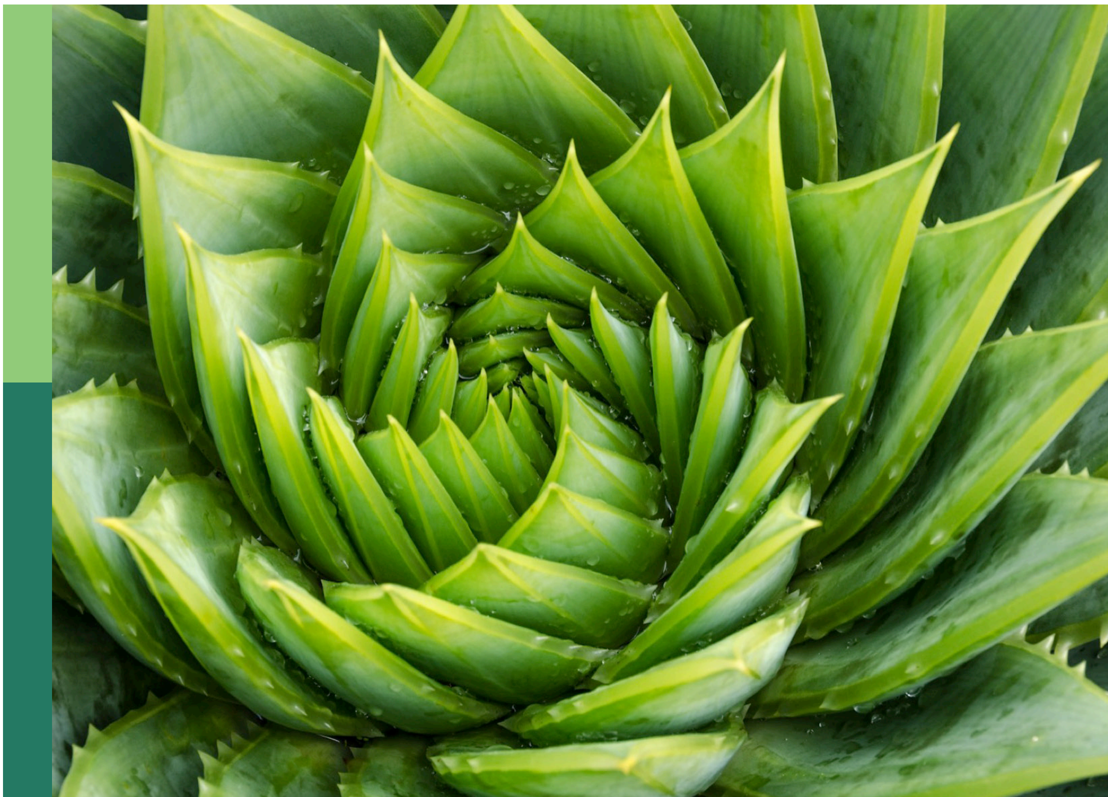
frontiers | Research Topics

Mobile elements and plant genome evolution, comparative analyses and computational tools, volume II

Edited by
Ruslan Kalendar and Gennady I. Karlov

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-4016-9
DOI 10.3389/978-2-8325-4016-9

About Frontiers

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

Frontiers journal series

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

Dedication to quality

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

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Mobile elements and plant genome evolution, comparative analyses and computational tools, volume II

Topic editors

Ruslan Kalendar — University of Helsinki, Finland

Gennady I. Karlov — All-Russia Research Institute of Agricultural Biotechnology, Russia

Citation

Kalendar, R., Karlov, G. I., eds. (2023). *Mobile elements and plant genome evolution, comparative analyses and computational tools, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4016-9

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 10 November 2023
DOI 10.3389/fpls.2023.1308536

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Atsushi Fukushima,
Kyoto Prefectural University, Japan

*CORRESPONDENCE
Ruslan Kalendar
ruslan.kalendar@helsinki.fi

RECEIVED 06 October 2023
ACCEPTED 07 November 2023
PUBLISHED 10 November 2023

CITATION
Kalendar R and Karlov GI (2023) Editorial: Mobile elements and plant genome evolution, comparative analyses and computational tools, volume II. *Front. Plant Sci.* 14:1308536. doi: 10.3389/fpls.2023.1308536

Editorial: Mobile elements and plant genome evolution, comparative analyses and computational tools, volume II

Ruslan Kalendar[1,2*] and Gennady I. Karlov[3]

[1]Institute of Biotechnology, Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland, [2]National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan, [3]All-Russia Research Institute of Agricultural Biotechnology, Russian Academy of Sciences, Moscow, Russia

KEYWORDS

plant genome evolution, interspersed repetitive sequences, transposable element, retrotransposon, endogenous viral element, repeatome-based molecular markers

Editorial on the Research Topic
Mobile elements and plant genome evolution, comparative analyses and computational tools, volume II

Mobile element and host genome evolution

The genomes of eukaryotes are mostly composed of diverse families of interspersed repetitive sequences, including retrotransposons and transposable and endogenous viral elements. The prevailing view is that the diverse families of the genome repeatome should be considered only as parasites or “junk DNA” (Bourque et al., 2018). However, it is possible to follow genealogical trees, or pathways of evolutionary development and distribution of these elements, due to which, our understanding should be completely revised. The repeatome elements play a role that, in the sense of systems biology and medicine, goes far beyond “junk DNA” and viral fossils (Wells and Feschotte, 2020). Recent studies increasingly show that essential components, if not the most basic components of our genome, are of viral origin and that viruses as mobile genetic mediators have always played a crucial role in genetic evolution (Cosby et al., 2019). The evolution of genomes is associated with overcoming and fixing integrated events. With each important evolutionary step, the number of mobile genetic elements in the genome increased dramatically. Since the beginning of life, there has not been an organism that did not contain all these diverse mobile elements. In the formation of the genome, we can trace numerous processes involving mobile elements with their countless different appearances. Genomes are not the end product of innumerable accidental mutations and their selection, but a kind of living deposit from originally external, viral influences that is constantly being recycled and, like a chronicle, reinterpreted (Vassilieff et al., 2023). To be able to develop at all, mobile elements must have a coevolutionary relationship with their host genome (Gebrie, 2023). Evolutionary phylogenetic trees of mobile elements and the host genome show strong correlations (Kalendar et al., 2004; Kalendar et al., 2008; Moisy et al., 2014; Kalendar et al., 2020). Endogenous retroviruses, to which retrotransposons also belong, are single-stranded enveloped RNA viruses that are characterized by the fact that their genetic information, by means of reverse transcriptase, is rewritten into the DNA of the host genome and thus multiplies with each cell cycle (Johnson, 2019). If these retrotransposons enter the germline directly, they are not only passed on with each cell division but are also inherited and remain an integral part of the species genome. Retrotransposons inhabit almost all eukaryotic organisms without exception, and they can probably even be found as part of the genome of giant viruses. In most cases, retrotransposon-related elements live in the host genome and help the host resist infections and various forms of stress (Lanciano and Mirouze, 2018). Mobile genetic elements are crucial for species diversity and the evolution of the host genome. New genetic combinations always reveal the basis for new things for adaptive and developmental processes (Klein and Anderson, 2022). Diversity is always an indicator of vital and healthy ecosystems; this is true for the genome with its numerous families of interspersed repetitive elements. Thus, diverse families of repetitive elements are, genetically, one of the decisive factors in evolutionary innovation and species diversity.

We continued in the Research Topic "*Mobile Elements and Plant Genome Evolution, Comparative Analyses, and Computational Tools II*" to explore the effectiveness of new genomic tools to detect repetitive elements and highlighted some recent studies on the role of repetitive elements in host genome evolution, comparative analysis, and genome-wide profiling of retrotransposons and transposable and endogenous viral elements (Kalendar et al., 2021).

Plant genome evolution has mainly been determined by polyploidization and amplification or loss of retrotransposon-related elements. Research conducted by Mascagni et al. revealed that repetitive DNA within the *Olea* taxa constitutes a significant portion (ranging from 59% to 73%) of the total genome. This finding showcases substantial variations in terms of composition among these taxa. Notably, an intriguing observation emerged, namely the abundance of tandem repeats exhibited an inverse correlation with retrotransposons. For example, *Olea paniculata*, closest to *O. ancestor*, has few tandem repeats but abundant long terminal repeat retrotransposons, suggesting tandem repeat expansion post-divergence. This research unveiled the temporal dynamics that have played a pivotal role in shaping the genome structure throughout *Olea* speciation. This also provides a unique and insightful model for understanding the evolution of genomes in higher plants.

The genome of *Humulus scandens*, which is an important dioecious plant with XX/XY1Y2 chromosomes, was annotated with the repetitive portion of both the male and female genomes and compared with the different groups of repetitive sequences among the male and female genomes and with a close relative, *H. lupulus*. Zhang et al. analyzed the distribution of retrotransposons and satellite DNAs and determined the orientation position of the pseudoautosomal regions and indicated that the XX-XY1Y2 sex chromosomes of *H. scandens* might have originated from a centric fission event. Thus, this study revealed the nature of the origin and evolution of the sex chromosome of *H. scandens*.

Discovery and comparative analysis of transposable elements

Endogenous viral elements (EVRs) are derived from DNA viruses of the family Caulimoviridae and abundant in plant genomes. de Tomás and Vicient analyzed 278 genome assemblies corresponding to 267 plant species to identify conserved domains of the reverse transcriptase of *Caulimoviridae*. These discovered EVRs were grouped in 57 clusters and classified in 13 genera, including a newly proposed genus *Wendovirus*. Comparing plant genomes, important differences between the plant families and genera in the number and type of endogenous pararetrovirus were found. In general, florendoviruses are the most abundant and widely distributed endogenous pararetrovirus.

The cold seasonal *Loliinae* subtribe includes taxa distributed worldwide and has a striking two-fold difference in genome size between the broad-leaved and fine-leaved *Loliinae* diploids and a general trend of genome reduction of some high polyploids. Moreno-Aguilar et al. used genome skimming data to uncover the composition, abundance, and potential phylogenetic signal of repetitive elements across 47 representatives of the main *Loliinae* lineages. The evolution of the *Loliinae* repeatome suggests a plausible scenario of recurrent allopolyploidizations followed by diploidizations that generated the large genome sizes of broad-leaved diploids and large genomic rearrangements in highly hybridogenous lineages that caused massive repeatome and genome contractions in the *Schedonorus* and *Aulaxyper* polyploids.

Genome-wide profiling for transposable element analysis of repetitive elements

Interspersed repetitive elements are ideal for studying genetic variability in the genome and are crucial for studying the evolution of the host genome. Therefore, diverse high-throughput genotyping and sequencing applications have been developed. Arvas et al. described the main trends on promising directions of molecular marker technologies directly related to deployment of high-throughput genotype sequencing platforms.

The EG4 strain of rice is a unique material in that the transposon mPing has high transpositional activity and high copy numbers under natural conditions. Monden et al. identified the candidate genes and transposon mPing insertion sites that drive the high protein content of rice. The identified high-protein lines can lead to development of rice cultivars by introducing valuable traits, such as high and stable yield, disease resistance, and rich nutrient content.

Bread wheat genome evolution is largely dependent on a large number of diverse families of transposable elements (TE), which constitute approximately 80% of the genome. Bariah et al. found that about 36% of the 70 818 genes in bread wheat contained at least one TE insertion within the gene body, mostly in triads. TE insertions within the exon or in the untranslated regions of one or more of the homoeologs in a triad were significantly associated with homoeolog expression bias. A significant association was observed between the presence of TE insertions from specific superfamilies and the expression of genes associated with biotic and abiotic stress responses.

Ubi et al. studied 52 miniature inverted-repeat transposable element (MITE) insertion polymorphism markers for genetic studies in wheat and related species. Phylogenetic analysis of these MITEs insertions were consistent with the evolutionary history of these wheat species, which clustered mainly according to ploidy and genome types (SS, AA, DD, AABB, and AABBDD). The MITE insertion site polymorphisms uncovered in this study are very promising as high-potential evolutionary markers for genomic studies in wheat.

Bioinformatic tools

To study and identify interspersed repetitive sequences and endogenous viral elements, specialized databases of these elements and bioinformatics tools for their *de novo* identification are needed.

A review of strategies used to identify transposition events in plant genomes is presented in a paper by Bajus et al. The authors described the basis of their operational principles to capture real cases of actively transposing elements and conceivable strategies. Combinations of methods resulting in improved performance are also proposed.

Argentin et al. performed a comparative analysis and classification of transposable element distribution in all plant species available in Ensembl plant genomes browser. The new classification of transposable elements was used to comparatively analyze the distribution of repetitive elements in 53 species (Wicker et al., 2007).

Mokhtar et al. developed the PlantLTRdb database containing retrotransposon sequences for 195 plant species. PlantLTRdb allows researchers to search, visualize, and analyze plant retrotransposons. PlantLTRdb can contribute to the understanding of structural variations, genome organization, functional genomics, and development of transposable elements targeting markers for molecular plant breeding.

Conclusion

The prospects and challenges facing the exploration of repetitive DNA sequences and their role in genome evolution are both promising and intricate. Recent research has illuminated the pivotal role of repetitive elements in shaping evolution, driving genetic diversity, and regulating gene expression. However, the origins of transposable elements and their influence on genome evolution remain a significant puzzle in the realm of evolutionary biology. The co-evolutionary relationship between transposable elements and their host genomes stands as a key driver of genome size evolution, with the dynamics of this interplay potentially governing genome expansion and contraction. In-depth molecular studies have underscored the functional significance of repetitive elements, highlighting their necessity in orchestrating the expression of unique coding sequences and organizing essential functions crucial for genome operation. The repetitive genome component assumes a prominent architectural role in structuring higher-order genomic organization, while repetitive elements serve as invaluable tools for deciphering comparisons between sequenced genomes. The investigation of repetitive DNA sequences and their role in genome evolution is an intricate and ongoing discipline, offering tremendous potential for unravelling the origins and evolution of life on our planet.

Author contributions

RK: Conceptualization, Funding acquisition, Investigation, Writing – original draft, Writing – review & editing. GK: Investigation, Resources, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (AP19678041) for RK.

Acknowledgments

We thank all authors and reviewers for their contributions to this Research Topic and the support of the editorial office.

Conflict of interest

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

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., et al. (2018). Ten things you should know about transposable elements. *Genome Biol.* 19, 199. doi: 10.1186/s13059-018-1577-z

Cosby, R. L., Chang, N. C., and Feschotte, C. (2019). Host-transposon interactions: conflict, cooperation, and cooption. *Genes Dev.* 33, 1098–1116. doi: 10.1101/gad.327312.119

Gebrie, A. (2023). Transposable elements as essential elements in the control of gene expression. *Mob DNA* 14, 9. doi: 10.1186/s13100-023-00297-3

Johnson, W. E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat. Rev. Microbiol.* 17, 355–370. doi: 10.1038/s41579-019-0189-2

Kalendar, R., Raskina, O., Belyayev, A., and Schulman, A. H. (2020). Long tandem arrays of Cassandra retroelements and their role in genome dynamics in plants. *Int. J. Mol. Sci.* 21, 2931. doi: 10.3390/ijms21082931

Kalendar, R., Sabot, F., Rodriguez, F., Karlov, G. I., Natali, L., and Alix, K. (2021). Editorial: mobile elements and plant genome evolution, comparative analyzes and computational tools. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.735134

Kalendar, R., Tanskanen, J., Chang, W., Antonius, K., Sela, H., Peleg, O., et al. (2008). Cassandra retrotransposons carry independently transcribed 5S RNA. *Proc. Natl. Acad. Sci. U S A* 105, 5833–5838. doi: 10.1073/pnas.0709698105

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166, 1437–1450. doi: 10.1534/genetics.166.3.1437

Klein, S. P., and Anderson, S. N. (2022). The evolution and function of transposons in epigenetic regulation in response to the environment. *Curr. Opin. Plant Biol.* 69, 102277. doi: 10.1016/j.pbi.2022.102277

Lanciano, S., and Mirouze, M. (2018). Transposable elements: all mobile, all different, some stress responsive, some adaptive? *Curr. Opin. Genet. Dev.* 49, 106–114. doi: 10.1016/j.gde.2018.04.002

Moisy, C., Schulman, A. H., Kalendar, R., Buchmann, J. P., and Pelsy, F. (2014). The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years. *Theor. Appl. Genet.* 127, 1223–1235. doi: 10.1007/s00122-014-2293-z

Vassilieff, H., Geering, A. D. W., Choisne, N., Teycheney, P. Y., and Maumus, F. (2023). Endogenous caulimovirids: fossils, zombies, and living in plant genomes. *Biomolecules* 13(7), 1069. doi: 10.3390/biom13071069

Wells, J. N., and Feschotte, C. (2020). A field guide to eukaryotic transposable elements. *Annu. Rev. Genet.* 54, 539–561. doi: 10.1146/annurev-genet-040620-022145

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

frontiers
in Plant Science

ORIGINAL RESEARCH
published: 31 March 2022
doi: 10.3389/fpls.2022.869048

The Singular Evolution of *Olea* Genome Structure

Flavia Mascagni[1], Elena Barghini[1], Marilena Ceccarelli[2], Luciana Baldoni[3], Carlos Trapero[4,5], Concepción Muñoz Díez[5], Lucia Natali[1], Andrea Cavallini[1] and Tommaso Giordani[1]*

[1]Department of Agriculture, Food and Environment, University of Pisa, Pisa, Italy, [2]Department of Chemistry, Biology and Biotechnology, University of Perugia, Perugia, Italy, [3]CNR, Institute of Biosciences and BioResources, Perugia, Italy, [4]CSIRO Agriculture & Food, Narrabri, NSW, Australia, [5]Agronomy Department, University of Cordoba, Cordoba, Spain

OPEN ACCESS

Edited by:
Ruslan Kalendar,
University of Helsinki, Finland

Reviewed by:
Francisco Luque,
University of Jaén, Spain
Miroslav Plohl,
Rudjer Boskovic Institute, Croatia
Matej Lexa,
Masaryk University, Czechia

****Correspondence:***
F. Mascagni
flavia.mascagni@unipi.it

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

Received: *03 February 2022*
Accepted: *07 March 2022*
Published: *31 March 2022*

Citation:
Mascagni F, Barghini E, Ceccarelli M, Baldoni L, Trapero C, Díez CM, Natali L, Cavallini A and Giordani T (2022) The Singular Evolution of Olea Genome Structure.
Front. Plant Sci. 13:869048.
doi: 10.3389/fpls.2022.869048

The current view of plant genome evolution proposes that genome size has mainly been determined by polyploidisation and amplification/loss of transposons, with a minor role played by other repeated sequences, such as tandem repeats. In cultivated olive (*Olea europaea* subsp. *europaea* var. *europaea*), available data suggest a singular model of genome evolution, in which a massive expansion of tandem-repeated sequences accompanied changes in nuclear architecture. This peculiar scenario highlights the importance of focusing on *Olea* genus evolution, to shed light on mechanisms that led to its present genomic structure. Next-generation sequencing technologies, bioinformatics and *in situ* hybridisation were applied to study the genomic structure of five related *Olea* taxa, which originated at different times from their last common ancestor. On average, repetitive DNA in the *Olea* taxa ranged from ~59% to ~73% of the total genome, showing remarkable differences in terms of composition. Among repeats, we identified 11 major families of tandem repeats, with different abundances in the analysed taxa, five of which were novel discoveries. Interestingly, overall tandem repeat abundance was inversely correlated to that of retrotransposons. This trend might imply a competition in the proliferation of these repeat classes. Indeed, *O. paniculata*, the species closest to the *Olea* common ancestor, showed very few tandem-repeated sequences, while it was rich in long terminal repeat retrotransposons, suggesting that the amplification of tandem repeats occurred after its divergence from the *Olea* ancestor. Furthermore, some tandem repeats were physically localised in closely related *O. europaea* subspecies (i.e., cultivated olive and *O. europaea* subsp. *cuspidata)*, which showed a significant difference in tandem repeats abundance. For 4 tandem repeats families, a similar number of hybridisation signals were observed in both subspecies, apparently indicating that, after their dissemination throughout the olive genome, these tandem repeats families differentially amplified maintaining the same positions in each genome. Overall, our research identified the temporal dynamics shaping genome structure during *Olea* speciation, which represented a singular model of genome evolution in higher plants.

Keywords: *Olea* evolution, tandem repeats, retrotransposons, genome landscape, NGS analyses, genome evolution

INTRODUCTION

The current view of plant genome evolution proposes that genome size is determined by processes such as polyploidisation and amplification/loss of transposable elements (TEs), especially retrotransposons (REs; Proost et al., 2011; Catlin and Josephs, 2022). The genome of most plant clades has been shaped during evolution by many polyploidisation events, with each new episode superimposed on genomic remnants from earlier rounds of duplication. At the same time, the bulk of non-coding DNA in plant genomes consists of active, silenced or degenerating mobile elements, which vary widely in composition and abundance among populations (Garrido-Ramos, 2015; Wendel et al., 2018).

Mobile elements can affect genomes either during mobilisation events or after their insertion. Mobilisation of a TE and its insertion within the coding sequence of a gene, or nearby the promoter, can lead to a loss of function or altered expression of that gene (Dubin et al., 2018). Furthermore, TE proliferation, or loss, produces changes in genome size. Notable examples are *Oryza australiensis*, where amplification of specific retrotransposon lineages has led to the doubling of its genome size within the last 3 million years (Piegu et al., 2006), and the legume tribe *Fabeae*, where genome dynamics, are dominated by a single lineage of REs that accounts for 57% of the variation in genome size in this clade (Macas et al., 2015). The impact of TEs on the genomic landscape continues after insertion, contributing to the organisation of the genome through epigenetic regulation (Lippman et al., 2004; Hollister and Gaut, 2009; Usai et al., 2021), or by still affecting gene expression after becoming transcriptionally inactive (Marcon et al., 2015; Sanseverino et al., 2015).

Transposable elements are classified into two different classes, according to whether their transposition intermediate is RNA (Class I or REs) or DNA (Class II or DNA transposons; Wicker et al., 2007). In plants, REs are the most common class of elements, representing the core of many genomes (Lisch, 2013; Vitte et al., 2014), and are further classified into five taxonomic orders (Wicker et al., 2007). The most abundant REs in plants, long terminal repeat retrotransposons (LTR-REs), are organised into two major superfamilies, called *Gypsy* and *Copia*, which differ in the position of protein domains within their encoded polyprotein (Wicker et al., 2007). In turn, the superfamilies can be classified into several major evolutionary lineages (Wicker and Keller, 2007; Llorens et al., 2010), seven lineages for *Copia* and three main lineages for *Gypsy* (Buti et al., 2017; Neumann et al., 2019; Vangelisti et al., 2019; Mascagni et al., 2020).

Other types of repeated sequences generally have a minor role in shaping plant genome structure and size, accounting for a small portion of the genomes sequenced so far. Among these, tandem repeats (TRs) are arranged in tandem repeating units, where individual copies lie adjacent to one another, and usually show different GC content from the rest of the genomic DNA (Szybalski, 1968).

Precise molecular mechanisms leading to TR proliferation in individual species and/or to their rapid turnover have not yet been clearly identified. Several mechanisms have been proposed for the generation of short arrays of TRs, including unequal crossing over of random sequences (Smith, 1976), slipped-strand mispairing (Levinson and Gutman, 1987) and sequence-directed mutagenesis (Fieldhouse and Golding, 1991). In addition, tandem duplications of varying length can also result from aberrant replication and replication stress (Mazurczyk and Rybaczek, 2015; Nikolov and Taddei, 2016).

Initially isolated from satellite bands in gradient centrifugation experiments, TRs are commonly known as satellite DNA (Schmidt and Heslop-Harrison, 1998). Satellite arrays are generally found in heterochromatic regions and may form essential chromosome structures such as centromeres and telomeres (Garrido-Ramos, 2017; Hartley and O'Neill, 2019). Apart from their common key role in these critical structures, TR families are characterised by a huge variety of sequences (Melters et al., 2013) differing in location, repeat unit length and abundance, suggesting they undergo rapid evolution (Thakur et al., 2021). Being one of the most dynamic components of eukaryotic genomes, most satellite repeat families are usually species- or genus-specific (Garrido-Ramos, 2015).

On the other hand, evidence of sequence conservation of satellite families for long evolutionary periods among species has also been reported (Quesada del Bosque et al., 2013, 2014; Cafasso and Chinali, 2014; Mehrotra et al., 2014), supporting the hypothesis of a possible functional role for these sequences in the genomes (Pezer et al., 2012; Plohl et al., 2012). Therefore, related species may share an ancestral set of satellite families with specific levels of conservation and amplification.

In the cultivated olive (*Olea europaea* subsp. *europaea* var. *europaea*), available data suggest a singular model of genome evolution, in which polyploidisation and amplification/loss of TEs were accompanied by a massive expansion of the tandemly repeated fraction. As a result, TRs compose almost one-third of the current olive genome, a much larger portion than in the vast majority of plant genomes (Barghini et al., 2014).

Several studies were conducted to elucidate the TR fraction of olive, with six TR families being isolated from genomic libraries, and in some case, localised by cytological hybridisation (Katsiotis et al., 1998; Bitonti et al., 1999; Minelli et al., 2000; Lorite et al., 2001; Contento et al., 2002; Barghini et al., 2014).

A first genome sequence for *Olea europaea* subsp. *europaea* var. Farga was released in 2016 (Cruz et al., 2016) with a limited characterisation of the repeated component; then, a genome sequence and annotation of the wild olive tree (*Olea europaea* subsp. *europaea* var. *sylvestris*; Unver et al., 2017) resulted in contrast with previous studies showing a significantly lower abundance of TRs than expected. The most recent studies related to the genome of cultivated olive, although revealed a great genetic variability as result of a significant activation of TEs during the domestication process (Jiménez-Ruiz et al., 2020), made only little progress in deciphering the complex structure of its repetitive component (Rao et al., 2021).

The difficulty in identifying satellite sequences might be explained by repeat collapse, which causes common mis-assembly due to the incorrect gauging of the number of repeat copies in a genome, and ultimately providing a reference with too few repeat copies (Phillippy et al., 2008).

New possibilities for investigating repetitive sequences in genomes were provided by massive parallel DNA sequencing techniques. In fact, the use of these technologies within a computational framework led to the identification of the different types of repetitive elements, allowing us to address many features of the dynamics which have changed the repetitive component of the *Olea* genome.

In this study, we aimed at characterising the repetitive component of a range of taxa representative of the *Olea* genus, including plants from different geographical origins. We also included *O. paniculata* as representative species of the subgenus *Paniculatae*, the closest relative of the *Olea* last common ancestor. This analysis represents the most comprehensive study of the evolutionary dynamics of repetitive elements within *Olea* genus, evaluating with different methodologies (bioinformatic, cytophotometric and cytological) how the genome structure has evolved and shedding light on mechanisms of genome expansion.

MATERIALS AND METHODS

Plant Material, DNA Isolation, and Illumina Sequencing

For this study, the following species of *Olea* were chosen, *O. paniculata*, a representative of the subgenus *Paniculatae*, and four taxa of the subgenus *Olea*, *O. exasperata* (section *Ligustroides*), *O. europaea* subsp. *europaea* (cv. Leccino), *O. europaea* subsp. *cuspidata* and *O. europaea* subsp. *guanchica* (**Table 1**). Plant material (leaves and root apices, the latters collected from potted plants or cuttings) was provided by the Olive Collection of CNR—Institute of Biosciences and Bioresources, Division of Perugia (Perugia, Italy), by the IFAPA World Olive Germplasm Bank and Agronomy Department of University of Cordoba (Cordoba, Spain) and by CSIRO Agriculture & Food (Narrabri, NSW, Australia).

Genomic DNA was extracted from young leaves using a GenElute Plant Genomic DNA Miniprep kit (Sigma-Aldrich) and following the manufacturer's instructions. Paired-end libraries were prepared as recommended by Illumina Inc. (San Diego, CA), with minor modifications, and sequencing was performed for all taxa samples.

Whole-genome shotgun sequences described are available on NCBI Sequence Read Archive under the accession number SRX465835 (*O. europaea* subsp. *europaea* cv. Leccino) and BioProject PRJNA810942 for the other analysed taxa.

Paired reads were first tested for quality and trimmed at 100 nt in length, using Trimmomatic (Bolger et al., 2014) with the parameters, leading:20 trailing:20 slidingwindow:4:20 crop:100 minlen:100. Duplicated reads and those containing organelle DNA sequences were removed using CLC-BIO Genomic Workbench 9.5.3 (CLC-BIO, Aarhus, Denmark).

Repeat Characterisation From NGS Reads

In order to perform a comparative analysis of the repetitive components of five taxa of the genus *Olea*, RepeatExplorer (Novák et al., 2013), a sequence similarity-based clustering method was applied allowing *de novo* identification of repeats and an estimation of their proportion in each genome. A random set of 1,500,000 sequences was used for each species, and these were analysed individually to maximise the number of analysed reads and the sensitivity and accuracy of the repeat data obtained allowing the identification of less abundant repeat families. Because of the large amount of satellite DNA sequence recovered by the software, after preliminary analysis, a filtering of abundant satellite repeats was performed. Using custom libraries, we filtered large satellite repeats from our data to allow more reads to be analysed during repeat identification.

RepeatExplorer output was parsed to collect the clusters identified as repeats. To increase the number of annotated clusters, similarity searches on the remaining unknown clusters were performed by BLASTN and tBLASTX against a library of 254 putative full-length REs of olive (Barghini et al., 2014).

Putative satellite repeats identified *via* graph-based clustering by RepeatExplorer were collected for each species. The validation of monomer sequences of selected satellites was performed by dot plot analysis of the contigs assembled and by using tandem repeat finder (Benson, 1999) and CAP3 (Huang and Madan, 1999) tools.

TR sequences were collected per species and the database was cleaned of redundant sequences by using CD-HIT (Li and Godzik, 2006) with a threshold identity of 95%. A subset of unique sequences was also obtained after grouping the entire collection of TRs.

Mapping Procedure for Abundance Estimation

Abundance values of sequences were estimated for each taxon by counting the number of reads mapping into clusters of interspersed repeated sequences or into the library of tandem repeat sequences, per million total reads. This method had already been used for many plant species (Swaminathan et al., 2007; Tenaillon et al., 2011; Natali et al., 2013; Mascagni et al., 2015, 2017a, 2018a) including olive (Barghini et al., 2014, 2015). CLC-BIO Genomic Workbench was used to perform mapping with the following parameters: mismatch cost = 1, deletion cost = 1, insertion cost = 1, similarity = 0.7 and length fraction = 0.7.

Phylogenetic Trees

A multiple sequence alignment of the TR sequences was performed using Clustal Omega (McWilliam et al., 2013), and phylogenetic trees were built using a neighbour joining clustering method (NJ; 1,000 bootstrap replications).

A dendrogram, based on the genome proportions, using data of each isolated TR, was built by using the R package pvclust version 1.3–2 (Suzuki and Shimodaira, 2006), which allowed the assignment of the uncertainty in hierarchical cluster analysis *via* multiscale bootstrap resampling with 10,000 bootstrap replications.

RE Insertion Time Analysis

Domain-based ANnotation of Transposable Elements (DANTE) was used to identify and extract conserved regions of reverse transcriptase (RT) protein domains for *Gypsy* and *Copia* RE superfamilies. Timing of LTR-REs proliferation bursts of the analysed species was measured according to Piegu et al. (2006), Buti et al. (2011) and Mascagni et al. (2017b, 2018b), through analysis of the distribution of divergence values between pairwise comparisons of sequences belonging to the same lineage. After collecting all RT domain-related sequences from RepeatExplorer results, cluster mapping was performed using CLC-BIO Genomic Workbench to isolate reads homologous to RT for each species. Then, paralogous reads were pairwise compared using MEGA version 7 (Kumar et al., 2016) within each species and Kimura distances (Kimura, 1980) were calculated. Kimura distances were converted to times, expressed as millions of years ago (MYA), using a substitution rate of 1.3×10^{-8} defined in rice, as described by Ma and Bennetzen (2004).

Genome Size Estimation

Root apices were collected from five *O. paniculata* plants and one rooted cutting of cv. Leccino, and fixed in ethanol:acetic acid (3:1 v/v). The apices were washed in an aqueous solution of 6 mM sodium citrate, 4 mM citric acid, treated with a mixture of 8% pectinase (Sigma), 2% macerozyme (Serva) and 7% cellulase (Calbiochem) in citrate buffer pH 4.6 for 45 min at 37°C, and then squashed under a coverslip in a drop of 60% acetic acid. The coverslips were removed after freezing at −80°C. The air-dried preparations (three slides for each *O. paniculata* plant and three for cv. Leccino) were simultaneously Feulgen stained after hydrolysis in 1 N HCl at 60°C for 8 min. After staining, the slides were subjected to three 10-min washes in SO_2 water prior to dehydration and mounting in distyrene-dibutylphthalatexylene (DPX; BDH Chemicals). For each slide, 30 prophase nuclei were measured. Feulgen stained DNA in individual prophase nuclei was measured in images captured by a charge-coupled-device camera on a Leica DMRB microscope, using a Leica Q500MC image analyser. Results are given as average of 4C-DNA absorption value ± standard error (in arbitrary units).

Fluorescence *in situ* Hybridisation

The *Copia-SIRE* probe, a 406 bp-long *Copia* fragment belonging to the *SIRE* lineage, was amplified by polymerase chain reaction (PCR) from both 50 ng of genomic DNA from *O. paniculata* and cv. Leccino. Primers were designed to an RNAse H encoding sequence (forward primer: 5′-TTGATCGAAAAAGCACTAGCGGAAC-3′ and reverse primer: 5′-AGTCCTCTACGAATAAATGAAAAACG-3′) of a *SIRE*-related cluster from the graph-based clustering analysis. PCR conditions were 94°C for 4 min, followed by 30 cycles of 94°C for 30 s, 58°C for 30 s and 72°C for 40 s. A final extension was performed at 72°C for 7 min. PCR products were purified with a Wizard SV Gel and PCR Clean-Up System (Promega), and cloned into the pGEM-T Easy plasmid vector (Promega). The cloned fragments were sequenced. For each probe, one clone was selected (GenBank accession number OM829845 for *Copia-SIRE* probe of *O. paniculata* and OM829844 for *Copia-SIRE* probe of cv.Leccino) and used for FISH analysis.

Six olive probes designed on the sequences of TRs families specific for *O. europaea* were also used as: O-51 (905 bp, GenBank accession number OM829846), O-80 (879 bp, GenBank accession number OM829847), O-86 (889 bp, GenBank accession number OM829848), O-178 (1,025 bp, GenBank accession number OM829849), O-179 (1,145 bp, GenBank accession number OM829850) and O-218 (1,289 bp, GenBank accession number OM829851).

Primers used for O-51 were 5′-CCTATTGATGCTGTGTTGACC-3′ and 5′- GGATAGACTTTGTCCCGTGA-3′, for O-80 were 5′-GAAAAATGACGAAATTGCCCCCGA-3′ and 5′-TCGACTGTGTCGGAATTGGCTGAAATTTG-3′, for O-86 were 5′-TTTTTTCGTTTTTGGCGAATTGCT-3′ and 5′-CAGGGTTTTCCCAGTCACGACGT-3′, for O-178 were 5’-CGAAGAAGATTTGAGTTCAATCCA-3′ and 5’-GAAGAATGAGCACTTTATATTTAGA-3′, for O-179 were 5′-ATAGAGAATAAGCAAAAGTCTACC-3′ and 5′-TGATGGTTTTAATATTGGAGCTT-3′ and for O-218 were 5′-CATTCCGACACCGATAAGAC-3′ and 5′-GGCCGAAATTTTGTAAGTTGT-3′. PCR conditions and cloning procedure were as described above.

Probes were labelled by nick translation using DIG-Nick Translation Mix (Roche) or Biotin-Nick Translation Mix (Roche).

In situ hybridisation was performed as described in Ceccarelli et al. (2010). Slides were prepared using root apices from potted plants for *O. paniculata*, or from cuttings for both cv. Leccino and *O. europaea* subsp. *cuspidata*. The apices were treated with a saturated aqueous solution of alpha-bromonaphtalene for 4 h at room temperature, fixed in ethanol:acetic acid (3:1 v/v) and processed as described above (see Genome Size Estimation). DNA of nuclei was denatured in a thermal cycler for 8 min at 70°C and the preparations were then incubated overnight at 37°C with 2 ng/μl of heat-denatured DNA probes. The digoxigenin and biotin at the hybridisation sites were detected by using sheep anti-digoxigenin-fluorescein (Roche) and streptavidin-Cy-3 (Sigma), respectively. Nuclei were then counterstained using 0.2 μg/ml 4,6-diamino-2-phenylindole (DAPI) in McIlvaine buffer pH 7.0, mounted in AF1 antifade solution (Citifluor) and examined with a Leica DMRB fluorescence microscope. At least ten metaphase plates were analysed for each probe and images were captured using an ILCE-7 camera (SONY) and optimised using Adobe Photoshop 5.0.

RESULTS

Characterisation of the Repetitive Component in the Genus *Olea*

Genome structure of the genus *Olea* was studied in four taxa of the subgenus *Olea*, i.e. the cultivated olive (*O. europaea* subsp. *europaea*, cv. Leccino); *O. europaea* subsp. *cuspidata; O. europaea* subsp. *guanchica; O. exasperata*; and in *O. paniculata*, belonging to the subgenus *Paniculatae* (**Table 1**).

TABLE 1 | *Olea* taxa analysed and number of Illumina reads used for the analyses.

Subgenus	Section	Species	Subspecies	Origin	Sample source	Raw reads	Trimmed reads
Olea	*Olea*	*O. europaea*	*europaea*	Italy	CNR-IBBR[1]	71,624,494	47,023,392
Olea	*Olea*	*O. europaea*	*guanchica*	Canary Islands	IFAPA[2]	16,457,568	13,478,858
Olea	*Olea*	*O. europaea*	*cuspidata*	Ethiopia	IFAPA[2]	20,368,004	16,175,030
Olea	*Ligustroides*	*O. exasperata*	–	South Africa	IFAPA[2]	15,348,186	12,211,284
Paniculatae	-	*O. paniculata*	–	Australia	CSIRO[3]	20,622,182	17,243,520

[1] *Olive Collection of CNR—Institute of Biosciences and Bioresources, Division of Perugia (Perugia, Italy).*

[2] *IFAPA World Olive Germplasm Bank (Cordoba, Spain).*

[3] *CSIRO Agriculture & Food (Narrabri, NSW, Australia).*

TABLE 2 | Genome proportion of repetitive sequence classes among the analysed taxa.

Repeats in the genome	*O. europaea* subsp. *europaea*	*O. europaea* subsp. *guanchica*	*O. europaea* subsp. *cuspidata*	*O. exasperata*	*O. paniculata*
DNA-TE%	2.06	2.41	1.64	2.74	2.59
RE%	28.84	27.89	18.28	32.85	51.59
TR%	23.89	23.35	50.44	26.43	1.94
rDNA%	0.37	0.72	1.45	1.24	0.50
Not classified%	1.51	1.64	1.24	1.53	2.62
TOTAL%	56.67	56.01	73.04	64.80	59.25

In order to identify different families of repeats, resulting samples of 100 nt paired-end reads were analysed with the RepeatExplorer2 tool. On average, repetitive DNA in *Olea* species ranged from 56% in *O. europaea* subsp. *guanchica* to 73% in *O. europaea* subsp. *cuspidata*, showing remarkable differences in terms of composition (**Table 2**). Our analysis indicated that the peculiar structure of the olive genome with the characteristic abundance of TR sequences (Barghini et al., 2014) was also present in other *Olea* taxa. In fact, the analysed genomes showed a massive occurrence of DNA satellites in the form of TRs, accounting from 23% in *O. europaea.* Subsp. *guanchica* to 50% in *O. europaea* subsp. *cuspidata*, with the notable exception of *O. paniculata*, for which TRs only amounted to 1.94% of the genome. For interspersed repeats, DNA TEs were poorly represented among the analysed taxa, while REs accounted for a considerable part of the repetitive component, ranging from 18.28% in *O. europaea* subsp. *cuspidata* to 51.59% in *O. paniculata.*

Analysis of Tandem Repeats

Clusters of *Olea* sequenced reads classified as putative satellites were inspected manually, in order to validate monomer consensus sequences. Overall, we identified 91 different sequences of TRs, organised in 11 major families (**Figure 1** and **Supplementary Figure S1**). Among these major families, six had previously been identified in cultivated olive (Katsiotis et al., 1998; Bitonti et al., 1999; Minelli et al., 2000; Lorite et al., 2001; Barghini et al., 2014), even if their homologues were not found in all species by clustering analysis. In addition, five new species-specific families, three in *O. exasperata* and two in *O. paniculata* were identified by graph-based cluster analysis (**Supplementary Figure S2**). As already reported for cultivated olive, besides TR families with a typical monomer length of more than a hundred base pairs, some families were detected with repeat units of either 51-bp or 47-bp. TRs O-80, O-178 and O-218 constituted heavy satellite families, having a GC content around 44% or higher. By contrast, O-47, O-121 and O-51 had a GC content around 22, 27 and 32%, respectively, representing light satellite families (**Supplementary Table S1**).

TR families showed great variability in terms of abundance across the genus *Ole*a. Mapping results indicated the presence of all sequences in all analysed taxa, highlighting a great genomic variability since some families were barely represented in one species while being highly abundant in another (**Supplementary Table S1**; **Figure 2**). Abundance data concerning TR families were also used to produce a phylogenetic tree (**Figure 2**). The dendrogram is consistent with the phylogeny of the genus *Olea* (Besnard et al., 2009), supporting separation among the three different sections analysed, with O. *paniculata*, the species closest to the *Olea* common ancestor, showing a TR abundance pattern quite different from the other species.

Analysis of LTR-Retrotransposons

Besides TRs, LTR-RE-related clusters composed the bulk of highly and moderately repeated sequences in *Olea* genomes. After annotation against a library of 254 putative full-length REs of olive (Barghini et al., 2014), these elements were studied at the lineage level (**Table 3**). Seven lineages (plus one group that could not be annotated) were identified among *Copia* retrotransposons (*AleI-Retrofit*, *AleII*, *Angela*, *Bianca*, *Ivana-Oryco*, *SIRE* and *TAR/Tork*), and three lineages (plus one group that could not be annotated) were identified among *Gypsy* elements (*Athila*, *Chromovirus* and *Ogre/Tat*).

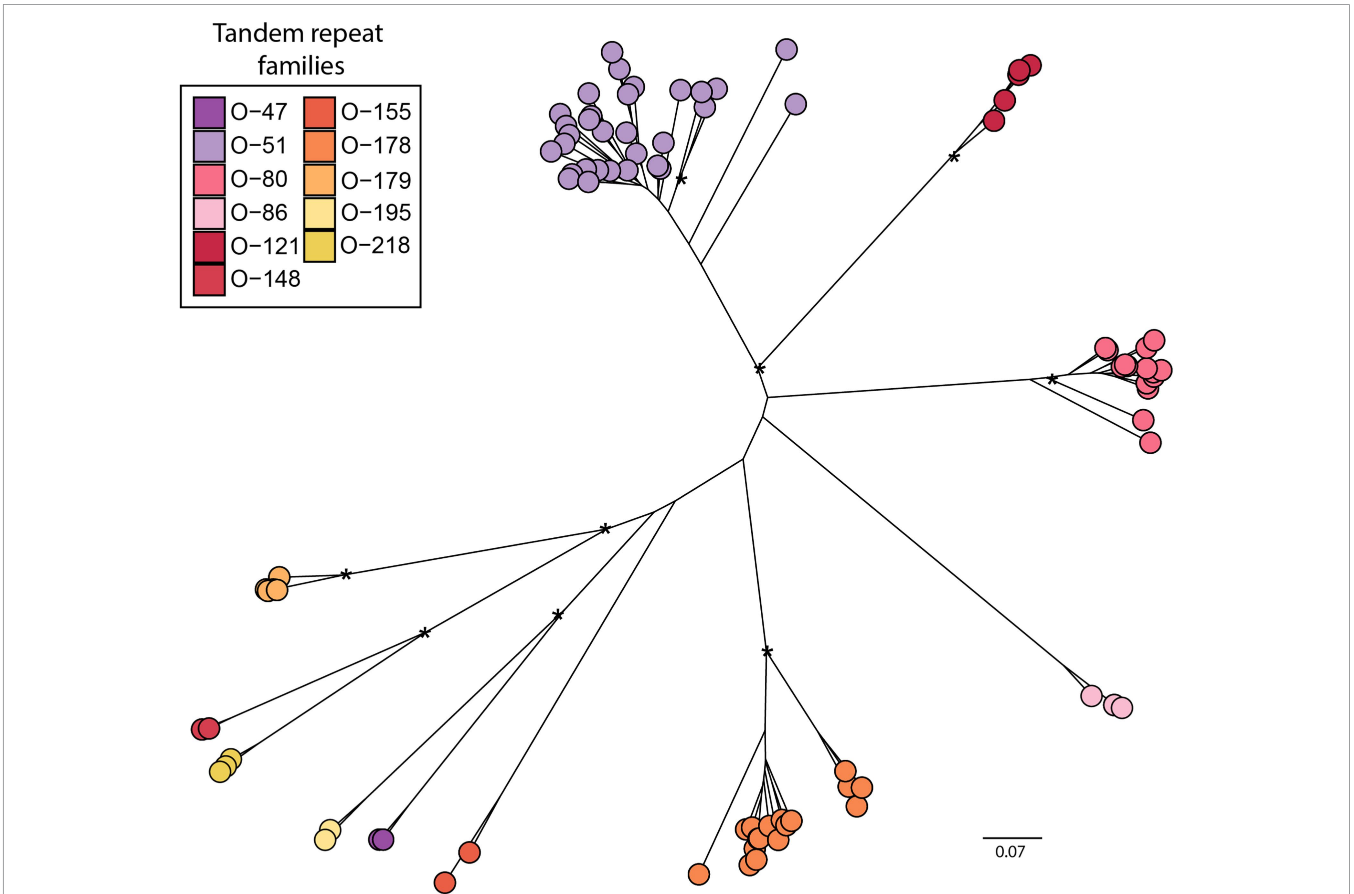


FIGURE 1 | Distance tree of 11 TR families identified across the genus *Olea* (91 representative sequences). Bootstrap values higher than 0.6 are shown. Bar shows the nucleotide distance.

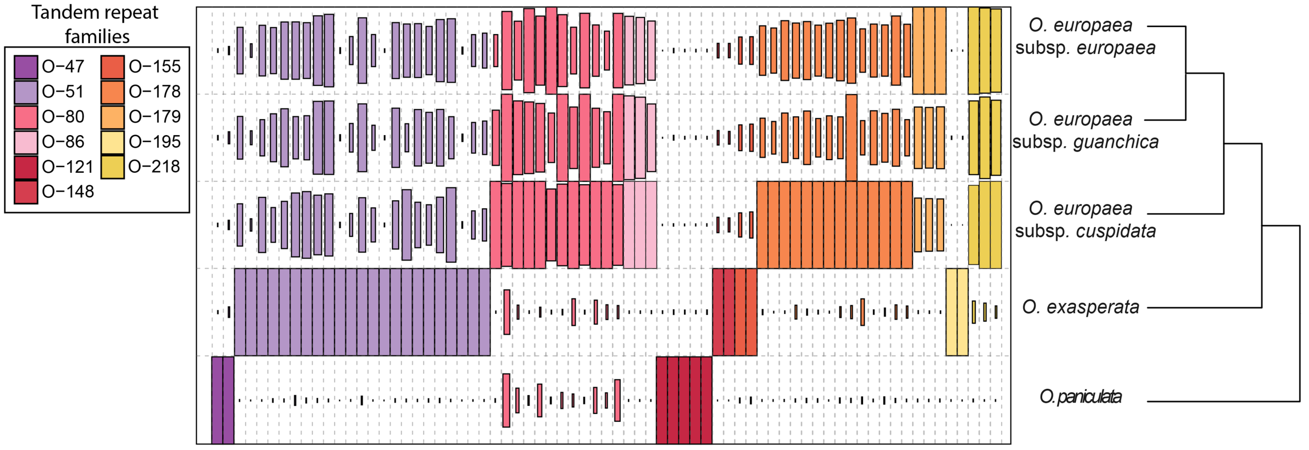


FIGURE 2 | Sequence composition of TR sequences isolated from the analysed species. The size of the rectangle is proportional to the genome proportion of a cluster for each species. The colours of the rectangles correspond to the different TR families.

Abundance of *Gypsy* LTR-REs ranged from 10.06% in *O. europaea* subsp. *cuspidata* to 26.07% in *O. paniculata*, and they were overrepresented compared to *Copia* elements, which ranged from 6.88% in *O. europaea* subsp. *cuspidata* to 20.54% in *O. paniculata*. The ratios of the genomic proportions of *Gypsy* and *Copia* elements differed among species, from 1.27 in

TABLE 3 | Genome proportion of LTR-RE sequences and maximum percentage of variation among the analysed taxa.

Superfamily	Lineage	Genomic abundance					Max. percentage of variation
		O. europaea* subsp. *europaea	***O. europaea* subsp. *guanchica***	***O. europaea* subsp. *cuspidata***	***O. exasperata***	***O. paniculata***	
Copia	*Alel-Retrofit*	0.007	0.014	<0.005	<0.005	<0.005	68.65
	AleII	0.41	0.36	0.22	0.29	1.09	79.79
	Angela	3.55	3.48	2.23	1.93	5.28	63.36
	Bianca	0.67	0.74	0.38	0.59	0.60	47.73
	Ivana-Oryco	0.27	0.24	0.11	0.27	0.34	68.88
	Maximus/SIRE	1.07	1.03	0.64	0.49	6.13	92.00
	TAR/Tork	5.12	4.54	3.05	4.26	6.60	53.73
	Unknown	0.37	0.38	0.24	0.22	0.50	52.50
	Total	11.46	10.76	6.88	8.05	20.54	66.52
Gypsy	*Athila*	3.34	3.09	2.02	5.46	7.07	71.43
	Chromovirus	5.27	4.88	3.00	4.98	10.11	70.34
	Ogre/Tat	4.96	4.75	4.10	10.48	7.62	60.89
	Unknown	1.33	1.62	0.95	1.74	1.27	45.42
	Total	14.89	14.34	10.06	22.65	26.07	61.39
LTR-RE unclassified		0.38	0.35	0.20	0.27	2.18	91.03
LTR-*Gypsy*/LTR-*Copia*		1.30	1.33	1.46	2.81	1.27	54.88

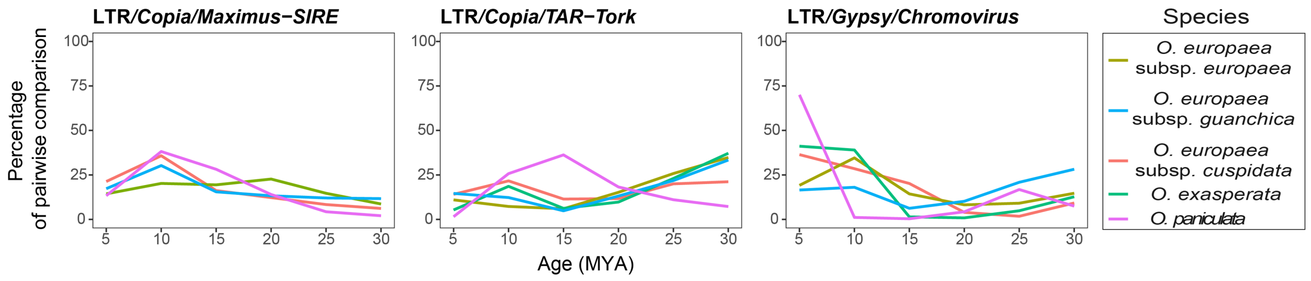


FIGURE 3 | Timing of the LTR/*Copia/Maximus-SIRE*, *TAR-Tork* and LTR/*Gypsy/Chromovirus* retrotranspositional activity in the analysed taxa. The y-axis shows the percentage number of pairwise comparisons of reads matching the RE-RT-specific domain.

O. paniculata to 2.81 in *O. exasperata*. Clusters that remained un-annotated composed a very small fraction of the analysed genomes, ranging from 0.20% in *O. exasperata* to 2.18% in *O. paniculata*.

Furthermore, to elucidate the possible role of LTR-RE dynamics during *Olea* taxa separation, we also analysed RE insertion time (**Figure 3**). Although RE insertion times, calculated by comparing coding sequences (Ammiraju et al., 2007), should be taken cautiously, the results showed a similar proliferation profile for all the analysed taxa, except for *O. paniculata*, in which the proliferation burst of three major families of REs started in the last 25/20 million years (MY) and reached its apex in the last 15/5 MY.

Repeats Composition Variation in *Olea* Taxa

Comparing the abundance of RE and TR families retrieved in the 5 taxa analysed, it can be seen that in four of them TR abundance was inversely correlated with that of REs (**Figure 4**). The opposing trend was observed for *O. paniculata*, potentially the oldest species, originated around 24 million years ago (MYA) from the *Olea* common ancestor (Besnard et al., 2009), which had very few tandem-repeated sequences, while being rich in LTR-REs.

Cytological Analyses

The differences in repeat organisation between *O. paniculata* and the other taxa were confirmed by cytological analyses. Image cytometry of prophase nuclei was used to estimate the genome size of *O. europaea* subsp. *europaea* and *O. paniculate*. The analyses returned a 4C-DNA absorption value of 207,067 ± 5,673 for *O. europaea* subsp. *europaea* and 376,475 ± 46,638 for *O. paniculate*, respectively, indicating that *O. paniculata* genome size was larger than that of *O. europaea* subsp. *europaea*, showing an increase of 44.9%.

The variation in genome size was reflected in the chromatin organisation. Indeed, *O. paniculata* interphase nucleus, largely occupied by LTR-REs, showed an eureticulate structure, characterised by dense, conspicuous and regular chromatin reticulum with barely visible chromocenters (DAPI positive heterochromatic regions; **Figure 5A**), while cultivated olive had

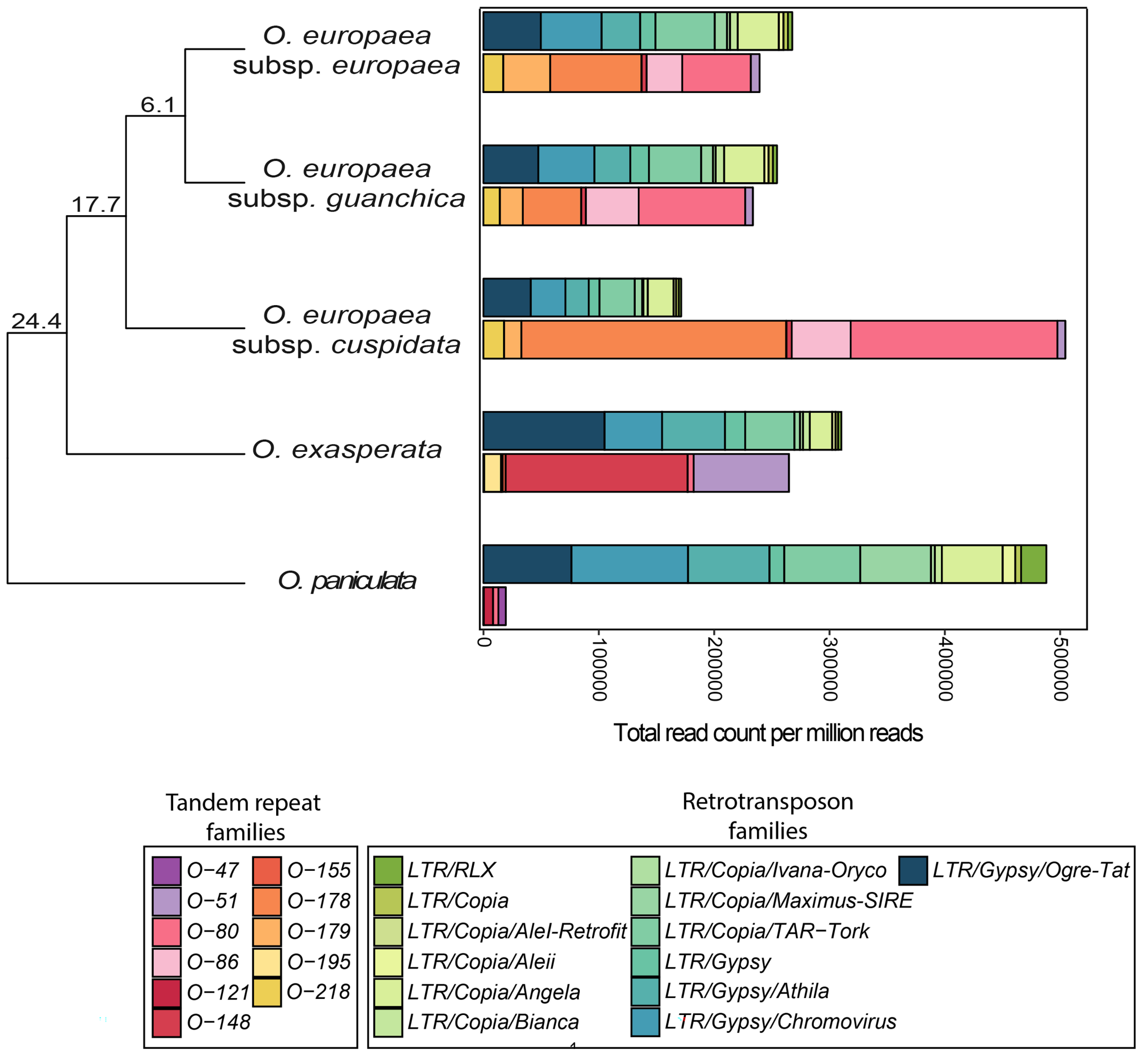


FIGURE 4 | Stacked bar plots comparing the genome proportion of LTR-RE families and TR families in *Olea*. Abundance values were measured by counting the number of reads (per million) mapping the set of repetitive sequences collected in the reference library. Phylogenetic tree reports the estimated divergence times (in MY) from the common ancestor for the *Olea* taxa used in this study, according to Besnard et al. (2009).

an areticulate or chromocentric nucleus, with prominent chromocenters standing out on a barely visible euchromatin reticulum (**Figure 5D**). Fluorescence *in situ* hybridisation (FISH) of a fragment belonging to a family of *Copia-SIRE* LTR-REs confirmed their massive presence in *O. paniculata*, being the hybridisation signal largely scattered across the nucleus (**Figure 5B**). By contrast, the signal from hybridisation of a TR fragment from the family O-80 (*Oe*Taq80) formed a few small clusters corresponding to as many chromocenters (**Figure 5C**). The opposite results were obtained in the nuclei of cultivated olive, where no signal was observed after FISH with the *Copia*-SIRE probe (**Figure 5E**), but intense hybridisation signals of *Oe*Taq80 were localised at the DAPI positive chromocenters (**Figure 5F**).

Finally, FISH experiments were carried out to highlight possible differences in TRs chromosomal localization between cultivated olive and *O. europaea* subsp. *cuspidata*, for which molecular analyses indicated a TR abundance of 50% of the genome. Six different probes were designed on the sequences of TRs families specific for *O. europaea* and hybridised in root-tips chromosomes of the two subspecies. O-51 and O-179 families had never been hybridised before, whereas the chromosomal localization of the remaining TRs was already studied by Katsiotis et al. (1998) and Minelli et al. (2000) in different olive cultivars. Metaphase plates hybridised with O-51 and O-178 were reported in **Figure 6**; those hybridised with O-80, O-86, O-179 and O-218 were reported in **Supplementary Figure S3**.

The maximum number of chromosome pairs showing signals after hybridisation with each probe, and minimum and maximum number of hybridisation signals counted on metaphase plates in the two subspecies were reported in **Table 4**. Differences

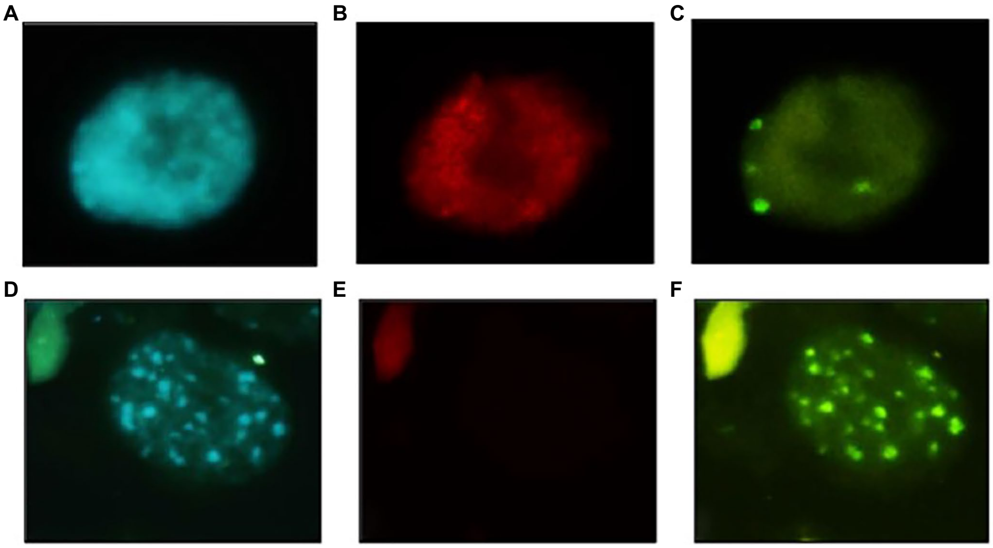


FIGURE 5 | Interphase nuclei in the shoot meristem of *Olea paniculata* **(A–C)** and *Olea europaea* subsp. *europaea* **(D–F)**. Images after DAPI staining **(A,D)**, after hybridisation with the *O. paniculata Copia-SIRE* probe **(B,E)** and after hybridisation with *Oe*Taq80 DNA repeats **(C,F)**. Images similar to **(B,E)** were obtained with the *O. europaea* subsp. *europaea Copia-SIRE* probe (data not shown). Bar = 10 µm.

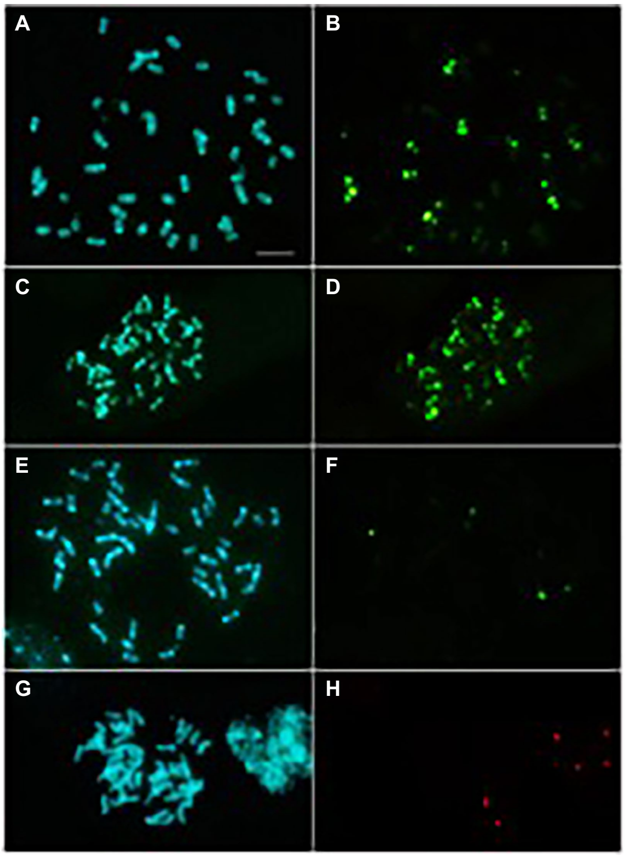


FIGURE 6 | Metaphase plates of *O. europaea* subsp. *europaea* [cv. Leccino; **(A,B,E,F)** and *O. europaea* subsp. *cuspidata* **(C,D,G,H)** after DAPI staining **(A,C,E,G)** and hybridisation with O-178 (**B,D**; fluorescein) or O-51 (**F,H**; fluorescein) repeats. Bar = 10 µm.

TABLE 4 | Maximum number of chromosome pairs showing signals after hybridisation with each probe, and minimum and maximum number of hybridisation signals counted on metaphase plates in the two subspecies.

	Chromosomes pairs		Hybridisation signals	
Probe	***O. europaea* subsp. *europaea***	***O. europaea* subsp. *cuspidata***	***O. europaea* subsp. *europaea***	***O. europaea* subsp. *cuspidata***
O-51	2	1	3–4	2
O-80	23	23	63–66	54–62
O-86	13	13	29–37	27–35
O-178	10	15	22–30	47–50
O-179	17	17	37–40	40–42
O-218	10	10	18–20	15–19

At least ten metaphase plates for each probe and subspecies were analysed.

in chromosomal distribution of O-178 and O-51 related sequences were found between the two taxa. Ten chromosome pairs of the cultivated olive complement showed O-178 hybridisation signals versus the 15 chromosome pairs in *O. europaea* subsp. *cuspidata*. In total, 47 to 50 hybridisation signals were counted on *O. europaea* subsp. *cuspidata* chromosomes while only 22 to 30 signals were found in cultivated olive (**Table 4**). On the contrary, O-51 probe found nucleotide sequence homology in two chromosome pairs of the cv. Leccino complement and only in one pair in *O. europaea* subsp. *cuspidata* (**Table 4**).

Any noticeable difference was found between the two subspecies regarding the chromosomal distribution of the other TRs (**Table 4**; **Supplementary Figure S3**). O-80-related sequences were found in all the chromosome pairs in both taxa. Structural heterozygosity of the chromosome pair I, already described in cultivated olive (cv. Coratina; Minelli et al., 2000), was also observed in both cv. Leccino and subsp. *cuspidata*. O-86 repeats hybridised on 13 chromosome pairs. The O-179 probe found related sequences in 17 pairs of both chromosome complements. A slightly higher number of weak hybridisation signals related to O-218 sequences was observed in cv. Leccino, the two complements substantially showing the same number of signals of major and minor intensity (**Table 4**).

DISCUSSION

Repetitive sequences represent one of the most cryptic components of eukaryotic genomes (Garrido-Ramos, 2015, 2017; Bourque et al., 2018). For a long time, this fraction was considered of little importance, and it still remains ill-defined because of the technical issues associated with reliable characterising representative sets of sequence and also for the great variability in terms of abundance and/or sequence conservation at interspecific and intraspecific levels (Mascagni et al., 2015, 2017a; Robledillo et al., 2018).

In order to clarify the processes that led to the present structure of the cultivated olive genome, a deep characterisation of the repetitive fraction of olive was performed in comparison with four other taxa belonging to the genus *Olea*, through bioinformatics, cytophotometric and cytological analyses. To achieve this, first, a graph-based clustering approach, already applied in several species (Novák et al., 2014; Barghini et al., 2015; Usai et al., 2017), including cultivated olive (Barghini et al., 2014a), was used. Results confirmed the peculiar genomic structure of cultivated olive, with its high composition of TRs (accounting for ~24%). The high abundance of TRs was also shown to be a general feature of all the analysed species of the subgenus *Olea*, with *O. europaea* subsp. *cuspidata* having a TR abundance of 50% of the genome. These data confirmed the singular evolution of the subgenus *Olea* since, in other taxa, TRs usually account for <10% of the genome, with some exceptions like cucumber or *Fritillaria falcata*, whose genomes comprise ~23 and 36% of these sequences, respectively (Huang et al., 2009; Ambrožová et al., 2010).

The TR families identified in the analysed genomes showed low sequence similarity and great variability in terms of genomic abundance, suggesting their independent origins. In plants, it is a common feature of related species to share a set of TR families, with one or a few predominant TR species-specific families (King et al., 1995). However, TR sequences are usually considered fast-evolving components that can also cause reproductive barriers between organisms, thus promoting species separation (Schmidt and Heslop-Harrison, 1993; Garrido-Ramos, 2017). In fact, while some TR sequences can exhibit conservation of the monomer sequence for long evolutionary periods (Cafasso and Chinali, 2014; Mehrotra and Goyal, 2014), other TRs are subjected to different constraints. Low preservation of sequence similarity or abundance is reported for several plant groups, where some monomers may be preferred over others at the evolutionary level (Flavell, 1982; Cafasso and Chinali, 2014; Mehrotra and Goyal, 2014). Recently, the hypothesis of a possible contribution to TR evolution and mobility by TEs has been proposed (Meštrović et al., 2015; Vondrak et al., 2020). In the genomes of *Chenopodium sensu stricto*, TEs may act as a substrate for TRs, generating a sort of 'library' of tandemly arranged sequences that, after being dispersed through the genome through transposition, may be amplified into long arrays of new TR families (Belyayev et al., 2020).

Since relative abundance of well-represented repeats is a representation of general genome composition, we used genome-wide abundance of TRs as continuously varying characters in order to build a phylogenetic tree. This methodology can be particularly useful in groups showing little genetic differentiation in classic phylogenetic markers, actually providing information for phylogenetic inference (Dodsworth et al., 2014). The dendrogram obtained from our data supported the separation among the three sections of *Olea* considered in this study (Besnard et al., 2009), highlighting the differences in the genome composition of O. *paniculata*, the closest species to the *Olea* common ancestor.

In *O. paniculata*, as typical of many plant species, interspersed REs accounted for the vast majority of the repetitive component, while TRs were barely present, consistently with the results reported for a TR family by Bitonti et al. (1999). In this species, our data indicated that massive RE proliferation started around ~20 MYA and reached its apex in the last 15–5 MY, i.e., after separation of the subgenus *Olea*. Concurrently, the other *Olea* species originating from the same ancestor (Besnard et al., 2009) had a huge increase in TR abundance which can be explained by the so-called 'library model' (Fry and Salser, 1977). In this hypothesis of TR evolution, closely related species share a set of conserved TR families each of which is differentially amplified in each species forming a sort of library accompanied by rapid evolution of nucleotide sequences and copy number change (Cesari et al., 2003; Thakur et al., 2021). In Olea, the partial replacement of an RE increase by TR accumulation, during subgenus *Olea* species separation, was a fairly unique event. Interestingly, in all species overall, TR abundance was inversely correlated to that of REs. This trend might imply a direct competition in the proliferation of these two classes of repeats, suggesting that the species of the subgenus *Olea* underwent amplification of TRs and a reduced proliferation of retrotransposons.

Cytological analyses underlined the differences in genome size and organisation of *O. paniculata* compared to *O. europaea* subsp. *europaea*. The genome size of *O. paniculata* was about 50% larger than that of cultivated olive. Such a difference between species with the same chromosome number is usually attributed to variations in the abundance of repetitive DNA (Flavell, 1986). In this case, supported by RE insertion timing data and by *in situ* hybridisation results, the genome expansion of *O. paniculata* might be derived from a massive amplification through retrotransposition of major individual RE families in the last ~20 MY, while TRs remained below 2% of the genome. A similar case is represented by a study on the genus *Passifora*, where *Passifora quadrangularis*, the species with the largest genome, presents a higher accumulation of REs compared to *Passifora organensis*, whose genome shows a greater diversity and the highest proportion of satellites (Sader et al., 2021).

Accordingly, there are reports of how the amplification of one or a few specific repeats led to an increase in genome size. In maize, almost 25% of the genome is represented by five LTR-RE families (SanMiguel et al., 1996). In five species of iris (*Iris* ser. *Hexagonae*), a characteristic RE type accounts for 6–10% of the genome (Kentner et al., 2003). Finally, in *Vicia pannonica*, a single family of *Gypsy* elements caused the expansion of the genome by 50% (Neumann et al., 2006).

The different composition of the *O. paniculata* genome also reflects in the organisation of its genetic material. Indeed, interphase nuclei are arranged in distinct reticulate structures (eureticulate type; Delay, 1946-1947, 1948) confirming the absence of highly repetitive TR families. In *O. europaea* subsp. *europaea*, the proliferation of TRs, which still represents an important part of its repetitive component, could have preserved the genome from massive expansion. Moreover, the great amount of TRs, which are the main component of heterochromatin, regulating its formation and preserving its structure (Grewal and Elgin, 2007; Garrido-Ramos, 2015), results in the occurrence of chromocenters, nuclear regions containing just highly repetitive, tandemly arranged DNA sequences (Botchan et al., 1971; Gall et al., 1971; Peacock et al., 1974; Guenatri et al., 2004). This phenomenon is not limited to plant kingdom: even in some animal genomes, it is possible to observe cases in which TEs likely affected the formation of TRs and the conversion of euchromatic chromosomes into heterochromatic ones (Bachtrog et al., 2019; Palacios-Gimenez et al., 2020).

Finally, FISH experiments highlight that some TRs were physically localised in the genome of closely related species (i.e., *O. europaea* subsp. *europaea* and subsp. *cuspidata*) significantly differing in TRs abundance. The results suggested a different evolutionary model for the various families within *O. europaea*. A higher number of hybridisation signals was observed for O-178 in *O. europaea* subsp. *cuspidata* rather than in subsp. *europaea*. In this case, it is clear that O-178 dissemination in a genome (involving TEs or other mechanisms) occurred more extensively than in the other one. On the contrary, O-51 showed 2 hybridisation signals in *O. europaea* subsp. *europaea* versus only one in *O. europaea* subsp. *cuspidata*. However, it is to be considered that O-51 accounted only for a minimal portion of the genomes. Concerning the other TRs, regardless of their genome abundance, a similar number of hybridisation signals were observed for O-80, O-86, O-179 or O-218 families in the two subspecies. It can be assumed that, after their dissemination throughout the *O. europaea* genome, these TR families differentially amplified in the two subspecies, maintaining the same positions in each genome. However, it cannot be ruled out that differences in genomic abundance not revealed by cytological observations could be due to the greater distribution in a genome of short arrays whose copy number is below the sensitivity FISH threshold (Ruiz-Ruano et al., 2016). In conclusion, the current study shed light on the evolution of the genus *Olea*, highlighting the prominent role of TRs in fostering genome structure variation. After the separation of the subgenus *Olea* (24.4 MYA), tandemly arranged sequences underwent a massive proliferation, leading to the peculiar genomes of cultivated olive and its related species. By contrast, in *O. paniculata*, the closest species to the *Olea* common ancestor, the TR proliferation burst never occurred, opening the way for REs amplification, which resulted in an expansion of the genome. Based on the huge difference in repetitive fraction composition, combined with the notable TR abundance of some species, the genus *Olea* represents a quite singular model of genome evolution in higher plants. Studies, using new long-molecule sequencing methods, will further decipher the structure of TR loci and help to clarify the amplification mechanisms of these sequences.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/, PRJNA810942, SRX465835 https://www.ncbi.nlm.nih.gov/genbank/, OM829845 https://www.ncbi.nlm.nih.gov/genbank/, OM829844 https://www.ncbi.nlm.nih.gov/genbank/, OM829846 https://www.ncbi.nlm.nih.gov/genbank/, OM829847 https://www.ncbi.nlm.nih.gov/genbank/, OM829848 https://www.ncbi.nlm.nih.gov/genbank/, OM829849 https://www.ncbi.nlm.nih.gov/genbank/, OM829850 https://www.ncbi.nlm.nih.gov/genbank/, OM829851.

AUTHOR CONTRIBUTIONS

FM, AC, and LN planned and designed the project. TG and LB performed nucleic acid extractions. MC and CT performed the cytological analyses. FM and EB performed the bioinformatics analyses. FM, EB, MC, LB, CT, CD, TG, LN, and AC discussed the data, wrote the manuscript, and contributed to its final form. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Department of Agriculture, Food and Environment of the University of Pisa, Italy, Project 'Plantomics'.

ACKNOWLEDGMENTS

We thank Brandon Gaut for his critical reading of the manuscript. RepeatExplorer2 computational resources were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.869048/full#supplementary-material

REFERENCES

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I. J., Koblížková, A., et al. (2010). Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria lilies*. *Ann. Bot.* 107, 255–268. doi: 10.1093/aob/mcq235

Ammiraju, J. S., Zuccolo, A., Yu, Y., Song, X., Piegu, B., Chevalier, F., et al. (2007). Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. *Plant J.* 52, 342–351. doi: 10.1111/j.1365-313X.2007.03242.x

Bachtrog, D., Mahajan, S., and Bracewell, R. (2019). Massive gene amplification on a recently formed *drosophila* Y chromosome. *Nat. Ecol. Evol.* 3, 1587–1597. doi: 10.1038/s41559-019-1009-9

Barghini, E., Mascagni, F., Natali, L., Giordani, T., and Cavallini, A. (2015). Analysis of the repetitive component and retrotransposon population in the genome of a marine angiosperm, *Posidonia oceanica* (L.) Delile. *Mar. Genomics* 24, 397–404. doi: 10.1016/j.margen.2015.10.002

Barghini, E., Natali, L., Cossu, R. M., Giordani, T., Pindo, M., Cattonaro, F., et al. (2014a). The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome. *Genome Biol. Evol.* 6, 776–791. doi: 10.1093/gbe/evu058

Barghini, E., Natali, L., Giordani, T., Cossu, R. M., Scalabrin, S., Cattonaro, F., et al. (2014). LTR retrotransposon dynamics in the evolution of the olive (*Olea europaea*) genome. *DNA Res.* 22, 91–100. doi: 10.1093/dnares/dsu042

Belyayev, A., Josefiová, J., Jandová, M., Mahelka, V., Krak, K., and Mandák, B. (2020). Transposons and satellite DNA: on the origin of the major satellite DNA family in the *Chenopodium* genome. *Mob. DNA* 11, 1–10. doi: 10.1186/s13100-020-00219-7

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

Besnard, G., Rubio de Casas, R., Christin, P. A., and Vargas, P. (2009). Phylogenetics of *Olea* (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: tertiary climatic shifts and lineage differentiation times. *Ann. Bot.* 104, 143–160. doi: 10.1093/aob/mcp105

Bitonti, M., Cozza, R., Chiappetta, A., Contento, A., Minelli, S., Ceccarelli, M., et al. (1999). Amount and organization of the heterochromatin in *Olea europaea* and related species. *Heredity* 83, 188–195. doi: 10.1046/j.1365-2540.1999.00564.x

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Botchan, M., Kram, R., Schmid, C. W., and Hearst, J. E. (1971). Isolation and chromosomal localization of highly repeated DNA sequences in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* 68, 1125–1129. doi: 10.1073/pnas.68.6.1125

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., et al. (2018). Ten things you should know about transposable elements. *Genome Biol.* 19:199. doi: 10.1186/s13059-018-1577-z

Buti, M., Giordani, T., Cattonaro, F., Cossu, R., Pistelli, L., Vukich, M., et al. (2011). Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions. *Theor. Appl. Genet.* 123, 779–791. doi: 10.1007/s00122-011-1626-4

Buti, M., Moretto, M., Barghini, E., Mascagni, F., Natali, L., Brilli, M., et al. (2017). The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to *Fragaria vesca* (the woodland strawberry). *GigaScience* 7:giy010. doi: 10.1093/gigascience/giy010

Cafasso, D., and Chinali, G. (2014). An ancient satellite DNA has maintained repetitive units of the original structure in most species of the living fossil plant genus zamia. *Genome* 57, 125–135. doi: 10.1139/gen-2013-0133

Catlin, N. S., and Josephs, E. B. (2022). The important contribution of transposable elements to phenotypic variation and evolution. *Curr. Opin. Plant Biol.* 65:102140. doi: 10.1016/j.pbi.2021.102140

Ceccarelli, M., Sarri, V., Polizzi, E., Andreozzi, G., and Cionini, P. G. (2010). Characterization, evolution and chromosomal distribution of two satellite DNA sequence families in *Lathyrus* species. *Cytogenet. Genome Res.* 128, 236–244. doi: 10.1159/000298852

Cesari, M., Luchetti, A., Passamonti, M., Scali, V., and Mantovani, B. (2003). PCR amplification of the Bag320 satellite family reveals the ancestral library and past gene conversion events in *bacillus rossius* (*Insecta Phasmatodea*). *Gene* 312, 289–295. doi: 10.1016/S0378-1119(03)00625-5

Contento, A., Ceccarelli, M., Gelati, M., Maggini, F., Baldoni, L., and Cionini, P., et al. (2002). Diversity of *Olea* genotypes and the origin of cultivated olives. *Theor. Appl. Genet.* 104, 1229–1238. doi: 10.1007/s00122-001-0799-7

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., et al. (2016). Genome sequence of the olive tree, *Olea europaea*. *Gigascience* 5, s13016–s13742. doi: 10.1186/s13742-016-0134-5

Delay, C. (1946-1947). Recherches sur la structure des noyaux quiescents chez les Phanerogames. *Rev Cytol Cytophysiol Veg.* 9, 169–222.

Delay, C. (1948). Recherches sur la structure des noyaux quiescents chez les Phanerogames. *Rev. Cytol. Cytophysiol. Veg.* 10, 103–228.

Dodsworth, S., Chase, M. W., Kelly, L. J., Leitch, I. J., Macas, J., Novák, P., et al. (2014). Genomic repeat abundances contain phylogenetic signal. *Syst. Biol.* 64, 112–126. doi: 10.1093/sysbio/syu080

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

Fieldhouse, D., and Golding, B. (1991). A source of small repeats in genomic DNA. *Genetics* 129, 563–572. doi: 10.1093/genetics/129.2.563

Flavell, R. (1982). "Sequence amplification, deletion and rearrangement: major sources of variation during species divergence," in *Genome Evolution*. eds. G. A. Dover and R. B. Flavell (New York: Academic Press), 301–323.

Flavell, R. B. (1986). Repetitive DNA and chromosome evolution in plants. *Phil. Trans. R. Soc. Lond. B* 312, 227–242. doi: 10.1098/rstb.1986.0004

Fry, K., and Salser, W. (1977). Nucleotide sequences of HS-α satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. *Cell* 12, 1069–1084. doi: 10.1016/0092-8674(77)90170-2

Gall, J. G., Cohen, E. H., and Polan, M. L. (1971). Repetitive DNA sequences in drosophila. *Chromosoma* 33, 319–344. doi: 10.1007/BF00284948

Garrido-Ramos, M. (2015). Satellite DNA in plants: more than just rubbish. *Cytogenet. Genome Res.* 146, 153–170. doi: 10.1159/000437008

Garrido-Ramos, M. (2017). Satellite DNA: an evolving topic. *Genes* 8:230. doi: 10.3390/genes8090230

Grewal, S. I. S., and Elgin, S. C. R. (2007). Transcription and RNA interference in the formation of heterochromatin. *Nature* 447, 399–406. doi: 10.1038/nature05914

Guenatri, M., Bailly, D., Maison, C., and Almouzni, G. (2004). Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J. Cell Biol.* 166, 493–505. doi: 10.1083/jcb.200403109

Hartley, G., and O'Neill, R. J. (2019). Centromere repeats: hidden gems of the genome. *Genes* 10:223. doi: 10.3390/genes10030223

Hollister, J. D., and Gaut, B. S. (2009). Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* 19, 1419–1428. doi: 10.1101/gr.091678.109

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., et al. (2009). The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* 41, 1275–1281. doi: 10.1038/ng.475

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877. doi: 10.1101/gr.9.9.868

Jiménez-Ruiz, J., Ramírez-Tejero, J. A., Fernández-Pozo, N., Leyva-Pérez, M. D. L. O., Yan, H., de la Rosa, R., et al. (2020). Transposon activation is a major driver in the genome evolution of cultivated olive trees (*Olea europaea* L.). *Plant Genome* 13:e20010. doi: 10.1002/tpg2.20010

Katsiotis, A., Hagidimitriou, M., Douka, A., and Hatzopoulos, P. (1998). Genomic organization, sequence interrelationship, and physical localization using in situ hybridization of two tandemly repeated DNA sequences in the genus *Olea*. *Genome* 41, 527–534. doi: 10.1139/g98-045

Kentner, E. K., Arnold, M. L., and Wessler, S. R. (2003). Characterization of high-copy-number retrotransposons from the large genomes of the Louisiana iris species and their use as molecular markers. *Genetics* 164, 685–697. doi: 10.1093/genetics/164.2.685

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/BF01731581

King, K., Jobst, J., and Hemleben, V. (1995). Differential homogenization and amplification of two satellite DNAs in the genus *Cucurbita* (Cucurbitaceae). *J. Mol. Evol.* 41, 996–1005. doi: 10.1007/BF00173181

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Levinson, G., and Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4, 203–221.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430, 471–476. doi: 10.1038/nature02651

Lisch, D. (2013). How important are transposons for plant evolution? *Nat. Rev. Genet.* 14, 49–61. doi: 10.1038/nrg3374

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., et al. (2010). The gypsy database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39, D70–D74. doi: 10.1093/nar/gkq1061

Lorite, P., Garcia, M. F., Carrillo, J. A., and Palomeque, T. (2001). A new repetitive DNA sequence family in the olive (*Olea europaea* L.). *Hereditas* 134, 73–78. doi: 10.1111/j.1601-5223.2001.00073.x

Ma, J., and Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 12404–12410. doi: 10.1073/pnas.0403715101

Macas, J., Novak, P., Pellicer, J., Čížková, J., Koblížková, A., Neumann, P., et al. (2015). In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe *Fabeae*. *PLoS One* 10:e0143424. doi: 10.1371/journal.pone.0143424

Marcon, H. S., Domingues, D. S., Silva, J. C., Borges, R. J., Fa, M., Filippi, B., et al. (2015). Transcriptionally active LTR retrotransposons in *eucalyptus* genus are differentially expressed and insertionally polymorphic. *BMC Plant Biol.* 15:198. doi: 10.1186/s12870-015-0550-1

Mascagni, F., Barghini, E., Giordani, T., Rieseberg, L. H., Cavallini, A., and Natali, L. (2015). Repetitive DNA and plant domestication: variation in copy number and proximity to genes of LTR-retrotransposons among wild and cultivated sunflower (*Helianthus annuus*) genotypes. *Genome Biol. Evol.* 7, 3368–3382. doi: 10.1093/gbe/evv230

Mascagni, F., Cavallini, A., Giordani, T., and Natali, L. (2017b). Different histories of two highly variable LTR retrotransposons in sunflower species. *Gene* 634, 5–14. doi: 10.1016/j.gene.2017.08.014

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

Mascagni, F., Usai, G., Natali, L., Cavallini, A., and Giordani, T. (2018b). A comparison of methods for LTR-retrotransposon insertion time profiling in the *Populus trichocarpa* genome. *Caryologia* 71, 85–92. doi: 10.1080/00087114.2018.1429749

Mascagni, F., Vangelisti, A., Giordani, T., Cavallini, A., and Natali, L. (2018a). Specific LTR-Retrotransposons show copy number variations between wild and cultivated sunflowers. *Genes* 9:433. doi: 10.3390/genes9090433

Mascagni, F., Vangelisti, A., Usai, G., Giordani, T., Cavallini, A., and Natali, L. (2020). A computational genome-wide analysis of long terminal repeats retrotransposon expression in sunflower roots (*Helianthus annuus* L.). *Genetica* 148, 13–23. doi: 10.1007/s10709-020-00085-4

Mazurczyk, M., and Rybaczek, D. (2015). Replication and re-replication: different implications of the same mechanism. *Biochimie* 108, 25–32. doi: 10.1016/j.biochi.2014.10.026

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., et al. (2013). Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.* 41, W597–W600. doi: 10.1093/nar/gkt376

Mehrotra, S., Goel, S., Raina, S. N., and Rajpal, V. R. (2014). Significance of satellite DNA revealed by conservation of a widespread repeat DNA sequence among angiosperms. *Appl. Biochem. Biotechnol.* 173, 1790–1801. doi: 10.1007/s12010-014-0966-3

Mehrotra, S., and Goyal, V. (2014). Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics Proteomics Bioinformatics* 12, 164–171. doi: 10.1016/j.gpb.2014.07.003

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., et al. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* 14:R10. doi: 10.1186/gb-2013-14-1-r10

Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E., and Plohl, M. (2015). Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosom. Res.* 23, 583–596. doi: 10.1007/s10577-015-9483-7

Minelli, S., Maggini, F., Gelati, M., Angiolillo, A., and Cionini, P. (2000). The chromosome complement of *Olea europaea* L.: characterization by differential staining of the chromatin and in-situ hybridization of highly repeated DNA sequences. *Chromosom. Res.* 8, 615–619. doi: 10.1023/A:1009286008467

Natali, L., Cossu, R. M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., et al. (2013). The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC Genomics* 14:686. doi: 10.1186/1471-2164-14-686

Neumann, P., Koblizkova, A., Navrátilová, A., and Macas, J. (2006). Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. *Genetics* 173, 1047–1056. doi: 10.1534/genetics.106.056259

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

Nikolov, I., and Taddei, A. (2016). Linking replication stress with heterochromatin formation. *Chromosoma* 125, 523–533. doi: 10.1007/s00412-015-0545-6

Novák, P., Hřibová, E., Neumann, P., Koblížková, A., Doležel, J., and Macas, J. (2014). Genome-wide analysis of repeat diversity across the family *Musaceae*. *PLoS One* 9:e98918. doi: 10.1371/journal.pone.0098918

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

Palacios-Gimenez, O. M., Koelman, J., Palmada-Flores, M., Bradford, T. M., Jones, K. K., Cooper, S. J., et al. (2020). Comparative analysis of morabine grasshopper genomes reveals highly abundant transposable elements and rapidly proliferating satellite DNA repeats. *BMC Biol.* 18, 1–21. doi: 10.1186/s12915-020-00925-x

Peacock, W. J., Brutlag, D., Goldring, E., Appels, R., Hinton, C. W., and Lindsey, D. L. (1974). The organization of highly repeated DNA sequences in *Drosophila melanogaster* chromosomes. *Cold Spring Harbor Lab. Press* 38, 405–416. doi: 10.1101/SQB.1974.038.01.043

Pezer, Ž., Brajković, J., Feliciello, I., and Ugarković, D. (2012). "Satellite DNA-mediated effects on Genome Regulation," in *Repetitive DNA*. ed. M. A. Garrido-Ramos (Basel: Karger Publishers), 153–169.

Phillippy, A. M., Schatz, M. C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* 9:R55. doi: 10.1186/gb-2008-9-3-r55

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16, 1262–1269. doi: 10.1101/gr.5290206

Plohl, M., Meštrović, N., and Mravinac, B. (2012). "Satellite DNA evolution," in *Repetitive DNA*. ed. M. A. Garrido-Ramos (Basel: Karger Publishers), 126–152.

Proost, S., Pattyn, P., Gerats, T., and Van de Peer, Y. (2011). Journey through the past: 150 million years of plant genome evolution. *Plant J.* 66, 58–65. doi: 10.1111/j.1365-313X.2011.04521.x

Quesada del Bosque, M. E., López-Flores, I., Suárez-Santiago, V. N., and Garrido-Ramos, M. A. (2014). Satellite-DNA diversification and the evolution of major lineages in *Cardueae* (*CarduoideaeAsteraceae*). *J. Plant Res.* 127, 575–583. doi: 10.1007/s10265-014-0648-9

Quesada del Bosque, M. E., López-Flores, I., Suárez-Santiago, V. N., and Garrido-Ramos, M. A. (2013). Differential spreading of Hin fI satellite DNA variants during radiation in *Centaureinae*. *Ann. Bot.* 112, 1793–1802. doi: 10.1093/aob/mct233

Rao, G., Zhang, J., Liu, X., Lin, C., Xin, H., Xue, L., et al. (2021). De novo assembly of a new *Olea europaea* genome accession using nanopore sequencing. *Horticult. Res.* 8:64. doi: 10.1038/s41438-021-00498-y

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

Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., and Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci. Rep.* 6:28333. doi: 10.1038/srep28333

Sader, M., Vaio, M., Cauz-Santos, L. A., Dornelas, M. C., Vieira, M. L. C., Melo, N., et al. (2021). Large vs small genomes in *Passiflora*: the influence of the mobilome and the satellitome. *Planta* 253, 1–18. doi: 10.1007/s00425-021-03598-0

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274, 765–768. doi: 10.1126/science.274.5288.765

Sanseverino, W., Hénaff, E., Vives, C., Pinosio, S., Burgos-Paz, W., Morgante, M., et al. (2015). Transposon insertions, structural variations, and SNPs contribute to the evolution of the melon genome. *Mol. Biol. Evol.* 32, 2760–2774. doi: 10.1093/molbev/msv152

Schmidt, T., and Heslop-Harrison, J. S. (1993). Variability and evolution of highly repeated DNA sequences in the genus *Beta*. *Genome* 36, 1074–1079. doi: 10.1139/g93-142

Schmidt, T., and Heslop-Harrison, J. (1998). Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends Plant Sci.* 3, 195–199. doi: 10.1016/S1360-1385(98)01223-0

Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science* 191, 528–535. doi: 10.1126/science.1251186

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542. doi: 10.1093/bioinformatics/btl117

Swaminathan, K., Varala, K., and Hudson, M. E. (2007). Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics* 8:132. doi: 10.1186/1471-2164-8-132

Szybalski, W. (1968). Use of cesium sulfate for equilibrium density gradient centrifugation. *Methods Enzymol.* 12, 330–360. doi: 10.1016/0076-6879(67)12149-6

Tenaillon, M. I., Hufford, M. B., Gaut, B. S., and Ross-Ibarra, J. (2011). Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol.* 3, 219–229. doi: 10.1093/gbe/evr008

Thakur, J., Packiaraj, J., and Henikoff, S. (2021). Sequence, chromatin and evolution of satellite DNA. *Int. J. Mol. Sci.* 22:4309. doi: 10.3390/ijms22094309

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 114, E9413–E9422. doi: 10.1073/pnas.1708621114

Usai, G., Mascagni, F., Natali, L., Giordani, T., and Cavallini, A. (2017). Comparative genome-wide analysis of repetitive DNA in the genus *Populus* L. *Tree Genet Genom.* 13:96. doi: 10.1007/s11295-017-1181-5

Usai, G., Vangelisti, A., Simoni, S., Giordani, T., Natali, L., Cavallini, A., et al. (2021). DNA modification patterns within the transposable elements of the fig (*Ficus carica* L.) genome. *Plan. Theory* 10:3. doi: 10.3390/plants10030451

Vangelisti, A., Mascagni, F., Giordani, T., Sbrana, C., Turrini, A., Cavallini, A., et al. (2019). Arbuscular mycorrhizal fungi induce the expression of specific retrotransposons in roots of sunflower (*Helianthus annuus* L.). *PloS One* 14:e0212371. doi: 10.1371/journal.pone.0212371

Vitte, C., Fustier, M.-A., Alix, K., and Tenaillon, M. I. (2014). The bright side of transposons in crop evolution. *Brief. Funct. Genomics* 13, 276–295. doi: 10.1093/bfgp/elu002

Vondrak, T., Ávila Robledillo, L., Novák, P., Koblížková, A., Neumann, P., and Macas, J. (2020). Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *Plant J.* 101, 484–500. doi: 10.1111/tpj.14546

Wendel, J. F., Lisch, D., Hu, G., and Mason, A. S. (2018). The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr. Opin. Genet. Dev.* 49, 1–7. doi: 10.1016/j.gde.2018.01.004

Wicker, T., and Keller, B. (2007). Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res.* 17, 1072–1081. doi: 10.1101/gr.6214107

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

Conflict of Interest: CT was employed by CSIRO Agriculture & Food, Narrabri, NSW (Australia).

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 01 July 2022
doi: 10.3389/fpls.2022.901733

Evolutionary Dynamics of the Repeatome Explains Contrasting Differences in Genome Sizes and Hybrid and Polyploid Origins of Grass Loliinae Lineages

María Fernanda Moreno-Aguilar[1]*, Luis A. Inda*[1,2]*, Aminael Sánchez-Rodríguez*[3]*, Itziar Arnelas*[3] *and Pilar Catalán*[1,4]*

[1] Escuela Politécnica Superior de Huesca, Universidad de Zaragoza, Huesca, Spain, [2] Instituto Agroalimentario de Aragón, Universidad de Zaragoza, Centro de Investigación y Tecnología Agroalimentaria, Zaragoza, Spain, [3] Departamento de Ciencias Biológicas y Agropecuarias, Universidad Técnica Particular de Loja, Loja, Ecuador, [4] Grupo de Bioquímica, Biofísica y Biología Computacional, Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Unidad Asociada al CSIC, Zaragoza, Spain

OPEN ACCESS

Edited by:
Ruslan Kalendar,
University of Helsinki, Finland

Reviewed by:
Ales Kovarik,
Academy of Sciences of the Czech Republic (ASCR), Czechia
Sebastián Pita,
Universidad de la República, Uruguay

***Correspondence:**
Pilar Catalán
pcatalán@unizar.es

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

Received: *22 March 2022*
Accepted: *25 May 2022*
Published: *01 July 2022*

Citation:
Moreno-Aguilar MF, Inda LA, Sánchez-Rodríguez A, Arnelas I and Catalán P (2022) Evolutionary Dynamics of the Repeatome Explains Contrasting Differences in Genome Sizes and Hybrid and Polyploid Origins of Grass Loliinae Lineages. Front. Plant Sci. 13:901733. doi: 10.3389/fpls.2022.901733

The repeatome is composed of diverse families of repetitive DNA that keep signatures on the historical events that shaped the evolution of their hosting species. The cold seasonal Loliinae subtribe includes worldwide distributed taxa, some of which are the most important forage and lawn species (fescues and ray-grasses). The Loliinae are prone to hybridization and polyploidization. It has been observed a striking two-fold difference in genome size between the broad-leaved (BL) and fine-leaved (FL) Loliinae diploids and a general trend of genome reduction of some high polyploids. We have used genome skimming data to uncover the composition, abundance, and potential phylogenetic signal of repetitive elements across 47 representatives of the main Loliinae lineages. Independent and comparative analyses of repetitive sequences and of 5S rDNA loci were performed for all taxa under study and for four evolutionary Loliinae groups [Loliinae, Broad-leaved (BL), Fine-leaved (FL), and Schedonorus lineages]. Our data showed that the proportion of the genome covered by the repeatome in the Loliinae species was relatively high (average $\sim$ 51.8%), ranging from high percentages in some diploids (68.7%) to low percentages in some high-polyploids (30.7%), and that changes in their genome sizes were likely caused by gains or losses in their repeat elements. Ty3-gypsy Retand and Ty1-copia Angela retrotransposons were the most frequent repeat families in the Loliinae although the relatively more conservative Angela repeats presented the highest correlation of repeat content with genome size variation and the highest phylogenetic signal of the whole repeatome. By contrast, Athila retrotransposons presented evidence of recent proliferations almost exclusively in the *Lolium* clade. The repeatome evolutionary networks showed an overall topological congruence with the nuclear 35S rDNA phylogeny and a geographic-based structure for some lineages. The evolution of the Loliinae repeatome suggests a plausible scenario of recurrent allopolyploidizations followed by diploidizations that generated the large genome sizes of BL diploids as well as large genomic rearrangements in highly hybridogenous lineages that caused massive repeatome and genome contractions in the Schedonorus and Aulaxyper polyploids. Our study has contributed to disentangling the impact of the repeatome dynamics on the genome diversification and evolution of the Loliinae grasses.

Keywords: diploidized paleo-allopolyploids, genome size diversification, Festuca, Lolium, phylogenetic signal, repeatome, transposable elements, 5S loci

INTRODUCTION

Comparative genomic studies have demonstrated that the repetitive DNA fraction is largely present in the nuclear genome of most plants (Pellicer et al., 2018). It is composed of diverse families of mobile elements (retrotransposons and transposons), which constitute the bulk of the predominant repeats, and of tandem satellite repeats, which can make up 10–20% of the genome (Macas et al., 2015). Although the constitution of the repetitive elements is complex and differs, sometimes by some orders of magnitude, among taxa (Hidalgo et al., 2017), there is an overall agreement on the impact that the dynamics of the repetitive elements have had in the variation of the genome size and its evolution across the angiosperms (Dodsworth et al., 2015; Pellicer et al., 2018). Alternative hypotheses have been launched to explain both the causes and the mechanisms of the plant repeatome turnovers. The "polyploid genome shock" hypothesis that postulates genomic reshuffling and mobility of the repetitive elements in hybrid and polyploid plants as a response to the sudden combination of distinct genomes and multiple copies of them (McClintock, 1984) has resulted, in some cases, in a rapid increase of repeats in the genomes after rounds of polyploidizations. The resulting polyploid genomes show additive patterns and equivalent genome size expansions (McCann et al., 2018). However, other plants do not show a proliferation of the repetitive elements in the allopolyploids, or only a gradual and low increase or decrease in their derived subgenomes (Chen et al., 2020). In contrast, other plant groups have experienced the opposite trend, with high-level polyploids exhibiting a drastic reduction in genome size and a considerable shrinkage of their repeatome relative to that of their diploid and low-level polyploid relatives (Chen, 2007; Parisod et al., 2010). The removal of the repetitive elements from the genome, attributed to several recombination mechanisms, and the driven forces that balance the expansions and contractions of the repeatome are still poorly known (Fedoroff, 2012; Drouin et al., 2021). In some exhaustively studied plants (*Gossypium*, *Brachypodium*) the abundance of some retrotransposon families and their apparent facility to proliferate (e.g., centromeric transposons) are interpreted as causing increased genome size, while the ability of other families to recombine and lose repeats are considered potential mechanisms for maintaining reduced genome size (Chen et al., 2020; Stritt et al., 2020). The dynamics of some repetitive elements, especially transposable elements (TEs) insertions, has been also related to the expression of some core or dispensable genes, although their mobility does not seem to substantially affect their regulation (Gordon et al., 2017) but can be affected by epigenetic effects (Chen, 2007; Fedoroff, 2012; Negi et al., 2016).

A comprehensive repetitive DNA analysis of plant genomes is still hampered by the unavailability of assembled and annotated genomes for many groups with complex and large genomes (Michael, 2014). In most cases it has been circumvented by using genome skim approaches and repeatome graph-topology analysis (Weiss-Schneeweiss et al., 2015; Garcia et al., 2020). Several studies have demonstrated that similarity-based clustering of low coverage genome sequencing reads, which confidentially represent 0.50–0.01× of the total haploid genome coverage, is proportional to the genomic abundance and longitude of the corresponding repeat-types (Macas et al., 2015; Pellicer et al., 2018) and could therefore be used to quantify them. The utility of the Repeat Explorer 2 bioinformatics tools for the quantification and annotation of repeats in plants (Novák et al., 2020) has been implemented by phylogenetic and distance-based network methods and by multivariate statistical methods that have corroborated the phylogenetic signal of the repeatome in various groups of angiosperm (Vitales et al., 2020a,b; Herklotz et al., 2021). It has also been supplemented by 5S rDNA graph-based clustering methods which have successfully corroborated the identity of the ancestral progenitor genomes of several polyploid plants (Garcia et al., 2020; Vozárová et al., 2021).

The grass subtribe Loliinae (*Festuca* and other close genera, like *Lolium*) constitutes one of the main lineages of the temperate pooids, both in number of species and in ecological and economic importance (Catalán, 2006; Kopecký and Studer, 2014). The Loliinae include more than 600 accepted species, Catalán (2006; Plants of the World On-line[1], accessed 3rd May 2022) which are distributed in cool seasonal and tropical mountainous regions of the five continents (Minaya et al., 2017; Moreno-Aguilar et al., 2020). The Loliinae species have large genomes ranging from 4.1 Gbp/2C to 23.6 Gbp/2C (Loureiro et al., 2007; Šmarda et al., 2008). Although these taxa show a uniform chromosome base number of $x = 7$ and ploidy levels ranging from diploids to dodecaplois, they exhibit striking differences in monoploid genome sizes, showing a 2.5-fold range decrease in chromosome size and C-values from more ancestral BL lineages (Drymanthele, Scariosae, Subbulbosae) to more recently evolved FL lineages (Festuca, Aulaxyper) (Catalán, 2006; Šmarda et al., 2008). In contrast, the heterochromatin pattern is inversely correlated with the genome size pattern, showing a rank increase of 7.5 between the same groups. However, this pattern is not homogeneous, as the early diverging fine-leaved Eskia lineage and the recently evolved broad-leaved Schedonorus-Lolium lineage

[1] http://www.plantsoftheworldonline.org/taxon/urn:lsid:ipni.org:names:328907-2

revealed independent intermediate karyotype patterns between the BL and FL groups (Catalán, 2006). Genome size analyses of Loliinae and other close Poeae suggested that the ancestor of Loliinae probably underwent a two-fold genome size enlargement (and parallel GC enrichment) relative to its close relatives, which was later followed by dramatic reductions, especially in the rapidly evolving FL Loliinae group (Šmarda et al., 2008). Nonetheless, alternative scenarios could involve large genome size increase only in the BL lineage or parallelisms in the most ancestral BL and FL lineages (Catalán, 2006). A genome downsizing trend has been detected in the fine-leaved Loliinae and in the polyploids, for which more pronounced genome losses have been hypothesized to have occurred in allopolyploids with large progenitor genomes than in autopolyploids with small progenitor genomes (Loureiro et al., 2007; Šmarda et al., 2008). However, none of these hypotheses have been tested yet through genomic analyses. There is a general lack of knowledge on the repetitive elements of the Loliinae genomes except for some chromosome barcoding markers in meadow fescue (Křivánková et al., 2017; Ebrahimzadegan et al., 2019) and the characterization of repeats and centromeric elements in eight species of tall fescues and relatives (Zwyrtková et al., 2020). Apart from these works, no other study has exhaustively explored the composition and dynamics of repetitive elements through a complete representation of the Loliinae.

Here, we have investigated the repeatome of 47 representatives of all the phylogenetic lineages recognized so far within the Loliinae (Inda et al., 2008; Minaya et al., 2017; Moreno-Aguilar et al., 2020) aiming to elucidate the potential role of repeats in the striking differences in genome size and in the evolution of both genomes and species. The objectives of our study are: (i) to characterize and quantify the repetitive elements of representatives of the BL and FL Loliinae and identify single or preponderant repeats in some groups; (ii) to test the plausible correlation between genome size and abundance of the repeats; (iii) to identify repeat types that could have contributed to the expansions or contractions of genomes and their relationships with the ploidy levels, the nature of the polyploidy and the phylogenetic positions of the groups; (iv) to assess the phylogenetic value of repeats using phylogenetic reconstructions and phylogenetic signal approaches; and (v) to test alternative hypotheses about which lineages were affected by repeat proliferation or contraction and the putative paleo-hybrid origin of BL diploids with large genome sizes using mobile and satellite repeat data analysis.

MATERIALS AND METHODS

Sampling, Cytogenetic Data and Genome Skim Sequencing

Forty-seven samples of diploid and polyploid taxa of Loliinae, representing its main broad-leaved (BL, 13 samples), fine-leaved (FL, 17) and Schedonorus (17) groups, were used in the study [**Table 1** and **Supplementary Table 1** (taxonomic ranks and authorships)]. Classification of samples into groups was based on previous phylogenetic frameworks (Minaya et al., 2017; Moreno-Aguilar et al., 2020). The sampling included taxa analyzed genomically for the first time within the BL (*Festuca scabra*, South African lineage; *F. mekiste*, Tropical Africa lineage) and FL (*F. rubra*, Aulaxyper lineage) groups plus the genome skim data generated in a previous study for representatives of other BL and FL lineages (Moreno-Aguilar et al., 2020). We obtained a large taxonomic representation of the Schedonorus group through the additional sequencing of species not studied molecularly (*F. dracomontana*, *F. gudoschnikovii*, *Lolium saxatile*) or genomically (*F. gigantea*, *F. simensis*, *Micropyropsis tuberosa*) before, and from a wide coverage of other tall fescues (*F. arundinacea*, *F. atlantigena*) and raygrasses (*L. canariense, L. perenne, L. persicum, L. rigidum*) (**Table 1** and **Supplementary Table 1**). The 47 selected taxa represent the 20 evolutionary lineages currently recognized within the Loliinae (Minaya et al., 2017; Moreno-Aguilar et al., 2020). They constitute a suitable test-bed case for investigating the putative role of repeat type dynamism in the genomic evolution of the major Loliinae lineages and their contrasting changes in genome size (Catalán, 2006; Šmarda et al., 2008). They could be also used to assess the potential phylogenetic value of the repeat elements at the subtribal level.

Cytogenetic knowledge of Loliinae taxa varies enormously. Besides relatively well scrutinized groups of economic importance, like some members of the Schedonorus, Aulaxyper, and Festuca lineages (Catalán et al., 2004; Šmarda et al., 2008; Minaya et al., 2017), cytogenetic data are missing for other species, especially for taxa from poorly studied taxonomic groups or less explored areas (Catalán, 2006). Chromosome number (2*n*) and genome size (2C/pg) data were estimated for some of the studied samples using DAPI-stained meristematic root cells and flow cytometry analysis following the protocols of Jenkins and Hasterok (2007) and Doležel et al. (2007), respectively. Chromosome staining was performed with the DAPI fluorescent marker (4′,6-diamino-2 phenylindole) and counts were done using a Motic BA410 fluorescence microscope. The nuclear DNA content of *F. asplundii, F. caldasii, F. chimborazensis, F. fontqueri* and *F. procera* were calculated from silica gel dried leaves using nuclei isolated from similarly processed leaves of *Pisum sativum* L. "Ctirad" (9,09 pg/2C) as standard. Nuclei were stained with propidium iodide and samples were analyzed using a CyFlow Ploidy Analyser SYSMEX. At least 5,000 nuclei were analyzed per sample and each sample (two replicates) was analyzed three times. Only measurements with coefficient of variation < 3.5% were recorded. Ploidy levels were inferred from chromosome counts (2*n*) and GS estimations performed in the same accessions used in our genomic study and through contrasted GS and 2*n* values obtained in conspecific accessions that showed similar values. However, cytogenetic data is still lacking for some unstudied species that could only be analyzed genomically using museomic approaches (Moreno-Aguilar et al., 2020; **Table 1** and **Supplementary Table 1**).

Total DNA for the 15 newly sampled Loliinae taxa was extracted from herbarium specimens (MHU, PRE, UZ, VLA) and silica gel dried leaf tissues from plants growing in the University of Zaragoza – High Polytechnic School of Huesca common garden (**Supplementary Table 1**). Isolation of DNA and its concentration quantification and quality evaluation for genome skimming sequencing was performed following the procedures indicated in Moreno-Aguilar et al. (2020). PCR free libraries were quantified by Library Quantification Kit for Illumina Platforms (Roche Kapa Biosystems). Genomic sequencing of a multiplexed pool of KAPA libraries was performed on a HiSeq4000 or HiSeq 2500 (TruSeq SBS Kit v4, Illumina, Inc.) in paired-end mode (2 × 100 bp) in the Centro Nacional de Análisis Genómicos (CNAG, Barcelona) as described in Moreno-Aguilar et al. (2020). Illumina paired-end (PE) reads were checked using FASTQC and the adapters and low quality sequences were trimmed and removed using TRIMMOMATIC (Bolger et al., 2014). The Loliinae genomic samples used in downstream analysis contained between 6.1 and 40.6 million reads (average 18.0 million reads) with insert sizes ranging between 190 and 300 bp (**Supplementary Table 2**).

Repeat Clustering and Annotation, and 5S rDNA Graph-Clustering Analysis

Identification of the composition and proportion of repetitive elements in the 47 Loliinae species studied was performed from similarity graph-based clustering analysis of filtered PE reads using the Repeat Explorer pipeline of RepeatExplorer2 (RE2)[2]. It was performed through the Galaxy platform as described by Novák et al. (2020). The clustering analysis of individual samples was fed with 500000 PE reads per sample in order to attain the recommended genome coverage (0.1–0.5×) of each taxon (**Supplementary Table 2**). The clustering was conducted employing default RE2 settings (90% similarity, minimum overlap = 55; cluster size threshold = 0.01%) and long queue (max runtime). Automated RE2 annotation of clusters was used to quantify the clusters and to calculate the proportions of repetitive elements in each sample. Plastid and mitochondrial DNA clusters were removed prior to downstream analyses. Comparative clustering analysis was performed for four evolutionary groups (Loliinae, BL, FL, Schedonorus) due to the impossibility of computing it for all the studied samples (47) in a single run of Galaxy employing the same RE2 configuration used for the individual analyses. The Loliinae group was reduced to 38 samples, representing all its main lineages, while the BL, FL and Schedonorus groups contained the same samples used in the individual analysis except the BL group which had two additional Schedonorus samples (**Table 1** and **Supplementary Tables 1, 2**). The comparative clustering analyses were conducted using the maximum number of randomly sampled PE reads that could be processed, representing ∼0.08–0.2× of genome coverage for each species (**Supplementary Table 2**). Automated RE2 repeat annotation was used to quantify the clusters and to estimate the proportions of repeats among the compared samples within each group. Plastid and mitochondrial DNA clusters were also removed from each group prior to downstream analyses.

Sequences of 5S ribosomal DNA genes from 43 out of the 47 studied Loliinae samples were searched using the TAREAN pipeline of RE2 (Garcia et al., 2020; Novák et al., 2020). The input for the 5S rDNA clustering analysis consisted of 500000 PE reads per sample, covering the expected lengths of the 5S rDNA for most of the Loliinae genomes ranging 4.2–20.7 Gbp (**Supplementary Table 1**). The clustering was performed using default TAREAN tool settings (BLAST threshold of 90%, similarity across 55% of the read to identify reads to each cluster, minimum overlap = 55, cluster threshold = 0.01%, minimum overlap for assembly = 40). The 5S rDNA clusters were found in the TAREAN tandem reports. Their shapes were characterized by a connected component index parameter (C) and their k-mer score was calculated as the sum of frequencies of all k-mers used for consensus sequence reconstruction (Garcia et al., 2020). The 5S rDNA cluster graph topologies were visually inspected and classified into graph groups (type 1, simple circular-shaped graph; type 2, complex graph with two or more loops where the interconnected loops represent IGS spacers) (Garcia et al., 2020). We examined the 5S graphs to detect potential variation of 5S rDNA loci and to identify presumable hybrids and allopolyploids. A RE2 5S rDNA sequence of *Festuca pratensis* (360 bp) was used as reference for a Geneious Prime read-mapping assembly of the 5S rDNA of the four Loliinae species (*F. caldasii, F. gigantea, F. gracillima, F. gudoschnikovii*) that could not be retrieved directly from TAREAN due to insufficient number of reads in the cluster for graphical analysis (see **Table 4**). Newly generated 5S rDNA sequences of Loliinae were deposited in GenBank under accessions codes ON248974–ON249019.

Plastome and Nuclear rDNA Phylogenies of Loliinae

Genome skimming PE reads were used to assemble and annotate the plastomes and the nuclear 35S rDNA of the newly sequenced Loliinae samples (**Table 1**). Plastome assembly was performed with Novoplasty v.2.7.1 (Dierckxsens et al., 2017) following the procedures indicated in Moreno-Aguilar et al. (2020) and using as reference the *Festuca pratensis* plastome sequence (JX871941). The 35S rDNA cistron (transcribed region ETS-18S-ITS1-5.8S-ITS2-25S) was assembled using the read-mapping and merging strategy of Moreno-Aguilar et al. (2020) using Geneious Prime and the *F. ovina* 35S rDNA sequence (MT145295) as reference. Newly generated plastome and 35S rDNA sequences of Loliinae were deposited in Genbank under accessions codes SAMN27777779–SAMN27777788 and ON243855–ON243864 (**Table 1**). Multiple sequence alignments (MSAs) of these sequences, together with those of the previously studied Loliinae samples and the *Oryza sativa* and *Brachypodium distachyon* outgroups (**Supplementary Table 1**), were performed with MAFFT v.7.031b (Katoh et al., 2002), visually inspected with Geneious Prime and debugged with trimAl v.1.2rev59 (imposing parameter-*automated*1) (Capella-Gutiérrez et al., 2009). The filtered plastome (133552 bp) and 35S rDNA cistron (6431 bp) MSA data sets were used to compute Maximum likelihood (ML) phylogenetic trees with IQTREE (Nguyen et al., 2015). Independent ML searches were performed imposing the best-fit nucleotide substitution model selected by ModelFinder for each partition, according to the Bayesian Information Criterion (BIC), and branch support for the best tree was estimated from

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

1,000 ultrafast bootstrap replicates (BS) (Chernomor et al., 2016; Kalyaanamoorthy et al., 2017).

The well resolved plastome and 35S ML trees were topologically contrasted to each other using the Kishino-Hasegawa (KH), Shimodaira-Hasegawa (SH), and Shimodaira Approximately Unbiased (AU) tests with resampling estimated log-likelihood (RELL) optimization and one million bootstrap replicates in PAUP* (Swofford, 2003). As all the pairwise tests showed that each topology did not significantly differ ($p < 0.001$) from the other topology, we constructed a combined ML plastome + 35S tree with IQTREE imposing the respective nucleotide substitution model to each partition and the procedures indicated above. To account for potential incomplete lineage sorting (Kubatko and Degnan, 2007) and to investigate the possibility that a single concatenated plastome + 35S data set could generate topological errors in the phylogeny, we run a parallel phylogenetic analysis with the same data set but modeling the coalescence process using the Singular Value Decomposition quartets (SVDq) approach implemented in Paup*, which uses a variant of Quartet FM (Reaz et al., 2014) to combine quartet trees into a species tree. We imposed the SVDQuartets nquartets = all seed = 2 nthreads = 4 bootstrap = 1000 options with a multispecies coalescent tree model and the quartet assembly algorithm QFM. Bootstrap support of branches was shown on the tree obtained from SVDquartests + Paup* analysis. Since the topology of the SVDq tree (**Supplementary Figure 1A**) was equal to that of the ML tree (**Supplementary Figure 1B**), we selected the strong to relatively well supported ML tree for downstream analysis. Different ML subtrees were computed from the whole combined plastome + 35S data matrix using the respective subsets of taxa of each of the four Loliinae evolutionary groups employed in the repeatome analyses (Loliinae, BL, FL, Schedonorus). These ML tree cladograms were used to estimate the phylogenetic signal of the repeats of each partition (see below). A MSA was also generated for the 5S rDNA sequences of Loliinae and close outgroups (**Supplementary Table 1**) and a ML phylogenetic tree was computed with this data set following the procedures indicated above.

Repeatome Trees and Evolutionary Networks of Loliinae, Phylogenetic Signal of Repeats

Evolutionary analyses were performed with the repeat data obtained from the comparative clustering of repeats for the Loliinae, BL, FL and Schedonorus groups. Distance-based phylogenetic trees and networks were computed from pairwise genetic distances between the repeat contents of the species included in the datasets. First, calculated repeat sequence similarity matrices for the observed/expected number of interspecies edges for each of the most abundant repeat clusters selected by RE2 were converted to Euclidean distances via the *dist* option of the *proxy* package in R (Euclidean matrices). Second, the same repeat sequence similarity matrices were transformed into distance matrices by calculating the inverse of their values as described by Vitales et al. (2020b) (inverse matrices). In both cases, the clusters with incomplete information (NA or zero values) for the similarity comparisons between species pairs were discarded from the analysis. Next, Neighbor-Joining phylogenetic trees were constructed for each repetitive element using either the Euclidean or the inverse distance matrices and the *NJ* function of *ape* package (Paradis et al., 2004) in R. Finally, consensus networks were built from all the repeat NJ trees with SplitsTree4 (Huson and Bryant, 2006) for each group.

The combined plastome + 35S ML subtrees were used to test the potential phylogenetic signal of different types of repeats of each group using Blomberg's K (Blomberg et al., 2003) with the *phylosig* function of the package *phytools* (Revell, 2012) in R. For these tests, K values > 1 indicate that the repeatome traits have more phylogenetic signal than expected, values ~ 1 that traits are consistent with the tree topology (phylogenetic signal), and values ~ 0 that there is no influence of shared ancestry on trait values (phylogenetic independence).

Correlations of Repeat Amounts and Genome Size Variation and Global Diversity Analysis of Repeat Types in Loliinae

The potential contribution of the various groups of repeat types and the repeatome to the variation in genome size (1Cx) observed between and within Loliinae lineages was tested using the data from the comparative analysis and by linear regression model analyses (Pearson correlation coefficient) with the *ggscatter* function from the *ggpubr* package in R. The respective contributions of repeats to pairwise differences in genome sizes were estimated following Macas et al. (2015). To correct for potential phylogeny-based bias, phylogenetically independent contrasts (PIC) methods were previously applied to the data using the *pic* option of the *ape* package in R. Correlations could be only performed for the 23 Loliinae species with known genome size (**Table 1**), representing all the main subtribal groups, and using absolute amounts (Mbp) of repeats calculated for individual species (**Supplementary Table 1**). In addition, we also tested whether there were significant differences in repeat amount for different repeat families obtained from the individual analysis through Kruskal–Wallis rank tests using the *multcompView* and *ggpubrr* packages in R. Furthermore, to investigate the levels of conservatism or diversity of the repeat types that most contributed to genome size variation in Loliinae (23 species with known genome sizes) we performed a genome landscape search for the global variability of these individual repeat types across the Loliinae genomes. We pooled the pairwise similarity values of reads, retrieved from the RE2 outputs (hitsort files), for each species and repeat type in a separate dataset and evaluated their similarities with respect to similarities of reads from the same repeat in all other species following Macas et al. (2015). We calculated intraspecific versus interspecific similarity hit ratios (Hs/Ho ratios) considering that conservative sequence repeats will produce similarity hits with about the same frequency for Hs and Ho, while diversified sequence repeats will generate similarity hits with different frequencies. We also calculated similarity hit ratios for the 5S tandem-repeat rDNA to compare its gene-conserved vs.

IGS-variable Hs/Ho ratios with those obtained from the other repeat elements analyzed.

RESULTS

Multiple Polyploidizations and Genome Size Diversification Across the Phylogeny of Loliinae

Chromosome counts and genome size data obtained for, respectively, 41 and 23 out of the 47 Loliinae taxa studied (**Table 1**) corroborated previous records but also revealed new findings about contrasting genome sizes between and within the BL and FL Loliinae lineages when mapped to the combined Loliinae tree (**Figure 1** and **Supplementary Figure 1B**). The inferred ploidy levels for the newly analyzed South American *F. asplundii* (6x), *F. caldasii* (4x), *F. chimborazensis* (subsp. *micacochensis*, 6x) and *F. procera* (4x) species (**Table 1**) confirmed the lack of Loliinae diploids in the southern hemisphere (Dubcovsky and Martínez, 1992; Catalán, 2006). Genome sizes ranged from 4.3 Gbp (*L. canariense*-2x; Schedonorus) and 4.82 Gbp (*F. ovina*-2x; FL) to 21.23 Gbp (*F. asplundii*-6x; FL), representing a near 5-fold (x4.9) increase within the Loliinae and the FL group. Monoploid genome sizes ranged from 2.02 Gbp (*V. ciliata*-4x; FL) to 4.98 Gbp (*F. caldasii*-4x; BL), representing a ×3.7 increase within the Loliinae (**Table 1** and **Supplementary Table 1**). Within the diploids, the broad-leaved species showed 2C genome sizes (*F. triflora*, 7.67 Gbp; *F. paniculata*, 7.48) 1.5x larger than those of the fine-leaved *Festuca* (*F. ovina*, 4.71) and some *Lolium* (*L. perenne*, 4.2) species, while the early diverging fine-leaved *F. eskia* (5.57) and other Schedonorus species (*F. fontqueri*, 5.52; *F. pratensis*, 6.36; *L. perenne*, 5.39; *L. rigidum*, 5.4; *L. persicum*, 6.26) displayed intermediate GS values between them (**Table 1** and **Supplementary Table 1**). A general trend of reduction in monoploid genome size was observed in some polyploid FL and Schedonorus taxa, showing lower values as ploidy level increased (FL: Aulaxyper: *F. rubra*-6x, 2.23 Gbp; American I: *F. chimborazensis*-6x, 2.2; Schedonorus: *F. arundinacea*-6x, 2.84; *F. atlantigena*-8x, 2.0; *F. letourneuxiana*-10x, 1.93). However, large 1Cx sizes were also detected among polyploid South-American Loliinae species nested either within the BL (Central and South American: *F. caldasii*-4x, 4.98) or the FL (American II: *F. procera*-4x, 3.64; *F. asplundii*-6x, 3.46) clades (**Table 1**, **Supplementary Table 1**, and **Figure 1**).

The combined plastome + 35S rDNA ML tree (**Figure 1** and **Supplementary Figure 1B**) was overall congruent with the phylogenies of Minaya et al. (2017) and Moreno-Aguilar et al. (2020) for the divergences of the main Loliinae lineages. The combined tree retrieved a robust topology which was also congruent with those of the well supported plastome and less supported 35S rDNA trees (**Supplementary Figures 1B–D**). The Loliinae phylogeny showed the split of the sister BL and FL clades (**Figure 1**) and divergences within the clades similar to those indicated in Moreno-Aguilar et al. (2020) except for the position of the BL Subulatae-Hawaiian lineage which was nested within the FL clade in the current tree (**Figure 1**). The largely sampled Schedonorus clade showed the branching-off of the 'Mahgrebian' and 'European' sister clades; the latter included the split of the *Festuca* gr. *arundinacea* allopolyploids from the rest, although their respective nesting positions swapped between their 'European' plastome and 'Mahgrebian' 35S rDNA trees (**Figure 1** and **Supplementary Figures 1B–D**). The remaining Schedonorus lineages of the 'European' clade showed the early divergences of diploids followed by those of polyploids and a reversal trend to diploidization in the recently split *Lolium* clade (**Figure 1**). Diploid and polyploid lineages were spread across the BL and FL clades of the Loliinae tree (**Figure 1**). Although several of the early diverging BL lineages are predominantly or uniquely made up of diploids (Drymanthele, Lojaconoa, Subbulbosae), other early splits contain exclusively low-to-high polyploids (South African, Central-South American). A similar trend of more ancient to more recent origins of polyploids could be observed within the Schedonorus and FL clades. Low-to-high polyploids have evolved in all FL lineages and several of them are formed exclusively by polyploids (American-Neozeylandic, American I, American-Pampas, Psilurus-Vulpia, Subulatae-Hawaiian, American II, Afroalpine) (**Figure 1**).

The Loliinae Repeatome

The annotated repeats found by RE2 in the individual analyses showed large differences in repeat types and amounts among the 47 Loliinae samples and lineages (**Table 2**, **Supplementary Table 2**, **Figure 1**, and **Supplementary Figure 1E**). The proportion of the holoploid genome occupied with repeats ranged from 30.69% (*F. letourneuxiana*-10x) to 68.8% (*L. persicum*-2x), with a mean across Loliinae of 51.8% (**Table 2**, **Figure 1**, and **Supplementary Figure 1E**). The highest percentages corresponded to diploid taxa of the Schedonorus group (e.g., *Lolium* spp., *M. tuberosa*, *F. simensis*; >60%) and diploid or polyploid taxa of the BL group (e.g., *F. lasto*-2x; *F. triflora*-2x, *F. scabra*-4x, Central-South American spp.-4x-6x, *F. africana*-10x, plus FL *F. molokaiensis*; >57%) and the lowest to high-polyploid taxa of the Schedonorus group (Mahgrebian-4x-10x, *F. arundinacea*-6x; <40%) and to diploid and high-polyploid species of the FL Aulaxyper group (*F. francoi*-2x, *F. rubra*-6x; <46%) (**Table 2**; **Figure 1**, and **Supplementary Figure 1E**). LTR-Gypsy and LTR-Copia retrotransposons represented the major fractions of repeatome in the studied genomes followed by Class II TIR-transposons and Satellite repeats (**Table 2** and **Supplementary Figure 1E**).

LTR-Gypsy Retand elements were the most represented repeats in almost all genomes, especially within the BL and Schedonorus groups, where they covered >10% and up to 20% of several Subbulbosae, Leucopoa, Central-South American, 'European,' *F.* gr. *arundinacea* and *Lolium* genomes, as well as two genomes of the BL and FL groups (*F. molokaiensis*, *V. ciliata*). Only the BL Tropical-South African and the FL American II and Aulaxyper genomes showed low coverages (<2%) of Retand repeats (**Table 2** and **Figure 1**). The more heterogeneous LTR-Gypsy Tekay and Athila elements were also well represented in some genomes, the former in the BL genomes (*F. scabra* 14%,

TABLE 1 | Taxa included in the repeatome analysis of Loliinae.

Taxon	Group	Locality	*2n*	Ploidy	2C(pg)	1Cx (pg)	1Cx (Mbp)	GenBank accession no.		
								Plastome	35S rDNA	5S rDNA
Festuca africana	BL	Uganda: Gahinga	70	10x	–	–	–	SAMN14647044	MT145277	**ON248974**
Festuca amplissima	BL	Mexico: Chihuahua	42	6x	–	–	–	SAMN14647045	MT145278	**ON248975**
Festuca caldasii	BL	Ecuador: Catamayo	**28**	4x	**20.36**	5.09	4978.02	SAMN14647047	MT145280	**ON248977**
Festuca durandoi	BL	Portugal: Serra Arga	14	2x	14.66 (4x)	3.66	3584.86	SAMN14647050	MT145283	**ON248980**
Festuca lasto	BL	Cádiz: Jerez	14	2x	–	–	–	SAMN14647058	MT145291	**ON248989**
Festuca mekiste	BL	Kenya: Mt. Elgon	–	–	–	–	–	**SAMN27777779**	**ON243855**	**ON248992**
Festuca molokaiensis	BL	United States: Hawai: Molokai	–	–	–	–	–	SAMN14647061	MT145294	**ON248993**
Festuca paniculata	BL	Spain: Caceres	14	2x	7.65	3.83	3740.85	SAMN14647064	MT145297	**ON248996**
Festuca parvigluma	BL	China: Baotianman	28	4x	–	–	–	SAMN14647065	MT145298	**ON248997**
Festuca scabra	BL	S Africa: Cathedral P.	28	4x	–	–	–	**SAMN27777781**	**ON243857**	**ON249003**
Festuca spectabilis	BL	Bosnia-H: Troglav	42	6x	–	–	–	SAMN14647071	MT145304	**ON249004**
Festuca superba	BL	Argentina: Jujuy	56	8x	–	–	–	SAMN14647072	MT145305	**ON249005**
Festuca triflora	BL	Morocco: Rif Mnts.	**14**	2x	**7.84**	3.92	3833.76	SAMN14647073	MT145306	**ON249006**
Festuca abyssinica	FL	Tanzania: Kilimanjaro	28	4x	–	–	–	SAMN14647043	MT145276	**ON248973**
Festuca asplundii	FL	Ecuador: Saraguro	42	6x	**21.23**	3.54	3460.49	SAMN14647046	MT145279	**ON248976**
Festuca capillifolia	FL	Morocco: Ifrane	14	2x	–	–	–	SAMN14647048	MT145281	**ON248978**
Festuca chimborazensis	FL	Ecuador: Chimborazo	**42**	6x	**13.48**	2.25	2197.24	SAMN14647049	MT145282	**ON248979**
Festuca eskia	FL	Spain: Picos de Europa	14	2x	5.7	2.85	2787.3	SAMN14647051	MT145284	**ON248981**
Festuca fimbriata	FL	Argentina: Apóstoles	42	6x	–	–	–	SAMN14647053	MT145286	**ON248983**
Festuca francoi	FL	Portugal: Terceira	12	2x	–	–	–	SAMN14647057	MT145290	**ON248984**
Festuca gracillima	FL	Argentina: Trra.Fuego	42	6x	–	–	–	SAMN14647055	MT145288	**ON248986**
Festuca holubii	FL	Ecuador: Saraguro	–	–	–	–	–	SAMN14647056	MT145289	**ON248988**
Festuca ovina	FL	Rusia: Gatchinskii Ra.	14	2x	4.82	2.41	2356.98	SAMN14647062	MT145295	**ON248994**
Festuca pampeana	FL	Argentina: Ventana	56	8x	–	–	–	SAMN14647063	MT145296	**ON248995**
Festuca procera	FL	Ecuador: Chimborazo	28	4x	**14.88**	3.72	3638.16	SAMN14647067	MT145299	**ON248999**
Festuca pyrenaica	FL	Spain: Tobacor	28	4x	–	–	–	SAMN14647068	MT145300	**ON249000**
Festuca pyrogea	FL	Argentina: Trra.Fuego	–	–	–	–	–	SAMN14647069	MT145302	**ON249001**
Festuca rubra	FL	Argentina: Trra.Fuego	42	**6x**	13.68	2.28	2229.84	**SAMN27777780**	**ON243856**	**ON249002**
Megalachne masafuerana	FL	Chile: Masafuera	–	–	–	–	–	SAMN14647075	MT145308	**ON249018**
Vulpia ciliata	FL	Spain: Ontígola	28	4x	8.28	2.07	2024.46	SAMN14647076	MT145309	**ON249009**
Festuca a. arundinacea	Sch	Spain: Ferrol	42	6x	17.46	2.91	2845.98	**SAMN27777774**	**ON243850**	**ON249007**
Festuca a. atlantigena	Sch	Morocco: Atlas Mnts	56	8x	16.22	2.03	1982.895	**SAMN27777775**	**ON243851**	**ON248990**
Festuca a. letourneuxiana	Sch	Morocco: Atlas Mnts	70	10x	19.7	1.97	1926.66	SAMN14647059	MT145292	**ON249010**
Festuca dracomontana	Sch	SAfrica:Haernertsburg	–	–	–	–	–	**SAMN27777776**	**ON243852**	**ON249011**
Festuca fenas	Sch	Spain	28	4x	10.48	2.62	2562.36	SAMN14647052	MT145285	**ON248982**
Festuca fontqueri	Sch	Morocco: Rif Mnts	14	2x	**5.54**	2.77	2709.06	SAMN14647054	MT145287	**ON249008**
Festuca gigantea	Sch	Norway	42	6x	20.75	3.46	3382.25	**SAMN27777777**	**ON243853**	**ON248985**
Festuca gudoschnikovii	Sch	Russia: Yermakovskii	28	4x	–	–	–	**SAMN27777778**	**ON243854**	**ON248987**
Festuca mairei	Sch	Morocco: Atlas Mnts	28	4x	10.04	2.51	2454.78	SAMN14647060	MT145293	**ON248991**
Festuca pratensis	Sch	United Kingdom: England	14	2x	6.5	3.25	3178.5	SAMN14647066	MT145301	**ON248998**
Festuca simensis	Sch	Kenya: Mt. Kenya	28	4x	–	–	–	**SAMN27777782**	**ON243858**	**ON249012**
Lolium canariense	Sch	Spain: Canary Islands	14	2x	4.3	2.15	2102.7	**SAMN27777783**	**ON243859**	**ON249013**
Lolium perenne	Sch	United Kingdom: Wales	14	2x	5.51	2.76	2694.39	**SAMN27777784**	**ON243860**	**ON249014**
Lolium persicum	Sch	Georgia	14	2x	6.4	3.2	3129.6	**SAMN27777785**	**ON243861**	**ON249015**
Lolium rigidum	Sch	Turkey	14	2x	5.49	2.75	2684.61	**SAMN27777786**	**ON243862**	**ON249017**
Lolium saxatile	Sch	Spain: Fuerteventura	14	2x	–	–	–	**SAMN27777787**	**ON243863**	**ON249016**
Micropyropsis tuberosa	Sch	Spain: Almonte	14	2x	–	–	–	**SAMN27777788**	**ON243864**	**ON249019**

*Loliinae group (BL, broad-leaved Loliinae; FL, fine-leaved Loliinae; Sch, Schedonorus), chromosome number (2n), ploidy level, genome size (2C, pg), monoploid genome size (1Cx, pg; 1Cx, Mbp) and GenBank accession codes for plastome and nuclear ribosomal 35S and 5S genes are given for each sample. Values in bold correspond to new data generated in this study. Hyphens indicate lack of 2n and/or 2C/1Cx data for some taxa. See **Supplementary Table 1** for additional information on taxonomic ranks and taxon authorship, detailed localities and vouchers, and sources of cytogenetic and genomic data.*

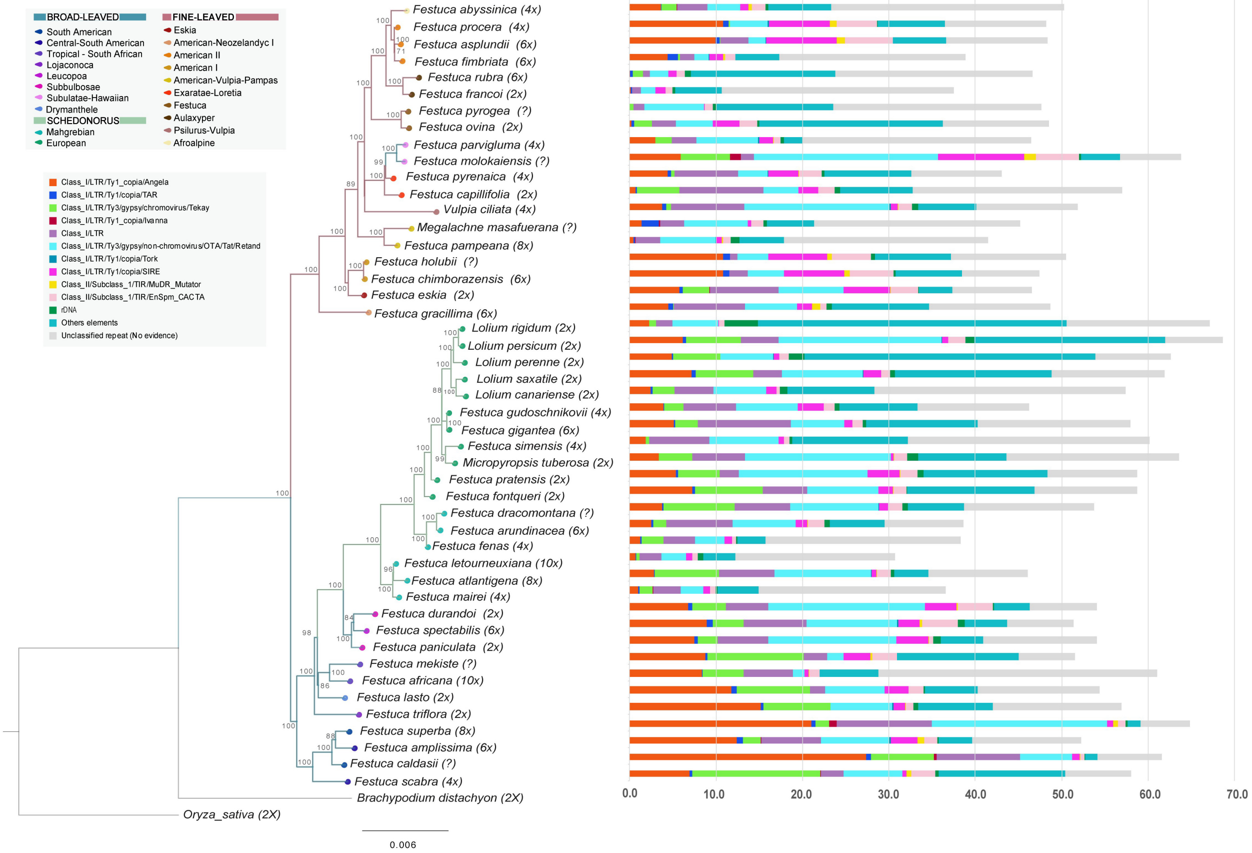


FIGURE 1 | Histograms of repeat contents per holoploid genome (1C) retrieved from the individual Repeat Explorer 2 analyses of the studied Loliinae samples mapped onto the Maximum Likelihood combined phylogenomic tree (plastome + nuclear 35S rDNA cistron) of Loliinae (color codes of Loliinae lineages are indicated in the chart). Color codes for repeat families are indicated in the corresponding inset charts. Scale bar: number of mutations per site.

F. mekiste 11%) and the latter in the *Lolium* genomes (*L. perenne*, 25%; *L. rigidum* 23%). In contrast, those elements generally had low coverages (<2%) in FL genomes (**Table 2** and **Figure 1**). Other LTR-Gypsy families were only moderately represented in some groups, such as Ogre in the Tropical and South African genomes (e.g., *F. mekiste*, 7.9%; *F. africana*, *F. scabra*, 4.6%) and *L. rigidum* (4.8%), and CRM in several Schedonorus genomes (e.g., *L. persicum* 5.1%, *F. pratensis* 4.3%) although they showed low coverages (<2%) in most of the remaining genomes. The LTR-Gypsy OTA, Reina and Tat families were only residually present in a few genomes (**Table 2**).

LTR-Copia Angela elements were the second most frequent repeat family in all Loliinae genomes. They were highly represented in the genomes of Central-South American taxa in both the BL (12–27%) and FL (9.8–10.8%) groups, relatively abundant in all remaining BL genomes (6.6–8.8%), moderately abundant in Schedonorus genomes (except the 'Mahgrebian' taxa, <2%) and in FL *F. eskia* and BL *F. molokaiensis* (5.7–7.2%), and poorly represented in the remaining FL genomes (<2%) (**Table 2** and **Figure 1**). LTR-Copia SIRE elements showed moderate to low frequency in all genomes except in *F. molokaiensis* (10%) and FL Eskia, American I and American II genomes (5.4–7.5%). Other LTR-Copia families (Ale, Ikeros, Ivanna, TAR, Tork) were only residually represented in a few Loliinae genomes (**Table 2** and **Figure 1**). TIR Class II transposons were found less frequently in Loliinae genomes; only CACTA elements were present in all taxa although they were only moderately represented in some FL American I, American II and Hawaiian genomes and in BL Subbulbosae and Leucopoa genomes (4–5.5%). Representation of other transposon elements (Mutator, Harbinger, hAT) in Loliinae genomes was only residual (**Table 2** and **Figure 1**). Some of the less frequent Class I and Class II repetitive elements were only represented in a very small fraction of some particular genomes (e.g., Reina in *L. saxatile*; hAT in *M. masafuerana*; Tat in *F. simensis*; **Table 2**). Tandem satellite repeats were generally moderately to poorly represented in most Loliinae genomes, except for their relatively high representation in FL *F. procera* and *F. pyrogea* (13.3%) and Schedonorus *F. simensis* (12%) and its moderate representation in FL Exaratae, Festuca and Aulaxyper genomes (4.2–5.9%). Kruskal–Wallis rank tests performed for each of the Loliinae repeat elements found significant differences for Retand, CRM, Tekay, Angela, Ivanna, Ale, LTR, CACTA, Mutator, Harbinger, rDNA

TABLE 2 | Genome proportion of repeats estimated by Repeat Explorer 2 for individual Loliinae samples (estimations per holoploid genome, 1C).

Loliinae taxon and phylogenetic group	Class I/LTR/Ty1_copia								Class I/LTR/Ty3_gypsy									Class II/Subclass_1/TIR											
	Ale	Angela	Ikeros	Ivanna	SIRE	TAR	Tork	Ty1_Copia	OTA	Athila	Tat	Ogre	Retand	CRM	Tekay	Reina	Ty3_Gypsy	EnSpm_ CACTA	Hat	MuDR-Mutator	PIF – Harbinger	rDNA (5S-45S)	Satellite	Mobile element	Class I/ LTR (conflict evidence)	Class I/ LINE	Repeat (conflicting evidences)	Unclassified (No evidence)	Total (%)
Broad-Leaved																													
Festuca africana Tropical-South African	0	8.42	0	0	0.46	0.06	0	0	0	0.3	0	4.63	1.32	0.14	4.75	0	0	1.14	0	0.13	0	0.04	1.7	0	5.74	0	0	32.26	61.1
Festuca amplissima Central-South American	0	12.43	0.05	0.11	3.04	0.71	0.23	0	0	0.26	0	0.02	7.87	0.67	2.05	0	0	1.54	0	0.79	0	0.12	2.72	0	6.89	0.16	0	12.58	52.25
Festuca caldasii Central-South American	0.03	27.45	0	0.33	0.82	0.48	0.02	0	0	0.26	0	0.13	6.03	0.17	7.32	0	0	0.49	0	0.11	0	0.06	0.88	0	9.64	0	0	7.4	61.63
Festuca durandoi Subbulbosae	0	6.81	0	0	3.6	0.47	0	0	0	0.17	0	0.16	18.2	1.84	3.87	0	0	4.04	0	0.19	0.09	0.1	0.96	0	4.9	0.02	0.87	7.81	54.12
Festuca lasto Drymanthele	0	11.83	0.09	0	2.73	0.59	0.02	0	0	3.42	0	0	6.85	1.76	8.51	0	0	1.72	0	0.03	0	0.17	0.84	0	1.76	0.01	0	14.11	54.46
Festuca mekiste Tropical-South African	0	8.79	0	0.01	3.08	0.24	0	0	0	2.77	0	7.92	1.91	0.35	11.14	0	0	2.86	0	0.27	0	0.03	3.01	0	2.68	0	0	6.5	51.57
Festuca molokaiensis Subulatae-Hawaiian	0	5.94	0.03	1.26	9.96	0.03	0	0	0	0.01	0	1.55	21.35	0.26	5.71	0	0	4.95	0	1.37	0	0.25	1.12	0	1.49	0.02	1.49	7.09	63.85
Festuca paniculata Subbulbosae	0	7.51	0	0	3.75	0.43	0.02	0	0	0.45	0	0	14.83	0.81	2.3	0	0	0.51	0	0.02	0.03	0.89	0.82	2.77	5.82	0.03	0	13.16	54.14
Festuca parvigluma Subulatae-Hawaiian	0	2.99	0.12	0	1.57	0.04	0.15	0	0	0.03	0	0.01	7.16	0.65	1.93	0	0	0.82	0	0.07	0	0.38	1.34	0	2.79	0	0	26.43	46.47
Festuca scabra South African	0	6.95	0.09	0.11	0.47	0.36	0	0	0	5.69	0	4.6	6.86	1.28	14.78	0	0	2.81	0	0.54	0	0.4	2.96	0	2.57	0	0	7.61	58.08
Festuca spectabilis Leucopoa	0	8.94	0	0	2.47	0.73	0.1	0	0	0.07	0	0.12	10.48	2.38	3.54	0	0	4.13	0	0.35	0.05	0.77	2.04	0	7.32	0.2	0	7.71	51.41
Festuca superba Central-South American	0.03	21.07	0	0.9	0.68	0.49	0.01	0	0	0	0	0.01	20.31	0.05	1.54	0	0	0.92	0	0.51	0	0.26	1.4	0	11.01	0	0	5.71	64.9
Festuca triflora Lojaconoa	0	15.24	0	0	1.34	0.33	0.14	0	0	5.98	0	0.11	7.15	0.78	7.71	0	0	0.85	0	0.13	0	0.51	1.78	0	0	0	0	14.93	56.98
Schedonorus																													
Festuca a. arundinacea F.gr.arundinacea	0	2.52	0.06	0	1.36	0.29	0.01	0	0	3.07	0	0.08	7.31	1.27	1.47	0	0	1.92	0	0.03	0.07	0.63	1.53	0	7.66	0.09	0.24	9.09	38.67
Festuca a. Atlantigena F.gr.arundinacea	0	2.84	0.02	0	0.6	0.08	0.01	0	0	0.14	0	0	11.16	1.49	7.5	0	0	1.62	0	0.06	0.03	0.37	2.13	0	6.42	0.08	0.09	11.43	46.09
Festuca dracomontana F.gr.arundinacea	0	3.79	0.03	0	1.05	0.18	0.01	0	0	1.58	0	0.13	10.24	2.45	8.23	0	0	1.68	0	0	0	0.59	1.49	0	6.44	0.04	0.79	15.09	53.82
Festuca fenas Mahgrebian	0	1.29	0.02	0	0.83	0.16	0	0	0	1.15	0	0	3.4	0.8	2.5	0	0	0.45	0	0	0	0.21	1.21	0	3.69	0.02	0	22.64	38.38
Festuca fontqueri European	0	7.31	0.09	0	1.65	0.28	0.01	0	0	7.55	0	0.63	8.21	2	7.9	0	0	1.54	0	0.08	0.01	0.09	3.25	0	5.12	0.03	1.11	11.96	58.82
Festuca gigantea European	0	5.16	0	0	0.98	0.13	0.01	0	0	0.98	0	0	6.19	3.73	2.62	0	0	1.19	0	0	0.03	0.4	8.06	0	10.76	0.1	0	17.62	57.96
Festuca gudoschnikovii European	0	3.96	0	0	3.02	0.12	0	0	0	0.19	0	0	7.17	3.37	2.22	0	0	1.24	0	0	0.02	0.59	5.32	0	6.04	0.07	0	12.93	46.25
Festuca a. letourneuxiana Mahgrebian	0	0.73	0.01	0	0.71	0.08	0	0.12	0.01	1.13	0	0	2.85	0.8	0.43	0	0.01	0.62	0	0	0.01	0.63	1.61	0	2.53	0.02	0	18.39	30.7
Festuca mairei Mahgrebian	0	1.02	0.03	0.02	0.82	0.18	0	0.1	0	1.32	0	0	2.59	0.99	1.51	0	0	0.62	0	0	0	0.3	2.28	0	3.19	0	0	21.58	36.57
Festuca pratensis European	0.04	5.41	0.01	0	3.77	0.19	0	0	0	7.18	0	0.66	14.88	4.26	4.89	0	0	2.02	0	0.01	0	0.69	1.81	0	2.17	0.01	0.4	10.29	58.72
Festuca simensis European	0	1.9	0	0	0.62	0.01	0	0.02	0.02	0.37	0.02	0	8.04	0.91	0.4	0	0	0.66	0	0	0	0.3	12.01	0	6.95	0.01	0	28.01	60.23

(Continued)

TABLE 2 | (Continued)

Loliinae taxon and phylogenetic group	**Class I/LTR/Ty1_copia**								**Class I/LTR/Ty3_gypsy**									**Class II/Subclass_1/TIR**											
	Ale	**Angela**	**Ikeros**	**Ivanna**	**SIRE**	**TAR**	**Tork**	**Ty1_Copia**	**OTA**	**Athila**	**Tat**	**Ogre**	**Retand**	**CRM**	**Tekay**	**Reina**	**Ty3_Gypsy**	**EnSpm_ CACTA**	**Hat**	**MuDR-Mutator**	**PIF – Harbinger**	**rDNA (5S-45S)**	**Satellite**	**Mobile element**	**Class I/ LTR (conflict evidence)**	**Class I/ LINE**	**Repeat (conflicting evidences)**	**Unclassified (No evidence)**	**Total (%)**
Lolium canariense Lolium	0.11	2.49	0	0	1.22	0.23	0	0	0	0.05	0	0.39	6.04	2.6	2.53	0	0	0.41	0	0.02	0	0.81	6.93	0	4.53	0	0	29.09	57.46
Lolium perenne Lolium	0.07	4.9	0	0	0.64	0.17	0.01	0	0	25.26	0	2.79	6.12	1.75	5.47	0	0	1.09	0	0.04	0.09	1.83	2.03	0	0	0.04	1.64	8.68	62.63
Lolium persicum Lolium	0.11	6.2	0	0	0.73	0.41	0.04	0	0	9.18	0	1.15	18.86	5.15	6.34	0	0	1.97	0	0	0.19	1.02	4.87	0	4.33	0	1.52	6.65	68.71
Lolium rigidum Lolium	0.1	2.29	0	0	0.14	0.04	0	0	0	23.1	0	4.86	5.3	2.68	0.79	0	0	0.63	0	0	0.06	3.83	2.53	0	1.85	0	2.42	16.53	67.15
Lolium saxatile Lolium	0.18	7.25	0.04	0	2.13	0.44	0.02	0.01	0	7.23	0	0	9.39	1.57	6.67	0.01	0	1.03	0	0	0.65	0.56	1.76	0	3.29	0.02	6.64	13.03	61.91
Micropyropsis tuberosa European	0.05	3.38	0	0	0.33	0.01	0	0	0	0.02	0	0.05	16.89	3.58	3.93	0	0	1.5	0	0.02	1.02	1.3	3.99	0	6.07	0	1.47	20.01	63.64
Fine-Leaved																													
Festuca abyssinica Afroalpine	0	3.65	0	0.07	0.96	0.08	0	0	0	0.15	0	1.85	3.83	0.34	1.78	0	0	1.56	0	0.41	0	0.31	4.93	0	3.45	0	0	26.92	50.27
Festuca asplundii American II	0	9.97	0.52	0	8.17	0.53	0.13	0	0	1.31	0	0.82	1.96	0.67	0.02	0	0	5.52	0	0.93	0.01	0.04	1.83	0	3.27	0	0.97	11.75	48.41
Festuca capillifolia Exaratae	0	0.75	0	0	2.32	0.14	0.01	0	0	1.58	0	0.22	4.03	1.41	4.87	0	0	1.86	0	0	0	0.65	5.16	0	9.77	0	0	24.23	57.02
Festuca chimborazensis American I	0	10.89	0.06	0.02	6.98	0.67	0	0	0.01	1.06	0	0.06	4.17	1.16	0	0	0	5.04	0	0.65	0.09	0.3	4.09	0	2.15	0	1.07	8.99	47.45
Festuca eskia Eskia	0	5.77	0.02	0.06	5.36	0.4	0	0	0	0.73	0	0.18	7.51	0.72	3.13	0	0	3.18	0	0.09	0.03	0.11	1.11	0	7.92	0.06	1.04	9.18	46.59
Festuca fimbriata American II	0.01	4.47	0.15	0.04	1.6	1.18	0.12	0.43	0	0.04	0	0.34	1.66	0.09	0.21	0	0	1.18	0	0.2	0	0.05	3.91	0	1.62	0.02	0	21.54	38.87
Festuca francoi Aulaxyper	0	0.13	0.02	0	1.17	0.16	0.01	0.02	0.09	0.85	0	0.07	1.62	0.23	0.03	0	0	0.81	0	0.02	0	0.27	4.16	0	1.07	0	0	26.83	37.56
Festuca gracillima American-Neozeylandic	0	4.5	0.15	0.02	1.74	0.61	0	0	0	0.87	0	1.22	6.05	1.54	0.01	0	0	0.76	0	0.92	0	0.61	1.45	0	8.23	0	6.04	13.95	48.68
Festuca holubii American I	0	10.87	0.07	0	6.86	0.76	0	0	0	0.61	0	0.02	3.52	1.1	0.01	0	0	4.54	0	0.54	0.01	0.42	5.96	0	0.9	0.01	0.99	13.32	50.52
Festuca ovina Festuca	0.03	0.26	0.02	0	3.16	0.33	0	0	0	7.18	0	0.59	4.26	1.04	2.04	0	0	2.01	0	0	0.01	0.28	5.22	0	2.75	0.03	7.1	12.28	48.58
Festuca pampeana American Pampas	0	0.52	0	0.06	0.63	0.1	0.06	0	0	0	0	0.33	6.47	0.03	0	0	0	0.85	0	0.19	0	1.02	4.77	0	2.85	0	0	23.59	41.48
Festuca pirenaica Exaratae	0	10.88	0.36	0	7.1	0.6	0.12	0	0.01	0.87	0	0.34	4.32	1.48	0.21	0	0	4.67	0	0.8	0	0.05	2.66	0	0	0	2.09	11.66	48.21
Festuca procera American II	0.01	4.47	0.11	0	3.53	0.36	0.03	0	0	1.31	0	0.26	3.46	1.41	0.42	0	0	2.62	0	0.01	0	0.37	5.94	0	7.32	0.04	0.96	10.43	43.08
Festuca pyrogea Festuca	0.04	0.02	0	0	0.08	0	0	0.03	0	0	0	0	6.91	0.02	0.49	0	0	0.95	0	0	0	0.42	13.4	0	1.24	0	0	24.13	47.73
Festuca rubra Aulaxyper	0	0.08	0	0	0.96	0.3	0.01	0	0	8.55	0	0.61	2.11	0.69	1.24	0	0	0.95	0	0.01	0	0.7	6.56	0	0.76	0	0.32	22.79	46.65
Megalachne masafuerana American Pampas	0	1.44	0	0.18	0.38	1.98	0	0.02	0	0.1	0	0	7.4	3.31	0	0	0	1.4	0.03	0	0	0.4	1.88	0	2.74	0.04	0.14	23.85	45.28
Vulpia ciliata Psilurus-Vulpia	0.14	3.81	0	0	0.78	0.49	0.16	0	0.01	0.22	0	0.08	16.79	0.86	0.56	0	0	1.6	0	0.11	0	0.7	2.76	0	8.45	0.33	2.33	11.74	51.91
Mean±SD	**0.02**	**5.94**	**0.05**	**0.07**	**2.26**	**0.35**	**0.03**	**0.02**	**0.00**	**2.86**	**0.00**	**0.79**	**7.68**	**1.42**	**3.31**	**0.00**	**0.00**	**1.84**	**0.00**	**0.21**	**0.05**	**0.53**	**3.41**	**0.06**	**4.43**	**0.03**	**0.89**	**15.61**	**51.85**
Kruskal Wallis test	**30.99**	**37.17**	19.01	**35.24**	21.49	21.30	19.04	20.78	9.81	19.63	12.89	20.52	**31.43**	**31.25**	**30.49**	8.40	14.67	**24.84**	22.50	**32.30**	**23.49**	**24.54**	**23.56**	14.67	**28.22**	14.33	22.30	21.92	
Kruskal Wallis test *p.value*	**0.01**	**0.00**	0.16	**0.00**	0.09	0.09	0.16	0.11	0.78	0.14	0.53	0.11	**0.00**	**0.01**	**0.01**	0.87	0.40	**0.04**	0.07	**0.00**	**0.05**	**0.04**	**0.05**	0.40	**0.01**	0.43	0.07	0.08	

Kruskal–Wallis tests for significant differences in repeat proportions for each repetitive element across the studied samples. Significant values are highlighted in bold.

and satellite repeats when examined in the entire group of samples (**Table 2**).

Regression model analysis of repeat content and monoploid genome sizes differences among the 23 Loliinae species with known 2C data, after PIC correction, showed a strong correlation when data from all main repeats were combined (R^2 = 0.83, p = 1.8E-09), accounting for 65.2% differences in genome size between species (**Table 3** and **Figure 2**). Angela repeats presented the highest correlation (R^2 = 0.71, p = 5.44E-07), followed by TAR (R^2 = 0.54, p = 5.85E-05), Tekay (R^2 = 0.38, p = 0.0018), Ivanna (R^2 = 0.35, p = 0.002), LTR (R^2 = 0.27, p = 0.011) and Retand (R^2 = 0.21, p = 0.02) repeats, while the other repetitive elements did not show significant correlations. The Angela family also showed the highest contribution to pairwise differences in genome sizes (19.6%), followed by Retand (10.7%), Tekay (6.47%) and LTR (5.49%), while the contributions of the other families were <5% (**Table 3** and **Supplementary Figure 2**). Our genome landscape analysis of global variability of these individual repeat types across the Loliinae genomes showed different histogram profiles of Hs/Ho hit ratios (**Figure 3**). The histogram of control 5S rDNA sequences comprised a narrow major peak near zero on the log(Hs/Ho) x-axis, indicating that the ratios of intraspecific Hs to interspecific Ho hit frequencies were close to one, and thus reflected the high sequence conservation of the 5S genes. In contrast, this 5S rDNA histogram also included a wide right-hand tail of log(Hs/Ho) hit values ranging from 0.1 to 3, accounting for the high divergence of intergenic spacer sequences (IGS) of 5S rDNA. However, the histogram patterns of the ten repeats analyzed showed general Gaussian distributions for log(Hs/Ho) hit values (**Figure 3**). Among the repeats that contributed the most to genome size variation (**Table 3** and **Supplementary Figure 2**), Angela elements generated main peaks of log(Hs/Ho) values closer to zero in the histogram than those of Retand, LTR and Tekay elements (**Figure 3**), suggesting a slightly higher conservatism of the Angela sequences and a higher diversification of the Retand, LTR and Tekay sequences in the Loliinae genome landscape.

Repeatome Phylogenies of Loliinae and Phylogenetic Signal of Repeats

The results of the RE2 comparative analysis of Loliinae repeats recovered different types and numbers of shared or sample specific repetitive elements in each of the four

TABLE 3 | Pearson linear correlation of repeat abundance with genome size variation (1Cx) in Loliinae, after PIC correction, and contribution of individual repeats to the genome size differences between species.

Repeat type	Correlation to genome size		Abundance in the analyzed genomes [Mbp/1Cx]		Average contribution to pairwise differences in genome sizes [%]
	R^2	p-Value	Min	Max	
Angela	**0.71**	**5.44E-07**	1.775	1366.503	**19.6**
TAR	**0.54**	**5.85E-05**	1.172	24.058	**0.642**
Tekay	**0.38**	**0.00187**	0	364.516	**6.47**
Ivanna	**0.35**	**0.00281**	0	16.597	**0**
LTR	**0.27**	**0.0111**	0	480.094	**5.49**
Retand	**0.21**	**0.0265**	46.947	652.52	**10.7**
Tork	0.16	0.0566	0	5.454	0.0376
SIRE	0.14	0.0784	3.791	282.611	2.8
MuDR_Mutator	0.11	0.131	0	32.148	0.0986
EnSpm_CACTA	0.09	0.165	8.715	190.978	2.27
Ty1_Copia	0.08	0.18	0	2.514	0
Ty3_Gypsy	0.08	0.197	0	0.208	0
Mobile_element	0.06	0.257	0	103.646	0
Ikeros	0.05	0.285	0	17.96	0
LINE	0.05	0.314	0	6.74	0
OTA	0.03	0.397	0	0.379	0
Unclassified	0.03	0.438	197.426	611.73	4.08
CRM	0.03	0.443	8.348	161.049	0.751
Repeat	0.01	0.61	0	167.43	0
Ale	0.01	0.716	0	3.465	0
PIF_Harbinger.	0.01	0.737	0	5.893	0
rDNA_5S-45S	0.00	0.789	1.446	102.852	–0.152
Athila	0.00	0.852	1.146	680.565	0.778
Satellite	0.00	0.863	30.69	272.468	–0.0164
Ogre	0.00	0.93	0	130.467	0.183
All repeats	**0.83**	**1.8E-09**	591.539	3067.826	**65.2**

Only the most represented repeat types of Loliinae are shown. Significant values are highlighted in bold.

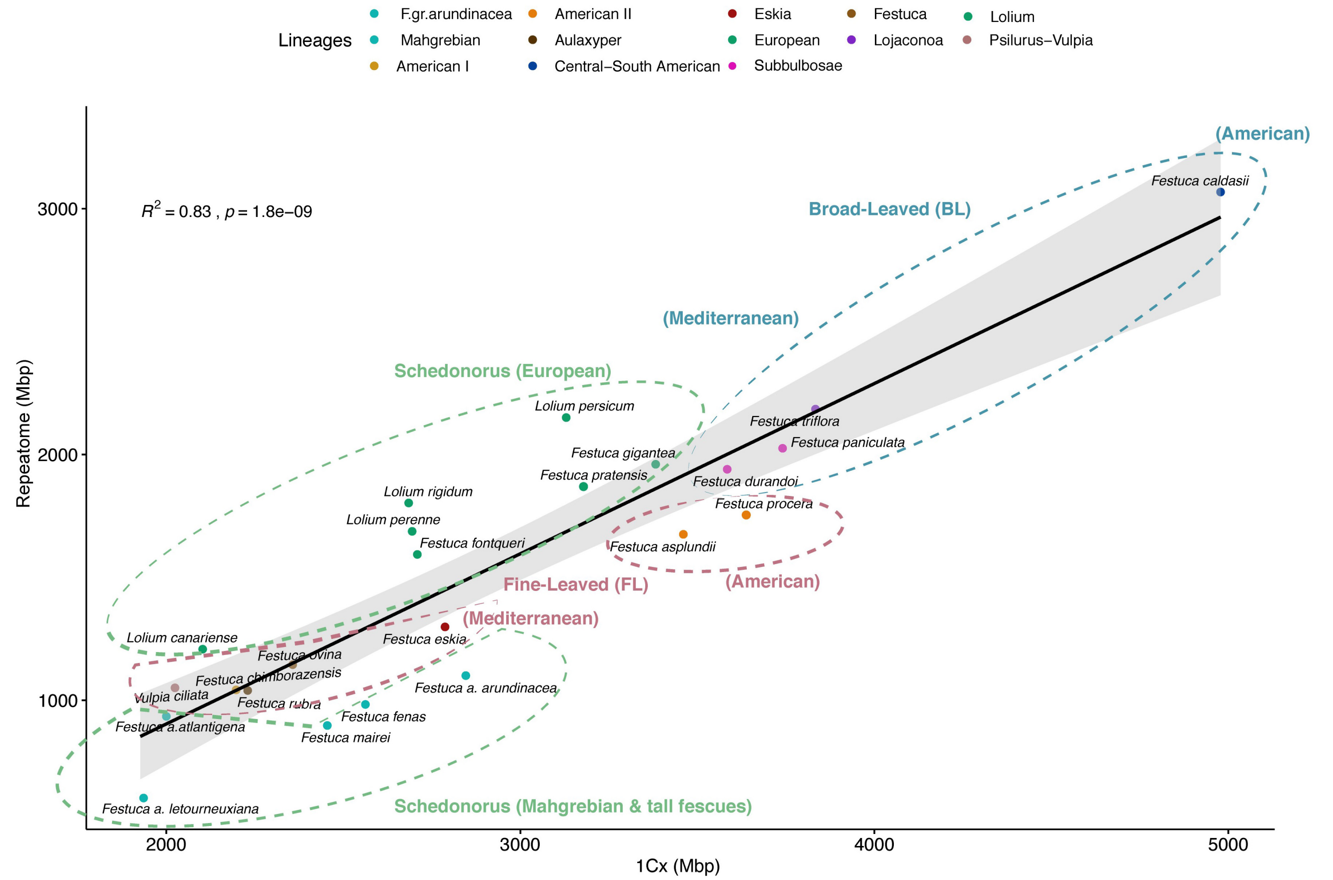


FIGURE 2 | Correlation plot of repeat content and genome size variation (1Cx) for the 23 Loliinae taxa with known genome sizes. Summed abundance values of the most represented repeat types obtained from the individual RE2 analysis. Pearson correlation analysis (R^2 = 0.83, p = 1.8e–09). Ellipses with dashed lines encircle the main Loliinae groups and subgroups [broad-leaved Loliinae (BL), blue; Schedonorus, green; fine-leaved Loliinae (FL), magenta]. Color codes of Loliinae lineages correspond to those indicated in **Figure 1**.

Loliinae evolutionary groups studied (**Supplementary Table 3**). RE2 annotated different numbers of tops clusters in each group [Loliinae: 337 clusters (total number of reads 2,659,145 (57%); minimum number of reads 468); FL: 308 (2,245,911 (57%); 395); BL: 336 (2,841,940 (64%); 443); Schedonorus: 270 (1,771,749 (65%); 274)] (**Supplementary Tables 3A–D**) representing presumably orthologous repeat families from different samples that were grouped together due to their high repeat sequence similarity (Macas et al., 2015). The number of top clusters used to build the NJ trees and networks was reduced in all groups after discarding clusters with NA or zero read values for some samples (Loliinae: 38 clusters; BL: 96; FL: 122; Schedonorus: 167) (**Supplementary Tables 4A–D**). Networks constructed from distance-based NJ trees computed with the Euclidean distances (**Figures 4A–D**) showed better resolutions than those obtained from NJ trees computed with the inverse distances (**Supplementary Figures 3A–D**); therefore, descriptions of repeatome phylogenies were based on the Euclidean networks. The unrooted Loliinae network showed three divergent groups corresponding to each of the main BL, FL and Schedonorus lineages (**Figure 4A**). In this network, the Schedonorus group was highly isolated from the others and, in contrast to its position in the Loliinae tree (**Figure 1**), it was closer to the FL group than to the BL group. Similarly, the fine-leaved *F. eskia* was closer to the BL group than to its own FL group. The unrooted BL network (**Figure 4B**) inferred a topology congruent with that of the BL lineage in the Loliinae tree except for the sister relationship of South African *F. scabra* with the other Tropical and South African taxa and the sister relationship of the two Subbulbosae species (*F. paniculata*/*F. durandoi*), resolutions that, however, matched those recovered from the 35S Loliinae tree (**Supplementary Figure 1C**). The unrooted FL network (**Figure 4C**) was generally consistent with the combined Loliinae tree except for the positions of the American I and American-Pampas taxa, which were closely related to the American II taxa; Afroalpine *F. abyssinica* was also close to them (**Figure 4C**). These phylogenetic topologies were also congruent with those retrieved in the 35S Loliinae tree (**Supplementary Figure 1C**).

The potential phylogenetic signal of the abundance of the repeat clusters (**Supplementary Tables 4A–D**) evaluated in different Loliinae subtrees, rendered significant *K* values for distinct clusters in each group (**Supplementary Table 5** and

Supplementary Figure 4). Within the Loliinae group, nine clusters (1 LTR, 4 Angela, 1 SIRE, 3 CACTA) had significant *K* values on the Loliinae tree cladogram, although only the *K* values of the four Angela clusters were >0.5. In contrast, within the FL group only four clusters (1 Angela, 2 Tekay, 1 repeat) had significant *K* values on the FL tree cladogram but all of them were ∼1. The BL and Schedonorus groups had 17 clusters that carried phylogenetic signal on their respective tree cladograms; however, whereas all the BL clusters (1 LTR, 3 Angela, 8 Tekay, 4 Athila, 1 Mutator) had *K* values close to 1, only nine out of the 17 Schedonorus clusters had *K* values ∼1 (3 LTR, 3 Repeat, 1 CRM, 1 Mutator, 1 Tekay) while the remaining eight cluster (6 LTR, 1 Athila, 1 Mutator) carried more phylogenetic signal than expected (*K* values > 1) (**Supplementary Table 5** and **Supplementary Figure 4**).

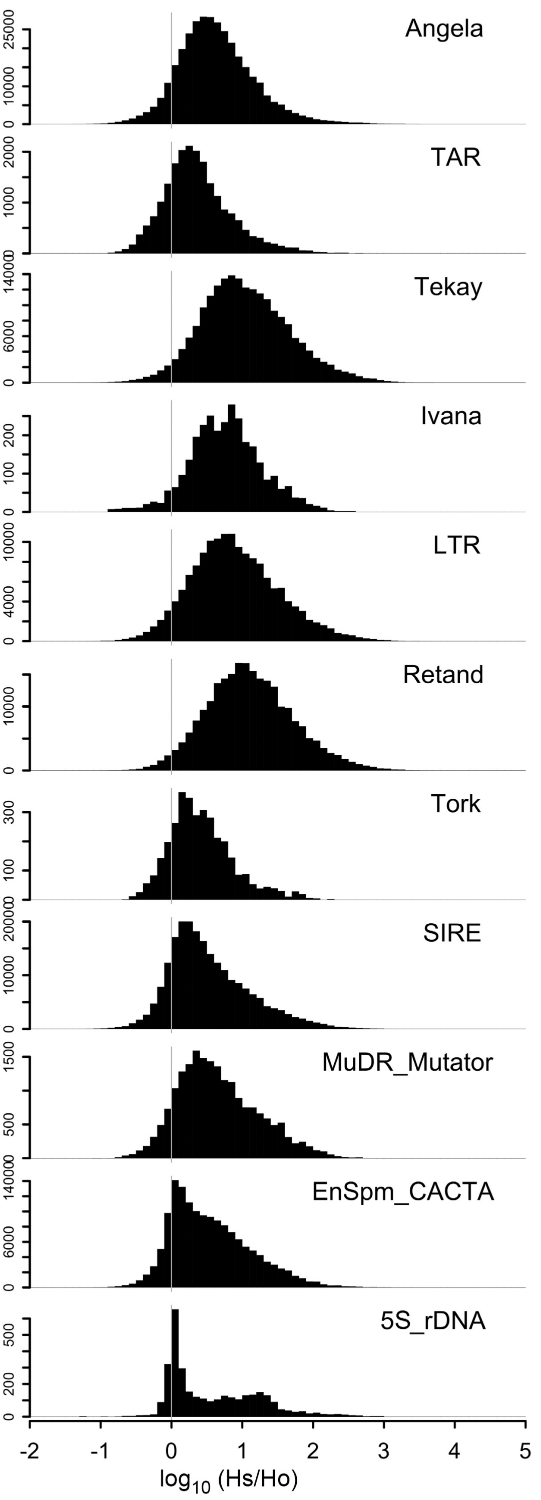


FIGURE 3 | Global variability of main repeat types and their sequence conservation across the Loliinae genome landscapes. Histograms show distributions of read similarity Hs/Ho hit ratios [frequencies of read similarity hits to reads from the same species (Hs) or to reads from all other species (Ho) (log scale, *x*-axis) and number of reads (*y*-axis)]. Hs/Ho ratios close to one (0 on the logarithmic scale) indicate sequence conservatism while larger values indicate sequence diversification.

5S rDNA Graph-Clusters of Loliinae

The Loliinae 5S rDNA region ranged from 245 to 316 bp in the Loliinae [a 120 bp 5S gene conserved in all taxa plus a variable IGS for specific taxa (range 125–196 bp); **Supplementary Table 1**]; the 5S MSA consisted of 316 bp (120 bp 5S gene; 196 bp IGS). The Loliinae 5S ML tree (**Supplementary Figure 5**) had poor support for most of its branches and was topologically incongruent with both the combined Loliinae tree (**Figure 1** and **Supplementary Figure 1B**) and the separate plastome and nuclear 35S rDNA trees (**Supplementary Figures 1C,D**). The only supported lineage was the Schedonorus clade (**Supplementary Figure 5**) although its internal resolution also departed from those of the other trees and was not considered further.

Analysis of the 5S rDNA clusters of 47 Loliinae species studied produced different types of simple and complex graphs that did not always match the expected shapes for their respective ploidy levels (**Table 4** and **Figure 5**). As expected, most graph topologies of diploid taxa corresponded to a simple circular graph that likely represents a single 5S gene family and locus. This was observed for most FL (*F. eskia, F. capillifolia, F. ovina*) and Schedonorus (*F. pratensis, F. fontqueri*, *M. tuberosa*, all five *Lolium* species) diploids. However, within the BL diploids one species showed a simple graph (*F. lasto*) but two species (*F. triflora, F. paniculata*) had complex graphs with two IGS loops interconnected by a junction section (coding region of the 5S gene), suggesting that the latter species could have two 5S ribotypes (**Figure 5**). Within Loliinae polyploids, 5S graph topologies ranged from those taxa showing complex graphs with a number of loops corresponding to their assumed number of 5S loci (tetraploid *F. pyrenaica*, two loops), to high polyploids with lower number of loops than expected based on their ploidy levels (decaploids *F. africana* and *F. letourneuxiana*, two loops), and low-to-high polyploids showing a simple graph (tetraploids *V. ciliata*, *F. parvigluma, F. procera, F. abyssinica, F. simensis, F. fenax, F. mairei, F. mekiste*; hexaploids *F. rubra, F. chimborazensis, F. asplundii, F. fimbriata, F. amplissima*; octoploids *F. pampeana, F. spectabilis, F. atlantigena, F. superba*). Loliinae species from the southern hemisphere with unknown ploidy level displaying complex 5S graphs (e.g., *F. pyrogea*, *M. masafuerana*; two loops) were identified as polyploids, while those displaying a single

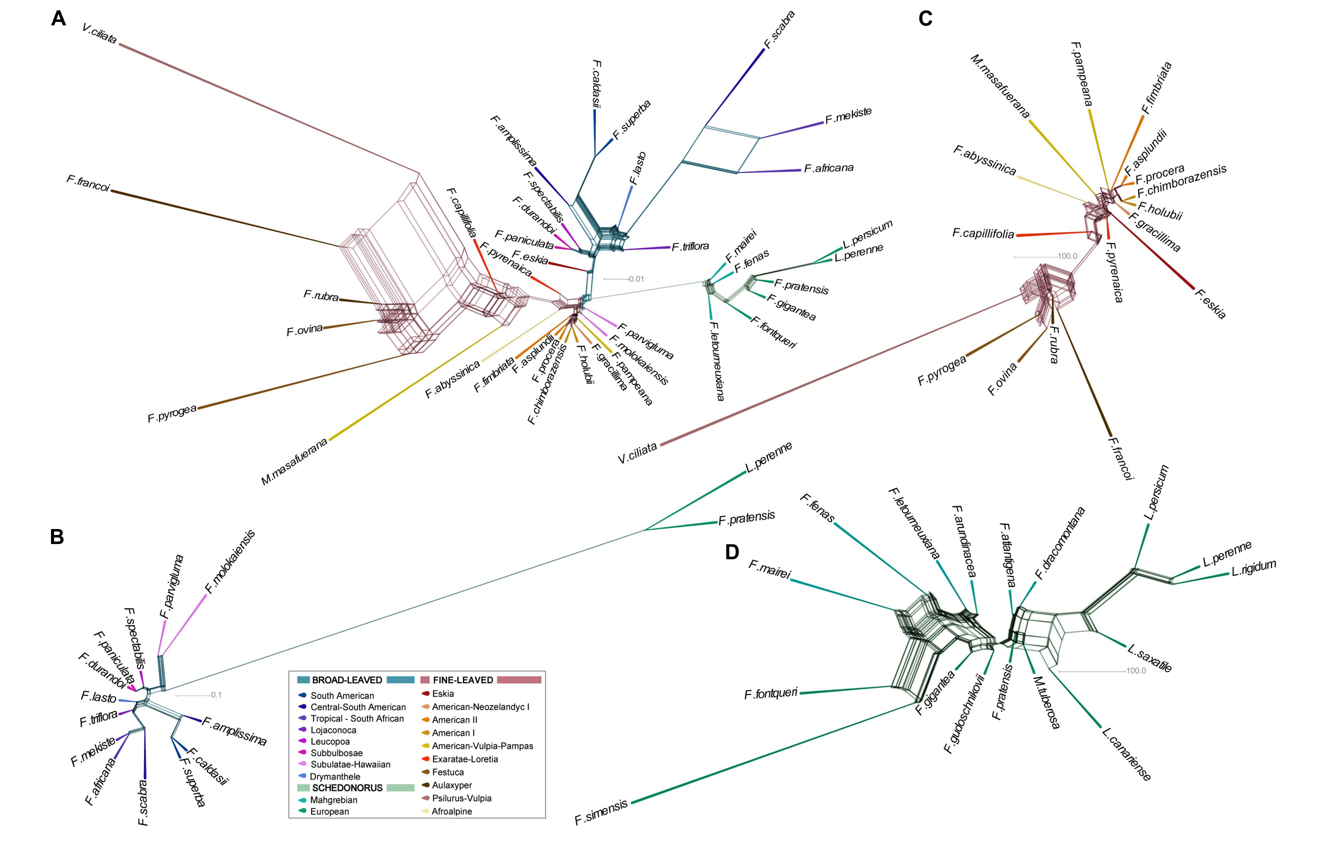


FIGURE 4 | Evolutionary networks based on standardized repeat data sets obtained from the comparative RE2 analysis of the four Loliinae evolutionary groups: **(A)** Loliinae, **(B)** broad-leaved (BL) Loliinae, **(C)** fine-leaved (FL) Loliinae, and **(D)** Schedonorus. The networks were constructed from distance-based NJ trees computed with pairwise Euclidean distances between samples (see text). Color codes of Loliinae lineages are indicated in the respective charts. Scale bar: number of mutations per site.

graph (e.g., *F. dracomontana, F. holubii, F. molokaiensis*) could not be classified as such (**Figure 5**).

DISCUSSION

Characterization of the Loliinae Repeatome and Its Impact on the Diversification of the Genome Size of Its Lineages

Our large-scale exploratory analysis of the Loliinae repeatome has uncovered the abundance and composition of the repetitive DNA across the genome landscape of all the subtribal lineages, confirming the substantial contribution of the repeatome to the genome size diversification of the studied Loliinae genomes (**Table 2**, **Figure 1**, and **Supplementary Figure 1E**). The repetitive elements represent more than half of the holoploid genome of most surveyed Loliinae taxa and accounted for the largest percentages (>60%) in the BL and *Lolium* genomes (**Table 2**, **Figure 1**, and **Supplementary Figure 1E**). Our data has demonstrated that the 1.5- to 3-fold downsizing monoploid genome trend observed by previous authors between BL and FL Loliinae lineages (Catalán, 2006; Šmarda et al., 2008) can be attributed to proportional amounts of their respective repetitive elements (**Tables 2**, **3**, **Figure 1**, and **Supplementary Figure 1E**). Unlike other studies that found no evidence of repeat activity causing large variation in genome size among diploid species (e.g., *Anacyclus*; Vitales et al., 2020a), our analyses have corroborated that striking differences in the 1.5-fold increase in genome size between BL and FL Loliinae diploid genomes was caused by significant differences in the repeat contents of the more abundant Retand and Angela retrotransposons (**Tables 2**, **3**, **Figures 1**, **2**, and **Supplementary Figure 2**). In general, the Loliinae diploid genomes, -either BL, FL or Schedonorus-, showed higher proportions of repeats than the allopolyploid genomes except for some of the South American BL and FL polyploid genomes (**Tables 1**, **2**, **Figure 1**, and **Supplementary Table 1**). Thus, our data partially rejects the "polyploid genome shock" hypothesis that predicts increased genome sizes (and correlated repeat expansions) in polyploids, as well as the additive pattern of diploid repeat contents in the derived allopolyploids (e.g., *Melampodium*; McCann et al., 2018). In contrast, it supports the alternative hypothesis that predicts a trend for genome (and repeatome) reduction after polyploidization due to genomic losses of duplicated genome

TABLE 4 | Ploidy levels and genomic pair-end read features of 5S rDNA loci and cluster graph parameters of the studied Loliinae taxa.

Taxon	Ploidy level	N. reads in cluster	Genome proportion (%)	Repeat size (bp)	k-mer coverage	Connected component index	Graph shape (type)
Festuca abyssinica	4x	180	0.036	316	0.885	0.967	1
Festuca africana	10x	214	0.043	317	0.488	0.879	2
Festuca amplissima	6x	158	0.032	318	0.744	0.994	1
Festuca a. arundinacea	6x	369	0.074	315	0.78	0.987	1
Festuca a. letourneuxiana	10x	532	0.11	307	0.721	0.974	2
Festuca a. atlantigena	8x	428	0.086	307	0.791	0.981	2
Festuca asplundii	6x	110	0.022	318	0.894	0.982	1
Festuca caldasii	4x	–	–	–	–	–	–
Festuca capillifolia	2x	340	0.068	318	0.9	0.976	1
Festuca chimborazensis	6x	179	0.036	319	0.845	0.899	2
Festuca dracomontana	–	629	0.13	307	0.75	0.936	1
Festuca durandoi	2x	520	0.1	318	0.812	0.994	2
Festuca eskia	2x	525	0.1	319	0.873	0.989	2
Festuca fenas	4x	222	0.044	307	0.781	0.973	1
Festuca fimbriata	6x	104	0.021	317	0.8	0.923	1
Festuca fontqueri	2x	470	0.094	296	0.824	0.977	2
Festuca francoi	2x	632	0.13	317	0.748	0.981	2
Festuca gigantea	6x	–	–	–	–	–	–
Festuca gracillima	6x	–	–	–	–	–	–
Festuca gudoschnikovii	4x	–	–	–	–	–	–
Festuca holubii	–	179	0.036	318	0.863	0.944	2
Festuca lasto	2x	470	0.094	296	0.824	0.977	1
Festuca mairei	4x	330	0.066	315	0.791	0.921	1
Festuca mekiste	–	109	0.022	317	0.619	0.917	1
Festuca molokaiensis	–	208	0.042	316	0.666	0.861	2
Festuca ovina	2x	331	0.066	316	0.952	0.985	1
Festuca pampeana	8x	402	0.08	317	0.812	0.98	1
Festuca paniculata	2x	269	0.054	318	0.781	0.978	2
Festuca parvigluma	4x	190	0.038	316	0.711	0.884	1
Festuca pratensis	2x	447	0.089	545	0.832	0.911	2
Festuca procera	4x	165	0.033	317	0.863	0.976	2
Festuca pyrenaica	4x	204	0.041	316	0.62	0.941	2
Festuca pyrogea	–	850	0.17	326	0.602	0.955	2
Festuca rubra	6x	338	0.068	316	0.737	0.87	2
Festuca scabra	4x	232	0.046	301	0.782	0.978	2
Festuca simensis	4x	412	0.082	296	0.675	0.951	2
Festuca spectabilis	6x	1128	0.23	316	0.791	0.99	2
Festuca superba	8x	184	0.037	316	0.772	0.995	1
Festuca triflora	2x	217	0.043	262	0.498	0.982	2
Lolium canariense	2x	306	0.061	294	0.842	0.974	1
Lolium perenne	2x	447	0.089	307	0.868	0.982	1
Lolium persicum	2x	1154	0.23	307	0.832	0.976	1
Lolium rigidum	2x	892	0.18	307	0.809	0.983	1
Lolium saxatile	2x	157	0.031	308	0.914	0.975	2
Megalachne masafuerana	–	690	0.14	224	0.438	0.997	2
Micropyropsis tuberosa	2x	911	0.18	307	0.865	0.98	1
Vulpia ciliata	4x	414	0.083	315	0.916	0.993	2

Graph shape types (type 1, simple circular-shaped graph with one loop; type 2, complex graph with two loops where the interconnected loops represent IGS spacers). 5S clustering analysis of F. caldasii, F. gigantea, F. gracillima and F. gudoschnikovii could not be performed due to insufficient number of 5S reads in the clusters. Hyphens, missing data.

fragments (e.g., *Spartina* and several sequenced plants; Chen, 2007; Parisod et al., 2010; Michael, 2014). The significantly lower genome sizes and correlated lower repeat contents of Old World Loliinae polyploids relative to diploids (**Tables 1**, **2**, **Figures 1**, **2**, and **Supplementary Figure 2**) could be attributed to the relatively ancestral DNA ages of some of these polyploid lineages [e.g., Schedonorus Mahgrebian (6.3 Ma) and FL Aulaxyper (6.1 Ma) clades; Moreno-Aguilar et al., 2020], which might have eliminated duplicated repeats over time. Furthermore, the high level of ploidy (6x-8x-10x) of these allopolyploids, which have apparently lost more redundant repeats compared to their closely related diploids or lower polyploids, could have resulted from

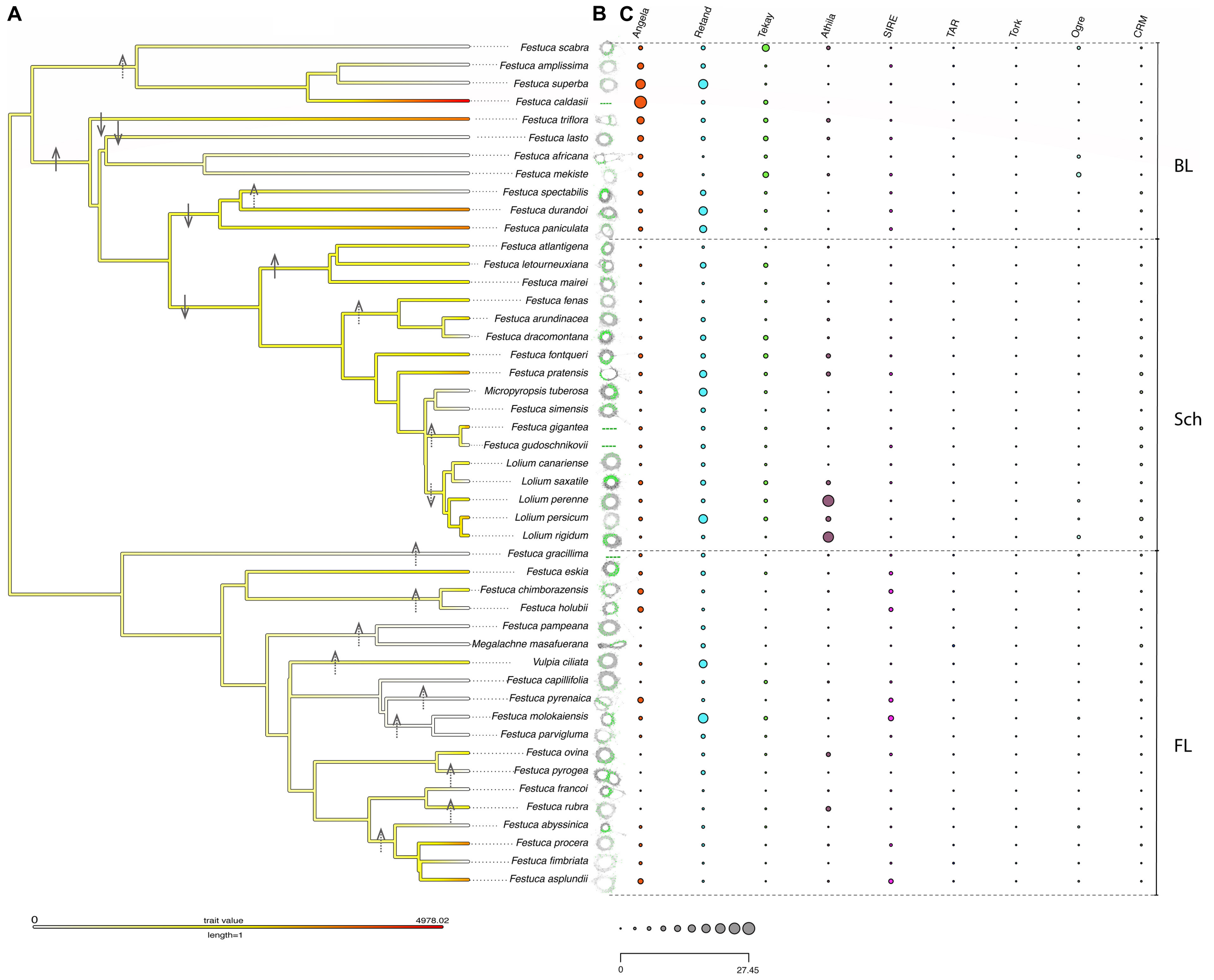


FIGURE 5 | (A) Maximum Likelihood phylogenetic tree (combined plastome + nuclear 35S rDNA data) of the 47 studied Loliinae samples showing their genome sizes by the colors of the terminal branches (color gradients indicate inferred genome size changes); white, missing data. **(B)** 5S clustering graph plots generated by RE2. **(C)** Proportions of the most abundant repeat elements (standardized values) obtained from the individual RE2 analysis of repeats are shown for each taxon. Hypothesized scenarios of allopolyploidization and diploidization events mapped onto the tree branches (ancestral allopolyploidization: solid arrow up; ancestral diploidization: solid arrow down; recent allopolyploidization: dashed arrow up; recent diploidization: dashed arrow down). BL, broad-leaved Loliinae; Sch, Schedonorus; FL, fine-leaved Loliinae.

a selective process to limit repetitive DNA damaging activity (Wang et al., 2021). Alternatively, some of these high polyploids could have originated through autopolyploidy or a combination of autopolyploidy and allopolyploidy; those scenarios would better explain the simple 5S graph patterns observed in many of these taxa (**Figure 5**). However, all thoroughly investigated Loliinae polyploids have been shown to be allopolyploids (Catalán, 2006, and references therein). The considerable reductions in retrotransposon and transposon contents detected in high polyploid Loliinae species are consistent with parallel losses of 35S rDNA loci in the same taxa (e.g., BL *F. africana*-10x, Namaganda, 2007; Schedonorus *F. atlantigena*-8x and *F. letourneuxiana*-10x, Ezquerro-López et al., 2017), suggesting that the two types of repetitive DNA reductions might have occurred after large genomic rearrangements in these high polyploids. In contrast, the large repeat contents of some Old World Loliinae diploids could be explained by the dynamic activity of young repeat types that have proliferated in recent diploid lineages (e.g., Athila in *Lolium*; **Table 2** and **Figures 1**, **2**; Zwyrtková et al., 2020).

As in many angiosperms (Eickbush and Malik, 2002), the retrotransposons LTR-Gypsy Retand (1.6–21.3%) and LTR-copia Angela (0.02–27.5%) were the most widely represented repeat family in the Loliinae genomes (**Table 2** and **Figure 1**). The Tekay, Athila and SIRE elements followed, while other retrotransposons (Ogre, CRM) and transposons (CACTA) were less common (**Table 2** and **Figure 1**). Together, they showed a strong correlation with genome size (R^2 = 0.83, p = 1.8E-09) and a considerable contribution to the differences in genome sizes (65.2%) between Loliinae lineages (**Table 3** and **Figure 2**), although these contributions varied for the most abundant types. The Retand repeats contributed significantly to the larger genome sizes of the BL and Schedonorus genomes compared to the FL genomes (**Table 2**), while the Angela repeats also contributed to the large sizes of the BL genomes and, notably, to some relatively large genomes of FL American I and American II genomes (**Table 2**). The Angela elements showed the highest correlation of repeat content with genome size (R^2 = 0.71) and also explained the greatest differences in genome size between species (19.6%), in contrast to the Retand repeats that presented lower correlation and contribution values (R^2 = 0.21; 10.7%) (**Table 3** and **Supplementary Figure 2**). The important role of Angela retrotransposons in genome size diversification of Loliinae genomes is likely related to the relatively higher conservatism of these repeats, compared to the more variable behavior of Retand and other repeat elements (**Figure 3**). In agreement with other studies that have also detected older and less active Angela copies in Fabaceae (Macas et al., 2015) and Triticeae (Wicker et al., 2017, 2018), but in contrast to the finding of a high turnover of Angela families in *Brachypodium distachyon* (Stritt et al., 2020), our data indicated that Angela repeats also tend to be relatively conserved in Loliinae and have probably better fitted long-term genomic diversification trends of their ancestral genomes (19.4 Ma; Moreno-Aguilar et al., 2020). In contrast, young and highly heterogeneous Athila families likely experienced a recent burst within the *Lolium* clade and especially in the allogamous *L. perenne* and *L. rigidum* genomes (23–25%) and were moderately abundant in other studied ray-grasses and their close *F. pratensis* and *F. fontqueri* relatives (7–8%) (**Table 2** and **Figure 1**). Noticeably, Athila elements also proliferated in recent FL *F. rubra* (8.5%) and *F. ovina* (7.1%) genomes, constituting the best represented annotated family in the red and sheep fescues (**Table 2** and **Figure 1**).

Phylogenetic Value of the Loliinae Repeatome and Deconvolution of the Origins of Some Genomes From 5S Cluster Graphs

In agreement with previous studies from other angiosperms (Dodsworth et al., 2015; McCann et al., 2018, 2020; Vitales et al., 2020b; Herklotz et al., 2021), the different amounts of shared repeats retrieved from comparative RE2 analyses of Loliinae have been shown to contain phylogenetic information at different systematic levels across the four Loliinae evolutionary groups. All evolutionary analyses have confirmed their ability to recover deep-to-shallow evolutionary relationships that were highly or relatively consistent with those based on the 35S rDNA and the plastome and combined data sets, respectively (**Tables 1**, **4**, **Figures 4**, **5**, **Supplementary Tables 3, 4**, and **Supplementary Figures 1, 3**). Some of the networks have, however, uncovered repeatome-specific topological features, which were not observed in the MSA trees (**Figure 4**).

The unrooted Loliinae and BL repeatome networks have demonstrated the high isolation of Schedonorus from the remaining Loliinae lineages (**Figures 4A,C**). This large divergence was based on the uniqueness of the Schedonorus repeat amounts within the representatives of the subtribe (**Supplementary Table 3**). Although Schedonorus has traditionally been considered a recent split within the broad-leaved Loliinae in all previous evolutionary studies (Minaya et al., 2017; Moreno-Aguilar et al., 2020, and references therein), and in the current combined tree of Loliinae (**Figure 1** and **Supplementary Figure 1B**), this position is mostly based in the strong plastome topology (**Supplementary Figure 1C**) and its large sequence dataset. By contrast, the weak nuclear 35S ML topology showed extremely low support for the potentially basal paraphyletic divergences of the BL lineages and an unclear position for Schedonorus within them (**Supplementary Figure 1D**). The repeatome network placed Schedonorus more closely related to the FL than to the BL group (**Figure 4A**). More reliable phylogenies based on single-copy nuclear genes would be needed to decipher the evolution of Schedonorus and other Loliinae nuclear genomes. Here, the phylogeny of tall fescues and ray-grasses has been enriched with three new taxa, showing the sister relationships of the eastern Canary Islands endemic *Lolium saxatile*-2x (Scholz and Scholz, 2005) to *L. canariense*-2x, of Siberian *F. gudoschnikovii*-4x (Stepanov, 2015; Probatova et al., 2017) to its morphologically close Eurosiberian relative *F. gigantea*-6x, and of previously unstudied South African *F. dracomontana* (Linder, 1986) to *F. arundinacea*-6x (plastome tree) or to the 'European' clade (35S tree) (**Figure 1** and **Supplementary Figures 1A–D**). A notable geographical signal of the repeatome was observed in the close relationships of NW African *F. fontqueri*-2x and Tropical African *F. simensis*-4x with Mahgrebian *F. mairei*-4x (**Figure 4D**), in contrast to their nesting positions within the predominantly diploid "European" clade in the plastome, 35S and combined trees (**Supplementary Figures 1B–D**). Also, the position of *F. dracomontana* in the repeatome network suggest that this austral Schedonorus species could be a polyploid close to the tall fescues (**Figure 4D** and **Supplementary Figures 1B–D**).

Geographically based evolutionary patterns of repetitive elements, congruent with those of the nuclear 35S rDNA tree, have been also observed in the FL and BL repeatome networks (**Figures 4B,C** and **Supplementary Figure 1D**). Within the FL network group, South American representatives of the American I, American-Pampas and American II lineages are closely related to each other (**Figure 4B** and **Supplementary Figure 1D**), while interspersed with other FL lineages in the plastome and combined Loliinae trees (**Supplementary Figures 1B,C**). These lineages are characterized by similar levels of Angela, Retand and LTR repeats (**Table 2** and **Figure 1**) and were inferred to be of similar age (late Miocene_Pliocene transition, 3.4–5.4 Ma; Minaya et al., 2017). They are probably the descendants of the same paternal lineage, which probably evolved *in situ* but crossed with distinct maternal FL lineages giving rise to these close but separate allopolyploid clades (**Supplementary Figures 1B,C**). Within the BL group, the close relationships between South African *F. scabra* and Tropical and South African *F. africana*/*F. mekiste* and between Mediterranean-European *F. spectabilis* (Leucopoa) and *F. paniculata*/*F. durandoi* (Subbulbosae) based on shared repeat contents are more similar to those recovered in the 35S tree than in the plastome tree (**Figure 4C** and

Supplementary Figure 1A–C), also suggesting a concerted evolution of nuclear repetitive DNA families and different hybridizations or chloroplast capture events with other BL lineages. In contrast, the close relationship of Central-American *F. amplissima* to the South American *F. superba*/*F. caldasii* lineage shown in the repeatome network is more similar to that observed in the plastome and combined Loliinae trees than in the 35S tree, probably due to the lower resolution of the nuclear topology (**Figure 4C** and **Supplementary Figures 1A–C**). Interestingly, these Central and South American taxa show some of the highest Loliinae genomic repeat contents (**Tables 1**, **2**, **Figure 1**, and **Supplementary Figure 1E**) despite their high 6x-8x ploidy-levels. It could be a consequence of their relatively young ages (∼5 Ma; Moreno-Aguilar et al., 2020) and the lack of a time course to purge the excess of repetitive DNA (Michael, 2014), or a recent bloating of repeats. The phylogenetic value of the Loliinae repetitive elements has been further corroborated by the significant phylogenetic signals carried by different repeat clusters when tested on the respective tree cladograms of each of the four Loliinae groups (**Supplementary Table 5** and **Supplementary Figure 4**). In most of the groups, the conservative Angela clusters had significant *K* values above 0.5 and close to 1, indicating their strong phylogenetic signal at different taxonomic levels.

Although tandem-repeated 5S rDNA did not retrieve a congruent evolutionary history for Loliinae (**Supplementary Figure 5**), their cluster graph topologies revealed their presumable number of loci (**Figure 5**), indicative of their potential hybridization events (Vozárová et al., 2021) and ploidy levels (Garcia et al., 2020). In contrast to the instability of 35S rDNA loci, the maintenance of 5S rDNA loci in high allopolyploid Loliinae species (Ezquerro-López et al., 2017) is consistent with their conserved patterns in other angiosperm allopolyploids (Garcia et al., 2017). Studies of allopolyploids with known subgenomes have demonstrated that species showing complex graphs with two IGS loops correspond to allotetraploids and those showing three loops to allohexaploids (Garcia et al., 2020), while in highly hybridogenous diploid rose species graphs with two loops probably correspond to ancient 5S rDNA families (Vozárová et al., 2021). Within the Loliinae studied, several polyploid taxa displayed 5S graphs with fewer loops than expected for their ploidy level (**Figure 5**), suggesting the existence of convergent evolution to one or few ribotypes. In contrast, three diploid species, BL *F. triflora* and *F. paniculata* and FL *F. francoi*, showed a 5S graph pattern typical of allotetraploids (**Figure 5**), supporting the hypothesis of their putative paleo-polyploid hybrid origin.

Recurrent Rounds of Allopolyploidizations and Diploidizations Within Loliinae Lineages Revealed by Their Repeats

The widely accepted evolutionary scenario for the origin of the angiosperms, consisting of several rounds of hybridizations and allopolyploidizations followed by a return to the diploid state (Soltis et al., 2016) has been also inferred for the grasses and their main lineages. Evidence suggests that protograss whole genome duplication (WGD) was likely followed by later diploidizations that ended in current paleo-ancestral diploid karyotypes for temperate and tropical grasses (Salse et al., 2008). These involved distinct and profound genomic rearrangements, such as nested chromosome fusions, chromosome inversions and paleocentromere inactivation, along with differential losses of heterologous duplicated copies in subgenomes of divergent lineages (Murat et al., 2010). In contrast, new allopolyploidization events apparently led to the emergence of grass mesopolyploids, originated some million years ago, and grass neopolyploids, considered to have emerged during or after the Quaternary glaciations (Stebbins, 1985; Marcussen et al., 2014). Our data allow us to hypothetize that the evolution of Loliinae could have resulted from relatively rapid recurrent rounds of allopolyploidizations and diploidizations during the last 19–22 Ma (Minaya et al., 2017; Moreno-Aguilar et al., 2020) that have leaved their signatures on their repeats (**Figure 1** and **Supplementary Figure 1E**) and 5S graph topologies (**Figure 5**). We postulate that the large genomes of the early diverging BL diploids (Lojaconoa, Drymanthele, Subulbosae; 7.5–5 Ma, Minaya et al., 2017) likely resulted from WGD of ancestral interspecific hybrids that later reverted to the diploid state with large chromosomes (Catalán, 2006), relatively large monoploid genome sizes and repeat contents (**Table 2**, **Figures 1**, **2**, and **Supplementary Figure 1E**) and complex 5S graphs indicative of putative allotetraploids (**Figure 5**). This polyploid hybrid origin could also explain the potential heterosis of these robust broad-leaved fescues (Catalán, 2006). We also hypothetize that the large genomes and repeatomes of the basal BL polyploid lineages (Central-South American, South African) may have resulted from more recent allopolyploidizations (5–2.5 Ma, Minaya et al., 2017), with genomes that still maintain large sizes and proportions of repeats, and retain traces of more than one 5S ribotype (**Table 2**, **Figures 1**, **2**, **5** and **Supplementary Figure 1E**).

Our findings are not fully compatible with the hypotheses of drastic genome contractions from a hypothetical large-genome Loliinae ancestor to the FL Loliinae lineage and in allopolyploids with large progenitor genomes but not in autopolyploids with small progenitor genomes (Loureiro et al., 2007; Šmarda et al., 2008). The observed reduction in repeat content and correlated genome size from the large BL Loliinae, through intermediate Schedonodorus and *F. eskia*, to the small FL Loliinae genomes (**Figures 2**, **5**) could have resulted from independent genome size diversifications along the major Loliinae lineages (**Figures 1**, **5** and **Supplementary Figure 1**). Our data also support an alternative scenario of independent hybridization and polyploidization events across FL Loliinae, which are similar in age (∼16 Ma, Minaya et al., 2017) to BL Loliinae. Their small chromosomes and genome sizes (Catalán, 2006), especially for the taxa of the core Eurasian and Mediterranean Vulpia, Festuca and Aulaxyper (plus Exaratae) lineages (**Tables 1**, **2** and **Figures 1**, **2**, **5**), are similar to those of the close subtribes Parapholiinae, Cynosuriinae, and Dactylidiinae with which they also share 35S rDNA families (Catalán et al., 2004). Therefore, it could be hypothesized that the ancestor of these FL Loliinae did not undergo the same double genome enlargement as the ancestor of BL Loliinae. In addition, the various polyploid New World FL lineages (American I, American-Pampas, Subulatae-Hawaiian, American II), which show larger genome sizes and geographically structured repeat contents (**Tables 1**, **2**, **Figures 1**, **4A,C**, **5**) are probably the results of recent allopolyploidizations (5–2.5 Ma, Minaya et al., 2017) that have not yet experienced considerable purging in their repeats.

The isolated Schedonorus lineage emerges as a highly dynamic repeat-driven evolving group, also accumulating evidence of various allopolyploidizations and diploidizations. A distinctive feature is the bloating of Athila repeats in the recently evolved diploid clade *Lolium*, especially in allogamous ray-grasses (**Table 2**, **Figures 1**, **2**, and **Supplementary Figure 2**; Zwyrtková et al., 2020). In contrast, the Mahgrebian clade constitute a relatively ancestral lineage with unknown diploid relatives (Inda et al., 2014), although it shows signatures of ancient hybridizations in its 5S graph topologies (**Figure 5**). The Schedonorus Mahgrebian and the FL Aulaxyper allopolyploid lineages have experienced the most pronounced reductions in their repeats and genome sizes of all Loliinae studied (**Table 2** and **Figures 1**, **2**, **5**). Interestingly, these two lineages also exhibit the highest and most extensive hybridization rates among the Loliinae, producing both intra- and intergeneric hybrids (Catalán, 2006). Schedonorus *Festuca* taxa spontaneously hybridize with each other and with close species of *Lolium* (x *Festulolium*) while Aulaxyper *Festuca* taxa (*F.* gr. *rubra*) also interbreed with each other and with close species of *Vulpia* (x *Festulpia*) (Catalán, 2006, and references therein). Therefore, it might be plausible that these two highly hybridogenous allopolyploid lineages have undergone large genome reshufflings to accommodate their highly divergent heterologous subgenomes and avoid DNA damage (Michael, 2014; Wang et al., 2021). These genomic rearrangements would have caused more severe losses in their respective repeats and genome sizes than those of other high polyploid American BL and FL Loliinae of similar ancestry that resulted from crosses of genomically similar progenitor species and presumably did not experience large repeat contractions (**Table 2** and **Figures 1**, **2**, **5**).

DATA AVAILABILITY STATEMENT

The newly studied grass plastome and 35S and 5S rDNA cistron sequences have been deposited in the Genbank data base under accession numbers SAMN27777779–SAMN27777788, ON243855–ON243864 and ON248974–ON249019, and at the Github repository (https://github.com/Bioflora/Loliinae_Repeatome).

AUTHOR CONTRIBUTIONS

PC designed the study. MM-A, IA, LI, and PC collected the samples. MM-A and LI developed the experimental work. PC, MM-A, LI, IA, and PC analyzed the data and interpreted the results. PC and MM-A prepared the manuscript. PC, MM-A, LI, IA, and AS-R revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Spanish Aragon Government grant project LMP82 _21, the Spanish Aragon Government and European Social Fund Bioflora research group grant A01-20R, and the European and Spanish Government (SEPIE) Erasmus + KA107 mobility grant 2019-1-ES01-KA107-062605. MM-A was supported by a University of Zaragoza Bioflora research contract.

ACKNOWLEDGMENTS

We thank Nina Probatova and Alan Stewart for sending us materials of *Festuca gudoschnikovii* and *Lolium saxatile*, the MHU, PRE, and VLA herbaria for their herbarium loans, Petr Novák and Jiri Macas for sharing their scripts for analysis of repeats and genome size variation and of repeat sequence similarity Hs/Ho ratios, Jaume Pellicer and Sònia García for valuable advice on genome size and 5S rDNA diversifications, the reviewers for their valuable comments to an early version of the manuscript, and the Ministerio del Ambiente of Ecuador and the South African National Parks for giving permissions to collect Loliinae samples in the Ecuadorian paramos (MAE-DNB-CM-2015-0016) and the Western Cape and Drakensberg Mountains Parks. The genome skimming data of the 47 Loliinae samples was generated at the Centro Nacional de Análisis Genómicos (CNAG, Barcelona, Spain). Computational resources for Repeat Explorer analysis were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure. The bioinformatic and evolutionary analyses were performed at the High Polytechnic School of Huesca (University of Zaragoza, Spain) Bioflora laboratory.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.901733/full#supplementary-material

Supplementary Table 1 | Taxa included in the repeatome analysis of Loliinae. Taxonomic rank, taxon authorship, detailed localities and vouchers, and source of cytogenetic and genomic data. Group: BL, broad-leaved Loliinae; FL, fine-leaved Loliinae; Sch, Schedonorus. Chromosome number (2n), ploidy, genome size (2C, pg), monoploid genome size (1Cx, pg; 1Cx, Mbp) and GenBank accession codes for plastome and nuclear ribosomal 35S and 5S genes are given for each sample. Values in bold correspond to new data generated in this study. Outgroups used in the phylogenomic analyses: *Oryza sativa, Brachypodium distachyon*.

Supplementary Table 2 | Loliinae samples used in the repetitive DNA analysis. Genome skimming paired-end (PE) reads per sample and PE reads selected by Repeat Explorer 2 per sample in each of the comparative analyses of the four Loliinae groups: Loliinae, BL (broad-leaved Loliinae), FL (fine-leaved Loliinae), Schedonorus.

Supplementary Table 3 | Repeat Explorer 2 comparative analysis. Repeat content data for top clusters (repeat families) in each of the four evolutionary groups of Loliinae: **(A)** Loliinae; **(B)** broad-leaved (BL) Loliinae; **(C)** fine-leaved (FL) Loliinae; **(D)** Schedonorus.

Supplementary Table 4 | Repeat Explorer 2 comparative analysis. Repeat content data for phylogenetically analyzed clusters (repeat families) in each of the four evolutionary groups of Loliinae: **(A)** Loliinae; **(B)** broad-leaved (BL) Loliinae; **(C)** fine-leaved (FL) Loliinae; **(D)** Schedonorus.

Supplementary Table 5 | Phylogenetic signal based on Blomberg's *K* values of repeat cluster contents obtained from the comparative RE2 analysis of Loliinae samples assessed in each of the four Loliinae groups: **(A)** Loliinae (38 samples, 38 clusters), **(B)** Broad-leaved (BL) Loliinae (15 samples, 96 clusters), **(C)** fine-leaved (FL) Loliinae (17 samples, 122 clusters), **(D)** Schedonorus (16 samples, 167 clusters), using the *phylosig* option of the *phytools* R package. Cluster abundance values (number of PE reads) are indicated in **Supplementary Table 4**. *K* values close to one indicate phylogenetic signal, values close to zero phylogenetic independence, and values >1 more phylogenetic signal than expected. *p*-Values based on 1000 randomizations. Significant values are highlighted in bold.

Supplementary Figure 1 | (A) Combined (plastome + 35S rDNA) Loliinae coalescent species tree computed through Singular Value Decomposition quartets (SVDq) analysis showing bootstrap support values on branches. **(B–D)** Maximum Likelihood phylogenomic trees of 47 Loliinae samples based on **(B)** Combined (plastome + 35S rDNA) data, **(C)** plastome data, **(D)** nuclear 35S rDNA data, **(E)** Histograms of repeat contents per holoploid genome (1C) retrieved from the individual Repeat Explorer 2 analyses of the studied Loliinae samples mapped onto the Maximum Likelihood combined phylogenomic tree (plastome + nuclear 35S rDNA cistron) of Loliinae. Ultrafast bootstrap support values are indicated on branches. *Oryza sativa* and *Brachypodium distachyon* outgroups were used to root the trees. Color codes of Loliinae lineages are indicated in the charts. Scale bar: number of mutations per site.

Supplementary Figure 2 | Correlation plots of repeat content and genome size variation (1Cx) for the 23 Loliinae taxa with known genome sizes. Individual plots for the most represented repeat types found across the 23 Loliinae taxa with known genome size data (see **Table 2** and **Figure 2**). Color codes of Loliinae lineages correspond to those indicated in **Figure 1**.

Supplementary Figure 3 | Evolutionary networks based on standardized repeat data sets obtained from the comparative RE2 analysis of the four Loliinae evolutionary groups: **(A)** Loliinae, **(B)** broad-leaved (BL) Loliinae, **(C)** fine-leaved (FL) Loliinae, **(D)** Schedonorus. The networks were constructed from distance-based NJ trees computed with pairwise inverse distances between samples (see text). Color codes of Loliinae lineages are indicated in the respective charts. Scale bar: number of mutations per site.

Supplementary Figure 4 | Maximum Likelihood Loliinae tree cladograms (combined plastome + nuclear 35S rDNA cistron) showing the relationships among the studied samples in each of the four evolutionary groups of Loliinae and phyloheatmaps of normalized values for different sets of repeat clusters retrieved by RE2 from the comparative analysis of each group: **(A)** Loliinae (38 samples, 38 clusters), **(B)** broad-leaved (BL) Loliinae (15 samples, 96 clusters), **(C)** fine-leaved (FL) Loliinae (17 samples, 122 clusters), **(D)** Schedonorus (16 samples, 167 clusters). Repeat clusters showing significant phylogenetic signal are highlighted with dotted lines.

Supplementary Figure 5 | Maximum Likelihood nuclear 5S rDNA cistron tree showing the relationships among the 47 studied Loliinae samples. Ultrafast bootstrap support values are indicated on branches. *Oryza eichingeri* and *Brachypodium distachyon* outgroups were used to root the tree. Color codes of Loliinae lineages are indicated in the chart. Scale bar: number of mutations per site.

REFERENCES

Blomberg, S. P., Garland, T. J., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57, 717–745. doi: 10.1111/j.0014-3820.2003.tb00285.x

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Catalán, P. (2006). "Phylogeny and evolution of *Festuca* L. and related genera of subtribe Loliinae (Poeae, Poaceae)," in *Plant Genome: Biodiversity and Evolution*, ed. A. S. A. K. Sharma (Enfield, NH: Science Publishers), 255–303.

Catalán, P., Torrecilla, P., López Rodríguez, J. Á., and Olmstead, R. G. (2004). Phylogeny of the festucoid grasses of subtribe Loliinae and allies (Poeae, Pooideae) inferred from ITS and trnL-F sequences. *Mol. Phylogenet. Evol.* 31, 517–541. doi: 10.1016/j.ympev.2003.08.025

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

Chen, Z. J., Sreedasyam, A., Ando, A., Song, Q., De Santiago, L. M., Hulse-Kemp, A. M., et al. (2020). Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* 52, 525–533. doi: 10.1038/s41588-020-0614-5

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* 65, 997–1008. doi: 10.1093/sysbio/syw037

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45:e18. doi: 10.1093/nar/gkw955

Dodsworth, S., Chase, M. W., Kelly, L. J., Leitch, I. J., Macas, J., Novak, P., et al. (2015). Genomic repeat abundances contain phylogenetic signal. *Syst. Biol.* 64, 112–126. doi: 10.1093/sysbio/syu080

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

Drouin, M., Hénault, M., Hallin, J., and Landry, C. R. (2021). Testing the genomic shock hypothesis using transposable element expression in yeast hybrids. *Front. Fungal Biol.* 2:729264. doi: 10.3389/ffunb.2021.729264

Dubcovsky, J., and Martínez, A. (1992). Distribución geográfica de los niveles de ploidía en *Festuca*. *Parodiana* 7, 91–99.

Ebrahimzadegan, R., Houben, A., and Mirzaghaderi, G. (2019). Repetitive DNA landscape in essential A and supernumerary B chromosomes of *Festuca pratensis* Huds. *Sci. Rep.* 9:19989. doi: 10.1038/s41598-019-56383-1

Eickbush, T. H., and Malik, H. S. (2002). "Origins and evolution of retrotransposons," in *Mobile DNA II*, eds N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz (Washington DC: ASM Press), 1111–1146. doi: 10.1128/9781555817954.ch49

Ezquerro-López, D., Kopecký, D., and Inda, L. A. (2017). Cytogenetic relationships within the Maghrebian clade of *Festuca* subgen. *Schedonorus* (Poaceae), using flow cytometry and FISH. *Anal. Jard. Bot. Madrid* 74, 1–9. doi: 10.3989/ajbm.2455

Fedoroff, N. V. (2012). Transposable elements, epigenetics, and genome evolution. *Science* 338, 758–767. doi: 10.1126/science.338.6108.758

Garcia, S., Kovařík, A., Leitch, A. R., and Garnatje, T. (2017). Cytogenetic features of rRNA genes across land plants: analysis of the Plant rDNA database. *Plant J.* 89, 1020–1030. doi: 10.1111/tpj.13442

Garcia, S., Wendel, J. F., Borowska-Zuchowska, N., Aïnouche, M., Kuderova, A., and 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. doi: 10.3389/fpls.2020.00041

Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., et al. (2017). Extensive gene content variation in the *Brachypodium* distachyon pan-genome correlates with population structure. *Nat. Commun.* 8:2184. doi: 10.1038/s41467-017-02292-8

Herklotz, V., Kovařík, A., Wissemann, V., Lunerová, J., Vozárová, R., Buschmann, S., et al. (2021). Power and weakness of repetition – evaluating the phylogenetic signal from repeatomes in the family rosaceae with two case studies from genera prone to polyploidy and hybridization (*Rosa* and *Fragaria*). *Front. Plant Sci.* 12:738119. doi: 10.3389/fpls.2021.738119

Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A. R., and Leitch, I. J. (2017). Is there an upper limit to genome size? *Trends Plant Sci.* 22, 567–573. doi: 10.1016/j.tplants.2017.04.005

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

Inda, L. A., Segarra-Moragues, J. G., Müller, J., Peterson, P. M., and Catalán, P. (2008). Dated historical biogeography of the temperate Loliinae (Poaceae, Pooideae) grasses in the northern and southern hemispheres. *Mol. Phylogenet. Evol.* 46, 932–957. doi: 10.1016/j.ympev.2007.11.022

Inda, L. A., Sanmartin, I., Buerki, S., and Catalán, P. (2014). Mediterranean origin and Miocene-Holocene Old World diversification of meadow fescues and ryegrasses (*Festuca* subgen. *Schedonorus* and *Lolium*). *J. Biogeogr.* 41, 600–614. doi: 10.1111/jbi.12211

Jenkins, G., and Hasterok, R. (2007). BAC "landing" on chromosomes of *Brachypodium* distachyon for comparative genome alignment. *Nat. Protoc.* 2, 88–98. doi: 10.1038/nprot.2006.490

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kopecký, D., and Studer, B. (2014). Emerging technologies advancing forage and turf grass genomics. *Biotechnol. Adv.* 32, 190–199. doi: 10.1016/j.biotechadv.2013.11.010

Křivánková, A., Kopecký, D., Stočes, Š, Doležel, J., and Hřibová, E. (2017). Repetitive DNA: a versatile tool for karyotyping in *Festuca* pratensis huds. *Cytogenet. Genome Res.* 151, 96–105. doi: 10.1159/000462915

Kubatko, L. S., and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56, 17–24. doi: 10.1080/10635150601146041

Linder, H. P. (1986). POACEAE. *Bothalia* 16, 59–61. doi: 10.4102/abc.v16i1.1072

Loureiro, J., Kopecký, D., Castro, S., Santos, C., and Silveira, P. (2007). Flow cytometric and cytogenetic analyses of Iberian Peninsula *Festuca* spp. *Plant Syst. Evol.* 269, 89–105. doi: 10.1007/s00606-007-0564-8

Macas, J., Novak, P., Pellicer, J., Cizkova, J., Koblizkova, A., Neumann, P., et al. (2015). In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe fabeae. *PLoS One* 10:e0143424. doi: 10.1371/journal.pone.0143424

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., et al. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science* 345:1250092. doi: 10.1126/science.1250092

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

McCann, J., Macas, J., Novák, P., Stuessy, T. F., Villaseñor, J. L., and Weiss-Schneeweiss, H. (2020). Differential genome size and repetitive DNA evolution in diploid species of *Melampodium* sect. *Melampodium* (Asteraceae). *Front. Plant Sci.* 11:362. doi: 10.3389/fpls.2020.00362

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

Michael, T. P. (2014). Plant genome size variation: bloating and purging DNA. *Brief. Funct. Genomics Proteomics* 13, 308–317. doi: 10.1093/bfgp/elu005

Minaya, M., Hackel, J., Namaganda, M., Brochmann, C., Vorontsova, M. S., Besnard, G., et al. (2017). Contrasting dispersal histories of broad- and fine-leaved temperate Loliinae grasses: range expansion, founder events, and the roles of distance and barriers. *J. Biogeogr.* 44, 1980–1993. doi: 10.1111/jbi.13012

Moreno-Aguilar, M. F., Arnelas, I., Sánchez-Rodríguez, A., Viruel, J., and Catalán, P. (2020). Museomics unveil the phylogeny and biogeography of the neglected juan fernandez archipelago megalachne and *Podophorus* endemic grasses and their connection with relict pampean-ventanian fescues. *Front. Plant Sci.* 11:819. doi: 10.3389/fpls.2020.00819

Murat, F., Xu, J. H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., et al. (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* 20, 1545–1557. doi: 10.1101/gr.109744.110

Namaganda, M. (2007). A Taxonomic Review of the Genus Festuca in Uganda: AFLP Fingerprinting, Chromosome Numbers, Morphology and Anatomy. Ph.D. thesis. Ås: Norwegian University of Life Sciences.

Negi, P., Rai, A. N., and Suprasanna, P. (2016). Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. *Front. Plant Sci.* 7:1448. doi: 10.3389/fpls.2016.01448

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Novák, P., Neumann, P., and Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat. Protoc.* 15, 3745–3776. doi: 10.1038/s41596-020-0400-y

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

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

Probatova, N. S., Barkalov, V. Y., and Stepanov, N. V. (2017). Chromosome numbers in some vascular plant species from Siberia and the Russian Far East. *Bot. Pacifica* 6, 51–55. doi: 10.17581/bp.2017.06103

Reaz, R., Bayzid, M. S., and Rahman, M. S. (2014). Accurate phylogenetic tree reconstruction from quartets: a heuristic approach. *PLoS One* 9:e104008. doi: 10.1371/journal.pone.0104008

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

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U. M., et al. (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 20, 11–24. doi: 10.1105/tpc.107.056309

Scholz, S., and Scholz, H. (2005). A new species of *Lolium* (Gramineae) from Fuerteventura and Lanzarote (Canary Islands, Spain). *Willdenowia* 35, 281–286. doi: 10.3372/wi.35.35208

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

Soltis, D. E., Visger, C. J., Blaine Marchant, D., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103, 1146–1166. doi: 10.3732/ajb.1500501

Stebbins, G. L. (1985). Polyploidy, hybridization and the invasion of new habitats. *Ann. Missouri Bot. Gard.* 72, 824–832.

Stepanov, N. V. (2015). *About Three New Species of Vascular Plants From the Western Sayan*. Tomsk: Sistematicheskie Zametki po Materialam Gerbarii Imeni P. N. Krylova pri Tomskom Gosudarstvennom Universitete, 3–15. doi: 10.17223/20764103.111.1

Stritt, C., Wyler, M., Gimmi, E. L., Pippel, M., and Roulin, A. C. (2020). Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass *Brachypodium* distachyon. *New Phytol.* 227, 1736–1748. doi: 10.1111/nph.16308

Swofford, D. L. (2003). *Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4*. Sunderland, MA: Sinauer Associates. doi: 10.1111/j.0014-3820.2002.tb00191.x

Vitales, D., Álvarez, I., Garcia, S., Hidalgo, O., Feliner, G. N., Pellicer, J., et al. (2020a). Genome size variation at constant chromosome number is not correlated with repetitive DNA dynamism in *Anacyclus* (Asteraceae). *Ann. Bot.* 125, 611–623. doi: 10.1093/aob/mcz183

Vitales, D., Garcia, S., and Dodsworth, S. (2020b). Reconstructing phylogenetic relationships based on repeat sequence similarities. *Mol. Phylogenet. Evol.* 147:106766. doi: 10.1016/j.ympev.2020.106766

Vozárová, R., Herklotz, V., Kovaøík, A., Tynkevich, Y. O., Volkov, R. A., Ritz, C. M., et al. (2021). Ancient origin of two 5S rDNA families dominating in the genus *Rosa* and their behavior in the Canina-type meiosis. *Front. Plant Sci.* 12:643548. doi: 10.3389/fpls.2021.643548

Wang, X., Morton, J. A., Pellicer, J., Leitch, I. J., and Leitch, A. R. (2021). Genome downsizing after polyploidy: mechanisms, rates and selection pressures. *Plant J.* 107, 1003–1015. doi: 10.1111/tpj.15363

Weiss-Schneeweiss, H., Leitch, A. R., Mccann, J., Jang, T. S., and Macas, J. (2015). "Employing next generation sequencing to explore the repeat landscape of the plant genome," in *Next Generation Sequencing in Plant Systematics. Regnum Vegetabile 157*, eds E. Hörandl and M. Appelhans (Königstein: Koeltz Scientific Books).

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R. H., et al. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* 19:103. doi: 10.1186/s13059-018-1479-0

Wicker, T., Schulman, A. H., Tanskanen, J., Spannagl, M., Twardziok, S., Mascher, M., et al. (2017). The repetitive landscape of the 5100 Mbp barley genome. *Mob. DNA* 8, 1–17. doi: 10.1186/s13100-017-0102-3

Zwyrtková, J., Němečková, A., Čížková, J., Holušová, K., Kapustová, V., Svačina, R., et al. (2020). Comparative analyses of DNA repeats and identification of a novel Fesreba centromeric element in fescues and ryegrasses. *BMC Plant Biol.* 20:280. doi: 10.1186/s12870-020-02495-0

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 02 September 2022
DOI 10.3389/fpls.2022.995586

Exploiting the miniature inverted-repeat transposable elements insertion polymorphisms as an efficient DNA marker system for genome analysis and evolutionary studies in wheat and related species

Benjamin Ewa Ubi[1,2]*, Yasir Serag Alnor Gorafi[3,4], Beery Yaakov[5], Yuki Monden[6], Khalil Kashkush[7] and Hisashi Tsujimoto[1]*

[1]Molecular Breeding Laboratory, Arid Land Research Center, Tottori University, Tottori, Japan, [2]Department of Biotechnology, Ebonyi State University, Abakaliki, Abakaliki, Ebonyi, Nigeria, [3]International Platform for Dryland Research and Education, Tottori University, Tottori, Japan, [4]Agricultural Research Corporation, Wad Medani, Sudan, [5]French Associates Institute for Agriculture and Biotechnology of Drylands, Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Beer-Sheva, Israel, [6]Graduate School of Environmental and Life Science, Okayama University, Okayama, Japan, [7]Department of Life Sciences, Ben-Gurion University, Beer-Sheva, Israel

OPEN ACCESS

EDITED BY
Ruslan Kalendar,
University of Helsinki,
Finland

REVIEWED BY
Yolanda Loarce,
University of Alcalá,
Spain
Eva Satovic Vuksic,
Rudjer Boskovic Institute,
Croatia

*CORRESPONDENCE
Benjamin Ewa Ubi
ubi.benjamin@tottori-u.ac.jp
Hisashi Tsujimoto
tsujim@tottori-u.ac.jp

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

RECEIVED 16 July 2022
ACCEPTED 09 August 2022
PUBLISHED 02 September 2022

CITATION
Ubi BE, Gorafi YSA, Yaakov B, Monden Y, Kashkush K and Tsujimoto H (2022) Exploiting the miniature inverted-repeat transposable elements insertion polymorphisms as an efficient DNA marker system for genome analysis and evolutionary studies in wheat and related species.
Front. Plant Sci. 13:995586.
doi: 10.3389/fpls.2022.995586

Transposable elements (TEs) constitute ~80% of the complex bread wheat genome and contribute significantly to wheat evolution and environmental adaptation. We studied 52 TE insertion polymorphism markers to ascertain their efficiency as a robust DNA marker system for genetic studies in wheat and related species. Significant variation was found in miniature inverted-repeat transposable element (MITE) insertions in relation to ploidy with the highest number of "full site" insertions occurring in the hexaploids (32.6±3.8), while the tetraploid and diploid progenitors had 22.3±0.6 and 15.0±3.5 "full sites," respectively, which suggested a recent rapid activation of these transposons after the formation of wheat. Constructed phylogenetic trees were consistent with the evolutionary history of these species which clustered mainly according to ploidy and genome types (SS, AA, DD, AABB, and AABBDD). The synthetic hexaploids sub-clustered near the tetraploid species from which they were re-synthesized. Preliminary genotyping in 104 recombinant inbred lines (RILs) showed predominantly 1:1 segregation for simplex markers, with four of these markers already integrated into our current DArT-and SNP-based linkage map. The MITE insertions also showed stability with no single excision observed. The MITE insertion site polymorphisms uncovered in this study are very promising as high-potential evolutionary markers for genomic studies in wheat.

KEYWORDS

transposable elements, mite, DNA markers, Dryland, wheat, genome analysis

Introduction

Wheat (*Triticum* spp.) is a cereal crop of major importance globally which provides about 20% of the calories consumed by man (FAO, 2015); and is a foremost source of vegetable protein in the human diet relative to other major cereal crops such as maize or rice (Wheat-Wikipedia, n.d.). The increased production of cereals such as wheat is urgently needed to meet the demand gap for global food supply by the year 2050 (Tester and Langridge, 2010) when 60%–70% increase in the food production is required to feed the projected rapid increase in population (Silva, 2018). Allohexaploid bread wheat (*Triticum aestivum*, $2n = 6x = 42$, AABBDD) is of relatively recent origin, having evolved ~8,500 years ago following two separate interspecific hybridization events involving three diploid donor species ($n = 7$; Leach et al., 2014). First, the tetraploid (pasta) wheat (*T. turgidum* L. ssp. *durum*; $2n = 4x = 28$, AABB) arose from a hybridization event between *T. urartu* ($2n = 2x = 14$, AA) and another yet unknown wild diploid relative of *Aegilops speltoides* ($2n = 2x = 14$, BB) about 0.5 million years ago (Dvorak and Akhunov, 2005; Yaakov et al., 2012; Ogbonnaya et al., 2013). The cultivated *T. turgidum* then crossed with *Aegilops tauschii*, a wild diploid relative ($2n = 2x = 14$, DD) resulting in the modern day allohexaploid bread wheat with its 42 chromosomes distributed in the A, B and D homoeologous sets contributed by the three diploid progenitors.

Several phylogenetic studies have been undertaken over the years to characterize the taxonomic relationships between members of the *Triticum* (wheat)—*Aegilops* complex. Polymorphic transposable element (TE) insertion sites have recently been shown to be a promising tool for the analysis of the phylogenetic relationships in wheat (Konovalov et al., 2010; Yaakov et al., 2012, 2013). TE-based DNA marker systems showed relative superiority over other marker systems in resolving phylogenetic relationships due largely to their high variability and informativeness (Konovalov et al., 2010), but such relative superiority might be dependent on the TE activity level during the course of evolution and their ability to generate insertion site polymorphisms.

In plant species, DNA-marker systems based on different TEs [e.g., miniature inverted-repeat transposable element (MITEs), long terminal repeat (LTR) elements, CACTA transposons, etc.] have been exploited for various genetic studies (Hirsch and Springer, 2017; Morata et al., 2018; Quesneville, 2020). MITEs are thought to be a peculiar type of non-autonomous Class II TE, activated by transposases encoded by their related autonomous elements, which might have enhanced their rapid amplification to potentially generate high copy numbers, though the mechanism of their amplification is yet unknown (Casacuberta and Santiago, 2003; Fattash et al., 2013; Chen et al., 2014). However, recent efforts by Castanera et al. (2021) suggested a replicative mechanism underlying the amplification dynamics of MITEs. MITEs typically have relatively short sequences (generally < 600 bp), large copy numbers, an AT-rich sequence, contain terminal inverted repeats (TIRs) and flanked by two short direct repeats referred to as target site duplications (TSD; Casacuberta and Santiago, 2003; Yaakov et al., 2012, 2013; Fattash et al., 2013; Chen et al., 2014; Li et al., 2014). Plant MITEs are categorized into two main families, *Habinger*/*Tourist*-like and *Mariner*/*Stowaway*-like, besides several other minor families. Recent genome-wide analysis of MITEs based on genome drafts of four wheat and related species [polyploids: *T. aestivum* and *T. turgidum* ssp. *dicoccoides*, and diploids: *Ae. tauschii* and *T. urartu*] involving 239,126 retrieved MITE insertions showed the *Stowaway-like* superfamily as the most abundant (83.4%) in the wheat genome, followed by *Tourist-like* superfamily (4.9%), *Mutator* (2.7%), with 8.9% being unknown; and novel wheat-unique family named "Inbar" belonging to the Stowaway-like superfamily, was also identified in this large-scale study (Keidar-Friedman et al., 2018). The relative on the most abundant MITEs in the wheat genome was found to be the *Stowaway*-like family (62.6%), followed by the *Tourist*-like family (12.1%), while all other families were not found. The relatively small size and high copy numbers of MITEs facilitate their invasiveness and frequent insertion into genomic regions such as promoters, untranslated regions, introns or coding sequences of genes; though they are thought to predominate in the non-coding regions of eukaryotic genes (Han and Wessler, 2010; Li et al., 2014). MITEs are often inserted within gene-rich euchromatic regions and frequently found associated with genes (Wessler et al., 1995; Yasuda et al., 2013; Chen et al., 2014).

Insertion site polymorphisms generated by MITEs can be a valuable molecular marker system for exploitation in various genetic and breeding studies (Monden et al., 2009; Shirasawa et al., 2012; Yaakov et al., 2012; Yaakov and Kashkush, 2012; Mondal et al., 2014; Wang et al., 2020). The simple inheritance of MITE insertion site polymorphisms, their low cost of detection, their dominant and/or co-dominant nature, and their ability to generate polymorphisms even between closely related genomes makes them suitable DNA markers for studying genetic diversity, association mapping and trait mapping.

In this study, 52 TE insertion polymorphism markers selected from 13 *Stowaway*-like MITE families (Yaakov et al., 2012) were investigated to ascertain their efficiency for exploitation as a robust DNA marker system for genetic diversity, linkage analysis and evolutionary studies in wheat. Furthermore, our investigation of the excision frequency of 16 polymorphic MITE markers in one of the parents of our mapping population (*T. aestivum* cv. Chinese Spring) revealed a putative stable inheritance, which is promising for genetic linkage analysis, evolutionary studies and as a useful tool for wheat molecular breeding.

Materials and methods

Plant materials and DNA extraction

In this study, we used 17 *Triticum* and *Aegilops* accessions (Table 1). In the initial amplification and isolation of MITE

TABLE 1 List of plant materials used for the amplification and isolation of MITE fragments and study of the distribution of MITE insertions and evolutionary relationships in the *Triticum*–*Aegilops* complex.

Species/genome	Genotype or accession	Abbreviation
T. aestivum, 2*n* = 6x = 42, AABBDD	Chinese Spring	CS
T. aestivum, 2*n* = 6x = 42, AABBDD	Norin 61	N61
T. aestivum, 2*n* = 6x = 42, AABBDD	Synthetic hexaploid wheat ABD. No.4	SHW ABD4
*T. aestivum**, 2*n* = 6x = 42, AABBDD	Synthetic 72	Syn72
*T. aestivum**, 2*n* = 6x = 42, AABBDD	Multiple synthetic derivative Original #1	MSD-original #1
*T. aestivum**, 2*n* = 6x = 42, AABBDD	MSD—2 (Waxless subpopulation)	MSD-2 waxless
*T. aestivum**, 2*n* = 6x = 42, AABBDD	MSD—5 (Heat-tolerant subpopulation)	MSD-5 heat-tolerant
*T. aestivum**, 2*n* = 6x = 42, AABBDD	Cytoplasmic substitution line 3–1	Cyto subst. line3-1
T. turgidum ssp. *carthricum*, 2*n* = 4x = 28, AABB	34H188, KU-138	T. tur
T. durum, 2*n* = 4x = 28, AABB	Langdon	Langdon
T. dicoccoides, 2*n* = 4x =28, AABB	KU-110	-
T. urartu, 2*n* = 2x = 14, AA	KU-199-1	-
T. urartu, 2*n* = 2x = 14, AA	KU-199-10	-
Ae. tauschii, 2*n* = 2x = 14, DD	34H203, KU-20-2	Ae. tau.
Ae. aucheri, 2*n* = 2x = 14, SS	KU-1-3	-
Ae. speltoides, 2*n* = 2x = 14, SS	KU-12962	-
Ae. speltoides, 2*n* = 2x = 14, SS	KU-14602	-

*These plants are artificially produced experimental materials having the same genomes of *T. aestivum*.

fragments, we used five accessions that are indicated by bold interface in Table 1. After the initial screening and isolation, we used an additional 12 accessions to study the variation in MITE insertions and phylogenetic analysis. *T. carthlicum* and *Ae. tauschii* are the parents of SHW ABD4 (Table 1). Syn.72 is an amphidiploid between Langdon and *Ae. tauschii* acc. PI508262. MSD-original #1, MSD-2 waxless and MSD-5 heat tolerant are multiple synthetic derivative lines selected from the original multiple synthetic derivatives population, the waxless and heat-tolerant subpopulations, respectively (Tsujimoto et al., 2015; Elbashir et al., 2017). The genomic DNA of all plants was isolated from fresh young leaf tissue (~0.5 g) using a modified CTAB-based miniprep extraction method. Briefly, 0.5 g of freshly harvested young wheat leaf samples were collected into a 2-ml eppendorf tube (frozen in liquid nitrogen) and kept at −80°C until ground into a fine powder (under liquid nitrogen) using the MicroMixer. A 1-ml pre-heated (65°C) 3% CTAB extraction buffer [containing 3% (w/v) CTAB, 1.4 M NaCl, 0.1 M Tris–HCl (pH 8.0), 0.02 M EDTA (pH 8.0), and 1% (v/v) β-Mercaptoethanol] was added to the ground frozen tissue and mixed briefly by tube inversions; and then tissue homogenization was carried out in a water bath set at 55°C for 30 min. Following tissue homogenization, 800 µl of Chloroform: Isoamyl alcohol (CI, 24:1) was added and gently but thoroughly mixed for 5 min, before centrifugation at 5,000 rpm for 5 min at room temperature. The supernatant was transferred to a new tube; and DNA precipitated with 0.8 × volume isopropanol, and hooked with a Pasteur pipette into a fresh 1.5-ml eppendorf tube. The hooked DNA was washed with 1-ml 70% ethanol and centrifuged at 8,000 rpm for 5 min at room temperature using a microcentrifuge. The supernatant was decanted, and the tubes air-dried. The isolated DNA was dissolved in 500 µl of 0.1 × Tris–HCl (pH 8.0), and 1.5 µl RNAse A (20 mg/ml stock) was added to each tube and incubated at 37°C for 30 min. An equal volume (500 µl) of Chloroform: Isoamyl alcohol (CI, 24:1) was further added and thoroughly mixed by gentle inversion several times. Centrifugation was carried out at 8,000 rpm for 10 min at room temperature and the supernatant carefully transferred to fresh tubes. DNA was precipitated by adding 2× vol of ice-cold absolute ethanol, and the tubes mixed well (by several inversions) and placed at −20°C for 1 h or overnight. The DNA was hooked with a Pasteur pipette into 1.5-ml eppendorf tube and washed with 1-ml 70% ethanol; and centrifuged at 7,000 rpm for 5 min at room temperature. The supernatant was removed, and the DNA air-dried and resuspended in 100 ul of 0.1 × Tris–HCl (pH 8.0). DNA concentration and quality were determined using a Nanodrop spectrophotometer and further confirmed by agarose gel electrophoresis.

PCR amplification of MITE fragments

A total of 52 primer pairs designed from flanking sequences surrounding intact MITEs (Yaakov et al., 2012) were used for the amplification of MITE fragments in this study (see Supplementary Table S1). PCR for amplification of MITE fragments was performed in a total reaction volume of 25 µl, containing 12.5 µl PCR Master Mix (Promega), 1.0 µl of genomic DNA (~50 ng/µl), 1.25 µl of each site-specific primer (6.1 pmol/µl) and 9.0 µl MilliQ water. The PCR was performed in a thermal cycler (GeneAmp PCR System 9,700, Applied Biosystems) using touchdown annealing temperature conditions as follows: initial denaturation at 94°C for 3 min; then 35 cycles with annealing decreasing by 2°C: 5 cycles of 94°C for 1 min, 60°C for 1 min, 72°C for 90 s; 5 cycles of 94°C for 1 min, 58°C for 1 min, 72°C for 90 s; 5 cycles of 94°C for1 min, 56°C for 1 min, 72°C for 90 s; followed

TABLE 2 Number of PCR–SCAR MITE markers obtained from 13 Stowaway-like MITE families.

S/No.	Stowaway-like MTE family	Approximate size (bp)	No. of PCR-SCAR MITE markers obtained
1	Thalos	164	8
2	Fortuna	327	4
3	Athos	85	6
4	Oleus	152	7
5	Minos	240	4
6	Eos	353	3
7	Pan	127	3
8	Aison	219	1
9	Icarus	112	2
10	Phoebus	322	4
11	Polyphemus	232	3
12	Victor	276	1
13	Xados	116	2
Total			48

by 20 cycles of 94°C for 1 min, 54°C for 1 min, 72°C for 90 s; and a final extension step at 72°C for 4 min. MITEs amplicons were size-separated using 1.5% agarose (Nippon gene, Japan) gel at 100 V for 25 min and stained with ethidium bromide. Stained gels were visualized under UV light and photographed using a gel documentation system.

Genotyping of RILs with polymorphic MITE fragments

Preliminary genotyping of PCR-SCAR MITE markers in a wheat recombinant inbred lines (RILs) mapping population and its parental genotypes (CS and SHW ABD4) was performed using five polymorphic primers from four *Stowaway*-like MITE families: *Thalos*, *Athos*, *Minos*, and *Eos*). PCR for this analysis was carried out using a total reaction volume of 11.0 µl containing 5.5 µl PCR Master Mix, 4.0 µl of genomic DNA (~3.125 ng/µl), 0.3 µl of each polymorphic MITE primer (6.1 pmol/µl), and 0.9 µl MilliQ water. The thermal cycling conditions were as described above, and amplicons were separated on 1.5% agarose gel at 100 V for 25 min and documented as described above. The easily scorable bands were analyzed and integrated into our wheat genetic linkage map.

Determining MITEs excision frequency

MITEs excision was assessed in 129 plants of *T. aestivum* cv. CS by PCR using 16 polymorphic CS-specific insertion sites from 9 MITEs families: Thal-EU835982, Thal-EU835981, Thal-CQ169689; Fort-AY663392, Fort-EU835980; Atho-AM932680, Atho-AB201447, Atho-DQ517494; Oleu-AF325198; Mino-FN564434; Eos-FN564434; Pan-DQ871219; Pan-FN564434; Phoebus-102; Polyphemus-110, and Polyphemus-111 (see Supplementary Table S1 for details of these primer sequences). PCR was carried out in a total reaction volume of 13 µl containing 6.5 µl PCR Master Mix, 2.0 µl of genomic DNA (~12.5 ng/µl), 0.625 µl of each polymorphic insertion site - primer (6.1 pmol/µl) and 2.25 µl MilliQ water. The thermal cycling conditions were as described above, and amplicons were separated on 1.5% agarose gel at 100 V for 25 min and documented as described above.

Statistical analysis

Analysis of similarity (ANOSIM) was conducted to confirm the statistical differences between the different genome types and ploidy levels. Phylogenetic analysis was performed based on Jaccard similarity and then a group average hierarchical clustering was conducted based on a SIMPROF test with 99,999 simulations (alpha < 0.05) using PRIMER6 (PRIMER-E).[1] Furthermore, a principal component analysis (PCA) was conducted on the similarity to statistically reveal the degree of dissimilarity between hexaploids, tetraploids, and diploids.

Results

Amplification of MITE fragments and their isolation in wheat and related species

Five genotypes, CS, SHW ABD4, Norin 61, 34H188, and 34H203 (Table 1) were initially used to detect the presence/absence of MITE fragments and facilitate their isolation *via* PCR. Of the 52 tested MITE primer pairs selected from 13 *Stowaway*-like families (Yaakov et al., 2012; Supplementary Table S1), 48 primer pairs produced amplified products in at least one of the tested genotypes, while four primer pairs did not yield amplified products in any of the tested genotypes which suggested a lack of insertion sites. As has been previously reported (Yaakov et al., 2012), the majority of amplified MITE sequences used in this study were from the B genome (59%, i.e., 17 of the 29 MITE sequences with known chromosomal location), while 12 of the 29 MITE sequences were from the A (6 MITE sequences, 21%) and D (6 MITE sequences, 21%) genomes. The DNA fragments produced from the 48 amplified PCR-SCAR MITE markers contained DNA sequences from 13 *Stowaway*-like MITE families (Yaakov et al., 2012; Supplementary Table S1), as shown in Table 2.

The primer pairs used in this study were designed by Yaakov et al. (2012) to amplify the MITE insertions and their flanking host sequences to produce the expected full amplicon

[1] https://onlinelibrary.wiley.com/doi/10.1111/j.1442-9993.1993.tb00438.x

size, which is termed "full site" (i.e., the size of the MITE insertion plus the flanking sequences), compared to an "empty site," i.e., without a MITE insertion, in which the amplicon will be relatively shorter in size, consisting of only the flanking sequence. An example of a site-specific PCR for *Thalos* (Thal-GQ169689 in Supplementary Table S1), which was inserted in the 11**th** intron of the *plastid glutamine synthetase* 2 (*GS2*) gene, is shown in Figure 1. In this case (Figure 1), the expected **"**full site**"** is 519 bp-long, while the **"**empty site**"** is 367 bp-long. While the hexaploids CS and N61 had the **"**full site**"** fragment containing the MITE insertion, ABD 4, *T. durum* and *Ae tauschii* lacked the MITE insertion. Very faint bands corresponding to the size of the empty site were observed in the hexaploid species (CS and N61), which could be a footprint probably due to the loss of the fragment in a small percentage of the cells in the tissue. Sequence comparison of the isolated site-specific PCR fragment from CS ("full site"), *Ae tauschii* ("empty site") and a database sequence (bread wheat, Thal-GQ169688) confirmed that the fragment differences in the gel are due to the presence or absence of this *Thalos* element (Figure 2.). Other examples of MITE amplified fragments observed in this study are shown in Figure 3.

Distribution of MITEs in 17 accessions of wheat and related species and evolutionary relationships inferred by MITE insertion polymorphisms

To assess MITE dynamics and determine whether the proliferation of MITEs was of recent origin in allohexaploid wheat, 17 accessions of wheat and related species comprising hexaploids, tetraploids and their diploid progenitors were genotyped with the 43 MITE primer pairs. Of the 43 markers studied, five (12%) were monomorphic, indicating that they may be of fossil origin. Different patterns of MITE insertion polymorphisms in relation to ploidy and/or genome type were also observed (Figure 4). MITE insertion numbers (i.e., abundance) increased with ploidy level (Table 3). The total number of amplified "full sites" ranged from 26 [in SHW ABD4 (AABBDD)] to 37 [in CS (AABBDD) and its cytoplasmic substitution line 3-1(AABBDD)], 22 [in 34H188 (*T. turgidum*, AABB) and Langdon (*T. durum*, AABB)] to 23 [in KU-110 (*T. dicoccoides*, AABB)], and 10 [in KU-1-3 (*Ae. aucheri*, SS)] to 18 [in KU-199-1 and KU-199-10 (*T. urartu*, AA); and 34H203 (KU-20-2, *Ae. tauschii*, DD)] as shown in Table 3. Significant variation in *"full site"* fragments was found among the hexaploids and the BB genome group of the diploids, unlike the AABB and AA/DD genome groups which showed little or no variation in full sites (Table 3). The proportion of polymorphic bands among these accessions ranged from 44 to 76% in *Ae. speltoides* (KU-14602) and *T. aestivum* cv. CS, respectively. An analysis of similarity (ANOSIM) was conducted to further investigate differences at the genomes or polyploid levels. ANOSIM revealed a high degree of dissimilarity between some hexaploids, tetraploids and diploids ($p < 5\%$ and $R > 0.75$). A high MITE proliferation was observed in allohexaploid bread wheat (32.6 ± 3.8 insertion sites) relative to the tetraploid (22.3 ± 0.6 sites) and diploid (15.0 ± 3.5 sites) progenitors, which lends support to the notion that rapid activation of transposons occurred recently after the formation of wheat.

The observed promising amplification of the MITE fragments in the subset of five accessions (as indicated above) enabled us to subsequently extend a survey of the MITE insertion polymorphisms in a relatively larger set of accessions of wheat and related species shown in Table 1; and different patterns of MITE insertion polymorphisms observed in the 17 accessions of wheat and related species of ploidy and/or genome types shown in Figure 4. Figure 4B, for instance, showed a ~450 bp-long fragment unique to the two *T. urartu* accessions, which might possibly be due to element dimer, which are known to form rapidly during

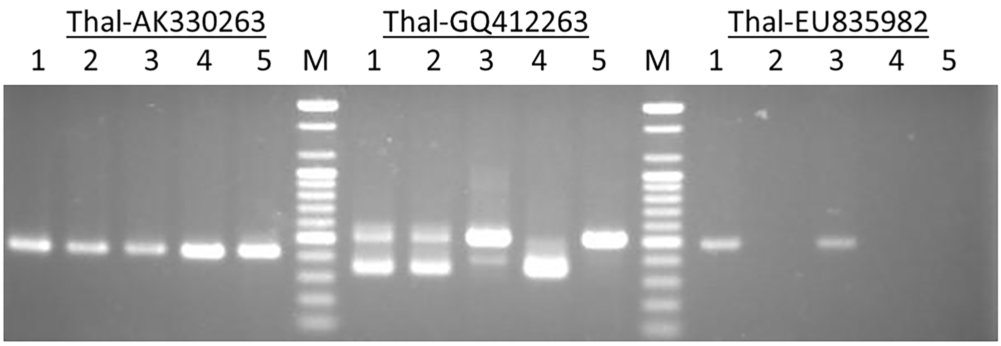


FIGURE 1
An example of a site-specific PCR for *Thalos* (Thal-GQ169689) that was inserted in the 11th intron of the *plastid glutamine synthetase* 2 (*GS2*) gene in five accessions. The expected "full site," i.e., the larger band of ~519bp was found only in the hexaploid Chinese Spring and Norin 61 (Lanes 1 and 3). The Synthetic hexaploid wheat (Lane 2) along with the tetraploid *T. turgidum* (Lane 4) and the diploid *Ae. tauschii* (Lane 5) lacked the full insertion site and only the "empty site," i.e., the lower band of ~367bp was found.

```
CS                        1 GCCACACAAATTACAGGTTCCACTCTTTTCTGTTAATATTTATTTATCCC 50
Ae. tauschii              1 GCCACACAAATTACAGGTTCCACTCTTTTCTGTTAATATTTATTTATCCC 50
Source_Thal-GQ169688.1    1 GCCACACAAATTACAGGTTCCACTCTTTTCTGTTAATATTTATTTATCCC 50

CS                       51 GCATTACTTTTGCAAAGTATATCTTGCTGTATATTTTCTTCGAGAAACCT 100
Ae. tauschii             51 GCATTACTTTTACAAAGTATATCTTGCTGTATATTTTTTTCGAGAAACCT 100
Source_Thal-GQ169688.1   51 GCATTACTTTTGCAAAGTATATCTTGCTGTATATTTTCTTCGAGAAACCT 100

CS                      101 ATATTAGAAAAATTCAGAAAACCTATTATAGCTGCGTTGAAGTAAATACA 150
Ae. tauschii            101 ATATTAAAAAATTTCAGAAAACCTATTATAGCTGCGTTGAAGTAAATACA 150
Source_Thal-GQ169688.1  101 ATATTAGAAAAATTCAGAAAACCTATTATAGCTGCGTTGAAGTAAATACA 150

CS                      151 ACGGGGATTTTGTGGGAAGACAACATATGCTGATACTAACAGACAATGTT 200
Ae. tauschii            151 ACGGGGATTTTGTGGGAAGATAACATATGCTGATACTAACAGACAATGTT 200
Source_Thal-GQ169688.1  151 ACGGGGATTTTGTGGGAAGACAACATATGCTGATACTAACAGACAATGTT 200

CS                      201 CCTCGAAAGACTATATTTAAACTTAAGCCCTGTTCGGATCCACTCCGCTC 250
Ae. tauschii            201 CCTCGAAAGACTATATTTAAACTTA------------------------- 225
Source_Thal-GQ169688.1  201 CCTCGAAAGACTATATTTAAACTTAAGCCCTGTTCGGATCCACTCCGCTC 250

CS                      251 CACAGCTGCAACTCCCGGAGCGGAGGGAGCGACAGCGCAATCGCACGGAG 300
Ae. tauschii            225 -------------------------------------------------- 225
Source_Thal-GQ169688.1  251 CACAGCTGCAACTCCCGGAGCGGAGGGAGCGACAGCGCAATCGCACGGAG 300

CS                      301 CTGCTCAGACCCAGCTCCTCGCACGGAGCGGAGTTTTGAAATGGGAGAAG 350
Ae. tauschii            225 -------------------------------------------------- 225
Source_Thal-GQ169688.1  301 CTGCTCAGACCCAGCTCCTCGCACGGAGCGGAGTTTTGAAATGGGAGAAG 350

CS                      351 TACCGAACAGGCACTTATTATCAAGTGTTATTCAGAATAGACATGTCTTC 400
Ae. tauschii            226 -----------------TTATCAAGTGTTATTCAGAATAGACATGTCTTC 258
Source_Thal-GQ169688.1  351 TACCGAACAGGCACTTATTATCAAGTGTTATTCAGAATAGACATGTCTTC 400

CS                      401 AGTATAGTTACTAACCTTTTTGGTAGTTTTTTCTCAATGCTTGATATAGT 450
Ae. tauschii            259 AGTATAGTTACTAACCTTTTTGGTAGTTTTTTCTCAATGCTTGATATAGT 308
Source_Thal-GQ169688.1  401 AGTATAGTTACTAACCTTTTTGGTAGTTTTTTCTCAATGCTTGATATAGT 450

CS                      451 TTGTCCTTAATTTGCAAGTGAGAAACAATCTTTTCTTGTTGTTGCAAATG 500
Ae. tauschii            309 TTGTCCTTAATTTGCAAGTGAGAAACAATCTTTTCTTGTTGTTGCAAATG 358
Source_Thal-GQ169688.1  451 TTGTCCTTAATTTGCAAGTGAGAAACAATCTTTTCTTGTTGTTGCAAATG 500

CS                      501 TAGCACATTGAGCATGCGT                                519
Ae. tauschii            359 TAGCACATTGAGCATGCGT                                377
Source_Thal-GQ169688.1  501 TAGCACATTGAGCATGCGC                                519
```

FIGURE 2
Multiple sequence alignment of sequenced amplified fragments (see Figure 3) corresponding to *Triticum aestivum* cv. Chinese Spring, *Ae. tauschii* and the best match NCBI database (Thal-GQ169688) bread wheat sequence. The *Thalos* element is indicated in blue letters, while the flanking sequences are indicated in black letters. The two *T. aestivum* cultivars Chinese Spring and the NCBI database (bread wheat) have the element, while the *Ae. tauschii* accession lacked it.

periods of active transposition (McGurk and Barbash, 2018), thereby leading to site duplication in *T. urartu*.

To further confirm the suitability of the MITE markers to explain the polymorphism between the different accessions tested, phylogenetic trees were produced and PCA analysis was conducted based on the polymorphic "full sites" (Figures 5A,B), which revealed different strata of evolutionary relationships in the *Triticum*–*Aegilops* complex. Generally, the accessions were grouped in five groups in relation to their ploidy and genome constitution: AA, BB or SS, DD, AABB, AABBDD (Figures 5A,B). Within the hexaploids group (Group 1), an original accession of multiple synthetic derivatives was neatly separated from the other accessions including its MSD-2 waxless and MSD-5 heat-tolerant offshoots that are in sub-cluster along with Norin 61, CS and its cytoplasmic substitution line 3–1 (Figures 5A,B). The synthetics were closer to the tetraploids from which they were re-synthesized in Group 2. Group 3 is made up only of diploid accessions comprising of three distantly separated sub-clades in relation to their genome constitution: *Ae tauschii* (DD), *T. urartu* (AA) and *Ae. aucheri* and *Ae. speltoides* (SS). Among the diploid progenitors, our results indicated that, *Ae. tauschii* (DD) is more genetically distant from the other diploid species, followed by *T. urartu* (AA) and then the SS or BB genome type (*Ae. aucheri* and *Ae. speltoides*, which were sub-clustered together; Figures 5A,B). The principal component analysis clearly revealed the relationship among the accessions in relation to their genome groups and ploidy; with the first two principal components explaining 59.7% of the total variation (Figure 5B; Supplementary Table S2). The genomic distribution, as well as the genetic relationships detected by these MITEs insertion polymorphisms, is consistent with the known evolutionary history and phylogenetic relationships in wheat.

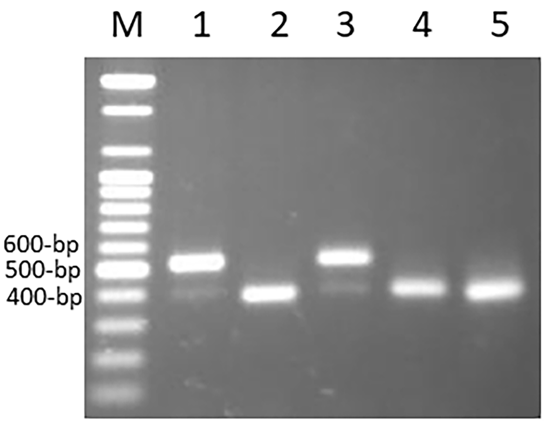


FIGURE 3
An example of amplified MITE fragments observed in five *Triticum–Aegilops* accessions with three MITE primers (Thal-AK330263, Thal-GQ412263, andThal-EU835982). The five accessions are: (1) Chinese Spring; (2) SHW ABD 4; (3) Norin 61; (4) *T. turgidum*; (5) *Ae. tauschii.*

Stability of MITE insertions in the genome of wheat cv. CS

As stable inheritance of transposable elements is a desirable feature for their usefulness as a valuable DNA marker system, we investigated the possible excision of MITEs in the wheat genome using CS, one of the parents of our mapping population. As shown in Table 3, the highest frequency of polymorphic MITE insertion sites was observed in this cultivar. With 16 polymorphic MITE markers ×129 plants amounting to a total of 2,064 sites investigated for possible excision, no single excision was observed in this wheat accession, i.e., the frequency of insertion stability was 100% (data not shown). This observation indicates the value of these MITEs as valuable molecular markers in wheat.

MITE insertion polymorphism rate in parents of our mapping cross, preliminary genotyping in a RIL population, and linkage mapping

Of a total of 48 amplified MITE markers assayed, 15 primer pairs yielded polymorphic bands between the two parents of our intraspecific mapping cross (CS and ABD4), indicating a 31% polymorphism rate. Of these polymorphic primer pairs, 10 and four showed presence/absence and co-dominant inheritance, respectively, while one primer pair produced both presence/absence/co-dominant fragments between the two parents. Preliminary genotyping using five polymorphic PCR-SCAR MITE primers in our mapping cross-population of 104 RILs yielded six scorable bands with five showing the expected 1:1 segregation ratio for simplex markers, while one of the two markers scored from Athos-DQ5176494 with an approximate size of 277 bp fitted a 3:1 ratio for a duplex × nulliplex marker (Table 4). An example of the segregation of MITE markers in a RIL mapping population is shown in Supplementary Figure 1. Interestingly, a high frequency of simplex alleles (83%) was observed with these MITE markers, which indicates the potential usefulness of these stable MITEs for efficient mapping of this complex allohexaploid genome. Four of the six scorable MITE markers obtained for the preliminary genotyping of a RIL mapping population (Table 4) have already been integrated into our current DArT-and SNP-based linkage mapping constructed. The reported chromosomal location of these tested markers was confirmed by our linkage mapping effort, while the location of a hitherto unassigned MITE marker [TE 30–2, Minos-EF567062] which mapped to chromosome 5D was established from our linkage mapping studies.

Discussion

We utilized a set of published MITE markers from 13 *Stowaway*-like families, whose main advantage is derived from the presence or absence of a small-sized element that can be easily assayed (Shirasawa et al., 2012; Yaakov et al., 2012) for genomic studies in wheat and related species. The simple inheritance of MITEs, their relatively inexpensive and co-dominant assay method, as well as their relatively high polymorphism rate (even between closely related taxa) make MITE markers highly promising for exploitation in different aspects of genomic studies. The potential of MITE insertion polymorphisms as an efficient DNA marker system has been demonstrated in several plant species including wheat (Yaakov et al., 2012, 2013), groundnut (Shirasawa et al., 2012; Gayathri et al., 2018), rice (Monden et al., 2009), barley (Lyons et al., 2008) and *Brassica* species (Sampath et al., 2014). The MITE markers tested in this study have been reportedly shown to be in association with genes or coding sequences with insertions occurring in introns of known genes, repetitive regions or in intergenic regions (Yaakov et al., 2012; Supplementary Table S1). In this study, similarly to the previous report in wheat by Yaakov et al. (2012), we found these MITE insertional polymorphisms as highly efficient evolutionary markers suited for inferring evolutionary relationships in the *Triticum–Aegilops* complex, genome analysis and linkage mapping in wheat.

MITEs have been known to be found often in close proximity to or within gene-rich euchromatic regions, where they might alter the expression and function of the associated gene(s) (Jiang et al., 2004; Naito et al., 2009). An *in-silico* study by Sabot et al. (2005) showed that ~43% of the MITE insertions occurred in association with wheat genes. Moreover, Yaakov et al. (2012) found 60% of the wheat MITEs used in this study to be associated with genes, with ~51% of the insertions occurring in the introns (Supplementary Table S1). Such insertions of active MITEs in gene-rich regions might drive the evolution of novel genes in wheat, thereby leading to altered phenotypes. For instance, Gorafi

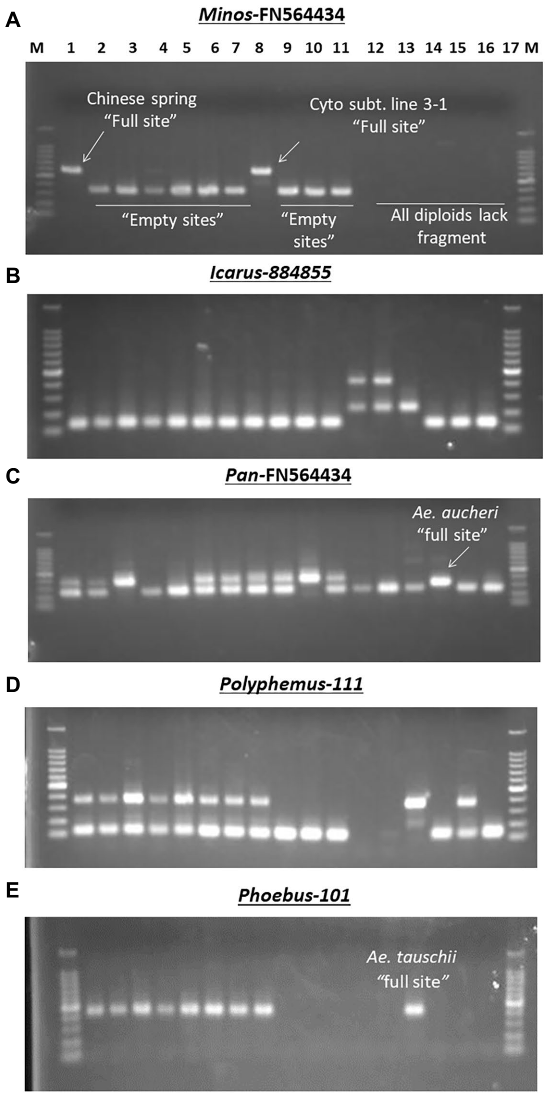


FIGURE 4
Examples of different MITE insertion polymorphism patterns observed in 17 accessions of wheat and related species in relation to ploidy and/or genome types. **(A)** Amplification patterns observed with *Minos*-FN564434: in this case, only the hexaploid cv. Chinese Spring (Lane 1) and its Cytoplasmic substitution line 3–1 (lane 8) had the *Minos* element, termed as "full site"; the rest of the hexaploid accessions and the tetraploids had only the "empty site," while all the diploids (lanes 12–17) showed band absence. **(B)** Amplification patterns observed with *Icarus*-884,855: all the hexaploid and tetraploid accessions (lanes 1–11) and the SS diploid types (lanes 15–17) lacked the *Icarus* element with only the "empty site" present; in addition to the Icarus element present in *Ae. tauschii* (DD, lane 14), the two *T. urartu* (AA, lanes 12 and 13) possessed an additional ~450bp-long unique fragment. **(C)** Amplification patterns observed with *Pan*–FN564434; all the hexaploid and tetraploid accessions [lanes 1–11, except Norin 61 and MSD-Original #1 (lanes 4 and 5, respectively)] had the *Pan* element; while all the diploids (lanes 12–17) except *Ae. aucheri* diploid accession (KU-1-3, lane 15) showed band absence. The relatively lower fragment position of the *Pan* element in KU-1-3 seems to suggest some sequence deletion in this accession. **(D)** Amplification patterns observed with Polyphemus-111; all the hexaploid accessions (lanes 1–8), *Ae. tauschii* (lane 14) and an accession of *Ae. speltoides* (lane 16) had the *Polyphemus* element; the tetraploid accessions (lanes

(Continued)

Figure 4 Continued
9–11) and *Ae. aucheri* (lane 15) had only the empty site, while the two *T. urartu* accessions (lanes 12 and 13) accessions showed band absence. **(E)** Amplification patterns observed with Phoebus 101; all the hexaploid accessions (lanes 1–8), along with the diploid *Ae. tauschii* (lane 14) had the *Phoebus* element; the tetraploid accessions (lanes 9–11) and the other diploid accessions lacked the element.

et al. (2016) showed a MITE insertion in the promoter region of *Vrn-1A* allele annulled the vernalization requirement of a wheat introgression line. Dai et al. (2021) revealed significant marker-trait associations for five agronomically important traits uncovered by 10 polymorphic markers generated from six MITE specific primer pairs in a diversity panel of 126 *Brassica napus* genotypes. A recent study on the variability of root system architecture in five subspecies of Spanish *T. turgidum* L. revealed differences in unique MITE insertion in the *TtDro1B* gene useful for the reliable differentiation of the subspecies *turgidum* from the *durum* and *polonicum* types (González et al., 2021). Thus, further genetic studies with these MITE markers will provide insights for uncovering useful variation in the genes associated with the activity of these MITE insertions (as shown in Supplementary Table S1) for the molecular breeding of wheat.

The distribution of MITE insertions in relation to different ploidy and genome types (Table 3), revealed the activity of MITEs during evolution. The large difference in observed MITE insertion sites among the hexaploids (26–37 insertion sites), tetraploids (22–23 insertion sites) and their diploid progenitors (10–18 insertions sites), as well as between the diploid genome types (e.g., 10–14 full sites in the SS versus 18 full sites in the AA and DD genome types) which suggest that MITEs might have undergone a recent rapid activation in wheat following the allopolyploidization events. This observation is consistent with the recent report of Keidar-Friedman et al. based on whole genome analysis where the distribution of a total of 239,126 retrieved MITE insertions were found to be 48.2% (in hexaploids), 32,7% (in tetraploid T. turgidum) and 9.6% (in diploids, av. from *T. urartu* and *Ae. tauschii*). Based on the abundance of *Stowaway-like* MITEs in wheat group 7 chromosomes, where more of the 2026 MITEs analyzed were found in 7D (35.79%) relative to the 7A (28.87%) and 7B (35.24%), Lu et al. (2014) suggested that the A and B sub-genomes might have eliminated some repetitive elements during the double hybridization events in allohexaploid wheat. Our present study also revealed more *Stowaway-like* MITE insertions in the D sub-genome relative to the S (or B) sub-genome, but similar numbers of insertion sites were found between the D and A sub-genomes. However, a *de novo* search for MITEs of the entire assembled wheat genome v2 using the MITETracker software enabled the discovery of 6,013 MITE families in the wheat genome, with the MITEs distributed along the chromosomes and associated with gene-rich regions (Crescente et al., 2018). Of the 125,800 different MITEs discovered across the wheat genome based on the MITETracker, the B sub-genome was more

TABLE 3 Distribution of amplified MITE fragments in 17 accessions of wheat and related species.

	No. amplified fragments			No. polymorphic fragments		
Genotype	**Total sites**	**Full sites**	**Empty sites**	**[a]Total sites**	**[b]Full sites**	**[c]Empty sites**
A: Hexaploids						
Chinese spring	48	37 (77.1%)	11 (22.9%)	41 (85.4%)	31 (75.6%)	10 (24.4%)
Synthetic hexaploid wheat ABD. No.4	39	26 (66.7%)	13 (33.3%)	32 (82. %)	20 (62.5%)	12 (37.5%)
Synthetic 72	41	29 (70.7%)	12 (29.3%)	34 (82.9%)	23 (67.7%)	11 (32.4%)
Norin 61	44	32 (72.7%)	12 (27.3%)	37 (84.1%)	26 (70.3%)	11 (32.4%)
Multiple synthetic derivative Original #1	44	32 (72.7%)	12 (27.3%)	37 (84.1%)	26 (70.3%)	11 (29.7%)
MSD—2 (Waxless subpopulation)	48	34 (70.8%)	14 (29.2%)	40 (83.3%)	28 (70.0%)	13 (30.0%)
MSD—5 (Heat-tolerant subpopulation)	46	34 (73.9%)	12 (26.1%)	39 (84.8%)	28 (71.8%)	11 (28.2%)
Cytoplasmic substitution line 3–1	49	37 (75.5%)	12 (24.5%)	42 (85.7%)	31 (73.8%)	11 (26.2%)
B: Tetraploids						
34H188, KU-138	35	22 (62.9%)	13 (37.1%)	28 (80.0%)	16 (57.1%)	12 (42.9%)
Langdon	33	22 (66.7%)	11 (33.3%)	26 (78.8%)	16 (61.5%)	10 (38.5%)
KU-110	35	23 (65.7%)	12 (34.3%)	28 (80.0%)	17 (60.7%)	11 (39.3%)
C: Diploids						
KU-199-1	25	18 (72.0%)	7 (28.0%)	18 (72.0%)	12 (66.7%)	6 (33.3%)
KU-199-10	25	18 (72.0%)	7 (28.0%)	18 (72.0%)	12 (66.7%)	6 (33.3%)
34H203, KU-20-2	26	18 (69.2%)	8 (30.8%)	19 (73.1%)	12 (63.2%)	7 (36.8%)
KU-1-3	16	10 (62.5%)	6 (37.5%)	9 (56.3%)	4 (44.4%)	5 (55.6%)
KU-12962	24	14 (58.3%)	10 (41.7%)	16 (66.7%)	8 (50.0%)	8 (50.0%)
KU-14602	23	12 (52.2%)	11 (47.8%)	16 (69.6%)	7 (43.8%)	9 (56.2%)
Total sites scored (43 primers)	64	46	18	56	41	15
Mean distribution by ploidy						
Hexploids	44.9 ± 3.6	32.6 ± 3.8	12.3 ± 0.9	37.8 ± 3.5	26.6 ± 3.8	11.3 ± 0.9
Tetraploids	34.3 ± 1.2	22.3 ± 0.6	12.0 ± 1.0	27.3 ± 1.2	16.3 ± 0.6	11.0 ± 1.0
Diploids	23.2 ± 3.7	15.0 ± 3.5	8.2 ± 1.9	16.0 ± 3.6	9.2 ± 3.4	6.8 ± 1.5

[a]Values in parenthesis indicate the percentage of total polymorphic sites relative to total amplified sites.
[b]Values in parenthesis indicate the percentage of polymorphic "full" sites relative to the total polymorphic sites.
[c]Values in parenthesis indicate the percentage of polymorphic "empty" sites relative to the total polymorphic sites.

MITEs-rich (40.14%) followed by the A sub-genome (32.81%) with the D sub-genome (27.05%) being the least (Crescente et al., 2018). It seems that this situation of the relative abundance of MITEs accords with the general tendency of TEs in allohexaploid wheat, as Wicker et al. (2022) analyzed long terminal repeat (LTR) retrotransposons (full-length) and found the B sub-genome to be relatively more abundant in the TEs studied, followed by the A sub-genome and then D sub-genome. In future studies, it would be worthwhile to clarify the evolutionary consequences of the relative abundance of MITE insertions in the different sub-genomes in allohexaploid wheat and the potential implications of the likely altered function of associated genes in future wheat breeding.

Phylogenetic and PCA analysis based on the detected MITE insertion site polymorphisms revealed high genetic divergence that clearly classified the accessions in the *Triticum–Aegilops* complex consistent with the known evolutionary history of wheat, and sub-grouped the different accessions according to their specific genome types (SS, AA, DD, AABB, AABBDD). This study showed the efficiency of these evolutionary markers in producing high-resolution phylogenetic trees that sub-grouped the accessions according to their specific genome types, consistent with the findings of Yaakov et al. (2012), thereby lending further support to the hypothesis that MITEs were recently active and proliferated in a species-unique fashion. Our observation indicated MITE activity in recently synthesized allohexaploid wheat (including the multiple synthetic derivatives, MSD; Table 1) high polymorphisms in MITE insertion sites especially between the MSD (32–34 sites), a primary synthetic wheat allohexaploid

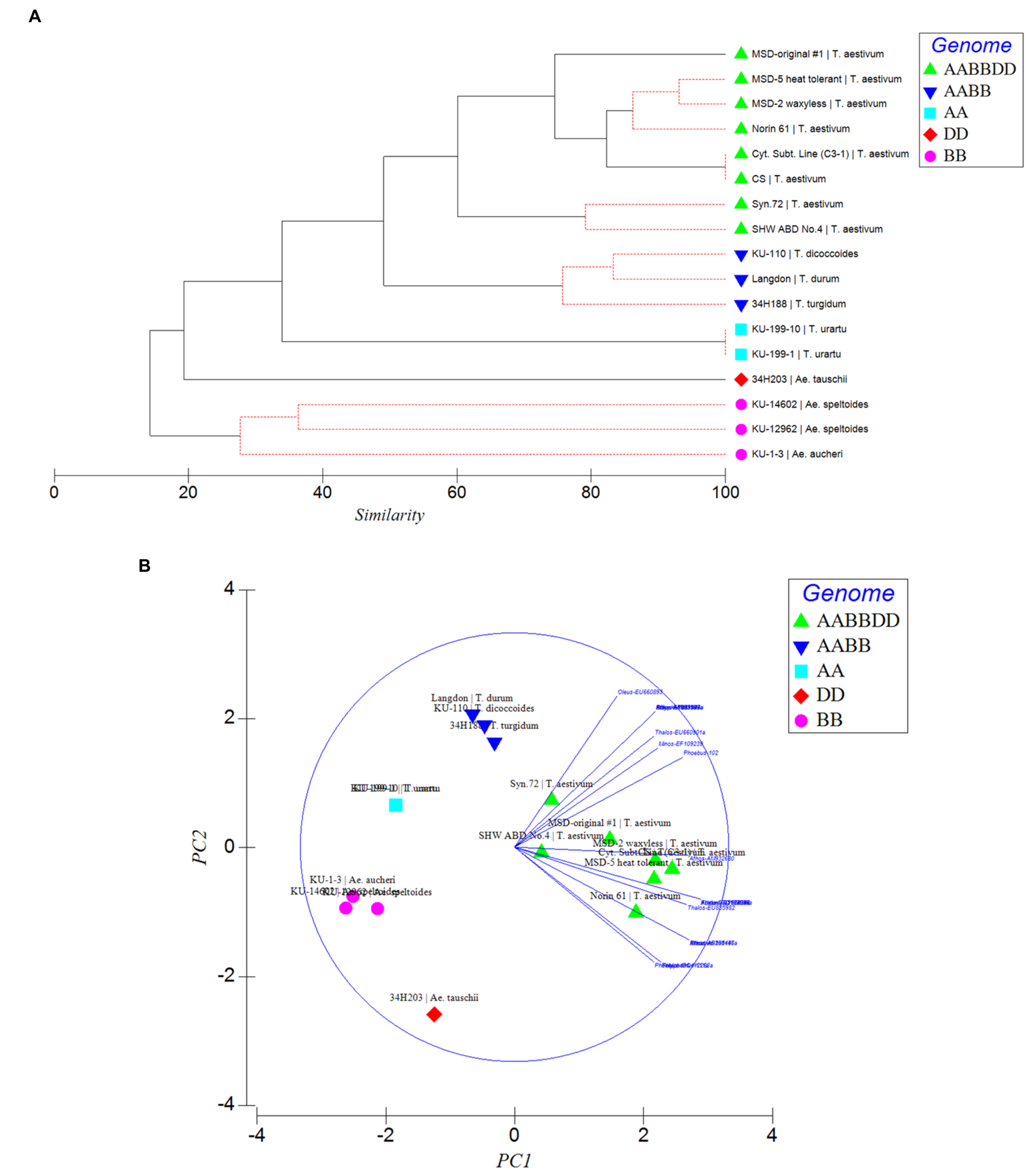


FIGURE 5
Phylogenetic relationships **(A)** and Principal component analysis (PCA; **B**) inferred by MITE insertion polymorphisms in the *Triticum–Aegilops* complex based on polymorphic "full" sites. Different strata of evolutionary relationships were inferred according to ploidy and genome types.

(Syn. 72, 29 sites) and synthetic hexaploid wheat (SHW ABD4, 26 sites). As shown in the phylogenetic tree and PCA (Figures 5A,B), the MSD accessions were clustered in Group 1 with the original MSD line #1 being uniquely separated, while the Syn.72 and ABD4 were sub-clustered in a group close to the tetraploid accessions (Group 2) from which these two accessions were derived. Collectively, our results suggest that MITE transposition activity occurred throughout the course of wheat evolution with rapid activation occurring more recently and might provide further insights in efforts at studying wheat biodiversity and TE-associated gene introgression (Yaakov et al., 2012).

TABLE 4 MITE marker identity, size, chromosomal location, and segregation ratio tested in a Chinese Spring (P_1) × SHW ABD No.4 (P_2) RIL mapping population.

Marker ID	Size (bp)	Location	N[ǂ]	No. present	No. absent	Genetic ratio tested
TE 9–1 (Thal-GQ169689)	598	Chr. 2D	104	52	52	1:1, $\chi^2 = 0.000$
TE 15–1 (Atho-DQ5176494)	356	Chr. 3B	100	56	44	1:1, $\chi^2 = 1.440$
TE 15–2 (Atho-DQ5176494)	277	Chr. 3BL	103	77	26	3:1, $\chi^2 = 0.115$
TE 29–1 (Mino-FN564434)	579	Chr. 3B	103	55	48	1:1, $\chi^2 = 0.476$
TE 30–2 (Mino-EF567062)	559	Chr. 5D*	102	42	60	1:1, $\chi^2 = 3.177$
TE 34–1 (Eos-FN564434)	822	Chr. 3B	101	55	46	1:1, $\chi^2 = 0.802$

[ǂ]Number of plants scored.
*Newly assigned chromosomal location from this study.

Our study on the stability of the PCR–SCAR MITE markers investigated in this study using *T. aestivum* cv. CS, one of the parents of our mapping population, revealed that they are quite stable in the wheat genome. Similar findings on the stable inheritance of *AhMITE1* (frequency of *de novo* excision = 0.00023, Shirasawa et al., 2012) and the rice *mPing* (0.00023, in 96 EG4 plants under normal conditions, Monden et al., 2009), which indicated their value as excellent molecular markers for linkage analysis, had been reported. No single element excision was found in 129 next-generation cv. CS plants investigated with 16 polymorphic MITE primer pairs, which suggested that element excision will not be a concern in the utility of these markers for linkage analysis in wheat.

Our field observations showed that the cv. CS usually presents stable phenotypes which might explain the lack of observed MITE excision; however, the wheat cv. Norin 33 tends to show a sort of genetic instability (Watanabe, 1962), that may be due to TE activity. As a future strategy, it will be useful to characterize the transposition activity in this known genetically unstable cv. Norin 33 relative to other more stable genotypes such as CS (and possibly SHW ABD4) using the multiplexed transposon display technique or high-throughput sequencing to identify new active transposons or their insertions in wheat.

The PCR-SCAR MITE markers demonstrated effectiveness in detecting insertion length polymorphisms in a CS × SHW ABD4 RIL mapping population developed from an intraspecific cross. The two parental lines were genetically partitioned into main clusters on the dendrogram, thereby suggesting a high degree of genetic divergence between them. The observed MITEs polymorphism rate of 31% detected between the two *T. aestivum* parents highlights the high potential of the MITEs in uncovering polymorphisms for the linkage mapping of bread wheat. Moreover, the stable MITEs showed seemingly simple inheritance and with a very high frequency of simplex markers (83% markers with a 1:1 segregation ratio, Table 3) that are potentially useful for the efficient mapping of the complex allohexaploid wheat. It generally seems that TE-based markers such as those based on insertional polymorphisms detectable by PCR-SCAR primers designed from their conserved flanking sequences or their multiplexed assay derivatives [e.g., MITE transposon display, Sequence-specific amplified polymorphism (SSAP), etc.], have a tendency to generate simplex markers that are even more suitable for the molecular mapping of complex polyploid genomes. A study in sweet potato (Monden et al., 2015) based on *Rtsp*-1 retrotransposon insertion polymorphisms found an abundance of simplex markers (~90%) which enhanced the mapping efficiency of the genetically complex autohexaploid sweet potato. In bent grass, a MITE-display analysis using four selective primer pairs (Amundsen et al., 2011) uncovered a total of 139 polymorphic markers, of which 28 markers fitted the expected 1:1 or 3:1 genetic ratio with the simplex marker types being the most abundant (~81.4%). Based on a RIL population of 104 individuals, four simplex MITE markers developed from our preliminary genotyping effort (TE 29–1, Minos-FN564434–Chr. 3B; TE 34–1, Eos-FN564434–Chr. 3B; TE 15–1, Athos-DQ5176494–Chr. 3B; and TE 30–2, Minos-EF567062–Chr. 5D) were integrated into our current DArT-and SNP-based linkage map being constructed. The reported chromosomal location of Minos-FN564434, Eos-FN564434 and Athos-DQ517494 in chromosome 3B was confirmed by our linkage mapping effort, while the chromosomal location of one of the MITE markers [Minos-EF567062, mapped to chromosome 5D (data not shown)] was, to the best of our knowledge, established for the first time from our linkage mapping studies.

Overall, these MITE markers which are simply inherited and well resolved on short gel runs in 1.5% normal agarose gels, that can be substituted with an automated system to increase the efficiency and reduce time, are very promising as cost-effective markers for exploitation in the genome analysis and evolutionary studies in wheat. Several sequences of these MITE fragments have been used to design transposon display primers to extend the frontiers of this research for the rapid genotyping of the wheat genome. Our preliminary transposon display-based MITE genotyping results using two selective primer pairs (data not shown) generated a large number of bands in the accessions tested. Thus, the MITE markers used in this study are promising and will potentially serve as a robust resource for several applications in wheat genetics and molecular breeding including biodiversity and evolutionary studies, linkage analysis, association mapping and MITE-associated modification of gene expression.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

BU, YG, YM, and HT: conceptualization. BU: experimentation. BY, KK, and BU: annotation of sequences and primer design. BU and YM: methodology. BU and YG: data curation. BU, BY, and YG: formal analysis. BY and KK: software. BU, YG, BY, YM, KK, and HT: interpretation of data. BU: writing—original draft. BU, YG, YM, BY, KK, and HT: writing—review and editing. HT: genetic resources, funding acquisition, and project administration. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.995586/full#supplementary-material

SUPPLEMENTARY FIGURE 1
An example of the segregation of a MITE marker (Minos-FN564434) in 104 Chinese Spring (P1) x SHW ABD. No.4 (P2) RIL mapping population. The genotypes of the two parental lines (CS and SHW) are shown.

References

Amundsen, K., Rotter, D., Huaijun, M. L., Messing, J., Jung, G., Belanger, F., et al. (2011). Miniature inverted-repeat transposable element identification and genetic marker development in Agrostis. *Crop. Sci.* 51, 854–861. doi: 10.2135/cropsci2010.04.0215

Casacuberta, J. M., and Santiago, N. (2003). Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene.* 311, 1–11. doi: 10.1016/S0378-1119(03)00557-2

Castanera, R., Vendrell-Mir, P., Amelie Bardil, A., Marie-Christine Carpentier, M.-C., Panaud, O., and Casacuberta, J. M. (2021). Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. *Plant J.* 107, 118–135. doi: 10.1111/tpj.15277

Chen, J., Hu, Q., Zhang, Y., Lu, C., and Kuang, H. (2014). P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res.* 42, D1176–D1181. doi: 10.1093/nar/gkt1000

Crescente, J. M., Zavallo, D., Helguera, M., and Vanzetti, L. S. (2018). MITE tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics* 19:348. doi: 10.1186/s12859-018-2376-y

Dai, S., Hou, J., Qin, M., Dai, Z., Jin, X., Zhao, S., et al. (2021). Diversity and association analysis of important agricultural trait based on miniature inverted-repeat transposable element specific marker in *Brassica napus* L. *Oil Crop Sci.* 6, 28–34. doi: 10.1016/j.ocsci.2021.03.004

Dvorak, J., and Akhunov, E. D. (2005). Tempos of gene deletions and duplications and their relationship to recombination rate during diploid and polyploidy evolution in the *Aegilops* – *Triticum* alliance. *Genetics.* 171, 323–332. doi: 10.1534/genetics.105.041632

Elbashir, A. A. E., Gorafi, Y. S. A., Tahir, I. S. A., Kim, J.-S., and Tsujimoto, H. (2017). Wheat multiple synthetic derivatives: a new source for heat stress tolerance adaptive traits. *Breed. Sci.* 67, 248–256. doi: 10.1270/jsbbs.16204

FAO (2015). Food and Agriculture Organization of the United Nations. Available at: http://faostat3 fao.org/home/index.html

Fattash, I., Rooke, R., Wong, A., Hui, C., Luu, T., Bhardwaj, P., et al. (2013). Miniature inverted-repeat transposable elements MITEs0: discovery, distribution and activity. *Genome* 56, 475–486. doi: 10.1139/gen-2012-0174

Gayathri, M., Shirasawa, K., Varshney, R. K., Pandey, M. K., and Bhat, R. S. (2018). Development of *AhMITE1* markers through genome-wide analysis in peanut (*Arachis hypogaea* L.). *BMC. Res. Notes* 11:10. doi: 10.1186/s13104-017-3121-8

González, J. M., Rodrigo Cañas, R., Cabeza, A., Ruiz, M., Giraldo, P., and Loarce, Y. (2021). Study of variability in root system architecture of Spanish *Triticum turgidum* L. subspecies and analysis of the presence of a MITE element inserted in the *TtDro1B* gene: evolutionary implications. *Agronomy* 11:2294. doi: 10.3390/agronomy11112294

Gorafi, Y. S. A., Eltayeb, E. L., and Tsujimoto, H. (2016). Alteration of wheat vernalization requirement by alien chromosome-mediated transposition of MITE. *Breed. Sci.* 66, 181–190. doi: 10.1270/jsbbs.66.181

Han, Y., and Wessler, S. R. (2010). MITE-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* 38:e199. doi: 10.1093/nar/gkq862

Hirsch, C. D., and Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta.* 1860, 157–165. doi: 10.1016/j.bbagrm.2016.05.010

Jiang, N., Feschotte, C., Zhang, X. Y., and Wessler, S. R. (2004). Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr. Opin. Plant Biol.* 7, 115–119. doi: 10.1016/j.pbi.2004.01.004

Keidar-Friedman, D., Bariah, I., and Kashkush, K. (2018). Genome-wide analyses of miniatureinverted-repeattransposableelementsrevealsnewinsightsinto the evolutionof theTriticum-Aegilops group. *PLoS One* 13:e0204972. doi: 10.1371/journal.pone.0204972

Konovalov, F. A., Goncharov, N. P., Goryunova, S., Shaturova, A., Proshlyakova, T., and Kudryavtsev, A. (2010). Molecular markers based on LTR retrotransposons BARE-1 and Jeli uncover different strata of evolutionary relationships in diploid wheats. *Mol. Genet. Genomics*. 283, 551–563. doi: 10.1007/s00438-010-0539-2

Leach, L., Belfield, E. J., Jiang, C., Brown, C., Mithani, A., and Harberd, N. P. (2014). Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. *BMC Genomics* 15:276. doi: 10.1186/1471-2164-15-276

Li, J., Wang, Z., Peng, H., and Liu, Z. (2014). A MITE insertion into the 3′-UTR regulates the transcription of *TaHSP16.9* in common wheat. *Crop J.* 2, 381–387. doi: 10.1016/j.cj.2014.07.001

Lu, Y.-Z., Wang, L., Yue, H., Wang, M.-X., Deng, P.-C., Edwards, D., et al. (2014). Comparative analysis of *stowaway*-like miniature inverted repeat transposable elements in wheat group 7 chromosomes: abundance, composition and evolution. *J. Syst. Evol.* 52, 743–749. doi: 10.1111/jse.12113

Lyons, M., Cardle, L., Rostoks, N., Waugh, R., and Flavell, A. J. (2008). Isolation, analysis and marker utility of novel miniature inverted repeat transposable elements from the barley genome. *Mol. Genet. Genomics*. 280, 275–285. doi: 10.1007/s00438-008-0363-0

McGurk, M. P., and Barbash, D. A. (2018). Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. *Genome Res.* 28, 714–725. doi: 10.1101/gr.231472.117

Mondal, S., Hande, P., and Badigannavar, A. M. (2014). Identification of transposable element markers for a rust (*Puccinia arachidis* Speg.) resistance gene in cultivated peanut. *J. Phytopathol.* 162, 548–552. doi: 10.1111/jph.12220

Monden, Y., Hara, T., Okada, Y., Jahana, O., Kobayashi, A., Tabuchi, H., et al. (2015). Construction of a linkage map based on retrotransposon insertion polymorphisms in sweet potato via high-throughput sequencing. *Breed. Sci.* 65, 145–153. doi: 10.1270/jsbbs.65.145

Monden, Y., Naito, K., Okumoto, Y., Saito, H., Oki, N., Tsukiyama, T., et al. (2009). High potential of a transposon *mPing* as a marker system in *japonica* × *japonica* cross in Rice. *DNA Res.* 16, 131–140. doi: 10.1093/dnares/dsp004

Morata, J., Marin, F., Payet, J., and Casacuberta, J. M. (2018). Plant lineage-specific amplification of transcription factor binding motifs by miniature inverted-repeat transposable elements (MITEs). *Genome Biol. Evol.* 10, 1210–1220. doi: 10.1093/gbe/evy073

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., et al. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461, 1130–1134. doi: 10.1038/nature08479

Ogbonnaya, F., Abdalla, O., Mujeeb-Kazi, A., Kazi, A. G., Xu, S. S., Gosman, N., et al. (2013). Synthetic Hexaploids: harnessing species of the primary gene pool for wheat improvement. *Plant Breed. Rev.* 37, 35–122. doi: 10.1002/9781118497869.ch2

Quesneville, H. (2020). Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. *Mob. DNA* 11:28. doi: 10.1186/s13100-020-00223-x

Sabot, F., Guyot, R., Wicker, T., Chantret, N., Laubin, B., Chalhoub, B., et al. (2005). Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. *Mol. Genet. Genomics*. 274, 119–130. doi: 10.1007/s00438-005-0012-9

Sampath, P., Murukarthick, J., Izzah, N. K., Lee, J., Choi, H.-I. I., Shirasawa, K., et al. (2014). Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in *Brassica rapa* and *B. oleracea*. *PLoS One* 9:e94499. doi: 10.1371/journal.pone.0094499

Shirasawa, K., Hirakawa, H., Tabata, S., Hasegawa, M., Kiyoshima, H., Suzuki, S., et al. (2012). Characterization of active miniature inverted-repeat transposable elements in the peanut genome. *Theor. Appl. Genet.* 124, 1429–1438. doi: 10.1007/s00122-012-1798-6

Silva, G. (2018). *Feeding the World in 2050 and Beyond – Part 1: Productivity Challenges.* U.S.A.: Michigan State University Extension.

Tester, M., and Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. *Science*. 327, 818–822. doi: 10.1126/science.1183700

Tsujimoto, H., Sohail, Q., and Matsuoka, Y. (2015). "Broadening the genetic diversity of common and durum wheat for abiotic stress tolerance breeding," in *Advances in Wheat Genetics: From Genome to Field.* eds. Y. Ogihara, S. Takumi and H. Handa (Tokyo: Springer), 233–238.

Wang, J., Lu, N., Yi, F., and Xiao, Y. (2020). Identification of transposable elements in conifer and their potential application in breeding. *Evol. Bioinforma.* 16, 117693432093026–117693432093024. doi: 10.1177/1176934320930263

Watanabe, Y. (1962). Studies on the cytological instabilities of common wheat. (in Japanese). *Rep. Tohoku Agric. Exp. Stat.* 23, 69–152.

Wessler, S. R., Bureau, T. E., and White, S. E. (1995). LTR-retrotranposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* 5, 814–821. doi: 10.1016/0959-437X(95)80016-X

Wheat-Wikipedia Available at: https://en.wikipedia.org/wiki/Wheat#cite_note-7 (Accessed May 2016).

Wicker, T., Stritt, C., Sotiropoulous, A. G., Poretti, M., Pozniak, C., Walkowiak, S., et al. (2022). Transposable element populations shed light on the evolutionary history of wheat and the complex co-evolution of autonomous and non-autonomus retrotransposons. *Adv. Genet* 3:2100022. doi: 10.1002/ggn2.202100022

Yaakov, B., Ben-David, S., and Kashkush, K. (2013). Genome-wide analysis of *stowaway*-like MITEs in wheat reveals high sequence conservation, gene association, and genomic diversification. *Plant Physiol.* 161, 486–496. doi: 10.1104/pp.112.204404

Yaakov, B., Ceylan, E., Domb, K., and Kashkush, K. (2012). Marker utility of miniature inverted-repeat transposable elements for wheat biodiversity and evolution. *Theor. Appl. Genet.* 124, 1365–1373. doi: 10.1007/s00122-012-1793-y

Yaakov, B., and Kashkush, K. (2012). Mobilization of stowaway-like MITEs in newly formed allohexaploid wheat species. *Plant Mol. Biol.* 80, 419–427. doi: 10.1007/s11103-012-9957-3

Yasuda, K., Ito, M., Sugita, T., Tsukiyama, T., Saito, H., Naito, K., et al. (2013). Utilization of transposable element *mping* as a novel genetic tool for modification of the stress response in rice. *Mol. Breed.* 32, 505–516. doi: 10.1007/s11032-013-9885-1

TYPE Original Research
PUBLISHED 02 September 2022
DOI 10.3389/fpls.2022.969582

OPEN ACCESS

EDITED BY
Gennady I. Karlov,
All-Russia
Research Institute of Agricultural
Biotechnology, Russia

REVIEWED BY
Wen Yao,
Henan Agricultural University,
China
Peijian Cao,
Zhengzhou Tobacco Research Institute of
CNTC, China

*CORRESPONDENCE
Takanori Yoshikawa
yoshikawa.takanori.5s@kyoto-u.ac.jp

SPECIALTY SECTION
This article was submitted to
Plant Systems and Synthetic Biology,
a section of the journal
Frontiers in Plant Science

RECEIVED 15 June 2022
ACCEPTED 05 August 2022
PUBLISHED 02 September 2022

CITATION
Monden Y, Tanaka H, Funakoshi R, Sunayama S, Yabe K, Kimoto E, Matsumiya K and Yoshikawa T (2022) Comprehensive survey of transposon *mPing* insertion sites and transcriptome analysis for identifying candidate genes controlling high protein content of rice. *Front. Plant Sci.* 13:969582. doi: 10.3389/fpls.2022.969582

Comprehensive survey of transposon *mPing* insertion sites and transcriptome analysis for identifying candidate genes controlling high protein content of rice

Yuki Monden[1], Hirona Tanaka[2], Ryota Funakoshi[3], Seiya Sunayama[3], Kiyotaka Yabe[3], Eri Kimoto[4], Kentaro Matsumiya[4] and Takanori Yoshikawa[4]*

[1]Graduate School of Environmental and Life Science, Okayama University, Okayama, Japan, [2]Faculty of Agriculture, Okayama University, Okayama, Japan, [3]Faculty of Agriculture, Kyoto University, Kyoto, Japan, [4]Graduate School of Agriculture, Kyoto University, Kyoto, Japan

Rice is the most important crop species in the world, being staple food of more than 80% of people in Asia. About 80% of rice grain is composed of carbohydrates (starch), with its protein content as low as 7–8%. Therefore, increasing the protein content of rice offers way to create a stable protein source that contributes to improving malnutrition and health problems worldwide. We detected two rice lines harboring a significantly higher protein content (namely, HP5-7 and HP7-5) in the EG4 population. The EG4 strain of rice is a unique material in that the transposon *mPing* has high transpositional activity and high copy numbers under natural conditions. Other research indicated that *mPing* is abundant in the gene-rich euchromatic regions, suggesting that *mPing* amplification should create new allelic variants, novel regulatory networks, and phenotypic changes in the EG4 population. Here, we aimed to identify the candidate genes and/or *mPing* insertion sites causing high protein content by comprehensively identifying the *mPing* insertion sites and carrying out an RNA-seq-based transcriptome analysis. By utilizing the next-generation sequencing (NGS)-based methods, *ca.* 570 *mPing* insertion sites were identified per line in the EG4 population. Our results also indicated that *mPing* apparently has a preference for inserting itself in the region near a gene, with 38 genes in total found to contain the *mPing* insertion in the HP lines, of which 21 and 17 genes were specific to HP5-7 and HP7-5, respectively. Transcriptome analysis revealed that most of the genes related to protein synthesis (encoding glutelin, prolamin, and globulin) were up-regulated in HP lines relative to the control line. Interestingly, the differentially expressed gene (DEG) analysis revealed that the expression levels of many genes related to photosynthesis decreased in both HP lines; this suggests the amount of starch may have decreased, indirectly contributing to the increased protein content. The high-protein lines studied here are expected to contribute to the development of high protein-content rice by introducing valuable phenotypic traits such as high and stable yield, disease resistance, and abundant nutrients.

KEYWORDS

amplicon sequencing, protein, rice, RNA-seq, starch, transposon

Introduction

The United Nations expects the world population to reach 9.6 billion by 2050. The current world population is 7.3 billion and its demand for protein is 202 million tons, but this is predicted to increase to 267–286 million tons in 2050 (Henchion et al., 2017). Protein is a polymer of amino acids and an indispensable component of body tissues, enzymes, hormones, etc., as well as a substantial source of essential nutrients and energy. Rice is cultivated worldwide, and 60% of the world's population depends upon it as a staple food, with more than 80% of people in Asia eating it (Kawakatsu et al., 2008). Rice also supplies 21% of the world's caloric intake; *ca.* 80% of rice consists of carbohydrates (starch), which is of great value as a staple food, yet its protein content can be as low as 7–8% (Kubota, 2016). Therefore, increasing the protein content of rice would create a stable protein source that can help to overcome malnutrition and health problems not only in Japan but also globally (Chen et al., 2018).

The endosperm of rice is composed of 70–80% starch, 7–10% protein, and 1% lipid (Martin and Fitzgerald, 2002). Rice has four types of seed storage proteins (SSP): albumin, glutelin, prolamin, and globulin. Glutelin is encoded by 15 genes and constitutes 60–80% of the total protein content and is classified into four subfamilies: GluA, GluB, GluC, and GluD (Kawakatsu et al., 2008). Prolamin, encoded by a multigene family of 34 gene copies, makes up 20–30% of total protein and may be categorized as 10 kDa prolamin (RP10), 13 kDa prolamin (RM1, RM2, RM4, and RM9), and 16 kDa prolamin (RP16) according to its molecular weight (Yamagata et al., 1982; Kawakatsu et al., 2008). Globulin is a protein representing 8–10% of total protein; it occurs as two types of polypeptides, 23–27 kDa and 16 kDa, which are structurally homologous to wheat grain glutenin called α-globulin (Ellepola et al., 2006; Kawakatsu et al., 2008). Both albumin and globulin are concentrated in the bran but polishing during the milling process removes a major portion of these proteins (Shewry, 2007). In recent years, enhancing the seed storage proteins to improve rice's nutritive value has emerged as a key target in rice quality breeding (Jiang et al., 2014). Both the content and composition of protein is crucial to quality of rice grain and its nutritional value (Lin et al., 2005; Chen et al., 2017).

Transposable elements (TEs) are mobile genetic elements in the eukaryotic genome, now recognized as an important source of genome evolution and diversification. TEs are a major component of higher plant genomes, accounting for 35% of the rice genome (Turucotte et al., 2001). TEs may alter the expression of neighboring genes *via* insertion into promoter regions, or disrupt the function of protein-coding genes when inserted into the genes, or even change gene structure by altering its splicing and polyadenylation patterns (Kumar and Bennetzen, 1999; Feschotte et al., 2002; Wessler, 2006; Feschotte and Pritham, 2007; Feschotte, 2008; Butelli et al., 2012). TEs are divided into two classes according to whether their transposition involves either RNA intermediates (Class I) or DNA intermediates (Class II; Kumar and Bennetzen, 1999; Feschotte et al., 2002; Wessler, 2006; Feschotte, 2008). Class I elements are transposed by a "copy and paste" mechanism, which involves the reverse transcription of RNA and the integration of a cDNA fragment. Class II elements are excised and integrated into new genomic locations by a 'cut and paste' mechanism. In this respect, Miniature Inverted-repeat Transposable Elements (MITEs) are non-autonomous Class II DNA transposons of small size (< 600 bp) that harbor short terminal inverted repeats (TIRs), capable of attaining high copy numbers in eukaryotic genomes (Feschotte et al., 2002; Feschotte and Pritham, 2007). MITEs have been classified into two superfamilies based on the similarity of their TIRs and their target site duplication (TSD): *Tourist*-like MITEs and *Stowaway*-like MITEs (Feschotte et al., 2002; Feschotte and Pritham, 2007).

mPing is the first active MITE through animal and plant genomes, which was discovered independently by three different assays: long-term rice cell culture (Jiang et al., 2003), short-term anther culture (Kikuchi et al., 2003), and plants derived from gamma-irradiated seeds of the rice cultivar (Nakazaki et al., 2013). The *mPing* element is short (430 bp) with 15 bp TIRs and belongs to the *Tourist* family (Jiang et al., 2003). Because *mPing* has no inherent capacity for transposition, the transposase is provided by two related autonomous elements, *Ping* and *Pong*. The autonomous *Ping* and *Pong* elements are members of the *PIF/Harbinger* superfamily that is widespread in both plants and animals (Hancock et al., 2010). Like most members of that superfamily, *Ping* and *Pong* have two open reading frames (ORFs: ORF1 and ORF2), both of which are required for *mPing* transposition from one place to another on the genome (Yang et al., 2007; Hancock et al., 2010). The ORF1 protein contains a conserved *Myb*-like domain whose involvement in DNA binding was hypothesized (Yang et al., 2007; Hancock et al., 2010). The ORF2 encodes the transposase, which contains a putative Asp-Asp-Glu (DDE) motif that is a signature for transposase catalytic centers (Yang et al., 2007; Hancock et al., 2010).

Although *mPing* is clearly an active MITE, its copy numbers are relatively low, with less than 10 copies found in the subspecies *indica* and ~ 50 copies in the subspecies *japonica*, including the sequenced rice genome (Nipponbare; Naito et al., 2006). Another study revealed that *mPing* had amplified to over 1,000 copies in a few *japonica* rice strains (EG4, HEG4, and related landraces), and is still actively transposing and increasing its copy number, by about 20 copies per plant per generation, without radiation (Naito et al., 2006). A comprehensive survey of *mPing* insertion sites in EG4 strain revealed that *mPing* is enriched in euchromatic, gene-rich regions but rarely present in heterochromatic regions (Naito et al., 2009). Considering both the high activity and insertion preference of *mPing* in the EG4 strain, it is reasonable to envision that *mPing* amplification could create new allelic variants and novel regulatory networks, which may generate plants with more phenotypic diversity and/or novel phenotypic traits.

In this study, we investigated the protein content of the EG4 strain known to exhibit high *mPing* activity, and then used next-generation sequencing (NGS) to comprehensively analyze *mPing* insertion sites in the selected EG4 strains with high protein content. Furthermore, an RNA-seq analysis was performed to quantify the expression levels of genes related to protein synthesis. Using the above information, we sought to identify the *mPing* insertion sites and related genes responsible for high protein content of rice.

Materials and methods

Plant materials

A total of 396 lines of the EG4 population were grown in the experimental paddy field of Kyoto University, Japan. Genomic DNA was extracted from all plants by using the DNeasy Plant Mini Kit following the manufacturer's instructions (QIAGEN, Hilden, Germany). The total RNA was extracted from immature seeds on the 7th day after flowering with three biological replicates per line using the RNeasy Plant Mini Kit according to its manufacturer's instructions (QIAGEN, Hilden, Germany). The extracted RNA was digested with DNase (TAKARA, Shiga, Japan) to remove the remaining gDNA. The yield and quality of the extracted DNA and RNA were confirmed using a NanoDrop 2000 instrument (Thermo Fisher Scientific, Wilmington, DE, United States).

Measurement of protein content

In this experiment, brown rice was used to quantify the protein content of rice seeds. For practical purposes, it is often desirable to use seeds post-milling to qualify their protein content. However, in this study, brown rice was instead used for two main reasons. First, it is difficult to obtain the minimum amount of brown rice required for the rice milling process. Second, an early study showed a 10% reduction in protein content after milling brown rice, but a strong positive correlation was nonetheless detected between the protein content of brown rice and white rice (Higashi et al., 1974). Hence, it was considered sufficient to quantify the protein content in seeds using brown rice.

In 2013, 396 lines of the EG4 population were cultivated for the primary screening of their protein content. To do this, four seeds per panicle from each line were crushed, and the protein content of their rice powder was measured once by the bicinchoninic acid (BCA) assay method. Based on the measurement results, 25 lines with diverse protein content values were selected and subjected to secondary screening in which total protein content was measured from brown rice by applying the improved Dumas method (Jung et al., 2003). This method burns and reduces the sample at a high temperature, and then measures the amount of nitrogen in the generated nitrogen gas (Jung et al., 2003). This method is quick, requiring just a few minutes to analyze each sample, without the use of any deleterious reagents. Protein content was measured in 700–800 mg of brown rice, three times per line, with SUMIGRAPH® (NC-TRINITY, Sumika Chemical Analysis Service, Ltd, Tokyo, Japan).

Preparation of the amplicon sequencing library for *mPing* insertion sites

To determine the *mPing* insertion sites in a comprehensive manner, flanking regions of *mPing* insertion sites were amplified by PCR, and the ensuing products were sequenced on an Illumina platform. An amplicon sequencing library was constructed according to previously described methods (Monden et al., 2014, 2015). First, genomic DNA was fragmented using gTUBE (~ 6 kb; Covaris Inc., MA, United States), and forked adaptors were ligated to the fragmented DNA. These forked adaptors were prepared by annealing two different oligos (Forked_Type1 and Forked_Com; Supplementary Table 1). Primary PCR amplification was performed with *mPing*-specific (*mPing*_1st) and adaptor-specific (AP2-Type1) primer combination, which used the adaptor-ligated DNA as the template (Supplementary Table 1). Nested PCR amplification was carried out using the tailed PCR primers (D501-D503 and D701-D712) with primary PCR products serving as the template. The tailed PCR primers contain the P5 or P7 sequence (Illumina) for hybridization on the sequencing flow cell, and several barcodes for multiplex sequencing. Thus, *mPing*-specific primers (i.e., D501–D503) consisted of a P5 sequence, a barcode sequence, and the *mPing* end sequence, while the adapter-specific primers (i.e., D701–D712) consisted of a P7 sequence, a barcode sequence, and an adapter sequence. The primer combinations of each sample can be found in Supplementary Table 2. The ensuing PCR products were size-selected (400–600 bp) on agarose gels and purified with a QIAquick Gel Extraction Kit (QIAGEN, Hilden, Germany). Purified products were then quantified using a Qubit fluorometer (Invitrogen, Carlsbad, CA, United States) and the size selection range was confirmed in an Agilent 2,200 TapeStation system (Agilent, Santa Clara, CA, United States). The MiSeq sequencing library was prepared by pooling equal amounts of purified barcoded products from each line.

Data analysis

The resulting paired-end reads (150 bp) were analyzed in two ways. The first method followed procedures described in our previous studies (Monden et al., 2014, 2015; Sasai et al., 2019; Hirata et al., 2020), which can detect genome-wide insertion sites of a known TE without any requirement for whole genome sequence information (Monden et al., 2014). The obtained reads were analyzed using Maser, a pipeline execution system of the Cell Innovation Program at the National Institute of Genetics[1]. Adaptor trimming and quality filtering (QV ≥ 30) were performed using cutadapt (Martin, 2011). Filtered reads were trimmed to a specific length that covered most of the sequences. Those reads with ≥10 identical sequences were reduced to a single sequence, in FASTA format, and then clustered using the BLAT self-alignment program (Kent, 2002) under these parameter settings: "-tileSize" = 8, "-minMatch" = 1, "-minScore" = 10, "-repMatch" = −1, and "-oneOff" = 2. This clustering analysis produced many clusters, each corresponding to a separate *mPing* insertion site. An optimal threshold was then set to evaluate the presence or absence of *mPing* insertions: if the number of reads on a given cluster at a specific insertion site comprise < 0.01% of the entire reads for that line, then *mPing* was considered absent from that site. This yielded genotyping information for the presence (1) versus absence (0) of *mPing* insertions in every line.

A second, different analytical method was adopted because the rice reference genome sequence was the first to be deciphered among staple crop species (Eckardt, 2000; Jackson, 2016). Given its subsequent improvements, a highly accurate reference genome sequence of rice is now available (Kawahara et al., 2013; Sakai et al., 2013). The 150-bp paired-end reads were mapped onto the Nipponbare reference genome sequence,[2] to identify the *mPing* insertion sites. Galaxy/NAAC (Advanced Analysis Center, NARO) was used for the follow-up data analysis. First, low quality reads were removed with Trimmomatic (Bolger et al., 2014). After removing low quality reads, high-quality reads were aligned to the Nipponbare reference genome sequence using BWA (Galaxy Version 1.2.3) software (Li and Durbin, 2009). The number of reads within 500 bp from the start site of each aligned read was counted, using an in-house Perl script, and each site with 50 or more aligned reads were designated a *mPing* insertion site. The MiSeq reads for analyzing the *mPing* insertion sites were deposited under the accession number DDBJ: DRA014320.

Experimental validation of *mPing* insertion sites

To verify the existence of *mPing* insertions identified by the above data analysis, PCR primers were designed based on the genomic sequence flanking a *mPing* insertion site (Monden et al., 2009; Supplementary Table 3). Each PCR was run in 10-μl reaction volume that contained 5 μl of 2 × Gflex PCR buffer (Mg^{2+}, dNTP plus), 0.2 μl of Tks Gflex DNA Polymerase, 1 μl of 10 μM primer (forward and reverse) and 1 μl of genomic DNA (l00 ng/μl). The cycling conditions were as follows: 94°C for 2 min; 35 cycles of 98°C for 10 s, 60°C for 30 s, and 68°C for 30 s; then 68°C for 3 min. Amplified products were visualized using electrophoresis on a 1.5% agarose gel.

RNA-sequencing

The RNA-sequencing (RNA-seq) library was sequenced using the NovaSeq system (Illumina, San Diego, CA, United States). The paired-end short reads with a read length of 150 bp were analyzed as follows. Adaptor trimming and quality filtering (QV ≥ 30) were performed using Trim Galore,[3] after which the remaining clean reads were aligned to the Nipponbare reference genome sequence, by using HISAT2 (Kim et al., 2019), and their expression levels were calculated by StringTie (Pertea et al., 2015, 2016). To determine the differentially expressed genes (DEGs), the obtained expression levels were normalized and log-transformed using DESeq2 (Love et al., 2014). Principal component analysis (PCA) was implemented using the "prcomp" function. A biplot graph was visualized with "ggfortify" package in R (Tang et al., 2016). A heatmap was generated by "clustermap" in the seaborn statistical data visualization library in Python (Waskom, 2021). Gene ontology (GO) enrichment analysis was carried out using ShinyGO (Ge et al., 2020). All RNA-seq reads were deposited under the accession number DDBJ: DRA014328.

Quantitative real-time PCR

The cDNA was synthesized from the total RNA, using the ReverTra Ace® qPCR RT Master Mix (TOYOBO, Osaka, Japan). Next, the qPCR was performed in a Roche LightCycler® 480II system with KOD SYBR® qPCR Mix (TOYOBO, Osaka, Japan). The reaction solution contained 5.0 μl of qPCR Mix, 4.0 μl of cDNA and 1 μl of each primer set (10 μM per forward and reverse). The cycling conditions were as follows: 98°C for 2 min, then 45 cycles of 98°C for 10 s, 60°C for 10 s, and 68°C for 30 s. The expression levels of genes were normalized to the level of constitutive *Actin* expression. All primers used in the qPCR are listed in Supplementary Table 4.

[1] http://cell-innovation.nig.ac.jp/index_en.html(Accessed July 17, 2019).

[2] https://rapdb.dna.affrc.go.jp/download/irgsp1.html

[3] https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

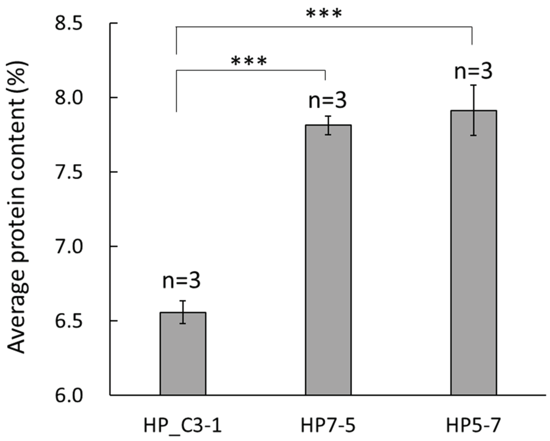


FIGURE 1
Average protein content in the selected EG4 lines. Protein content was evaluated using three independent plants per line. Data shown are the mean±SD of three replicates (n=3). A statistically significant difference between the mean values was inferred from Student's t-test (***p<0.001).

TABLE 1 The identified *mPing* insertion sites in the three lines of rice.

Subject	C3-1	HP5-7	HP7-5
No. of *mPing* insertion sites based on a clustering analysis	581	589	495
No. of *mPing* insertion sites according to the alignment	617	653	538
No. of common insertion sites	552 (95.0%)	572 (97.1%)	478 (96.6%)
No. of *mPing* insertion sites where reads occurred below the threshold (< 50)	18	8	7
Total number of *mPing* insertion sites	**570**	**580**	**485**

Results

Evaluation of protein content in the EG4 rice population

In the primary screening step, protein content was measured using the BCA method in 396 lines of the EG4 population. Based on these results, 25 lines with diverse protein content were selected. Secondary screening was performed using the improved Dumas method, and protein content was analyzed in detail. The protein content of the randomly selected EG4 lines was 6.56 to 6.99%, while that of two lines was higher, at 7.91 ± 0.17% and 7.81 ± 0.06%, respectively (Figure 1). Because these two lines had a significantly higher protein content than randomly selected EG4 lines, both (hereon HP5-7 and HP7-5) were selected for use as high-protein lines. In addition, one line (C3-1) featuring normal protein content (6.56 ± 0.08%) was chosen as a control for the subsequent analyses (Figure 1).

Sequencing and data analysis of *mPing* insertion sites

A total of 16,233,064 paired-end reads of 150 bp were obtained by MiSeq sequencing (min: 533,443, average: 649,323, max: 839,733 reads per line; Supplementary Table 5). After preprocessing, 2,995,488 reads remained overall (Supplementary Table 6). These reads were used for the clustering analysis with the BLAT self-alignment program (Kent, 2002), which suggested 3,268 independent insertion sites in 25 lines (Supplementary Table 6). Next, we determined the genotype (presence or absence) per insertion site and calculated the total number of insertion sites for each line. For the 25 lines, their number of *mPing* insertion sites averaged 573.1, ranging from 495.0 to 655.0 (Supplementary Table 7). The HP5-7 and HP7-5 (high-protein lines) had 589 and 495 *mPing* insertion sites, respectively, while the C3-1 line (control) had 581. These three lines were focused on subsequent analyses.

Given that a highly accurate reference genome sequence is available for rice, we also identified the *mPing* insertion sites in another way, by aligning the MiSeq reads to the Nipponbare reference genome sequence. This yielded 653, 538, and 617 *mPing* insertion sites identified in HP5-7, HP7-5, and C3-1, respectively (Table 1). Comparing the *mPing* insertion sites identified by the clustering-based versus the alignment-based methods in the three lines, 572 (97.1%), 478 (96.6%), and 552 (95.0%) sites were detected in common for HP5-7, HP7-5, and C3-1, respectively (Table 1). Hence, most of the insertion sites (≥95%) can be detected by both analytical methods: one which identified insertion sites by clustering the reads based on the sequence similarity, without reliance on the reference genome sequence, and the other doing so by aligning the reads to the rice reference genome sequence. Further, despite having aligned reads < 50 (see "Materials and methods")—the threshold for determining the presence of absence of insertion—there were some insertion sites where those alignment reads were confirmed. Because these insertion sites were identified by the clustering-based analysis, we considered them actually present. After including these insertion sites, the total number of *mPing* insertion sites per line was estimated to be 580, 485, and 570 in HP5-7, HP7-5, and C3-1, respectively (Table 1).

To verify the *mPing* insertion sites detected above, an experimental validation was performed using several selected insertion sites. For this, PCR primers were designed based on the flanking sequences of *mPing* insertion sites according to our previously described methodology (Monden et al., 2009). For all insertion sites, the PCR bands of expected size were detected (Supplementary Figure 1). Therefore, the analytical methods in this study were considered highly reliable.

Comparing the *mPing* insertion sites identified in HP5-7, HP7-5, and C3-1, 367 of them (> 60%) were common to these three lines (Figure 2). By contrast, 141, 93, and 132 *mPing* insertion sites were only detected in HP5-7, HP7-5, and C3-1, respectively (Figure 2), which indicated they occurred specifically in each line. These line-specific insertion sites may have arisen very recently.

To gain insight into the effects of *mPing* insertions upon genes, we investigated whether *mPing* insertion sites were located inside genes, near genes (within 3 kb from genes), or in intergenic regions. This revealed 14.0–14.6% of the *mPing* insertions positioned inside genes, 52.8–55.7% of them near genes, and 30.3–32.8% of them situated in intergenic regions (Table 2, Supplementary Table 8). The proportion of inside genes, near genes, and intergenic regions in the rice reference genome was 27.5, 31.0, and 41.5%, respectively (Supplementary Table 9); hence, the frequency of *mPing* insertion is significantly lower inside genes and intergenic regions and higher near genes. In addition, we investigated detailed information on *mPing* insertion sites inside or near genes in the rice genome (Supplementary Table 10). Of the *mPing* insertions located near genes, more *mPing* insertions were detected upstream of the gene as compared to downstream of the gene. Of the *mPing* insertions located inside genes, *mPing* is enriched in introns and 3’ UTR regions. It is possible the *mPing* insertion specific to the HP lines (HP5-7 and HP7-5) might affect their phenotype to increase protein content, so we investigated whether HP line-specific *mPing* insertions occurred inside or near genes. Of the HP5-7 specific insertion sites, 21 insertions were inside a gene and 80 insertions were near a gene (Supplementary Table 11); of the HP7-5 specific insertion sites, 17 insertions were inside a gene and 43 insertions were near a gene (Supplementary Table 12). Therefore, a total of 38 genes were detected that contained an *mPing* insertion unique to HP lines, and we posited they may have lost their function due to *mPing* insertion.

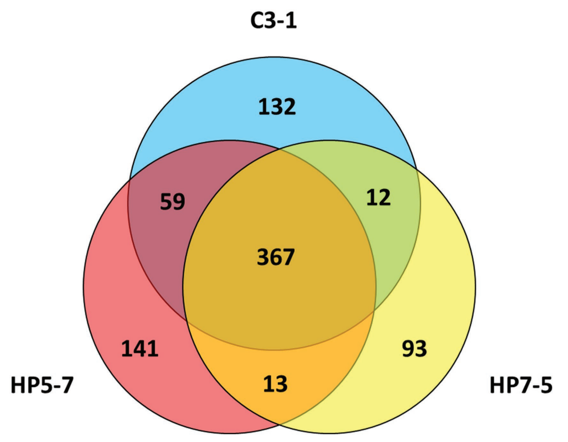


FIGURE 2
Venn diagram showing the number of *mPing* insertion sites and their overlap among the three lines.

Transcriptome analysis

To investigate how *mPing* insertion affected gene expression, and how the latter is to protein content, RNA-seq-based transcriptome analysis was carried out using immature seeds 7 days post-pollination. In the PCA of expression levels of 37,871 genes, the HP5-7, HP7-5, and C3-1 samples clustered separately in the PCA biplot, while their three replicates per line were close to one another (Figure 3). This indicated the expression patterns of plants belonging to the same line were similar, but differed for plants among the lines. Interestingly, HP5-7 and HP7-5 were distributed in the positive direction of the first principal component (PC1) whereas C3-1 was distributed in the negative direction of PC1 (Figure 3). Accordingly, PC1 was inferred as the axis that reflected protein content of rice. Along PC2, however, HP7-5 and HP5-7 were distributed opposite directions (Figure 3); hence, PC2 was designated an axis that explaining features other than protein content. As expected, principal component scores of seed storage protein genes, when plotted, were strongly biased in the positive direction of PC1 (Figure 3). Therefore, we investigated the expression levels of the genes related to protein synthesis in HP5-7, HP7-5, and C3-1 based on the RNA-seq data.

The RNA-seq results showed that most of the genes encoding prolamin, glutelin, and globulin were clearly up-regulated in HP5-7 and HP7-5 compared to C3-1 (Figure 4; Supplementary Figure 2). To confirm those results, three genes, respectively, encoding prolamin, glutelin and globulin were randomly selected and subjected to qRT-PCR. Significant differences in expression were confirmed for all genes, and all validated genes had expression patterns similar to those of the RNA-seq data (Figure 5).

Using eXpress and edgeR software tools, differential expression analysis for HP5-7 versus C3-1 revealed 568 up-regulated and 1,910 down-regulated DEGs in HP5-7 (Supplementary Figure 3). Likewise, there were 550 up-regulated and 2,043 down-regulated DEGs in HP7-5 versus C3-1 (Supplementary Figure 3). These results revealed that approximately 80% of DEGs were down-regulated in the HP lines when compared to C3-1. Hierarchical clustering and heatmap expression analyses were then performed using the 1,278 DEGs commonly detected in the HP lines (Supplementary Figure 4). Evidently, expression patterns of these DEGs clearly differed between the C3-1 and both HP lines, whereas those of HP5-7 and HP7-5 were much more similar (Supplementary Figure 4). Compared with C3-1, there were 106 DEGs whose expression was significantly increased in the HP lines (orange cluster in Supplementary Figure 4) and 128 DEGs whose expression was

TABLE 2 Distribution of *mPing* insertion sites in the rice genome.

Line	Inside gene		Near gene		Intergenic region		Total
	Number	Proportion (%)	Number	Proportion (%)	Number	Proportion (%)	
C3-1	83	14.6	306	53.7	181	31.8	570
HP5-7	81	14.0	323	55.7	176	30.3	580
HP7-5	70	14.4	256	52.8	159	32.8	485

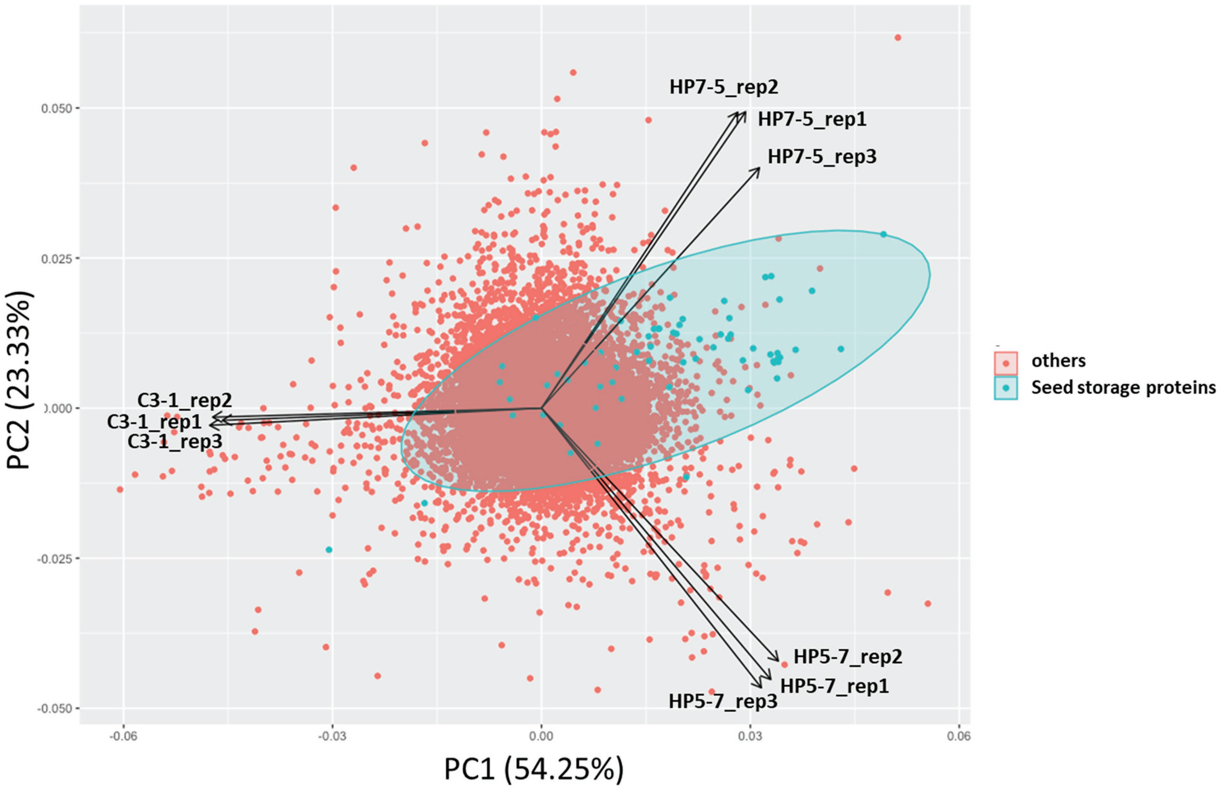


FIGURE 3
Results of principal component analysis (PCA) based on the expression levels of 37,871 genes. Principal component scores of seed storage protein genes are plotted as green dots.

significantly decreased in the HP lines (green cluster in Supplementary Figure 4). We performed a GO enrichment analysis, using ShinyGO, to functionally annotate these DEGs. The top-enriched GO terms for the up-regulated genes in HP lines were "UTP metabolic process," "UTP biosynthetic process," "GTP metabolic process," "GTP biosynthetic process," and "Guanosine-containing compound biosynthetic process" (Supplementary Figure 5). On the other hand, the top-enriched GO terms for the down-regulated genes in HP lines were "Photosynthesis, light harvesting in photosystem I," "Regulation of photosynthesis, dark reduction," "Regulation of reductive pentose-phosphate cycle," "Negative regulation of reductive pentose-phosphate cycle," and "Photosynthesis, light harvesting" (Figure 6). This indicated a tendency for decreased expression levels of photosynthesis-related genes in the HP lines, suggesting that the amount of starch, a major product of photosynthesis, might be reduced in the HP lines.

Differentially expressed genes and *mPing* insertion sites

Finally, to investigate the relationships between gene expression levels of DEGs and *mPing* insertion sites in the HP lines, we extracted those DEGs containing *mPing* insertions or near them (within 3 kb of a gene). Of the DEGs in HP5-7 and HP7-5, an *mPing* insertion was detected near 12 and 10 DEGs, respectively, and one of these DEGs (*Os3g0758551*) was common to both lines (Supplementary Table 13). In addition, for two DEGs (*Os08g0233900* and *Os08g0260400*), the *mPing* insertion inside

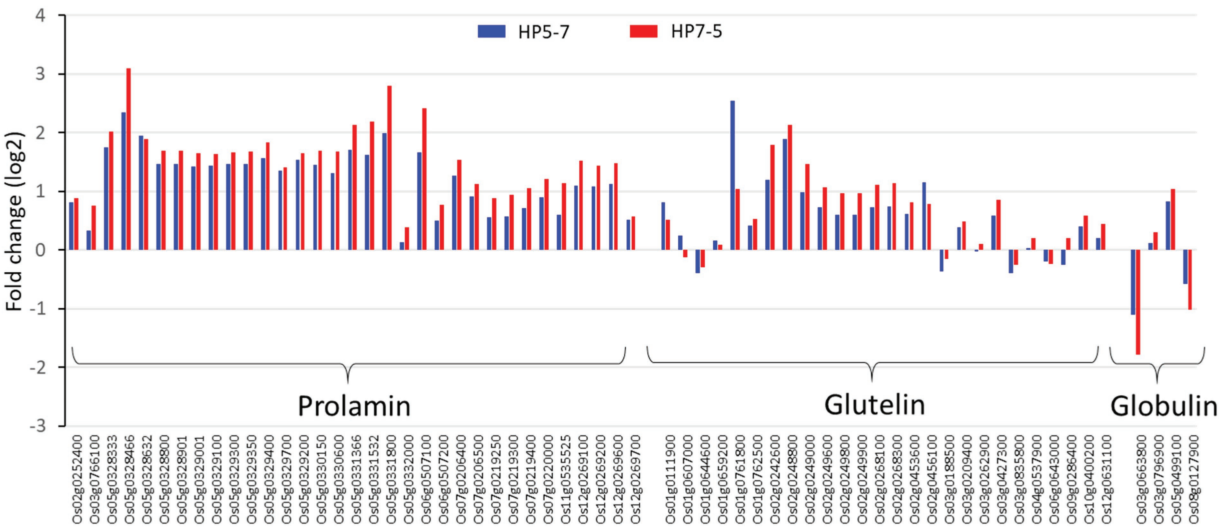


FIGURE 4
Expression levels of the genes encoding glutelin, prolamin, and globulin, based on the RNA-seq results. The log2[fold-change] value was calculated by comparison with the expression levels of C3-1 (the control).

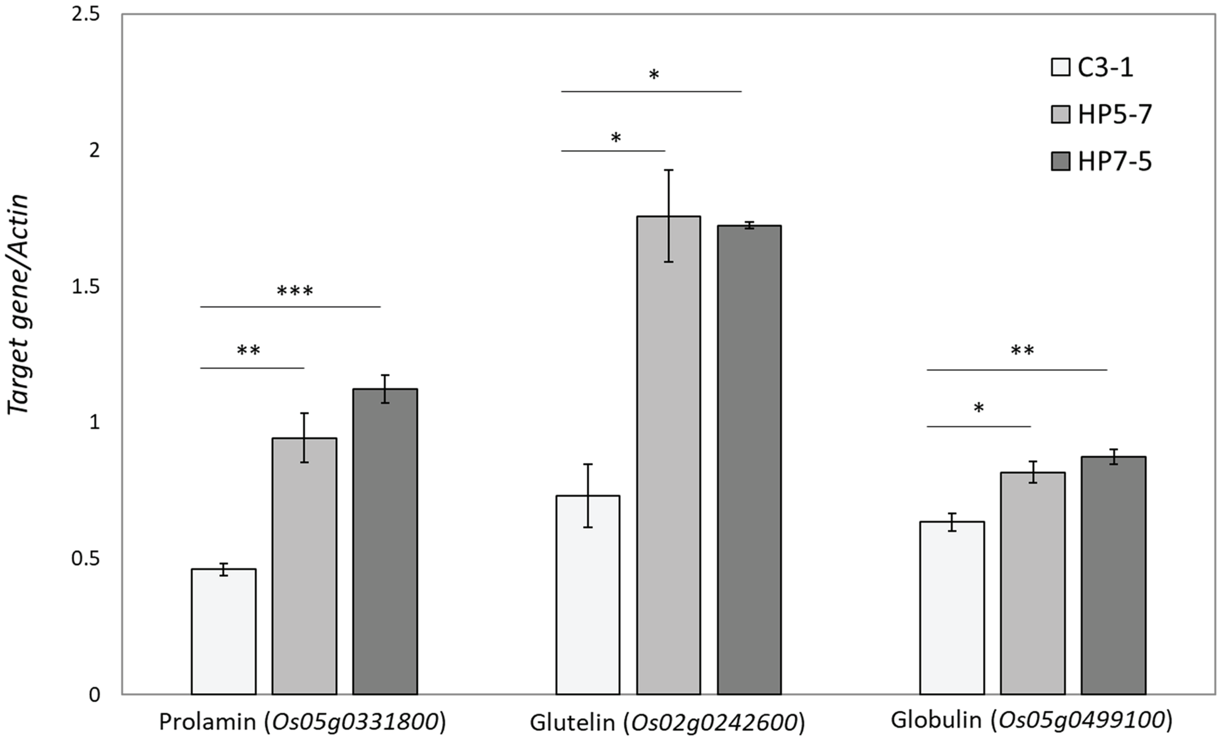


FIGURE 5
Validation of expression patterns of the selected three genes by qRT-PCR. The expression level of each gene was calculated relative to that of the *Actin* gene that served as an internal standard. Data shown are the mean±SD of three replicates. A statistically significant difference between the mean values was inferred from Student's *t*-test (**p*<0.05, ***p*<0.01, and ****p*<0.001).

the gene were detected in HP7-5 (Supplementary Table 13). Against the number of *mPing* insertions near or inside genes (404 and 326 *mPing* insertions in HP5-7 and HP7-5, respectively; Table 2), only 23 DEGs have *mPing* insertions near or inside them. Those results indicated that most of the *mPing* insertions did not affect the expression levels of neighboring genes. Intriguingly, the expression levels of most DEGs with or near an *mPing* insertion were down-regulated, with only two DEGs (*Os10g0532150* and

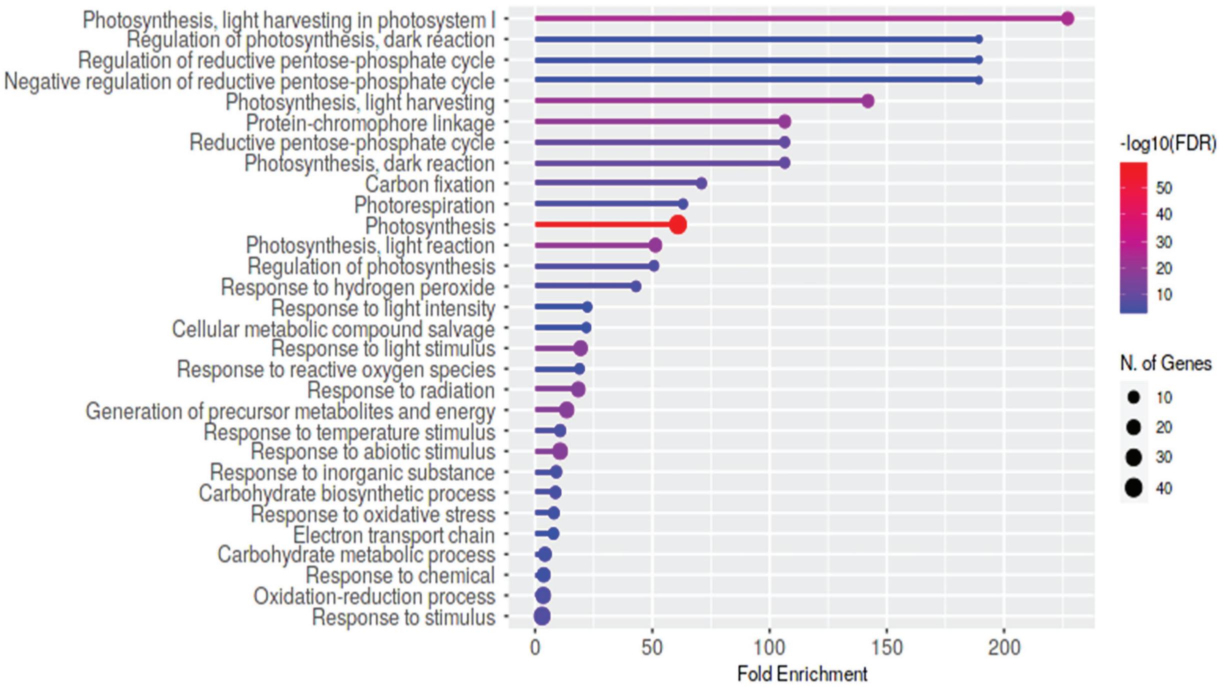


FIGURE 6
The enriched Gene Ontology (GO) terms of the identified 1,278 DEGs commonly detected in the HP lines. The top-ranked GO terms for the down-regulated DEGs in HP lines compared with C3-1.

Os04g0415100) near an *mPing* insertion found up-regulated (Supplementary Table 13). This suggested *mPing* insertions are more likely to reduce the expression levels of neighboring genes when affecting gene expression. The gene annotations of these DEGs were examined carefully; unfortunately, among them no candidate gene governing the high protein content of HP lines was found.

Discussion

In this study, two high protein lines (i.e., HP5-7 and HP7-5) were screened from EG4 rice population, and candidate genes and/or *mPing* insertion sites related to high protein content were explored by genome-wide *mPing* insertion sites and an RNA-seq based transcriptome analysis. *mPing* has high transpositional activity and high copy numbers in EG4 and related rice strains under natural conditions (Naito et al., 2009). Considering that *mPing* tends to be inserted into gene-rich euchromatic regions, it is reasonable to suppose that altered gene expression and/or a gene knockout *via mPing* insertion and associated phenotypic changes are likely to occur in the EG4 population. In screening for high-protein content lines from that population based on two independent approaches (primary screening by the BCA assay and secondary screening by the improved Dumas method), we detected two lines (HP5-7 and HP7-5) characterized by high protein content. Utilizing an NGS-based method, a genome-wide analysis of *mPing* insertion sites in EG4 population was completed. These results uncovered approximately 570 *mPing* insertion sites per line for the 25 EG4 lines having variable protein content. The transcriptome results revealed that most of the genes related to storage protein content—encoding glutelin, prolamin, and globulin—were up-regulated in HP lines compared to the control line. We found a total of 38 genes containing an *mPing* insertion that were restricted to the HP lines, consisting of 21 and 17 genes specific to HP5-7 and HP7-5, respectively (Supplementary Tables 11, 12). Focusing on the DEGs, a total of 23 DEGs were detected to have *mPing* insertions inside or near them (Supplementary Table 13). These genes and/or DEGs may be responsible for causing the high protein content in the two selected HP lines. Further research is needed to find genes and/or mutations governing that high protein content.

Rice is a typical self-fertilizing crop, and its genome has been fixed over generations. Generally, the phenotype of these self-fertilizing plants has rarely changed during cultivation, resulting in a uniform population. TEs are considered a pivotal factor for inducing somatic mutations, nucleotide changes, and phenotypic variation in self-fertilizing plants. In the EG4 population, *mPing* is known to be actively transposing and capable of producing approximately 20 new copies per generation without any particular stress (Naito et al., 2006), which should lead to the generation of new allelic mutations and regulatory networks. This study comprehensively investigated the *mPing* insertion sites of several lines from the EG4 population by using the NGS platform.

From the rice reference genome sequence, we categorized three regions: inside genes, near genes (within 3 kb from a gene) and intergenic regions, whose corresponding proportions of *mPing* insertion sites in the 12 rice chromosomes were 27.5, 31.0, and 41.5%, respectively (Supplementary Table 9). Yet 14.0–14.6%, 52.8–55.7% and 30.3–32.8% of *mPing* insertion sites were, respectively, located inside genes, near genes, and intergenic regions (Table 2; Supplementary Table 8). Our results suggest *mPing* is more apt to insert itself into the region near a gene. These results are consistent with other studies finding *mPing* enriched in euchromatic, gene-rich regions yet infrequent in heterochromatic regions (Naito et al., 2009, 2014).

Previous studies have analyzed *mPing* insertion sites by a variety of methods. Before the advent of NGS, the copy number of *mPing* was estimated using an experimental method called transposon display (Naito et al., 2006; Takagi et al., 2007). Transposon display, a modified method of amplified fragment length polymorphism (AFLP), has been used to generate and display hundreds of genomic fragments that flank specific transposable elements (Casa et al., 2000). To know the sequence information of a flanking region for the insertion site, the amplified products of interest were excised and purified from polyacrylamide gels. After cloning the purified products, these were individually sequenced piecemeal, using the Sanger method (Naito et al., 2006). But the advent of NGS technology now makes it possible to analyze the *mPing* insertion sites comprehensively at once. Naito et al. (2009) was the first to report on the identification of genome-wide *mPing* insertion sites using high-throughput sequencing technology in EG4 rice and related strains. That paper identified *mPing* insertion sites by amplifying the flanking DNA fragments of *mPing* insertion sites by applying vectorette PCR (Arnold and Hodgson, 1991) and pyrosequencing in the Roche 454 platform. Later, Chen et al. (2019) also characterized *mPing* insertion sites in EG4 and related strains, by using high-throughput short-reads sequencing data. In that paper, the authors analyzed the whole-genome sequencing reads obtained from the Illumina platform using a tool they developed, RelocaTE2 (Chen et al., 2017). RelocaTE2 detects the insertion sites of a known TE using resequencing data, by searching junction reads which contain parts of the TE sequence and parts of the unique host genomic sequence (Chen et al., 2017). In Chen et al. (2019), the copy number of *mPing* was estimated to be 437 in EG4. By contrast, our study found an average of 573.1 (min: 495, max: 655) *mPing* insertion sites identified in each line derived from the EG4 population (Supplementary Table 7). Therefore, our methods can detect more *mPing* insertion sites than that previous study. Moreover, we applied and compared two different analytical methods, clustering-based and alignment-based, to identify the *mPing* insertion sites. More than 95% of *mPing* insertion sites were identified by both methods in the tested three lines (Table 1). These identified insertion sites were amplified as expected by PCR and verified experimentally, confirming our analytical methods are highly effective and reliable.

In this study, we aimed to identify the causal genes underpinning the high-protein phenotype in the HP lines (HP5-7 and HP7-5). Unfortunately, we could not detect any candidate genes. Nevertheless, most of the genes encoding glutelin, prolamin, and globulin were clearly up-regulated in both HP lines (Figure 4). Accordingly, it seems the expression of multiple genes may contribute to the greater protein content, rather than one major gene *per se* being responsible for increasing the protein content. In addition, given the reduced expression of many genes related to photosynthesis, as inferred from the DEG analysis (Figure 6), the starch content likely decreased in HP5-7 and HP7-5. Photosynthesis is a process whereby plants convert light energy into chemical energy; the former is used to convert water, carbon dioxide, and minerals into oxygen and energy-rich organic compounds. In most green plants, carbohydrates, especially starch and the sugar sucrose, are the direct, major organic products of photosynthesis. In a previous study, nitrogen fertilization reduced the expression of genes related to starch synthesis and decreased the storage starch content, while increasing the expression of genes related to amino acid biosynthesis and increasing the storage protein content, implying a trade-off between protein and starch synthesis (Midorikawa et al., 2014). Therefore, a decreased starch synthesis in the HP lines may have increased protein content. When we quantified the starch content in rice landraces varying in their protein content, a strong negative correlation was clearly confirmed between the starch and protein content (*unpublished data*). Based on the above, we posit that the expression of those genes involved in photosynthesis may indirectly contribute to the improved protein content of rice. In the near future, we plan to further investigate both HP lines, in terms of their starch and protein synthesis, which should point to the network of causal genes responsible for their high protein content. Still, for unknown reasons, many DEGs were down-regulated in the HP lines compared with C3-1 (Supplementary Figure 3). It is interesting that thousands of DEGs were detected, even though HP lines and C3-1 have the same genetic background derived from the EG4 strain. Of the detected DEGs, 0.48% (12/2,478) and 0.39% (10/2,593) of DEGs have an *mPing* insertion within 3 kb in HP5-7 and HP7-5, respectively, which indicates that the vast majority of DEGs were not affected by an *mPing* insertion. Accordingly, researchers should also consider the possibility that other factors besides *mPing* might cause an increase in the protein content of HP lines of rice.

Conclusion

In this study, we selected two rice lines (HP5-7 and HP7-5) with high protein content from the unique EG4 population in which *mPing* is actively transposing under natural conditions, and analyzed their *mPing* insertion sites and their transcriptome. Given that the *mPing* insertion sites identified by NGS were confirmed experimentally, we consider our analytical methods to be highly effective and reliable. Many of the detected *mPing* insertion sites were positioned near the gene (i.e., within 3 kb), which suggests that *mPing* tends to affect the transcription activity of those genes. Transcriptomics revealed that most genes encoding glutelin, prolamin, and globulin were up-regulated in both HP5-7 and HP7-5 lines. Conversely, many genes had expression levels that were lower in the HP lines than the control C3-1 line, and most of the DEGs near *mPing* insertion sites were also down-regulated. In particular, there tends to be reduced expression of photosynthesis-related genes in the HP lines, which suggests that decreased starch content may contribute to greater protein content. In the future, we plan to identify the causal genes responsible for the high protein content of HP lines, by considering the involvement of genes related to photosynthesis and starch synthesis. Further, we anticipate the high-protein lines detected here could lead to the development of high protein rice cultivars by introducing valuable traits such as high and stable yield, disease resistance, and rich nutrient content.

Data availability statement

The datasets presented in this study can be found in the DDBJ repository, accession numbers DRA014320 and DRA014328.

Author contributions

TY and YM conceived and designed the experiments. YM, HT, and TY conducted the experiments and data analysis to identify the *mPing* insertion sites. RF and TY performed the transcriptome experiments and analyzed that data. EK and KM measured the protein content of rice seeds. SS performed the experiments that validated the *mPing* insertion sites and transcriptome analysis. KY cultivated the experiments' rice plants. YM wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Japanese Society for the Promotion of Science (JSPS) [KAKENHI, grant number 22H02311 (to TY)] and by the Japanese Science and Technology (JST) agency [COI grant number JPMJCE1307 (to TY)].

Acknowledgments

We would like to express our thanks to Tsuyoshi Tanaka (NARO) for providing the sequence information of the rice reference genome, and to Tao Watanabe (Graduate School of Agriculture Kyoto University) for generating the gene expression heatmap. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.969582/full#supplementary-material

References

Arnold, C., and Hodgson, I. J. (1991). Vectorette PCR: a novel approach to genomic walking. *PCR Methods Appl.* 1, 39–42. doi: 10.1101/gr.1.1.39

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackey, S., Bailey, P., et al. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* 24, 1242–1255. doi: 10.1105/tpc.111.095232

Casa, A. M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S., et al. (2000). The MITE family *heartbreaker* (*Hbr*): molecular markers in maize. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10083–10089. doi: 10.1073/pnas.97.18.10083

Chen, L., Lu, L., Benjamin, J., Diaz, S., Hancock, C. N., Jason, E. S., et al. (2019). Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nat. Commun.* 10:641. doi: 10.1038/s41467-019-08451-3

Chen, P., Shen, Z., Ming, L., Li, Y., Dan, W., Lou, G., et al. (2018). Genetic basis of variation in rice seed storage protein (albumin, globulin, prolamin, and glutelin) content revealed by genome-wide association analysis. *Front. Plant Sci.* 9:612. doi: 10.3389/fpls.2018.00612

Chen, J., Wrightsman, T. R., Wessler, S. R., and Stajich, J. E. (2017). RelocaTE2: A high resolution transposable element insertion site mapping tool for population resequencing. *PeerJ.* 5:e2942. doi: 10.7717/peerj.2942

Eckardt, N. A. (2000). Sequencing the rice genome. *Plant Cell* 12, 2011–2017. doi: 10.1105/tpc.12.11.2011

Ellepola, S., Choi, S., Phillips, D., and Ma, C. (2006). Raman spectroscopic study of rice globulin. *J. Cereal Sci.* 43, 85–93. doi: 10.1016/j.jcs.2005.06.006

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405. doi: 10.1038/nrg2337

Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* 3, 329–341. doi: 10.1038/nrg793

Feschotte, C., and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* 41, 331–368. doi: 10.1146/annurev.genet.40.110405.090448

Ge, S. X., Jung, D., and Yao, R. (2020). ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* 36, 2628–2629. doi: 10.1093/bioinformatics/btz931

Hancock, C. N., Zhang, F., and Wessler, S. R. (2010). Transposition of the *tourist*-MITE *mPing* in yeast: an assay that retains key features of catalysis by the class 2 *PIF/harbinger* superfamily. *Mob. DNA* 1:5. doi: 10.1186/1759-8753-1-5

Henchion, M., Hayes, M., Mullen, A. M., Fenelon, M., and Tiwari, B. (2017). Future protein supply and demand: strategies and factors influencing a sustainable equilibrium. *Foods* 6:6. doi: 10.3390/foods6070053

Higashi, T., Kushibuchi, K., and Ito, R. (1974). Studies on breeding for high protein rice: I. Protein content of different rice varieties and their relations with some agronomic traits including yield. *Japan. J. Breed.* 24, 88–96.

Hirata, C., Waki, T., Shimomura, K., Wada, T., Tanaka, S., Ikegami, H., et al. (2020). DNA markers based on retrotransposon insertion polymorphisms can detect short DNA fragments for strawberry cultivar identification. *Breed. Sci.* 70, 231–240. doi: 10.1270/jsbbs.19116

Jackson, S. A. (2016). Rice: the first crop genome. *Rice* 9:14. doi: 10.1186/s12284-016-0087-4

Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S. R., McCouch, S. R., et al. (2003). An active DNA transposon family in rice. *Nature* 421, 163–167. doi: 10.1038/nature01214

Jiang, C., Cheng, Z., Zhang, C., Yu, T., Zhong, Q., Shen, J. Q., et al. (2014). Proteomic analysis of seed storage proteins in wild rice species of the *Oryza* genus. *Proteome Sci.* 12:51. doi: 10.1186/s12953-014-0051-4

Jung, S., Rickert, D. A., Deak, N. A., Aldin, E. D., Recknor, J., Johnson, L. A., et al. (2003). Comparison of kjeldahl and dumas methods for determining protein contents of soybean products. *J. Amer. Oil Chem. Soc.* 80:1169. doi: 10.1007/s11746-003-0837-3

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4. doi: 10.1186/1939-8433-6-4

Kawakatsu, T., Yamamoto, M. P., Hirose, S. Y., and Takaiwa, F. (2008). Characterization of a new rice glutelin gene *GluD-1* expressed in the starchy endosperm. *J. Exp. Bot.* 59, 4233–4245. doi: 10.1093/jxb/ern265

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Kikuchi, K., Terauchi, K., Wada, M., and Hirano, H. Y. (2003). The plant MITE *mPing* is mobilized in anther culture. *Nature* 421, 167–170. doi: 10.1038/nature01218

Kim, D., Paggi, J. M., Park, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genoty*Ping* with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Kubota, M. (2016). Novel physiological functions of rice protein. *J. Jpn. Soc. Nutr. Food Sci.* 69, 283–288. doi: 10.4327/jsnfs.69.283

Kumar, A., and Bennetzen, J. L. (1999). Plant retrotransposons. *Annu. Rev. Genet.* 33, 479–532. doi: 10.1146/annurev.genet.33.1.479

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

Lin, S. K., Chang, M. C., Tsai, Y. G., and Lur, H. S. (2005). Proteomic analysis of the expression of proteins related to rice quality during caryopsis development and the effect of high temperature on expression. *Proteomics* 5, 2140–2156. doi: 10.1002/pmic.200401105

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10. doi: 10.14806/ej.17.1.200

Martin, M., and Fitzgerald, M. A. (2002). Proteins in rice grains influence cooking properties. *J. Cereal Sci.* 36, 285–294. doi: 10.1006/jcrs.2001.0465

Midorikawa, K., Kuroda, M., Terauchi, K., Hoshi, M., Ikenaga, S., Ishimaru, Y., et al. (2014). Additional nitrogen fertilization at heading time of rice down-regulates cellulose synthesis in seed endosperm. *PLoS One* 9:e98738. doi: 10.1371/journal.pone.0098738

Monden, Y., Hara, T., Okada, Y., Jahana, O., Kobayashi, A., Tabuchi, H., et al. (2015). Construction of a linkage map based on retrotransposon insertion polymorphisms in sweet potato via high-throughput sequencing. *Breed. Sci.* 65, 145–153. doi: 10.1270/jsbbs.65.145

Monden, Y., Naito, K., Okumoto, Y., Saito, H., Oki, N., Tsukiyama, T., et al. (2009). High potential of a transposon *mPing* as a marker system in *japonica* × *japonica* cross in rice. *DNA Res.* 16, 131–140. doi: 10.1093/dnares/dsp004

Monden, Y., Yamamoto, A., Shindo, A., and Tahara, M. (2014). Efficient DNA fingerprinting based on the targeted sequencing of active retrotransposon insertion sites using a bench-top high-throughput sequencing platform. *DNA Res.* 21, 491–498. doi: 10.1093/dnares/dsu015

Naito, K., Cho, E., Yang, G., Campbell, M. A., Yano, K., Okumoto, Y., et al. (2006). Dramatic amplification of a rice transposable element during recent domestication. *Proc. Natl. Acad. Sci. U. S. A.* 103, 17620–17625. doi: 10.1073/pnas.0605421103

Naito, K., Monden, Y., Yasuda, K., Saito, H., and Okumoto, Y. (2014). mPing: the bursting transposon. *Breed. Sci.* 64, 109–114. doi: 10.1270/jsbbs.64.109

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., et al. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461, 1130–1134. doi: 10.1038/nature08479

Nakazaki, T., Okumoto, Y., Horibata, A., Yamahira, S., Teraishi, M., Nishida, H., et al. (2013). Mobilization of a transposon in the rice genome. *Nature* 421, 170–172. doi: 10.1038/nature01219

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotech.* 33, 290–295. doi: 10.1038/nbt.3122

Sakai, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013). Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54:e6. doi: 10.1093/pcp/pcs183

Sasai, R., Tabuchi, H., Shirasawa, K., Kishimoto, K., Sato, S., Okada, Y., et al. (2019). Development of molecular markers associated with resistance to *Meloidogyne incognita* by performing quantitative trait locus analysis and genome-wide association study in sweet potato. *DNA Res.* 26, 399–409. doi: 10.1093/dnares/dsz018

Shewry, P. R. (2007). Improving the protein content and composition of cereal grain. *J. Cereal Sci.* 46, 239–250. doi: 10.1016/j.jcs.2007.06.006

Takagi, K., Ishikawa, N., Maekawa, M., Tsugane, K., and Iida, S. (2007). Transposon display for active DNA transposons in rice. *Genes Genet. Syst.* 82, 109–122. doi: 10.1266/ggs.82.109

Tang, Y., Horikoshi, M., and Li, W. (2016). Ggfortify: unified Interface to visualize statistical result of popular R packages. *R J.* 8, 474–485. doi: 10.32614/RJ-2016-060

Turucotte, K., Srinivsan, S., and Bureau, T. (2001). Survey of transposable elements from rice genomic sequences. *Plant J.* 25, 169–179. doi: 10.1046/j.1365-313x.2001.00945.x

Waskom, M. L. (2021). Seaborn: statistical data visualization. *J. Open Source Soft.* 6:3021. doi: 10.21105/joss.03021

Wessler, S. R. (2006). Transposable elements and the evolution of eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.* 103, 17600–17601. doi: 10.1073/pnas.0607612103

Yamagata, H., Sugimoto, T., Tanaka, K., and Kasai, Z. (1982). Biosynthesis of storage proteins in develo*Ping* rice seeds. *Plant Physiol.* 70, 1094–1100. doi: 10.1104/pp.70.4.1094

Yang, G., Zhang, F., Hancock, C. N., and Wessler, S. R. (2007). Transposition of the rice miniature inverted repeat transposable element *mPing* in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 10962–10967. doi: 10.1073/pnas.0702080104

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 01 December 2022
DOI 10.3389/fpls.2022.1080993

OPEN ACCESS

EDITED BY
Ruslan Kalendar,
University of Helsinki, Finland

REVIEWED BY
Tony Heitkam,
Technical University
Dresden, Germany
Kenji K. Kojima,
Genetic Information Research
Institute, United States
David Roquis,
Technical University of
Munich, Germany

*CORRESPONDENCE
Miroslav Baránek
miroslav.baranek@mendelu.cz

SPECIALTY SECTION
This article was submitted to
Technical Advances in Plant Science,
a section of the journal
Frontiers in Plant Science

RECEIVED 26 October 2022
ACCEPTED 17 November 2022
PUBLISHED 01 December 2022

CITATION
Bajus M, Macko-Podgórni A,
Grzebelus D and Baránek M (2022) A
review of strategies used to identify
transposition events in plant genomes.
Front. Plant Sci. 13:1080993.
doi: 10.3389/fpls.2022.1080993

A review of strategies used to identify transposition events in plant genomes

Marko Bajus[1], Alicja Macko-Podgórni[2], Dariusz Grzebelus[2] and Miroslav Baránek[1]*

[1]Mendeleum—Institute of Genetics, Faculty of Horticulture, Mendel University in Brno, Lednice, Czechia, [2]Department of Plant Biology and Biotechnology, Faculty of Biotechnology and Horticulture, University of Agriculture in Krakow, Kraków, Poland

Transposable elements (TEs) were initially considered redundant and dubbed 'junk DNA'. However, more recently they were recognized as an essential element of genome plasticity. In nature, they frequently become active upon exposition of the host to stress conditions. Even though most transposition events are neutral or even deleterious, occasionally they may happen to be beneficial, resulting in genetic novelty providing better fitness to the host. Hence, TE mobilization may promote adaptability and, in the long run, act as a significant evolutionary force. There are many examples of TE insertions resulting in increased tolerance to stresses or in novel features of crops which are appealing to the consumer. Possibly, TE-driven *de novo* variability could be utilized for crop improvement. However, in order to systematically study the mechanisms of TE/host interactions, it is necessary to have suitable tools to globally monitor any ongoing TE mobilization. With the development of novel potent technologies, new high-throughput strategies for studying TE dynamics are emerging. Here, we present currently available methods applied to monitor the activity of TEs in plants. We divide them on the basis of their operational principles, the position of target molecules in the process of transposition and their ability to capture real cases of actively transposing elements. Their possible theoretical and practical drawbacks are also discussed. Finally, conceivable strategies and combinations of methods resulting in an improved performance are proposed.

KEYWORDS

transposable elements, transposon mobilization, course of transposition, detection methods, eccDNA, bioinformatics tools

Introduction

Transposable elements (TEs) were found and described in the early 1950s by Barbara McClintock in maize, as entities causing chromosome breakage, with breaking points capable of changing their chromosomal positions (Mc Clintock, 1950). The importance of her observation has eventually been recognized as fundamental and finally, more than

30 years after publishing her seminal paper, McClintock was awarded the Nobel prize (Ravindran, 2012).

TEs are abundant structural genome components inhabiting genomes throughout the course of life evolution (Chuong et al., 2017). Initially, TEs were considered unnecessary or even harmful components of the genome (Sotero-Caio et al., 2017). At present, it is commonly accepted that their interactions with the host genome are far more complex and still not fully understood. In plants, TEs are important drivers of genome evolution, propelling phenotypic variability in the course of crop domestication and improvement. Their representation in plant genomes varies, ranging from approximately 20% in small genomes, such as Arabidopsis to more than 80% in maize (Kim, 2017).

TEs are divided into two classes, according to their mechanism of transposition: Class I (retrotransposons) and Class II (DNA transposons). Retrotransposons use an RNA intermediate to be copied and subsequently inserted as a novel copy at a new position in the genome, which results in an increase of their copy numbers (Feschotte et al., 2002). Retrotransposons are further subdivided into those harboring long terminal repeats (long terminal repeat retrotransposons, LTR-RTs) and non-LTR retrotransposons, including Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs). LTR-RTs are predominant in the TE landscape of plant genomes (Satheesh et al., 2021). In contrast, most DNA transposons physically excise and reinsert (a 'cut and paste' mechanism), while those classified as Helitrons utilize a 'rolling circle' mechanism for their transposition. Thus, transposition of Class II TEs does not involve any RNA intermediate. DNA transposons are widespread and active across many bacterial, archaeal and eukaryotic species, while their activity in mammals is low (Rodriguez-Terrones and Torres-Padilla, 2018). The distribution of TEs in plant genomes has been reviewed in more detail by Sahebi et al. (2018).

Most successful TE mobilization events are neutral or even deleterious to the host. They can cause changes in the pattern of gene expression and alter gene function by up- or down-regulating adjacent genes following insertion into promoter regions, introns, exons or downstream regions (Makarevitch et al., 2015; Deneweth et al., 2022). Also, they may become a source of small interfering RNAs (siRNAs) (Piriyapongsa and Jordan, 2008; Gill et al., 2021). In order to protect integrity of the host genome, TEs are silenced and the state is epigenetically heritable (Fultz et al., 2015). In general, *de novo* silencing of active TE involves DNA methylation and repressive modifications of histones. These epigenetic marks are maintained across subsequent mitotic divisions and transmitted from generation to generation. Importantly, precise mechanisms resulting in TE inactivation depend on the location of a TE copy in the genomic context (Sigman and Slotkin, 2016)

In order to recognize TEs showing ongoing activity, it is necessary to use tools targeting one of the molecules produced in the course of mobilization, i.e. RNA transcripts, extrachromosomal linear DNA (eclDNA), extrachromosomal circular DNA (eccDNA), small RNA or TE-encoded proteins (Figure 1). It is also important to monitor whether mobilized copies are competent to successfully reintegrate with the host genome to produce novel insertion sites.

The approach used by B. McClintock can be viewed as the first method of monitoring TE activity, as she observed that Ac as an activator autonomous TE mobilized non-autonomous Ds elements resulting in chromatid breakage. Fortunately, we have come a long way since then, and new possibilities and approaches are constantly emerging. The subject of the review is to summarize methods used for the analysis of TE activity and to discuss their advantages and specific applications. Special attention is paid to the LTR-RTs, which are considered the most abundant TEs in plant genomes (Deniz et al., 2019). The described methods are divided on the basis of their operational principles, the position of target molecules in the process of transposition and their ability to capture real cases of actively transposing elements. Their possible theoretical and practical drawbacks are also discussed. Finally, conceivable strategies and combinations of methods resulting in an improved performance are proposed.

Detection of TE-derived transcripts

As LTR-RTs require the formation of an RNA intermediate, it is the first target usable for the evaluation of their activity. Generally, LTR-RT-derived RNAs can be identified using tools similar to those used for monitoring gene expression, i.e. techniques based on nucleic acid hybridization (northern blotting, microarrays), PCR (RT-qPCR), or transcriptome sequencing (RNA-seq).

Historically, northern blotting was used as the first method of choice (Manninen and Schulman, 1993; Meyer et al., 1994; Pozueta-Romero et al., 1995). With the development of new technologies, its significance gradually declined due to the complexity of protocols and necessity to ensure high amounts of input RNA. Subsequently, methods based on RT-qPCR started to be utilized to monitor TE activity in plants (Marcon et al., 2015; Paz et al., 2015; Jiang et al., 2016; Voronova, 2019; Usai et al., 2020). An important limitation of RT-qPCR is that it targets individual copies or TE families grouping very similar copies and specificity is provided by primers used for qPCR. Hence, the assay requires prior knowledge about propensity of the studied copy to be mobilized. On the other hand, it may be problematic to design specific primers to investigate TEs from different families (Morillon et al., 2002). Another limitation is the fact that the target sequence may include nucleotide substitutions and/or indels in transcripts produced from

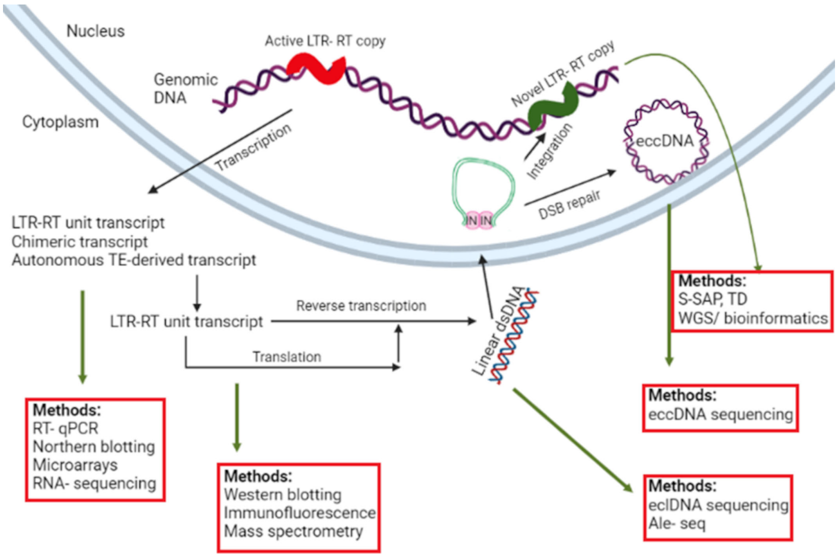


FIGURE 1
An overview of target molecules generated in the course of LTR-RT transposition and methods suitable for their detection. The meaning of individual abbreviations is as follows: LTR-RT, Long Terminal Repeat Retrotransposon; RT-qPCR, Reverse Transcription – quantitative PCR; IN, integrase; DSB, Double Strand Break; eclDNA, Extrachromosomal Linear DNA; eccDNA, extrachromosomal circular DNA; S-SAP, Sequence-Specific Amplification Polymorphism; TD, transposon display; WGS, Whole Genome Sequencing; ALE-Seq, Amplification of LTR of eclDNAs followed by Sequencing.

different copies. In such case, northern blotting seems to be a good complementary method, as it may reveal the size distribution of TE-derived transcripts, including full length TEs (Böhrer et al., 2020).

A global analysis of TE-derived transcripts can be produced with microarrays (Picault et al., 2009; Rocheta et al., 2016). Comprehensive information about the whole spectrum of actively transcribed TEs can also be captured by RNA-seq based on massive parallel DNA sequencing technologies (Gürkök, 2017; Oberlin et al., 2017; Qiu and Ungerer, 2018; Vangelisti et al., 2019; Jiménez-Ruiz et al., 2020; Kirov et al., 2020). RNA-seq data have been utilized and interpreted differently in reports aiming at the description of global activity of TEs. While some reports simply presented a spectrum of TEs captured in RNA-seq reads (Gürkök, 2017; Jiménez-Ruiz et al., 2020), in other reports, especially those concerning plant species for which high quality reference genomes were available, TE-derived transcripts were mapped to the reference genome assembly (Li et al., 2010; Hollister et al., 2011; Valdebenito-Maturana and Riadi, 2018). However, owing to the fact that some TE families comprise numerous copies and the evolutionary relationships among TE families can be complex, interpretation of the RNA-seq data remains challenging. Different strategies have been implemented, solely or in combination, to confirm TE expression from RNA-seq data, i.e. mapping TE-derived reads to a reference genome, a TE pseudogenome and a model transcriptome (Lanciano and Cristofari, 2020). Precision of the mapping process can be significantly improved by using longer reads provided by PacBio or Oxford Nanopore technologies (Sexton and Han, 2019). When using them it is much easier to predict if the sequenced TE-derived transcript has a potential to complete its full life cycle, or vice versa, whether it does not contain signs of inactive forms such as chimeric transcripts. Available bioinformatic tools and techniques for TE mapping to reference genomes were recently reviewed by O'Neill et al. (2020).

In general, with respect to all TE-derived transcript targeting techniques, it is necessary to be aware that there are issues that can impact clarity of results when the primary interest is to investigate only actively transposing TEs. It is because a significant share of TEs is transcribed by PolII and processed into 21~24 nt siRNA, involved in epigenetic silencing of TEs (Tang et al., 2022). Moreover, stress-dependent genome demethylation (Pandey et al., 2017; Liang et al., 2019) may result in increased expression of TEs. Also, transcripts containing sequences derived from TEs may also include chimeric transcripts containing both TE and genic fragments, e.g. those resulting from the initiation of transcription from a TE promoter or from exonization of intronic TE insertions. Such transcripts are obviously not an indication of ongoing transposition activity, but still they can be abundant in RNA samples. Besides, active post-transcriptional suppression mechanisms by TE-derived sequences was also described

(Fultz et al., 2015). The above-described drawbacks and the fact that transcription is only an initial step in the process of transposition suggest that monitoring TE-derived transcripts is not an optimal strategy aiming at the identification of TEs capable of completing new insertion. There is a serious risk of misinterpretations and incorrect conclusions deeply discussed also by Deininger et al. (2017). However, expression-based assays can be used to support results concerning TE mobility produced by using other approaches.

Detection of TE-encoded proteins

One of the possible manifestations of TE mobilization is translation of TE-encoded proteins constituting an essential transposition machinery. Thus, theoretically such proteins can also be used for monitoring an ongoing process of TE mobilization. It is necessary to emphasize that some types of TEs, e.g. SINEs or MITEs, referred to as non-autonomous, do not encode any proteins and utilize transposition machinery provided by their autonomous counterparts, LINEs and related DNA transposons, respectively. Historically, proteomic studies related to TE activity were based on western blotting. Western blot is an analytical technique used to detect a specific protein in a mixture of all proteins extracted from a tissue sample. Thus, TE mobilization-related experiments focus on a limited group of TE-derived proteins, such as transposases (Torres et al., 2013). The advantage of western blotting is that it can reveal events where internal mutations within coding regions of a TE prevent protein translation and subsequently hamper TE transposition. Such cases remained unrevealed by the analysis of TE-derived transcripts. Drawbacks of western blotting include limited availability and sensitivity of reagents, potential nonspecific activity of antibodies between related families of TEs, and necessity to produce large quantities of the starting material.

One of the most promising approaches for proteomic analysis is the application of methods based on mass spectrometry (MS) that may provide broad-spectrum results. Generally, MS is used to determine the mass of particles in order to determine the elemental composition and chemical structure of molecules, including complex substances, such as peptides. In the case of peptide analysis, combination of liquid chromatography (LC) with MS (LC-MS or LC-MS/MS), allowing for broad-spectrum analyses even down to the level of their amino acid sequences, are the most frequently used techniques. Obtained sequences can subsequently be evaluated with respect to the presence and the type of TE-derived proteins in analysed samples (Maringer et al., 2017). For example, Vuong et al. (2019) used MS to identify proteins of human TEs belonging to the L1 family of LINEs. In turn, Wang et al. (2008) used LC-MS/MS to study proteins activated by the moss *Physcomitrella patens* upon high salinity stress, revealing TE-derived proteins as being differentially expressed. Matrix Assisted Laser Desorption Ionization - Time of Flight (MALDI-TOF-TOF) combined with MS was also used to reveal proteomic background of sporadic flowering in bamboo species, suggesting a direct relationship of TE activation and the induction of flowering (Louis et al., 2015).

With respect to the fact that proteins are synthetized in initial stages of the TE transposition process, it is necessary to realize that proteomics, while allowing for detection of actively transposing TE, also bears some limitations. Feschotte and Pritham (2007) reported that ancient TEs were less likely to be actively transposing, however they might still express proteins, especially when they originated from domesticated TEs, and at present those proteins fulfill essential host cell functions. Altogether, proteomic techniques may provide unique insights to investigations on the TE activity, e.g. involvement of TE-derived proteins in the assembly of protein complexes. However, the employment of complementary strategies is needed to obtain a comprehensive landscape of actively transposing TEs. Proteomics Informed by Transcriptomics (PIT) may be one such prospective strategy. In this method, proteomic MS/MS spectra are searched against open reading frames derived from assembled RNA-Seq transcripts. This approach can reveal previously unknown translated genomic elements or can also identify hotspots of incomplete genome annotation. PIT was initially generated in general principle, however, it can be easily tuned to investigate TE ongoing activity (Davidson et al., 2017; Maringer et al., 2017).

Detection of extrachromosomal linear DNA

The formation of extrachromosomal linear DNA (eclDNA) molecules is inherent to the process of LTR-RT mobilization. LTR-RTs contain two ORFs, Gag encoding a coat protein, and Pol encoding a polyprotein comprising four domains, i.e. reverse-transcriptase (RT), RNase H (RH), aspartic protease (AP) and integrase (INT). The life cycle of LTR-RTs begins with transcription of an active LTR-RT copy by a host-encoded RNA polymerase II, followed by synthesis of LTR-RT-encoded proteins, formation of virus-like particles (VLPs) encapsulating the RNA template, and its reverse transcription resulting in the formation of eclDNA. Subsequently, eclDNA enters the nucleus and integrates with the host genome (Havecker et al., 2004). Thus, the detection of eclDNAs seems to be an exquisite approach to mine for actively transposing LTR-RTs (Grandbastien, 2015), as they represent the final intermediates in LTR-RT retrotransposition (Figure 1). However, eclDNA can occur in cells also as a result of other events, such as cell lysis-originating eclDNA, as cells are constantly being lysed, or extrachromosomal linear microDNA interspersed with microRNAs (Sun et al., 2019). All these eclDNA sources may contain LTR-RT sequences, but only in the case of linear products resulting from the transposition process, the identified fragment is expected to be terminated with LTR sequences, without additional fragments of genomic DNA sequence. Thus, a stage allowing selection of LTR-RTs should be included. A strategy based on PCR amplification utilizing a primer annealing to the tRNA primer binding site (PBS) could be used. It was originally applied to generate PCR-based iPBS molecular markers (Kalendar et al., 2010), while later it became the basis of SIRT (Sequence-Independent Retrotransposon Trapping) – the first method using LTR-RT-derived eclDNAs as targets (Griffiths et al., 2018). It took advantage from the fact that eclDNA ends are blunt-ended and competent for ligation of synthetic adaptors. Subsequently, using PCR primers complementary to the adaptor and to the PBS, a segment comprising the 5´LTR was amplified. When compiling complementary PBS primers, they used the fact that actively transposing LTR-RTs described in plants use predominantly as the initiator methionine tRNA (Met-iCAT) (Wicker et al., 2007; Kalendar et al., 2010). Thus, PBS sequences consist of 12 nucleotides complementary to the terminal nucleotides of the MET-iCAT tRNA. To ensure specific PCR amplification, the PBS-specific primers were therefore extended using the knowledge that two terminal nucleotides of 5′ LTR mostly end in cytidine and adenosine (Griffiths et al., 2018). The disadvantage of the SIRT method is that it utilizes Sanger sequencing and that PBS-anchored primers are specific to particular LTR-RTs, which limits its usefulness for a global analysis of all LTR-RT families. It also turned out that the concept cannot be applied to large and TE-rich genomes,

To eliminate these disadvantages, the ALE-Seq (amplification of LTR of eclDNAs followed by sequencing) approach was developed (Cho et al., 2019). In comparison to SIRT, the ALE-Seq protocol utilizes more versatile primers complementary to PBS (or their combinations), high throughput sequencing, and is more elaborate as it includes adapter ligation, transcription and reverse transcription targeted to PBS domains. On the other hand, the ALE-Seq protocol is markedly more selective and efficient than SIRT, which relies on the single PCR amplification (Cho et al., 2019). The method is relatively recent, its applicability has been proved by the identification of actively transposing LTR-RTs in rice and tomato. On the basis of subsequent clustering of sequenced reads some retroelements were recognized as newly identified families for the respective genome. To summarize, ALE-Seq has potential for future use allowing reference-free annotation of new, active retroelements, what is especially important in plant species for which no reference genome assemblies are available (Satheesh et al., 2021).

Detection of extrachromosomal circular DNA

Some LTR-RT-derived eclDNA molecules were shown to be circularized. As integrase (IN) molecules are attached to LTRs of eclDNAs, their homodimerization causes the formation of a pseudocircular but unclosed structures. Following their recognition as double strand breaks by DNA repair machineries in the nucleus, they are ligated resulting in closed extrachromosomal circular DNA (eccDNA) molecules (Figure 1). As such, they do not directly participate in the process of transposition and can be seen as mobilization by-products, however, their presence provides information about actively transposing LTR-RTs (Lanciano et al., 2017).

It should be stressed LTR-RT transposition is not a sole source of eccDNAs; they can also occur as a result of other cellular processes. They are common in eukaryotes and can be very heterogenic in number, length, origin, and role as reviewed by Cao et al. (2021).

The first methods of eccDNA detection, i.e. inverse PCR amplification of LTR-LTR junctions and electron microscopy, suggested that some circles originated from TEs, mostly LTR-RTs (Hirochika and Otsuki, 1995) and Mutator-like class II elements (Sundaresan and Freeling, 1987). Advances in sequencing techniques contributed to the development of efficient eccDNA detection methods along with the bioinformatics tools for analysis of such data.

The first high-throughput method of sequencing eccDNA, Circle-Seq, was developed for yeast and consisted of alkaline-based extraction of circular DNA, followed by digestion of linear DNA, eccDNA amplification using φ29 DNA polymerase and sequencing on the Illumina platform using SE mode (Møller et al., 2015). Soon after, based on similar assumptions, a standardized Mobilome-seq protocol of extraction and Illumina SE sequencing of eccDNA from plant tissues was established (Lanciano et al., 2017). Another approach, CIDER-Seq (Circular DNA Enrichment sequencing) method, originally developed for analysis of plants infected with viruses, utilizes electrophoresis-based size-selection as the first step of sample preparation, followed by random amplification of circular DNA with φ29 DNA polymerase, repair by DNA polymerase I and sequencing using Single Molecule Real Time sequencing (Pacific Biosciences) (Mehta et al., 2019).

The production of large amounts of sequencing data raises the need for simultaneous development of analytical tools. Circle-Map (Prada-Luengo et al., 2019) and Circle_finder (Kumar et al., 2020) were developed for identification of human tumor related eccDNA sequenced using short-reads technology. The downside to these tools is that they both require a reference genome as an input file and they were not tested on plant data. Short reads can be also analysed using ECCsplorer (Mann et al., 2022), a tool for mapping reads to the reference genome, identifying genomic origin of eccDNAs on the basis of read distribution, coverage, discordant mapping, and split reads, but also enabling reference-free clustering of reads. This helps to identify and annotate LTR-RTs enriched in eccDNA libraries. eccDNA analysis from long reads is possible using the CIDER-seq2 (Mehta et al., 2020). Although the method was developed for identification and characterization of plant virus genomes, and includes the 'annotate' module that is restricted to viruses annotation, part of the pipeline that outputs eccDNA candidates and their genomic localization can be used for the identification of LTR-RTs. Other long-reads based tools, such as CReCIL (Wanchai et al., 2022) allow not only efficient identification of circular DNA but also annotation and Circos-based visualization of assembled circles, but its performance was tested only on long-reads from mammals eccDNA sequencing. Another tool, ecc_finder (Zhang et al., 2021) is based on a pipeline applied for the analysis of Mobilome-seq data originated from plant tissues (Lanciano et al., 2017). The pipeline allows analysis of both short and long reads and can be run in the reference genome and reference-free modes.

The eccDNA identification was reported to be useful for monitoring mobilization of previously known actively transposing TEs in Arabidopsis, rice and tomato (Lanciano et al., 2017; Benoit et al., 2019; Lanciano et al., 2021; Roquis et al., 2021; Wang et al., 2021; Zhang et al., 2021; Mann et al., 2022) and *de novo* identification of mobilized LTR-RTs, as shown for potato (Esposito et al., 2019), poplar (Sow et al., 2021) and carrot (Kwolek et al., 2022).

Mapping eclDNA or eccDNA sequencing reads to the reference genome may provide a clue as to what is the TE copy that has been undergoing mobilization. Ideally, a reference genome highly related to the individual used for eclDNA or eccDNA should be used. However, the typical properties of TEs, such as their highly repetitive character and the fact that TE families can be highly interrelated within a given species may complicate conclusions driven from such analyses.

Identification of novel insertion sites produced by actively transposing elements

The life cycle of a TE is completed upon its insertion into a new position in the host genome (Figure 1). Such *de novo* insertions are thus present in the progeny while they are absent in the ancestral plants. In earlier studies, these uncommon events were recognized only when they resulted in changed phenotypes. Obviously, these events represent a very small proportion of the total number of successful transpositions resulting in the integration occurring in genic regions.

Historically, the principles of positional (genetic map-based) cloning were used to identify insertional polymorphisms in the genome. However, mapping with high resolution requires numerous mapping populations and many genetic markers, thus it is costly and time consuming. It is therefore not suitable for mapping newly transposed TEs, although one can find some examples here as well (Bortiri et al., 2006). Identification of TE insertion sites and resulting transposon insertion polymorphisms (TIPs) can be also performed using marker systems derived from conservative sequences specific to certain TEs (Kalendar and Schulman, 2006) or by a modification of the amplified fragment length polymorphism (AFLP) protocol (Vos et al., 1995). It is based on comparing the distribution of copies of a particular TE family in a collection of closely related accessions and works especially well for TE families with a number of copies highly uniform in their sequence, which is a proxy for recent or ongoing transposition. Two AFLP modifications aiming at the identification of TIPs have been developed, i.e. sequence-specific amplification polymorphism (S-SAP), used for the identification of LTR-RT insertions, where the final amplification is performed with a retrotransposon-specific and a *Mse*I-adaptor-specific primer (Waugh et al., 1997), and transposon display (TD) using two rounds of PCR with nested transposon-specific primers (Casa et al., 2000; Grzebelus et al., 2007) and applied mostly to identify TIPs produced by DNA transposons. Those methods have often been used to identify TIPs derived from few known TE families. One of the first attempts where the S-SAP method was successfully applied to identify a newly inserted LTR-RT was reported by Tahara et al. (2004). They identified Ty1-copia retrotransposons in sweet potato activated in the callus. Similar approach was used by Yamashita and Tahara (2006), where a polymorphic S-SAP product was identified as a LINE retroelement activated in meristem stem cells. There are examples of S-SAP being successfully used also to identify ongoing transpositions upon stress other than *in vitro* cultures. For example, Woodrow et al. (2010) identified Ty1-copia transposition in durum wheat under salt and light stress. The effect of interspecific hybridization and polyploidization on the actively transposing LTR-RT using S-SAP was evaluated by Gantuz et al. (2022). Another TIP identification system named palindromic sequence-targeted PCR (PST-PCR v.2) was proposed by Kalendar et al. (2021). It relies on the use of capturing primers targeting palindromic sequences arbitrarily present in natural DNA templates in combination with a sequence –specific primer. PST-PCR v.2 consists of two rounds of PCR. The first round utilizes a combination of one sequence-specific primer with one capturing (PST) primer. The second round uses a combination of a single (preferred) or two universal primers; one anneals to a 5′ tail attached to the sequence-specific primer and the other anneals to a different 5′ tail attached to the PST primer. The key advantage of PST-PCR v.2 is to quickly produce amplified PCR fragments containing a portion of the template flanked by the sequence-specific and capturing primers. The approach allowed characterization of Ac transposon integration sites (Kalendar et al., 2021). Lack of restriction digestion and adapter ligation, i.e. steps required in S-SAP or TD, reduces the cost and time of identifying new insertion sites.

All wet-lab methods are primarily useful for monitoring the mobilization of previously identified TEs, e.g. under stress conditions or in a range of genetically diverse accessions, since they require the use of primers with a sequence specific to the sequence of the investigated TE. Moreover, the specificity of the amplification and the reliability of the new insertion sites should be confirmed by sequencing.

In 2004, the 454 technology became commercially available next generation sequencing (NGS) platform. Since then, NGS began to be widely applied to study plant TEs. In the early stages, they were usually combined with other techniques based on PCR amplification of regions specific to TEs. As an example, Monden et al. (2014), produced a LTR-RT libraries derived from eight strawberry cultivars, based on the primer binding site (PBS) adjacent to the conserved 5′ LTR motif and sequenced them using Illumina HiSeq2000. It allowed detection of cultivar-specific LTR-RT insertion sites.

Another approach for genome-wide TIPs detection produced by a single TE family includes AFLP-based enrichment of DNA fragments in TE sequences followed by Illumina library preparation and sequencing. The recently published TEAseq pipeline (Lyu et al., 2021) developed for maize Ds transposons consists of samples barcoding, TE enrichment, library preparation and Illumina sequencing. The bioinformatics workflow for sequencing data analysis starts from de-barcoding, next reads containing the TE sequence are identified, the TE-portion of the read is trimmed and the remaining portion of the sequence is mapped against the reference genome to identify the insertion site. The method was successfully used for the identification of 35,696 putative germinal insertion sites in over 1,600 Ds insertional mutants. The major advantage of such approach is not only more detailed information about the number of TE insertions and the level of polymorphism among tested individuals but also the availability of sequences of regions flanking insertions, that is vital for verification of novel insertion sites and their downstream analyses.

With the advent of high throughput sequencing technologies, strategies have been developed to mine for TE insertion sites using raw reads and a suite of bioinformatics tools is currently available (Serrato-Capuchina and Matute, 2018; Vendrell-Mir et al., 2019; Fan et al., 2022). Depending on the purpose of the analysis and the type of investigated TEs, different tools and approaches are being developed. Some tools like the TRACKPOSON (Carpentier et al., 2019) can identify TIPs very quickly and efficiently using discordant reads identified in the process of reads mapping against a TE sequence for the identification of insertions based on their position in the reference genome. It shortens time of the analysis at the expense of the precise determination of the site of insertion. Nevertheless, the identification of 'insertion signatures', i.e. TE sequences in specified genomic windows rather than their precise locations, might be the first choice for large-scale analysis of LTR-RTs, including thousands of re-sequenced genomes, as shown for the analysis of 3,000 rice genomes (Carpentier et al., 2019). The method reports both reference and non-reference insertions and does not require any prior TE annotation in the reference genome.

Tools based on the usage of discordant reads and split-reads report precise localization of insertion sites. That group of tools often requires high quality annotation of TEs in the reference genome, which in case of non-model organisms may limit their utility. In spite of higher computation demands, they can be efficiently used for large-scale population studies. Evaluation of this type of analysis make easier if another selective step is included in the experiment, such as the principle of TE sequence capture described firstly by Baillie et al. (2011) on the example of retrotranspositions registered in the human brain. Subsequently, this principle was used by Quadrana et al. (2016) in mining of transposition events within sequencing data for 211 *Arabidopsis thaliana* accessions. The SPLITREADER used here (Quadrana et al., 2016) was utilised for a global analysis of LTR-RTs in 602 tomato accessions and TIP-based GWAS (TE-GWAS; TIP-GWAS), that allowed identification of retrotransposon insertions associated with important phenotypic traits, such as flavor (Domínguez et al., 2020), while insertional polymorphism of class II MITEs in 3,000 rice genomes was analysed using PoPoolationTE2 (Kofler et al., 2016) and TIP-based GWAS showed association of particular MITE copies with MITE copy number, suggesting that MITE subfamilies originate from few "master" copies (Castanera et al., 2021). Another short read based method, RelocaTE2 (Chen et al., 2017) was used to analyse copy number and distribution of mPing, Ping and Pong class II elements actively transposing in rice in 3,000 rice genomes (Chen et al., 2019) and to detect *de novo* insertions of mPing in 272 rice recombinant inbred lines (RILs) developed from a cross between Nipponbare and HEG4 known to carry active mPing (Chen et al., 2020).

The obvious prerequisite for their utilization is availability of a high quality reference genome. The combination of high throughput sequencing and *in silico* discovery of new TE insertion events currently seems to be the most efficient strategy. Nevertheless, the risk that some new insertions are not being recorded still remains, but can be reduced by sufficient amount of reads i.e. it is necessary to achieve a high sequencing coverage.

It is also possible to utilize a pan-genome approach, i.e. to compare two or more genome assemblies representing the same or closely related species, with the intention of finding TIPs differentiating those genomes. However, availability of multiple genome assemblies limits the usage of such approach to the identification of TIPs and analyses of contribution of TEs to genome organization, as shown for four maize genotypes (Anderson et al., 2019), rather than for tracking or identification of active TEs.

A further significant improvement in the identification of TIPs may be achieved by the use of long-read NGS techniques, such as the Oxford Nanopore technology (ONT) (Ellison and Cao, 2020; Ewing et al., 2020). While short reads technologies work well in identifying insertion sites of small TEs, such as MITEs, long reads significantly improve the efficiency of analysis of longer elements, especially LTR-RTs that are the most abundant TEs in plant genomes. For example, the utility ONT was shown for detection of novel insertions of actively transposing LTR-RTs in Arabidopsis; EVD (Debladis et al., 2017) and ONSEN (Kirov et al., 2021), as well as for the identification of TIPs in collections of insertional mutants of Medicago and soybean (Song et al., 2021). Along with the development of long read sequencing, tools dedicated to the identification of insertions in such data are becoming available. The first tool identifying TIPs in long read data was PALMER (Pre-mAsking Long reads for Mobile Element insertion), based on the alignment of reads to the genome and masking reference insertions of the investigated TE family in the reads sequence. Subsequently, the TE sequence is identified in the unmasked part of the read and, based on the presence of specific features, the software identifies ends of TE, and the remaining part of the read is used to detect non-reference insertion sites (Zhou et al., 2020). The method was successfully applied to the human genome and it was adjusted to the most common human TEs (L1, Alu, SVA). Hence it may not work for other types of TEs, e.g. those abundant in plant genomes. Another pipeline, also developed to screen actively transposing human TEs, utilizes a slightly different strategy, as in the reads the portion mapped to the genome is masked, while the remaining part is mapped to a TE library, TE sequences are reconstructed and the remaining part of the sequence is re-mapped to the reference genome to identify non-reference insertions. In addition to TIPs identification, this pipeline allows analysis of TEs methylation, that is called by the software dedicated for identification of CpG methylation in ONT reads (Ewing et al., 2020). The long read sequencing methods produce reads overlapping full TE sequences and their flanking regions, providing opportunity for comprehensive characterization of those sequences. They also allow identification of TEs insertions within repetitive regions. However, for the identification of novel insertions of actively transposing elements, especially in plants, the Illumina platform is still a method of choice, as efficient bioinformatic tools have been available and the cost of sequencing is still much lower. The Cas9-targeted sequence capture to enrich library with TE sequences, in combination with long read sequencing, may be an alternative solution, that would reduce the cost of sequencing while still benefiting from the advantage provided by long reads (McDonald et al., 2021).

Long read sequencing also improves genome assemblies in TE-rich regions, TE detection, annotation and identification of TIPs (Shahid and Slotkin, 2020), opening new perspectives for better understanding of the TE biology and activity.

Based on the information provided, a screening was carried out to estimate the popularity of selected perspective approaches in the last period (see Figure 2). Here it is confirmed that the frequency of their use is generally increasing, especially in the last 2 years, while the use of Oxford Nanopore technology seems to be as most frequently used from compared approaches. Finally, the most important advantages and disadvantages of all discussed detection techniques were summarized (see Table 1).

Concluding remarks and future perspectives

Historically, the importance of TEs in plant genomes has been neglected. However, it turned out that their presence affects many areas important for the life and development of plants, as well as in terms of their possible use in the field of plant breeding. It puts pressure on the availability of suitable analytical methods to trace the

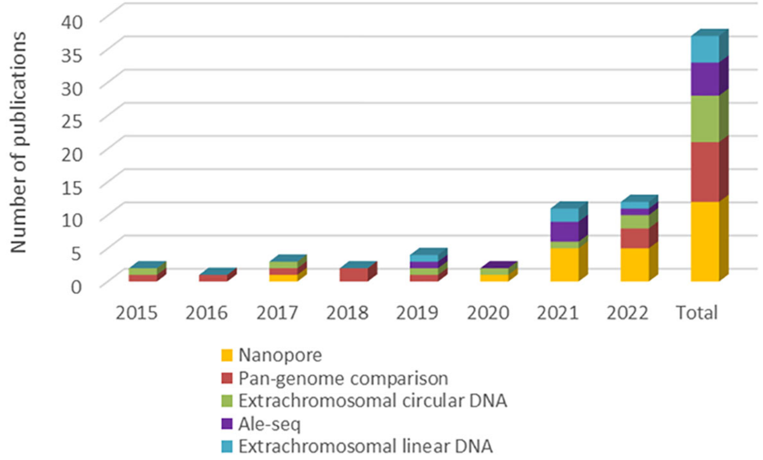


FIGURE 2
Popularity estimation of selected perspective approaches based on the frequency of their use in recent scientific articles. * The number of publications was generated by a search combining core keywords "*plant + transposable + activ**" and keywords corresponding to individual perspective approaches.

TABLE 1 Summary of approaches used to identify actively transposing elements.

Strategy used to identify actively transposing TEs	Main drawbacks	Recommendations for efficient targeting actively transposing TEs
Targeting TE-derived transcripts	- existence of TE-derived transcripts not competent for transposition (chimeric transcripts; transcripts involved in epigenetic silencing of TEs; post-transcriptional suppression mechanisms by TE-derived sequences)	- combine with another technique targeting products from the final phases of the transposition process (e.g. eclDNA, eccDNA)
Targeting TE-derived proteins	- non-transposing TEs can still express proteins - requires equipment that is not so common in molecular genetics laboratories	- combine with another technique targeting products from the final phases of the transposition process (e.g. eclDNA, eccDNA) - PIT (Proteomics Informed by Transcriptomics)
Targeting eclDNA	- eclDNA can occur as a result of other cellular processes (e.g. cell lysis, existence of micro-eclDNA)	- include a selective step to enrich TE-derived eclDNAs (e.g. PBS complementary to MET-iCAT tRNA) - combine with high throughput sequencing (ALE-Seq)
Targeting eccDNA	- eccDNA does not directly participate in the process of transposition - eccDNA can occur as a result of other cellular processes	- combine with high throughput sequencing to identify novel insertion sites
Identification of novel insertion sites by using TE-based genotyping platforms	- laborious and time consuming and error-prone	- use PST-PCR v.2 as a less laborious method
High throughput sequencing	- availability of a high quality reference genome or a large set of resequenced genomes of related accessions - inaccuracies related to short reads provided by the Illumina technology (problems with longer TEs, such as LTR-RTs; insertions in repetitive regions)	- use technologies producing long reads, e.g. Oxford Nanopore

pathways of actively transposing TEs. However, the interpretation of results produced by the above-presented methods can be difficult owing to the inherent properties of TEs. This review seeks to present techniques that can be used to obtain information about mobilized TEs and some pitfalls associated with the interpretation of results. The methods were divided on the basis of the context of their use with respect to the process of transposition.

Apparently, the use of some of the older methods mentioned above can be expedient in some specific cases and can bring unique information at relatively low price and experimental demands. The most comprehensive results are seemingly achievable by the methods based on massive parallel sequencing, however, they have also their limits. One such limitation is the fact that the created evaluation tools detect only a limited part of TEs. Related to this is also the need for thorough genomic TE annotation as an important prerequisite for appropriate detection of new copies. Some of shortcomings in the accuracy in bioinformatics data interpretation can be significantly improved by NGS techniques producing long reads. Generally, the strengths of one method are usually offset by other shortcomings. To obtain a comprehensive picture, a combination of methods based on different principles, seems to be the most effective. One of such examples is a strategy combining RNA-seq and MS, for which the designation Proteomics Informed by Transcriptomics is used. From the principle of the matter, a combination of methods targeting molecules originating from the final stages of the transposition process of actively transposing TEs seems to be the most suitable. Namely, it means to focus on methods aimed at detecting novel insertion sites, eclDNA and eccDNA. From this perspective, coupling WGS and analysis of the intermediates or by-signals of actively transposing TEs, such as eccDNA, ALE-Seq or multi-genomic comparisons, seems to be a promising approach to reveal complete information regarding TEs activity and their impact on host genome.

Author contributions

MBaj wrote first draft of the manuscript and perform graphical support. AP wrote sections of the manuscript that referred to current bioinformatics tools. DG contributed to conception, compiled and revised author contributions. MBar established conception and design, wrote some parts of manuscript and revised author contributions. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by Internal Grant of Mendel University (IGA-ZF/2021-SI1007) and by the project CZ.02.2.69/0.0/0.0/16_018/0002333 Research Infrastructure for

Young Scientists, this is co-financed from Operational Programme Research, Development and Education.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anderson, S. N., Stitzer, M. C., Brohammer, A. B., Zhou, P., Noshay, J. M., O'Connor, C. H., et al. (2019). Transposable elements contribute to dynamic genome content in maize. *Plant J.* 100 (5), 1052–1065. doi: 10.1111/tpj.14489

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., et al. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479 (7374), 534–537. doi: 10.1038/nature10531

Benoit, M., Drost, H.-G., Catoni, M., Gouil, Q., Lopez-Gomollon, S., Baulcombe, D., et al. (2019). Environmental and epigenetic regulation of rider retrotransposons in tomato. *PloS Genet.* 15 (9), e1008370. doi: 10.1371/journal.pgen.1008370

Böhrer, M., Rymen, B., Himber, C., Gerbaud, A., Pflieger, D., Laudencia-Chingcuanco, D., et al. (2020). "Integrated genome-scale analysis and northern blot detection of retrotransposon siRNAs across plant species," in *RNA Tagging* (Humana, New York, NY, Springer), 387–411.

Bortiri, E., Jackson, D., and Hake, S. (2006). Advances in maize genomics: the emergence of positional cloning. *Curr. Opin. Plant Biol.* 9 (2), 164–171. doi: 10.1016/j.pbi.2006.01.006

Cao, X., Wang, S., Ge, L., Zhang, W., Huang, J., and Sun, W. (2021). Extrachromosomal circular DNA: Category, biogenesis, recognition, and functions. *Front. Vet. Sci.* 8. doi: 10.3389/fvets.2021.693641

Carpentier, M.-C., Manfroi, E., Wei, F.-J., Wu, H.-P., Lasserre, E., Llauro, C., et al. (2019). Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nat. Commun.* 10 (1), 1–12. doi: 10.1038/s41467-018-07974-5

Casa, A. M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S., et al. (2000). The MITE family heartbreaker (Hbr): molecular markers in maize. *Proc. Natl. Acad. Sci.* 97 (18), 10083–10089. doi: 10.1073/pnas.97.18.10083

Castanera, R., Vendrell-Mir, P., Bardil, A., Carpentier, M. C., Panaud, O., and Casacuberta, J. M. (2021). Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. *Plant J.* 107 (1), 118–135. doi: 10.1111/tpj.15277

Chen, J., Lu, L., Benjamin, J., Diaz, S., Hancock, C. N., Stajich, J. E., et al. (2019). Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nat. Commun.* 10 (1), 1–10. doi: 10.1038/s41467-019-08451-3

Chen, J., Lu, L., Robb, S. M., Collin, M., Okumoto, Y., Stajich, J. E., et al. (2020). Genomic diversity generated by a transposable element burst in a rice recombinant inbred population. *Proc. Natl. Acad. Sci.* 117 (42), 26288–26297. doi: 10.1073/pnas.2015736117

Chen, J., Wrightsman, T. R., Wessler, S. R., and Stajich, J. E. (2017). RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. *PeerJ* 5, e2942. doi: 10.7717/peerj.2942

Cho, J., Benoit, M., Catoni, M., Drost, H.-G., Brestovitsky, A., Oosterbeek, M., et al. (2019). Sensitive detection of pre-integration intermediates of long terminal repeat retrotransposons in crop plants. *Nat. Plants* 5 (1), 26–33. doi: 10.1038/s41477-018-0320-9

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* 18 (2), 71–86. doi: 10.1038/nrg.2016.139

Davidson, A. D., Matthews, D. A., and Maringer, K. (2017). Proteomics technique opens new frontiers in mobilome research. *Mob Genet. Elements* 7 (4), 1–9. doi: 10.1080/2159256X.2017.1362494

Debladis, E., Llauro, C., Carpentier, M.-C., Mirouze, M., and Panaud, O. (2017). Detection of active transposable elements in arabidopsis thaliana using Oxford nanopore sequencing technology. *BMC Genomics* 18 (1), 1–8. doi: 10.1186/s12864-017-3753-z

Deininger, P., Morales, M. E., White, T. B., Baddoo, M., Hedges, D. J., Servant, G., et al. (2017). A comprehensive approach to expression of L1 loci. *Nucleic Acids Res.* 45 (5), e31. doi: 10.1093/nar/gkw1067

Deneweth, J., Van de Peer, Y., and Vermeirssen, V. (2022). Nearby transposable elements impact plant stress gene regulatory networks: A meta-analysis in a. thaliana and s. lycopersicum. *BMC Genomics* 23 (1), 1–19. doi: 10.1186/s12864-021-08215-8

Deniz, Ö., Frost, J. M., and Branco, M. R. (2019). Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* 20 (7), 417–431. doi: 10.1038/s41576-019-0106-6

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J. M., Colot, V., et al. (2020). The impact of transposable elements on tomato diversity. *Nat. Commun.* 11 (1), 1–11. doi: 10.1038/s41467-020-17874-2

Ellison, C. E., and Cao, W. (2020). Nanopore sequencing and Hi-c scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of drosophila melanogaster. *Nucleic Acids Res.* 48 (1), 290–303. doi: 10.1093/nar/gkz1080

Esposito, S., Barteri, F., Casacuberta, J., Mirouze, M., Carputo, D., and Aversano, R. (2019). LTR-TEs abundance, timing and mobility in solanum commersonii and s. tuberosum genomes following cold-stress conditions. *Planta* 250 (5), 1781–1787. doi: 10.1007/s00425-019-03283-3

Ewing, A. D., Smits, N., Sanchez-Luque, F. J., Faivre, J., Brennan, P. M., Richardson, S. R., et al. (2020). Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol. Cell* 80 (5), 915–928.e915. doi: 10.1016/j.molcel.2020.10.024

Fan, W., Wang, L., Chu, J., Li, H., Kim, E. Y., and Cho, J. (2022). Tracing mobile DNAs: From molecular to population scales. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.837378

Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* 3 (5), 329–341. doi: 10.1038/nrg793

Feschotte, C., and Pritham, E. J. (2007). DNA Transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* 41, 331–368. doi: 10.1146/annurev.genet.40.110405.090448

Fultz, D., Choudury, S. G., and Slotkin, R. K. (2015). Silencing of active transposable elements in plants. *Curr. Opin. Plant Biol.* 27, 67–76. doi: 10.1016/j.pbi.2015.05.027

Gantuz, M., Morales, A., Bertoldi, M. V., Ibañez, V. N., Duarte, P. F., Marfil, C. F., et al. (2022). Hybridization and polyploidization effects on LTR-retrotransposon activation in potato genome. *J. Plant Res.* 135 (1), 81–92. doi: 10.1007/s10265-021-01354-9

Gill, R. A., Scossa, F., King, G. J., Golicz, A., Tong, C. B., Snowdon, R. J., et al. (2021). On the role of transposable elements in the regulation of gene expression and subgenomic interactions in crop genomes. *Crit. Rev. Plant Sci.* 40 (2), 157–189. doi: 10.1080/07352689.2021.1920731

Grandbastien, M. A. (2015). LTR Retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim. Biophys. Acta* 1849 (4), 403–416. doi: 10.1016/j.bbagrm.2014.07.017

Griffiths, J., Catoni, M., Iwasaki, M., and Paszkowski, J. (2018). Sequence-independent identification of active LTR retrotransposons in arabidopsis. *Mol. Plant* 11 (3), 508–511. doi: 10.1016/j.molp.2017.10.012

Grzebelus, D., Jagosz, B., and Simon, P. W. (2007). The DcMaster transposon display maps polymorphic insertion sites in the carrot (Daucus carota l.) genome. *Gene* 390 (1-2), 67–74. doi: 10.1016/j.gene.2006.07.041

Gürkök, T. (2017). Transcriptome-wide identification and expression analysis of brachypodium distachyon transposons in response to viral infection. *Turkish J. Agriculture-Food Sci. Technol.* 5 (10), 1156–1160. doi: 10.24925/turjaf.v5i10.1156-1160.1260

Havecker, E. R., Gao, X., and Voytas, D. F. (2004). The diversity of LTR retrotransposons. *Genome Biol.* 5 (6), 1–6. doi: 10.1186/gb-2004-5-6-225

Hirochika, H., and Otsuki, H. (1995). Extrachromosomal circular forms of the tobacco retrotransposon ttol. *Gene* 165 (2), 229–232. doi: 10.1016/0378-1119(95)00581-P

Hollister, J. D., Smith, L. M., Guo, Y.-L., Ott, F., Weigel, D., and Gaut, B. S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between arabidopsis thaliana and arabidopsis lyrata. *Proc. Natl. Acad. Sci.* 108 (6), 2322–2327. doi: 10.1073/pnas.1018222108

Jiang, S., Cai, D., Sun, Y., and Teng, Y. (2016). Isolation and characterization of putative functional long terminal repeat retrotransposons in the pyrus genome. *Mob DNA* 7 (1), 1. doi: 10.1186/s13100-016-0058-8

Jiménez-Ruiz, J., Ramírez-Tejero, J. A., Fernández-Pozo, N., Leyva-Pérez, M. D. L. O., Yan, H., Rosa, R. D. L., et al. (2020). Transposon activation is a major driver in the genome evolution of cultivated olive trees (Olea europaea l.). *Plant Genome* 13 (1), e20010. doi: 10.1002/tpg2.20010

Kalendar, R., Antonius, K., Smýkal, P., and Schulman, A. H. (2010). iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. *Theor. Appl. Genet.* 121 (8), 1419–1430. doi: 10.1007/s00122-010-1398-2

Kalendar, R., and Schulman, A. H. (2006). IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. *Nat. Protoc.* 1 (5), 2478–2484. doi: 10.1038/nprot.2006.377

Kalendar, R., Shustov, A. V., and Schulman, A. H. (2021). Palindromic sequence-targeted (PST) PCR, version 2: an advanced method for high-throughput targeted gene characterization and transposon display. *Front. Plant Sci.* 12, 691940. doi: 10.3389/fpls.2021.691940

Kim, N. S. (2017). The genomes and transposable elements in plants: are they friends or foes? *Genes Genomics* 39 (4), 359–370. doi: 10.1007/s13258-017-0522-y

Kirov, I., Merkulov, P., Dudnikov, M., Polkhovskaya, E., Komakhin, R. A., Konstantinov, Z., et al. (2021). Transposons hidden in arabidopsis thaliana genome assembly gaps and mobilization of non-autonomous LTR retrotransposons unravelled by nanotei pipeline. *Plants* 10 (12), 2681. doi: 10.3390/plants10122681

Kirov, I., Omarov, M., Merkulov, P., Dudnikov, M., Gvaramiya, S., Kolganova, E., et al. (2020). Genomic and transcriptomic survey provides new insight into the organization and transposition activity of highly expressed LTR retrotransposons of sunflower (Helianthus annuus l.). *Int. J. Mol. Sci.* 21 (23), 9331. doi: 10.3390/ijms21239331

Kofler, R., Gómez-Sánchez, D., and Schlötterer, C. (2016). PoPoolationTE2: comparative population genomics of transposable elements using pool-seq. *Mol. Biol. Evol.* 33 (10), 2759–2764. doi: 10.1093/molbev/msw137

Kumar, P., Kiran, S., Saha, S., Su, Z., Paulsen, T., Chatrath, A., et al. (2020). ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines. *Sci. Adv.* 6 (20), eaba2489. doi: 10.1126/sciadv.aba24

Kwolek, K., Kędzierska, P., Hankiewicz, M., Mirouze, M., Panaud, O., Grzebelus, D., et al. (2022). Diverse and mobile–eccDNA-based identification of carrot low-copy LTR retrotransposons active in callus cultures. *Plant J* 110, 1811–1828. doi: 10.1111/tpj.15773

Lanciano, S., Carpentier, M. C., Llauro, C., Jobet, E., Robakowska-Hyzorek, D., Lasserre, E., et al. (2017). Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PloS Genet.* 13 (2), e1006630. doi: 10.1371/journal.pgen.1006630

Lanciano, S., and Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nat. Rev. Genet.* 21 (12), 721–736. doi: 10.1038/s41576-020-0251-y

Lanciano, S., Zhang, P., Llauro, C., and Mirouze, M. (2021). "Identification of extrachromosomal circular forms of active transposable elements using mobilome-seq," in *Plant transposable elements* (Humana, New York, NY, Springer), 87–93.

Liang, X., Hou, X., Li, J., Han, Y., Zhang, Y., Feng, N., et al. (2019). High-resolution DNA methylome reveals that demethylation enhances adaptability to continuous cropping comprehensive stress in soybean. *BMC Plant Biol.* 19 (1), 1–17. doi: 10.1186/s12870-019-1670-9

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26 (4), 493–500. doi: 10.1093/bioinformatics/btp692

Louis, B., Waikhom, S. D., Goyari, S., Jose, R. C., Roy, P., and Talukdar, N. C. (2015). First proteome study of sporadic flowering in bamboo species (Bambusa vulgaris and dendrocalamus manipureanus) reveal the boom is associated with stress and mobile genetic elements. *Gene* 574 (2), 255–264. doi: 10.1016/j.gene.2015.08.010

Lyu, M., Liu, H., Waititu, J. K., Sun, Y., Wang, H., Fu, J., et al. (2021). TEAseq-based identification of 35,696 dissociation insertional mutations facilitates functional genomic studies in maize. *J. Genet. Genomics* 48 (11), 961–971. doi: 10.1016/j.jgg.2021.07.010

Møller, H. D., Parsons, L., Jørgensen, T. S., Botstein, D., and Regenberg, B. (2015). Extrachromosomal circular DNA is common in yeast. *Proc. Natl. Acad. Sci.* 112 (24), E3114–E3122. doi: 10.1073/pnas.150882511

Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross-Ibarra, J., et al. (2015). Transposable elements contribute to activation of maize genes in response to abiotic stress. *PloS Genet.* 11 (1), e1004915. doi: 10.1371/journal.pgen.1004915

Manninen, I., and Schulman, A. H. (1993). BARE-1, a copia-like retroelement in barley (Hordeum vulgare l.). *Plant Mol. Biol.* 22 (5), 829–846. doi: 10.1007/BF00027369

Mann, L., Seibt, K. M., Weber, B., and Heitkam, T. (2022). ECCsplorer: a pipeline to detect extrachromosomal circular DNA (eccDNA) from next-generation sequencing data. *BMC Bioinf.* 23 (1), 1–15. doi: 10.1186/s12859-021-04545-2

Marcon, H. S., Domingues, D. S., Silva, J. C., Borges, R. J., Matioli, F. F., Fontes, M. R., et al. (2015). Transcriptionally active LTR retrotransposons in eucalyptus genus are differentially expressed and insertionally polymorphic. *BMC Plant Biol.* 15 (1), 198. doi: 10.1186/s12870-015-0550-1

Maringer, K., Yousuf, A., Heesom, K. J., Fan, J., Lee, D., Fernandez-Sesma, A., et al. (2017). Proteomics informed by transcriptomics for characterising active transposable elements and genome annotation in aedes aegypti. *BMC Genomics* 18 (1), 101. doi: 10.1186/s12864-016-3432-5

Mc Clintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A.* 36 (6), 344–355. doi: 10.1073/pnas.36.6.344

McDonald, T. L., Zhou, W., Castro, C. P., Mumm, C., Switzenberg, J. A., Mills, R. E., et al. (2021). Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat. Commun.* 12 (1), 1–13. doi: 10.1038/s41467-021-23918-y

Mehta, D., Cornet, L., Hirsch-Hoffmann, M., Zaidi, S. S. A., and Vanderschuren, H. (2020). Full-length sequencing of circular DNA viruses and extra-chromosomal circular DNA using CIDER-seq. *Nat. Protoc.* 15, 1673–1689. doi: 10.1038/s41596-020-0301-0

Mehta, D., Hirsch-Hoffmann, M., Were, M., Patrignani, A., Zaidi, S.S.E.A., and Were, H. (2019). A new full-length circular DNA sequencing method for viral-sized genomes reveals that RNAi transgenic plants provoke a shift in geminivirus populations in the field. *Nucleic Acids Res.* 47 (2), e9. doi: 10.1093/nar/gky914

Meyer, C., Pouteau, S., Rouze, P., and Caboche, M. (1994). Isolation and molecular characterization of dTnp1, a mobile and defective transposable element of nicotiana plumbaginifolia. *Mol. Gen. Genet.* 242 (2), 194–200. doi: 10.1007/BF00391013

Monden, Y., Fujii, N., Yamaguchi, K., Ikeo, K., Nakazawa, Y., Waki, T., et al. (2014). Efficient screening of long terminal repeat retrotransposons that show high insertion polymorphism *via* high-throughput sequencing of the primer binding site. *Genome* 57 (5), 245–252. doi: 10.1139/gen-2014-0031

Morillon, A., Benard, L., Springer, M., and Lesage, P. (2002). Differential effects of chromatin and Gcn4 on the 50-fold range of expression among individual yeast Ty1 retrotransposons. *Mol. Cell. Biol.* 22 (7), 2078–2088. doi: 10.1128/Mcb.22.7.2078-2088.2002

O'Neill, K., Brocks, D., and Hammell, M. G. (2020). Mobile genomics: tools and techniques for tackling transposons. *Philos. Trans. R Soc. Lond B Biol. Sci.* 375 (1795), 20190345. doi: 10.1098/rstb.2019.0345

Oberlin, S., Sarazin, A., Chevalier, C., Voinnet, O., and Mari-Ordonez, A. (2017). A genome-wide transcriptome and translatome analysis of arabidopsis transposons identifies a unique and conserved genome expression strategy for Ty1/Copia retroelements. *Genome Res.* 27 (9), 1549–1562. doi: 10.1101/gr.220723.117

Pandey, G., Yadav, C. B., Sahu, P. P., Muthamilarasan, M., and Prasad, M. (2017). Salinity induced differential methylation patterns in contrasting cultivars of foxtail millet (Setaria italica l.). *Plant Cell Rep.* 36 (5), 759–772. doi: 10.1007/s00299-016-2093-9

Paz, R. C., Rendina Gonzalez, A. P., Ferrer, M. S., and Masuelli, R. W. (2015). Short-term hybridisation activates Tnt1 and Tto1 copia retrotransposons in wild tuber-bearing solanum species. *Plant Biol. (Stuttg)* 17 (4), 860–869. doi: 10.1111/plb.12301

Picault, N., Chaparro, C., Piegu, B., Stenger, W., Formey, D., Llauro, C., et al. (2009). Identification of an active LTR retrotransposon in rice. *Plant J.* 58 (5), 754–765. doi: 10.1111/j.1365-313X.2009.03813.x

Piriyapongsa, J., and Jordan, I. K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA* 14 (5), 814–821. doi: 10.1261/rna.916708

Pozueta-Romero, J., Klein, M., Houlne, G., Schantz, M. L., Meyer, B., and Schantz, R. (1995). Characterization of a family of genes encoding a fruit-specific wound-stimulated protein of bell pepper (Capsicum annuum): identification of a new family of transposable elements. *Plant Mol. Biol.* 28 (6), 1011–1025. doi: 10.1007/BF00032663

Prada-Luengo, I., Krogh, A., Maretty, L., and Regenberg, B. (2019). Sensitive detection of circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads. *BMC Bioinf.* 20 (1), 1–9. doi: 10.1186/s12859-019-3160-3

Qiu, F., and Ungerer, M. C. (2018). Genomic abundance and transcriptional activity of diverse gypsy and copia long terminal repeat retrotransposons in three wild sunflower species. *BMC Plant Biol.* 18 (1), 6. doi: 10.1186/s12870-017-1223-z

Quadrana, L., Silveira, A. B., Mayhew, G. F., LeBlanc, C., Martienssen, R. A., Jeddeloh, J. A., et al. (2016). The arabidopsis thaliana mobilome and its impact at the species level. *eLife* 5, e15716. doi: 10.7554/eLife.15716.046

Ravindran, S. (2012). Barbara McClintock and the discovery of jumping genes. *Proc. Natl. Acad. Sci. U.S.A.* 109 (50), 20198–20199. doi: 10.1073/pnas.1219372109

Rocheta, M., Coito, J. L., Ramos, M. J. N., Carvalho, L., Becker, J. D., Carbonell-Bejerano, P., et al. (2016). Transcriptomic comparison between two vitis vinifera l. varieties (Trincadeira and touriga nacional) in abiotic stress conditions. *BMC Plant Biol.* 16 (1), 1–19. doi: 10.1186/s12870-016-0911-4

Rodriguez-Terrones, D., and Torres-Padilla, M. E. (2018). Nimble and ready to mingle: Transposon outbursts of early development. *Trends Genet.* 34 (10), 806–820. doi: 10.1016/j.tig.2018.06.006

Roquis, D., Robertson, M., Yu, L., Thieme, M., Julkowska, M., and Bucher, E. (2021). Genomic impact of stress-induced transposable element mobility in arabidopsis. *Nucleic Acids Res.* 49 (18), 10431–10447. doi: 10.1093/nar/gkab828

Sahebi, M., Hanafi, M. M., van Wijnen, A. J., Rice, D., Rafii, M. Y., Azizi, P., et al. (2018). Contribution of transposable elements in the plant's genome. *Gene* 665, 155–166. doi: 10.1016/j.gene.2018.04.050

Satheesh, V., Fan, W., Chu, J., and Cho, J. (2021). Recent advancement of NGS technologies to detect active transposable elements in plants. *Genes Genomics* 43 (3), 289–294. doi: 10.1007/s13258-021-01040-z

Serrato-Capuchina, A., and Matute, D. R. (2018). The role of transposable elements in speciation. *Genes* 9 (5), 254. doi: 10.3390/genes9050254

Sexton, C. E., and Han, M. V. (2019). Paired-end mappability of transposable elements in the human genome. *Mob DNA* 10 (1), 29. doi: 10.1186/s13100-019-0172-5

Shahid, S., and Slotkin, R. K. (2020). The current revolution in transposable element biology enabled by long reads. *Curr. Opin. Plant Biol.* 54, 49–56. doi: 10.1016/j.pbi.2019.12.012

Sigman, M. J., and Slotkin, R. K. (2016). The first rule of plant transposable element silencing: location, location, location. *Plant Cell* 28 (2), 304–313. doi: 10.1105/tpc.15.00869

Song, R., Wang, Z., Wang, H., Zhang, H., Wang, X., Nguyen, H., et al. (2021). InMut-finder: a software tool for insertion identification in mutagenesis using nanopore long reads. *BMC Genomics* 22 (1), 1–7. doi: 10.1186/s12864-021-08206-9

Sotero-Caio, C. G., Platt, R. N.2nd, Suh, A., and Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* 9 (1), 161–177. doi: 10.1093/gbe/evw264

Sow, M. D., Le Gac, A. L., Fichot, R., Lanciano, S., Delaunay, A., Le Jan, I., et al. (2021). RNAi suppression of DNA methylation affects the drought stress response and genome integrity in transgenic poplar. *New Phytol.* 232 (1), 80–97. doi: 10.1111/nph.17555

Sundaresan, V., and Freeling, M. (1987). An extrachromosomal form of the mu transposons of maize. *Proc. Natl. Acad. Sci.* 84 (14), 4924–4928. doi: 10.1073/pnas.84.14.4924

Sun, T., Wang, K., Liu, C., Wang, Y., Wang, J., and Li, P. (2019). Identification of extrachromosomal linear microDNAs interacted with microRNAs in the cell nuclei. *Cells* 8 (2), 111. doi: 10.3390/cells8020111

Tahara, M., Aoki, T., Suzuka, S., Yamashita, H., Tanaka, M., Matsunaga, S., et al. (2004). Isolation of an active element from a high-copy-number family of retrotransposons in the sweetpotato genome. *Mol. Genet. Genomics* 272 (1), 116–127. doi: 10.1007/s00438-004-1044-2

Tang, Y., Yan, X., Gu, C., and Yuan, X. (2022). Biogenesis, trafficking, and function of small RNAs in plants. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.825477

Torres, A. R., Rodrigues, E. P., Batista, J. S., Gomes, D. F., and Hungria, M. (2013). Proteomic analysis of soybean [Glycine max (L.) Merrill] roots inoculated with bradyrhizobium japonicum strain CPAC 15. *Proteomics Insights* 6, 7–11. doi: 10.4137/PRI.S13288

Usai, G., Mascagni, F., Vangelisti, A., Giordani, T., Ceccarelli, M., Cavallini, A., et al. (2020). Interspecific hybridisation and LTR-retrotransposon mobilisation-related structural variation in plants: A case study. *Genomics* 112 (2), 1611–1621. doi: 10.1016/j.ygeno.2019.09.010

Valdebenito-Maturana, B., and Riadi, G. (2018). TEcandidates: Prediction of genomic origin of expressed transposable elements using RNA-seq data. *Bioinformatics* 34 (22), 3915–3916. doi: 10.1093/bioinformatics/bty423

Vangelisti, A., Mascagni, F., Giordani, T., Sbrana, C., Turrini, A., Cavallini, A., et al. (2019). Arbuscular mycorrhizal fungi induce the expression of specific retrotransposons in roots of sunflower (Helianthus annuus l.). *PloS One* 14 (2), e0212371. doi: 10.1371/journal.pone.0212371

Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J. M., and Castanera, R. (2019). A benchmark of transposon insertion detection tools using real data. *Mobile DNA* 10 (1), 1–19. doi: 10.1186/s13100-019-0197-9

Voronova, A. (2019). Retrotransposon expression in response to *in vitro* inoculation with two fungal pathogens of scots pine (Pinus sylvestris l.). *BMC Res. Notes* 12 (1), 243. doi: 10.1186/s13104-019-4275-3

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T., Hornes, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23 (21), 4407–4414. doi: 10.1093/nar/23.21.4407

Vuong, L. M., Pan, S., and Donovan, P. J. (2019). Proteome profile of endogenous retrotransposon-associated complexes in human embryonic stem cells. *Proteomics* 19 (15), e1900169. doi: 10.1002/pmic.201900169

Wanchai, V., Jenjaroenpun, P., Leangapichart, T., Arrey, G., Burnham, C. M., Tümmler, M.C., et al. (2022). CReSIL: Accurate identification of extrachromosomal circular DNA from long-read sequences. *Brief. Bioinformatics*. 23 (6), bbac422. doi: 10.1093/bib/bbac422

Wang, K., Tian, H., Wang, L., Wang, L., Tan, Y., Zhang, Z., et al. (2021). Deciphering extrachromosomal circular DNA in arabidopsis. *Comput. Struct. Biotechnol. J.* 19, 1176–1183. doi: 10.1016/j.csbj.2021.01.043

Wang, X., Yang, P., Gao, Q., Liu, X., Kuang, T., Shen, S., et al. (2008). Proteomic analysis of the response to high-salinity stress in physcomitrella patens. *Planta* 228 (1), 167–177. doi: 10.1007/s00425-008-0727-z

Waugh, R., McLean, K., Flavell, A., Pearce, S., Kumar, A., Thomas, B., et al. (1997). Genetic distribution of bare–1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Mol. Gen. Genet. MGG* 253 (6), 687–694. doi: 10.1007/s004380050372

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8 (12), 973–982. doi: 10.1038/nrg2165

Woodrow, P., Pontecorvo, G., Fantaccione, S., Fuggi, A., Kafantaris, I., Parisi, D., et al. (2010). Polymorphism of a new Ty1-copia retrotransposon in durum wheat under salt and light stresses. *Theor. Appl. Genet.* 121 (2), 311–322. doi: 10.1007/s00122-010-1311-z

Yamashita, H., and Tahara, M. (2006). A LINE-type retrotransposon active in meristem stem cells causes heritable transpositions in the sweet potato genome. *Plant Mol. Biol.* 61 (1), 79–84. doi: 10.1007/s11103-005-6002-9

Zhang, P., Peng, H., Llauro, C., Bucher, E., and Mirouze, M. (2021). Ecc_finder: A robust and accurate tool for detecting extrachromosomal circular DNA from sequencing data. *Front. Plant Sci.* 12, 743742. doi: 10.3389/fpls.2021.743742

Zhou, W., Emery, S. B., Flasch, D. A., Wang, Y., Kwan, K. Y., Kidd, J. M., et al. (2020). Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* 48 (3), 1146–1163. doi: 10.1093/nar/gkz1173

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 December 2022
DOI 10.3389/fpls.2022.1011565

Check for updates

OPEN ACCESS

EDITED BY
Gennady I. Karlov,
All-Russia Research Institute of
Agricultural Biotechnology, Russia

REVIEWED BY
Katja R. Richert-Pöggeler,
Julius Kühn-Institut, Germany
Osamah Alisawi,
University of Kufa, Iraq

*CORRESPONDENCE
Carlos M. Vicient
carlos.vicient@cragenomica.es

SPECIALTY SECTION
This article was submitted to
Plant Bioinformatics,
a section of the journal
Frontiers in Plant Science

RECEIVED 04 August 2022
ACCEPTED 15 November 2022
PUBLISHED 14 December 2022

CITATION
de Tomás C and Vicient CM (2022)
Genome-wide identification of
Reverse Transcriptase domains of
recently inserted endogenous plant
pararetrovirus (*Caulimoviridae*).
Front. Plant Sci. 13:1011565.
doi: 10.3389/fpls.2022.1011565

Genome-wide identification of Reverse Transcriptase domains of recently inserted endogenous plant pararetrovirus (*Caulimoviridae*)

Carlos de Tomás and Carlos M. Vicient*

Structure and Evolution of Plant Genomes Group, Centre for Research in Agricultural Genomics, CSIC-IRTA-UAB-UB, Edifici CRAG, Bellaterra, Barcelona, Spain

Endogenous viral elements (EVEs) are viral sequences that have been integrated into the nuclear chromosomes. Endogenous pararetrovirus (EPRV) are a class of EVEs derived from DNA viruses of the family *Caulimoviridae*. Previous works based on a limited number of genome assemblies demonstrated that EPRVs are abundant in plants and are present in several species. The availability of genome sequences has been immensely increased in the recent years and we took advantage of these resources to have a more extensive view of the presence of EPRVs in plant genomes. We analyzed 278 genome assemblies corresponding to 267 species (254 from *Viridiplantae*) using tBLASTn against a collection of conserved domains of the Reverse Transcriptases (RT) of *Caulimoviridae*. We concentrated our search on complete and well-conserved RT domains with an uninterrupted ORF comprising the genetic information for at least 300 amino acids. We obtained 11.527 sequences from the genomes of 202 species spanning the whole Tracheophyta clade. These elements were grouped in 57 clusters and classified in 13 genera, including a newly proposed genus we called *Wendovirus*. Wendoviruses are characterized by the presence of four open reading frames and two of them encode for aspartic proteinases. Comparing plant genomes, we observed important differences between the plant families and genera in the number and type of EPRVs found. In general, florendoviruses are the most abundant and widely distributed EPRVs. The presence of multiple identical RT domain sequences in some of the genomes suggests their recent amplification.

KEYWORDS

pararetrovirus, Reverse Transcriptase (RT), *Caulimoviridae*, endogenous, virus

Abbreviations: EPRV, Endogenous pararetrovirus; RT, Reverse Transcriptase; EVEs, Endogenous viral elements; OUT, operational taxonomic unit.

Introduction

Endogenous viral elements (EVEs) are viral sequences that have been integrated into the nuclear chromosomes, enabling their vertical transmission and potential fixation in host populations (Feschotte and Gilbert, 2012). Viral integration within eukaryotic genomes is a widely recognized phenomenon described in many species thanks to the sequencing of whole genomes. Some of these EVEs are the consequence of a mandatory genome integration stage in the life cycle of reverse-transcribing viruses, such as retroviruses (Katzourakis and Gifford, 2010), but for others, such as all plant viruses, hepadnaviruses or the SARS-CoV-2, the integration in the host genome is not part of the virus life cycle and the mechanisms of integration are few well understood (Kojima et al., 2021; Zhang et al., 2021).

The first described plant EVE was a *Geminiviridae* element (Bejarano et al., 1996). EVEs derived from *Caulimoviridae* are abundant in plants (Diop et al., 2018). EVEs derived from another non-retroviral dsRNA, ssRNA, or ssDNA viruses have also been described as, for example, from *Narnaviridae* (Choi et al., 2021), *Partitiviridae, Betarhabdovirinae and Betaflexiviridae* (Chiba et al., 2011). The genome integration mechanism of the EVEs remains largely uncharacterized and different mechanisms for the integration were proposed. The most accepted theory is that endogenization results from a non-homologous recombination between virus and host genomes, usually in the context of either a double-stranded DNA break repair or a transposon-mediated process (Richert-Pöggeler et al., 2021).

If EVEs are integrated into or near host genes, this will be generally detrimental, and they will be removed from host population by purifying selection. In the rare cases that the integration of an EVE is beneficial, it will be fixed in the host population by positive selection, in the same way that occurs with other types of genomic elements like transposons (Catlin and Josephs, 2022). However, most of the EVEs are neutral and will become degraded due to the accumulation of disruptive mutations, insertions or deletions. Due to the random nature of these mutations, it is possible to reconstruct the sequences of the infectious viruses based on the EVEs sequences, particularly for high copy number EVEs (Aiewsakun and Katzourakis, 2015). In consequence, EVEs can be considered as genomic "fossils" and be employed for investigating viral origins and diversity and become the main tool for a new emerging field called Paleovirology. Paleovirology is the study of the ancient evolution of viruses through analyzing endogenous viral elements in the host genomes (Etienne, 2017). Due to the increasing number of sequenced genomes, numerous EVEs can be uncovered, and some of them are distinct from the currently known episomal viruses (Johnson, 2019). Another important property of EVEs is that they can be used to calibrate the timing of virus evolution. If an EVE is orthologous across several species, this gives a minimum estimate for the age of the virus that integrated into the genome (Aiewsakun and Katzourakis, 2015).

Caulimoviridae is a family of double-stranded DNA (dsDNA) viruses infecting plants that contain a reverse transcription stage in their replication cycle (International Committee on Taxonomy of Viruses, ICTV, https://ictv.global/). Although integration into the genome is not an essential part of their replication cycle, there are much evidence of their presence as integrated forms among genomes of the plant kingdom (Geering et al., 2014; Diop et al., 2018) and they have been included as a new category in some repetitive DNA sequence databases like Repbase (Bao et al., 2015). *Caulimoviridae* can be classified into 11 genera based on their genome organization (number of open reading frames and the arrangement of protein domains within them) and the morphology of their virus particles (ICTV, https://ictv.global/). Some of these genera have been reported as EVEs in plant genomes (Endogenous pararetrovirus, EPRVs) but, in addition, many of the EPRVs belong to a genus for which so far no episomal counterparts have been described (Geering et al., 2014; Chen and Kishima, 2016; Diop et al., 2018).

The integration of an EVE into or near a gene can potentially modify gene transcription or modify mRNA processing, resulting in mutant phenotypes. Most of the described EPRVs are inserted in intergenic regions and have no apparent deleterious effect on the host. However, there are examples of EPRVs inserted inside genes with potential effects on gene expression as, for example, in the case of *Vitis vinifera*, which has several EPRVs inserted in introns (Geering et al., 2014).

Most of the EPRVs are transcriptionally or translationally inactive because they are partial and/or comprise rearranged sequences and/or inactivating mutations. Often EPRVs form clusters resulting from the simultaneous integration of several complete or partial copies in tandem or nested (Richert-Pöggeler et al., 2003). Infrequently, these integrated sequences are transcriptionally active and the resulting RNAs can serve as precursors of extrachromosomal viral DNA and lead to systemic and vertically transmitted infections (Hohn et al., 2008; Gayral et al., 2008). Transcriptional activation can be driven by viral promoters present within the integrated element or plant promoters in the vicinity of the EPRV sequence (Lockhart et al., 2000; Kuriyama et al., 2020). On the other hand, EPRV derived RNAs can also be inducers for RNA interference (RNAi) and gene silencing mechanisms through the generation of small interfering RNAs (siRNAs) (Bertsch et al., 2009; Ricciuti et al., 2021).

RNA-directed DNA polymerase (Reverse Transcriptase, RT) coding sequences are present in a wide variety of genetic elements and contains a relatively well conserved central domain, allowing its use for phylogenetic analyses (Hansen and Heslop-Harrison, 2004) and for searches for homologues of, for example, EPRVs in genome sequences (Diop et al., 2018).

Previous studies have examined the EPRVs diversity in plant genomes based on the limited number of genome sequences available in each case (Geering et al., 2014; Diop et al., 2018) Nowadays, the number of sequenced plant genomes have increased significantly, and we decided to screen them for the presence of EPRVs, obtaining a broader picture of the distribution of these endogenous elements. We identified the major EPRV lineages and analyzed their distribution in the different plant orders and genera. We also describe a new possible genus of *Caulimoviridae* present only as EPRVs we called *Wendovirus*.

Materials and methods

Discovery and analyses of recently inserted endogenous *Caulimoviridae*

We built a library containing an assortment of 182 RT central domain amino acid sequences (Supplementary Data 1). This collection includes one sequence from *Retroviridae*, 14 from Ty3/Gypsy LTR retrotransposons of the six most abundant genera in plants (Athila, CRM, Galadriel, Ogre, Reina, Retand and Tekay), 104 from the eleven genera of *Caulimoviridae* (*Badnavirus*, *Caulimovirus*, *Vaccinivirus*, *Soymovirus*, *Cavemovirus*, *Solendovirus*, *Dioscovirus*, *Rosadnavirus*, *Tungrovirus*, *Petuvirus* and *Ruflodivirus*), and 63 from six groups of exclusively endogenous *Caulimoviridae* (*Florendovirus*, *Xendovirus*, *Yendovirus*, *Zendovirus*, *Gymnendovirus* and *Fernendovirus*) (hereafter referred to as operational taxonomic units (OTUs) following the nomenclature proposed by Diop et al., 2018. For further analyses, we selected ten sequences representatives of the *Caulimoviridae* groups (Supplementary Data 2).

We selected 278 genome assemblies corresponding to 267 species (Supplementary Data 3): two from *Bacteria*, one from *Chromista*, two from *Protozoa*, 13 from *Animal*, six from *Fungi* and 254 from *Plantae* kingdom. *Plantae* kingdom's genomes include three *Rodophyta*, seven *Chlorophyta*, three *Bryophyta*, one *Marchantiophyta* and 240 *Tracheophyta* genomes. *Tracheophyta* includes one *Lycopodiopsida*, four *Pinopsida*, 35 *Liliopsida* (11 families) and 200 *Magnoliopsida* (46 families) genomes. The genomes outside the *Plantae* kingdom were used as negative controls.

We compared the ten RT sequences with the 278 genome assemblies using tBLASTn with default parameters (except –e option set to 1e–10). Only the hits with at least 300 amino acid residues and no stop codons nor frameshifts were selected for further analysis. To avoid the inclusion in the selection of tandem duplications, we removed a hit if it was located less than 1500 bp to another (Supplementary Data 3). For each genome assembly, the selected set of RT sequences were clustered with the 182 RT selected reference domains and those having higher similarity with retrotransposons were removed from the analyses. RT sequences having higher similarity with *Caulimoviridae* were used for further analyses (Supplementary Data 4).

For cluster determination, the selected sequences from the genome assemblies were grouped using CD-HIT with a sequence identity cut-off of 60% (Cluster60) or of 100% (Cluster100), a bandwidth of alignment of 20 and a length of sequence to skip of 10. One sequence was then selected to be representative of each cluster60 (Supplementary Data 5). Only in the case of cluster60-8 we selected two sequences because the sequences in this cluster were clearly divided in two groups.

The cluster representative sequences were aligned with the representative sequences of episomal or endogenous *Caulimoviridae* (Supplementary Data 1) using MEGA-X (Kumar et al., 2018). The resulting alignment was then used to build a phylogenetic reconstruction using the maximum likelihood (ML) method and 500 bootstrap replicates using MEGA-X. The resulting tree was then used as a reference to classify the EPRV-RTs found in the genome assemblies.

The minimum ages of the integration events reported in this study were inferred by identifying the most distantly related pair of host species sharing a particular cluster of EPRVs and applying the estimated species divergence dates in TimeTree (http://www.timetree.org/) (Kumar et al., 2017).

Potential ORFs were predicted using ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/) and the presence of Pfam domains in their encoded polypeptides was confirmed using MOTIF Search (https://www.genome.jp/tools/motif/).

Results

Distribution of genomic sequences encoding Reverse Transcriptase domains of recently inserted endogenous pararetroviruses (*Caulimoviridae*)

The objective of the work was to determine the presence of sequences encoding complete conserved RT domains corresponding to endogenous pararetrovirus (*Caulimoviridae*) within a collection of publicly available genome sequence assemblies from plant species and using some non-plant genome assemblies as negative controls. To identify them, we used a custom designed tBLASTn-based discovery pipeline, using as a probe a collection of 10 representative RT sequences of the different *Caulimoviridae* genera and OTUs (Supplementary Data 2). To give priority to the recently inserted copies, we only select sequences encoding RT domains of at least 300 amino acids that contain uninterrupted reading frames. Frequently EPRVs are inserted in tandemly arranged structures. To remove these duplications, when a RT coding region was located less than 1500 bp of another we only kept one of them. Due to their high sequence similarity, this first selection also contained RT sequences from Ty3/gypsy LTRretrotansposons (*Metaviridae)*. To remove them, EPRVs were confirmed by phylogenetic analyses. They were aligned with RT sequences of representative *Caulimoviridae* and LTR retrotransposons (Supplementary Data 1). Those sequences showing higher similarity with the *Metaviridae* than with *Caulimoviridae* were removed. Finally, we obtained 11.527 RT-EPRV sequences (Supplementary Data 4).

None of the analyzed genomes outside *Plantae* Kingdom contain RT- EPRV sequences and among the genomes of the *Plantae* kingdom, we did not find RT-EPRVs in *Chlorophyta*, *Rodophyta, Bryophyta* or *Marchantiophyta*. Among the *Tracheophyta* species, we did not find RT-EPRVs in the class *Lycopodiopsida* (*Selaginella moellendorffii*) but we found RT-EPRVs in genomes of all *Tracheophyta* classes (*Pinopsida*, *Liliopsida* and *Magnoliopsida*), confirming previous results (Gong and Han, 2018). All the four *Pinopsida* genomes analyzed contain RT-EPRV sequences (between 4 and 46). We included 35 genomes of species of the class *Liliopsida* and we found RT-EPRV sequences in 22 of them (63%) (between 1 and 63). Finally, we found RT-EPRV sequences in 180 of the 201 *Magnaliopsida* genomes (88%) (between 1 and 1186).

When comparing the results with the genomes of species belonging to the same genus, or varieties of the same species, the results obtained are, in general, similar. For example, the genomes of the two species of *Kalanchoe* contain 20 and 24, the two of *Vitis* contain 24 and 29 and the three of *Solanum* between 29 and 35. However, this is not always the case, and we can observe important differences in the number of RT-EPRVs in species of the same genus. For example, in the genera *Arachis* (between 56 and 473), *Prunus* (between 3 and 144), *Rosa* (between 76 and 340), *Citrus* (between 63 and 306) and *Nicotiana* (between 12 and 130). Some of these differences can be due to differences in the quality of the genome assemblies. For example, the presence of undetermined nucleotides can give rise to a reduction in the number of RT-EPRVs we detected. However, there are cases in which the best quality genome is the one with the least number of sequences. For example, we included three species of the genera *Arabidopsis* and the genome with the least number of sequences is the one with the best quality *(Arabidopsis thaliana*). All these results suggest that in some of the species there have been very recent integrations of EPRVs.

Classification of the RT-EPRVs present in plant genomes

To provide a classification, RT-EPRV sequences with at least 60% amino acid identity to each other were grouped, yielding a total of 57 clusters. The total number of sequences and genomes represented in each cluster varies greatly (Table 1). We performed a phylogenetic analysis using representative sequences of each cluster (Supplementary Data 5) and representatives of all *Caulimoviridae* genera and OTUs (Supplementary Data 1). Our phylogenetic analysis clustered together all the previous known sequences corresponding to the same genera and OTU of the *Caulimoviridae*, confirming the robustness of the analysis (Figure 1). This phylogenetic reconstruction allowed us to determine the diversity and nature of our collection of RT-EPRV sequences (Table 2). They were separated into 13 *phyla*. 30 of the clusters were associated with sequences of *Caulimoviridae* with episomal forms: 10 *Petuvirus*, 5 *Dioscovirus*, 5 *Soymovirus*, 5 *Tungrovirus*, 2 *Badnavirus*, 2 *Caulimovirus* and 1 *Solendovirus*. We did not find any representative of the genera *Cavemovirus*, *Rosadnavirus* or *Vaccinivirus*, and neither from the recently proposed genera *Ruflodivirus*. This result suggests that the virus species of these genera do not carry out endogenization, at least not recently or as frequently, or they only do it in a small range of species whose complete genomic sequence is not yet available. Of the rest, 20 clusters corresponded to OTUs from which only endogenous forms have been found: 11 *Florendovirus*, 3 *Xendovirus*, 3 *Yendovirus*, 3 *Zendovirus* and 1 *Gymnendovirus*. As we will describe later in detail, the remaining 6 clusters were associated with each other, forming a new OTU we called *Wendovirus* (Figure 1).

We observed important differences between genera for both the number of RT- EPRV sequences and the diversity of species in which they were found (Table 1). *Florendovirus* are clearly the most abundant followed by *Petuvirus*, *Solendovirus* and *Zendovirus*. However, whereas *Florendovirus* is present in genomes of 40 families of species, *Petuvirus* is present in 14 and *Solendovirus* and *Zendovirus* in only two. Interestingly, although we only detected 80 RT-EPRV sequences corresponding *Badnavirus*, they present a wide distribution (3 Classes, 10 Orders and 11 Families). On the opposite, *Gymnendovirus* are only present in *Pinopsida*.

If we look at the different classes of plants, we observed important differences. *Pinopsida* only contains *Gymnendovirus*. *Magnolids* contains *Badnavirus*, *Petuvirus*, *Solendovirus*, *Tungrovirus*, *Florendovirus* and *Yendovirus*. *Liliopsida* contains *Badnavirus*, *Dioscovirus*, *Florendovirus* and *Yendovirus*. Finally, *Magnaliopsida* contains all the genera except *Gymnendovirus*.

If we look at the distribution of the clusters in the different plant species, we observed a wide diversity (Table 2). Some of them are exclusively present in one class. For example, Gymnendovirus-1 is only present in *Pinopsida*, Tungrovirus-3 is only present in *Magnolids*, Badnavirus-2, Dioscovirus-2 and -5 and Yendovirus-1 are only present in *Liliopsida*, and many clusters are only present in *Magnaliopsida*. On the opposite, Badnavirus-1, Florendovirus-1 and Florendovirus-3 are present in *Magnolids*, *Liliopsida* and *Magnoliopsida*. Looking at more detail, 31 of the 57 clusters are present in genomes of only one family of plants, whereas two are present in genomes of more than 20 plant families (both florendovirus). These differences of distribution are reflected in the Maximum Age Value (Table 1),

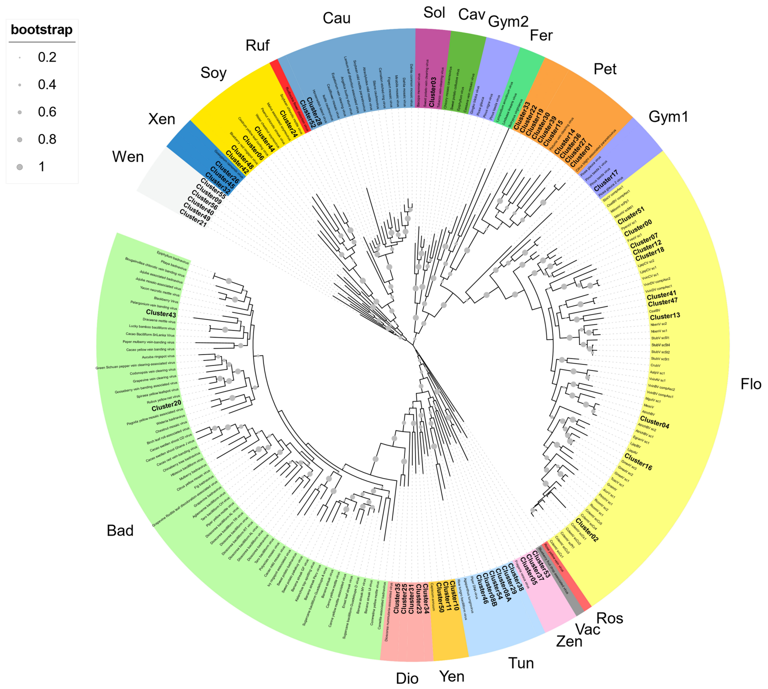


FIGURE 1
Phylogenetic relationships within the episomal and endogenous *Caulimoviridae*. Phylogram obtained from a maximum likelihood analysis with protein sequence data from RT conserved domains using 500 bootstrap replications. The size of the point indicated the bootstrap support of the tree branch. Known episomal and endogenous pararetrovirus are shown in grey and small letters. New endogenous Clusters60 are shown in bold letters. The color of the branch indicates the genus of *Caulimoviridae*; Bad, *Badnavirus*; Dio, *Dioscovirus*; Yen, *yendovirus*; Tun, *tungrovirus*; Zen, *zendovirus*; Vac, *vaccinivirus*; Ros, *rosadnavirus*; Flo, *florendovirus*; Gym1 and Gym2, *gymnendovirus*1 and 2; Pet, *petuvirus*; Fer, *fernendovirus*; Cav, *cavemovirus*; Sol, *solendovirus*; Cau, *caulimovirus*; Ruf, *ruflodivirus*; Soy, *soymovirus*; Xen, *xendovirus*; and Wen, *wendovirus*.

which depends on the maximum phylogenetic distance between the species present in the cluster.

Very recent EPRV amplification in plant genomes

The above results suggest that, at least in some species, there has been a recent amplification in the number of EPRV sequences inserted in their genomes. To try to delve further into this aspect, we decided to select those cases in which 100% identical RT-EPRV sequences were present in 10 or more copies in the same genome. Using this highly restrictive criterion, we detected 31 clusters grouping a total of 1534 sequences (Table 3). These clusters (clusters100) involve 19 genomes. Only one corresponds to a *Liliopsida (Hordeum vulgare)* and the remaining 18 are genomic sequences of *Magnaliophyta*. Nine EPRV OTUs are represented in the Clusters100 including *Caulimovirus*, *Dioscovirus*, *Florendovirus*, *Petuvirus*, *Solendovirus*, *Tungrovirus*, *Yendovirus*, *Zendovirus* and the newly proposed *Wendovirus*.

Cluster100-10 is particularly noteworthy as it includes 951 sequences present in the genome of pepper (*Capsicum annuum*). Another four groups also correspond to the same genome, with a total of 1014 sequences (962 are *Solendovirus*, 31 are *Florendovirus* and 21 *Yendovirus*). In total, we found 1183 RT-EPRV sequences in this genome and more than 81% are present in the Cluster100 selection. This is a very clear indication of a relatively recent proliferation of EPRVs in the pepper genome.

Next, we perform a phylogenetic analysis of representatives of each Cluster-100 and from the described OTUs from *Caulimoviridae* (Figure 2). The sequences of some of the clusters100 are very similar and, probably, they correspond to the same virus. This is the case of clusters100-1 and -26 (*Solendovirus* of *Capsicum annuum*), clusters100-11 and -13 (*Petuvirus* of *Atalantia buxifolia*) and Clusters100-5 and -6 (*Petuvirus* of *Citrus medica*). The sequences of clusters100-12 and -16 (*Florendovirus* of *Fortunella hindsii*) and of clusters100-19 and -24 (*Florendovirus* of *Atalantia buxifolia*) are also near identical. The sequences of the Clusters100-20 and 29, that correspond to two different but closely related species (*Nicotiana tabacum* and *Nicotiana sylvestris*), are also almost identical, which suggests that they could come from the same virus capable of infecting both species. Figure 2 also shows that some of the endogenous sequences grouped in Clusters100 are very similar to the sequences of episomal virus. For example, the

TABLE 1 Cluster60 statistics.

Cluster	Cluster N.	EPRV-RT-seqs	N.Classes	N.Orders	N.Families	N.Genus	N.Species	A	B	Max.Age (MY)
BADNAVIRUS		80	3	10	11	12	13			
Badnavirus-01	20	75	2	9	10	12	12	Dioscorea	Amborella	191
Badnavirus-02	43	5	1	2	2	2	2	Phalaenopsis	Musa	117
CAULIMOVIRUS		38	1	4	4	6	9			
Caulimovirus-01	28	36	1	3	3	5	8	Helianthus	Arabidopsis	118
Caulimovirus-02	52	2	1	1	1	1	1	Gossypium	Gossypium	0
DIOSCOVIRUS		144	2	5	5	7	9			
Dioscovirus-01	23	49	1	3	3	4	5	Cynara	Cajanus	118
Dioscovirus-02	25	43	1	1	1	1	2	Dioscorea	Dioscorea	0
Dioscovirus-03	31	24	1	1	1	2	2	Glycine	Vigna	23
Dioscovirus-04	34	16	1	1	1	1	1	Macadamia	Macadamia	0
Dioscovirus-05	35	12	1	1	1	1	2	Dioscorea	Dioscorea	0
PETUVIRUS		1693	2	14	16	47	66			
Petuvirus-01	1	1202	1	5	5	10	19	Arachis	Citrus	108
Petuvirus-02	14	131	1	9	9	16	18	Amborella	Helianthus	191
Petuvirus-03	15	129	1	3	4	6	9	Coffea	Gossypium	118
Petuvirus-04	19	78	1	1	1	11	13	Brassica	Rorippa	27
Petuvirus-05	22	52	1	1	1	1	1	Ipomoea	Ipomoea	0
Petuvirus-06	27	39	1	1	1	7	9	Arachis	Cicer	59
Petuvirus-07	30	24	1	3	3	4	6	Populus	Gossypium	108
Petuvirus-08	33	18	1	1	1	3	8	Citrus	Atalantia	18
Petuvirus-09	36	12	1	2	2	2	2	Durio	Macadamia	123
Petuvirus-10	39	8	1	1	1	1	1	Eucalyptus	Eucalyptus	0
SOLENDOVIRUS		1124	1	2	2	5	8			
Solendovirus-01	3	1124	1	2	2	5	8	Nymphaea	Nicotiana	179
SOYMOVIRUS		454	1	5	6	12	14			
Soymovirus-01	6	391	1	1	1	1	3	Arachis	Arachis	0
Soymovirus-02	24	49	1	4	5	6	6	Lactuca	Cleome	118
Soymovirus-03	42	6	1	1	1	1	1	Chenopodium	Chenopodium	0
Soymovirus-04	44	5	1	1	1	3	3	Brassica	Cakile	13
Soymovirus-05	48	3	1	1	1	1	1	Medicago	Medicago	0
TUNGROVIRUS		308	2	5	5	10	32			
Tungrovirus-01	8	251	1	3	3	10	29	Prunus	Vitis	117
Tungrovirus-02	29	32	1	1	1	1	1	Lindenbergia	Lindenbergia	0

(Continued)

TABLE 1 Continued

Cluster	Cluster N.	EPRV-RT-seqs	N.Classes	N.Orders	N.Families	N.Genus	N.Species	A	B	Max.Age (MY)
Tungrovirus-03	38	9	1	1	1	1	1	Cinnamomum	Cinnamomum	0
Tungrovirus-04	46	4	1	1	1	1	2	Malus	Malus	0
Tungrovirus-05	54	2	1	1	1	1	1	Citrus	Citrus	0
FLORENDOVIRUS		6162	3	29	40	91	151			
Florendovirus-01	0	3207	2	27	34	70	114	Asparagus	Amborella	191
Florendovirus-02	2	1188	1	6	8	21	35	Brassica	Nicotiana	118
Florendovirus-03	4	949	2	21	27	38	47	Asparagus	Amborella	191
Florendovirus-04	7	317	1	2	2	2	3	Coffea	Lindenbergia	77
Florendovirus-05	12	133	1	1	1	3	5	Arachis	Lotus	59
Florendovirus-06	13	132	1	2	2	5	8	Lindenbergia	Nicotiana	79
Florendovirus-07	16	120	1	8	9	13	18	Amborella	Brassica	191
Florendovirus-08	18	79	1	2	2	7	8	Glycine	Manihot	101
Florendovirus-09	41	7	1	1	2	2	2	Capsicum	Nicotiana	24
Florendovirus-10	47	4	1	1	1	2	2	Cucumis	Momordica	48
Florendovirus-11	51	2	2	2	2	2	2	Asparagus	Prunus	160
GYMNENDOVIRUS		95	1	1	1	2	3			
Gymnendovirus-1-1	17	95	1	1	1	2	2	Pinus	Picea	130
WENDOVIRUS		282	1	7	7	10	17			
Wendovirus-01	9	200	1	1	1	4	11	Citrus	Atalantia	18
Wendovirus-02	21	70	1	2	2	3	3	Helianthus	Coffea	101
Wendovirus-03	40	7	1	2	2	3	4	Citrus	Solanum	118
Wendovirus-04	49	3	1	1	1	1	1	Lindenbergia	Lindenbergia	0
Wendovirus-05	55	1	1	1	1	1	1	Olea	Olea	0
Wendovirus-06	56	1	1	1	1	1	1	Portulaca	Portulaca	0
XENDOVIRUS		65	1	6	6	8	10			
Xendovirus-01	26	41	1	4	4	6	8	Vaccinium	Rosa	118
Xendovirus-02	32	19	1	1	1	1	1	Olea	Olea	0
Xendovirus-03	45	5	1	1	1	1	1	Ipomoea	Ipomoea	0
YENDOVIRUS		334	2	6	7	17	23			
Yendovirus-01	10	190	1	1	1	9	11	Oryza	Eleusine	47
Yendovirus-02	11	142	2	5	5	8	12	Dioscorea	Solanum	160
Yendovirus-03	50	3	2	2	2	2	2	Ananas	Nymphaea	179
ZENDOVIRUS		781	1	2	2	5	19			
Zendovirus-01	5	768	1	1	1	4	18	Fragaria	Rubus	41
Zendovirus-02	37	11	1	1	1	2	4	Fragaria	Rosa	31
Zendovirus-03	53	2	1	1	1	1	1	Pistacia	Pistacia	0

TABLE 2 Distribution of Cluster60 in plant families.

			BADN		CAUL		DIOSCOVIRUS					PETUVIRUS										SL	SOYMOVIRUS					TUNGROVIR					FLORENDOVIRUS											G	WENDOVIRUS						XENDOV			YENDOV			ZENDOV		
Class	Order	Family	1	2	1	2	1	2	3	4	5	1	2	3	4	5	6	7	8	9	10	1	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	6	7	8	9	10	11	1	1	2	3	4	5	6	1	2	3	1	2	3	1	2	3
Pinopsida	Pinales	Pinaceae																																										24															
Liliopsida	Acorales	Acoraceae																															15																										
	Alismatales	Lemnaceae																																																								
		Zosteraceae																																																								
	Dioscoreales	Dioscoroceae	1					22			6																																												3				
	Asparagales	Asparagaceae																															19		40								1																
		Orchidaceae		2																															3																								
	Arecales	Arecaceae																															1		17																								
	Poales	Bromeliaceae																																																					1	2			
		Joinvillaceae																																																								
		Poacea																															1																					9					
	Zingiberales	Musaceae	5	1																																																							
Magnaliopsida	Amborellales	Amborellaceae	1										3																				59		8				10																				
	Nymphaeales	Nymphaeaceae	10																			6																																		1			
	Laurales	Lauraceae																												9			7		1																								
	Ranunculales	Papaveraceae																															2		34																								
		Ranunculaceae																																	38																								
	Proteales	Nelumbonaceae																																																								
		Proteaceae								16			27							11													70																										
	Saxifragales	Crassulaceae																																	27																								
	Vitales	Vitaceae																										8					18		1																								
	Celastrales	Celastraceae																															1		6																				4				
	Malpighiales	Euphorbiaceae										4																					33	33	17				1	1																			
		Linaceae																																																								
		Salicaceae			1													2																3																									
	Fabales	Fabaceae					1		1			2					2						23	1			1						24	1	2		8		3	4																			
	Rosales	Rhamnaceae	2																														18																										
		Moraceae										1																																															
		Rosaceae	1																									5			1		26		1								1								1						16	1	
	Urticales	Cannabaceae											2																																														
	Cucurbitales	Cucurbitaceae																															4		1							1																	
	Fagales	Juglandaceae	1																														49																										
		Fagaceae											10																				102																										
		Betulaceae																															63																										
	Myrtales	Lythraceae																						20									30						4																				
		Myrtaceae										1	2	1				3			4																		20																				
	Sapindales	Anacardiaceae											2																				1		1																								1
		Rutaceae	1									105							2									1				1	1	91					1						18		1												
	Malvales	Malvaceae				1							1	2				1		1													4						1																				
	Brassicales	Brassicaceae			1								1		3									1		1							2	1	1				1																				
		Cleomaceae																						3									3	7																									
		Caricaceae																																																								
	Caryophyllales	Amaranthaceae																							2								1																										
		Portulacaceae																																	4															1									
	Cornales	Hydrangeaceae																															155		185																14								
	Ericales	Ericaceae																															40		36																13								
		Theaceae	35																														71		4																								
	Gentianales	Rubiaceae												47																			2	2		150										2									29				
	Solanales	Convolvulaceae					21									26																																					3						
		Solanaceae												3								140											1	1	2			15			1						1								5				
	Lamiales	Lamiaceae																															96		58																								
		Oleaceae																																															1			19			31				
		Pedaliaceae																																	7																								
		Phrymaceae																																	31																								
	Scrophulariales	Scrophulariaceae																											32				42		3	17		12										3											
	Asterales	Asteraceae	3		6		1						3											1									1		15											13													
	Apiales	Apiaceae																															9		11																								

■ ≤1 ■ 2-5 ■ 6-10 ■ 11-40 ■ >40

Numbers are the average number of RT-EPRV sequences per genome of each cluster 60. The genomes are grouped according to the plant families. BADN, *Badnavirus*; CAUL, Caulimovirus; SL, Solendovirus; G, Gymnendovirus.

10.3389/fpls.2022.1011565

TABLE 3 Cluster 100 with 10 or more copies.

Cluster 100%	Num. Seq.	Genome	EPRV group
1	951	*Capsicum annuum*	Solendovirus-01
2	77	*Lotus japonicus*	Florendovirus-01
3	53	*Citrus maxima*	Petuvirus-01
4	43	*Hydrangea quercifolia*	Florendovirus-01
5	27	*Citrus medica*	Petuvirus-01
6	26	*Citrus medica*	Petuvirus-01
7	24	*Salvia splendens*	Florendovirus-03
8	22	*Ipomoea triloba*	Petuvirus-05
9	21	*Capsicum annuum*	Yendovirus-02
10	20	*Capsicum annuum*	Florendovirus-03
11	20	*Atalantia buxifolia*	Petuvirus-01
12	19	*Fortunella hindsii*	Florendovirus-02
13	19	*Atalantia buxifolia*	Petuvirus-01
14	16	*Helianthus annuus*	Wendovirus-02
15	16	*Ipomoea triloba*	Dioscovirus-01
16	14	*Fortunella hindsii*	Florendovirus-02
17	13	*Lactuca sativa*	Florendovirus-03
18	12	*Castanea dentata*	Florendovirus-01
19	12	*Atalantia buxifolia*	Florendovirus-02
20	12	*Nicotiana tabacum*	Solendovirus-01
21	12	*Lindenbergia philippensis*	Tungrovirus-02
22	11	*Lactuca sativa*	Caulimovirus-01
23	11	*Lotus japonicus*	Florendovirus-01
24	11	*Atalantia buxifolia*	Florendovirus-02
25	11	*Capsicum annuum*	Florendovirus-03
26	11	*Capsicum annuum*	Solendovirus-01
27	10	*Fragaria nilgerrensis*	Florendovirus-01
28	10	*Arachis hypogaea*	Florendovirus-01
29	10	*Nicotiana sylvestris*	Solendovirus-01
30	10	*Hordeum vulgare*	Yendovirus-01
31	10	*Rosa chinensis*	Zendovirus-01

RT sequence of the citrus blight associated virus is highly similar to the sequences of cluster100-3, -5 and -6, all of them belonging to genomes of the genus *Citrus*, and the sequence of the tobacco vein clearing virus is similar to clusters100-20 and -29, belonging to genomes of the genus *Nicotiana*.

Wendovirus, a new group of Caulimoviridae

Six of the Cluster60 and one of the Cluster100 correspond to a new group of endogenous *Caulimoviridae* with distinctive characteristics that, following the nomenclature proposed by Diop et al. (2018) (*Zendovirus*, *Xendovirus* and *Yendovirus*), we have called them *Wendovirus* (Supplementary Data 4 and Table 3).

We were able to reconstruct the structure of the *Wendovirus* for seven genomes corresponding to Cluster60 (Figure 3; Supplementary Data 6). The structure was very similar in all of them, with four partially overlapping ORFs. Comparisons with protein motif databases allowed us to find different conserved domains (Supplementary Data 6). The ORF1 encodes for a zinc finger motif, which is typical of the *Caulimoviridae* coat proteins. The ORF2 encodes for a movement protein and an aspartic proteinase. The ORF3 encodes a second aspartic proteinase, the RT and the RNAseH. Finally, the ORF4 encodes a protein without significant homologies to other reference proteins and without known protein domains but that is well-conserved in all the wendovirus elements. The most noticeable aspect of these structures is the presence of two aspartic proteinase domains instead of one, as usual. They are located close to each other, but in two different ORFs (2 and 3). In the case of the HelAnn-006 element (Wendovirus2 cluster), although the domains and their order are conserved, the ORF2 is shorter and the ORF3 is

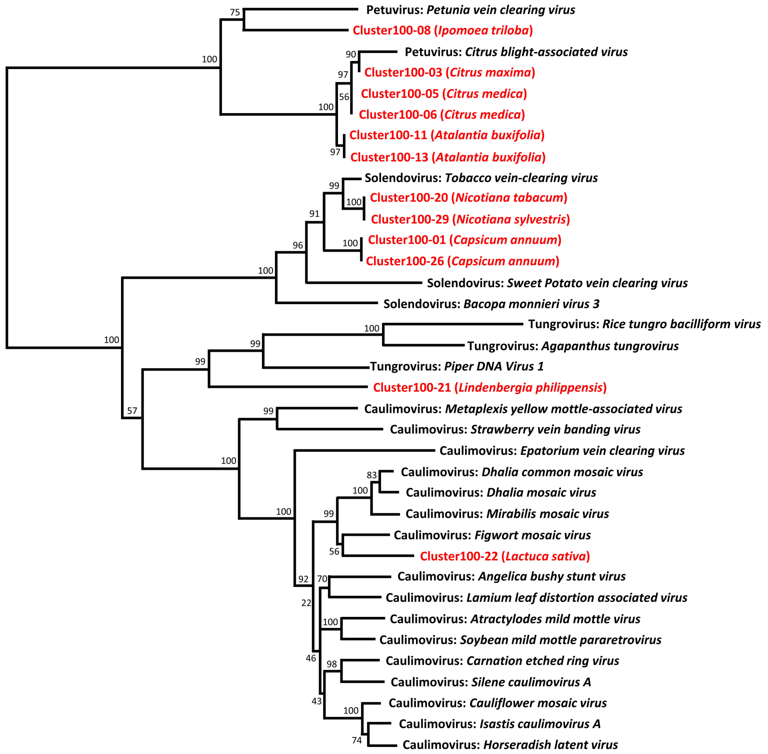

75
Petuvirus: *Petunia vein clearing virus*
Cluster100-08 (*Ipomoea triloba*)
100
90
Petuvirus: *Citrus blight-associated virus*
97
Cluster100-03 (*Citrus maxima*)
56
Cluster100-05 (*Citrus medica*)
100
Cluster100-06 (*Citrus medica*)
Cluster100-11 (*Atalantia buxifolia*)
97
Cluster100-13 (*Atalantia buxifolia*)
99
Solendovirus: *Tobacco vein-clearing virus*
100
Cluster100-20 (*Nicotiana tabacum*)
91
Cluster100-29 (*Nicotiana sylvestris*)
100
Cluster100-01 (*Capsicum annuum*)
96
Cluster100-26 (*Capsicum annuum*)
100
Solendovirus: *Sweet Potato vein clearing virus*
Solendovirus: *Bacopa monnieri virus 3*
100
Tungrovirus: *Rice tungro bacilliform virus*
100
99
Tungrovirus: *Agapanthus tungrovirus*
99
Tungrovirus: *Piper DNA Virus 1*
Cluster100-21 (*Lindenbergia philippensis*)
57
99
Caulimovirus: *Metaplexis yellow mottle-associated virus*
Caulimovirus: *Strawberry vein banding virus*
Caulimovirus: *Epatorium vein clearing virus*
100
83
Caulimovirus: *Dhalia common mosaic virus*
100
Caulimovirus: *Dhalia mosaic virus*
100
99
Caulimovirus: *Mirabilis mosaic virus*
Caulimovirus: *Figwort mosaic virus*
56
Cluster100-22 (*Lactuca sativa*)
92
70
Caulimovirus: *Angelica bushy stunt virus*
22
Caulimovirus: *Lamium leaf distortion associated virus*
100
Caulimovirus: *Atractylodes mild mottle virus*
Caulimovirus: *Soybean mild mottle pararetrovirus*
46
98
Caulimovirus: *Carnation etched ring virus*
Caulimovirus: *Silene caulimovirus A*
43
Caulimovirus: *Cauliflower mosaic virus*
100
Caulimovirus: *Isastis caulimovirus A*
74
Caulimovirus: *Horseradish latent virus*

FIGURE 2
Phylogenetic relationships of representative sequences of the Cluster100. Representative sequences of the RT-EPRV Cluster100 (in red) were aligned with RT sequences of pararetroviral elements (in black), and a phylogenetic tree was constructed using the NJ method and 1000 bootstrap replications.

divided in two. When compared to databases, the highest similarities of these two aspartic proteinase domains are with members of *Caulimoviridae*.

Discussion

Endogenous viral elements (EVEs) are viral sequences integrated in host genomes that are inherited as host DNA sequences (Holmes, 2011). Some of the EVEs, are derived from viruses in which integration into the genome is part of their replication cycle, for example, mammalian retroviruses. However, many viruses in which integration into the genomic DNA is not a part of their normal replication cycle can also be found as EVEs, as is the case of the endogenous *Caulimoviridae* (Endogenous Pararetrovirus, EPRVs). The presence of EPRVs has been described in the genomes of different plant species (Hohn et al., 2008). In this work we have focused on determining the presence of EPRV sequences relatively recently integrated, based on the selection of elements with complete and conserved RT domains.

Based on the RT domain sequence similarity we detected 11.527 sequences distributed in 57 clusters corresponding to 13 OTUs. Twelve of these groups had already been described (Diop et al., 2018) and one is shown here for first time, we called *Wendovirus*. Contrary to what has been observed in other plant viruses as *Geminivirus* or *Nanovirus* (Nino Barreat and Katzourakis, 2021), EVEs from *Caulimoviridae* are exclusively present in plants. Recently integrated RT-EPRVs are present in genomes of *Lycopodiopsida*, *Pinopsida*, *Liliopsida* and *Magnoliopsida*, but not necessary in all the genomes of these groups. For example, they are not present in the genomes of *Arabidopsis thaliana*, *Zea mays*, *Triticum aestivum*, *Phaseolus vulgaris*, *Theobroma cacao* or *Spinacia oleracea*. They are also absent in the *Selaginella moellendorffii* (*Marchantiophyta*) and in *Rhodophyta*, *Chlorophyta* or *Bryophyta*.

We have found that, in some cases, the integration events can be considered very recent. Once in the genome, the EPRV sequences begin to accumulate point random mutations, so, if the sequences are identical that means that they probably integrated recently in the genome. We have found multiple sequences encoding identical RT domains in different species being the most extreme case *Capsicum annuum* in whose genome we found up to 951 sequences encoding identical RT domains. Recent genome integrations of *Caulimoviridae* sequences have been described in some species, such as banana (Gayral et al., 2010). It is interesting to note that, in some cases, these identical RT sequences correspond to

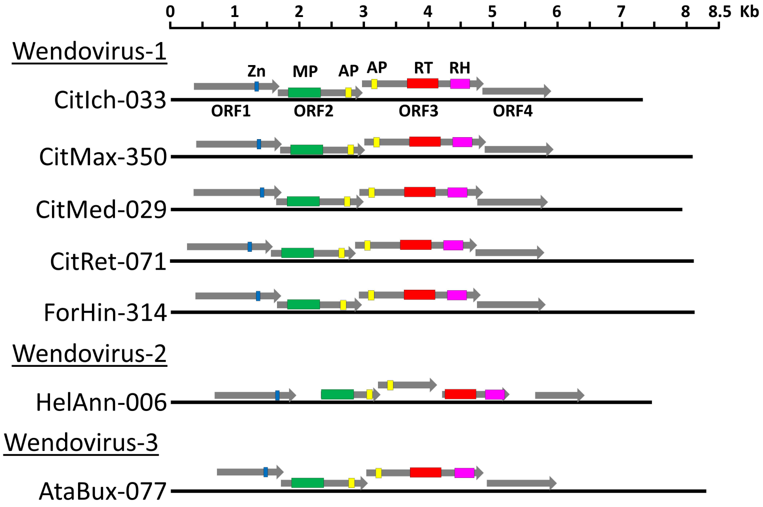


FIGURE 3
Schematic representation of *wendovirus* endogenous pararetrovirus. A scaled linear view of the genome organization of *Wendovirus*. The name of the sequences is the same as in Supplementary Data 4. Grey arrows mark open reading frames and colored regions within ORFs are conserved protein domains: blue, zinc finger typically present in the coat proteins; green, Movement Protein; yellow, Aspartic Proteinase; red, Reverse Transcriptase; pink, RNaseH.

groups that have only been detected as endogenous forms (*Florendovirus*, *Yendovirus*, *Zendovirus*, *Wendovirus*) suggesting that probably at least some of them may have their corresponding episomal virus species that have not been yet identified.

The distribution of the different clusters of EPRVs between species shows a great diversity. Some clusters are present exclusively in certain plants as, for example, *Gymnendovirus* in *Pinopsida*, *Zendovirus*1 in the tribus *Potentilleae* and *Roseae*, *Soymovirus*1 in the genus *Arachis* or *Wendovirus*1, only present in *Rutaceae*. In other cases, such as *Florendovirus*1 and 3, the distribution is very wide, including *Lilipsida* and *Magnoliopsida*. In general, the distribution of the different groups of EPRVs is consistent with the phylogeny, but not always. For example, *Petuvirus*2 are present in *Amborella trichopoda* and in eight *Magnoliopsida* orders, *Florendovirus*7 are present in *Amborella trichopoda* and in seven *Magnoliopsida* orders and *Solendovirus*1 are present in *Nymphaea colorata* and in *Solanaceae*. A possible explanation for these species distributions is the horizontal transmission of the virus between species. There are data suggesting multiple viral jumps between different animal species in *Hepadnavirus* (Dill et al., 2016), and previous data also suggests such horizontal transfers can occur for EPRVs in plants (Diop et al., 2018; Gong and Han, 2018).

We have detected differences in the number of EPRVs in the different genomes. Sometimes the differences are also observed comparing the genomes of species of the same genus or varieties of the same species. The number of EPRVs observed results from the combination of the virus integration and the mechanisms of amplification or reduction of the integrated sequences. First, *Caulimoviridae* integration requires the presence of viruses that are infectious for the species and that the defense mechanisms of the plant are not able to eliminate, or not completely. Second, the main integration mechanism is thought to involve illegitimate recombination, which requires the existence of DNA double-strand breaks and subsequent repair mechanisms (Richert-Pöggeler et al., 2021). Furthermore, to be transmitted, integration must occur in reproductive cells. Third, once integrated, EPRVs, copies are inactivated by sequence degeneration or fragmentation, or by the insertion of transposable elements, and subjected to epigenetic silencing (reviewed by Richert-Pöggeler et al., 2021). All these processes lead to the degeneration of the coding sequences. Finally, it has also been proposed that once integrated, the sequences can be amplified, and different mechanisms have been suggested such as transposition like retroelements, rolling circle amplification, unequal meiotic crossing-over of tandem arrays, or ectopic recombination between EPRV clusters on non-homologous chromosomes (reviewed by Richert-Pöggeler et al., 2021). Variations in any of these processes together with the time elapsed since the last event of integration could explain the observed differences in the number of EPRVs in the analyzed genomes. Nor can we rule out that the different quality of the genome assemblies may also affect.

We have identified a new putative genus of the *Caulimoviridae*, tentatively named '*Wendovirus*'. *Wendovirus* genomes are about 7,7 Kb long and are present in the genomes of different *Magnaliopsida* species, especially in *Rutaceae* and in sunflower. Our phylogenetic analysis shows that *wendovirus* are related to *Xendovirus* and *Soymovirus*. They contain four ORFs that encode the typical protein domains in *Caulimoviridae*: Zinc-Finger, Movement Protein, Aspartic Proteinase, Reverse Transcriptase and RNAseH. A remarkable feature of *wendovirus* is the presence of two protease coding domains located in two different ORFs (Figure 3). Although both encode aspartyl proteases, the domains are different (PF13975 in ORF2 and PF00077 in ORF3), so the hypothesis that their origin was a genomic duplication can be discarded. When compared to protein bases, all these described domains, including the two aspartic proteinase domains, show the greatest similarities against other members of *Caulimoviridae*. Therefore, it seems to be ruled out that the second proteinase domain could come from some other families of viruses. Recombination between EPRV fragments has been observed (Chabannes and Iskra-Caruana, 2013) and many viruses have modularly acquired domains and ORFs (Smyshlyaev et al., 2013; Koonin et al., 2015). Encapsidation of genomes (or genome fragments) of different species of *Caulimoviriridae* in the same capsid can lead to recombination and formation of chimeric genomes. Virus-like particles (VLPs) containing host RNAs were found to be produced during agroinfiltration of cucumber necrosis virus, some of them corresponding to retrotransposon or retrotransposon-like RNA sequences (Ghoshal et al., 2015). On the other hand, template switching between two RNA molecules during reverse transcription has been shown for retroviruses, LTR retrotransposons and is proposed for *Caulimoviridae* (Froissart et al., 2005; Tromas et al., 2014; Sanchez et al., 2017; Richert-Pöggeler et al., 2021). Such an acquisition of ORFs likely contributed to the evolution of the *Wendovirus*, although the possible functions of this second proteinase domain remain unknown.

Data availability statement

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

Author contributions

All authors contributed equally to the design and processing data. All authors have read and approved the final manuscript.

Funding

This work was funded by grants from the Spanish Ministerio de Economía, Industria y Competitividad (AGL2016-78992-R/FEDER), Ministerio de Ciencia e Innovación (PID2019-106374RB-I00/AEI/10.13039/501100011033), by "ERDF A way of making Europe" and by the CERCA Programme of the Generalitat de Catalunya. We also acknowledge financial support from the Spanish Ministerio de Economía y Competitividad through the "Severo Ochoa Programme for Centres of Excellence in R&D" 2016-2019 (SEV-2015-0533) and CEX2019-000902-S funded by MCIN/AEI/10.13039/501100011033.

Acknowledgments

We thank all groups members whose comments contributed to this study.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1011565/full#supplementary-material

SUPPLEMENTARY DATA SHEET 1
RT domain sequences of *Caulimoviridae*, retrovirus and LTR-retrotransposons used in this study.

SUPPLEMENTARY DATA SHEET 2
Caulimoviridae RT sequences used in the tBLASTn analyses.

SUPPLEMENTARY DATA SHEET 3
Genome assemblies used in the tBLASTn analyses.

SUPPLEMENTARY DATA SHEET 4
RT-EPRV sequences found.

SUPPLEMENTARY DATA SHEET 5
Representative RT-EPRV sequences for the clusters60 used in the phylogenetic analysis.

SUPPLEMENTARY DATA SHEET 6
Wendovirus sequences and polypeptides encoded.

References

Aiewsakun, P., and Katzourakis, A. (2015). Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* 479-480, 26–37. doi: 10.1016/j.virol.2015.02.011

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6, 11. doi: 10.1186/s13100-015-0041-9

Bejarano, E. R., Khashoggi, A., Witty, M., and Lichtenstein, C. (1996). Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proc. Natl. Acad. Sci. U.S.A.* 93, 759–764. doi: 10.1073/pnas.93.2.759

Bertsch, C., Beuve, M., Dolja, V. V., Wirth, M., Pelsy, F., Herrbach, E., et al. (2009). Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. *Biol. Direct* 4, 21. doi: 10.1186/1745-6150-4-21

Catlin, N. S., and Josephs, E. B. (2022). The important contribution of transposable elements to phenotypic variation and evolution. *Curr. Opin. Plant Biol.* 65, 102140. doi: 10.1016/j.pbi.2021.102140

Chabannes, M., and Iskra-Caruana, M. L. (2013). Endogenous pararetroviruses–a reservoir of virus infection in plants. *Curr. Opin. Virol.* 3, 615–620. doi: 10.1016/j.coviro.2013.08.012

Chen, S., and Kishima, Y. (2016). Endogenous pararetroviruses in rice genomes as a fossil record useful for the emerging field of palaeovirology. *Mol. Plant Path.* 17, 1317–1320. doi: 10.1111/mpp.12490

Chiba, S., Kondo, H., Tani, A., Saisho, D., Sakamoto, W., Kanematsu, S., et al. (2011). Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PloS Pathog.* 7, e1002146. doi: 10.1371/journal.ppat.1002146

Choi, I. S., Wojciechowski, M. F., Ruhlman, T. A., and Jansen, R. K. (2021). In and out: Evolution of viral sequences in the mitochondrial genomes of legumes (*Fabaceae*). *Mol. Phylog. Evol.* 17, 107236. doi: 10.1016/j.ympev.2021.107236

Dill, J. A., Camus, A. C., Leary, J. H., Di Giallonardo, F., Holmes, E. C., and Ng., T. F. F. (2016). Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses. *J. Virol.* 90, 7920–7933. doi: 10.1128/JVI.00832-16

Diop, S. I., Geering, A. D. W., Alfama-Depauw, F., Loaec, M., Teycheney, P. Y., and Maumus, F. (2018). Tracheophyte genomes keep track of the deep evolution of the caulimoviridae. *Sci. Rep.* 8, 572. doi: 10.1038/s41598-017-16399-x

Etienne, L. (2017). Paleovirology: looking back in time to better understand and control modern viral infections. *Virologie (Montrouge)* 21, 245–246. doi: 10.1684/vir.2017.0715

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* 13, 283–296. doi: 10.1038/nrg3199

Froissart, R., Roze, D., Uzest, M., Galibert, L., Blanc, S., and Michalakis, Y. (2005). Recombination every day: Abundant recombination in a virus during a single multi-cellular host infection. *PloS Biol.* 3, e89. doi: 10.1371/journal.pbio.0030089

Gayral, P., Blondin, L., Guidolin, O., Carreel, F., Hippolyte, I., Perrier, X., et al. (2010). Evolution of endogenous sequences of banana streak virus: What can we learn from banana (Musa sp.) evolution? *J. Virol.* 84, 7346–7359. doi: 10.1128/JVI.00401-10

Gayral, P., Noa-Carrazana, J. C., Lescot, M., Lheureux, F., Lockhart, B. E. L., Matsumoto, T., et al. (2008). A single banana streak virus integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *J. Virol.* 82, 6697–6710. doi: 10.1128/JVI.00212-08

Geering, A. D., Maumus, F., Copetti, D., Choisne, N., Zwickl, D. J., Zytnicki, M., et al. (2014). Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* 10, 5269. doi: 10.1038/ncomms6269

Ghoshal, K., Theilmann, J., Reade, R., Maghodia, A., and Rochon, D. (2015). Encapsidation of host RNAs by cucumber necrosis virus coat protein during both agroinfiltration and infection. *J. Virol.* 89, 10748–10761. doi: 10.1128/jvi.01466-15

Gong, Z., and Han, G. Z. (2018). Euphyllophyte paleoviruses illuminate hidden diversity and macroevolutionary mode of *Caulimoviridae*. *J. Virol.* 92, e02043-17. doi: 10.1128/JVI.02043-17

Hansen, C., and Heslop-Harrison, J. S. (2004). Sequences and phylogenies of plant pararetroviruses, viruses, and transposable elements. *Adv. Bot. Res.* 41, 165–193. doi: 10.1016/S0065-2296(04)41004-0

Hohn, T., Richert-Pöggeler, K. R., Staginnus, C., Harper, G., Schwarzacher, T., Teo, C. H., et al. (2008). ""Evolution of integrated plant viruses."," in *Plant virus evolution*. Ed. ,. M. J. Roossinck (Berlin: Springer), 53–81. doi: 10.1007/978-3-540-75763-4_4

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host Microbe* 10, 368–377. doi: 10.1016/j.chom.2011.09.002

Johnson, W. E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat. Rev. Microbiol.* 17, 355–370. doi: 10.1038/s41579-019-0189-2

Katzourakis, A., and Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PloS Genet.* 6, e1001191. doi: 10.1371/journal.pgen.1001191

Kojima, S., Yoshikawa, K., Ito, J., Nakagawa, S., Parrish, N. F., Horie, M., et al. (2021). Virus-like insertions with sequence signatures similar to those of endogenous nonretroviral RNA viruses in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2010758118. doi: 10.1073/pnas.2010758118

Koonin, E. V., Dolja, V. V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* 479-480, 2–25. doi: 10.1016/j.virol.2015.02.039

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Kuriyama, K., Tabara, M., Moriyama, H., Kanazawa, A., Koiwa, H., Takahashi, H., et al. (2020). Disturbance of floral colour pattern by activation of an endogenous pararetrovirus, petunia vein clearing virus, in aged petunia plants. *Plant J.* 103, 497–511. doi: 10.1111/tpj.14728

Lockhart, B. E., Menke, J., Dahal, G., and Olszewski, N. E. (2000). Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *J. Gen. Virol.* 81, 1579–1585. doi: 10.1099/0022-1317-81-6-1579

Nino Barreat, J. G., and Katzourakis, A. (2021). Paleovirology of the DNA viruses of eukaryotes. *Trends Microbiol.* 30, 281–292. doi: 10.1016/j.tim.2021.07.004

Ricciuti, E., Laboureau, N., Noumbissié, G., Chabannes, M., Sukhikh, N., Pooggin, M. M., et al. (2021). Extrachromosomal viral DNA produced by transcriptionally active endogenous viral elements in non-infected banana hybrids impedes quantitative PCR diagnostics of banana streak virus infections in banana hybrids. *J. Gen. Virol.* 102, 11. doi: 10.1099/jgv.0.001670

Richert-Pöggeler, K. R., Noreen, F., Schwarzacher, T., Harper, G., and Hohn, T. (2003). Induction of infectious petunia vein clearing (*pararetro*) virus from endogenous provirus in petunia. *EMBO J.* 22, 4836–4845. doi: 10.1093/emboj/cdg443

Richert-Pöggeler, K. R., Vijverberg, K., Alisawi, O., Chofong, G. N., Heslop-Harrison, J. S., and Schwarzacher, T. (2021). Participation of multifunctional RNA in replication, recombination and regulation of endogenous plant pararetroviruses (EPRVs). *Front. Plant Sci.* 21(12). doi: 10.3389/fpls.2021.689307

Sanchez, D. H., Gaubert, H., Drost, H. G., Zabet, N. R., and Paszkowski, J. (2017). High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. *Nat. Commun.* 8, 1374. doi: 10.1038/s41467-017-01374-x

Smyshlyaev, G., Voigt, F., Blinov, A., Barabas, O., and Novikova, O. (2013). Acquisition of an archaea-like ribonuclease h domain by plant L1 retrotransposons supports modular evolution. *Proc. Nat. Acad. Sci. U.S.A.* 110, 20140–20145. doi: 10.1073/pnas.1310958110

Tromas, N., Zwart, M. P., Poulain, M., and Elena, ,. S. F. (2014). Estimation of the *in vivo* recombination rate for a plant RNA virus. *J. Gen. Virol.* 95, 724–732. doi: 10.1099/vir.0.060822-0

Zhang, L., Richards, A., Barrasa, M. I., Hughes, S. H., Young, R. A., and Jaenisch, R. (2021). Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2105968118. doi: 10.1073/pnas.2105968118

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fpls.2022.1072232

Transposable elements are associated with genome-specific gene expression in bread wheat

Inbar Bariah, Liel Gribun and Khalil Kashkush*

Department of Life Sciences, Ben-Gurion University, Beer-Sheva, Israel

Introduction: Recent studies in wheat emphasized the importance of TEs, which occupy ~85% of the wheat genome, as a major source of intraspecific genetic variation due to their recent activity and involvement in genomic rearrangements. The contribution of TEs to structural and functional variations in bread wheat genes is not fully understood.

Methods: Here, publicly available RNA-Seq databases of bread wheat were integrated to identify TE insertions within gene bodies (exons\ introns) and assess the impact of TE insertions on gene expression variations of homoeologs gene groups. Overall, 70,818 homoeologs genes were analyzed: 55,170 genes appeared in each one of the three subgenomes (termed ABD), named triads; 12,640 genes appeared in two of the three subgenomes (in A and B only, termed AB; or in A and D only, termed AD; or in B and D only, termed BD);, named dyads; and 3,008 genes underwent duplication in one of the three subgenomes (two copies in: subgenome A, termed AABD; subgenome B, termed ABBD; or subgenome D, termed ABDD), named tetrads.

Results: To this end, we found that ~36% of the 70,818 genes contained at least one TE insertion within the gene body, mostly in triads. Analysis of 14,258 triads revealed that the presence of TE insertion in at least one of the triad genes (7,439 triads) was associated with balanced expression (similar expression levels) between the homoeolog genes. TE insertions within the exon or in the untranslated regions (UTRs) of one or more of the homoeologs in a triad were significantly associated with homoeolog expression bias. Furthermore, we found a statistically significant correlation between the presence\absence of TEs insertions belonging to six TE superfamilies and 17 TE subfamilies and the suppression of a single homoeolog gene. A significant association was observed between the presence of TE insertions from specific superfamilies and the expression of genes that are associated with biotic and abiotic stress responses.

Conclusion: Our data strongly indicate that TEs might play a prominent role in controlling gene expression in a genome-specific manner in bread wheat.

KEYWORDS

transposable elements, wheat, genome evolution, allopolyploidy, genome-specific, *Triticum aestivum*, gene expression, copy number variation

Check for updates

OPEN ACCESS

EDITED BY
Gennady I. Karlov,
All-Russia Research Institute of
Agricultural Biotechnology, Russia

REVIEWED BY
Zujun Yang,
University of Electronic Science and
Technology of China, China
Nikolai Borisjuk,
Huaiyin Normal University, China

*CORRESPONDENCE
Khalil Kashkush
kashkush@bgu.ac.il

SPECIALTY SECTION
This article was submitted to
Plant Systems and Synthetic Biology,
a section of the journal
Frontiers in Plant Science

RECEIVED 17 October 2022
ACCEPTED 20 December 2022
PUBLISHED 12 January 2023

CITATION
Bariah I, Gribun L and Kashkush K
(2023) Transposable elements are
associated with genome-specific gene
expression in bread wheat.
Front. Plant Sci. 13:1072232.
doi: 10.3389/fpls.2022.1072232

10.3389/fpls.2022.1072232

1 Introduction

Transposable elements (TEs) are a major component of plant genomes (Mhiri et al., 2022), e.g., they account for ~85% of the bread wheat genome (Appels et al., 2018; Wicker et al., 2018). Once thought of as "junk DNA" and "parasites", today, a growing body of evidence reveals that TEs have a prominent role in genome evolution (Avni et al., 2017; Bourque et al., 2018; Dubin et al., 2018). TEs are capable of moving and increasing their copy number within the host genome mainly through copy and paste (Class I, retrotransposons) or cut and paste (Class II) mechanisms (Wicker et al., 2007). The highly repetitive nature, high abundance, and activity of TEs might trigger massive structural genomic rearrangements (Gray, 2000; Bourque et al., 2018; Krasileva, 2019). They are considered a great source for genetic variation, mainly creating new alleles by transposing within gene bodies (Bourque et al., 2018; Dubin et al., 2018).

The high abundance of TEs near and within plant genes might impact the function of those genes by influencing both transcriptional and post-transcriptional levels and lead to the creation of novel transcripts (Schrader and Schmitz, 2019; Bariah et al., 2020; Zhang et al., 2021). The mere presence of TE, adjacent or within the transcribed region, might result in gene downregulation or silencing due to epigenetic modifications or interfering with enhancers or regulatory promoter elements (Dubin et al., 2018). Furthermore, TEs contain regulatory sequences such as promoters, transcription factors binding sits, and target sites for post-transcriptional degradation, which might affect adjacent gene expression or even modulate gene expression through complex transcriptional regulatory networks (Bourque et al., 2018; Dubin et al., 2018; Qiu and Köhler, 2020; Zhang et al., 2021). Additionally, the insertion of TE into a gene body might result in the creation of new isoforms through exonization, truncation, alternative splicing, or even by the domestication of TE-derived coding sequences into host genes, potentially altering the gene function (Keidar et al., 2018; Poretti et al., 2019; Crescente et al., 2022).

Wheat (*Triticum- Aegilops* group) is among the world's most widely grown crops, providing a significant portion of daily human caloric intake (Shewry and Hey, 2015; Levy and Feldman, 2022). The most widely grown bread wheat, *Triticum aestivum*, is a relatively new polyploid species that has been generated by two subsequent allopolyploidization events between members of two closely related genera, *Triticum* and *Aegilops* (Avni et al., 2017; Appels et al., 2018; Levy and Feldman, 2022). Allopolyploidization is the only mechanism that enables the formation of a new species in one step (Feldman and Levy, 2005). The rapid genomic structural and functional alterations accompanied with an allopolyploidization process have been intensively studied in recent years (Ramírez-González et al., 2018; Salina and Adonina, 2018; Fox et al., 2020; Juery et al., 2020). While currently there is still a debate regarding the extent of TE activity following allopolyploidization in wheat (amplification bursts vs. slow accumulation), there is no doubt that rapid TE turnover occurred during the wheat group evolution (Wicker et al., 2018; Bariah et al., 2020). The great contribution of TEs to genome plasticity might affect the ability of the new polyploid species to survive and rapidly adapt to various biological, environmental, and even cultivation environment stress (Van De Peer et al., 2017; Levy and Feldman, 2022).

The huge number of TE insertions adjacent to wheat genes led researchers to investigate the possible role of TEs in gene regulation (Wicker et al., 2018; Keidar-Friedman et al., 2020). Wicker et al. (2018) found no strong associations between specific TE families found near promoters and various expression modules. Additionally, a study by Ramírez-González et al., 2018 focused on the effect of TE insertions within triads (homeologs with a 1:1:1 correspondence across the three bread wheat sub-genomes- ABD) genes promoters and found no correlation between the presence of TEs in gene promoters and altered expression patterns between the three homeolog genes. However, Ramírez-González et al., 2018 did observe that higher TE abundance in the vicinity of the translation start site correlated with triads that showed more dynamic expression patterns across different tissues. This observation led Ramírez-González et al., 2018 to suggest a possible role for TEs in gene regulation as cis-regulatory elements or through other epigenetic mechanisms in a tissue-specific manner. Moreover, recent studies showed that TEs, specifically MITEs (Miniature Inverted repeat TEs), which are prevalent in the vicinity of wheat genes, might act as miRNAs precursors in wheat and thus can potentially shape regulatory gene networks (Poretti et al., 2019; Crescente et al., 2022). While the effect of TE insertions into promoter regions in wheat has been well-investigated, very little is known about the possible effect of TE insertions within gene bodies (Li et al., 2014; Xi et al., 2016; Keidar et al., 2018; Keidar-Friedman et al., 2018; Domb et al., 2019; Jiang et al., 2019). Here, the analysis of a large amount of publicly available databases in bread wheat facilitated the assessment of the functional impact of TE insertions within gene bodies in a genome-specific manner.

2 Methods

2.1 Identification of TE insertions within gene bodies

To identify TE insertions within gene bodies (exons\introns) in the *Chinese Spring* bread wheat cultivar (*CS42*), we integrated data from two publicly available databases (Appels et al., 2018; Juery et al., 2020). The name, homoeologous group IDs, assignment to one of the five chromosomal regions (R1, R2a, C, R2b, and R3), and the start and stop positions of 70,818 wheat genes belonging to 6,320 dyads (12,640 genes belongs to homoeologous groups that underwent elimination of a single gene), 18,390 triads (55,170 genes), and 752 tetrads (3,008 genes, belongs to homoeologous groups that underwent duplication of a single gene) were retrieved from Juery et al. (2020) and integrated with the IWGSC RefSeq v1.0 assembly coordinates for TEs (Appels et al., 2018) using python 3.7 (Guttag, 2021) scripts. Prior to the data integration, the IWGSC RefSeq v1.0 assembly annotations for TEs were organized using pandas, a Python package used for data analysis (Reback et al., 2021), and filtered to include only repeats defined as "repeat region" (nested repeats and repeat fragments were removed). Next, overlaps between repeat regions and each of the 70,818 genes were detected based on the genes and TEs coordinates and summarized in Supplementary Table S1. To compare the proportions of TE-containing genes between dyads, triads, and tetrads, the Chi-square test of independence of variables was performed using the *chi2_contingency* function from Python *SciPy* (RRID : SCR_008058). *Stats* module (Virtanen et al., 2020).

2.2 Polymorphic TE insertions within gene bodies

Following the identification of TE insertions within gene bodies and the characterization of TE insertions distribution, we wanted to assess the polymorphism(s) generated by TE insertions between the homoeologous copies in dyads, triads, and tetrads. For this, we used the *pandas* Python package (RRID : SCR_018214) (Reback et al., 2021) to organize the genes (see Supplementary Table S1) as homoeologous groups (Supplementary Table S2) according to the homoeologous group IDs and to sum the number of genes which contained one or more TE insertion within the gene body in each expressed homoeologous group (a group that includes one or more expressed gene, not expressed groups were removed from the analysis). For each of the homoeologous groups, we determined whether it was a polymorphic or monomorphic group. If all the homoeologs in a specific homoeologous group contained TE insertion (not depending on TE type or insertion location within the gene), the homoeologous group was considered as monomorphic. However, if one or more, but not all, of the homoeologs in the group contained TE insertion, the homoeologous group was considered polymorphic.

To test whether TE insertions were randomly distributed between the genes or rather tend to be more\ less polymorphic than expected, we focused only on homoeologous groups that included TE insertions in one or more of the homoeologs gene bodies (referred to as homoeologous groups that include TE insertions) and were determined to be expressed (include one or more expressed gene). Then, we performed the Chi-square Goodness of Fit Test separately for dyads, triads, and tetrads, to test whether the numbers of monomorphic and polymorphic homoeologous groups fit the expected numbers calculated based on the proportions of gene bodies that contain TE insertions. The expected number of monomorphic and polymorphic homoeologous groups was calculated assuming a binomial distribution of the presence \ absence of TE insertions within a gene body. First, the probability of a single gene including TE insertion was calculated based on the number of genes containing TE insertions and belonging to TE containing homoeologous group and the total number of genes belonging to TE containing homoeologous group. Then, the expected number of monomorphic and polymorphic homoeologous groups was calculated according to binomial distribution using the probability of a single gene including TE insertion and then divided by the probability of a homoeologous group to have at least one TE insertion (conditional probability) and multiplied by the number of groups containing one or more TE insertion. The observed numbers of polymorphic and monomorphic homoeologous groups were compared with the calculated expected numbers using the *chisquare* function from Python *SciPy* (RRID : SCR_008058). *Stats* module (Virtanen et al., 2020).

2.3 Correlation between polymorphic TE insertions within gene bodies and homoeolog expression bias

To assess the possible impact of TE insertional polymorphism within gene bodies on the relative gene expression in the homoeologous groups, we used summarized data "relative contribution category in brief" retrieved from Juery et al. (2020), on the assignment of each of the homoeologous groups to relative contribution categories. Based on this analysis, if all the homoeologs in a specific homoeologous group had similar relative abundance, the group was assigned to the balanced category, while groups in which different relative abundance was observed between the homoeologs were assigned to one of the non-balanced categories (Juery et al., 2020). Specifically, triads were assigned to the balanced category, homoeolog-suppressed category, or homoeolog-dominant category, dyads were assigned to a balanced category or homoeolog-suppressed category, and tetrads were assigned to either one of the following categories: balanced category, tetrads with one suppressed copy, tetrads with two suppressed copies, and tetrads with one dominant copy (Juery et al., 2020). Additionally, some of the homoeologous groups were referred to as not expressed and thus were excluded from further analysis. The assignment of the homoeologous groups to relative contribution categories was performed according to the calculation method described by Ramírez-González et al. (2018) and based on the same RNA-seq data used by Ramírez-González et al. (2018) for 123 samples of bread wheat (*Chinese Spring*) taken from 15 different tissues under non-stress conditions (Juery et al., 2020). The dependency between polymorphism and balanced\ non-balanced expression of the homoeologs was tested using the Chi-square test of independence of variables with the *chi2_contingency* function from Python *SciPy. Stats* module (Virtanen et al., 2020).

2.4 Correlation between TE insertion within gene bodies and homoeolog expression bias in triads

Here we used data on the relative expression abundance of the homoeologs in each of the triads (Ramírez-González et al., 2018) and the IWGSC RefSeq v1.0 assembly annotations for genes and TEs (Appels et al., 2018) to assess the possible impact of TE insertions on gene expression variations of homoeologous groups. We used 55,422 genes that had a 1:1:1 correspondence across the three homoeologous subgenomes (A, B, and D) of bread wheat (18,474 homoeolog triads) from Ramírez-González et al. (2018) and identified TE insertion within each of the genes bodies as described in Supplementary Table S1 (see Supplementary Table S3). For each of the triads, the TE classification (superfamily and subfamily) was determined for elements that were found to be inserted within the gene bodies of the genes in the triad (Supplementary Tables S4, S5). TE subfamily names were according to the ClariTeRep naming system (Wicker et al., 2018), in which the three first letters of the subfamily name represent the TE superfamily, and the number at the end of the name represents the family and in some cases is followed by a dot and a number, which represents specific subfamily within the TE family.

In addition to the identification of TE insertions within gene bodies, we identified TE insertions found specifically within exons and within the UTRs using a similar approach, combining the exons, 5’ UTRs, and 3’ UTRs coordinates for each gene according to IWGSC RefSeq v1.0 HC genes annotations with TEs coordinates (Appels et al., 2018). Then, we integrated data from files dividing the homoeolog triads into seven relative contribution categories (Ramírez-González et al., 2018), to create Supplementary Table S3.

The seven files divided the triads into contribution categories as follows: triads for which a similar abundance of transcripts was observed from each of the three homoeologs were assigned to a balanced category, while triads that showed a higher or lower abundance of transcripts from a single homoeolog relative to the other two, were assigned to one of six non-balanced categories. The non-balanced categories include three homoeolog-dominant categories (A dominant, B dominant, and D dominant) and three homoeolog-suppressed categories (A suppressed, B suppressed, and D suppressed) (Ramírez-González et al., 2018). Each triad was attributed to one of the above categories based on ternary diagrams representing the relative expression of each homoeolog and by comparison to the ideal normalized expression bias for the seven categories as described by Ramírez-González et al., (Ramírez-González et al., 2018). The analysis was performed for RNA-seq data from several different studies (Ramírez-González et al., 2018) (a total of 850 wheat RNA-sequencing samples), which were organized into partly overlapping datasets. Here we focused on data generated from 123 RNA-Seq samples of bread wheat (*Chinese Spring*) (Ramírez-González et al., 2018). The 123 samples were derived from 15 different tissues under non-stress conditions. For our analysis, we focused on 14,258 triads which were found to be syntenic and expressed in at least 6 out of 15 tissues tested for this dataset (see Supplementary Table S3) (Ramírez-González et al., 2018).

For the following analysis, the three homoeolog genes in each one of the 14,258 triads were combined, meaning that a triad was referred to as a triad that included TE insertions if one or more TE insertions were found within the sequence of at least one of the triad genes. The dependency between the presence\ absence of TE insertions within the gene bodies, exons, or UTRs and different contribution categories was analyzed using the Chi-square test of independence of variables with the *chi2_contingency* function from Python *SciPy. Stats* module (Virtanen et al., 2020). The dependency between the TE superfamilies\ subfamilies from which insertions were present\ absent in at least one of the triad genes and the different contribution categories was analyzed using the Chi-square test of independence of variables with the *chi2_contingency* function from Python *SciPy. Stats* module and corrected for multiple testing using the *multipletests* function from the Python *statsmodels* module (RRID : SCR_016074) with the Benjamini/ Hochberg Procedure (non-negative) (Virtanen et al., 2020).

2.5 Gene ontology enrichment analysis

Gene ontology (GO) provides structured, computable knowledge regarding the functions of genes and gene products in three non-overlapping domains of molecular biology (Carbon et al., 2019). The three domains are Biological Process (BP), which refers to a biological objective to which the gene or gene product contributes, Molecular Function (MF), defined as the biochemical activity of a gene product and Cellular Component (CC), which refers to the location in the cell where a gene product is active (Ashburner et al., 2000). GO enrichment analysis is used to find over-represented GO terms in a gene set compared to a reference set.

Here, we performed GO enrichment analysis for triads, including TE insertions from each of the 14 TE superfamilies (see Table 1). Additionally, we selected triads that belonged to

TABLE 1 Analysis of the 7,439 expressed and syntenic triads which contained TE insertions belonging to 14 different TE superfamilies.

					GO classes[3]		
Code[1]	Class	Order	Superfamily	Triads[2]	BP	CC	MF
RLG	Class I (retrotransposons)	LTR	*Gypsy*	1,134	438	70	148
RLC			*Copia*	1,359	371	67	157
RLX			Unclassified LTR-retrotransposons	474	132	15	84
RIX		non-LTR (LINE)	Long interspersed nuclear elements	1,458	416	97	209
SIX		non-LTR (SINE)	Short interspersed nuclear elements	20	19	11	23
DTC	Class II (DNA transposons)	TIR	*CACTA*	2,960	734	140	327
DTM			*Mutator*	428	185	11	80
DTX			unknown	2,099	606	89	248
DTH			*Harbinger*	456	142	21	94
DTT			*Mariner*	4,576	956	156	400
DTA			*hAT*	9	7	1	4
DXX		unknown	unknown	105	68	15	22
DHH		Helitron	*Helitron*	2	–	–	–
XXX	unknown	unknown	unknown	1,347	266	16	225

[1] The three letters code represents the class (first letter), order (second letter) and superfamily (third letter) of the TE (Wicker et al., 2007).

[2] Number of triads in which at least one of the genes includes TE insertion from the specific superfamily. Note that the sum of the triad column is larger than 7,439. This is since some triads include insertions from more than one subfamily.

[3] Number of significantly enriched GO terms found in GO SEA preformed for triads which include TE insertions from mentioned superfamily in each of the three biological objective to which the gene or gene product contributes: BP, Biological Process, CC, Cellular Component and MF, Molecular Function."-" notes missing values due to short query list which did not met the criteria for enrichment analysis.

specific relative contribution categories and included TE insertions from superfamilies that showed a correlation to the mentioned category (see Table 2). The reference set for all the GO enrichment analyses performed in this study was the whole set of 14,258 expressed and syntenic triads. GO Singular Enrichment Analysis (SEA) was performed using the AgriGO toolkit (RRID : SCR_006989) (Tian et al., 2017) with Fisher's exact test to identify enriched GO terms.

Following the GO SEA, enriched GO terms for each GO category (Biological function, Cellular component, and Molecular function) were visualized as a scatter plot generated by REVIGO (RRID : SCR_005825) (Supek et al., 2011). REVIGO summarizes the GO terms lists generated from the GO SEA by reducing functional redundancies based on the value provided and visualizes the remaining GO terms as a scatterplot, where more semantically similar GO terms are found closer to each other in the plot. For each GO SEA, we provided REVIGO, a list of GO terms that were found to be significantly enriched with false discovery rate (FDR) less or equal to 0.05 and their FDR value which is an adjusted p-value that enables us to have less false positive results then if the p-value was used. The scatterplots generated by REVIGO were imported into R, where wanted labels were added, and others were moved manually to slightly different coordinates to better visualize all the labels.

3 Results

3.1 Different TE insertion patterns within gene bodies in dyads, triads, and tetrads

In order to perform a genome-wide analysis of TE insertions within wheat gene bodies, 70,818 bread wheat genes belonging to 6,320 dyads, 18,390 triads, and 752 tetrads were analyzed. TE insertions within gene bodies were identified based on the IWGSC RefSeq v1.0 assembly coordinates for HC genes and TEs (Appels et al., 2018). We found that ~36% of the 70,818 genes (25,811 genes) contain at least one TE insertion within the gene body, with higher proportions of TE containing genes observed for triads (20,975 genes, 38.02%) relative to dyads (3,972 genes, 31.42%) and tetrads (864 genes, 28.72%) (Figure 1A; Supplementary Figure S1). The difference in the proportions of TE containing genes between the dyads, triads, and tetrads genes was statistically significant (χ^2 = 273.99, p < 0.001). TE insertions were found either in all the homoeologous copies in the group (i.e., for triads: monomorphic insertion in the three sub-genomes) or only in some of the homoeolog genes (i.e., for triads: polymorphic insertion in the three sub-genomes).

The differences in TE abundance between the dyads, triads, and tetrads categories might be the result of the different

TABLE 2 TE superfamilies for which the presence\absence of TE insertion in at least one of the triad genes correlated with specific triad expression patterns.

		TE insertion			
TE superfamily[1]	corrected p-values[2]	Yes	No	Yes	No
		Balanced triads[3]		Non-balanced triads[4]	
RLC	0.008837323	1132	5252	227	828
DTT	6.26E-06	4001	2383	575	480
DTM	0.000236883	338	6046	90	965
DTX	0.01607886	1837	4547	262	793
SIX	0.008282664	12	6372	8	1047
RLX	0.00382206	382	6002	92	963
XXX	0.01015199	1123	5261	224	831
		Suppressed triads[5]		**Not suppressed triads[6]**	
RLC	0.014455	192	690	1167	5390
DTT	0.000675	488	394	4088	2469
DTM	0.023678	68	814	360	6197
DTX	0.047574	220	662	1879	4678
RLX	0.000675	83	799	391	6166
XXX	0.006588	194	688	1153	5404
		Dominant triads[7]		**Not dominant triads[8]**	
DTT	0.013897	87	86	4489	2777
DTM	0.001364	22	151	406	6860

[1] The three letters code represents the class (first letter), order (secuned letter) and superfamily (third letter) of the TE (Wicker et al., 2007).
[2] χ^2 corrected p-values for multiple tests using Benjamini/Hochberg Procedure (non-negative).
[3,5,7] Number of triads belonging to the mentioned category (balanced, suppressed, or dominant), in which at least one homoeolog contains TE insertion from the mention superfamily (Yes) or none of the homoeologs contain TE insertion from the mention superfamily (No).
[4,6,8] Number of triads that does not belong to the mentioned category, in which at least one homoeolog contains TE insertion from the mention superfamily (Yes) or none of the homoeologs contain TE insertion from the mention superfamily (No). For the balanced category it will refer to the number of triads from the homoeolog-dominant or homoeolog-suppressed categories, for the suppressed categories it will refer to the number of triads from balanced or to one of the homoeolog-dominant categories and for the dominant categories it will include triads belong to either the balanced or to one of the homoeolog- suppressed categories.

chromosomal distribution patterns of the genes between the categories. While triads are more abundant in the proximal region (R2a, C and R2b), which contains a higher proportion of TEs, dyads, and tetrads are most abundant in the distal region (R1 and R3), which was found to have lower TE density (Wicker et al., 2018; Juery et al., 2020). The proportions of TEs containing genes belonging to each of the three categories in each of the five chromosomal regions are shown in Figure 1A. To test whether the difference in TEs abundant between the dyads, triads, and tetrads is mainly due to the chromosomal distribution of the genes, we performed the analysis separately for the proximal and distal regions. Significant differences in TEs abundant within gene bodies from dyads, triads, and tetrads were observed for each region separately, displaying the same pattern observed for the whole genome. Out of the 41,503 genes (4,781 belonging to dyads, 35,447 to triads, and 1,275 to tetrads) found in the proximal region, 40.29% included TE insertion within the gene body, with significant differences in proportion (χ^2 = 135.78, $p < 0.001$) between dyads (1,665 genes, 34.83%), triads (14,675 genes, 41.40%), and tetrads (380 genes, 29.80%) genes. A lower proportion of TE containing genes was observed in the distal regions, where only 31.01% of the 29,315 genes (7,859 belonging to dyads, 19,723 to triads, and 1,733 to tetrads) included TE insertion. However, the significant differences in proportion (χ^2 = 25.77, $p < 0.001$) of TE containing genes were still observed in the distal region, with a higher proportion of TE insertion in triad genes (6,300 genes, 31.94%) relative to dyads (2,307 genes, 29.35%) and tetrads (484 genes, 27.93%).

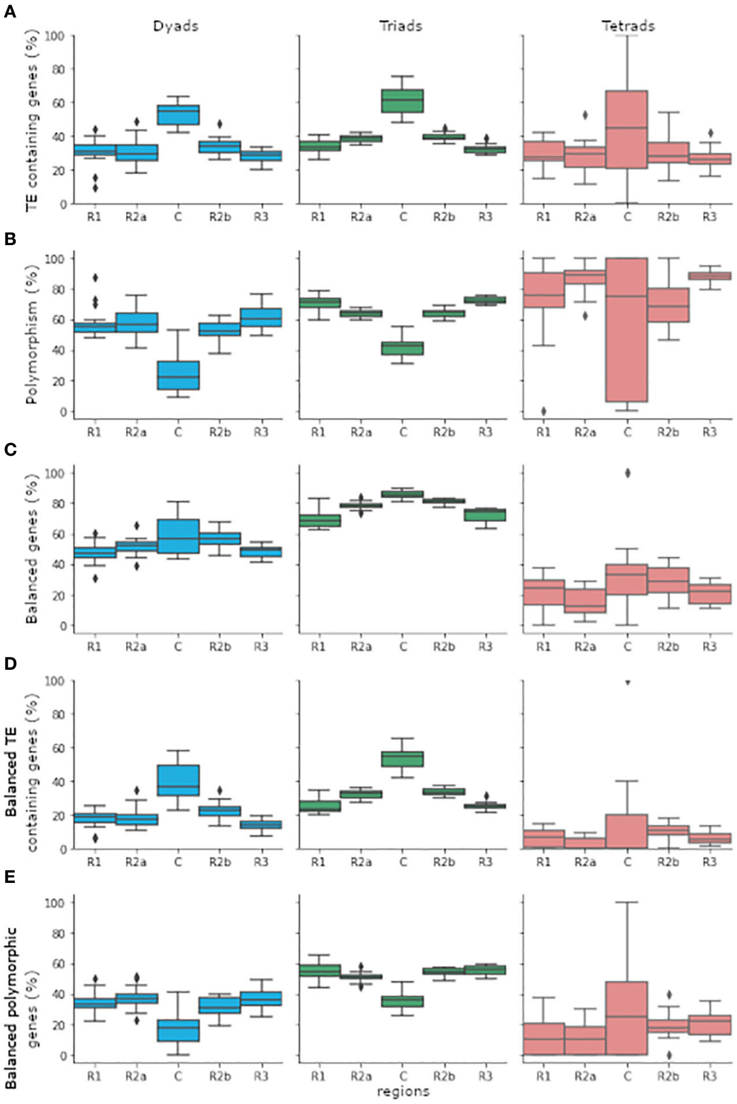


FIGURE 1
The distribution of genes belonging to the dyads, triads, or tetrads categories in the five chromosomal regions (R1, R2a, C, R2b, and R3). Gene distribution was calculated individually for genes that belong to the same category and found in the same and chromosomal region in each one of 18 bread wheat chromosomes. Genes found in chromosomes 1A, 1B, and 1D were eliminated from the analysis due to the lack of tetrads genes in the centromeres of chromosomes 1A and 1B. **(A)** Percentage of TE containing genes out of total genes. **(B)** Percentage of genes belonging to polymorphic group out of genes that found in TE containing group. **(C)** Percentage of genes belonging to balanced group out of total genes. **(D)** Percentage of genes belonging to balanced group out of genes that found in TE containing group. **(E)** Percentage of genes belonging to balanced and polymorphic group out of genes that found in TE containing group. The boxplots depict the first quartile (Q1) and the third quartile (Q3) of the data with the median between them. The whiskers extend from the box to 1.5x the interquartile range (IQR). Rhombuses represent values found past the end of the whiskers. boxplots were drawn using the *boxplot* function from the seaborn python package (Waskom, 2021).

3.2 Polymorphic TE insertions within gene bodies and homoeologous group expression patterns

To assess the associations between TE insertion patterns and gene expression in dyads, triads, and tetrads, we first grouped the homoeologs copies from each homoeologous group and determined for each expressed group (a group that includes one or more expressed gene) whether it is monomorphic or polymorphic. Homoeologous group was considered as monomorphic if all the homoeologs copies in the group included at least one TE insertion or polymorphic if at least one but not all the homoeologs copies in the group included TE insertion within the gene body (Figure 1B, Supplementary Table S2). Then, the number of monomorphic and polymorphic homoeologous groups was compared with the numbers predicted by our module, which was based on the assumption that the presence\ absence of TE insertions within a gene is random. Out of 5,059 expressed dyads, 47.16% (2,386 dyads) included TE insertions, meaning at least one TE insertion was present within one or more of the homoeologs gene bodies. Focusing only on the 2,386 dyads that include TE insertions, we found that 54.99% (1,312 dyads) were polymorphic and included TE insertions in only one of the homoeologs gene bodies, a result that did not fit our module (χ^2 = 136.78, p < 0.001) which predicted that only 43.13% (~1,029 dyads) would be polymorphic. While the percentage of polymorphic dyads was higher than our module anticipated, the opposite trend was observed for triads and tetrads. Out of 17,676 expressed triads, 59.24% (10,471 triads) included TE insertions, with 64.08% of the triads identified as polymorphic (6,710 triads), a distribution that did not match our module (χ^2 = 212.36, p < 0.001), which predict that 70.57% of the triads (~7390 triads) will be polymorphic. For tetrads, we found that out of the 667 expressed tetrads, 58.17% (388 tetrads) included TE insertions, and 82.73% (321 tetrads) of the TE insertions containing tetrads were found to be polymorphic. This result also did not fit our module (χ^2 = 43.47, p < 0.001), which predicts that 91.88% of the tetrads (~356 tetrads) will be polymorphic.

To reduce the effect of the different chromosomal distribution patterns of dyads, triads, and tetrads on our analysis, we reperformed the analysis focusing on homoeologous groups in which all the homoeologs copies were found in the same chromosomal region (proximal\ distal). Thus, the numbers of monomorphic and polymorphic homoeologous groups and the predicted distribution were counted and calculated separately for dyads, triads, and tetrads in each chromosomal region. Similar patterns to those observed for the whole chromosome were observed separately for the proximal and distal regions. Out of the 2,386 expressed dyads that include TE insertions, 872 included only genes found in the proximal region, and 1,288 included only genes found in the distal region. The numbers of polymorphic dyads in both the proximal and distal regions were higher than expected by our module. They did not match our predictions (proximal region: χ^2 = 33.11, p < 0.001, distal region: χ^2 = 95.11, p < 0.001), with 48.62% polymorphic dyads (424 dyads) at the proximal region and 59.24% polymorphic dyads (763) at the distal region. In contrast, our module predicted that 39.19% of the dyads found in the proximal region (~341 dyads) and 45.70% of the dyads in the distal region would be polymorphic. For triads, we found that out of the 10,471 expressed TE containing groups, 6,654 triads included only genes located in the proximal region and 2,901 triads included only genes located in the distal region. The number of polymorphic triads was lower than the number predicted by our module (proximal region: χ^2 = 154.72, p < 0.001, distal region: χ^2 = 28.96, p < 0.001), with 59.78% polymorphic triads (3,978 triads) at the proximal region and 72.84% polymorphic triads (2,113 triads) in the distal region, versus 66.96% in the proximal region (~4455 triads) and 77.04% in the distal region (~2,235 triads) predicted by our module. The distribution of polymorphic tetrads also did not fit the numbers predicted by our module (proximal region: χ^2 = 25.53, p < 0.001, distal region: χ2 = 16.65, p < 0.001). Out of the 388 expressed tetrads which include TE insertions, 132 tetrads included only genes located in the proximal region, and 79.55% (105 tetrads) of them were found to be polymorphic, while our module predicted that 91.69% of the tetrads (~121 tetrads) in the proximal region would be polymorphic. Moreover, out of the 181 TE containing expressed tetrads that included only genes located in the distal region, only 83.43% were found to be polymorphic, while our module predicted that 91.76% of the tetrads in the distal region (~166 tetrads) would be polymorphic.

Next, we aimed to assess whether polymorphic TE insertions affect the relative expression within the homoeologous group. The data on the relative expression within each homoeologous group was retrieved from Juery et al. (2020), which performed the analysis for 123 RNA-Seq samples of bread wheat (*Chinese Spring*) taken from 15 different tissues under non-stress conditions, and was integrated into Supplementary Table S2. Expressed homoeologous groups were assigned as balanced if all the homoeologs showed similar transcript abundance or as homoeolog-dominant or homoeolog-suppressed (non-balanced), based on the relative higher\ lower transcript abundance of each homoeolog (Figure 1C) (Juery et al., 2020). To learn about the possible effect of TE insertions within gene bodies on the relative expression, we focused only on TE containing homoeologous groups and tested the correlation between polymorphism and expression patterns separately for dyads, triads, and tetrads (Figures 1B, D, E). Our analysis revealed that a higher percentage of the polymorphic homoeologous groups belonged to one of the non-balanced expression categories relative to monomorphic groups, with a significant difference in proportions for dyads and triads (dyads: χ^2 = 11.34, p < 0.001, triads: χ^2 = 73.45, p < 0.001), while for tetrads the differences in proportions were not statistically significant (χ^2 = 0.32, p = 0.57). However, different results were obtained when the analysis was performed separately for the proximal and the distal chromosomal regions. For dyads, 37.88% (497 dyads) of the 1,312 polymorphic groups showed non-balanced expression, while only 31.19% (335 dyads) of the 1,074 monomorphic groups showed non-balanced expression. A statistically significant correlation (χ^2 = 9.96, p = 0.002 < 0.05) was also identified between polymorphism and relative expression pattern for dyads that contained only genes found in the proximal region (872 dyads), with 32.55% of the polymorphic dyads (138 dyads) and 22.77% of the monomorphic dyads (102 dyads) found in one of the non-balanced categories. However, no significant correlation was found between polymorphism and relative expression patterns for dyads which include only genes located in the distal region (χ^2 = 0.05, p = 0.82 > 0.05), although higher proportions of polymorphic dyads were found in non-balanced categories (37.75%, 288 dyads) relative to monomorphic dyads (36.95%, 194 dyads). Similar to dyads, for triads, 18.21% (1,222 triads) of the 6,710 polymorphic groups were classified as non-balanced, compared to 11.81% (444 triads) of the 3,761 monomorphic triads that were classified as non-balanced. The correlation between non-balanced expression and polymorphism in triads was also observed separately for triads that include only genes located in the proximal region (χ^2 = 52.98, p < 0.001) and for triads that include only genes located at the distal region (χ^2 = 5.54, p = 0.02 < 0.05). At the proximal region, 16.37% of the polymorphic triads (651 triads) and 10.05% of the monomorphic triads (269 triads) were classified as non-balanced, and at the distal region, 21.15% of the polymorphic triads (447 triads) and 17.13% of the monomorphic triads (135 triads) were classified as non-balanced. Finally, for tetrads, 77.26% (248 tetrads) of the 321 polymorphic groups and 73.13% (49 tetrads) of the 67 monomorphic groups were assigned to one of the non-balanced categories. No statistically significant dependency between polymorphism and homoeologous group expression pattern was identified upon performing the analysis separately for tetrads which include genes located only at the proximal (χ^2 = 0.43, p = 0.51 > 0.05) or only at the distal (χ^2 = 0.02, p = 0.89 > 0.05) chromosome regions.

3.3 TE content within gene bodies and triad expression patterns

Here, we aimed to study the possible effect of TE insertions on gene expression in wheat. We analyzed 14,258 expressed and syntenic triads that were assigned to 7 relative contribution categories according to the calculation method described by Ramírez-González et al. (2018). Most of the 14,258 triads (11,834 triads, 83%) showed balanced expression, meaning a similar relative abundance of transcripts was observed for the three homoeologs. The remaining 2,424 triads were divided between 6 non-balanced categories, with 13.99% of the triads (1,995 triads) assigned to one of the homoeolog-suppressed relative contribution categories (5.16% of the triads belonged to the A suppressed category, 5.31% to the B suppressed category and 3.52% to the D suppressed category) and 3.01% of the triads (429 triads) assigned to one of the homoeolog-dominant relative contribution categories (0.90% of the triads belonged to the A dominant category, 1.05% to the B dominant category and 1.07% to the D dominant category). TE insertions within the gene bodies of triads genes were identified based on the IWGSC RefSeq v1.0 assembly coordinates for high confidence (HC) genes and TEs (Appels et al., 2018).

The analysis of the 14,258 expressed and syntenic triads revealed that the presence of TE insertions in at least one of the triad genes (7,439 triads) correlated to balanced expression between the homoeolog genes. Out of the 14,258 expressed and syntenic homoeolog triads, 52.17% (7,439 triads) contain one or more TE insertions (based on repeat regions coordinates) within the gene body sequence of at least one of the genes in the triad (triads that include TE insertions). A higher proportion of triads that include TE insertions are found in the balanced expression category (6,384, 85.82%) relative to triads that don't include TE insertions (5,450, 79.92%) with a statistically significant difference in proportions (χ^2 = 87.18, p < 0.001).

The TEs that were found to be within gene bodies represented all the 14 TE superfamilies identified in the wheat genome (see Table 1) and belonged to 455 subfamilies out of the 570 subfamilies annotated by the IWGSC as "repeat region", as was counted from the annotation file (Appels et al., 2018). To learn about the possible association between TE type and the relative expression contribution of each of the homoeologs in the triad, we tested separately for each TE superfamily and subfamily whether the presence\ absence of TE insertions from said type within gene bodies correlated with balanced, suppressed, or dominant relative expression of the homoeologs. Here, we focused only on TE groups (superfamily or subfamily) that had sufficient sample size, mining 5 or more cases were observed for all the combinations of the tested conditions for the group with the examined relative expression category. For instance, the number of triads in which TE insertions from specific TE superfamily were presence\ absent must be five or higher both in balanced and non-balanced categories for the superfamily to be included in the analysis against the balanced relative expression category. Out of the 14 TEs superfamilies, 12 were found adequate for analysis against the balanced expression category (DTA and DHH were removed from the analysis), and 7 TEs superfamilies showed a statistically significant correlation (Chi-square corrected p-value ≤ 0.05, Table 2) with balanced\ non-balanced expression categories. The correlation between superfamily and balanced expression was negative for 5 (SIX, DTM, RLX, RLC, and XXX, Table 2) of the 7 superfamilies and positive for the remaining 2 superfamilies (DTT and DTX). The same 12 superfamilies that were found adequate for analysis against the balanced expression category were also found adequate for comparison against homoeolog-suppressed\ non-suppressed expression categories, with the remaining 2 superfamilies (DTA and DHH, Table 2) excluded from the analysis due to a low number of cases. Specific superfamilies also showed a statistically significant correlation with homoeolog-suppressed\ non-suppressed expression categories. In total, 6 superfamilies showed statistically significant correlation with homoeolog-suppressed\ non-suppressed expression categories (Chi-square corrected p-value ≤ 0.05, Table 2), all of them also showed correlation with balanced\ non-balanced expression categories. TEs superfamilies that showed a positive correlation with balanced expression showed a negative correlation with suppressed expression (DTT and DTX, Table 2), while TEs superfamilies that showed a negative correlation with balanced expression showed a positive correlation with suppressed expression (DTM, RLX, RLC and XXX, Table 2). Finally, only 10 of the 14 superfamilies were found fitted for analysis against the homoeolog-dominant\ non-dominant expression categories (SIX, DTA, DXX, and DHH were removed from the analysis), with only 2 TE superfamilies (DTM and DTT) showing statistically significant correlation with homoeolog-dominant\ non-dominant expression categories (Chi-square corrected p-value ≤ 0.05, Table 2), both also found to correlate with balanced expression significantly. The DTM superfamily showed a positive correlation with dominant expression and a negative correlation with balanced expression, while the DTT superfamily showed a negative correlation with dominant expression and a positive correlation with balanced expression.

Next, we performed a similar analysis for TE subfamilies. The majority of the 455 TE subfamilies found within gene bodies were excluded from the analysis due to the small sample size: out of the 455 TE subfamilies, 303 subfamilies were eliminated from the analysis for the balanced expression category, 323 subfamilies were excluded from the analysis for the suppressed expression categories, and 433 subfamilies were excluded from the analysis for the dominant expression categories. Out of the TEs subfamilies which were found adequate for analysis, 19 subfamilies showed a statistically significant correlation (Chi-square corrected p-value ≤ 0.05, Table 3) with balanced\ non-balanced expression categories, 17 subfamilies showed a statistically significant correlation (Chi-square corrected p-value ≤ 0.05, Table 4) with homoeolog-suppressed\ non-suppressed expression categories and none of the TE subfamilies showed statistically significant correlation with homoeolog-dominant\ non-dominant expression categories. Fourteen of the subfamilies that showed a significant correlation between presence\absence of TE insertions and suppression of a single homoeolog gene also showed a correlation with balanced relative expression of the homoeologs, while the other subfamilies were found in correlation only to suppressed (3 subfamilies) or balanced (5 subfamilies) relative expression. Of the 17 subfamilies that showed statistically significant correlation with homoeolog-suppressed expression categories, only the *DTT_famn14* subfamily showed a negative correlation with homoeolog-suppressed expression, while insertions of the remaining 16 subfamilies appeared in higher proportions than expected in the homoeolog-suppressed categories. Similarly, 17 of the 19 subfamilies that showed a statistically significant correlation with homoeolog-balanced expression showed a negative correlation with balanced expression, and only two subfamilies, *RIX_famc8* and *DTT_famn14*, showed a positive correlation with homoeolog-balanced expression.

3.4 Triads which include TE insertions belonging to specific TE superfamilies were associated with various GO terms

To assess the association between the presence of TE insertions from specific types within gene bodies and gene function, we tested whether triads that include TE insertions from each of the 14 TEs superfamilies were associated with specific cellular functions. GO SEA conducted by AgriGO toolkit (Tian et al., 2017) against the database of the 14,258 expressed and syntenic triads revealed that triads which include TEs from each of 13 specific superfamilies were enriched for numerous GO terms from the BP, MF, and CC domains (Table 5; Figure 2 and Supplementary Figures S2-S13, Supplementary Tables S6-S25). The DTA superfamily was excluded from the analysis due to the small sample size.

In the BP domain, a significantly correlation was found between the presence of TEs from specific superfamilies within the triad and basic cell processes like gene silencing by RNA, cell cycle, organelle organization, recombinational repair, DNA recombination, telomere organization, DNA-templated DNA replication, and DNA methylation. Additionality, a significant correlation was found between the presence of TE insertions from specific superfamilies and response to biotic and abiotic stress, such as response to virus, response to nematode, vernalization response, and response to symbiotic fungus. Interestingly, triads that include TEs from specific superfamilies were also found to be associated with GO terms from the BP domine associated with the transposition mechanisms of the two TE classes, including transposition, RNA-mediated, and DNA-mediated (Table 5; Figure 2 and Supplementary Figures S2-S13, Supplementary Tables S6-S25). For the MF domain association was observed between the presence of TE insertions from specific superfamilies within the triad and enzymes activities and that carry out basic cell processes, including ligase activity, helicase activity, and DNA-directed DNA polymerase activity and with DNA repair, including DNA insertion or deletion binding (tale 5). Similarly, for the CC domain, an association was observed with the RNA polymerase I complex, responsible for basic cell activity, and with the DNA repair complex (Table 5; Supplementary Figures S2-S13, Supplementary Tables S6-S25). In addition, our analysis revealed enrichment in terms associated with the regulation of gene expression, such as the RISC complex (CC) and RNAi effector complex (CC), and transposase activity (MF) (Table 5; Supplementary Figures S2-S13, Supplementary Tables S6-S25).

While some of the TE superfamilies were found to be significantly enriched for most of the mentioned GO terms,

TABLE 3 TE subfamilies for which the presence/absence of TE insertion in at least one of the triad genes correlated with balanced relative expression of the three homoeologs.

CLARITE name[1]	corrected p-values[2]	TE insertion			
		Yes	No	Yes	No
		Balanced triads[3]		Non-balanced triads[4]	
DTC_famc11.1	0.046018	23	6361	11	1044
RLC_famc6	0.000736	18	6366	13	1042
RLC_famc1.6	0.00292	6	6378	7	1048
RLC_famc20	9.35E-11	22	6362	24	1031
RLC_famc7.1	0.034407	7	6377	6	1049
DTC_famc4.3	0.00766	5	6379	6	1049
DTM_famc9	0.04758	20	6364	10	1045
RLG_famc1.1	0.002862	13	6371	10	1045
RIX_famc1	0.004706	212	6172	59	996
RLC_famc8	0.001236	14	6370	11	1044
RIX_famc8	0.029024	693	5691	82	973
DTM_famc8	0.04758	20	6364	10	1045
DTT_famn14	0.000511	721	5663	72	983
XXX_famc13	0.000302	99	6285	38	1017
SIX_famc1	0.026102	9	6375	7	1048
XXX_famc16	3.10E-05	119	6265	46	1009
RLX_famc22	0.001572	29	6355	16	1039
XXX_famc112	0.001236	14	6370	11	1044
RIX_famc15	3.23E-10	24	6360	24	1031

[1] According to Wicker et al. (Wicker et al., 2018). TE names were selected based on the ClariTeRep naming system, which assigns simple numbers to individual families and subfamilies.

[2] χ^2 corrected p-values for multiple tests using Benjamini/Hochberg Procedure (non-negative).

[3] Number of triads belonging to the homoeolog-balanced category, in which at least one homoeolog contains TE insertion from the mention subfamily (Yes) or none of the homoeologs contain TE insertion from the mention subfamily (No).

[4] Number of triads belonging to one of the non-balanced categories, meaning to one of the homoeolog-dominant or homoeolog-suppressed categories, in which at least one homoeolog contains TE insertion from the mention subfamily (Yes) or none of the homoeologs contain TE insertion from the mention subfamily (No).

others showed enrichment for only a few of the GO terms we decided to focus on or even only for one of the mentioned terms (Table 5). For instance, triads that included TE insertions belonging to the DTT superfamily (Figure 2A) were found to be enriched for all the GO terms mentioned in Table 5 except for transposition, RNA-mediated (BP), and chromosome (CC), while triads that included TE insertions belonging to the DTH superfamily showed association with only 2 of the GO terms from Table 5, gene silencing by RNA (BP) and DNA methylation (BP).

Following the GO SEA performed for triads that included TE insertions from specific superfamilies within the gene bodies, we further examined whether triads that showed a specific relative expression pattern and included TE insertions from specific superfamilies would associate with different GO terms relative to all the triads which include TE insertions from the same superfamily. For this purpose, we focused on triads that include TE insertions from superfamilies that we found that their presence within a triad correlated with specific relative expression patterns and are found in the relevant expression category (shown in Table 2). For example, triads that include DTT insertions were significantly more likely to be found in the balanced relative expression category compering to triads that included TE insertions but did not include insertions of DTT TEs, and thus, the analysis was performed for triads that included insertions belonging to the DTT superfamily and showed balanced expression of the homoeologs (Figure 2A; Supplementary Figures S2-S4, S11-S13, Supplementary Tables S12 and S20). However, for triads that include insertions belonging to the DTM superfamily, the analysis was

TABLE 4 TE subfamilies for which the presence/absence of TE insertion in at least one of the triad genes correlated with the suppression of a single homoeolog gene.

CLARITE name[1]	corrected p-values[2]	TE insertion			
		Yes	No	Yes	No
		Suppressed triads[3]		Not suppressed triads[4]	
DTC_famc11.1	0.028255	10	872	24	6533
RLC_famc6	0.013022	10	872	21	6536
RLC_famc1.6	0.000371	7	875	6	6551
RLC_famc20	3.85E-10	21	861	25	6532
XXX_famc33	0.013022	5	877	5	6552
RLC_famc7.1	0.008919	6	876	7	6550
RLG_famc1.1	0.018009	8	874	15	6542
RIX_famc1	0.017541	49	833	222	6335
RLC_famc8	0.008831	9	873	16	6541
DTT_famn14	0.000371	57	825	736	5821
RLG_famc15	0.026756	8	874	16	6541
XXX_famc13	3.68E-05	35	847	102	6455
XXX_famc16	1.02E-06	43	839	122	6435
XXX_famc140	0.018009	8	874	15	6542
RLX_famc22	0.000371	15	867	30	6527
XXX_famc112	7.95E-05	11	871	14	6543
RIX_famc15	1.88E-10	22	860	26	6531

[1] According to Wicker et al. (Wicker et al., 2018). TE names were selected based on the ClariTeRep naming system, which assigns simple numbers to individual families and subfamilies.
[2] χ^2 corrected p-values for multiple tests using Benjamini/Hochberg Procedure (non-negative).
[3] Number of triads belonging to one of the homoeolog-suppressed categories, in which at least one homoeolog contains TE insertion from the mention subfamily (Yes) or none of the homoeologs contain TE insertion from the mention subfamily (No).
[4] Number of triads belonging to the balanced category or to one of the homoeolog-dominant categories, in which at least one homoeolog contains TE insertion from the mention subfamily (Yes) or none of the homoeologs contain TE insertion from the mention subfamily (No).

performed separately for triads that belonged to the suppressed and the dominant relative expression categories since triads that include DTM insertions were significantly more likely to be found in suppressed or dominant relative expression category in comparison to triads that included TE insertions but did not include insertions of DTM TEs (Figure 2B; Supplementary Figures S2-S10, Supplementary Tables S6, S21, S22).

Generally, similar GO terms were found to be significantly enriched for the same TE superfamily when all the TE containing triads were tested and upon focusing on triads from a specific relative expression contribution category (Supplementary Figures S2-S13; Supplementary Tables S6-S25). However, we noticed that in some cases, specific terms were found to be enriched by the analysis performed for all the triads with TE insertions from specific TE superfamily and were missing from the results when the analysis was performed only for triads from specific relative expression category, or the other way around. For example, for the DTM superfamily, while the GO terms production of siRNA involved in RNA interference (GO:0030422), regulation of DNA methylation (GO:0044030), and posttranscriptional gene silencing by RNA (GO:0035194) were found to be significantly enriched when the analysis was performed for all the DTM insertions containing triads they were missing from the results of the analysis for only triads from the dominant relative expression categories, and from the results of the analysis for only triads from the suppressed relative expression categories (Figure 2B; Supplementary Figures S2-S10, Supplementary Tables S6, S21, S22). However, significant association with GO terms that were not found to be enriched for all the DTM continuing triads was identified for the triads that include DTM insertions and belonging to one of the homoeolog-dominant expression categories, mainly associated with response to biotic and abiotic factors and aging (aging (GO:0007568), leaf senescence (GO:0010150), organ senescence (GO: 0010260), response to metal ion (GO:0010038), response to oxidative stress (GO:0006979), response to biotic stimulus

TABLE 5 Significantly enriched GO terms found in GO SEA preformed for triads which include TE insertions from mentioned superfamily in each of the three GO domains classes: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF).

Class	GO term	GO ID	RLG[1*]	RLC[1*]	RLX[1*]	RIX[1*]	SIX[1*]	DTC[1*]	DTM[1*]	DTX[1*]	DTH[1*]	DTT[1*]	DTA[1*]	DXX[1*]	XXX[1*]
BP	gene silencing by RNA	GO:0031047	+	+	–	+	–	+	–	+	+	+	–	–	+
	Cell cycle	GO:0007049	+	+	–	+	–	+	–	+	–	+	–	–	+
	Organelle organization	GO:0006996	+	+	+	+	–	+	–	+	–	+	–	–	–
	Recombinational repair	GO:0000725	+	+	–	+	–	+	–	+	–	+	–	–	+
	DNA recombination	GO:0006310	+	+	–	+	–	+	–	+	–	+	–	+	+
	Telomere organization	GO:0032200	+	+	–	+	–	+	–	+	–	+	–	–	–
	DNA-templated DNA replication	GO:0006261	–	–	–	–	–	+	–	+	–	+	–	–	–
	Transposition, RNA-mediated	GO:0032197	–	+	–	–	–	+	–	–	–		–	–	–
	DNA methylation	GO:0006306	+	+	–	+	–	+	–	+	+	+	–	–	+
	Transposition, DNA-mediated	GO:0006313	–	–	–	–	–	–	–	+	–	+	–	–	–
	response to virus	GO:0009615	–	+	–	+	–	+	–	+	–	+	–	–	+
	Response to nematode	GO:0009624	+		–	–	+	–	+	–	–	+	–	–	+
	vernalization response	GO:0010048	+	+	+	–	–	+	–	–	–	+	–	–	–
	Response to symbiotic fungus	GO:0009610	–	–	–	–	–	–	–	+	–	+	–	+	–
MF	transposase activity	GO:0004803	–	–	–	–	–	–	–	+	–	+	–	–	–
	ligase activity	GO:0016874	+	+	–	+	–	+	–	+	–	+	–	–	–
	helicase activity	GO:0004386	+	+	+	+	–	+	–	+	–	+	–	–	+
	DNA-directed DNA polymerase activity	GO:0003887	–	–	–	+	–	–	–	+	–	+	–	–	–
	DNA insertion or deletion binding	GO:0032135	–	–	–	–	–	+	–	+	–	+	–	–	–
CC	DNA repair complex	GO:1990391	+	+	–	+	–	+	–	+	–	+	–	–	–
	RISC complex	GO:0016442	+	+	–	–	–	+	–	+	–	+	–	–	–
	RNAi effector complex	GO:0031332	+	+	–	–	–	+	–	+	–	+	–	–	–
	chromosome	GO:0005694	+	+	–	+	–	+	–	+	–		–	–	–
	RNA polymerase I complex	GO:0005736	–	–	+	–	–	–	–	+	–	+	–	–	–

[1] A code for each of the 13 TE superfamilies, representing the class (first letter), order (second letter) and superfamily (third letter) (Wicker et al., 2007). "+" notes that the respective GO term was found to be enriched in triads that include TE insertion from the mention superfamily. "-" notes that the respective go term was not found to be enriched in triads which include TE from the specific superfamily. GO SEA was preformed using AgriGO toolkit (Tian et al., 2017) with Fisher's exact test (FDR ≤ 0.05).

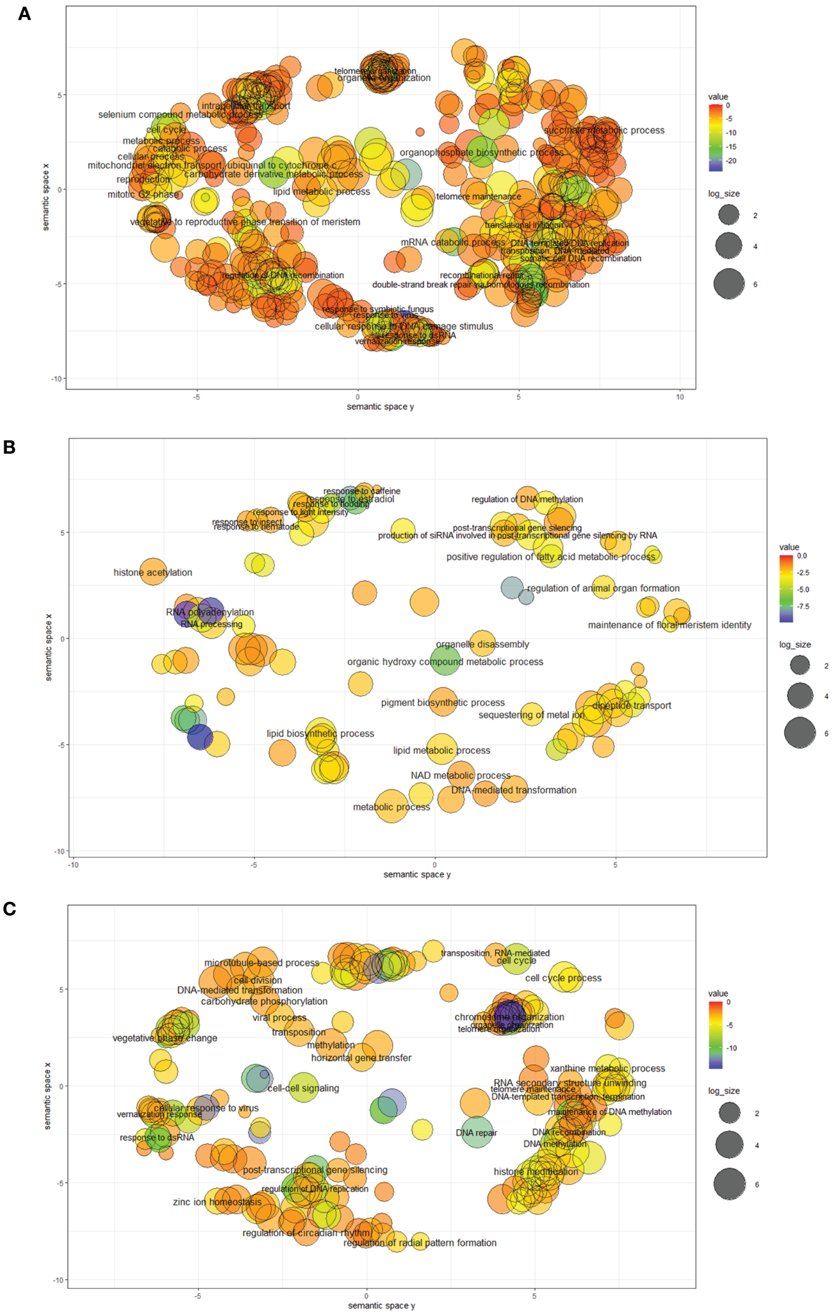


FIGURE 2

Significantly enriched GO BP (Biological prosses) terms found in GO SEA preformed for triads which include TE insertions from the DTT **(A)**, DTM **(B)** and RLC **(C)** superfamilies (TE codes are based on Wicker et al., 2007). GO SEAs were preformed using AgriGO toolkit with Fisher's exact test (FDR ≤ 0.05). Following the GO SEA, the enriched GO terms for BP for each superfamily were visualized as a scatter plot generated by REVIGO. Closer GO terms in the plot are showing higher semantically similarity to each other. The bubble color indicates the FDR value and the size indicates the frequency of the GO term in the GOA database, bubbles of more general terms are larger.

(GO:0009607), innate immune response (GO:0045087)) (Figure 2B; Supplementary Figures S2-S10, Supplementary Tables S6, S21). Similarly, significant association with GO terms that were not found to be enriched for all the DTM continuing triads was identified for the triads that include DTM insertions and belonging to one of the homoeolog-suppressed expression categories, including GO terms related to pollen formation and circadian rhythm regulation (negative regulation of circadian rhythm (GO:0042754), pollen exine formation (GO:0010584), pollen wall assembly (GO:0010208), pollen development (GO:0009555)) (Figure 2B, Supplementary Figures S2-S10; Supplementary Tables S6, S22). Another example is the differences observed in the GO SEA results performed for all triads that include TE insertions from unknown class (XXX- unclassified repeats) versus the results obtained only for triads which include TE insertions from unknown class (XXX) and assigned to one of the homoeolog-suppressed expression categories (Supplementary Figures S2-S4 , S8-S10, Supplementary Tables S18, S25). Go terms directly related to the regulating gene expression, DNA modifications, and response to biotic and abiotic stress (cellular response to glucose stimulus (GO:0071333)gene silencing by RNA (GO:0031047), chromatin silencing (GO:0006342), DNA methylation (GO:0006306), gene silencing(GO:0016458), RNA interference (GO:0016246), response to dsRNA(GO:0043331), response to virus (GO:0009615), response to nematode (GO:0009624)) were found to be significantly enriched in the analysis performed for all the XXX insertions containing triads and missing from the results obtained from the GO SEA for triads belonging to an homoeolog-suppressed expression category that include XXX insertions (Supplementary Figures S2-S4, S8-S10, Supplementary Tables S18, S25). Meanwhile, GO terms related to aging and some abiotic stress (aging (GO:0007568), negative regulation of growth (GO:0045926), leaf senescence (GO:0010150) response to salt stress (GO:0009651), cellular response to alcohol (GO:0097306)) were found to be enriched in the list of triads belonging to one of the homoeolog-suppressed expression categories and containing XXX insertions, and missing from the GO SEA results obtained for all the XXX insertions containing triads (Supplementary Figures S2-S4, S8-S10, Supplementary Tables S18, S25).

3.5 TE insertions site within the gene body and triad expression patterns

Next, we tested for a possible association between TE insertion context within the gene body (exon, 5' UTR, or 3' UTR) of one or more of the homoeologs in a triad and homoeolog expression bias. Of the 7,439 triads that include TE insertions, 10.55% (785) include TE insertion within an exon of at least one of the genes (Supplementary Table S3). A lower proportion of the triads that include TE insertion within an exon was found in the balanced expression category (608, 77.45%) compared to triads that contain TE insertions but did not contain TE insertions within exons (5,776, 86.80%) with a significant difference in proportions (χ^2 = 49.70, p < 0.001). The significant difference in proportions of the triads that include TE within an exon in balanced vs. non-balanced relative contribution categories was also observed separately for homoeolog-suppressed vs. non-suppressed (balanced and dominant relative contribution) categories (χ^2 = 30.65, p < 0.001) and for homoeolog-dominant vs. non-dominant (balanced and suppressed relative contribution) categories (χ^2 = 18.64, p < 0.001). Out of the triads that contained TE insertions but did not contain insertions within exons, only 11.14% (741) were found in one of the homoeolog-suppressed categories, and 2.06% (137) were found in one of the homoeolog-dominant categories, while triads that include TE insertion within an exon were more likely to be found in both the homoeolog-suppressed (17.96%, 141 triads) and the homoeolog-dominant (4.59%, 36 triads) categories.

More specifically, of the 608 triads that include TE insertions within an exon of at least one of the genes in the triad, 10.53% (64) include TE insertions within the 5' UTR and 61.02% (371) include TE insertions within the 3' UTR (Table S3). A lower proportion of the triads that include TE insertions within the 5' UTR (46, 71.88%) and 3' UTR (299, 80.59%) was found in the balanced expression category compared to triads that contain TE insertions but did not contain insertions within the 5' UTRs (6,338, 85.94%) or within the 3' UTR (6085, 86.09%), respectively. The difference in proportions of the triads that include TE insertions within the 5' UTR (χ^2 = 9.19, p = 0.0024 < 0.05) and within the 3' UTR (χ^2 = 8.31, p = 0.0039 < 0.05) in the balanced expression category relative to the proportions of triads with no TE insertions in mentioned regions, were statistically significant. We observed that triads that include TE insertions within the 5' UTR were more likely to be found both in the homoeolog-suppressed (15.63%, 10 triads) and the homoeolog-dominant (12.50%, 8 triads) categories, relative to triads that contain TE insertions but did not contain insertions the 5' UTRs, with only 11.82% of the triads (872) assigned to one of the homoeolog-suppressed categories and 2.24% of the triads (165) assigned to one of the homoeolog-dominant categories. Similar results were observed for triads which include TE insertions within the 3' UTR, which were also found in higher proportions in the homoeolog-suppressed (15.09%, 56 triads) and the homoeolog-dominant (4.31%, 16 triads) categories, in comparison to triads that contain TE insertions but did not contain insertions the 3' UTRs, for which only 11.69% of the triads (826) were assigned to one of the homoeolog-suppressed categories and only 2.22% of the triads (157) were assigned to one of the homoeolog-dominant categories. While a significant difference was observed for the proportions of triads which include TE insertions within the UTRs and triads that contain TE insertions but did not contain insertions in the UTRs within homoeolog-dominant categories (for 5' UTR: χ^2 = 25.08, p < 0.001 and for 3' UTR: $\chi2$ = 5.90, p = 0.0152 < 0.05), the difference in proportion in the homoeolog-suppressed categories were not statistically significant (for 5' UTR: χ^2 = 0.55, p = 0.46 > 0.05 and for 3' UTR: χ^2 = 3.60, p = 0.058 > 0.05).

4 Discussion

As an allohexaploid species, the bread wheat contains three subgenomes, A, B, and D, which originated from three diploid genome donors that diverged from a common progenitor ~7 MYA (Million Years Ago) (Levy and Feldman, 2022). While high conservation in gene content and order was observed between the three subgenomes, almost no sequence conservation was found in the intergenic regions, containing mostly TEs (Appels et al., 2018; Wicker et al., 2018). The contribution of TEs to the differentiation between the three subgenomes of the young allohexaploid bread wheat might facilitate genetic and cytological diploidization, which is essential for the survival of the new species.

4.1 TE insertions are highly abundant within gene bodies

Together, dyads (11.7%), triads (51.1%), and tetrads (2.8%) are accounted for 65.6% of all bread wheat HC genes, while the rest of the genes deviate from these rations (Juery et al., 2020). While triads genes were kept in a 1:1:1 ratio between the three bread wheat subgenomes, dyads and tetrads homoeologous groups underwent copy number variations during the wheat group evolution (Juery et al., 2020). A study by Juery et al. (2020) showed that triads are diverse from dyads and tetrads in various characteristics, including conservation, chromosomal distribution, epigenetic modification, gene ontology, and expression patterns. Their findings led Juery et al. (2020) to suggest that the highly conserved triads belong to the bread wheat core genome, while dyads and tetrads are parts of the dispensable genome.

To address the possible effect of TE insertions within gene bodies on gene expression, we first identified TE insertions within dyads, triads, and tetrads genes. We found that a high percentage of the examined genes contained TE insertions within exons and introns, with the highest proportions of TE insertions found in triads genes. Additionally, genes found in the proximal region were more likely to include TE insertions within the gene body, suggesting that TE distribution within the gene body is in accordance with TE distribution across the chromosome, with lower density in the distal regions. However, the higher percentage of TE insertions in triads genes relative to dyads and tetrads genes persist throughout the different chromosomal regions. Therefore, the difference in the abundance of TE insertions within gene bodies between triads and dyads and tetrads genes is not only due to the higher abundance of triads genes in the TE rich proximal regions.

The higher abundance of TE insertions within triads genes relative to dyads and tetrads genes might be attributed to some of the distinguish characteristics of each of the categories. For instance, triads genes were found to be enriched in the H3K9ac active euchromatin mark and expressed at a higher level and higher breadth relative to dyads and tetrads genes, which were more associated with repressive H3K27me3 modification (Juery et al., 2020). There is evidence that some TEs are preferentially inserted into transcriptionally active regions near active histone marks (Bennetzen, 2000; Hirsch and Springer, 2017; Sultana et al., 2017), conditions that fit better to triads genes. Specifically, TEs belonging to the *Mariner* superfamily (DTT), the most abundant superfamily in triad genes, are known to be enriched in genes with high expression (Sultana et al., 2017). Moreover, the higher conservation of triads genes might contribute to the persistence of the TE insertion, provided that the insertion did not result in loss of fitness. Alternatively, the presence of TE insertion within the gene body might impact various characteristics of the gene and maybe ultimately on gene conservation. Further study is necessary for a better understanding of the processes leading to the unique TE distribution pattern observed in this study.

4.2 Polymorphic TE insertions within gene bodies associated with non-balanced expression within the homoeologous group

Since the three diploid genome donors of bread wheat originated from a common ancestor, it is not surprising that a high percentage of wheat HC genes are conserved and syntenic between the three subgenomes (Appels et al., 2018). Similarly, the abundances of 76% of TEs families were found to be similar between the A, B, and D sub-genomes of bread wheat and TE families distribution in promoter regions was found to be highly conserved between subgenomes (Wicker et al., 2018). However, almost no conserved TE insertions were observed between the three subgenomes, and more specifically, no conservation of TE insertions was observed between homeologous promoters (Wicker et al., 2018).

Here, we assigned homoeologous groups as monomorphic or polymorphic based only on the presence of TE insertions within all or only some of the gene bodies of the homoeologs in the group. While the TE insertions found within the homoeologs in a monomorphic group were not necessarily inherited from the common ancestor, did not necessarily belong to the same TE family, and might have been inserted in different locations in the sequence, in this part of our analysis we focused on the effect of the mere presence of TE insertion on the relative expression within the homoeologous group. However, the fact that the proportion of polymorphic groups did not match a module describing the random distribution of presence\ absence of TE insertions leads us to suggest that some of the TE insertions are indeed having a common origin, or alternatively, that common characteristics of the homoeologs affected the probability of TEs to insert into each one of the genes in the homoeologous group. The percentage of polymorphic groups was lower than expected for triads and tetrads and higher than expected for dyads, which were found to be the least conserved relative to triads and tetrads (Juery et al., 2020). Additionally, our analysis revealed a strong significant correlation between polymorphic TE insertions and non-balanced expression patterns of triads. We suggest that the mentioned correlation might be a result of either the effect of TE insertions on gene expression and \ or TE target site preference influenced by gene expression patterns and expression breadth.

4.3 Strong association between TE insertions within gene bodies and homoeolog expression bias

Since TE insertions were found to be abundant in triads, and a clear correlation was observed between TE insertion pattern and relative expression of the homoeolog within the triad, we focused on triads to further learn about the possible impact of TE insertions on gene expression, using existing data regarding the relative contribution of each homoeolog to the overall triad expression. Our analysis revealed that a great variety of TEs inserted within wheat gene bodies, both into introns and exons. Here, we observed a strong correlation between the presence of TE insertions in gene bodies and the balanced expression of the three homoeologs in the triad. Similar to the differences in TE abundant in triads vs. dyads and tetrads genes, the unique characteristics of each relative expression category might explain the difference in TE content. Syntenic triads that were classified as balanced showed higher expression levels and had higher levels of active histone markers than syntenic triads in the homoeolog-dominant and homoeolog-suppressed categories (Ramírez-González et al., 2018). As we suggested for triads vs. dyads and tetrads, those characteristics, together with the balanced triads over representation in low recombination regions (Ramírez-González et al., 2018), might lid to higher insertion rate and higher persistent of TE insertions in the balanced triads genes relative to triads from the non-balanced categories. This claim is supported by the very high abundant of insertions from the *Mariner* superfamily (DTT), which was found to be enriched in genes with high expression (Sultana et al., 2017), in balanced triads and by the strong correlation observed specifically between the presence of TE insertions from the DTT superfamily and balanced homoeologs expression. Additionally, we suggest that the presence of TE insertions within gene bodies might result in a change in gene expression, resulting in balanced expression of the homoeologs.

Generally, a strong correlation was observed between the presence of TE insertions within the triad and balanced expression. However, while considering the insertion site and TE type, a more complex relationship between the presence of TE insertion and homoeolog expression bias is revealed. We found that specific TE superfamilies and families were enriched in triads which showed specific relative expression patterns. Furthermore, the presence of TE insertions from 13 out of the 14 TE superfamilies within a triad associated with multiple GO terms enriched both in basic cellular functions and in response to environmental factors. Triads that contained TE insertions from each one of the 13 different TE superfamilies showed enrichment for a unique set of GO terms. Triads which include insertions of DTT and DTX, superfamilies for which a positive correlation was identified between TE presence within gene bodies and balanced expression of the triad, are found in association with numerous GO terms related to basic cell processes (Figure 2A). Additionally, triads which include insertions belonging to TE superfamilies that showed a positive correlation between TE presence within gene bodies and suppressed or dominant relative expression of the homoeologs (DTM, RLX, RLC, and XXX) were enriched for multiple GO terms. Specifically, triads that contained TE insertions from each of those 4 superfamilies (DTM, RLX, RLC, and XXX) were enriched for GO terms related to response to biotic and \ or abiotic stimulus (Figures 2B, C).

5 Conclusions

In this study, the integration and analysis of data from several publicly available databases revealed significant correlations between the presence of TE insertions within gene bodies, gene expression and gene function in a genome-specific manner in wheat. We found that TE insertion site within the gene (exon\ intron) and TE type (superfamily\ subfamily) correlate strongly with homoeolog expression bias. Additionally, presence of TE insertion from all tested TE superfamilies were found to associate with numerous gene functions. Future studies are needed to decipher the causes for such correlations. In addition, comparative analysis between bread wheat accessions might shed light on the evolutionary time frame for TE insertions into gene bodies and on the involved mechanisms connecting TE presence within the gene body, gene expression, and gene function.

Resource identification initiative

SciPy (RRID : SCR_008058)
Pandas (RRID : SCR_018214)
statsmodel (RRID : SCR_016074)
agriGO (RRID : SCR_006989)
REViGO (RRID : SCR_005825)

Data availability statement

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

Author contributions

IB: generated data, data analysis, wrote manuscript. LG: data analysis, wrote manuscript. KK: data analysis, wrote and edit the manuscript, corresponding author, submitted manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by a grant from the Israel Science Foundation (grant # 1311/21) to KK.

Acknowledgments

We would like to thank Alon Ziv and Vadim Khasdan for their assistance with the manuscript preparation.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1072232/full#supplementary-material

References

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191. doi: 10.1126/science.aar7191

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25 (1), 25–29. doi: 10.1038/75556

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., et al. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357, 93–97. doi: 10.1126/science.aan0032

Bariah, I., Keidar-Friedman, D., and Kashkush, K. (2020). Where the wild things are: Transposable elements as drivers of structural and functional variations in the wheat genome. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.585515

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42 (1), 251–269. doi: 10.1023/A:1006344508454

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., et al. (2018). Ten things you should know about transposable elements 06 biological sciences 0604 genetics. *Genome Biol.* 19 (1), 199. doi: 10.1186/s13059-018-1577-z

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., et al. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi: 10.1093/NAR/GKY1055

Crescente, J. M., Zavallo, D., del Vas, M., Asurmendi, S., Helguera, M., Fernandez, E., et al. (2022). Genome-wide identification of MITE-derived microRNAs and their targets in bread wheat. *BMC Genomics* 23, 1–14. doi: 10.1186/S12864-022-08364-4/FIGURES/4

Domb, K., Keidar-Friedman, D., and Kashkush, K. (2019). A novel miniature transposon-like element discovered in the coding sequence of a gene that encodes for 5-formyltetrahydrofolate in wheat. *BMC Plant Biol.* 19, 1–11. doi: 10.1186/s12870-019-2034-1

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

Feldman, M., and Levy, A. A. (2005). Allopolyploidy - a shaping force in the evolution of wheat genomes. *Cytogenet. Genome Res.* 109, 250–258. doi: 10.1159/000082407

Fox, D. T., Soltis, D. E., Soltis, P. S., Ashman, T. L., and Van de Peer, Y. (2020). Polyploidy: A biological force from cells to ecosystems. *Trends Cell Biol.* 30 (9), 688–694. doi: 10.1016/j.tcb.2020.06.006

Gray, Y. H. M. (2000). It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements. *Trends Genet.* 16, 461–468. doi: 10.1016/S0168-9525(00)02104-1

Guttag, J. V. (2021). *Introduction to Computation and Programming Using Python, Third Edition*. MIT Press.

Hirsch, C. D., and Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1860, 157–165. doi: 10.1016/J.BBAGRM.2016.05.010

Jiang, Y. F., Chen, Q., Wang, Y., Guo, Z. R., Xu, B. J., Zhu, J., et al. (2019). Re-acquisition of the brittle rachis trait *via* a transposon insertion in domestication gene q during wheat de-domestication. *New Phytol.* 224, 961–973. doi: 10.1111/nph.15977

Juery, C., Concia, L., De Oliveira, R., Papon, N., Ramírez-González, R., Benhamed, M., et al. (2020). New insights into homoeologous copy number variations in the hexaploid wheat genome. *Plant Genome.* 14(1). doi: 10.1002/tpg2.20069

Keidar, D., Doron, C., and Kashkush, K. (2018). Genome-wide analysis of a recently active retrotransposon, au SINE, in wheat: content, distribution within subgenomes and chromosomes, and gene associations. *Plant cell rep.* 37 (2), 193–208. doi: 10.1007/s00299-017-2213-1

Keidar-Friedman, D., Bariah, I., Domb, K., and Kashkush, K. (2020). The evolutionary dynamics of a novel miniature transposable element in the wheat genome. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.01173

Keidar-Friedman, D., Bariah, I., and Kashkush, K. (2018). Genome-wide analyses of miniature inverted-repeat transposable elements reveals new insights into the evolution of the triticum-aegilops group. *PloS One* 13, e0204972. doi: 10.1371/journal.pone.0204972

Krasileva, K. V. (2019). The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. *Curr. Opin. Plant Biol.* 48, 18–25. doi: 10.1016/J.PBI.2019.01.004

Levy, A. A., and Feldman, M. (2022). Evolution and origin of bread wheat. *Plant Cell.* 34(7), 2549–2567. doi: 10.1093/PLCELL/KOAC130

Li, J., Wang, Z., Peng, H., and Liu, Z. (2014). A MITE insertion into the 3′-UTR regulates the transcription of TaHSP16.9 in common wheat. *Crop J.* 2, 381–387. doi: 10.1016/j.cj.2014.07.001

Mhiri, C., Borges, F., and Grandbastien, M. A. (2022). Specificities and dynamics of transposable elements in land plants. *Biol* 11, 488. doi: 10.3390/BIOLOGY11040488

Poretti, M., Praz, C., Meile, L., Kälin, C., Schaefer, L. K., Schläfli, M., et al. (2019). Domestication of high-copy transposons underlays the wheat small RNA response to an obligate pathogen. *Mol. Biol. Evol.* 37(3), 839–848. doi: 10.1093/molbev/msz272

Qiu, Y., and Köhler, C. (2020). Mobility connects: transposable elements wire new transcriptional networks by transferring transcription factor binding motifs. *Biochem. Soc Trans.* 48, 1005–1017. doi: 10.1042/BST20190937

Ramírez-González, R. H., Borrill, P., Lang, D., Harrington, S. A., Brinton, J., Venturini, L., et al. (2018). The transcriptional landscape of polyploid wheat. *Science* 361 (6403). doi: 10.1126/science.aar6089

Reback, J., McKinney, W., Van den Bossche, J., Augspurger, T., Cloud, P., Clein, A., et al. (2021). pandas-dev/pandas: Pandas 1.2.4. doi: 10.5281/ZENODO.4681666

Salina, A.E., and Adonina, I. G. (2018). Cytogenetics in the study of chromosomal rearrangement during wheat evolution and breeding. *Cytogenet. - Past, Present Furth. Perspect*, 10. doi: 10.5772/INTECHOPEN.80486.

Schrader, L., and Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. *Mol. Ecol.* 28, 1537–1549. doi: 10.1111/mec.14794

Shewry, P. R., and Hey, S. J. (2015). The contribution of wheat to human diet and health. *Food Energy Secur.* 4, 178–202. doi: 10.1002/FES3.64

Sultana, T., Zamborlini, A., Cristofari, G., and Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* 18 (5), 292–308. doi: 10.1038/nrg.2017.7

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One* 6, e21800. doi: 10.1371/JOURNAL.PONE.0021800

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017).). agriGO v2.0: a GO analysis toolkit for the agricultural community 2017 update. *Nucleic Acids Res.* 45, W122–W129. doi: 10.1093/NAR/GKX382

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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17 (3), 261–272. doi: 10.1038/s41592-019-0686-2

Waskom, M. (2021). Seaborn: statistical data visualization. *J. Open Source Software* 6, 3021. doi: 10.21105/joss.03021

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-Gonzalez, R. H., et al. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* 19, 103. doi: 10.1186/s13059-018-1479-0

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

Xi, X., Li, N., Li, S., Chen, W., Zhang, B., Liu, B., et al. (2016). The characteristics and functions of a miniature inverted-repeat transposable element TaMITE81 in the 5′ UTR of TaCHS7BL from triticum aestivum. *Mol. Genet. Genomics* 291, 1991–1998. doi: 10.1007/s00438-016-1234-8

Zhang, Y., Li, Z., Zhang, Y., Lin, K., Peng, Y., Ye, L., et al. (2021). Evolutionary rewiring of the wheat transcriptional regulatory network by lineage-specific transposable elements. *Genome Res.* 31, 2276–2289. doi: 10.1101/GR.275658.121

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 06 March 2023
DOI 10.3389/fpls.2023.1134627

OPEN ACCESS

EDITED BY
Ruslan Kalendar,
University of Helsinki, Finland

REVIEWED BY
Flavia Mascagni,
University of Pisa, Italy
Margaret Woodhouse,
Agricultural Research Service (USDA),
United States
Kenji K. Kojima,
Genetic Information Research Institute,
United States

*CORRESPONDENCE
Achraf El Allali
achraf.elalalli@um6p.ma
Morad M. Mokhtar
morad.mokthar@um6p.ma

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

RECEIVED 30 December 2022
ACCEPTED 16 February 2023
PUBLISHED 06 March 2023

CITATION
Mokhtar MM, Alsamman AM
and El Allali A (2023) PlantLTRdb: An
interactive database for 195 plant
species LTR-retrotransposons.
Front. Plant Sci. 14:1134627.
doi: 10.3389/fpls.2023.1134627

PlantLTRdb: An interactive database for 195 plant species LTR-retrotransposons

Morad M. Mokhtar*, Alsamman M. Alsamman
and Achraf El Allali*

African Genome Center, Mohammed VI Polytechnic University, Benguerir, Morocco

LTR-retrotransposons (LTR-RTs) are a large group of transposable elements that replicate through an RNA intermediate and alter genome structure. The activities of LTR-RTs in plant genomes provide helpful information about genome evolution and gene function. LTR-RTs near or within genes can directly alter gene function. This work introduces PlantLTRdb, an intact LTR-RT database for 195 plant species. Using homology- and *de novo* structure-based methods, a total of 150.18 Gbp representing 3,079,469 pseudomolecules/scaffolds were analyzed to identify, characterize, annotate LTR-RTs, estimate insertion ages, detect LTR-RT-gene chimeras, and determine nearby genes. Accordingly, 520,194 intact LTR-RTs were discovered, including 29,462 autonomous and 490,732 nonautonomous LTR-RTs. The autonomous LTR-RTs included 10,286 *Gypsy* and 19,176 *Copia*, while the nonautonomous were divided into 224,906 *Gypsy*, 218,414 *Copia*, 1,768 BARE-2, 3,147 TR-GAG and 4,2497 unknown. Analysis of the identified LTR-RTs located within genes showed that a total of 36,236 LTR-RTs were LTR-RT-gene chimeras and 11,619 LTR-RTs were within pseudo-genes. In addition, 50,026 genes are within 1 kbp of LTR-RTs, and 250,587 had a distance of 1 to 10 kbp from LTR-RTs. PlantLTRdb allows researchers to search, visualize, BLAST and analyze plant LTR-RTs. PlantLTRdb can contribute to the understanding of structural variations, genome organization, functional genomics, and the development of LTR-RT target markers for molecular plant breeding. PlantLTRdb is available at https://bioinformatics.um6p.ma/PlantLTRdb.

KEYWORDS

LTR-retrotransposons, plant genomes, database, insertion age, LTR-RT gene chimeras

1 Introduction

Long terminal repeats (LTR) have attracted considerable interest in recent years because of their potential impact on the genome structure of most eukaryotic organisms (Jedlicka et al., 2020). LTR-retrotransposons (LTR-RTs) are a large and diverse group of transposable elements (TE) that replicate *via* an RNA intermediate (Jedlicka et al., 2020). LTR-RTs are divided into autonomous and nonautonomous groups. Wicker et al. (2007)

defined LTR-RT as autonomous if it encodes all domains essential for its mobilization without the element being either functional or active. This is in contrast to nonautonomous LTR-RT, which are defined as elements that lack some (or all) of the domains necessary for mobilization. Features of an autonomous LTR-RT include two identical LTRs, a primer binding site (PBS), a polypurine tract (PPT), *GAG* and *Pol* genes (Kumar and Bennetzen, 1999; Gao et al., 2003; Havecker et al., 2004; Eickbush and Jamburuthugoda, 2008; Bennetzen and Wang, 2014). The *Pol* gene is located at the 3' end of *GAG* and encodes reverse transcriptase (RT), protease (PROT), RNase H (RH), and integrase (INT), all of which are involved in retrotransposon DNA replication and the transposition system (Gao et al., 2003). *Copia*, *Gypsy*, and *BEL–Pao* superfamilies represent LTR-RT classes according to the arrangement of internal domains (Wicker et al., 2007). There is further evidence that nonautonomous elements can retrotranspose actively or inactively (Sabot et al., 2006). Examples of nonautonomous groups include LArge Retrotransposon Derivatives (LARD) (Kalendar et al., 2004), Terminal Repeat Retrotransposons with *GAG* domain (TR-GAG) (Chaparro et al., 2015), Terminal Repeats In Miniature (TRIM) (Witte et al., 2001), and Barley RetroElement-2 (BARE-2) (Tanskanen et al., 2007).

During the replication cycle of LTR-RT, the newly inserted copy has two identical LTRs at the two ends of the element. The accumulation of mutations between the two LTRs of an LTR-RT was used to estimate the elapsed time after insertion (Zhou et al., 2021). They also show divergence caused by mutations acquired over time proportional to the age of insertion (Kijima and Innan, 2009; Neumann et al., 2019). Unequal crossovers between LTRs result in loss of internal sequence and formation of solo-LTRs (Cossu et al., 2017). In species that allow LTR-RTs accumulation, the activity of LTR-RTs is a critical factor in genome evolution, causing extremely large genome sizes (Kelly et al., 2015; Wicker et al., 2018). Genomic studies have established that LTR-RTs account for a considerable proportion of many plant genomes, including 19% of peach genome (Alseekh et al., 2020), 62% of tomato genome (Sato et al., 2012), 53% of potato genome (Diambra, 2011), and more than 70% of maize genome (Schnable et al., 2009). Tracking LTR-RT activities in plant genomes provides useful information on genome evolution and consequently gene function. The activity of LTR-RT near or within genes not only provides the raw material for structures such as centromeres and introns, but also directly alters gene function (Bennetzen and Wang, 2014; Vitte et al., 2014). LTR-RTs can influence gene regulation processes such as alternative splicing, alternative promoter control, and gene silencing (Kashkush et al., 2003; Qu et al., 2019; Yamamoto et al., 2021). Their influence on gene activity may affect the agronomic traits of various crops. According to TE-genome-wide association studies, the activity of LTR-RT is associated with several important agronomic traits, including fruit weight of tomato and width of rice grains (Akakpo et al., 2020; Alseekh et al., 2020). In addition, the activation of TE can also be triggered by environmental stress, for example, the biotic stress-responsive *Tnt1* and *Tto1* families in tobacco (Grandbastien, 2015), the heat-responsive retrotransposons *Go-on* in rice (Cho et al., 2019), the cold-responsive *Tcs* family in citrus (Butelli et al., 2012), and *ONSEN* in *Arabidopsis* (Ito et al., 2013).

Advances in genome sequencing technologies have opened new avenues for the study of LTR-RTs and for understanding their role in plant evolution. Several efforts have been made to provide a stable and well-documented LTR data resource that can be used to support current and future plant functional genomic research. Several databases for TEs in plants have been developed with general and specific research tools. These databases include Repbase (Bao et al., 2015), TREP (Wicker et al., 2002), RetrOryza (Chaparro et al., 2006), MASiVEdb (Bousios et al., 2012), MnTEdb (Ma et al., 2015), DPTEdb (Li et al., 2016), GrTEdb (Xu et al., 2017), PlaNC-TE (Pedro et al., 2018), ConTEdb (Yi et al., 2018b), SPTEdb (Yi et al., 2018a), REXdb (Neumann et al., 2019), RepetDB (Amselem et al., 2019), and CicerSpTEdb (Mokhtar et al., 2021). Although these databases provide unique and useful information about LTR-RTs in different plant genomes, they lack important details and features. Therefore, there is a need for robust, publicly available LTR-RT databases to address the growing interest in the impact of LTR-RTs on genome evolution and functionality. Such databases would be beneficial in current and future efforts to incorporate LTR-RTs annotation as a potential component for understanding the hidden dynamics of the gene regulatory system. Several studies have used such data to guide annotation in gene expression experiments (Bui and Grandbastien, 2012) or to identify retrotransposon structures such as extrachromosomal circular DNA (Mann et al., 2022).

Here, we introduce PlantLTRdb, a comprehensively designed database to expand the understanding of plant genome organization and its structural variations. PlantLTRdb provides online and searchable data resources for LTR-RT genomic information and a reliable and powerful computational service. PlantLTRdb contains detailed information on LTR-RTs in 195 plant species, both model and non-model organisms. These results are easily accessible and can be displayed using various statistical and genome-wide visualization tools. Users can download annotation files for use in advanced genomic procedures. In addition, the website provides online identification analysis for LTR-RTs *via* LTR_FINDER (Xu and Wang, 2007), which supports the standard input sequence format (FASTA).

2 Materials and methods

2.1 Plant genome data

Genome sequences of plant species and their annotations were retrieved from the NCBI database (https://www.ncbi.nlm.nih.gov/). Only genome sequences annotated and labeled as reference or representative genomes, including model and non-model plant species, were used for this analysis. The resulting dataset included 201 plant species divided into 180 Streptophyta, 18 Chlorophyta, and 3 Rhodophyta. The species name, taxonomy ID, phylum, family, assembly level, genome coverage, GenBank accession number, and genome size of all plant species can be found in Table S1.

2.2 LTR-RT identification and classification

The intact LTR-RTs were identified and classified using the EDTA pipeline (Ou et al., 2019), LTRdigest (Steinbiss et al., 2009), and TEsorter (Zhang et al., 2022) in 201 plant species. The intact LTR-RT element consists of two identical or very similar LTRs, TG-CA terms of the LTRs, and a target site duplication (TSD) (Du et al., 2010a; Dai et al., 2018). The EDTA pipeline integrates the structure-based, homology-based, and *de novo* intact LTR-RT identification tools such as LTR_FINDER (Xu and Wang, 2007), LTRharvest (Ellinghaus et al., 2008) and LTR_retriever (Ou and Jiang, 2017). The parameters of LTR_FINDER were *maximum distance between LTRs: 15000, minimum distance between LTRs: 1000, maximum LTR Length: 7000, minimum LTR Length: 100, length of exact match pairs: 20, match score: 0.85 and output format: 2.* LTRharvest parameters were *minimum LTR Length: 100, maximum LTR Length: 7000, minimum length for each TSD: 4, maximum length for each TSD: 6, motif: TGCA, maximum number of mismatches in motif: 1, similarity threshold: 85, number of nucleotides to be searched for TSDs: 10, minimum seed length for exact repeats: 20 and use sequence descriptions in GFF3 output: yes.*

LTRdigest (Steinbiss et al., 2009) was used to identify and annotate primer binding sites, polypurine tract, and tRNAs of LTR-RT elements. The tRNA sequences of the 201 plant species were retrieved from a tRNA database of plant species (Mokhtar and EL Allali, 2022). TEsorter (Zhang et al., 2022) was used to annotate coding regions and classify LTR-RTs into clades using the REXdb database. In addition, TEsorter used the 80-80-80 (identity-coverage-length) unified classification system proposed by Wicker et al. (2007) to classify identified elements. To assess the quality of assembled repetitive sequences, the LTR Assembly Index (LAI) (Ou et al., 2018) was estimated using LAI within LTR_retriever (v2.9.0) with default parameters.

Classification of the identified LTR-RTs into putative autonomous and nonautonomous elements was based on the complete structure of the elements. Elements with a complete structure of LTR-RT were classified as autonomous, whereas incomplete elements were considered nonautonomous. The structures of autonomous *Copia* and *Gypsy* were classified using the domain orders TSD-LTR-PBS-GAG-PROT-INT-RT-RH-PPT-LTR-TSD and TSD-LTR-PBS-GAG-PROT-RT-RH-INT-PPT-LTR-TSD, respectively. Nonautonomous elements contain LARD, TR-GAG, TRIM and BARE-2, classified according to the criteria presented by Kalendar et al. (2004); Chaparro et al. (2015), Witte et al., (2001) and Tanskanen et al. (2007), respectively (Figure 1). The intact LTR-RT elements that were not subject to the previous conditions and defined as *Copia* or *Gypsy* were classified as nonautonomous *Copia* and nonautonomous *Gypsy*, respectively. The unknown element is defined as an intact LTR-RT that contains PBS and PPT or has lost one or both of these elements and does not contain any of the *GAG* and *Pol* domains.

LTR-RT insertion age was determined only for intact LTR-RT elements. A comparison of the 5’ and 3’ semi-identical LTRs for

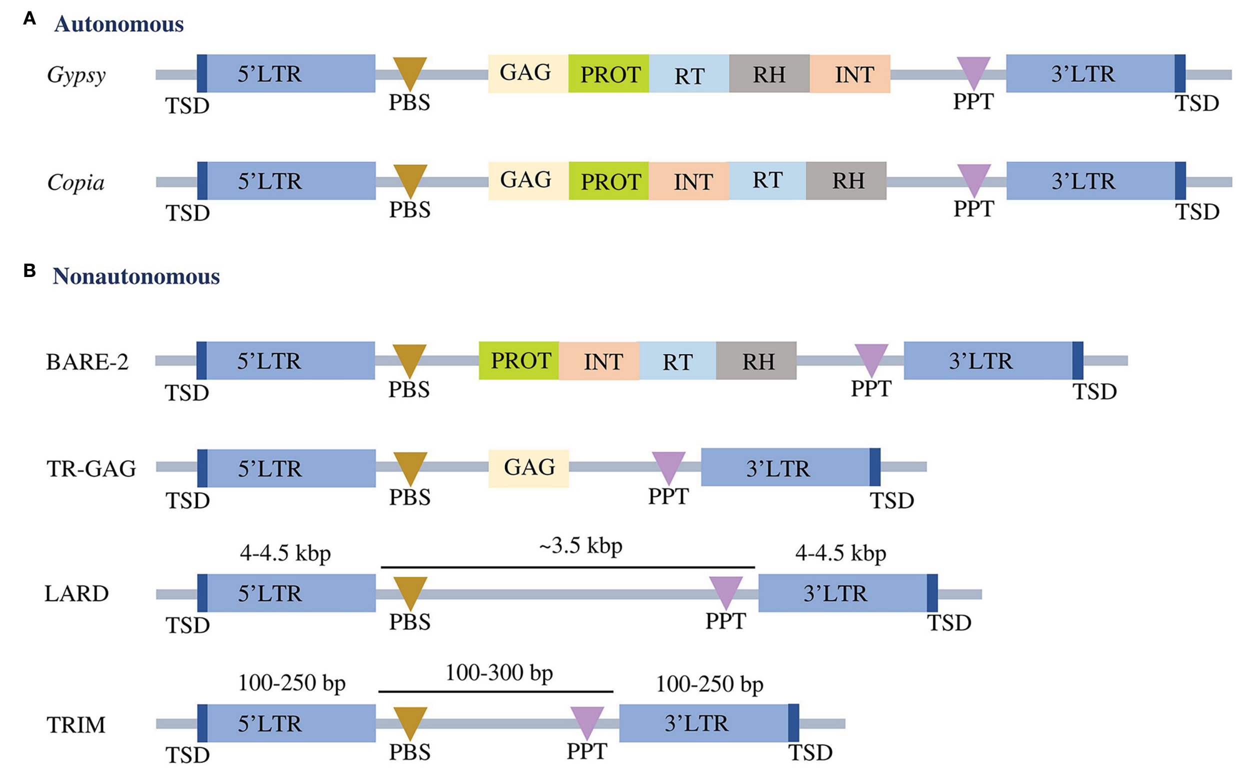


FIGURE 1
Conserved structures of autonomous **(A)** and nonautonomous **(B)** LTR-RTs. LTR refers to long terminal repeats. TSD is the target site duplication. PBS represents primer binding site, GAG represents capsid proteins, PROT represents protease, RT represents reverse transcriptase, RH represents RNase H, INT represents integrase, and PPT represents polypurine tract. The structures are not drawn to scale.

each LTR-RT element was used to calculate the insertion age. This comparative analysis was carried out using ClustalW (Thompson et al., 2003) to obtain a local alignment of the two LTRs. The estimation of the insertion age based on the method of Tajima and Nei (Tajima and Nei, 1984) and the Kimura-2 parameter model (Kimura, 1980) was performed with REANNOTATE (Pereira, 2008). Nucleotide substitutions per site (K) between LTRs were estimated using the Kimura-2 parameter model. The estimated age was calculated using the formula T= K/2r. The evolutionary rate (r) of 1.3×10^{-8} substitutions per site per year was used for grass plants (Kimura, 1980; Ma and Bennetzen, 2004), whereas a substitution rate of 1.5×10^{-8} was used for other species as reported in the literature (Koch et al., 2000; Gonzalez and Deyholos, 2012; Marcon et al., 2015). Here, we used a substitution rate of 1.5×10^{-8} for plants other than grasses because an average substitution rate is not available for many plants.

Based on genomic position, identified LTR-RTs were classified into LTR-RT-gene chimeras by comparing the start and end coordinates of genes and LRT-RTs within the genome. LTR-RT was considered an LTR-RT-gene chimera if it was within the gene start and end coordinates. A gene ontology was assigned to all genes that contained LTR-RT elements or were in close proximity using STRINGdb (Szklarczyk et al., 2015). Gene enrichment analysis was performed using *Arabidopsis thaliana* and *Medicago truncatula* as model plants. Figure 2 shows the workflow and procedure used in the data analysis. Statistical correlations between plant genome size, LTR-RT length, and insertion age were performed for all diploid plant species with LAI ≥10. LAI is the ratio of the length of intact LTR-RTs to the total LTR length (Ou et al., 2018). Scripts used for data analysis are available on GitHub for public use at https://github.com/agc-bioinformatics/PlantLTRdb.

2.3 Database development

The PlantLTRdb was created as a hub and interactive web interface using a variety of programming languages including Perl, Python, R, MongoDB, PHP, CSS, HTML, and JavaScript. In addition, PlantLTRdb includes an implementation of a simple interface for the software LTR_FINDER (Xu and Wang, 2007). The easy-to-use LTR_FINDER interface allows users to identify LTR-RT elements using the standard input sequence format (FASTA). PlantLTRdb is hosted on a server with 32 GB of memory, 16-core CPUs, and a 10 TB disk; running Linux 5.4.0-

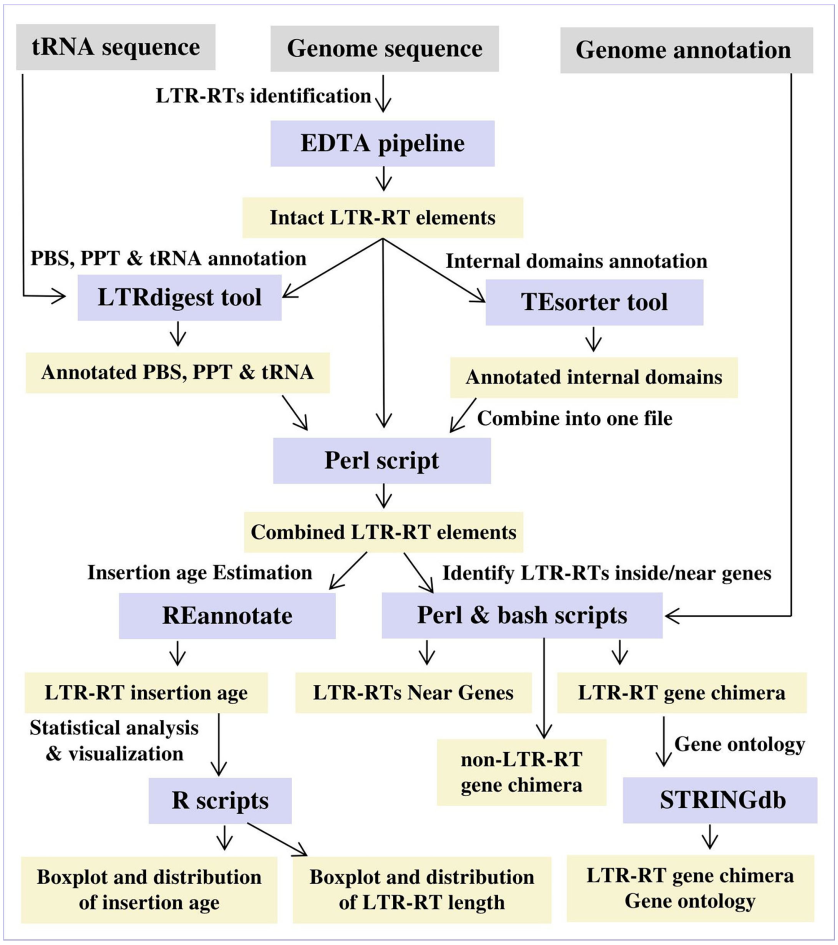


FIGURE 2
The workflow and procedure for identifying and characterizing LTR-RT in 201 plant species.

89-generic x86 64, Apache 2.4, MongoDB and PHP 7.4.3. The online tools require Python (v3.8.10), Perl (v5.30.0), and R (v4.1.2). HTCondor (v9.5.0) is used to manage and schedule submitted tasks and processes. The frontend of the website is built using the Vue (3.2) framework, while the backend is built using Laravel (8.75). Users can select, filter, and visualize available LTR-RT information based on the processed plant species using several online tools. Several tools were used for data visualization, including JBrowse (Buels et al., 2016), ggplot2 (R package), and Google Charts (https://developers.google.com/chart).

3 Results and discussion

3.1 Identification and classification of LTR-RT elements

In the last decade, the complete genome sequences of hundreds of plant species have been published (Mokhtar and Atia, 2018). Access to these extensive data has paved the way for the study of LTR-RTs at the genome level. Over the past decade, LTR-RTs from several plant species have been identified and classified using homology, structural, and *de novo* investigation methods (Wicker et al., 2018; Yi et al., 2018b; Neumann et al., 2019; Mokhtar et al., 2021; Zhou et al., 2021). The development of a unified, well-maintained, effective resource for plant LTR-RT is a prerequisite to support progress in understanding the functional effects of these factors on genomic structure and functionality. Ou et al. (2018) reported that more intact LTR-RTs were identified from complete genome assemblies compared with draft genomes. In addition, other reports indicated a correlation between sequencing technique and the number of intact LTR-RTs detected (Al-Dous et al., 2011; Jiao et al., 2017; Ou et al., 2018). These reports suggest that more intact LTR-RTs are detected from continuous genome assembly. In the current study, we used only annotated genomes because annotation is required to identify LTR-RT-gene chimeras and LTR-RT nearby genes.

We used established and validated LTR-RT tools to create a workflow for the identification and classification of LTR-RTs in different plant species (Figures 1, 2). The resulting data were used to create a user-friendly public resource for intact LTR-RT in plants. The PlantLTRdb was developed by processing the entire genomic sequences of 201 plant species, totaling 150.18 Gbp. These sequences represent genomic data from 3,079,469 pseudomolecules/scaffolds. However, 6 genomes, including *Chloropicon primus*, *Cyanidioschyzon merolae*, *Galdieria sulphuraria*, *Genlisea aurea*, *Micromonas commoda*, and *Monoraphidium neglectum*, failed the EDTA filtering step and no intact LTR-RTs were found. These genomes were excluded from the analysis and only 195 plant species that passed the filtering step were used for further analysis.

As a result, 2,722,415 LTR candidates were identified in the studied species. The identified LTRs were filtered based on the intact LTR-RT structure (TSD-LTR-[internal sequence]-LTR-TSD) and candidates with missing components were excluded from further analysis. Only 528,891 candidates passed filtering and had intact LTR-RT structures. The candidates that include nested LTRs or other TEs insertions were excluded using LTR_retriever module 6 (Ou et al., 2018). The remaining 520,194 elements were then annotated using LTRdigest and TEsorter to identify PBS, PPT, *GAG*, and *Pol* regions and classify them into lineages. Table S2 shows the lineages and their corresponding totals in the database. Based on the structure of the autonomous LTR-RT, the identified intact LTR-RTs were classified into putative 29,462 autonomous and 490,732 non-autonomous LTR-RTs. The 29,462 autonomous LTR-RTs include 10,286 from the *Gypsy* superfamily and 19,176 from the *Copia* superfamily. Further analyses were performed to classify non-autonomous elements using the criteria presented by Witte et al., (2001); Kalendar et al. (2004); Tanskanen et al. (2007); Chaparro et al. (2015). In addition, incomplete *Copia* and *Gypsy* elements were classified as nonautonomous. All non-autonomous elements not subject to any of the above structures were defined as unknown elements. Based on these criteria, 490,732 nonautonomous elements were divided into 224,906 *Gypsy*, 218,414 *Copia*, 1,768 BARE-2, 3,147 TR-GAG, and 42,497 unknown, while LARD and TRIM elements were not present (Table S3). The number of LTR-RTs detected ranged from 1 (*Micractinium conductrix*) to 33,245 (*Aegilops tauschii*). After excluding outliers by boxplot analysis of LTR-RT length for all 195 plant species, the results showed that the minimum, first quartile, median, third quartile, and maximum lengths were 1,140, 5,398, 8,273, 11,061, and 19,555, respectively (Figure 3). The analysis also showed that most plant species had a wide range of LTR-RT lengths (Figure S1).

The differences in the length of LTR-RT are primarily due to divergence in the size of LTR and the existence and size of spacer regions between internal domains rather than *GAG/Pol* coding regions (Zhou et al., 2017). Figure 4 shows that after excluding outliers by boxplot analysis, the first quartile, median, and third quartile of the autonomous LTR-RTs were 4,920, 5,267, 8,757 bp for *Copia* and 6,337, 10,420, 11,873 bp for *Gypsy*, respectively. For nonautonomous LTR-RTs, the first quartile, median, and third quartile were 4,971, 6,522, 9,213 bp for *Copia*, 7,832, 10,237, 12,340 bp for *Gypsy*, 3,608, 4,837, 8,054 bp for TR-GAG, 4,220, 4,555, 5,310 bp for BARE-2 and 5,746, 7,112, 7,985 bp for unknown, respectively. The first quartile, median, third quartile, minimum, and maximum of LTR-RT length for the 195 plant species were listed in Table S4. In the present study, based on the parameters used to identify LTR-RTs, and without excluding outliers by boxplot analysis, the length of autonomous and nonautonomous *Copia* elements ranged from 1,140 to 25,333 bp, whereas the length of *Gypsy* ranged from 1,182 to 25,575 bp. This is consistent with previous studies by Du et al. (2010b); Ma et al. (2019); Neumann et al. (2019); Li et al. (2022), which found that a number of *Gypsy* elements were smaller than 4kb and a number of *Copia* elements were larger than 15kb. For example, Li et al. (2022) examined LTR-RTs of 16 *Cucurbitaceae* species and reported that the length of LTR-RTs ranged from 1,173 to 28,350 bp. Boxplot analysis showed *Gypsy* elements smaller than 4kb and *Copia* elements larger than 15kb.

The insertion age of the plant species studied reflects the evolutionary rate associated with the uniqueness of their genomic content. This evolutionary difference can help researchers

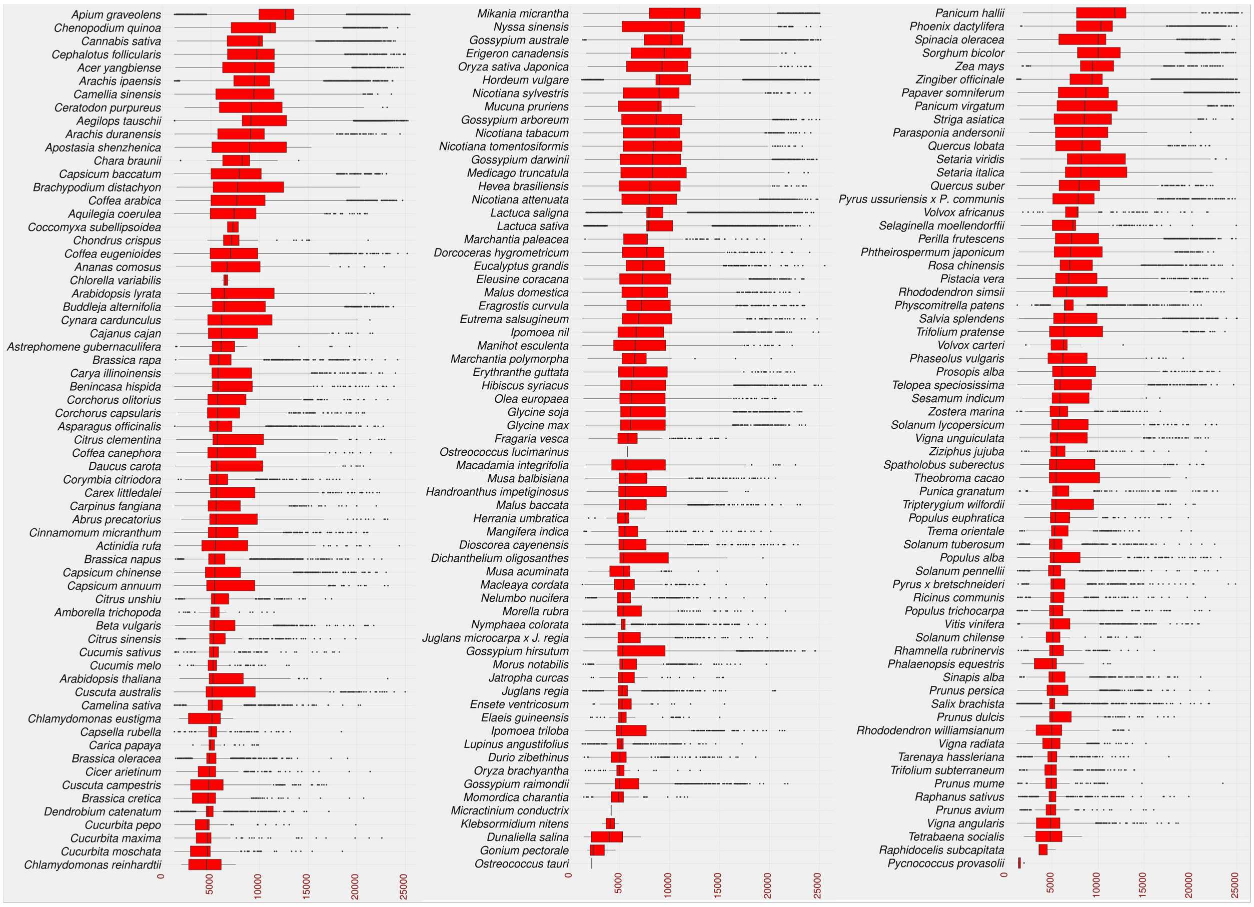


FIGURE 3
Statistical overview of the LTR-RT length by base pair. The boxplot of the LTR-RT length in the studied plant species. Species sorted in descending order by the median value. The LTR-RT lengths are shown in bp scale (x-axis).

understand the relationship between genetic and phenotypic variation (Barghini et al., 2014). Recently, TE-family has proven particularly useful in understanding the evolutionary mechanisms involved in species divergence (Liu and Yang, 2014). Transposition of LTR-RTs results in identical sequences of 5' and 3' LTRs, and the accumulation of nucleotide substitutions/divergences between the two arms of LTR-RTs is used to calculate the insertion age (da Costa et al., 2019; Mokhtar et al., 2021). Figure 5 shows that the maximum assumed age after exclusion of outliers by boxplot analysis was 5.1 million years (MY) for *Dichanthelium oligosanthes*. Several plant

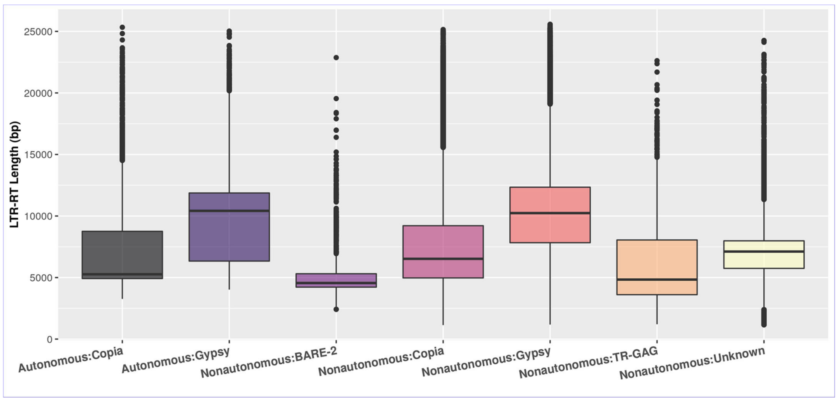


FIGURE 4
The boxplot of the LTR-RT length in the studied plant species of the autonomous and nonautonomous elements. The LTR-RT lengths are shown in bp scale (Y-axis).

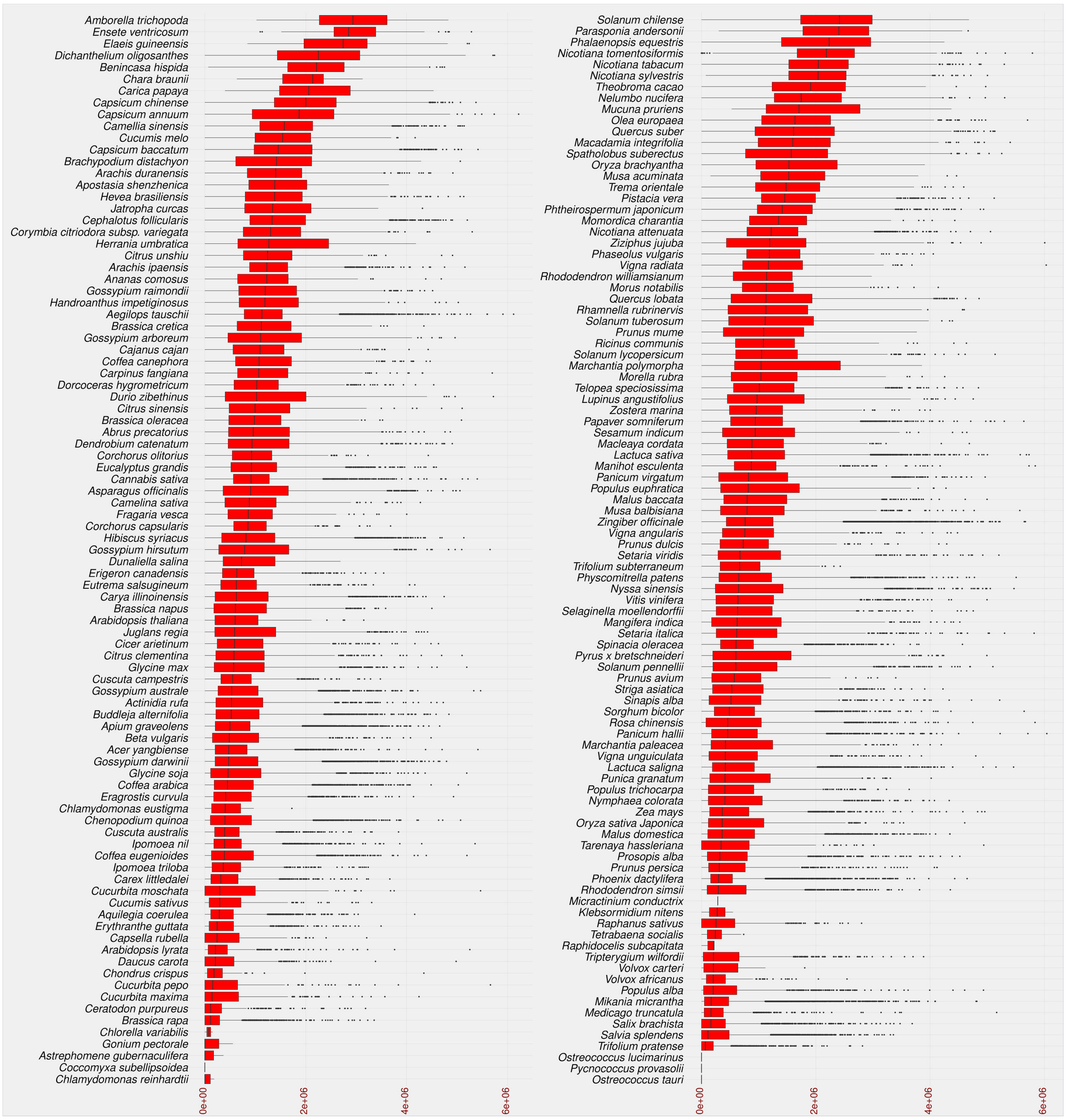


FIGURE 5

Statistical overview of the age of LTR-RT insertion in the studied plant species using boxplot analysis. Species are sorted in descending order by median value. Values for age of LTR-RT insertion are in years (x-axis).

species have a high rate of young LTR-RTs in their genome, such as *Phoenix dactylifera*, *Cucumis sativus*, *Arabidopsis lyrata*, *Daucus carota*, *Medicago truncatula*, and *Brassica rapa*. In addition, other plant species show a homogeneous collection of LTR-RTs with different insertion ages, such as *Solanum chilense*, *Carica papaya*, *Theobroma cacao*, *Capsicum annuum*, and *Mucuna pruriens*. Our results are consistent with previous findings for some of these plant species. According to a previous analysis, most LTR-RTs identified in *Medicago truncatula* are relatively recent and were inserted in the last 0.52 MY, with possibly more than 10 million bp lost due to deletion of LTR elements and removal of full-length structures (Wang and Liu, 2008). Ibrahim et al. (2021) examined LTR-RTs in numerous palm genomes and concluded that the *Elaeis guineensis* genome has undergone several LTR-RT events with different temporal patterns of transposition activity. In our study, the first quartile of insertion age in *Ensete ventricosum*, *Amborella trichopoda* and *Elaeis guineensis* shows the highest values (1.9-2.5 MY), while the maximum third quartile has the lowest values in *Chlamydomonas reinhardtii*, *Chlorella variabilis*, *Astrephomene gubernaculifera*, *Trifolium pratense*, and *Raphidocelis subcapitata*

(0.11-0.21 MY). The third quartile of LTR-RT insertion age ranges from 0.11 MY in *Chlamydomonas reinhardtii* to 3.6 MY in *Amborella trichopoda*.

The different distribution of LTR-RT insertion age among the studied species suggests that there is a relationship between the overall insertion age and the insertion age of each LTR-RT type. Most plant species with a high first quartile of insertion age, such as *Elaeis guineensis*, have a wide range of insertion age values, whereas those with a high rate of young LTR-RTs have a narrow range of insertion age. On the other hand, some species such as *Capsicum annuum* and *Ziziphus jujuba* have narrow and wide ranges of insertion age in different parts of the LTR-RT distribution (Figure S2).

The LAI was estimated for studied plant genomes using the LAI program within LTR_retriever (Ou et al., 2018). Only the diploid plant species with LAI ratio greater than 10 (50 species) were subjected to correlation analysis between genome size, LTR-RT estimated insertion age and LTR-RT length. The correlation between genome size and LTR-RT length was 0.4 (R), with a p-value of 0.0035, indicating that it was significant. This suggests that although there is a clear correlation between genome size and intact LTR-RT length, the effect of genome size on length is weak (Figure 6A). The correlation between the genome size and LTR-RT insertion age was 0.12 (R), with a p-value of 0.43 indicating a non-significant association. This suggests that there is no relationship between genome size and insertion age. It also suggests that genome size has little or no effect on LTR-RT insertion age (Figure 6B). Some plant species, such as *Amborella trichopoda* and *Elaeis guineensis* have medium or small genome sizes with a high LTR-RT insertion age. The correlation between total LTR-RTs and genome size is 0.85 (R), with a p-value of 5.2×10^{-15}, showing a strong positive correlation between them (Figure 6C).

LTR-RT transposition can affect the expression of both housed LTR-RT-gene chimeras and nearby genes. LTR-RTs influence genes through the processes of movement, duplication, and recombination constructing or modifying gene structure (Zhao et al., 2016). Further analysis was performed on the plant species studied to identify LTR-

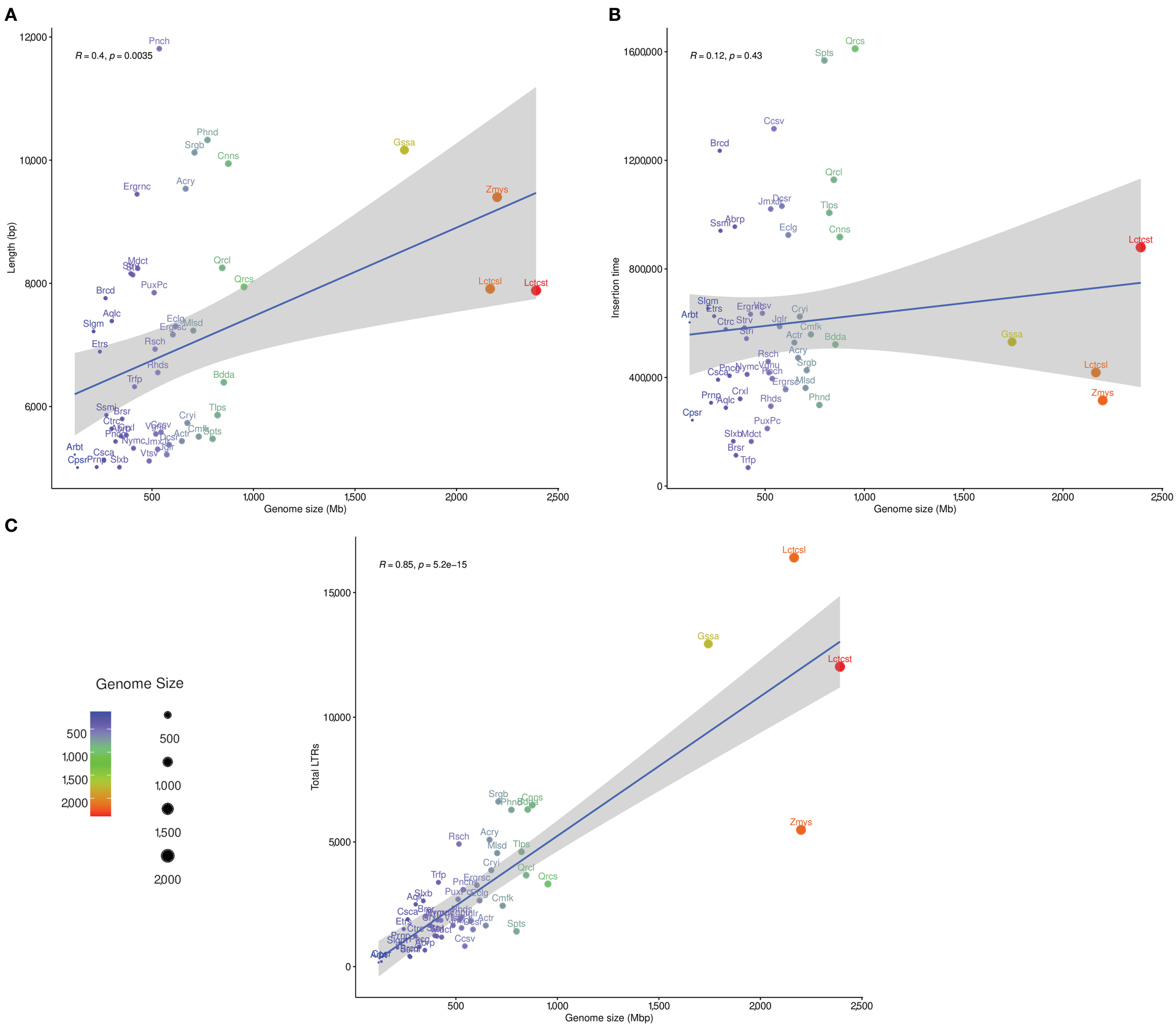


FIGURE 6
The statistical correlation between plant genome size, and LTR-RT length **(A)**, insertion age **(B)**, and total number of LTR-RT **(C)**.

RTs located within or near genes. In several plant species, the interaction of LTR-RT activity in genic regions could result in a hybrid of LTR-RT-gene structures or LTR-RT-gene chimeras (Jiang et al., 2004; Wang et al., 2006). A total of 37,206 LTR-RTs were classified as LTR-RT-gene chimeras in all species studied, and 11,844 LTR-RTs were found within pseudo-genes (Table S5). In addition, 300,613 genes were up to 10 Kbp from LTR-RTs. Table S5 shows that of the 300,613 genes, 50,026 were located up to 1 Kbps away and 250,587 were within 1 to 10 Kbp. The *Copia* superfamily, found within genes, was more prevalent than *Gypsy* elements in the current study, consistent with previous studies in some plant species (Bennetzen, 1996; Rossi et al., 2001; Lockton and Gaut, 2009; Mokhtar et al., 2021). Using the gene ontology of two different model plant species, gene enrichment analysis was performed for genes located within or near LTR-RTs in the plant genomes studied. Gene ontologies such as binding, cell membrane, and catalytic activity were highly enriched in LTR-RTs associated genes. The high frequency of genes associated with catalytic activity and binding may be related to the biological activity of LTR-RTs within the plant genome to promote gene expression, such as those associated with stress response (Bui and Grandbastien, 2012).

3.2 Technical validation

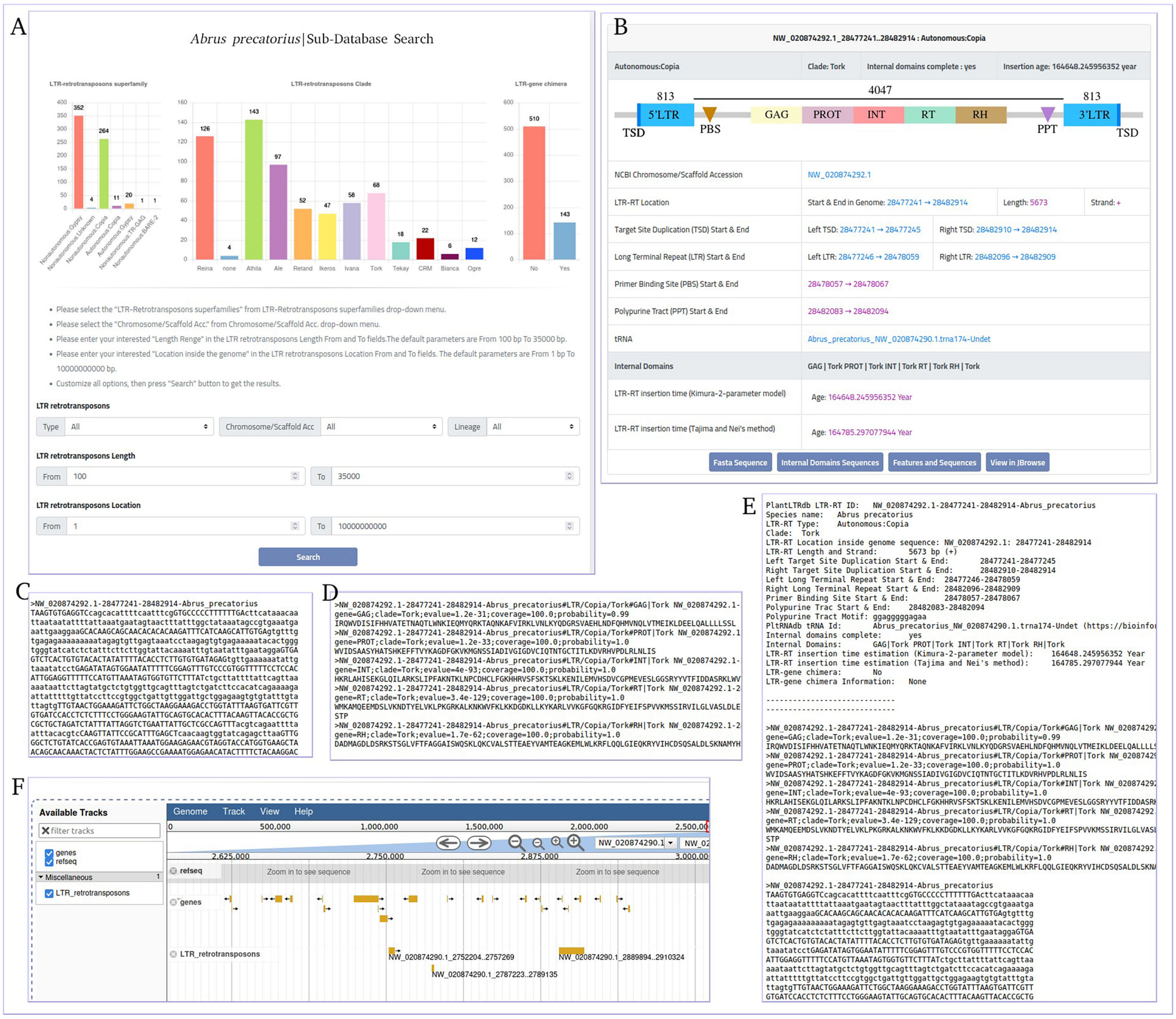
To verify the quality of the identified intact LTR-RTs in the current study, a manually curated LTR-RTs library of rice (*Oryza sativa. ssp. japonica*) was used for comparison with our *Oryza sativa* dataset. *Oryza sativa* was selected for comparison because its genome sequence is well structured and arranged in chromosomes and has a high LAI score of 22.41. The curated rice library was presented in a previous study by Ou and Jiang (2017) and included 897 LTR-RTs elements. TEsorter (Zhang et al., 2022) was used to annotate the *GAG*- and *Pol* domains of this library (897 LTR-RTs) using the REXdb database (Neumann et al., 2019) based on a unified LTR-RTs classification rule (80-80-80) as proposed by Wicker et al. (2007). Of the 897 LTR-RTs, 242 elements have a complete *GAG*- and *Pol* domains, which were used for comparison with the currently identified LTR-RTs of *Oryza sativa* using the OrthoFinder tool (Emms and Kelly, 2019). In our results, the LTR-RTs dataset of *Oryza sativa* contains 1,496 LTR-RT elements divided into 54 autonomous and 1,442 nonautonomous. Using OrthoFinder, 1,114 elements were assigned to the curated library ortho-groups (Table S6). The remaining 382 elements also have strong evidence as they include 204 elements that have all necessary domains for their transposition, 17 contain four domains, six contain three domains, 25 contain two domains, 97 contain one domain, and 33 elements are unknown (Table S6). This comparison verifies the reliability of the identified LTR-RTs in the current investigation.

3.3 PlantLTRdb as a resource for intact LTR-RTs in plants

The data generated by the LTR-RT analysis workflow during the processing of 195 plant genomes were used to build a flexible, efficient, and well-maintained database of LTR-RT elements and associated information in the plant species studied (Figure S3). Our plant LTR-RTs database (PlantLTRdb) is accessible through an easy-to-use web interface. Through the public website portal, users can search, visualize, BLAST, and analyze plant LTR-RT elements. The PlantLTRdb search dropdown menu provides users with access to two separate search pages. The first allows a general search for LTR-RTs from all plant species studied, and the second allows a search for detected LTR-RT-gene chimeras and nearby genes. On the LTR-RTs general search page, users can view bar charts summarizing the number of identified LTR-RTs in the species searched. In addition, several search options are available, including searching by LTR-RT superfamily, pseudomolecules/scaffolds, LTR-RT length, LTR-RT position in the genome, and searching by all of the above criteria (Figure 7A). All results are displayed on a separate page with additional information about LTR-RTs including, NCBI accession, LTR-RT position in genome, length, type, target site duplication, long terminal repeat, primer binding site, polypurine tract, tRNA, internal domains, LTR-RT insertion age, JBrowse link and download buttons for LTR-RT FASTA file, internal domains FASTA file, and LTR-RT features (Figures 7B–F).

The LTR-RT gene interaction search page provides users with bar charts summarizing the number of detected LTR-RT-gene chimeras and neighboring genes within the searched species. In addition, several search options are available, including searching by LTR-RT superfamily, gene category, NCBI gene ID/locus tag, and/or protein ID. Generated results include NCBI gene ID, gene start and end, gene description, distance between LTR-RT and gene, protein ID, gene ontology, superfamily, LTR-RT position in genome, length, type, target site duplication, long terminal repeat, primer binding site, polypurine tract, tRNA, internal domains, LTR-RT insertion age, JBrowse link and download buttons for LTR-RT FASTA sequence, internal domains FASTA sequence and LTR-RT features (Figure S4). Users can visualize LTR-RTs at the genome level with JBrowse, which is integrated into PlantLTRdb. The JBrowse page displays information about the selected plant species, such as the reference genome sequence, genome annotations (genes), and any LTR-RT coordinates that have been identified. Users can view the details of LTR-RT elements and evaluate LTR-RT nearby genes. The statistics page provides the user with interactive graphs for the LTR-RTs superfamily statistics, LTR-RT-gene chimera statistics, LTR-RTs clade, statistical overview of the LTR insertion age, the LTR-RT length by bp, and the gene ontology of the LTR-RT-gene chimeras and nearby genes for each plant species. Several statistical plots are generated to cover all aspects of the results (Figure 8).

PlantLTRdb provides powerful tools for searching for specific LTR-RT elements in processed genomic data or user-supplied data. Our online BLASTN allows users to align their LTR-RT sequences against the LTR-RTs of the specified plant species. Results from this tool include the known BLAST results and a link to similar LTR-RT details stored in PlantLTRdb, such as pseudomolecules/scaffold accession, LTR-RT genomic position, length, insertion age, sequence, and JBrowse profile link. In addition, an online version of LTR_Finder (Xu and Wang, 2007) has been integrated into the PlantLTRdb platform. This tool allows users to explore LTR-RTs in



- Please select the "LTR-Retrotransposons superfamilies" from LTR-Retrotransposons superfamilies drop-down menu.
- Please select the "Chromosome/Scaffold Acc." from Chromosome/Scaffold Acc. drop-down menu.
- Please enter your interested "Length Renge" in the LTR retrotransposons Length From and To fields.The default parameters are From 100 bp To 35000 bp.
- Please enter your interested "Location inside the genome" in the LTR retrotransposons Location From and To fields. The default parameters are From 1 bp To 10000000000 bp.
- Customize all options, then press "Search" button to get the results.

LTR retrotransposons

Type: All | Chromosome/Scaffold Acc: All | Lineage: All

LTR retrotransposons Length

From: 100 | To: 35000

LTR retrotransposons Location

From: 1 | To: 10000000000

NW_020874292.1_28477241..28482914 : Autonomous:Copia

Autonomous:Copia	Clade: Tork	Internal domains complete : yes	Insertion age: 164648.245956352 year

NCBI Chromosome/Scaffold Accession	NW_020874292.1		
LTR-RT Location	Start & End in Genome: 28477241 → 28482914	Length: 5673	Strand: +
Target Site Duplication (TSD) Start & End	Left TSD: 28477241 → 28477245	Right TSD: 28482910 → 28482914	
Long Terminal Repeat (LTR) Start & End	Left LTR: 28477246 → 28478059	Right LTR: 28482096 → 28482909	
Primer Binding Site (PBS) Start & End	28478057 → 28478067		
Polypurine Tract (PPT) Start & End	28482083 → 28482094		
tRNA	Abrus_precatorius_NW_020874290.1.trna174-Undet		
Internal Domains	**GAG | Tork PROT | Tork INT | Tork RT | Tork RH | Tork**		
LTR-RT insertion time (Kimura-2-parameter model)	Age: 164648.245956352 Year		
LTR-RT insertion time (Tajima and Nei's method)	Age: 164785.297077944 Year		

```
>NW_020874292.1-28477241-28482914-Abrus_precatorius
TAAGTGTGAGGTCcagcacattttcaatttcgGTGCCCCCTTTTTTGActtcataaacaa
ttaataatattttattaaatgaatagtaactttatttggctataaatagccgtgaaatga
aattgaaggaaGCACAAGCAGCAACACACACAAGATTTCATCAAGCATTGTGAgtgtttg
tgagagaaaaaaaaatagagtgttgagtaaatcctaagagtgtgagaaaaatacactggg
tgggtatcatctctatttcttcttggtattacaaaatttgtaatatttgaataggaGTGA
GTCTCACTGTGTACACTATATTTTACACCTCTTGTGTGATAGAGtgttgaaaaaatattg
taaatatcctGAGATATAGTGGAATATTTTTCGGAGTTTGTCCCGTGGTTTTTCCTCCAC
ATTGGAGGTTTTTCCATGTTAAATAGTGGTGTTCTTTATctgcttattttattcagttaa
aaataattcttagtatgctctgtggttgcagtttagtctgatcttccacatcagaaaaga
attatttttgttatccttccgtggctgattgttggattgctggagaagtgtgtatttgta
ttagtgTTGTAACTGGAAAGATTCTGGCTAAGGAAAGACCTGGTATTTAAGTGATTCGTT
GTGATCCACCTCTCTTTCCTGGGAAGTATTGCAGTGCACACTTTACAAGTTACACCGCTG
CGCTGCTAGATCTATTTATTAGGTCTGAATTATTGCTCGCCAGTTTacgtcagaatttta
atttacacgtcCAAGTTATTCCGCATTTGAGCTcaacaagtggtatcagagcttaaGTTG
GGCTCTGTATCACCGAGTGTAAATTAAATGGAAGAGAACGTAGGTACCATGGTGAAGCTA
ACAGCAACAAACTACTCTATTTGGAAGCCGAAAATGGAGAACATACTTTTCTACAAGGAC
```

```
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#GAG|Tork NW_020874292.1-
gene=GAG;clade=Tork;evalue=1.2e-31;coverage=100.0;probability=0.99
IRQWVDISIFHHVATETNAQTLWNKIEQMYQRKTAQNKAFVIRKLVNLKYQDGRSVAEHLNDFQHMVNQLVTMEIKLDEELQALLLLSSL
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#PROT|Tork NW_020874292.1
gene=PROT;clade=Tork;evalue=1.2e-33;coverage=100.0;probability=1.0
WVIDSAASYHATSHKEFFTVYKAGDFGKVKMGNSSIADIVGIGDVCIQTNTGCTITLKDVRHVPDLRLNLIS
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#INT|Tork NW_020874292.1-
gene=INT;clade=Tork;evalue=4e-93;coverage=100.0;probability=1.0
HKRLAHISEKGLQILARKSLIPFAKNTKLNPCDHCLFGKHHRVSFSKTSKLKENILEMVHSDVCGPMEVESLGGSRYYVTFIDDASRKLWV
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#RT|Tork NW_020874292.1-2
gene=RT;clade=Tork;evalue=3.4e-129;coverage=100.0;probability=1.0
WMKAMQEEMDSLVKNDTYELVKLPKGRKALKNKWVFKLKKDGDKLLKYKARLVVKGFGQKRGIDFYEIFSPVVKMSSIRVILGLVASLDLE
STP
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#RH|Tork NW_020874292.1-2
gene=RH;clade=Tork;evalue=1.7e-62;coverage=100.0;probability=1.0
DADMAGDLDSRKSTSGLVFTFAGGAISWQSKLQKCVALSTTEAEYVAMTEAGKEMLWLKRFLQQLGIEQKRYVIHCDSQSALDLSKNAMYH
```

```
PlantLTRdb LTR-RT ID:   NW_020874292.1-28477241-28482914-Abrus_precatorius
Species name:   Abrus precatorius
LTR-RT Type:    Autonomous:Copia
Clade:  Tork
LTR-RT Location inside genome sequence: NW_020874292.1: 28477241-28482914
LTR-RT Length and Strand:       5673 bp (+)
Left Target Site Duplication Start & End:       28477241-28477245
Right Target Site Duplication Start & End:      28482910-28482914
Left Long Terminal Repeat Start & End:  28477246-28478059
Right Long Terminal Repeat Start & End: 28482096-28482909
Primer Binding Site Start & End:        28478057-28478067
Polypurine Trac Start & End:    28482083-28482094
Polypurine Tract Motif: ggaggggagaa
PltRNAdb tRNA Id:       Abrus_precatorius_NW_020874290.1.trna174-Undet (https://bioinfor
Internal domains complete:      yes
Internal Domains:       GAG|Tork PROT|Tork INT|Tork RT|Tork RH|Tork
LTR-RT insertion time estimation (Kimura-2-parameter model):    164648.245956352 Year
LTR-RT insertion time estimation (Tajima and Nei's method):     164785.297077944 Year
LTR-gene chimera:       No
LTR-gene chimera Information:   None

---------------------------
---------------------------

>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#GAG|Tork NW_02087429
gene=GAG;clade=Tork;evalue=1.2e-31;coverage=100.0;probability=0.99
IRQWVDISIFHHVATETNAQTLWNKIEQMYQRKTAQNKAFVIRKLVNLKYQDGRSVAEHLNDFQHMVNQLVTMEIKLDEELQALLLLS
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#PROT|Tork NW_02087429
gene=PROT;clade=Tork;evalue=1.2e-33;coverage=100.0;probability=1.0
WVIDSAASYHATSHKEFFTVYKAGDFGKVKMGNSSIADIVGIGDVCIQTNTGCTITLKDVRHVPDLRLNLIS
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#INT|Tork NW_02087429
gene=INT;clade=Tork;evalue=4e-93;coverage=100.0;probability=1.0
HKRLAHISEKGLQILARKSLIPFAKNTKLNPCDHCLFGKHHRVSFSKTSKLKENILEMVHSDVCGPMEVESLGGSRYYVTFIDDASRK
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#RT|Tork NW_020874292.
gene=RT;clade=Tork;evalue=3.4e-129;coverage=100.0;probability=1.0
WMKAMQEEMDSLVKNDTYELVKLPKGRKALKNKWVFKLKKDGDKLLKYKARLVVKGFGQKRGIDFYEIFSPVVKMSSIRVILGLVASL
STP
>NW_020874292.1-28477241-28482914-Abrus_precatorius#LTR/Copia/Tork#RH|Tork NW_020874292.
gene=RH;clade=Tork;evalue=1.7e-62;coverage=100.0;probability=1.0
DADMAGDLDSRKSTSGLVFTFAGGAISWQSKLQKCVALSTTEAEYVAMTEAGKEMLWLKRFLQQLGIEQKRYVIHCDSQSALDLSKNA

>NW_020874292.1-28477241-28482914-Abrus_precatorius
TAAGTGTGAGGTCcagcacattttcaatttcgGTGCCCCCTTTTTTGActtcataaacaa
ttaataatattttattaaatgaatagtaactttatttggctataaatagccgtgaaatga
aattgaaggaaGCACAAGCAGCAACACACACAAGATTTCATCAAGCATTGTGAgtgtttg
tgagagaaaaaaaaatagagtgttgagtaaatcctaagagtgtgagaaaaatacactggg
tgggtatcatctctatttcttcttggtattacaaaatttgtaatatttgaataggaGTGA
GTCTCACTGTGTACACTATATTTTACACCTCTTGTGTGATAGAGtgttgaaaaaatattg
taaatatcctGAGATATAGTGGAATATTTTTCGGAGTTTGTCCCGTGGTTTTTCCTCCAC
ATTGGAGGTTTTTCCATGTTAAATAGTGGTGTTCTTTATctgcttattttattcagttaa
aaataattcttagtatgctctgtggttgcagtttagtctgatcttccacatcagaaaaga
attatttttgttatccttccgtggctgattgttggattgctggagaagtgtgtatttgta
ttagtgTTGTAACTGGAAAGATTCTGGCTAAGGAAAGACCTGGTATTTAAGTGATTCGTT
GTGATCCACCTCTCTTTCCTGGGAAGTATTGCAGTGCACACTTTACAAGTTACACCGCTG
```

FIGURE 7
Abrus precatorius sub-database search page as an example of PlantLTRdb search. **(A)** general search page, **(B)** example of general search results, **(C)** LTR-RT FASTA sequence, **(D)** internal domains FASTA sequence, **(E)** LTR-RT features, and **(F)** JBrowse example.

their own genomic data. The LTR_FINDER webserver (http://tlife.fudan.edu.cn/ltr_finder/) has been decommissioned for unknown reasons, and only the local version is available for use. LTR_FINDER will help many researchers to investigate LTR-RT elements using their genomic data.

3.4 Comparison with other TEs databases

Existing databases provide useful information on LTR-RTs in various plant genomes, but they still lack important details and features (Table 1). For example, Repbase (Bao et al., 2015) is a dataset for eukaryotic repetitive sequences including LTR-RTs. The LTR-RTs have been classified into superfamilies and lineages, but details of internal structure are lacking. In addition, Repbase requires a subscription for access. REXdb (Neumann et al., 2019) contains LTR-RTs protein domains from 80 species. The LTR-RT elements are detected using LTR_FINDER and the data are not available as a stand-alone database, but can only be downloaded from the RepeatExplorer website. In addition, PlaNC-TE (Pedro et al., 2018) contains only overlapping regions between TEs and non-coding RNAs (ncRNAs) in 40 plant species. RepetDB (Amselem et al., 2019) contains TEs for 23 species. MnTEdb (Ma et al., 2015), DPTEdb (Li et al., 2016), ConTEdb (Yi et al., 2018b), SPTEdb (Yi et al., 2018a), and CicerSpTEdb (Mokhtar et al., 2021) are databases containing only TEs from 1 to 8 genomes, while MASiVEdb (Bousios et al., 2012), GrTEdb (Xu et al., 2017), and RetrOryza (Chaparro et al., 2006) are no longer available.

PlantRep (Luo et al., 2022), InpactorDB (Orozco-Arias et al., 2021), and APTEdb (Pedro et al., 2021) databases were recently published and contain TEs of 459, 195, and 67 plant species, respectively. Although these databases provide useful information on plant LTRs, they lack important features such as the classification of LTR as intact LTR-RT and into autonomous and nonautonomous elements. InpactorDB, for example, has only a single function, which is to search using a set of

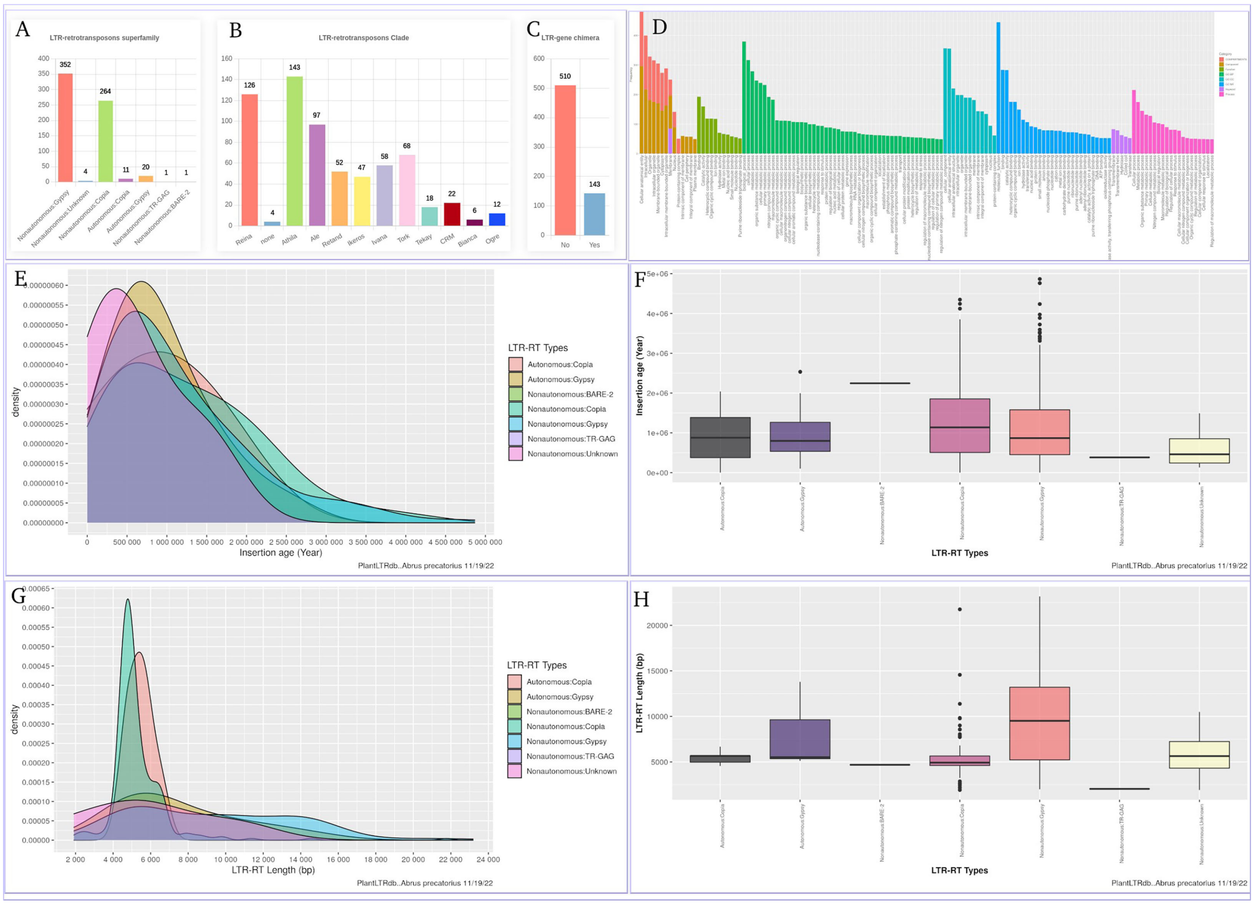


FIGURE 8
An example of the statistics page using Abrus precatorius. **(A)** LTR-RTs superfamily statistics, **(B)** LTR-RTs clade, **(C)** LTR-RT-gene chimera statistics, **(D)** Gene ontology, **(E, F)** statistical overview of the LTR insertion age, **(G, H)** statistical overview of the LTR-RT length by bp.

TABLE 1 Comparison of online TE databases based on the number of species and their LTR-RT related features.

Database	Species	Search	Insertion time	LTR-RT in/near genes	Buit-in tools	Availability
PlantLTRdb	195	✓	✓	✓	JBrowse, Blast, LTR Finde	https://bioinformatics.um6p.ma/PlantLTRdb
REXdb	80	×	×	×	–	http://repeatexplorer.org/?pageid=918
TREP	60	✓	×	×	Blast	http://botserv2.uzh.ch/kelldata/trep-db
PlaNC-TE	40	✓	×	×	JBrowse	http://planc-te.cp.utfpr.edu.br/
RepetDB	23	✓	×	×	Blast	https://urgi.versailles.inra.fr/repetdb/begin.do
MASiVEdb	11	✓	×	×	–	No longer Available
DPTEdb	8	✓	×	×	JBrowse, Blast, GetORF, Hmm, Cut sequence	http://genedenovoweb.ticp.net:81/DPTEdb
CicerSpTEdb	3	✓	×	✓	JBrowse	http://cicersptedb.easyomics.org
ConTEdb	3	✓	×	×	JBrowse, Blast, GetORF, Hmm, Cut sequence	http://genedenovoweb.ticp.net:81/conTEdb

(Continued)

TABLE 1 Continued

Database	Species	Search	Insertion time	LTR-RT in/near genes	Buit-in tools	Availability
SPTEdb	3	✓	×	×	JBrowse, Blast, GetORF, Hmm, Cut sequence	http://genedenovoweb.ticp.net:81/SPTEdb
GrTEdb	1	✓	×	×	–	No longer Available
RetrOryza	1	✓	×	×	–	No longer Available
MnTEdb	1	✓	×	×	JBrowse, Blast, GetORF, Hmm, Cut sequence	https://morus.swu.edu.cn/mntedb/

parameters, and does not provide visualization, bulk downloading, or built-in tools for manipulating the data. APTEdb contains a small number of plant genomes and only allows downloading the LTRs of each genome as gff and fasta files. PlantRep contains genomes that are not annotated and does not include tools such as visualization and searching. Because gene annotation information is limited, users of the existing databases may not be able to understand the impact of LTR-RTs on the plant genome and their association with specific genes or biological processes. In addition to limitations in the data, some databases have limited features and are rarely maintained and updated. Finally, most databases do not include an online LTR-RT identification tool that can be used to analyze user-specific data. Such tools would benefit those attempting to annotate newly sequenced genomic data. Compared to previously published LTR-RT databases, PlantLTRdb has unique features that can contribute to the understanding of the structural variations and organization of LTR-RTs in the genome.

4 Conclusions and future directions

PlantLTRdb is a hub portal of LTR-RTs in plant species. For all plant species studied, various analyzes were performed to identify, characterize, and annotate LTR-RTs, as well as to estimate insertion ages, detect LTR-RT-gene chimeras, and determine nearby genes. The PlantLTRdb contains 520,194 intact LTR-RTs, including 29,462 autonomous and 490,732 nonautonomous LTR-RTs. In addition, the website portal allows users to search, visualize, BLAST, and analyze plant LTR-RT elements. PlantLTRdb will be continuously updated with newly annotated genomes. PlantLTRdb is an important database that can contribute to the understanding of structural variations, genome organization, and the development of LTR-RT target markers for molecular plant breeding.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

MMM and AE Conceptualization, Formal analysis. MMM, AMA and AE. Data curation, Methodology, Visualization, wrote and reviewed the manuscript. AE. Supervision and Resources. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors acknowledge the African Supercomputing Center at Mohammed VI Polytechnic University for supercomputing resources (https://ascc.um6p.ma/) made available for conducting the research reported in this paper.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1134627/full#supplementary-material

References

Akakpo, R., Carpentier, M.-C., Ie Hsing, Y., and Panaud, O. (2020). The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytol.* 226, 44–49. doi: 10.1111/nph.16356

Al-Dous, E. K., George, B., Al-Mahmoud, M. E., Al-Jaber, M. Y., Wang, H., Salameh, Y. M., et al. (2011). *De novo* genome sequencing and comparative genomics of date palm (phoenix dactylifera). *Nat. Biotechnol.* 29, 521–527. doi: 10.1038/nbt.1860

Alseekh, S., Scossa, F., and Fernie, A. R. (2020). Mobile transposable elements shape plant genome diversity. *Trends Plant Sci.* 25, 1062–1064. doi: 10.1016/j.tplants.2020.08.003

Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V., et al. (2019). RepetDB: a unified resource for transposable element references. *Mobile DNA* 10, 1–8. doi: 10.1186/s13100-019-0150-y

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6, 1–6. doi: 10.1186/s13100-015-0041-9

Barghini, E., Natali, L., Giordani, T., Cossu, R. M., Scalabrin, S., Cattonaro, F., et al. (2014). LTR Retrotransposon dynamics in the evolution of the olive (Olea europaea) genome. *DNA Res.* 22, 91–100. doi: 10.1093/dnares/dsu042

Bennetzen, J. L. (1996). The contributions of retroelements to plant genome organization, function and evolution. *Trends Microbiol.* 4, 347–353. doi: 10.1016/0966-842X(96)10042-1

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 65, 505–530. doi: 10.1146/annurev-arplant-050213-035811

Bousios, A., Minga, E., Kalitsou, N., Pantermali, M., Tsaballa, A., and Darzentas, N. (2012). MASiVEdb: the sirevirus plant retrotransposon database. *BMC Genomics* 13, 158. doi: 10.1186/1471-2164-13-158

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17, 66. doi: 10.1186/s13059-016-0924-1

Bui, Q. T., and Grandbastien, M.-A. (2012). LTR Retrotransposons as Controlling Elements of Genome Response to Stress?. In: M. A. Grandbastien and J. Casacuberta (eds) *Plant Transposable Elements. Topics in Current Genetics*, vol 24. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-31842-9_14

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., et al. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* 24, 1242–1255. doi: 10.1105/tpc.111.095232

Chaparro, C., Gayraud, T., de Souza, R. F., Domingues, D. S., Akaffou, S., Laforga Vanzela, A. L., et al. (2015). Terminal-repeat retrotransposons with gag domain in plant genomes: a new testimony on the complex world of transposable elements. *Genome Biol. Evol.* 7, 493–504. doi: 10.1093/gbe/evv001

Chaparro, C., Guyot, R., Zuccolo, A., Piegu, B., and Panaud, O. (2006). RetrOryza: a database of the rice LTR-retrotransposons. *Nucleic Acids Res.* 35, D66–D70. doi: 10.1093/nar/gkl780

Cho, J., Benoit, M., Catoni, M., Drost, H.-G., Brestovitsky, A., Oosterbeek, M., et al. (2019). Sensitive detection of pre-integration intermediates of long terminal repeat retrotransposons in crop plants. *Nat. Plants* 5, 26–33. doi: 10.1038/s41477-018-0320-9

Cossu, R. M., Casola, C., Giacomello, S., Vidalis, A., Scofield, D. G., and Zuccolo, A. (2017). Ltr retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. *Genome Biol. Evol.* 9, 3449–3462. doi: 10.1093/gbe/evx260

da Costa, Z. P., Cauz-Santos, L. A., Ragagnin, G. T., Van Sluys, M.-A., Dornelas, M. C., Berges, H., et al. (2019). Transposable element discovery and characterization of LTR-retrotransposon evolutionary lineages in the tropical fruit species passiflora edulis. *Mol. Biol. Rep.* 46, 6117–6133. doi: 10.1007/s11033-019-05047-4

Dai, X., Wang, H., Zhou, H., Wang, L., Dvořak, J., Bennetzen, J. L., et al. (2018). Birth and death of ltr-retrotransposons in aegilops tauschii. *Genetics* 210, 1039–1051. doi: 10.1534/genetics.118.301198

Diambra, L. A. (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195. doi: 10.1038/nature10158

Du, J., Tian, Z., Bowen, N. J., Schmutz, J., Shoemaker, R. C., and Ma, J. (2010a). Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through ltr swapping in soybean. *Plant Cell* 22 (1), 48–61. doi: 10.1105/tpc.109.068775

Du, J., Tian, Z., Hans, C. S., Laten, H. M., Cannon, S. B., Jackson, S. A., et al. (2010b). Evolutionary conservation, diversity and specificity of ltr-retrotransposons in flowering plants: Insights from genome wide analysis and multi-specific comparison. *Plant J.* 63 (4), 584–598. doi: 10.1111/j.1365-313X.2010.04263.x

Eickbush, T. H., and Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res.* 134, 221–234. doi: 10.1016/j.virusres.2007.12.010

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf.* 9, 1–14. doi: 10.1186/1471-2105-9-18

Emms, D. M., and Kelly, S. (2019). Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 1–14. doi: 10.1186/s13059-019-1832-y

Gao, X., Havecker, E. R., Baranov, P. V., Atkins, J. F., and Voytas, D. F. (2003). Translational recoding signals between gag and pol in diverse LTR retrotransposons. *RNA* 9, 1422–1430. doi: 10.1261/rna.5105503

Gonzalez, L. G., and Deyholos, M. K. (2012). Identification, characterization and distribution of transposable elements in the flax (Linum usitatissimum l.) genome. *BMC Genomics* 13, 644. doi: 10.1186/1471-2164-13-644

Grandbastien, M.-A. (2015). LTR Retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim. Biophys. Acta (BBA) - Gene Regul. Mech.* 1849, 403–416. doi: 10.1016/j.bbagrm.2014.07.017

Havecker, E. R., Gao, X., and Voytas, D. F. (2004). The diversity of ltr retrotransposons. *Genome Biol.* 5, 1–6. doi: 10.1186/gb-2004-5-6-225

Ibrahim, M. A., Al-Shomrani, B. M., Alharbi, S. N., Elliott, T. A., Alsuabeyl, M. S., Alqahtani, F. H., et al. (2021). Genome-wide comparative analysis of transposable elements in palmae genomes. *Front. bioscience (Landmark edition)* 26, 1119–1131. doi: 10.52586/5014

Ito, H., Yoshida, T., Tsukahara, S., and Kawabe, A. (2013). Evolution of the onsen retrotransposon family activated upon heat stress in brassicaceae. *Gene* 518, 256–261. doi: 10.1016/j.gene.2013

Jedlicka, P., Lexa, M., and Kejnovsky, E. (2020). What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00644

Jiang, N., Bao, Z., Zhang, X., Eddy, S. R., and Wessler, S. R. (2004). Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431(7008), 569–573. doi: 10.1038/nature02953

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature* 546, 524–527. doi: 10.1038/nature22971

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large Retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166, 1437–1450. doi: 10.1534/genetics.166.3.1437

Kashkush, K., Feldman, M., and Levy, A. A. (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* 33, 102–106. doi: 10.1038/ng1063

Kelly, L. J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., 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 Phytol.* 208, 596–607. doi: 10.1111/nph.13471

Kijima, T. E., and Innan, H. (2009). On the estimation of the insertion time of LTR retrotransposable elements. *Mol. Biol. Evol.* 27, 896–904. doi: 10.1093/molbev/msp295

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/BF01731581

Koch, M. A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in arabidopsis, arabis, and related genera (Brassicaceae). *Mol. Biol. Evol.* 17, 1483–1498. doi: 10.1093/oxfordjournals.molbev.a026248

Kumar, A., and Bennetzen, J. L. (1999). Plant retrotransposons. *Annu. Rev. Genet.* 33, 479–532. doi: 10.1146/annurev.genet.33.1.479

Li, S.-F., She, H.-B., Yang, L.-L., Lan, L.-N., Zhang, X.-Y., Wang, L.-Y., et al. (2022). Impact of ltr retrotransposons on genome structure, evolution, and function in curcurbitaceae species. *Int. J. Mol. Sci.* 23, 10158. doi: 10.3390/ijms231710158

Li, S.-F., Zhang, G.-J., Zhang, X.-J., Yuan, J.-H., Deng, C.-L., Gu, L.-F., et al. (2016). DPTEdb, an integrative database of transposable elements in dioecious plants. *Database* 2016, baw078. doi: 10.1093/database/baw078

Liu, Y., and Yang, G. (2014). Tc 1-like transposable elements in plant genomes. *Mobile DNA* 5, 17. doi: 10.1186/1759-8753-5-17

Lockton, S., and Gaut, B. S. (2009). The contribution of transposable elements to expressed coding sequence in arabidopsis thaliana. *J. Mol. Evol.* 68, 80–89. doi: 10.1007/s00239-008-9190-5

Luo, X., Chen, S., and Zhang, Y. (2022). Plantrep: a database of plant repetitive elements. *Plant Cell Rep.* 41, 1163–1166. doi: 10.1007/s00299-021-02817-y

Ma, J., and Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci.* 101, 12404–12410. doi: 10.1073/pnas.0403715101

Ma, B., Kuang, L., Xin, Y., and He, N. (2019). New insights into long terminal repeat retrotransposons in mulberry species. *Genes* 10, 285. doi: 10.3390/genes10040285

Ma, B., Li, T., Xiang, Z., and He, N. (2015). MnTEdb, a collective resource for mulberry transposable elements. *Database* 10, 1194 . doi: 10.1093/database/bav004

Mann, L., Seibt, K. M., Weber, B., and Heitkam, T. (2022). ECCsplorer: a pipeline to detect extrachromosomal circular DNA (eccDNA) from next-generation sequencing data. *BMC Bioinf.* 23, 40. doi: 10.1186/s12859-021-04545-2

Marcon, H. S., Domingues, D. S., Silva, J. C., Borges, R. J., Matioli, F. F., de Mattos Fontes, M. R., et al. (2015). Transcriptionally active LTR retrotransposons in eucalyptus genus are differentially expressed and insertionally polymorphic. *BMC Plant Biol.* 15, 1–16. doi: 10.1186/s12870-015-0550-1

Mokhtar, M. M., Alsamman, A. M., Abd-Elhalim, H. M., and El Allali, A. (2021). CicerSpTEdb: A web-based database for high-resolution genome-wide identification of transposable elements in cicer species. *PLos One* 16, 1–21. doi: 10.1371/journal.pone.0259540

Mokhtar, M. M., and Atia, M. A. M. (2018). SSRome: an integrated database and pipelines for exploring microsatellites in all organisms. *Nucleic Acids Res.* 47, D244–D252. doi: 10.1093/nar/gky998

Mokhtar, M. M., and EL Allali, A. (2022). Pltrnadb: Plant transfer rna database. *PLos One* 17, 1–12. doi: 10.1371/journal.pone.0268904

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

Orozco-Arias, S., Jaimes, P. A., Candamil, M. S., Jiménez-Varón, C. F., Tabares-Soto, R., Isaza, G., et al. (2021). Inpactordb: a classified lineage-level plant ltr retrotransposon reference library for free-alignment methods based on machine learning. *Genes* 12, 190. doi: 10.3390/genes12020190

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the ltr assembly index (lai). *Nucleic Acids Res.* 46, e126–e126. doi: 10.1093/nar/gky730

Ou, S., and Jiang, N. (2017). LTR Retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20, 1–18. doi: 10.1186/s13059-019-1905-y

Pedro, D. L. F., Amorim, T. S., Varani, A., Guyot, R., Domingues, D. S., and Paschoal, A. R. (2021). An atlas of plant transposable elements. *F1000Research* 10. doi: 10.12688/f1000research.74524.1

Pedro, D. L. F., Lorenzetti, A. P. R., Domingues, D. S., and Paschoal, A. R. (2018). PlaNC-TE: a comprehensive knowledgebase of non-coding RNAs and transposable elements in plants. *Database* 2018, bay078. doi: 10.1093/database/bay078

Pereira, V. (2008). Automated paleontology of repetitive DNA with REANNOTATE. *BMC Genomics* 9, 614. doi: 10.1186/1471-2164-9-614

Qu, D., Sun, W.-W., Li, L., Ma, L., Sun, L., Jin, X., et al. (2019). Long noncoding RNA MALAT1 releases epigenetic silencing of HIV-1 replication by displacing the polycomb repressive complex 2 from binding to the LTR promoter. *Nucleic Acids Res.* 47, 3013–3027. doi: 10.1093/nar/gkz117

Rossi, M., Araujo, P. G., and Van Sluys, M.-A. (2001). Survey of transposable elements in sugarcane expressed sequence tags (ESTs). *Genet. Mol. Biol.* 24, 147–154. doi: 10.1590/S1415-47572001000100020

Sabot, F., Sourdille, P., Chantret, N., and Bernard, M. (2006). Morgane, a new ltr retrotransposon group, and its subfamilies in wheats. *Genetica* 128, 439–447. doi: 10.1007/s10709-006-7725-5

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635. doi: 10.1038/nature11119

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *science* 326, 1112–1115. doi: 10.1126/science.1178534

Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of *de novo* predicted ltr retrotransposons. *Nucleic Acids Res.* 37, 7002–7013. doi: 10.1093/nar/gkp759

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 43 (D1), D447–D452. doi: 10.1093/nar/gku1003

Tajima, F., and Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 1 (3), 269–285. doi: 10.1093/oxfordjournals.molbev.a040317

Tanskanen, J. A., Sabot, F., Vicient, C., and Schulman, A. H. (2007). Life without gag: The bare-2 retrotransposon as a parasite's parasite. *Gene* 390, 166–117. doi: 10.1016/j.gene.2006.09.009

Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2003). Multiple sequence alignment using clustalw and clustalx. *Curr. Protoc. Bioinf.* 00, 2.3.1–2.3.22. doi: 10.1002/0471250953.bi0203s00

Vitte, C., Fustier, M.-A., Alix, K., and Tenaillon, M. I. (2014). The bright side of transposons in crop evolution. *Briefings Funct. Genomics* 13, 276–295. doi: 10.1093/bfgp/elu002

Wang, H., and Liu, J.-S. (2008). LTR Retrotransposon landscape in medicago truncatula: more rapid removal than in rice. *BMC Genomics* 9, 382. doi: 10.1186/1471-2164-9-382

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., et al. (2006). High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* 18, 1791–1802. doi: 10.1105/tpc.106.041905

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-Gonzalez, R. H., et al. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* 19, 103. doi: 10.1186/s13059-018-1479

Wicker, T., Matthews, D. E., and Keller, B. (2002). TREP: a database for triticeae repetitive elements. *Trends in plant science.* 7(12), 561–2. doi: 10.1016/S1360-1385(02)02372-5

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

Witte, C. P., Le, Q. H., Bureau, T., and Kumar, A. (2001). Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences*. 98(24), 13778–83. doi: 10.1073/pnas.241341898

Xu, Z., Liu, J., Ni, W., Peng, Z., Guo, Y., Ye, W., et al. (2017). GrTEdb: the first web-based database of transposable elements in cotton (Gossypium raimondii). *Database* 2017, bax013. doi: 10.1093/database/bax013.Bax013

Xu, Z., and Wang, H. (2007). LTR FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yamamoto, G., Miyabe, I., Tanaka, K., Kakuta, M., Watanabe, M., Kawakami, S., et al. (2021). SVA retrotransposon insertion in exon of MMR genes results in aberrant RNA splicing and causes lynch syndrome. *Eur. J. Hum. Genet.* 29, 680–686. doi: 10.1038/s41431-020-00779-5

Yi, F., Jia, Z., Xiao, Y., Ma, W., and Wang, J. (2018a). Sptedb: a database for transposable elements in salicaceous plants. *Database* 2018, bay024. doi: 10.1093/database/bay024

Yi, F., Ling, J., Xiao, Y., Zhang, H., Ouyang, F., and Wang, J. (2018b). ConTEdb: a comprehensive database of transposable elements in conifers. *Database* 2018, bay131. doi: 10.1093/database/bay131

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., et al. (2022). Tesorter: an accurate and fast method to classify ltr-retrotransposons in plant genomes. *Horticulture Res.* 9, uhac017. doi: 10.1093/hr/uhac017

Zhao, D., Ferguson, A. A., and Jiang, N. (2016). What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochim. Biophys. Acta (BBA)-Gene Regul. Mech.* 1859, 366–380. doi: 10.1016/j.bbagrm.2015.12.005

Zhou, M., Hu, B., and Zhu, Y. (2017). Genome-wide characterization and evolution analysis of long terminal repeat retroelements in moso bamboo (phyllostachys edulis). *Tree Genet. Genomes* 13, 1–12. doi: 10.1007/s11295-017-1114-3

Zhou, S.-S., Yan, X.-M., Zhang, K.-F., Liu, H., Xu, J., Nie, S., et al. (2021). A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes. *Sci. Data* 8, 174. doi: 10.1038/s41597-021-00968-x

TYPE Review
PUBLISHED 25 April 2023
DOI 10.3389/fpls.2023.1174339

The power of retrotransposons in high-throughput genotyping and sequencing

Yunus Emre Arvas[1], Sevgi Marakli[2], Yılmaz Kaya[3,4] and Ruslan Kalendar[5,6*]

[1]Department of Biology, Faculty of Sciences, Karadeniz Technical University, Trabzon, Türkiye, [2]Department of Molecular Biology and Genetics, Faculty of Arts and Sciences, Yildiz Technical University, Istanbul, Türkiye, [3]Agricultural Biotechnology Department, Faculty of Agriculture, Ondokuz Mayıs University, Samsun, Türkiye, [4]Department of Biology, Faculty of Science, Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan, [5]Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan, [6]Institute of Biotechnology, Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland

Check for updates

OPEN ACCESS

EDITED BY
Diego H. Sanchez,
Instituto de Investigaciones Fisiológicas y Ecologicas Vinculadas a la Agricultura (IFEVA) (CONICET), Argentina

REVIEWED BY
Khalil Kashkush,
Ben-Gurion University of the Negev, Israel
Yuki Monden,
Okayama University, Japan
Matthias Benoit,
INRAE Occitanie Toulouse, France

*CORRESPONDENCE
Ruslan Kalendar
ruslan.kalendar@helsinki.fi

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 26 February 2023
ACCEPTED 11 April 2023
PUBLISHED 25 April 2023

CITATION
Arvas YE, Marakli S, Kaya Y and Kalendar R (2023) The power of retrotransposons in high-throughput genotyping and sequencing.
Front. Plant Sci. 14:1174339.
doi: 10.3389/fpls.2023.1174339

The use of molecular markers has become an essential part of molecular genetics through their application in numerous fields, which includes identification of genes associated with targeted traits, operation of backcrossing programs, modern plant breeding, genetic characterization, and marker-assisted selection. Transposable elements are a core component of all eukaryotic genomes, making them suitable as molecular markers. Most of the large plant genomes consist primarily of transposable elements; variations in their abundance contribute to most of the variation in genome size. Retrotransposons are widely present throughout plant genomes, and replicative transposition enables them to insert into the genome without removing the original elements. Various applications of molecular markers have been developed that exploit the fact that these genetic elements are present everywhere and their ability to stably integrate into dispersed chromosomal localities that are polymorphic within a species. The ongoing development of molecular marker technologies is directly related to the deployment of high-throughput genotype sequencing platforms, and this research is of considerable significance. In this review, the practical application to molecular markers, which is a use of technology of interspersed repeats in the plant genome were examined using genomic sources from the past to the present. Prospects and possibilities are also presented.

KEYWORDS

molecular markers, interspersed repeats, amplification profiling, inter-retrotransposon amplified polymorphism, transposable elements, retrotransposon

Abbreviations: TEs, Transposable Elements; LTR, Long Terminal Repeats; MAS, Marker-Assisted Selection; RFLP, Restriction Fragment Length Polymorphism; SSR, Simple Sequence Repeats; IRAP, Inter-Retrotransposon Amplified Polymorphism; iPBS, Inter-Primer Binding Site; RAPD, Random Amplification of Polymorphic DNA; TD, Transposon Display; AFLP, Amplified Fragment Length Polymorphism; S-SAP, Retrotransposon-Based Sequence-Specific Amplification; NGS, Next-Generation Sequencing; SNPs, single nucleotide polymorphisms; InDels,Insertion-deletion mutations.

Introduction

The genomes of eukaryotic organisms mostly consist of interspersed repetitive sequences, in particular transposable elements (TEs). In most species studied, interspersed repeats are rather unevenly distributed, with some of them clustered around telomeres or centromeres (Kalendar et al., 2021a). Although TEs exhibit considerable sequence diversity, they can be divided into two well-defined classes according to their structure and propagation strategies. Retrotransposons are the most common type of TEs that belong to class I. Retrotransposons are retrovirus-related genetic elements that amplify through the process of reverse transcription use an RNA-mediated process to transpose, in contrast to class II transposons, which do not require an RNA intermediate (Papolu et al., 2022). Depending on their structure and transposition cycle, retrotransposons can be divided into two main subclasses, long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons, which are based on the presence or absence of LTR at their ends (Wicker et al., 2007). Non-LTR retrotransposons can be further subdivided into long interspersed nucleotide elements (LINE) and short interspersed elements (SINE). Retrotransposons and retroviruses share similarities such as common structural features and basic life cycle stages (Figure 1). Retrovirus-like or LTR retrotransposons include endogenous pararetroviruses (Valli et al., 2023), which are remnants of previous rounds of germline virus infection that have lost their ability to reinfect and are fixed in the genome. Each group of transposons has corresponding non-autonomous forms, which are missing one or more genes necessary for transposition. In the case of class II transposons, these non-autonomous forms are known as miniature inverted-repeat transposable elements (MITE), while SINEs are non-autonomous forms of non-LTR retrotransposons. LTR retrotransposons have the following two types of non-autonomous forms: terminal-repeat retrotransposons in miniature (TRIM) and large retrotransposon derivatives (LARD) (Kalendar et al., 2004; Kalendar et al., 2008; Kalendar et al., 2020).

The number of retrotransposons in an organism's genome is directly proportional to the genome size (Hawkins et al., 2006; Holligan et al., 2006; Jaillon et al., 2007; Ming et al., 2008; Wang and Liu, 2008; Huang et al., 2009; Qiu et al., 2009; Wicker et al., 2009; Kovach et al., 2010; Schmutz et al., 2010; Argout et al., 2011; Shulaev

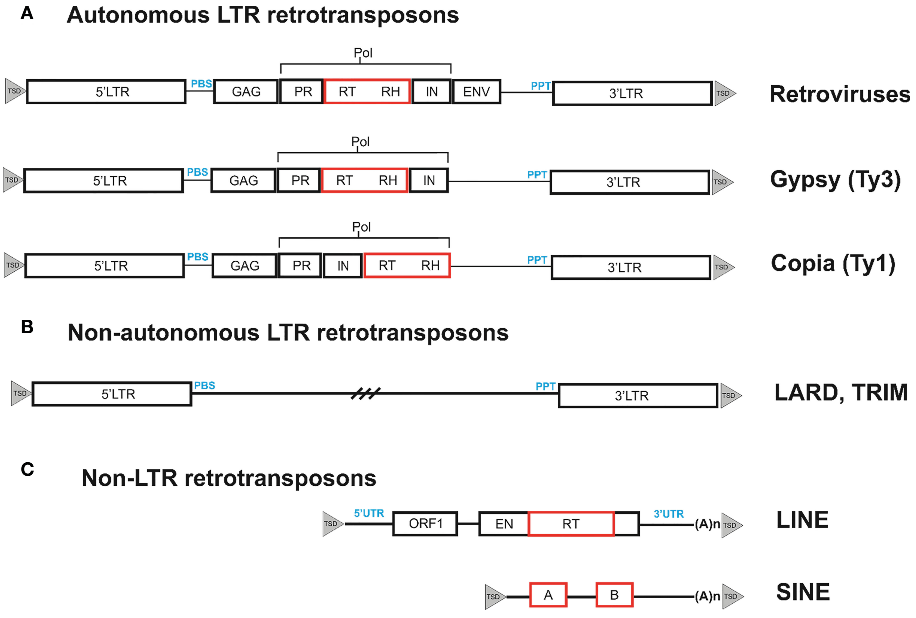


FIGURE 1
Organization of different types of retrotransposons. **(A)**. Retrotransposons can be classified based on their structural features. One such type is bounded by long terminal repeats (LTRs), which contain the transcriptional promoter and terminator. These LTRs contain short inverted repeats at either end, shown as filled triangles. During reverse transcription, the primer binding site (PBS) and polypurine tract (PPT) domains prime the synthesis of the complementary DNA (cDNA) strand on the (–) and (+) strands, respectively. The internal region of the retrotransposon codes for the proteins necessary for the retrotransposon life cycle. These include the capsid protein (GAG), the aspartic proteinase (AP) that cleaves the polyprotein (AP), the integrase (IN) that inserts the cDNA copy into the genome, and the reverse transcriptase (RT) and RNaseH (RH) that together copy the transcript into cDNA. The internal region also contains evolutionarily conserved domains (noted below the element as black boxes) that are necessary for function and can be used to isolate retrotransposons from previously unstudied plant species. These conserved domains can be targeted for amplification and sequencing of retrotransposons from a variety of species. The LTRs are generally well-conserved within families and can serve as targets for primer design to generate DNA footprints. DNA footprints are useful for studying the evolutionary history and diversity of retrotransposons within and between species. **(B)**. Non-autonomous LTR retrotransposons have two types of non-autonomous forms: Terminal-Repeat Retrotransposons in Miniature (TRIMs) and Large Retrotransposon Derivatives (LARDs). **(C)** Non-LTR retrotransposons, such as LINEs and SINEs, are terminated by a 3′ poly(A) stretch. LINEs have two open reading frames (ORFs) that encode a nucleocapsid protein (gag), an endonuclease (EN), and a reverse transcriptase (RT). The ORFs are flanked by untranslated regions (UTR), with the broken line indicating the 5′ truncations found in many LINEs. SINEs are composed of a tRNA-derived region, an unrelated DNA sequence. The two black boxes labeled A and B depict regions with homology to RNA polymerase III promoters.

et al., 2011). LTR retrotransposons are the most abundant subclass of retrotransposons in eukaryotic genomes (Bennetzen and Wang, 2014). Retrotransposons are believed to have a vital role in regulating chromosomal structure and structural genes due to their repetitive structure and presence of regulatory signals (Huang et al., 2012; Zervudacki et al., 2018). Eighty percent of the plant genomes of grasses (wheat, barley, and rye) consist of TEs and other repeats (Wicker et al., 2018). The corresponding value for the human genome is 45%; for *Arabidopsis* (Zhang and Wessler, 2004), retrotransposons form 14% of the genome (Turnbull et al., 2018; Akakpo et al., 2020).

Numerous studies have investigated the activities of LTR retrotransposons under stress conditions (Grandbastien, 1998; Pecinka et al., 2010; Tittel-Elmer et al., 2010; Ito et al., 2011; Butelli et al., 2012; Gaubert et al., 2017; Benoit et al., 2019; Cho et al., 2019; Papolu et al., 2021; Ramakrishnan et al., 2022). Mobile elements and endogenous retroviruses are important in the distribution of cis-regulatory elements, which contribute to genetic control in both the short (accumulation of hidden variability and processes of selection) and the long (adaptation and separation) term.

Transposons can insert into different positions within a genome; this leads to changes in the DNA sequence and consequently mutations (Mcclintock, 1950). Retrotransposons can also alter the amount of DNA in the genome by increasing the number of copies of the TE. Transposons have insertional, transcriptional, and translational properties, implying that transposon movements may change depending on the organism, environment, and even tissue specific. However, the movement mechanisms are not completely understood (Rebollo et al., 2011; Schrader and Schmitz, 2019).

TEs have long been considered as agents that give rise to mutations that disrupt gene function after inserting into coding or promoter sequences. These elements play a role in evolution but can also have negative effects on the host organism by disrupting normal gene function (Kalendar et al., 2020). This was demonstrated by Barbara McClintock, who showed that TEs can cause pigmentation loss in corn kernels (Mcclintock, 1950). There has been some opposition to this view; transposition may be beneficial to an organism by mediating epigenetic factors or by acting as cis-acting regulatory regions that exhibit alternative promoters that regulate gene expression (Mirouze and Vitte, 2014; Hirsch and Springer, 2017). Furthermore, amplification of TEs can drive significant biological novelty [i.e. placental pregnancy (Lynch et al., 2011) and innate immunity (Chuong et al., 2016)]; hence, transposition may drive eukaryotic evolution by reshaping gene networks that result in novel features. In crops, TEs have a role in variation of agronomic features, including tomato shape (Xiao et al., 2008) and red pigmentation of apples (Zhang et al., 2019). However, the degree that genetic variation caused by TE can be used for agronomic applications has not been fully verified for all crops (Ramakrishnan et al., 2023).

Marker-assisted selection (MAS) and historical developments

Molecular genetic markers are short DNA sequences that can be used to identify specific regions of the genome. They can be used in a variety of applications, including genome mapping, disease diagnosis, and classification of individuals or populations. The genetic marker might be either a gene or a sequence that possesses no known functions. Currently, genetic polymorphisms in DNA sequences are analyzed by several strategies, such as various PCR-based methods that detect polymorphisms and PCR-based genome profiling applications: various platforms for hybridization (NanoString Technologies, multiplex ligation-dependent probe amplification [MLPA], microarrays), and next-generation sequencing (NGS). Current analysis strategies have been developed depending on the main methods (Table 1).

Molecular MAS is a method that improves the efficiency of plant breeding by relying on DNA markers to score for specific traits or characteristics. This method allows for earlier selection and reduces the population size of plants, thus saving time and effort. MAS can be used to screen for traits that are difficult or expensive to score phenotypically, such as disease resistance or fruit quality. The use of DNA markers also allows for detection of heredity patterns at the genomic level and the cloning of genes important for natural resistance to disease. This can result in more “green” and cost-effective solutions for disease control. MAS also has the potential to pyramid multiple desirable genes in a new plant variety. By increasing precision in the selection, unwanted side effects in future plant generations can be reduced.

The selection identified by MAS includes scoring in the absence or presence of a plant phenotype of interest that is based upon the DNA banding pattern of connected markers on an autoradiogram or a gel relying on the market framework. The rationale is that the banding pattern provides information about the parental source of the bands in segregants at a marker locus, which illustrates the absence or presence of a specifical chromosomal fragment harboring the allele of interest. The effectiveness of screening in breeding methods is improved in the following several regards. Segregants can be graded at the seedling phase for features that are expressed late in the progress of the plant; this involves features including photoperiod sensitivity, male sterility, and grain quality. It is likely that screening for features that are exceedingly time-consuming, difficult, or expensive to score and determine, such as resistance to biotypes of insects or diseases or certain types or nematodes, tolerance to root morphology, toxicity, mineral, salt deficiencies, and drought. Selection can be applied to certain features concurrently, which is difficult or is not possible *via* traditional means. Heterozygotes are readily defined and separated from homozygotes without referring to progeny testing, which saves effort and time. MAS is a promising choice to be used in improving many phenotypic features of interest in which the evaluation is usually unreliable or expensive. MAS may also increase the efficacy of selection by permitting earlier selection and decreasing the plant population size at the time of selection. Cultivators can rapidly detect heredity patterns at the genome level by directly analyzing the genetic makeup of empirical plants at the seedling stage. It may be useful for features that cannot be defined before plant maturity, such as fruit qualifications and for the features that are difficult to test (such as disease resistance). The selection of resistant plants is performed by using a DNA marker that is connected to the feature that controls the gene rather than turning the resistance of plants to disease into account. To achieve the most economical, environmentally safe, and

TABLE 1 Main strategies for detection of molecular genetic polymorphisms.

Marker system	Polymorphism detection strategy	Principle
PCR-based methods detection polymorphism		
Short Tandem Repeat (STR) analysis: variable number tandem repeat (VNTR)	Single-loci polymorphic DNA markers: Simple Sequence Repeats (SSR)	A STR is a microsatellite with repeating units between 2 and 5 bp in length; the number of repeats varies between individuals. This method detects differences in STRs based on PCR product length. Microsatellites and VNTRs can be highly polymorphic and are essential for utility as genetic markers.
Exon-Primed Intron-Crossing (EPIC), Intron Targeting Polymorphism (ITP)	Multiple-loci polymorphic DNA markers	The method relies on the design of primers selected to anneal to highly conserved regions, for example to exons. For illustrative applications, this has been applied to analyze conserved domains within eukaryotic 18S and 28S ribosomal genes and prokaryotic 16S and 23S ribosomal genes to amplify variable intergenic regions known as internal transcribable spacers (ITS) containing the 5,8S ribosomal gene. Intronic regions selected for the determination of polymorphisms are amplified by primers designed for regions close to the exon.
Nucleotide-Binding Site (NBS) profile	Multiple-loci polymorphic DNA markers	Genomic DNA is digested with restriction enzymes after being attached to adapters. Fingerprints of resistance gene regions are generated with the use of adapter-specific and R-gene-specific primers.
Resistance Gene Analog Polymorphism (RGAP)	Multiple-loci polymorphic DNA markers	Analog fingerprints based on resistance genes are amplified with either degenerate specific primers or primer pairs. The primers are designed by targeting the conserved regions of the R genes.
PCR-based genome profiling applications (multiple-loci polymorphic DNA markers)		
Random amplification of polymorphic DNA (RAPD)		Method is based on the use of a single primer (short or standard length) for universal amplification of prokaryotic or eukaryotic genomes (genome profiling). The primer sequence is not essential, thus virtually any primers can be used, including those used for specific amplification of a particular locus. However, PCR amplification conditions should facilitate the formation of multiple amplicons. Generates anonymous markers.
Inter-simple sequence repeat (ISSR)		RAPD-like specific amplification technique for genome profiling; this is a PCR method that uses a single specific primer complementary to the microsatellite sequence. The complementary sequences of two neighboring microsatellite loci are used as primers for PCR; the variant region between them is amplified.
Inter-Primer Binding Site (iPBS)		RAPD-like specific amplification technique for genome profiling; this is a PCR method based on the actually universal occurrence of complement tRNA as a binding site for the reverse transcriptase site in LTR retrotransposons. Primers are annealed to the PBS region of LTR retrotransposons, which are found head-to-head. Amplified products contain LTR and genetic regions.
Inter-Repeat Amplified Polymorphism (IRAP)		RAPD-like specific amplification technique for genome profiling; this is a PCR method based on using a single repeat-specific primer. Many kinds of repeats are dispersed and clustered in the genome, which makes this PCR possible.
Retrotransposon Microsatellite Amplified Polymorphism (REMAP)		RAPD-like specific amplification technique for DNA fingerprinting; this is a PCR method where one of the two primers matches a microsatellite motif with the second specific primers associated with retrotransposons (or any type of repeat sequence). In REMAP, anchored nucleotides (one or more) are used on the 3′ ends of the simple sequence repeat primer to avoid primer drift within the microsatellite sequence.
Palindromic Sequence-Targeted (PST) PCR; Transposon Display (TD)		PCR-based methods combine sequence-specific primers with a universal primer that can anneal to unknown DNA targets, thus ensuring rapid and efficient PCR. This method is based on targeting universal primers to palindromic sequences occurring randomly in natural DNA sequences. PST-PCR involves two rounds of PCR. The first round utilizes a combination of one sequence-specific primer and one universal primer (PST). The second round involves a combination of single- or two-tailed primers; one anneals on a 5′-tail attached to the sequence-specific primer and the other anneals on another 5′-tail attached to the PST primer. The main benefit of PST-PCR is the convenience of using a single-tailed primer for all types of target sequences.
Amplified Fragment Length Polymorphism (AFLP)		A DNA fingerprinting method that utilizes an amplification technique which selectively amplifies a specific subset of digested DNA fragments, resulting in distinctive fingerprints that can be used to compare and analyze genomes of interest. The AFLP protocol involves several key steps. First, the genomic DNA is digested using restriction enzymes, and then adaptors are ligated to the restricted fragments. Next, a preselective PCR amplification is performed to amplify a subset of the restricted fragments. This is followed by a selective PCR amplification, which amplifies only the fragments that have the adaptors and primers that are specific to the target genome of interest. Finally, the amplified DNA fragments are separated using electrophoresis. Variations of the standard AFLP methodology have been developed to target additional levels of diversity, such as transcriptomic variation and DNA methylation polymorphism.

(Continued)

TABLE 1 Continued

Marker system	Polymorphism detection strategy	Principle
Retrotransposon-Based Sequence-Specific Amplification (S-SAP) Polymorphism (Transposon Display)		S-SAP is a derivative of the Amplified Fragment Length Polymorphism (AFLP) technique and generates amplified fragments containing a retrotransposon LTR sequence at one end and a host restriction site at the other end. Genomic DNA is completely digested, preferably with two different enzymes (usually *Mse*I and *Pst*I) to generate a target for amplification between the retrotransposon sequence and adaptors that are ligated after digestion, using selective bases in the adapter.
Platforms for hybridization		
NanoString Technologies	nCounter	The nCounter technology employs distinctive optical barcodes that hybridize with each target, allowing for the precise digital counting of individual oligonucleotides without the need for any enzymatic steps. These barcodes consist of six fluorophores, which enable highly multiplexed, single-molecule counting of the targets.
Multiplex Ligation-dependent Probe Amplification (MLPA)		Involves a multiplex PCR assay that can employ as many as 50 probes, with each probe specific to a distinct target DNA sequence. These probes consist of two half-probes, namely a 5′ and 3′ half-probe, which comprise a target-specific sequence and a universal primer sequence. This design permits the simultaneous multiplex PCR amplification of all probes. The assay follows several steps, which include DNA denaturation and probe hybridization, followed by ligation and PCR amplification. The amplification products are separated using electrophoresis.
Molecular Inversion Probes (MIP) Technology		The described method involves single-stranded DNA molecules that possess sequences complementary to two areas flanking the target, spanning several hundred base pairs. Once the MIPs bind to the target and hybridize, gap-filling and ligation occur, producing circular DNA molecules that include the target's sequence, along with adaptors and barcodes. These circularized DNA molecules are then available for subsequent analyses.
Next-Generation Sequencing (Genotyping-by-Sequencing)		
GoldenGate Assay		The extension and amplification steps of the genomic DNA involve a high degree of loci multiplexing (1536-plex) to minimize time, reagent volumes, and material requirements. For each SNP locus, three oligonucleotides are designed: two are specific to each allele of the SNP site, called Allele-Specific Oligos (ASOs), and a third oligo called the Locus-Specific Oligo (LSO) hybridizes several bases downstream from the SNP site. All three oligonucleotides contain regions of genomic complementarity and universal PCR primer sites, while the LSO also has a unique address sequence that targets a particular bead type. During the primer hybridization process, the assay oligonucleotides hybridize to the genomic DNA sample bound to paramagnetic particles. Following hybridization, several wash steps are performed to reduce noise by removing excess and mis-hybridized oligonucleotides. The extension of the appropriate ASO and ligation of the extended product to the LSO join the genotype present at the SNP site to the address sequence on the LSO. The joined, full-length products serve as a template for PCR using universal PCR primers P1, P2, and P3, where P1 and P2 are Cy3- and Cy5-labeled, respectively. After downstream processing, the single-stranded, dye-labeled DNAs are hybridized to their complement bead type through their unique address sequences. Hybridization of the GoldenGate Assay products onto the Array Matrix or BeadChip allows for separation of the assay products in solution onto a solid surface for each individual SNP genotype readout. After hybridization, the fluorescence signal on the Sentrix Array Matrix or BeadChip is analyzed using the BeadArray Reader, which is in turn analyzed using software for automated genotype clustering and calling. No amplification bias can be introduced into the assay, as hybridization occurs before any amplification steps.
Genotyping-in-Thousands by sequencing (GT-seq)		GT-seq utilizes next-generation sequencing of multiplex PCR amplicons to produce genotypes from relatively small panels (50-500) of target single nucleotide polymorphisms for thousands of individuals on a single Illumina HiSeq lane.
Diversity Arrays Technology (DArT)		Typical procedures include reducing the complexity of genomic DNA using specific restriction enzymes, selecting different fragments to represent parental genomes, PCR amplification, and inserting the fragments into a vector to be inserted as probes in a microarray. Fluorescent targets from the reference sequence can then hybridize with the probes and run through the imaging system.
digitalMLPA		digitalMLPA is a semi-quantitative technology that is used to detect relative copy number variation and identify specific (SNP/InDels) mutations. With digitalMLPA, up to 1000 target sequences can be identified in a single multiplex PCR-based reaction. digitalMLPA produces PCR amplicons that are quantified using Illumina NGS platforms. Sequencing is utilized to detect the number of reads of each digitalMLPA probe amplicon.

effective outcomes in disease control, natural resistance genes should be marked in different plant varieties. This approach offers a "green" solution that eliminates the need for expensive chemicals used in disease control. To enhance precision in selection, a uniform practice for scoring involves determining which fragment of each chromosome belongs to each parent and identifying how many genes come from each parent. By increasing precision, undesired side effects can be reduced in the next generation of plants. Additionally, using MAS can help in pyramiding two or more desirable genes in a new plant variety. This approach can further enhance the efficacy of disease control in a cost-effective and environmentally friendly way (Das et al., 2017).

Interspersed repeats-based genome profiling to study genetic polymorphisms

Interspersed repeats-based genome profiling is a range of different approaches that utilizes the polymorphic nature of TEs to study genetic variations in different plant populations. This approach involves identifying and analyzing the repetitive elements that are interspersed throughout the genome. As mentioned, TEs can insert into different positions within a genome, leading to changes in the DNA sequence and mutation. The emerging heterogeneity in the location of distinct TEs has been exploited for specific molecular marker methods focused on repetitive elements (Bennetzen and Wang, 2014). If integration occurs in the cell line from which pollen or eggs eventually originate, a new polymorphism is formed. These new integrated copies are useful for distinguishing breeding lines, varieties, or plant populations. Changes in the copy number of these repeats and internal rearrangements on both homologous chromosomes occur after the induction of recombination processes during prophase of meiosis (Belyayev et al., 2010). TEs create genomic variation in plants, which has been revealed by certain studies (Kalendar et al., 2000; Belyayev et al., 2010; Wicker et al., 2018; Kwolek et al., 2022). According to a comparative study on different plant genomes, most sequences associated with TEs come from modern insertions (El Baidouri and Panaud, 2013). This implies that the ancient TEs were removed from the genomes and, consequently, there must have been a force that balanced the expansion of the genome caused by TE. This was formulated previously in the "increase/decrease" model (Vitte and Panaud, 2005), which has recently been considered more explicitly with mathematical models (Dai et al., 2018).

Applications of inter-retrotransposon amplified polymorphisms

Methods of detection of hidden (phenotypically invisible) genetic variations, such as the molecular marker system for genome profiling, were developed based on sequences of multiple families of complex interspersed genomic repeats (Figure 2). These genome-profiling PCR techniques for the study of genetic variability in eukaryotes that utilize multicopy and genomic diversity abundance of TEs and endogenous viruses can increase knowledge of genetic relationships and assess the genetic diversity of specific species. Interspersed repeats-based genome profiling applications are a simple PCR method and a cost-effective technique to study individual genetic polymorphisms. Genome profiling is essential as TEs, particularly LTR retrotransposons, are widely distributed throughout the genome and can facilitate recombination events during meiosis (Kent et al., 2017).

The basic principle on which numerous PCR methods for genome profiling have been developed is the use of a single primer specific to the high-copy-number retrotransposon sequences or any other sequences for multiple families of complex interspersed genomic repeats. A second strategy is to use a specific primer to retrotransposon sequences in combination with an anchored primer that may be of varied origin, also including other sequences of interspersed genomic repeats (Figure 2).

Eukaryotic genomes harbor many retrotransposon elements, each with their own unique history and level of relatedness. As such, for closely related species, the sequences of a particular retrotransposon will be similar, which reflects the degree of relatedness between species. However, as species become more distantly related, the sequence of a particular retrotransposon will diverge, including the most conserved regions. In the case of closest species, such as within the grass families *Triticum* and *Aegilops*, the sequences of specific mobile elements are almost exactly the same. For species more distant from *Triticum*, such as *Hordeum*, sequences of mobile elements will differ but still retain more than 90% similarity. Therefore, the similarity of mobile elements can be used to study related but distinct species. PCR primers corresponding to the most conserved regions of these mobile elements can also be used (Kalendar et al., 2004; Kalendar et al., 2008; Moisy et al., 2014; Kalendar et al., 2020). One prominent example of such an application is PCR with a single specific primer corresponding to conserved sequences in LTR retrotransposons, which is used for interspersed repeats-based genome profiling. By comparing the sequences of mobile elements across different species, researchers elucidate the evolutionary relationships between those species and the processes that shaped their genomes. In particular, PCR approaches that rely on the identification of transposition element insertion or mutation site polymorphisms include Inter-retrotransposon amplified polymorphism (IRAP) (Kalendar and Schulman, 2006). IRAP is based on PCR amplification of the genomic region between two adjacent LTR retrotransposons oriented in opposite directions. This technique requires a single LTR primer for use in PCR. The PCR products are run on an agarose gel and polymorphisms among the samples are identified based on the banding pattern. IRAP is a technique that is straightforward and rapid. However, the sequences of LTR retrotransposons must be known for this method.

The Inter-Primer Binding Site (iPBS) amplification method is a powerful genomic profiling technique that does not require prior sequence identification of the retrotransposon. The iPBS amplification technique for DNA fingerprinting is a PCR method based on the universal presence of complement tRNA as a binding site for the reverse transcriptase site (PBS) in LTR retrotransposons (Kalendar et al., 2010). Primers are annealed to the PBS region of LTR retrotransposons, which are found head-to-head. Amplified products contain LTR and genetic regions. The LTR primers used in other marker methods are challenging to design, as retrotransposons have no conserved sequences in the LTR regions. On the contrary, many LTR retrotransposons include evolutionarily conserved PBS sequences. The most significant advantage of this method is that it does not need TE sequence information for primer design (Kalendar et al., 2022b). Many studies on TEs are on the determination of relationships using these marker techniques (Monden et al., 2014).

The second strategy for genome profiling and transposon display applications is based on using a specific primer for the

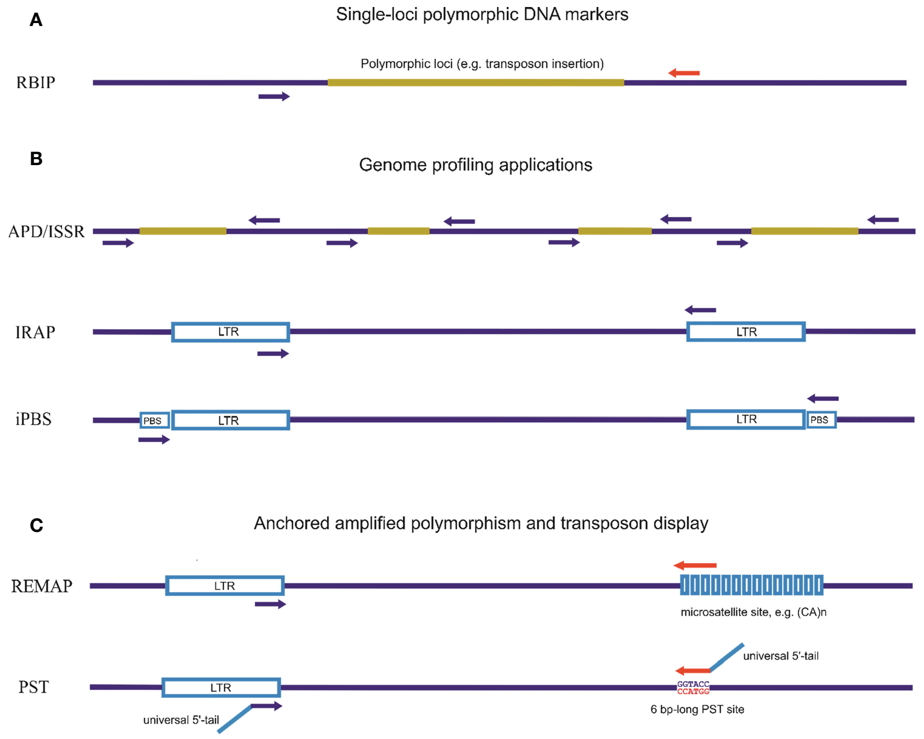

A
Single-loci polymorphic DNA markers
RBIP
Polymorphic loci (e.g. transposon insertion)
B
Genome profiling applications
APD/ISSR
IRAP
LTR
LTR
iPBS
PBS
LTR
LTR
PBS
C
Anchored amplified polymorphism and transposon display
REMAP
LTR
microsatellite site, e.g. (CA)n
PST
LTR
universal 5′-tail
universal 5′-tail
GGTACC
CCATGG
6 bp-long PST site

FIGURE 2
Different approaches for detecting inter-retrotransposon amplified polymorphisms using PCR techniques. **(A)**. Single-loci polymorphic DNA markers. Retrotransposon-Based Insertional Polymorphism (RBIP) is a codominant marker system that uses PCR primers designed for flanking retrotransposon DNA to study insertional polymorphisms in individual retrotransposons. This method detects the presence of mobile element insertions or differences in tandem repeats based on PCR product length. **(B)**. PCR-based genome profiling applications. Random amplification of polymorphic DNA (RAPD) is based on the use of a single primer (short or standard length) for amplification. Inter-simple sequence repeat (ISSR) is a PCR method for DNA fingerprinting using a single primer complementary to the microsatellite sequence. Inter-Repeat Amplified Polymorphism (IRAP) is a PCR method for DNA fingerprinting based on using a single repeat-specific primer. Inter-Primer Binding Site (iPBS) is a PCR method for DNA fingerprinting based on the actually universal occurrence of complement tRNA as a binding site for the reverse transcriptase site in LTR retrotransposons. **(C)**. Anchored genome profiling and transposon display applications using one specific primer for repeat elements in combination with an anchored non-specific primer. Retrotransposon Microsatellite Amplified Polymorphism (REMAP) is a PCR method for DNA fingerprinting where one of the two primers matches a microsatellite motif with second primers associated with retrotransposons (or any type of repeat sequence). Palindromic Sequence-Targeted (PST) and Transposon Display (TD) are PCR-based methods that combine repeat-specific primers with a universal primer able to anneal to palindromic sequences. PST-PCR involves two rounds of PCR. The first round utilizes a combination of one sequence-specific primer and one universal primer (PST). The second round involves a combination of single- or two-tailed primers; one anneals to a 5′-tail attached to the sequence-specific primer and the other anneals to another 5′-tail attached to the PST primer.

repeat element in combination with an anchored non-specific primer. The Retrotransposon Microsatellite Amplified Polymorphism (REMAP) technique for genome profiling is a PCR method in which a specific primer (out of two primers) matches the LTR retrotransposon sequence (or any repeating sequence can be used) and the second anchored primer is annealed to the microsatellite sequence (Kalendar and Schulman, 2006). In REMAP, anchored nucleotides (one or more) are used on the 3′ ends of the microsatellite primer to avoid primer drift within the microsatellite sequence. REMAP and IRAP share similar working principles, as both of require retrotransposon-specific primers whose sequences are known (Hosid et al., 2012).

New methods for high-throughput targeted gene characterization and transposon display have been added to current methodologies, which may be modified to include high-throughput sequencing technologies, among other techniques. For example, Palindromic Sequence-Targeted (PST) PCR utilizes a pair of primers, one of which is complementary to 6-bp long palindromic sequence (PST site) and the other to conserved TE sequences (Kalendar et al., 2019; Kalendar et al., 2021b). The PST-PCR technique allows genome walking and profiling that can be used for the primary characterization of intraspecific and interspecific genetic variability and for screening lines and genotypes.

Retrotransposon-Based Insertional Polymorphism (RBIP) is primarily a codominant marker system that uses one or two pairs of PCR primers designed from combinations of sequences for the retrotransposon and its flanking DNA to study insertional polymorphisms in individual retrotransposons. RBIP is a marker method based on PCR and is used to determine the polymorphisms among two alleles. The comparison of two distinct PCRs permits determination of the polymorphisms. One of the PCRs uses primers that are specific to either retrotransposons or a genomic region near retrotransposons. At the end of the reaction, the interested transposon LTR region is amplified. In the other reaction, two primers specific to the genomic DNA surrounding the retrotransposon are used. The band profiles among the two reactions indicate whether a retrotransposon insertion occurred in the region of interest. RBIP requires sequence information for primer design (Flavell et al., 1998; Kalendar, 2022).

The retrotransposon-based Sequence-Specific Amplification Polymorphism (S-SAP), also known as transposon display, is a modification of the Amplified Fragment Length Polymorphism (AFLP) technique (Vos et al., 1995). This method involves complete digestion of genomic DNA with two different restriction enzymes, typically *Mse*I and *Pst*I, to generate a target for amplification between the retrotransposon sequence and adaptors that are ligated after digestion. The adaptors contain selective bases to facilitate amplification of specific regions. PCR is performed using two primers that are specific to the adapter sequence and the specific mobile element, allowing detection of variations in DNA flanking the mobile element insertion site. The technique uses a multiplex marker system to analyze band profiles among different samples and to detect polymorphisms (Queen et al., 2004). Compared to other retrotransposon marker techniques, S-SAP is costlier and is more difficult to perform. Additionally, the technique requires sequence information for primer generation, which must be designed specifically for the LTR site. However, S-SAP offers greater resolution and accuracy in detecting polymorphisms, making it a valuable tool for genetic analysis.

Prospects, challenges, and discussion

In plant genetics research and breeding practices, molecular techniques such as genetic characterization, genome profiling, genetic integrity, genetic mapping, feature mapping, MAS, and molecular breeding are widely used. Continuous improvements in molecular marker technology, such as high-throughput genotyping platforms, has led to development of new methods such as the GoldenGate assay, Genotyping-in-Thousands by sequencing (GT-seq) (Campbell et al., 2015), Diversity Arrays Technology (DArT) (Alheit et al., 2011), and NGS-based high-throughput hybridization platform systems (digitalMLPA [Multiplex Ligation-dependent Probe Amplification] and Molecular Inversion Probes [MIP]) (Table 1). These advancements have increased efficiency and reduced costs (Elbaidouri et al., 2013; Mir and Varshney, 2012). With such NGS-based high-throughput technological developments, low-throughput molecular markers, such as Kompetitive allele-specific PCR (KASP) (Makhoul et al., 2020), nevertheless remain indispensable for tracking specific genomic regions in molecular breeding programs when analyzing large numbers of samples (Kalendar et al., 2022a; Kalendar et al., 2022c). Therefore, single nucleotide polymorphism (SNP) markers continue to be the most preferred marker systems for development of high-throughput genotypic platforms for genome-wide marker scanning. However, detection of SNP-based markers alone drastically limits the potential to study diversity and genetic polymorphism. One type of polymorphism is the insertion/deletion polymorphism (InDel), which involves the addition or removal of a sequence of different lengths and origins. InDels can have important functional consequences for the chromosome and therefore studying them is crucial. Therefore, other techniques based on mobile element sequences should also be developed, as they can provide complementary information about genomic diversity and evolution.

Interspersed repeats-based genome profiling is a powerful tool for studying genetic polymorphisms and identifying markers for crop improvement programs. By analyzing the distribution and frequency of TEs within the genome, this method can provide valuable insights into the genetic diversity and evolution of plant populations. In addition to whole-genome sequencing, NGS platforms can also be used for targeted sequencing approaches that focus on specific regions of interest, such as the TEs interspersed throughout the genome. Practically all existing PCR approaches that involve identifying and analyzing repetitive elements can be adapted for use on modern NGS platforms (Figure 3). However, when designing primers for NGS-based analysis of repetitive elements, it is important to take into account the specificities of the platform being used. For example, the Illumina HiSeq platform requires the incorporation of adapter sequences in the 5′ tail structure of the primer to facilitate the attachment of the DNA fragments to the sequencing flowcell. One approach that has been successfully adapted for use on NGS platforms is the sequence-specific amplification polymorphism (SSAP) technique, which targets specific retrotransposon insertion sites. The SSAP method involves PCR amplification of the region between two specific primers, one of which is anchored within the retrotransposon and the other in the flanking genomic DNA. The resulting fragments can be sequenced on an NGS platform to identify polymorphisms. Another NGS-based approach for studying TEs is the retrotransposon insertion polymorphisms (RIPs) technique (Du et al., 2022), which uses PCR to amplify specific regions of the genome flanking retrotransposon insertions. The resulting fragments can be sequenced on an NGS platform to identify insertion and deletion events. Overall, the use of NGS platforms for studying TEs offers several advantages over traditional PCR-based methods. NGS allows for the simultaneous analysis of multiple samples and provides greater resolution and sensitivity in detecting polymorphisms. Additionally, NGS platforms can provide information on the structure and organization of TEs within the genome, which can aid in the identification of functional elements and the study of genome evolution.

NGS-based technologies rely on reduced representation sequencing (RRS) techniques (Andrews et al., 2016), including reduced-representation libraries (RRLs), complexity reduction of polymorphic sequences (CRoPS) (Van Orsouw et al., 2007), restriction site-associated DNA (RAD) sequencing (Baird et al., 2008), and low-coverage genotyping. The latter includes multiplexed shotgun genotyping (MSG) and genotyping by sequencing (GBS), which are innovative methods used in genomics (Elshire et al., 2011). These methods can provide informative results even when the reference genome is not available. The choice of method depends on the specific requirements of the study (Mir and Varshney, 2013). As more genome sequences become available, it will be necessary to develop new technologies that allow rapid exploration of the diversity of allelic gene variants in cultivated species that correspond to important plant physiological traits (Springer and Jackson, 2010). In conclusion, NGS-based analysis of repetitive elements is a

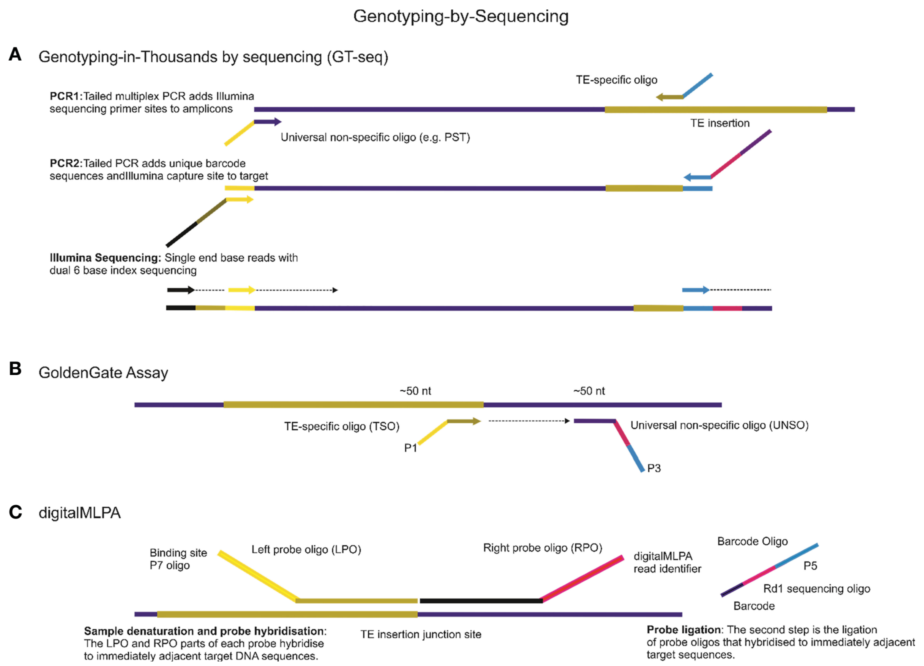

Genotyping-by-Sequencing
A Genotyping-in-Thousands by sequencing (GT-seq)
PCR1:Tailed multiplex PCR adds Illumina sequencing primer sites to amplicons
TE-specific oligo
TE insertion
Universal non-specific oligo (e.g. PST)
PCR2:Tailed PCR adds unique barcode sequences andIllumina capture site to target
Illumina Sequencing: Single end base reads with dual 6 base index sequencing
B GoldenGate Assay
~50 nt
~50 nt
TE-specific oligo (TSO)
P1
Universal non-specific oligo (UNSO)
P3
C digitalMLPA
Binding site P7 oligo
Left probe oligo (LPO)
Right probe oligo (RPO)
digitalMLPA read identifier
Barcode Oligo
P5
Rd1 sequencing oligo
Barcode
Sample denaturation and probe hybridisation: The LPO and RPO parts of each probe hybridise to immediately adjacent target DNA sequences.
TE insertion junction site
Probe ligation: The second step is the ligation of probe oligos that hybridised to immediately adjacent target sequences.

FIGURE 3
Genotyping-by-Sequencing (GBS) is a powerful technique for detecting genetic variations using Next-Generation Sequencing (NGS) platforms. This technique has several approaches, some of which are highlighted below. **(A)**. One such approach is Genotyping-in-Thousands by sequencing (GT-seq), which uses NGS of multiplex PCR amplicons to produce genotypes from a small panel of mobile element insertions for thousands of individuals on Illumina NGS platforms. The basic principles and steps involved in this technology are also applicable to other NGS applications. **(B)** Another approach is the GoldenGate Assay, which can be designed for genotyping mobile element insertion sites involving a high degree of loci multiplexing (1536-plex). For all mobile element insertions sites, two oligonucleotides are designed, one is TE-specific to the mobile element site, called TE-Specific Oligos (TSO), and the second oligo, called the Universal Non-Specific Oligo (UNSO), for example, a universal primer able to anneal to palindromic sequences (PST), hybridizes downstream from the TE site. All oligonucleotides contain regions of genomic complementarity and universal PCR primer sites, while the UNSO also has a unique address sequence that targets a particular bead type. The extension of the appropriate TSO and ligation of the extended product to the UNSO join the genotype present at the TE site to the address sequence on the UNSO. **(C)** digitalMLPA is a semi-quantitative technology that is used to detect relative copy number variation and identify specific polymorphisms. With digitalMLPA, up to 1000 target sequences can be identified in a single multiplex PCR-based reaction.

powerful tool for genome profiling and can provide valuable insights into the genetic diversity and evolution of plant populations. However, careful consideration must be given to the specificities of the sequencing platform and the design of the primers used to ensure accurate and efficient sequencing.

Multiple NGS platforms and "omics" (i.e. genomics, proteomics, transcriptomics, epigenomics, and metabolomics) technologies offer many advantages and can therefore be used as high-throughput genotyping platforms. NGS platforms have revolutionized genomic approaches and have drastically reduced the time and cost required to obtain a DNA sequence. Markers can then be used for various applications, such as population genetics, association studies, and GWAS. The discovery of high-throughput genetic markers and the use of restriction enzymes for genotyping have several advantages and will become the methods of choice for marker research (Kalendar et al., 2022b). One of the advantages of these methods is that they can be used both for model organisms with high-quality reference genome sequences and for non-model species that do not have existing genomic data. Using these evolving technological methods, linkage mapping or (Quantitative Trait Locus) QTL may identify recombination breakpoints and genomic regions that are differentially expressed among populations for quantitative genetic research, genotype progenies for MAS, or resolve phylogeographies of wild populations.

Omics technologies that promise to detect tissue-specific changes with increased sensitivity and allow simultaneous analysis of thousands of genes, proteins, or metabolites (Kroeger, 2006) will increasingly provide sufficient data to create many digital platforms. The integration of new omics technologies with traditional breeding methods is important for seed production in the agriculture industry, as it helps plant growers make informed decisions based on genetic and molecular information. This combination of technologies will assist in meeting commercial criteria for seed production (Flavell, 2010; Li et al., 2018). Traditional plant breeding methods that rely exclusively on phenotypic mapping of desirable traits have limitations in defining gene-trait relationships (Flavell, 2010). The integration of modern omics technologies with traditional breeding can help identify specific gene functions related to seed development, leading to improved seed quality in economically important crops. The use of these cutting-edge technologies can benefit the modern seed industry by providing better tools for seed production (Toubiana and Fait, 2012). Studying the activity of genes at specific plant growth stages, such as grain filling or embryogenesis, may reveal critical components that regulate important metabolic processes that can be used to improve seed quality (Thompson et al., 2009).

Conclusion

TEs are a core component of all eukaryotic genomes, each with its own unique history and level of relatedness within the same species and between related species. Retrotransposons are widely present throughout the genome, and their replicative transposition allows them to insert themselves into the genome without removing the original elements. For closely related species, the sequences of a particular retrotransposon will be similar, reflecting the degree of relatedness between these species. Various applications of molecular markers have been developed to exploit the fact that these genetic elements are ubiquitous and their ability to be stably integrated into dispersed chromosomal localities that are polymorphic within a species. The ongoing development of molecular marker technologies is directly related to the deployment of high-throughput genotype sequencing platforms, and this research is of considerable significance. Digital NGS-based platforms can be used to study the transposition and site-specific recombination of TEs. Genotyping by sequencing includes a wide range of approaches for detecting genetic variations, and each of these approaches has unique advantages and limitations. Nonetheless, the advances in NGS technologies have greatly improved the ability to investigate the diverse nature of polymorphisms and their role in phenotypic variation. These platforms allow for high-throughput and cost-effective analysis of large amounts of genomic data, which can be used to identify and characterize TEs and their impact on the genome. This information can then be used for various applications, such as genetic characterization, genetic mapping, and marker-assisted selection.

Author contributions

RK, SM, YA, and YK wrote the manuscript. All authors reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP14869076).

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akakpo, R., Carpentier, M. C., Hsing, Y. I., and Panaud, O. (2020). The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytol.* 226, 44–49. doi: 10.1111/nph.16356

Alheit, K. V., Reif, J. C., Maurer, H. P., Hahn, V., Weissmann, E. A., Miedaner, T., et al. (2011). Detection of segregation distortion loci in triticale (x triticosecale wittmack) based on a high-density DArT marker consensus genetic linkage map. *BMC Genomics* 12, 380. doi: 10.1186/1471-2164-12-380

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

Argout, X., Salse, J., Aury, J. M., Guiltinan, M. J., Droc, G., Gouzy, J., et al. (2011). The genome of theobroma cacao. *Nat. Genet.* 43, 101–108. doi: 10.1038/ng.736

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3, e3376. doi: 10.1371/journal.pone.0003376

Belyayev, A., Kalendar, R., Brodsky, L., Nevo, E., Schulman, A. H., and Raskina, O. (2010). Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. *Mobile DNA* 1, 6. doi: 10.1186/1759-8753-1-6

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 65, 505–530. doi: 10.1146/annurev-arplant-050213-035811

Benoit, M., Drost, H. G., Catoni, M., Gouil, Q., Lopez-Gomollon, S., Baulcombe, D., et al. (2019). Environmental and epigenetic regulation of rider retrotransposons in tomato. *PLoS Genet.* 15, e1008370. doi: 10.1371/journal.pgen.1008370

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., et al. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* 24, 1242–1255. doi: 10.1105/tpc.111.095232

Campbell, N. R., Harmon, S. A., and Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Mol. Ecol. Resour* 15, 855–867. doi: 10.1111/1755-0998.12357

Cho, J., Benoit, M., Catoni, M., Drost, H. G., Brestovitsky, A., Oosterbeek, M., et al. (2019). Sensitive detection of pre-integration intermediates of long terminal repeat retrotransposons in crop plants. *Nat. Plants* 5, 26–33. doi: 10.1038/s41477-018-0320-9

Chuong, E. B., Elde, N. C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* 351, 1083–1087. doi: 10.1126/science.aad5497

Dai, X., Wang, H., Zhou, H., Wang, L., Dvorak, J., Bennetzen, J. L., et al. (2018). Birth and death of LTR-retrotransposons in aegilops tauschii. *Genetics* 210, 1039–1051. doi: 10.1534/genetics.118.301198

Das, G., Patra, J. K., and Baek, K. H. (2017). Insight into MAS: a molecular tool for development of stress resistant and quality of rice through gene stacking. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00985

Du, Z., D'alessandro, E., Zheng, Y., Wang, M., Chen, C., Wang, X., et al. (2022). Retrotransposon insertion polymorphisms (RIPs) in pig coat color candidate genes. *Anim. (Basel)* 12, 969. doi: 10.3390/ani12080969

Elbaidouri, M., Chaparro, C., and Panaud, O. (2013). "Use of next generation sequencing (NGS) technologies for the genome-wide detection of transposition," In: Peterson, T. (eds) *Plant transposable elements* (Springer) 1057, 265–274. doi: 10.1007/978-1-62703-568-2_19

El Baidouri, M., and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biol. Evol.* 5, 954–965. doi: 10.1093/gbe/evt025

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Flavell, R. (2010). From genomics to crop breeding. *Nat. Biotechnol.* 28, 144–145. doi: 10.1038/nbt0210-144

Flavell, A. J., Knox, M. R., Pearce, S. R., and Ellis, T. H. (1998). Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. *Plant J.* 16, 643–650. doi: 10.1046/j.1365-313x.1998.00334.x

Gaubert, H., Sanchez, D. H., Drost, H. G., and Paszkowski, J. (2017). Developmental restriction of retrotransposition activated in arabidopsis by environmental stress. *Genetics* 207, 813–821. doi: 10.1534/genetics.117.300103

Grandbastien, M. (1998). Activation of plant retrotransposons under stress conditions. *Trends Plant Sci.* 3, 181–187. doi: 10.1016/S1360-1385(98)01232-1

Hawkins, J. S., Kim, H., Nason, J. D., Wing, R. A., and Wendel, J. F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in gossypium. *Genome Res.* 16, 1252–1261. doi: 10.1101/gr.5282906

Hirsch, C. D., and Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta Gene Regul. Mech.* 1860, 157–165. doi: 10.1016/j.bbagrm.2016.05.010

Holligan, D., Zhang, X., Jiang, N., Pritham, E. J., and Wessler, S. R. (2006). The transposable element landscape of the model legume lotus japonicus. *Genetics* 174, 2215–2228. doi: 10.1534/genetics.106.062752

Hosid, E., Brodsky, L., Kalendar, R., Raskina, O., and Belyayev, A. (2012). Diversity of long terminal repeat retrotransposon genome distribution in natural populations of the wild diploid wheat aegilops speltoides. *Genetics* 190, 263–274. doi: 10.1534/genetics.111.134643

Huang, C. R. L., Burns, K. H., and Boeke, J. D. (2012). Active transposition in genomes. *Annu. Rev. Genet.* 46, 651–675. doi: 10.1146/annurev-genet-110711-155616

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., et al. (2009). The genome of the cucumber, cucumis sativus l. *Nat. Genet.* 41, 1275–1281. doi: 10.1038/ng.475

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I., and Paszkowski, J. (2011). An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* 472, 115–119. doi: 10.1038/nature09861

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Kalendar, R. (2022). A guide to using FASTPCR software for PCR, in silico PCR, and oligonucleotide analysis. *Methods Mol. Biol.* 2392, 223–243. doi: 10.1007/978-1-0716-1799-1_16

Kalendar, R., Antonius, K., Smykal, P., and Schulman, A. H. (2010). iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. *Theor. Appl. Genet.* 121, 1419–1430. doi: 10.1007/s00122-010-1398-2

Kalendar, R., Baidyussen, A., Serikbay, D., Zotova, L., Khassanova, G., Kuzbakova, M., et al. (2022a). Modified "Allele-specific qPCR" method for SNP genotyping based on FRET. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.747886

Kalendar, R., Hunter, C., and Orbovic, V. (2022b). Editorial: innovative applications of sequencing technologies in plant science. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1058347

Kalendar, R., Raskina, O., Belyayev, A., and Schulman, A. H. (2020). Long tandem arrays of Cassandra retroelements and their role in genome dynamics in plants. *Int. J. Mol. Sci.* 21, 2931. doi: 10.3390/ijms21082931

Kalendar, R., Sabot, F., Rodriguez, F., Karlov, G. I., Natali, L., and Alix, K. (2021a). Editorial: mobile elements and plant genome evolution, comparative analyzes and computational tools. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.735134

Kalendar, R., and Schulman, A. (2006). IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. *Nat. Protoc.* 1, 2478–2484. doi: 10.1038/nprot.2006.377

Kalendar, R., Shustov, A. V., Akhmetollayev, I., and Kairov, U. (2022c). Designing allele-specific competitive-extension PCR-based assays for high-throughput genotyping and gene characterization. *Front. Mol. Biosci.* 9. doi: 10.3389/fmolb.2022.773956

Kalendar, R., Shustov, A., and Schulman, A. (2021b). Palindromic sequence-targeted (PST) PCR, version 2: an advanced method for high-throughput targeted gene characterization and transposon display. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.691940

Kalendar, R., Shustov, A. V., Seppänen, M. M., Schulman, A. H., and Stoddard, F. L. (2019). Palindromic sequence-targeted (PST) PCR: a rapid and efficient method for high-throughput gene characterization and genome walking. *Sci. Rep.* 9, 17707. doi: 10.1038/s41598-019-54168-0

Kalendar, R., Tanskanen, J., Chang, W., Antonius, K., Sela, H., Peleg, O., et al. (2008). Cassandra Retrotransposons carry independently transcribed 5S RNA. *Proc. Natl. Acad. Sci. United States America* 105, 5833–5838. doi: 10.1073/pnas.0709698105

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

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large Retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166, 1437–1450. doi: 10.1534/genetics.166.3.1437

Kent, T. V., Uzunovic, J., and Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philos. Trans. R Soc. Lond B Biol. Sci.* 372, 20160458. doi: 10.1098/rstb.2016.0458

Kovach, A., Wegrzyn, J. L., Parra, G., Holt, C., Bruening, G. E., Loopstra, C. A., et al. (2010). The pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genomics* 11, 420. doi: 10.1186/1471-2164-11-420

Kroeger, M. (2006). How omics technologies can contribute to the '3R'principles by introducing new strategies in animal testing. *Trends Biotechnol.* 24, 343–346. doi: 10.1016/j.tibtech.2006.06.003

Kwolek, K., Kedzierska, P., Hankiewicz, M., Mirouze, M., Panaud, O., Grzebelus, D., et al. (2022). Diverse and mobile: eccDNA-based identification of carrot low-copy-number LTR retrotransposons active in callus cultures. *Plant J.* 110, 1811–1828. doi: 10.1111/tpj.15773

Li, Y., Xiao, J., Chen, L., Huang, X., Cheng, Z., Han, B., et al. (2018). Rice functional genomics research: past decade and future. *Mol. Plant* 11, 359–380. doi: 10.1016/j.molp.2018.01.007

Lynch, V. J., Leclerc, R. D., May, G., and Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* 43, 1154. doi: 10.1038/ng.917

Makhoul, M., Rambla, C., Voss-Fels, K. P., Hickey, L. T., Snowdon, R. J., and Obermeier, C. (2020). Overcoming polyploidy pitfalls: a user guide for effective SNP conversion into KASP markers in wheat. *Theor. Appl. Genet.* 133, 2413–2430. doi: 10.1007/s00122-020-03608-x

Mcclintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. United States America* 36, 344–355. doi: 10.1073/pnas.36.6.344

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. doi: 10.1038/nature06856

Mir, R. R., and Varshney, R. K. (2012). "Future prospects of molecular markers in plants," in *Molecular markers in plants*. Ed. R. J. Henry (Wiley-Blackwell Publication), 169–190. doi: 10.1002/9781118473023.ch10

Mirouze, M., and Vitte, C. (2014). Transposable elements, a treasure trove to decipher epigenetic variation: insights from arabidopsis and crop epigenomes. *J. Exp. Bot.* 65, 2801–2812. doi: 10.1093/jxb/eru120

Moisy, C., Schulman, A. H., Kalendar, R., Buchmann, J. P., and Pelsy, F. (2014). The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years. *Theor. Appl. Genet.* 127, 1223–1235. doi: 10.1007/s00122-014-2293-z

Monden, Y., Yamaguchi, K., and Tahara, M. (2014). Application of iPBS in high-throughput sequencing for the development of retrotransposon-based molecular markers. *Curr. Plant Biol.* 1, 40–44. doi: 10.1016/j.cpb.2014.09.001

Papolu, P. K., Ramakrishnan, M., Mullasseri, S., Kalendar, R., Wei, Q., Zou, L. H., et al. (2022). Retrotransposons: how the continuous evolutionary front shapes plant genomes for response to heat stress. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1064847

Papolu, P. K., Ramakrishnan, M., Wei, Q., Vinod, K. K., Zou, L. H., Yrjala, K., et al. (2021). Long terminal repeats (LTR) and transcription factors regulate PHRE1 and PHRE2 activity in moso bamboo under heat stress. *BMC Plant Biol.* 21, 585. doi: 10.1186/s12870-021-03339-1

Pecinka, A., Dinh, H. Q., Baubec, T., Rosa, M., Lettner, N., and Mittelsten Scheid, O. (2010). Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in arabidopsis. *Plant Cell* 22, 3118–3129. doi: 10.1105/tpc.110.078493

Qiu, D., Gao, M., Li, G., and Quiros, C. (2009). Comparative sequence analysis for brassica oleracea with similar sequences in b. rapa and arabidopsis thaliana. *Plant Cell Rep.* 28, 649–661. doi: 10.1007/s00299-008-0661-3

Queen, R., Gribbon, B., James, C., Jack, P., and Flavell, A. (2004). Retrotransposon-based molecular markers for linkage and genetic diversity analysis in wheat. *Mol. Genet. Genomics* 271, 91–97. doi: 10.1007/s00438-003-0960-x

Ramakrishnan, M., Papolu, P. K., Mullasseri, S., Zhou, M., Sharma, A., Ahmad, Z., et al. (2023). The role of LTR retrotransposons in plant genetic engineering: how to control their transposition in the genome. *Plant Cell Rep.* 42, 3–15. doi: 10.1007/s00299-022-02945-z

Ramakrishnan, M., Zhang, Z., Mullasseri, S., Kalendar, R., Ahmad, Z., Sharma, A., et al. (2022). How stress memory is regulated for cold and heat stress responses in plants: scope for future climate resilient agriculture. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1075279

Rebollo, R., Karimi, M. M., Bilenky, M., Gagnier, L., Miceli-Royer, K., Zhang, Y., et al. (2011). Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS Genet.* 7, e1002301. doi: 10.1371/journal.pgen.1002301

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Schrader, L., and Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. *Mol. Ecol.* 28, 1537–1549. doi: 10.1111/mec.14794

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (Fragaria vesca). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Springer, N. M., and Jackson, S. A. (2010). Realizing the potential of genomics for crop improvement. *Briefings Funct. Genomics* 9, 93–94. doi: 10.1093/bfgp/elq006

Thompson, R., Burstin, J., and Gallardo, K. (2009). Post-genomics studies of developmental processes in legume seeds. *Plant Physiol.* 151, 1023–1029. doi: 10.1104/pp.109.143966

Tittel-Elmer, M., Bucher, E., Broger, L., Mathieu, O., Paszkowski, J., and Vaillant, I. (2010). Stress-induced activation of heterochromatic transcription. *PLoS Genet.* 6, e1001175. doi: 10.1371/journal.pgen.1001175

Toubiana, D., and Fait, A. (2012). "Metabolomics-assisted crop breeding towards improvement in seed quality and yield," in *Seed development: OMICS technologies toward improvement of seed quality and crop yield* (Dordrecht: Springer), 453–475. doi: 10.1007/978-94-007-4749-4_22

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., et al. (2018). The 100 000 genomes project: bringing whole genome sequencing to the NHS. *BMJ* 361, k1687. doi: 10.1136/bmj.k1687

Valli, A. A., Gonzalo-Magro, I., and Sanchez, D. H. (2023). Rearranged endogenized plant pararetroviruses as evidence of heritable RNA-based immunity. *Mol. Biol. Evol.* 40, msac240. doi: 10.1093/molbev/msac240

Van Orsouw, N. J., Hogers, R. C., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., et al. (2007). Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One* 2, e1172. doi: 10.1371/journal.pone.0001172

Vitte, C., and Panaud, O. (2005). LTR Retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic Genome Res.* 110, 91–107. doi: 10.1159/000084941

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Van De Lee, T., Hornes, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23, 4407–4414. doi: 10.1093/nar/23.21.4407

Wang, H., and Liu, J. S. (2008). LTR Retrotransposon landscape in medicago truncatula: more rapid removal than in rice. *BMC Genomics* 9, 382. doi: 10.1186/1471-2164-9-382

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-Gonzalez, R. H., et al. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* 19, 103. doi: 10.1186/s13059-018-1479-0

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

Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M., et al. (2009). A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.* 59, 712–722. doi: 10.1111/j.1365-313X.2009.03911.x

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., and van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* 319, 1527–1530. doi: 10.1126/science.1153040

Zervudacki, J., Yu, A., Amesefe, D., Wang, J., Drouaud, J., Navarro, L., et al. (2018). Transcriptional control and exploitation of an immune-responsive family of plant retrotransposons. *FEBS J.* 37, e98482. doi: 10.15252/embj.201798482

Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., et al. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nat. Commun.* 10, 1494. doi: 10.1038/s41467-019-09518-x

Zhang, X., and Wessler, S. R. (2004). Genome-wide comparative analysis of the transposable elements in the related species arabidopsis thaliana and brassica oleracea. *Proc. Natl. Acad. Sci. United States America* 101, 5589–5594. doi: 10.1073/pnas.0401243101

TYPE Original Research
PUBLISHED 14 July 2023
DOI 10.3389/fpls.2023.1103035

OPEN ACCESS

EDITED BY
Ruslan Kalendar,
University of Helsinki, Finland

REVIEWED BY
Khalil Kashkush,
Ben-Gurion University of the Negev, Israel
Panagiotis Ioannidis,
Foundation for Research and Technology Hellas (FORTH), Greece

*CORRESPONDENCE
Joris Argentin
✉ joris.argentin@chu-angers.fr
Dan Bolser
✉ dan@geromics.co.uk

RECEIVED 19 November 2022
ACCEPTED 14 June 2023
PUBLISHED 14 July 2023

CITATION
Argentin J, Bolser D, Kersey PJ and Flicek P (2023) Comparative analysis of repeat content in plant genomes, large and small. *Front. Plant Sci.* 14:1103035.
doi: 10.3389/fpls.2023.1103035

Comparative analysis of repeat content in plant genomes, large and small

Joris Argentin[1]*, Dan Bolser[2]*, Paul J. Kersey[2,3] and Paul Flicek[2]

[1]Institut de Biologie en Santé, Centre Hospitalier Universitaire (CHU) d'Angers, Angers, France, [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom, [3]Digital Revolution, Royal Botanic Gardens, Kew, Richmond, United Kingdom

The DNA Features pipeline is the analysis pipeline at EMBL-EBI that annotates repeat elements, including transposable elements. With Ensembl's goal to stay at the cutting edge of genome annotation, we proved that this pipeline needed an update. We then created a new analysis that allowed the Ensembl database to store the repeat classification from the PGSB repeat classification (Recat). This new dataset was then fetched using Perl scripts and used to prove that the pipeline modification induced a gain in sensitivity. Finally, we performed a comparative analysis of transposable element distribution in all plant species available, raising new questions about transposable elements in certain branches of the taxonomic tree.

KEYWORDS

transposons, plants, pipeline, annotation, comparative analysis

1 Introduction

Transposable elements (TEs) are a major group of genomic repetitive elements. TEs encompass many genomic structures that all have in common the ability to move from one genomic location to another in a process called transposition. Transposition itself covers various mechanisms.

Approximately 3% to 80% of a plant's genome is composed of TEs. By their nature as repetitive sequences, TEs are major contributors, with whole genome duplications, to the large genome size reported in plant genomes (Muotri et al., 2007). The predominance of transposons makes repeat content detection essential. Each species has its own history of repeat expansions and removals, which poses intriguing questions about evolution, host control, transposon countermeasures, and other factors that influence genomic repeats.

1.1 Transposition mechanisms

There are two main ways for a repeat element to move in the genome. These two ways to perform a transposition will define the main classes of repeat elements.

The replicative transposition, or retrotransposition, implies a reverse transcription of the TEs. LTR-Retrotransposons are flanked by long terminal repeat (LTR) and code for their own transposition proteins. As for the non-LTR coding elements, long interspersed nuclear elements (LINEs) also code for their own transposition proteins, while short interspersed nuclear elements (SINEs) are non-autonomous. Both LTR and non-LTR transpositions are considered a "copy–paste system" and result in the duplication of the repeat element.

The other mechanism, similar to a "cut–paste" system, is called conservative transposition. It involves the transposase, coded by the gene in the transposon sequence, and inverted tandem repeats (ITRs). The transposase will then bind to both ITRs, cleave the DNA, forming a circular structure, and transport the TEs to the target site.

1.2 Classifications of transposable elements

TEs are not under large selection pressure, so multiple copies accumulate mutations, to the point of losing all transposition activities. This accumulation of mutations can also play beneficial roles in evolutionary processes (Chénais et al., 2012), creating variety in genetic portions that can be transferred with TEs. Therefore, transposable elements have a relatively short transposition activity, and active elements in modern genomes are rare. This degeneration can also happen with repeat elements getting inserted within other elements, ultimately leading to complex, nested, and degenerated structures, making homology-defined families not straightforward.

Wicker et al. (2007) defined a transposable element family with these criteria: "two elements belong to the same family if they share 80% (or more) sequence identity in at least 80% of their coding or internal domain, or within their terminal repeat regions, or in both".

1.3 TE detection and annotation

During gene annotation processes, repeat elements are masked to minimize unwanted transposon-related gene calls due to the repetitive nature of transposable elements. This detection is mainly performed by searching the genome sequence against a reference library, like RepeatMasker[1]. libraries are automatically built from motif discovery tools based on repetitiveness (Benson, 1999), specific TE structures, or comparative genomics (Ou et al., 2019). However, these automated methods have flaws in accuracy and still need manual annotation. EMBL-EBI, displaying annotation information for scientists worldwide in the Ensembl browser, must be on the cutting edge of transposable element annotation. In 2020, repeat elements at EBI were annotated by the DNA Features pipeline. This pipeline ran RepeatMasker with the Repbase (Bao et al., 2015) repeat library.

In the current work, we aimed to extend our existing pipeline for repeat annotation to produce a comprehensive catalog of repeat families across the complete range of sequenced plant genomes. We ran the existing repeat annotation pipeline across all genomes in Ensembl Plants and compared the results to the literature. This was necessary to assess the need to implement a new, more specific repeat library. We had to extend the pipeline to apply and import repeat classification from the PGSB repeat classification (REcat; Nussbaumer et al., 2013), similar to the way repeat classification is added from Repbase. The new data generated using the REcat were used to quantify the improvement in TE detection. REcat integrates existing classifications for repetitive elements into a more detailed hierarchical tree structure. The resulting catalog of classified repeats was then compared across the taxonomic space to establish the evolutionary trends of repeat expansion and to extend understanding of chromosomal architecture in plants.

2 Materials and methods

2.1 Literature monitoring

2.1.1 Bibliographic research

The repeat distribution table, for the barley (*Hordeum vulgare*) genome described in Mascher et al. (2017), was used as the working base in our repeat statistics spreadsheet. For its completeness, this table was also used as the standard of quality for other repeat distribution tables. To find genome-wide repeat distribution reports, two queries on PubMed were made: one using Mesh terms (*((Genus + species name[All Fields]) AND Interspersed Repetitive Sequences[MeSH Terms]) AND plants, genetics[MeSH Terms]*) and the other for repeat distribution tables in genome-wide assembly reports *via* the linked articles in the National Center for Biotechnology Information (NCBI) Genome website (*Genus species[orgn]*).

2.1.2 Quality control of repeat distribution tables

All repeat distribution tables found using both of the methods described in the previous paragraph had to have a quality at least equivalent to the previous standard. A table could still pass quality control if in a repeat type a superfamily-related row was missing, but all other family rows for this type were present. In that particular case, statistics for the missing row were considered zero. The various classifications used in the articles were normalized using the PGSB repeat classification (Nussbaumer et al., 2013). Due to the quality control, processing of annotation statistics was only performed on eight of the 53 species (including a cultivar) present in the database: two genome-wide repeat distribution studies were found for *Brachypodium distachyon* (Initiative, 2010) and *Amborella trichopoda* (Zuccolo et al., 2011); four assembly reports that comprised relevant repeat annotation statistics were found for Japanese and Indian rice (Mahesh et al., 2016) (*Oryza sativa* sp. *japonica* cv. *Nipponbare* and sp. *indica* cv.

[1] Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org

93-11), soybean (Schmutz et al., 2010) (*Glycine max*), cacao (Motamayor et al., 2013) (*Theobroma cacao*), and maize (Jiao et al., 2017) (*Zea mays*) respectively.

2.2 Comparison of repeat distributions between the DNA Features pipeline and scientific articles

The statistics were stored in a Google Sheets spreadsheet (Bolser et al., 2015). This spreadsheet comprised six metrics (percentage of the genome covered, percentage of total transposable element length, base pairs covered, number of features, size in Mbp, and average length in bp) for classes, superfamilies, and the main families of transposable elements, similar to the statistics presented in Mascher et al. (2017). Repeat distribution statistics from the literature were also stored in this spreadsheet, next to their corresponding distribution from the pipeline. The percentage of the genome covered and the number of features for all transposable elements (or the "Transposable elements" repeat sequence group) were used as metrics to compare annotation performances between the initial and modified DNA Features pipeline and the literature used as reference.

2.3 Statistics and software

2.3.1 Cluster computing

Data processing of the pipeline was performed on the EBI cluster monitored by the LSF[2] and eHive[3] systems.

2.3.2 Annotation of mobile elements in the pipeline

What is referred to as the "initial pipeline" is the DNA Features pipeline in its March 2020 version (Figure 1). The initial pipeline run used RepeatMasker with default parameters and the Repbase repeat library on all 53 plant genomes of version 39, release 92, of Ensembl. What is referred to as the "updated pipeline" is the DNA Features pipeline in its May 2020 version. A run of the updated pipeline was made with RepeatMasker on low-sensitivity parameters and used REdat as an additional repeat library.

2.3.3 Comparative analysis of repeat elements distributions

The file containing all repeat element features extracted from the Ensembl database was post-processed by a Perl script to remove every line that was not a transposable element. All REcat keys that had four levels of classification (group, class, type, superfamily, or "unclassified") were then extended with an additional "unclassified" level, and every REcat key with six levels (group, class, type, superfamily, family, and "unclassified") was trimmed of their "unclassified" final classification level, using a second Perl script. This modification led to 100 unique REcat keys with five levels. Finally, the processed repeat feature data were treated by a third Perl script. This script used a multi-dimensional hash table as a data structure, with the REcat keys as keys and the species name as value, with this species name also a key for an array of four key metrics, as follows: a binary value for the presence or absence of a key in the given species, number of copies, feature coverage in bp, and feature coverage in the percentage of the genome covered. To compute and visualize the distributions of repeat elements in plant species, all four types of values for every key/species couple were stored in four R vectors and then converted into four matrices of 53 (for 53 species) rows by 100 (for 100 unique REcat keys) columns. The *dist* R module set up distance matrices for the initial 53 by 100 matrices. This module was used with default parameters, except for the "presence/absence" matrix, where the distance parameter used was "binary", as the values for this particular matrix were binary. Then, the distance matrices were processed with the *hclust* module, also with default parameters, to build clusters from the distance values and then creating views in the form of dendrograms (Figure 2). The values of the distance matrices have also been visualized in a heatmap (Figure 3).

3 Results

3.1 Presentation of results in the Ensembl Genome web interface

The data generated from the updated pipeline run have been used as a testing set by the Ensembl Genome team when setting up a web sandbox, and they were made available in the public Ensembl 93 release, with a tag indicating "REdat" data source, to distinguish between Repbase and REdat annotations.

3.2 Initial pipeline run and comparison to reported values

To determine if the DNA Features pipeline was sensitive enough to compete or at least come close to current specialized TE detection tools, we compared repeat distributions produced by the pipeline with repeat distributions from the literature that passed through the established quality control.

Total genome coverage and the total number of TE features detected were used as comparison metrics between data sources. The fractions between the pipeline metrics and the article metrics, converted in percentage, were used to determine differences between the pipeline statistics and scientific report statistics (Bolser et al., 2015). On average, the DNA Features pipeline masks 50.32% of reported sequences and detects 116% of reported features. RepeatMasker, with the default library, found too many repetitive structures when compared to what is considered standard. Worse

[2] IBM's LSF product page, https://www.ibm.com/products/hpc-workload-management

[3] eHive production system page on Ensembl documentation, https://www.ensembl.org/info/docs/eHive.html

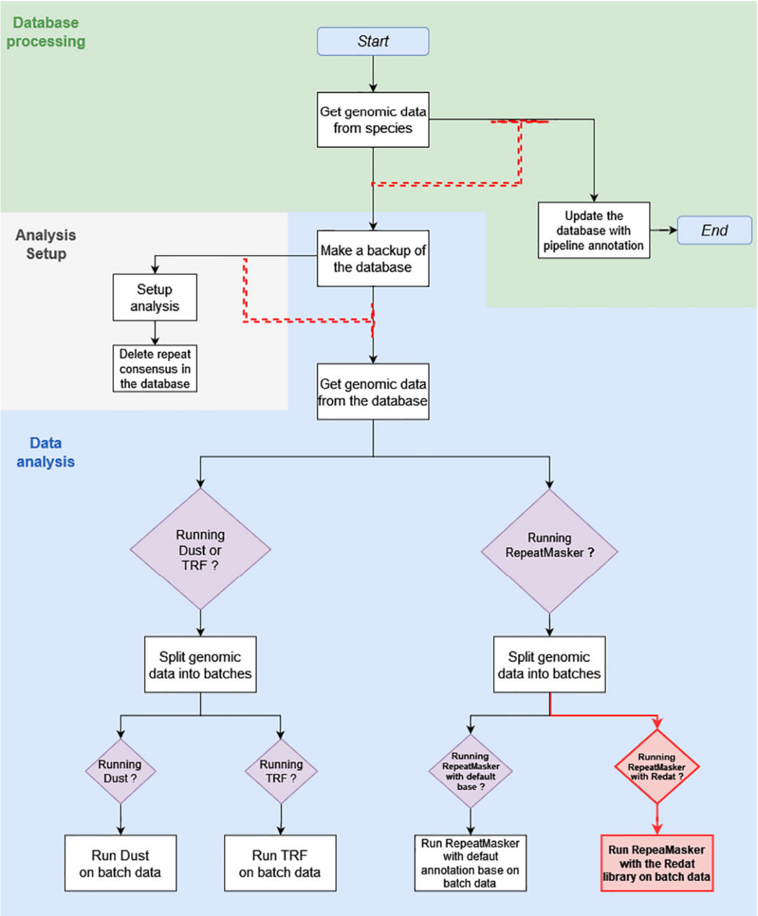

Database processing
Start
Get genomic data from species
Update the database with pipeline annotation
End
Analysis Setup
Make a backup of the database
Setup analysis
Delete repeat consensus in the database
Get genomic data from the database
Data analysis
Running Dust or TRF ?
Running RepeatMasker ?
Split genomic data into batches
Split genomic data into batches
Running Dust ?
Running TRF ?
Running RepeatMasker with default base ?
Running RepeatMasker with Redat ?
Run Dust on batch data
Run TRF on batch data
Run RepeatMasker with defaut annotation base on batch data
Run RepeaMasker with the Redat library on batch data

FIGURE 1
This diagram shows the March 2020 Repeat Features pipeline structure, with modifications made in May 2020 highlighted in red. The white square boxes are the pipeline analyses. Each box is associated with a module, written for the most part in Perl. Boxed purple diamonds are conditional structures. Analyses following these structures are only executed if the condition specified in the diamond is met, in this case when the module specified by the condition has been activated upon pipeline initialization. Black and red arrows show the sequence of analyses. Red dashed arrows are dependent dataflows, where the analysis at the head does not run as long as there are jobs pending in the analysis, or a group of analyses under the arrow base.

still, this overabundance of detected repetitive structures provides masking that is below this very standard.

The very high value for detected features is mostly due to the extremely high value of features detected for *Z. mays*, as the pipeline detects 490% more features than the literature used as standard. This fivefold increase in the number of reported features results in a genome coverage 20% higher than the standard (Figure 4A). When *Z. mays* is removed from the dataset, the pipeline detects, on average, 48.86% fewer features than the reports in the literature. In contrast to the extreme values, the genome coverage and the number of features detected for *O. sativa* sp. *indica* cv. *93-11* are unexpectedly low, with 0.25% of genome masking and 0.74% of feature detection when compared to reports. There is a significant difference in *O. sativa* sp. *indica* between the pipeline statistics and the reference article and also between *O. sativa* sp. *indica* and *O. sativa* sp. *japonica* pipeline statistics. *O. sativa* sp. *japonica* and *O. sativa* sp. *indica*, being cultivars of the same species, should have highly similar repeat distributions.

We suspect that these differences come from the Repbase species-specific annotation, meaning that if few repeats from the Repbase dataset are labeled in the EMBL file as having an "Oryza indica" species annotation, only a few of these repeats are mapped on the *O. sativa* sp. *indica* genome, leading to underestimated statistics. The very high values for *Z. mays* might be the opposite of the same bias, as TEs in *Z. mays* are widely studied. With a plant genome masking 60% under the values considered the golden standard, it has been determined that modifications to the pipeline were relevant. However, these metrics could be restrictive and hide class- or type-specific variations that could only be detected by Repbase. Subsequent updates of Repbase and RepeatMasker could also reduce the significant differences in the considered metrics.

3.3 Pipeline extension, test, and rerun

This pipeline extension implemented a new RepeatMasker analysis, similar to the analysis with Repbase (or custom libraries). This new analysis used REdat as a repeat library. Then,

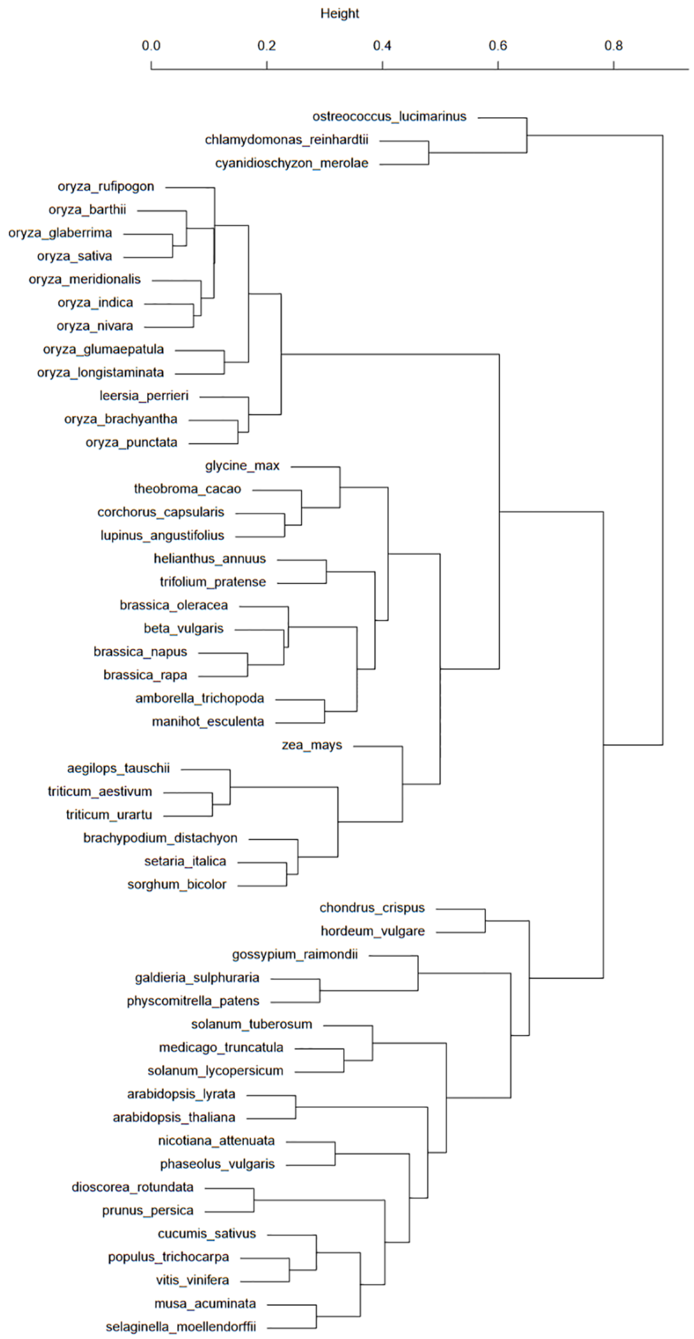


FIGURE 2
This taxonomic tree was computed from the presence/absence binary matrix of 100 transposons types in the 53 plant species available in the Ensembl database. The top scale shows the relative distances (from 0 to 1) between these species.

the RepeatMasker output, when used with REdat, could be parsed to provide a TE classification.

A new run was performed on the 53 species present in the Ensembl database, with the new analysis. RepeatMasker was used with low sensitivity. The intent was to determine if the implementation showed a significant improvement in the pipeline annotation capabilities. This run using the REdat library increased the average proximity to reference articles by 39% for genome coverage and by 13% for the number of features detected (Figure 4B). When compared to the initial pipeline run, the mean genome coverage increased by 22.7% (from 30.16% to 39.02%) and the average number of detected features by 55.36% (from 276,714 to 619,930). This means that running RepeatMasker with REdat on low sensitivity gives better results than RepeatMasker with Repbase on medium sensitivity.

If the extremely low values for Indian rice seem to have been solved, the extremely high values for maize remain after the update. This invalidates the hypothesis of the species-specific system in Repbase and raises a new hypothesis: it could be due to the variation in the number of reported transposable elements in a given species.

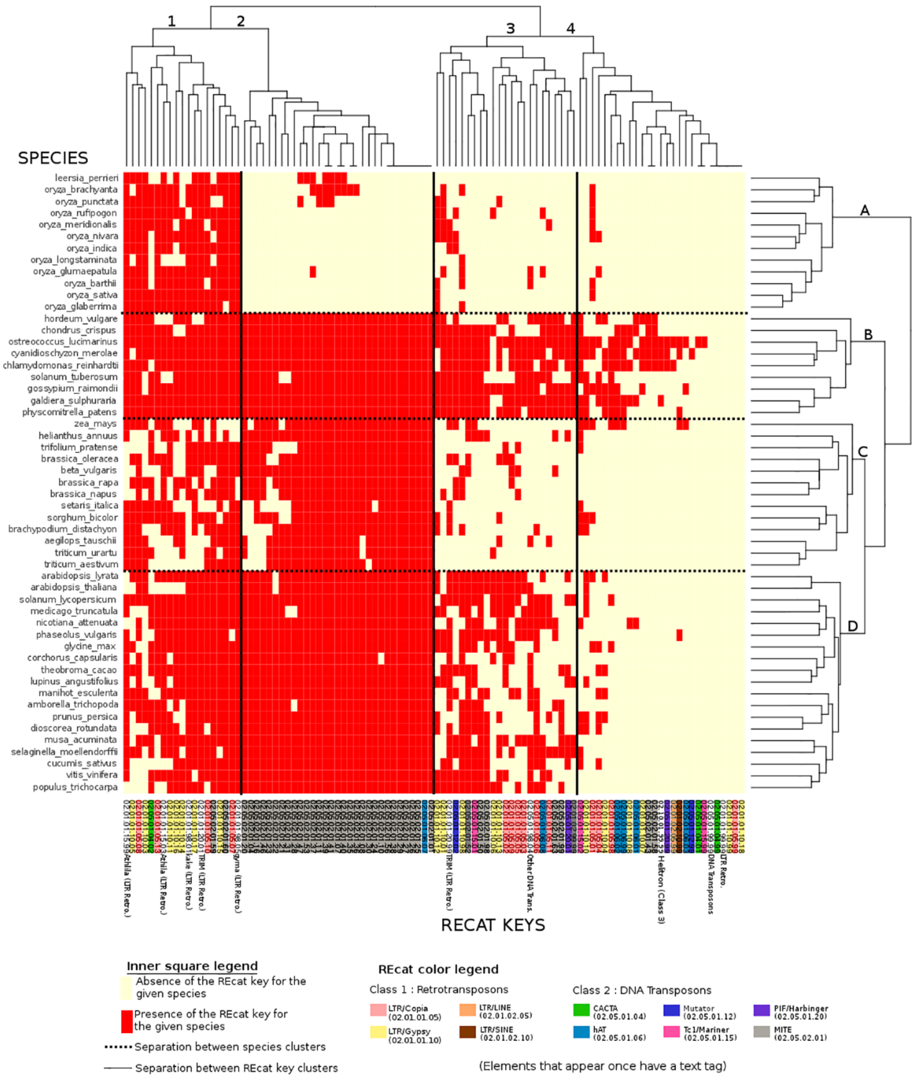


FIGURE 3
Binary heatmap, where each red point represents the presence of a REcat key on the x-axis in a given species on the y-axis. Each REcat has its five-digit code at the bottom. On the left side is the list of plant species in the Ensembl core database. The top and the left side are trees showing computed relations between keys and between species.

The high number of species that have values superior to 100% raises the question of the specificity behind the sensitivity or the number of false positives in the updated run and the need for manual validation. It could also benefit in the long run with higher RepeatMasker sensitivity.

3.4 Comparison between species in the context of the known taxonomy

Two figures, a dendrogram (Figure 2) and a heatmap (Figure 3), were produced from the comparative analysis of repeat distributions, using the presence/absence metric. Analysis heatmaps and dendrograms were produced for the three other metrics (copy number, feature coverage in bp, and percentage of the genome covered) but did not show significant results.

Figure 2 shows a good classification of rice and grasses in a common branch. However, the fact that this common branch is also populated with a large group of eudicots raises some questions about the TE history of these elements. One particular case of this separation of eudicots is about the Brassicaceae, with the *Brassica* genus in the branch comprising monocots and eudicots and the *Arabidopsis* genus in the "eudicots-only" branch. These species are separated by many events of whole genome duplications (Chalhoub et al., 2014). This study asks questions about the impact of whole genome duplications on transposon distribution and activity. Another case worth investigating is the presence of *H. vulgare* and *Gossypium raimondii* among algae and mosses.

If the dendrogram bootstrap has not been performed, its strength can nonetheless be assessed with the clusters from Figure 2. As grasses are grouped with eudicots, this branch

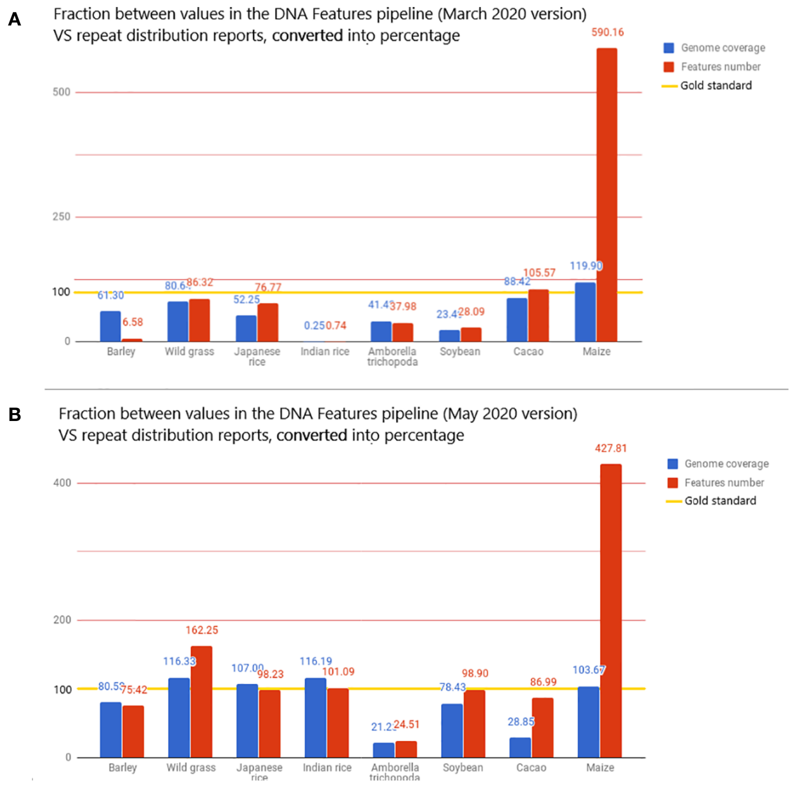


FIGURE 4
The value displayed is the fraction, converted in percentage, between the metrics from the DNA Features pipeline and repeat distributions from the literature. The reported values are then used here as the gold standard for transposon annotation quality: for each species, the value of genome coverage or the number of features reported in the related article is considered 100% in the bar chart and is highlighted with a gold line. These metrics are the total genome coverage for all transposable elements (blue) and the total number of transposable elements detected (red). The full dataset is available in the Google Docs spreadsheet (Bolser et al., 2015).

position could be considered unstable. *H. vulgare* is still grouped with mosses and algae, allowing us to reiterate our questions about barley TE history.

In the binary heatmap, the LTR/copia and LTR/gypsy superfamilies are spread over clusters 1, 2, and 4. Moreover, it is worth noting that cluster 4, which represents the most keys distributed around species, is mainly composed of LTR, which can be explained by the retroviral origin of these elements (Hayward, 2017). Cluster 4 also has two types of miniature inverted-repeat transposable elements (MITEs) that are known to have a large history of horizontal transfers (Zhang et al., 2018). However, the presence of a DNA Transposon/CACTA superfamily in this cluster is left unexplained.

MITEs, DNA transposons in general, are mostly absent from species cluster A. However, they are present in eudicots, algae, mosses, and other monocots, raising questions about the genetic appearance or removal event that occurred with MITEs and DNA transposons in rice.

Finally, this analysis is based on a binary matrix, and it could benefit from a deeper analysis using non-binary values. Moreover, the REcat key system has been altered to overcome Perl limitations. If this alteration still provides a solid analysis, with a hundred keys taken into account, an analysis using an imposed hierarchy tree and every REcat key available could provide more precise information.

4 Conclusions

The high number of repeat elements in plant genomes was a significant challenge in Ensembl's quest to annotate and align genomes. The detection of these elements by the DNA Features pipeline also had phylogenetic implications in the determination of repeat expansions and their subsequent removals. However, using RepeatMasker with Repbase, a library of eukaryotes, showed limitations. The implementation of the REdat repeat library proved to be needed and efficient, compared to repeat distribution from reference scientific articles. The new classification associated with REdat, REcat, also allowed a comparative analysis of the repeat element distributions in the 53 species available in the Ensembl Genome in 2020. The dendrogram from this comparative analysis showed promising results (Figure 2), in particular with monocots. However, strong discrepancies with the expectations, especially with *H. vulgare*, or the Brassicaceae, need to be investigated. The heatmap associated with this analysis shows the absence of MITEs in most species of rice, the presence of LTRs in every species cluster, and DNA transposons in a cluster comprising mosses, algae, "outliers", and *H. vulgare*. These particular clusterings need to be investigated, in addition to the differences between taxonomic space and repeat distributions.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://docs.google.com/spreadsheets/d/1kMMckERzqy9gwsFVWELfuj9q0dSKxh5IXL1D_S9wOug/edit#gid=359252355.

Author contributions

Bioinformatics execution, figure rendering, code editing and writing was made by JA. DB was Ensembl Plants team leader, JA's intership supervisor and provided scientific input, advise and data. PK and PF were respectively leaders of the Ensembl non-vertebrate genomics and vertebrate genomics teams. All authors contributed to the article and approved the submitted version.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6. doi: 10.1186/s13100-015-0041-9

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

Bolser, D., Naamati, G., and Argentin, J. (2015) *Repeat feature stats for 39 plant species*. Available at: https://docs.google.com/spreadsheets/d/1kMMckERzqy9gwsFVWELfuj9q0dSKxh5IXL1D_S9wOug.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-neolithic brassica napus oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435

Chénais, B., Caruso, A., Hiard, S., and Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* 509, 7–15. doi: 10.1016/j.gene.2012.07.042

Hayward, A. (2017). Origin of the retroviruses: when, where, and how? *Curr. Opin. Virol.* 25, 23–27. doi: 10.1016/j.coviro.2017.06.006. Animal models for viral diseases • Paleovirology.

Initiative, T. I. B. (2010). enGenome sequencing and analysis of the model grass brachypodium distachyon. *Nature* 463, 763–768. doi: 10.1038/nature08747

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., et al. (2017). enImproved maize reference genome with single-molecule technologies. *Nature* 546, 524–527. doi: 10.1038/nature22971

Mahesh, H. B., Shirke, M. D., Singh, S., Rajamani, A., Hittalmani, S., Wang, G.-L., et al. (2016). Indica rice genome assembly, annotation and mining of blast disease resistance genes. *BMC Genomics* 17, 242. doi: 10.1186/s12864-016-2523-7

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). enA chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, III, D., Cornejo, O., et al. (2013). The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol.* 14, r53. doi: 10.1186/gb-2013-14-6-r53

Muotri, A. R., Marchetto, M. C. N., Coufal, N. G., and Gage, F. H. (2007). enThe necessary junk: new functions for transposable elements. *Hum. Mol. Genet.* 16, R159–R167. doi: 10.1093/hmg/ddm196

Nussbaumer, T., Martis, M. M., Roessner, S. K., Pfeifer, M., Bader, K. C., Sharma, S., et al. (2013). MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.* 41, D1144–D1151. doi: 10.1093/nar/gks1153

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20, 275. doi: 10.1186/s13059-019-1905-y

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). enGenome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

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

Zhang, H.-H., Zhou, Q.-Z., Wang, P.-L., Xiong, X.-M., Luchetti, A., Raoult, D., et al. (2018). Unexpected invasion of miniature inverted-repeat transposable elements in viral genomes. *Mobile DNA* 9, 5–9. doi: 10.1186/s13100-018-0125-4

Zuccolo, A., Bowers, J. E., Estill, J. C., Xiong, Z., Luo, M., Sebastian, A., et al. (2011). A physical map for the amborella trichopoda genome sheds light on the evolution of angiosperm genome structure. *Genome Biol.* 12, R48. doi: 10.1186/gb-2011-12-5-r48

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 16 October 2023
DOI 10.3389/fpls.2023.1230250

OPEN ACCESS

EDITED BY
Ruslan Kalendar,
University of Helsinki, Finland

REVIEWED BY
Vratislav Peska,
Academy of Sciences of the Czech Republic, Czechia
André Luís Laforga Vanzela,
State University of Londrina, Brazil

*CORRESPONDENCE
Guo-Jun Zhang
zgj003@126.com
Shu-Fen Li
lishufen83@163.com

RECEIVED 28 May 2023
ACCEPTED 04 October 2023
PUBLISHED 16 October 2023

CITATION
Zhang G-J, Jia K-L, Wang J, Gao W-J and Li S-F (2023) Genome-wide analysis of transposable elements and satellite DNA in *Humulus scandens*, a dioecious plant with XX/XY$_1$Y$_2$ chromosomes.
Front. Plant Sci. 14:1230250.
doi: 10.3389/fpls.2023.1230250

Genome-wide analysis of transposable elements and satellite DNA in *Humulus scandens*, a dioecious plant with XX/XY$_1$Y$_2$ chromosomes

Guo-Jun Zhang[1,2*], Ke-Li Jia[2,3], Jin Wang[2], Wu-Jun Gao[2] and Shu-Fen Li[2*]

[1]School of Basic Medical Sciences, Xinxiang Medical University, Xinxiang, China, [2]College of Life Sciences, Henan Normal University, Xinxiang, China, [3]SanQuan Medical College, Xinxiang Medical University, Xinxiang, China

Transposable elements (TEs) and satellite DNAs, two major categories of repetitive sequences, are expected to accumulate in non-recombining genome regions, including sex-linked regions, and contribute to sex chromosome evolution. The dioecious plant, *Humulus scandens*, can be used for studying the evolution of the XX/XY$_1$Y$_2$ sex chromosomes. In this study, we thoroughly examined the repetitive components of male and female *H. scandens* using next-generation sequencing data followed by bioinformatics analysis and florescence *in situ* hybridization (FISH). The *H. scandens* genome has a high overall repetitive sequence composition, 68.30% in the female and 66.78% in the male genome, with abundant long terminal repeat (LTR) retrotransposons (RTs), including more Ty3/*Gypsy* than Ty1/*Copia* elements, particularly two Ty3/*Gypsy* lineages, Tekay and Retand. Most LTR-RT lineages were found dispersed across the chromosomes, though CRM and Athila elements were predominately found within the centromeres and the pericentromeric regions. The Athila elements also showed clearly higher FISH signal intensities in the Y$_1$ and Y$_2$ chromosomes than in the X or autosomes. Three novel satellite DNAs were specifically distributed in the centromeric and/or telomeric regions, with markedly different distributions on the X, Y$_1$, and Y$_2$ chromosomes. Combined with FISH using satellite DNAs to stain chromosomes during meiotic diakinesis, we determined the synapsis pattern and distinguish pseudoautosomal regions (PARs). The results indicate that the XY$_1$Y$_2$ sex chromosomes of *H. scandens* might have originated from a centric fission event. This study improves our understanding of the repetitive sequence organization of *H. scandens* genome and provides a basis for further analysis of their chromosome evolution process.

KEYWORDS

***Humulus scandens*, repetitive sequences, satellites, sex chromosome evolution, transposable elements (TEs)**

Introduction

Plant genomes typically consist of a large number of various repetitive DNA sequences. For example, they represent approximately 85% and more than 90% of the maize and onion genomes, respectively (Schnable et al., 2009; Fu et al., 2019). According to their structural arrangement and sequence composition, repetitive DNAs can be divided into two major groups: tandem repeats and transposable elements (TEs) (reviewed in Biscotti et al., 2015). Tandem repeats are arrays of non-coding sequences arranged in tandem. According to the monomer length, tandem repeats are usually classified into microsatellites (2–7 bp), minisatellites (tens of bp), and satellites (hundreds of bp). The satellite DNAs predominately cluster at specific positions on the chromosomes, such as (peri)centromeres, (sub)telomeres, and other heterochromatic regions, making them ideal markers for cytogenetic analysis (He et al., 2015; Kirov et al., 2017; Lang et al., 2019; Li et al., 2019). TEs are elements that have the unique ability to mobilize from one position of a genome to another. Based on whether using RNA as a transposition intermediate, TEs are divided into two major classes: retrotransposons transposing via copy-and-paste mechanism and DNA transposons using cut-and-paste transposition mode. Due to the transposition mechanism, retrotransposons are the most prevalent TEs in plant genomes. Of these, the long terminal repeat (LTR) retrotransposons, primarily Ty1/*Copia* and Ty3/*Gypsy*, were shown to be the most frequent in the plant genomes (Neumann et al., 2019).

According to popular belief, it is impossible to understand how eukaryotes' complex genomes are shaped by evolutionary mechanisms without extensive examination of repetitive genomic sequences (reviewed in Shapiro and von Sternberg, 2005; Slotkin and Martienssen, 2007). Angiosperm plants are particularly prone to this due to the large proportion of repetitive sequences in the genome and the significant contribution of them to the extraordinary variation in genome sizes among different taxa (Piegu et al., 2006; Pellicer et al., 2021; Sader et al., 2021). Even though most repetitive elements tend to be selfish, growing evidence presents that repetitive sequences play a variety of roles, such as influencing the organization and stability of the genome (Bennetzen and Wang, 2014), mediating chromosomal rearrangement (reviewed in George and Alani, 2012), chromatin modulation (reviewed in Ohtani and Iwasaki, 2021), modification of gene expression (Wyler et al., 2020), and shaping phenotypic variation (Cai et al., 2022).

The majority of flowering plants are hermaphrodites, e.g., individual plants contain bisexual flowers harboring both pistil and stamen, and with just around 6% being dioecious, that is, plants with unisexual flowers on different individuals (reviewed in Renner and Ricklefs, 1995). These dioecious plants have multiple sex-determining mechanisms, including the XY and the ZW sex chromosome systems, as well as the sex index, i.e., the ratio of X chromosomes to autosome sets (X:A) (Baránková et al., 2020). In either case, the sex chromosomes are thought to be evolved from a pair of autosomes. Several events occurred during the sex chromosome evolutionary process, such as the formation of sex-determining gene(s), recombination inhibition, TE amplification, and Y chromosome degeneration (reviewed in Charlesworth, 2015). Among them, the accumulation of repetitive sequences is a dominant feature of the non-recombining region of the sex chromosome. This phenomenon has been observed in a variety of dioecious plants, for instance in *Carica papaya* (Wang et al., 2012; VanBuren and Ming, 2013), *Rumex acetosa* (Jesionek et al., 2021), and *Spinacia oleracea* (Li et al., 2019). These repetitive sequences have been postulated to be important driving forces for the evolution of sex chromosomes (Yang et al., 2021).

Humulus scandens is a climbing dioecious herbaceous plant belonging to the Cannabaceae family. The chromosome numbers of males and females are different; females have 2n = 16 = 14 + XX, whereas males have 2n = 17 = 14 + XY_1Y_2. The gender of *H. scandens* is governed by the X:A ratio; with an X:A ratio of 1.0 for females, and a ratio of 0.5 for males (Shephard et al., 2000). The multiple sex chromosome system (XX/XY_1Y_2) and the X:A ratio sex determination system make *H. scandens* an ideal species for examining sex chromosome evolution and sex determination mechanism. The two Y chromosomes are approximately the same size as X chromosome (Grabowska-Joachimiak et al., 2006; Alexandrov et al., 2012). GISH painting of the Y chromosomes showed that the male-specific region of the Y chromosome (MSY) spans the bulk of the Y chromosomes, suggesting advanced phases of sex chromosome evolution in *H. scandens*, as well as its relatives in the same family, *H. lupulus* and *Cannabis sativa* (Razumova et al., 2023). To date, limited studies have reported the repetitive elements of *H. scandens* (Alexandrov et al., 2012). The abundance and distribution of repetitive sequences, as well as role of repetitive sequences in the sex chromosome evolution of *H. scandens* has not been thoroughly examined. To gain a better understanding of the genome structure of *H. scandens*, we extensively analyzed the repetitive components of male and female *H. scandens* by combining next-generation sequencing, bioinformatics analysis, and florescence *in situ* hybridization (FISH). We first used a graph-based clustering strategy to identify and annotate repetitive sequences based on next-generation sequencing data. Then we examined the distribution patterns of different groups of TEs and satellite DNAs using FISH analysis. Our research offers a crucial foundation for comprehending the repeat elements in *H. scandens*.

Materials and methods

Plant material and DNA extraction

Plants of *H. scandens* were cultivated in a garden field at Henan Normal University. The genders were determined by observation of the morphology of flowers. Whole genomic DNA was extracted from three male and three female individuals using the cetyl trimethylammonium bromide method (Doyle and Doyle, 1987).

DNA library preparation, high-throughput sequencing and repetitive DNA identification

At least 3 μg genomic DNA of each sample was fragmented, end repaired, phosphorylated, and ligated with adapters. Then 300–400 bp fragments were selected and PCR amplified to construct paired-end library. The libraries were sequenced on Illumina NovaSeq 6000 platform in paired-end, 150-bp mode. The raw reads (DRR024400, DRR024402, DRR024404, DRR024405, DRR024456, and DRR024457) of *H. lupulus*, a close relative of *H. scandens*, were downloaded from the NCBI SRA database. For the preprocessing of raw sequence data, the FASTQ files were quality-controlled, filtered using HTQC with default parameters (v1.92.1) (Yang et al., 2013), then a custom perl script was used to filter out the reads with N and transform into FASTA format. Then, for each sample, a randomly chosen dataset comprising 2,000,000 paired-end reads, which represented approximately 0.34× of the genome, was used for further analysis. We clustered, assembled, and annotated all selected reads with the help of RepeatExplorer platform (http://www.repeatexplorer.org, Novák et al., 2020) using the Green Plants (Viridiplantae) database (Neumann et al., 2019). To categorize LTR retrotransposons into different lineages, their RT sequences were analyzed using DANTE within the RepeatExplorer platform. After eliminating duplicated sequences using CD-hit (Li and Godzik, 2006), these RT sequences were aligned with MUSCLE (Edgar, 2004), and phylogenetic trees were generated using FastTree (Price et al., 2010). FigTree software was used to draw and modify the trees.

Satellite DNA identification

The TAREAN tool embedded in RepeatExplorer was adopted to detect satellite DNA using the same samples described above (Novák et al., 2017). The high-confidence satellites were identified. The logo for the satellites was drawn by Web-Logo (Crooks et al., 2004).

FISH probe design

For LTR retrotranposons, the RT domains of each lineage were amplified using specific primers (Table S1). The gel electrophoresis, cloning, and sequence validation were performed following a previous study (Li et al., 2019). Finally, clones with high sequence similarity to the respective contigs were amplified and labeled with Texas-red-dCTP (PerkinElmer, Waltham, Massachusetts, USA) utilizing the nick translation approach. For satellite DNA, the monomers were identified, and 50 bp of the monomer was randomly chosen and tagged directly with Texas Red-X (Invitrogen, Shanghai, China) in the synthesis process. 45S rDNA was also labeled with Chroma Tide Alexa Fluor 488-5-dUTP (Invitrogen) to aid in chromosome identification.

Mitotic and meiotic chromosome spreads preparation

For mitotic chromosome spread preparation, the branches of both male and female plants were cut and placed in water until the new roots were grown. When the roots were grown to 1–1.5 cm, they were cut and treated in nitrous oxide at 10.9 atm pressure, fixed in a 90% acetic acid solution for 10 min. For meiotic chromosome spread preparation, male inflorescences were fixed in Carnoy's fixative and stored in 70% ethanol. The immature flowers with a diameter of 1.3–1.5 mm were selected, and the anthers were picked out for further analysis. The enzyme digestion and drop spreading were completed according to earlier instructions (Li et al., 2019).

FISH assays using LTR-RT domain and satellite probes

FISH analysis was carried out as previously described (Li et al., 2019). First, the slides with well-spread chromosomes or desired meiotic stages were UV cross-linked. Then the hybridization mixture (2×SSC, 1×TE, 200 ng labeled probe) was added to the slides, followed by denaturation in boiled water for 5 min. After that, the denatured chromosome slides with probes were incubated overnight at 37°C. Next, the slides were washed in 2×SSC for 5 min, and counterstained with DAPI solution. FISH signals were detected under an Olympus BX 63 fluorescence microscope with an ANDOR CCD.

Statistical analysis

Pairwise comparisons were used to assess how the data groups differed from one another using Excel software. Least significance difference (LSD) analysis was performed on the data with a significance threshold of 0.05.

Results

Repeat proportion in *H. scandens* genome

Next-generation sequencing produced a total of approximately 20 Gb of raw sequencing data for three male and three female *H. scandens*. For each individual, a subset of 2,000,000 clean paired reads, representing about 34% of the genome, was chosen for repetitive sequence analysis. We also used sequencing reads of the *H. lupulus* genome for comparison analysis. Table 1 shows the comparative proportions of different repeat types in the three genomes. The findings revealed a high composition of repetitive sequences in the *H. scandens* genome, with 68.30% in the female genome and 66.78% in the male genome. These values were significantly higher than those of the *H. lupulus* genome (59.45%) (Figure 1A). Similar to most other plant species, LTR-retrotransposons were the most prevalent group of *H. scandens* repeats, accounting for nearly 60% of the genome (when male and female values were averaged), followed by tandem repeats (3.06%), DNA transposons (1.16%), and long interspersed nuclear elements (LINEs) (<0.01%). Within the LTR retrotransposon superfamily, the proportion of Ty3/*Gypsy* elements was more than 30-fold greater than Ty1/*Copia* elements (Table 1; Figure 1B). In addition, about 3.35% of organelle DNAs were identified in the *H. scandens* genome.

For the comparison of different groups of repetitive sequences, the proportions of LTR-retrotransposons and tandem repeats of the

TABLE 1 Repeat proportions (%) estimated in the *Humulus scandens* and *Humulus lupulus* genomes.

Repeats		Lineage/ class	Hs-female				Hs-male				Hl			
			Hs-F1	Hs-F2	Hs-F3	Ave.	Hs-M1	Hs-M2	Hs-M3	Ave.	Hl-1	Hl-2	Hl-3	Ave.
LTR retroelements	Ty1/*Copia*	Ale	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.05	0.04	0.03	0.04
		Angela	0.27	0.13	0.25	0.22	0.28	0.20	0.25	0.24	3.27	3.25	3.81	3.44
		Ikeros	0.11	0.08	0.10	0.10	0.05	0.08	0.06	0.06	0.22	0.14	0.23	0.20
		SIRE	1.56	1.38	1.47	1.47	1.24	1.31	1.16	1.24	0.58	0.61	0.65	0.61
		TAR	0.10	0.09	0.12	0.10	0.09	0.12	0.08	0.10	1.99	2.00	1.87	1.95
		Tork	0.08	0.02	0.05	0.05	0.01	0.03	0.03	0.02	0.00	0.00	0.00	0.00
		Subtotal	**2.12**	**1.71**	**1.99**	**1.94**	**1.66**	**1.73**	**1.58**	**1.66**	**6.11**	**6.04**	**6.59**	**6.25**
	Ty3/*Gypsy*	Athila	8.24	8.04	7.84	8.04	8.66	8.94	8.61	8.74	0.79	0.80	0.88	0.82
		Ogre/Tat	0.02	0.05	0.02	0.03	0.03	0.01	0.03	0.02	0.00	0.00	0.00	0.00
		Retand	21.98	25.11	22.24	23.11	19.11	19.42	20.35	19.63	14.84	15.76	14.97	15.19
		CRM	0.16	0.14	0.16	0.15	0.16	0.17	0.16	0.16	0.35	0.36	0.35	0.35
		Galadriel	0.07	0.10	0.09	0.06	0.09	0.10	0.11	0.10	0.18	0.19	0.19	0.19
		Tekay	23.01	25.20	23.61	23.94	25.26	26.12	25.96	25.78	21.39	21.10	21.19	21.23
		Subtotal	**53.47**	**58.62**	**53.94**	**55.34**	**53.31**	**54.75**	**55.21**	**54.42**	**37.55**	**38.21**	**37.58**	**37.78**
	Unclassified LTR RTs		2.38	2.95	2.55	2.63	3.14	3.10	3.13	3.12	3.79	1.93	3.90	3.21
Other	LINE		0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	DNA transposons	EnSpm_CACTA	0.45	0.40	0.38	0.41	0.39	0.42	0.37	0.39	1.83	1.32	1.60	1.58
		hAT	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.05	0.04	0.04	0.04
		MuDR_Mutator	1.22	0.89	0.73	0.95	0.35	0.63	0.57	0.52	0.22	0.27	0.20	0.23
		Subtotal	**1.69**	**1.31**	**1.21**	**1.40**	**0.75**	**1.06**	**0.95**	**0.92**	**2.10**	**1.63**	**1.84**	**1.86**
	Tandem repeats	rDNA	1.06	1.53	1.34	1.31	0.90	0.93	0.93	0.92	0.16	0.15	0.15	0.15
		Satellite	2.26	1.77	2.04	2.02	2.05	1.71	1.85	1.87	0.85	0.80	0.66	0.77
		Subtotal	**3.32**	**3.30**	**3.38**	**3.33**	**2.95**	**2.64**	**2.78**	**2.79**	**1.01**	**0.95**	**0.81**	**0.92**
Total annotated repeats			**62.99**	**67.90**	**63.00**	**64.63**	**61.80**	**63.27**	**63.65**	**62.91**	**50.56**	**48.76**	**50.72**	**50.01**
Unclassified repeat			3.78	3.37	3.86	3.67	4.29	4.11	3.20	3.87	9.12	10.28	8.92	9.44
Total			**66.77**	**71.26**	**66.86**	**68.30**	**66.10**	**67.38**	**66.85**	**66.78**	**59.68**	**59.04**	**59.64**	**59.45**

Hs-female and Hs-male represent the female and male genomes of *H. scandens*, respectively. Hl represents the genome of *H. lupulus*. The bold values indicate the total value of "Ty1-*Copia*", "Ty3-*Gypsy*", "DNA transposons", "Tandem repeats", "Annotated repetitive sequences", and all the repetitive sequences identified, respectively. "Ave." means "the average value".

H. scandens genome were all significantly higher than those of the *H. lupulus* genome (Table 1; Figure 1B). However, the DNA transposons showed a little higher proportions in the *H. lupulus* genome than in the *H. scandens* genome. Out of the LTR retrotransposons, the proportions of Ty3/*Gypsy* elements in the *H. scandens* genome were significantly higher than those in *H. lupulus*. By contrast, the *H. lupulus* genome showed higher proportions of Ty1/*Copia* elements than the *H. scandens* genome (Table 1; Figure 1B). Thus, the ratio of Ty3/*Gypsy* to Ty1/*Copia* in the *H. lupulus* genome was 6:1, compared with 30:1 in the *H. scandens* genome (Table 1). We also compared the proportions of various lineages of LTR retrotransposons (Figure 2). In the *H. scandens* genome, Tekay and Retand lineages belonging to the Ty3/*Gypsy* superfamily were prevalent, together accounting for nearly 80% of the LTR retrotransposable elements. The Athila lineage belonging to the Ty1/*Copia* superfamily ranked third (Figures 2A,

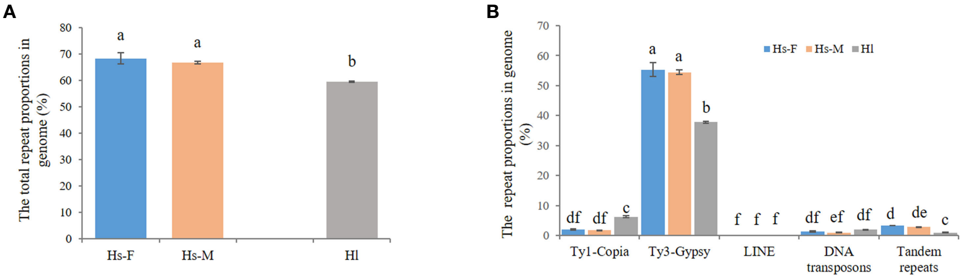


FIGURE 1
Comparison of repetitive sequences between the *H. scandens* and *H. lupulus* genomes. **(A)** The total repeat proportions in the *H. scandens* and *H. lupulus* genomes. Hs-F, female genome of *H. scandens*; Hs-M, male genome of *H. scandens*; Hl, *H. lupulus* genome. **(B)** The repeat proportion in the *H. scandens* and *H. lupulus* genomes. "a, b, c, d, e, f" means significance level by multiple comparison, $p< 0.05$.

B). In the *H. lupulus* genome, the Tekay and Retand lineages were also the more abundant elements, similar to the *H. scandens* genome; however, the third-ranked lineage was Angela (Figure 2C).

Comparison of transposable elements between male and female *H. scandens* genomes

Generally, a conserved repeat proportion pattern was observed between the male and female *H. scandens* genomes. However, slight but significant differences were still observed in the proportions of several repeat groups between the two genomes with different genders, such as SIRE and Retand lineages, which all showed a higher proportion in female than in male genomes. By contrast, two other lineages, including Athila and Tekay, showed a higher proportion of males than females (Figure 3).

Phylogenetic analysis of the lineages of LTR-RTs in *H. scandens* genome

Two phylogenetic trees based on the RT sequences of the Ty1/*Copia* and Ty3/*Gypsy* LTR-RTs were created to ascertain their evolutionary relationships. As shown in the evolutionary dendrograms (Figure 4), the Ty1/*Copia* elements could be classified into three clades: one included elements belonging to Angela and Ikeros lineages, another consisted of TAR and Tork lineages, whereas SIRE, Ale, Alesia, and Ivana lineages grouped together to form the third clade. Among the Ty3/*Gypsy* elements, there were also three clades: Athila and Retand/Ogre each formed one clade, while the other three lineages, including CRM, Tekay, and Galadriel, formed the third clade. The Retand/Ogre lineage showed a high degree of variability and could be further segregated into two distinct groups.

Chromosome distributions patterns of LTR-retrotransposons

We examined the chromosomal distribution patterns of each lineage of LTR-RTs by using mitotic-FISH analysis. The findings demonstrated that different lineages has varied patterns of chromosomal distribution. Most of the lineage-based elements were dispersed over all of the chromosomes; these included all eight lineages of the *Copia* superfamily and four lineages of the *Gypsy* superfamily (Ogre, Retand, Galadriel, and Tekay) (Figures 5, 6). The other two lineages (Athila and CRM) mainly occupied the

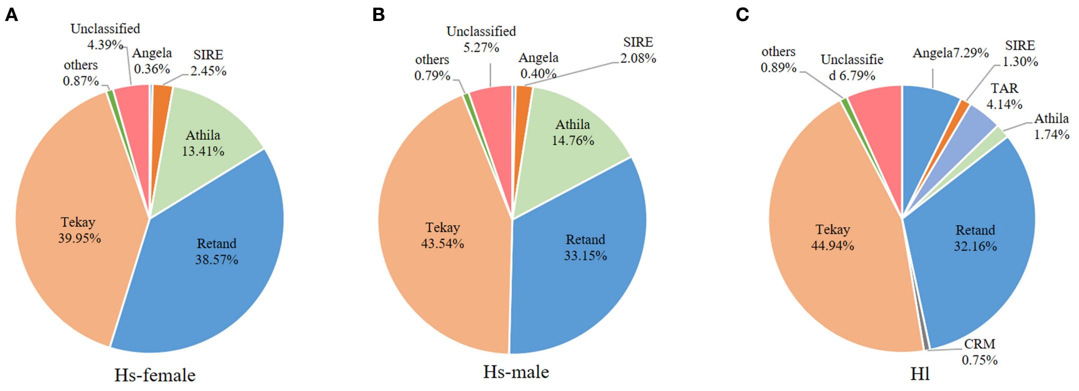


FIGURE 2
The proportions of different LTR-RT lineages in the *H. scandens* female **(A)** and male **(B)** genomes as well as the *H. lupulus* genome **(C)**.

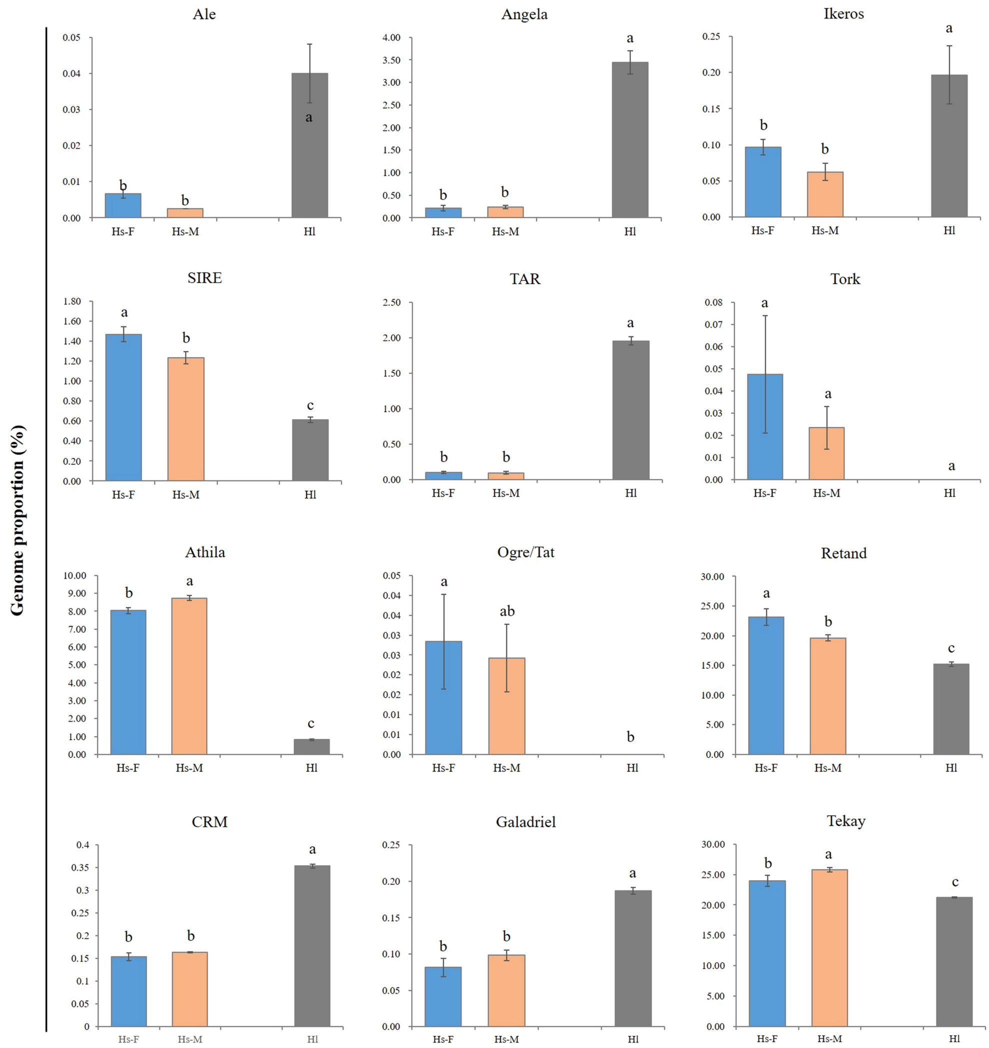


FIGURE 3
Comparison of the major LTR-RT lineages between male and female *H. scandens* as well as *H. lupulus* genomes. "a and b" mean significance level by multiple comparison, $p< 0.05$.

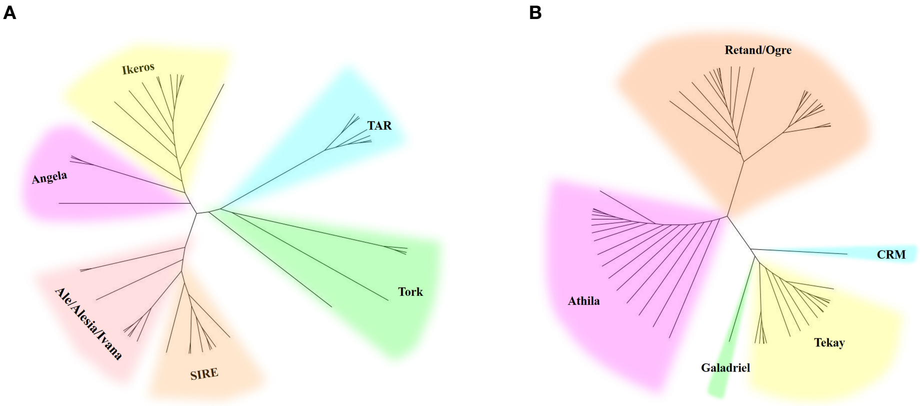


FIGURE 4
Phylogenetic analysis of Ty1/*Copia* **(A)** and Ty3/*Gypsy* **(B)** LTR-RT lineages represented by the LTR-RT domain protein sequences.

pericentromeric regions of all chromosomes. In addition, the Athila elements showed clearly higher signal intensity in the Y_1 and Y_2 chromosomes than in the X and autosomes.

Identification of satellites and analysis of chromosomal locations

By using TAREAN, three clusters (CL33, CL37, and CL111) were identified as highly reliable satellite DNAs. Each of them was given the designations Hssat1, Hssat2, and Hssat3, respectively. The three clusters were all characterized by star-like graph topologies (Figure S1). The monomer of Hssat1 was about 312 bp. A search of GenBank yielded no matches to other known sequences. The other two satellites, Hssat2 and Hssat3, demonstrated ~380 bp and ~122 bp, respectively. Like Hssat1, the clusters were all unknown or *H. scandens*-specific sequences.

The results showed that the signal of Hssat1 is distributed at the centromere of all chromosomes except for Y_1 and Y_2 chromosomes. Hssat2 was distributed specifically at the terminal position on metaphase chromosomes, but there is a large variation between male and female chromosomes. In the female plant, there were 4 pairs of chromosomes with signals at both ends, while in the male plant, there were 5 pairs. Even the chromosomes with signals at one end are inconsistent; for example, there were significant differences in the signals at the ends of chromosomes 6 and 7. In addition, it is worth noting that the signals of the two Y chromosomes were obviously inconsistent, one of which had a signal at one end while the other Y chromosome had no hybridization signal. The signal of Hssat3 showed more complex distribution characteristics, both at the end position and at the centromere position. In female *H. scandens*, the end positions of all chromosomes except one pair of autosomes could be seen, and obvious signals could be seen at the centromere positions of this

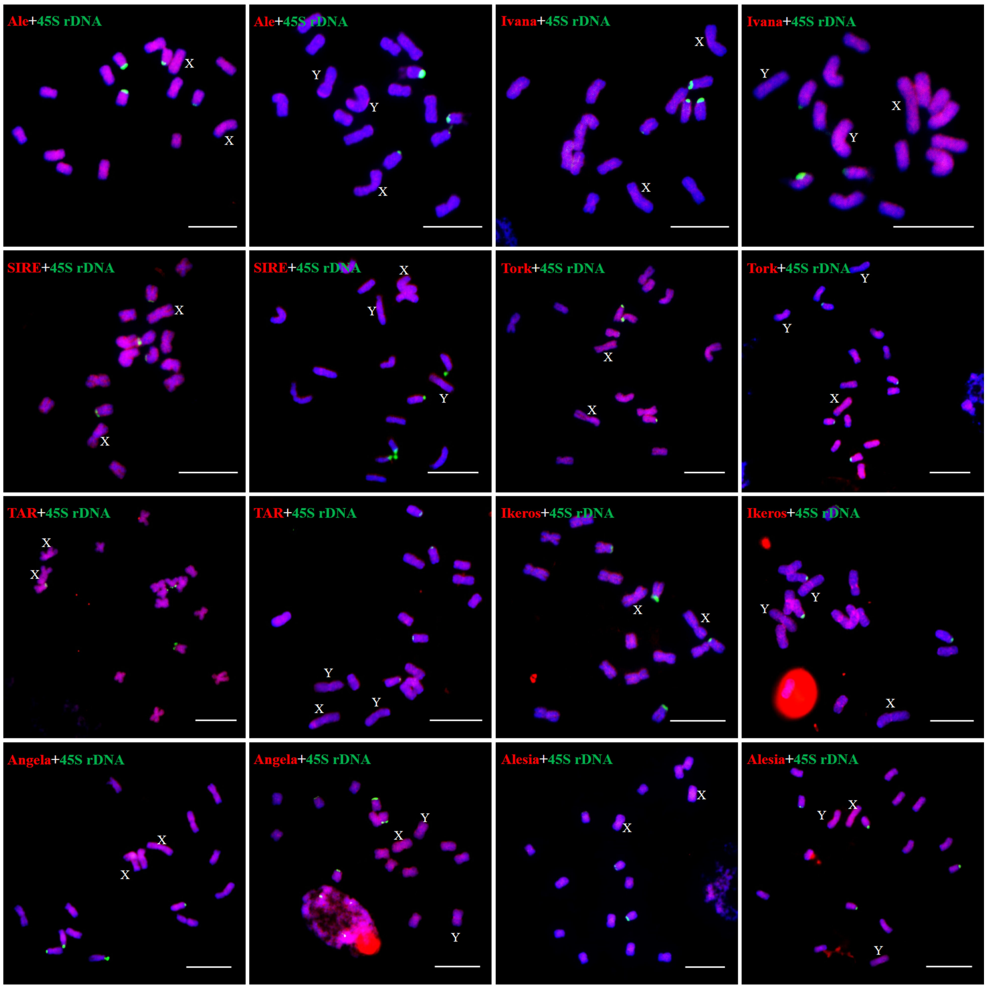


FIGURE 5
Distribution patterns of different LTR-RT lineages of Ty1/*Copia* elements on metaphase chromosomes of *H. scandens*. The lineage names and genders of individuals are presented inside each figure. The RT sequences of each lineage were labeled with Texas red (red), 45S rDNA was labeled with Chroma Tide Alexa Fluor 488 (green), and the chromosomes were counterstained with DAPI (blue). Arrows indicate the sex chromosomes. (Bars = 10 µm).

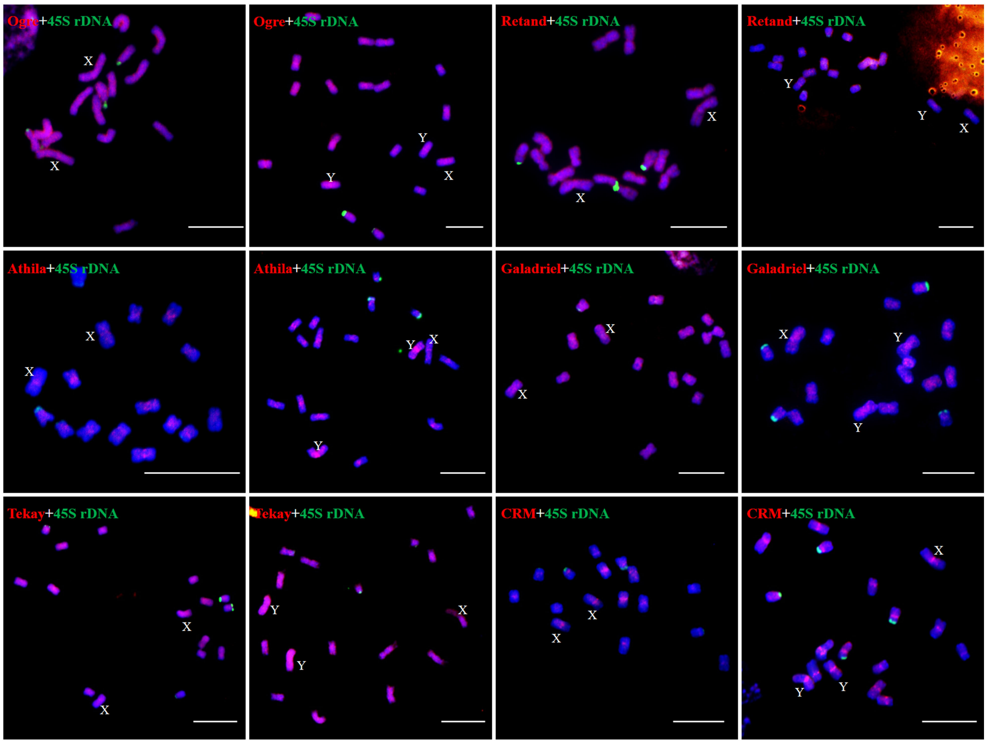


FIGURE 6
Distribution patterns of different LTR-RT lineages of Ty3/*Gypsy* elements on metaphase chromosomes of *H. scandens*. (Bars = 10 μm).

pair of autosomes. There was no signal distribution on one Y chromosome of male *H. scandens*, and the signal of the other Y chromosome was also distributed at the end of one side, like the X chromosome (Figures 7, 8).

Discussion

TE annotation of *H. scandens* genome

Repetitive sequences represent approximately 68% of the *H. scandens* genome, and this value is slightly but significantly higher than that of its close relative, the *H. lupulus* genome. Generally, the amount of repetitive sequences, particularly TEs, is positively correlated with the genome size of plants (Li et al., 2017). *H. scandens* (1.8 Gb) and *H. lupulus* (2.5 Gb) possess relatively large genomes, and the TE fraction proportions are generally in line with the trend. However, the genome of *H. scandens* is smaller than that of *H. lupulus*, whereas the repetitive sequence fraction of *H. scandens* is higher than that of *H. lupulus*. This suggests that TEs were amplified more extensively in the *H. scandens* genome. Similar to other plant genomes, in the *H. scandens* genome, LTR-RTs were more abundant than DNA transposons, LINEs, and tandem repeats. The diversity and proliferation ability of LTR-RTs make them an important contributor to the structure, function, and evolution of the *H. scandens* genome. In particular, two Ty3/*Gypsy* lineages, Retand and Tekay, proliferated massively. This is similar to the phenomenon in many other plant genomes, in which one or several TE groups are significantly amplified. For example, differential lineage-specific proliferation of distinct families of transposable elements contributes greatly to genome size differences in *Gossypium* (Hawkins et al., 2006).

Repetitive sequences and sex chromosome evolution of *H. scandens*

According to the general model of sex chromosome evolution, X and Y sex chromosomes are gradually differentiated from a pair of autosomes due to the suppression of recombination around the sex-determining locus. The accumulation of repetitive sequences can facilitate the recombination suppression of sex-determining loci and adjacent regions between X and Y chromosomes, eventually forming the sex-specific region. Furthermore, the formation of sex-specific regions can further recruit more repetitive sequences. Thus, repetitive sequence accumulation was a conspicuous feature of the sex chromosomes in both plants and animals (reviewed in Li et al., 2016). For instance, these findings were observed in papaya (VanBuren and Ming, 2013; VanBuren et al., 2015), *Rumex acetosa* (Mariotti et al., 2009), *Salix viminalis* (Almeida et al., 2020), *Salix dunnii* (He et al., 2021), so on and so forth. The TEs and TE-derived repetitive sequences are thought to be involved in almost all of the major evolutionary phases of sex chromosome evolution (reviewed in Li et al., 2016). Furthermore, TEs and associated repetitive sequences may influence plant sex determination and differentiation. For example, in *Populus*

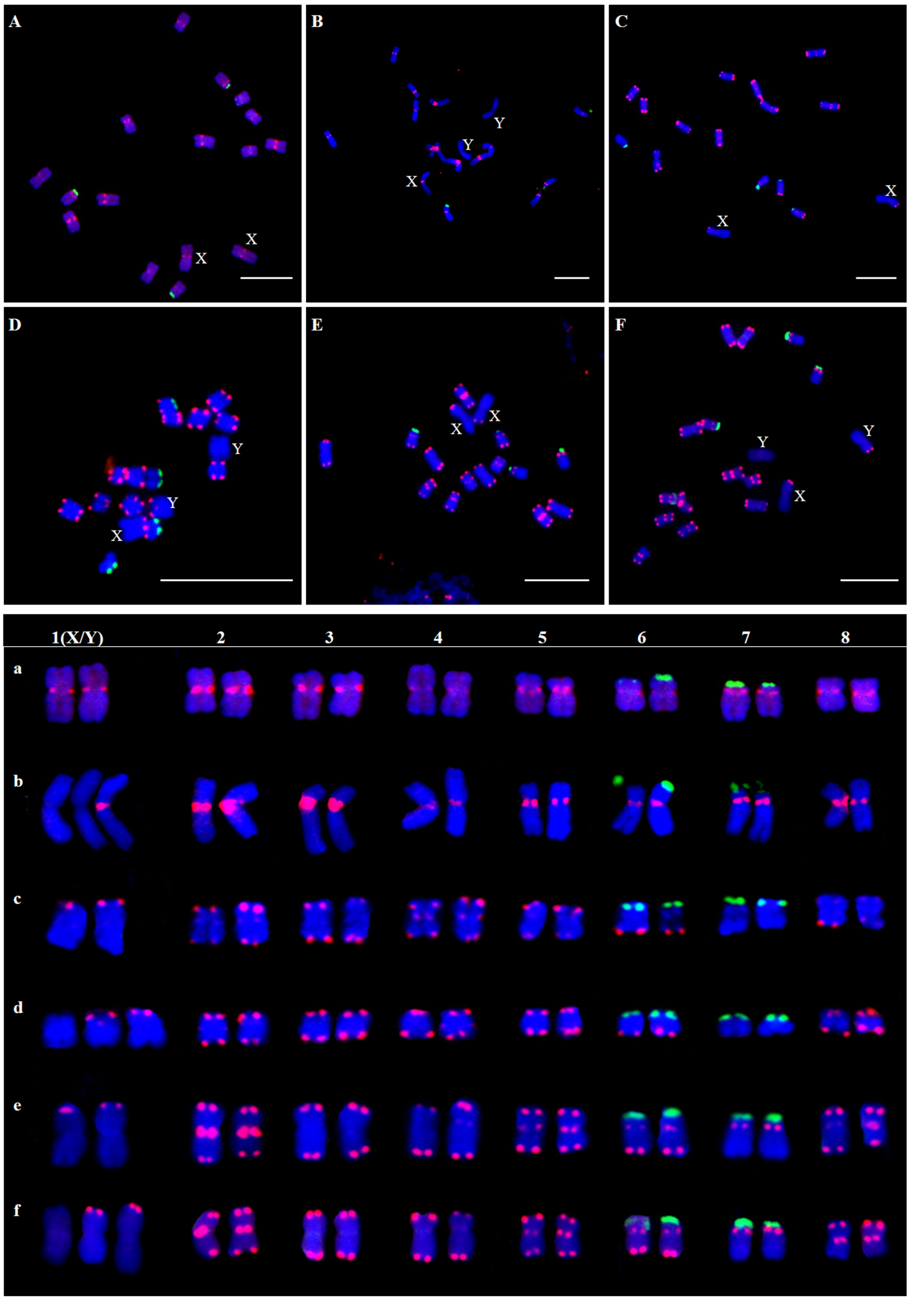


FIGURE 7
Localization of three satellites on metaphase chromosomes of female and male *H. scandens* using FISH analysis and karyotype of mitotic chromosomes based on FISH analysis of satellite DNAs in *H. scandens*. **(A)** Hssat1 on female chromosomes; **(B)** Hssat1 on male chromosomes; **(C)** Hssat2 on female chromosomes; **(D)** Hssat2 on male chromosomes; **(E)** Hssat3 on female chromosomes; **(F)** Hssat3 on male chromosomes. a-f show the karyograms of the metaphase chromosome spreads represented by the corresponding upper letters. (Bars = 10 μm).

deltoids, a Ty3/*Gypsy* transposable element family member within the Y-linked region can generate long non-coding RNAs and act as a male promoter (Xue et al., 2020). In *H. scandens*, FISH analysis showed that the Athila elements accumulated more intensively in the Y_1 and Y_2 chromosomes. The Y chromosome-biased TEs may be involved in sex chromosome evolution in *H. scandens*.

However, the satellites did not show such an accumulation pattern. The signals of the three satellites are absent from one or two Y chromosomes. Such a phenomenon was also observed in other dioecious plant species. For instance, one family of the Ogre/Tat lineage is found on all autosomes and the X chromosome but not on the Y chromosome in *Silene latifolia* (Kubat et al., 2014). Due to this, it can be challenging to comprehend the way repetitive sequences and the evolution of plant sex chromosomes are related. According to the few reports that are currently available, the accumulation or depletion of some repetitive sequences appears

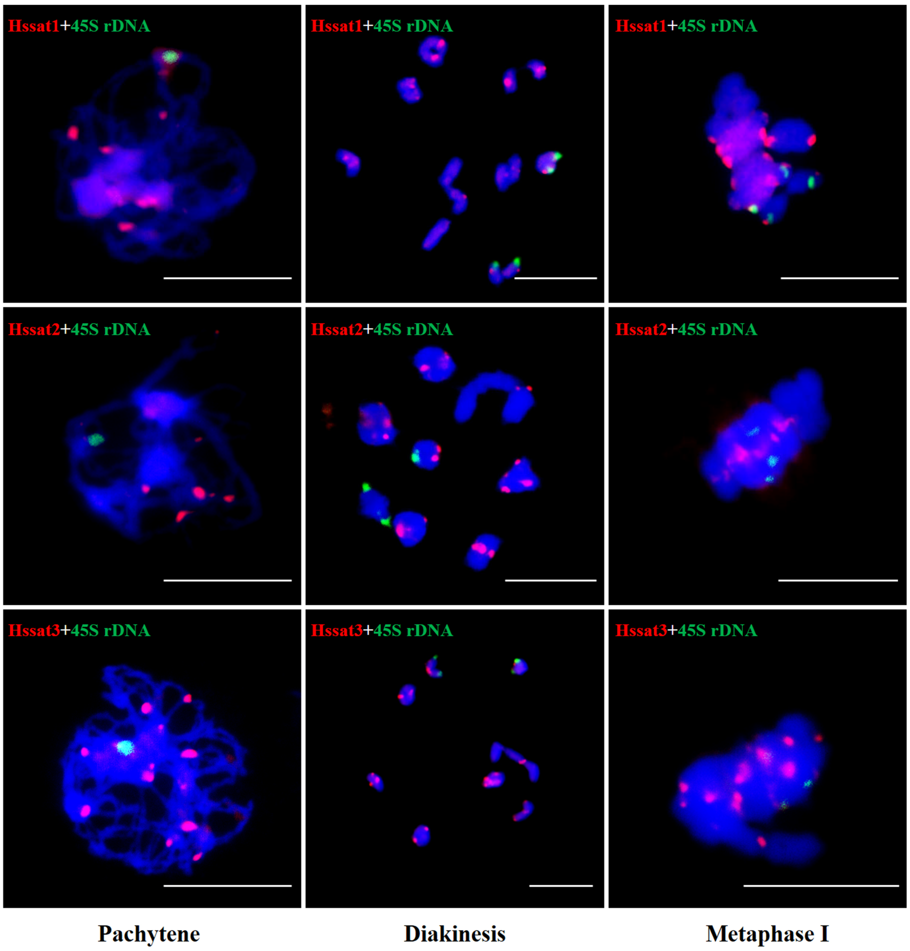


FIGURE 8
FISH analysis of three satellite DNAs on meiotic chromosomes in *H. scandens*. Three typical phases, including pachytene, diakinesis, and metaphase I are shown. (Bars = 10 µm).

to be species-specific, which is in keeping with the fact that sex chromosomes have evolved multiple times independently in different lineages of plants.

The satellite DNA probes showed typical signals located at the centromeric and/or telomeric regions. The X and Y chromosomes showed different signal distribution patterns, so the satellite DNA can be used as cytogenetic markers for identifying the X and Y chromosomes. Sex chromosome identification is crucial for cytological examinations and subsequent studies of sex chromosome evolution. Based on the cytogenetic markers, at the meiotic diakinesis stage, we observed obvious trivalents, showing Y_1-X-Y_2 connection mode. The synaptic region was the telomere position of two adjacent sex chromosomes connected end-to-end Figure 9. These results suggested that the sex-linked regions of *H. scandens* are large, which is in accordance with a recent study showing that the MSY covers the majority of the Y chromosomes (Razumova et al., 2023). These findings suggest advanced phases of sex chromosome evolution in *H. scandens.* In addition, the two Y chromosomes showed large differences because of the fact that the two Y chromosomes only pair at one telomeric end, and the majority of Y chromosomes are chromosome-specific regions. The FISH analysis using satellite probes also supports this perspective. Furthermore, combined with the fact that Hssat1 was specifically distributed on the centromeric regions of all the chromosomes except for the two Y chromosomes, we speculated that the XX-XY_1Y_2 sex chromosomes of *H. scandens* might have originated from a centric fission event. The centromere-specific satellite DNA might be lost during the centric fission event. However, we still need more evidence to support this speculation.

Conclusions

In conclusion, this work allowed us to have a comprehensive view of the repetitive fraction of the nuclear genome of *H. scandens*, which is an important dioecious plant with XX/XY_1Y_2 chromosomes. We annotated the repetitive portion of both the male and female *H. scandens* genomes based on the RepeatExplorer platform and extensively compared the different groups of repetitive sequences among the male and female genomes of *H. scandens*, as well as a close relative, *H. lupulus*. We also analyzed the distribution

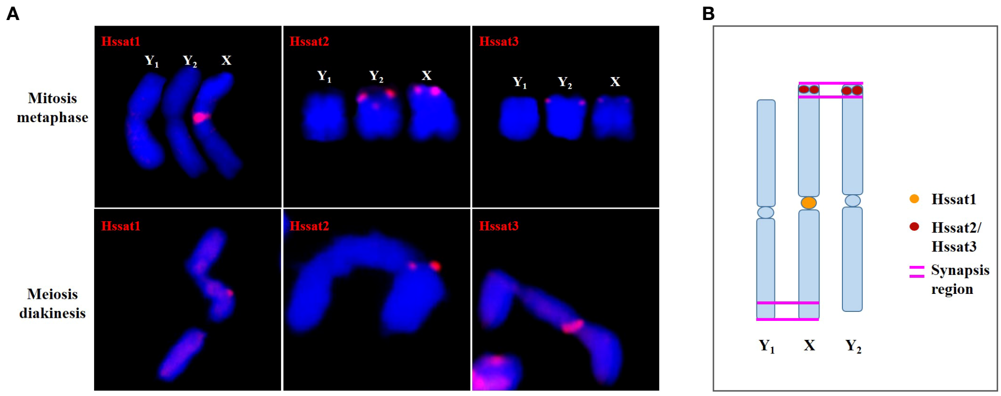


FIGURE 9
The synapsis pattern of the X, Y_1, and Y_2 sex chromosomes based on FISH signals of satellite DNAs. **(A)** The X, Y_1, and Y_2 sex chromosomes at mitosis metaphase and meiosis diakinesis stages with FISH signals of satellite DNAs. **(B)** Schematic diagram of synapsis pattern of the X, Y_1, and Y_2 chromosomes with FISH signals of satellite DNAs.

patterns of major LTR-RT lineages and three satellite DNAs using FISH analysis. Based on the FISH results of satellite DNAs, we were able to determine the orientation position of the PARs, and the results also indicated that the XX-XY_1Y_2 sex chromosomes of *H. scandens* might have originated from a centric fission event. Our findings shed light on the genome structure and evolution of *H. scandens* and laid a foundation for future research into the sex chromosome evolution of *H. scandens*.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI SRA database. BioProject accession number is PRJNA978042.

Author contributions

S-FL and W-JG designed the experiments. G-JZ, K-LJ and JW conducted the study and processed the data. G-JZ wrote the manuscript. All authors discussed the results and revised the manuscript. All authors have read and approved the final manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was financially supported by grants from the National Natural Science foundation of China (32170336 and 31970240), the Natural Science foundation of Henan province (222300420053), and the Program for Science & Technology Innovation Talents in the Universities of Henan Province (23HASTIT035).

Acknowledgments

We would like to thank all members who took part in this study.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1230250/full#supplementary-material

References

Alexandrov, O. S., Divashuk, M. G., Yakovin, N. A., and Karlov, G. I. (2012). Sex chromosome differentiation in *Humulus japonicus* Siebold & Zuccarini 1846 (Cannabaceae) revealed by fluorescence *in situ* hybridization of subtelomeric repeat. *Comp. Cytogen.* 47, 239–247. doi: 10.3897/CompCytogen.v6i3.3261

Almeida, P., Proux-Wera, E., Churcher, A., Soler, L., Dainat, J., Pucholt, P., et al. (2020). Genome assembly of the basket willow, *Salix viminalis*, reveals earliest stages of sex chromosome expansion. *BMC Biol.* 18, 78. doi: 10.1186/s12915-020-00808-1

Baránková, S., Pascual-Díza, J. P., Sultana, N., Alonso-Lifante, M. P., Balant, M., Barros, K., et al. (2020). Sex-chrom, a database on plant sex chromosomes. *New Phytol.* 227, 1594–1604. doi: 10.1111/nph.16635

Bennetzen, J. L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 65, 505–530. doi: 10.1146/annurev-arplant-050213-035811

Biscotti, M. A., Olmo, E., and Heslop-Harrison, J. S. (2015). Repetitive DNA in eukaryotic genomes. *Chromosome Res.* 23, 415–420. doi: 10.1007/s10577-015-9499-z

Cai, X., Lin, R., Liang, J., King, G. J., Wu, J., and Wang, X. (2022). Transposable element insertion: a hidden major source of domesticated phenotypic variation in *Brassica rapa*. *Plant Biotechnol. J.* 20, 1298–1310. doi: 10.1111/pbi.13807

Charlesworth, D. (2015). Plant contributes to our understanding of sex chromosome evolution. *New Phytol.* 208, 52–65. doi: 10.1111/nph.13497

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15. doi: 10.1016/0031-9422(80)85004-7

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.2460/ajvr.69.1.82

Fu, J., Zhang, H., Guo, F., Ma, L., Wu, J., Yue, M., et al. (2019). Identification and characterization of abundant repetitive sequences in *Allium cepa*. *Sci. Rep.* 9, 16756. doi: 10.1038/s41598-019-52995-9

George, C. M., and Alani, E. (2012). Multiple cellular mechanisms prevent chromosomal rearrangements involving repetitive DNA. *Crit. Rev. Biochem. Mol.* 47, 297–313. doi: 10.3109/10409238.2012.675644

Grabowska-Joachimiak, A., Śliwińska, E., Pigula, M., Skomra, U., and Joachimiak, A. J. (2006). Genome size in *Humulus lupulus* L. and *H. japonicus* Siebold & Zucc. (Cannabaceae). *Acta Soc Bot. Pol.* 75, 207–214. doi: 10.5586/ASBP.2006.024

Hawkins, J. S., Kim, H., Nason, J. D., Wing, R. A., and Wendel, J. F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.* 16, 1252–1261. doi: 10.1101/gr.5282906

He, Q., Cai, Z., Hu, T., Liu, H., Bao, C., Mao, W., et al. (2015). Repetitive sequence analysis and karyotyping reveals centromere-associated DNA sequences in radish (*Raphanus sativus* L.). *BMC Plant Biol.* 15, 105. doi: 10.1186/s12870-015-0480-y

He, L., Jia, K. H., Zhang, R. G., Wang, Y., Shi, T. L., Li, Z. C., et al. (2021). Chromosome-scale assembly of the genome of *Salix dunnii* reveals a male-heterogametic sex determination system on chromosome 7. *Mol. Ecol. Resour.* 21, 1966–1982. doi: 10.1111/1755-0998.13362

Jesionek, W., Bodláková, M., Kubát, Z., Čegan, R., Vyskot, B., and Vrána, J. (2021). Fundamentally different repetitive element composition of sex chromosomes in *Rumex acetosa*. *Ann. Bot.* 127, 33–47. doi: 10.1093/aob/mcaa160

Kirov, I. V., Kiseleva, A. V., Laere, K. V., Roy, N. V., and Khrustaleva, L. I. (2017). Tandem repeats of *Allium fistulosum* associated with major chromosomal landmarks. *Mol. Genet. Genomics* 292, 453–464. doi: 10.1007/s00438-016-1286-9

Kubat, Z., Zluvova, J., Vogel, I., Kovacova, V., Cermak, T., Cegan, R., et al. (2014). Possible mechanisms responsible for absence of a retrotransposon family on a plant Y chromosome. *New Phytol.* 202, 662–678. doi: 10.1111/nph.12669

Lang, T., Li, G., Wang, H., Yu, Z., Chen, Q., Yang, E., et al. (2019). Physical location of tandem repeats in the wheat genome and application for chromosome identification. *Planta* 249, 663–675. doi: 10.1007/s00425-018-3033-4

Li, W., and Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, S. F., Guo, Y. J., Li, J. R., Zhang, D. X., Wang, B. X., Li, N., et al. (2019). The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (*Spinacia oleracea* L.). *Mobile DNA* 10, 3. doi: 10.1186/s13100-019-0417-6

Li, S. F., Su, T., Cheng, G. Q., Wang, B. X., Li, X., Deng, C. L., et al. (2017). Chromosome evolution in connection with repetitive sequences and epigenetics in plants. *Genes* 8, 290. doi: 10.3390/genes8100290

Li, S. F., Zhang, G., Yuan, J. H., Deng, C. L., and Gao, W. ,. J. (2016). Repetitive sequences and epigenetic modification: inseparable partners play important roles in the evolution of plant sex chromosomes. *Planta* 243, 1083–1095. doi: 10.1007/s00425-016-2485-7

Mariotti, B., Manzano, S., Kejnovsky, E., Vyskot, B., and Jamilena, M. (2009). Accumulation of Y-specific satellite DNAs during the evolution of *Rumex acetosa* sex chromosomes. *Mol. Genet. Genomics* 281, 249–259. doi: 10.1007/s00438-008-0405-7

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

Novák, P., Neumann, P., and Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat. Protoc.* 15, 3745–3776. doi: 10.1038/s41596-020-0400-y

Novák, P., Robledillo, L.Á., Koblížková, A., Vrbová, I., Neumann, P., and Macas, J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* 45, e111. doi: 10.1093/nar/gkx257

Ohtani, H., and Iwasaki, Y. W. (2021). Rewiring of chromatin state and gene expression by transposable elements. *Dev. Growth Differ.* 63, 262–273. doi: 10.1111/dgd.12735

Pellicer, J., Fernández, P., Fay, M. F., Michálková, E., and Leitch, I. J. (2021). Genome size doubling arises from the differential repetitive DNA dynamics in the genus *Heloniopsis* (Melanthiaceae). *Front. Genet.* 12. doi: 10.3389/fgene.2021.726211

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16, 1262–1269. doi: 10.1101/gr.5290206

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS One* 5, e9490. doi: 10.1371/journal.pone.0009490

Razumova, O. V., Divashuk, M. G., Alexandrov, O. S., and Karlov, G. I. (2023). GISH painting of the Y chromosomes suggests advanced phases of sex chromosome evolution in three dioecious Cannabaceae species (*Humulus lupulus*, *H. japonicus*, and *Cannabis sativa*). *Protoplasma* 260, 249–256. doi: 10.1007/s00709-022-01774-x

Renner, S. S., and Ricklefs, R. E. (1995). Dioecy and its correlates in the flowering plants. *Am. J. Bot.* 82, 596–606. doi: 10.1002/j.1537-2197.1995.tb11504.x

Sader, M., Vaio, M., Cauz-Santos, L. A., Dornelas, M. C., Vieira, M. L. C., Melo, N., et al. (2021). Large vs small genomes in *Passiflora*: the influence of the mobilome and the satellitome. *Planta* 253, 86. doi: 10.1007/s00425-021-03598-0

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.11785

Shapiro, J. A., and von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biol. Rev.* 80, 227–250. doi: 10.1017/s1464793104006657

Shephard, H. L., Parker, J. S., Darby, P., and Ainsworth, C. C. (2000). Sexual development and sex chromosomes in hop. *New Phytol.* 148, 397–411. doi: 10.1046/j.1469-8137.2000.00771.x

Slotkin, R. K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 8, 272–285. doi: 10.1038/nrg2072

VanBuren, R., and Ming, R. (2013). Dynamic transposable element accumulation in the nascent sex chromosome of papaya. *Mob. Genet. Elements* 3, e23462. doi: 10.4161/mge.23462

VanBuren, R., Zeng, F., Chen, C., Zhang, J., Wai, C. M., Han, J., et al. (2015). Origin and domestication of papaya Y^h chromosome. *Genome Res.* 25, 524–533. doi: 10.1101/gr.183905.114

Wang, J. P., Na, J. K., Yu, Q. Y., Gschwend, A. R., Han, J., Zeng, F., et al. (2012). Sequencing papaya X and Y^h chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13710–13715. doi: 10.1073/pnas.1207833109

Wyler, M., Stritt, C., Walser, J. C., Baroux, C., and Roulin, A. C. (2020). Impact of transposable elements on methylation and gene expression across natural accessions of *Brachypodium distachyon*. *Genome Biol. Evol.* 12, 1994–2001. doi: 10.1093/gbe/evaa180

Xue, L., Wu, H., Chen, Y., Li, X., Hou, J., Lu, J., et al. (2020). Evidences for a role of two Y-specific genes in sex determination in *Populus deltoides*. *Nat. Commun.* 11, 5893. doi: 10.1038/s41467-020-19559-2

Yang, X., Liu, D., Wu, J., Zou, J., Xiao, X., Zhao, F., et al. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinf.* 14, 33. doi: 10.1186/1471-2105-14-33

Yang, W., Wang, D., Li, Y., Zhang, Z., Tong, S., Li, M., et al. (2021). A general model to explain repeated turnovers of sex determination in the Salicaceae. *Mol. Biol. Evol.* 38, 968–980. doi: 10.1093/molbev/msaa261

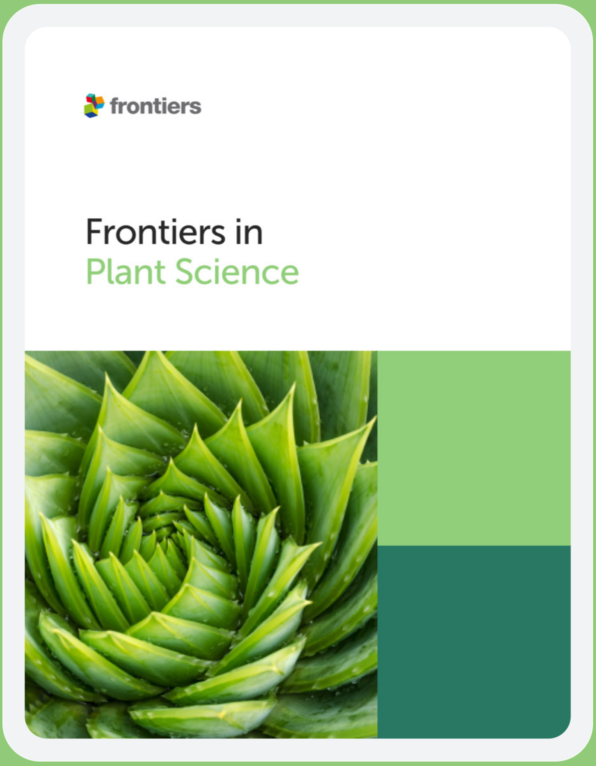
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact