gl1-1* haploid offspring and control diploid.

Besides haploids, 5% of aneuploid seeds were detected by FC (Table 1). However, the control crosses of wild-type Col-0 with *gl1-1 Ler* under the same conditions did not produce haploid plants. Thus, short temperature stress increases the haploidization frequency if *knl2* was used as a female crossing partner.

A point mutation at the CENPC-k motif of KNL2 results in haploid induction on outcrossing

We previously identified a conserved CENPC-k motif in the KNL2 protein and showed that deletion of this motif or mutagenesis of its conserved amino acids Arg-546 and Trp-555 abolishes the centromeric localization of KNL2 (Sandmann *et al.*, 2017). To test whether the introduction of a point mutation into the CENPC-k motif would be sufficient for haploid induction, the genomic KNL2 fragment with the endogenous promoter was cloned into the pDONR221 vector. To substitute the conserved Thr-555 by Arg, PCR-based site-directed mutagenesis was performed (Figure 4). The resulting clone and wild-type KNL2 were subcloned into pGWB640 vector in fusion with EYFP and used for the transformation of the *knl2* mutant. The selected transgenic plants were analyzed for the subcellular localization of KNL2-EYFP fusion protein. In *knl2* mutant complemented with the unmodified KNL2-EYFP construct, fluorescence signals were detected in the nucleoplasm and at chromocenters (Figure 4a), while in *knl2* expressing KNL2-EYFP with the point mutation within the CENPC-k motif, the EYFP signals were detected only in the nucleoplasm (Figure 4b). Immunostaining of root tip nuclei of both variants of transformants with anti-KNL2 antibodies confirmed the centromeric localization of unmutated KNL2 and the nucleoplasmic localization of the variant with a point mutation (Figure 4). Three transgenic lines per construct were selected for the haploid induction experiment under the 30°C degree short temperature stress condition as female crossing partners.

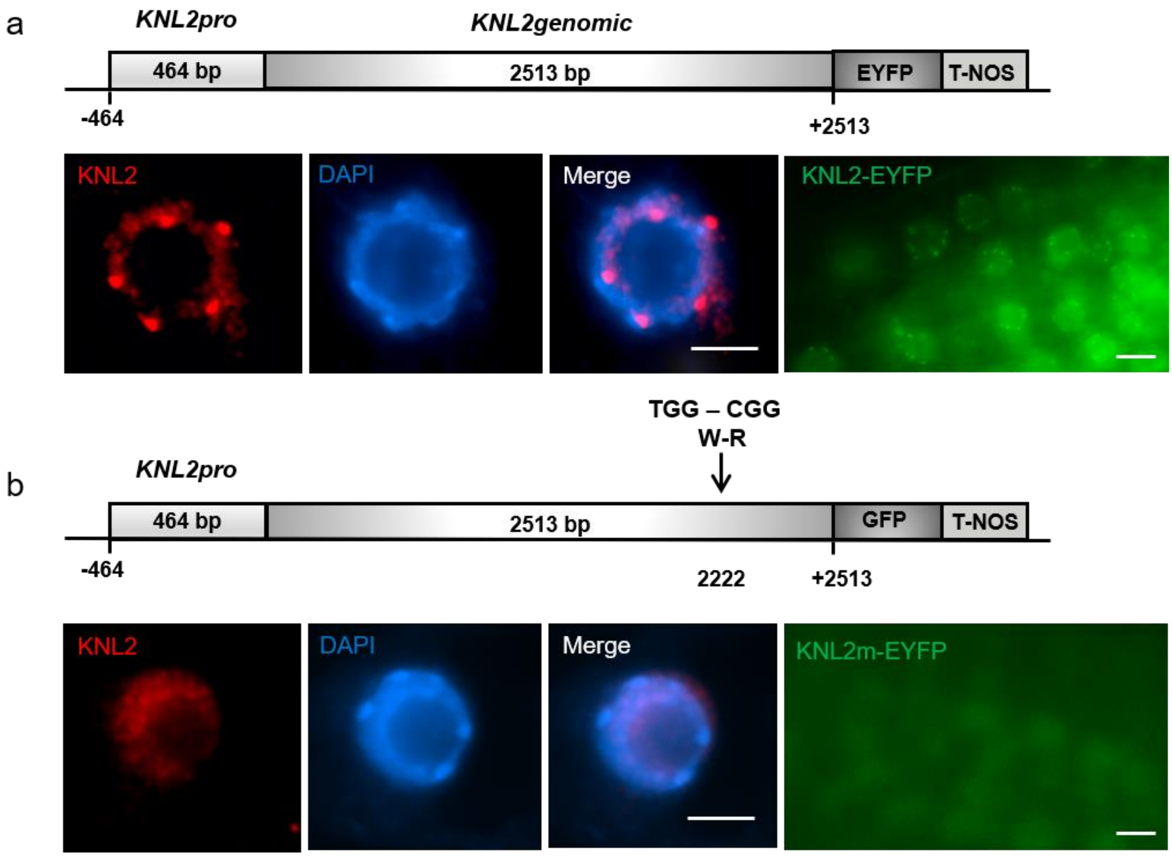


Figure 4 | Substitution of the amino acid Trp by Arg within the conserved CENPC-k motif abolished centromeric localization of KNL2.

a-b Schematic representation of the genomic KNL2-EYFP fusion construct (upper parts) unmodified (a) or carrying the W to R mutation within the CENPC-k motif (b) and the subcellular localization of the corresponding fusion proteins in root tip nuclei of *Arabidopsis* immunostained with anti-KNL2 antibodies (lower parts, panels 1-3) or analyzed by confocal microscopy (lower parts, panel 4). The unmutated KNL2-EYFP fusion protein showed centromeric and nucleoplasmic localization (a) while the variant with point mutation can be detected only in the nucleoplasm (b). The scale bars are 5 µm.

The haploid induction efficiency of the *knl2* mutant complemented by KNL2-EYFP with the point mutation varied from 0.8 to 5.6%. In contrast, no haploids were detected in the case of *knl2* expressing the wild-type KNL2 control construct was used (Table 3).

Table 3 | Phenotype-based selection of plants derived from the reciprocal cross of *knl2 mutant* with *gl1-1*

Cross (♀ x ♂)	Total seeds	No of haploids	Haploids in (%)	No of aneuploids	Aneuploids in (%)	Conditions
KNL2gen(W-R) Line 1 x gl1-1	**126**	**1**	**0.8**	**2**	**1.5**	**30°C[1]**
KNL2gen(W-R) Line 2 x gl1-1	**108**	**6**	**5.6**	**4**	**3.7**	**30°C[1]**
KNL2gen(W-R) Line 4 x gl1-1	**120**	**1**	**0.8**	**2**	**1.7**	**30°C[1]**
KNL2gen Line1 x gl1-1	60	0	0	0	0	30°C[1]
KNL2gen Line2 x gl1-1	108	0	0	0	0	30°C[1]
KNL2gen Line2 x gl1-1	54	0	0	0	0	30°C[1]

[1] The plants treated under temperature stress for short time before crossing

Analysis of the paternal haploid plants did not reveal any traces of the maternal genome

Next, a PCR-based marker analysis of three double haploid plants was performed to confirm whether only the chromosomes of the pollen donor remained. One genotype-specific marker per chromosome was employed (Figure S3), and in all cases, PCR amplicons were found corresponding to the *gl1-1* mutant. To exclude the presence of small chromosome fragments as a byproduct of the haploidization process as reported by (Tan *et al.*, 2015), a Single Nucleotide Polymorphism (SNP) analysis was performed based on Next-Generation Sequence (NGS) reads of DNA samples isolated from three double haploids, one hybrid, and two parental plants. The SNP analysis clearly showed that *gl1-1* double haploid plants do not contain any residues of the maternal *knl2* chromosome complement (Figure 5). The hybrid, by contrast, was heterozygous throughout its genome. A read depth analysis did not show any chromosomal aberrations as observed by (Tan *et al.*, 2015)(Figure S4).

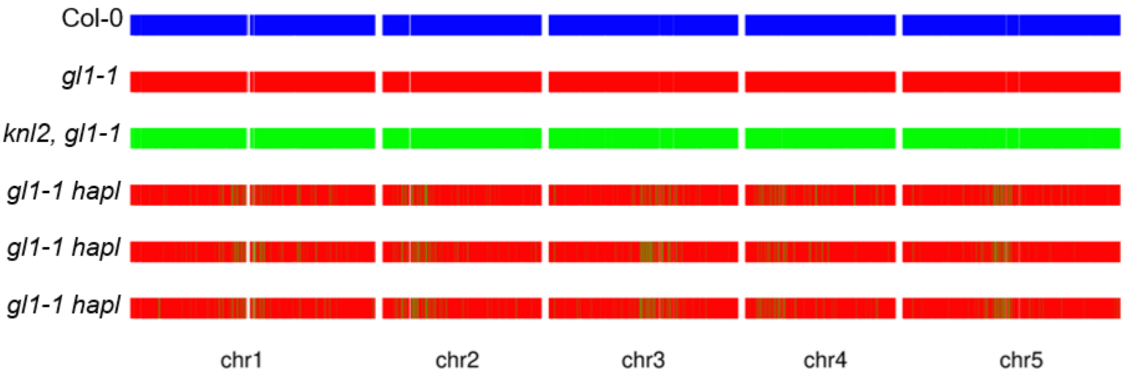
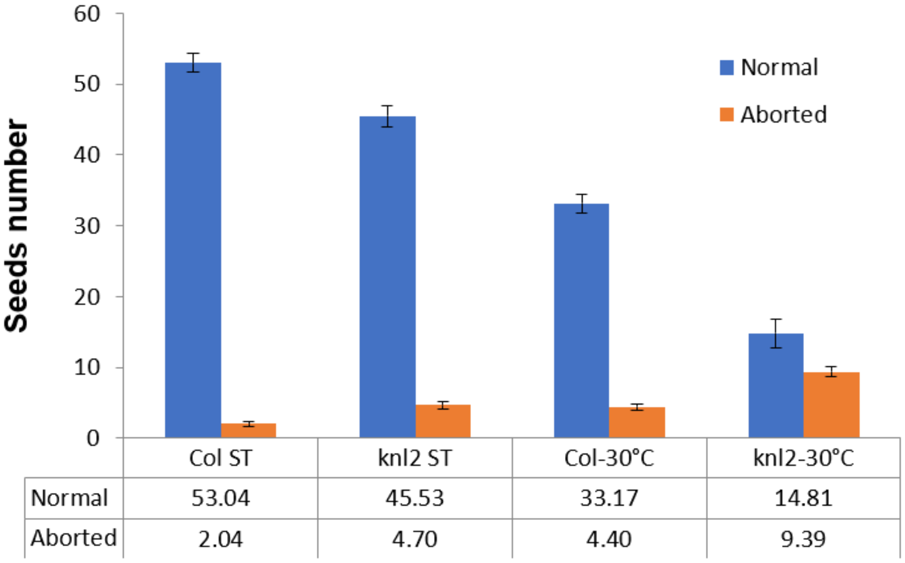


Figure 5 | Confirmation of haploid progeny obtained by crossing of *knl2* with the trichome-less *gl1-1* mutant.

Single Nucleotide Polymorphism (SNP) analysis of three *gl1-1* double haploids, *knl2*, *gl1-1* hybrid, *gl1-1*, and Col plants. The results displayed that the hybrid plants were completely heterozygous whereas *gl1-1* double haploid plants do not contain any residues of the maternal *knl2* mutant genome.

Plants exposed to high-temperature show reduced seed setting as a result of increased mitotic and meiotic abnormalities

In *Arabidopsis* wild-type and *knl2* mutant, exposure to high temperature (30°C) using the regime indicated (Figure 2b) resulted in decreased seed setting and an increased number of aborted seeds after selfing. However, this effect was more pronounced in *knl2* mutant compared to wild-type (Figure 6). Thus, after exposure to high temperature, the average seed number per silique was reduced from 53 to 33 in Col and 45 to 14 in *knl2* mutant, while the number of aborted seeds was increased from 2 to 4 in Col and 5 to 9 in *knl2* mutant, respectively. To address whether the reduction in fertility was based on defects during meiosis, male meiotic chromosome spread analysis was performed in wild-type and *knl2* plants exposed to the same growth conditions as mentioned above.



	Col ST	knl2 ST	Col-30°C	knl2-30°C
Normal	53.04	45.53	33.17	14.81
Aborted	2.04	4.70	4.40	9.39

Figure 6 | Exposure of *Arabidopsis knl2* mutant and wild-type to high temperature resulted in a decreased seed setting and an increased number of aborted seeds.

Seed setting analysis was performed on selfed plants either continuously grown under standard growth conditions or exposed for 4 days to 30/25°C (day/night) as it is indicated in Figure 2b.

No meiotic defects were observed neither at 21°C nor at 30°C (two plants each) in wild-type plants, i.e. homologous chromosomes undergo synapsis at pachytene, five bivalents are inevitably found at metaphase I, homologous chromosomes segregate during the first meiotic division, and the sister chromatids are separated during the second meiotic division (Figure 7). In *knl2* plants grown at 21°C (2 out of 4 plants) and 30°C (4 out of 4 plants), meiotic defects were detected including synapsis defects (asynapsis and interlocks), as well as lagging chromosomes and chromosome fragmentation during the first and second meiotic divisions (Figure 7).

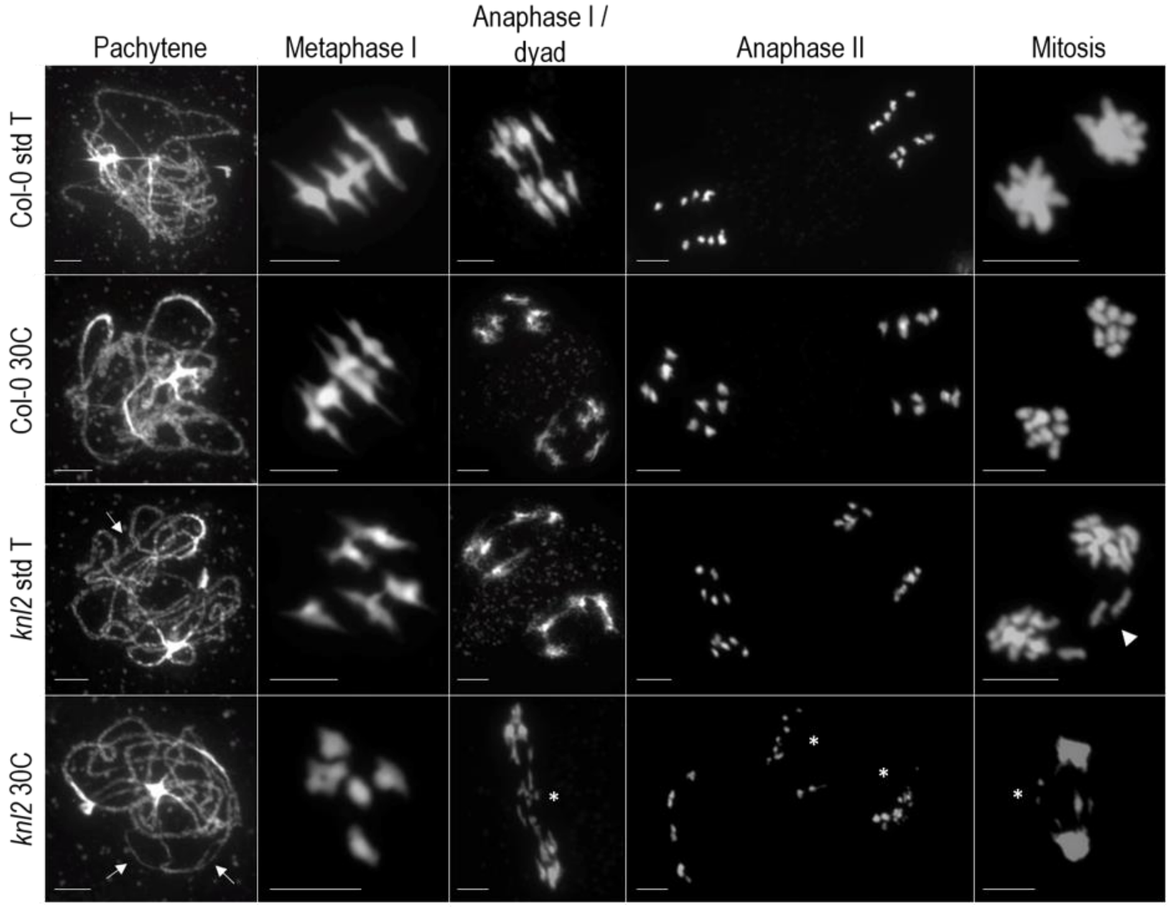
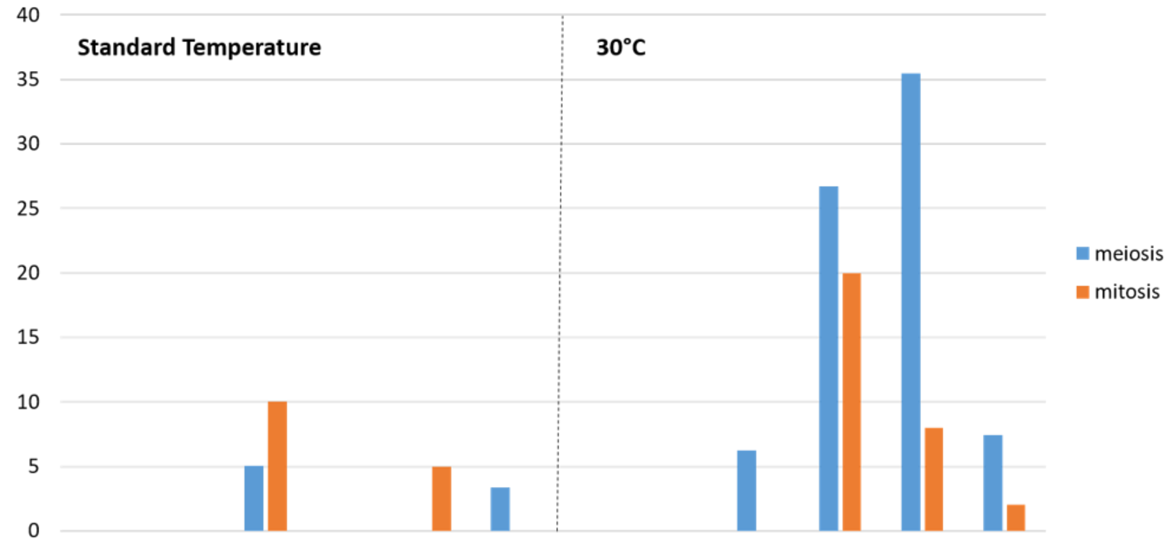


Figure 7 | Mitotic and meiotic defects under high temperature in the *knl2* mutant

Male meiosis and mitosis (from tapetum cells) in wild-type and knl-2 plants grown at 21°C or exposed to 30°C as it is indicated in Figure 2b. In wild-type plants grown at either temperature show no errors in meiosis and mitosis progress. During meiosis, homologous chromosomes undergo synapsis at pachytene, form five bivalents at metaphase I, segregate to opposite poles at anaphase I/dyad, and separate sister chromatids during anaphase II. During mitosis, sister chromatids separate to opposite poles. In knl2 plants, meiotic and mitotic defects are found, including asynapsis and interlocks during pachytene (arrows), lagging chromosomes (arrowhead), and chromosome fragmentation (asterisks) during mitotic and meiotic divisions. The chromosomes were stained with DAPI (blue). The bar represents 5 µm.

The degree of observed defects varied among plants and was more pronounced at 30°C than at 21°C (Figure 8). Due to observed meiotic chromosome fragmentation, mitotic divisions of tapetum cells from the same plants were studied. Similar to meiosis, mitotic defects were observed in *knl2* plants grown at 21°C (2 out of 4 plants) and 30°C (3 out of 4 plants), which include lagging chromosomes, anaphase bridges, and chromosome fragmentation (Figure 7) and were more pronounced at a higher temperature. No obvious mitotic defects were found in wild-type grown at either temperature (two plants each). In a nutshell, mitotic and meiotic defects were found to varying degrees in *knl2* plants that were more frequent in plants grown at 30°C, while no noticeable defects were observed in wild-type plants at both temperature regimes.



	#1	#2	#1	#2	#3	#4	#1	#2	#1	#2	#3	#4
	Col-0		*knl-2*				Col-0		*knl-2*			
Meiosis N	0% 46	0% 109	5.1% 79	0% 31	0% 30	3.3% 90	0% 48	0% 96	6.3% 32	26.7% 45	35.5% 31	7.4% 81
Mitosis N	0% 50	0% 50	10% 50	0% 50	5% 60	0% 50	0% 50	0% 50	0% 50	20% 50	8% 50	2% 50

Figure 8 | Meiotic and mitotic phenotype of independent wild-type and *knl-2* plants at 21°C and 30°C.

The percentage of observed cells with defects per independent plant (meiosis: synapsis defects, lagging chromosomes, and chromosome fragmentation; mitosis: lagging chromosomes, anaphase bridges, and chromosome fragmentation) and the number of cells analyzed were indicated.

A *cenh3-4* mutation induces haploid formation under temperature stress

The *cenh3-4*, a point mutation of cenH3 (G→A amino acid substitution in the third exon of cenH3) showed a substantially reduced level of cenH3 at the centromere and causes defects in the mitotic spindle. Nevertheless, the reduced level of cenH3 induced haploid plants only with a very low frequency (0.2%) when crossed with wild-type plants. Thus the smaller centromere size was not efficient to trigger haploidization in *Arabidopsis* (Capitao *et al.,* 2021). Considering the effect of heat stress on the *knl2* mutant, we tested the haploid induction rate with *cenh3-4* mutants cultivated under heat stress. In our initial experiment, *cenh3-4* plants were exposed to increasing temperatures for a longer period before pollination, but transferred to standard conditions (22°C) immediately after the pollination. The *cenh3-4* mutant and trichome-less *gl1-1* plants were grown under 22°C for two weeks, plants at the earlier rosette stage were transferred to the chambers containing different temperatures varying from 16ºC to 30ºC and cultivated for additional two to three weeks until they formed flowers (Table 4). Then, the *cenh3-4* mutant plants were pollinated with pollen from trichome-less *gl1-1* plants grown under 22°C, and the pollinated plants were transferred to 22°C. The haploid induction rate was determined based on the trichome-less phenotype. Out of 181 progeny plants grown at 26ºC and 253 plants grown at 30ºC, one and two haploid plants were identified, respectively (Table 4). No haploid plants were found among more than 3197 progeny plants at lower temperatures (16, 22, 24°C). This data supports the notion that increased temperature promotes haploid induction in centromere-impaired plants.

Table 4 | Analysis of haploid induction in cenH3 mutants based on phenotype and ploidy levels by flow cytometry

Cross (♀ x ♂)	Total seeds/plants	No of haploids	Haploids in (%)	No of aneuploids	Aneuploids in (%)	Conditions
cenh3-4 x gl1-1	1023	0	0	-	-	16°C [1]
cenh3-4 x gl1-1	998	0	0	-	-	22°C [1]
cenh3-4 x gl1-1	1176	0	0	-	-	24°C [1]
cenh3-4xgl1-1	**181**	**1**	**0.55**	-	-	**26°C [1]**
cenh3-4 x gl1-1	86	0	0	-	-	28°C [1]
cenh3-4 x gl1-1	**253**	**2**	**0.79**	-	-	**30°C [1]**
cenh3-4 x gl1-1	**336**	**11**	**3.3**	**9**	**2.7**	**30°C [2]**
gl1-1 x cenh3-4	90	0	0	1	1.1	30°C [2]
cenh3-4 x gl1-1	**96**	**4**	**4.1**	-	-	**30°C [3]**
cenH3 RNAi x gl1-1	120	0	0	0	0	30°C [2]
gl1-1 x cenH3 RNAi	66	0	0	0	0	30°C [2]

[1] *cenh3-4* plants treated under temperature stress continuously and analyzed based on phenotypic differences
[2] *cenh3-4* plants treated under temperature stress shortly before crossing and analyzed by flow cytometry
[3] *cenh3-4* plants treated under temperature stress shortly before crossing and analyzed based on phenotypic differences

Nevertheless, the HIR in *cenh3-4* plants at continuous heat stress regime was substantially lower than the one achieved with *knl2* mutants using heat treatment described in Figure 2a. Therefore, we recapitulated the experiment using the same conditions as for the *knl2* plants. The *cenh3-4* plants were subjected to 30°C for 2 days (Figure 2b), then pollinated with *gl1-1* pollen, cultivated for additional 2 days at 30°C followed by 3 days at 25°C before transferring to standard conditions. Analysis of 56 seed pools (6 seeds per pool) with a total of 336 seeds by flow cytometry revealed a haploid and aneuploid induction rate of 3.3% and 2.7%, respectively (Table 4). Moreover, the haploid induction frequency of 4.1 % was determined based on the trichomeless phenotype of *gl1-1* (Table 4). This result indicates that sustaining the temperature stress for several days after pollination further improves HIR.

Additionally, cenH3 RNAi transformants that revealed a substantial reduction of cenH3 at centromeres (Lermontova *et al.*, 2011) were tested as haploid inducers in combination with heat stress. Reciprocal crosses of RNAi with *gl1-1* were performed under short-term 30°C-stress conditions (Figure 2b). In contrast to *cenh3-4*, FC analysis of seeds did not reveal either haploids or aneuploids (Table 4).

Discussion

In most eukaryotes, kinetochore assembly is primed by cenH3 and multiple kinetochore protein complexes are required for accurate chromosome segregation. We showed that the disruption of the cenH3 loading machinery via the inactivation of the centromere licensing factor KNL2 of *Arabidopsis* resulted in the generation of haploids (HIR – 1%) on outcrossing with wild-type. To enhance the efficiency of haploid induction, the *knl2* mutant plants were subjected to various stress conditions, such as increased temperature and light intensity as we found that deregulated expression of KNL2 leads to differential expression of many stress-responsive genes (Boudichevskaia *et al.*, 2019). However, cultivation of *knl2* plants under long-term stress conditions (25/20°C day/night, 100 µmol $m^{-2}sec^{-1}$ light intensity or 21/18°C day/night and 400 µmol $m^{-2}sec^{-1}$ light intensity) did not increase HIR in the reciprocal crosses of *knl2* with wild-type. Assuming that the applied cultivation regimes either did not cause severe stress or that the plants had adapted to the long-term treatment, short-term treatment with high temperature (30/25°C day/night) has been applied for 2-3 days prior to the crossing experiments. Indeed, exposure of *knl2* to high temperatures for a short period allowed us to increase the HIR to up to 10 %.

Moreover, the same heat stress treatment applied to the *cenh3-4* mutant (Capitao *et al.*, 2021) resulted in an increase in HIR from 0.2% under standard conditions to 4.1%. Interestingly, our data also show that the more efficient HIR is achieved when heat stress is prolonged to the postfertilization period. This indicates that the haploid induction in centromere impaired mutants is conditioned by temperature stress during both ovule development as well as early embryogenesis. Originally, it was thought that haploids could only be obtained by crossing the *cenh3* mutant lines complemented by a modified version of cenH3 with the wild-type, as the elimination of one of the genomes is caused by competition between two structurally different cenH3 variants for the deposition to centromeres (Ravi *et al.,* 2014, Thondehaalmath *et al.*, 2021). To understand the mechanism of genome elimination in such crosses, Marimuthu *et al.* (2021) analyzed the distribution of the altered cenH3 (GFP-tail swap) variant in gametes and at different developmental stages of hybrid zygotes. It has been shown that altered cenH3 is selectively removed from mature *Arabidopsis* eggs and early hybrid zygotes while at the later zygotic stages, cenH3 and GFP-tail swap preferentially can be loaded into the centromere of the wild-type parent, whereas the cenH3-depleted mutant chromosomes are not able to reconstitute new cenH3 chromatin and undergo elimination. However, Wang *et al.* (2021) and Lv *et al.* (2020) have demonstrated that heterozygous cenH3 mutants of maize and wheat, respectively, can also function as efficient haploid inducers in crosses with the wild-type in both directions, despite the lack of competition between the two structurally different cenH3 variants. In these cases, it was assumed that weak centromeres are formed due to cenH3 dilution that occurs as a result of postmeiotic divisions in gametogenesis. Since female haploid spores undergo three mitotic divisions and male only two, the level of cenH3 in female gametes is expected to be lower (Wang *et al.*, 2021). When a heterozygous *cenh3-1* mutant of *Arabidopsis* was used as an HI in crosses with wild-type, haploids could be generated at a frequency of ~1%, indicating that also in *Arabidopsis* haploids can be produced without alteration of cenH3 (Marimuthu *et al.*, 2021).

Using *knl2* and *cenh3-4* mutants of *Arabidopsis*, we further demonstrated that a competition of structurally different variants of cenH3 is not always a prerequisite for haploid induction. Thus, similar to maize and wheat, the centromere size model of (Wang and Dawe, 2018) can be applied to the haploid induction approach based on *knl2* and *cenh3-4* mutants. We suggest that applying temperature stress to *knl2* and *cenh3-4* mutants can weaken their centromeres additionally compared to standard growth conditions and therefore lead to an increase in HIR (Figure 9).

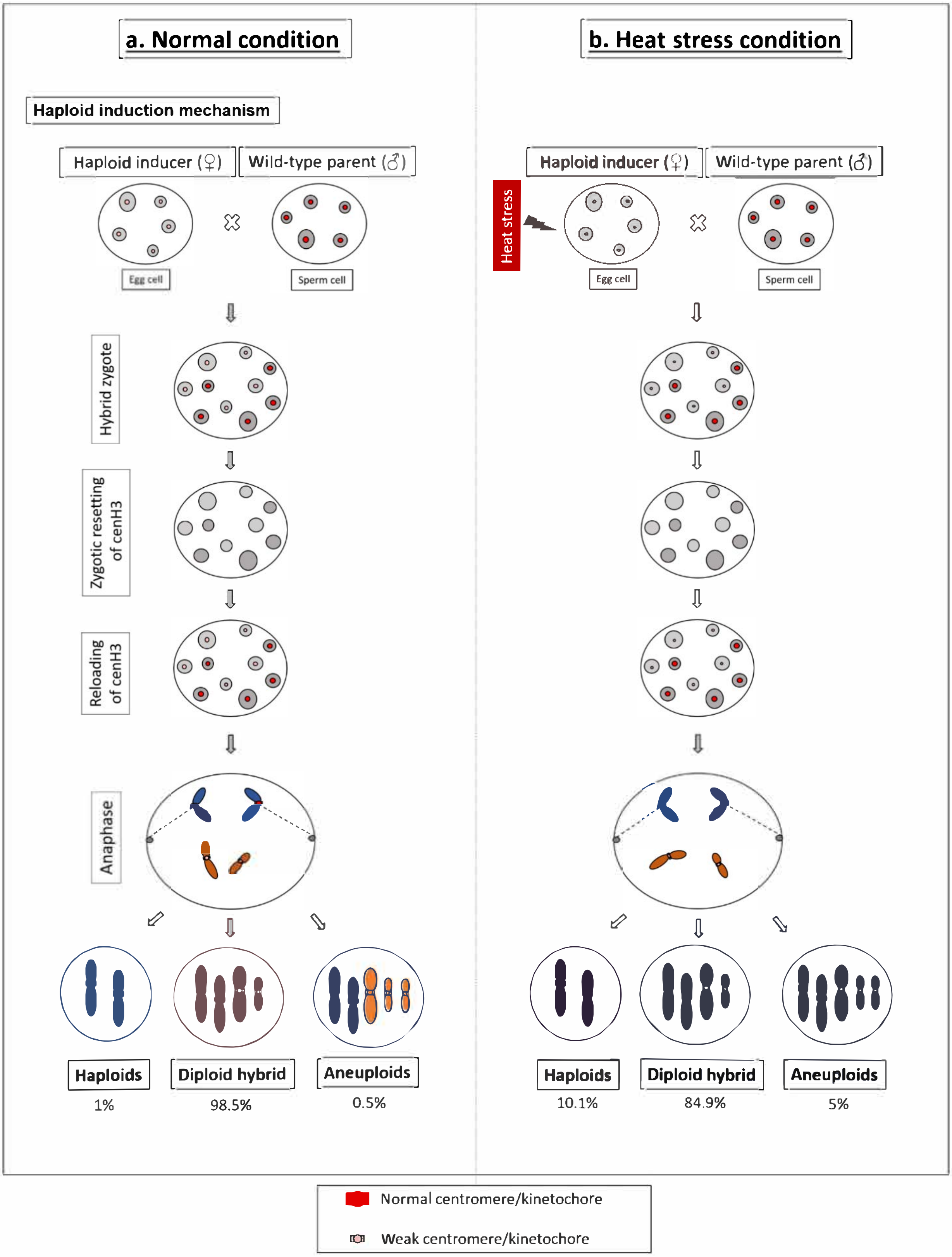

a. Normal condition
b. Heat stress condition
Haploid induction mechanism
Haploid inducer (♀)
Wild-type parent (♂)
Heat stress
Egg cell
Sperm cell
Hybrid zygote
Zygotic resetting of cenH3
Reloading of cenH3
Anaphase
Haploids
1%
Diploid hybrid
98.5%
Aneuploids
0.5%
Haploids
10.1%
Diploid hybrid
84.9%
Aneuploids
5%
Normal centromere/kinetochore
Weak centromere/kinetochore

Figure 9 | A model comparing a uniparental chromosome elimination mechanism in crosses of the *knl2* mutant × wild-type under standard (a) and heat stress condition (b).

Model explains the elimination of uniparental chromosomes in haploid inducer *knl2* mutant crossed with wild-type under standard condition (left panel) and heat stress condition (right panel). (a) In standard conditions, the combination of small size centromeres of the haploid inducer (*knl2* mutant) with 'normal' wild-type centromeres in the hybrid zygote leads to centromere competition, followed by complete or partial elimination of the genome and formation of haploid (1%) and aneuploid (0.5%) progeny of wild-type, respectively. (b) Under heat stress conditions, the haploid inducer (*knl2*) chromosomes have severe mitotic and meiotic defects as shown in (Figure 7), which leads to the centromere inactivity and making very small centromeres. Therefore, the haploid induction rate was increased to 10% when heat stressed *knl2* mutant crossed with wild-type plants.

The bioinformatic analysis revealed a reduced expression of genes coding for the kinetochore proteins such as cenH3, KNL2, and CENP-C under stress conditions. And while this reduction is not critical in the wild-type, in *knl2* and *cenh3-4* it amplifies the effect of mutations. This suggestion can be supported by our data showing that the frequency of mitotic and meiotic defects in *knl2* can be increased after short-term heat stress treatment while wild-type plants cultivated under the same conditions did not show any mitotic or meiotic abnormalities. Our RNAseq data analysis has revealed that a high number of transposable elements was activated in seedlings and flower buds of the *knl2* mutant cultivated under standard growth conditions (Boudichevskaia *et al.*, 2019). Thus, it can be assumed that exposure of *knl2* to heat stress might result in even an increased number of active transposons compared to the standard conditions and disturb chromatin organization consequently (Probst and Mittelsten Scheid, 2015). These are some of the possibilities behind the temperature stress-induced haploid formation but, a clear mechanism is still needed to be elucidated.

It is important to understand why haploids cannot be obtained when heat-treated mutants are used as pollen donors. There are relatively few examples in which the effects of temperature stress on female reproductive organs have been investigated, but much more is known of the effects of temperature stress on male reproductive structures (Hedhly *et al.*, 2009). Using tomato male-sterile and male-fertile lines, (Peet *et al.*, 1998) have demonstrated that stress applied to the pollen donor plant before and during pollen release decreased seed number and fruit set more severely than heat stress applied to the developing ovule. Thus, heat stress treatment of *knl2* and *cenh3-4* mutants as pollen donors can lead to a decrease in pollen viability and the inability to fertilize the egg. At the same time, fertilization of the ovule with viable pollen will not lead to the process of genome elimination. In contrast, heat stress treatment of mutants as maternal crossing partners can only lead to a weakening of the centromeres but does not affect the viability of the ovules.

The most notorious phenotypes in the *knl2* mutant plant's meiosis were the presence of lagging chromosomes and fragmentation. Lagging chromosomes could appear as a consequence of weaker or defective centromere activity in *knl2* plants. Fragmentation appeared during mitosis and meiosis. Interestingly, two plants grown at 30°C showed fragmentation in anthers during the first meiotic division, and the other two plants showed during the second meiotic division. This suggests that fragments could be originated by different mechanisms. One possibility is that interlocks observed during pachytene are not properly repaired and another would be the defective DNA repair in pathways involving both inter-sister and inter-homolog events. Besides, anaphase bridges were also observed due to the effect of heat stress in the *knl2* mutant. In a recent publication describing the comparative analysis of seedlings and flower bud transcriptomes of *knl2* mutant and wild-type, it has been shown that genes coding for proteins involved in DNA repair were overrepresented among down-regulated *knl2* genes (Boudichevskaia *et al.*, 2019).

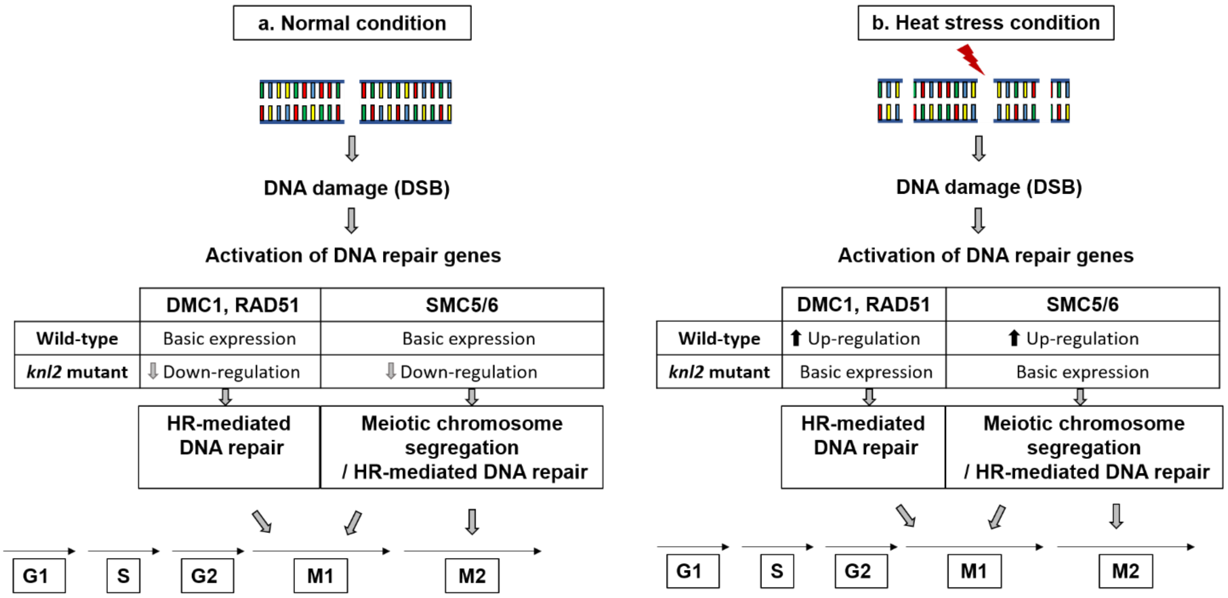


Figure 10 | Effect of DNA repair pathway in wild-type and *knl2* mutant under normal (a) and heat stress conditions (b)

In standard growth conditions (left panel), *knl2* mutant shows a reduced expression of genes encoding the DNA repair proteins (Boudichevskaia *et al.*, 2019) that might correlate with the mitotic and meiotic abnormalities in the mutant. Exposure of *knl2* and wild-type to heat stress (right panel) results in increased DNA damage. In wild-type, this is accompanied by increased expression of DNA repair genes (Han *et al.*, 2020), whereas in *knl2* these genes are down-regulated under standard growth conditions and therefore, their expression under heat stress cannot be sufficiently increased. In the *knl2* mutant, heat stress increases DNA damage, mitotic and meiotic abnormalities, and reduced expression of cenH3 and other kinetochore related genes. These factors could be the reason for the increase in HIR when heat stressed *knl2* mutant used as a haploid inducer.

For instance, the down-regulated genes were KU70 (AT1G16970), KU80 (AT1G48050), and LIGASE 4 (AT5G57160), the key players participating in the canonical non-homologous end joining, RAD51 (AT5G20850), essential for meiotic repair of DSBs caused by AtSPO11-1 (Li *et al.*, 2004), DMC1 (AT3G22880), known to promote interhomolog recombination, SMC6A (At5G07660) and SMC6B (At5G61460), two components of the SMC5/6 complex, engaged in DNA repair, meiotic synapsis, genome organization, and stability. Previously it was shown that high temperatures disturb genome integrity by causing strand breakages and impending DNA repair (Kantidze *et al.*, 2016) and crosstalk between heat stress and genotoxic stress in *Arabidopsis* has been demonstrated Han *et al.* (2020). We speculate that because of reduced expression of genes encoding components of the DNA repair mechanism, the *knl2* mutant cannot cope with heat-induced DNA damage as efficiently as the wild-type and therefore increased mitotic and meiotic defects in *knl2* after exposure to high temperature has been detected (Figure 10).

In principle, we can expect that the cenH3 RNAi transformants with strongly reduced cenH3 levels can also work as efficient haploid inducers (Lermontova *et al.*, 2011). However, subjecting cenH3 RNAi transformants to heat stress did not result in haploid formation when crossed with the untreated wild-type. Previously, it has been shown that the level of cenH3 in the cenH3 RNAi transformants was strongly reduced in leaves than in root tips enriched in meristematic cells (Lermontova *et al.*, 2011). Based on previously published data, we hypothesized that this may be due to a decreased activity of the CaMV 35S promoter (Holtorf *et al.*, 1995) and suppression of post-translational gene silencing in meristems, which may cause the ineffective function of the RNAi machinery in these tissues (Mitsuhara *et al.*, 2002). Using maize cenH3 RNAi lines complemented by the *AcGREEN-tail swap-CENH3*, Kelliher *et al.* (2016) have demonstrated that in crosses with wild-type these lines can generate 0.24% maternal and 0.07% paternal haploids.

Thus, the method of obtaining haploid inducers through reduction of cenH3 levels in plants by expressing cenH3 RNAi constructs appeared to be inefficient.

The introduction of point mutations into cenH3 and the conserved CENPC-k motif of KNL2 has shown to be sufficient to generate haploid inducer lines. Thus, *knl2* and *cenh3* mutants for crop species can be obtained via the chemical ethyl methanesulfonate (EMS) mutagenesis to avoid using transgenic plants at all steps of haploid production. Alternatively, mutants can be produced by targeted mutagenesis using the CRISPR-Cas9 approach. In either case, complementation with altered cenH3 variants is not required, making the production of haploid inducers much easier. Moreover, under standard growth conditions, the growth rate of the *cenh3-4* mutant is similar to that of the wild-type, while the growth rate of *knl2* is slightly reduced. Thus, we believe that obtaining vigorous haploid inducers and short-term exposure of them to heat stress before crossing with the wild-type has great potential for application in plant breeding.

Acknowledgements

This research was supported by the German Federal Ministry of Education and Research (Plant 2030, Project 031B0192NN, HaploTools), by the Deutsche Forschungsgemeinschaft (LE2299/3-1; LE2299/5-1) and by the European Regional Development Fund-Project 'REMAP' (No. CZ.02.1.01/0.0/0.0/15_003/0000479) to KR. We thank Heike Kuhlmann and Oda Weiss for excellent technical assistence. We also acknowledge the Plant Sciences Core Facility of CEITEC MU for support with plant cultivation.

References

Allshire, R.C. (1997) Centromeres, checkpoints and chromatid cohesion. *Current opinion in genetics & development*, **7**, 264-273.

Armstrong, S.J., Sanchez-Moran, E. and Franklin, F.C. (2009) Cytological analysis of Arabidopsis thaliana meiotic chromosomes. *Methods Mol Biol*, **558**, 131-145.

Boudichevskaia, A., Houben, A., Fiebig, A., Prochazkova, K., Pecinka, A. and Lermontova, I. (2019) Depletion of KNL2 Results in Altered Expression of Genes Involved in Regulation of the Cell Cycle, Transcription, and Development in Arabidopsis. *Int J Mol Sci*, **20**.

Britt, A.B. and Kuppu, S. (2016) Cenh3: An Emerging Player in Haploid Induction Technology. *Front Plant Sci*, **7**, 357.

Capitao, C., Tanasa, S., Fulnecek, J., Raxwal, V.K., Akimcheva, S., Bulankova, P., Mikulkova, P., Crhak Khaitova, L., Kalidass, M., Lermontova, I., Mittelsten Scheid, O. and Riha, K. (2021) A CENH3 mutation promotes meiotic exit and restores fertility in SMG7-deficient Arabidopsis. *PLoS Genet*, **17**, e1009779.

Clough, S.J. and Bent, A.F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *The plant journal*, **16**, 735-743.

Galbraith, D.W., Harkins, K.R., Maddox, J.M., Ayres, N.M., Sharma, D.P. and Firoozabady, E. (1983) Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science*, **220**, 1049-1051.

Han, S.H., Park, Y.J. and Park, C.M. (2020) HOS1 activates DNA repair systems to enhance plant thermotolerance. *Nat Plants*, **6**, 1439-1446.

Hedhly, A., Hormaza, J.I. and Herrero, M. (2009) Global warming and sexual plant reproduction. *Trends Plant Sci*, **14**, 30-36.

Holtorf, S., Apel, K. and Bohlmann, H. (1995) Comparison of different constitutive and inducible promoters for the overexpression of transgenes in Arabidopsis thaliana. *Plant Mol Biol*, **29**, 637-646.

Kalinowska, K., Chamas, S., Unkel, K., Demidov, D., Lermontova, I., Dresselhaus, T., Kumlehn, J., Dunemann, F. and Houben, A. (2019) State-of-the-art and novel developments of in vivo haploid technologies. *Theor Appl Genet*, **132**, 593-605.

Kantidze, O.L., Velichko, A.K., Luzhin, A.V. and Razin, S.V. (2016) Heat Stress-Induced DNA Damage. *Acta Naturae*, **8**, 75-78.

Karimi-Ashtiyani, R., Ishii, T., Niessen, M., Stein, N., Heckmann, S., Gurushidze, M., Banaei-Moghaddam, A.M., Fuchs, J., Schubert, V., Koch, K., Weiss, O., Demidov, D., Schmidt, K., Kumlehn, J. and Houben, A. (2015) Point mutation impairs centromeric CENH3 loading and induces haploid plants. *Proc Natl Acad Sci U S A*, **112**, 11211-11216.

Kelliher, T., Starr, D., Wang, W., McCuiston, J., Zhong, H., Nuccio, M.L. and Martin, B. (2016) Maternal Haploids Are Preferentially Induced by CENH3-tailswap Transgenic Complementation in Maize. *Front Plant Sci*, **7**, 414.

Kuppu, S., Ron, M., Marimuthu, M.P.A., Li, G., Huddleson, A., Siddeek, M.H., Terry, J., Buchner, R., Shabek, N., Comai, L. and Britt, A.B. (2020) A variety of changes, including CRISPR/Cas9-mediated deletions, in CENH3 lead to haploid induction on outcrossing. *Plant Biotechnol J*.

Kuppu, S., Tan, E.H., Nguyen, H., Rodgers, A., Comai, L., Chan, S.W. and Britt, A.B. (2015) Point mutations in centromeric histone induce post-zygotic incompatibility and uniparental inheritance. *PLoS Genet*, **11**, e1005494.

Lermontova, I., Koroleva, O., Rutten, T., Fuchs, J., Schubert, V., Moraes, I., Koszegi, D. and Schubert, I. (2011) Knockdown of CENH3 in Arabidopsis reduces mitotic divisions and causes sterility by disturbed meiotic chromosome segregation. *The Plant Journal*, **68**, 40-50.

Lermontova, I., Kuhlmann, M., Friedel, S., Rutten, T., Heckmann, S., Sandmann, M., Demidov, D., Schubert, V. and Schubert, I. (2013) Arabidopsis kinetochore null2 is an upstream component for centromeric histone H3 variant cenH3 deposition at centromeres. *The Plant Cell*, **25**, 3389-3404.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987-2993.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **1**, 7.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

Li, W., Chen, C., Markmann-Mulisch, U., Timofejeva, L., Schmelzer, E., Ma, H. and Reiss, B. (2004) The Arabidopsis AtRAD51 gene is dispensable for vegetative development but required for meiosis. *Proceedings of the National Academy of Sciences*, **101**, 10596-10601.

Lv, J., Yu, K., Wei, J., Gui, H., Liu, C., Liang, D., Wang, Y., Zhou, H., Carlin, R., Rich, R., Lu, T., Que, Q., Wang, W.C., Zhang, X. and Kelliher, T. (2020) Generation of paternal haploids in wheat by genome editing of the centromeric histone CENH3. *Nat Biotechnol*, **38**, 1397-1401.

Marimuthu, M.P.A., Maruthachalam, R., Bondada, R., Kuppu, S., Tan, E.H., Britt, A., Chan, S.W.L. and Comai, L. (2021) Epigenetically mismatched parental centromeres trigger genome elimination in hybrids. *Sci Adv*, **7**, eabk1151.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, **17**, pp. 10-12.

Mitsuhara, I., Shirasawa-Seo, N., Iwai, T., Nakamura, S., Honkura, R. and Ohashi, Y. (2002) Release from post-transcriptional gene silencing by cell proliferation in transgenic tobacco plants: possible mechanism for noninheritance of the silencing. *Genetics*, **160**, 343-352.

Peet, M., Sato, S. and Gardner, R. (1998) Comparing heat stress effects on male-fertile and male-sterile tomatoes. *Plant, cell & environment*, **21**, 225-231.

Probst, A.V. and Mittelsten Scheid, O. (2015) Stress-induced structural changes in plant chromatin. *Curr Opin Plant Biol*, **27**, 8-16.

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016.

Ravi, M. and Chan, S.W. (2010) Haploid plants produced by centromere-mediated genome elimination. *Nature*, **464**, 615-618.

Ravi, M., Marimuthu, M.P., Tan, E.H., Maheshwari, S., Henry, I.M., Marin-Rodriguez, B., Urtecho, G., Tan, J., Thornhill, K., Zhu, F., Panoli, A., Sundaresan, V., Britt, A.B., Comai, L. and Chan, S.W. (2014) A haploid genetics toolbox for Arabidopsis thaliana. *Nat Commun*, **5**, 5334.

Ravi, M., Shibata, F., Ramahi, J.S., Nagaki, K., Chen, C., Murata, M. and Chan, S.W. (2011) Meiosis-specific loading of the centromere-specific histone CENH3 in Arabidopsis thaliana. *PLoS Genet*, **7**, e1002121.

Sandmann, M., Fuchs, J. and Lermontova, I. (2016) Immunolabeling of Nuclei/Chromosomes in Arabidopsis thaliana. *Methods Mol Biol*, **1370**, 127-135.

Sandmann, M., Talbert, P., Demidov, D., Kuhlmann, M., Rutten, T., Conrad, U. and Lermontova, I. (2017) Targeting of Arabidopsis KNL2 to Centromeres Depends on the Conserved CENPC-k Motif in Its C Terminus. *Plant Cell*, **29**, 144-155.

Schalch, T. and Steiner, F.A. (2017) Structure of centromere chromatin: from nucleosome to chromosomal architecture. *Chromosoma*, **126**, 443-455.

Tan, K.-K., Tan, Y.-C., Chang, L.-Y., Lee, K.W., Nore, S.S., Yee, W.-Y., Mat Isa, M.N., Jafar, F.L., Hoh, C.-C. and AbuBakar, S. (2015) Full genome SNP-based phylogenetic analysis reveals the origin and global spread of Brucella melitensis. *BMC genomics*, **16**, 1-11.

Thondehaalmath, T., Kulaar, D.S., Bondada, R. and Maruthachalam, R. (2021) Understanding and exploiting uniparental genome elimination in plants: insights from Arabidopsis thaliana. *J Exp Bot*, **72**, 4646-4662.

Verdaasdonk, J.S. and Bloom, K. (2011) Centromeres: unique chromatin structures that drive chromosome segregation. *Nature Reviews Molecular Cell Biology*, **12**, 320-332.

Wang, N. and Dawe, R.K. (2018) Centromere Size and Its Relationship to Haploid Formation in Plants. *Mol Plant*, **11**, 398-406.

Wang, N., Gent, J.I. and Dawe, R.K. (2021) Haploid induction by a maize cenh3 null mutant. *Sci Adv*, **7**.

Zhang, H., Zhang, F., Yu, Y.M., Feng, L., Jia, J.B., Liu, B., Li, B.S., Guo, H.W. and Zhai, J.X. (2020) A Comprehensive Online Database for Exploring similar to 20,000 Public Arabidopsis RNA-Seq Libraries. *Mol Plant*, **13**, 1231-1233.

Materials and Methods

Plasmid construction, plant transformation, and plant growth conditions

To analyze whether complementation of the *knl2* mutant with the genomic KNL2::KNL2-EYFP fusion construct would abolish the ability of *knl2* to induce haploids, genomic KNL2 fragment (464 up to +2513 relative to the transcriptional KNL2 start site) was amplified by PCR from Col-0 genomic DNA using KNL2-attB1gensh and KNL2-attB2 primers (Supplemental Table S2) and cloned into the pDONR221 vector. These constructs were used to generate KNL2:KNL2-EYFP fusion construct using the pGWB640 vector (https://novoprolabs.com/vector/Vgy4dmna). The substitution of conserved amino acid Trp by Arg within the CENPC-k motif of KNL2 was performed by PCR using the Phusion site-directed mutagenesis kit (Thermo Fisher Scientific). A

KNL2::KNL2-EYFP/pDONR221 construct was PCR mutagenized using the following primer pairs: KNL2gen_W_R_f and KNL2gen_W_R_r for the substitution of Trp by Arg (Supplemental Table S2). *Arabidopsis thaliana* plants were transformed according to the flower dip method (Clough and Bent, 1998). T1 transformants were selected on Murashige and Skoog medium containing 20 mg L-1 phosphinothricin (PPT). The plants were propagated under short- or long-day conditions in a cultivation room at 8 h light/20°C:16 h dark/18°C and 16 h light/20°C:8 h dark/18°C, respectively.

Immunostaining and microscopy analysis of fluorescent signals

Immunostaining of nuclei/chromosomes was performed as described previously (Sandmann *et al.*, 2016). Wide-field fluorescence microscopy was used to evaluate and image the nuclei preparations with an Olympus BX61 microscope (Olympus, Tokio, Japan) and an ORCA-ER CCD camera (Hamamatsu, Japan). For the life cell imaging, *Arabidopsis* seeds of lines harbouring mutagenized KNL2::KNL2-EYFP/pGWB640 variants or non-mutagenized control were germinated in agar medium in coverslip chambers (Nalge Nunc). Roots growing parallel to the coverslip bottom were analyzed in an LSM 510META confocal microscope (Carl Zeiss) using a 63x oil immersion objective (NA 1.4). EYFP was excited with a 488-nm laser line and fluorescence was recorded with a 505- to 550-nm band-pass filter. Images were analyzed with the LSM software release 3.2.

Whole-mount preparation

Siliques of different developmental stages were fixed in ethanol-acetic acid (9:1) overnight at 4°C and dehydrated in 70% and 90% ethanol, for 1 h each. The preparation was then cleared in chloral hydrate (chloral hydrate:water:glycerol=8:2:1) overnight at 4°C. Seeds in siliques were counted under a binocular (Carl Zeiss, Germany).

Cytogenetic techniques

A. thaliana inflorescences were fixed in freshly prepared ice-cold ethanol: acetic acid 3:1 and stored at 4°C for chromosome preparations by spreading technique (Armstrong *et al.*, 2009). After cell wall digestion, individual buds were dissected on slides, treated with 60% acetic acid, and spread for 30 seconds on a hot plate at 45°C stirring the meiocytes suspension. Post-fixation was done by applying ice-cold 3:1. Air-dried slides were counterstained with DAPI and mounted in Vectashield. Images were acquired in a Nikon Eclipse Ni equipped with a Nikon DS-Qi2 camera and a NIS Elements v. 4.60 software.

Single Nucleotide Polymorphism (SNP) analysis

Next-Generation Sequencing of genomic DNA was carried out by Eurofins Genomics Europe Shared Services GmbH (Konstanz, Germany) using a Genome Sequencer Illumina NovaSeq 6000 Sequencing System with approximately $5x10^6$ reads for each sample. After adapter trimming with cutadapt (Martin, 2011) version 1.15, next-generation sequence reads were aligned to the TAIR10 assembly with minimap2 (Li, 2018) version 2.17. Alignment records were converted to BAM format with SAMtools (Li *et al.*, 2009) and sorted with Novosort (http://www.novocraft.com/products/novosort/). SNP calling was done with BCFtools (Li, 2011) version 1.9 (command 'mpileup' and 'call') using the parameters '-a DP,DV' to record allelic depths. Only reads with a mapping quality >= Q20 were considered for variant calling. Allelic depths for each sample at bi-allelic SNP sites with a quality score >= Q40 were written to tabular format and read into R (R Core Team, 2017) for further processing. Homozygous genotypes call for the reference (alternative) allele were made if <= 10 % (>= 90 %) of reads supported the variant allele. Heterozygous calls were made if 40-60 % of reads supported the variant allele. If allelic ratios were outside these ranges or the total read depth was < 5, genotype calls were set to missing. SNPs at which the parents carried opposite homozygous alleles were selected to plot graphical genotypes of the progeny along the genome.

Gene expression analyses

The transcriptome data was retrieved from the *Arabidopsis* RNA-seq Database available on (http://ipf.sustech.edu.cn/pub/athrdb/) (Zhang *et al.*, 2020). The gene expression profiles of cenH3, CENP-C, KNL2, and NASP were extracted for different stress treatments including temperature and light stress. Down-regulated treatments among these genes were used for comparative co-expression analysis.

Flow cytometric ploidy measurements of seeds

To measure the ploidy of seeds, six seeds per pool were chopped together in 500 µl nuclei isolation buffer (Galbraith *et al.*, 1983) supplemented with propidium iodide (50 µg/ml) and DNase-free RNase (50µg/ml) in a Petri dish using a sharp razor blade. The resulting nuclei suspensions were filtered a 50 µm mesh (CellTrics, Sysmex-Partec) and measured on a CyFlow Space flow cytometer (Sysmex-Partec), a FACSAria cell sorter (BD Biosciences), or an Influx cell Sorter (BD Biosciences). The presence of haploid/aneuploid seeds within the pools was determined by evaluating the PI fluorescence intensity on a linear scale. Since the precise number of haploids/aneuploids per pool cannot be determined unequivocally, we considered only one seed per pool as being deviating from the diploid status if in addition peak was found. Thus, the number of haploids/aneuploids is an underestimation rather than an overestimation.

8 LIST OF ABBREVIATIONS

bp	Base pair
BAMM	Bayesian analysis of macroevolutionary mixtures
BI	Bayesian inference
BiSSE	Binary state speciation and extinction
ChIP	Chromatin immunoprecipitation
cp	Chloroplast
DAPI	4′,6-diamidino-2-phenylindole
DNA	Deoxyribonucleic acid
FISH	Fluorescent *in situ* hybridization
GMO	Genetically modified organisms
GS	Genome size
HGP	Human genome project
Hi-C	High-throughput chromosomal conformation capture
IGS	Intergenic spacer
INT	Integrase
IR	Inverted repeat
ITR	Interstitial telomeric repeat
ITS	Internal transcribed spacer
KNL2	KINETOCHORE NULL2
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
Mb	Megabase pairs
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
mtDNA	Mitochondrial DNA
NGS	Next-generation sequencing
NJ	Neighbor-joining
NOR	Nucleolar organizer region
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PBS	Primer binding site
PCG	Protein-coding gene

PPD	Post-polyploid diploidization
PPT	Polypurine tract
PROT	Proteinase
RH	Ribonuclease H
RT	Reverse transcriptase
SINE	Short interspersed nuclear element
SSC	Small single copy
SSR	Simple sequence repeat
TAREAN	Tandem repeat analyzer
TE	Transposable element
TRF	Tandem repeat finder
WGD	Whole genome duplication
WT	Wild-type

9 CURRICULUM VITAE

Sheng Zuo (左 胜)

Email: sheng.zuo@ceitec.muni.cz; shengzuo@outlook.com

Kamenice 5, CEITEC, Masaryk University, 62500 Brno, Czech Republic

ORCID: 0000-0002-2104-3726

EDUCATION

2018 - 2022 Ph.D. in Bio-omics, CEITEC and Faculty of Science, Masaryk University, Brno, Czech Republic, Advisor: Prof. Martin A. Lysak; Dr. Inna Lermontova

2015 - 2018 M.S. in Biochemistry and Molecular Biology, Fujian Agriculture and Forestry University, Fujian, China, Advisor: Prof. Kai Wang

2011 - 2015 B.S. in Biotechnology, Wenzhou University, Zhejiang, China

RESEARCH INTERESTS

Genome and Repeatome Evolution; Centromere and Kinetochore; Plant Systematics; Genomics and Phylogenetics

BACKGROUND AND SKILLS

Working with Nanopore, PacBio and Illumina sequencing data during Ph.D. studies; Working with DNase-Seq and ChIP-Seq data during master studies; Used to work with Linux in a cluster environment; Using Python and R for scripting

SELECTED PUBLICATIONS

5. **Sheng Zuo**[#], Ramakrishna Yadala[#], Fen Yang, Paul Talbert, Joerg Fuchs, Veit Schubert, Ulkar Ahmadli, Ales Pecinka, Martin A. Lysak, Inna Lermontova. Recurrent plant-specific duplications of KNL2 and its conserved function as kinetochore assembly factor. **Molecular Biology and Evolution**. 2022 https://doi.org/10.1093/molbev/msac123 (IF = 16.24)

4. **Sheng Zuo**#, Xinyi Guo#, Terezie Mandáková, Mark Edginton, Ihsan A. Al-Shehbaz, Martin A. Lysak. Genome diploidization associates with cladogenesis, trait disparity and plastid gene evolution. **Plant Physiology**. 2022 https://doi.org/10.1093/plphys/kiac268 (IF = 8.34)

3. **Sheng Zuo**, Terezie Mandáková, Michaela Kubová, Martin A. Lysak. Genomes, repeatomes and interphase chromosome organization in the meadowfoam family (Limnanthaceae, Brassicales). **The Plant Journal**. 2022, 110:1462-1475. (IF = 6.417)

2. Yinjia Li#, **Sheng Zuo**#, Zhiliang Zhang#, Zhanjie Li, Jinlei Han, Zhaoqing Chu, Robert Hasterok, Kai Wang. Centromeric DNA characterization in the model grass *Brachypodium distachyon* provides insights on the evolution of the genus. **The Plant Journal**. 2018, 93:1088-1101. (co-first author)

1. Wenpan Zhang#, **Sheng Zuo**#, Zhanjie Li, Zhuang Meng, Jinlei Han, Junqi Song, Yongbao Pan, Kai Wang. Isolation and characterization of centromeric repetitive DNA sequences in *Saccharum spontaneum*. **Scientific Reports**. 2017, 7: 41659. (co-first author)

PRESENTATIONS

6. Poster presentation, ELIXIR CZ Annual Conference 2022: Plant Cytogenomics, September 2022

5. Oral presentation, 2022 CEITEC Ph.D. Conference, Masaryk University, May 2022

4. Oral presentation, 7th European Workshop on Plant Chromatin 2022, May 2022

3. Poster presentation, International Plant Systems Biology 2021, EMBO workshop, April 2021

2. Oral presentation, Chromosome Biology Retreat 2020, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), November 2020

1. Poster presentation, Graduate Academic Forum of Plant Science, Institute of Botany, Chinese Academy of Sciences, August 2018

Professional Experience

4. 1/4/2022 – 30/4/2022. Ph.D. Internship. Domestication Genomics Group, Leibniz Institute of Plant Genetics and Crop Plant Research, Germany. Advisor: Dr. Martin Mascher

3. 1/3/2022 – 30/3/2022. Ph.D. Internship. Kinetochore Biology Group, Leibniz Institute of Plant Genetics and Crop Plant Research, Germany. Advisor: Dr. Inna Lermontova
2. 21/5/2019 – 23/5/2019. 8th RepeatExplorer Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis, Institute of Plant Molecular Biology, České Budějovice, Czech Republic
1. 1/2015 – 5/2015. Graduate trainee, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China

AWARDS &FELLOWSHIPS

2018-2022	Doctoral scholarship from Ministry of Education, Youth and Sports, Czech Republic
2018-2022	CEITEC PhD School scholarship, Masaryk University
2017-2018	Extraordinary scholarship of the concept of Life Sciences Program for outstanding creative activity, Fujian Agriculture and Forestry University
2015-2016	Graduate student study fellowship, Fujian Agriculture and Forestry University

RESEARCH SUPPORT

2020-2022 Czech Science Foundation (GACR): Post-polyploid genome evolution in Microlepidieae species (Brassicaceae). Participant

Other Activity

07/2012-08/2012 Volunteer activity Team Leader

Organizing a team went to autistic children's school as a volunteer

02/2013-02/2014 Vice Chairman of the Student Union of the College

Responsible for schedule of the Student Union; Organizing college debates events

Languages: Chinese (native), English (fluent)