School of Doctoral Studies in Biological Sciences
University of South Bohemia in České Budějovice
Faculty of Science

Analysis of repeat-rich regions of plant genomes using long read sequencing technologies

Ph.D. Thesis

MSc. Tihana Vondrak

Supervisor:
RNDr. Jiří Macas, Ph.D.
Institute of Plant Molecular Biology
Biology Centre
Czech Academy of Sciences

České Budějovice 2022

This thesis should be cited as:

Vondrak T. 2022. Analysis of repeat-rich regions of plant genomes using long read sequencing technologies. University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic, 156pp.

ANNOTATION

Satellite DNA has been identified in varying proportions in many eukaryotic genomes. It consists of monomeric units arranged in tandem into long, homogeneous arrays. Due to its repetitive nature, satellite DNA is difficult to assemble and analyze, and has therefore been largely neglected in research. With the development of next generation sequencing technologies such as Illumina, research on satellite DNA has intensified and our understanding of it has improved. However, the information we obtain with Illumina reads is limited by their short length. While we can characterize the sequence of satellite DNA and its abundance in genomes, it is not possible to study the long-range organization of satellite DNA , which limits our understanding of the origin and evolution of satellite DNA. This limitation can be addressed by using the latest generation of sequencing technologies that generate much longer reads of tens to hundreds of kilobases.

The goal of this work was to develop bioinformatics approaches for analyzing the properties of satellite DNA arrays from long sequence reads or from genome assemblies generated with these reads. These were then used to analyze populations of satellite repeats throughout the genome or in the specific type of chromatin in three plant species that differ in the organization of their centromeres.

Declaration

I hereby declare that I am the author of this dissertation and that I have used only those sources and literature detailed in the list of references.

České Budějovice, 30.6.2022

Vondrak
....
Tihana Vondrak

This thesis originated from a partnership of the Faculty of Science, University of South Bohemia and Institute of Plant Molecular Biology, Biology Centre of the Czech Academy of Sciences, supporting doctoral studies in Molecular and cell biology and genetics.

Přírodovědecká
fakulta
Faculty
of Science

Jihočeská univerzita
v Českých Budějovicích
University of South Bohemia
in České Budějovice

BIOLOGY
CENTRE
ASCR

Financial support

This work was supported by the Czech Science Foundation (grant no. 2024252S), the ELIXIR CZ Research Infrastructure Project (Czech Ministry of Education, Youth and Sports; Grant No. LM2015047 and LM2018131), including access to computing and storage facilities, and the ERDF/ESF project ELIXIR-CZ Capacity building (No. CZ.02.1.01/0.0/0.0/16_013/0001777)

Acknowledgements

I would primarily like to thank my supervisor, Jiří Macas for his patient guidance and support throughout my Ph.D. journey. Jirka, thank you for sharing your time and knowledge with me and giving me the opportunity to be part of such an amazing group of researchers.

Many thanks to Petr Novák and Pavel Neumann for all of their enormous help and useful advice. And of course I am so grateful to Laura Ávila Robledillo and Ludmila Oliveira, who helped me overcome so much during the last years. I wouldn’t have been able to do it without you.

A big thank you to all of the friends I have made during the Ph.D. and especially to Hynek, for making Budějovice feel like home.

I na kraju, najveća hvala mojoj mami, tati i bratu. Samnom ste proživljavali sve moje uspone i padove, bez vas ništa od ovog ne bi bilo moguće. Mojoj teti, koja mi je cijeli život bila uzor. Baki Kaji i Zlati, dedi Vasilju i Đuri, u mislima ste uvijek samnom. I Arčiju koji je uvijek bio terapija.

Volim vas najviše.

This thesis is based on the following papers:

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

(IF = 6.141; citations = 45)

J.M. conceived the study and drafted the manuscript. T.V. and P.No. developed the scripts for the bioinformatic analysis, and T.V., P.No., P.Ne. and J.M. analyzed the data. A.K. isolated the HMW genomic DNA and cloned the FISH probes. J.M. performed the nanopore sequencing. L.A.R. conducted the FISH experiments. All authors reviewed and approved the final manuscript. (TV 60%)

II. **Vondrak T**, Oliveira L, Novák P, Koblížková A, Neumann P, Macas J. 2021. Complex sequence organization of heterochromatin in the holocentric plant *Cuscuta europaea* elucidated by the computational analysis of nanopore reads. *Computational and Structural Biotechnology Journal*, 19*:*2179-2189. DOI:10.1016/j.csbj.2021.04.011

(IF = 7.271; citations = 4)

J.M. conceived the study and drafted the manuscript. T.V. and P.No. developed the scripts for the bioinformatic analysis, and T.V., P.No., P.Ne. and J.M. analyzed the data. A.K. isolated the HMW genomic DNA and cloned the FISH probes. J.M. performed the nanopore sequencing. L.O. conducted the FISH experiments. All authors reviewed and approved the final manuscript. (TV 70%)

III. Hofstatter GP, Thangavel G, Lux T, Neumann P, **Vondrak T**, Novak P, Zhang M, Costa L, Castellani M, Scott A, Toegelová H, Fuchs J, Mata-Sucre Y, Dias Y, Vanzela LLA, Hüttel B, Almeida SCC, Šimková H, Souza G, Pedrosa-Harand A, Macas J, Mayer XFK, Houben A, Marques A. 2022. All around centromeres: repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell*

(IF = 41.58, citations = 0)

A.M. conceived the initial project idea. A.P.H., J.M., K.F.X.M., and A.H. contributed to the project design. P.G.H. and A.M. performed k-mer analyses, genome assemblies, Omni-C scaffolding, and characterization of end-to-end fusions. P.G.H. and G.S. performed phylogenetic and dating analyses. G.T. performed all ChIP experiments. G.T., P.G.H. and A.M. performed ChIP and DNA methylation analyses. T.L. and K.F.X.M. performed gene annotations, synteny and whole-genome alignments. T.L. and A.S. performed OrthoFinder analyses. A.M., T.V., Pa.N., L.C., Pe.N., and J.M. performed repeat annotation and Tyba evolution and dynamics analyses. Pa.N. performed the TCR1 and TCR2 characterization. T.V. and Pe.N. performed analysis of emergence and loss of CENH3-binding domains. H.T. and H.S. performed Bionano optical mapping and initial assembly scaffolding. M.Z. performed statistical chromatin contact counting analyses. G.T., Y.M.S. and Y.M. performed FISH and immunostaining analysis. M.C. and J.F. performed flow cytometry. C.C.S.A. performed assembly and annotation of organelle genomes. A.L.L.V. and A.P.H. provided plant material. B.H. assembled all the sequencing libraries and performed most of the sequencing. P.G.H., G.T., and A.M. drafted the first version of the M.S. with input from the other authors. All authors approved the manuscript before submission. (TV 20%)

Co-authors agreement

The senior and corresponding authors of the manuscripts included in this thesis, hereby confirm that Tihana Vondrak contributed significantly to these publication, according to the statement above:

....

RNDr. Jiří Macas, Ph.D.

Max Planck Institute
for Plant Breeding Research
Dept. of Chromosome Biology
Evolution of Meiosis Group
Group leader: Dr. André Marques
Carl-von-Linné-Weg 10
50829 Cologne - Germany

....

André Marques, Ph.D.

Table of Contents

Introduction 1

1. Origin and evolution of satellite DNA 4

2. Satellite DNA and the centromere 6

3. Methods of satellite DNA research 7

References 10

Objective and aims of the thesis 19

Scope of the thesis 20

Chapter I 22

Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats 23

Supplementary information 40

Chapter II 51

Complex sequence organization of heterochromatin in the holocentric plant *Cuscuta europaea* elucidated by the computational analysis of nanopore reads 52

Supplementary information 63

Chapter III 68

All around centromeres: Repeat-based holocentromeres influence genome architecture and karyotype evolution 69

Supplementary information 128

Summary 142

Curriculum vitae 144

Introduction

Eukaryotic genomes vary immensely in size, and plant genomes are no different. The smallest haploid plant genomes of *Genlisea aurea* and *Genlisea tuberosa* range from 0.063 to 0.067 Gbp (Fleischmann et al. 2014) and the largest of *Paris japonica* is 150 Gbp (Pellicer et al. 2010). However, genome coding capacity does not correlate with genome size, which is known as the C-value paradox. Noncoding repetitive DNA is one of the main causes of the C-value paradox, as it is widely distributed in eukaryotic genomes and occurs in varying levels of abundance (Gregory 2005). Depending on its distribution and organization within the genome, repetitive DNA can be classified as either tandem repeats or dispersed repeats.

Tandem repeats consist of basic monomeric units that repeat head-to-tail in a tandem fashion. Three dimensions can be identified in their organization: the nucleotide sequence of the monomers, the complexity and size of the arrays, and their placement in the genome. Tandem repeats used to be classified based on the sizes of their monomers as either microsatellites, minisatellites, or satellites. Microsatellite monomers are usually thought to be between 1-10bp, while minisatellite monomers are between 10-100bp in length and satellite monomers are longer than 100bp. Microsatellite and minisatellite arrays are also shorter and simpler than satellite arrays and are mainly found in euchromatin. In contrast, large and complex satellite arrays, sometimes exceeding one megabase (Mb) in length (Lower et al. 2018) are mainly part of heterochromatin. However, with the discovery of large extended tandem arrays with very small monomers in heterochromatin, such as the tandem repeat families of *Drosophila* (Talbert et al. 2018) or the centromeric and pericentromeric repeats of lupin (Hufnagel et al. 2020), it became clear that classification of tandem repeats based on monomer size is inappropriate. Rather, they should be classified as satellites based on their organization in large extended arrays, regardless of monomer size.

Satellite DNA (SatDNA) appears to be a pervasive part of eukaryotic genomes, but it is present in various abundances that can change rapidly and without correlation to genome size, even between closely related species. For example, in the genus *Fritillaria* the FriSAT1 family amplified rapidly and differentially, accounting for between 0.1% and 36% of total nuclear DNA in different species (Ambrožová et al. 2011). The disparity between genome size and satDNA abundance can also be observed in the *Olea* genus, where *Olea europaea* subsp. *Cuspidata* has 50% of its entire genome made up of satDNA while the larger genome of *Olea paniculata* has only 1.94% (Mascagni et al. 2022).

Some nucleotide sequence features of satDNA monomers in plants appear to be conserved. Such as monomer size, which in satDNA generally range between 130-190bp and 310-370 bp in length (Macas et al. 2002) and correlates with the amount of DNA in a nucleosome or its multiple. Nevertheless, much shorter and longer monomers have been described, such as the simple sequence repeats of *Luzula elegans* between 4bp and 6bp (Heckmann et al. 2013) or the 4.7kb Sobo of *Solanum tuberosum (Tek et al. 2005)*. Moreover, the analysis of 152 monomers belonging to different satDNA families showed that their AT-content averaged at 58%, even though it ranged considerably from 22% to 75%. When analyzing the monomers for conserved nucleotide sequence characteristics they appeared to particularly overrepresent AA/TT dinucleotides as well as the CAAAA motif (Macas et al. 2002). This preference for certain nucleotides may be related to the formation of secondary and tertiary DNA structures, which are thought to facilitate nucleosome positioning (Melters et al. 2013) . Nonetheless nucleotide sequences of satDNA was found to change rapidly and apparently randomly, leading to the frequent occurrence of genus- or even species-specific satDNA families. Such an emergence and accumulation of a species-specific satellite was the reason for a 2.7-fold variation in DNA content between cultivated and wild diploid rice species (Uozu et al. 1997). Further studies confirmed these findings. For example, a study of 15 species of the genus *Heliophila* identified 108 satDNA families, of which only 16 were shared between two or more species (Dogan et al. 2021). The large-scale study of 23 species of the monophyletic legume tribe Fabeae identified only a few satDNA families to be shared by most species, wile the majority of 384 identified satDNA families was restricted to a single or a few closely related species (Macas et al. 2015). This lack of conservation in the amount of satDNA or its nucleotide sequence makes this genome fraction the most rapidly evolving.

Dispersed repeats, unlike tandem repeats, are scattered throughout the host genome. This dispersed organization is a consequence of their ability to move across and between genomes. Similar to satDNA, they come in various sizes ranging from 100bp-30kb (Arkhipova et al. 2019; Wells and Feschotte 2020), but they all rely on protein machinery necessary for transposition. These transposable elements (TEs) are classified into one of two classes depending on their mode of transposition. Class I or retrotransposons use a copy-and-paste mechanism with an RNA intermediate in which the RNA molecule is reverse transcribed into a cDNA and integrated into the genome, leaving the original element intact. Class II or DNA transposons use a cut-and-paste mechanism with a DNA intermediate in which the transposon itself is excised and moved to a new genomic locus (Finnegan 1989). Classes are further subdivided into subclasses based on the integration mechanism. Each subclass consists of elements that occur in a variety of host organisms and can be divided into families and subfamilies based on their phylogenetic relationships (Bourque et al. 2018). In addition to classification based on transposition mechanism, they can also be divided into autonomous and non-autonomous elements. Autonomous elements still have coding sequences that are required to establish the transposition machinery, whereas non-autonomous elements such as SINEs (Deragon and Zhang 2006) and MITEs (Feschotte et al. 2002) only have non-coding sequences and therefore rely on the autonomous equivalents for transposition. Non-autonomous elements usually arise from mutation or the complete elimination of their protein-coding sequences (Kejnovsky et al. 2012).

TEs can be a source of genomic variation in the host genome and were first discovered by McClintock in the 1940s, due to genetic variation the Activator/Dissociation (Ac/Ds) system causes in the maize genome, which shows up as brown or purple spots on colorless grains (McClintock 1951). This genomic variation associated with TE activity can be potentially beneficial or detrimental. On the one hand, stressful conditions are known to be associated with activation of TEs, and McClintock was the first to propose that their activity could provide the genomic variation that species need to survive these conditions (McClintock 1984). Although her theory was met with skepticism due to the possibility that TE activity was caused by the breakdown of regulatory machinery, later studies linked TE induced mutations to an increase in fitness under stress (Casacuberta and González 2013). On the other hand, TE insertions often have detrimental effects, either by inserting into coding sequences and generating mutant transcripts, inserting into regulatory sequences and disrupting gene expression, affecting epigenetic regulation of adjacent sequences, or indirectly by providing the similarity between different genomic loci required for ectopic recombination (Kumar 2020). Therefore, most TE copies are silenced by host silencing mechanisms (Kejnovsky et al. 2012).

Where retrotransposons are found in genomes depends on their insertion preference and post-insertion genome dynamics (Sultana et al. 2019) which varies between different types of retrotransposons. An apparent lack of insertion specificity can be observed in members of the Ty1/Copia retrotransposon superfamily in *A. thaliana*. These elements insert indiscriminately into chromosomes and then are passively retained in proximal chromosomal regions as they are removed from euchromatin by purifying selection (Vini Pereira 2004). Even in the same order of retrotransposons, different insertion preferences can be observed in different species. For example, LINE elements appear to be distributed fairly evenly across chromosomes in maize (Baucom et al. 2009), whereas a primarily centromeric distribution was observed for sunflower LINEs (Nagaki et al. 2015). A very clear preference for insertion into centromeric heterochromatin is found in plant-specific LTR-retrotransposon elements belonging to the chromovirus CRM clade. A group of elements from the CRM clade possess a chromodomain in their integrase domain that is supposed to target them into centromeres. Due to their widespread occurrence in plants and plant centromeres, elements belonging to the CRM clade have been extensively studied (Neumann et al. 2011).

Long homogeneous arrays of satDNA, occasionally interrupted by TE insertions, are the major components and are most commonly found in constitutive heterochromatin. Constitutive heterochromatin is a genomic compartment that remains compact and transcriptionally repressed regardless of circumstances. Heterochromatin can also be facultative, meaning that it differs in its compaction and expression levels depending on various conditions, such as developmental stage, cell cycle stage, or nuclear location. Principally heterochromatin is characterized by gene scarcity, reduced recombination (Janssen et al. 2018) and the presence of DNA methylation at CG, CHG and CHH motifs (Widman et al. 2009) as well as histone H3K9 and H3K27 mono- and dimethylation (Liu et al. 2010). Depending on the location of heterochromatin on chromosomes, it can be classified as subtelomeric, interstitial, or centromeric and pericentromeric. In smaller plant genomes, the majority of heterochromatin is located in the centromeres and pericentromeres, whereas plant species with larger genomes may have additional interstitial and subtelomeric bands (Vanrobays et al. 2017). Since heterochromatin lacks coding capacity, it was considered useless. This opinion has since changed and we now know that heterochromatin is involved in the maintenance of genome architecture, sister chromatid cohesion, kinetochore formation and gene regulation (Vanrobays et al. 2017).

Despite the importance of heterochromatin in eukaryotic genomes, our knowledge of the underlying nucleotide sequences, the relationships they may have with each other, or the mechanisms acting on them is limited. This is primarily due to the repetitive structure of heterochromatin, which makes it very difficult to assemble. Fortunately, with a combination of short-read sequencing technologies and cytogenetic techniques, it is possible to characterize the underlying nucleotide sequence and map its distribution on chromosomes. However, we lack an intermediate level of information to study the more detailed organization of satDNA and TEs and their patterns in heterochromatin. Thus, until we are able to easily look at the long range organization of repetitive DNA in eukaryotes, we will not fully understand the biology of repetitive DNA and, consequently, eukaryotic genomes.

1. Origin and evolution of satellite DNA

Individual evolutionary mechanisms acting on tandem repeats have been explored, but there is a lack of understanding of how these mechanisms interact with each other under different circumstances to create the repeat landscapes we observe.

Computer simulations from the 1970s and 1980s focused particularly on the relationship between unequal crossing over and the evolution of tandem repeats. In particular, the work of Smith demonstrates the possibility of the emergence of satDNA from random, non-tandemly organized sequences through the joint action of unequal crossing over (UCO) and mutation (Smith 1976). In his model, Smith assumed that UCO occurs between sister chromatids without the pressure of natural selection. However, Smith did not consider the cases of intrastrand recombination, which certainly lead to deletions and loss of tandem arrangements. Based on this, further studies showed that recombination alone is not only insufficient for satDNA conservation but would also lead to its elimination, suggesting that other mechanisms contribute to satDNA evolution (Walsh 1987). Consistent with this, Charlesworth postulated that suppressed recombination rates lead to expansion of repeats (Charlesworth et al. 1994), which is supported by the frequent colocalization of long tandem repeat arrays and heterochromatin (Thakur et al. 2021).

In vitro studies have shown that replication slippage is an important mechanism in the formation of short simple tandem repeats (Schlotterer and Tautz 1992). With DNA polymerase stalling and disassociation (Viguera et al. 2001), there is a preference for insertion of di- and tri-nucleotide sequences. While the speed of replication slip and the size of insertion depend on the disassociation propensity of DNA polymerase in the first case and on the nature of the underlying sequence in the second case, it appears that the rate of slippage may be higher in AT -rich sequences (Schlotterer and Tautz 1992).

The results of numerous studies show that TEs have an important influence on the origin and evolution of satDNA (Sharma et al. 2013; Belyayev et al. 2020; Huang et al. 2021). In some cases microsatellite and minisatellites were found embedded in TEs (Inukai 2004; Smýkal et al. 2009). In addition, a PisTR-A tandem repeat with a bimodal distribution of array sizes was found in *Pisum sativum* as it was organized in short arrays in 3' untranslated regions (3’UTRs) of Ogre retrotransposons and as extended arrays in different genomic loci (Macas et al. 2009). These studies provide evidence for the role of the transposition machinery in generating new tandem repeats and mobilizing them throughout the genome, whereupon the tandem repeats have the potential to expand into satDNA. TEs can also serve as template sequences for the creation of new satDNA, evidenced by Sobo satDNA of *Solanum bulbocastanum*. Based on the monomer's size of 4.7kb and similarity to LTR sequences of a Ty3/Gypsy retrotransposon, it is unlikely that the monomer was formed by repplication slippage. This leads to the hypothesis that it is a product of extrachromosomal circular DNA (EccDNA) amplification and insertion (Tek et al. 2005). However, it is important to note that dispersed repeats were not found in EccDNA based on 2D gel analyzes (Cohen and Segal 2009).

High similarity between satDNA monomers independent of the rDNA locus and the intergenic spacer (IGS) of rDNA can be observed in numerous plant species, suggesting that the IGS can also act as a seed sequence for the formation of new satDNA. Such satDNA was found in *Vicia faba* (Maggini et al. 1991), *Vigna radiata* (Unfried et al. 1991), *Phaseolus vulgaris* (Falquet et al. 1997), *Solanum bulbocastanum* (Stupar et al. 2002), *Vicia sativa* (Macas et al. 2003), and tobacco (Lim et al. 2004). Unfortunately, the present results do not allow us to determine whether these satDNA sequences integrate into the IGS or whether they are derived from the IGS and spread to new loci. Nevertheless, Macas *et al.* (Macas et al. 2003) have provided arguments favouring the latter hypothesis. First, that the IGS-like satDNA is present in those species that have a similar IGS structure, such as *V. faba* and *V. radiata*. Second, although the IGS within the *Vicia*, *Vigna*, and *Phaseolus* genera has a complex structure with at least two subrepeats, the satDNA shows homology only with the subrepeat adjacent to the 3' end of the 25S gene where transcription termination sequences were mapped. This suggests that a possible mechanism for the formation of satDNA may have been reverse transcription and reintegration into new loci. Finally, large monomeric tandem repeats homologous to IGS and sharing the same complex structure were identified in *Solanum bulbocastanum* (Stupar et al. 2002). Thus, the probability that such large monomeric tandem repeats did not arise from IGS is quite low (Macas et al. 2003).

After their emergence and potential expansion into long homogeneous arrays of satDNA, their monomers do not randomly accumulate mutations and change independently. SatDNA arrays follow a kind of nonindependent evolution called concerted evolution (Garrido-Ramos 2017). Concerted evolution is thought to be a consequence of DNA repair and replication mechanisms (Elder and Turner 1995). In this way, new monomer variants spread horizontally to all members of the satDNA family faster than new mutations accumulate, resulting in high intraspecies and low interspecies similarity of satDNA. These mechanisms interact in different ways to generate different homogenization patterns, with arrays homogenizing into either long stretches of identical monomers or higher order repeats (HOR). HORs are formed when two or more different monomer types are homogenized as a unit (Plohl et al. 2012), with monomers within one HOR having less similarity than corresponding monomers of neighboring HOR (Willard and Waye 1987).

One of the mechanisms thought to be involved in the expansion and homogenization of satDNA arrays is EccDNA amplification and integration. EccDNA molecules are most likely the products of intrastrand recombinations within satDNA arrays and could act as templates for rolling circle amplification. The linear product, containing multiple copies of the template EccDNA, is then possibly reintegrated into the genome thereby increasing the satDNA copy number. EccDNAs have been identified in a number of plant species and can vary in size from 500bp to 20kb (Cohen et al. 2008). Navrátilová *et al.* examined 10 plant species and found EccDNA products from 9 satDNA families and 3 subfamilies, often in multiples of the corresponding monomers and open circles (Navrátilová et al. 2008). The influence of EccDNA on satDNA evolution is unclear since their reinsertion has not been detected. Thus, EccDNA may only be a product of satDNA removal from the genome (Navrátilová et al. 2008).

Gene conversion is attributed to satDNA homogenization. It involves a nonreciprocal transfer where one DNA sequence replaces a homologous one, so that the sequences become identical. It is thought to be a normal product of homologous recombination in which the heteroduplex is resolved as a noncrossover. However, it is more difficult to detect than crossing over because the converged sequence is usually less than 2kb in length (Talbert and Henikoff 2010). Nevertheless, gene conversion has been confirmed to facilitate homogenization of satDNA in *Arabidopsis* and *Zea mays* centromeres (Kawabe and Charlesworth 2007; Shi et al. 2010). Lastly, segmental duplication events were also found to modulate the expansion of centromeric repeats. Particularly in rice where segmental duplications contributed more to centromeric retrotransposon accumulation than element insertion (Ma and Jackson 2006).

Concerted evolution can also be disrupted, which leads to divergence of satDNA families and formation of new families. Several factors such as chromosome location (Macas et al. 2006), recombination rate between homologous and non-homologous chromosomes (Navajas-Pérez et al. 2005; Navajas-Pérez et al. 2009), or the number of reproducing individuals within a population (Andrea et al. 2006) have the potential to disrupt or enhance concerted evolution.

2. Satellite DNA and the centromere

The centromere is a chromosomal region to which the multiprotein kinetochore complex assembles, the microtubule spindle is attached, and which is responsible for the cohesion of sister chromatids. Therefore, the centromere ensures the correct segregation of genetic material to daughter cells during mitotic and meiotic divisions (Jiang et al. 2003; Kursel and Malik 2016; Comai et al. 2017).

Three main types of centromere organizations have been discovered: monocentromeres, meta-polycentromeres, and holocentromeres (Schubert et al. 2020). The most obvious difference between these centromere types is the number of microtubule attachment sites and their distribution along the chromosomes. Monocentromeres have only one site or region for microtubule attachment. The simple "dot" monocentromere of *Saccharomyces cervisiae* was the first centromere whose nucleotide sequence was characterized by John Carbon and Louise Clarke as only 125bp long and as such is necessary and sufficient for centromere formation and function (Clarke and Carbon 1985). However, the short and simple centromere of *S. cervisiae* is the exception rather than the rule, as most higher plants have large "regional" monocentromeres that can span megabases, as observed in rice (Li et al. 2021), maize (Jiao et al. 2017), or potato (Pham et al. 2020). In metaphase, these centromeres are visible as primary constrictions surrounded by pericentromeric heterochromatin bearing specific post-translational marks (Fransz et al. 2006). Holocentromeres differ from monocentromeres in that centromeric activity is extended along the entire length of chromosomes that have no visible primary constriction in metaphase (Schubert et al. 2020). Holocentricity has evolved at least 13 times convergently in different lineages, including insects, nematodes, arachnids, and plants (Melters et al. 2013; Kasinathan and Henikoff 2018) . Moreover, it is thought to be a monocentric-derived trait. In what way or with what goal the transition from monocentricity to holocentricity may have occurred is unclear, but Neumann *et al.* have argued that meta-polycentromeres may be an evolutionary link between monocentricity and holocentricity

(Neumann et al. 2012). Meta-polycentromeres were first observed in *Pisum sativum* and later in *Lathyrus sativus*, where there may be between 2 and 5 regions of microtubule attachment and the primary constrictions are much more extensive (Neumann et al. 2012; Neumann et al. 2015).

Although the function of the centromere is largely conserved in eukaryotes, its nucleotide sequence is not, which has been termed the centromere paradox (Henikoff et al. 2001). Nevertheless, centromeres often exhibit a repetitive structure. An example are the centromeres of *Arabidopsis thaliana*, whose major component is the 180 bp pAL1 satellite (Nagaki et al. 2003), which is present in all centromeres, but in varying proportions (Copenhaver et al. 1999), and in long continuous arrays occasionally interspersed with an Athila retrotransposon (Thompson et al. 1996). Although the centromeres of *A. thaliana* are uniform in sequence composition, plant species have been found with more diversified centromere sequence composition. For example, an extraordinary diversity of 13 centromeric satellites, all differentially distributed among the chromosomes, was discovered in *P. sativum* (Neumann et al. 2012). While in *Vicia faba*, three distinct centromeric satellites were found on chromosome one and one chromosome-specific centromeric satellite was found on four other centromeres, as well as one centromere without any tandem repeats (Ávila Robledillo et al. 2018). Repeat-based and repeatless centromeres also coexist in the genome of *Solanum tuberosum* (Gong et al. 2012). In contrast, holocentromeres were assumed to lack centromeric repeats (Heckmann et al. 2013) until the centromeres of *Rhynchospora pubera* were characterized. *R. pubera* was the first holocentromeric species in which centromeric sequences were detected. Similar to monocentromeres or meta-polycentromeres, they consist mainly of a tandem repeat, a centromeric retrotransposon, and two other mobile elements (Marques et al. 2015). Unlike how monocentromeres or meta-polycentromeres organize into large units that then affect large-scale chromosome organization (Muller et al. 2019), the small centromeric units of holocentromeres are interspersed with euchromatin (Marques et al. 2015). How these small repetitive centromeric units are spread throughout the genome and influence genome architecture is not yet known.

CENH3 is used as a marker for centromeric activity since the nucleotide sequence of centromeres is not conserved and therefore cannot be used. CENH3 is a centromeric histone variant of histone H3 and an inner kinetochore protein that is conserved in presence and function across eukaryotic species (Karpen and Allshire 1997).

In the holocentric plant *Cuscuta europaea*, CENH3 has lost its function and instead of colocalizing along chromosomes to centromeres, it is deposited on large DAPI-positive heterochromatic bands. The finding of these bands is unusual, as holocentric plants usually do not have them. Even more surprising, however, is the complex structure of these heterochromatin bands, which includes a CUS-TR24 satDNA, a simple sequence repeat, and TEs (Oliveira et al. 2020). However, it was not possible to study the pattern organization of the heterochromatin components.

3. Methods of satellite DNA research

Due to its tandem repetitive nature and sequence composition, satDNA often separates from the rest of the genome when centrifuged in a cesium chloride gradient. On this basis it was first discovered in mouse in 1961 (Kit 1961). Subsequently, various techniques such as C_0t analysis, restriction endonuclease analysis, and Southern blot have been used to identify and characterize satDNA of various species. However, less abundant satDNA families or those that do not contain restriction sites are left undetected with these methods (Garrido-Ramos 2017; Novák et al. 2017). Even with the advancements in genome sequencing and the development of the Sanger sequencing technique, satDNA has been largely neglected. This is primarily due to the repetitive nature of satDNA, that makes assembling the contigs containing them difficult and tedious. The technical limitations of studying satDNA have led to a lack of understanding of the biology of satDNA and centromeres.

SatDNA research intensified with the development of Next Generation Sequencing (NGS). The first NGS techniques were a big improvement over Sanger sequencing due to their massively parallel sequencing by synthesis. This not only increased throughput and accuracy, but also reduced the cost of sequencing (Schatz et al. 2010). The first commercially available NGS technology was Roche 454, which uses pyrosequencing. Pyrosequencing is based on the detection of a pyrophosphate light signal, a byproduct of nucleotide incorporation (Liu et al. 2012). Later, other NGS technologies entered the market, with Illumina quickly becoming the most widely used. In this process, library DNA fragments are provided with adapters that allow the template to hybridize to flow cell oligos, and a polymerase creates a copy of the template. This bound template copy is then amplified by bridge amplification, creating clusters of clonal DNA fragments. After clustering is complete, sequencing begins. One by one, the different nucleotides carrying a fluorescent dye are incorporated, and the fluorescent signal is detected (Bronner et al. 2013). This significant advance in sequencing technology has led to an increase in sequencing projects for various plant and animal species (Schatz et al. 2010).

Nevertheless, the disadvantage of all NGS sequencing methods at that time was their short read length, e.g., standard Illumina reads had a length of up to 300nt. Since NGS reads are not able to span satDNA arrays or TEs, repetitive DNA causes significant problems in genome assemblies as highly repetitive genome parts are left as gaps and different repeats are therefore unrepresented (Treangen and Salzberg 2012). Thus, identification of new repeats by their assembly is not possible. Additionally, *de novo* identification of satDNA even between closely related species is complicated by similarity searches of existing repeat databases, as satDNA is evolutionarily dynamic.

RepeatExplorer is a graph based clustering pipeline developed with the motivation to *de novo* identify repeats from unassembled Illumina reads (Novák et al. 2013; Novák et al. 2020). The input to RepeatExplorer are paired-end low pass shotgun Illumina reads between 100-300 nt in length. When the genome coverage is low, between 0.1-0.5X genome coverage, a random sample of the genome is generated. The short unassembled reads are all compared and repeats are identified based on the frequent similarity hits between reads since the number of similarity hits is proportional to genome copy number of the sequences that the reads represent. Therefore the number of similarity hits should be smallest for low copy or unique sequences and largest for repetitive sequences. The similarities can be represented as edges of a graph, connecting vertices representing individual reads. Based on the identification of the clusters of frequently connected nodes in the graph, different types of repetitive elements can be identified, and the number of reads in each cluster allows estimation of the repeat abundance (Novák et al. 2010; Novák et al. 2013; Novák et al. 2020). Although RepeatExplorer was developed primarily for identifying repeats in plants, it has been used for both plant and animal species analysis (Weiss-Schneeweiss et al. 2015;

Ruiz-Ruano et al. 2016; Pamponét et al. 2019). Nevertheless, there were limitations with respect to identification and to the analysis of satDNA, such as the need for visual inspection of cluster graphs and the inability to infer the most abundant monomer consensus sequences.

The TAREAN program proved to be a significant improvement in the unsupervised identification of tandem repeats and the reconstruction of consensus monomer sequences. The first step of TAREAN analysis the same to that of Repeat Explorer, in which low-pass shotgun Illumina reads are all compared to form clusters. To identify which clusters belong to tandem repeats, the read clusters are converted into directed graphs. Based on the connectivity of the directed graphs, they can be classified as belonging or not belonging to a tandem repeat. To reconstruct the consensus sequence of the most abundant monomers, the occurrence of k-mers in the oriented reads must be counted from the directed graphs. The most abundant k-mers are then used to reconstruct the most representative monomers for each satellite (Novák et al. 2017). These bioinformatics tools have been successfully used to identify and classify novel repetitive DNA elements in various organisms (Pamponét et al. 2019; González et al. 2020; Boštjančić et al. 2021; Valeri et al. 2021).

Although RepeatExplorer and TAREAN successfully use short reads to identify and quantify repetitive DNA, the shortness of the reads means that information on how these repeats are organized in the genome and in relation to each other is lacking.

The recent development of long read sequencing technologies offers significant improvement in read length. The available long read technologies are Single Molecule Real Time (SMRT) sequencing from Pacific Biosciences (PacBio) and Oxford Nanopore sequencing. However, while PacBio is based on sequencing by synthesis (Rhoads and Au 2015), Oxford Nanopore sequencing developed a new approach. A double stranded DNA molecule is unwound and passed through a protein pore embedded in a synthetic membrane. As the intact single-stranded DNA passes through the pore, the change of the ion current is recorded as a “squiggle” signal, which is then used for basecalling to determine the nucleotide sequence. Oxford Nanopore technology is very valuable because of their potential to generate extremely long reads that can be hundreds of kilobases in length, especially in conjunction with improved protocols (Jain et al. 2016).

These long-read technologies are capable of capturing highly repetitive genomic regions and improving the quality of genome assemblies. By combining long Nanopore reads and short reads of a higher accuracy, the human Y centromere was sequenced (Jain et al. 2018), while the combination of multiple long read technologies produced a telomere to telomere sequence of the human X chromosome (Miga et al. 2020) and most recently the complete sequence of a human genome (Nurk et al. 2022). In addition, unassembled long reads have been successfully used to analyze tandem repeats (Cechova et al. 2019; Harris et al. 2019).

Because of their large size, long reads therefore provide an unprecedented opportunity to observe the patterns that repetitive DNA generates on a longer scale than Illumina reads allow to investigate. By observing these patterns, we are able to describe the evolutionary mechanisms and genome dynamics that lead to their formation, leading to a better understanding of satDNA biology.

References

Ambrožová K, Mandáková T, Bureš P, Neumann P, Leitch IJ, Koblížková A, Macas J, Lysak MA. 2011. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. *Annals of Botany* 107:255–268. DOI:10.1093/aob/mcq235

Andrea L, Marini M, Mantovani B. 2006. Non-concerted evolution of the RET76 satellite DNA family in *Reticulitermes* taxa (Insecta, Isoptera). *Genetica* 128:123–132. DOI:10.1007/s10709-005-5540-z

Arkhipova IR, Yushenova IA, Angert E. 2019. Giant Transposons in Eukaryotes: Is Bigger Better? *Genome Biology and Evolution* 11:906–918. DOI:10.1093/gbe/evz041

Ávila Robledillo L, Koblížková A, Novák P, Böttinger K, Vrbová I, Neumann P, Schubert I, Macas J. 2018. Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Scientific Reports* 8:5838. DOI:10.1038/s41598-018-24196-3

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genetics* 5:e1000732. DOI:10.1371/journal.pgen.1000732

Belyayev A, Josefiová J, Jandová M, Mahelka V, Krak K, Mandák B. 2020. Transposons and satellite DNA: On the origin of the major satellite DNA family in the *Chenopodium* genome. *Mobile DNA* 11:20. DOI:10.1186/s13100-020-00219-7

Boštjančić LL, Bonassin L, Anušić L, Lovrenčić L, Besendorfer V, Maguire I, Grandjean F, Austin CM, Greve C, Hamadou A Ben, *et al.* 2021. The *Pontastacus leptodactylus* (Astacidae) Repeatome Provides Insight Into Genome Evolution and Reveals Remarkable Diversity of Satellite DNA. *Frontiers in Genetics* 11:611745. DOI:10.3389/fgene.2020.611745

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, *et al.* 2018. Ten things you should know about transposable elements. *Genome Biology* 19:1–12. DOI:10.1186/s13059-018-1577-z

Bronner IF, Quail MA, Turner DJ, Swerdlow H. 2013. Improved Protocols for Illumina Sequencing. *Current Protocols in Human Genetics* 79:18–2. DOI:10.1002/0471142905.hg1802s79

Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. *Molecular Ecology* 22:1503–1517. DOI:10.1111/mec.12170

Cechova M, Harris RS, Tomaszkiewicz M, Arbeithuber B, Chiaromonte F, Makova KD. 2019. High Satellite Repeat Turnover in Great Apes Studied with Short- And Long-Read Technologies. *Molecular Biology and Evolution* 36:2415–2431. DOI:10.1093/molbev/msz156

Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371:215–220. DOI:10.1038/371215a0

Clarke L, Carbon J. 1985. The structure and function of yeast centromeres. *Annual Review of Genetics* 19:29–56. DOI:10.2106/00004623-196244040-00020

Cohen S, Houben A, Segal D. 2008. Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants. *The Plant Journal* 53:1027–1034. DOI:10.1111/j.1365-313X.2007.03394.x

Cohen S, Segal D. 2009. Extrachromosomal circular DNA in eukaryotes: Possible involvement in the plasticity of tandem repeats. *Cytogenetic and Genome Research* 124:327–338. DOI:10.1159/000218136

Comai L, Maheshwari S, Marimuthu MPA. 2017. Plant centromeres. *Current Opinion in Plant Biology* 36:158–167. DOI:10.1016/j.pbi.2017.03.003

Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S, Lin X, Bevan M, Murphy G, Harris B, Parnell LD, et al. 1999. Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* 286:2468–2474. DOI:10.1126/science.286.5449.2468

Deragon JM, Zhang X. 2006. Short interspersed elements (SINEs) in plants: Origin, classification, and use as phylogenetic markers. *Systematic Biology* 55:949–956. DOI:10.1080/10635150601047843

Dogan M, Pouch M, Mandáková T, Hloušková P, Guo X, Winter P, Chumová Z, Van Niekerk A, Mummenhoff K, Al-Shehbaz IA, et al. 2021. Evolution of Tandem Repeats Is Mirroring Post-polyploid Cladogenesis in *Heliophila* (Brassicaceae). *Frontiers in Plant Science* 11:607893. DOI:10.3389/fpls.2020.607893

Elder JF, Turner BJ. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. *Quarterly Review of Biology* 70:297–320. DOI:10.1086/419073

Falquet J, Creusot F, Dron M. 1997. Molecular analysis of *Phaseolus vulgaris* rDNA unit and characterization of a satelliteDNA homologous to IGS subrepeats. *Plant Physiology and Biochemistry* 35:611–622.

Feschotte C, Xiaoyu Z, Wessler SR. 2007. Miniature Inverted-Repeat Transposable Elements and Their Relationship to Established DNA Transposons. *Mobile DNA II* :1147–1158.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends in genetics* 5:103–107.

Fleischmann A, Michael TP, Rivadavia F, Sousa A, Wang W, Temsch EM, Greilhuber J, Müller KF, Heubl G. 2014. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany* 114:1651–1663. DOI:10.1093/aob/mcu189

Fransz P, Ten Hoopen R, Tessadori F. 2006. Composition and formation of heterochromatin in *Arabidopsis thaliana*. *Chromosome Research* 14:71–82. DOI:10.1007/s10577-005-1022-5

Garrido-Ramos MA. 2017. Satellite DNA: An evolving topic. *Genes* 8:230. DOI:10.3390/genes8090230

Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, Iovene M, Neumann P, Zhang W, Novák P, Buell CR, et al. 2012. Repeatless and Repeat-Based Centromeres in Potato: Implications for Centromere Evolution. *The Plant cell* 24:3559–3574. DOI:10.1105/tpc.112.100511

González ML, Chiapella J, Topalian J, Urdampilleta JD. 2020. Genomic differentiation of *Deschampsia antarctica* and *D. cespitosa* (Poaceae) based on satellite DNA. *Botanical Journal of the Linnean Society* 194:326–341. DOI:10.1093/botlinnean/boaa045

Gregory TR. 2005. The C-value enigma in plants and animals: A review of parallels and an appeal for partnership. *Annals of Botany* 95:133–146. DOI:10.1093/aob/mci009

Harris RS, Cechova M, Makova KD. 2019. Noise-cancelling repeat finder: Uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics* 35:4809–4811. DOI:10.1093/bioinformatics/btz484

Heckmann S, Macas J, Kumke K, Fuchs J, Schubert V, Ma L, Novák P, Neumann P, Taudien S, Platzer M, et al. 2013. The holocentric species *Luzula elegans* shows interplay between centromere and large-scale genome organization. *The Plant Journal* 73:555–565. DOI:10.1111/tpj.12054

Henikoff S, Ahmad K, Malik HS. 2001. The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. *Science* 293:1098–1102. DOI:10.1126/science.1062939

Huang Y, Ding W, Zhang M, Han J, Jing Y, Yao W, Hasterok R, Wang Z, Wang K. 2021. The formation and evolution of centromeric satellite repeats in *Saccharum* species. *The Plant Journal* 106:616–629. DOI:10.1111/tpj.15186

Hufnagel B, Marques A, Soriano A, Marquès L, Divol F, Doumas P, Sallet E, Mancinotti D, Carrere S, Marande W, et al. 2020. High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. *Nature Communications* 11:1–12. DOI:10.1038/s41467-019-14197-9

Inukai T. 2004. Role of transposable elements in the propagation of minisatellites in the rice genome. *Molecular Genetics and Genomics* 271:220–227. DOI:10.1007/s00438-003-0973-5

Jain M, Olsen HE, Paten B, Akeson Mark, Branton D, Daniel B, Deamer Dw, Andre M, Hagan B, Benner S, et al. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* 17:1–11. DOI:10.1186/s13059-016-1103-0

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel K V., Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018. Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology* 36:321–323. DOI:10.1038/nbt.4109

Janssen A, Colmenares SU, Karpen GH. 2018. Heterochromatin: Guardian of the Genome. *Annual Review of Cell and Developmental Biology* 34:265–288. DOI:10.1146/annurev-cellbio-100617-062653

Jiang J, Birchler J a, Parrott W a, Kelly Dawe R. 2003. A molecular view of plant centromeres. *Trends in Plant Science* 8:570–575. DOI:10.1016/j.tplants.2003.10.011

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C, et al. 2017. Improved maize reference genome with single-molecule technologies. *Nature* 546:524–527. DOI:10.1038/nature22971

Karpen GH, Allshire RC. 1997. The case for epigenetic effects on centromere identity and function. *Trends in Genetics* 13:489–496. DOI:10.1016/s0168-9525(97)01298-5

Kasinathan S, Henikoff S. 2018. Non-B-form DNA is enriched at centromeres. *Molecular Biology and Evolution* 35:949–962. DOI:10.1093/molbev/msy010

Kawabe A, Charlesworth D. 2007. Patterns of DNA variation among three centromere satellite families in *Arabidopsis halleri* and *A. lyrata*. *Journal of Molecular Evolution* 64:237–247. DOI:10.1007/s00239-006-0097-8

Kejnovsky E, Hawkins JS, Feschotte C. 2012. Plant Transposable Elements: Biology and Evolution. *Plant Genome Diversity Volume 1*:1–279. DOI:10.1007/978-3-7091-1130-7

Kit S. 1961. Equilibrium Sedimentation in Density Gradients of DNA Preparations from Animal Tissues. *Journal of Molecular Biology* 3:711–716. DOI:10.1016/S0022-2836(61)80075-2

Kumar A. 2020. Jump around: Transposons in and out of the laboratory. *F1000Research* 9:1–12. DOI:10.12688/f1000research.21018.1

Kursel LE, Malik HS. 2016. Centromeres. *Current Biology* 26:R487–R490. DOI:10.1016/j.cub.2016.05.031

Li K, Jiang W, Hui Y, Kong M, Feng LY, Gao LZ, Li P, Lu S. 2021. Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. *Molecular Plant* 14:1745–1756. DOI:10.1016/j.molp.2021.06.017

Lim KY, Skalicka K, Koukalova B, Volkov RA, Matyasek R, Hemleben V, Leitch AR, Kovarik A. 2004. Dynamic Changes in the Distribution of a Satellite Homologous to Intergenic 26-18S rDNA Spacer in the Evolution of *Nicotiana*. *Genetics* 166:1935–1946. DOI:10.1534/genetics.166.4.1935

Liu C, Lu F, Cui X, Cao X. 2010. Histone methylation in higher plants. *Annual Review of Plant Biology* 61:395–420. DOI:10.1146/annurev.arplant.043008.091939

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* 251364:. DOI:10.1155/2012/251364

Lower SS, McGurk MP, Clark AG, Barbash DA. 2018. Satellite DNA evolution: old ideas, new approaches. *Current Opinion in Genetics and Development* 49:70–78. DOI:10.1016/j.gde.2018.03.003

Ma J, Jackson SA. 2006. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Research* 16:251–259. DOI:10.1101/gr.4583106

Macas J, Koblížková A, Navrátilová A, Neumann P. 2009. Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. *Gene* 448:198–206. DOI:10.1016/j.gene.2009.06.014

Macas J, Mészáros T, Nouzová M. 2002. PlantSat: A specialized database for plant satellite repeats. *Bioinformatics* 18:28–35. DOI:10.1093/bioinformatics/18.1.28

Macas J, Navrátilová A, Koblížková A. 2006. Sequence homogenization and chromosomal localization of VicTR-B satellites differ between closely related *Vicia* species. *Chromosoma* 115:437–447. DOI:10.1007/s00412-006-0070-8

Macas J, Navrátilová A, Mészáros T. 2003. Sequence subfamilies of satellite repeats related to rDNA intergenic spacer are differentially amplified on *Vicia sativa* chromosomes. *Chromosoma* 112:152–158. DOI:10.1007/s00412-003-0255-3

Macas J, Novak P, Pellicer J, Čížková J, Koblížková A, Neumann P, Fuková I, Doležel J, Kelly LJ, Leitch IJ. 2015. In Depth Characterization of Repetitive DNA in 23 Plant Genomes Revelas Sources of Genome Size Variation in the Legume Tribe *Fabae*. *PLoS ONE* 10:1–23. DOI:10.1371/journal.pone.0143424

Maggini F, Cremonini R, Zolfino C, Tucci GF, D’Ovidio R, Delre V, DePace C, Scarascia Mugnozza GT, Cionini PG. 1991. Structure and chromosomal localization of DNA sequences related to ribosomal subrepeats in *Vicia faba*. *Chromosoma* 100:229–234. DOI:10.1007/BF00344156

Marques A, Ribeiro T, Neumann P, Macas J, Novák P, Schubert V, Pellino M, Fuchs J, Ma W, Kuhlmann M, et al. 2015. Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. *Proceedings of the National Academy of Sciences of the United States of America* 112:13633–13638. DOI:10.1073/pnas.1512255112

Mascagni F, Barghini E, Ceccarelli M, Baldoni L, Trapero C, Díez CM, Natali L, Cavallini A, Giordani T. 2022. The Singular Evolution of *Olea* Genome Structure. *Frontiers in Plant Science* 13:869048–869048. DOI:10.3389/fpls.2022.869048

McClintock B. 1951. Chromosome organization and genic expression. *Cold Spring Harbor symposia on quantitative biology* 16:13–47. DOI:10.1101/SQB.1951.016.01.004

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

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology* 14:1–20. DOI:10.1186/gb-2013-14-1-r10

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585:79–84. DOI:10.1038/s41586-020-2547-7

Muller H, Gil J, Drinnenberg IA. 2019. The Impact of Centromeres on Spatial Genome Architecture. *Trends in Genetics* 35:565–578. DOI:10.1016/j.tig.2019.05.003

Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang J. 2003. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics* 163:1221–1225. DOI:10.1093/genetics/163.3.1221

Nagaki K, Tanaka K, Yamaji N, Kobayashi H, Murata M. 2015. Sunflower centromeres consist of a centromere-specific LINE and a chromosome-specific tandem repeat. *Frontiers in Plant Science* 6:912. DOI:10.3389/fpls.2015.00912

Navajas-Pérez R, De La Herrán R, González GL, Jamilena M, Lozano R, Rejón CR, Rejón MR, Garrido-Ramos MA. 2005. The evolution of reproductive systems and sex-determining mechanisms within *Rumex* (polygonaceae) inferred from nuclear and chloroplastidial sequence data. *Molecular Biology and Evolution* 22:1929–1939. DOI:10.1093/molbev/msi186

Navajas-Pérez R, Schwarzacher T, Rejón MR, Garrido-Ramos MA. 2009. Molecular cytogenetic characterization of *Rumex papillaris*, a dioecious plant with an XX/XY1Y2 sex chromosome system. *Genetica* 135:87–93. DOI:10.1007/s10709-008-9261-y

Navrátilová A, Koblížková A, Macas J. 2008. Survey of extrachromosomal circular DNA derived from plant satellite repeats. *BMC Plant Biology* 8:1–13. DOI:10.1186/1471-2229-8-90

Neumann P, Navrátilová A, Koblížková A, Kejnovsky E, Hřibová E, Hobza R, Widmer A, Doležel J, Macas J. 2011. Plant centromeric retrotransposons: A structural and cytogenetic perspective. *Mobile DNA* 2:1–16. DOI:10.1186/1759-8753-2-4

Neumann P, Navrátilová A, Schroeder-Reiter E, Koblížková A, Steinbauerová V, Chocholová E, Novák P, Wanner G, Macas J. 2012. Stretching the rules: Monocentric Chromosomes with Multiple Centromere Domains. *PLoS Genetics* 8:e1002777. DOI:10.1371/journal.pgen.1002777

Neumann P, Pavlíková Z, Koblížková A, Fuková I, Jedličková V, Novák P, Macas J. 2015. Centromeres Off the Hook: Massive Changes in Centromere Size and Structure Following

Duplication of CenH3 Gene in *Fabae* Species. *Molecular Biology and Evolution* 32:1862–1879. DOI:10.1093/molbev/msv070

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* 11:1–12. DOI:10.1186/1471-2105-11-378

Novák P, Neumann P, Macas J. 2020. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols* 15:3745–3776. DOI:10.1038/s41596-020-0400-y

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

Novák P, Robledillo LÁ, Koblížková A, Vrbová I, Neumann P, Macas J. 2017. TAREAN: A computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research* 45:e111–e111. DOI:10.1093/nar/gkx257

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, *et al.* 2022. The complete sequence of a human genome. *Science* 376:44–53. DOI:10.1126/science.abj6987

Oliveira L, Neumann P, Jang T, Klemme S, Schubert V. 2020. Mitotic spindle attachment to the holocentric chromosomes of *Cuscuta europaea* does not correlate with the distribution of CENH3 chromatin. *Frontiers in Plant Science* 10:1–11. DOI:10.3389/fpls.2019.01799

Pamponét VCC, Souza MM, Silva GS, Micheli F, De Melo CAF, De Oliveira SG, Costa EA, Corrêa RX. 2019. Correction to: Low coverage sequencing for repetitive DNA analysis in *Passiflora edulis* Sims: Citogenomic characterization of transposable elements and satellite DNA. *BMC Genomics* 20:1–17. DOI:10.1186/s12864-019-5678-1

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* 164:10–15. DOI:10.1111/j.1095-8339.2010.01072.x

Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, Vaillancourt B, Ou S, Jiang J, Robin Buell C. 2020. Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience* 9:1–11. DOI:10.1093/gigascience/giaa100

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA Evolution. DOI:https://doi.org/10.1159/000337122

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics* 13:278–289. DOI:10.1016/j.gpb.2015.08.002

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports* 6:1–14. DOI:10.1038/srep28333

Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research* 20:1165–1173. DOI:10.1101/gr.101360.109

Schlotterer C, Tautz D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* 20:211–215. DOI:10.1093/nar/20.2.211

Schubert V, Neumann P, Marques A, Heckmann S, Macas J, Pedrosa-Harand A, Schubert I, Jang T-S, Houben A. 2020. Super-Resolution Microscopy Reveals Diversity of Plant Centromere Architecture. *International journal of molecular sciences* 21:3488. DOI:10.3390/ijms21103488

Sharma A, Wolfgruber TK, Presting GG. 2013. Tandem repeats derived from centromeric retrotransposons. *BMC Genomics* 14:1–10. DOI:10.1186/1471-2164-14-142

Shi J, Wolf SE, Burke JM, Presting GG, Ross-Ibarra J, Dawe RK. 2010. Widespread Gene Conversion in Centromere Cores. *PLoS Biology* 8:1–10. DOI:10.1371/journal.pbio.1000327

Smith GP. 1976. Evolution of Repeated DNA Sequences by Unequal Crossover. *Science* 191:528–535. DOI:https://doi.org/10.1126/science.1251186

Smýkal P, Kalendar R, Ford R, Macas J, Griga M. 2009. Evolutionary conserved lineage of *Angela*-family retrotransposons as a genome-wide microsatellite repeat dispersal agent. *Heredity* 103:157–167. DOI:10.1038/hdy.2009.45

Stupar RM, Song J, Tek AL, Cheng Z, Dong F, Jiang J. 2002. Highly Condensed Potato Pericentromeric Heterochromatin Contains rDNA-Related Tandem Repeats. *Genetics* 162:1435–1444. DOI:10.1093/genetics/162.3.1435

Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger L, Nigumann P, Saccani S, Andrau JC, et al. 2019. The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Molecular Cell* 74:555-570.e7. DOI:10.1016/j.molcel.2019.02.036

Talbert PB, Henikoff S. 2010. Centromeres Convert but Don’t Cross. *PLoS Biology* 8:1–5. DOI:10.1371/journal.pbio.1000326

Talbert PB, Kasinathan S, Henikoff S. 2018. Simple and Complex Centromeric Satellites in *Drosophila* Sibling Species. *Genetics* 208:977–990. DOI:10.1534/genetics.117.300620

Tek AL, Song J, Macas J, Jiang J. 2005. Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences. *Genetics* 170:1231–1238. DOI:10.1534/genetics.105.041087

Thakur J, Packiaraj J, Henikoff S. 2021. Sequence, Chromatin and Evolution of Satellite DNA. *International Journal of Molecular Sciences* 22:4309. DOI:10.3390/ijms22094309

Thompson H, Schmidt R, Brandes A, Heslop-Harrison P, Dean C. 1996. A novel repetitive sequence associated with the centromeric regions of *Arabidopsis thaliana* chromosomes. *Molecular Genetics and Genomics* 253:247–252. DOI:10.1007/s004380050319

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics* 13:36–46. DOI:10.1038/nrg3117

Unfried K, Schiebel K, Hemleben V. 1991. Subrepeats of rDNA intergenic spacer present as prominent independent satellite DNA in *Vigna radiata* but not in *Vigna angularis*. *Gene* 99:63–68. DOI:10.1016/0378-1119(91)90034-9

Uozu S, Ikehashi H, Ohmido N, Ohtsubo H, Ohtsubo E, Fukui K. 1997. Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus *Oryza*. *Plant Molecular Biology* 35:791–799. DOI:10.1023/A:1005823124989

Valeri MP, Dias GB, do Espírito Santo AA, Moreira CN, Yonenaga-Yassuda Y, Sommer IB, Kuhn GCS, Svartman M. 2021. First Description of a Satellite DNA in Manatees' Centromeric Regions. *Frontiers in Genetics*:1537. DOI:10.3389/fgene.2021.694866

Vanrobays E, Thomas M, Tatout C. 2018. Heterochromatin Positioning and Nuclear Architecture. *Annual Plant Reviews online*. :157–190. DOI:10.1002/9781119312994.apr0502

Viguera E, Canceill D, Ehrlich SD. 2001. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO Journal* 20:2587–2595. DOI:10.1093/emboj/20.10.2587

Vini Pereira. 2004. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome biology* 5:1–10. DOI:10.1186/gb-2004-5-10-r79

Walsh JB. 1987. Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics* 115:553–567. DOI:https://doi.org/10.1093/genetics/115.3.553

Weiss-Schneeweiss H, Leitch AR, McCann J, Jang T, Macas J. 2015. Employing next generation sequencing to explore the repeat landscape of the plant genome. *Next-Generation Sequencing in Plant Systematics*. *Regnum Vegetabile* 157:155-179. DOI:10.14630/000006

Wells JN, Feschotte C. 2020. A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics* 54:539–561. DOI:10.1146/annurev-genet-040620-022145

Widman N, Jacobsen SE, Pellegrini M. 2009. Determining the conservation of DNA methylation in *Arabidopsis*. *Epigenetics* 4:119–124. DOI:10.4161/epi.4.2.8214

Willard HF, Waye JS. 1987. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends in Genetics* 3:192–198. DOI:10.1016/0168-9525(87)90232-0

Objective and aims of the thesis

The overall objective of this work was to use state of the art, long reads to explore patterns of satellite DNA genome organization to advance our understanding of the origin and evolution of satellite DNA.

For this purpose long unassembled nanopore reads belonging to two plant species with different centromere organizations were analyzed: the meta-polycentric *Lathyrus sativus* and the holocentric *Cuscuta europaea*. *L. sativus* was selected because of the large number of different satellite DNA families in its genome, which make it a good model for testing the suitability of unassembled nanopore reads for satellite DNA analysis. Because of its relatively large genome size and low chromosome number, it is also a good cytogenetic model, offering the possibility of using cytogenetic techniques to verify and complement bioinformatic results. The holocentric *C. europaea* was chosen because of its unusual heterochromatin bands. Not only are heterochromatin bands an unusual occurrence in holocentric species, but these bands also contain three repeat types: the CUS-TR24 satellite, simple sequence repeats, and additional repeats. This composition indicates a complex structure, in contrast to the usually expected homogeneous arrays of satellites sparsely interspersed with transposable elements. Moreover, the heterochromatin domain accumulates CENH3, a typical centromeric marker. Therefore, we decided to use the long reads to study the organization of repeats within the bands and infer the evolutionary mechanisms that led to their formation.

Finally, *Rynchospora pubera* was the first holocentric species found to have centromeric repeats consisting of Tyba satellites, centromeric retrotransposons, and two unidentified mobile elements. However, there was no information on the size of the Tyba arrays or how these elements were organized relative to each other. When a chromosome scale whole-genome assembly was created using PacBio HiFi reads, two whole genome duplication events were discovered that resulted in the formation of paralogous centromeric loci. Hence, taking advantage of the available genome assembly we compared the centromeric paralogs to find polymorphisms and to understand what changes these loci undergo and how they spread through the genome.

We aimed to:

1. Develop a bioinformatic pipeline for the annotation of repeats in long unassembled nanopore reads.

2. Utilize the previously developed pipeline to investigate the long-range organization of satellite DNA in *L. sativus* and *C. europaea*. Interpret these results to infer the origin and evolutionary patterns of selected satellite DNA families.

3. Utilize the genome assembly of *Rhynchospora pubera* to identify paralogous centromeric loci and investigate the mechanisms acting on their gain and removal in the genome.

Scope of the thesis

Chapter I addresses the genome-wide analysis of satellite DNA arrays enabled by nanopore sequencing and the development of a bioinformatics pipeline for their annotation. *Lathyrus sativus* was selected for analysis because it has a relatively large genome, a small chromosome number (2n=14), meta-polycentric chromosomes, and a large number of different satellite DNA families. Therefore, we wanted to investigate the size distribution of satellite DNA families and their possible correlation. A total of 11 satellite DNA families were analyzed, all representing at least 0.1% of the genome. Unexpectedly, two patterns of array size distribution were found. Only 2 of the 11 families analyzed appeared to consist mainly of long arrays, while the remaining satellite DNA families contained mainly short arrays (< 5kb). To investigate the possible association between the different satellite DNA families, the 10kb neighborhood on either side of each satellite array was examined. In the case of the satellite DNA arrays belonging to the long array group, the arrays were mostly surrounded by satellite DNA arrays of the same family due to breaks in the otherwise long arrays. However, for the other groups of satellites, frequent association with LTR Ogre retrotransposons was observed. Based on a subsequent neighborhood analysis of the satellite arrays associated with Ogre elements, which revealed an accurate and expected arrangement of protein domain spacing and orientations, we concluded that these short arrays are embedded in the Ogre retrotransposons. Furthermore, the results of the FISH experiments showed that the extended satellite arrays were located in the primary constrictions of the chromosomes. This suggests that the Ogre elements facilitate the movement of the short arrays, while the pericentromeres might promote the expansion of the satellite arrays. Furthermore, these results prove that unassembled long reads can be reliably used to study repetitive DNA.

In contrast to the extended homogeneous satellites of the pericentromeric and subtelomeric heterochromatin of *L. sativus*, the heterochromatin bands of holocentric *Cuscuta europaea* were known to have a complex structure containing the CUS-TR24 satellite, simple sequence repeats (SSR), and other repetitive elements. In addition, the heterochromatic bands were known to accumulate CENH3. Inspired by these unusual features in **Chapter II**, we set out to investigate the content and long-range organization of the heterochromatic bands. Nanopore sequencing yielded ultralong reads, and inspection of these reads using self-similarity dot plots revealed a complex structure of heterochromatin. A composite repeat structure of CUS-TR24 arrays regularly interspersed with SSR and a transposable element prompted us to analyze the array size distributions of the different components and assess their association with the same pipeline as in the previous chapter. It was found that the CUS-TR24 arrays are mostly short (< 10kb), but apart from that, there seem to be two types of array termination. A larger portion of the CUS-TR24 arrays is a multiple of the consensus monomer, while the other portion appears to be terminated by a truncated monomer of ~120bp. A 10kb window neighborhood analysis revealed that 40% of the CUS-TR24 arrays are terminated by SSR, leading to termination of arrays with a truncated monomer. 20% of CUS-TR24 arrays were terminated by transposable elements identified as LINE retrotransposons belonging to the L1- CS lineage. While three lineages of LINEs were identified in the *C. europaea* genome, only L1- CS was identified to specifically target CUS-TR24, while the remaining elements were found scattered throughout the genome. The insertion sites of LINEs were mapped to specific positions within the CUS-TR24 monomer, suggesting a possible nucleotide sequence-specific targeting of the insertion. Considering all these results, we propose an explanation for the origin of this pattern in which the CUS-TR24 monomer played a central role in providing seed sequences for SSR expansion and targeted insertion of the L1- CS LINE retrotransposon.

Rhynchospora pubera was the first holocentric species to have centromeric repeats identified, which were the Tyba satellite, a centromeric retrotransposon, and two other unidentified mobile elements. However, there was no information on the array sizes or the organisation of these repeats in the genome and in relation to each other. Therefore, we focused on the evolutionary dynamics of these centromeric repeats in **Chapter III**. Chromosome-scale genome assembly of *R. pubera* was performed along with two other holocentric and a closely related monocentric species to compare their genomes and potentially identify the changes in the transition from monocentricity to holocentricity. In addition, *R. pubera* was found to be an autooctoploid due to two genome duplications, providing a unique opportunity to study Tyba polymorphisms as well as differences in CENH3 deposition in four paralogous copies of the genome. The paralogous regions were identified as such based on the presence of two paralogous genes on either side of a CENH3 region. Most paralogous loci were found to be unaltered with respect to the presence of Tyba arrays or CENH3 deposits. However, some paralogous loci showed an increase in CENH3 deposition associated with the acquisition of a new Tyba array. Similar to *L. sativus*, where Ogre elements carry short arrays of tandem repeats, a Helitron DNA transposon insertion was frequently found to cause the emergence of new Tyba arrays. This suggests that the Helitron aids in the spread of Tyba. Loss of CENH3 loci in the four paralogous regions was also detected, mainly due to complete removal or degradation of the underlying Tyba arrays. These results suggest a clear link between CENH3 deposition and the Tyba satellite. Nevertheless, the question remains whether the transition to holocentricity was caused by the spread of Tyba to new loci, possibly due to Helitron activity, or whether the transition to holocentricity preceded the spread of Tyba and provided an environment favourable for Tyba accumulation.

Chapter I

Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats.

the plant journal SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2020) **101**, 484–500 doi: 10.1111/tpj.14546

TECHNICAL ADVANCE

Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats

Tihana Vondrak[1,2], Laura Ávila Robledillo[1,2], Petr Novák[1], Andrea Koblížková[1], Pavel Neumann[1] and Jiří Macas[1,*]
[1]*Biology Centre, Czech Academy of Sciences, Branišovská 31, České Budějovice CZ-37005, Czech Republic, and*
[2]*Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic*

Received 3 July 2019; revised 9 September 2019; accepted 12 September 2019; published online 27 September 2019.
*For correspondence (e-mail macas@umbr.cas.cz).

SUMMARY

Amplification of monomer sequences into long contiguous arrays is the main feature distinguishing satellite DNA from other tandem repeats, yet it is also the main obstacle in its investigation because these arrays are in principle difficult to assemble. Here we explore an alternative, assembly-free approach that utilizes ultra-long Oxford Nanopore reads to infer the length distribution of satellite repeat arrays, their association with other repeats and the prevailing sequence periodicities. Using the satellite DNA-rich legume plant *Lathyrus sativus* as a model, we demonstrated this approach by analyzing 11 major satellite repeats using a set of nanopore reads ranging from 30 to over 200 kb in length and representing 0.73× genome coverage. We found surprising differences between the analyzed repeats because only two of them were predominantly organized in long arrays typical for satellite DNA. The remaining nine satellites were found to be derived from short tandem arrays located within LTR-retrotransposons that occasionally expanded in length. While the corresponding LTR-retrotransposons were dispersed across the genome, this array expansion occurred mainly in the primary constrictions of the *L. sativus* chromosomes, which suggests that these genome regions are favourable for satellite DNA accumulation.

Keywords: satellite DNA, *Lathyrus sativus*, long-range organization, sequence evolution, nanopore sequencing, centromeres, heterochromatin, fluorescence *in situ* hybridization (FISH), technical advance.

INTRODUCTION

Satellite DNA (satDNA) is a class of highly repeated genomic sequences characterized by its occurrence in long arrays of almost identical, tandemly arranged units called monomers. It is ubiquitous in animal and plant genomes, where it can make up to 36% or 18 Gbp/1C of nuclear DNA (Ambrožová *et al*., 2010). The monomer sequences are typically hundreds of nucleotides long, although they can be as short as simple sequence repeats (<10 bp) (Heckmann *et al*., 2013) or reach over 5 kb (Gong *et al*., 2012). Thus, satDNA is best distinguished from other tandem repeats like micro- or minisatellites by forming much longer arrays (tens of kilobases up to megabases) that often constitute blocks of chromatin with specific structural and epigenetic properties (Garrido-Ramos, 2017). This genomic organization and skewed base composition have played a crucial role in satDNA discovery in the form of additional (satellite) bands observed in density gradient centrifugation analyses of genomic DNA (Kit, 1961). Thanks to a number of studies in diverse groups of organisms, the initial view of satellite DNA as genomic 'junk' has gradually shifted to an appreciation of its roles in chromosome organization, replication and segregation, gene expression, disease phenotypes and reproductive isolation between species (reviewed in Plohl *et al*., 2014; Garrido-Ramos, 2015, 2017; Hartley *et al*., 2019). Despite this progress, there are still serious limitations in our understanding of the biology of satDNA, especially with respect to the molecular mechanisms underlying its evolution and turnover in the genome.

Although the presence of satDNA is a general feature of eukaryotic genomes, its sequence composition is highly variable. Most satellite repeat families are specific to a

© 2019 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd
This is an open access article under the terms of the Creative Commons Attribution License,
which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

single genus or even a species (Macas *et al.*, 2002), which makes satDNA the most dynamic component of the genome. A theoretical framework for understanding satDNA evolution was laid using computer simulations (reviewed in Elder and Turner, 1995). For example, the computer models demonstrated the emergence of tandem repeats from random non-repetitive sequences by a joint action of unequal recombination and mutation (Smith, 1976), predicted satDNA accumulation in genome regions with suppressed meiotic recombination (Stephan, 1986) and evaluated possible impacts of natural selection (Stephan and Cho, 1994). It was also revealed that recombination-based processes alone cannot account for the persistence of satDNA in the genome, which implied that additional amplification mechanisms need to be involved (Walsh, 1987). These models are of great value because, in addition to predicting conditions that can lead to satDNA origin, they provide testable predictions regarding tandem repeat homogenization patterns, the emergence of higher order repeats (HORs) and the gradual elimination of satDNA from the genome. However, their utilization and further development have been hampered by the lack of genome sequencing data revealing the long-range organization and sequence variation within satDNA arrays that were needed to test their predictions.

A parallel line of research has focused on elucidating satDNA evolution using molecular and cytogenetic methods. These studies confirmed that satellite repeats can be generated by tandem amplification of various genomic sequences, for example, parts of dispersed repeats within potato centromeres (Gong *et al.*, 2012) or a single-copy intronic sequence in primates (Valeri *et al.*, 2018). An additional putative mechanism of satellite repeat origin was revealed in DNA replication studies, which showed that repair of static replication forks leads to the generation of tandem repeat arrays (Kuzminov, 2016). SatDNA can also originate by expansion of existing short tandem repeat arrays present within rDNA spacers (Macas *et al.*, 2003) and in hypervariable regions of LTR retrotransposons (Macas *et al.*, 2009). Moreover, there may be additional links between the structure or transpositional activity of mobile elements and satDNA evolution (Meštrović *et al.*, 2015; McGurk and Barbash, 2018). Once amplified, satellite repeats usually undergo a fast sequence homogenization within each family, resulting in high similarities of monomers within and between different arrays. This process is termed concerted evolution (Elder and Turner, 1995) and is supposed to employ various molecular mechanisms, such as gene conversion (Schindelhauer and Schwarz, 2002), segmental duplication (Ma and Jackson, 2006) and rolling-circle amplification of extrachromosomal circular DNA (Cohen *et al.*, 2005; Navrátilová *et al.*, 2008). However, little evidence has been gathered thus far to evaluate real importance of these mechanisms for satDNA evolution. Since each of these mechanisms leaves specific molecular footprints, this question can be tackled by searching for these patterns within satellite sequences. However, obtaining such sequence data from a wide range of species has long been a limiting factor in satDNA investigation.

The introduction of next generation sequencing (NGS) technologies (Metzker, 2009) marked a new era in genome research, including the characterization of repetitive DNA (Weiss-Schneeweiss *et al.*, 2015). Although the adoption of short-read technologies like Illumina resulted in a boom of genome assembly projects, such assemblies are of limited use for satDNA investigation because they exclude repeat-rich regions that cannot be efficiently resolved with the short reads (Peona *et al.*, 2018). On the other hand, the short-read data are successfully utilized by bioinformatic pipelines specifically tailored to the identification of satellite repeats employing assembly-free algorithms (Novák *et al.*, 2010; Ruiz-Ruano *et al.*, 2016; Novák *et al.*, 2017). Although these approaches proved to be efficient in satDNA identification and revealed a surprising diversity of satellite repeat families in some plant and animal species (Macas *et al.*, 2015; Ruiz-Ruano *et al.*, 2016; Ávila Robledillo *et al.*, 2018), they, in principle, could not provide much insight into their large-scale arrangement in the genome. In this respect, the real breakthrough was recently made by the so-called long-read sequencing technologies that include the Pacific Biosciences and Oxford Nanopore platforms. Especially the latter has, due to its principle of reading the sequence directly from a native DNA strand during its passage through a molecular pore, a great potential to generate "ultra-long" reads reaching up to one megabase (van Dijk *et al.*, 2018). Different strategies utilizing such long reads for satDNA investigation can be envisioned. First, they can be combined with other genome sequencing and mapping data to generate hybrid assemblies in which satellite arrays are faithfully represented and then analyzed. This approach has already been successfully used for assembling satellite-rich centromere of the human chromosome Y (Jain *et al.*, 2018) and for analyzing homogenization patterns of satellites in *Drosophila melanogaster* (Khost *et al.*, 2017). Alternatively, it should be possible to infer various features of satellite repeats by analyzing repeat arrays or their parts present in individual nanopore reads. Since only a few attempts have been made to adopt this strategy (Cechova and Harris, 2018) it has yet to be fully explored, which is the subject of the present study.

In this work, we aimed to characterize the basic properties of satellite repeat arrays in a genome-wide manner by employing bioinformatic analyses of long nanopore reads. As the model for this study, we selected the grass pea (*Lathyrus sativus* L.), a legume plant with a relatively large

© 2019 The Authors.

genome (6.52 Gbp/C) and a small number of chromosomes (2*n* = 14) which are amenable to cytogenetic experiments. The chromosomes have extended primary constrictions with multiple domains of centromeric chromatin (meta-polycentric chromosomes) (Neumann *et al.*, 2015; Neumann *et al.*, 2016) and well distinguishable heterochromatin bands indicative of the presence of satellite DNA. Indeed, repetitive DNA characterization from low-pass genome sequencing data revealed that the *L. sativus* genome is exceptionally rich in tandem repeats that include 23 putative satDNA families, which combined represent 10.7% of the genome (Macas *et al.*, 2015). Focusing on the fraction of the most abundant repeats, we developed a workflow for their detection in nanopore reads and subsequent evaluation of the size distributions of their arrays, their sequence homogenization patterns and their interspersion with other repetitive sequences. This work revealed surprising differences of the array properties between the analyzed repeats, which allowed their classification into two groups that differed in origin and amplification patterns in the genome.

RESULTS

For the present study, we chose a set of 16 putative satellites with estimated genome proportions exceeding a threshold of 0.1% and reaching up to 2.6% of the *L. sativus* genome (Table 1). These sequences were selected as the most abundant from a broader set of 23 tandem repeats that were previously identified in *L. sativus* using graph-based clustering of Illumina reads (Macas *et al.*, 2015). The clusters selected from this study were further analyzed using the TAREAN pipeline (Novák *et al.*, 2017), which confirmed their annotation as satellite repeats and reconstructed consensus sequences of their monomers (Data S1). The monomers were 32–660 bp long and varied in their AT/GC content (46.3–76.6% AT). Mutual sequence similarities were detected between some of the monomers, which suggested that they represented variants (sub-families) of the same repeat family (Figure S1). These included three variants of the satellite families FabTR-51 and FabTR-53 and two variants of FabTR-52 (Table 1). Except for the FabTR-52 sequences, which were found to be up to 96% identical to the repeat pLsat described by (Ceccarelli *et al.*, 2010), none of the satellites showed similarities to sequences in public sequence databases. We assembled a reference database of consensus sequences and additional sequence variants of all selected satellite repeats to be used for similarity-based detection of these sequences in the nanopore reads. The reference sequences were put into the same orientation to allow for evaluation of the orientation of the arrays in the nanopore reads.

We conducted two sequencing runs on the Oxford Nanopore MinION device utilizing independent libraries prepared from partially fragmented genomic DNA using a 1D ligation sequencing kit (SQK-LSK109). The two runs resulted in similar size distributions of the reads (Figure S2, panel a) and combined produced a total of 8.96 Gbp of raw read data. Following quality filtering, the reads shorter than 30 kb were discarded because we aimed to analyze only a fraction of the longest reads. The remaining 78 563 reads ranging from 30 to 348 kb in length (N50 = 67 kb)

Table 1 Characteristics of the investigated satellite repeats

Satellite family Subfamily	Monomer [bp]	AT [%]	Genomic abundance [%]	 [Mbp/1C]	FISH probe
FabTR-2	49	71.4	1.700	110.8	LASm3H1
FabTR-51			3.101	202.2	
FabTR-51-LAS-A	80	46.3	2.500	163.0	LASm1H1
FabTR-51-LAS-B	79	51.9	0.560	36.5	LasTR6_H1
FabTR-51-LAS-C	118	50.0	0.041	2.7	
FabTR-52			2.019	131.6	
FabTR-52-LAS-A	55	47.3	2.000	130.4	LASm2H1
FabTR-52-LAS-B	32	50.0	0.019	1.2	
FabTR-53			2.600	169.5	c1644 + c1645
FabTR-53-LAS-A	660	76.6	n.d.		
FabTR-53-LAS-B	368	76.4	n.d.		
FabTR-53-LAS-C	565	75.9	n.d.		
FabTR-54	104	51.0	0.840	54.8	LasTR5_H1
FabTR-55	78	55.1	0.480	31.3	LasTR7_H1
FabTR-56	46	60.9	0.250	16.3	LasTR8_H1
FabTR-57	61	65.6	0.130	8.5	LasTR9_H1
FabTR-58	86	59.3	0.140	9.1	LasTR10_H1
FabTR-59	131	49.6	0.110	7.2	LasTR11_H1
FabTR-60	86	52.3	0.110	7.2	LasTR12_H1

© 2019 The Authors.

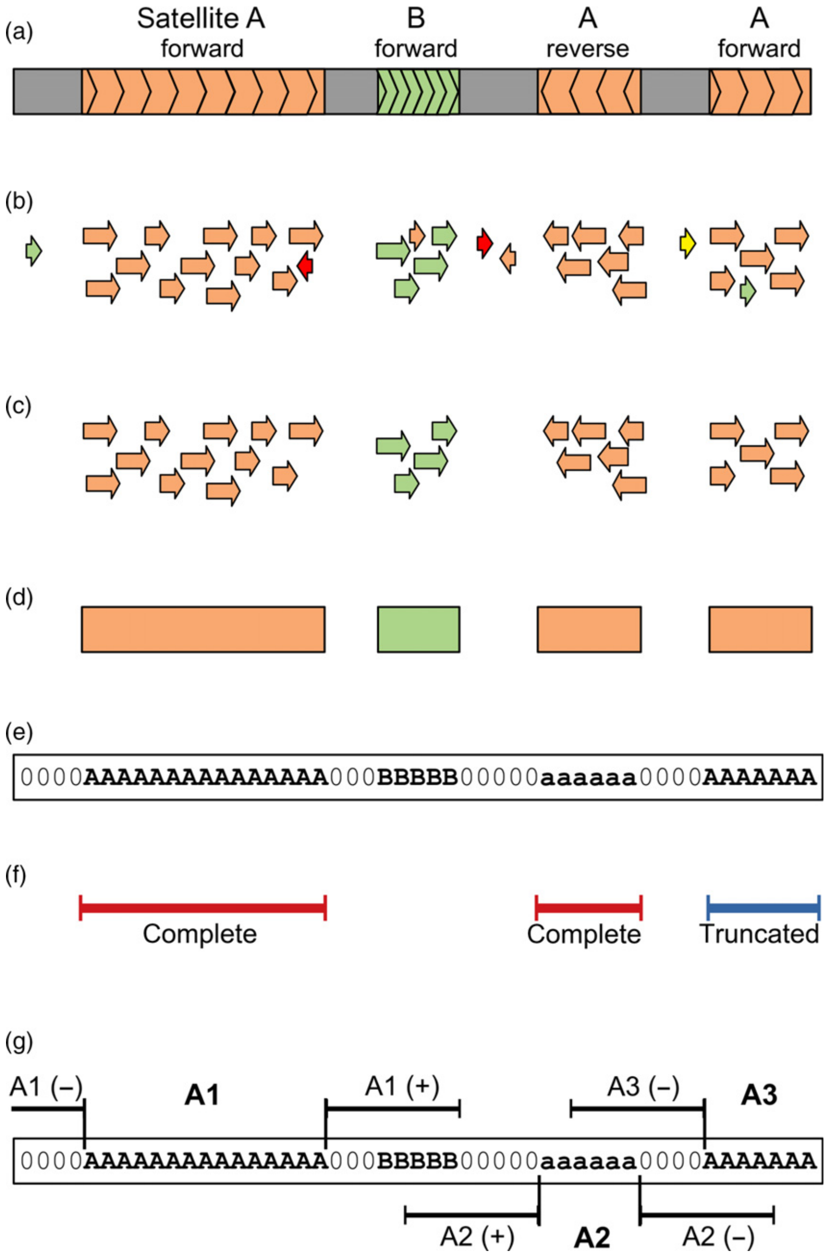


Figure 1. Schematic representation of the analysis strategy. (a) Nanopore read (grey bar) containing arrays of satellites A (orange) and B (green). The orientations of the arrays with respect to sequences in the reference database are indicated. (b) LASTZ search against the reference database results in similarity hits (displayed as arrows showing their orientation, with colours distinguishing satellite sequences) that are quality-filtered to remove non-specific hits (c). The filtered hits are used to identify the satellite arrays as regions of specified minimal length that are covered by overlapping hits to the same repeat (d). The positions of these regions are recorded in the form of coded reads where the sequences are replaced by satellite codes and array orientations are distinguished using uppercase and lowercase characters (e). The coded reads are then used for various downstream analyses. (f) Array lengths are extracted and analyzed regardless of orientation of the arrays but while distinguishing the complete and truncated arrays (here it is shown for satellite A). (g) Analysis of the sequences adjacent to the satellite arrays includes 10 kb regions upstream (−) and downstream (+) of the array. This analysis is performed with respect to the array orientation (compare the positions of upstream and downstream regions for arrays in forward (A1, A3) versus reverse orientation (A2)).

provided a total of 4.78 Gbp of sequence data, which corresponded to 0.73× coverage of the *L. sativus* genome.

Detection of the satellite arrays in nanopore reads revealed repeats with contrasting array length distributions

The strategy for analyzing the length distribution of the satellite repeat arrays in the genome using nanopore reads is schematically depicted in Figure 1. The satellite arrays in the nanopore reads were identified by similarity searches against the reference database employing the LASTZ program (Harris, 2007). Using a set of nanopore reads with known repeat compositions, we first optimized the LASTZ parameters towards high sensitivity and specificity. Under these conditions, the satDNA arrays within nanopore reads typically produced a series of short overlapping similarity hits that were filtered and parsed with custom scripts to detect the contiguous repeat regions longer than 300 bp. Then, the positions and orientations of the detected repeats were recorded, while distinguishing whether they were complete or truncated by the read end. In the latter case, the recorded array length was actually an underestimation of the real size.

When the above analyses were applied to the whole set of nanopore reads, the detected array lengths were pooled for each satellite repeat, and their distributions were

© 2019 The Authors.

visualized as weighted histograms with a bin size of 5 kb, distinguishing complete and truncated satellite arrays (Figure 2). This type of visualization accounts for the total lengths of the satellite sequences that occur in the genome as arrays of the lengths specified by the bins. Alternatively, the array size distributions were also plotted as histograms of their counts (Figure S3). As a control for the satellite repeats, we also analyzed the length distribution of 45S rDNA sequences, which typically form long arrays of tandemly repeated units (Copenhaver and Pikaard, 1996). Indeed, the plots revealed that most of the 45S rDNA repeats were detected as long arrays ranging up to >120 kb. A similar pattern was expected for the satellite repeats; however, it was found for only two of them, FabTR-2 and FabTR-53 (Figure 2a). Both of these repeats were almost exclusively present as long arrays that extended beyond the lengths of most of the reads. To verify these results, we analyzed randomly selected reads using sequence self-similarity dot-plots, which confirmed that most of the arrays spanned entire reads or were truncated at only one of their ends (Figure S4a,e). However, all nine remaining satellites generated very different array length distribution profiles that consisted of relatively large numbers of short (<5 kb) arrays and comparatively fewer longer arrays (Figure 2b; Figure S3b). The proportions of these two size classes differed between the satellites, for example, while for FabTR-58, most of the arrays (98%) were short and only a few were expanded over 5 kb, FabTR-51 displayed a gradient of sizes from <5 to 174 kb. To check whether these profiles could have partially been due to differences in the lengths of the reads containing these satellites, we also analyzed their size distributions. However, the read length distributions were similar between the different repeats, and there was no bias towards shorter read lengths (Figure S2, panel b). Thus, we concluded that nine of 11 analyzed satellites occurred in the *L. sativus* genome predominantly as short tandem arrays, and only a fraction of them expanded to form long arrays typical of satellite DNA. This conclusion was also confirmed by the dot-plot analyses of the individual reads, which revealed reads carrying short or intermediate-sized arrays and a few expanded ones (Figure S4i–n).

Analysis of genomic sequences adjacent to the satellite arrays identified a group of satellites that originated from LTR-retrotransposons

Next, we were interested in whether the investigated satellites were frequently associated in the genome with each other or with other types of repetitive DNA. Using a reference database for the different lineages of LTR-retrotransposons, DNA transposons, rDNA and telomeric repeats compiled from *L. sativus* repeated sequences identified in our previous study (Macas *et al.*, 2015), we detected these repeats in the nanopore reads using LASTZ along with the analyzed satellites. Their occurrences were then analyzed within 10-kb regions directly adjacent to each satellite repeat array, and the frequencies at which they were associated with individual satDNA families were plotted with respect to the oriented repeat arrays (Figure 3). When performed for the control 45S rDNA, this analysis revealed that they were mostly surrounded by arrays of the same sequences oriented in the same direction. This pattern emerged due to short interruptions of otherwise longer arrays. Similar results were found for FabTR-2 and FabTR-53 (Figure 3a) which also formed long arrays in the genome. Notably, the adjacent regions could be analyzed for only 33 and 35% of the FabTR-2 and FabTR-53 arrays, respectively, because these repeats mostly spanned entire reads. Substantially different profiles were obtained for the remaining nine satellites (Figure 3b), revealing their frequent association with Ogre LTR-retrotransposons. No other repeats were detected at similar frequencies, except for unclassified LTR-retrotransposons that probably represented less-conserved Ogre sequences. At a much smaller frequency (~0.1), the FabTR-54 repeat was found to be adjacent to the FabTR-56 satellite arrays. Based on its position and size in relation to FabTR-56, the detected pattern corresponded to short FabTR-54 arrays attached to FabTR-56 in a direction-specific manner. Inspection of the individual reads confirmed that short arrays of these satellites occurred together in a part of the reads (Figure S4l). A peculiar pattern was revealed for FabTR-58 that consisted of a series of peaks that suggested interlacing FabTR-58 and Ogre sequences at fixed intervals (Figure 3). This pattern was found to be due to occurrence of complex arrays consisting of multiple short arrays of FabTR-58 arranged in the same orientation and embedded into Ogre sequences (Figure S4q). Upon closer inspection, this organization was found in numerous reads.

Ogre elements represent a distinct phylogenetic lineage of Ty3/gypsy LTR-retrotransposons (Neumann *et al.*, 2019) that were amplified to high copy numbers in some plant species including *L. sativus*. Because they comprise 45% of the *L. sativus* genome (Macas *et al.*, 2015), the frequent association of Ogres with short array satellites could simply be due to their random interspersion. However, we noticed from the structural analysis of the reads that these short arrays were often surrounded by two direct repeats, which is a feature typical of LTR-retrotransposons. This finding could mean that the arrays are actually embedded within the Ogre elements and were not only frequently adjacent to them by chance. To test this hypothesis, we performed an additional analysis of the array neighbourhoods, but this time, we specifically detected parts of the Ogre sequences coding for the retroelement protein domains GAG, protease (PROT), reverse transcriptase (RT), RNase H (RH), archeal RNase H (aRH) and integrase (INT). If the association of Ogre sequences with the satellite

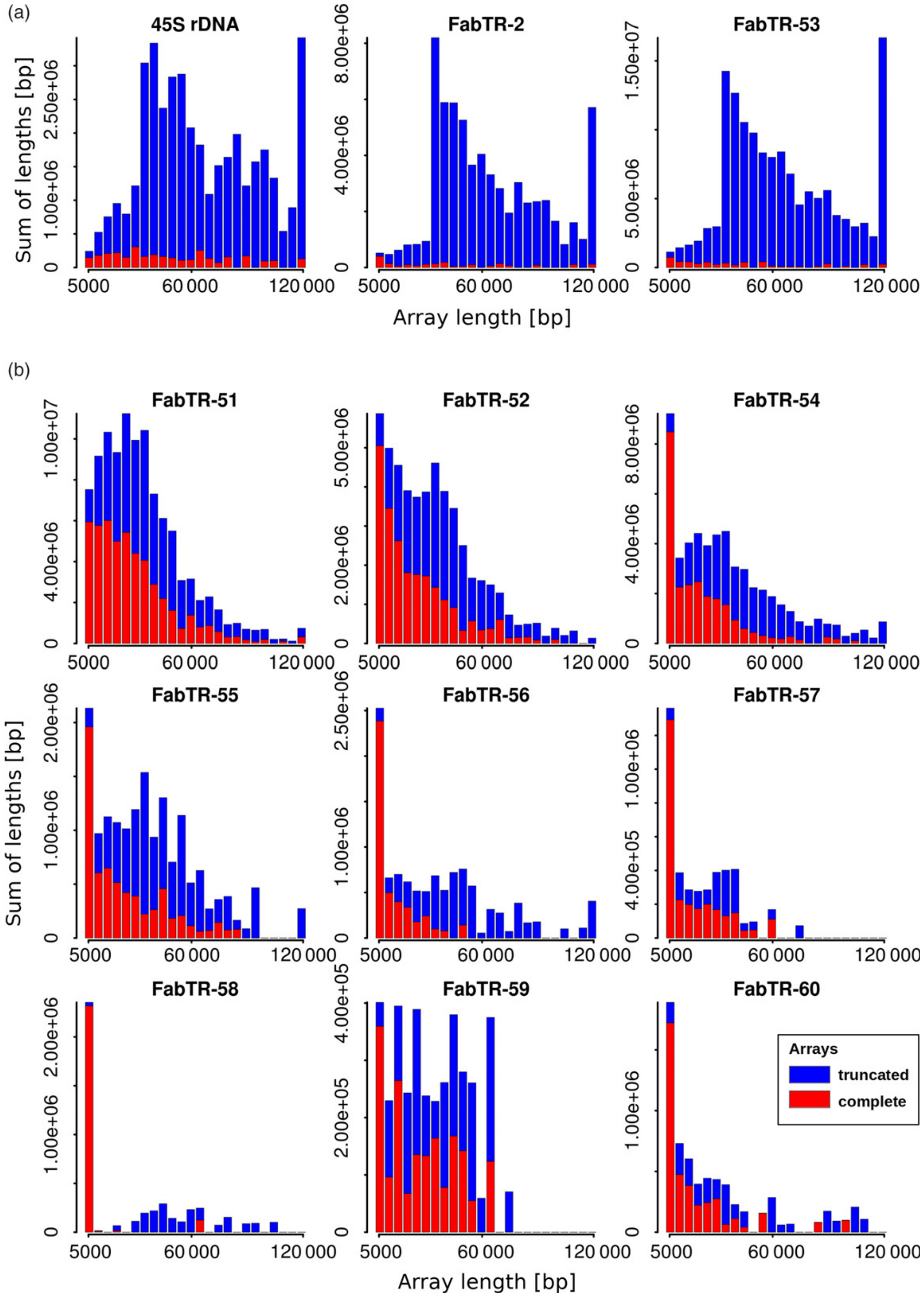


Figure 2. Length distributions of the satellite repeat arrays. The lengths of the arrays detected in the nanopore reads are displayed as weighted histograms with a bin size of 5 kb; the last bin includes all arrays longer than 120 kb. The arrays that were completely embedded within the reads (red bars) are distinguished from those that were truncated by their positions at the ends of the reads (blue bars). Due to the array truncation, the latter values are actually underestimations of the real lengths of the corresponding genomic arrays and should be considered as lower bounds of the respective array lengths. Tandem repeats forming long arrays are shown in panel (a), while the remaining repeats forming predominantly short arrays are in panel (b).

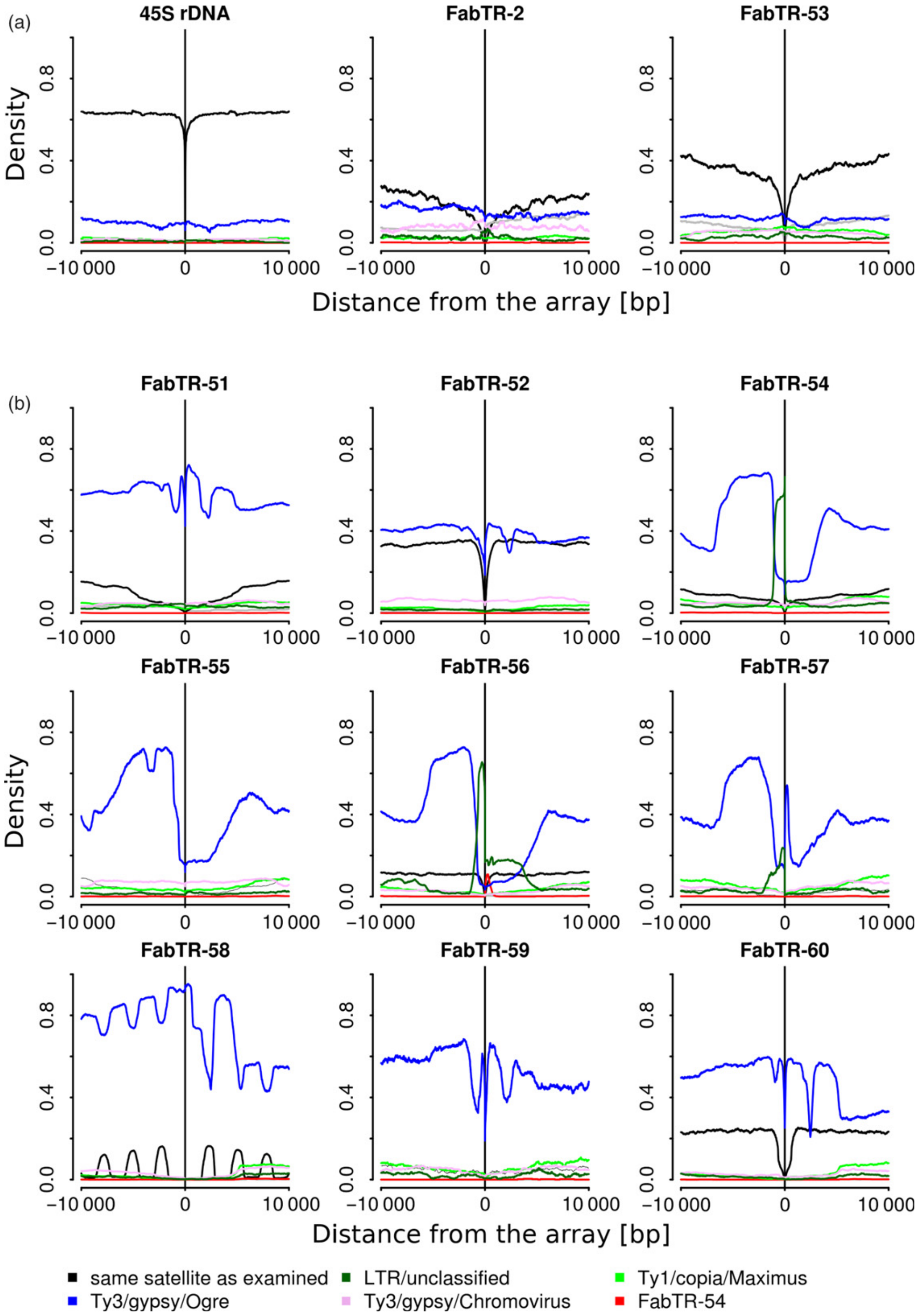


Figure 3. Sequence composition of the genomic regions adjacent to the satellite repeat arrays. The plots show the proportions of repetitive sequences identified within 10 kb regions upstream (positions −1 to −10 000) and downstream (1 to 10 000) of the arrays of individual satellites (the array positions are marked by vertical lines, and the plots are related to the forward-oriented arrays). Only the repeats detected in proportions exceeding 0.05 are plotted (coloured lines). The black lines represent the same satellite as examined. Tandem repeats forming long arrays are shown in panel (a), while the remaining repeats forming predominantly short arrays are in panel (b).

© 2019 The Authors.

arrays was random, these domains would be detected at various distances and orientations with respect to the arrays. In contrast, finding them in a fixed arrangement would confirm that the tandem arrays were in fact parts of the Ogre elements and occurred there in specific positions. As evident from Figure 4(a), that latter explanation was confirmed for all nine satellites. We found that their arrays occurred downstream of the Ogre *gag-pol* region including the LTR-retrotransposon protein coding domains in the expected order and orientation (see the element structure in Figure 4b). In two cases (FabTR-54 and 57), some protein domains were not detected, and major peaks corresponded to the GAG domain which was relatively close to the tandem arrays. These patterns were explained by the frequent occurrence of these tandem arrays in non-autonomous elements lacking their *pol* regions due to large deletions. In approximately half of the satellites (*e.g.*, FabTR-51 and 52), we detected additional smaller peaks corresponding to the domains in both orientations located approximately 7–10 kb from the arrays. Further investigation revealed that these peaks represented Ogre elements that were inserted into the expanded arrays of corresponding satellites (Figure S4k). Consequently, they were detected only in satellites such as FabTR-51 and 52 in which the proportions of expanded arrays were relatively large and not FabTR-58 in which the expanded arrays were almost absent.

The finding that the nine satellite sequences are also present as short tandem arrays within Ogre elements can be explained by either of the two principally different scenarios: (1) the long satellite arrays originated by expansion of tandem sequences originally present only within Ogre elements, or (2) the long satellite arrays are ancestral and unrelated to Ogre sequences but their fragments were captured by some element copies and subsequently dispersed in the genome along with the element amplification. Although the array size distributions (Figure 2b; Figure S3b) suggest gradual expansion of the arrays from their short precursors and thus support the first scenario, we set to further investigate this question employing an alternative, phylogeny-based approach. Using the repeat sequencing and annotation data generated previously for a group of *Fabeae* species (Macas *et al.*, 2015), we tested the presence of these satellite sequences in two related *Lathyrus* species, *L. vernus* and *L. latifolius*. No similarity hits to repeat clusters annotated as satellite repeats were detected, thus revealing that these sequences occur as amplified satellite DNA only in *L. sativus*. However, significant similarity hits to clusters annotated as Ogre elements or putative LTR-retrotransposons were found for three of the tested repeats, FabTR-54, FabTR-55 and FabTR-57 in both species (Table S1). Detailed inspection of these clusters confirmed their annotation and revealed that all of them also included tandem subrepeats, some of which matched the query sequences. Thus, at least for these three repeats it was demonstrated that while the elements carrying their short arrays occur in all three *Lathyrus* species, the corresponding satellite repeats were detected in *L. sativus* only, thus supporting the model of satellite DNA evolution from the tandem subrepeats within Ogre sequences.

Satellites with mostly expanded arrays show higher variation in their sequence periodicities

The identification of large numbers of satellite arrays in the nanopore reads provided sequence data for investigating the conservation of monomer lengths and the eventual occurrence of additional monomer length variants and HORs. To this purpose we designed a computational pipeline that extracted all satellite arrays longer than 30 kb and subjected them to a periodicity analysis using the fast Fourier transform algorithm (Venables and Ripley, 2002). The analysis revealed the prevailing monomer sizes and eventual additional periodicities in the tandem repeat arrays as periodicity spectra containing peaks at positions corresponding to the lengths of the tandemly repeated units. These periodicity spectra were averaged for all arrays of the same satellite (Figure 5) or plotted separately for the individual arrays to explore the periodicity variations (Figure S5). As an alternative approach, we also visualized the array periodicities using nucleotide autocorrelation functions (Herzel *et al.*, 1999; Macas *et al.*, 2006). In selected cases, we verified the periodicity patterns within arrays using dot-plot analyses (Figure S4b–d and f,h).

As expected, the periodicity spectra of all satellites contained peaks corresponding to their monomer lengths (Figure 5 and Table 1). In the nine Ogre-derived satellite repeats, the monomer periods were the longest detected. There were only a few additional peaks detected with shorter periods that corresponded to higher harmonics (see Experimental procedures) or possibly reflected short subrepeats or underlying single-base periodicities. In contrast, FabTR-2 and FabTR-53 repeats, which occur in the genome as the expanded arrays, displayed more periodicity variations. Various HORs that probably originated from multimers of the 49 bp consensus were detected in the FabTR-2 arrays. Closer examination of the individual arrays revealed that the multiple peaks evident in the averaged periodicity spectrum (Figure 5) originated as combinations of several simpler HOR patterns that differed between individual satellite arrays (Figure S5). In FabTR-53, the HORs were not detected, but a number of shorter periodicities were revealed, which suggests that the current monomers of 660, 368 and 565 bp (sub-families A, B and C, respectively) actually originated as HORs of shorter units that are represented by the peaks on the left from the monomer peaks (Figure 5). An additional analysis using autocorrelation functions generally agreed with the fast Fourier

© 2019 The Authors.

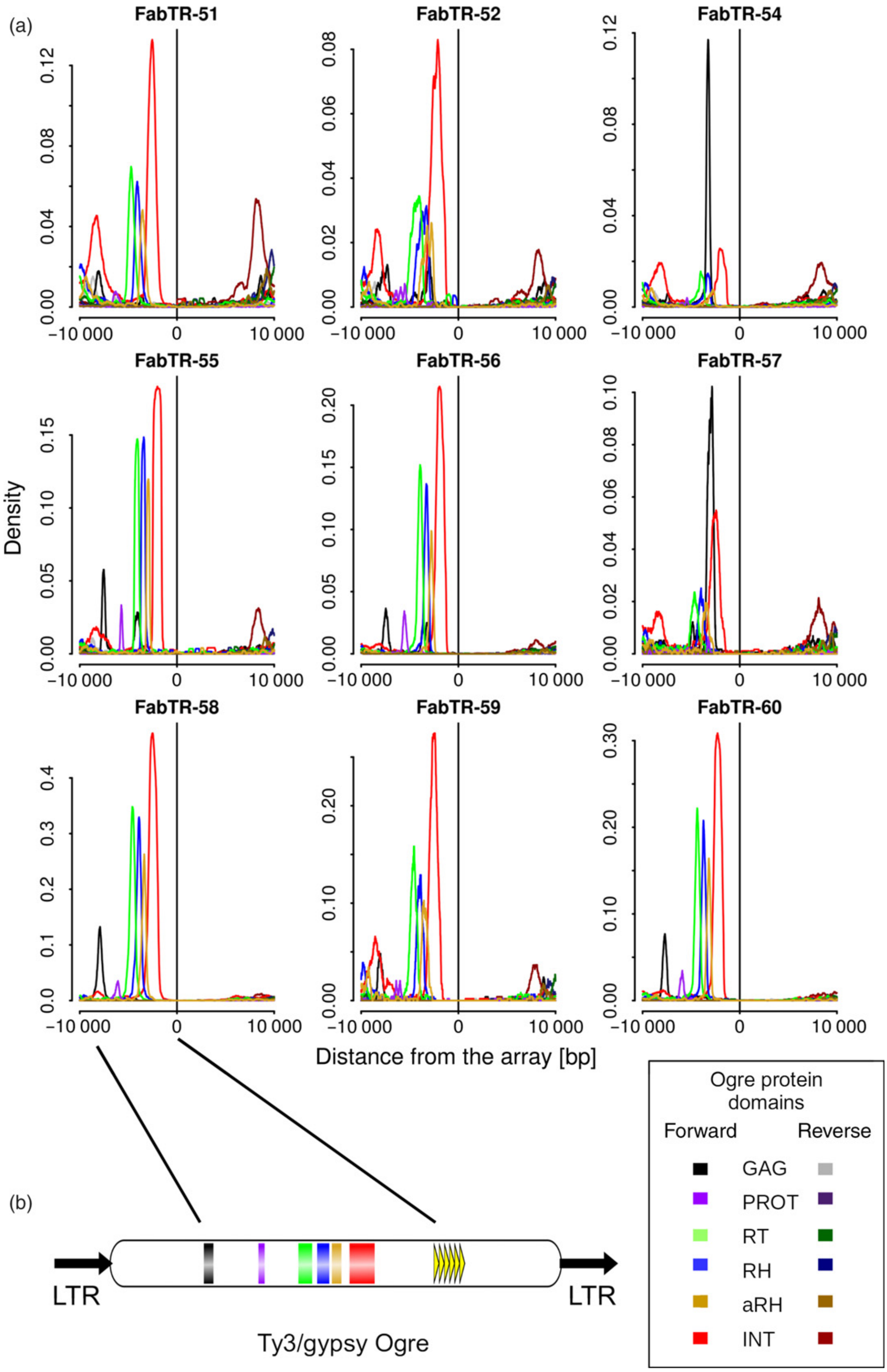


Figure 4. Detection of the Ogre sequences coding for the retrotransposon conserved protein domains in the genomic regions adjacent to the satellite repeat arrays. (a) The plots show the proportions of similarity hits from the individual domains and their orientation with respect to the forward-oriented satellite arrays. (b) A schematic representation of the Ogre element with the positions of the protein domains and short tandem repeats downstream of the coding region.

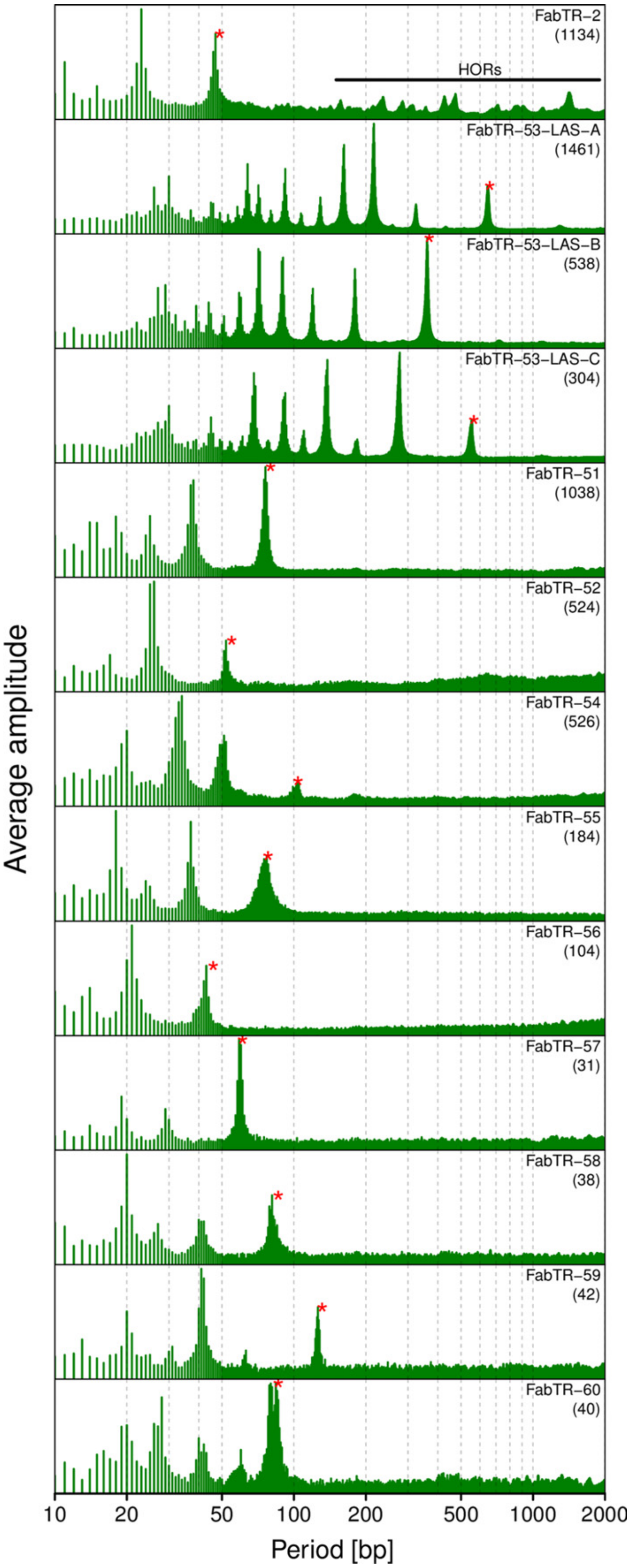


Figure 5. Periodicity spectra revealed by the fast Fourier transform analysis of the satellite repeat arrays. Each spectrum is an average of the spectra calculated for the individual arrays longer than 30 kb of the same satellite family or subfamily. The numbers of arrays used for the calculations are in parentheses. The peaks corresponding to the monomer lengths listed in Table 1 are marked with red asterisks. The peaks in the FabTR-2 spectrum corresponding to higher-order repeats are indicated by the horizontal line.

© 2019 The Authors.

transform approach and confirmed the high variabilities in FabTR-2 and FabTR-53 (Figure S5).

Array expansion of the retrotransposon-derived satellites occurred preferentially in the pericentromeric regions of *L. sativus* chromosomes

To complement the analysis of satellite arrays with the information about their genomic distribution, we performed their detection on metaphase chromosomes using fluorescence *in situ* hybridization (FISH) (Figure 6). Labelled oligonucleotides corresponding to the most conserved parts of the monomer sequences were used as hybridization probes in all cases except for FabTR-53 for which a mix of two cloned probes was used instead due to its relatively long monomers (Table 1 and Data S2). Although each satellite probe generated a different labelling pattern, most of them were located within the primary constrictions. The exception was FabTR-53, which produced strong hybridization signals that overlapped with most of the subtelomeric heterochromatin bands (Figure 6a). The other distinct pattern was revealed for FabTR-2, which produced a series of dots along the periphery of the primary constrictions on all chromosomes (Figure 6b). This pattern was identical to that obtained using an antibody to centromeric histone variant CenH3 (Neumann *et al.*, 2015; Neumann *et al.*, 2016), which suggests that FabTR-2 is the centromeric satellite. The remaining nine probes corresponding to Ogre-derived satellites mostly produced bands at various parts of primary constrictions (Figure 6c–f; Figure S6). For example, the bands of FabTR-54 occurred within or close to the primary constrictions of all chromosomes and produced a labelling pattern which, together with the chromosome morphology, allowed us to distinguish all chromosome types within the *L. sativus* karyotype (Figure 6c). A peculiar pattern was generated by the FabTR-51-LAS-A subfamily probe, which painted whole primary constrictions of one pair of chromosomes (chromosome 1, Figure 6d); a similar pattern was produced by the FabTR-52-LAS-A probe, but it labelled the entire primary constrictions of a different pair (chromosome 7, Figure 6e).

Although the FISH signals of the Ogre-derived satellites were supposed to originate from their expanded and sequence-homogenized arrays, we had to consider the possibility that the probes had also cross-hybridized to the short repeat arrays within the elements; therefore these FISH patterns may have reflected the genome distribution of Ogre elements. Thus, we investigated the Ogre distribution in the *L. sativus* genome using a probe designed from the major sequence variant of the integrase coding domain of the elements carrying the satellite repeats (see the element scheme in Figure 4b). The probe produced signals dispersed along the whole chromosomes that differed from the locations of the bands in the primary constrictions revealed by the satellite repeat probes (Figure 6g–i). Thus, these results confirmed that, while the Ogre elements carrying short tandem repeat arrays were dispersed throughout the genome, these arrays expanded and gave rise to long satellite arrays only within the primary constrictions.

DISCUSSION

In this work, we demonstrated that the detection and analysis of satellite repeat arrays in the bulk of individual nanopore reads is an efficient method to characterize satellite DNA properties in a genome-wide manner. This is an addition to an emerging toolbox of approaches utilizing long sequence reads for investigating satellite DNA in complex eukaryotic genomes. Currently, these approaches have primarily been based on generating improved assemblies of satellite-rich regions and their subsequent analyses (Weissensteiner *et al.*, 2017; Jain *et al.*, 2018). Alternatively, satellite array length variation was analyzed using the long reads aligned to the reference genome (Mitsuhashi *et al.*, 2019) or by detecting a single specific satellite locus in the reads (Roeck *et al.*, 2018). Compared to these approaches, our strategy does not distinguish individual satDNA arrays in the genome. Instead, our approach applies statistics to partial information gathered from individual reads to infer the general properties of the investigated repeats. As such, this approach can analyze any number of different satellite repeats simultaneously and without the need for a reference genome. However, the inability to specifically address individual repeat loci in the genome may be considered a limitation of our approach. For example, we could not precisely measure the sizes of the arrays that were longer than the analyzed reads and instead provided lower bounds of their lengths. On the other hand, we could reliably distinguish tandem repeats that occurred in the genome predominantly in the form of short arrays from those forming only long contiguous arrays and various intermediate states between these extremes. Additionally, we could analyze the internal arrangements of the identified arrays and characterized the sequences that frequently surrounded the arrays in the genome. This analysis was achieved with a sequencing coverage that was substantially lower compared with that needed for genome assembly. Thus, this approach could be of particular use when analyzing very large genomes, genomes of multiple species in parallel or simply whenever sequencing resources are limited. However, it could be valuable even for the genome assembly projects as it provides information that is complementary to that obtained from the assembly-based methods.

We found that only two of the 11 most abundant satellite repeats occurred in the genome exclusively as long tandem arrays typical of satellite DNA. Both occupied specific genome regions, FabTR-2 was associated with centromeric chromatin, and FabTR-53 made up subtelomeric

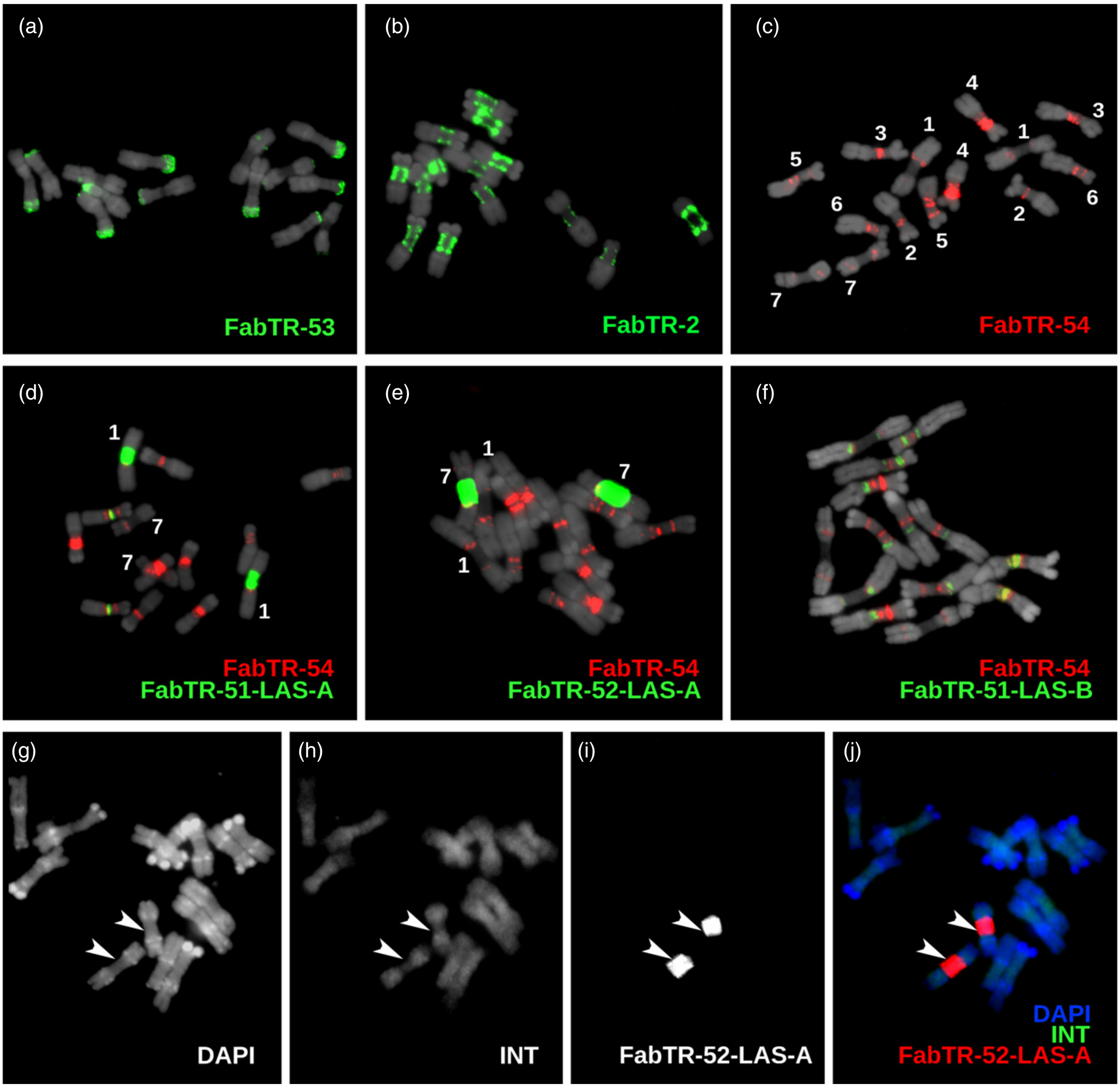


Figure 6. Distribution of the satellite repeats on the metaphase chromosomes of *Lathyrus sativus* ($2n = 14$). (a–f) The satellites were visualized using multi-colour FISH, with individual probes labelled as indicated by the colour-coded descriptions. The chromosomes counterstained with DAPI are shown in grey. The numbers in panel (c) correspond to the individual chromosomes that were distinguished using the hybridization patterns of the FabTR-54 sequences. This satellite was then used for chromosome discrimination in combination with other probes. (g–i) Simultaneous detection of the Ogre integrase probe (INT) and the satellite FabTR-52-LAS-A demonstrates the different distribution of these sequences in the genome. The probe signals and DAPI counterstaining are shown as separate grayscale images (g–i) and a merged image (j). The arrowheads point to the primary constrictions of chromosomes 7.

heterochromatic bands on mitotic chromosomes. Both are also present in other *Fabeae* species (Macas *et al.*, 2015), which suggests that they are phylogenetically older compared with the rest of the investigated *L. sativus* satellites. The other feature common to these satellites was the occurrence of HORs that emerge when a satellite array becomes homogenized by units longer than single monomers. The factors that trigger this shift are not clear, however, it is likely that chromatin structure plays a role in this process by exposing only specific, regularly-spaced parts of the array to the recombination-based homogenization. There are examples of HORs associated with specific types of chromatin (Henikoff *et al.*, 2015) or chromosomal locations (Macas *et al.*, 2006), but data from a wider range of species and diverse satellite repeats are needed to provide a better insight into this phenomenon. The

© 2019 The Authors.

methodology presented here may be instrumental in this task because both the fast Fourier transform and the nucleotide autocorrelation function algorithms employed for the periodicity analyses proved to be accurate and capable of processing large volumes of sequence data provided by nanopore sequencing.

One of the key findings of this study is that the majority of *L. sativus* satellites originated from short tandem repeats present in the 3′ untranslated regions (3′UTRs) of Ogre retrotransposons. These hypervariable regions made of tandem repeats that vary in sequences and lengths of their monomers are common in elements of the Tat lineage of plant LTR-retrotransposons, including Ogres (Macas *et al.*, 2009; Neumann *et al.*, 2019). These tandem repeats were hypothesized to be generated during element replication by illegitimate recombination or abnormal strand transfers between two element copies that are co-packaged in a single virus-like particle (Macas *et al.*, 2009); however, the exact mechanism is yet to be determined. The same authors also documented several cases of satellite repeats that likely originated by the amplification of 3′UTR tandem repeats. In addition to proving this mechanism by detecting various stages of the retroelement array expansions in the nanopore reads, the present work on *L. sativus* also revealed that this phenomenon can be responsible for the emergence of many different satellites within a species. Considering the widespread occurrence and high copy numbers of Tat/Ogre elements in many plant taxa (Neumann *et al.*, 2006; Macas and Neumann, 2007; Kubát *et al.*, 2014; Macas *et al.*, 2015), it can be expected that they play a significant role in satDNA evolution by providing a template for novel satellites that emerge by the expansion of their short tandem repeats. Additionally, similar tandem repeats occur in other types of mobile elements; thus, this phenomenon is possibly even more common. For example, tandem repeats within the DNA transposon *Tetris* have been reported to give rise to a novel satellite repeat in *Drosophila virilis* (Dias *et al.*, 2014).

The other important observation presented here is that the long arrays of all nine Ogre-derived satellites are predominantly located in the primary constrictions of metaphase chromosomes. This implies that these regions are favourable for array expansion, perhaps due to specific features of the associated chromatin. Indeed, it has been shown that extended primary constrictions of *L. sativus* carry a distinct type of chromatin that differs from the chromosome arms by the histone phosphorylation and methylation patterns (Neumann *et al.*, 2016). However, it is not clear how these chromatin features could promote the amplification of satellite DNA. An alternative explanation could be that the expansion of the Ogre-derived tandem arrays occurs randomly at different genomic loci, but the expanded arrays persist better in the constrictions compared with the chromosome arms. Because excision and eventual elimination of tandem repeats from chromosomes is facilitated by their homologous recombination (Navrátilová *et al.*, 2008), this explanation would be supported by the absence of meiotic recombination in the centromeric regions. The regions with suppressed recombination have also been predicted as favourable for satDNA accumulation by computer models (Stephan, 1986). These hypotheses can be tested in the future investigations of properly selected species. For example, the species known to carry chromosome regions with suppressed meiotic recombination located apart from the centromeres would be of particular interest. Such regions occur, for instance, on sex chromosomes (Vyskot and Hobza, 2015), which should allow for assessments of the effects of suppressed recombination without the eventual interference of the centromeric chromatin. In this respect, the spreading of short tandem arrays throughout the genome by mobile elements represents a sort of natural experiment, providing template sequences for satDNA amplification, which in turn, could be used to identify genome and chromatin properties favouring satDNA emergence and persistence in the genome.

EXPERIMENTAL PROCEDURES

DNA isolation and nanopore sequencing

Seeds of *Lathyrus sativus* were purchased from Fratelli Ingegnoli S.p.A. (Milano, Italy, cat. no. 455). High molecular weight (HMW) DNA was extracted from leaf nuclei isolated using a protocol adapted from (Vershinin and Heslop-Harrison, 1998) and (Macas *et al.*, 2007). Five grams of young leaves were frozen in liquid nitrogen, ground to a fine powder and incubated for 5 min in 35 ml of ice-cold H buffer (1× HB, 0.5 M sucrose, 1 mM phenylmethyl-sulphonylfluoride (PMSF), 0.5% (v/v) Triton X-100, 0.1% (v/v) 2-mercaptoethanol). The H buffer was prepared fresh from 10× HB stock (0.1 M Tris–HCl pH 9.4, 0.8 M KCl, 0.1 M EDTA, 40 mM spermidine, 10 mM spermine). The homogenate was filtered through 48 μm nylon mesh, adjusted to 35 ml volume with 1× H buffer, and centrifuged at 200 ***g*** for 15 min at 4°C. The pelleted nuclei were resuspended and centrifuged using the same conditions after placement in 35 ml of H buffer and 15 ml of TC buffer (50 mM Tris–HCl pH 7.5, 75 mM NaCl, 6 mM $MgCl_2$, 0.1 mM $CaCl_2$). The final centrifugation was performed for 5 min only, and the nuclei were resuspended in 2 ml of TC. HMW DNA was extracted from the pelleted nuclei using a modified CTAB protocol (Murray and Thompson, 1980). The suspension of the nuclei was mixed with an equal volume of 2× CTAB buffer (1.4 M NaCl, 100 mM Tris–HCl pH 8.0, 2% CTAB, 20 mM EDTA, 0.5% (w/v) $Na_2S_2O_5$, 2% (v/v) 2-mercaptoethanol) and incubated at 50°C for 30–40 min. The solution was extracted with chloroform: isoamylalcohol (24:1) using MaXtract^{TM} High Density Tubes (Qiagen, Hilden, Germany) and precipitated with a 0.7 volume of isopropanol using a sterile glass rod to collect the DNA. Following two washes in 70% ethanol, the DNA was dissolved in TE and treated with 2 μl of RNase Cocktail^{TM} Enzyme Mix (Thermo Fisher Scientific) for 1 h at 37°C. The DNA integrity was checked by running a 200 ng aliquot on inverted field gel electrophoresis (FIGE Mapper, Bio-Rad, Hercules, CA, USA). Because intact HMW DNA gave poor yields when used

© 2019 The Authors.

with the Oxford Nanopore Ligation Sequencing Kit, the DNA was mildly fragmented by slowly passing the sample through a 0.3 × 12 mm syringe to get a fragment size distribution ranging from ~30 kb to over 100 kb. Finally, the DNA was further purified by mixing the sample with a 0.5 volume of CU and a 0.5 volume of IR solution from the Qiagen DNeasy PowerClean Pro Clean Up Kit (Qiagen, Hilden, Germany), centrifugation for 2 min at 24 000 ***g*** at room temperature and DNA precipitation from the supernatant using a 2.5 volume of 96% ethanol. The DNA was dissolved in 10 mM Tris–HCl pH 8.5 and stored at 4°C.

The sequencing libraries were prepared from 3 μg of the partially fragmented and purified DNA using a Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's protocol. Briefly, the DNA was treated with 2 μl of NEBNext formalin-fixed paraffin-embedded (FFPE) DNA Repair Mix and 2 μl of NEBNext Ultra II End-prep enzyme mix in a 60 μl volume that also included 3.5 μl of FFPE and 3.5 μl of End-prep reaction buffers (New England Biolabs, Ipswisch, MA, USA). The reaction was performed at 20°C for 5 min and 65°C for 5 min. Then, the DNA was purified using a 0.4× volume of AMPure XP beads (Beckman Coulter, Brea, CA, USA); because long DNA fragments caused clumping of the beads and were difficult to detach, the elution was performed with 3 mM Tris–HCl (pH 8.5) and was extended up to 40 min. Subsequent steps including adapter ligation using NEBNext Quick T4 DNA Ligase and the library preparation for the sequencing were performed as recommended. The whole library was loaded onto FLO-MIN106 R9.4 flow cell and sequenced until the number of active pores dropped below 40 (21–24 h). Two sequencing runs were performed, and the acquired sequence data were first analyzed separately to examine eventual variations. However, because the runs generated similar read length profiles and analysis results, the data were combined for the final analysis.

Bioinformatic analysis of the nanopore reads

The raw nanopore reads were basecalled using Oxford Nanopore basecaller Guppy (ver. 2.3.1). Quality filtering of the resulting FastQ reads and their conversion to the FASTA format were performed with BBDuk (part of the BBTools, https://jgi.doe.gov/data-and-tools/bbtools/) run with the parameter maq = 8. Reads shorter than 30 kb were discarded. Unless stated otherwise, all bioinformatic analyses were implemented using custom Python and R scripts and executed on a Linux-based server equipped with 64 GB RAM and 32 CPUs.

Satellite repeat sequences were detected in the nanopore reads by similarity searches against a reference database compiled from contigs assembled from clusters of *L. sativus* Illumina reads in the frame of our previous study (Macas *et al.*, 2015). Additionally, the database included consensus sequences and their most abundant sequence variants calculated from the same Illumina reads using the TAREAN pipeline (Novák *et al.*, 2017) executed with the default parameters and cluster merging option enabled. For each satellite, the reference sequences in the database were placed in the same orientation to allow for the evaluation of the orientations of the satellite arrays in the nanopore reads. The sequence similarities between the reads and the reference database were detected using LASTZ (Harris, 2007). The program parameters were fine-tuned for error-prone nanopore reads using a set of simulated and real reads with known repeat contents while employing visual evaluation of the reported hits using the Integrative Genomics Viewer (Thorvaldsdóttir *et al.*, 2013). The LASTZ command including the optimized parameters was "lastz nanopore_reads[multiple, unmask] reference_database -format=general: name1,size1,start1, length1,strand1,name2,size2,start2,length2,strand2,identity,score –ambiguous=iupac --xdrop=10 --hspthresh=1000". Additionally, the hits with bit scores below 7000 and those with lengths exceeding 1.23× the length of the corresponding reference sequence were discarded (the latter restriction was used to discard the partially unspecific hits that spanned a region of unrelated sequence embedded between two regions with similarities to the reference). Because the similarity searches typically produced large numbers of overlapping hits, they were further processed using custom scripts to detect the coordinates of contiguous repeat regions in the reads (Figure 1). The regions longer than 300 bp (satellite repeats) or 500 bp (rDNA and telomeric repeats) were recorded and further analyzed. The positions and orientations of the detected satellites were recorded in the form of coded reads where nucleotide sequences were replaced by characters representing the codes for the detected repeats and their orientations, or "0" and "X", which denoted no detected repeats and annotation conflicts, respectively. In the case of the analysis of repeats other than satellites, the reference databases were augmented for assembled contig sequences representing the following most abundant groups of *L. sativus* dispersed repeats: Ty3/gypsy/Ogre, Ty3/gypsy/Athila, Ty3/gypsy/Chromovirus, Ty3/gypsy/other, Ty1/copia/Maximus, Ty1/copia/other, LTR/unclassified and DNA transposon. These repeats were not arranged nor scored with respect to their orientations. In cases of annotation conflicts of these repeats with the selected satellites, they were scored with lower priority.

Detection of the retrotransposon protein coding domains in the read sequences was performed using DANTE, which is a bioinformatic tool available on the RepeatExplorer server (https://repeatexplorer-elixir.cerit-sc.cz/) employing the LAST program (Kielbasa *et al.*, 2011) for similarity searches against the REXdb protein database (Neumann *et al.*, 2019). The hits were filtered to pass the following cutoff parameters: minimum identity = 0.3, min. similarity = 0.4, min. alignment length = 0.7, max. interruptions (frameshifts or stop codons) = 10, max. length proportion = 1.2, and protein domain type = ALL. The positions of the filtered hits were then recorded in coded reads as described above.

Analysis of the association of the satellite arrays with other repeats was performed by summarizing the frequencies of all types of repeats detected within 10 kb regions directly adjacent to all arrays of the same satellite repeat family. Visual inspection of the repeat arrangement within the individual nanopore reads using self-similarity dot-plot analysis was performed using the Dotter (Sonnhammer and Durbin, 1995) and Gepard (Krumsiek *et al.*, 2007) programs.

Periodicity analysis was performed for the individual satellite repeat arrays longer than 30 kb that were extracted from the nanopore reads and plotted for each array separately or averaged for all arrays of the same satellite. The analysis was performed using the fast Fourier transform algorithm (Venables and Ripley, 2002) as implemented in R programming environment. Briefly, a nucleotide sequence X was converted to its numerical representation $\hat{X}$ where

$$\hat{X}(i) = \begin{cases} 1 \text{ if } X(i) = A \\ 2 \text{ if } X(i) = C \\ 3 \text{ if } X(i) = G \\ 4 \text{ if } X(i) = T \end{cases}$$

For the resulting sequences of integers, fast Fourier transform was conducted, and the frequencies f from the frequency spectra were converted to periodicity T as:

$$\mathrm{T} = \frac{\mathrm{L}}{\mathrm{f}}$$

where L is the length of the analyzed satellite array. The analysis reveals the lengths of monomers and other tandemly repeated units like HORs as peaks at the corresponding positions on the resulting periodicity spectrum. However, it should be noted that, while these sequence periodicities will always be represented by peaks, some additional peaks with shorter periods could have merely reflected higher harmonics that are present due to the non-sine character of the numerical representation of nucleotide sequences (Li, 1997; Sharma *et al.*, 2004). Alternatively, periodicity was analyzed using the autocorrelation function as implemented in the R programming environment (McMurry and Politis, 2010). The nucleotide sequence, X, was first converted to four numerical representations: $\widehat{X}_A, \widehat{X}_C, \widehat{X}_T, \widehat{X}_G$ where:

$$\widehat{X}_N = \begin{cases} 1 \text{ if } X(i) = N \\ 0 \text{ if } X(i) \neq N \end{cases}$$

The resulting numerical series were used to calculate the autocorrelations with a lag ranging from 2 to 2000 nucleotides.

Chromosome preparation and fluorescence *in situ* hybridization

Mitotic chromosomes were prepared from root tip meristems synchronized using 1.18 mM hydroxyurea and 15 μM oryzalin as described previously (Neumann *et al.*, 2015). Synchronized root tip meristems were fixed in a 3:1 v/v solution of methanol and glacial acetic acid for 2 days at 4°C. Then the meristems were washed in ice-cold water and digested in 4% cellulase (Onozuka R10, Serva Electrophoresis, Heidelberg, Germany), 2% pectinase and 0.4% pectolyase Y23 (both MP Biomedicals, Santa Ana, CA, USA) in 0.01 M citrate buffer (pH 4.5) for 90 min at 37°C. Following the digestion, the meristems were carefully washed in ice-cold water and post-fixed in the 3:1 fixative solution for 1 day at 4°C. The chromosome spreads were prepared by transferring one meristem to a glass slide, macerating it in a drop of freshly made 3:1 fixative and placing the glass slide over a flame as described in (Dong *et al.*, 2000). After air-drying, the chromosome preparation were kept at −20°C until used for FISH.

Oligonucleotide FISH probes were labelled with biotin, digoxigenin or rhodamine-red-X at their 5′ ends during synthesis (Integrated DNA Technologies, Leuven, Belgium). They were used for all satellite repeats except for FabTR-53, for which two genomic clones, c1644 and c1645, were used instead. The clones were prepared by PCR amplification of *L. sativus* genomic DNA using primers LASm7c476F (5′-GTTTCTTCGTCAGTAAGCCACAG-3′) and LASm7c476R (5′-TGGTGATGGAGAAGAAACATATTG-3′), cloning the amplified band and sequence verification of randomly picked clones as described (Macas *et al.*, 2015). The same approach was used to generate probe corresponding to the integrase coding domain of the Ty3/gypsy Ogre elements. The PCR primers used to amplify the prevailing variant A (clone c1825) were PN_ID914 (5′-TCTCMYTRGTGTACGGTATGGAAG-3′) and PN_ID915 (5′-CCTTCRTARTTGGGAGTCCA-3′). The sequences of all probes are provided in Data S2. The clones were biotin-labelled using nick translation (Kato *et al.*, 2006). FISH was performed according to (Macas *et al.*, 2007) with hybridization and washing temperatures adjusted to account for the AT/GC content and hybridization stringency while allowing for 10–20% mismatches. The slides were counterstained with 4′,6-diamidino-2-phenylindole (DAPI), mounted in Vectashield mounting medium (Vector Laboratories, Burlingame, CA, USA) and examined using a Zeiss AxioImager.Z2 microscope with an Axiocam 506 mono camera. The images were captured and processed using ZEN pro 2012 software (Carl Zeiss GmbH).

AVAILABILITY OF SOURCE CODE AND REQUIREMENTS

- Project Name: nanopore-read-annotation
- Project homepage: https://github.com/vondrakt/nanopore-read-annotation
- Operating system(s): Linux
- Programming language: python3, R
- Other requirements: R packages: TSclust, Rfast, Biostrings (Bioconductor),
- License: GPLv3

AVAILABILITY OF SUPPORTING DATA AND MATERIALS

Raw nanopore reads are available in the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under run accession numbers ERR3374012 and ERR3374013.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This work was supported supported by the ERDF/ESF project ELIXIR-CZ - Capacity building (No. CZ.02.1.01/0.0/0.0/16_013/0001777) and the ELIXIR-CZ research infrastructure project (MEYS No: LM2015047) including access to computing and storage facilities.

AUTHORS' CONTRIBUTIONS

JM conceived the study and drafted the manuscript. TV and PNo developed the scripts for the bioinformatic analysis, and TV, PNo, PNe and JM analyzed the data. AK isolated the HMW genomic DNA and cloned the FISH probes. JM performed the nanopore sequencing. LAR conducted the FISH experiments. All authors reviewed and approved the final manuscript.

ACKNOWLEDGEMENTS

We thank Ms. Vlasta Tetourová and Ms. Jana Látalová for their excellent technical assistance.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

SUPPORTING INFORMATION

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

Figure S1. Dot-plot sequence similarity comparison of consensus monomer sequences.

Figure S2. Length distributions of nanopore reads.

Figure S3. Length distributions of satellite repeat arrays (histograms of counts).

Figure S4. Self-similarity dot-plot of selected nanopore reads.

Figure S5. Detailed periodicity analysis of FabTR-2 and FabTR-53 arrays.

Figure S6. Distribution of the satellite repeats on the metaphase chromosomes of *L. sativus*.

© 2019 The Authors.

Table S1. Similarity hits of *L. sativus* satellite repeats to the repeat clustering data from two related *Lathyrus* species.

Data S1. Consensus sequences of satellite repeat monomers.

Data S2. Sequences of FISH probes.

REFERENCES

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I.J., Koblížková, A., Macas, J. and Lysák, M.A. (2010) Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. *Ann. Bot.* **107**, 255–268.

Ávila Robledillo, L., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., Schubert, I. and Macas, J. (2018) Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Sci. Rep.* **8**, 5838.

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.

Cechova, M. and Harris, R.S. (2018) High inter- and intraspecific turnover of satellite repeats in great apes. *bioRxiv*. https://doi.org/10.1101/470054.

Cohen, S., Agmon, N., Yacobi, K., Mislovati, M. and Segal, D. (2005) Evidence for rolling circle replication of tandem genes in *Drosophila*. *Nucleic Acids Res.* **33**, 4519–4526.

Copenhaver, G.P. and Pikaard, C.S. (1996) Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. *Plant J.* **9**, 273–282.

Dias, G.B., Svartman, M., Delprat, A., Ruiz, A. and Kuhn, G.C.S. (2014) Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. *Genome Biol. Evol.* **6**, 1302–1313.

van Dijk, E.L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018) The third revolution in sequencing technology. *Trends Genet.* **34**, 666–681.

Dong, F., Song, J., Naess, S.K., Helgeson, J.P., Gebhardt, C. and Jiang, J. (2000) Development and applications of a set of chromosome-specific cytogenetic DNA markers in potato. *Theor. Appl. Genet.* **101**, 1001–1007.

Elder, J.F. and Turner, B.J. (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. *Q. Rev. Biol.* **70**, 297–320.

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

Garrido-Ramos, M.A. (2017) Satellite DNA: An evolving topic. *Genes (Basel)*, **8**, 230.

Gong, Z., Wu, Y., Koblížková, A. *et al.* (2012) Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell*, **24**, 3559–3574.

Harris, R.S. (2007) Improved pairwise alignment of genomic. DNA. Doctoral Thesis, The Pennsylvania State University.

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

Heckmann, S., Macas, J., Kumke, K. *et al.* (2013) The holocentric species *Luzula elegans* shows interplay between centromere and large-scale genome organization. *Plant J.* **73**, 555–565.

Henikoff, J.G., Thakur, J., Kasinathan, S. and Henikoff, S. (2015) A unique chromatin complex occupies young alpha-satellite arrays of human centromeres. *Sci. Adv.* **1**, e1400234.

Herzel, H., Weiss, O. and Trifonov, E.N. (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics*, **15**, 187–193.

Jain, M., Olsen, H.E., Turner, D.J. *et al.* (2018) Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323.

Kato, A., Albert, P.S., Vega, J.M. and Birchler, J.A. (2006) Sensitive fluorescence *in situ* hybridization signal detection in maize using directly labeled probes produced by high concentration DNA polymerase nick translation. *Biotech. Histochem.* **81**, 71–78.

Khost, D.E., Eickbush, D.G. and Larracuente, A.M. (2017) Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res.* **27**, 709–721.

Kielbasa, S.M., Wan, R., Sato, K., Kiebasa, S.M., Horton, P. and Frith, M.C. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493.

Kit, S. (1961) Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *J. Mol. Biol.* **3**, 711–716.

Krumsiek, J., Arnold, R. and Rattei, T. (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, **23**, 1026–1028.

Kubát, Z., Zlůvová, J., Vogel, I., Kováčová, V., Cermák, T., Cegan, R., Hobza, R., Vyskot, B. and Kejnovský, E. (2014) Possible mechanisms responsible for absence of a retrotransposon family on a plant Y chromosome. *New Phytol.* **202**, 662–678.

Kuzminov, A. (2016) Chromosomal replication complexity: a novel DNA metrics and genome instability factor. *PLOS Genet.* **12**, e1006229.

Li, W. (1997) The study of correlation structures of DNA sequences: a critical review. *Comput. Chem.* **21**, 257–271.

Ma, J. and Jackson, S.A. (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.* **16**, 251–259.

Macas, J. and Neumann, P. (2007) Ogre elements – a distinct group of plant Ty3/gypsy-like retrotransposons. *Gene*, **390**, 108–16.

Macas, J., Mészáros, T. and Nouzová, M. (2002) PlantSat: a specialized database for plant satellite repeats. *Bioinformatics*, **18**, 28–35.

Macas, J., Navrátilová, A. and Mészáros, T. (2003) Sequence subfamilies of satellite repeats related to rDNA intergenic spacer are differentially amplified on *Vicia sativa* chromosomes. *Chromosoma*, **112**, 152–158.

Macas, J., Navrátilová, A. and Koblížková, A. (2006) Sequence homogenization and chromosomal localization of VicTR-B satellites differ between closely related *Vicia* species. *Chromosoma*, **115**, 437–447.

Macas, J., Neumann, P. and Navrátilová, A. (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics*, **8**, 427.

Macas, J., Koblížková, A., Navrátilová, A. and Neumann, P. (2009) Hypervariable 3′UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. *Gene*, **448**, 198–206.

Macas, J., Novák, P., Pellicer, J. *et al.* (2015) In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe *Fabeae*. *PLoS One*, **10**, e0143424.

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.

McMurry, T.L. and Politis, D.N. (2010) Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *J. Time Ser. Anal.* **31**, 471–482.

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.

Metzker, M.L. (2009) Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46.

Mitsuhashi, S., Frith, M.C., Mizuguchi, T. *et al.* (2019) Tandem-genotypes : robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **20**, 58.

Murray, M.G. and Thompson, W.F. (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4326.

Navrátilová, A., Koblížková, A. and Macas, J. (2008) Survey of extrachromosomal circular DNA derived from plant satellite repeats. *BMC Plant Biol.* **8**, 90.

Neumann, P., Koblížková, 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**, 1047–56.

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

Neumann, P., Schubert, V., Fuková, I., Manning, J.E., Houben, A. and Macas, J. (2016) Epigenetic histone marks of extended meta-polycentric centromeres of *Lathyrus* and *Pisum* chromosomes. *Front. Plant Sci.* **7**, 234.

© 2019 The Authors.

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.

Novák, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

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

Peona, V., Weissensteiner, M.H. and Suh, A. (2018) How complete are 'complete' genome assemblies? - an avian perspective. *Mol. Ecol. Resour.* **18**, 1188–1195.

Plohl, M., Meštrović, N. and Mravinac, B. (2014) Centromere identity from the DNA point of view. *Chromosoma*, **123**, 313–325.

De Roeck, A., De Coster, W., Bossaerts, L. ***et al.*** (2018) Accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *bioRxiv*, 439026. https://doi.org/10.1101/439026

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.

Schindelhauer, D. and Schwarz, T. (2002) Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous alpha-satellite DNA array. *Genome Res.* **12**, 1815–1826.

Sharma, D., Issac, B., Raghava, G.P.S. and Ramaswamy, R. (2004) Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics*, **20**, 1405–1412.

Smith, G.P. (1976) Evolution of repeated DNA sequences by unequal crossover. *Science*, **191**, 528–535.

Sonnhammer, E.L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1-10.

Stephan, W. (1986) Recombination and the evolution of satellite DNA. *Genet. Res.* **47**, 167–174.

Stephan, W. and Cho, S. (1994) Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics*, **136**, 333–341.

Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192.

Valeri, M.P., Dias, G.B., Pereira, V.D.S., Campos Silva Kuhn, G. and Svartman, M. (2018) An eutherian intronic sequence gave rise to a major satellite DNA in *Platyrrhini*. *Biol. Lett.* **14**, 20170686.

Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics with S*. 4th edn. New York, NY: Springer.

Vershinin, A.V. and Heslop-Harrison, J.S. (1998) Comparative analysis of the nucleosomal structure of rye, wheat and their relatives. *Plant Mol. Biol.* **36**, 149–161.

Vyskot, B. and Hobza, R. (2015) The genomics of plant sex chromosomes. *Plant Sci.* **236**, 126–135.

Walsh, J.B. (1987) Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics*, **115**, 553–567.

Weissensteiner, M.H., Pang, A.W.C., Bunikis, I., Höijer, I., Vinnere-Petterson, O., Suh, A. and Wolf, J.B.W. (2017) Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res.* **27**, 697–708.

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* (Hörandl, E. and Appelhans, M., eds). Königstein, Germany: Koeltz Scientific Books, pp. 155–179.

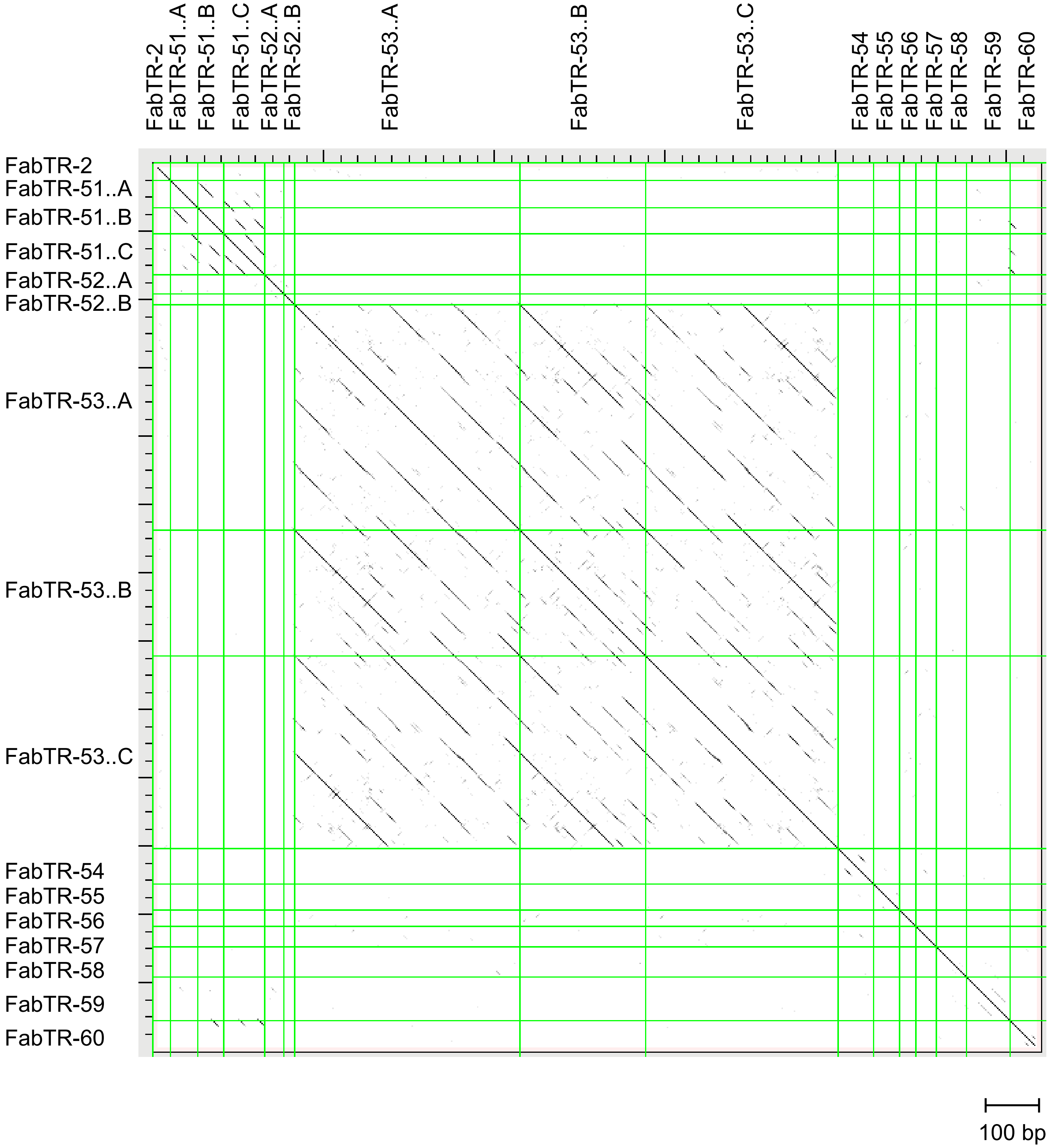
© 2019 The Authors.

Supplementary information

Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats.

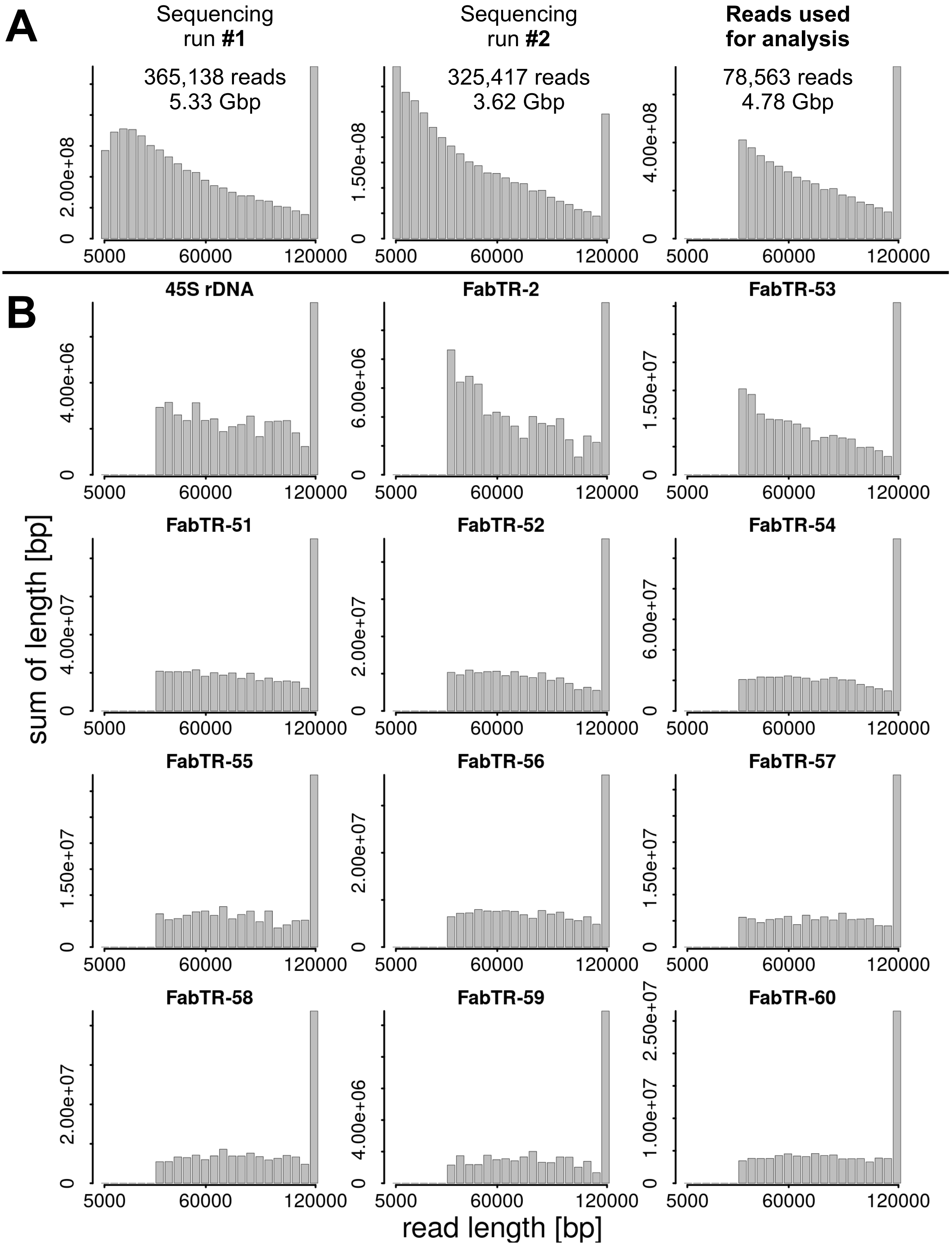
Tihana Vondrak, Laura Ávila Robledillo, Petr Novák, Andrea Koblížková, Pavel Neumann and Jiří Macas.

Supplementary Fig. S1



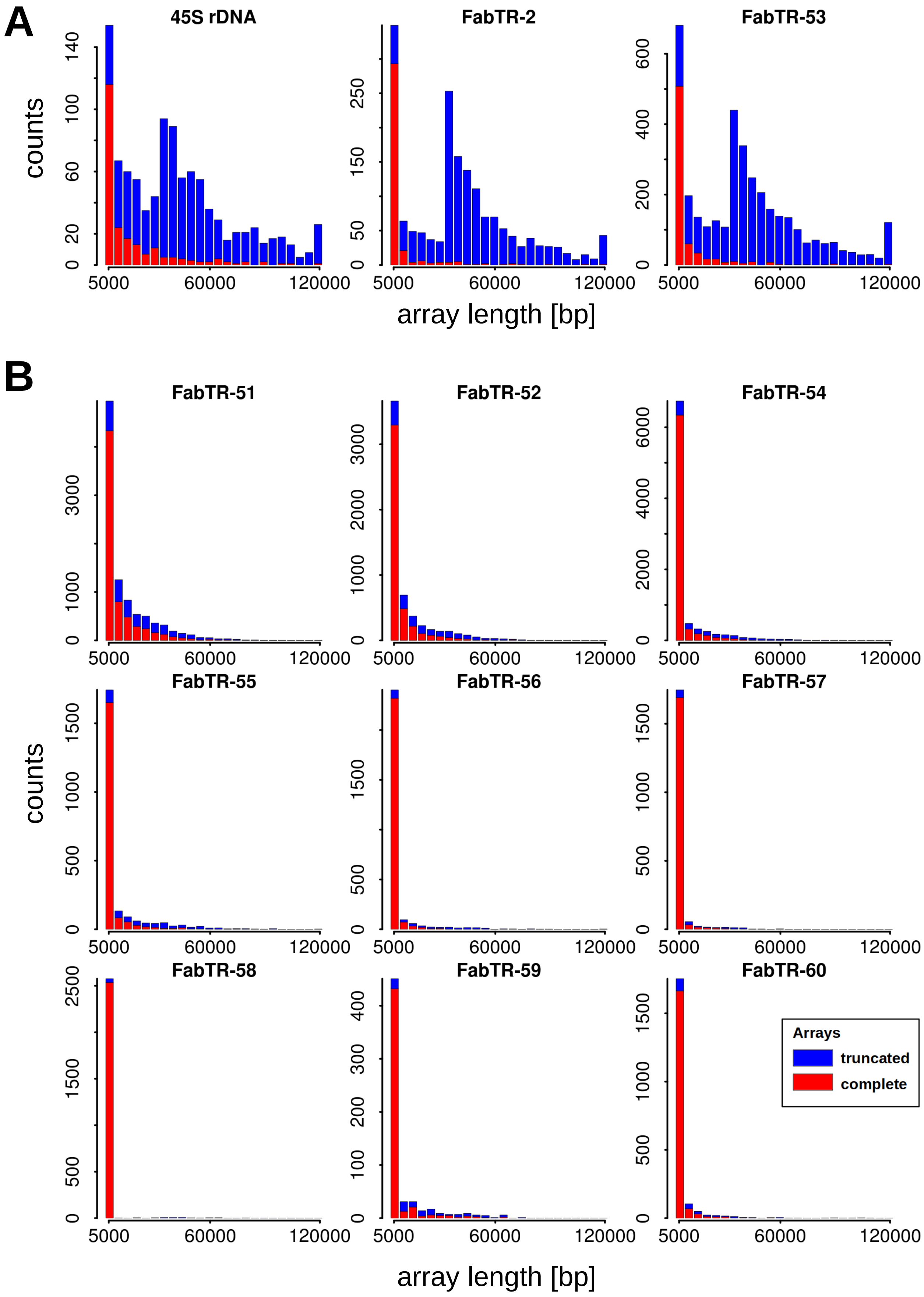
Supplementary Fig. S1. Dot-plot sequence similarity comparison of consensus monomer sequences. The sequences are separated by green lines and their similarities exceeding 40% over a 100 bp sliding window are displayed as black dots or diagonal lines.

Supplementary Fig. S2



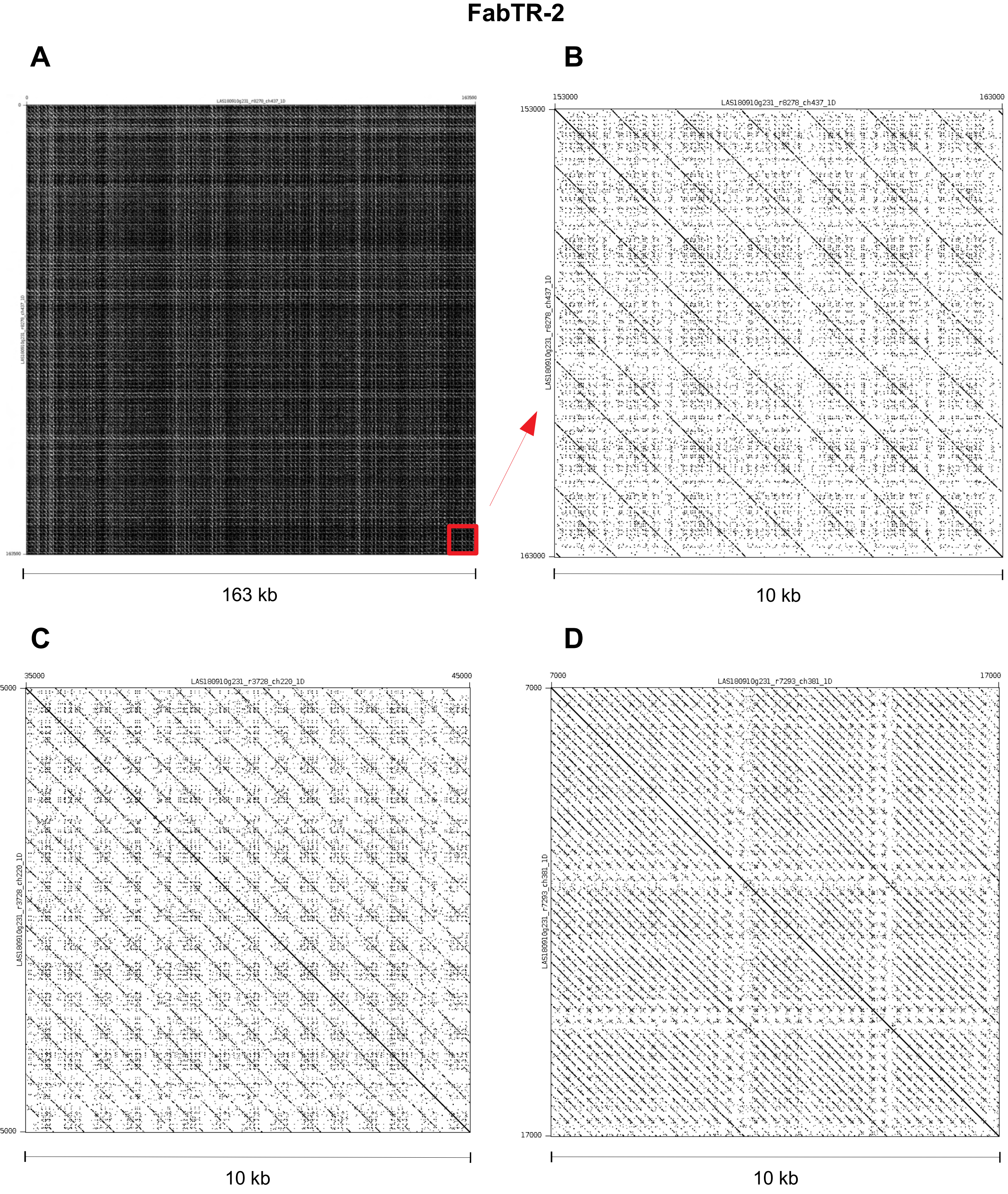
Supplementary Fig. S2. Length distributions of nanopore reads displayed as weighted histograms with bin size of 5 kb, with the last bin including all reads longer than 120 kb. (**A**) Length distributions of raw reads from two sequencing runs and the final set of quality-filtered and size-selected (>30kb) reads used for analysis. (**B**) Length distributions of nanopore reads containing rDNA and satellite repeats.

Supplementary Fig. S3



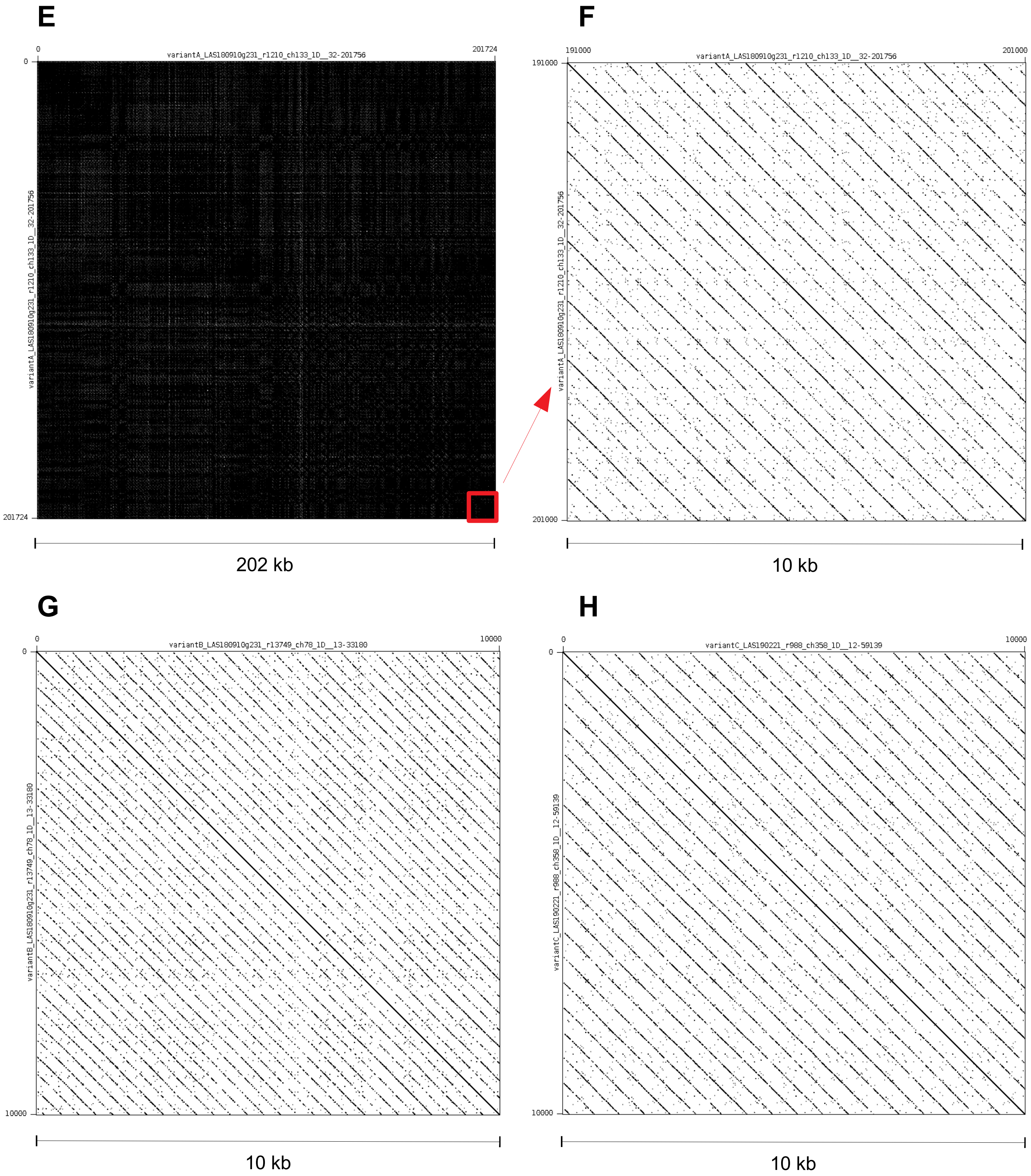
Supplementary Fig. S3. Length distributions of satellite repeat arrays displayed as histograms with bin size of 5 kb, with the last bin including all arrays longer than 120 kb. Arrays which were completely embedded within the reads (red bars) are distinguished from those truncated due to their positions at the ends of the reads (blue bars). Tandem repeats forming long arrays are shown in panel **A**, while the remaining repeats forming predominantly short arrays are in panel **B**.

Supplementary Fig. S4



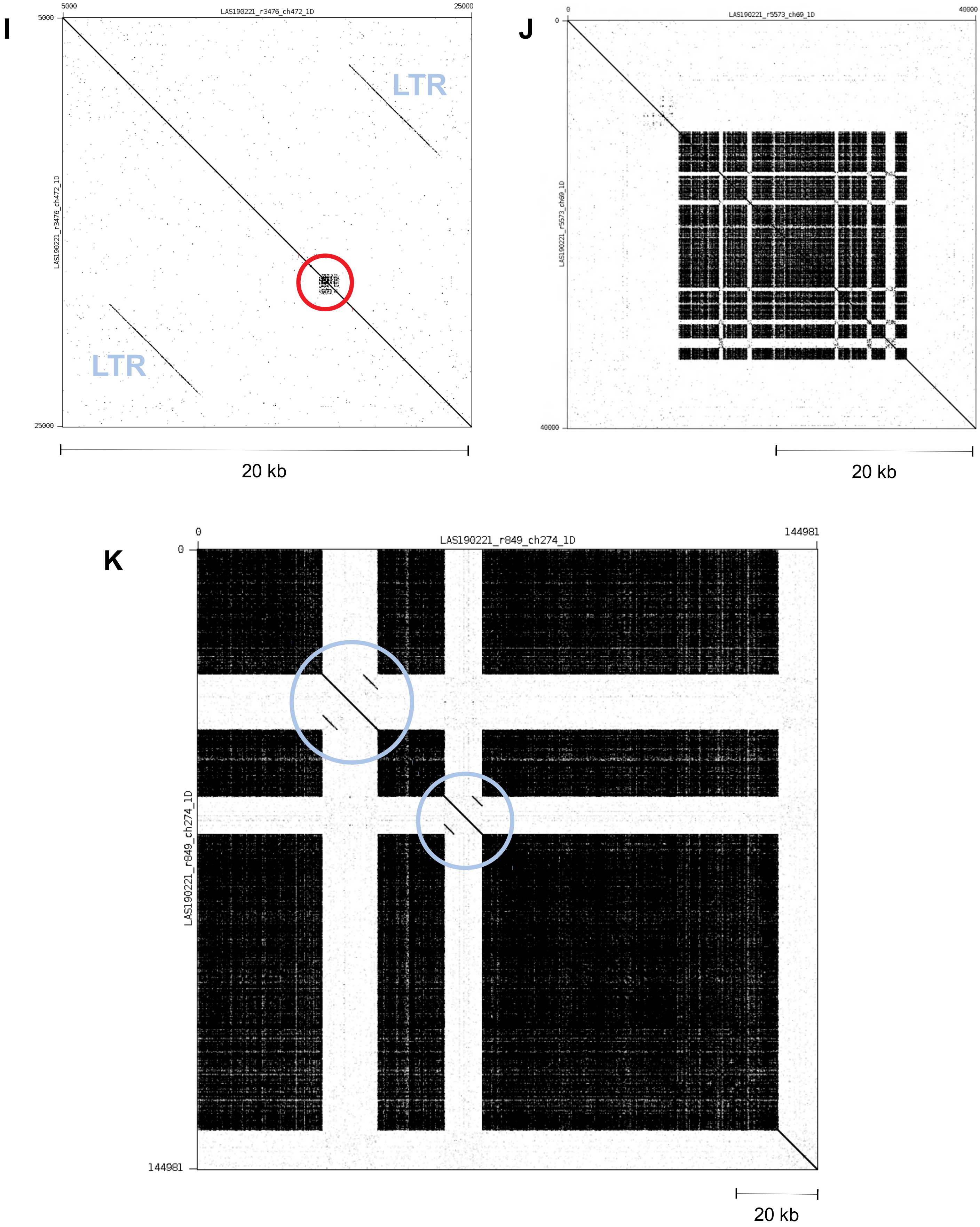
Supplementary Fig. S4 A-D. Self-similarity dot-plot visualization of FabTR-2 arrays. Tandem repeats are revealed as diagonal lines with spacing corresponding to monomer length. (**A**) Example of a 163 kb read completely made of FabTR-2 array (the periodicity pattern is obscured by the high density of lines). (**B**) Magnification of the 10 kb region highlighted by a red square on panel A. This array is homogenized as ~1300 bp HOR. (**C,D**) Examples of other FabTR-2 periodicities detected in different reads (only 10 kb regions were used for dot-plots to make periodicity patterns comparable with other plots).

FabTR-53

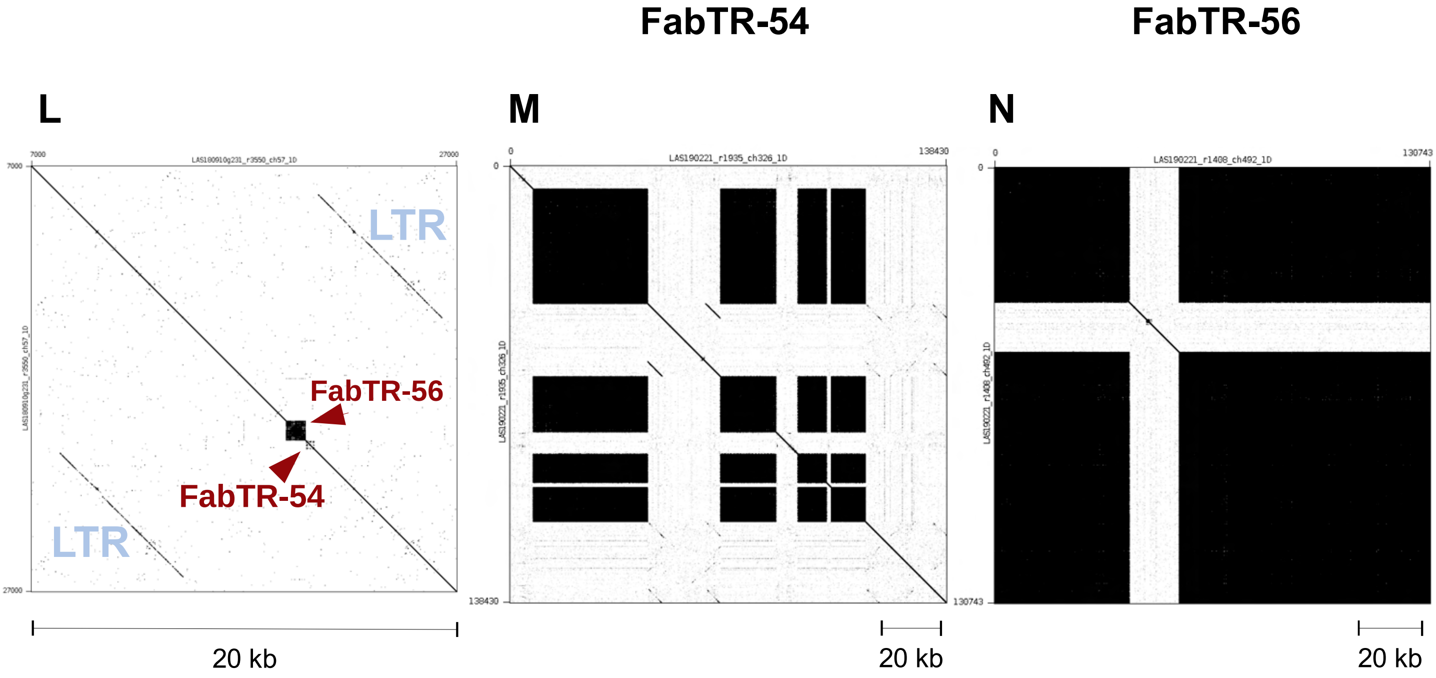


Supplementary Fig. S4 E-H. Self-similarity dot-plot visualization of FabTR-53 arrays. (**E**) Example of a 202 kb read completely made of FabTR-2 array (the periodicity pattern is obscured by the high density of lines). (**F**) Magnification of the 10 kb region highlighted by a red square on panel A. (**G,H**) Examples of other FabTR-53 periodicities detected in different reads (only 10 kb regions were used for dot-plots to make periodicity patterns comparable with other plots).

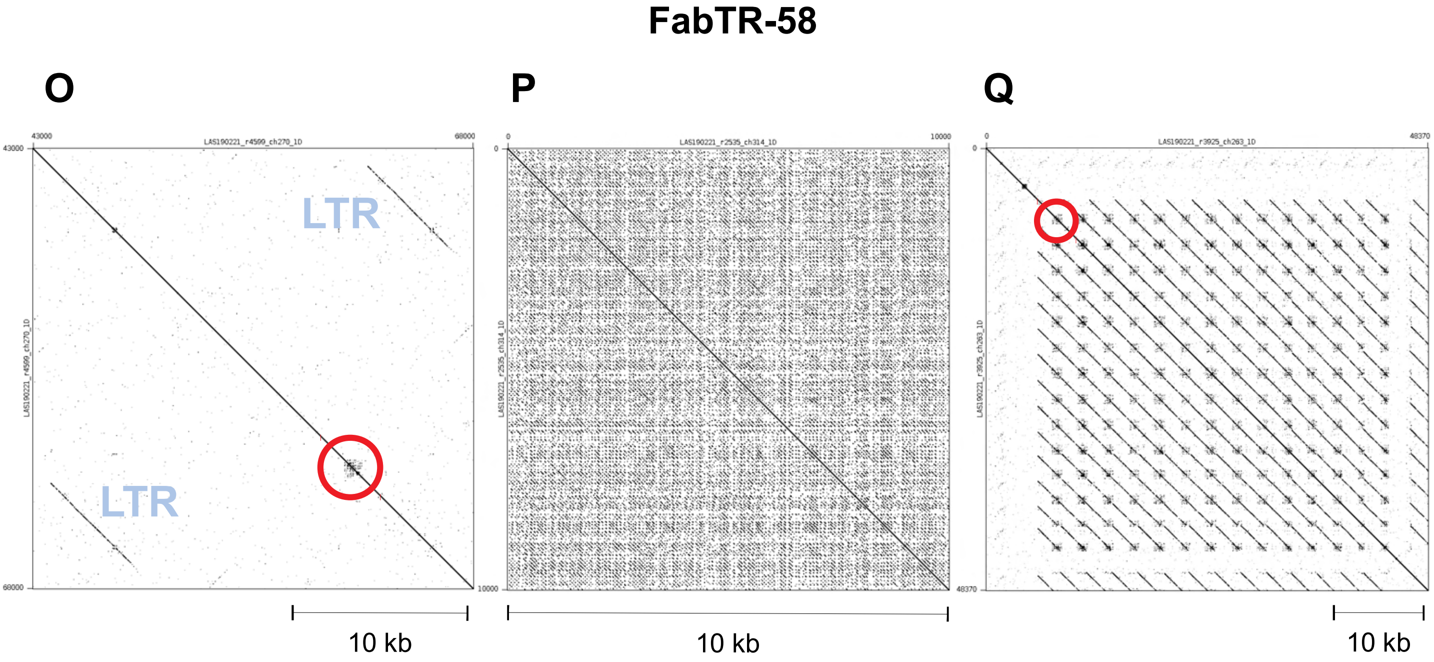
FabTR-52



Supplementary Fig. S4 I-K. Dot-plots demonstrating length distribution of FabTR-52 arrays, ranging from short arrays (red circle) embedded within LTR-retrotransposon sequences (**I**) and partially expanded arrays (**J**) to the arrays >100 kb in length which are interrupted by insertions of LTR-retrotransposons (blue circles) (**K**).

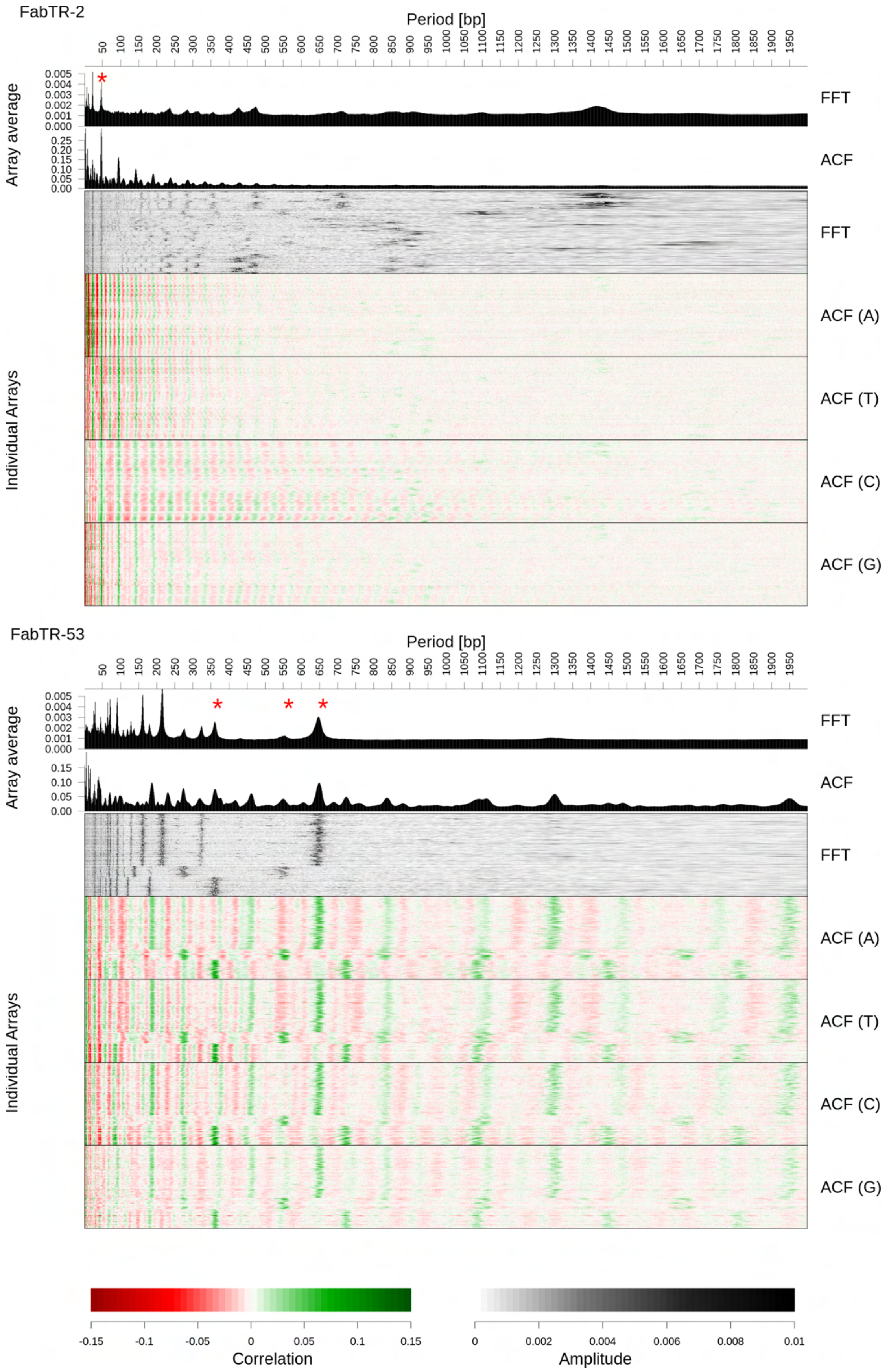


Supplementary Fig. S4 L-N. (**L**) Example of LTR-retrotransposon carrying short FabTR-54 and FabTR-56 arrays. Reads with those tandem repeats expanded to long arrays are shown on panels **M** (FabTR-54) and **N** (FabTR-56). The expanded tandem arrays appear as black squares on the dot-plots due to high density of lines.



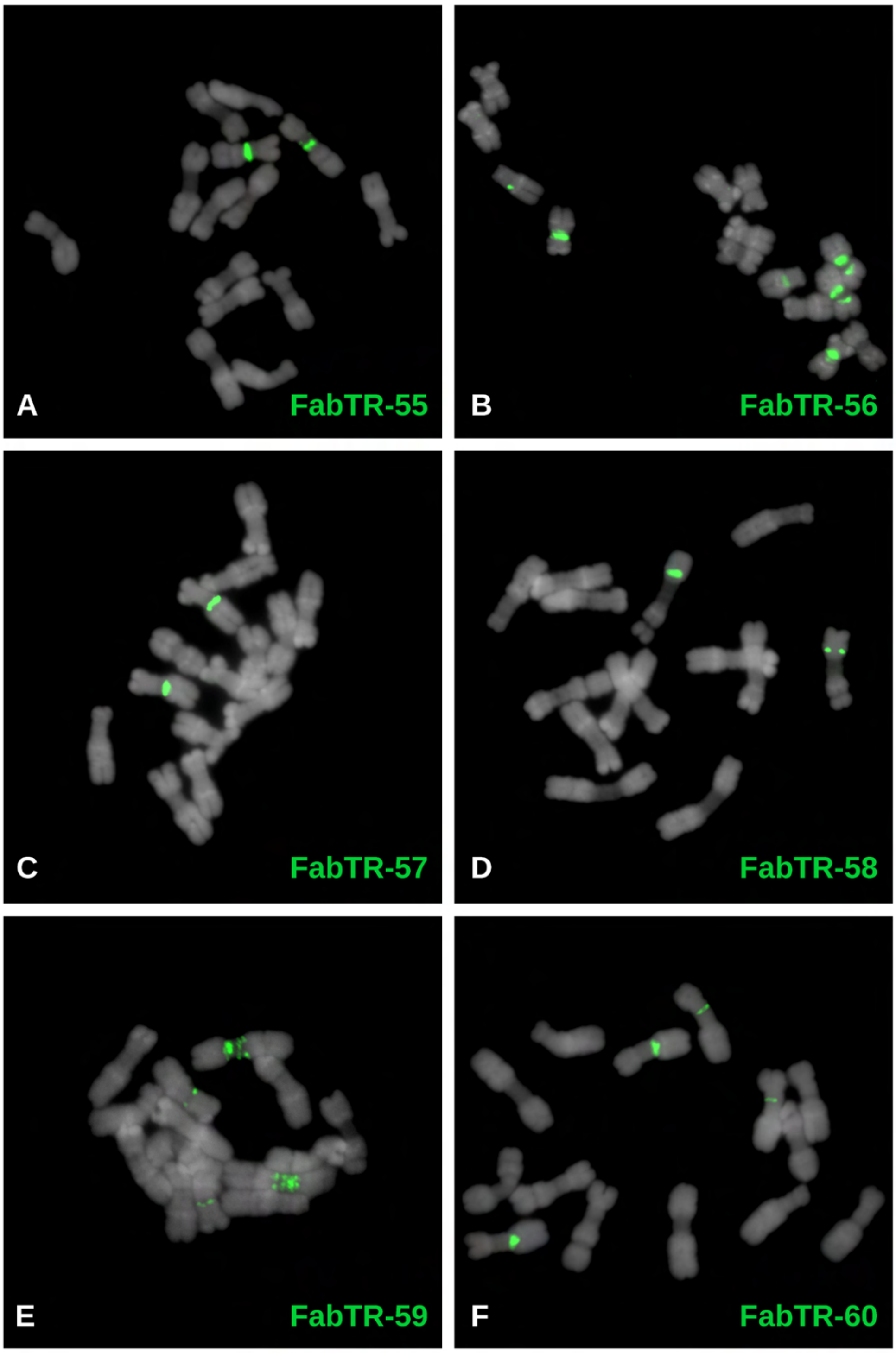
Supplementary Fig. S4 O-Q. Three types of genome organization of FabTR-58 repeats: (O) short array (marked by red circle) within LTR-retrotransposon, (P) expanded array, (Q) short arrays embedded within a longer tandem repeat monomer.

Supplementary Fig. S5



Supplementary Fig. S5. Detailed periodicity analysis of FabTR-2 and FabTR-53 arrays. Periodicity analysis using fast Fourier transform (FFT) and autocorrelation function (ACF) are shown as averages of spectra calculated on individual satellite arrays longer than 30 kb. Periodicity spectra from individual arrays are shown as heatmaps with rows corresponding to individual arrays. Autocorrelations are shown separately for individual nucleotides. The array average graphs of FabTR-53 were calculated with all subfamilies combined and the FFT peaks corresponding to different monomer lengths of the three subfamilies are indicated with asterisks.

Supplementary Fig. S6



Supplementary Fig. S6. **Distribution of the satellite repeats on the metaphase chromosomes of *L. sativus* (2n = 14)**. The satellites were visualized using FISH, with individual probes labeled as indicated by the color-coded descriptions. The chromosomes counterstained with DAPI are shown in gray.

Supplementary Tab. 1. Similarity hits of *L. sativus* satellite repeats to the repeat clustering data (Macas et al., 2015) from two related *Lathyrus* species

	L. vernus				***L. latifolius***			
Satellite repeat	Hit score [(a)]	Cluster[(b)]	Annotation [(b)]	tandem subrepeats [(c)]	Hit score [(a)]	Cluster [(b)]	Annotation [(b)]	tandem subrepeats [(c)]
FabTR-54	3e-05, 24/24 (100%)	CL87	Putative LTR-retrotransposon	Yes	1e-06, 26/26 (100%)	CL135	Dispersed repeat	Yes
FabTR-55	3e-14, 92/113 (81%)	CL87	Putative LTR-retrotransposon	Yes	3e-64, 145/152 (95%)	CL150	Dispersed repeat	Yes
FabTR-57	2e-33, 99/107 (92%)	CL82	LTR/gypsy/ Ogre	Yes	1e-54, 120/123 (97%)	CL5	Putative LTR-retrotransp.	Yes

(a) BLASTn hit score is provided as E-value, number of identities/hit length (% similarity)

(b) Cluster numbers and their annotations correspond to the repeat analysis described in Macas et al. (2015)

(c) Presence of short, tandem subrepeats in contigs assembled from the repeat clusters

Chapter II

Complex sequence organization of heterochromatin in the holocentric plant *Cuscuta europaea* elucidated by the computational analysis of nanopore reads.

Computational and Structural Biotechnology Journal 19 (2021) 2179–2189

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Complex sequence organization of heterochromatin in the holocentric plant *Cuscuta europaea* elucidated by the computational analysis of nanopore reads

Tihana Vondrak [a,b], Ludmila Oliveira [a], Petr Novák [a], Andrea Koblížková [a], Pavel Neumann [a], Jiří Macas [a,*]

[a] *Biology Centre, Czech Academy of Sciences, Institute of Plant Molecular Biology, Branišovská 31, České Budějovice CZ-37005, Czech Republic*
[b] *University of South Bohemia, Faculty of Science, České Budějovice, Czech Republic*

ARTICLE INFO

Article history:
Received 12 February 2021
Received in revised form 31 March 2021
Accepted 3 April 2021
Available online 22 April 2021

Keywords:
Satellite DNA
Heterochromatin
Oxford Nanopore sequencing
Fluorescence *in situ* hybridization
Holocentric chromosomes
LINE elements

ABSTRACT

Repeat-rich regions of higher plant genomes are usually associated with constitutive heterochromatin, a specific type of chromatin that forms tightly packed nuclear chromocenters and chromosome bands. There is a large body of cytogenetic evidence that these chromosome regions are often composed of tandemly organized satellite DNA. However, comparatively little is known about the sequence arrangement within heterochromatic regions, which are difficult to assemble due to their repeated nature. Here, we explore long-range sequence organization of heterochromatin regions containing the major satellite repeat CUS-TR24 in the holocentric plant *Cuscuta europaea*. Using a combination of ultra-long read sequencing with assembly-free sequence analysis, we reveal the complex structure of these loci, which are composed of short arrays of CUS-TR24 interrupted frequently by emerging simple sequence repeats and targeted insertions of a specific lineage of LINE retrotransposons. These data suggest that the organization of satellite repeats constituting heterochromatic chromosome bands can be more complex than previously envisioned, and demonstrate that heterochromatin organization can be efficiently investigated without the need for genome assembly.

© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Heterochromatin is a tightly packed, fundamental form of chromatin organization in eukaryotic nuclei exhibiting a unique combination of post-translational histone modifications [1,49]. In higher plants, cytologically defined constitutive heterochromatin is mostly associated with large tracks of highly repetitive satellite DNA (satDNA) and forms densely stained bands on mitotic chromosomes or chromocenters in interphase nuclei [13]. In plants with monocentric chromosomes and small genomes, this heterochromatin is usually confined to centromeric and pericentric regions [49]. In species with larger genomes, however, it can be found in additional subtelomeric and interstitial chromosomal loci [12], whereas plants with holocentric chromosomes usually lack distinguishable heterochromatic bands [19]. Heterochromatin is supposed to play an important role in chromosome segregation, gene regulation and the maintenance of genome stability [49], yet the processes shaping its distribution throughout the genome, and the role of underlying repetitive sequences, remain poorly understood [13]. This is in part due to our limited knowledge of the long-range sequence arrangement of repeat-rich heterochromatic regions which are in principle difficult to assemble [38].

SatDNA is organized in the genome in long arrays of almost identical, tandemly arranged units called monomers. Monomer sequences are typically hundreds of nucleotides long [27], although they can be as short as simple sequence repeats (<10 bp) [19] or reach over 5 kb [15]. Since monomer arrays can extend megabases in length, they present a significant challenge for even the most advanced genome assembly projects. Consequently, sequence composition of plant heterochromatin is traditionally elucidated by mapping repeats to heterochromatic chromosome bands using fluorescence *in situ* hybridization (FISH) [21]. However, this approach requires prior knowledge of the repeated sequences to be used as FISH probes. Despite the recent introduction of bioinformatic tools designed to retrieve satellite DNA sequences from short next generation sequencing (NGS) reads [34,41], this reverse approach does not ensure that all repeats present in heterochromatic regions are revealed. Moreover, FISH-based methods have relatively limited resolution and are unable

---

[*] Corresponding author.
E-mail address: macas@umbr.cas.cz (J. Macas).

https://doi.org/10.1016/j.csbj.2021.04.011
2001-0370/© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

to reveal details of the internal structure of highly repetitive regions.

It has recently been demonstrated that repeat-rich genome regions, such as centromeres, can be efficiently assembled using long-read sequencing technologies that include the Pacific Biosciences and Oxford Nanopore platforms [26,30]. The latter platform can generate "ultra-long" reads of up to 1 Mb [8] allowing for investigation of the long-range organization of genomic loci made of satellite DNA. In addition to greatly improving genome assembly [38] , unassembled nanopore reads can also be utilized to examine the properties of satellite repeat arrays using dedicated bioinformatic tools [50] .

One of the most interesting satellite-rich heterochromatic genome regions has recently been described in the holocentric plant *Cuscuta europaea* [36] . Mitotic chromosomes of this species display large 4′,6-diamidino-2-phenylindole (DAPI)-positive heterochromatic bands (schematically depicted in Fig. 1), which are atypical for holocentric plants. Moreover, most of these heterochromatic bands are unique in their association with CENH3, a specific variant of canonical histone H3 that usually marks the position of active centromeres [44] . *C. europaea* CENH3 may have lost this function, however, as the mitotic spindle is able to attach to chromosomes at CENH3-free sites in this species [36] . The mechanism driving CENH3 deposition at heterochromatic bands in this species is currently unknown.

We have shown previously that heterochromatic bands on *C. europaea* chromosomes consist of 389 bp CUS-TR24 satellite repeats amplified to approximately 466,000 copies, accounting for 15.5% of the genome [33,36]. FISH mapping of other *C. europaea* tandem repeats showed that heterochromatic regions also accumulated the simple sequence repeat (SSR) (TAA)n. Moreover, bioinformatic analysis of low-pass shotgun sequencing reads using the RepeatExplorer pipeline showed that the CUS-TR24 satellite can be interspersed with additional repeats [33] . Taken together, these findings indicated that the structure of *C. europaea* heterochromatic genome regions is complex.

In the present work, we have used ultra-long read sequencing to investigate the internal structure of the heterochromatic regions of *C. europaea* chromosomes. We adopted an assembly-free strategy, developed for the characterization of satDNA in the repeat-rich genome of *Lathyrus sativus* [50] , for the genome-wide characterization of satellite arrays. This strategy employed a custom-made reference database for the identification of satellite arrays in individual nanopore reads. Nanopore reads representing significant genome coverage were then analyzed, revealing the prevalent length of arrays in the genome, sequence variations, and patterns of interspersion with other repetitive elements.

2. Material and methods

2.1. Genomic DNA isolation and nanopore sequencing

Seeds of *Cuscusta europaea* (serial number: 0101147) were obtained from the Royal Botanic Garden (Ardingly, UK). The plants were cultivated in the greenhouse and propagated vegetatively, using *Urtica dioica* as their host. High molecular weight nuclear DNA was isolated from young shoots of *C. europaea* employing the protocol described previously [50] . Five grams of tissue was frozen in liquid nitrogen, ground to a fine powder and incubated for 5 min in 35 ml ice-cold H buffer (1 × HB, 0.5 M sucrose, 1 mM phenylmethyl-sulphonylfluoride (PMSF), 0.5% (v/v) Triton X-100, 0.1% (v/v) 2-mercaptoethanol) prepared fresh from a 10 × HB stock (0.1 M TRIS-HCl pH 9.4, 0.8 M KCl, 0.1 M EDTA, 40 mM spermidine, 10 mM spermine). The homogenate was filtered through 48 μm nylon mesh, adjusted to 35 ml with 1 × H buffer, and centrifuged at 230 × g for 15 min at 4°C. The pelleted nuclei were resuspended in 35 ml H buffer, centrifuged at 230 × g for 15 min at 4°C, and the resulting pellet was resuspended in 15 ml TC buffer (50 mM TRIS-HCl pH 7.5, 75 mM NaCl, 6 mM $MgCl_2$, 0.1 mM $CaCl_2$). A final centrifugation was performed at 400 × g for 5 min, and the nuclei were resuspended in 2 ml TC. The suspension of nuclei was mixed with an equal volume of 2 × CTAB buffer (1.4 M NaCl, 100 mM Tris-HCl pH 8.0, 2% CTAB, 20 mM EDTA, 0.5% (w/v) $Na_2S_2O_5$, 2% (v/v) 2-mercaptoethanol) and incubated at 50°C for 30–40 min. The solution was extracted with chloroform: isoamylalcohol (24:1) using MaXtract™ High Density Tubes (Qiagen) and precipitated with a 0.7 × volume of isopropanol using a sterile glass rod to collect the DNA. Following two washes in 70% ethanol, the DNA was dissolved in TE and treated with 2 μl RNase Cocktail™ Enzyme Mix (Thermo Fisher Scientific) for 1 h at 37°C. Finally, the DNA was further purified by mixing the sample with a 0.5 × volume of CU and a 0.5 × volume

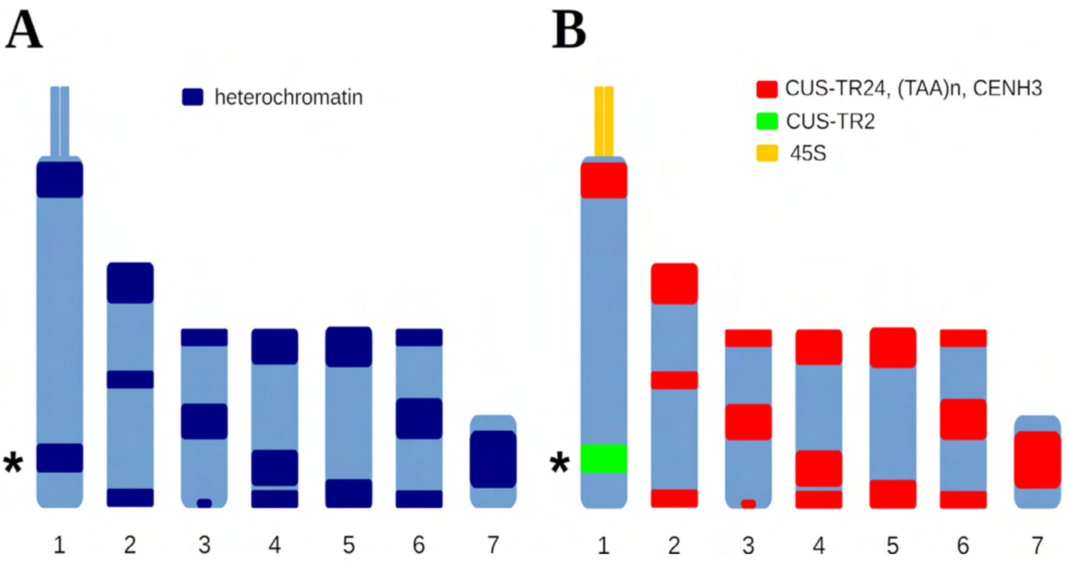


Fig. 1. Schematic representation of the *Cuscuta europaea* karyotype (n = 7) with distribution of DAPI-positive heterochromatin bands (A) and tandem repeats (B). The loci containing CUS-TR24 repeats are associated with the CENH3 protein. The band on chromosome 1 that lacks CUS-TR24 but is composed of the satellite CUS-TR2 is marked with the asterisk. Adapted from [36]

of IR solution from the Qiagen DNeasy PowerClean Pro Clean Up Kit (Qiagen), centrifugation for 2 min at 15,000 rpm at room temperature and DNA precipitation from the supernatant using a 2.5 × volume of 96% ethanol. The DNA was dissolved in 10 mM TRIS-HCl pH 8.5 and stored at 4°C.

Sequencing libraries were prepared from 3 μg of the purified, partially fragmented DNA (from ~20 kb to >100 kb) using a Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies), following the manufacturer's protocol. Briefly, the DNA was treated with 2 μl NEBNext FFPE DNA Repair Mix and 3 μl NEBNext Ultra II End-prep enzyme mix in a 60 μl volume that included 3.5 μl FFPE and 3.5 μl End-prep reaction buffers (New England Biolabs). The reaction was performed at 20°C for 5 min and 65°C for 5 min. Subsequently, the DNA was purified using a 0.4 × volume of AMPure XP beads (Beckman Coulter); because long DNA fragments caused clumping of the beads and were difficult to detach, elution was performed with 5 mM TRIS-HCl (pH 8.5) for 40 min. Subsequent steps, including adapter ligation using NEBNext Quick T4 DNA Ligase and library preparation for sequencing, were performed as recommended. The whole library was loaded onto a MinION FLO-MIN106 R9.4.1 flow cell and sequenced until the number of active pores dropped below 40 (19–20 h). Two independent sequencing runs were performed, and the resulting raw reads were deposited into the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under run accession numbers ERR5237073 and ERR5237074.

2.2. Bioinformatic analysis of nanopore reads

Raw nanopore reads were basecalled using the Oxford Nanopore basecaller Albacore (ver. 2.3.4). Quality-filtering of the resulting FastQ reads, and their conversion to the FASTA format, was performed with BBDuk (part of the BBTools, https://jgi.doe.gov/data-and-tools/bbtools/) run with the parameter maq = 8. Reads shorter than 30 kb were discarded. Unless stated otherwise, all bioinformatic analyses were implemented using custom Python and R scripts, and executed on a Linux-based server equipped with 64 GB RAM and 32 CPUs.

Self-similarity dot-plot analysis of individual nanopore reads was done using the Gepard [23] and Dotter programs [46], and the annotated dot-plots used for the figures were generated using FlexiDot [45]. Repeat annotation in nanopore reads and subsequent analysis of the length distribution of tandem repeat arrays and their interspersion with other repetitive sequences followed the procedures described previously [50]. Briefly, the repeats were identified and annotated in the nanopore reads based on their similarities to a custom-made reference database. The database included consensus sequences that were representative of all major repeat groups identified in the *C. europaea* genome using the RepeatExplorer analysis of Illumina reads [33]. For each family of tandem repeats and LINE elements, the reference sequences in the database were placed in the same orientation to allow for the evaluation of their mutual orientations in the nanopore reads. Sequence similarities were detected using LASTZ [18]. The search parameters and processing of the resulting similarity hits were as described previously [50]. The reference database and custom scripts used for the analysis are available from GitHub (https://github.com/vondrakt/nanopore-read-annotation-Cuscuta-europaea.git).

2.3. Analysis of LINE sequences

Consensus sequences of full-length LINE elements were reconstructed from contigs produced by the RepeatExplorer [33]. The positions of regions coding for retrotransposon proteins in these sequences were detected by DANTE (https://repeatexplorer-elixir.cerit-sc.cz/) based on their similarity to the REXdb protein database [32]. Phylogenetic analysis of LINEs was performed using reverse transcriptase (RT) protein domain sequences extracted by DANTE from *C. europaea* contigs. These sequences were supplemented with a set of 71 randomly selected reference RT domains representing different lineages of plant LINEs from Eudicot plant species [20]. Multiple sequence alignment of RT sequences was done using the Muscle alignment program [11] and refined by manual inspection using Seaview [16]. A Neighbor-Joining phylogenetic tree was calculated using Geneious Prime 2020.1.1 (https://www.geneious.com) with default parameters.

Associations of the individual LINE lineages with CUS-TR24 repeats were investigated by extracting all identified LINE sequences from the nanopore reads and dividing them into two groups based on their presence in a 10 kb region adjacent to CUS-TR24 arrays. The elements located within these regions were labeled as associated while those located >10 kb from CUS-TR24 were classified as not associated. Sequences from both groups were assigned to a LINE lineage using the LASTZ program, which compared each sequence with a set of full-length reference LINE sequences. To obtain a unique hit for each sequence, the best hit for each sequence was identified based on the highest bitscore. The LASTZ command for running the alignment was 'lastz query[multiple,unmask] database –format = general:name1,size1,start1, length1,strand1,name2,size2,start2,length2,strand2,identity,score --ambiguous = iupac --xdrop = 10 –hspthresh = 1000'.

Insertion sites of LINEs were mapped to a dimer of CUS-TR24 consensus sequence for full-length (5–7 kb) LINE elements. A 200 bp window was extracted from each side of the LINE. Windows that were shorter than 190 bp or that had <190 bp annotated as CUS-TR24 were discarded. These windows were then aligned to the CUS-TR24 dimer using the LASTZ alignment program and the command described above. The alignment was filtered by bitscore so that each window had a unique hit to the CUS-TR24 dimer, and the insertion sites were recorded. The insertion site frequencies from the identical parts of CUS-TR24 the dimer were merged to produce a monomer insertion site profile.

2.4. Chromosome preparation and FISH

Mitotic chromosomes for FISH experiments were prepared from shoot apical meristems fixed in a 3:1 solution of methanol: glacial acetic acid for at least 24 h, without previous treatment. The fixed meristems were washed three times in distilled water for 5 min. To remove the cell wall, washed meristems were incubated in a solution of 2% cellulase and 2% pectinase in PBS for 70 min at 37°C, followed by two washes with cold distilled water. Slides were prepared using the flame-drying method; meristems were macerated in a drop of cold 3:1 ethanol: glacial acetic acid fixative solution using fine-pointed forceps on a glass slide, which was subsequently warmed over an alcohol flame and air-dried before immediate use or storage at 4°C. An oligonucleotide probe for CUS-TR24 (5′-AGT GTC ACA AAT ACT TAG CCT TAT CTC TAT GAT TTA GCG TTT TCA GCG AA-3′) was labeled with fluorescein isothiocyanate (FITC) at its 5′ ends during synthesis (Integrated DNA Technologies, Leuven, Belgium). Fragments of other probes were PCR-amplified from genomic DNA of *C. europaea* and cloned into pCR4-TOPO vector (Thermo Fisher Scientific). PCR primer sequences were 5′-CCT CTT TGA TAT TGG AGA TAA TAA ATC-3′ and 5′-GGC AAG GTC ATA ATC AGC A-3′ for L1-CS_cl3, 5′-GTT TGA TAT TGG GGA TGA CAA-3′ and 5′-AAC ACC TCC CAA GAA AAT ATT AGA T-3′ for L1-CS_cl48, and 5′-AGG CAG ATC TTC CGA GGT A-3′ and 5′-AAA GTC AAG CAC AAG CAT CC-3′ for the RTE probe; the sequences of the cloned probes are available from GenBank under accession numbers MN625503, MN625506 and MN625501, respectively. These probes were labeled with biotin-16-dUTP (Roche, Mannheim, Germany) using nick translation

[22]. FISH was performed as described previously [28] with a hybridization and washing temperature of 37°C. Slides were counterstained with DAPI, mounted in Vectashield mounting medium (Vector Laboratories, Burlingame, CA, USA), and examined using a Zeiss AxioImager.Z2 microscope with an Axiocam 506 mono camera. Images were captured and processed using ZEN pro 2012 software (Carl Zeiss GmbH).

3. Results

3.1. Nanopore sequencing and initial analysis of the reads provides the first insight into the complex structure of the CUS-TR24 loci

The sequencing of high molecular weight nuclear DNA from *C. europaea* was performed on the Oxford Nanopore MinION device using a 1D ligation sequencing kit. Quality-filtered reads were pooled from two independent sequencing runs and filtered for a minimum length of 30 kb, resulting in the selection of 96,528 reads for further analysis. Selected reads were up to 239 kb in length (N50 = 56.9 kb) and represented 5.9 Gbp of sequence data (5-fold coverage of 1169 Gb/1C *C. europaea* genome [33]).

Initial sequence analysis of randomly selected reads containing CUS-TR24 sequences was performed using self-similarity dot-plots to investigate their internal structure (Fig. 2). The dot-plots revealed that these reads had a complex and variable structure composed of arrays of tandemly repeated CUS-TR24 monomers frequently interrupted with short regions of simple sequence repeats (SSRs). These SSRs were mostly (TAA)n motifs, confirming our previous finding, from FISH experiments, that (TAA)n repeats co-localize with CUS-TR24 [36]. In addition, other, less frequent motifs were detected, including a (TGA)n motif and irregular repetitions of SSR-like sequences. It was also evident from the dot-plots that CUS-TR24 arrays are often interrupted with common sequences identified as fragments of mobile elements with multiple copies found within and between reads (Fig. 2). Structure of these elements and sequences of their open reading frames coding for reverse transcriptase and endonuclease proteins led to their classification as LINE retrotransposons.

3.2. Computational analysis of all nanopore reads reveals a general pattern of sequence arrangement in CUS-TR24-containing heterochromatin

To investigate if the patterns uncovered by the dot-plot analysis of selected reads represented general features of the genomic loci containing CUS-TR24 repeats, we performed a computational analysis of their properties across the whole set of nanopore reads. A reference database containing a representative set of CUS-TR24, SSRs and LINE sequences was assembled and used to identify regions containing these repeats in individual nanopore reads.

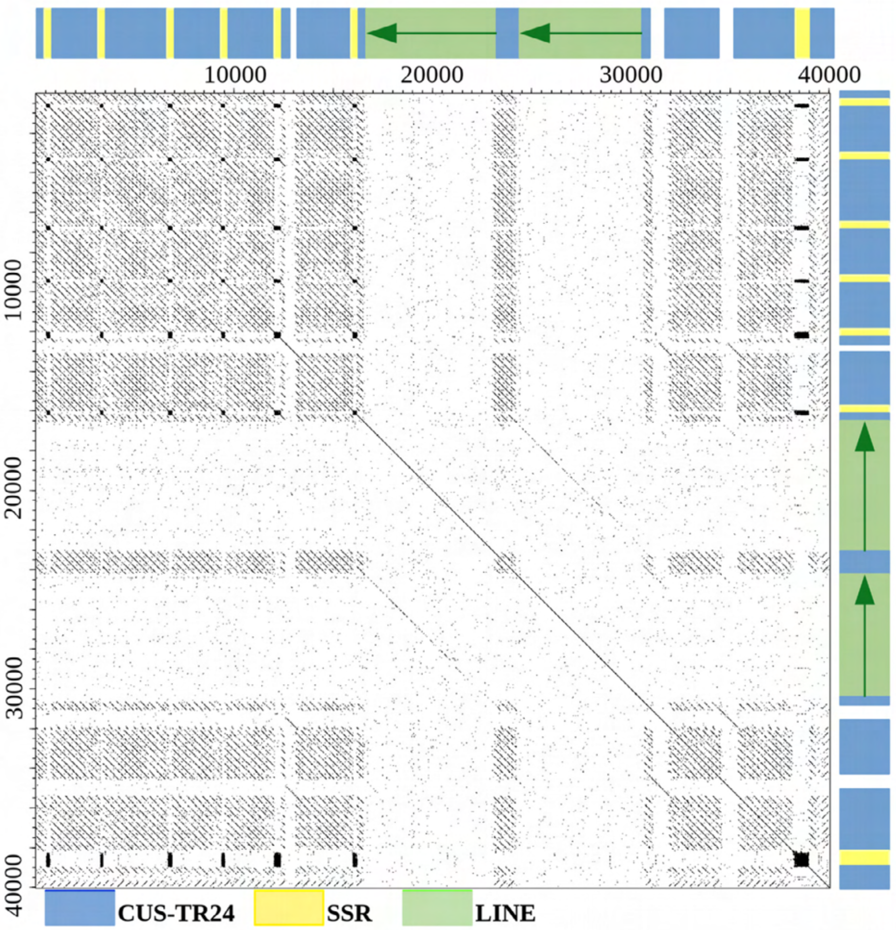


Fig. 2. Sequence organization of CUS-TR24 loci revealed by self-similarity dot-plot analysis of individual nanopore reads. A typical sequence arrangement is demonstrated here on a dot-plot from a 40 kb portion of a 98 kb read. Sequence annotation within the read is provided along the dot-plot axes, with colored rectangles representing CUS-TR24 satellite arrays (blue), SSRs (yellow) and LINEs (green, with the arrow showing the 5′→3′ orientation). Dot-plot of the entire read and additional dot-plot examples are provided in Supplementary Fig. S1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The lengths of these regions and their mutual interspersion were then evaluated. The reference database was also supplemented with other *C. europaea* tandem repeats including the abundant CUS-TR2 satellite and 45S rDNA sequences, and by representatives of major groups of mobile elements that were previously characterized from the *C. europaea* genome [33] .

The analysis of the length distribution of tandem repeat arrays revealed remarkable differences between the investigated repeats. The array length distributions were visualized as weighted histograms with a bin size of 5 kb, distinguishing complete and truncated satellite arrays (Fig. 3). While 45S rDNA and CUS-TR2 sequences were almost exclusively present as long contiguous arrays of up to hundreds of kilobases that extended beyond the lengths of most reads (Fig. 3A,B), CUS-TR24 arrays were much shorter, with over 96% of them not exceeding 10 kb (Fig. 3C). A detailed plot of CUS-TR24 array length distribution showed a series of peaks ranging from ~200 bp to 4 kb (Fig. 3D). The occurrence of these peaks and their spacing suggested that the CUS-TR24 arrays are not terminated at random positions but instead differ by multiples of the consensus monomer length (389 bp). The observed pattern of prominent peaks interlaced by smaller ones could then be explained by the presence of two variants of array termination in the genome: the prominent peaks represented arrays containing multiple complete monomers terminated by a truncated monomer sequence of ~120 bp, while a series of smaller peaks corresponded to multiples of full-length monomers (Fig. 3D). The size distribution of SSR arrays (Fig. 3E) did not show any regular pattern and were mostly of a short length (<400 bp).

Next, we investigated patterns of interspersion of CUS-TR24 sequences with other repeats by examining the presence and orientation of repeats within 10 kb regions directly adjacent to each CUS-TR24 array. Results were pooled from all reads, and the frequencies at which different repeats were associated with CUS-TR24 arrays were summarized (Fig. 4A). This analysis revealed that about 40% of CUS-TR24 arrays are terminated by short SSR repeats (30% in forward and 10% in reverse orientation with respect to the CUS-TR24 arrays). However, their broader neighborhood (1–10 kb) was most frequently (40–45%) occupied by another CUS-TR24 array in the same orientation, while CUS-TR24 sequences in the opposite orientation were less frequent (10–15%). Up to 20% of CUS-TR24 arrays were directly adjacent to LINE elements, with the LINE elements frequently in reverse orientation to the CUS-TR24 consensus. Similar analysis of LINE elements revealed that up to 50% of the genome regions directly adjacent to LINE sequences consisted of CUS-TR24 in a reverse orientation (Fig. 4B). SSR arrays were found to be similarly surrounded by CUS-TR24 sequences and, to a lesser extent, by further SSR sequences (Fig. 4C). The distinct peaks evident in the CUS-TR24 and SSR density plots reflect the interlaced pattern of these repeats, with SSRs separated by CUS-TR24 arrays of various lengths that differ by multiples of CUS-TR24 monomer size (Fig. 4C). In contrast to CUS-TR24, another highly amplified satellite, CUS-TR2, did not show preferential association with other repetitive sequences (Fig. 4D), consistent with the observation that this satellite usually forms long, homogeneous arrays (Fig. 3B).

3.3. *CUS-TR24 sequences are interspersed with a specific lineage of LINEs due to its insertional target site preference*

The observed association of LINEs with CUS-TR24 arrays prompted us to perform detailed characterization of these sequences in the *C. europaea* genome. Using previously published data on repeat variation in *Cuscuta* [33] , we defined three major LINE element groups in the *C. europaea* genome. These groups corresponded to sequence clusters or super-clusters generated by the

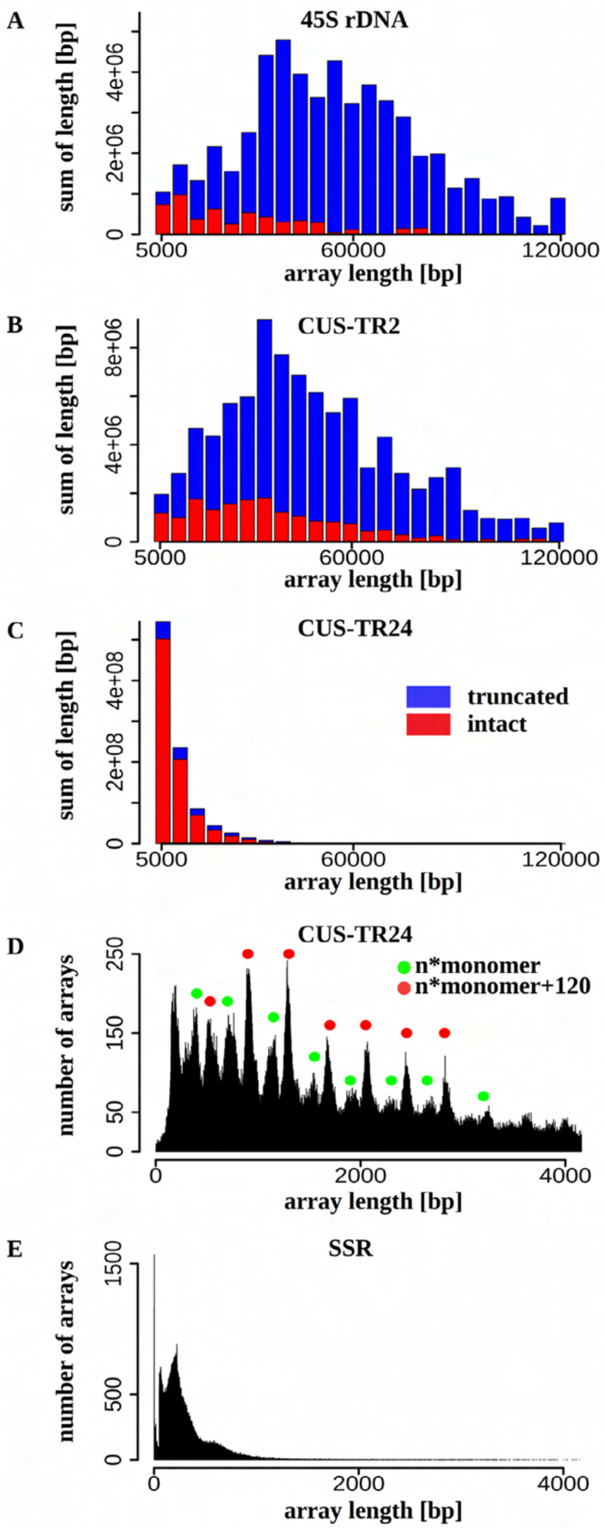


Fig. 3. Length distribution of the satellite repeat arrays. (A-C) The lengths of the arrays detected in nanopore reads are displayed as weighted histograms with a bin size of 5 kb; the last bin includes all arrays longer than 120 kb. Arrays completely embedded within a read (red bars) are distinguished from truncated arrays positioned at the end of a read (blue bars). Due to array truncation, the latter values are underestimation of the lengths of corresponding genomic arrays and should be considered as lower bounds of the respective array lengths. (D-E) The distribution of CUS-TR24 and SSR array length plotted in 1 bp resolution. The formulas provided in (D) explain the prevalent array lengths represented by the peaks marked with corresponding colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

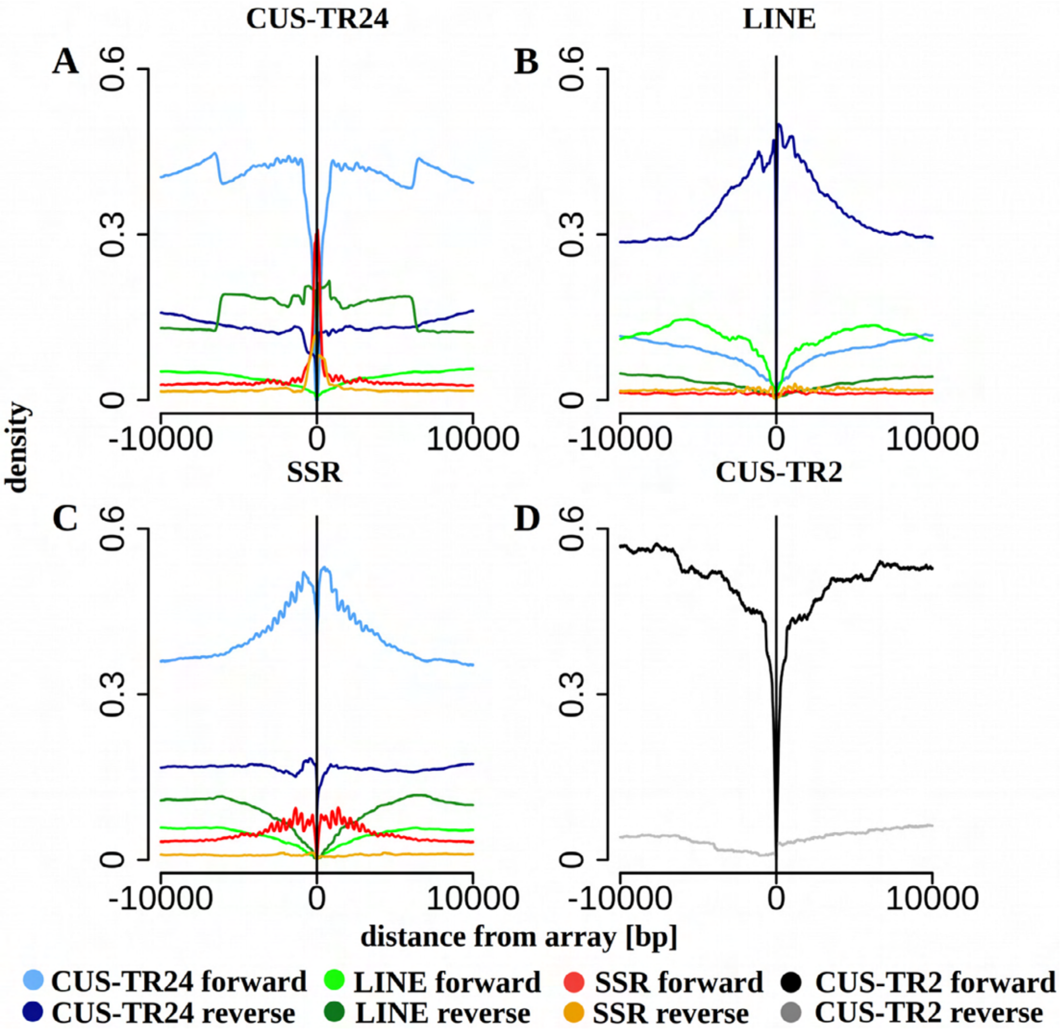


Fig. 4. Sequence composition of genomic regions adjacent to CUS-TR24 arrays (A), LINE elements (B), SSRs (C) and CUS-TR2 arrays (D). The plots show the proportions of repetitive sequences identified within 10 kb regions upstream (positions −1 to −10,000) and downstream (1 to 10,000) of the arrays of tandem repeats (A, C, D) or insertion of LINE elements (B). The vertical line shows the array or LINE position, and the plots are relative to the forward-oriented sequences. Only the repeats detected in proportions exceeding 0.05 are plotted (colored lines).

similarity-based repeat clustering of Illumina reads employing the RepeatExplorer pipeline [35] . Representative full-length elements were reconstructed for each group using consensus sequences produced by the RepeatExplorer. The structure of these elements is provided in Fig. 5A, showing positions of the regions coding for reverse transcriptase (RT), RNase-H (RH), and endonuclease (ENDO) protein domains. We used protein sequences obtained by conceptual translations of the RT-coding regions to assign the identified elements to the phylogenetic lineages of plant LINEs defined by Heitkam et al. [20] . A Neighbor-Joining tree constructed from multiple alignment of RT sequences sampled from various plant species [20] revealed that the three groups of *C. europaea* LINEs belong to the three major branches of the tree, representing L1 LINE-CS (L1-CS), L1-Llb, and RTE lineages (Fig. 5B). Repeat clustering data estimated the proportions of these lineages in the *C. europaea* genome to be 4.26%, 0.08%, and 0.45%, respectively.

We re-analyzed the nanopore read data taking the classification of LINE element by lineage into account, examining LINE sequences located in proximity (up to 10 kb) to CUS-TR24 arrays and comparing them with all remaining LINE elements detected in the nanopore reads. This analysis revealed that 91% of L1-CS elements were associated with CUS-TR24 arrays, while the other two lineages showed no such strong association (Table 1). To verify these results, we designed hybridization probes for L1-CS and RTE sequences and visualized their distribution on metaphase chromosomes of *C. europaea* using FISH (the L1-Llb elements were not examined due to their low proportion in the genome). Two different probes were used for L1-CS to account for sequence variation among these elements. The FISH signals of both probes were much stronger in the DAPI-positive heterochromatic bands than in the euchromatic chromosome regions (Fig. 6A,B). In addition, only bands known to contain CUS-TR24 repeats were strongly labeled, while a band on chromosome 1 consisting of CUS-TR2 (Fig. 1) lacked these strong FISH signals. Conversely, the RTE probe generated labeling patterns that were uniformly scattered along whole chromosomes (Fig. 6C), suggesting that these elements are evenly dispersed in the genome. These experiments thus confirmed that CUS-TR24 loci are specifically enriched with LINEs of the L1-CS lineage.

The specific association of L1-CS elements with CUS-TR24 repeats prompted us to investigate if this association might result from an insertional preference for this LINE lineage. LINEs insert into the genome via target-site primed reverse transcription, generating target site duplication (TSD) upon their insertion [29] . Selective insertional targeting to specific sequence motifs has been described for some LINE families [7] . If this mechanism was also

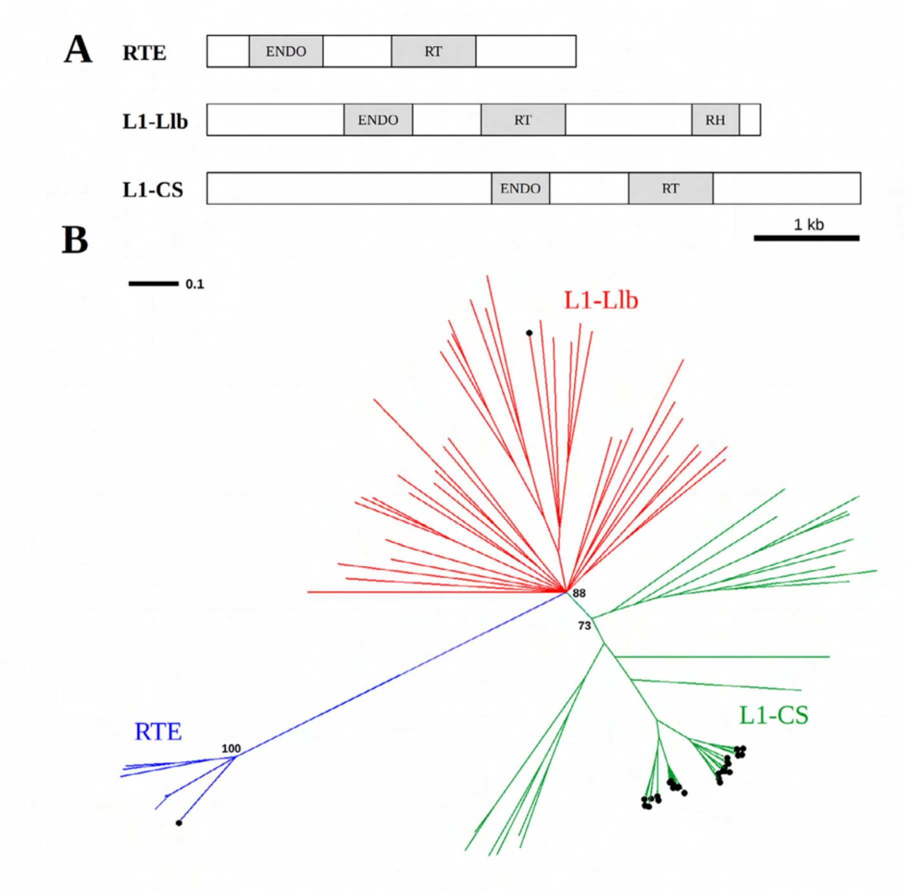


Fig. 5. Structure of the reconstructed consensus sequences representing three distinct LINE element groups identified in the *C. europaea* genome (A). Groups were assigned to phylogenetic lineages defined by Heitkam et al. [20] according to similarities of their RT domain sequences (B). The branches of the neighbor-joining tree labeled with circles represent RT sequences extracted from *C. europaea* elements. The remaining branches represent reference sequences collected from various plant species [20] . Bootstrap values are provided for the major nodes and the scale bar indicates numbers of changes per site.

Table 1
Estimated proportions of LINE elements associated with CUS-TR24 repeats.

Lineage	Associated with CUS-TR24		Elements
	YES	NO	scored
L1-CS	91%	9%	97,860
L1-Llb	32%	68%	7302
RTE	14%	86%	21,645

functional in the *C. europaea* L1-CS elements, it could explain the observed interspersion patterns, if the LINE target sequence was conserved in CUS-TR24 monomers. Indeed, the mapping of LINE insertions with respect to the CUS-TR24 consensus monomer showed clear preference for two adjacent regions of the monomer (Figs. 7 and 8A). These two insertion sites had consensus sequences of 5′-TTCTA-3′ and 5′-TTTCAA-3′, similar to the cannonical cleavage site of mammalian L1 elements (5′-TTTTAA-3′) [47] .

3.4. The model for the origin of the complex structure of CUS-TR24 loci

Taken together, our findings indicated that the complex sequence arrangement of heterochromatic loci containing CUS-TR24 repeats resulted from a combined action of several processes, outlined in Fig. 8B. It appears that the nucleotide sequence of the CUS-TR24 monomer played a crucial role in these processes by providing target sequences for the L1-CS element insertions and hotspots for the emergence of SSR arrays (Fig. 8A). In the proposed model, we presume that ancestral arrays of CUS-TR24 were amplified in the *C. europaea* genome (Fig. 8B). The frequent occurrence of (TAA) motifs within the monomer sequences (highlighted in Fig. 8A) provided a template for their occasional conversion and/or expansion into SSR arrays, possibly via the replication strand slippage mechanism known to generate microsatellite sequences [43]. The AT-rich sequences may also constitute fragile sites that are prone to DNA breakage and structural rearrangements [2] . Our detailed inspection of the CUS-TR24/SSR boundaries in multiple nanopore reads revealed that the presence of expanded (TAA)n motifs within CUS-TR24 arrays was frequently associated with the truncation of neighboring monomer sequences (Fig. 8A). The length of truncated monomers varied between ~120–150 bp, which roughly corresponds to the observed size distribution pattern of CUS-TR24 arrays (Fig. 3D), consisting of multiple full-length monomers terminated by the truncated monomer sequence of ~120 bp.

Concurrent with the emergence of SSRs, the CUS-TR24 monomers were specifically targeted by L1-CS lineage LINEs (Fig. 8B). Since these CUS-TR24-associated LINEs are relatively heterogeneous in their nucleotide sequences (Supplementary Fig. S2 and Fig. 2), it is likely that they originated from the retrotransposition of multiple master elements. Finally, the CUS-TR24 loci were probably shaped by additional processes including segmental duplications, inversions (both are evident from the dot-plot analysis; Supplementary Fig. S1), and possibly recombination-based deletions, resulting in the present complex structure of these loci.

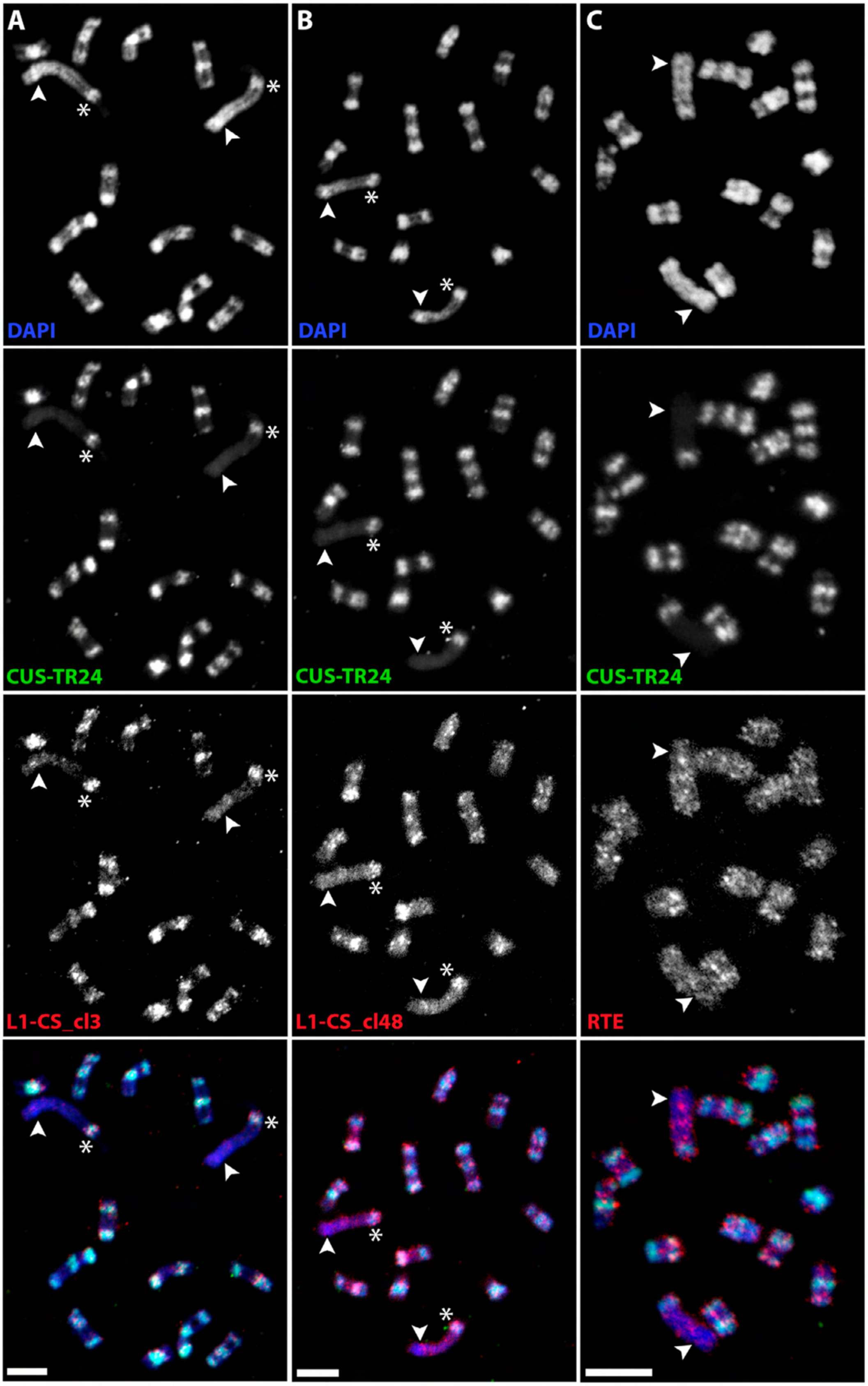


Fig. 6. Distribution of LINE sequences on metaphase chromosomes of *C. europaea*. Two-color FISH experiments were performed to detect LINEs (red channel) and CUS-TR24 sequences (green). The chromosomes were counterstained with DAPI (blue). Individual channels and corresponding merged color images are shown for experiments including LINE probes L1-CS_cl3 (A), L1-CS_cl48 (B) and RTE (C). Arrowheads mark the position of DAPI-positive heterochromatic band on the chromosome 1 that lacks CUS-TR24 repeats (for comparison, the CUS-TR24-containing band on the same chromosome that is also enriched for L1-CS LINEs is marked with asterisk). See also Fig. 1 for a schematic of this karyotype. Bar = 5 µm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

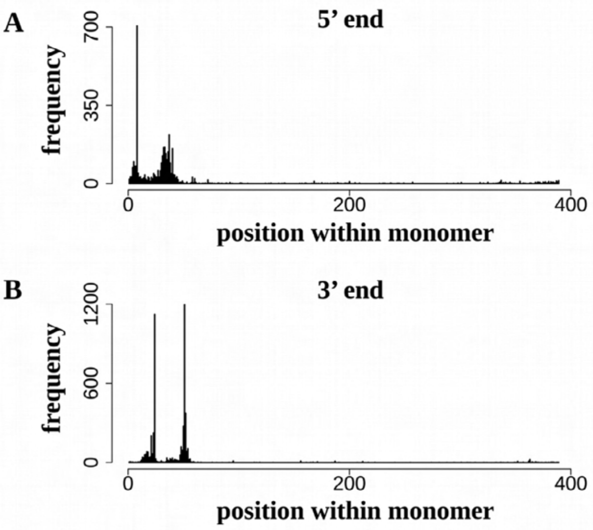


Fig. 7. Analysis of the insertional target sites of LINE elements within CUS-TR24 monomers. Plots show the frequency of 5′ (A) and 3′ ends (B) of LINEs mapped to individual positions along a CUS-TR24 consensus monomer (the monomer sequence is provided in Fig. 8A). Due to target site duplication generated upon element insertion, the mapped positions of 5′ and 3′ ends are shifted by approximately 13–16 bp. An example of CUS-TR24 sequence with LINE insertion is provided in Supplementary Fig. S3, including also target site duplication generated upon LINE insertion.

4. Discussion

In this study, we uncovered the complex sequence structure of genomic loci containing the satellite CUS-TR24, which constitute most of the heterochromatic bands on holocentric chromosomes of *C. europaea*. Satellite DNA is known to be the major component of constitutive heterochromatin in eukaryotic genomes [14,40], being supposedly arranged in long contiguous arrays that are only sparsely interrupted by random insertions of mobile elements. Such arrangement has been documented for some human and plant satellites [25,30,50], and it has also been found for the other abundant satellite family in *C. europaea*, CUS-TR2. This contrasts with the genome organization of the CUS-TR24 repeats, which are highly fragmented due to their interspersion with short SSR arrays and insertions of LINE elements. Such a complex structure was unexpected for highly amplified satellite repeat, especially considering that its amplification in *C. europaea* occurred relatively recently as judged from the absence of the CUS-TR24 repeats from closely related species *C. epithymum* [33]. On the other hand, arrays of abundant satellite repeats in maize were found to be highly fragmented by retrotransposon insertions [26]. Mobile elements were also proposed to generate complex arrangements and even facilitate genomic dispersal of satellite repeats in other species [37,42,50]. Since detailed studies of satellite repeat arrays are still scarce, it is yet to be elucidated what is the prevailing type of satDNA organization and how it is affected by various factors like the age of the arrays or their location in the genome.

To explain the origin of a complex pattern shared among CUS-TR24 loci, we considered two different scenarios, proposing that either (1) there was an ancestral, low-copy repeat composed of adjacent CUS-TR24, SSR, and LINE sequences that became amplified and spread throughout the genome as a new compound monomeric unit; or (2) the pattern resulted from ongoing and concurrent processes of CUS-TR24 amplification, the emergence of SSRs from proto-SSR units, and the insertional targeting of LINEs during their genomic proliferation.

The compound monomers proposed in the first scenario have already been described for several satellite repeats. For example, a 4.7 kb-long monomer of the *Sobo* satellite from *Solanum bulbocastanum* originated from part of an LTR-retrotransposon and a genomic tandem repeat [48]. Similar satellites with long monomers consisting of unrelated, repeated and/or low-copy genomic sequences have been described from *Solanum tuberosum* [15] and *Secale cereale* [24]. However, the monomer sequences of these satellites are highly homogenized throughout the genome, with up to 99% similarity between copies [48], and therefore the arrangement of the original sequence components is identical in all monomers. No such conserved arrangement of CUS-TR24, SSR, and LINE sequences occurs at CUS-TR24 loci, making it unlikely that they were amplified as a single conserved monomer unit. In addition, there is variation in the presence and length of the SSR arrays in CUS-TR24 monomers (Fig. 2 and Supplementary Fig. S1), and considerable LINE sequence diversity (Supplementary Fig. S2), which suggests that they do not originate from a single insertion into an ancestral satellite array. However, the association of heterogeneous LINE sequences with CUS-TR24 can be explained by recurrent retrotransposition of multiple template elements. Despite their sequence variation, these elements belong to the same phylogenetic lineage of LINEs and share insertional target sites (Fig. 7). Considering these facts, we favor the explanation provided by the second scenario, which was included into the proposed model of the evolution of CUS-TR24 loci (Fig. 8).

A notable feature of the CUS-TR24 loci is their association with the CENH3 protein [36] which serves as an epigenetic marker of active centromeres in all plant species studied so far [6,44]. However, *C. europaea* CENH3 may have lost this function: the distribution of CENH3 on chromosomes does not correlate with the attachment of the mitotic spindle [36]. It is supposed that CENH3 deposition to plant centromeres is independent of the underlying centromeric repeats; instead, it is a part of an epigenetically determined self-propagation loop based on the interactions of CENH3 chaperones and the additional constitutive centromere-associated network (CCAN) proteins [17]. However, the mechanism driving CENH3 deposition in *C. europaea* is unknown. It is possible that there is a sequence-dependent interaction between CENH3 (or its chaperone) and CUS-TR24 repeats, in a manner similar to the interaction of human centromeric protein CENP-B with a 17 bp CENP-B-box sequence within centromeric alpha satellites [10]. However, such sequence-specific deposition of CENH3 has not been reported in any plant species.

Although repeat-rich regions of the genome are generally transcriptionally silent, it has been reported that transcriptional activity at centromeric repeats plays an important role in CENH3 deposition [9,39]. In this respect, the accumulation of LINE elements in the CUS-TR24 loci may be of interest, as these elements could initiate transcription of adjacent sequences, which in turn may promote CENH3 deposition. In support of this hypothesis is the finding that LINE-L1 transcripts are an essential component of human neocentromeres [5]. LINEs represent major repeats associated with centromeric chromatin in *Drosophila* [4] and centromere-specific LINE elements have been reported in the sunflower genome [31]. Although we currently cannot provide an explanation for the observed co-localization of CENH3 with heterochromatin containing CUS-TR24 repeats, the findings discussed above warrant further investigation of kinetochore composition and centromere determination in the holocentric *Cuscuta* species.

This work provides evidence for a new type of highly complex sequence arrangement in plant constitutive heterochromatin. It also demonstrated the potential of long-read sequencing technologies to fill gaps in our knowledge of the satellite DNA-rich regions of eukaryotic genomes that are otherwise hard to investigate.

60

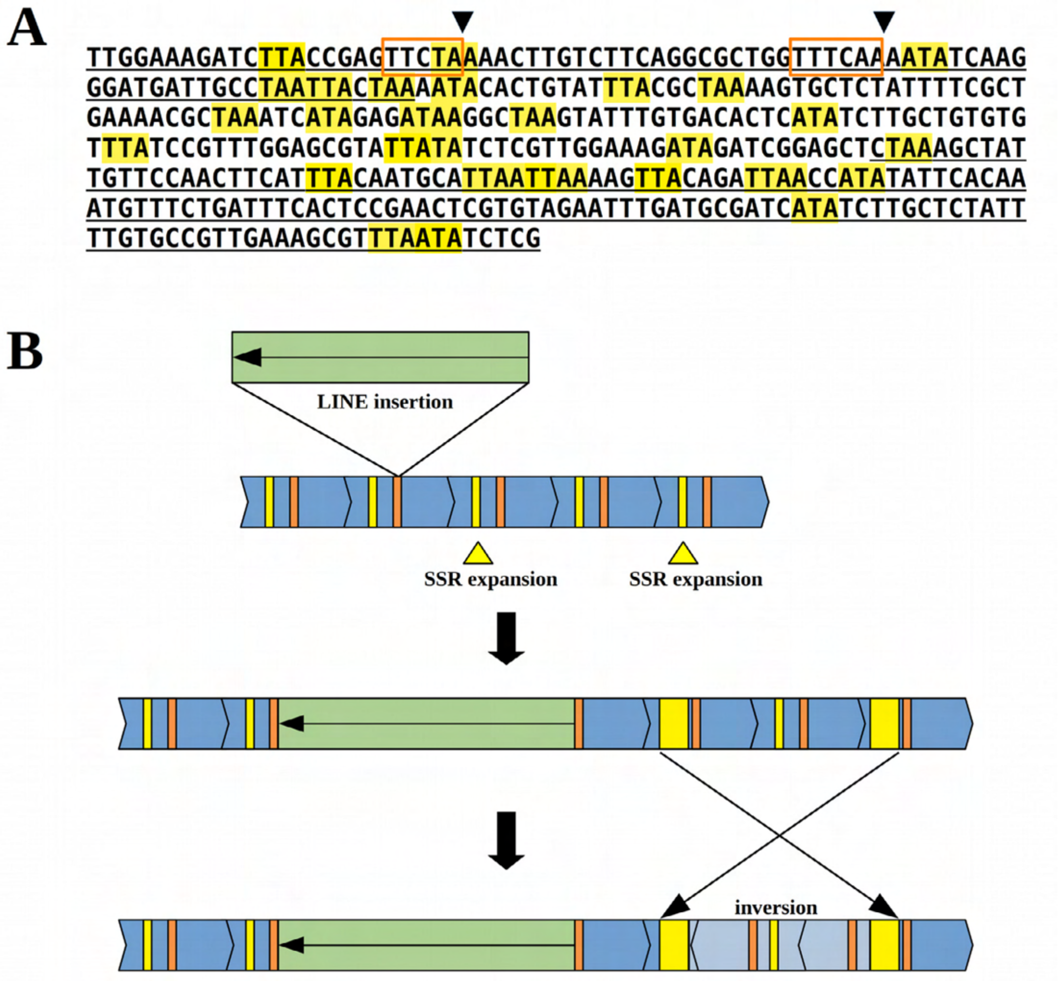

A
TTGGAAAGATCTTACCGAGTTCTAAAACTTGTCTTCAGGCGCTGGTTTCAAAATATCAAG
GGATGATTGCCTAATTACTAAAATACACTGTATTTACGCTAAAAGTGCTCTATTTTCGCT
GAAAACGCTAAATCATAGAGATAAGGCTAAGTATTTGTGACACTCATATCTTGCTGTGTG
TTTATCCGTTTGGAGCGTATTATATCTCGTTGGAAAGATAGATCGGAGCTCTAAAGCTAT
TGTTCCAACTTCATTTACAATGCATTAATTAAAAGTTACAGATTAACCATATATTCACAA
ATGTTTCTGATTTCACTCCGAACTCGTGTAGAATTTGATGCGATCATATCTTGCTCTATT
TTGTGCCGTTGAAAGCGTTTAATATCTCG
B
LINE insertion
SSR expansion
SSR expansion
inversion

Fig. 8. (A) Consensus sequence of a CUS-TR24 monomer. The target sequences for LINE insertion are marked with orange rectangles (the putative cleavage sites are marked with arrowheads). The (TAA) motifs and their variants are highlighted in yellow, and the monomer region frequently lost at CUS-TR24/SSR junctions is underlined. (B) A model to represent the processes leading to the complex structure of CUS-TR24 loci. The ancestral CUS-TR24 monomer arrays (blue) contain hotspots for SSR emergence from (TAA) motifs (yellow) and LINE target sites (orange). These arrays become fragmented by concurrent SSR expansion and insertion of new LINE elements, and undergo further rearrangements, including segmental duplications and inversions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Although the ultra-long sequence reads are mostly used to improve whole genome assemblies [30] , this work, and work previously reported [3,50] paves the way for their use in assembly-free bioinformatic approaches to provide a unique insight into the origin and structure of satellite repeats.

CRediT authorship contribution statement

Tihana Vondrak: Investigation, Formal analysis, Software, Writing – Original Draft, Writing - review & editing. **Ludmila Oliveira:** Investigation, Writing - review & editing. **Petr Novák:** Formal analysis, Writing - review & editing. **Andrea Koblížková:** Resources. **Pavel Neumann:** Formal analysis, Writing - review & editing. **Jiří Macas:** Conceptualization, Supervision, Investigation, Writing – Original Draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Ms. Vlasta Tetourová and Ms. Jana Látalová for their excellent technical assistance. This work was supported by the Czech Science Foundation (grant no. 20-24252S) and the ELIXIR CZ Research Infrastructure Project (Czech Ministry of Education, Youth and Sports; grant no. LM2018131), including access to computing and storage facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.04.011.

References

[1] Allshire RC, Madhani HD. Ten principles of heterochromatin formation and function. Nat Rev Mol Cell Biol 2018;19:229–44.
[2] Burrow AA, Marullo A, Holder LR, Wang Y-H. Secondary structure formation and DNA instability at fragile site FRA16B. Nucleic Acids Res 2010;38:2865–77.
[3] Cechova M, Harris RS, Tomaszkiewicz M, Arbeithuber B, Chiaromonte F, Makova KD. High satellite repeat turnover in great apes studied with short- and long-read technologies. Mol. Biol. Evol. 2019;36:2415–31.

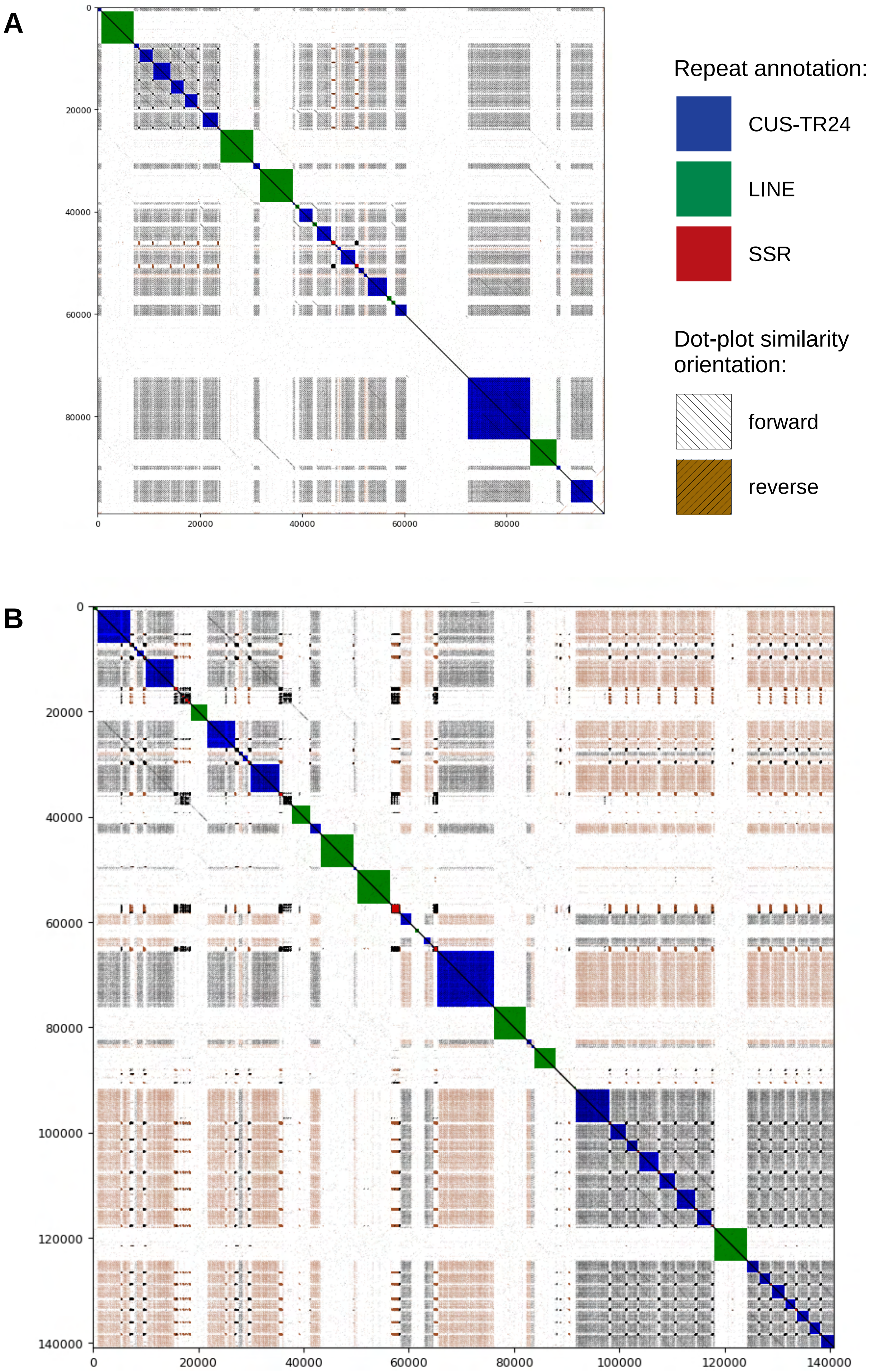
[4] Chang C-H, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, et al. Islands of retroelements are the major components of *Drosophila* centromeres. PLoS Biol 2019;17:e3000241.
[5] Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KHA, Wong LH, Bickmore WA. LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. PLoS Genet 2009;5:e1000354.
[6] Comai L, Maheshwari S, Marimuthu MP. Plant centromeres. Curr Opin Plant Biol 2017;36:158–67.
[7] Cost GJ, Boeke JD. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry 1998;37:18081–93.
[8] van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet 2018;34:666–81.
[9] Duda Z, Trusiak S, O'Neill R. Centromere Transcription: Means and Motive. In: Centromeres and Kinetochores, 56. Progress iCham: Springer; 2017. p. 257–81.
[10] Dumont M, Fachinetti D. DNA Sequences in Centromere Formation and Function. In: Centromeres and Kinetochores, 112. Cham: Springer; 2017. p. 305–36.
[11] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf 2004;5:1–19.
[12] Fuchs J, Strehl S, Brandes A, Schweizer D, Schubert I. Molecular-cytogenetic characterization of the *Vicia faba* genome–heterochromatin differentiation, replication patterns and sequence localization. Chromosome Res 1998;6:219–30.
[13] Garrido-Ramos M. Satellite DNA: An evolving topic. Genes 2017;8:230.
[14] Garrido-Ramos MA. Satellite DNA in plants: More than just rubbish. Cytogenet Genome Res 2015;146:153–70.
[15] Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, Iovene M, et al. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell 2012;24:3559–74.
[16] Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 2010;27:221–4.
[17] Hara M, Fukagawa T. Critical Foundation of the Kinetochore: The Constitutive Centromere-Associated Network (CCAN). In: Centromeres and Kinetochores, 112. Cham: Springer; 2017. p. 29–57.
[18] Harris RS. Improved Pairwise Alignment of Genomic DNA. University Park, PAUnited States: Pennsylvania State University; 2007ISBN:978-0-549-43170-1.
[19] Heckmann S, Macas J, Kumke K, Fuchs Jörg, Schubert V, Ma L, et al. The holocentric species *Luzula elegans* shows interplay between centromere and large-scale genome organization. Plant J 2013;73:555–65.
[20] Heitkam T, Holtgräwe D, Dohm JC, Minoche AndréE, Himmelbauer H, Weisshaar B, et al. Profiling of extensively diversified plant LINEs reveals distinct plant-specific subclades. Plant J 2014;79:385–97.
[21] Jiang J. Fluorescence in situ hybridization in plants: recent developments and future applications. Chromosom Res 2019;27:153–65.
[22] Karafiátová M, Bartoš J, Doležel J. Localization of Low-Copy DNA Sequences on Mitotic Chromosomes by FISH. Methods Mol Biol 2016;1429:49–64.
[23] Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 2007;23:1026–8.
[24] Langdon T, Seago C, Jones RN, Ougham H, Thomas H, Forster JW, et al. De novo evolution of satellite DNA on the rye B chromosome. Genetics 2000;154:869–84.
[25] Lee H-R, Neumann P, Macas J, Jiang J. Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice. Mol Biol Evol 2006;23:2505–20.
[26] Liu J, Seetharam AS, Chougule K, Ou S, Swentowsky KW, Gent JI, et al. Gapless assembly of maize chromosomes using long-read technologies. Genome Biol 2020;21:121.
[27] Macas J, Meszaros T, Nouzova M. PlantSat: a specialized database for plant satellite repeats. Bioinformatics 2002;18:28–35.
[28] Macas J, Neumann P, Navrátilová A. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. BMC Genomics 2007;8:1–16.
[29] Martin SL, Li W-L-P, Furano AV, Boissinot S. The structures of mouse and human L1 elements reflect their insertion mechanism. Cytogenet Genome Res 2005;110:223–8.
[30] Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature 2020;585:79–84.
[31] Nagaki K, Tanaka K, Yamaji N, Kobayashi H, Murata M. Sunflower centromeres consist of a centromere-specific LINE and a chromosome-specific tandem repeat. Front Plant Sci 2015;6:1–12.
[32] Neumann P, Novák P, Hoštáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob DNA 2019;10:1–17.
[33] Neumann P, Oliveira L, Čížková J, Jang T, Klemme S, Novák P, et al. Impact of parasitic lifestyle and different types of centromere organization on chromosome and genome evolution in the plant genus *Cuscuta*. New Phytol 2021;229:2365–77.
[34] Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res 2017;45:1–10.
[35] Novák P, Neumann P, Macas J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nat Protoc 2020;15:3745–76.
[36] Oliveira L, Neumann P, Jang T-S, Klemme S, Schubert V, Koblížková A, et al. Mitotic Spindle Attachment to the Holocentric Chromosomes of *Cuscuta europaea* Does Not Correlate With the Distribution of CENH3 Chromatin. Front Plant Sci 2020;10:1799.
[37] Paço A, Adega F, Chaves R. LINE-1 retrotransposons: from 'parasite' sequences to functional elements. J Appl Genet 2015;56:133–45.
[38] Peona V, Weissensteiner MH, Suh A. How complete are "complete" genome assemblies?-An avian perspective. Mol Ecol Resour 2018;18:1188–95.
[39] Perea-Resa C, Blower MD. Centromere Biology: Transcription Goes on Stage. Mol Cell Biol 2018;38.
[40] Plohl M, Luchetti A, Meštrović N, Mantovani B. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 2008;409:72–82.
[41] Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep 2016;6:1–14.
[42] Satović E, Vojvoda Zeljko T, Luchetti A, Mantovani B, Plohl M. Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. BMC Genomics 2016;17:997.
[43] Schlotterer C. Evolutionary dynamics of microsatellite DNA. Chromosoma 2000;109:365–71.
[44] Schubert V, Neumann P, Marques A, Heckmann S, Macas J, Pedrosa-Harand A, et al. Super-Resolution Microscopy Reveals Diversity of Plant Centromere Architecture. Int J Mol Sci 2020;21:3488.
[45] Seibt KM, Schmidt T, Heitkam T. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. Bioinformatics 2018;34:3575–7.
[46] Sonnhammer ELL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 1995;167:GC1–GC10.
[47] Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. Molecular archeology of L1 insertions in the human genome. Genome Biol 2002;3:1–18.
[48] Tek AL, Song J, Macas J, Jiang J. Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences. Genetics 2005;170:1231–8.
[49] Vanrobays E, Thomas M, Tatout C. Heterochromatin Positioning and Nuclear Architecture. In: Annual Plant Reviews, 46. Chichester, UK: John Wiley & Sons, Ltd.; 2013. p. 157–90.
[50] Vondrak T, Ávila Robledillo L, Novák P, Koblížková A, Neumann P, Macas J. Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. Plant J 2020;101:484–500.

Supplementary information

Complex sequence organization of heterochromatin in the holocentric plant *Cuscuta europaea* elucidated by the computational analysis of nanopore reads.

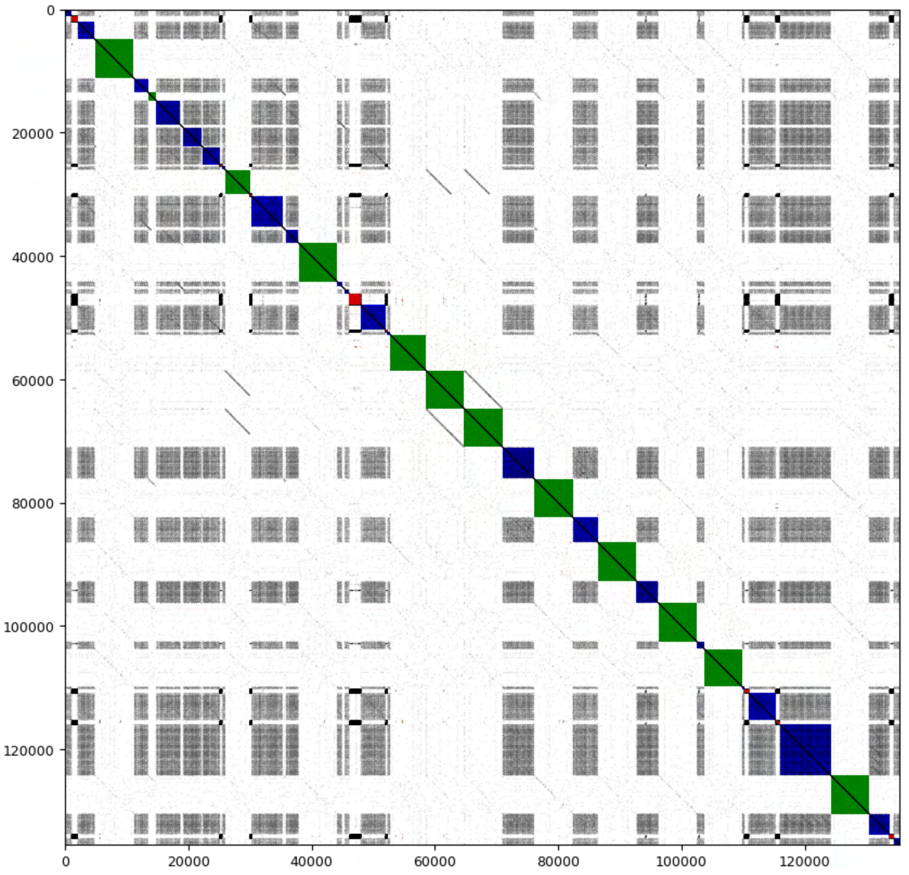
Tihana Vondrak, Ludmila Oliveira, Petr Novák, Andrea Koblížková, Pavel Neumann, Jiří Macas

Supplementary Fig. S1 A,B

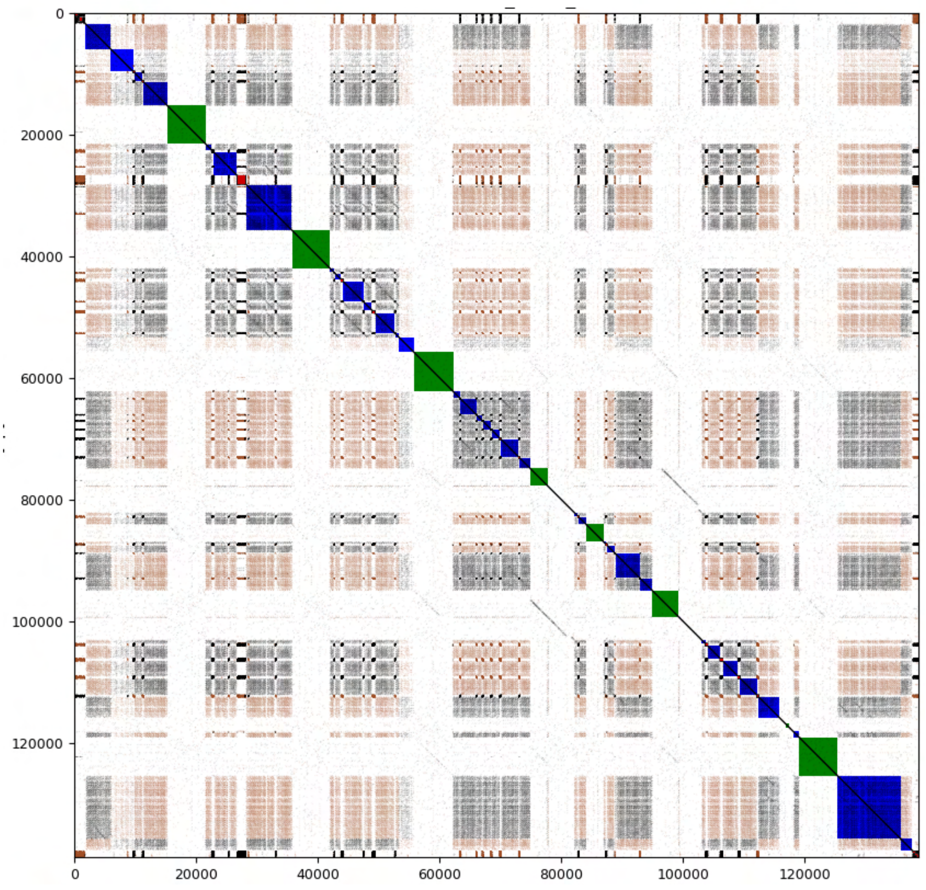


Supplementary Fig. S1 C,D

C

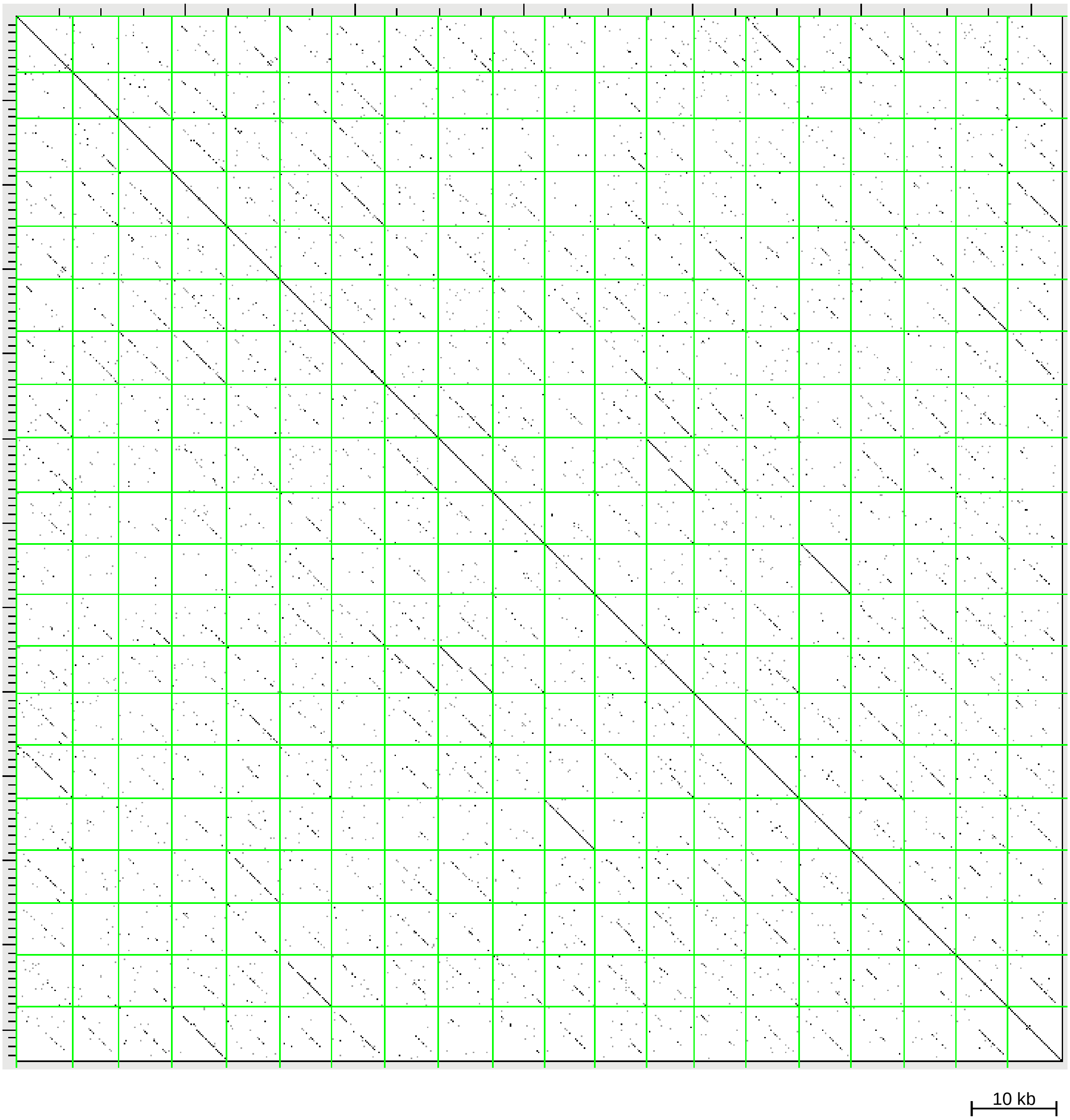


D



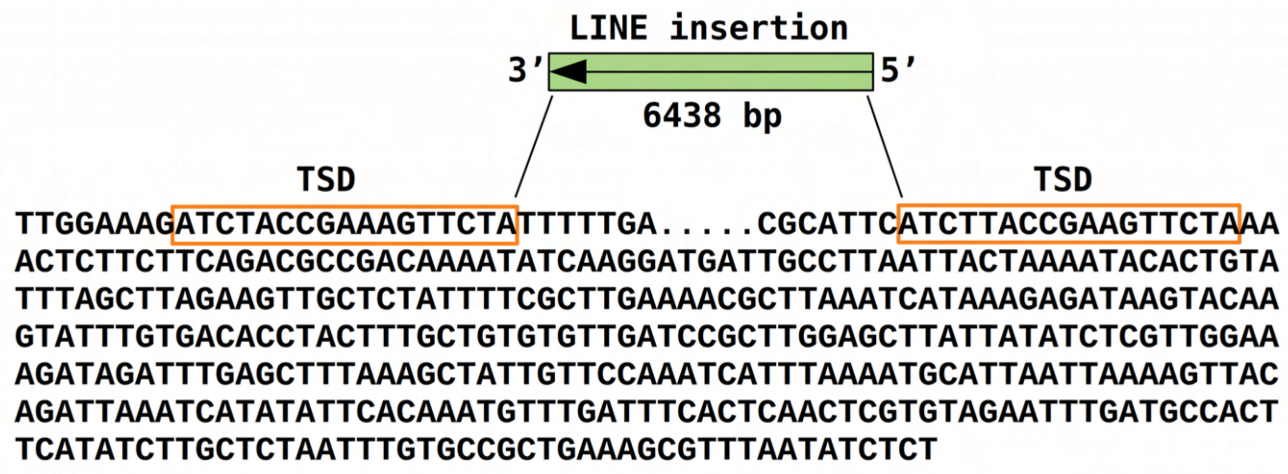
Supplementary Fig. S1. Examples of self-similarity dot-plots of individual nanopore reads (A-D) containing CUS-TR24 arrays. The plots were generated and annotated using FlexiDot (Seibt et al., 2018). The read used to generate dot-plot for Fig. 2 is shown on panel (A).

Supplementary Fig. S2



Supplementary Fig. S2. Sequence similarities of L1-CS LINE elements. The dot-plot shows all-to-all sequence comparison of twenty randomly sampled elements (the elements are separated by green lines). The similarities were scored within a sliding window of 100 bp and dots or lines were drawn when at least 80 matching bases were detected.

Supplementary Fig. S3


LINE insertion
3’
5’
6438 bp
TSD
TSD
TTGGAAAGATCTACCGAAAGTTCTATTTTTGA.....CGCATTCATCTTACCGAAGTTCTAAA
ACTCTTCTTCAGACGCCGACAAAATATCAAGGATGATTGCCTTAATTACTAAAATACACTGTA
TTTAGCTTAGAAGTTGCTCTATTTTCGCTTGAAAACGCTTAAATCATAAAGAGATAAGTACAA
GTATTTGTGACACCTACTTTGCTGTGTGTTGATCCGCTTGGAGCTTATTATATCTCGTTGGAA
AGATAGATTTGAGCTTTAAAGCTATTGTTCCAAATCATTTAAAATGCATTAATTAAAAGTTAC
AGATTAAATCATATATTCACAAATGTTTGATTTCACTCAACTCGTGTAGAATTTGATGCCACT
TCATATCTTGCTCTAATTTGTGCCGCTGAAAGCGTTTAATATCTCT

Supplementary Fig. S3. Example of the target site duplication (TSD) generated upon insertion of LINE element into CUS-TR24 monomer. The sequence was retrieved from a nanopore read, therefore the CUS-TR24 monomer sequence differs from the consensus provided in Fig. 8A. Only 5’ and 3’ terminal sequences are shown for the LINE element.

Chapter III

All around centromeres: repeat-based holocentromeres influence genome architecture and karyotype evolution

(Accepted manuscript)

All around centromeres: Repeat-based holocentromeres influence genome architecture and karyotype evolution

(Accepted manuscript)

Author list:

Paulo G. Hofstatter[1,‡], Gokilavani Thangavel[1,‡], Thomas Lux[2,‡], Pavel Neumann[3], Tihana Vondrak[3,4], Petr Novak[3], Meng Zhang[1], Lucas Costa[5], Marco Castellani[1], Alison Scott[1], Helena Toegelová[6], Joerg Fuchs[7], Yennifer Mata-Sucre[5], Yhanndra Dias[5], André L. L. Vanzela[8], Bruno Huettel[9], Cicero C. S. Almeida[10], Hana Šimková[6], Gustavo Souza[5], Andrea Pedrosa-Harand[5], Jiri Macas[3], Klaus F. X. Mayer[2,11], Andreas Houben[7] & André Marques[1,12,*]

Affiliations:

[1]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, NRW, 50829, Germany

[2]Plant Genome and Systems Biology, German Research Center for Environmental Health, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

[3]Biology Centre, Czech Academy of Sciences, Institute of Plant Molecular Biology, České Budějovice, 37005 Czech Republic

[4]Faculty of Science, University of South Bohemia, České Budějovice 37005 Czech Republic

[5]Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Centre of Biosciences, Federal University of Pernambuco, Recife, Pernambuco, 50670-901 Brazil

[6]Institute of Experimental Botany of the Czech Academy of Sciences, Centre of Plant Structural and Functional Genomics, Olomouc, 779 00, Czech Republic

[7]Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Saxony-Anhalt, 06466, Germany

[8]Laboratory of Cytogenetics and Plant Diversity, State University of Londrina, 86097-570, Paraná, Brazil

[9]Max Planck Genome-Centre Cologne, Max Planck Institute for Plant Breeding Research, Cologne, NRW, 50829, Germany

[10]School of Agronomical Sciences, Campus Arapiraca, Federal University of Alagoas, Arapiraca, 57309-005, Brazil

[11]School of Life Sciences Weihenstephan, Technical University of Munich, Alte Akademie 8, 85354 Freising, Germany

[12]Lead contact

[‡]These authors contributed equally to the work

*Corresponding author: amarques@mpipz.mpg.de

Summary

The centromere represents a single region in most eukaryotic chromosomes. However, several plant and animal lineages assemble holocentromeres along the entire chromosome length. Here, we compare genome organization and evolution as a function of centromere type by assembling the first chromosome-scale holocentric genomes with repeat-based holocentromeres from three beak-sedges (*Rhynchospora pubera*, *R. breviuscula*, and *R. tenuis*) and their closest monocentric relative, *Juncus effusus*. We demonstrate that transition to holocentricity affected 3D genome architecture by redefining genomic compartments, while distributing centromere function to thousands of repeat-based centromere units genome-wide. We uncover a complex genome organization in *R. pubera* that hides its unexpected octoploidy and describe a marked reduction in chromosome number for *R. tenuis*, which has only two chromosomes. We show that chromosome fusions, facilitated by repeat-based holocentromeres, promoted karyotype evolution and diploidization. Our study thus sheds light on several important aspects of genome architecture and evolution influenced by centromere organization.

Keywords: spatial genome organization, genome regulation, holocentric chromosomes, *Rhynchospora*, whole-genome duplication, dysploidy

Introduction

Most eukaryotes are monocentric, meaning that their centromeres are restricted to single regions on each chromosome. These centromeric regions can range from kilobases (kb) to megabases (Mb) in length and comprise specific repeats (Gohard et al., 2014). Holocentromeres, by contrast, consist of multiple centromeric units distributed along the poleward surface of metaphase chromosomes, extending from one telomere to the other, and are thus typically visible as a line on each chromatid (Heckmann et al., 2013; Senaratne et al., 2021; Steiner and Henikoff, 2014). Holocentromeres are hypothesized to stabilize chromosomal fragments and fusions that favor karyotype rearrangements and speciation (Mandrioli and Manicardi, 2020), directly influencing chromosome evolution (Schubert and Lysak, 2011). This hypothesis is supported by the fact that holocentromeres have evolved independently several times in different plant and animal lineages (Escudero et al., 2016; Melters et al., 2012).

Aside from their function in cell division, centromeres have an evolutionarily conserved role in determining large-scale genome architecture and chromatin composition (Muller et al., 2019). Centromeres in monocentric chromosomes influence the distribution of genes, eu- and heterochromatin-specific post-translational histone modification domains, transposable elements, and meiotic crossovers (Fernandes et al., 2019; Fuchs et al., 2006; Muller et al., 2019; Naish et al., 2021). However, genome organization and chromatin composition of organisms with holocentric chromosomes is poorly understood, and it is likely that holocentric species differ markedly from the monocentric paradigm.

The beak-sedge *Rhynchospora pubera* (Cyperaceae, sedges) has repeat-based holocentromeres (Marques et al., 2015), as do other species from the same genus (Costa et al., 2021; Ribeiro et al., 2017). *R. pubera* holocentromeres are associated with a single tandem repeat family (the centromeric 172-bp unit *Tyba* repeat) and the centromeric retroelement of *Rhynchospora* (*CRRh*), giving rise to thousands of small centromere units across the genome (Marques et al., 2015). The lack of a *Rhynchospora* reference genome has, however, hampered detailed studies about its intriguing centromere organization.

Here, we combined genomic and chromatin analyses to elucidate genomic adaptations related to different centromere organizations. We report the full characterization of a holocentric genome containing thousands of repeat-based centromere units. We show that this centromere organization influences the 3D genome architecture by redefining the extent of genomic compartments due to the lack of centromere clustering. Strikingly, despite substantial genome restructuring, the epigenetic regulation of centromere units in beak-sedges resembles that of monocentric centromeres, as in *Arabidopsis thaliana* (Naish et al., 2021). This observation suggests evolutionarily conserved epigenetic regulation of repeat-based centromeres in both monocentric and holocentric organisms. We further reveal that chromosome fusions facilitated by repeat-based holocentromeres reduce chromosome number and can act as an alternative to diploidization after genome doubling without the need for genome downsizing. Our work sheds light on the role of centromeres in overall genome organization and chromosome evolution.

Results

Holocentricity affects spatial genome organization

To identify the genomic adaptations related to the transition to holocentricity, we constructed chromosome-scale reference genomes using PacBio HiFi sequencing and Dovetail Omni-C (DNase-based Hi-C) for three holocentric *Rhynchospora* species, *R. pubera* (*n*=5; 1C=1.61 Gb), *R. breviuscula* (*n*=5; 1C=415 Mb), and *R. tenuis* (the plant with the fewest known chromosomes; *n*=2; 1C=394 Mb) (Castiglione and Cremonini, 2012; Vanzela et al., 1996), as well as their closest monocentric relative, the rush *J. effusus* (*n*=21; 1C=271 Mb) (Guerra et al., 2019) (**Figure 1; Figure 2; Figure S1A–B; Table S1; STAR Methods**).

J. effusus showed a typical monocentric configuration of chromatin interaction within A (euchromatin) and B (heterochromatin) compartments, including some degree of a telomere-to-centromere axis (**Figure 2A–B**) (see (Hoencamp et al., 2021).

The concept of chromosome arms does not apply to holocentric species, as centromeres are ubiquitous. Consequently, we observed no large-scale compartmentalization or telomere-to-centromere axes, as evidenced by the chromatin configuration capture (Hi-C) contact matrices of our three *Rhynchospora* species (**Figure 2C–D; Figure S1A–B**). Further quantification of intrachromosomal (*cis*) and interchromosomal (*trans*) chromatin contacts revealed a significantly higher ratio ($p < 4.04e\text{-}05$) of *cis* versus *trans* interactions in all *Rhynchospora* species compared to the monocentric *J. effusus* (**Figure S1C**). Thus, holocentric beak-sedges are characterized by higher intra-chromosomal spatial genome organization and lack of centromere clustering.

The distribution of genomic features differed markedly between holocentric *Rhynchospora* and monocentric *J. effusus* (**Figure 2E–F**). *Rhynchospora* had a uniform distribution of genes, transcriptional activity, *Tyba* centromeric repeats, transposable elements (TEs), and DNA methylation (**Figure 2F; Figure S1D–E**). By contrast, *J. effusus* genes were concentrated toward telomeric regions, while TEs and tandem repeats clustered towards centromeric regions (**Figure 2E**). Genome-wide gene distribution and transcriptional activity were positively correlated, while repeat distribution was positively correlated with overall DNA methylation levels (**Figure 2E**). Genome-wide CpG methylation (mCpG) was lower in *R. pubera* than in *J. effusus*, whereas CHG methylation was higher and CHH methylation was the same in both species (**Figure S1F–G**). Thus, transition to holocentricity likely affects 3D genome architecture by redefining the extents of genomic compartments and their relationships to each other.

Genetic and epigenetic composition of repeat-based holocentromeres

We analyzed the sequence organization and chromatin structure of the *Rhynchospora* repeat-based holocentromeres. The contiguity of our assemblies coupled with the short array size of centromeric *Tyba* repeats allowed us to resolve mostly complete *Tyba* arrays in the three *Rhynchospora* genomes. While total number and amount of *Tyba* arrays increased with chromosome size (**Figure 3A–B**), the density of arrays decreased (**Figure 3C**). Average array sizes of 20.3, 20.5, and 19.8 kb, and average spacing between two consecutive arrays of 368, 492, and 424 kb were found in *R. breviuscula*, *R. pubera*, and *R. tenuis*, respectively (**Figure 3D–E**). These results confirm a similar overall organization of centromeric *Tyba* repeats among the three *Rhynchospora* species. In common with monocentric centromeric repeats (Kasinathan and Henikoff, 2018), we also found a high frequency of dyad symmetries in the *Tyba* consensus sequences of all three *Rhynchospora* species (**Figure 3F**).

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) confirmed the highest enrichment of CENTROMERIC HISTONE H3 (CENH3) for the *Tyba* repeats and lower enrichment for *CRRh* (centromeric retrotransposon of *Rhynchospora*) throughout the entire *R. pubera* and *R. breviuscula* genomes (**Figure 4A–C**; **Figure S1H**; **Table S2**). We detected 2,753 and 995 CENH3-binding regions (hereafter CENH3 domains) evenly distributed across the five chromosomes of *R. pubera* and *R. breviuscula*, respectively. In both species, length, density and spacing of CENH3 domains followed a similar pattern to the number of *Tyba* arrays detected (**Figure 3A–E**). Considering that one CENH3 domain is equivalent to one centromere unit, on average, each *R. pubera* chromosome carried 600 centromere units (1.88 domains/Mb), while the smaller chromosomes of *R. breviuscula* carried 200 centromere units each on average (2.69 domains/Mb) (**Figure 3A–C**). Thus, genome/chromosome size may be negatively correlated with centromere unit density in beak-sedges. Genome-wide there was a significant association between CENH3 domains and *Tyba* repeats for both species ($p < 0.05$), confirming that *Tyba* repeats are the main CENH3-binding sites. Therefore, repeat-based holocentromeres are likely to be conserved and associated with *Tyba* repeats in beak-sedges.

In the monocentric *J. effusus*, the histone mark H3K4me3 (euchromatin-specific) showed dispersed labeling along chromosome arms, while H3K9me2 (heterochromatin-specific) was concentrated at pericentromeric regions and co-localized with chromocenters in interphase (**Figure S1I**). By contrast, in the holocentric *R. pubera*, both eu- and heterochromatin-specific histone marks were intermingled all along the chromosomes with a constant density even towards subtelomeric and central chromosomal regions (**Figure 4A–C**). Locally, H3K4me3 was mostly highly enriched at the promoter regions of protein-coding genes, whereas H3K9me2 was enriched on small heterochromatic islands, typically resembling TEs (**Figure 4C**). H3K4me3 was depleted at CENH3 domains, while H3K9me2 showed residual enrichment. We noticed a slight increase in H3K9me2 enrichment flanking CENH3 domains relative to the core region, mimicking the pericentromeric chromatin composition in monocentromeres (**Figure 4C**).

Irrespective of centromere type, gene bodies were highly enriched for mCpG in both *R. pubera* and *J. effusus*, with a sharp decrease at promoters and terminal regions. Methylation in the CHH and CHG contexts was much lower for the gene bodies than for intergenic regions (**Figure 4D–E**), as previously reported for other plants (Feng et al., 2020). Remarkably, despite the differences in chromosome organization, both the *Tyba* repeats in *R. pubera* and tandem repeats in centromeric regions of *J. effusus* chromosomes were highly enriched for mCpG at similar levels to those for TEs (**Figure 4D–E**). mCHG was sharply enriched flanking CENH3-binding regions in *R. pubera*, resembling the H3K9me2 pattern (**Figure 4C**). We obtained a similar pattern for mCHG at centromeric repeats in *J. effusus* (**Figure 2D**; **Figure 4D–E**). TEs showed the highest enrichment for mCHG and mCHH, while *Tyba* repeats displayed lower levels of mCHH, similar to genes (**Figure 4D–E**). Our results argue for the presence of a pericentromere-like chromatin state around the ends of centromere units in *Rhynchospora* that may mark the borders for CENH3 loading.

A typical centromere unit in *R. pubera* comprised a single *Tyba* array surrounded by genes and TEs (**Figure 4F**). We detected CENH3 domains all along the chromosomes, even in *Tyba* arrays located near telomeres like those at both ends of *Rp*Chr2 (**Figure 4F–G**), confirming the telomere-to-telomere centromere activity of holocentric chromosomes. Notably, we observed an enrichment for H3K4me3 and actively transcribed genes close to centromere units, with an average distance of 6.3 kb (**Figure 4H–I**). We identified 313 genes that showed at least a 1-bp overlap with CENH3 domains. We even detected actively transcribed genes with typical H3K4me3 enrichment inside CENH3 domains (**Figure 4H**), a characteristic only rarely observed in monocentric organisms (Mizuno et al., 2011; Schotanus et al., 2021). Both CENH3 association and transcription were frequently reduced in genic regions inserted into centromere units, compared to genic regions residing outside the core centromere unit (**Figure 4H**), reflecting the precise regulation of chromatin composition of the *R. pubera* genome. *CRRh* was frequently inserted into *Tyba* arrays enriched for CENH3, but also H3K9me2 and some level of H3K4me3, suggesting a different epigenetic regulation of this retroelement compared to *Tyba* repeats (**Figure 4I**). Our results thus point to fine-scale epigenetic regulation of genomes with repeat-based holocentromeres.

Transposition partially explains genome-wide *Tyba* dispersal and expansion

Tyba repeats in *R. pubera* can be flanked by *TCR1* and *TCR2* repeats, suggesting that some *Tyba* arrays are part of larger repetitive elements (Marques et al., 2015). The consensus full-length *TCR1* element contained a *Tyba* array with a 5′ sequence of approximately 4.8-kb and a 136-bp 3′ sequence. The element possessed no open reading frame (ORF), lacked terminal repeats, and its 5′ and 3′ ends harbored the ATC and CTAGT sequence motifs, respectively, suggesting that *TCR1* is a non-autonomous *Helitron* transposable element (Thomas and Pritham, 2015), from the same family as a fully autonomous *Helitron* element (*Helitron-27*) in the *R. pubera* genome. Despite sharing conserved terminal sequences, *TCR1* and *Helitron-27* exhibited no similarity in their internal regions. We identified three intact copies of the autonomous *Helitron-27* in the genome (**Table S3**) with high mutual similarity, each encoding a full *Helitron* helicase (1,340 amino acids), indicating that *TCR1* and *Helitron-27* elements are still capable of transposition. We further identified an additional 322 full-length elements (**Table S3**) with both *TCR1* termini as well as *Tyba* and another 146 partial elements with the 3′ terminal sequence and containing *Tyba* within the upstream 500-bp region. We conclude that at least 468 *Tyba*-containing loci in the genome resulted from the transposition activity of *TCR1* elements. The full-length *TCR1* sequences were 6.9–49.6 kb (24.8 kb average), containing 1.2–31.3 kb from *Tyba* (15.7 kb average). In many *TCR1* elements, *Tyba* arrays were split into multiple segments due to insertions of other sequences, showing that multiple *Tyba* loci can originate from a single *TCR1* insertion (**Figure 4J–K**). Importantly, a comparison of *TCR1*- and CENH3 domains revealed that the vast majority (98.7%) of full-length *TCR1* elements are embedded within or overlap with the centromere units (**Table S2**).

*Helitron*s with boundaries similar to *TCR1*/*Helitron-27* were present in *R. tenuis* and *R. breviuscula*, but all but one of the full-length elements in these two genomes lacked *Tyba*. The sole exception was a single element from *R. breviuscula* (Chr1:69162288–69195619) with 5′ and 3′ boundary sequences characteristic of this *Helitron* family as well as a *Tyba* array; however, the remaining sections lacked any similarity to the *TCR1* of *R. pubera*. These results suggest that *Tyba* was amplified as a part of a *TCR1 Helitron* only in the genome of *R. pubera*.

The *TCR2* element was found to be a miniature inverted-repeat transposable element (MITE) and ranged from 672 to 1,235 bp, likely originated from the DNA transposon MuDR with shared similarity (up to 97%) in the terminal inverted repeats. All 158 full-length *TCR2* elements identified in the *R. pubera* genome were in *Tyba* arrays, but none were characterized by *Tyba* insertions. Thus, *TCR2* elements did not contribute to the dispersal of *Tyba* in the *R. pubera* genome.

R. pubera is a cryptic auto-octoploid with *n*=5 chromosomes

The *R. pubera* genome is four times larger than that of its closely related species, despite sharing the same ancestral chromosome number (ACN, *x*=5) (Burchardt et al., 2020; Ribeiro et al., 2018) (**Figure S2A**). One explanation for this pronounced genome expansion would be a sudden and massive proliferation of repeat elements. However, we observed no accumulation of repeats when comparing repeat abundance profiles among closely related *Rhynchospora* species (**Figure S2A**). Thus, a different process must be responsible for the large genome size in *R. pubera*.

Completeness assessment of the *R. pubera* genome by calculating the Benchmarking Universal Single-Copy Orthologs (BUSCO) score revealed a surprisingly high level of gene duplications (96.0% duplicated BUSCOs) (**Figure S2B**). Annotation of the genome yielded far more high-confidence gene models (91,363) in *R. pubera* compared to the other species (**Figure 1D–E; Table S1**), confirming the high level of gene duplication (**Figure S2C**). Self-synteny analysis revealed that the *R. pubera* genome comprises two large syntenic blocks in four copies across the five chromosomes (**Figure S3A**). The larger syntenic block, named Block1, corresponded to the entire *R. pubera* chromosome 4 (*Rp*Chr4) and *Rp*Chr5 and contributed to a large fraction of both *Rp*Chr1 and *Rp*Chr2. We identified the smaller block, named Block2, twice in an inverted arrangement in *Rp*Chr3, as well as in *Rp*Chr1 and *Rp*Chr2 (**Figure S3A**).

The distribution of synonymous substitutions per synonymous site (Ks) for coding sequences over the intragenomic syntenic blocks in *R. pubera* had a large peak indicative of recent and successive whole-genome duplication (WGD) events. An additional small peak was also observed, indicating an ancient WGD (**Figure S3B**). By filtering out the sequences showing the lowest Ks values, we determined that Block1 from *Rp*Chr1 shows higher sequence identity to *Rp*Chr4, which we renamed Block1A1 and Block1A2, respectively. Similarly, Block1 from *Rp*Chr2 showed higher sequence identity to *Rp*Chr5, which were thus named Block1B1 and Block1B2, respectively (**Figure S3B–C**). We confirmed the relationships of the four Block1 copies by comparative phylogenetic analysis (**Figure S3D**). A similar analysis of Block2 copies was inconclusive (**Figure S3C;E**). Using k-mer analysis, which provides information on genome size, ploidy, and genome structure through scrutiny of heterozygous k-mer pairs (Ranallo-Benavidez et al., 2020), we detected a higher incidence of homozygous and duplicated k-mers, favoring an autopolyploidy genome model for *R. pubera* (**Figure S3F–G**). Importantly, this analysis accurately determined the diploid heterozygous state of *R. breviuscula* and *R. tenuis* (**Figure S4**). Thus, *R. pubera* has an auto-octoploid genome shaped by two rounds of genome doubling explaining its large genome size. Post-polyploid genome shuffling events considerably reduced the chromosome number to *n*=5.

Chromosome fusions explain karyotype evolution in beak-sedges

To explore the genome duplications seen in *R. pubera*, we compared its genome to its close relative *R. breviuscula*, which has the same chromosome number but a genome one-quarter the size (415 Mb) (**Figure S2A**). Assessment of the *R. breviuscula* genome revealed a high level of completeness, with a BUSCO score of 98.3%, and little gene duplication (2.1%) (**Figure S2B–C**), confirmed by the absence of self-synteny (**Figure S1D**). Gene annotation yielded 24,354 high-confidence gene models (**Figure 1D–E**; **Table S1**), four times fewer than in the *R. pubera* genome, as expected. Synteny analysis between both genomes illustrated how each *R. breviuscula* chromosome is present in four copies in the *R. pubera* genome (**Figure 5A**). Remarkably, *Rp*Chr1 and *Rp*Chr2 contained all five putative *R. breviuscula* chromosomes (*Rb*) in end-to-end configurations. *Rp*Chr3 contained *Rb*3 and *Rb*4 copied twice in an inverted order, comprising Block2, while *Rp*Chr4 and *Rp*Chr5 contained *Rb*1, *Rb*2, and *Rb*5, comprising Block1A and Block1B, respectively (**Figure 5A**). Thus, *R. breviuscula* likely conserved the ancestral karyotype, while *R. pubera* restored the ACN (*x*=5) of its clade due to descending dysploidy, which was mediated by a complex chain of chromosome fusions, e.g., end-to-end fusions (EEFs), with 15 EEF junctions detected. Remarkably, each chromosome pair had a unique combination of ancestral chromosomes. We conclude that descending dysploidy involving a unique combination of chromosomes may be a strategy to avoid meiotic pairing issues that could potentially arise from autopolyploidy, thereby acting as a rapid route to diploidization facilitated by holocentricity.

Because we detected several EEFs in multiple copies in *R. pubera*, we assessed whether they were derived from the same rearrangement or if they arose from multiple independent events. All duplicated EEFs in the *R. pubera* genome, e.g., *Rb*2/*Rb*5, *Rb*3/*Rb*4, *Rb*1/*Rb*2, and *Rb*1/*Rb*5 EEFs, share a fusion signature involving the same regions. This observation suggested that the *Rb*2/*Rb*5 and *Rb*3/*Rb*4 EEFs present four times in the *R. pubera* genome emerged only once—before the first WGD event (**Figure 5C; Figure S5**). The *Rb*1/*Rb*2 and *Rb*1/*Rb*5 EEFs, which were found twice, likely emerged after the first WGD event. Finally, we found the *Rb*1/*Rb*4, *Rb*2/*Rb*4, and *Rb*3/*Rb*3 EEFs only once, suggesting that they occurred after the second WGD (**Figure 5C**; **Figure S5**).

The *Rb*3/*Rb*4 EEF, which forms Block2 in the *R. pubera* genome, was likely maintained as a duplicated fused chromosome after the first WGD, which might have allowed a longer period of tetrasomic inheritance. This hypothesis might explain the fact that the sequences of the four copies from Block2 cannot be distinguished from each other, in contrast to Block1.

We attempted to date the duplication events using a set of conserved genes shared among the four copies of Block1, which revealed the first WGD event as occurring around 3.8 Mya followed by a second WGD event around 2.1 Mya (**Figure S3D**). Based on this analysis, we deduced the origin and evolution of the *R. pubera* karyotype (**Figure 5C**). These results further support an autopolyploid origin for *R. pubera* and confirm a short interval between the two rounds of WGDs, indicating rapid chromosome number reduction in this species.

We carried out a number of analyses to determine the origin of the reduced karyotype in *R. tenuis* (*n*=2). BUSCO analysis of its genome revealed high completeness (98.5% against the viridiplantae_odb10 dataset) and little duplication (3.7%) (**Figure S2B–C**). Gene annotation yielded 23,215 high-confidence gene models (**Figure 1D–E**). The absence of self-synteny in the *R. tenuis* genome ruled out large duplications (**Figure S2E**). Synteny comparison between *R. tenuis* and *R. breviuscula* genomes showed that again all *R. breviuscula* chromosomes were present in simple end-to-end configurations in the *R. tenuis* genome, explaining its karyotype by descending dysploidy from n=5 to n=2 (**Figure 5A**). Strikingly, we observed similar associations of syntenic *Rb* blocks as found in Block1 and 2 in both *R. pubera* and *R. tenuis*, where *Rt*Chr1 resembled Block1B and was composed of *Rb*2, *Rb*5, and *Rb*1, while *Rt*Chr2 resembled Block2 consisting of *Rb*3 and *Rb*4 (**Figure 5A**). However, the orientation of chromosome ends involved in the EEFs differed in the two instances, suggesting that the EEFs occurred independently (**Figure 5C**).

Despite their high chromosome number and centromere-type differences, *J. effusus* and the previously available genome for the sedge *Carex littledalei* (synonym *Kobresia littledalei*) (Can et al., 2020) showed a typical diploid gene content and no evidence of any recent WGD, outside of a shared ancient WGD between sedges and rushes (**Figure S4**). The *J. effusus* genome also revealed high completeness (100% viridiplantae_odb10 dataset) and little duplication (1.6%) (**Figure S2B–C**). Annotation of its genome yielded 18,942 high-confidence gene models (**Figure 1D–E; Table S1**). Synteny analysis further revealed that most *J. effusus* and *C. littledalei* chromosomes are present as highly collinear blocks across the five chromosomes of *R. breviuscula*, suggesting a high conservation of synteny although the group is ancient (78 Mya) (**Figure 5B–C**). Thus, neither the low nor the high chromosome numbers observed in many holocentric species necessarily reflect the absence or presence of recent polyploidy, respectively, and these numbers should be interpreted with caution in the absence of detailed genomic studies.

Tyba repeats are frequently present at the junctions of end-to-end fusions

Transposable elements can influence chromosomal rearrangements (Lonnig and Saedler, 2002). To assess their possible role in the EEFs observed in *Rhynchospora* genomes, we looked for enrichment of specific repeats at the end of *R. breviuscula* chromosomes and near the junctions of

EEFs in *R. pubera* and *R. tenuis*. We detected a high density of TEs in almost all subtelomeric regions of *R. breviuscula* chromosomes. These repeat-rich regions varied from 500 kb to 3 Mb in size, were mainly enriched for an LTR *Ty3/Gypsy* element of the *Athila* clade, were poorly enriched for genes, and lacked *Tyba* repeats (**Figure 5A–B; Figure S1D**). Notably, the *R. breviuscula* subtelomeric repeat-rich regions were largely missing at the junctions of fused chromosomes in both *R. pubera* and *R. tenuis* (**Figure 5D; Figure S6**). Remarkably, we detected *Tyba* repeats exactly at the EEF junctions in 10 out of the 15 EEFs of *R. pubera*, while we observed a small 45S rDNA remnant array (with only five 18S-5.8S-26S units) in one EEF (**Figure 5C–D; Figure S6A–G**). In *R. tenuis*, we also identified a *Tyba* repeat array in one out of three EEF junctions (*Rb*3/*Rb*4 junction on *Rt*Chr2), while an interstitial telomeric site (536 bp) was detected at the *Rb*5/*Rb*1 junction (**Figure 5C; Figure S6H**).

Emergence and loss of CENH3 domains related to *Tyba*

We used the duplicated genome copies of *R. pubera* to study cases of paralogous CENH3 domains and *Tyba* arrays. Of 660 groups of paralogous regions, 66% of the CENH3 domains were present in all four copies (**Table S5**). We also identified 50 groups of paralogous regions in which the CENH3 domain was lost in one of the paralogs. Most cases (88%) were associated with *Tyba* loss (**Figure 6A; Figure S7A**) or a reduced size of the *Tyba* region in loci devoid of CENH3 signal compared to their paralogous regions bound by CENH3. We observed the likely inverse event, the gain of a new CENH3 domain, in groups of paralogous regions in which we only identified the CENH3 domain in only one of the four paralogous regions. In the newly acquired CENH3 domain, there was either a new *Tyba* insertion, likely due to a new insertion of *TCR1* (**Figure 6B; Figure S7B**), a *Tyba* expansion, or the insertion of a new TE (most frequent). However, the ChIP/Input ratios within these potentially new CENH3 domains containing a new TE insertion (1.5) were significantly lower

($p < 2.2.e–16$) than the ChIP/Input ratios in potentially new CENH3 domains associated with a new insertion of a *Tyba* element (4.1).

DISCUSSION

Here, we report the first high-quality and contiguous chromosome-scale reference genomes for three species with repeat-based holocentromeres, *R. pubera*, *R. breviuscula*, and *R. tenuis*, and their closest monocentric relative, *J. effusus*. These newly assembled genomes provide a valuable resource for comparative biology and studies related to genome adaptation to different centromere types.

Repeat-based holocentromeres influence genome organization and regulation

Repeat-based holocentromeres in beak-sedges comprise small islands (20–25 kb) of centromeric *Tyba* repeats in which high mCpG, low H3K9me2, and depletion of H3K4me3 distinguish them from other holocentric genomes with and without repeat-based holocentromeres (Cortes-Silva et al., 2020; Despot-Slade et al., 2021; Gassmann et al., 2012; Nhim et al., 2021; Senaratne et al., 2021; Steiner and Henikoff, 2014). The association levels of H3K9me2 and mCHG at the core (low) and flanking (high) centromere units in *R. pubera* are strikingly similar to the recently reported *A. thaliana* centromeres (Naish et al., 2021). We also observed a similar pattern of mCHG methylation in monocentric *J. effusus*. Heterochromatinization of pericentromeres appears to be important for stabilizing the centromeric core, by preventing recombination between core repeats and stopping the spread of CENH3 into adjacent regions (Achrem et al., 2020; Wong et al., 2020). Thus, despite substantial genome restructuring, the epigenetic regulation of centromere units in beak-sedges resembles that in monocentric centromeres. This observation suggests an evolutionarily conserved epigenetic regulation of repeat-based centromeres in both mono- and holocentric organisms. Although rare, we observed active genes close to and even within centromere units, which is likely only possible with a plastic regulation of eu-/heterochromatic boundaries. We hypothesize that *R. pubera* achieves such a feat with a fine-scale epigenetic regulation of centromere units (**Figure 7A–B**).

Centromere units are regularly spaced (350–500 kb) in the *Rhynchospora* genomes, instead of randomly distributed. This specific spacing might point to a selection mechanism for establishing centromere units separated by an optimal spacing required to fold the chromatin during cell-cycle progression and for the recruitment of CENH3-positive nucleosomes to build the line-like holocentromere at metaphase (**Figure 7B**). *In silico* modeling based on polymer simulations of chromatin folding in holocentric chromosomes suggests that centromere units can act as anchors of loop extruders, facilitating the formation of line-like holocentromeres during chromosome condensation (Camara et al., 2021).

A mechanism for the formation of repeat-based holocentromeres

The repeat-based holocentromeres of the *Rhynchospora* species analyzed here are almost exclusively composed of *Tyba* repeats. We cannot conclude from the available data whether the accumulation of such repeats triggered the transition to holocentricity or whether CENH3 spreading preceded and/or also facilitated the subsequent expansion of holocentromeric repeats. However, we did demonstrate that a portion of *Tyba* arrays in *R. pubera* emerged in the genome as a result of amplification of *TCR1*-type *Helitron*s and that most (98.7%) full-length *TCR1* elements possessing *Tyba* are associated with CENH3-bound chromatin. This either indicates that centromere units are at least partially determined genetically by the nucleotide sequence of *Tyba* or that *TCR1* transposition involves the transfer of epigenetic centromere marks, e.g., CENH3, which remain associated with the new copy of the element. The presence of tandem repeats within *Helitron*s is common in both plants and animals (Thomas and Pritham, 2015), but *TCR1* is the first *Helitron* described to possess a centromeric satellite. The lack of *Tyba* repeats in *TCR1*-related elements in *R. breviuscula* and *R.*

tenuis suggests that *Tyba* was captured by *TCR1* after the ancestors of *R. pubera* and the two other *Rhynchospora* species diverged.

It is conceivable, however, that the amplification and dispersal of *Tyba* occurred via mobilization by transposable elements earlier in the evolution of *Rhynchospora* and that the signatures of such events have long since been lost due to the accumulation of mutations, insertions/deletions, and DNA rearrangements. We also observed such changes in many *TCR1* loci identified in the genome of *R. pubera* that contained either truncated *TCR1* elements or full-length elements with nested insertions of other sequences (**Figure 4K–L**). The existence of a single *TCR1/Helitron27*-related element in *R. breviuscula* that possesses *Tyba* but lacks overall similarity to *TCR1* suggests that *Tyba* capture by *Helitron*s occurs recurrently in the evolution of *Rhynchospora* species and may result in waves of *Tyba* amplification via *Helitron* transposition.

The effect of holocentricity on karyotype evolution and diploidization

Our results are consistent with dysploidy as the main driver of karyotype evolution in holocentric organisms (Guerra, 2016; Mayrose and Lysak, 2021), where strong descending dysploidy restored the ACN (x=5) in *R. pubera* and reduced the chromosome number in *R. tenuis*. In both cases, the same ancestral chromosomes were fused independently either without (*R. tenuis*) or following WGDs (*R. pubera*). Such tolerance of extensive chromosomal rearrangements seems to underlie rapid karyotype evolution, eventually leading to chromosomal speciation (Lucek et al., 2022; Lukhtanov et al., 2018).

Robertsonian translocations and chromosome fusions leading to descending dysploidy have been reported in some holocentric butterflies (Cicconardi et al., 2021; Hill et al., 2019). However, the incidence of EEFs as the sole mechanism of descending dysploidy in *Rhynchospora* is intriguing. Remarkably, meiotic pairing and segregation is not disturbed in the *R. pubera* genome (Marques et al., 2016), suggesting selection has produced a balanced set of fewer chromosomes. Since *R.*

pubera underwent two rounds of WGD, descending dysploidy by EEFs would be a way to effectively create chromosomes with different combinations of ancestral syntenic blocks, reducing the risk of meiotic multivalent pairing without need of rapid genome downsizing. EEFs in genomes with monocentric chromosomes is normally associated with the formation of typically unstable dicentric chromosomes but may represent a tolerable mechanism for chromosomal rearrangements when coupled with concurrent centromere elimination as part of structural diploidization after WGDs (Mandakova et al., 2010; Mandakova and Lysak, 2018; Murat et al., 2010). We argue that the prevalence of EEFs observed in *R. pubera* was facilitated by holocentricity, avoiding the deleterious effect of two centromeres after EEFs in monocentric species and likely promoting rapid structural diploidization.

In *Rhynchospora*, homologous chromatids are linked by terminal chromatin threads during inverted meiosis (Cabral et al. 2014). EEFs may occur with high(er) frequency in scenarios where chromatids of non-homologous chromosomes are erroneously connected via (repeat-based) chromatin threads. However, this notion does not exclude the possibility of EEFs occurring during interphase or mitosis. It is tempting to speculate that the repeat-rich regions observed at chromosome ends in *R. breviuscula* are involved in the formation of chromatin threads, which may act as substrates for ectopic recombination. *Tyba* repeats near these repeat-rich regions may be preferentially used as the site for recombination and may thus facilitate the occurrence of EEFs (**Figure 7C**). Alternatively, the recruitment of *Tyba* repeats as DNA templates to seal double-stranded breaks involved in EEFs may explain their pronounced association with EEFs (Vu et al., 2017).

Limitations of the study

The three *Rhynchospora* species analyzed in this study are characterized by repeat-based holocentromeres associated with *Tyba* repeats. However, some *Rhynchospora* species lack *Tyba* repeats (Ribeiro et al., 2017); thus, it is not clear whether repeat-based holocentromeres evolved in all species of the genus. Extending our approach to other holocentric species lacking *Tyba*-like repeats will certainly reveal new insights into the evolution of repeat-based holocentromeres. In addition, the presence of holocentric chromosomes in multiple genera of sedges as well as in closely related rushes (e.g., *Luzula* species), but not in *Juncus*, suggests transition to holocentricity occurred a long time ago (>60 Mya), which makes temporal tracking challenging. Indeed, our analyses of orthogroups did not identify a clear pattern related to different centromere types

Acknowledgements

We thank Martin Mascher, Raphael Mercier, Korbinian Schneeberger, Martin Lysak, and Ingo Schubert for the comments on the manuscript. We thank Saurabh Pophaly for assistance with NCBI sequence submission. We thank Neysan Donnelly for extensive manuscript editing. We thank Kristin Krause and Hequan Sun for support with ChIP experiments and k-mer frequency analysis, respectively. We thank the excellent technical assistance of Zdeňka Dubská, Jitka Weiserová and Eva Jahnová in preparing HMW DNA for optical mapping. This work was supported by the Max Planck Society, a PhD fellowship awarded to G.T. from the DAAD/India, and grants awarded to A.M. (DFG MA 9363/2-1), A.H., A.P-H., A.M (CAPES/DAAD PROBRAL 88881.144086/2017-01) and J.M. (GACR 20-24252S). We also thank the ELIXIR CZ Research Infrastructure Project (LM2018131) for providing computing and data storage facilities. This manuscript is dedicated to Prof. Marcelo Guerra for his long-standing work on beak-sedges cytogenetics, which paved the way for this work.

Author contributions

Conceptualization, A.M.; funding and resources, A.P-H, J.M., K.F.X.M., A.H., and A.M.; data production, P.G.H., G.T., B.H., H.S., and A.M.; formal analyses, investigation, and visualization, P.G.H., G.T., T.L., P. Neumann, T.V., P. Novak, M.Z., L.C., M.C., A.S., H.T., J.F., Y.M-S., Y.D., A.L.L.V., C.C.S.A., G.S. and A.M.; writing – original draft, P.G.H., G.T., and A.M.; writing - review & editing: all co-authors.

Declaration of interests

The authors declare no competing interests.

Inclusion and Diversity

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community.

Figure1

Click here to access/download;Figure;Figure_1.pdf

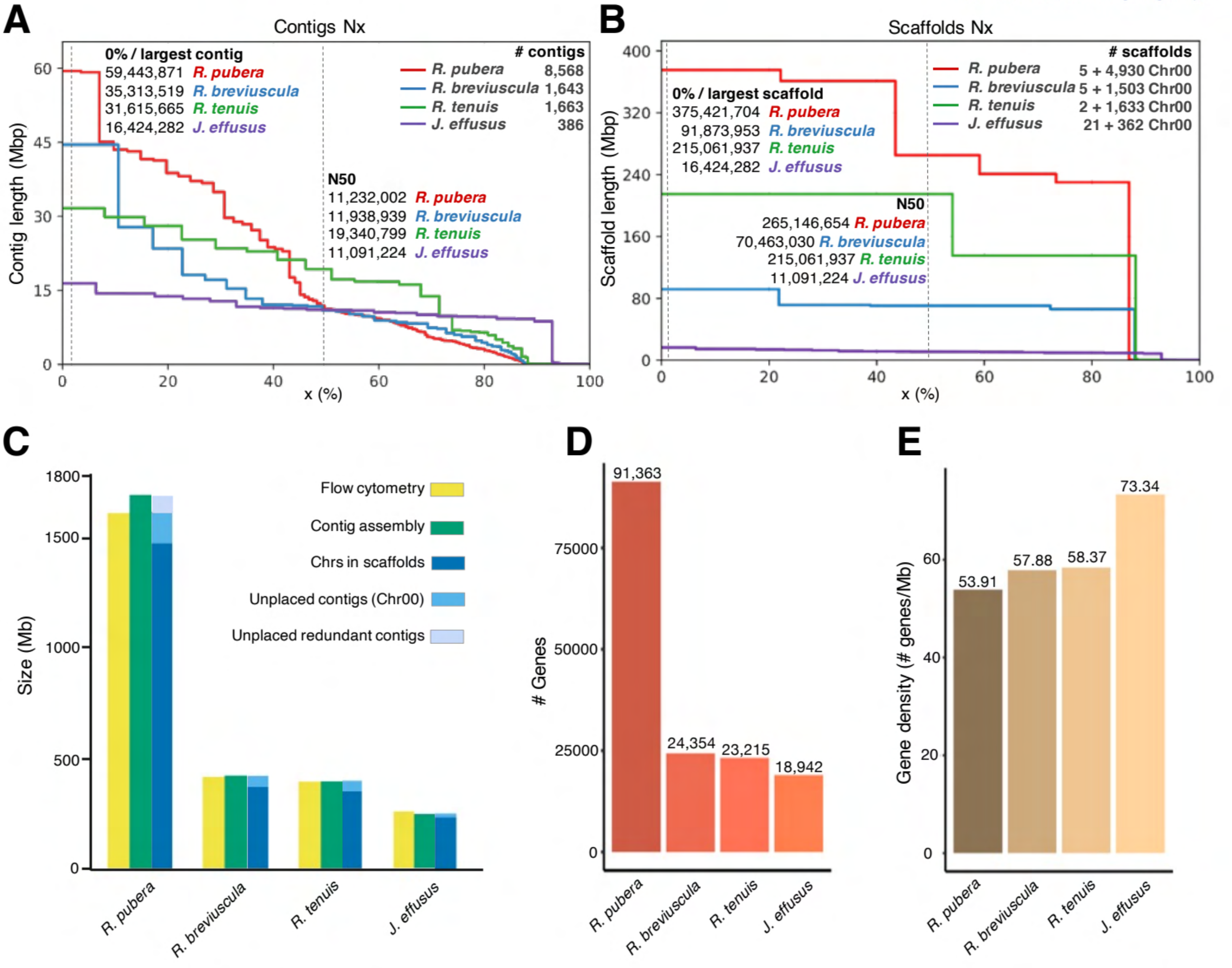


Figure 1. Summary of genome sizes, assemblies, scaffolding, and annotations.

(**A**) Assembly and (**B**) final scaffolding stats; (**C**) comparison between estimated genome size and assembly and scaffolding sizes; (**D**) total number of high-confidence annotated genes; (**E**) gene density per Mb. See also **Table S1**.

Click here to access/download;Figure;Figure_2.pdf

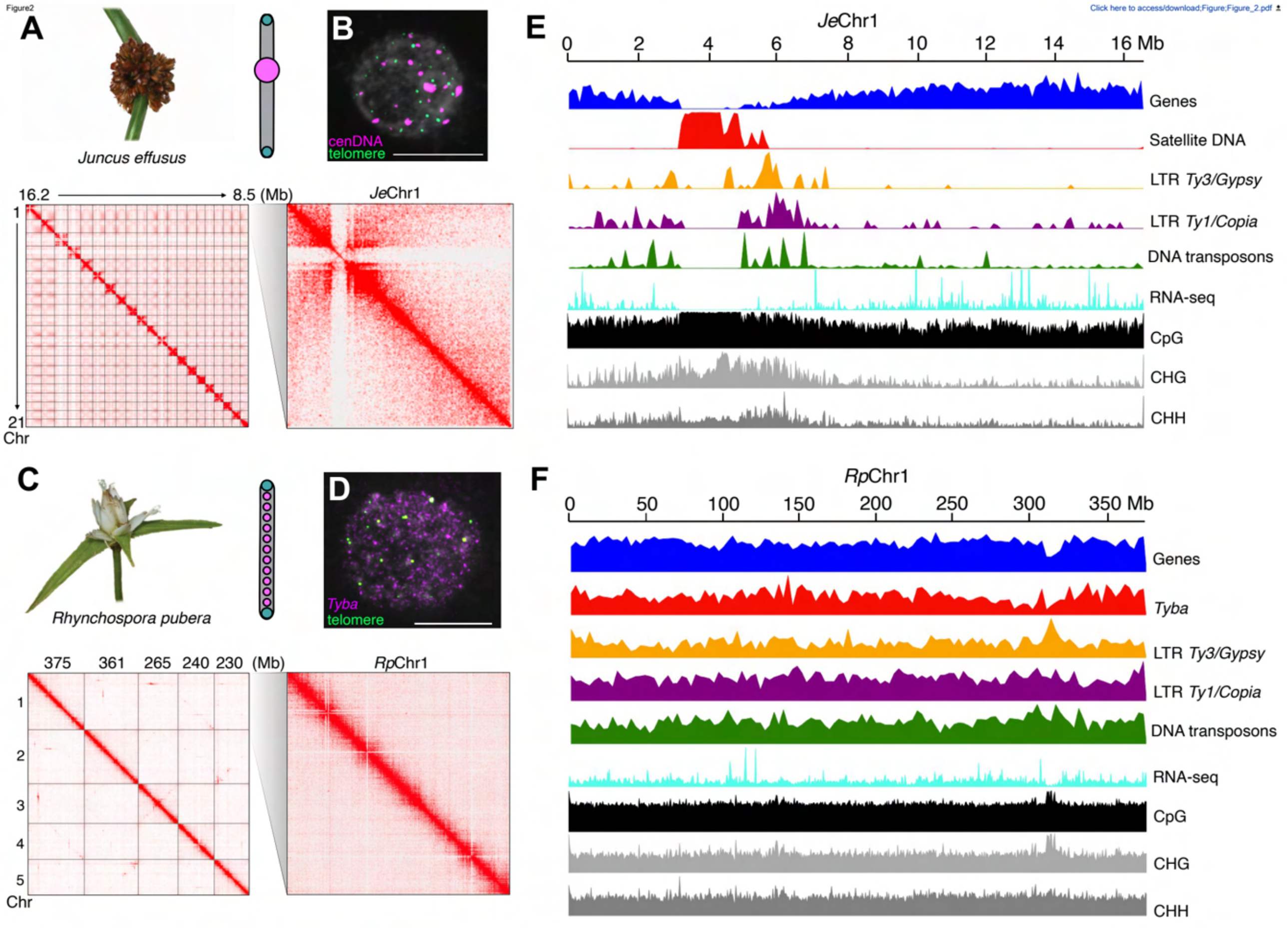


Figure 2. Spatial genome organization: monocentric *versus* holocentric chromosomes.

(**A**) *J. effusus* (top left) genome contact map (bottom left) and chromosome 1 (*Je*Chr1) detailed view (bottom right). Centromere organization in monocentric chromosomes (top right). (**B**) Interphase nucleus hybridized with DNA probes for the centromeric DNA (cenDNA) and telomeric sequence in *J. effusus*. (**C**) *R. pubera* (top left) genome contact map (bottom left) and *Rp*Chr1 detailed view (bottom right). Centromere organization in holocentric chromosomes (top right). (**D**) Interphase nucleus hybridized with DNA probes for the centromeric repeat *Tyba* and telomeric sequence in *Rhynchospora*. (**E**) *Je*Chr1 and (**F**) *Rp*Chr1 detailed view showing the clustered (*Je*Chr1) and uniform (*Rp*Chr1) distribution of main genomic features, typical for monocentric chromosomes and holocentric chromosomes in sedges, respectively. Window sizes for sequence type distribution density 100 kb (*J. effusus*) and 3 Mb (*R. pubera*). Centromeres and telomeres in chromosome models are represented by magenta and green circles, respectively. Scale bar = 10 µm. See also **Figure S1**.

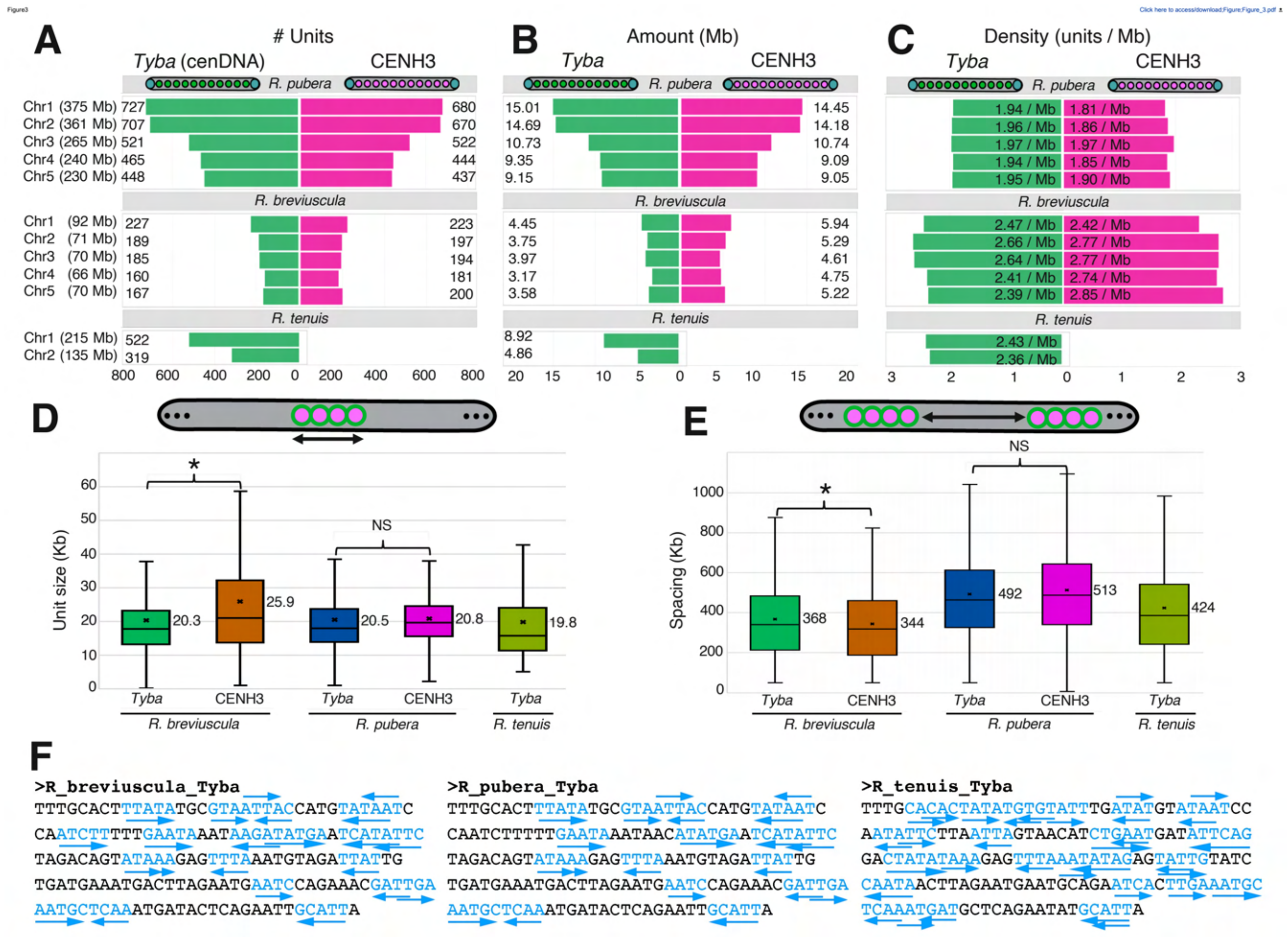


Figure 3. Features of *Tyba* centromeric DNA and CENH3 domains among *Rhynchospora* species.

(**A**) Total number per chromosome of annotated *Tyba* arrays and CENH3 domains, (**B**) total amount of bases associated with *Tyba* arrays and CENH3 domains, (**C**) density of *Tyba* arrays and CENH3 domains per chromosome. (**D**) Size distribution of *Tyba* arrays and CENH3 domains, and (**E**) spacing between two consecutive arrays/domains among *Rhynchospora* species. Asterisks indicate Dunn's test $P < 0.05$. (**F**) Patterns of DNA dyad symmetry in the *Tyba* consensus sequences of the three *Rhynchospora* species.

Click here to access/download;Figure;Figure_4.pdf

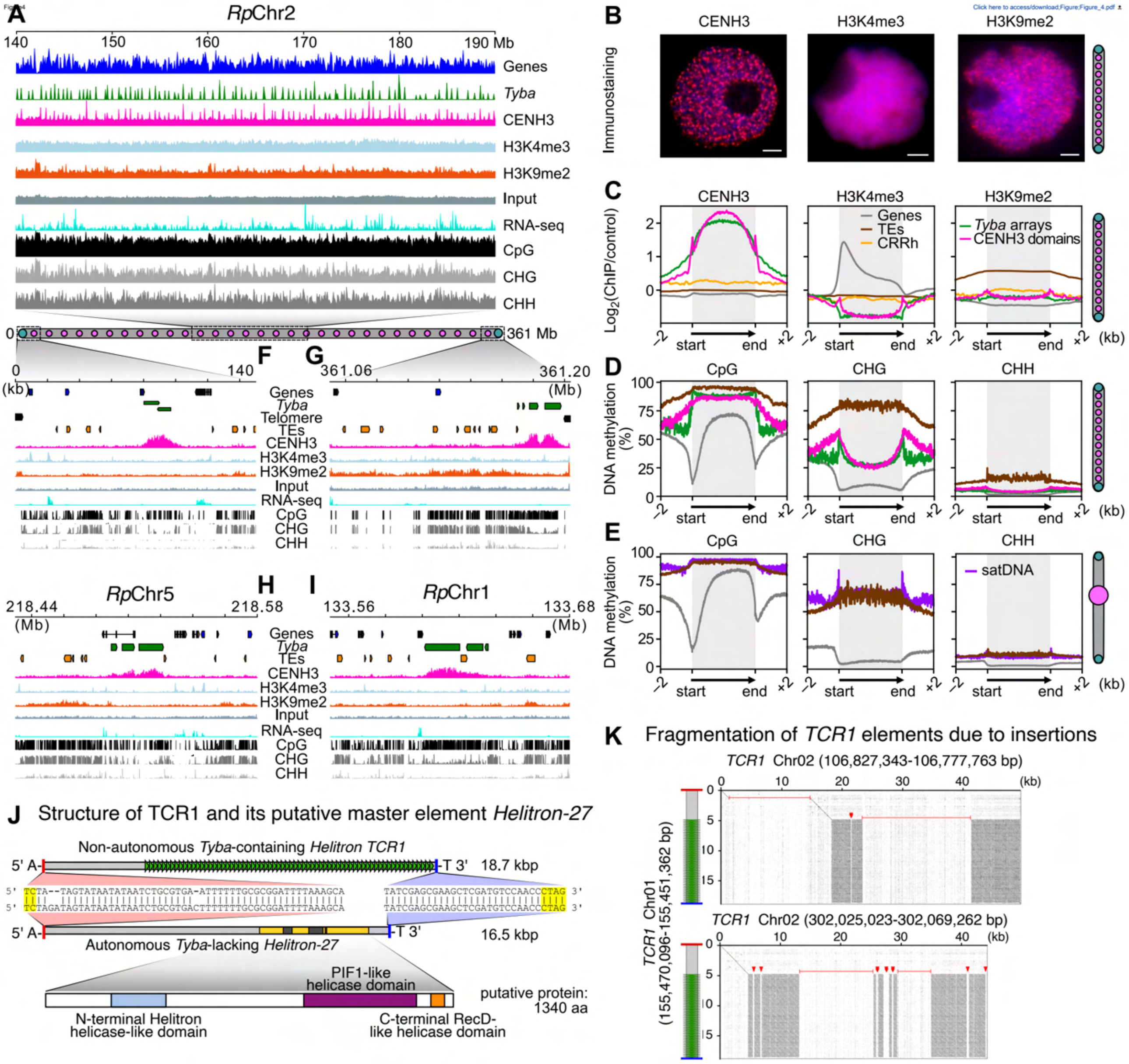


Figure 4. Genetic and epigenetic composition of repeat-based holocentromeres in *R. pubera*.

(**A**) Zoomed-in view of *Rp*Chr2 showing a 50-Mb region with multiple CENH3 domains that are closely correlated with *Tyba* repeat distribution. Gene and *Tyba* densities were calculated over 100-kb windows. (**B**) Immunostaining of *R. pubera* interphase nuclei for CENH3, H3K4me3, and H3K9me2. Scale bar = 2 µm. (**C**) Enrichment of CENH3, H3K4me3, and H3K9me2 from the *start* and *end* of different types of sequences: genes (gray line), TEs (brown), *CRRh* (yellow), *Tyba* repeats (green), and CENH3 domains (magenta). ChIP-seq signals are shown as Log2(normalized

RPKM ChIP/input). (**D–E**) Enrichment of DNA methylation in the CpG, CHG, and CHH contexts for the same sequence types as shown in **C** for *R. pubera* (**D**) and *J. effusus* (**E**), genes (gray line), satDNA (purple) and TEs (brown). Gray boxes in **C–E** highlight the modification enrichment over the body of each sequence type. (**F–G**) Close-up view of the first (**F**) and last (**G**) centromere units of *Rp*Chr2, which are composed of a *Tyba* repeat array very close to the telomere and showing the typical CENH3 enrichment. (**H**) A centromere unit where an active gene is intermingled with the *Tyba* repeat. (**I**) A *Tyba* array showing an insertion of the centromeric retrotransposon *CRRh* and CENH3-binding activity. (**J**) Structures of the typical non-autonomous *TCR1* element (Chr01:155470096–155451362) and its likely master element *Helitron-27* (Chr05:40901972–40918485). Similarity between *TCR1* and *Helitron-27* is mostly restricted to the terminal sequences. The 5′ and 3′ terminal sequences are in red and blue, respectively. Yellow: conserved *Helitron* sequence motifs in the alignment of *TCR1* and *Helitron-27* terminal sequences. Light gray: non-coding regions. Green triangles: *Tyba* array in *TCR1*. Yellow and dark gray: putative exons and introns in the *Helitron-27* coding region, respectively. (**K**) Dot-plot comparison of a typical *TCR1* element (vertical sequences) with two other elements (horizontal sequences) that have insertions of *TCR1*-unrelated sequences marked as red lines and triangles. See also **Figure S1**, **Table S2** and **S3**.

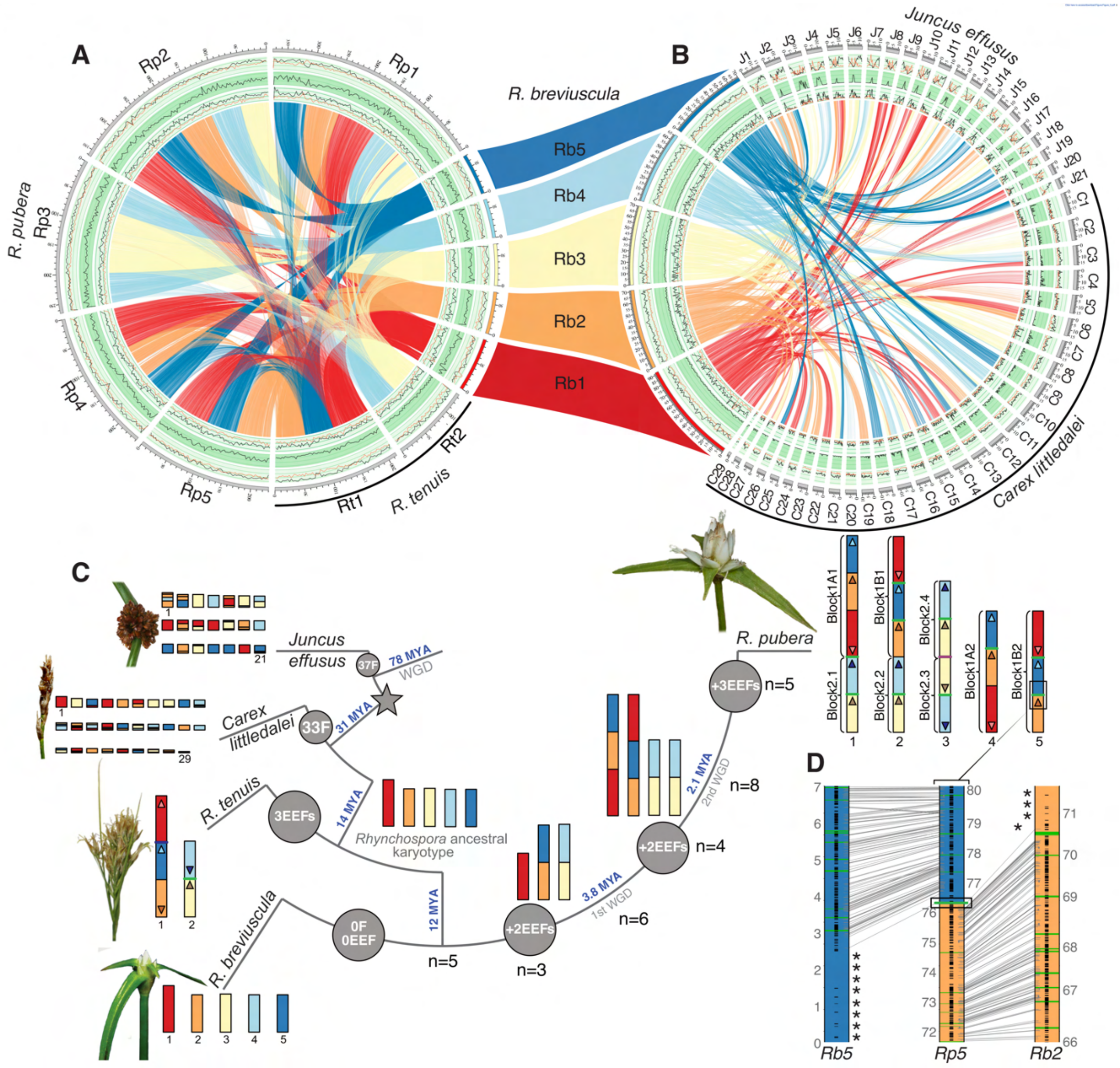


Figure 5. Genome organization and evolution of sedges and common rush.

(**A**) Circos plots of *R. breviuscula* synteny to *R. pubera* and *R. tenuis*. (**B**) Circos plots of *R. breviuscula* synteny to *J. effusus* and *C. littledalei*. For A and B, tracks from outside to inside: 1. Genes (black line) and TEs (red line), 2. *Tyba*/tandem repeats (black line), and 3. LTR *Ty1/Copia* (black line) and *Ty3/Gypsy* (red line) retroelement distribution. Distribution of the main sequence classes was calculated in 3-Mb windows for *R. pubera*, *R. tenuis*, and *R. breviuscula* (**A**), in a 1-Mb window for *R. breviuscula*, and in 500-kb windows for *C. littledalei* and *J. effusus* (**B**). (**C**)

Karyotype evolution and synteny conservation in sedges and common rush. Transition to holocentricity is indicated by a star. Hypothetical ancestral karyotype for *Rhynchospora* based on the simplest karyotype of *R. breviuscula* illustrates frequent end-to-end fusions (EEFs) in beak-sedges. For reconstruction of karyotype evolution in *R. pubera* see also **Figures S4 & S5**. Arrowheads: orientation of the *R. breviuscula* chromosomes in the *R. pubera* and *R. tenuis* ideograms. For both *J. effusus* and *C. littledalei*, ideograms indicate the syntenic blocks to *R. breviuscula* chromosomes. Numbers of putative EEFs or fission (F) events necessary to transform the hypothetical *Rhynchospora* ancestral karyotype into the extant genomes are within the gray circles. Repeat sequences at the junctions between *Rb* blocks are indicated by colored bars (*Tyba* = green; rDNA = purple; telomeric DNA = blue) in *R. tenuis* and *R. pubera* ideograms. (**D**) *R. pubera* Chr5 showing a *Tyba* array (black rectangle) at the junction between syntenic *Rb*2 and *Rb*5 blocks. Synteny from *Rp*Chr5 to *Rb*2 and *Rb*5 stops close to the last *Tyba* array, which is followed by a gene-poor, TE-enriched region, mainly LTR *Ty3/Gypsy* of the *Athila* clade (indicated by asterisks) that are frequently within *R. breviuscula* subtelomeric regions but absent in the fused chromosomes. Genes and *Tyba* arrays are annotated asblack stripes and green lines, respectively. See also **Figure S2**, **S3**, **S4** and **S5**, and **Table S4**.

Figure6 Click here to access/download;Figure;Figure_6.pdf

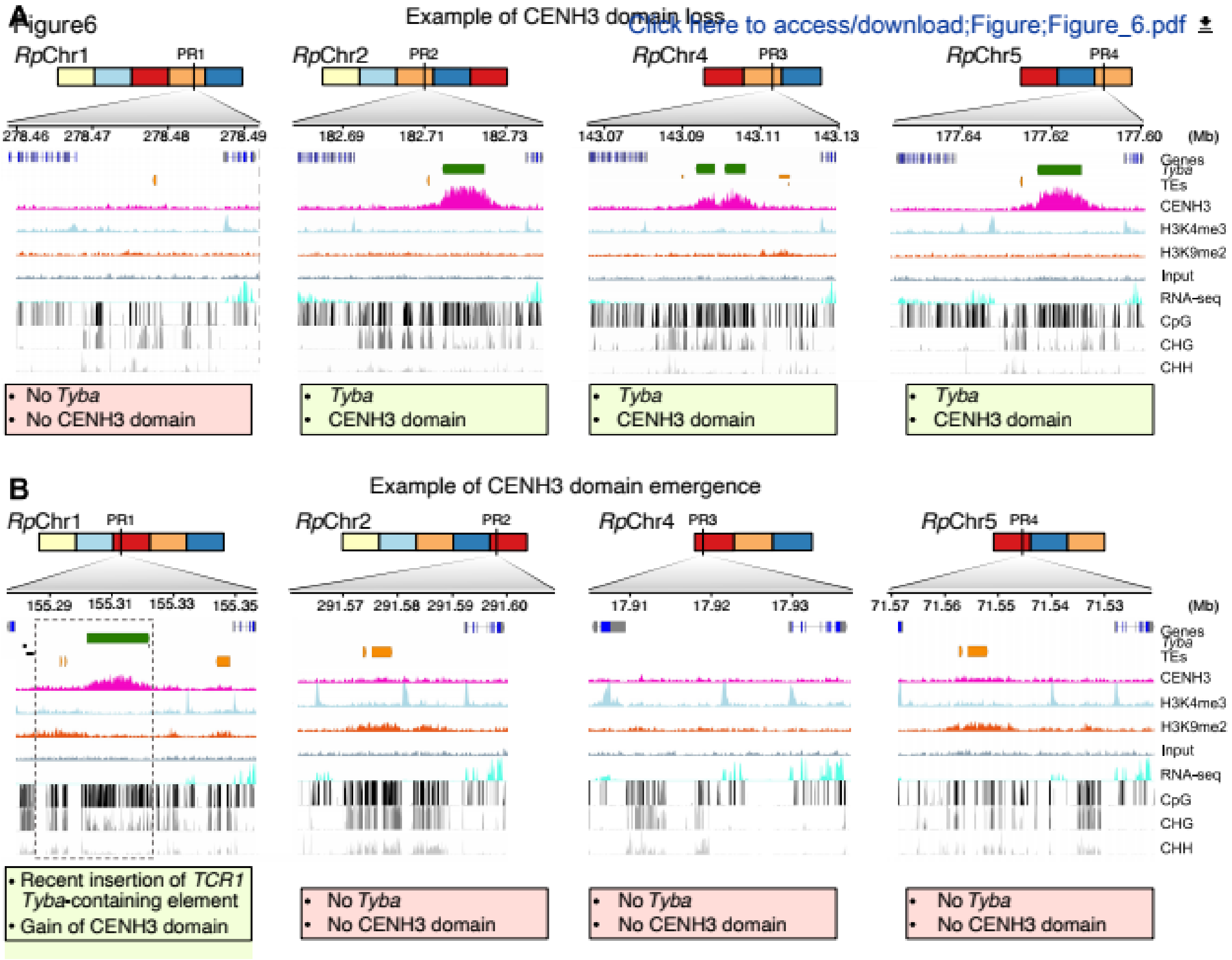


Figure 6. Emergence and loss of CENH3 domains in *R. pubera*.

(**A**) CENH3 domain with *Tyba* array loss in one of the four paralogous regions, while the other three copies retain the *Tyba* array. Zoomed-in view of all four regions demonstrates the CENH3 domain loss only in the *Rp*Chr1 copy. (**B**) CENH3 domain with *Tyba* array gain in one of four paralogous regions due to a transposition of a *Tyba*-containing *TCR1* in *Rp*Chr1, while the other three copies lack the *Tyba* array. The gained locus is indicated by the dashed box. Zoomed-in view of all four regions demonstrates the acquisition of a new CENH3 domain only in the *Rp*Chr1 copy. PR = paralogous region. Note that the four copies shared a similar chromatin composition. See also **Figure S5, Figure S7** and **TableS5**.

Figure7

Click here to access/download;Figure;Figure_7.pdf

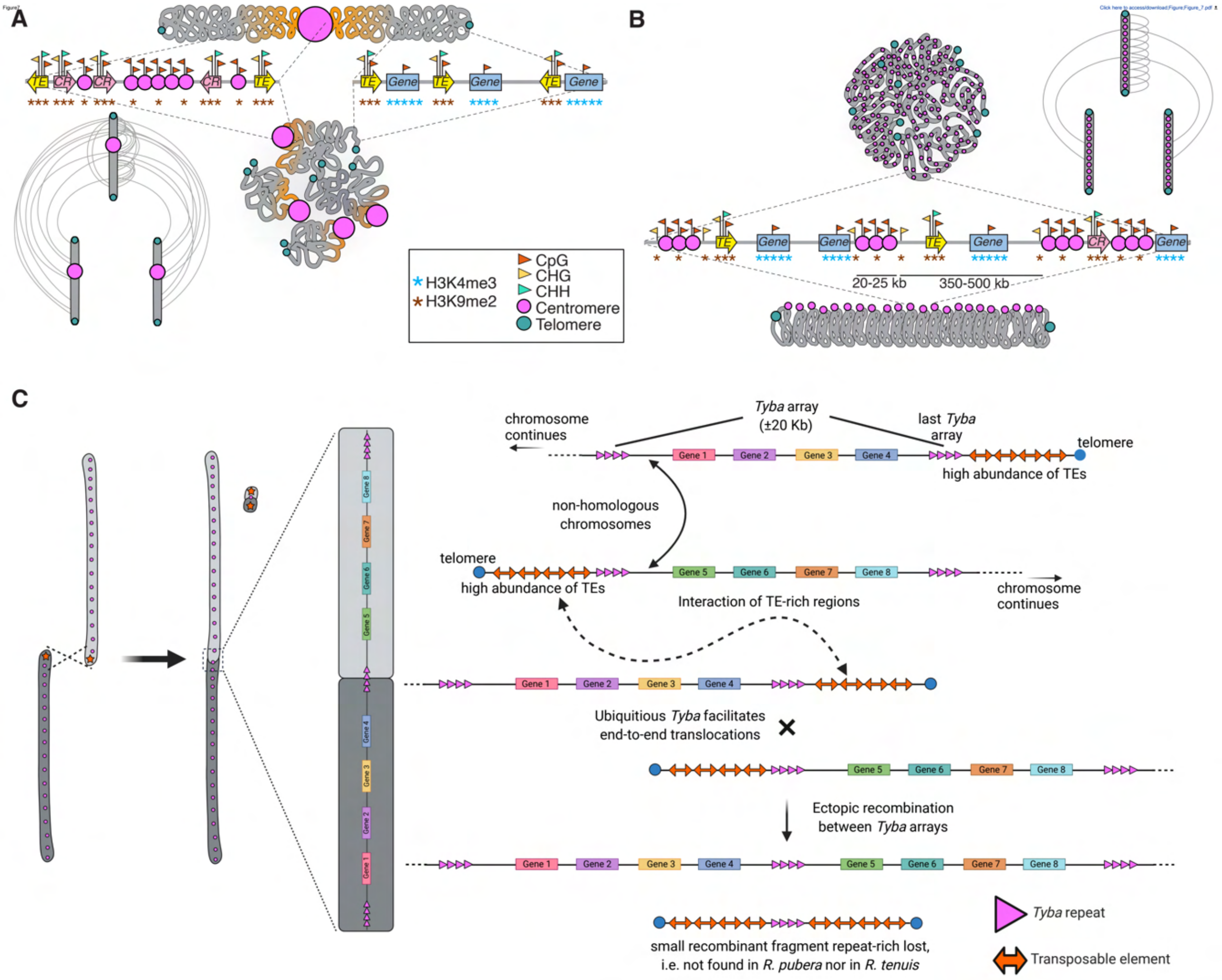


Figure 7. Genome organization in monocentric versus holocentric chromosomes and proposed model for end-to-end fusions.

(**A**) Typically, in a monocentric chromosome, compartments for deposition of more compacted and silenced chromatin states extend along large megabase-long regions around centromeres and pericentromeres, while genes concentrate at subtelomeric regions. A telomere-to-centromere axis is frequently observed in monocentric species due to the clustering of centromeres and telomeres, which increases the rate of interchromosomal chromatin contacts. (**B**) *Rhynchospora* holocentric genome revealed uniform deposition of epigenetic marks at the macro scale and fine epigenetic regulation of repeat-based centromere units and silenced and active chromatin states at the micro scale. The regular spacing between centromere units (350–500 kb) appears to be the distance necessary to loop the chromatin back, aligning centromere units (20–25 kb) at the outer surface of the condensed chromosome. A telomere-to-centromere axis is absent in holocentric species due to the lack of centromere clustering, affecting the spatial genome organization and decreases the rate of interchromosomal chromatin contacts. The model represents intra-/interchromosomal contacts among three different monocentric (**A**, bottom left) and holocentric (**B**, upper right) chromosomes. (**C**) Possible mechanism for the involvement of centromeric *Tyba* repeats in end-to-end fusions (EEFs). Interaction of highly repetitive regions close to the telomere could facilitate ectopic recombinations of *Tyba* repeats.

STAR Methods

RESOURCES AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, André Marques (amarques@mpipz.mpg.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All sequencing data used in this study have been deposited at NCBI under the Bioproject no. PRJNA784789 and are publicly available as of the date of publication. The reference genomes, annotations and all tracks presented in this work are made available at https://data.cyverse.org/dav-anon/iplant/home/dabitz66/marquesLabTrackHub/, the CoGe platform (https://genomevolution.org/coge) and the following UCSC Genome Browser hosted by CyVerse. All other data needed to evaluate the conclusions in the paper are provided in the paper and/or the Supplementary Materials.

This paper does not report original code.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Plant material

Plants from naturally occurring populations of *R. pubera* and *R. tenuis* growing in Curado (Recife), Northeastern Brazil, and *R. breviuscula* growing in Londrina (Paraná state), Southern Brazil were collected in 2013 and further cultivated under controlled greenhouse conditions (16h daylight, 26

°C, >70% humidity). As a monocentric outgroup an individual of the ornamental plant *Juncus effusus* var. *spiralis* was commercially obtained and cultivated under controlled greenhouse conditions (16h daylight, 20°C).

METHOD DETAILS

Genome size measurement by flow cytometry

The genome size of 1C=1.6 Gb for the *R. pubera* accession sequenced here has been previously measured (Marques et al., 2015). Thus, genome size estimations by flow cytometry were performed for the accessions of *R. breviuscula* and *R. tenuis* as well as for *Juncus effusus* var. *spiralis*. For that, roughly 0.5 cm^2 of young leaf tissue was chopped with a sharp razorblade in a Petri dish together with appropriate amounts of leaf tissue of the internal reference standard *Raphanus sativus* cv. Voran (2C=1.11 pg; Genebank Gatersleben, accession number: RA 34) using the 'CyStain PI Absolute P' nuclei extraction and staining kit (Sysmex-Partec). The resulting nuclei suspension was filtered through a 50-µm filter (CellTrics, Sysmex-Partec) and measured on a CyFlow Space flow cytometer (Sysmex-Partec). The absolute DNA content (pg/2C) was calculated based on the values of the G1 peak means and the corresponding genome size (Mb/1C) according to (Dolezel et al., 2003).

Library preparations and sequencing

DNA isolation: High-molecular-weight DNA was isolated from 1.5 g of material with a NucleoBond HMW DNA kit (Macherey Nagel). Quality was assessed with a FEMTO-pulse device (Agilent), and quantity was measured with a Quantus fluorometer (Promega).

Whole-genome shotgun sequencing (WGS): Genomic DNA from *R. breviuscula* and *R. alba* were deep-sequenced with an Illumina HiSeq 3000 in 150-bp paired-end mode. Alternatively, DNBseq short read sequencing (BGI Genomics, Hong Kong) of genomic DNA was performed for *R. pubera*, *R. tenuis*, and *R. tenerrima*. Available WGS short reads from *R. cephalotes* (SRX9381225), *R. ciliata* (Ribeiro et al., 2017), *R exaltata* (SRX9381226), *R. globosa* (Ribeiro et al., 2017), and *C. littledalei* (SRX5833125, SRX5833124) were used.

PacBio: A HiFi library was then prepared according to the "*Procedure & Checklist - Preparing HiFi SMRTbell® Libraries using SMRTbell Express Template Prep Kit 2.0*" manual with an initial DNA fragmentation by g-Tubes (Covaris) and final library size binning into defined fractions by SageELF (Sage Science). Size distribution was again controlled by FEMTO-pulse (Agilent). Size-selected libraries were then sequenced on a Sequel II device with Binding kit 2.0 and Sequel II Sequencing Kit 2.0 for 30 h (*Pacific Biosciences*). The numbers of SMRT cells for each species were as follows: *R. pubera* (3 cells), *R. breviuscula* (1 cell), *R. tenuis* (2 cells), and *J. effusus* (1 cell).

Omni-C: For each species, a single chromatin-capture library was prepared from 0.5 g fresh weight material input. All treatments were according to the recommendations of the kit vendor for plants (Omni-C, Dovetail). As a final step, an Illumina-compatible library was prepared (Dovetail) and paired-end 2 x 150 bp deep-sequenced on a HiSeq 3000 (Illumina) device for *R. breviuscula*, *R. tenuis*, and *J. effusus*. Alternatively, the *R. pubera* library was paired-end 2 x 150 bp deep-sequenced using DNBseq technology (BGI Genomics, Hong Kong).

ChIPseq: ChIP DNA was quality-controlled using the NGS-assay on a FEMTO-pulse (Agilent); then, an Illumina-compatible library was prepared with the Ovation Ultralow V2 DNA-Seq library preparation kit (Tecan Genomics) and single-end 1 x 150-bp sequenced on a HiSeq 3000 (Illumina) device. For each library, an average of 20 millions reads were obtained.

Enzymatic Methyl-seq: To investigate the methylome space in *R. pubera* and *J. effusus*, the relatively non-destructive NEBNext® Enzymatic Methyl-seq Kit was employed to prepare an Illumina-compatible library, followed by paired-end sequencing (2 x 150 bp) on a HiSeq 3000 (Illumina) device. For each library, 10 Gb of reads were generated.

RNAseq: Total RNA from root, leaves, and flower buds was isolated from *R. breviuscula*. For *R. tenuis*, total RNA was isolated from flower buds only. For *J. effusus*, RNAseq data from the NCBI (accession numbers SRX2268676, SRX2268675, and SRX1639021) were used to complement its genome annotation. For *R. pubera*, total RNA was extracted from six different tissues (i.e., roots, young leaves, old leaves, stem, early flower buds, and late flower buds). Poly-A RNA was enriched from 1 µg total RNA using the NEBNext® Poly(A) mRNA Magnetic Isolation Module. RNAseq libraries were prepared as described in the NEBNext Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs). A total of 11 cycles were applied to enrich library concentration. Sequencing was done at BGI Genomics (Hong Kong) with a BGISEQ-500 system in the DNBseq platform in paired-end mode 2 x 150 bp.

IsoSeq: For the proper annotation of the complex *R. pubera* genome, total RNA was extracted from six different tissues (i.e., roots, young leaves, old leaves, stem, early flower buds and late flower buds) and quality-assessed by a Nanochip (Agilent Bioanalyser, Santa Clara, U.S.A.). Next, cDNA was synthesized according to the TeloPrime Version 2 kit (Lexogen, Vienna, Austria). We exchanged the Lexogen first-strand synthesis oligo-dT primer for the (5'-AAGCAGTGGTATCAACGCAGAGTACT(30)VN-3') primer to introduce a 3' anchor base. Then, the optimal number of cycles was determined by qPCR (Viia7, Applied Biosystems) with the 1x Evagreen fluorochrome (Biotium, Fremont, U.S.A.), TeloPrime kit chemistry and 25% of the cDNA as input. The forward primer was FP from the TeloPrime kit, and the reverse primer was 5'-AAGCAGTGGTATCAACGCAGAGTAC-3'. The residual cDNA was mass-amplified with an extended Lexogen FP primer by adding 16mer barcodes as recommended by PacBio at the 5' end and a cycle number by which 80% of the maximal fluorescence signal was reached. The PCR products were bead-purified (Pronex beads, Promega) followed by PacBio library preparation with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, U.S.A.), and then quantity- (Quantus, Promega) and quality-assessed (Agilent Bioanalyser). Long-read sequencing was performed on a Sequel II sequencer with a Sequel II Binding kit 2.1, Sequel II Sequencing Kit 2.0 sequencing chemistry 2.0, and a single 8M SMRT cell (*Pacific Biosciences*, Menlo Park, U.S.A.). The movie time was 30 h after a 2-h immobilization step and 2-h pre-extension step to adjust for high-fidelity (HiFi) sequencing.

Genome size estimation using k-mer frequency

Genome sizes of the three *Rhynchospora* species and *J. effusus* were also confirmed by k-mer frequency analysis with the findGSE tool (Sun et al., 2018), after counting k-mers with Jellyfish (Marcais and Kingsford, 2011). High-coverage short reads were used as follows: *R. pubera* (60x), *R. breviuscula* (50x), and *R. tenuis* (130x). Since for *J. effusus* we did not have short-read data, we used our high-coverage HiFi PacBio reads (70x).

Sequence-based ploidy assessment

We used Smudgeplot (Ranallo-Benavidez et al., 2020) to visualize and estimate the ploidy and structure of the sequenced genomes. This tool can infer ploidy directly from the k-mers present in sequencing reads by analyzing heterozygous k-mer pairs.

Genome assembly

Reads obtained by the sequencing process were subjected to assembly using the HiCanu function of Canu (Nurk et al., 2020) for *R. pubera* with the HiCanu command line: *canu -assemble -p Hicanuassembly -d Hicanutest genomeSize=1.6g maxThreads=40 useGrid=false -pacbio-hifi *.fastq*. Alternatively, Hifiasm (Cheng et al., 2021) was used for the assembly of *R. breviuscula*, *R. tenuis,* and *J. effusus* with Hifiasm command: *hifiasm -o output.asm -t 40 reads.fq.gz.* Preliminary assemblies were evaluated for contiguity and completeness with BUSCO (Seppey et al., 2019).

Optical map and hybrid scaffolding

We developed an optical mapping strategy to help resolve the complexity of the *R. pubera* genome. High-molecular-weight DNA was prepared from young leaves of *R. pubera*. A total of 3.15 million cell nuclei were purified by flow cytometry, pelleted by centrifugation (30 min at 300 *g*), and embedded in four agarose plugs of 20-µL volume. The nuclear DNA was purified in the plugs as described by Šimková et al. (2003) with an increased concentration of proteinase K (1 mg/mL of lysis buffer). The proteinase- and RNase-treated DNA was isolated from the agarose gel, and the resulting 525 ng DNA was directly labeled at DLE-1 recognition sites following the standard Bionano Prep Direct Label and Stain (DLS) Protocol (Bionano Genomics, San Diego, USA) and analyzed on the Saphyr platform of Bionano Genomics. A total of 1.27 Tbp of single-molecule data with N50 of 236 kb, corresponding to effective coverage of 96.8x of the *R. pubera* genome, was used in *de novo* assembly by Bionano Solve 3.6.1_11162020, using a standard configuration file "optArguments_nonhaplotype_noES_noCut_DLE1_saphyr.xml" (**Table S6**). A p-value threshold of 1e–11 was used to build the initial assembly, a p-value of 1e–12 was used for extension and refinement steps (five rounds), and a p-value of 1e–16 was used for final map merging. To improve the contiguity of the sequence assembly, an automatic hybrid scaffold pipeline integrated in Bionano Solve 3.6.1_11162020 was run with the *de novo* optical map assembly. The default *DLE-1 Hybrid Scaffold* configuration file was used with the "Resolve conflict" option for conflict resolution. The conflicts between sequences and the optical map were manually curated, and the pipeline was re-run using the modified *conflict_cut_statu.txt* file (**Table S7**). The results obtained from the optical mapping scaffolding of the genome assembly of *R. pubera* were used as input for Omni-C scaffolding.

Omni-C scaffolding

Omni-C scaffolding was performed using the HiC-pipeline pre-processing scripts from https://github.com/esrice/hic-pipeline and SALSA2 (Ghurye et al., 2019) with default parameters. After testing several minimum mapping quality values of bam alignments, final scaffolding was performed with MAPQ10. Several rounds of assembly correction guided by Omni-C contact maps and manual curation of scaffolds were performed to obtain the pseudomolecules.

Assembly and scaffolding strategy

The rather homozygous genome of *J. effusus* was estimated to be close to 1C=271 Mb (**Figure 1; Figure S3Q**). Sequencing of *J. effusus* var. *spiralis* yielded 19 Gb of reads and an initial assembly of 258 Mb (79× coverage, N_{50} = 11 Mb, **Figure 1**), where 18 contigs corresponded to complete chromosomes. The assembly was further scaffolded to the expected 21 pseudomolecules (240 Mb), and unplaced contigs contained 18 Mb, corresponding to the complete haploid chromosome set of the species (**Figure 2A; Figure 1**). The sequencing of *R. pubera*, which is an inbred species, yielded 66 Gb of PacBio HiFi reads, and the initial assembly spanned 1.7 Gb (38× coverage, N_{50} = 11.2 Mb). After removing redundant sequences likely due to some small residual heterozygosity, the assembly closely matched its estimated haploid genome size (**Figure 1, Figure S3H**). A first scaffolding using optical mapping was followed by a second scaffolding using chromosome conformation capture (Omni-C, Dovetail™) of the genome assembly, which yielded five very large pseudomolecules (1.47 Gb, N_{50} = 361 Mb), while unplaced contigs contained 141 Mb (**Figure 3E, Figure 1**). The sequencing of *R. breviuscula* yielded 30 Gb of PacBio HiFi reads, resulting in an initial assembly that was 813 Mb in length. In contrast to *R. pubera*, *R. breviuscula* is outbred, which resulted in an assembly of its diploid genome size showing a high level of heterozygosity confirmed by k-mer analysis (**Figure S3K**). We pruned the resulting large contigs to the single largest representative haplotype (75× coverage, 421 Mb, N_{50} = 11 Mb; **Figure 1**) and then oriented and ordered it into five pseudomolecules using Omni-C scaffolding comprising 370 Mb (N_{50} = 71 Mb; **Figure S1A**). Unplaced contigs contained 50 Mb (**Figure 1**). The sequenced genome of *R. tenuis* yielded 45.9 Gb of PacBio HiFi reads resulting in an assembly of 770 Mb, which closely corresponds to its diploid genome size, showing a high level of heterozygosity (**Figure S3N**). We pruned the resulting large contigs to the single largest representative haplotype (120× coverage, 395 Mb, N_{50} = 19 Mb, **Figure 1**), which was oriented and ordered into two pseudomolecules of about 350 Mb (N_{50} = 215 Mb; **Figure S1B**). Unplaced contigs contained 47 Mb (**Figure 1**).

Generation of Hi-C maps

Final Hi-C maps of *R. pubera*, *R. breviuscula*, *R. tenuis*, and *J. effusus* were generated by Juicer (v1.6) (Durand et al., 2016) using the sequencing data from DNase *in situ* Hi-C (Omni-C) experiments. Specifically, technical replicates were aligned and deduplicated and then the results of each replicate were merged by MEGA from Juicer.

Quantitative analysis of Hi-C contacts

The python version of Straw (strawC v0.0.9) (Durand et al., 2016) was used to extract Hi-C counts from the illustrated Hi-C maps (**Figure 2**; **Figure S1**) in 1-Mb resolution and with the normalization approach of Vanilla Coverage (VC). To represent the intra- and interchromosomal interactions in an intuitive manner, the *cis* Hi-C contact of a chromosome was quantified as the sum of all Hi-C counts within the chromosome *per se*, while *trans* Hi-C contacts referred to the sum of Hi-C counts between the designated chromosome and all other chromosomes. The final intra- and interchromosomal contacts for each single chromosome were normalized through the percentages of Hi-C counts over the sum of all Hi-C signals in the corresponding Hi-C map. It is also noteworthy that the infinite extracted Hi-C counts through Straw were replaced by the mean of all other finite counts within the extracted chromosomal pair.

ChIP

ChIP experiments were performed following Reimer and Turck (2010), with adjustments for *R. pubera* and *R. breviuscula*. Unopened flower buds were harvested and frozen in liquid nitrogen until sufficient material was obtained. The samples were fixed in 4% formaldehyde for 30 min and the chromatin was sonicated for 25 min. Then, 7–85 µL of sonicated chromatin was incubated with 2 ng of respective antibody overnight. Immunoprecipitation was carried out for rabbit anti-*Rp*CENH3, for *R. pubera* and *R. breviuscula*, and for rabbit anti-H3K4me3 (abcam, ab8580), and mouse anti-H3K9me2 (abcam, ab1220). Recombinant rabbit IgG (abcam, ab172730) and no-antibody inputs were used as controls. Two experimental replications were also maintained for all the combinations. After overnight incubation of chromatin with antibody, protein beads (anti-mouse: Protein G Sepharose 4 Fast Flow, anti-rabbit: rProtein A Sepharose Fast Flow) were added to the chromatin-antibody mixture. The bound chromatin was finally eluted, de-crosslinked, precipitated, and sent for sequencing.

Synteny analysis

The synteny analysis shown in **Figure 5** was performed using the MCscan pipeline implemented in the JVCI utility libraries (Tang et al., 2008). For this analysis, CDS sequences of the longest transcript were used. Circular plots were drawn with the circos package (Krzywinski et al., 2009).

Whole-genome alignment (WGA)

A whole-genome alignment (WGA) between *R. pubera*, *R. tenuis*, *R. breviuscula, J. effusus*, and *C. littledalei* was generated using the Cactus pipeline (Version 1.0) (Paten et al., 2011). Prior to the alignment step, all nucleotide sequences were 20-kmer-softmasked to reduce complexity and facilitate construction of the WGA using the tallymer subtools from the genome tools package (Version 1.6.1) (Kurtz et al., 2008). The Cactus pipeline was run stepwise with the default settings described at https://github.com/ComparativeGenomicsToolkit/cactus#running-step-by-step.

Self-synteny

SyMAP v. 5.0.6 was applied to perform both synteny and self-synteny analyses (Soderlund et al., 2011; Soderlund et al., 2006). Circular self-synteny plots were obtained with SyMAP or RIdeogram software (Hao et al., 2020) using the synteny calculation blocks obtained from SyMAP.

Characterization of end-to-end fusions

For the characterization of the regions involved in EEFs observed in *R. pubera* and *R. tenuis*, we first compared the synteny between their genomes with *R. breviuscula* used as a reference. This allowed us to pin the putative regions around the borders of the fusion events. Afterwards, genes and *Tyba* repeats were loaded as annotation features on SyMAP. This further allowed us to detect the sequence types in the putative translocated regions. In *R. pubera*, we counted 15 potential EEF events, of which 11 regions had a *Tyba* array right in the middle between two ancestral syntenic chromosomes of *R. breviuscula*. Further inspection and characterization of such regions were done by checking the genome coordinates and annotation features with IGV and Geneious (Kearse et al., 2012), which revealed a remnant rDNA cluster involved in the EEF of two ancestral *Rb*3 in the

*Rp*Chr3. This detailed analysis further allowed us to reconstruct the karyotype history of *R. pubera* based on the shared EEF signatures found in the genome.

Whole-genome duplication analysis

To identify ancient WGD events, we performed Ka/Ks analysis on the fully annotated genomes with the SynMap available on CoGe (genomevolution.org).

Gene annotation

Structural gene annotation was done combining *de novo* gene calling and homology-based approaches with RNAseq, IsoSeq, and protein datasets.

Using evidence derived from expression data, RNAseq data were first mapped using STAR (Dobin et al., 2013) (version 2.7.8a) and subsequently assembled into transcripts by StringTie (Kovaka et al., 2019) (version 2.1.5, parameters -m 150-t -f 0.3). *Triticeae* protein sequences from available public datasets (UniProt, https://www.uniprot.org, 05/10/2016) were aligned against the genome sequence using GenomeThreader (Gremme et al., 2005) (version 1.7.1; arguments -startcodon -finalstopcodon -species rice -gcmincoverage 70 -prseedlength 7 -prhdist 4). Isoseq datasets were aligned to the genome assembly using GMAP (Wu and Watanabe, 2005) (version 2018-07-04). All transcripts from RNAseq, IsoSeq, and aligned protein sequences were combined using Cuffcompare (Ghosh and Chan, 2016) (version 2.2.1) and subsequently merged with StringTie (version 2.1.5, parameters --merge -m150) into a pool of candidate transcripts. TransDecoder (version 5.5.0; http://transdecoder.github.io) was used to find potential open reading frames and to predict protein sequences within the candidate transcript set.

Ab initio annotation was initially done using Augustus (Hoff and Stanke, 2019) (version 3.3.3). GeneMark (Ter-Hovhannisyan et al., 2008) (version 4.35) was additionally employed to further improve structural gene annotation. To avoid potential over-prediction, we generated guiding hints using the above described RNAseq, protein, and IsoSeq datasets as described by Hoff and Stanke (2019). A specific Augustus model for *Rhynchospora* was built by generating a set of gene models with full support from RNAseq and IsoSeq. Augustus was trained and optimized using the steps detailed by Hoff and Stanke (2019).

To maximize uniformity across all annotated species, Augustus was also run in comparative annotation mode (Nachtweide and Stanke, 2019). The generated WGA served as sequence input together with the mapping of RNAseq data as described above.

All structural gene annotations were joined using EVidenceModeller (Haas et al., 2008) (version 1.1.1), and weights were adjusted according to the input source: *ab initio* (Augustus: 5, GeneMark: 2), homology-based (10), and comparative *ab initio* (7). Additionally, two rounds of PASA (Haas et al., 2003) (version 2.4.1) were run to identify untranslated regions and isoforms using transcripts generated by a genome-guided TRINITY (Grabherr et al., 2011) (version 2.13.1) assembly derived from *Rhynchospora* RNAseq data and the above described IsoSeq datasets.

We used BLASTP (Altschul et al., 1990) (ncbi-blast-2.3.0+, parameters -max_target_seqs 1 -evalue 1e–05) to compare potential protein sequences with a trusted set of reference proteins (Uniprot Magnoliophyta, reviewed/Swissprot, downloaded on 3 Aug 2016; https://www.uniprot.org). This differentiated candidates into complete and valid genes, non-coding transcripts, pseudogenes, and transposable elements. In addition, we used PTREP (Release 19; http://botserv2.uzh.ch/kelldata/trep-db/index.html), a database of hypothetical proteins containing deduced amino acid sequences in which internal frameshifts have been removed in many cases. This step is particularly useful for the identification of divergent transposable elements with no significant similarity at the DNA level. Best hits were selected for each predicted protein from each of the three databases. Only hits with an e-value below 10e–10 were considered. Furthermore, functional annotation of all predicted protein sequences was done using the AHRD pipeline (https://github.com/groupschoof/AHRD).

Proteins were further classified into two confidence classes: high and low. Hits with subject coverage (for protein references) or query coverage (transposon database) above 80% were considered significant and protein sequences were classified as high-confidence using the following criteria: protein sequence was complete and had a subject and query coverage above the threshold in the UniMag database or no BLAST hit in UniMag but in UniPoa and not PTREP; a low-confidence protein sequence was incomplete and had a hit in the UniMag or UniPoa database but not in PTREP. Alternatively, it had no hit in UniMag, UniPoa, or PTREP, but the protein sequence was complete. In a second refinement step, low-confidence proteins with an AHRD-score of 3* were promoted to high-confidence.

BUSCO (Seppey et al., 2019) (version 5.1.2.) was used to evaluate the gene space completeness of the pseudomolecule assembly and structural gene annotation with the 'viridiplantae_odb10' database containing 425 single-copy genes.

Orthogroup analysis

Orthogroup assignments (**Table S4**) was performed with OrthoFinder (Emms and Kelly, 2019). For GO term enrichment, a GO annotation file (gaf; 2.1) was built using all GO terms assigned by the functional annotations of *R. pubera*, *R. breviuscula*, *R. tenuis*, and *J. effusus*. GO term enrichment was performed by feeding GO terms of the shared orthologos into Ontologiser (ontologiser.de). *P*-values were corrected using the Benjamini-Hochberg procedure. We used the UpSetR in the R package (http://gehlenborglab.org/research/projects/upsetr/) to analyze how many orthogroups are shared between the five species or are unique to a single species.

De novo *repeat discovery and annotation*

To identify the overall repetitiveness of genomes we performed *de novo* repeat discovery with RepeatExplorer2 (Novak et al., 2020) for nine species of *Rhynchospora*, *C. littledalei*, and *J.*

effusus. We used a repeat library obtained from the RepeatExplorer2 analysis of Illumina paired-end reads. All clusters representing at least 0.005% of the genomes were manually checked, and the automated annotation was corrected if needed. Contigs from the annotated clusters were used to build a repeat library. To minimize potential conflicts due to the occasional presence of contaminating sequences in the clusters, only contigs with average read depths ≥ 5 were included and all regions in these contigs that had read depths < 5 were masked. Genome assemblies were then annotated using custom RepeatMasker (REF - Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org) search with options -xsmall -no_is -e ncbi -nolow. Output from RepeatMasker was parsed using custom scripts (https://github.com/kavonrtep/repeat_annotation_pipeline) to remove overlapping and conflicting annotations.

Transposable element protein domains (Neumann et al., 2019) found in the assembled genomes were annotated using the DANTE tool available from the RepeatExplorer2 Galaxy portal. To find master *Helitron* elements related to *TCR1*, we first searched the genome assembly for *Helitron* helicase-coding sequences using DANTE (https://repeatexplorer-elixir.cerit-sc.cz/galaxy/) exploiting the REXdb database (Neumann et al., 2019) (Viridiplantae_version_3.0) and then manually identified boundaries of full-length *Helitron* elements. We identified 111 putative autonomous *Helitron*s and compared their terminal sequences with *TCR1*. This revealed that *TCR1* is most similar to the *Helitron-27*, sharing 90% and 100% identity over 30-bp sequences at the 5' and 3' ends, respectively (**Figure 4K–L**), meeting the criteria for classification of *TCR1* and *Helitron-27* into the same family (Thomas and Pritham, 2015). To find *TCR1* insertions in the *R. pubera* genome, we performed iterative blastn searches using 30-bp sequences from their 5' and 3' termini and consensus sequences of *Tyba*.

To obtain the average number of *Tyba* arrays for each *Rhynchospora* genome, we first removed spurious low-quality *Tyba* monomer annotations with less than 500 bp and merged with bedtools all adjacent *Tyba* monomers situated at a maximum distance of 50 kb into individual annotations to eliminate the gaps that arise because of fragmented *Tyba* arrays. Length and distance between *Tyba* arrays were then calculated using bedtools. Bar plots of the average distance and unit length used to compare the *Tyba* arrays among the three *Rhynchospora* species were made in RStudio using ggplot library.

Detection of dyad symmetries in Tyba repeats

Dyad simmetries detection was performed as reported in Kasinathan and Henikoff (2018). We used EMBOSS palindrome (Rice et al., 2000) to detect perfect dyad symmetries in the *Tyba* consensus of the three *Rhynchospora* species with the following parameters:

-minpallen 4 -maxpallen 100 -gaplimit 20 -nummismatches 0 –overlap

ChIP-seq analysis

Raw sequencing reads were trimmed by Cutadapt (Martin 2011) to remove low-quality nucleotides (with quality score less than 30) and adapters. Trimmed ChIPed 150-bp single-end reads were mapped to the respective reference genome with bowtie2 (Langmead and Salzberg, 2012) with default parameters, where all read duplicates were removed and only the single best matching read was kept on the final alignment BAM file. BAM files were converted into BIGWIG coverage tracks using the bamCompare tool from deeptools (Ramirez et al., 2016). The coverage was calculated as the number of reads per 50-bp bin and normalized by reads per kilobase per million mapped reads (RPKM). Plots of detailed chromosome regions showing multiple tracks presented in **Figure 4** and **Figure 6** were done with pyGenomeTracks (Lopez-Delisle et al., 2021).

CENH3 domains were identified by comparing the ChIPed and input data using MACS3 (Zhang et al., 2008). The parameters for MACS3 included -B –broad –g 1470000000 –trackline. As an alternative method for detection of CENH3 domains, we compared input and ChIP using the epic2 program for detection of diffuse domains (Stovner and Saetrom, 2019). Parameters for epic2 included --bin-size 2000. Only CENH3 domains detected with both methods were kept for further analysis.

To determine the sizes and positions of centromere units, we merged with bedtools CENH3 peaks that were separated by less than 50 kb to eliminate the gaps that arise because of fragmented *Tyba* arrays or due to insertion of TEs. Small CENH3 domains of less than 1 kb were discarded. Length and distance between *Tyba* arrays and between CENH3 domains were then calculated using bedtools.

Bar plots of the average distance and unit length used to compare CENH3 domains and *Tyba* arrays were made in RStudio using the ggplot library.

The obtained repeat annotation was used to evaluate the association of individual classes of repetitive sequences with the CENH3 domain in *R. pubera*. For each repeat type, we calculated the total abundance in the genome as a sum of repetitive element length and compared it with abundance of repetitive elements located within CENH3 domains. For each type of repetitive element, we calculated the observed/expected ratio using:

$$OE=\frac{\sum\left(R_{CENH3}\right)}{\sum\left(\frac{L_{CENH3}}{L_G}\right)R_G}$$

where R_{CENH3} is length of repeat located within CENH3 domains, L_{CENH3} is the length of CENH3-binding regions, L_G is total genome size, and R_G is total length of repeat type in the genome.

Identification of paralogous CENH3 domains

To identify groups of paralogous CENH3 domains within the blocks of homologous regions of *R. pubera*, we identified the two nearest paralogous genes on both sides of each CENH3 domain. Subsequently, the groups of four genes surrounding CENH3 domains were used to identify corresponding regions on the other homologous blocks where we checked for the presence of the CENH3 domain. Resulting groups of four homologous regions were manually inspected using dotplot (https://doi.org/10.1093/bioinformatics/btm039) and the IGV (https://doi.org/10.1038/nbt.1754) browser.

Methyl-seq analysis

To comparatively evaluate the DNA methylation context of a holocentric and monocentric genome, we applied Methyl-seq and used the Bismarck pipeline (Krueger and Andrews, 2011) to analyze the data. Individual methylation context files for CpG, CHG, and CHH were converted to BIGWIG format and used as input track for overall genome-wide DNA methylation visualization with pyGenomeTracks.

Metaplots

Analysis of the enrichment of all ChIP treatment files was performed as follows: BAM files of each ChIP treatment were normalized to the ChIP Input BAM file by RPKM using bamCompare available from deeptools. The generated normalized BIGWIG files were used to calculate the level of enrichment associated with gene bodies, *Tyba* repeats, CENH3 domains, and TEs using computeMatrix scale-regions (parameters: --regionBodyLength 4000 –beforeRegionStartLength 2000 –afterRegionStartLength 2000) also available from deeptools. Finally, metaplots for all ChIPseq treatment files were plotted with plotHeatmap available from deeptools (Ramirez et al., 2016). Additionally, coverage BIGWIG files of transcriptional activity (RNAseq) and all DNA methylation contexts were also used to calculated their enrichment on gene bodies, *Tyba* repeats, CENH3 domains, and TEs with computeMatrix and plotting with plotHeatmap.

Dating WGD events

To date the two rounds of duplication of the genome of *R. pubera*, a large tree of concatenated single copy genes was produced. For this analysis, each of the four homologous regions of *R. pubera* were separated and treated as a tip in the subsequent phylogeny reconstructions. Only coding sequences were used. We used BUSCO (*Poales* dataset) (Seppey et al. 2019) to look for conserved single-copy genes that are shared by all selected datasets. We performed this analysis in three different ways: solely the large syntenic block of *R. pubera*, solely the smaller syntenic bloc of *R. pubera*, and the two blocks combined. For the analyses, we included the following nine datasets: *J. effusus*, *Carex littledalei*, *R. tenuis*, *R. breviuscula,* and the four homologous blocks of *R. pubera*. BUSCO analyses were run for all datasets; all the resulting single-copy genes were selected for each dataset. The single-copy genes shared among all datasets were used for the analyses: 841 for the larger block1, 400 for the smaller block2, and 1,204 for the two blocks combined. All genes were then aligned with MAFFT (Katoh and Standley, 2013), trimmed with Trimal (Capella-Gutierrez et al., 2009), and concatenated into a single large multi-fasta alignment, and used as input for a ML tree built with IQtree (Minh et al., 2020).

A molecular clock analysis was performed to explore genome evolution in *Rhynchospora* and related genera. Divergence times were estimated using BEAST v.1.10.4 (Drummond and Rambaut, 2007) through the CIPRES Science Gateway fixing the tree topology from the Bayesian inference of the *Rhynchospora* concatenated 1,204 BUSCO gene alignment. Uncorrelated relaxed lognormal clock (Drummond and Rambaut, 2007) and Birth-Death speciation model (Gernhard, 2008) were applied. Two independent runs of 100,000,000 generations were performed, sampling every 10,000 generations. After removing 25% of samples as burn-in, the independent runs were combined and a maximum clade credibility (MCC) tree was constructed using TreeAnnotator v.1.10.4 (Drummond and Rambaut, 2007). To verify the effective sampling of all parameters and assess convergence of independent chains, we examined their posterior distributions in TRACER. The MCMC sampling was considered sufficient at effective sampling sizes (ESSs) equal to or higher than 200. The phylogeny was dated using both fossils and secondary calibration from published dated phylogenies. We chose three calibration points: i) Juncaceae/Cyperaceae divergence at 72.0 Mya (Bremer, 2002); ii) a fossil for *Carex* at 37.8 MYA (Smith et al., 2010), and iii) *R. pubera*/*R. tenuis* divergence at 32.0 Mya (Buddenhagen, 2016).

Fluorescence in situ *hybridization (FISH)*

Interphase nuclei were prepared using the air-drying method, after enzymatic digestion with 2% cellulase Onozuka and 20% pectinase Sigma (Ribeiro et al., 2017). Roots were fixed in Carnoy ethanol:acetic acid 3:1 (v/v) for 2 h and stored at –20 °C. The best slides were selected for FISH, performed as described by Pedrosa et al. (2002) and the slides were counterstained with 2 µg/mL DAPI in Vectashield (Vector) mounting buffer. *Juncus effusus* interphase nucleus was hybridized with directly labeled (FAM)TTTAGGG(8)-telomeric probe and a directly labeled (CY3) probe for its most abundant satellite repeat, while *R. breviuscula* nucleus was hybridized with the same telomeric probe and directly labeled *Tyba* (CY3) oligo-probe.

Immunostaining

Immunostaining was performed as described before by Marques et al. (2016). Rabbit anti-H3K4me3 (abcam, ab8580), mouse anti-H3K9me2 (abcam, ab1220), and previously generated *R. pubera* rabbit anti-CENH3 (Marques et al., 2015) were used for immunostaining.

QUANTIFICATION AND STATISTICAL ANALYSIS

Comparison of Hi-C contacts

The chromosomal interactions between holo- and monocentric plant species were compared by the ratios of *cis* and *trans* Hi-C contacts, i.e., for each species, we quatified the ratios of *cis* and *trans*

Hi-C counts for every chromosome and tested if they were significantly different across distinct species. For grouped comparison, we adopted the mutiple testing method of one-way ANOVA (Analysis of Variance), specifically the Kruskal-Wallis ranked test with Holm-Bonferroni correction, because the compared values and ratios of intra- and inter-chromosomal contacts were different in length among various species and were not supported by evidence such as normality. Pair-wise significance analysis was conducted using Dunn's *post hoc* test.

Tyba array and CENH3 domain size and spacing

The Dunn's test was used to compare pairwise distributions of values of interest between *Tyba* arrays and CENH3 domains size and spacing.

REFERENCES

Achrem, M., Szucko, I., and Kalinka, A. (2020). The epigenetic regulation of centromeres and telomeres in plants and animals. Comp Cytogenet *14*, 265-311.

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

Bremer, K. (2002). Gondwanan evolution of the grass alliance of families (Poales). Evolution *56*, 1374-1387.

Buddenhagen, C.E. (2016). A View of Rhynchosporeae (Cyperaceae) Diversification before and after the Application of Anchored Phylogenomics Across the Angiosperms. In Department of Biological Science (Florida: Florida State University).

Burchardt, P., Buddenhagen, C.E., Gaeta, M.L., Souza, M.D., Marques, A., and Vanzela, A.L.L. (2020). Holocentric Karyotype Evolution in Rhynchospora Is Marked by Intense Numerical, Structural, and Genome Size Changes. Frontiers in plant science *11*.

Camara, A.S., Schubert, V., Mascher, M., and Houben, A. (2021). A simple model explains the cell cycle-dependent assembly of centromeric nucleosomes in holocentric species. Nucleic acids research.

Can, M., Wei, W., Zi, H., Bai, M., Liu, Y., Gao, D., Tu, D., Bao, Y., Wang, L., Chen, S.*, et al.* (2020). Genome sequence of Kobresia littledalei, the first chromosome-level genome in the family Cyperaceae. Sci Data *7*, 175.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972-1973.
Castiglione, M.R., and Cremonini, R. (2012). A fascinating island: 2n=4. Plant Biosystems *146*, 711-726.

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

Cicconardi, F., Lewis, J.J., Martin, S.H., Reed, R.D., Danko, C.G., and Montgomery, S.H. (2021). Chromosome Fusion Affects Genetic Diversity and Evolutionary Turnover of Functional Loci but Consistently Depends on Chromosome Size. Molecular biology and evolution *38*, 4449-4462.

Cortes-Silva, N., Ulmer, J., Kiuchi, T., Hsieh, E., Cornilleau, G., Ladid, I., Dingli, F., Loew, D., Katsuma, S., and Drinnenberg, I.A. (2020). CenH3-Independent Kinetochore Assembly in Lepidoptera Requires CCAN, Including CENP-T. Curr Biol *30*, 561-572 e510.

Costa, L., Marques, A., Buddenhagen, C., Thomas, W.W., Huettel, B., Schubert, V., Dodsworth, S., Houben, A., Souza, G., and Pedrosa-Harand, A. (2021). Aiming off the target: recycling target capture sequencing reads for investigating repetitive DNA. Annals of botany.

Despot-Slade, E., Mravinac, B., Sirca, S., Castagnone-Sereno, P., Plohl, M., and Mestrovic, N. (2021). The Centromere Histone Is Conserved and Associated with Tandem Repeats Sharing a

Conserved 19-bp Box in the Holocentromere of Meloidogyne Nematodes. Molecular biology and evolution *38*, 1943-1965.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. Cytometry A *51*, 127-128; author reply 129.

Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology *7*, 214.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst *3*, 99-101.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome biology *20*, 238.

Escudero, M., Marquez-Corro, J.I., and Hipp, A.L. (2016). The Phylogenetic Origins and Evolutionary History of Holocentric Chromosomes. Systematic Botany *41*, 580-585.

Feng, S., Zhong, Z., Wang, M., and Jacobsen, S.E. (2020). Efficient and accurate determination of genome-wide DNA methylation patterns in Arabidopsis thaliana with enzymatic methyl sequencing. Epigenetics Chromatin *13*, 42.

Fernandes, J.B., Wlodzimierz, P., and Henderson, I.R. (2019). Meiotic recombination within plant centromeres. Current opinion in plant biology *48*, 26-35.

Fuchs, J., Demidov, D., Houben, A., and Schubert, I. (2006). Chromosomal histone modification patterns--from conservation to diversity. Trends in plant science *11*, 199-208.

Gassmann, R., Rechtsteiner, A., Yuen, K.W., Muroyama, A., Egelhofer, T., Gaydos, L., Barron, F., Maddox, P., Essex, A., Monen, J.*, et al.* (2012). An inverse relationship to germline transcription defines centromeric chromatin in *C. elegans*. Nature *484*, 534-537.

Gernhard, T. (2008). The conditioned reconstructed process. Journal of Theoretical Biology *253*, 769-778.

Ghosh, S., and Chan, C.K. (2016). Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol Biol *1374*, 339-361.

Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M., and Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol *15*, e1007273.

Gohard, F.H., Zhiteneva, A.A., and Earnshaw, W.C. (2014). Centromeres. In In: eLS John Wiley & Sons, Ltd: Chichester.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.D., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology *29*, 644-U130.

Gremme, G., Brendel, V., Sparks, M.E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. Information and Software Technology *47*, 965-978.
Guerra, M. (2016). Agmatoploidy and symploidy: a critical review. Genet Mol Biol *39*, 492-496.

Guerra, M., Ribeiro, T., and Felix, L.P. (2019). Monocentric chromosomes in Juncus (Juncaceae) and implications for the chromosome evolution of the family. Bot J Linn Soc *191*, 475-483.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., *et al.* (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research *31*, 5654-5666.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology *9*, R7.

Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., and Chen, J. (2020). RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci *6*, e251.

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

Hill, J., Rastas, P., Hornett, E.A., Neethiraj, R., Clark, N., Morehouse, N., de la Paz Celorio-Mancera, M., Cols, J.C., Dircksen, H., Meslin, C., *et al.* (2019). Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution. Sci Adv *5*, eaau3648.

Hoencamp, C., Dudchenko, O., Elbatsh, A.M.O., Brahmachari, S., Raaijmakers, J.A., van Schaik, T., Sedeno Cacciatore, A., Contessoto, V.G., van Heesbeen, R., van den Broek, B., *et al.* (2021). 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. Science *372*, 984-989.

Hoff, K.J., and Stanke, M. (2019). Predicting Genes in Single Genomes with AUGUSTUS. Curr Protoc Bioinformatics *65*, e57.

Kasinathan, S., and Henikoff, S. (2018). Non-B-Form DNA Is Enriched at Centromeres. Molecular biology and evolution *35*, 949-962.

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

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

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome biology *20*, 278.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571-1572.
Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome research *19*, 1639-1645.

Kurtz, S., Narechania, A., Stein, J.C., and Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC genomics *9*, 517.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Lonnig, W.E., and Saedler, H. (2002). Chromosome rearrangements and transposable elements. Annual Review of Genetics *36*, 389-410.

Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Gruning, B., Ramirez, F., and Manke, T. (2021). pyGenomeTracks: reproducible plots for multivariate genomic datasets. Bioinformatics *37*, 422-423.

Lucek, K., Augustijnen, H., and Escudero, M. (2022). A holocentric twist to chromosomal speciation? Trends Ecol Evol.

Lukhtanov, V.A., Dinca, V., Friberg, M., Sichova, J., Olofsson, M., Vila, R., Marec, F., and Wiklund, C. (2018). Versatility of multivalent orientation, inverted meiosis, and rescued fitness in holocentric chromosomal hybrids. Proceedings of the National Academy of Sciences of the United States of America *115*, E9610-E9619.

Mandakova, T., Joly, S., Krzywinski, M., Mummenhoff, K., and Lysak, M.A. (2010). Fast Diploidization in Close Mesopolyploid Relatives of Arabidopsis. The Plant cell *22*, 2277-2290.

Mandakova, T., and Lysak, M.A. (2018). Post-polyploid diploidization and diversification through dysploid changes. Current opinion in plant biology *42*, 55-65.

Mandrioli, M., and Manicardi, G.C. (2020). Holocentric chromosomes. PLoS genetics *16*.

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

Marques, A., Ribeiro, T., Neumann, P., Macas, J., Novak, P., Schubert, V., Pellino, M., Fuchs, J., Ma, W., Kuhlmann, M., *et al.* (2015). Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. Proceedings of the National Academy os Sciences of the USA *112*, 13633-13638.

Marques, A., Schubert, V., Houben, A., and Pedrosa-Harand, A. (2016). Restructuring of Holocentric Centromeres During Meiosis in the Plant Rhynchospora pubera. Genetics *204*, 555-568.

Mayrose, I., and Lysak, M.A. (2021). The Evolution of Chromosome Numbers: Mechanistic Models and Experimental Approaches. Genome biology and evolution *13*.

Melters, D.P., Paliulis, L.V., Korf, I.F., and Chan, S.W. (2012). Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. Chromosome Research *20*, 579-593.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution.

Mizuno, H., Kawahara, Y., Wu, J., Katayose, Y., Kanamori, H., Ikawa, H., Itoh, T., Sasaki, T., and Matsumoto, T. (2011). Asymmetric distribution of gene expression in the centromeric region of rice chromosome 5. Frontiers in plant science *2*, 16.

Muller, H., Gil, J., Jr., and Drinnenberg, I.A. (2019). The Impact of Centromeres on Spatial Genome Architecture. Trends Genet *35*, 565-578.

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

Nachtweide, S., and Stanke, M. (2019). Multi-Genome Annotation with AUGUSTUS. Methods Mol Biol *1962*, 139-160.

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A.J., Abramson, B.W., Schmucker, A., Mandakova, T., Jamge, B., Lambing, C., Kuo, P*., et al.* (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. Science *374*, eabi7489.

Neumann, P., Novak, P., Hostakova, 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.

Nhim, S., Gimenez, S., Nait-Saidi, R., Severac, D., Nam, K., d'Alençon, E., and Nègre, N. (2021). H3K9me2 genome-wide distribution in the holocentric insect <em>Spodoptera frugiperda</em> (Lepidoptera: Noctuidae). bioRxiv, 2021.2007.2007.451438.

Novak, P., Neumann, P., and Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nature protocols *15*, 3745-3776.

Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R., Logsdon, G.A., Grothe, R., Miga, K.H., Eichler, E.E., Phillippy, A.M., and Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome research *30*, 1291-1305.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., and Haussler, D. (2011). Cactus: Algorithms for genome multiple sequence alignment. Genome research *21*, 1512-1528.

Pedrosa, A., Sandal, N., Stougaard, J., Schweizer, D., and Bachmair, A. (2002). Chromosomal map of the model legume Lotus japonicus. Genetics *161*, 1661-1672.

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic acids research *44*, W160-165.

Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature communications *11*.

Reimer, J.J., and Turck, F. (2010). Genome-wide mapping of protein-DNA interaction by chromatin immunoprecipitation and DNA microarray hybridization (ChIP-chip). Part A: ChIP-chip molecular methods. Methods Mol Biol *631*, 139-160.

Ribeiro, T., Buddenhagen, C.E., Thomas, W.W., Souza, G., and Pedrosa-Harand, A. (2018). Are holocentrics doomed to change? Limited chromosome number variation in Rhynchospora Vahl (Cyperaceae). Protoplasma *255*, 263-272.

Ribeiro, T., Marques, A., Novak, P., Schubert, V., Vanzela, A.L., Macas, J., Houben, A., and Pedrosa-Harand, A. (2017). Centromeric and non-centromeric satellite DNA organisation differs in holocentric Rhynchospora species. Chromosoma *126*, 325-335.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277.

Schotanus, K., Yadav, V., and Heitman, J. (2021). Epigenetic dynamics of centromeres and neocentromeres in Cryptococcus deuterogattii. PLoS genetics *17*, e1009743.

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

Senaratne, A.P., Muller, H., Fryer, K.A., Kawamoto, M., Katsuma, S., and Drinnenberg, I.A. (2021). Formation of the CenH3-Deficient Holocentromere in Lepidoptera Avoids Active Chromatin. Curr Biol *31*, 173-+.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol *1962*, 227-245.

Šimková, H., Číhalíková, J., Vrána, J., Lysák, M.A., and Doležel, J. (2003). Preparation of HMW DNA from plant nuclei and chromosomes isolated from root tips. Biologia Plantarum *46*, 369-373.

Smith, S.Y., Collinson, M.E., Rudall, P.J., and Simpson, D.A. (2010). The Cretaceous and Paleogene fossil record of Poales: Review and current research. In Diversity, phylogeny, and evolution in monocotyledons, O. Seberg, G. Petersen, A. Barfod, and J.I. Davis, eds. (Aarhus, Denmark: Aarhus University Press), pp. 333–356.

Soderlund, C., Bomhoff, M., and Nelson, W.M. (2011). SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic acids research *39*, e68.

Soderlund, C., Nelson, W., Shoemaker, A., and Paterson, A. (2006). SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome research *16*, 1159-1168.

Steiner, F.A., and Henikoff, S. (2014). Holocentromeres are dispersed point centromeres localized at transcription factor hotspots. Elife *3*, e02025.

Stovner, E.B., and Saetrom, P. (2019). epic2 efficiently finds diffuse domains in ChIP-seq data. Bioinformatics *35*, 4392-4393.

Sun, H.Q., Ding, J., Piednoel, M., and Schneeberger, K. (2018). findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. Bioinformatics *34*, 550-557.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science *320*, 486-488.
Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome research *18*, 1979-1990.

Thomas, J., and Pritham, E.J. (2015). Helitrons, the Eukaryotic Rolling-circle Transposable Elements. Microbiol Spectr *3*.

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

Vu, G.T.H., Cao, H.X., Fauser, F., Reiss, B., Puchta, H., and Schubert, I. (2017). Endogenous sequence patterns predispose the repair modes of CRISPR/Cas9-induced DNA double-stranded breaks in Arabidopsis thaliana. Plant Journal *92*, 57-67.

Wong, C.Y.Y., Lee, B.C.H., and Yuen, K.W.Y. (2020). Epigenetic regulation of centromere function. Cellular and molecular life sciences : CMLS *77*, 2899-2917.

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

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Supplementary information

All around centromeres: repeat-based holocentromeres influence genome architecture and karyotype evolution

Paulo G. Hofstatter, Gokilavani Thangavel, Thomas Lux, Pavel Neumann, Tihana Vondrak, Petr Novak, Meng Zhang, Lucas Costa, Marco Castellani, Alison Scott, Helena Toegelová, Jörg Fuchs, Yennifer Mata-Sucre, Yhanndra Dias, André L. L. Vanzela, Bruno Hüttel, Cicero C. S. Almeida, Hana Šimková, Gustavo Souza, Andrea Pedrosa-Harand, Jiri Macas, Klaus F. X. Mayer, Andreas Houben & André Marques

Supplemental Figures

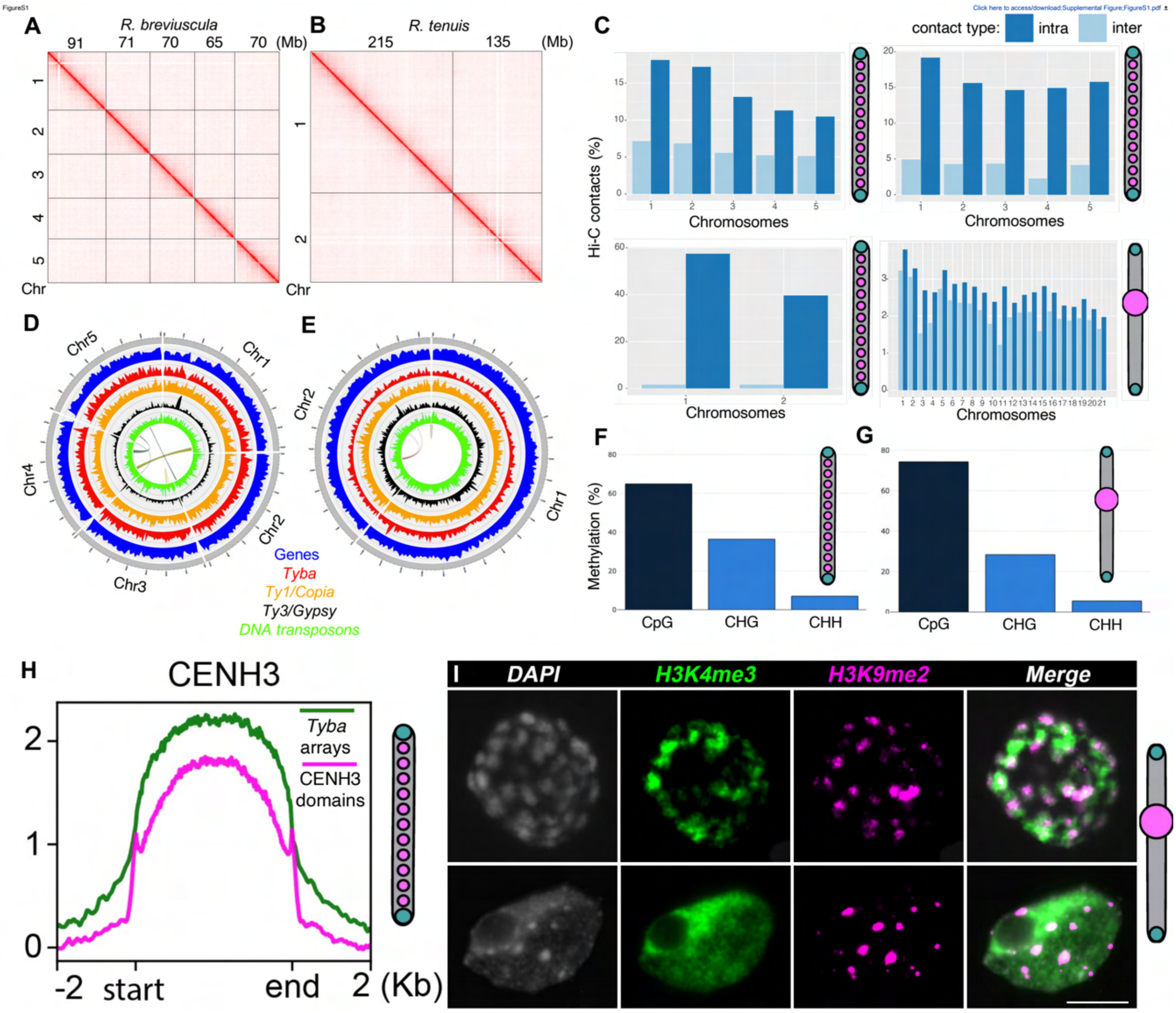


Figure S1. Related to Figures 2 and 4. Characterization of the *Rhynchospora* and *J. effusus* genomes.

(**A–B**) Contact maps for the five assembled pseudochromosomes of *R. breviuscula* (**A**) and the two assembled pseudochromosomes of *R. tenuis* (**B**). The intensity of pixels represents the normalized count of Hi-C links between 500-kb windows on a Log scale. (**C**) Hi-C contact counts (bin size = 1 Mb, normalization = VC) of intra- (*cis*) and interchromosomal (*trans*) chromatin contacts in the four species showing a significantly higher ratio (p < 4.04e–05) in holocentric compared to monocentric species, which implies relatively enriched *trans* interactions in the latter species. (**D–E**) Distribution of the main classes of sequence types in *R. breviuscula* (**D**) and *R. tenuis* (**E**) with a 1-Mb sliding window. Note the high peaks of LTR *Ty3/Gypsy* density at most subtelomeric regions in *R. breviuscula* chromosomes. Self-synteny of *R. breviuscula* (**D**) and *R. tenuis* (**E**) genomes is shown in the inner circle. (**F–G**) Summary of genome-wide DNA methylation contexts in *R. pubera* (**F**) and *J. effusus* (**G**). (**H**) Metaplot showing the enrichment of CENH3 on *Tyba* repeat arrays (green) and CENH3 domains (magenta) in *R. breviuscula*. (**I**) Immunostaining of metaphase chromosomes and an interphase nucleus of *J. effusus* for H3K4me3 and H3K9me2. Scale bar = 5 µm.

FigureS2

Click here to access/download;Supplemental Figure;FigureS2.pdf

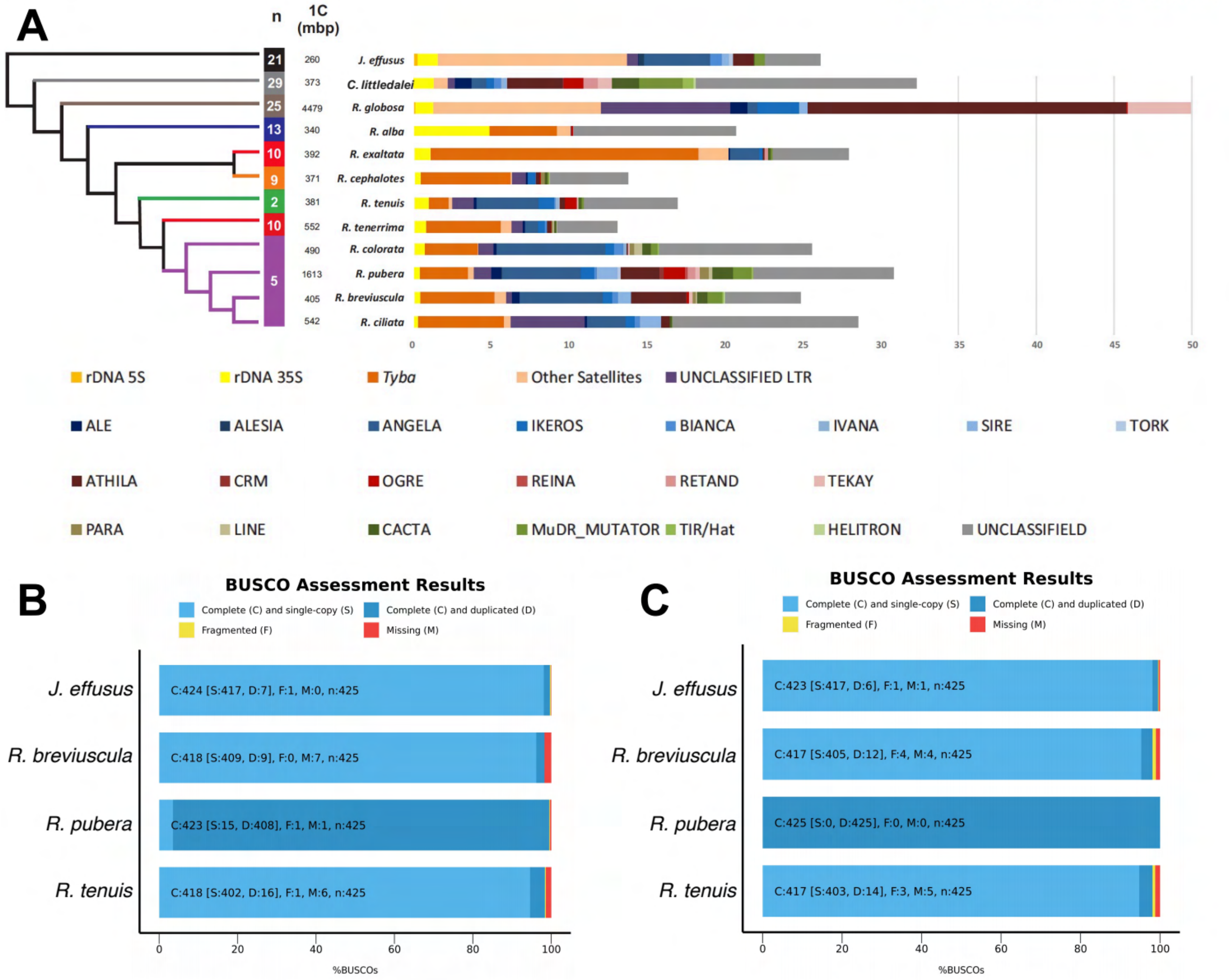


Figure S2. Related to Figure 5. Composition and evolution of sedges and rush genomes.

(**A**) Schematic phylogenetic tree and repeat composition of beak-sedge genomes and comparison with *C. littledalei* and *J. effusus*. (**B–C**) BUSCO assessment for completeness of genic space with the viridiplantae_odb10 dataset, using the entire genome assembly (**B**) or the longest transcript (**C**).

FigureS3

Click here to access/download;Supplemental Figure;FigureS3.pdf

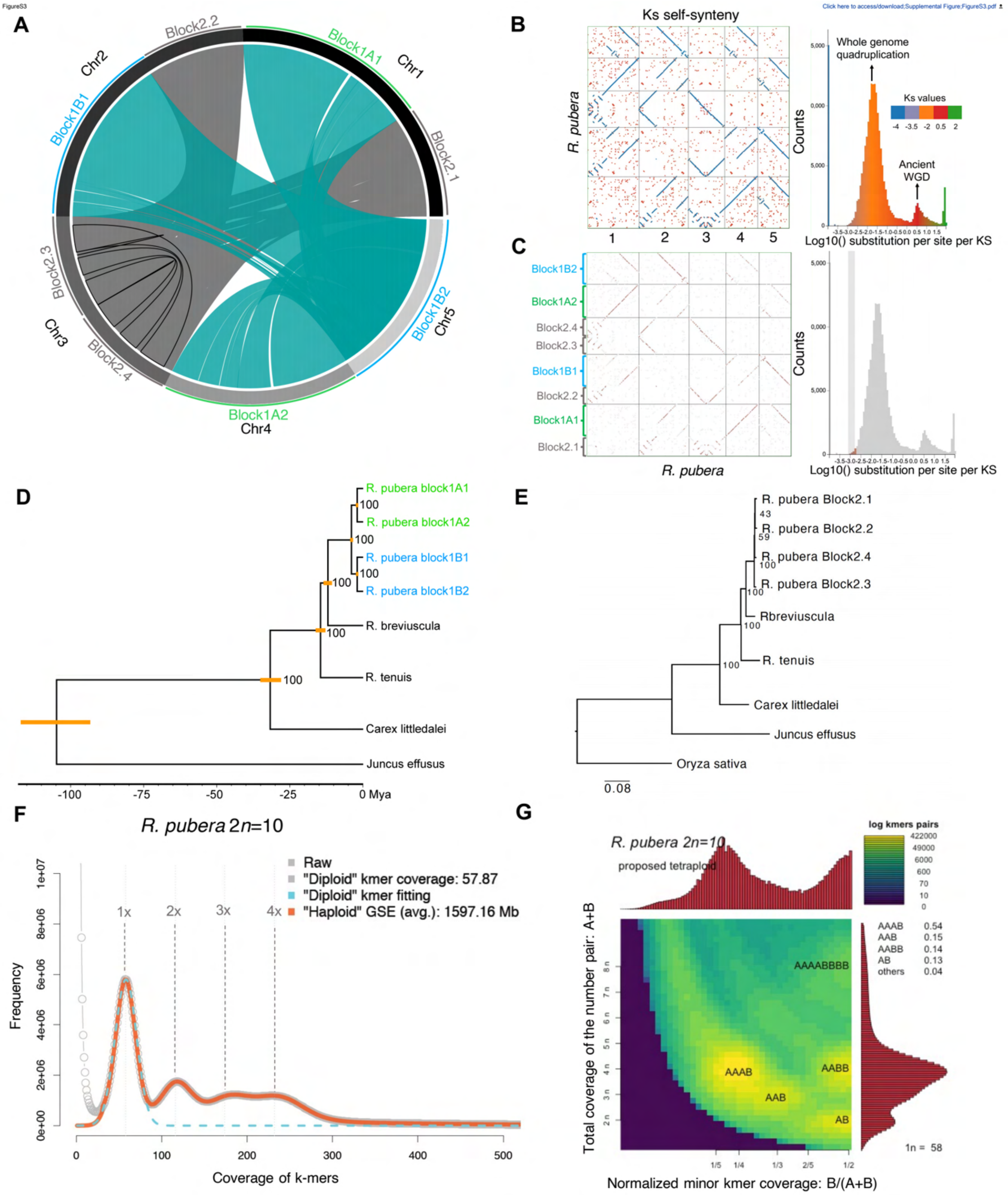


Figure S3. Related to Figure 5. Identification, characterization, and dating of WGDs in in *R. pubera*.

(**A**) SynMap self-synteny plot of *R. pubera*. Block structure is indicated by outer arcs. (**B**) SynMap self-synteny dot plot colored based on Ks values. Ks values on a Log scale are shown to the right of the dot plot. Note the large peak that correlates with the large duplication events in *R. pubera* and a second small peak most likely representing an ancient WGD. (**C**) Same plot as (**B**), but selecting only the sequences with the lowest number of synonymous substitutions, allowing the identification of intragenomic syntenic block relationships (Block1A and Block1B). We were unable to detect any relationships for Block2. The small colored block within the vertical gray bar represents the sequences with the lowest number of synonymous substitutions usSed in the dotplot to the left. Ks values are indicated by the color scale in B. (**D**) Based on the assessment of the relationships among the syntenic blocks of *R. pubera*, we selected 1,204 BUSCO genes (Poales dataset) uniquely present in each block and also shared with *R. breviuscula*, *R. tenuis*, *C. littledalei*, and *J. effusus* to build a phylogenetic tree from a concatenated alignment, which was further used for dating the duplication events in *R. pubera*. We confirmed the Block1A and Block1B relationships with 100% bootstrap support and also determined that a first WGD occurred around 3.8 Mya, followed by a second event around 2.1 Mya. Note that the second WGD closely overlaps in both Block1A and Block1B branches. Yellow bars indicate the dating time interval. (**E**) Phylogenetic analysis of Block2 genes did not resolve the relationships for this particular block and was not used for dating. (**F**) K-mer based estimation of genome size and heterozygosity and (**G**) Smudgeplot analysis of k-mer-based ploidy inference for *R. pubera* using 21-mers. GSE = genome size estimation. Smudgeplot infers ploidy directly from the k-mers present in sequencing reads by analyzing heterozygous k-mer pairs.

FigureS4

Click here to access/download;Supplemental Figure;FigureS4.pdf

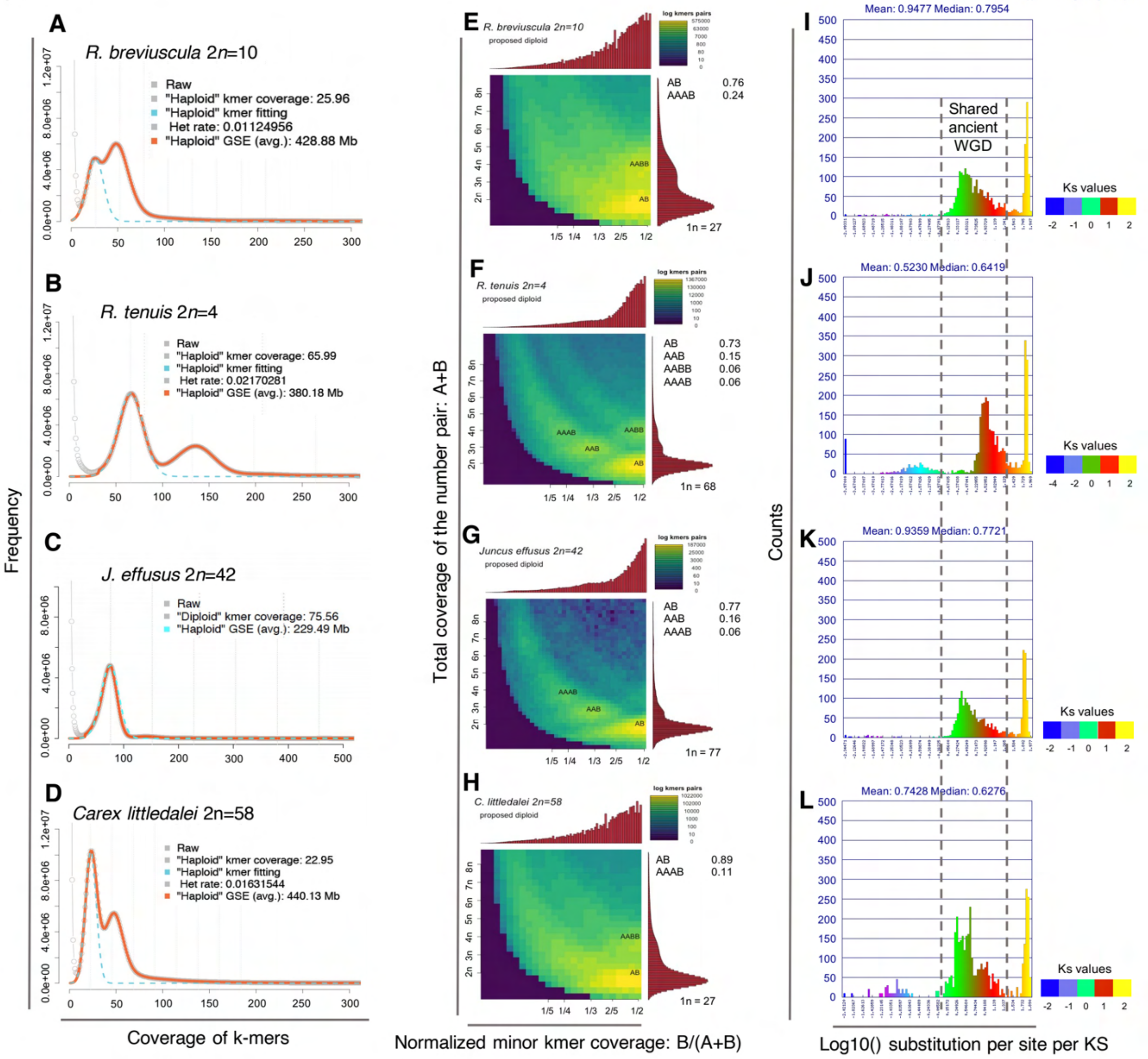


Figure S4. Related to Figure 5. K-mer based genome size estimation and ploidy inference and WGD identification in sedges and rushes.

(**A–D**) 21-mer based estimation of genome size and heterozygosity. GSE = genome size estimation. (**E–H**) Ploidy and genome structure inference based on 21-mer Smudgeplot analysis. (**I–L**) Ks values of coding sequences for each genome; a shared ancient WGD peak was observed for all species.

FigureS5

Click here to access/download;Supplemental Figure;FigureS5.pdf

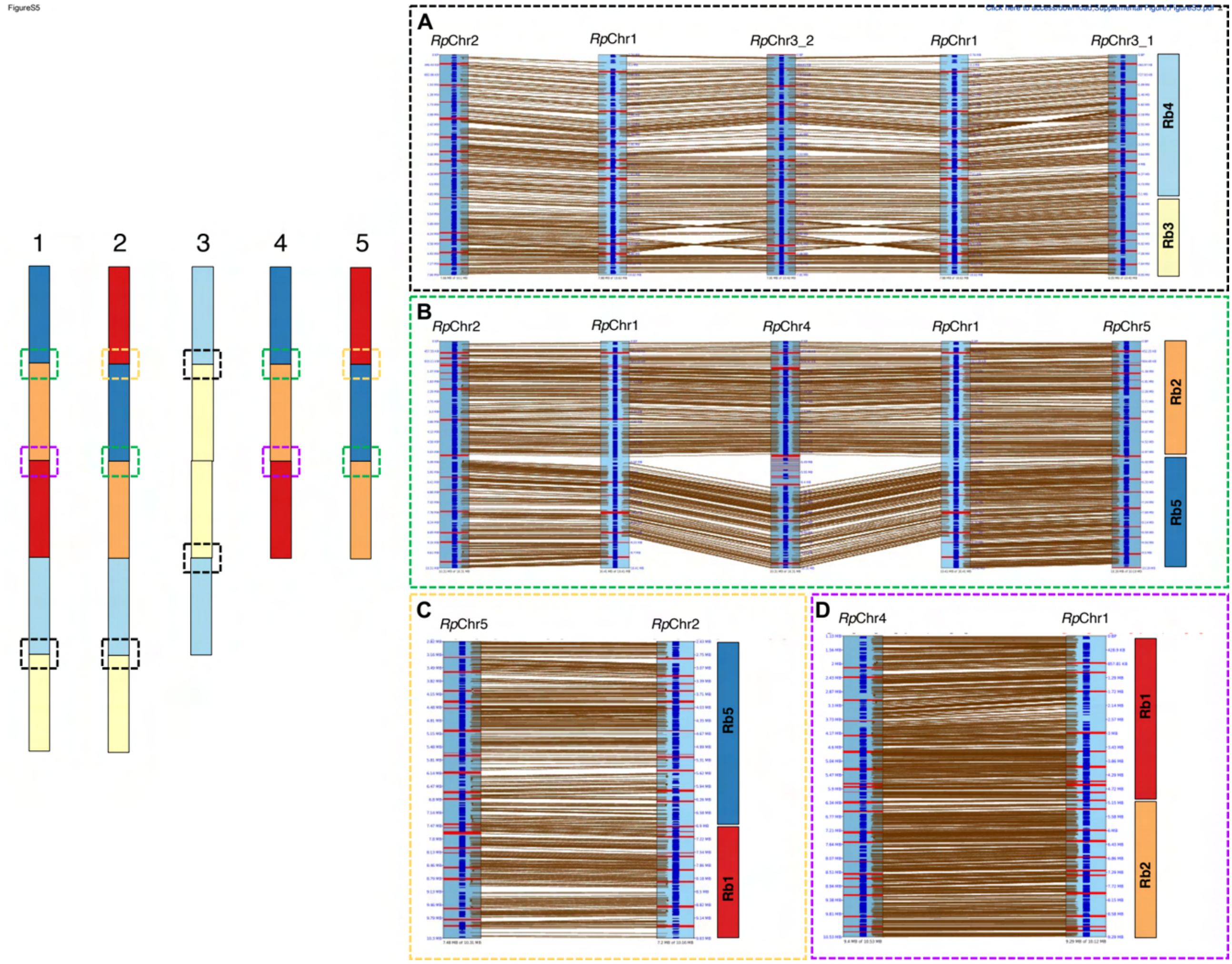


Figure S5. Related to Figure 5. Comparative alignment of the duplicated end-to-end fusion (EEF) transition regions in the *R. pubera* genome.

(Left) Ideogram model of *R. pubera* chromosomes, with the dashed boxes indicating the extracted and compared regions on the right. (**A**) Alignment of the EEF of *Rb*3 and *Rb*4 found once on *Rp*Chr1 and *Rp*Chr2 and twice on *Rp*Chr3, showing the same fusion signature. (**B**) Alignment of the EEF of *Rb*2 and *Rb*5, found on *Rp*Chr1, *Rp*Chr2, *Rp*Chr4, and *Rp*Chr5, also showing the same fusion signature. (**C**) Alignment of the EEF of *Rb*1 and *Rb*5, found on *Rp*Chr2 and *Rp*Chr5 with the same fusion signature. (**D**) Alignment of the EEF of *Rb*1 and *Rb*2, found on *Rp*Chr1 and *Rp*Chr4 with the same fusion signature. Colored boxes assign the synteny to *R. breviuscula* chromosomes. Red stripes on the synteny alignments depict *Tyba* repeats, while genes are annotated in dark blue.

Click here to access/download;Supplemental Figure;FigureS6.pdf

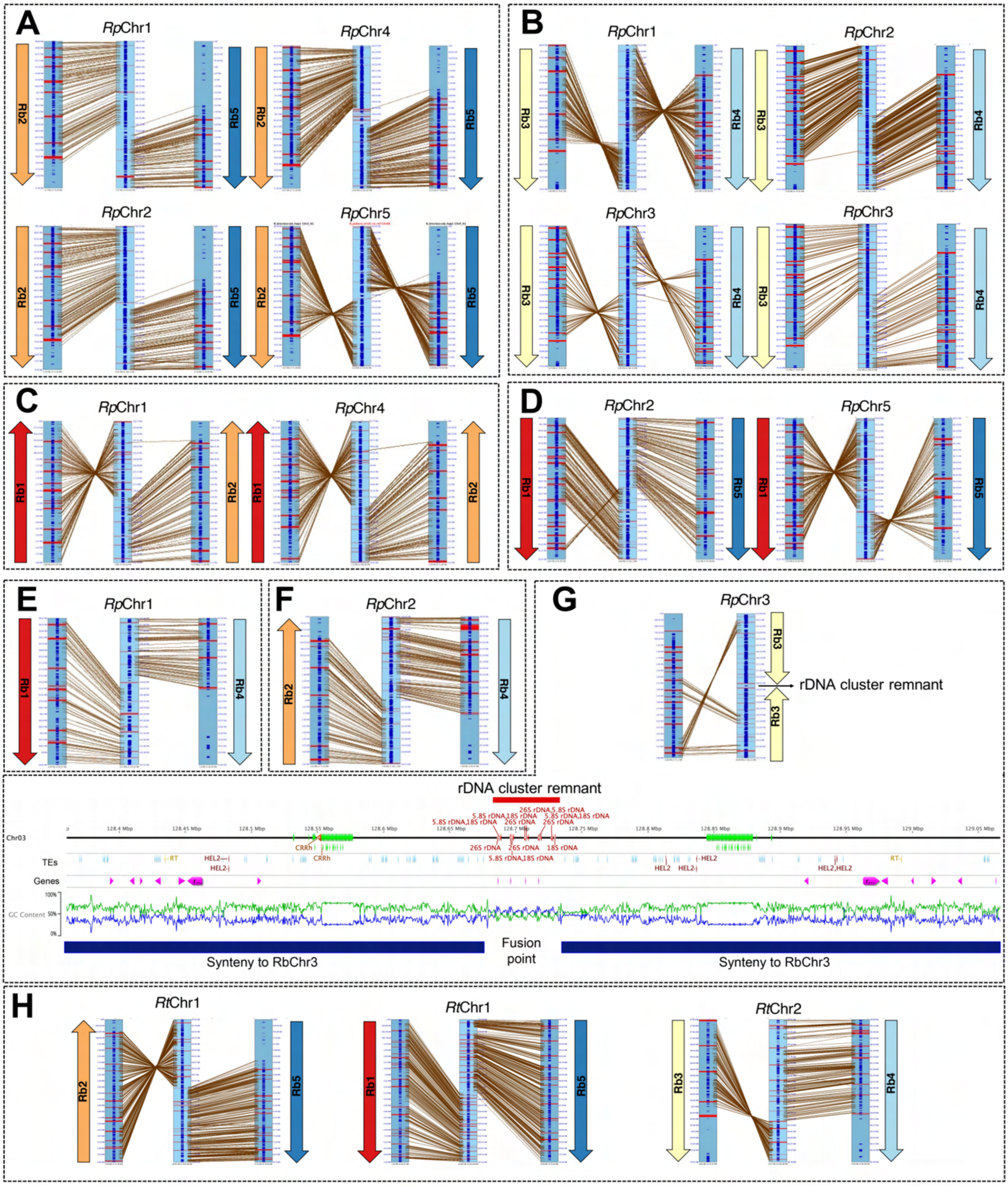


Figure S6. Related to Figure 5. Identification of the sequences underlying the transitions between the syntenic regions to *R. breviuscula* chromosomes in the end-to-end fusions (EEFs) found in the *R. pubera* and *R. tenuis* genomes.

(**A**) EEF of *Rb*2 and *Rb*5 found on *Rp*Chr1, *Rp*Chr2, *Rp*Chr4, and *Rp*Chr5. Similar fusion signatures are shared among the four chromosomes. In three of them, a *Tyba* repeat is found between them. (**B**) EEF of *Rb*3 and *Rb*4 found on *Rp*Chr1 and *Rp*Chr2 and twice on *Rp*Chr3 with the same fusion signature. A *Tyba* repeat array is found between the transitions in all cases. (**C**) EEF of *Rb*1 and *Rb*2 found on *Rp*Chr1 and *Rp*Chr4 with the same fusion signature, without a *Tyba* repeat in between. (**D**) EEF of *Rb*1 and *Rb*5 found on *Rp*Chr2 and *Rp*Chr5 with the same fusion signature, with a *Tyba* repeat in between. (**E**) EEF of *Rb*1 and *Rb*4 found only on *Rp*Chr1 with a *Tyba* repeat array in between. (**F**) EEF of *Rb*2 and *Rb*4 found only on *Rp*Chr2 with no *Tyba* repeat in between. (**G**) EEF of *Rb*3 and *Rb*3 found only on *Rp*Chr3 and with a remnant of a rDNA cluster in the transition region (with detailed annotation shown to the right). (**H**) Characterization of the three EEFs responsible for the chromosome reduction in *R. tenuis*. On *Rt*Chr1 we found an EEF involving *Rb*2 and *Rb*5, and a second event involving *Rb*5 and *Rb*1, while on *Rt*Chr2, we found a single EEF involving *Rb*3 and *Rb*4. Colored arrows indicate the *R. breviuscula* chromosomes and point to the telomeric region involved in the fusion event. Remarkably, although similar ancestral chromosome associations are found in *R. pubera* and *R. tenuis*, the chromosomal ends involved in the fusions are different. Red stripes on the synteny alignments depict *Tyba* repeats, while genes are annotated in dark blue.

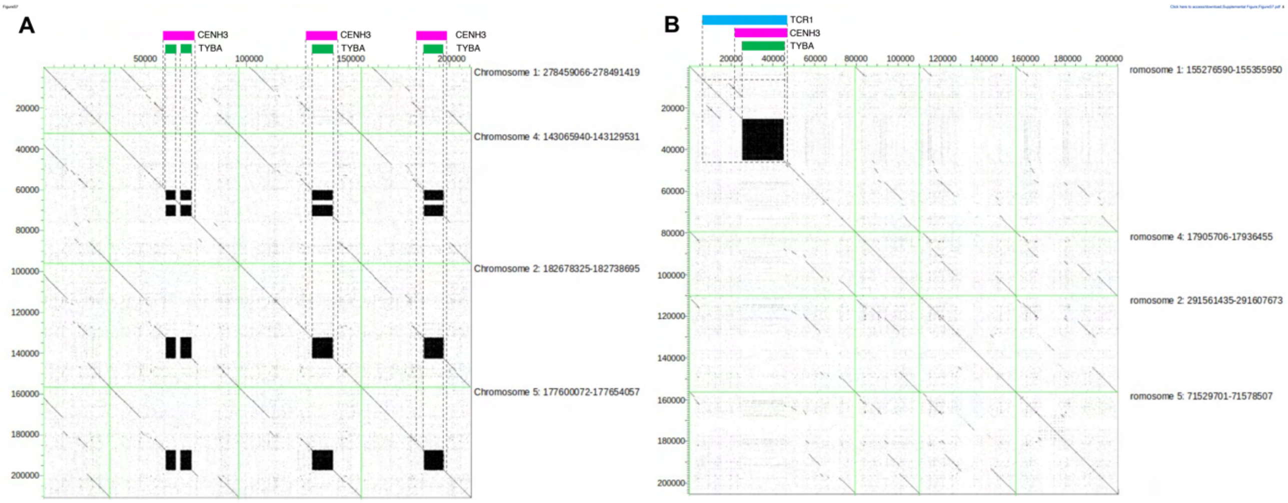


Figure S7. Related to Figure 6. Characterization of emergence and loss of CENH3-binding regions in *R. pubera*.

(**A**) Example of CENH3-binding region and *Tyba* array lost in one of four paralogous regions, while the other three copies retained the *Tyba* array and CENH3 binding. The conserved locus is indicated by the dashed box, along the x-axis of the dot plot, with rectangles marking the area associated with CENH3 (magenta) and the *Tyba* array (green). The genome positions of the extracted regions are given to the right. (**B**) Example of CENH3-binding region and *Tyba* array gain in one of four paralogous regions due to a transposition of *Tyba*-containing *TCR1* in *Rp*Chr1, while the other three copies lack the *Tyba* array. The gained locus is indicated by the dashed box, along the x-axis of the dot plot, with rectangles marking the *TCR1* element (blue), the area associated with CENH3 (magenta), and the *Tyba* array (green). The genome positions of the extracted regions are given to the right.

Supplemental Tables

TableS1. Related to Figure1. Summary of structural gene annotation per chromosome and total (only high-confidence genes).

	J. effusus	Chromome length	Genes /1mb	*R. tenuis*	Chromome length	Genes /1mb	*R. breviuscula*	Chromome length	Genes /1mb	*R. pubera*	Chromome length	Genes /1mb
Chr1	1,672	16,424,282	101.8	13,086	215,061,937	60.85	4,557	91,873,953	49.6	19,709	375,421,704	52.5
Chr2	1,462	14,398,045	101.54	7,984	135,357,977	58.98	4,176	71,359,411	58.52	18,971	361,203,633	52.52
Chr3	626	14,353,144	43.61				4,010	70,414,202	56.95	14,040	265,146,654	52.95
Chr4	629	13,791,583	45.61				3,447	65,946,630	52.27	12,078	240,819,442	50.15
Chr5	1,264	13,318,477	94.91				3,788	70,463,030	53.76	11,952	230,081,685	51.95
Chr6	1,082	12,805,256	84.5									
Chr7	834	11,678,469	71.41									
Chr8	955	11,338,180	84.23									
Chr9	760	11,435,724	66.46									
Chr10	870	11,091,224	78.44									
Chr11	872	11,033,772	79.03									
Chr12	698	10,907,635	63.99									
Chr13	783	10,565,805	74.11									
Chr14	655	10,411,571	62.91									
Chr15	605	10,062,658	60.12									
Chr16	646	9,765,450	66.15									
Chr17	609	9,619,279	63.31									
Chr18	620	9,689,104	63.99									
Chr19	611	9,280,436	65.84									
Chr20	720	9,245,364	77.88									
Chr21	807	8,729,336	92.45									
Chr00	1,162	18,332,129	63.39	2,145	47,328,916	45.32	4,376	50,710,879	86.29	14,613	222,205,218	65.76
Total number of genes	18,942	258,276,923	73.34	23,215	397,748,830	58.37	24,354	420,768,105	57.88	91,363	1,694,878,336	53.91
Number of monoexonic genes	2,576			4,300			4,544			18,918		
Number of transcripts	21,245			31,583			38,330			122,577		
Transcripts per gene[1]	1.12			1.36			1.57			1.34		
cDNA lengths (bp)[1]	1,659			1,804			1,869			1,832		
CDS lengths (bp)[1]	1,329			1,338			1,325			1,322		
Exons per transcript[1]	6.12			6.06			5.98			6.01		
Exon lengths (bp)[1]	271			298			313			305		

Table S2. Related to Figure 3 and Figure 4. Repeat characterization and respective association with CENH3 in *R. pubera*.

Repeat Type	Total size [bp]	Number of regions	Total size of elements overlaping with CENH3	Number of regions in any CENH3	Number of CENH3 regions containg repeat	Observed/Expected
All/repeat/satellite/TYBA	44,550,003	10,717	41,660,767	8,009	2,652	**23.04**
mobile_element/Class_II/Subclass_1/TIR/TCR2	4,678	38	1,620	10	8	**8.53**
mobile_element/Class_II/Subclass_2/Helitron/TCR1	225,651	1,503	55,393	137	109	**6.05**
mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/CRM/CRRh	580,406	1,830	103,346	208	157	**4.39**
mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/Tekay	7,279,641	15,909	332,179	654	191	1.12
mobile_element/Class_I/LTR	57,960,329	132,771	2,177,905	4,823	1,043	0.93
mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/CRM	2,346,954	3,693	70,478	103	33	0.74
mobile_element/Class_II/Subclass_2/Helitron	5,024,836	26,331	131,643	611	236	0.65
mobile_element/Class_I/LTR/Ty3_gypsy/non-chromovirus/OTA/Athila	48,902,271	66,490	1,200,879	1,731	438	0.61
mobile_element/Class_I/LTR/Ty1_copia/SIRE	22,984,586	31,735	540,490	827	208	0.58
mobile_element/Class_I/LTR/Ty1_copia/Ikeros	16,078,076	23,523	309,796	493	171	0.47
mobile_element/Class_I/LTR/Ty1_copia/Angela	79,399,181	124,150	1,483,218	2,902	671	0.46
mobile_element/Class_I/LTR/Ty1_copia/Ivana	4,033,882	5,665	69,822	78	36	0.43
Unknown repeat	57,088,116	241,989	956,523	3,667	1,160	0.41
mobile_element/Class_I/LINE	7,476,329	24,368	109,946	237	163	0.36
mobile_element/Class_I/LTR/Ty1_copia/Alesia	2,032,760	7,201	28,848	115	65	0.35
mobile_element/Class_I/LTR/Ty1_copia/Bianca	2,623,561	2,812	34,192	27	14	0.32
rDNA/5S_rDNA	445,568	844	5,715	3	3	0.32
mobile_element/Class_I/LTR/Ty3_gypsy/non-chromovirus/OTA/Tat/Ogre	60,730,039	105,591	752,647	1,331	252	0.31
mobile_element	94,639,533	463,223	1,099,339	5,484	1,507	0.29
mobile_element/Class_II/Subclass_1/TIR	88,511,105	436,511	959,789	5,041	1,692	0.27
mobile_element/Class_I/LTR/Ty3_gypsy/chromovirus/Reina	5,806,502	21,062	62,885	265	129	0.27
mobile_element/Class_II/Subclass_1/TIR/hAT	12,140,614	40,653	129,540	501	209	0.26
mobile_element/Class_I/LTR/Ty1_copia/Ale	19,209,093	34,307	178,590	295	95	0.23
mobile_element/Class_I/pararetrovirus	10,054,165	7,772	83,323	92	43	0.20
mobile_element/Class_I/LTR/Ty1_copia/Tork	4,987,506	5,329	38,075	29	16	0.19
mobile_element/Class_II/Subclass_1/TIR/MuDR_Mutator	34,872,900	84,633	263,668	748	299	0.19
mobile_element/Class_I/LTR/Ty3_gypsy/non-chromovirus/OTA/Tat/Retand	12,918,720	16,297	85,234	118	33	0.16
satellite	319,277	343	2,049	7	1	0.16

Table S4. Related to Figure 3. Summary of orthofinder results for all high-confidence genes.

	C. littledalei	*J. effusus*	*R. breviuscula*	*R. pubera*	*R. tenuis*
Number of genes	22,441	18,942	19,978	80,676	21,070
Number of genes in orthogroups	19,010	9,602	19,504	79,451	20,178
Number of unassigned genes	3,431	9,340	474	1,225	892
Percentage of genes in orthogroups	84.7	50.7	97.6	98.5	95.8
Percentage of unassigned genes	15.3	49.3	2.4	1.5	4.2
Number of orthogroups containing species	14,067	5,882	17,559	19,661	17,564
Percentage of orthogroups containing species	62.4	26.1	77.9	87.3	77.9
Number of species-specific orthogroups	962	1172	78	993	204
Number of genes in species-specific orthogroups	3,177	3,592	307	4,377	785
Percentage of genes in species-specific orthogroups	14.2	19	1.5	5.4	3.7

	TOTAL
Number of species	5
Number of genes	173,794
Number of genes in orthogroups	157,762
Number of unassigned genes	16,032
Percentage of genes in orthogroups	90.8
Percentage of unassigned genes	9.2
Number of orthogroups	22,503
Number of species-specific orthogroups	3,361
Number of genes in species-specific orthogroups	18,068
Percentage of genes in species-specific orthogroups	10.4
Mean orthogroup size	7
Median orthogroup size	7
G50 (assigned genes)	7
G50 (all genes)	7
O50 (assigned genes)	7,285
O50 (all genes)	8,430
Number of orthogroups with all species present	3,908
Number of single-copy orthogroups	6

Table S5. Related to Figure 4. Number of CENH3-binding domains paralogous groups

	Number of CENH3-binding domains paralogous groups
Total number of groups analyzed	660
No change – CENH3-binding domain in all four paralogous regions	403
Single change - gain of CENH3 locus in one paralog	169
- gain of CENH3 locus associated with gain of Tyba	22
- gain of CENH3 locus associated with expansion of Tyba	1
- gain of CENH3 locus associated TE insertion	129
- no change in sequence	17
Single change - loss of CENH3 locus in one paralog	50
- loss of CENH3 locus associated with loss of Tyba	19
- loss of CENH3 locus associated with shrinking of Tyba	25
- no change in sequence	6
CENH3 locus present in two of four paralogs	38

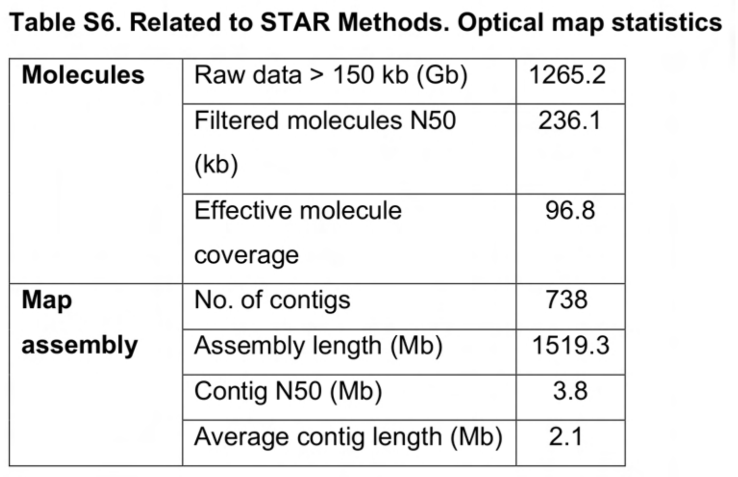


Table S6. Related to STAR Methods. Optical map statistics

Molecules	Raw data > 150 kb (Gb)	1265.2
	Filtered molecules N50 (kb)	236.1
	Effective molecule coverage	96.8
Map assembly	No. of contigs	738
	Assembly length (Mb)	1519.3
	Contig N50 (Mb)	3.8
	Average contig length (Mb)	2.1

Table S7: Sequence assembly and hybrid scaffold statistics

	Original sequence assembly (*Hicanu.contigs. assembly.cleaned.fasta*)	**Sequence assembly after hybrid scaffolding**	**Hybrid scaffolds only**
No of contigs/scaffolds	8656	8566	59
Total length (Mb)	1692.4	1694.9	1475.7
Contig/scaffold N50 (Mb)	9.7	49.1	50.4

Summary

The central objective of this work was to use long sequence reads to study the long-range organization of satellite DNA. Therefore we aimed to develop a bioinformatics pipeline for the satDNA annotation and analysis in ultra-long unassembled nanopore reads. This goal has been achieved, and two of the three chapters of this thesis demonstrate the accuracy and utility of this approach in revealing patterns of organization of repeats that can then be used to infer the evolutionary mechanisms that led to their formation. The main advantages of this approach is that neither a reference genome nor high coverage is needed to analyze the genome average organization of any number of repeats simultaneously. Nevertheless, some features of the analysis can be considered limitations. Namely read length, which limits the upper array size that can be determined and the reference database accuracy that determines the accuracy of the read annotation.

Using this approach the genomes of meta-polycentric *Lathyrus sativus* and holocentric *Cuscuta europaea* were analyzed. In *L. sativus* a surprising bimodal array length was detected, with short arrays found at the 3' ends of Ogre retrotransposons and long arrays located at pericentromeres. These results suggest that Ogres affect the *L. sativus* genome in two ways, accumulating new tandem repeats within their sequences and mobilizing them throughout the genome. It is not yet known how new tandem repeats originate in Ogres, but based on their widespread occurrence in plants, it is possible that this mechanism is more widespread than reported. The presence of long arrays in the pericentromere would also suggest that it is an environment favorable for array expansion and/or retention, possibly because of the low recombination rates characteristic of heterochromatin.

The analysis of *C. europaea* focused on resolving the previously mentioned complex structure of its heterochromatic bands. Annotation and analysis of Nanopore reads confirmed that the most abundant member of this domain is the CUS-TR24 satellite, whose arrays are interrupted by simple sequence repeats (SSRs) and L1- CS LINE retrotransposons. Because there are two periodicities in the size of the CUS-TR24 arrays, the size of the SSR arrays varies and the nucleotide sequences of L1-CS are diverse, this pattern is most likely the result of multiple simultaneous processes. These are processes in which the CUS-TR24 monomer acts as the basis for pattern formation by continuously amplifying and providing hotspots for SSR emergence while being targeted by L1-CS retrotransposons.

Lastly, a chromosome-scale whole genome assembly of holocentric *Rhynchospora pubera* and it being an auto-octoploid, provided a unique opportunity to identify paralogous centromeric loci in the genome assembly and to observe potential polymorphisms between them. While most centromeric loci did not exhibit any polymorphisms, gain and loss of centromeric loci in a subset of cases was observed. The gain of centromeric domains was connected to the gain of previously identified Tyba centromeric satellites that often happened by the mobilization of Tyba arrays by a Helitron transposon. The loss of centromeric loci corresponded to the loss of Tyba arrays either through deletion or gradual mutation. Unfortunately, the results gathered were insufficient to decide whether Tyba arrays spread throughout the chromosomes first through Helitron activity, which triggered centromeric activity or centromeric activity expanded along chromosomes initially and the Tyba arrays followed.

Usually, transposable elements and satellites are researched and considered independently because of their different structures and amplification mechanisms. However, in this thesis, it is clear throughout all three chapters that transposable elements and satellites often evolve together. In this context, transposable elements are involved in various stages of satellite evolution, from the origin of tandem repeats to acting as vehicles of satellite mobilization, while satellite arrays potentially act as targeting sequences for transposable element insertion. Only three plant species are presented here, but owing to the widespread nature of both types of repeats, this relationship can potentially be observed and further investigated across various eukaryotic genomes.

Curriculum vitae

Tihana Vondrak
Ph.D. Student

Department of Molecular Biology and Genetics; Faculty of Sciences; University of South Bohemia in České Budějovice, Czech Republic.

Institute of Plant Molecular Biology; Biology Centre; Czech Academy of Sciences; České Budějovice, Czech Republic.

+(420) 777654076

tihana.vondrak12@gmail.com

Born: Osijek, Croatia, 29th of November 1993

Nationality: Croatian

Languages: Croatian, English

Education:

March 2018 – Present: Ph.D. candidate at the University of South Bohemia, České Budějovice, Czech Republic.

September 2015 – December 2017: Masters degree in Molecular Biology, University of Zagreb, Zagreb (Croatia)

September 2012 – July 2015: Bachelors degree in Biology, University of Osijek, Osijek (Croatia)

Publications:

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

Vondrak T, Oliveira L, Novák P, Koblížková A, Neumann P, Macas J. 2021. Complex sequence organization of heterochromatin in the holocentric plant *Cuscuta europaea* elucidated by the computational analysis of nanopore reads. *Computational and Structural Biotechnology Journal* 19: 2179-2189. DOI:0.1016/j.csbj.2021.04.011

Hofstatter GP, Thangavel G, Lux T, Neumann P, **Vondrak T**, Novak P, Zhang M, Costa L, Castellani M, Scott A, Toegelová H, Fuchs J, Mata-Sucre Y, Dias Y, Vanzela LLA, Hüttel B, Almeida SCC, Šimková H, Souza G, Pedrosa-Harand A, Macas J, Mayer XFK, Houben A, Marques A. 2022. All around centromeres: repeat-based holocentromeres influence genome architecture and karyotype evolution, *Cell*

Conferences:

2019: Talk at the ELIXIR CZ Annual Conference in Kurdějov, Czech Republic

2021: Poster presentation at the Cytogenetics 2021 Meeting, Goerlitz, Germany