School of Doctoral Studies in Biological Sciences

University of South Bohemia in České Budějovice

Faculty of Science

Evolutionary dynamics of satellite DNA in plant genomes

Ph.D. Thesis

MSc. Laura Ávila Robledillo

Supervisor:

RNDr. Jiří Macas, Ph.D.

Institute of Plant Molecular Biology

Biology Centre

Czech Academy of Sciences

České Budějovice 2021

This thesis should be cited as:

Ávila Robledillo, L., 2021, Evolutionary dynamics of satellite DNA in plant genomes. University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic, 109pp.

ANNOTATION

Satellite DNA (satDNA) belongs to the highly repetitive fraction of eukaryotic genomes. It is best characterized by the formation of long arrays of almost identical sequences that are tandemly repeated. These repeats are widely distributed in plant species where they can make up a substantial proportion of their genomes. Despite the long history of satDNA research, the classic methodology did not allow for its comprehensive characterization. Consequently, the fragmentary information gathered during the last 60 years does not answer the many questions surrounding the evolution of these elements. The development of new techniques in sequencing, together with the availability of new bioinformatics tools for analyzing different genome fractions, has presented an opportunity to advance studies of tandem repeats.

This thesis describes the landscape characterization of satDNA in the genome of *Fabeae* species by exploring the diversity of satDNA within a genome, the association of these elements with functional centromeres, as well as their genome-wide organization. We employed new computational pipelines specifically designed for the analysis of tandem repeats from next generation sequencing data, and combined their results with molecular and cytogenetic methods to achieve comprehensive characterization of the satellite repeats.

Declaration [in Czech]

Prohlašuji, že svoji disertační práci jsem vypracovala samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své disertační práce, a to v úpravě vzniklé vypuštěním vyznačených částí archivovaných Přírodovědeckou fakultou elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

České Budějovice, 05.03.2021

…………………………………

Laura Ávila Robledillo

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 the 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 grants from the Czech Science Foundation (Grant No. 17-09750S; BP501/12/G090), the Czech Academy of Sciences (Grant No. RVO:60077344), the ELIXIR CZ Research Infrastructure Project (Czech Ministry of Education, Youth and Sports; Grant No. LM2015047), 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

First and foremost, I would like to thank my supervisor Jiří Macas for the opportunity to be a part of the team, but most importantly, for his patience and support over those years. Jiří, I really appreciate the knowledge and time that you have shared with me, and I hope one day I get to be as good a scientist as you are.

Sincere thanks to Pavel Neumann and Peter Novák for their enormous help and useful advice. My special thanks to Iva Vrbová and Andrea Koblížková for all their care and dedication; I have learned a lot from you. To Vlasta Látalová and Jana Tetourová for their excellent technical assistance, Děkuji! Of course, I would like to thank my colleagues Tihana Vondrak and Ludmila Oliveira for the friendship we built over the years, all the coffee breaks, and the dramas that you helped me to overcome… I can not thank you enough.

On a personal note, I would like to thank all of you, all the people I met on this journey. You made those years worth remembering and frame. But specially I would like to thank Honza, Javi, Guillem and Thomas for their contribution to my life in Budejovice. Thank you for hosting me so many times, thank you for becoming family.

A Carmen, Marina, Ana, Nibal, Luz, Lolu, LauraG. Con vosotras he compartido mucho más que clases, practicas, pisos y tapas, gracias por estar siempre, no se como hubiese sido todo este camino sin vosotras a mi lado. A Jane Chapman, por la alegría que siempre me ha transmitido, por tratarme como una hermana, por todas las traducciones... por todo, gracias. Espero poder celebrar con todas vosotras pronto. Nos veremos en Pedro Antonio!

Y por supuesto a mi familia, gracias por apoyar todos mis sueños. Vosotros me habéis dado la fuerza para llegar hasta aquí. Muchísimas gracias de todo corazón por todo lo que habéis hecho por mi.

This thesis is based on the following papers:

I. **Ávila Robledillo, L.**, Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., ... & Macas, J. (2018). Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Scientific Reports*, 8(1), 1-11. https://doi.org/10.1038/s41598-018-24196-3

   (IF = 3.998 ; citations = 24)

   L.A.R., A.K., K.B. and I.V. conducted the experiments. P.No. and J.M. carried out the bioinformatics analysis. L.A.R., P.Ne., I.S. and J.M. analyzed the results. J.M. conceived the study and drafted the manuscript. All authors reviewed the manuscript.

II. **Ávila Robledillo, L.**, Neumann, P., Koblížková, A., Novák, P., Vrbová, I., & Macas, J. (2020). Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe *Fabeae*. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msaa090

   (IF = 11.062; citations = 2)

   L.A.R., A.K., and I.V. conducted the experiments. P.No. and J.M. carried out the bioinformatics analysis. L.A.R., P.Ne., I.S. and J.M. analyzed the results. J.M. conceived the study and drafted the manuscript. All authors reviewed the manuscript.

III. 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. *Plant Journal*, 101(2), 484. https://doi.org/10.1111/tpj.14546

   (IF = 6.141 ; citations = 12)

   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. J.M. conceived the study and drafted the manuscript. All authors reviewed and approved the final manuscript.

Table of Contents

Introduction.....1

1. Sequence composition and abundance of satDNA in plant genomes.....1
2. Origin of satDNA.....3
3. SatDNA evolution.....4
4. Centromeric satellites.....4
5. Methodological approaches for identification and characterization of satDNA in plant genomes.....6

References.....9

AIMS OF THIS WORK.....19

SCOPE OF THE THESIS.....20

Chapter I.....22

Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing.....22

Supplementary Information.....34

Chapter II.....48

Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe *Fabeae*.....48

Supplementary Information.....65

Chapter III:.....76

Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats.....76

Supplementary Information.....94

SUMMARY.....106

CURRICULUM VITAE.....108

Introduction

Satellite DNA (satDNA) is a class of repetitive genomic sequences characterized by its organization into tandemly repeated units known as monomers. Tandem repeats have been traditionally placed into three classes according to the length of the monomere unit; micro- (<10 bp), mini- (>10 bp) and satellite repeats (>100 bp). However, the most distinguishable and accepted feature differentiating them is the size of the array of the tandem repeated element (Macas et al., 2002). Satellite DNA forms long arrays of tandem repeats that can be visualized in nuclei as chromocenters or heterochromatic bands in chromosomes. The overall distribution of satDNA on chromosomes is often used in cytogenetic studies because they can provide markers that differentiate pairs of chromosomes. The location of satDNA is predominantly subtelomeric and centromeric in most of the eukaryotic genomes studied to date, although interstitial satDNA families have also been described in plant species (Garrido-Ramos, 2015; Oliveira & Torres, 2018). SatDNA is a ubiquitous component of eukaryotic genomes where one single satDNA family can range from 0.1 up to 36% of the total nuclear DNA in *Fritillaria* genus (Ambrožová et al., 2011), or from 1.5 up to 25% in *Vicia* species(Macas et al., 2000). The same variability is observed in animals, ranging from 0.5 to 50% of the genome in some species (Garrido-Ramos, 2017). The satellite profile of one species can thus differ substantially from other close relatives with respect to its abundance and sequence composition. However, the molecular mechanisms responsible for these variations are not fully understood.

Similarly, the actual roles of satDNA in eukaryotic genomes remain unresolved. For decades satDNA has been considered as “junk DNA” due to its lack of coding potential. However, some authors considered that satDNA arrays could influence nearby gene expression since tandem repeats are targets for epigenetic silencing mechanisms (Pezer et al., 2012; Ugarkovic, 2005). In addition, satDNA structure could promote the organization and packing of the full complement within a single nucleus, avoiding the formation of micro-nuclei (Jagannathan et al., 2018). Moreover, the frequent association of satDNA and centromeric loci suggests some function in assuring proper chromosome segregation during cell division by a sequence-specific interaction of the repeats and the proteins involved in kinetochore formation (Henikoff et al., 2001). While attempts to elucidate the role of satDNA are numerous in the literature, its genomic organization into long arrays of almost identical monomers condemned satDNA to remain poorly characterized. This is true even for genomes that have been extensively studied, which has made predictions difficult to test.

1. Sequence composition and abundance of satDNA in plant genomes

Although no specific sequence motif is generally conserved in satellite DNA, these repeats share some common features. The main attribute of satDNA is its genomic organization into long arrays of relatively short monomers. The monomer sizes most often described in the literature range from 135–195 bp and 315–375 bp (Macas et al., 2002). This range of lengths correlate with that of the DNA wrapping a mono- or di-nucleosome particle, thus it has been tempting for many authors to infer that the periodicity could promote the spacing arrangement of nucleosomes along with the arrays of tandemly organized DNA sequences in heterochromatic regions (Lowman & Bina, 1990; Tommerup et al., 1994). The phasing of nucleosomes has been demonstrated for some repeats (Vershinin & Heslop-Harrison, 1998). However, it is unclear whether or not there is selection for some preferential monomer size as both long arrays of short monomers typical for micro- and mini-satellites (Ananiev et al., 2005; Heckmann et al., 2013), as well as monomer sequences up to 5 kb (Gong et al., 2012; Ruiz-Ruano et al., 2016), have been described.

SatDNA sequences generally have a nucleotide composition that is AT rich. An analysis of monomers belonging to 152 families of satDNA showed that AT content for most satellites is above 50% although this ranged from 22 to 75% (Macas et al., 2002). Different families of repetitive sequences show a consistent preference for motifs like AA/TT dinucleotides or the pentanucleotide CAAAA. The presence of specific motifs in unrelated satDNA monomers suggests their possible significance for molecular mechanisms underlying the amplification and maintenance of tandem repeats in eukaryotic genomes (Appels et al., 1986; Katsiotis et al., 1998; Macas et al., 2002; Mehrotra & Goyal, 2014). Moreover, the richness of AT as well as the existence of clusters of AA/TT periodically spaced have been related to the degree of DNA curvature (Palomeque & Lorite, 2008). Although the potential role of the DNA curvature is not well established, it could be implicated in chromatin organization or specific protein binding (Melters et al., 2013; Pezer et al., 2012). Moreover, inverted and palindromic motifs could operate as nucleosome-positioning signals (Barceló et al., 1998; Kasinathan & Henikoff, 2018).

The absence of any conserved sequences, apart from short AT-rich motifs, suggests that satellites can originate from any sequence. Whether or not satDNA is further subject to some sort of selection for monomer size or structural features is yet to be fully understood. It is clear however that satDNA displays an extraordinary sequence variation even between closely related species (Macas et al., 2015). The rapid elimination/amplification of their copy number or the different degree of similarity observed between members of the same satDNA family among species reveals the dynamism of this genome fraction (Charlesworth et al., 1994; Maumus & Quesneville, 2014). Although some satellites find representation in members of the same taxa (S. Sharma & Raina, 2005), most families remain species-specific (Macas et al., 2002). A significant variation of repetitive sequences can be observed in *Oryza sativa* when compared to close relatives. The 2.7 fold variation in nuclear DNA content between genomes of diploid cultivated rice and wild rice has been attributed to the amplification of species-specific satDNA families (Uozu et al., 1997). Recent studies demonstrated that even satDNA families shared among related species suffer a certain degree of variation on their monomer sequences that could lead to the generation of species-specific satDNAs (Belyayev et al., 2019). Among the variation observed, either within arrays of the same species or between conserved satellites in different species, the formation of high order repeats (HORs) is recurrent. The formation of longer monomers by the homogenization of dimers, trimers, k-mers, has been frequently reported in plant and animal kingdoms (Willard & Waye, 1987). In addition, the location of the satDNA family on the chromosomes seems to influence the fate of its sequences. The satDNA family VicTR-B is conserved in several species of the genus *Vicia* and the general monomer size of this satellite is 38 bp(Macas et al., 2000). However, the location of this satellite varies from interstitial to subtelomeric between species. The arrays of VicTR-B show more similarity between species were they are located within chromosome arms, whereas, the arrays forming HORs were found to be specific from the subtelomeric loci in *V. grandiflora* (Macas et al., 2006). In humans, the subtelomeric regions were found to be hotspots for interchromosomal recombination (Linardopoulou et al., 2005), and in plant chromosomes they were found to have a complex long-range structure, probably resulting from extensive rearrangements (Alkhimova et al., 2004), perhaps promoting the turnover of the located sequences. Nevertheless, it is not yet fully understood how different satDNA families diverge among species. On the other hand, some "frozen" satDNA families can persist immutable over long evolutionary times in different close relatives (Biscotti et al., 2015; Mravinac et al., 2002, 2005; Petraccioli et al., 2015).

Having detailed information on the satDNA landscapes of related species may provide answers as to why we observe high turnover of certain repeats while others seem to be frozen in time. It is unclear how much our present understanding is simply constrained by technical limitations, as recent studies using advanced NGS-based approaches have discovered many satellites within a species for which just a few were previously reported (Ruiz-Ruano et al., 2016).

2. Origin of satDNA

The mechanisms responsible for the variability described for different satDNA families are not fully understood. Smith (1976) developed a theoretical framework using computer simulations which significantly improved the predictions of the evolutionary dynamics of tandemly repeated elements (Smith, 1976). Smith's simulations predicted that satDNA could emerge from random non-repetitive sequences by the joint action of unequal recombination and mutation when natural selection is not considered. However, some authors objected to the oversimplification of models which did not include the possibility of recombination occurring within a chromatid, a process that would inevitably result in deletions. Assuming that only recombination mechanisms are implicated in satDNA evolution, tandemly repeated arrays will decrease over time. Consequently, recombination-based processes alone cannot account for the persistence of satDNA over generations (Walsh, 1987). In line with these assumptions, it was postulated that satDNA would accumulate in genomic regions with suppressed meiotic recombination (Charlesworth et al., 1986; Stephan, 1986). Further models evaluated the role of natural selection on the evolution of satDNA showing that it could act as a means of controlling the length of the array itself but not the nucleotide sequence (Stephan & Cho, 1994).

A parallel line of research using molecular and cytogenetic approaches found that satDNA could emerge from other repetitive elements of the genome. For example, several satDNA families seem to arise from the intergenic spacer (IGS) of the rDNA cassette. Evidence of this can be found in legumes where the satDNA S12 in *Vicia sativa* or pVf7 in *V. faba* have similarities to the IGS repeats (Falquet et al., 1997; Macas et al., 2003; Maggini et al., 1991). The same was also shown in tobacco (Lim et al., 2004; Volkov et al., 1999) and tomato (Stupar et al., 2002). However, similarities between IGS and satDNA arrays placed outside the rDNA loci can be explained either by the transposition of a preexistent satDNA into the IGS (Maggini et al., 1991), or by the reverse mechanism, i.e., the satellite is generated by the spread of the IGS to another genomic region (Macas et al., 2003; Unfried et al., 1991). Bioinformatic analysis of LTR-retrotransposons from plant genomic sequence data revealed the frequent occurrence of variable tandem repeats within the 3' UTR element (Cafasso et al., 2009; Macas et al., 2009). The observation of micro and minisatellites embedded within mobile elements is in agreement with the idea of TE facilitating their amplification and dispersion along the genome (Inukai, 2004; Smýkal et al., 2009). This event may indicate that the transposition machinery of the LTR-retrotransposons could be involved in the generation of novel satDNAs. LTR-retrotransposons can generate a library of short repeats that can subsequently be dispersed through the genome and eventually further amplified and homogenized into novel satellite repeats. Moreover, similarities between satDNA and parts of mobile elements which do not contain tandem repeats (A. Sharma et al., 2013; Tek et al., 2005), or to single-copy sequences, have also been reported (Pelizaro Valeri et al., 2018). These findings corroborate the idea that tandem repeats can arise from any sequence which is not constrained by evolution. However, the occurrence of these examples among eukaryotic genomes are scarce and must be further explored in order to find intermediary development stages of a satDNA family.

3. SatDNA evolution

Regardless of the primary origin of short tandemly repeated arrays, additional mechanisms are thought to mediate their expansion into long arrays of homogenized tandem repeats. Each satDNA family usually undergoes fast sequence homogenization resulting in high similarity of monomers within the array as well as between arrays members of the same family. This phenomenon is known as concerted evolution and refers to the fact that each repeated unit does not evolve independently, resulting in a greater similarity within monomers of the same species than with those present in close relatives. Concerted evolution is believed to be the result of a number of DNA repair and replication mechanisms that are involved in the amplification of the repetitive elements (Elder & Turner, 1995). However amplification is a broad term that can include several mechanisms. Some authors proposed the reinsertion of a circular molecule produced by intrastrand recombination processes as a means of amplification for repetitive families. The extra-chromosomal circular DNA (eccDNA) might contain the origin of replication, becoming a template for rolling circle replication and further reinsertion (Walsh, 1987). The presence of eccDNA derived from tandem repeats was demonstrated in *Arabidopsis* and *Brachycome dichromosomatica* (Cohen et al., 2008; Zellinger et al., 2007), as well as in various genera of higher plants (Navrátilová et al., 2008). However, there is no evidence of the reintegration of these extra-chromosomal elements into the genome. Segmental duplication and gene conversion are two other mechanisms capable of creating homogeneous arrays of tandem repeats. The former has been observed for the satDNA expansion in rice centromeres (Ma & Jackson, 2006) or telomeric tandem repeats in the human genome (Linardopoulou et al., 2005). Gene conversion refers to the events of nonreciprocal transfer between homologous sequences. Thus, one but not the other DNA is locally changed to the genotype of the other. This mechanism has been shown to be involved in the process of concerted evolution that tandemly repeated sequences display (Kawabe & Charlesworth, 2007; Shi et al., 2010; Sun et al., 2012; Talbert & Henikoff, 2010) Since each of these mechanisms leave specific molecular footprints, this question could be addressed by searching for these patterns of evolution within long tracks of satellite sequences.

4. Centromeric satellites

Although satDNA can occur at various chromosomal locations, it is predominantly found at centromeric regions (Oliveira & Torres, 2018). Centromeres are chromosomal regions responsible for the proper segregation of chromosomes during cell division. At centromeres, kinetochore proteins assemble, serving as an anchor point for the attachment of the microtubule fibers

(McKinley & Cheeseman, 2015). Centromeres have specific features that distinguish them from other regions of the chromosome. Perhaps the most general feature is the presence of a variant of the histone H3, CENH3 (CENP-A in mammals, CID in *Drosophila* or CSE4 in budding yeast (Earnshaw et al., 2013)) and other proteins of the constitutive centromere-associated network (Hara & Fukagawa, 2017; Jiang et al., 2003). CENH3, in contrast to the canonical histone H3, is not conserved between related species, and bears a highly variable N-terminal tail. In addition, the chromatin associated with CENH3 show signatures of suppressed meiotic recombination, as has been demonstrated for several plant species (Copenhaver et al., 1998; Gore et al., 2009; Saintenac et al., 2009; Tanksley et al., 1992), humans (Puechberty et al., 1999), and it also display specific profiles of epigenetic modifications (Fuchs & Schubert, 2012).

Whether and how these features drive the evolution of underlying centromeric sequences remains controversial. Centromeres in eurkaryotes have evolved different forms of organization: point centromeres (e.g. budding yeast), regional centromeres/monocentric (e.g. *Arabidopsis*), holocentromeres (e.g *Cuscuta europea*, *Rhynchospora*) and meta-polycentric (e.g. *Pisum sativum*) (Neumann et al., 2012; Plohl et al., 2014; Schubert et al., 2020). These types of centromeres are different in terms of their CENH3 organization along the chromosome. The simplest are the budding yeast point centromeres, which span only 125 bp and thus involve just a single nucleosome. Regional centromeres are more complex and can be from 4 kb up to several megabases long (Burrack & Berman, 2012; Pidoux & Allshire, 2004). In regional centromeres the chromatin containing CENH3 can be observed as a single compact domain at the primary constriction of metaphase chromosomes. The meta-polycentric chromosomes display an extended primary constriction where the CENH3 chromatin is located over more than one domain (Neumann et al., 2015). In contrast, holocentric species lack a primary constriction since the domains of CENH3 are distributed along nearly the whole length of the chromosome. With the exception of point centromeres that are determined by a specific sequence motif (Bloom & Carbon, 1982), the role of the centromeric DNA sequence is not properly understood. Despite centromeric regions playing an essential role in cell functioning, the underlying sequence is rarely conserved even between related species. Early studies demonstrated that centromeric satDNA diverge even between related taxa (Wang et al., 2009). This has opened a debate about the importance of satDNA for centromere establishment and function. One of the most influential hypotheses is that of centromere drive (Henikoff et al., 2001) which assumes an interaction between kinetochore proteins and the centromeric satellite in a sequence-specific manner. This interaction is supposed to occur in species with an asymmetric female meiois where the divided genetic material competes for inclusion into the egg cell. In this model, the expansion of a satellite array in one of the homologous chromosomes results in a stronger centromere, which would bind more kinetochore proteins facilitating its transmission to the egg. However, this process would be deleterious for male meiosis where the tendency of expansion could be restored by changes in CENH3 affinity. Centromeric satDNA consequently diversifies between related species and CENH3 or other kintetocore proteins undergo adaptive evolution. A single satellite whose sequence diverges between close relatives has been reported from rice (Lee et al., 2005), *Medicago* (Yu et al., 2017) and some *Brassicaceae* species (Lermontova et al., 2014), thus supporting the predictions of the centromere drive model. Moreover, adaptive evolution of CENH3 has been reported in some of these species as well as in some other taxa with asymmetric meiosis (Cooper & Henikoff, 2004; Hirsch et al., 2009; Zedek & Bureš, 2016).

On the other hand, some observations are not consistent with the hypothesized evolutionary arms race between CENH3 and its underlying centromeric satDNA (Kawabe et al., 2006)

(Masonbrink et al., 2014). First, it is unclear whether or not a specific satDNA array is needed for the centromere function (Roberti et al., 2019), since centromeres with more than one satDNA family or those lacking repeats in some chromosomes of the complement have been reported (Gong et al., 2012; Neumann et al., 2012). Moreover, maize lines carrying homologous chromosomes with different centromere sizes were used to test whether the differences would result in meiotic drive of the large centromeres in female meiosis. This effort demonstrated that centromere size does not drastically affect the segregation of the different chromosomes (Han et al., 2018). In addition, functional complementation is observed when CENH3 proteins from *Lepidium oleaceum* and *Zea mays* correctly assume their role in *Arabidopsis thaliana* centromeres (Maheshwari et al., 2017). This observation appears to be in conflict with the expected specificity of CENH3 toward certain types of repeats. This may indicate that the process of centromere drive is not as common as expected, or that it occurs during limited periods of centromere evolution.

Hypotheses accounting for other features rather than the specific underlying sequence explain the role of satDNA in centromere evolution from a more passive perspective. It has been proposed that the tandem structure itself might be favorable for the deposition of centromeric proteins. Hence, homologous recombination between identical repeats would promote the formation of loops that could be necessary for protein deposition (McFarlane & Humphrey, 2010). In addition, it has recently been proposed that proper levels of centromeric satDNA transcription are necessary for centromeric function (Duda et al., 2017). Transcripts derived from CentO, the centromeric repeat of *Oryza punctata* (Zhang et al., 2005), CentC in *Zea mays* (Topp et al., 2004), or cen180 in *Arabidopsis* (May et al., 2005) have been shown to be necessary for successful segregation of the chromosomes during cell division. Moreover, it has been speculated that centromere sequence composition can be driven by inbreeding and selection for centromere-linked genes (Schneider et al., 2016). Therefore, several forces might be shaping centromeric regions although the interplay of satDNA and centromeres is yet to be established. The evaluation of each scenario could be addressed by gathering information about centromeric satellite sequences and kinetochore proteins from a wide range of species and examining them in their phylogenetic context.

5. Methodological approaches for identification and characterization of satDNA in plant genomes

Even in the present post-genomic era, satDNA remains the least characterized component of most investigated genomes. The genomic organization of this repetitive fraction into arrays of highly homogenized monomers extending over hundreds of kilobases have made its characterization difficult. Historically, there have been a variety of experimental methods used to characterize satDNA, while more recent efforts have included bioinformatic techniques. The efficiency and specific limitations of each approach have impacted the quality of the information garnered (Garrido-Ramos, 2017).

SatDNA was discovered by density gradient centrifugation experiments (Kit, 1961). Differentiated satellite bands, from which satDNA derived its name, formed as a result of different buoyant densities compared to the bulk of genomic DNA, thus warranting further analysis of its composition. The establishment of C_ot analysis based on renaturation kinetics served as a tool by which different tandem repeats were able to be selectively cloned and characterized by their abundance and complexity (Britten & Kohne, 1968). Further methods for the characterization of satellite DNA included the digestion of nuclear DNA by restriction endonucleases (Singer, 1982) or self-priming of the repeated sequence in a modified PCR setup (Buntjer & Lenstra, 1998). Both methods have been usually followed by the analysis of the resulting pattern by electrophoresis in agarose gel. The approach based on the presence of specific restriction sites in the monomer sequence of a satDNA array provided information about the sequence length of the repeated unit (Singer, 1982). In addition, self-priming of repeated sequences in a mixture of sheared and high-molecular-weight genomic DNA can lead to the amplification of some satDNA families in the form of very long concatenated sequences. The result of the amplification can then be visualized after gel electrophoresis as a smear of high molecular weight DNA fragments (Buntjer & Lenstra, 1998) (Macas et al., 2000). Although these methods were of use for many years, the satellites lacking restriction sites and those present in small proportions in the genome were difficult to identify. Further cloning of the DNA excised from the agarose gel followed by Sanger sequencing of the independent clones provided information about the nucleotide composition of a few monomers per satellite analyzed from the vast amount that represent an array. However, this approach is in principle unable to identify satellite repeats lacking suitable restriction sites. Therefore, the structural features of satDNA, as with the continuity of the arrays or the arrangement of monomers along the array, remained unknown. The resulting data on the characterization of satDNA in different species were scattered and suffered from methodological bias. Consequently, investigations of repetitive elements in large and complex eukaryotic genomes have been constrained by the lack of representative sequencing data.

The situation improved with the introduction of next generation sequencing (NGS). Although NGS technologies can produce a high-throughput of genomic data, most of them suffer from the short length of the produced reads that limit their utilization for the assembly of repetitive regions such as satDNA arrays (Treangen & Salzberg, 2012). In this regard, most of the available tools able to process large amounts of NGS data either require complete genome assemblies for repeat identification or are based on similarity searches of databases of previously characterized satDNA. Although complete genome sequence assemblies are available for a number of model species, the repetitive fraction is generally underrepresented even for well-studied genomes such as the human genome (Altemose et al., 2014; Miga et al., 2020). Moreover, considering satDNA is among the most dynamic components of eukaryotic genomes (Macas et al., 2002), it is not possible to identify new satellite repeats by their similarity to known repeats from phylogenetically distant taxa. As such the search for satDNA repeats should be ideally performed in unassembled reads. Nevertheless, the short length of the reads provided by NGS technologies has been a limiting factor.

A breakthrough on the characterization of satDNA was provided by Novak et al. with the development of RepeatExplorer (Novák et al., 2010, 2013) a combination of software tools that analyze unassembled low-pass genome sequencing data making use of the similarity-based clustering algorithm. The authors demonstrated that the low-pass genome sequencing provided by a single 454 sequencing reaction is sufficient to capture information about all major repeat families, decreasing the cost of the identification of repeats (Novák et al., 2010). In addition, the introduction of similarity-based clustering algorithms performing all-to-all comparisons over whole-genome shotgun reads proved to be efficient for repeat identification from NGS data. The clustering algorithm employed by RepeatExplorer represents reads as nodes and sequence similarities as connecting edges in a virtual graph, and identifies clusters by examination of the graph topology. The shapes of the graphs reflect the genomic organization and sequence variability of the identified elements. Therefore, satDNAs are graphically represented with a circular or globular shape due to their tandem structure. This strategy turned out to be appropriate for the identification of tandem repeats, leading to similar strategies being employed by other authors (Kelly et al., 2015; Ruiz-Ruano et al., 2016, 2017) However, this process required visual inspection and did not provide the monomer sequence of the elements, which can be necessary for further analyses requiring the design of PCR primers. The monomer's reconstruction using multiple sequence alignments is often truncated when the size of the monomer expands over the length of the read. An alternative approach using k-mers frequencies to reconstruct the most frequent monomer from unassembled reads proved efficient in reconstructing the centromeric satellite CentO in rice (Macas et al., 2010). The development of TAREAN (Novák et al., 2017) facilitated the unsupervised identification and characterization of satDNA from unassembled sequencing reads. Based on the principles of RepeatExplorer, TAREAN examines the presence of circular-shaped graphs characteristic for tandem repeats. Then, the reads from these clusters are decomposed to k-mers and the most frequent k-mers are then used for reconstructing the most representative monomers for each satellite repeat. TAREAN automates the workflow for the identification of satDNA and implements alignment-free approaches that are suitable for monomer reconstruction from unassembled reads using k-mers frequency statistics. This method was shown to be efficient in several species (Ebrahimzadegan et al., 2019; Ribeiro et al., 2020; Saint-Oyant et al., 2018). Implementing these bioinformatic tools has created a more realistic picture of the diversity of satDNA within and between genomes (Hobza et al., 2017; Macas et al., 2015).

Although the short-read NGS platforms, in combination with the proper bioinformatic tools, are highly efficient at discovering novel satellite repeats, they do not provide any insight into their large-scale arrangement and sequence variability patterns. The recent introduction of the so-called long-read sequencing technologies by Pacific Bioscience and Oxford Nanopore has increased the potential of comprehensively studying tandem repeats. The benefit of this approach lies in the long length of the reads that are able to produce. The Pacific Bioscience long-read technology, termed SMRT (single-molecule real-time), uses a circular DNA created by ligating adapters to both ends of the dsDNA. Further replication of the circular DNA incorporates fluorescently labeled nucleotides, producing a fluorescent signal recorded by a camera in real-time. On the other hand, Oxford Nanopore sequencing reads the sequence directly from a native DNA strand during its passage through a molecular pore. In contrast to SMRT, the nanopore read length is not restricted by the method of sequencing *per se* but rather by the quality of the isolated DNA used. SatDNA research could benefit from incorporating these technologies since it should be possible to infer various features of satellite repeats by analyzing repeat arrays, or their parts, present in individual nanopore reads. Combining with other genome sequencing and mapping data can generate hybrid assemblies in which satellite arrays are faithfully represented and then analyzed. This approach has proven successful on the assembly of the centromere of human chromosome Y (Jain et al., 2015), as well as the whole chromosome X, telomere to telomere (Miga et al., 2020). Moreover, it has been used to investigate the homogenization patterns of satDNA in *Drosophila* (Khost et al., 2017), and calculate the expansion and methylation status of short monomer tandem repeats (Gießelmann et al., 2018). Alternatively, assembly-free strategies can also be applied to infer features of satellite repeats by analyzing repeat arrays or their parts present in individual long-reads (Cechova et al., 2019; Harris et al., 2019).

References

Alkhimova, O. G., Mazurok, N. A., Potapova, T. A., Zakian, S. M., Heslop-Harrison, J. S., & Vershinin, A. V. (2004). Diverse patterns of the tandem repeats organization in rye chromosomes. *Chromosoma*, *113*(1), 42–52. https://doi.org/10.1007/s00412-004-0294-4

Altemose, N., Miga, K. H., Maggioni, M., & Willard, H. F. (2014). Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Computational Biology*, *10*(5), e1003628. https://doi.org/10.1371/journal.pcbi.1003628

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I. J., Koblížková, A., Macas, J., & Lysak, M. A. (2011). Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria lilies*. *Annals of Botany*, *107*(2), 255–268. https://doi.org/10.1093/aob/mcq235

Ananiev, E. V., Chamberlin, M. A., Klaiber, J., & Svitashev, S. (2005). Microsatellite megatracts in the maize (*Zea mays* L.) genome. *Genome*, *48*(6), 1061–1069. https://doi.org/10.1139/g05-061

Appels, R., Moran, L. B., & Gustafson, J. P. (1986). Rye heterochromatin. I. Studies on clusters of the major repeating sequence and the identification of a new dispersed repetitive sequence element. *Canadian Journal of Genetics and Cytology*, *28*(5), 645–657. https://doi.org/10.1139/g86-094

Barceló, F., Gutiérrez, F., Barjau, I., & Portugal, J. (1998). A theoretical perusal of the satelliteDNA curvature in tenebrionid beetles. *Journal of Biomolecular Structure and Dynamics*, *16*(1), 41–50. https://doi.org/10.1080/07391102.1998.10508225

Belyayev, A., Josefiová, J., Jandová, M., Kalendar, R., Krak, K., & Mandák, B. (2019). Natural history of a satellite DNA family: From the ancestral genome component to species-specific sequences, concerted and non-concerted evolution. *International Journal of Molecular Sciences*, *20*(5). https://doi.org/10.3390/ijms20051201

Biscotti, M. A., Canapa, A., Forconi, M., Olmo, E., & Barucca, M. (2015). Transcription of tandemly repetitive DNA: functional roles. *Chromosome Research*, *23*(3), 463–477. https://doi.org/10.1007/s10577-015-9494-4

Bloom, K. S., & Carbon, J. (1982). Yeast centromere DNA is in a unique and highly ordered structure in chromosomes and small circular minichromosomes. *Cell*, *29*(2), 305–317. https://doi.org/10.1016/0092-8674(82)90147-7

Britten, R. J., & Kohne, D. E. (1968). Repeated Sequences in DNA. *Science*, *161*(3841), 529–540. https://doi.org/10.1126/science.161.3841.529

Buntjer, J. B., & Lenstra, J. A. (1998). Self-amplification of satellite DNA *in vitro*. *Genome*, *41*(3), 429–434. https://doi.org/10.1139/g98-036

Burrack, L. S., & Berman, J. (2012). Flexibility of centromere and kinetochore structures. *Trends in Genetics*, *28*(5), 204–212. https://doi.org/10.1016/j.tig.2012.02.003

Cafasso, D., Cozzolino, S., Vereecken, N. J., Luca, P., & Chinali, G. (2009). Organization of a dispersed repeated DNA element in the *Zamia* genome. *Biologia Plantarum*, *53*(1), 28–36. https://doi.org/10.1007/s10535-009-0005-3

Cechova, M., Harris, R. S., Tomaszkiewicz, M., Arbeithuber, B., Chiaromonte, F., & Makova, K. D. (2019). High satellite repeat turnover in great apes studied with short- and long-read technologies. *Molecular Biology and Evolution*, *36*(11), 2415–2431. https://doi.org/10.1093/molbev/msz156

Charlesworth, B., Langley, C. H., & Stephan, W. (1986). The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics*, *112*(4), 947–962.

Charlesworth, B., Sniegowski, P., & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, *371*(6494), 215–220. https://doi.org/10.1038/371215a0

Cohen, S., Houben, A., & Segal, D. (2008). Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants. *Plant Journal*, *53*(6), 1027–1034. https://doi.org/10.1111/j.1365-313X.2007.03394.x

Cooper, J. L., & Henikoff, S. (2004). Adaptive evolution of the histone fold domain in centromeric histones. *Molecular Biology and Evolution*, *21*(9), 1712–1718. https://doi.org/10.1093/molbev/msh179

Copenhaver, G. P., Browne, W. E., & Preuss, D. (1998). Assaying genome-wide recombination and centromere functions with *Arabidopsis* tetrads. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(1), 247–252. https://doi.org/10.1073/pnas.95.1.247

Duda, Z., Trusiak, S., & O’Neill, R. (2017). Centromere transcription: means and motive. In *B.E. Black (ed.), Centromeres and kinetochores, progress in molecular and subcellular biology* (pp. 257–281). Springer, Cham. https://doi.org/10.1007/978-3-319-58592-5_11

Earnshaw, W. C., Allshire, R. C., Black, B. E., Bloom, K., Brinkley, B. R., Brown, W., Cheeseman, I. M., Choo, K. H. A., Copenhaver, G. P., Deluca, J. G., Desai, A., Diekmann, S., Erhardt, S., Fitzgerald-Hayes, M., Foltz, D., Fukagawa, T., Gassmann, R., Gerlich, D. W., Glover, D. M., … Cleveland, D. W. (2013). Esperanto for histones: CENP-A, not CenH3, is the centromeric histone H3 variant. *Chromosome Research*, *21*(2), 101–106. https://doi.org/10.1007/s10577-013-9347-y

Ebrahimzadegan, R., Houben, A., & Mirzaghaderi, G. (2019). Repetitive DNA landscape in essential A and supernumerary B chromosomes of Festuca pratensis Huds. *Scientific Reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-56383-1

Elder, J. F., & Turner, B. J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology*, *70*(3), 297–320. http://www.ncbi.nlm.nih.gov/pubmed/7568673

Falquet, J., Creusot, F., & Dron, M. (1997). Molecular analysis of *Phaseolus vulgaris* rDNA unit and characterization of a satellite DNA homologous to IGS subrepeats. *Plant Physiology and Biochemistry*, *35*(8), 611–622.

Fuchs, J., & Schubert, I. (2012). Chromosomal distribution and functional interpretation of epigenetic histone marks in plants. In *Plant Cytogenetics: Genome Structure and Chromosome Function* (Vol. 9, pp. 231–253). Springer New York. https://doi.org/10.1007/978-0-387-70869-0_9

Garrido-Ramos, M. A. (2015). Satellite DNA in plants: More than just rubbish. *Cytogenetic and Genome Research*, *146*(2), 153–170. https://doi.org/10.1159/000437008

Garrido-Ramos, M. A. (2017). Satellite DNA: An evolving topic. *Genes*, *8*(9), 230. https://doi.org/10.3390/genes8090230

Gießelmann, P., Brändl, B., Raimondeau, E., Bowen, R., Rohrandt, C., Tandon, R., Kretzmer, H., Assum, G., Galonska, C., Siebert, R., Ammerpohl, O., Heron, A., Schneider, S., Ladewig, J., Koch, P., Schuldt, B., Graham, J., Meissner, A., & Müller, F.-J. (2018). Repeat expansion and methylation state analysis with nanopore sequencing. *BioRxiv (Preprint Server)*, *37*, 1478–1481. https://doi.org/10.1101/480285

Gong, Z., Wu, Y., Koblížková, A., Torres, G. a, Wang, K., Iovene, M., Neumann, P., Zhang, W., Novák, P., Buell, C. R., Macas, J., & Jiang, J. (2012). Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell*, *24*(9), 3559–3574. https://doi.org/10.1105/tpc.112.100511

Gore, M. A., Chia, J.-M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., McMullen, M. D., Grills, G. S., Ross-Ibarra, J., Ware, D. H., & Buckler, E. S. (2009). A first-generation haplotype map of maize. *Science (New York, N.Y.)*, *326*(5956), 1115–1117. https://doi.org/10.1126/science.1177837

Han, F., Lamb, J. C., McCaw, M. E., Gao, Z., Zhang, B., Swyers, N. C., & Birchler, J. A. (2018). Meiotic studies on combinations of chromosomes with different sized centromeres in maize. *Frontiers in Plant Science*, *9*, 785. https://doi.org/10.3389/fpls.2018.00785

Hara, M., & Fukagawa, T. (2017). Critical foundation of the kinetochore: The Constitutive Centromere-Associated Network (CCAN). In *Progress in molecular and subcellular biology* (Vol. 56, pp. 29–57). NLM (Medline). https://doi.org/10.1007/978-3-319-58592-5_2

Harris, R. S., Cechova, M., & Makova, K. D. (2019). Noise-cancelling repeat finder: Uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics*, *35*(22), 4809–4811. https://doi.org/10.1093/bioinformatics/btz484

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

Henikoff, S., Ahmad, K., & Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science (New York, N.Y.)*, *293*(5532), 1098–1102. https://doi.org/10.1126/science.1062939

Hirsch, C. D., Wu, Y., Yan, H., & Jiang, J. (2009). Lineage-specific adaptive evolution of the centromeric protein CENH3 in diploid and allotetraploid *Oryza* species. *Molecular Biology and Evolution*, *26*(12), 2877–2885. https://doi.org/10.1093/molbev/msp208

Hobza, R., Cegan, R., Jesionek, W., Kejnovsky, E., Vyskot, B., & Kubat, Z. (2017). Impact of repetitive elements on the Y chromosome formation in plants. *Genes*, *8*(11), 302. https://doi.org/10.3390/genes8110302

Inukai, T. (2004). Role of transposable elements in the propagation of minisatellites in the rice genome. In *Molecular Genetics and Genomics* (Vol. 271, Issue 2, pp. 220–227). Springer. https://doi.org/10.1007/s00438-003-0973-5

Jagannathan, M., Cummings, R., & Yamashita, Y. M. (2018). A conserved function for pericentromeric satellite DNA. *ELife*, *7*, 1–19. https://doi.org/10.7554/eLife.34122

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., & Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, *12*(4), 351–356. https://doi.org/10.1038/nmeth.3290

Jiang, J., Birchler, J. a., Parrott, W. a., & Dawe, R. K. (2003). A molecular view of plant centromeres. *Trends in Plant Science*, *8*(12), 570–575. https://doi.org/10.1016/j.tplants.2003.10.011

Kasinathan, S., & Henikoff, S. (2018). Non-B-Form DNA is enriched at centromeres. *Molecular Biology and Evolution*, *35*(April), 949–962. https://doi.org/10.1093/molbev/msy010

Katsiotis, A., Hagidimitriou, M., Douka, A., & Hatzopoulos, P. (1998). Genomic organization, sequence interrelationship, and physical localization using in situ hybridization of two tandemly repeated DNA sequences in the genus *Olea*. *Genome*, *41*(4), 527–534. https://doi.org/10.1139/g98-045

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*(2), 237–247. https://doi.org/10.1007/s00239-006-0097-8

Kawabe, A., Nasuda, S., & Charlesworth, D. (2006). Duplication of centromeric histone H3 (HTR12) gene in *Arabidopsis halleri* and *A. lyrata* , plant species with multiple centromeric satellite sequences. *Genetics*, *174*(4), 2021–2032. https://doi.org/10.1534/genetics.106.063628

Kelly, L. J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., Lysak, M. A., Day, P. D., Berger, M., Fay, M. F., Nichols, R. A., Leitch, A. R., & Leitch, I. J. (2015). Analysis of the giant genomes of *Fritillaria* (*Liliaceae*) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytologist*, *208*(2), 596–607. https://doi.org/10.1111/nph.13471

Khost, D. E., Eickbush, D. G., & Larracuente, A. M. (2017). Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila*. *Genome Research*, *27*(5), 709–721. https://doi.org/10.1101/gr.213512.116

Kit, S. (1961). Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *Journal of Molecular Biology*, *3*(6), 711-IN2. http://www.sciencedirect.com/science/article/pii/S0022283661800752

Lee, H.-R., Zhang, W., Langdon, T., Jin, W., Yan, H., Cheng, Z., & Jiang, J. (2005). Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(33), 11793–11798. https://doi.org/10.1073/pnas.0503863102

Lermontova, I., Sandmann, M., & Demidov, D. (2014). Centromeres and kinetochores of *Brassicaceae*. *Chromosome Research*, *22*(2), 135–152. https://doi.org/10.1007/s10577-014-9422-z

Lim, K. Y., Skalicka, K., Koukalova, B., Volkov, R. A., Matyasek, R., Hemleben, V., Leitch, A. R., & 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*(4), 1935–1946. https://doi.org/10.1534/genetics.166.4.1935

Linardopoulou, E. V., Williams, E. M., Fan, Y., Friedman, C., Young, J. M., & Trask, B. J. (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, *437*(7055), 94–100. https://doi.org/10.1038/nature04029

Lowman, H., & Bina, M. (1990). Correlation between dinucleotide periodicities and nucleosome positioning on mouse satellite DNA. *Biopolymers*, *30*(9–10), 861–876. https://doi.org/10.1002/bip.360300902

Ma, J., & Jackson, S. A. (2006). Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Research*, *16*(2), 251–259. https://doi.org/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*(2), 198–206. https://doi.org/10.1016/j.gene.2009.06.014

Macas, J., Meszaros, T., & Nouzova, M. (2002). PlantSat: a specialized database for plant satellite repeats. *Bioinformatics*, *18*(1), 28–35. https://doi.org/10.1093/bioinformatics/18.1.28

Macas, J., Navrátilová, A., & Koblízková, A. (2006). Sequence homogenization and chromosomal localization of VicTR-B satellites differ between closely related *Vicia* species. *Chromosoma*, *115*(6), 437–447. https://doi.org/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*(3), 152–158. https://doi.org/10.1007/s00412-003-0255-3

Macas, J., Neumann, P., Novák, P., & Jiang, J. (2010). Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data. *Bioinformatics*, *26*(17), 2101–2108. https://doi.org/10.1093/bioinformatics/btq343

Macas, J., Novák, P., Pellicer, J., Čížková, J., Koblížková, A., Neumann, P., Fuková, I., Doležel, J., Kelly, L. J., & Leitch, I. J. (2015). In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe *Fabeae*. *PLoS ONE*, *10*(11), 1–23. https://doi.org/10.1371/journal.pone.0143424

Macas, J., Požárková, D., Navrátilová, A., Nouzová, M., & Neumann, P. (2000). Two new families of tandem repeats isolated from genus *Vicia* using genomic self-priming PCR. *Molecular and General Genetics MGG*, *263*(5), 741–751. https://doi.org/10.1007/s004380000245

Maggini, F., Cremonini, R., Zolfino, C., Tucci, G. F., D'Ovidio, R., Delre, V., DePace, C., Scarascia Mugnozza, G. T., & Cionini, P. G. (1991). Structure and chromosomal localization of DNA sequences related to ribosomal subrepeats in *Vicia faba*. *Chromosoma*, *100*(4), 229–234. https://doi.org/10.1007/BF00344156

Maheshwari, S., Ishii, T., Brown, C. T., Houben, A., & Comai, L. (2017). Centromere location in *Arabidopsis* is unaltered by extreme divergence in CENH3 protein sequence. *Genome Research*, *27*(3), 471–478. https://doi.org/10.1101/gr.214619.116

Masonbrink, R. E., Gallagher, J. P., Jareczek, J. J., Renny-Byfield, S., Grover, C. E., Gong, L., & Wendel, J. F. (2014). CenH3 evolution in diploids and polyploids of three angiosperm genera. *BMC Plant Biology*, *14*(1), 1–11. https://doi.org/10.1186/s12870-014-0383-3

Maumus, F., & Quesneville, H. (2014). Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. *Nature Communications*, *5*(May). https://doi.org/10.1038/ncomms5104

May, B. P., Lippman, Z. B., Fang, Y., Spector, D. L., & Martienssen, R. A. (2005). Differential regulation of strand-specific transcripts from *Arabidopsis* centromeric satellite repeats. *PLoS Genetics*, *1*(6), e79. https://doi.org/10.1371/journal.pgen.0010079

McFarlane, R. J., & Humphrey, T. C. (2010). A role for recombination in centromere function. *Trends in Genetics : TIG*, *26*(5), 209–213. https://doi.org/10.1016/j.tig.2010.02.005

McKinley, K. L., & Cheeseman, I. M. (2015). The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm.2015.5

Mehrotra, S., & Goyal, V. (2014). Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. In *Genomics, Proteomics and Bioinformatics* (Vol. 12, Issue 4, pp. 164–171). Beijing Genomics Institute. https://doi.org/10.1016/j.gpb.2014.07.003

Melters, D. P., Bradnam, K. R., Young, H. a, Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J. F., Derisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., & Chan, S. W. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1), R10. https://doi.org/10.1186/gb-2013-14-1-r10

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., … Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Mravinac, B., Plohl, M., Mestrović, N., & UgarkovićD., D. (2002). Sequence of PRAT satellite DNA "frozen" in some coleopteran species. *Journal of Molecular Evolution*, *54*(6), 774–783. https://doi.org/10.1007/s0023901-0079-9

Mravinac, B., Plohl, M., & Ugarković, D. (2005). Preservation and high sequence conservation of satellite DNAs suggest functional constraints. *Journal of Molecular Evolution*, *61*(4), 542–550. https://doi.org/10.1007/s00239-004-0342-y

Navrátilová, A., Koblízková, A., & Macas, J. (2008). Survey of extrachromosomal circular DNA derived from plant satellite repeats. *BMC Plant Biology*, *8*, 90. https://doi.org/10.1186/1471-2229-8-90

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*(6), e1002777. http://dx.plos.org/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 *Fabeae* species. *Molecular Biology and Evolution*, *32*(7), 1862–1879. https://doi.org/10.1093/molbev/msv070

Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P., & Macas, J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*, *12*, 1–10. https://doi.org/10.1093/nar/gkx257

Novák, P., Neumann, P., & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, *11*(1), 378. https://doi.org/10.1186/1471-2105-11-378

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., & Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, *29*(6), 792–793. https://doi.org/10.1093/bioinformatics/btt054

Oliveira, L. C., & Torres, G. A. (2018). Plant centromeres: genetics, epigenetics and evolution. *Molecular Biology Reports*, *45*(5), 1491–1497. https://doi.org/10.1007/s11033-018-4284-7

Palomeque, T., & Lorite, P. (2008). Satellite DNA in insects: A review. *Heredity*, *100*(6), 564–573. https://doi.org/10.1038/hdy.2008.24

Pelizaro Valeri, M., Borges Dias, G., do Socorro Pereira, V., Campos Silva Kuhn, G., & Svartman, M. (2018). An eutherian intronic sequence gave rise to a major satellite DNA in *Platyrrhini*. *Biology Letters*, *14*(1), 20170686. https://doi.org/10.1098/rsbl.2017.0686

Petraccioli, A., Odierna, G., Capriglione, T., Barucca, M., Forconi, M., Olmo, E., & Biscotti, M. A. (2015). A novel satellite DNA isolated in *Pecten jacobaeus* shows high sequence similarity among molluscs. *Molecular Genetics and Genomics*, *290*(5), 1717–1725. https://doi.org/10.1007/s00438-015-1036-4

Pezer, Ž., Brajković, J., Feliciello, I., & Ugarković, Đ. (2012). Satellite DNA-mediated effects on genome regulation. *Garrido- Ramos MA (Ed): Repetitive DNA. Genome Dyn. Basel, Karger.*, *7*, 153–169.

Pidoux, A. L., & Allshire, R. C. (2004). Kinetochore and heterochromatin domains of the fission yeast centromere. *Chromosome Research*, *12*(6), 521–534. https://doi.org/10.1023/B:CHRO.0000036586.81775.8b

Plohl, M., Meštrović, N., & Mravinac, B. (2014). Centromere identity from the DNA point of view. *Chromosoma*, *123*(4), 313–325. https://doi.org/10.1007/s00412-014-0462-0

Puechberty, J., Laurent, A. M., Gimenez, S., Billault, A., Brun-Laurent, M. E., Calenda, A., Marçais, B., Prades, C., Ioannou, P., Yurov, Y., & Roizès, G. (1999). Genetic and physical analyses of the centromeric and pericentromeric regions of human chromosome 5: Recombination across 5cen. *Genomics*, *56*(3), 274–287. https://doi.org/10.1006/geno.1999.5742

Ribeiro, T., Vasconcelos, E., dos Santos, K. G. B., Vaio, M., Brasileiro-Vidal, A. C., & Pedrosa-Harand, A. (2020). Diversity of repetitive sequences within compact genomes of *Phaseolus* L. beans and allied genera *Cajanus* L. and *Vigna* Savi. *Chromosome Research*, *28*(2), 139–153. https://doi.org/10.1007/s10577-019-09618-w

Roberti, Bensi, Mazzagatti, Piras, Nergadze, Giulotto, & Raimondi. (2019). Satellite DNA at the centromere is dispensable for segregation fidelity. *Genes*, *10*(6), 469. https://doi.org/10.3390/genes10060469

Ruiz-Ruano, F. J., Cabrero, J., López-León, M. D., & Camacho, J. P. M. (2017). Satellite DNA content illuminates the ancestry of a supernumerary (B) chromosome. *Chromosoma*, *126*(4), 487–500. https://doi.org/10.1007/s00412-016-0611-8

Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., & Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, *6*(January), 1–14. https://doi.org/10.1038/srep28333

Saint-Oyant, L. H., Ruttink, T., Hamama, L., Kirov, I., Lakhwani, D., Zhou, N. N., Bourke, P. M., Daccord, N., Leus, L., Schulz, D., Van De Geest, H., Hesselink, T., Van Laere, K., Debray, K., Balzergue, S., Thouroude, T., Chastellier, A., Jeauffre, J., Voisine, L., … Foucher, F. (2018). A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nature Plants*, *4*(7), 473–484. https://doi.org/10.1038/s41477-018-0166-1

Saintenac, C., Falque, M., Martin, O. C., Paux, E., Feuillet, C., & Sourdille, P. (2009). Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (*Triticum aestivum* L.). *Genetics*, *181*(2), 393–403. https://doi.org/10.1534/genetics.108.097469

Schneider, K. L., Xie, Z., Wolfgruber, T. K., & Presting, G. G. (2016). Inbreeding drives maize centromere evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, E987–E996. https://doi.org/10.1073/pnas.1522008113

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*(10), 3488. https://doi.org/10.3390/ijms21103488

Sharma, A., Wolfgruber, T. K., & Presting, G. G. (2013). Tandem repeats derived from centromeric retrotransposons. *BMC Genomics*, *14*(1), 142. https://doi.org/10.1186/1471-2164-14-142

Sharma, S., & Raina, S. N. (2005). Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes. *Cytogenetic and Genome Research*, *109*(1–3), 15–26. https://doi.org/10.1159/000082377

Shi, J., Wolf, S. E., Burke, J. M., Presting, G. G., Ross-Ibarra, J., & Dawe, R. K. (2010). Widespread gene conversion in centromere cores. *PLoS Biology*, *8*(3), e1000327. https://doi.org/10.1371/journal.pbio.1000327

Singer, M. F. (1982). Highly repeated sequences in mammalian genomes. *International Review of Cytology*, *76*(C), 67–112. https://doi.org/10.1016/S0074-7696(08)61789-1

Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science (New York, N.Y.)*, *191*(4227), 528–535. 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*(2), 157–167. https://doi.org/10.1038/hdy.2009.45

Stephan, W. (1986). Recombination and the evolution of satellite DNA. *Genetical Research*, *47*(3), 167–174. https://doi.org/10.1017/S0016672300023089

Stephan, W., & Cho, S. (1994). Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics*, *136*(1), 333–341. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205784&tool=pmcentrez&rendertype=abstract

Stupar, R. M., Song, J., Tek, A. L., Cheng, Z., Dong, F., & Jiang, J. (2002). Highly condensed potato pericentromeric heterochromatin contains rDNA-related tandem repeats. *Genetics*, *162*(3), 1435–1444. https://www.genetics.org/content/162/3/1435.short

Sun, Y., Ambrose, J. H., Haughey, B. S., Webster, T. D., Pierrie, S. N., Muñoz, D. F., Wellman, E. C., Cherian, S., Lewis, S. M., Berchowitz, L. E., & Copenhaver, G. P. (2012). Deep genome-wide measurement of meiotic gene conversion using tetrad analysis in *Arabidopsis thaliana*. *PLoS Genetics*, *8*(10), e1002968. https://doi.org/10.1371/journal.pgen.1002968

Talbert, P. B., & Henikoff, S. (2010). Centromeres convert but don’t cross. *PLoS Biology*, *8*(3), 1–5. https://doi.org/10.1371/journal.pbio.1000326

Tanksley, S. D., Ganal, M. W., Prince, J. P., de Vicente, M. C., Bonierbale, M. W., Broun, P., Fulton, T. M., Giovannoni, J. J., Grandillo, S., & Martin, G. B. (1992). High density molecular linkage maps of the tomato and potato genomes. *Genetics*, *132*(4).

Tek, A. L., 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*(3), 1231–1238. https://doi.org/10.1534/genetics.105.041087

Tommerup, H., Dousmanis, A., & de Lange, T. (1994). Unusual chromatin in human telomeres. *Molecular and Cellular Biology*, *14*(9), 5777–5785. https://doi.org/10.1128/mcb.14.9.5777

Topp, C. N., Zhong, C. X., & Dawe, R. K. (2004). Centromere-encoded RNAs are integral components of the maize kinetochore. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(45), 15986–15991. https://doi.org/10.1073/pnas.0407154101

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. In *Nature Reviews Genetics* (Vol. 13, Issue 1, pp. 36–46). Nature Publishing Group. https://doi.org/10.1038/nrg3117

Ugarkovic, D. (2005). Functional elements residing within satellite DNAs. *EMBO Reports*, *6*(11), 1035–1039. https://doi.org/10.1038/sj.embor.7400558

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*(1), 63–68. https://doi.org/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*(6), 791–799. https://doi.org/10.1023/A:1005823124989

Vershinin, A. V, & Heslop-Harrison, J. S. (1998). Comparative analysis of the nucleosomal structure of rye, wheat and their relatives. *Plant Mol Biol*, *36*. https://doi.org/10.1023/A:1005912822671

Volkov, R. A., Borisjuk, N. V., Panchuk, I. I., Schweizer, D., & Hemleben, V. (1999). Elimination and rearrangement of parental rDNA in the allotetraploid *Nicotiana tabacum*. *Molecular Biology and Evolution*, *16*(3), 311–320. https://doi.org/10.1093/oxfordjournals.molbev.a026112

Walsh, J. B. (1987). Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics*, *115*(3), 553–567. http://www.ncbi.nlm.nih.gov/pubmed/3569882

Wang, G., Zhang, X., & Jin, W. (2009). An overview of plant centromeres. *Journal of Genetics and Genomics*, *36*(9), 529–537. https://doi.org/10.1016/S1673-8527(08)60144-7

Willard, H. F., & Waye, J. S. (1987). Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends in Genetics*, *3*(C), 192–198. https://doi.org/10.1016/0168-9525(87)90232-0

Yu, F., Dou, Q., Liu, R., & Wang, H. (2017). A conserved repetitive DNA element located in the centromeres of chromosomes in *Medicago* genus. *Genes and Genomics*, *39*(8), 903–911. https://doi.org/10.1007/s13258-017-0556-1

Zedek, F., & Bureš, P. (2016). CenH3 evolution reflects meiotic symmetry as predicted by the centromere drive model. *Scientific Reports*, *6*, 33308. https://doi.org/10.1038/srep33308

Zellinger, B., Akimcheva, S., Puizina, J., Schirato, M., & Riha, K. (2007). Ku suppresses formation of telomeric circles and alternative telomere lengthening in *Arabidopsis*. *Molecular Cell*, *27*(1), 163–169. https://doi.org/10.1016/j.molcel.2007.05.025

Zhang, W., Yi, C., Bao, W., Liu, B., Cui, J., Yu, H., Cao, X., Gu, M., Liu, M., & Cheng, Z. (2005). The transcribed 165-bp CentO satellite is the major functional centromeric element in the wild rice species *Oryza punctata*. *Plant Physiology*, *139*(1), 306–315. https://doi.org/10.1104/pp.105.064147

AIMS OF THIS WORK

The general objective of this work is to contribute to our understanding of the origin, evolution, and function of satellite DNA in plant genomes by the application of advanced sequencing and bioinformatic approaches combined with molecular cytogenetic experiments. The legume tribe *Fabeae* has been selected as the subject of this study because it includes a diverse set of species differing in their genome size, repeat content, and chromosome morphology. The work has the following aims:

- To perform a comprehensive characterization of sequence diversity and genomic distribution of satellite DNA in the repeat-rich model species *Vicia faba*.

- To identify and characterize satellite repeats associated with centromeric chromatin in a set of related species and investigate the diversity and evolution of these satellites in the phylogenetic context. Interpret the results with respect to the evolution of CENH3 proteins and predictions of the centromere drive model.

- To utilize ultra-long nanopore reads for elucidating the long-range arrangement of satDNA arrays and for investigating the origin of satellite DNA.

SCOPE OF THE THESIS

In **Chapter I**, we explore the satellite DNA landscape in *Vicia faba* (2n=12), a species that has long served as a cytogenetic model. The large genome size of the species (1C=13.41 Gbp), together with the low chromosome number, makes this species suitable for cytogenetic studies. We focused on the characterization of each identified satDNA, notably its sequence composition, abundance, and location on chromosomes, as well as previously described features of the particular loci. In the course of our study, we applied recently introduced bioinformatic tools to the identification of satDNA from NGS data. We found over 30 putative satellites. Surprisingly, the monomer sizes found in this study are far more diverse than those traditionally described, ranging from the prevalent length of 150-350 bp to the unusual 687 bp up to 2033 bp. Using *in situ* hybridization, we were able to verify and localize these satDNA families on metaphase chromosomes of *V. faba*. The results showed a frequent association of the satellites with unusually long monomers with the pericentromeric regions.

We used chromatin immunoprecipitation to identify the repeats associated with the centromeric histone H3 variant, CENH3. An extraordinary diversity of centromeric satDNA was found where each identified centromeric repeat was chromosome-specific. This chromosome-specific location of centromeric repeats is rare among the species described in the literature, which usually possess one satellite family associated with centromeres of all chromosomes. In order to explain this diversity, we analyzed different features of the centromeric elements. We noted that despite their sequence variability, they all follow the same dynamics of mid-S-phase replication while the rest of the satellites replicate in late S-phase. In addition to this, there was no other feature common to all centromeric satellites.

Prompted by the high diversity of centromeric repeats in *V. faba*, in **Chapter II**, we analyzed the composition of satellite DNA associated with CENH3 in different species of the *Fabeae* tribe. This chapter collects the most complete set of data on described centromeric satDNA, together with the CENH3 sequences in the context of their phylogenetic relationship. Following the same approach as in Chapter I, we analyzed 14 species of the tribe, finding a total of 64 centromeric satellites which differ in their nucleotide sequence and length. This work shows that the rare composition of *V. faba* centromeres is often found within the tribe, finding species bearing from 2 to 12 different centromeric satDNAs. Moreover, we tested whether these repeats are conserved between species, revealing that most of them were species-specific.

Furthermore, among the centromeric satDNAs shared by several species, the centromeric role was not necessarily preserved, being found in a pericentromeric or interstitial positions in related species. Even within a species, certain repeats were associated with CENH3 in some chromosomes but in a non-centromeric regions in others. As the centromere drive model proposed the adaptive evolution of CENH3 resulting from an arms race with centromeric repeats, we wanted to test if *Fabeae* CENH3 provides any evidence in support of this model. However, our data showed that CENH3 in the *Fabeae* tribe evolves mainly under purifying selection.

Taken together, these findings suggest that the evolution of *Fabeae* centromeres is not shaped by the co-evolution of a single centromeric satellite with its interacting CENH3, as proposed by the centromere drive model.

In **Chapter III**, we focused on the genome-wide organization of satDNA by making use of Oxford Nanopore sequencing technology. Firstly, we developed a workflow for the identification of satDNA arrays in long nanopore reads. We further analyzed the size distribution of the arrays, the patterns of sequence homogenization as well as their association with other repeats using a member of the legume tribe *Fabeae*, the grass pea *Lathyrus sativus*. The satellite fraction of this species was previously characterized by low-pass sequencing, using RepeatExplorer and TAREAN. These preliminary efforts revealed 23 putative satDNA families, which were used to identify repeats in the long nanopore reads. Moreover, the large genome size of the species (1C= 6.52Gbp), together with the low number of chromosomes (2n=14) makes *L. sativus* suitable for cytogenetic studies. The patterns observed by the analysis of nanopore reads were confirmed by *in situ* hybridization techniques on metaphase chromosomes. The work presented in Chapter III revealed different organizations for different families of satDNA, which allowed for their classification in two groups according to their mechanisms of origin and amplification. One of the most important findings in this chapter is that the majority of satDNA families in *L. sativus* originated from short tandem repeats present in the 3’ untranslated region (3’UTR) of Ogre retrotransposons. Moreover, this work appears to be a proof of concept for the usability of long-nanopore reads to study the long-scale organization of satDNA.

Chapter I

Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 24 November 2017
Accepted: 28 March 2018
Published online: 11 April 2018

Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing

Laura Ávila Robledillo[1,2], Andrea Koblížková[1], Petr Novák[1], Katharina Böttinger[1,2], Iva Vrbová[1], Pavel Neumann[1], Ingo Schubert[3] & Jiří Macas[1]

Satellite DNA, a class of repetitive sequences forming long arrays of tandemly repeated units, represents substantial portions of many plant genomes yet remains poorly characterized due to various methodological obstacles. Here we show that the genome of the field bean (*Vicia faba*, 2n = 12), a long-established model for cytogenetic studies in plants, contains a diverse set of satellite repeats, most of which remained concealed until their present investigation. Using next-generation sequencing combined with novel bioinformatics tools, we reconstructed consensus sequences of 23 novel satellite repeats representing 0.008–2.700% of the genome and mapped their distribution on chromosomes. We found that in addition to typical satellites with monomers hundreds of nucleotides long, *V. faba* contains a large number of satellite repeats with unusually long monomers (687–2033 bp), which are predominantly localized in pericentromeric regions. Using chromatin immunoprecipitation with CenH3 antibody, we revealed an extraordinary diversity of centromeric satellites, consisting of seven repeats with chromosome-specific distribution. We also found that in spite of their different nucleotide sequences, all centromeric repeats are replicated during mid-S phase, while most other satellites are replicated in the first part of late S phase, followed by a single family of FokI repeats representing the latest replicating chromatin.

Satellite DNA (satDNA) is a class of repetitive DNA characterized by its genomic organization into long arrays of tandemly arranged units called monomers. It is best distinguished from other tandemly repeated sequences by its formation of much larger arrays spanning up to megabases in length and often forming blocks of heterochromatin that appear as nuclear chromocenters and chromosomal bands. Although monomer sizes of 135–195 bp and 315–375 bp, corresponding to the length of DNA wrapped around mono- and di-nucleosome particles, were found to be predominant[1], the satellite monomers can range from lengths typical for microsatellites (2–7 bp) and minisatellites (tens of bp)[2] up to over five kilobases[3]. Except for the specific types of tandem repeats including rRNA gene arrays and telomeric motifs that have coding or structural roles[4,5], the function of satDNA in the genome is still a matter of debate. It has been proposed that satellite repeats may have a structural role in the genome[6] and that they affect expression of nearby genes by epigenetic modifications induced by specific changes in the environment[7]. Perhaps best documented is the frequent association of satellite repeats with centromeres, implying their importance for centromere determination or function[8]. On the other hand, it has been shown that neocentromeres may arise at satDNA-free regions[9], and some established centromeres may be free of satellite

[1]Biology Centre of the Czech Academy of Sciences, Institute of Plant Molecular Biology, České Budějovice, 37005, Czech Republic. [2]University of South Bohemia, Faculty of Science, České Budějovice, 37005, Czech Republic. [3]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466, Gatersleben, Stadt Seeland, Germany. Correspondence and requests for materials should be addressed to J.M. (email: macas@umbr.cas.cz)

repeats[3]. Thus, it is yet to be established whether satDNA is a key functional component of centromeric regions or whether centromeres merely provide favorable conditions for satDNA accumulation.

SatDNA belongs to the most dynamic components of eukaryotic genomes, and its high evolutionary rate results in considerable sequence diversification. Therefore, most satellite repeat families are species- or genus-specific[1]. However, precise molecular mechanisms leading to this rapid turnover and their eventual regulation in individual species have not yet been elucidated. A variety of mechanisms have been proposed to generate short arrays of tandem repeats that may provide a template for further expansion. Such mechanisms include unequal crossing over of random sequences[10], slipped-strand mispairing[11], and sequence-directed mutagenesis[12]. Tandem duplications of varying length can also result from aberrant replication and replication stress[13–15]. In addition, satellite repeat arrays were found to originate from amplification of short tandemly repeated regions present in rDNA intergenic spacers and 3′ untranslated regions of Ty3/gypsy LTR-retrotransposons[16,17]. Regardless of the primary origin of short tandemly repeated loci, it is supposed that there are additional mechanisms that mediate their expansion into long arrays and subsequent concerted evolution of monomer sequences resulting in their genome-wide homogenization[18–20]. One of the potential mechanisms mediating amplification and sequence homogenization of satellite DNA is recombination-based formation of extrachromosomal circular DNA (eccDNA) from tandem repeats which in turn could serve as a template for their rolling circle replication and subsequent re-integration of the products. Although populations of eccDNA molecules derived from satellite repeats were successfully detected in a number of plant species[21,22], the evidence for their amplification and re-integration into the genome is still missing. Other potential mechanisms of satDNA amplification include unequal chromatid exchange[10] and segmental duplication[23].

To gain better insight into the biology of satellite repeats, comprehensive analysis of sequence diversity, abundance, and homogenization of satDNA families within and between species is needed. In spite of the relatively long history of satDNA investigation, such knowledge is still limited in several ways. Recently introduced methodologies utilizing a combination of next-generation sequencing (NGS) with appropriate bioinformatics tools revealed that previously used experimental approaches suffered from relatively low sensitivity, resulting in efficient identification of only the most amplified satellite repeats with specific properties of their sequences. For example, the very discovery of satDNA was achieved by density gradient centrifugation, whereby it was revealed as satellite bands formed due to the different buoyant density of satellite repeats compared to the bulk of genomic DNA[24]. Alternatively, satellite repeats were often identified based on the presence of conserved restriction sites in their monomer sequences[25]. Consequently, the satellites lacking these features and those with small proportions in the genome were, in principle, hard to identify. On the other hand, novel sequencing technologies provide deep information about sequence composition of complex genomes of eukaryotes via generation of unprecedented amounts of sequence data. These data can then be utilized by bioinformatic pipelines specifically tailored to the identification of satellite repeats from NGS reads without the need for their assembly[26–29]. These approaches have proved to be very efficient and revealed surprising diversity of satellite repeat families in some plant and animal species[29–31].

In this work, we focus on the characterization of the satellite DNA population in the genome of *Vicia faba*, a species that has long served as a model for cytogenetic studies in plants[32–34]. Owing to its relatively large genome (1 C = 13.41 Gbp) and small chromosome number (2n = 12), *V. faba* chromosomes are large and easy to investigate with cytogenetic techniques. Consequently, a number of features like bands corresponding to different types of chromatin and epigenetic modifications have been revealed; however, only a few are associated with specific genomic sequences[35,36]. In our previous study of the repeat composition of species from the legume tribe *Fabeae*[31], *V. faba* was found to carry a large number of satellite repeats that together constituted 935 Mbp (7%) of its genome. Putative satellite repeats were identified based on the properties of cluster graphs obtained by similarity-based clustering of low-pass genome sequencing Illumina reads, as implemented in the RepeatExplorer pipeline[27]. These graphs represent the reads and their sequence similarities as nodes and connecting edges, respectively, and form globular or ring-like shapes in the case of tandem repeats. Such shapes, combined with other properties of the clusters, are reliable indicators of satellite repeats, regardless of their monomer lengths[26,28]. Over 30 putative families of satDNA were identified and partially characterized by these bioinformatics approaches[27,28], in contrast with only four satellites (FokI, pVf7, a 172 bp-subtelomeric repeat and TIII15) that had been previously reported in this species[37–40]. Here, we provide experimental verification for most of the satellite repeat families predicted by the bioinformatic analysis, their localization on chromosomes, and information about replication timing and association with different types of chromatin. Moreover, we employed ChIP-seq analysis using CenH3 antibody to identify centromeric satellites, revealing their surprising diversity.

Results

Large number and sequence diversity of satellite repeats in *V. faba*.

Clusters of NGS reads from low-pass sequencing of the *V. faba* genome that were classified as putative satellites in our previous study[31] were inspected manually, as well as by the TAREAN pipeline[28], to reconstruct consensus monomer sequences from tandem repeats. All novel satellite repeats with an abundance exceeding 0.1% of the *V. faba* genome and selected representatives of less abundant satellites (Table 1 and Supplementary Data S10) were subjected to detailed sequence analysis. This analysis focused on AT/GC content, distribution of nucleotides between complementary strands (Table 1), di- and tri-nucleotide frequencies (Supplementary Fig. S1), presence of subrepeats and detection of sequence similarities. In addition, distribution of all selected satellites in the genome was studied by fluorescence *in situ* hybridization (FISH) on mitotic chromosomes. The selected families differed in their nucleotide sequences, and their genomic abundance ranged from 0.008 up to 2.72% of the genome, corresponding to a physical size between 1.1 and 365.1 Mb/1 C. Besides repeat families with a typical monomer length of hundreds of base pairs, there were some with substantially smaller monomers and an unexpectedly large number of 17 families with long monomers ranging from 687 up to 2033 bp (Table 1). The majority of the satellite sequences had an

		Genomic abundance		Sequence characteristics					
Satellite	**monomer [bp]**	**[%]**	**[Mbp/1 C]**	**% AT**	**max A/T**	**max C/G**	**max Pu/Py**	**ChIP-seq (centromere)**	**Notes**
VfSat1	191	2.723	365.1	75.9	1.07	1.88	1.22	1.1	[Vf_TA_11]; similarity to TR-9 from *P. sativum*
FokI	59 / 57	2.322	311.4	59.3	1.44	2.00	1.11	1.0	[37]
VfSat2	~26	1.292	173.2	71.4	2.60	5.00	2.00	0.4	
VfSat3	702	0.329	44.1	68.1	1.24	1.13	1.21	0.3	[Vf_TA_39]
TIII15	58	0.222	29.8	53.4	1.39	2.86	1.90	1.9	[40]
VfSat4	38	0.199	26.7	73.7	1.80	2.33	1.92	0.6	similarity to VicTR-B
VfSat5	687	0.187	25.1	77.6	1.16	1.19	1.07	0.3	
pVf7	169	0.182	24.4	55	1.07	1.00	1.04	0.3	[38]
VfSat6	50	0.132	17.6	64	1.91	1.57	1.78	**103.6 (CEN1)**	similarity to TR-5 from *P. sativum*
VfSat7	44	0.102	13.7	70.5	1.21	1.17	1.2	**103.2 (CEN1)**	
VfSat8	2033	0.061	8.1	71.7	1.68	1.03	1.46	**91.3 (CEN4)**	
VfSat9	963	0.055	7.4	77.6	1.02	1.77	1.00	0.2	
VfSat10	1762	0.042	5.7	74.7	1.08	1.27	1.00	**41.2 (CEN1)**	
VfSat11	1619	0.040	5.3	75.7	1.08	1.23	1.01	0.3	
VfSat12	1004	0.038	5.1	74.2	1.12	1.27	1.02	0.2	
VfSat13	47	0.036	4.8	68.1	2.56	1.14	1.76	**149.2 (CEN5)**	
VfSat14	888	0.035	4.7	75.6	1.06	1.33	1.02	0.3	
VfSat15	942	0.035	4.7	76.5	1.11	1.03	1.08	0.3	similarity to TR-20 from *P. sativum*
VfSat16	1712	0.038	5.1	65.1	1.20	1.00	1.27	**109.9 (CEN6)**	[Vf_TA_157]
VfSat17	781	0.031	4.2	75	1.11	1.10	1.06	0.4	
VfSat18	1172	0.031	4.2	79.7	1.11	1.03	1.08	0.3	
VfSat19	1345	0.024	3.3	80.2	1.01	1.35	1.07	0.2	
VfSat20	924	0.022	3	74.5	1.23	1.29	1.24	0.7	
VfSat21	1057	0.017	2.3	74.3	1.02	1.35	1.10	0.2	
VfSat22	1834	0.016	2.2	71.8	1.25	1.22	1.11	0.3	
VfSat23	1325	0.008	1.1	73.3	1.17	1.41	1.18	**81.9 (CEN2)**	

Table 1. Satellite repeats investigated in this study. Novel satellite repeats are numbered with the prefix "VfSat"; references to previously described repeats are given in Notes. Names in square brackets refer to homologous repeats that were partially characterized by Novák *et al.*[28]. The column "ChIP-seq" provides ChIP/input ratios; the values of significant enrichment are highlighted and supplemented with repeat localization determined by FISH. Sequences of all newly identified satellites are provided in Supplementary Data S10.

elevated AT content (65–80%), and some were found to have asymmetrical distributions of A/T, C/G, or purine/pyrimidine bases between complementary strands (Table 1).

There were ten satellite repeats whose abundance exceeded 0.1% of the genome. They included the previously described repeats FokI, TIII15, and pVf7 and seven novel families (Table 1). One of them, VfSat1, was estimated to be of similar abundance to FokI, which is one of the most abundant satellites found in a plant species so far. Contrary to FokI, which is located in a number of bands within the long arms of all five acrocentric chromosomes, FISH with VfSat1 probe produced one major band near the centromere within the satellite arm of the metacentric chromosome 1 and additional minor signals in the pericentromeric region of all acrocentrics (Fig. 1a). VfSat1 was found to share 78% sequence similarity with the pea (*Pisum sativum*) satellite TR-9 (Supplementary Fig. S2), which occurs in terminal regions of three pairs of pea chromosomes[41]. The third most abundant satellite, VfSat2, had a prevailing monomer sequence TATTTGAC(GTT)6, which probably originated from a degenerated simple sequence repeat, $(GTT)_n$. Due to its simple sequence, this satellite showed high strand asymmetry values (Table 1). VfSat2 produced FISH signals that were partially co-localized or interlaced with FokI repeats except for an additional band adjacent to a heterochromatic DAPI-positive segment on chromosome 1 (Figs 1b and 2). There was another highly abundant satellite, VfSat3, with similar distribution to FokI, which also occurred in the same additional locus on chromosome 1 as VfSat2, but in this case its signal matched the position of the DAPI-positive band (Fig. 1c). Considering the presence of additional three satellites (pVf7, VfSat9, VfSat19; Fig. 2), this region of chromosome 1 together with heterochromatic loci within long arms of acrocentric chromosomes can be considered a hotspot of satellite DNA accumulation.

Satellites with large monomers are predominantly located in pericentric regions. A substantial fraction of putative satellite repeats (17 out of 26) had estimated monomer sizes between 687 and 2033 bp, thus being significantly larger than the previously reported preferred monomer length of 135–375 bp[1]. To validate the predicted monomer sequences and confirm their tandem arrangement, PCR was performed with *V. faba* genomic DNA as a template and with primers designed to face outwards from the reconstructed monomer

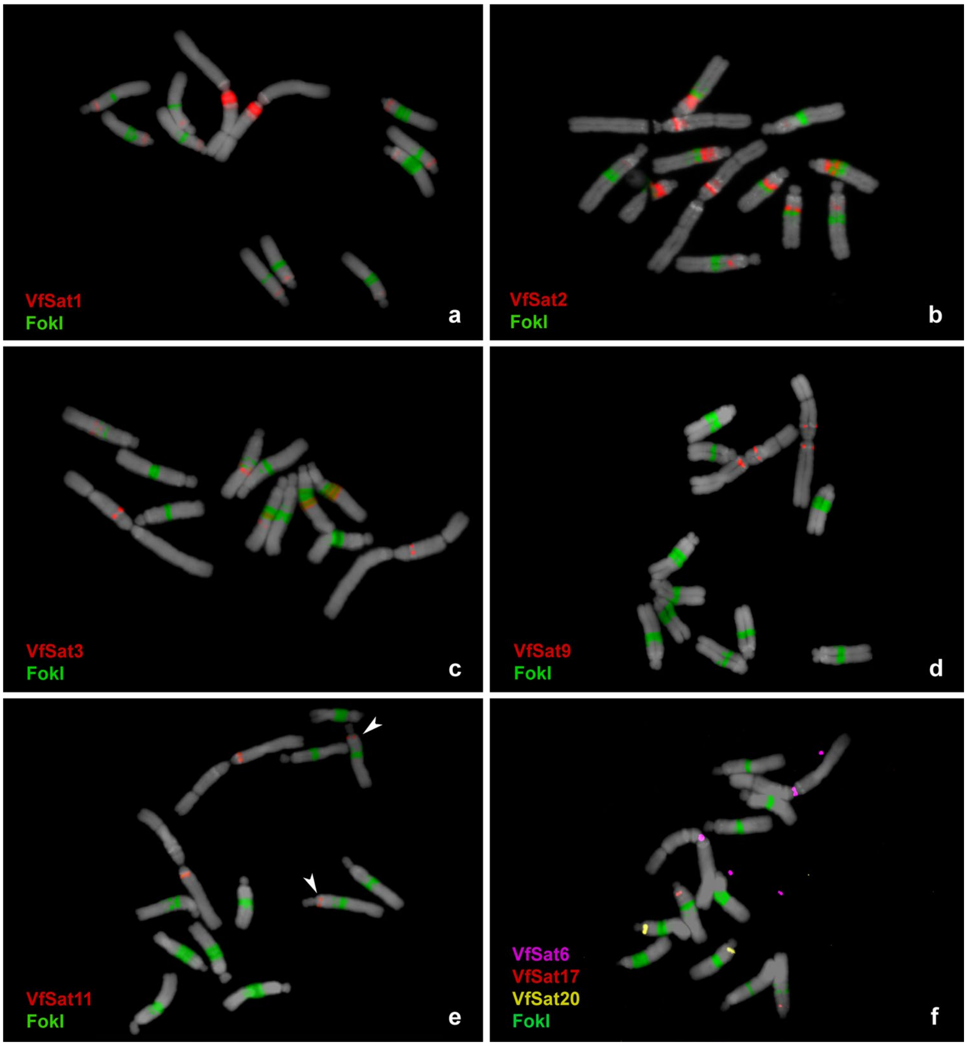


Figure 1. Distribution of selected satellite repeat families on metaphase chromosomes of *Vicia faba*. Satellites were visualized using multi-color FISH, with individual probes labeled as indicated by color-coded descriptions. Hybridization patterns of FokI repeats (green signals) were used for chromosome discrimination as shown in Supplementary Fig. S3. Chromosomes counterstained with DAPI are shown in gray. Arrowheads in (**e**) point to polymorphic VfSat11 signals on chromosome 2 (see Fig. 3b for comparison).

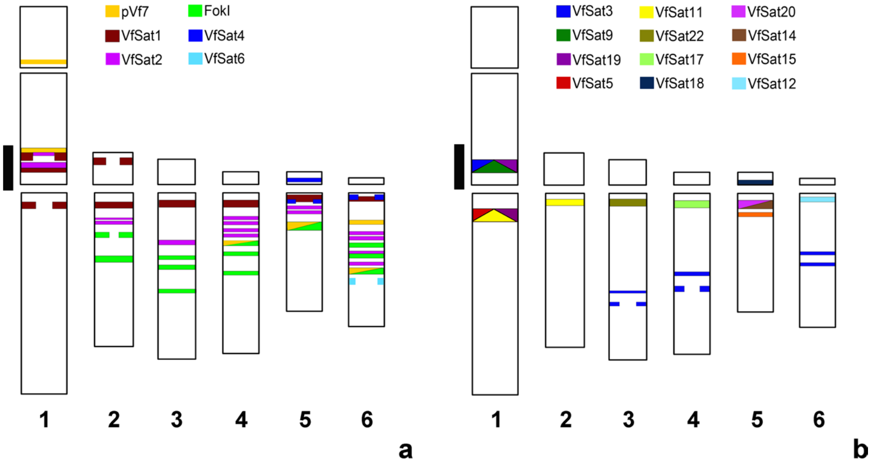


Figure 2. Schematic representation of genomic distributions of all non-centromeric satellites mapped by FISH. Satellites with short monomers are shown in (**a**), while those with long monomers exceeding 600 bp are in panel (**b**). The black line along chromosome 1 marks the region of accumulation of multiple satellite repeats.

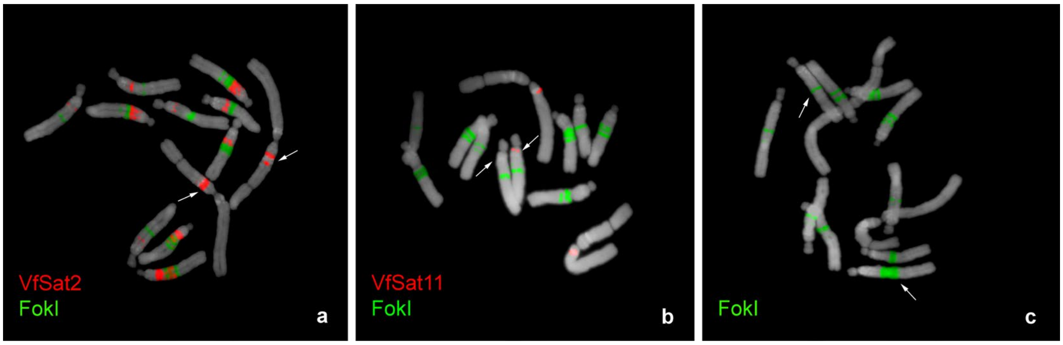


Figure 3. Variation in the number (**a**,**b**) or size (**c**) of FISH signals between homologous chromosomes detected for VfSat2, VfSat11, and FokI. Arrows indicate the positions of polymorphic loci on homologous chromosomes.

consensus. In this arrangement, the amplification can take place only between the primer pairs located in adjacent tandemly repeated monomers (Supplementary Fig. S4). All 17 putative repeats tested using this assay produced the expected amplification products, and their cloned sequences matched the predicted consensus with 92–99% similarity. Dot-plot analysis of the monomer sequences did not reveal any internal subrepeats that could explain their large size as a result of evolution via higher-order repeat intermediates. The only exception was VfSat8 which displayed irregular internal subrepeats (Supplementary Fig. S5). The analysis did not detect any significant similarities between different satellite repeats.

Selected clones were labeled and used as probes for FISH. Most of these long monomer satellites produced only a single hybridization signal, except for VfSat9 (one signal on each arm of chromosome 1, Fig. 1d) and VfSat11, which, in addition to labeling one locus on the long arm of chromosome 1, produced a hemizygous signal on the long arm of chromosome 2 (Figs 1e and 3b). Although the FISH signals of long monomer satellites occurred on different chromosomes, they were mostly located in similar positions within their long arms, close to the centromeres (Fig. 2b). Four of the repeats were located within primary constrictions, and further analysis confirmed that they represented centromeric satellites (see below).

Centromeric repeat composition differs between *V. faba* chromosomes. The association of repetitive sequences with centromeric chromatin was investigated via chromatin immunoprecipitation using the antibody against the centromeric histone H3 variant CenH3, followed by Illumina sequencing of retrieved DNA (ChIP-seq). The resulting reads were mapped to repeat clusters based on their sequence similarities, as were the reads obtained by sequencing DNA fragments extracted from chromatin preparations prior to ChIP (input control). A total of 21.3 million ChIP and 10.7 million input reads were mapped to reference clusters, and normalized ratios of ChIP to input reads were evaluated for the 500 largest clusters representing highly and moderately repeated sequences with genomic proportions of at least 0.002%. There were seven clusters that showed elevated ratios of ChIP/input reads (41- to 149-fold enrichment), indicative of their association with centromeric chromatin, whereas all other analyzed clusters showed ratios close to or below 1 (Table 1). All ChIP-enriched sequences represented satellite repeats, and their centromeric location was confirmed by FISH (Fig. 4). The seven satellites differed substantially in their monomer length (44–2033 bp), sequence composition, and distribution on chromosomes. Whereas three different satellites were found at the centromere of chromosome 1 (CEN1), four other centromeres contained a single chromosome-specific satellite, and no centromeric repeat was identified for chromosome 3. One of the CEN1 satellites, VfSat6, was found additionally at a non-centromeric locus on the long arm of chromosome 6. However, corresponding FISH signals were very weak and detectable only on a fraction of chromosomes, most likely due to the small size of these loci. No sequence similarities to repeats from other species were detected for *V. faba* centromeric satellites except for VfSat6, which was found to be related to the satellite TR-5 (88% similarity, Supplementary Fig. S2) located in the pericentromeric region of chromosome 2 of *Pisum sativum*[41].

Three satellites display supernumerary FISH signals or signal size polymorphisms between homologous chromosomes. FISH with the satellite sequences VfSat2, VfSat11, and FokI revealed differences between homologous chromosomes regarding the number of labeled loci or sizes of some signals (Fig. 3). These polymorphisms included VfSat2 loci on chromosome 1, which appeared as either two bands (stronger and weaker) separated by a gap of non-labeled chromatin or as two closely adjacent bands of equal strength (Fig. 3a). The observed genotypes were either homozygous for the former pattern or heterozygous. The other polymorphic site was the locus of VfSat11 on the long arm of chromosome 2, which was missing from one of the homologs in part of the examined individuals (Figs 1e and 3b). In the case of the FokI repeat, an expansion of the signal size was revealed at one locus of the acrocentric chromosome 5, which was paralleled by the size change of a corresponding DAPI-positive band and by an increase in chromosome size (Fig. 3c). The expanded band was observed only in heterozygous configuration, while a part of the genotypes was homozygous for the smaller variant of this FokI band.

Satellite repeats vary as to their replication time during S phase. The replication time of satDNAs was investigated by incorporating the thymidine analog EdU, employing 15 or 30 min pulses of exposure to EdU

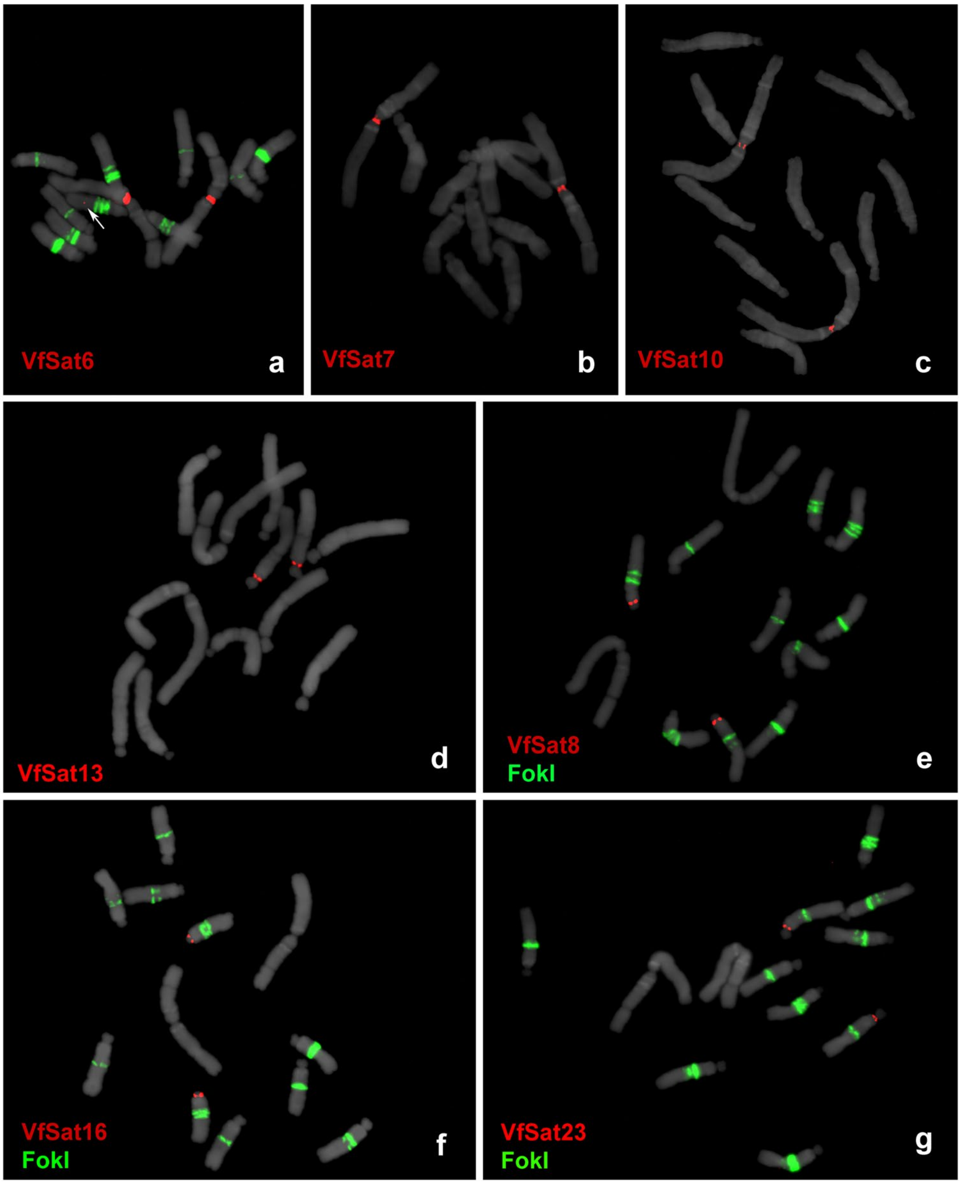


Figure 4. FISH localization of centromeric satellites. Three satellites located in the centromere of chromosome 1 (VfSat6, VfSat7 and VfSat10) are shown separately in panels a–c. Centromeres of four other chromosomes that each contain a single satellite repeat are labeled in panels d–g: (**d**) VfSat13 repeat located in centromere 5, (**e**) VfSat8 in centromere 4, (**f**) VfSat16 in centromere 6, and (**g**) VfSat23 in centromere 2. Arrow in (**a**) points to minor non-centromeric locus of VfSat6 which is detectable only on a fraction of chromosomes. Hybridization patterns of FokI repeats (green signals) were used for chromosome discrimination in (**e–g**). Chromosomes counterstained with DAPI are shown in gray.

followed by fixation of root meristems at various times after the pulse (1–9 h). Depending on the time elapsed since the labeling pulse, the fixed material displayed labeling of early-, middle-, or late-replicating chromatin. Examples of labeled chromosomes are shown in Fig. 5, and the observed labeling patterns are summarized in Supplementary Table S7.

The early replication pattern consisted of weak dispersed labeling with strongly labeled NORs and a few additional bands (Fig. 5a). The early replication pattern was gradually replaced with more uniform staining of whole chromosomes except for their heterochromatic regions. At this stage, corresponding to mid-S phase, there were also distinguishably brighter signals observed at all centromeres (Fig. 5b). The late S phase labeling signals appeared as sharp bands corresponding to most of the satellite repeat loci (Fig. 5c,d). Replication timing of individual satellite repeats was determined by associating their known chromosomal positions with the observed replication patterns and in some cases also by performing their FISH detection on EdU-labeled chromosomes. Taken together, these experiments revealed that the replication timing of *V. faba* satellites is not uniform (Supplementary Table S7). Mid-S phase replicating centromeric repeats and VfSat2 preceded most other satellite families, which replicate in late S phase. All these satellites except for FokI finished their replication before

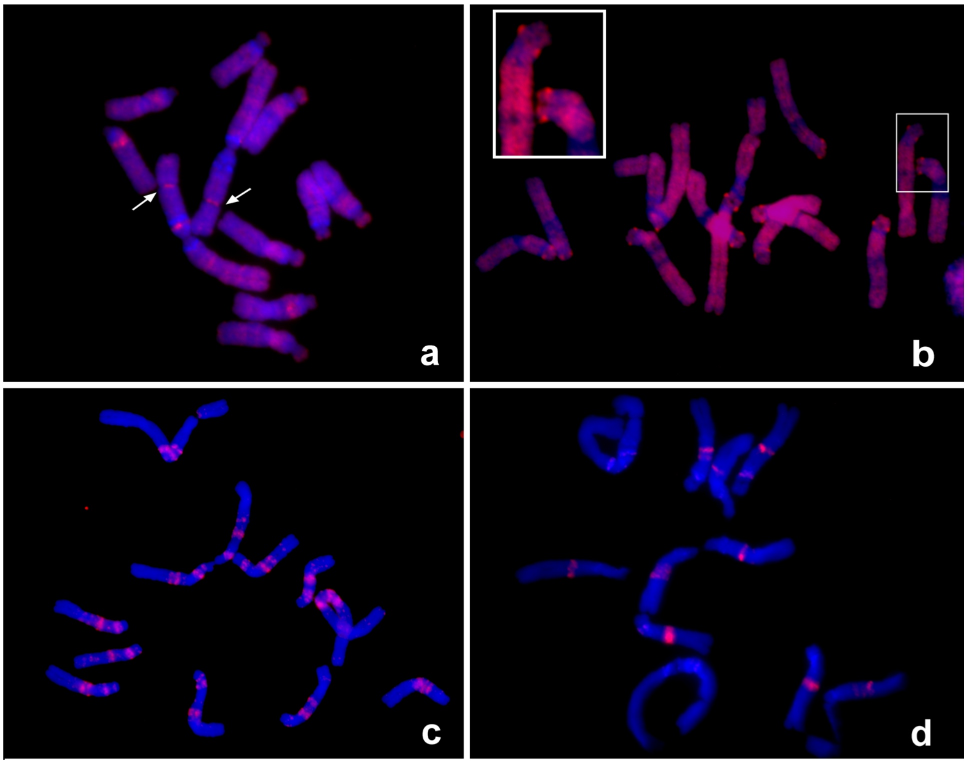


Figure 5. DNA replication assay. Examples of EdU labeling of early (**a**), mid (**b**) and late (**c–d**) replicated chromatin. Arrows in panel (a) show the positions of early replicating NORs. The inset in panel (b) shows a detail of two chromosomes, with bright spots corresponding to labeled centromeres. The two late replication patterns that could be distinguished consisted of labeling most satellite repeats (**c**), followed by exclusive labeling of FokI loci, which represented the last genomic sequences to be replicated (**d**).

the end of S phase, while the FokI sequences alone represented the latest replicating part of the genome (Fig. 5d). The EdU-labeling patterns were correlated with the presence of specific satellite sequences, as exemplified by the FokI bands that were polymorphic between homologous chromosomes 5 and consequently displayed EdU labeling of different intensities corresponding to the sizes of FokI bands. In addition, the earlier replication of VfSat2 was maintained on all genomic loci of this satellite, even those adjacent to the blocks of FokI that were the latest genomic sequences to be replicated (Supplementary Fig. S6). Thus, the replication timing appeared to be sequence dependent, not determined by the chromosomal position. We investigated whether there is a correlation between various characteristics of the nucleotide sequences (monomer length, AT/GC and di- and tri-nucelotide composition) and the replication timing of individual satellites; however, no statistically significant correlation was found.

Discussion

Satellite DNA still represents one of the most enigmatic components of eukaryotic genomes, which is in part due to the technical difficulties associated with reliable characterization of a representative set of satellite repeats from the genomes of interest. Here we demonstrated that such characterization can be achieved by employing next-generation sequencing combined with bioinformatics tools specifically tailored to this task. Application of this approach to the genome of *V. faba* generated a large body of new sequence and cytogenetic information surpassing the evidence that had been gathered so far about satellite DNA in this long used cytogenetic model. The reliability of bioinformatic identification and reconstruction of satellite repeats from NGS data was confirmed experimentally by successful FISH detection of all 25 selected repeats on *V. faba* chromosomes. The same strategy was recently used in several plant and animal species, where it revealed surprising diversity of satellite repeats, similar to that reported here for *V. faba*[29,30,42]. However, such satDNA diversity is not a common feature of all genomes because there were also species relatively poor in satellite DNA reported after examination using similar approaches[43], including several species of *Vicia*[31].

Satellite repeats usually show little or no sequence conservation between different taxa, owing to their rapid evolutionary turnover in the genome[6]. However, great sequence diversity can also be found between the repeats within a single species, as demonstrated here for *V. faba*. Except for their preference for AT-rich sequences, the families of *V. faba* satDNA did not show any conserved features or sequence similarities, pointing to their independent origin. Despite the wealth of repeat sequencing data from closely related genera of *Vicia*, *Lathyrus*, *Pisum*, and *Lens*[31], the majority of *V. faba* satellites did not show sequence similarities to repeats from other species. This suggests their species-specific origin or rapid sequence diversification. The exceptions included three satellites (VfSat1, VfSat6, and VfSat15) with similarities to satellite repeats from *Pisum sativum*, and VfSat4, which was related to VicTR-B repeats highly amplified in several *Vicia* species[44] (Table 1 and Supplementary Fig. S2). The existence of rapid turnover of satellite repeats in the *V. faba* genome was also supported by the occurrence of polymorphic or supernumerary loci of three satellite repeats, VfSat2, VfSat11, and FokI. This is in agreement with earlier reports of variability of heterochromatic Giemsa bands at chromosomal positions corresponding to FokI repeats observed within and between different accessions and karyotypes[32,35].

Satellites with long monomers (0.7–2.0 kb) were found to be surprisingly numerous in the *V. faba* genome. This contrasts with relatively few cases of such repeats reported to date in other plants. Several satellites with monomers ranging from 0.9 to 4.0 kb were found to accumulate on the B chromosomes of rye (*Secale cereale*)[45–47]. Long monomer satellites were also reported in *Solanum* species, including the Sobo satellite (4.7 kb monomer) of *S. bulbocastanum*[48] and a diverse group of centromeric satellites with monomers up to 5.4 kb from *S. tuberosum*[3] and *S. verrucosum*[49]. Most of these satellite sequences display similarities to various retrotransposons or have a complex structure indicating their origin from different genomic repeats[48,49]. These features indicate that the rye and *Solanum* long monomer satellites might be evolutionary young and mostly originate at specific genomic regions represented by dispensable B chromosomes and centromeres. However, none of the repeats reported here for *V. faba* had detectable sequence similarity to other genomic sequences, and only a small fraction was located in centromeres. On the other hand, most *V. faba* long monomer satellites displayed a preference for pericentric regions, the significance of which is yet to be investigated.

The four satellite repeats with long monomers (1.7–2.0 kb) that were localized in primary constrictions of metaphase chromosomes were proved to be associated with centromeric chromatin using ChIP-seq with the CenH3 antibody. However, the long monomer size was not a universal feature of *V. faba* centromeric repeats, as the other three centromeric satellites had extremely short monomers, ranging from 44 to 50 bp. There is mounting evidence that most eukaryotic centromeres are determined epigenetically, independent of the underlying DNA sequence[50,51]. The frequent accumulation of satDNA in centromeric regions is then explained by its positive role in stabilizing centromeres, promoting deposition of inner kinetochore proteins, or simply by passive accumulation due to the absence of recombination-based elimination mechanisms[8,52,53]. Most higher plant species investigated so far have a single or only a few centromeric satellites with monomers hundreds of nucleotides long that are shared by all chromosomes[54], an observation that is explained by their coevolution with kinetochore proteins[55]. Thus, the monomer length diversity as well as the overall number of different centromeric satellites, all of which are specific to a single *V. faba* chromosome, are unusual. Similar diversity has been reported in potato (*Solanum tuberosum*), where part of the centromeres contains chromosome-specific satellites, while the rest is free of satellite DNA[3]. Comparison of homeologous centromeres between potato and its wild relative *S. verrucosum* revealed that all but one of the centromeric satellites differ in their sequences, which, together with the absence of centromeric satellites on some chromosomes, indicates that centromeres in these species are evolutionarily novel and still undergo rapid cycles of satellite DNA expansion and contraction/elimination that precede fixation of a single satellite in most centromeres[49]. Even higher intra-specific diversity of centromeric satellites has been described in the pea (*Pisum sativum*), which was found to carry 13 sequence families differing in their genomic abundance and distribution on chromosomes[41]. The genus *Pisum* is closely related to *Vicia*, but chromosomes in *Pisum* and its sister genus *Lathyrus* exhibit a unique morphology of their centromeres, which are composed of multiple separated CenH3 loci arranged along extended primary constrictions[56]. While it was tempting to speculate that the extraordinary diversity of pea satellites originated from the evolutionary shift to its complex centromere structure, the diverse repeat composition of a simple *V. faba* centromere calls for the investigation of additional species from both genera to get more representative insight into evolution of their centromeres.

The replication of most *V. faba* satellites during the late and the centromeric repeats in mid-S phase is in agreement with observations from other plant species[45,57]. Our data also suggest that, although the replication patterns were found to be conserved for specific satellite sequences regardless of their chromosomal location, the nucleotide sequences alone are not likely determinant of the replication timing, as we did not identify any conserved sequence features among synchronously replicating satellite repeats. Thus, other features like epigenetic modifications of chromatin proteins that were found to correlate with replication timing may be more important[57,58]. Indeed, FokI repeats were previously shown to be distinguishable from the rest of the *V. faba* genome by a specific combination of epigenetic marks[36] which could also explain their partially different replication timing compared to other satellites.

The data described in the present work allow for reexamination of previously described cytogenetic features of *V. faba* chromosomes to investigate their eventual correlation with sequence composition and chromosomal distribution of novel satellite repeats. For example, previous studies revealed that the region located proximally on the NOR-bearing arm of chromosome 1 exhibits extreme sensitivity regarding misrepair of DNA damage, in particular after exposure to the mutagens mitomycin C and maleic hydracid, but also to other genotoxins[59,60]. This S phase-dependent misrepair yielded a high frequency of chromatid-type structural aberrations, such as isochromatid breaks, interstitial deletions, duplication deletions, and reciprocal translocations. The clustering of aberration breakpoints was likely due to misrepair of DNA double-strand breaks, which arise from the overlap of excision repair with replication during S phase. The misrepair results from the ligation of the wrong strand ends, favored by sequence homology with the break ends and/or by strand discontinuities at the border between regions with different replication timing[61]. The present study discovered that this aberration hotspot correlates with a region of unusual clustering of diverse satellite repeats, including five repeat families (VfSat1, pVf7, VfSat3, VfSat9, and VfSat19) mapped to two adjacent DAPI- and Giemsa-positive chromatin bands and one (VfSat2) located between them in a DAPI-negative area (Fig. 2). Also, other regions of frequent aberration breakpoints were reported. These were associated with FokI repeats in the middle of the long arms of all acrocentrics, in particular the region on the long arm of chromosome 5 with a largely expanded FokI region[62]. Many FokI loci were found adjacent to or interspersed with VfSat2, the satellite with the most contrasting replication pattern compared to FokI, supporting the idea that such adjacent loci with different replication timing may cause chromosomal instability. Furthermore, due to its chromosomal location, its abundance, and its strong asymmetry of A and T between both strands of the double helix, the non-centromeric VfSat2 may represent the physical basis for the asymmetric bands previously observed by the fluorescence-plus-Giemsa technique after BrdU incorporation for one S phase[34].

Methods

Plant material and DNA isolation.

Seeds of the *Vicia faba* cultivar Merkur were purchased from Osiva Boršov (Boršov nad Vltavou, Czech Republic). Total genomic DNA was isolated from young leaves as described by Dellaporta *et al.*[63].

Sequence analysis and cloning of satellite repeats.

Putative satellite repeats were identified in the course of our previous study[31] via graph-based clustering of *V. faba* genomic shotgun reads using the RepeatExplorer pipeline[27]. Reconstruction of monomer sequences of selected satellites was performed using TAREAN[28]. Reconstructed sequences were used to design oligonucleotide probes for hybridization (Supplementary Table S8) or PCR primers for amplification and cloning of corresponding repeats from genomic DNA (Supplementary Table S9). The latter option was used for satellites with long monomers that could not be efficiently detected using short oligonucleotide probes. The PCR reactions were performed in 30 μL volume containing 1 × PCR buffer, 0.2 mM dNTPs, 0.2 μM primers, and 2U of Platinum Taq Polymerase (Invitrogen). The amplification was carried out for 30 cycles of 94 °C for 1 min, 55 °C for 1 min, and 72 °C for 3 min. The amplicons were cloned using the TOPO-TA Cloning Kit for Sequencing (Invitrogen). Plasmid clones were verified by sequencing, and selected inserts were used as probes for *in situ* hybridization experiments. Sequences of all cloned probes were deposited in GenBank under accession numbers MF796528-MF796546.

Over- and underrepresentation of di- and tri-nucleotides in satellite repeats was calculated from unassembled sequence reads according to Burge *et al.*[64]. Correlation between sequence composition and replication timing of satellite repeats was tested using regression analysis in an R programming environment.

Fluorescence *in situ* hybridization (FISH).

Mitotic chromosomes were prepared from root tip meristems synchronized using 2.5 mM hydroxyurea and 2.5 μM amiprophos-methyl as described previously[56,65]. Probes were labeled with Alexa Fluor 568 or Alexa Fluor 488 (Thermo Fisher Scientific, Waltham, MA, USA) using nick translation[66]. The oligo-probes were labeled with biotin or fluorescein at their 5′ ends during synthesis (Integrated DNA Technologies, Leuven, Belgium). FISH was performed according to Macas *et al.*[67] with hybridization and washing temperatures adjusted to account for AT/GC content and hybridization stringency 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), 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).

Identification of centromeric repeats using chromatin immunoprecipitation.

Chromatin immunoprecipitation was performed with nuclei isolated from fresh leaves as described previously[41] using a custom-made antibody raised against a peptide designed according to the CenH3 protein sequence identified in *Vicia faba* (CenH3-2_VF)[56]. ChIPed DNA and input DNA control were sequenced on the Illumina platform (Global Biologics, LLC, Columbia, USA) in a single-end, 101 nt read mode. The resulting reads were trimmed to 100 nt by removing the first base and quality filtered to exceed the cutoff quality score of 10 over at least 95 nucleotides. Quality-filtered reads were mapped to reference contigs assembled from clusters of genome shotgun sequencing reads representing *V. faba* repetitive sequences produced and characterized in our previous work[31]. Similarity-based mapping of reads to repeat contigs was done using BLASTn[68] with the parameters "-m 8 -b 1 -e 1e-20 -W 9 -r 2 -q -3 -G 5 -E 2 -F F" and was followed by output parsing to ensure that each read was mapped to a maximum of one repeat cluster with the highest similarity score. The proportion of ChIP and input reads mapped to individual clusters was evaluated to identify repeats with a ChIP/input ratio >10, which were considered to represent repeats enriched in the ChIP sample.

DNA replication assay.

Root tip meristems were treated with thymidine analog 5-Ethynyl-2′-deoxyuridine (EdU) by submersing the roots of three-day-old seedlings for 15 or 30 min in a 10 μM EdU solution in Hoagland medium. The EdU treatment was followed by incubating the seedlings in Hoagland medium for various time intervals ranging from 1 to 9 hours (all incubations were done at 25 °C). Since *V. faba* S phase was reported to last 7.5 h, followed by 5 h of G2 phase before entering mitosis[69], this time sampling enabled observation of metaphase chromosomes with their DNA labeled at various stages of the S phase (late replicating chromatin was labeled in samples collected 1–3 h after EdU treatment while collecting tissues after 6–9 h provided information about early replicating chromatin). Tissue fixation was performed with methanol-acetic acid (3:1) and chromosome preparations were done as described above for FISH experiments. EdU detection was performed using EdU HTS kit (BaseClick GmbH, Neuried, Germany) according to the manufacturer's protocol, except for the washing procedure, which was done at 35 °C in 2 × saline sodium-citrate (SSC) buffer for 5 min, followed by 5 min in 50% formamide/2 × SSC, 10 min in 2 × SSC, and finally 5 min in 1 × BT buffer (0.1 M $NaHCO_3$, 0.05% Tween 20, pH 8.0) at room temperature.

Accession codes.

Cloned sequences of satellite repeats were deposited in GenBank under accession numbers MF796528-MF796546. Raw Illumina reads from ChIP-seq experiment will be available from European Nucleotide Archive under the study PRJEB5241.

References

1. Macas, J., Mészáros, T. & Nouzová, M. PlantSat: a specialized database for plant satellite repeats. *Bioinformatics* **18**, 28–35 (2002).
2. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).
3. Gong, Z. *et al*. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell* **24**, 3559–74 (2012).
4. Garcia, S., Garnatje, T. & Kovařík, A. Plant rDNA database: ribosomal DNA loci information goes online. *Chromosoma* **121**, 389–394 (2012).
5. Chan, S. R. W. L. & Blackburn, E. H. Telomeres and telomerase. *Philos. Trans. R. Soc. B Biol. Sci.* **359**, 109–122 (2004).
6. Garrido-Ramos, M. A. Satellite DNA in plants: more than just rubbish. *Cytogenet. Genome Res.* **146**, 153–70 (2015).
7. Feliciello, I., Akrap, I. & Ugarković, Đ. Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. *PLoS Genet.* **11**, e1005466 (2015).
8. Plohl, M., Meštrović, N. & Mravinac, B. Centromere identity from the DNA point of view. *Chromosoma* **123**, 313–325 (2014).
9. Tolomeo, D. *et al*. Epigenetic origin of evolutionary novel centromeres. *Sci. Rep.* **7**, 41980 (2017).
10. Smith, G. P. Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528–535 (1976).
11. Levinson, G. & Gutman, G. a. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221 (1987).
12. Fieldhouse, D. & Golding, B. A source of small repeats in genomic DNA. *Genetics* **129**, 563–572 (1991).
13. Nikolov, I. & Taddei, A. Linking replication stress with heterochromatin formation. *Chromosoma* **125**, 523–533 (2016).
14. Mazurczyk, M. & Rybaczek, D. Replication and re-replication: Different implications of the same mechanism. *Biochimie* **108**, 25–32 (2015).
15. Kuzminov, A. Chromosomal replication complexity: a novel DNA metrics and genome instability factor. *PLoS Genet.* **12**, e1006229 (2016).
16. Macas, J., Navrátilová, A. & Mészáros, T. Sequence subfamilies of satellite repeats related to rDNA intergenic spacer are differentially amplified on *Vicia sativa* chromosomes. *Chromosoma* **112**, 152–8 (2003).
17. Macas, J., Koblízková, A., Navrátilová, A. & Neumann, P. Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. *Gene* **448**, 198–206 (2009).
18. Dover, G. Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–7 (1982).
19. Kuhn, G. C. S., Heinrich, K., Moreira-Filho, O. & Heslop-Harrison, J. S. The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. *Mol. Biol. Evol.* **29**, 7–11 (2012).
20. Liao, D. Concerted evolution: molecular mechanism and biological implications. *Am. J. Hum. Genet.* **64**, 24–30 (1999).
21. Cohen, S., Houben, A. & Segal, D. Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants. *Plant J.* **53**, 1027–1034 (2008).
22. Navrátilová, A., Koblížková, A. & Macas, J. Survey of extrachromosomal circular DNA derived from plant satellite repeats. *BMC Plant Biol.* **8**, 90 (2008).
23. Ma, J. & Jackson, S. A. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.* **16**, 251–259 (2006).
24. Kit, S. Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *J. Mol. Biol.* **3**, 711–716 (1961).
25. Hemleben, V., Kovařík, A., Torres-Ruiz, R. A., Volkov, R. A. & Beridze, T. Plant highly repeated satellite DNA: Molecular evolution, distribution and use for identification of hybrids. *Syst. Biodivers.* **5**, 277–289 (2007).
26. Novák, P., Neumann, P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11**, 378 (2010).
27. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
28. Novák, P. *et al*. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **45**, e111 (2017).
29. Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J. & Camacho, J. P. M. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci. Rep.* **6**, 28333 (2016).
30. Heckmann, S. *et al*. The holocentric species *Luzula elegans* shows interplay between centromere and large-scale genome organization. *Plant J.* **73**, 555–565 (2013).
31. Macas, J. *et al*. 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 (2015).
32. Döbel, P., Schubert, I. & Rieger, R. Distribution of heterochromatin in a reconstructed karyotype of *Vicia faba* as identified by banding- and DNA-late replication patterns. *Chromosoma* **69**, 193–209 (1978).
33. Fuchs, J., Pich, U., Meister, A. & Schubert, I. Differentiation of field bean heterochromatin by*in situ* hybridization with a repeated FokI sequence. *Chromosom. Res.* **2**, 25–28 (1994).
34. Schubert, I. & Rieger, R. Asymmetric banding of *Vicia faba* chromosomes after BrdU incorporation. *Chromosoma* **70**, 385–391 (1979).
35. 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. *Chromosom. Res.* **6**, 219–30 (1998).
36. Fuchs, J. & Schubert, I. Chromosomal distribution and functional interpretation of epigenetic histone marks in plants. In *Plant Cytogenetics* (eds. Bass, H. W. & Birchler, J. A.) 231–253 (Springer Science + Business Media, LLC, 2012).
37. Kato, A., Yakura, K. & Tanifuji, S. Sequence analysis of *Vicia faba* repeated DNA, the FokI repeat element. *Nucleic Acids Res.* **12**, 6415–6426 (1984).
38. Maggini, F. *et al*. Structure and chromosomal localization of DNA sequences related to ribosomal subrepeats in *Vicia faba*. *Chromosoma* **100**, 229–234 (1991).
39. Houben, A., Brandes, A., Pich, U., Manteuffel, R. & Schubert, I. Molecular-cytogenetic characterization of a higher plant centromere/kinetochore complex. *Theor. Appl. Genet.* **93**, 477–484 (1996).
40. Nouzová, M. *et al*. Cloning and characterization of new repetitive sequences in field bean (*Vicia faba* L.). *Ann. Bot.* **83**, 535–541 (1999).
41. Neumann, P. *et al*. Stretching the rules: monocentric chromosomes with multiple centromere domains. *PLoS Genet.* **8**, e1002777 (2012).
42. Puterova, J. *et al*. Satellite DNA and transposable elements in Seabuckthorn (*Hippophae rhamnoides*), a dioecious plant with small Y and large x chromosomes. *Genome Biol. Evol.* **9**, 197–212 (2017).
43. Emadzade, K. *et al*. Differential amplification of satellite PaB6 in chromosomally hypervariable *Prospero autumnale* complex (Hyacinthaceae). *Ann. Bot.* **114**, 1597–608 (2014).
44. Macas, J., Požárková, D., Navrátilová, A., Nouzová, M. & Neumann, P. Two new families of tandem repeats isolated from genus *Vicia* using genomic self-priming PCR. *Mol. Gen. Genet.* **263**, 741–51 (2000).
45. Klemme, S. *et al*. High-copy sequences reveal distinct evolution of the rye B chromosome. *New Phytol.* **199**, 550–558 (2013).
46. Langdon, T. *et al*. De novo evolution of satellite DNA on the rye B chromosome. *Genetics* **154**, 869–84 (2000).
47. Martis, M. M. *et al*. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc. Natl. Acad. Sci. USA* **109**, 13343–13346 (2012).

48. Tek, A. L., 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* **170**, 1231–8 (2005).
49. Zhang, H. *et al.* Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* species: rapid evolution of DNA sequences associated with centromeres. *Plant Cell* **26**, 1436–1447 (2014).
50. McKinley, K. L. & Cheeseman, I. M. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol* **17**, 16–29 (2016).
51. Comai, L., Maheshwari, S. & Marimuthu, M. P. A. Plant centromeres. *Curr. Opin. Plant Biol.* **36**, 158–167 (2017).
52. Catania, S., Pidoux, A. L. & Allshire, R. C. Sequence features and transcriptional stalling within centromere DNA promote establishment of CENP-A chromatin. *PLOS Genet.* **11**, e1004986 (2015).
53. McFarlane, R. J. & Humphrey, T. C. A role for recombination in centromere function. *Trends Genet.* **26**, 209–13 (2010).
54. Wang, G., Zhang, X. & Jin, W. An overview of plant centromeres. *J. Genet. Genomics* **36**, 529–537 (2009).
55. Malik, H. S. & Henikoff, S. Major evolutionary transitions in centromere complexity. *Cell* **138**, 1067–82 (2009).
56. Neumann, P. *et al.* 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 (2015).
57. Wear, E. E. *et al.* Genomic analysis of the DNA replication timing program during mitotic S phase in maize (*Zea mays* L.) root tips. *Plant Cell.* **29**, 2126–2149 (2017).
58. Lee, T. J. *et al. Arabidopsis thaliana* chromosome 4 replicates in two phases that correlate with chromatin state. *PLoS Genet.* **6**, e1000982 (2010).
59. Rieger, M., Schubert, I., Dobel, P. & Jank, H.-W. Non-random intrachromosomal distribution of chromatid aberrations induced by X-rays, alkylating agents and ethanol in Vicia faba. *Mutat. Res.* **27**, 69–79 (1975).
60. Rieger, R., Michaelis, A., Schubert, I. & Kaina, B. Effects of chromosome repatterning in *Vicia faba* L. *Biol. Zent. Bl.* **96**, 161–182 (1977).
61. Schubert, I. *et al.* DNA damage processing and aberration formation in plants. *Cytogenet. Genome Res.* **104**, 104–108 (2004).
62. Schubert, I., Rieger, R., Fuchs, J. & Pich, U. Sequence organization and the mechanism of interstitial deletion clustering in a plant genome (*Vicia faba*). *Mutat. Res.* **325**, 1–5 (1994).
63. Dellaporta, S. L., Wood, J. & Hicks, J. B. A plant DNA minipreparation: Version II. *Plant Mol. Biol. Report.* **1**, 19–21 (1983).
64. Burge, C., Campbell, A. M. & Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci.* **89**, 1358–1362 (1992).
65. Neumann, P., Požárková, D., Vrána, J., Doležel, J. & Macas, J. Chromosome sorting and PCR-based physical mapping in pea (*Pisum sativum* L.). *Chromosom. Res.* **10**, 63–71 (2002).
66. Kato, A., Albert, P. S., Vega, J. M. & Birchler, J. A. 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 (2006).
67. 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* **8**, 427 (2007).
68. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402 (1997).
69. Evans, H. J. & Scott, D. Influence of DNA synthesis on the production of chromatid aberrations. *Genetics* **49**, 17–38 (1964).

Acknowledgements

The authors thank V. Tetourová and J. Látalová for excellent technical assistance. This research was financially supported by grants from the Czech Science Foundation (BP501/12/G090) and the Czech Academy of Sciences (RVO:60077344*)*. This work was supported by ELIXIR CZ research infrastructure project (MEYS Grant No: LM2015047) including access to computing and storage facilities.

Author Contributions

J.M. conceived the study and drafted the manuscript. L.A.R., A.K., K.B. and I.V. conducted the experiments. P.No. and J.M. carried out the bioinformatics analysis. L.A.R., P.Ne., I.S. and J.M. analyzed the results. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-24196-3.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

Supplementary Information

Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing.

Laura Ávila Robledillo, Andrea Koblížková, Petr Novák, Katharina Böttinger, Iva Vrbová, Pavel Neumann, Ingo Schubert & Jiří Macas

Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing

Supplementary Information

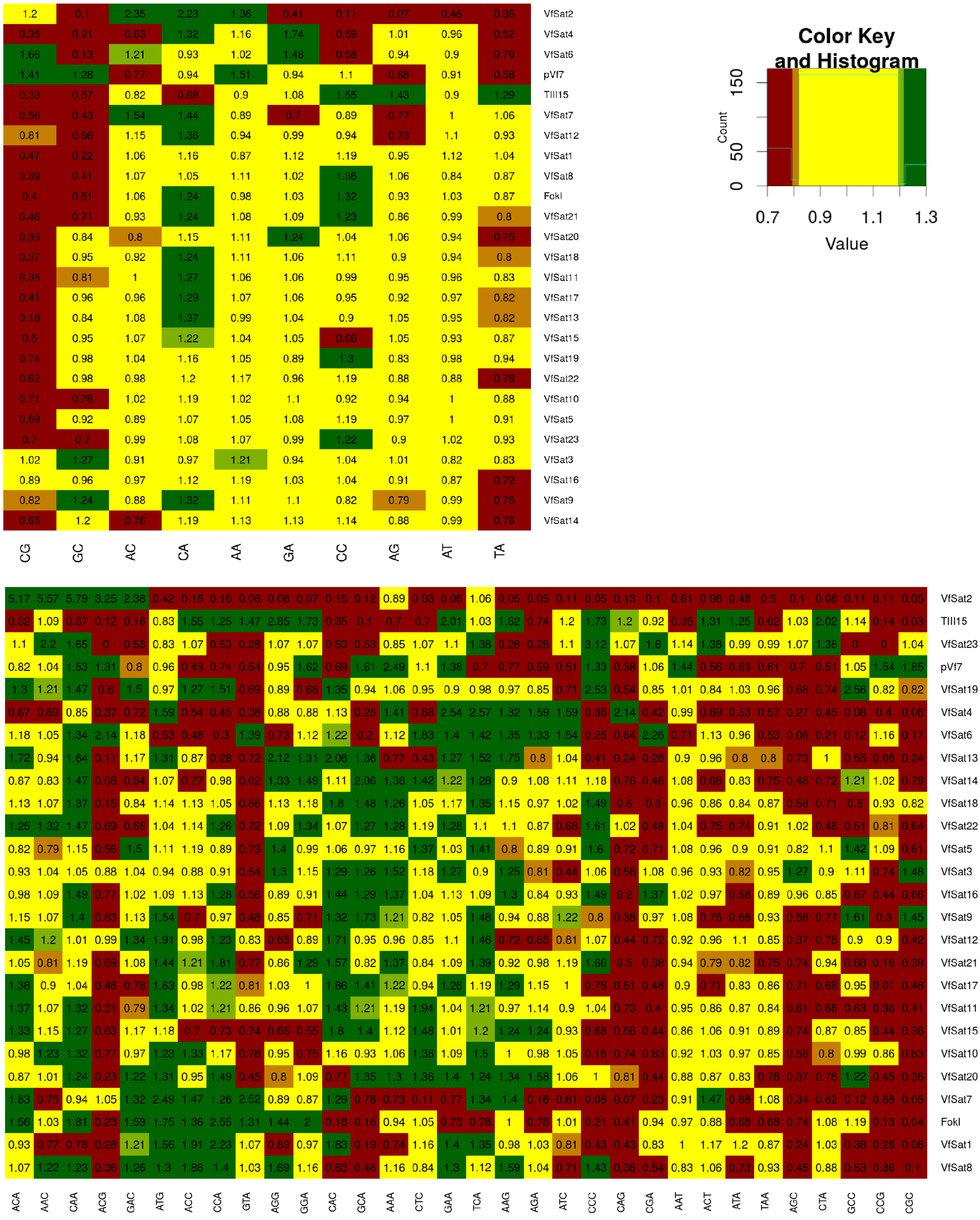
Laura Ávila Robledillo[1,2], Andrea Koblížková[1], Petr Novák[1], Katharina Böttinger[1,2], Iva Vrbová[1], Pavel Neumann[1], Ingo Schubert[3], and Jiří Macas[1,*]

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

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

[3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Stadt Seeland, Germany

[*] corresponding author (macas@umbr.cas.cz)



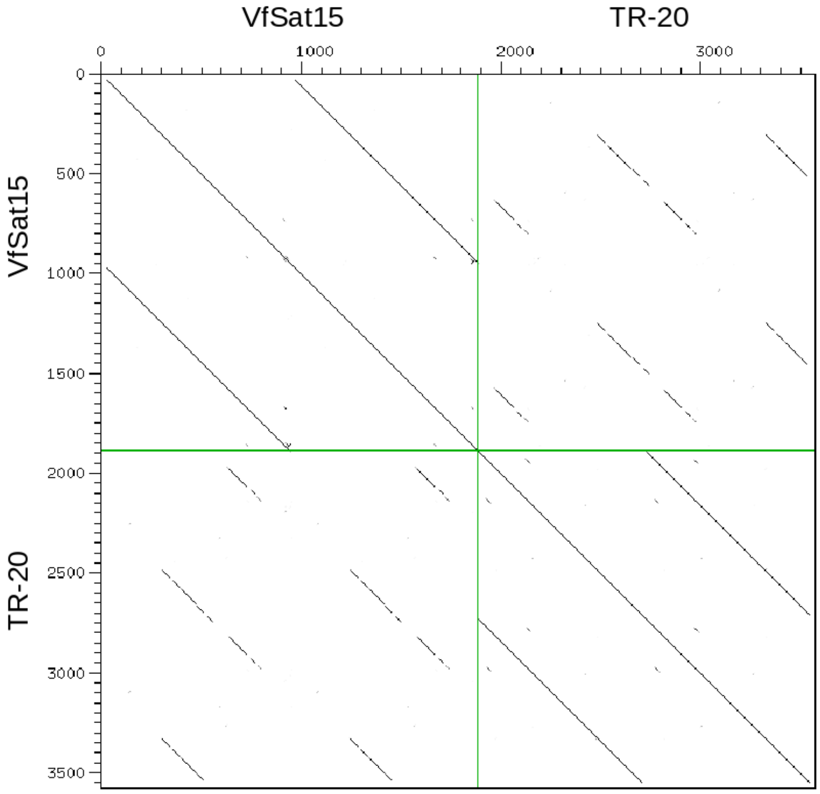
	CG	GC	AC	CA	AA	GA	CC	AG	AT	TA
VfSat2	1.2	0.1	2.35	2.23	1.36	0.41	0.11	0.07	0.46	0.36
VfSat4	0.35	0.21	0.63	1.32	1.16	1.74	0.59	1.01	0.96	0.52
VfSat6	1.66	0.13	1.21	0.93	1.02	1.48	0.58	0.94	0.9	0.76
pVf7	1.41	1.28	0.77	0.94	1.51	0.94	1.1	0.68	0.91	0.58
TIII15	0.33	0.57	0.82	0.68	0.9	1.08	1.55	1.43	0.9	1.29
VfSat7	0.56	0.43	1.54	1.44	0.89	0.7	0.89	0.77	1	1.06
VfSat12	0.81	0.66	1.15	1.36	0.94	0.99	0.94	0.73	1.1	0.93
VfSat1	0.47	0.22	1.06	1.16	0.87	1.12	1.19	0.95	1.12	1.04
VfSat8	0.39	0.41	1.07	1.05	1.11	1.02	1.36	1.06	0.84	0.87
FokI	0.4	0.51	1.06	1.24	0.98	1.03	1.32	0.93	1.03	0.87
VfSat21	0.46	0.71	0.93	1.24	1.08	1.09	1.23	0.86	0.99	0.8
VfSat20	0.35	0.84	0.8	1.15	1.11	1.24	1.04	1.06	0.94	0.75
VfSat18	0.37	0.95	0.92	1.24	1.11	1.06	1.11	0.9	0.94	0.8
VfSat11	0.36	0.81	1	1.27	1.06	1.06	0.99	0.95	0.96	0.83
VfSat17	0.41	0.96	0.96	1.29	1.07	1.06	0.95	0.92	0.97	0.82
VfSat13	0.18	0.84	1.08	1.37	0.99	1.04	0.9	1.05	0.95	0.82
VfSat15	0.5	0.95	1.07	1.22	1.04	1.05	0.66	1.05	0.93	0.87
VfSat19	0.74	0.98	1.04	1.16	1.05	0.89	1.3	0.83	0.98	0.94
VfSat22	0.62	0.98	0.98	1.2	1.17	0.96	1.19	0.88	0.88	0.76
VfSat10	0.71	0.76	1.02	1.19	1.02	1.1	0.92	0.94	1	0.88
VfSat5	0.69	0.92	0.89	1.07	1.05	1.08	1.19	0.97	1	0.91
VfSat23	0.7	0.7	0.99	1.08	1.07	0.99	1.22	0.9	1.02	0.93
VfSat3	1.02	1.27	0.91	0.97	1.21	0.94	1.04	1.01	0.82	0.83
VfSat16	0.89	0.96	0.97	1.12	1.19	1.03	1.04	0.91	0.87	0.72
VfSat9	0.82	1.24	0.88	1.32	1.11	1.1	0.82	0.79	0.99	0.76
VfSat14	0.65	1.2	0.76	1.19	1.13	1.13	1.14	0.88	0.99	0.76

	ACA	AAC	CAA	ACG	GAC	ATG	ACC	CCA	GTA	AGG	GGA	CAC	GCA	AAA	CTC	GAA
VfSat2	5.17	5.57	5.79	3.25	2.38	0.42	0.18	0.16	0.06	0.08	0.07	0.15	0.12	0.89	0.03	0.06
TIII15	0.32	1.09	0.37	0.12	0.16	0.83	1.55	1.25	1.47	2.85	1.73	0.35	0.1	0.7	0.7	2.01
VfSat23	1.1	2.2	1.65	0	0.53	0.83	1.07	0.53	0.28	1.07	1.07	0.53	0.53	0.85	1.07	1.1
pVf7	0.82	1.04	1.53	1.31	0.8	0.96	0.43	0.74	0.54	0.95	1.62	0.69	1.61	2.49	1.1	1.38
VfSat19	1.3	1.21	1.47	0.6	1.5	0.97	1.27	1.51	0.69	0.89	0.68	1.35	0.94	1.06	0.95	0.9
VfSat4	0.67	0.69	0.85	0.37	0.72	1.59	0.54	0.45	0.38	0.88	0.88	1.13	0.25	1.41	0.68	2.54
VfSat6	1.18	1.05	1.34	2.14	1.18	0.53	0.48	0.3	1.39	0.73	1.12	1.22	0.2	1.12	1.83	1.4
VfSat13	1.72	0.94	1.64	0.11	1.17	1.31	0.87	0.28	0.72	2.12	1.31	2.06	1.36	0.77	0.43	1.27
VfSat14	0.87	0.83	1.47	0.68	0.64	1.07	0.77	0.98	0.62	1.33	1.49	1.11	2.06	1.36	1.42	1.22
VfSat18	1.13	1.07	1.37	0.15	0.84	1.14	1.13	1.05	0.56	1.13	1.18	1.8	1.48	1.26	1.05	1.17
VfSat22	1.25	1.32	1.47	0.69	0.65	1.04	1.14	1.26	0.72	1.09	1.34	1.07	1.27	1.28	1.19	1.28
VfSat5	0.82	0.79	1.15	0.56	1.5	1.11	1.19	0.89	0.73	1.4	0.99	1.06	0.97	1.16	1.37	1.03
VfSat3	0.93	1.04	1.05	0.88	1.04	0.94	0.88	0.91	0.54	1.3	1.15	1.29	1.26	1.52	1.18	1.27
VfSat16	0.98	1.09	1.49	0.77	1.02	1.09	1.13	1.28	0.56	0.89	0.91	1.44	1.29	1.37	1.04	1.13
VfSat9	1.15	1.07	1.4	0.63	1.13	1.54	0.7	0.97	0.48	0.85	0.71	1.32	1.73	1.21	0.82	1.05
VfSat12	1.45	1.2	1.01	0.99	1.34	1.91	0.98	1.23	0.83	0.63	0.89	1.71	0.95	0.96	0.85	1.1
VfSat21	1.05	0.81	1.19	0.69	1.08	1.44	1.21	1.81	0.77	0.86	1.29	1.57	0.82	1.37	0.84	1.09
VfSat17	1.38	0.9	1.04	0.46	0.78	1.63	0.98	1.22	0.81	1.03	1	1.86	1.41	1.22	0.94	1.26
VfSat11	1.37	1.07	1.32	0.31	0.79	1.34	1.02	1.21	0.86	0.96	1.07	1.43	1.21	1.19	1.94	1.04
VfSat15	1.33	1.15	1.27	0.63	1.17	1.18	0.7	0.73	0.74	0.65	0.55	1.8	1.4	1.12	1.48	1.01
VfSat10	0.98	1.23	1.32	0.77	0.97	1.23	1.33	1.17	0.78	0.95	0.75	1.16	0.93	1.06	1.38	1.09
VfSat20	0.87	1.01	1.24	0.23	1.22	1.31	0.95	1.49	0.45	0.8	1.09	0.77	1.35	1.3	1.36	1.4
VfSat7	1.83	0.75	0.94	1.05	1.32	2.49	1.47	1.26	2.52	0.89	0.87	1.29	0.78	0.73	0.11	0.77
FokI	1.56	1.03	1.81	0.23	1.59	1.75	1.36	2.55	1.31	1.44	2	0.18	0.16	0.94	1.05	0.73
VfSat1	0.93	0.77	0.78	0.28	1.21	1.56	1.91	2.23	1.07	0.69	0.97	1.83	0.19	0.74	1.16	1.4
VfSat8	1.07	1.22	1.23	0.36	1.26	1.3	1.86	1.4	1.03	1.69	1.16	0.63	0.46	1.16	0.84	1.3

	TCA	AAG	AGA	ATC	CCC	CAG	CGA	AAT	ACT	ATA	TAA	AGC	CTA	GCC	CCG	CGC
VfSat2	1.06	0.05	0.05	0.11	0.05	0.13	0.1	0.61	0.08	0.48	0.5	0.1	0.08	0.11	0.11	0.05
TIII15	1.03	1.62	0.74	1.2	1.73	1.2	0.92	0.35	1.31	1.25	0.62	1.03	2.02	1.14	0.14	0.03
VfSat23	1.38	0.28	0.28	1.1	3.12	1.07	1.6	1.14	1.38	0.99	0.99	1.07	1.38	0	0	1.04
pVf7	0.7	0.77	0.59	0.51	1.33	0.38	1.06	1.44	0.56	0.63	0.61	0.7	0.51	1.05	1.54	1.85
VfSat19	0.98	0.97	0.85	0.71	2.53	0.54	0.85	1.01	0.84	1.03	0.96	0.68	0.74	2.56	0.82	0.82
VfSat4	2.57	1.32	1.59	1.59	0.36	2.14	0.42	0.99	0.69	0.53	0.57	0.27	0.45	0.08	0.4	0.06
VfSat6	1.42	1.35	1.33	1.54	0.25	0.64	2.26	0.71	1.13	0.96	0.53	0.06	0.21	0.12	1.16	0.17
VfSat13	1.52	1.75	0.8	1.04	0.41	0.24	0.26	0.9	0.96	0.8	0.8	0.73	1	0.55	0.06	0.24
VfSat14	1.28	0.9	1.08	1.11	1.18	0.76	0.46	1.08	0.68	0.83	0.75	0.46	0.72	1.21	1.02	0.79
VfSat18	1.35	1.15	0.97	1.02	1.49	0.6	0.3	0.96	0.86	0.84	0.87	0.58	0.71	0.5	0.93	0.82
VfSat22	1.1	1.1	0.87	0.68	1.61	1.02	0.48	1.04	0.75	0.74	0.91	1.02	0.48	0.51	0.81	0.64
VfSat5	1.41	0.8	0.89	0.91	1.6	0.72	0.71	1.08	0.96	0.9	0.91	0.82	1.1	1.42	1.09	0.61
VfSat3	0.9	1.25	0.81	0.44	1.06	0.56	1.08	0.96	0.93	0.82	0.95	1.27	0.9	1.11	0.74	1.48
VfSat16	1.09	1.3	0.84	0.93	1.49	0.2	1.37	1.02	0.97	0.58	0.89	0.96	0.85	0.67	0.44	0.66
VfSat9	1.48	0.94	0.88	1.22	0.8	0.38	0.97	1.08	0.76	0.66	0.93	0.58	0.77	1.61	0.3	1.45
VfSat12	1.46	0.72	0.65	0.81	1.07	0.44	0.72	0.92	0.96	1.1	0.85	0.37	0.78	0.9	0.9	0.42
VfSat21	1.39	0.92	0.98	1.19	1.66	0.5	0.38	0.94	0.79	0.82	0.75	0.74	0.94	0.68	0.16	0.39
VfSat17	1.19	1.29	1.15	1	0.75	0.51	0.48	0.9	0.71	0.83	0.86	0.71	0.68	0.95	0.01	0.46
VfSat11	1.21	0.97	1.14	0.9	1.04	0.73	0.4	0.95	0.86	0.87	0.84	0.61	0.66	0.63	0.36	0.41
VfSat15	1.2	1.24	1.24	0.93	0.68	0.56	0.44	0.86	1.06	0.91	0.89	0.74	0.87	0.85	0.44	0.36
VfSat10	1.5	1	0.98	1.05	0.16	0.74	0.63	0.92	1.03	0.97	0.85	0.56	0.8	0.99	0.86	0.63
VfSat20	1.24	1.34	1.58	1.06	1	0.81	0.44	0.88	0.87	0.83	0.78	0.37	0.76	1.22	0.45	0.36
VfSat7	1.34	1.4	0.16	0.61	0.08	0.07	0.23	0.91	1.47	0.68	1.08	0.34	0.62	0.12	0.68	0.05
FokI	0.78	1	0.78	1.01	0.21	0.41	0.94	0.97	0.88	0.68	0.68	0.74	1.08	1.19	0.13	0.04
VfSat1	1.35	0.98	1.03	0.81	0.43	0.43	0.83	1	1.17	1.2	0.87	0.24	1.03	0.38	0.28	0.08
VfSat8	1.12	1.59	1.04	0.71	1.43	0.36	0.54	0.83	1.06	0.73	0.93	0.45	0.88	0.53	0.36	0.1

Figure S1. Frequencies of di- and tri-nucleotides in *V. faba* satellite repeats. Representation values ρ^{*}_{XY} and ρ^{*}_{XYZ} for di- and tri- nucleotides respectively were calculated from sequence reads according to Karlin and Burge (1995) and are shown by numeric values and colors of the heatmap. A deviation of ρ^{*}_{XY} and ρ^{*}_{XYZ} value from 1 reflects marginal (1.20-1.22) or extreme(>1.22) over-representation, or marginal (0.79-0.82) or extreme (<0.79) under-representation.

```
VfSat4          CTGATGAAATTTGAAGTGAACATAAATCTGAAGAAAAT
VicTR-B         CTGATGAAATTTGAAGTGAATATAAGTCTTCAGAAAAT
                ******************** **** ***  *******
----------------------------------------------------------------------------

VfSat6          AAGATTTAACACGAACGAGTGTTT-GAATCAATACGGACGAGTAT---CAAAGA
PST_TR5         AATGATTAACACGGACGAGTGTTGAAAATCAATACGGACGAGTATTGACAAAGA
                **   ******** *********   *******************   ******
----------------------------------------------------------------------------

VfSat1          CAAATTTTAGGTTACTTCATCACTAAGAAACTAAGTT-AAAAGACTATTACTTAATGACA
PST_TR9         CAAATTTTTGGTTTCTTCATCACTAAGCAACAAAGTTAAAAAAACTATAATAGAATGATT
                ******** **** ************* *** ***** **** ***** *   *****

VfSat1          CATATTCCATATACATTTGAAATAATTCAAATTATCTAATGAGTCTCGATAGTATATTTA
PST_TR9         CATATTATATATAAATGGGTAACAAGTGAAATTACATAATCAATATCAATATTATTTGTA
                ******  ***** **  * ** ** * ******  **** * * ** *** *** * **

VfSat1          TTCACCATATTCATATTGTATTATGGTATAATAGATGTAAACAATTTCAATATTTTTCTT
PST_TR9         TTCGGCATATTCACATTGTATTATGGTATATTATATGTAGATAAATTCAATAAGT---TT
                ***  ******** **************** ** ***** * ** *******  *   **

VfSat1          CTTCTCCATCAC
PST_TR9         CTTCTACATCAC
                ***** ******

----------------------------------------------------------------------------
```



```
----------------------------------------------------------------------------
```

Figure S2. Sequence similarities of *V. faba* satellites to repeats from *Pisum sativum* (PST_TR5, PST_TR9 and TR-20) and *Vicia sativa* (VicTR-B). The similarities are shown as pairwise alignments of consensus monomer sequences, except for VfSat15/TR-20 which are due to long monomers and only partial similarity compared using dot-plot of dimer sequences with similarity threshold of 70 identities over 100 nucleotides.

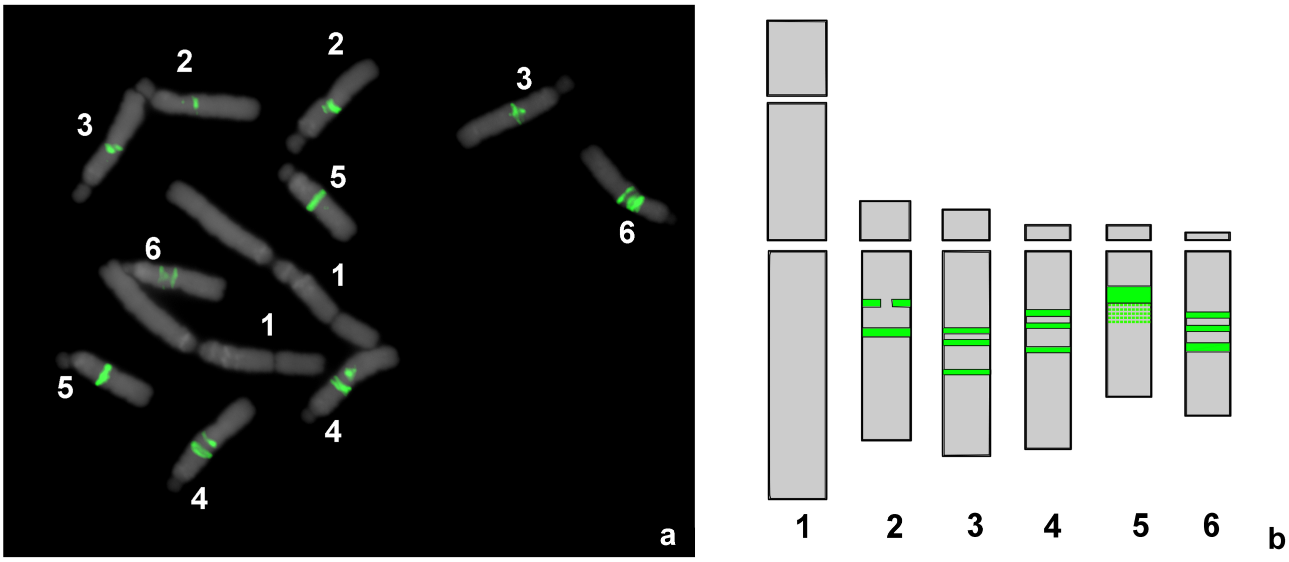


Figure S3. Discrimination of individual *V. faba* chromosomes based on their morphology and distribution patterns of FokI repeat. (a) FISH labeling of FokI repeats (green) on metaphase chromosomes. **(b)** Schematic representation of chromosome morphology and FokI patterns. The polymorphic FokI band on chromosome 5 is marked with dotted pattern.

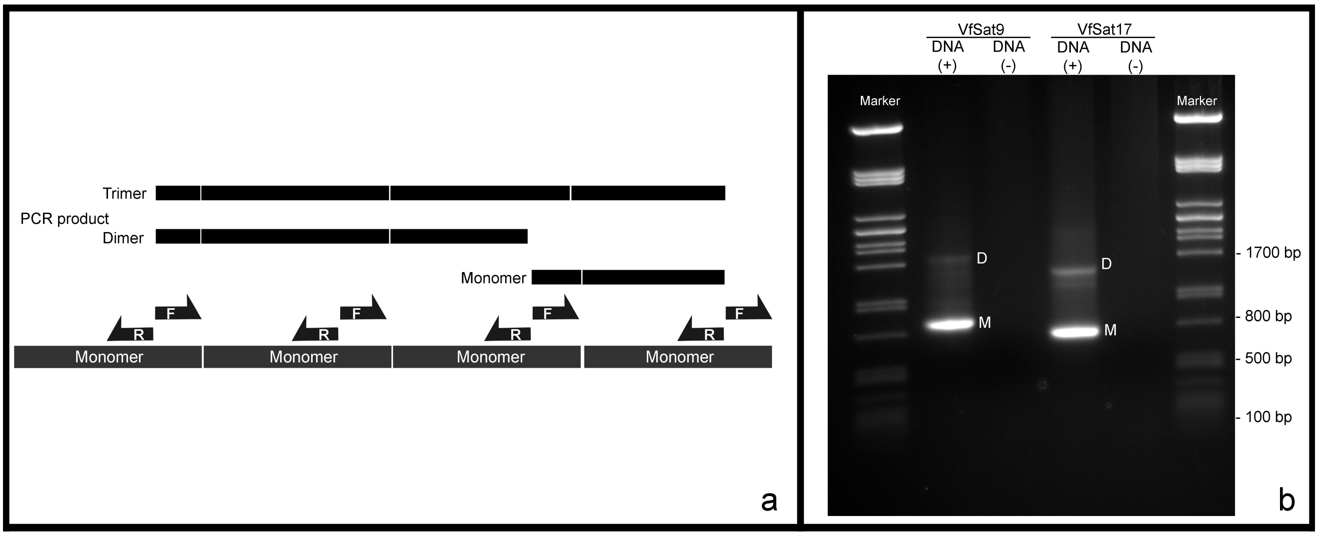


Figure S4. Design of primers used for PCR amplification of satellite repeats from genomic DNA. (**a**) Forward (F) and reverse (R) primers were facing outwards from predicted repeat monomers, thus generating amplification products only when their target sequences were arranged in tandem. (**b**) Example of agarose gel electrophoresis of amplification products from *V. faba* genomic DNA using primers for VfSat9 and VfSat17 repeats, showing bands corresponding to amplified monomer (M) and dimer (D) sequences.

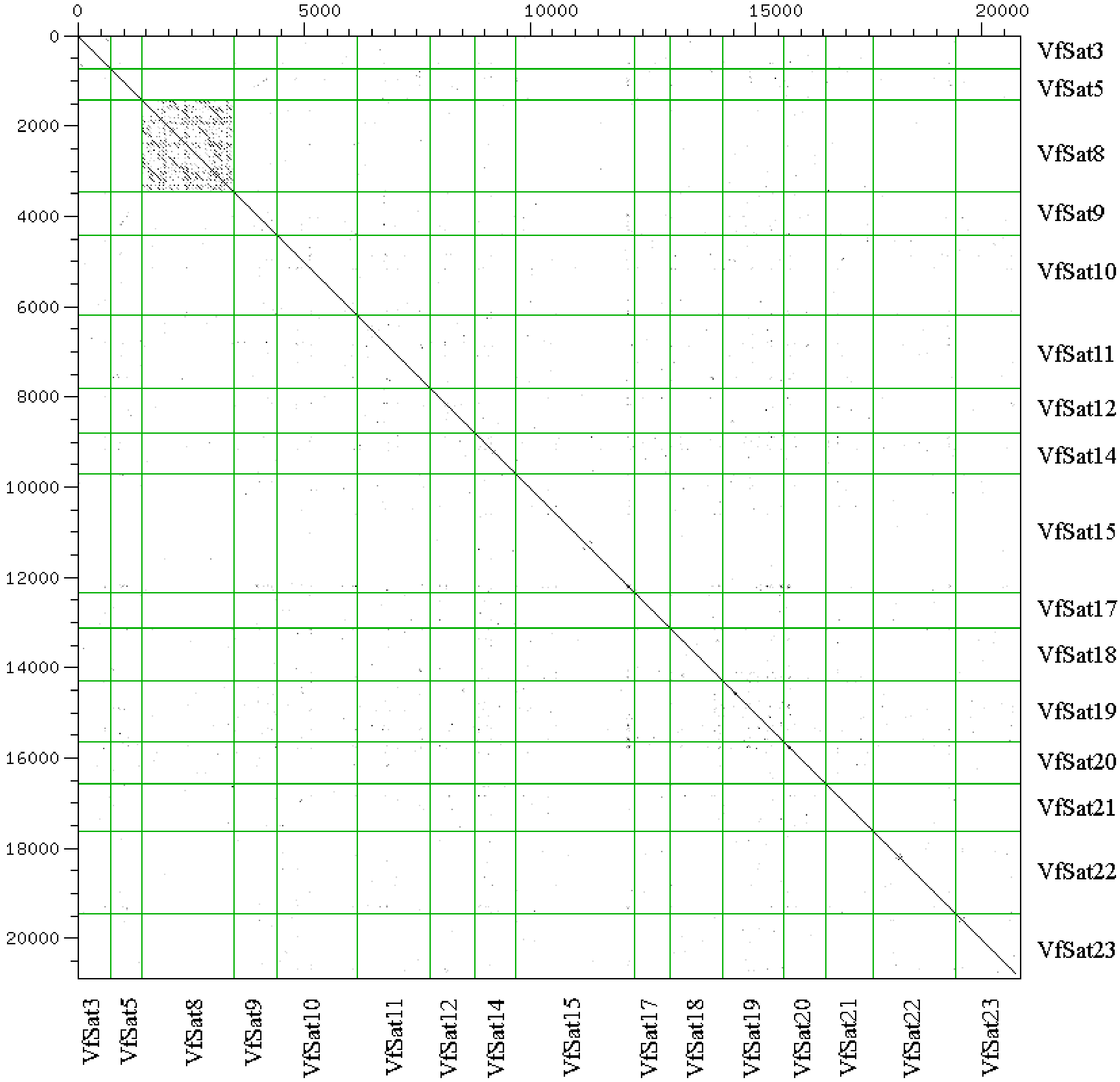


Figure S5. **Dot-plot comparison of satellites with long monomers.** Similarity threshold of 70 identities over 100 nucleotides was employed.

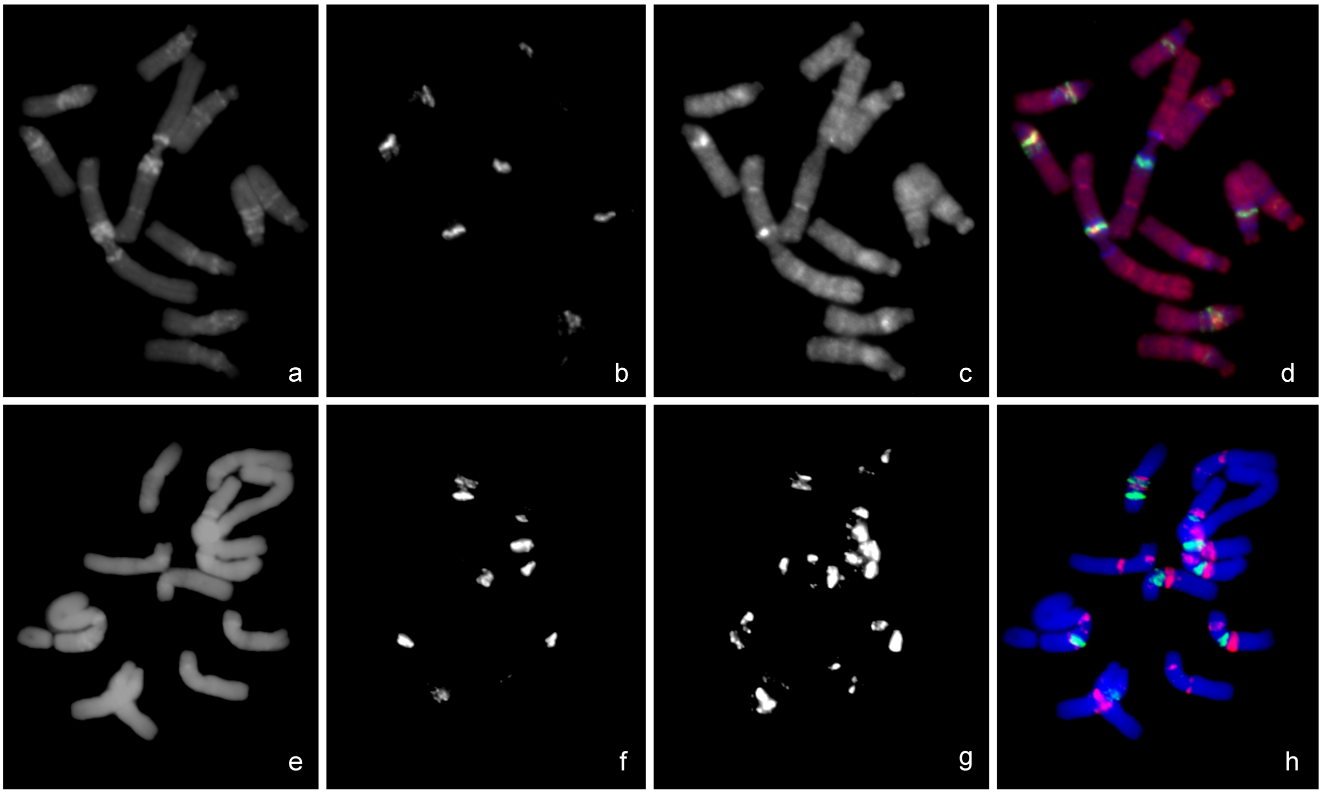


Figure S6. Combination of replication assay (EdU labeling, red) showing early (**a-d**) and late (**e-h**) replication patterns with FISH detection of VfSat2 (green). Chromosomes counterstained with DAPI are shown in blue.

Table S7. Replication timing of satellite repeats.

	Replication time (hours since the start of S phase)								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
VfSat2 included in dispersed labeling pattern*	+	+	+	+	(+)	(+)	-	-	-
Centromeric satellites	-	-	-	(+)	+	+	-	-	-
VfSat1	-	-	-	-	(+)	+	+	-	-
All remaining satellites	-	-	-	-	-	(+)	+	(+)	-
FokI	-	-	-	-	-	(+)	+	(+)	(+)

[*] VfSat2 did not replicate during late S phase (hours 7-9) but its exact replication timing was not determined due to the lack of distinct patterns on the background of dispersed labeling during early/mid S phase

Table S8. Sequences of oligonucleotide FISH probes.

ID	**Probe**	**Label**	**Sequence 5'->3'**
VfSat1	VFBm2H1	Biotin	CTTTTAACTTAGTTTCTTAGTGATGAAGTAACCTAAAATTTGGTGATGGA
FokI	VFBm3_Fok_H1	Fluorescein	CTACCTTCCATAATGACAAGGCTACCATCCATTGGAGTAACAAAAATCTC
VfSat2	VFBm15H1	Biotin	CAACAACAACAACAACAACGTCAAATAAACAACAACAACAACAACAAC
VfSat5	VFBm105H1	Biotin	AGCTCCCATCATCCAAGTAGGTAGTGCTATCTCACTCCT
pVf7	pVF7_TA_CL34	Biotin	TAAACCGAGGGCTTGTCGAAACGCTACGAAACTTTGGGGACACTCTCAAT
VfSat6	VFBm127H1	Biotin	ATCAAAGAAAGGTTTAACACGRACGAGTGTTTGAATCAATACGGACGAGT

Table S9. PCR primers used for amplification of satellite repeats from genomic DNA and GenBank accession numbers of cloned probes.

Satellite	Primer	Sequence 5'->3'	Accession number
VfSat3	Vf_TA_CL39_1 Vf_TA_CL39_2	AGCACGAATAAAACTAAAGTTC TACTTTTGAAGTGAAATGGAG	MF796528
VfSat4	VFBm102c102F VFBm102c102R	GCAGAAAATCTGATGAAAAATGATG TTGTTCACTTCAAATTTCGTCAG	MF796529
VfSat7	VFBm144c57F VFBm144c57R	TACCATAATGAATGGACCTTTATACT CGTTACATATTTTGACTAAGTACTTTTAATATG	MF796530
VfSat8	VFBm164c16F VFBm164c16R	CTAATCATGTTATGTCTCATGTAGTTTC GAAATGTTAATATCTTGTTAATCAAAGACT	MF796531
VfSat9	VFBm168c23F VFBm168c23R	CTATTTTCAAATGTATATTCGACATGC TAGGCCTTTTAGAATCAGTTATTGACA	MF796532
VfSat10	VFBm186c4F VFBm186c4R	AGGAAACAAATAACATTGCATTCTC ATTTTTACCGTCTCTACAAAGATTGAT	MF796533
VfSat11	VFBm187c11F VFBm187c11R	CCAAAACAATAACAAACAACATCAA CTTATGTTGTTTAGCGACATTGGA	MF796534
VfSat12	VFBm190c10F VFBm190c10R	TGTGTTTCAGTTCAAATGTGTGTCT AAATGTGAGATAACAACTACGGACA	MF796535
VfSat13	VFBm197C23F VFBm197C23R	GGTTATAAAACAACAAGCAAAGTAAG CCTTGCATGTTTCCCTTTAT	MF796536
VfSat14	VFBm198c11F VFBm198c11R	CTCTCTGTTCAATTTCTCAATCGTC GATTATATCTGCGAATGCCTGAA	MF796537
VfSat15	VFBm199F VFBm199R	TGAGAAGTCGTCCATCCTGA TTGCACAAAGAGAAACTTAAGGAA	MF796538
VfSat16	VFBm200C23F VFBm200C23R	ATCAAATTAGTTGGGGCTTG TTCGGCAATCGTAATCAAC	MF796539
VfSat17	VFBm205c11F VFBm205c11R	GGTATGAGAATGGTGTATCTTTTATCA AGAAAAGATATTTGGTTTCGAATGA	MF796540
VfSat18	VFBm207c9F VFBm207c9R	AAGATTCATCGGAAGTATTCCTTTT GAGAAATCACTTTGTAAAGAATTTGGA	MF796541
VfSat19	VFBm220c12F VFBm220c12R	TTCTGCACAAGTAAATGAATGGTTAT GGTTGAAGCCACTTATAAATCTCAA	MF796542
VfSat20	VFBm224c8F VFBm224c8R	ACTGGGCAGAAAAATGAGACTTA TTCAACTTTGCAAAAGGGGTTA	MF796543
VfSat21	VFBm233c2F VFBm233c2R	CACACTATTGTAATCTCCTTGCAAAT ACAAAATGGGGTAGCATGGA	MF796544
VfSat22	VFBm237c7F VFBm237c8R	TCAAATAGGACAACGTATTTAAGCAA TAATGCAGTGTTGTCAATGTTGG	MF796545
VfSat23	Vf_TA_CL281_2 Vf_TA_CL281_1	TAACCCAAGAGGACCCAATG GATACCTTCCTCACCCATACA	MF796546

Supplementary Data S10 - Reconstructed monomer sequences

```
>VfSat1_TA
CATCACCAAATTTTAGGTTACTTCATCACTAAGAAACTAAGTTAAAAGACTATTACTTAATGACACATATTCCATATACATTTGAAATAAT
TCAAATTATCTAATGAGTCTCGATAGTATATTTATTCACCATATTCATATTGTATTATGGTATAATAGATGTAAACAATTTCAATATTTTT
CTTCTTCTC

>FokI_TA
TCCATCGGAGTAACAAATCTCAACAACGAACTATCTCCCATAATGACGAGACTACCA

>VfSat2_TA
TTTATTTGACGTTGTTGTTGT

>VfSat3_TA
TAATAACTAAAAAGGAGGCAAACTAAATTGGTGGGTGTAATGAAATTTTCGCACAAACAGTAGCGTAGGTCCAAATAAACGCTTTGTCAAC
ACCAAGTATTTTTCTCGAAATCAAACCATTTTCAAAATATTTTCATAACAAATGCACTCGAAGGACACATGCATTAAGTAGTGAAGAAGGA
ATGATTAAATGTATGTCCCTTTGTTCTTAGTCCGACGACGCTGGCTGTATAGGTTATCGTTCGAATGCTCACCTTCACTCATGAACTTTAG
TTTTATTCGTGCTATATACTTTTGAAGTGAAATGGAGGAAAATGACTAGTTAGGCGTATGCTCCTCTTATAAAGATTCTAATAAAGACAAA
GAAGAGTGATCTCTTCTGTCCAAAAGGGTAAACCAACTTTGAAAGAGAAAATAGAGCAACAGACAATATGCCATAAAAGTTATTTAATAGC
ACTTTCTTTTGTTTTTGATAGCGCTGAAAAGCGCTATTAACCGTGCCACTAATGTGAACGATTTTGTTTAATAGCACTTTAAAAGAGCTAT
TAAAGTACTAGTTTTTAATATAAAAATAATAATTAATTAAAACAATAACTAATAAGGAGGCAAACGAAATTTGTTAGGGAAAATGAATATT
TTTTCATAAACAATAGCGTCGGCCTGAACCGTCGCTAATATAAAAAAATAAATAATTATTTAAAA

>TIII15_TA
TACTTCGAAATGAAAGCCTGGATAGTAGGGCTGTGAGAAGGAACCTGGATAGTAGGGT

>VfSat4_TA
CTGATGAAATTTGAAGTGAACATAAATCTGAAGAAAAT

>VfSat5_TA
GTCAATGATATGAGTAAATGGCTATTTCTATAGTTAATGTAATATTTAGTTTATAAATAAATAATTATATCATATTTTCGTTTACGGCCTA
ATTTTGTTCATTAAAACATAAAGGTACTTAGTTAGAATTGAGTTTGATTAAAAATGAATGAGTAAAAGTATGTATGTTTAAAGTATTTAAG
TTTGATTTTTATTTTGTGAGATTTATATTTTTAGTTTAAATAAATTATAAAAATTGCTTTTGAGTAATAAATACGATTAAATTTAGATAGT
AATTTAAATTCTAAAATTTAAAATACTATCATTCATAGTCAAAACTTGATTGAGGGTGAAAAATAAACTATTCAGCATATTTATTCTCATT
GATTTTTCAATTTGTCAATTTATTTAACATTACTTGAATTGTTAATTTACAAATGTTTAAGTGAAATAGACTTTTATATAGGTTTGTAGGC
TAATCAGACATTAAAAATGACTAGGCTCAGATCTACAAATAAGCTTATAATAGATTACCGGTTCATACTTTACATCCTTTATAGATTTAGA
ATTTTTTAGCTCCCATCATCCAAGTAGGTAGTGCCTATCTCACTCCTTCGAATTTTTTATTCCTTGTATATGTGAAAGTCATTAGTATAAT
TTATCATTCATTGAAAAATAAGTCACCTCTCTTTATAATTTGAATCTTCATATGACAG

>pVf7_TA
GAAATTCAAAATAAACCGAGGGCTTGTCGAAACGCTACGAAACTTTGGGGACACTCTCAATGTGTTATTTGAGATGTCCATGCAAAAAATC
AGCAGGAGATTATTTTCCTAAGGCGCGTTTGCCTCCGCTTCCGTTTTTGGCAAAAACGCAATTGCACGCGTACCGTCG

>VfSat6_TA
AAGATTTAACACGAACGAGTGTTTGAATCAATACGGACGAGTATCAAAGA

>VfSat7_TA
ATGTACATTGATGAATGAACCTTTACACTAAGTCAAAGTATGTA

>VfSat8_TA
CGATAAGGTTTAATCCTTTCAAATACTCATAGTATGTATAATATTCCTTTCTTTTCATTAGGTTTATTATGGTTTGAAATTAAGCTTAAAT
GGGACAGGCCTTTCAAATGCTTATACAATGTAACATAGTTTCTTTCTCTTTAATTGGTTTTTAATGACCGAGTATTTGGTCAACATGGATC
GGACCTTTCAGCAAATCATATTATGTAATCTACTTCTTACATGTTTATTTATTTTCTTTCGATTTAGTCTTTGATCAAGAAGCGTTTCTAT
TTTCAACTACTAATTTTATAAGATGCATTTATTTAGTCATCTTTTTTAGTCTAATTACTAGGTTTTTATCCATCAAGGAATGAATATTTCA
ACTACTCAAATCATGTAATGTACTTTATTTCCTAACTATTTGGTTTCCCATGATTATGGCTTTGACCAACCCAGACTAGACATTTCAACTA
CTCATTGTACGTAATGTACTTATTTATTTTTATTTGGTTTATTAGGGTTTAGTCTTTGATTAACAAGATATTAACATTTCAACTAATCATG
TTATGTCTCATGTA
GTTTCTTACTTTTAGTTGGGTTTTAAGGAATAGGTGTTTGGTCAGTAAAGAAGGTACCTTTGAAATACTCTTAGTATGTAAATCACTTATT
TTCTCTTCAATTGCTTCTTTATGGTTGATTCTTGGGTTGATAAGGATTGATTCTTTCAATTACTCATAGTATGTGTAATATTTCCTTTATT
TCCATTTGGTTTATAATGGTTTGAGATTTGTTTTACATGGAAAGGACCTTTAAACTACTCTTATTTTATAGTGTAGTTTCTATCTATTTAT
TTTGATTTTAATGACTAAGACTCTGACCTTTCAACTCATTATATTAGTTTACTACTTCTTACCTATTTATTTGGTTTCTTACCTTTAAGTC
TCTGACCAACAAGGGGTTCACTATTCAATTACTAATGTTATGTAATGTATTTCTTTTATCTTTATTTAGTTTCTAATTACTAGGTCTTTGT
TCATTAAGTAATTGACCTTTCAACTTCTCAAAGTTTGTAATGTACATTCTTTCCTTTTTATTTGGACTCTTAATGTAAGTCTTTGATCTAC
AAGGGTCATACATT
TCAACTTCTTATTATGTGTAATGTACTATGCTTCTCTTTAATTGGTTTTTTAGGGTTATATATATTTTATTAACAAGGTATGAACTTTTCA
ACTACACACATAATGTAATGTAGTTTGTTTCTTTCATTTGGTTTCGGAAAACTATGCCTTTGGTCAATAAAGAAGGAACCTTTTAATTACT
CATAGTATGTAAAGCACGTATTTTTCTTCATTTTCTTCTTTATGGTCGAGTCTTTGTTCGATAAGGTTTGTTCTATTCAACTGATCATAAT
ATGTGTAATATTTCCTTTCTTTTGATTTGGTTTATTATGGTTTGTGTTTGGTTTAATCGGGATGGACATTTCAACTACCAATATTATTTAA
TTAGTTCCTTTCTCTTTATTTGGTTTTTAATAACTAAGTCTTTAGTCAACAAGGACCACATGTTTCAACAATCATATTATGTAAAGTAATT
CTTACATGTTTATTTGGTTTCTTATGATTAAGTCGTTGATCAACAAGGGATTCTCTTTTCAACTACAAATGTTATGTAATGTATTTATTTC
ATCTTCATTTAGTT
```

TCCAAATACTAGTTCTTTGTTCATCAAGGAATGCACCTTTCAACTACACAAATAATGTAATGTACTTTATTTCCTCTTTATTTGGTTTCTT
ATGGTTAAGTCTTTGATAAACAAGAGCCTGGAATTTGGACTAATCATTTTATGTAATGAACTTTTTTTCATCTTTATTTGGTATATTAGGG
TTCACAAGGAAGGACCTTTCAACTACTAATGTTATGTAATGTATTTTATTTCTTTCATTTGATTCCTAATGGCCAGGTCTTTAGTCAACAA
GGAAGGGATATTTCATCAACTCATATTATTTAAAGGACTTGTTTTCT
CTTTATTAGATTCTTTTTTTGTTAATCTTTGGT

>VfSat9_TA
ATAAAATAAGCGACATAATCATCATTTTAAATCTTTATTATTTTTTTCACTCAAATAGTTGTTATCTTTTCTAAAAAATGTTTTCTAAAAA
ATATACCACAAAAAAATCATAAAACTTTTTACTTGAAAATTCTAAAAACATATCATCAACTCATTTCTGCAATCACTAGTTGCATATTAAT
AGAACTTTTGACATGAAAGACAACAACTCATTTGATATGTGAATCATAACATTTTCGCATCAAAAGATTGACATATTTAAATTCTTATTTA
AGATCTTATTTGGGCAATGCATGCTGGTCAATTTGTGTTCGATGGCAAGTTGAATGTGAATCCCATGATGTTTCTTAATTGCATCAAATTT
TTCTAATGACGCTCTTACCTTGTGTGGTAAACCATGTTATTCAAACATTAAGGAATAAAATTGTCAAATTAGATTTAGTTAGGATCATGAT
TAAATGATTGATAAAATGACAAATTAATTTTCCGAGAGACTTTTAATAAGTGTATTAACTATTATTAAATAATAATATTTAGAAGATTTAT
ATTCACATTGATTCTTTCAATTTTTACATGATTATTCAAATTTTACTTAGATAAGGAAACTAAGGCATGTCGAATATACATTTGAAAATAG
AGAATTTAAAATAACTTGTAAAATCTATCGAATAAAAGTAATTAGGCCTTTTAGAATCATTTATTGACAATTTTTTTATAAAACTATTAAA
ATAATAACATTTGATTAATTAAATCATAATATCAACGCAAACTAATAAAACAGTTAAACAATGTTGTTACACGTAGCACTTTTCAACTCAT
AATATTTTATTTTAGATTAATTATTTTTAAGATTATATCCTATATATTTTAAAGTCGCGTAAATGATTTATTTAAAAAGTATTGAGCTAAA
TTATCGTGTAGGATTATTAATAATAATTTGATTTTTATATTAAAACAAATAAA

>VfSat10_TA
ATATAAATATTACTTTTGAAAAGAAATACATTTCACATTGTAATTATAATTTCAAAATAAGAGAGTTCAAAATTCATTTAAAAATGCACAG
ATTTCATAGTTTTAAAAACCAAGATTAAATTTCATTCAAGTGATAAAACATATTTAATTCTTATGAACGCACAAATTTCTCTTAAATTTCT
CTTATATCATATAGGCAAAAAAATATGATAACATTGCTCTACTTATAAATAAAAGACTTATAAATAAACTTCAAATGACTACATTAAACAA
CAACTTATTCTTCACATGCCATTAAATCAAAAGAGATAATATAGAAAAATTAACCTAGAATATTCATTACTCTAATAAACTTCAAAACTCG
CCACCTCACTTGATATAAAGTCCAATGATATGTTTAGAGTAATTCATAAAACAATAAAGATAGTTTACATAAAAACAAAAATAAATTATAT
GTAAATTGAGATGTGAACTCACATATGATCACACAGAACAAGTGAACCACCATCAAAGCAAATAAAATAAATAAGTTCTTATATTTTATCT
ATTTTATTGAAAAATATTTTCTTTTGGCAACGTTGTTGTCGCCGCCACTTCAGTATCAGAAGCAATGACTCTGATCATCTGAGAATTTTCA
AAATACTACCAATACCAATCGTTAGTATTATTGGTTGCCAAATCGGTTGAATATGATGTTCTTTCGATTGAAATTACTAAATTTTTATTAT
TTTTATTTAAATTTTATTTATGTTAATAAATCCAGTTAATGAGAATGCAATGTTATTTGTTTCCTATCAAGTGTTTCTATGTGGCTGTGTT
GCAAGTACAATTATCAAGTCTAATTTTTACCGTCTCTACAAAGATTGATTGTGTTTTAAAAACTGTTCAATTGTTTCTATAACTTTAAGAA
TGATTTATCAGTTTTTTGTGACTACAAAAAGTAAAATAACAGAAAGTAAAATGAACTTTGGTGTTTAACCGTTTAGAAAACTTGATGGTTT
AAATATCATTGACATATTATCATATAGTATCCTTGATCAGTAATATGCAATTCTAATCAATTATTAATATTACCGCTCGAAACTCTCCTCA
CAACTCATATATGAGTCAACGACGTGAGATTATTCATATTACCTAATATATGATATCTCAACCTATTAACCATCACGATATCTTTAAGTTT
CAACAGTTCATACGAATATTGAACCTAAATCTATGTCTAGACTTTAGCATATCCAATAGATGAAAAAGTTGGACCTAGTAATTACAAGTAC
CTTTCGTATATCAAGTCATACTATGAAAACATGTTTTAGGTAGATAATCCAAATCAAGCAATAAGAACAAATAAGTCATCAATGTTAAAGC
TTAAGAAGAATTACATAGAACGCCATGACCGGGATCAAAACATGAGTATTAGGACAACCTCAATTTTTAAACTAATTTGTAATATAAAATA
TAAACTTATTTTAGGTATAAAAGTTGAGAAAATGATTTTACAATAGAAAAAAAAATGCTATAATGGAAAATCTATGGAAGATGTCCTGATAC
TTATAATAACTCATACGTATGATATTCTTCTTAATCATATTTTACATTTGTTAAAGGTTAACTCAGAAACTCATCCTTCATAATTTTAAAA
TAAAATACTACAGTTGCTTTGAAATTCTTATATTTATATTTTATTTTTTGTCAAATAATTTAGTTGGCTTTCAGCCTTATTTTATTTTTAT
TTACTATTAAATTATTCTTCATAATTATATTTA

>VfSat11_TA
CCAGTTTAAATATCGATGTATGATTTAGATTAAATTGCATTTGTTTTTTGTTTTCTTCAATTTTTGTAAGTACAAATTTTTTTAAGTTAAG
TACGTAGGGAAATAGAAGTTAATATACATGTAAGTACAATTTTTGTAATAAGATGAAAGGAGAATATACAGTGATAAGTGTGATGTGATGT
TCATTGATGTTGTTTGTTATTGTTTTGGTTTTGATGGGTTTTTTTTCTTCATTTTTTTCTTATGTTGTTTAGCGACATTGGAGAGAGCGGT
AGAGTCTGAAGCAAAATGAATGAGTATAACGAGAACAAACCAATTGAAATCTTAGTTTGATTTGATTTTATATATTTATGTTACATTGGAG
AGAGTCTCTGTCCTACCTTAATTTTTAATGGTTCTTTTAGGATTGCGTGATTCCTTATACAATTTACATGTCTTTCTCAAATATTAAAATA
TGCATGTTTTTGTTTATCTAATTTAGAAAAACAAATAAGAGTATAACCTGAATAAATTTGTTTTATTATTTCTATATATGTTACAAATCAC
AAAGAGAGCATAATCACAGTTTAATGATTAATAAGTTATATATTTAGTTAGTTTCACAATATTTTTTTATAAAAATGTTAACATAAATAGA
TTATACTAAAAGGGTACTTTAAATAAATAAAATAAAAAATAAAAAATCTCAACAAAATCCAAATAAGAAAAATGAATTATAATGCATGCTA
TTTATACACTCATATGTATCAATCTGTTGCAAAGAAAAATGTGAAGCAAAGTAGCAGAAGGTATTTTAATACTTCTTATACACTCATATAT
TAGTTCAATTCTATTCATATTGAGGTTTAAGTTGTCATATTTTATTTTAAGAGTTCATCTGAAACTGACATATTTTTTTTAAATATTATTT
TACAGTTTCTTTAAACCAATAAAGAGAACAATCGAATAAAAAATTACTATTTCATGATTTGGCAACCAAACTCACTCTAAAAGAATGCCAC
ATTGGTCAACTTGCCTATAACTATCATATTTTACATTTG
AAAAATAATGTATTTTGATTTTATTAAAATCTCTTTATAAGTGAGAGTAAAATGTAGCTTATATACTGCAATGTAAAATCAATTTTTCACT
CAATATAATGTGCAACTTGATTTGAAATGATACACGAGATGGAGAGACTTTTAATGTAGGGTAATCTCAATTTTCTAACTTTTAATTTCCA
TTGTTGGTATTTTTAAATCTTAAAGACATTTAAATGATAACTAGAATTATAATGGTCCAATTATCATTTAATATTATAATTGATGAGGATA
CAACTATTTATTACTCATTTTGATTCACAATGAATACATCTAAGTGAAAAATGAATAAATAAAAATGTAATGTTAAGTTTTTAAATATATA
TTTTATAAGTATGCAATTAATATAATAAAATAAATAGCATGCATTAAAATTATGTATAGTGTCTCTTCTTATTATTGTATATCTATCGTAT
AAGAAAACTCACTCCTTATTAATCACACCGTCTGCTTTGGATTCAAATTTGTCTCCTAACAACTTTGGTGAAAGACTAAACATTTATCGGT
AAATGCATCAGATTTGTAGAGGCTTCTAATTTT

>VfSat12_TA
AACTCGGAGACATTTTGATGAACAAGTGTCTTCACACAAACAAAAGTTAAGGGCTTTAAAAATTAATATTAACAATATAGTCTAGAGTTCA
TCATATGATATTCTTGATATTTTATTTCCTTCGTTGTAGTAATGGATTGTAAGAGACAACTTACATATAAACATCTTGAACACAATGTTTT
TGTTCATTAGTACTATTATTTTTTAATTAATTAATTTTTTTTCTCAAAGCTAGTTTACATCCAAGGACCTATTGGCATGGGCACACATCCC
TTCAACATCCCAACTTAAATTTTCATTAAACCTTAGAGGTTGTGTTATAAAAAATTCTAAAAGATAAATAAAATAAATAGATATCTTCACA
AAATATTCTTATTATTTTCCATAATGATAAATATAAAAGAATTAGAACAATAAATATTTACCTTAAACTTATAATTTCATATCATTAATAG
ATTAAGCATTGGTAACATAAAAAAATAAGTTTCATTTAGTAAAGAGTTTTCATCCGTATAACAATGAGTAATGAACCATTTAATTTATATG
CATCTTTATGTATGGTGATGATCATGCTAGTTCTAGAATAACAAACAACCAACATAAAAAATCTATTTCACTTACACACGTAAAATTCATA
ATGGGTGAATAGTCGAATATGAATAAATATGAAACATCATATATATGACTACCGTCACTAGTAAGTATTCACATGACATTTTATAATGACA

```
TAAATAAATATTTCTCGTATATCAACATATAAAATTTTTTGTGCATTGAATTTATAGACACACATTTGAACTGAAACACATATCCTTAATT
AAATGTGAGATAACAACTACGGACATACGACTACATATAACTAGTTATTGCACATTGAAACCAAAAGTATTGTTCCATTTATTGCAAATGA
CATCATAAGTTACACAAGTAATTTTTTTCAGTATTCATACATGTCACTTATATTCATTATGATGAACGATATTCCAAGCAAAACGTAGTTA
GAT

>VfSat13_TA
TGCTTGTTTAGGAATACTTATTGTGATGTCTTGATTCACTTTCCTAT

>VfSat14_TA
TTTAGAAATCTTCTATATCAAATATGAGTGTACGGGAAATGACTTATTCATTTCCATTATTAATATACATTCTGAAATATTATTATAAAAA
TAATTATAAAAACTAGCTAAGAAAAACAAAAAACAGAGTAAAAAATTGAAAACAATTTGTAAACAAGTCACCAAATTTAAGAAATTTAAAA
CATAATATATTTAGATAAAGAAACTAATCAAACAAAGTCAAAATTATAATACTTAATTAAAAATACAAATTTATTGTTGTTTGTTGTTTTC
CCTTATTATATTTTTTAATTTTTTCTTTCTACCTTTTTTTATTCTATGGACGAACTTTTTTTTGTTCACTTATTTTTTTTATAATACTTAA
TTTTGATTTTTACGTTATATTGGATTTTAATTGAGATTTTGATATGTGCAATTGCCTGCCATTTGCGATGGTTTTGATTTTTATGCTATTT
TAGCTGATGATTTTGATATGCAACTTACGTTTATTTTTTTTAATTGAATGTGCTTTGAGTATTTTTACATATCCTTGAATTGTGATATTTG
ATAGGATGAGTATCACATTAATTTTTTTAATTTAAACATCACAGGATTTTTGAATGGTACCTCTCTTTCTAAAGATTTGACAATGAAGAAG
AGAAAAACAAAATACTAAATAGTAATGCAAAAGATATGCCGGAAGATATTAGGGAGAGAAAGGAGAATTGCCCTATGAAGCAGGTTGGAAG
TGCATTTGATTGGAATATATTCAATAAAATAGGACGATTGAGAAATTGAACAGAGAGTAAAATTTAATGCAATTCTAATGATTATATCTGC
GAATGCCTGAATTTCCATTTTTAGAAACATAGTGATTATTAATCTTTAGAAACATAGTGGTTATCAACC

>VfSat15_TA
ATTTTTAAAATAAATTTTTAAAATTTAAATAACACATCATCTTGAATTCTATTCACCTAAGATGACACAATTACTAAGAGTCATTATCTCT
GATTTTGCAAAATTCATAGTAAATTCTAGCTTCTTTAAAAGTTTGATAAGTGTGTGTCTGTGAATATAATATGTAATGTGATATAACATCA
CTTTAAAATATATACAAACAATATGAAGATCAATAGAGAAGACAAAACATATTAGTGTGTGTCTTTAAAATACTCAAATCTTTAAAAAAAA
TCATTCATATTCTAATGAATCATGCAAACAAAATAATAATAATACACAATATAGAGATATGTTTGTTAGAAAAATTAGAATTGTTATTCAA
ACAAGTTTTATCTTAATATTTACATATACGAGAAGCAAAAATAAGAAGAAATCTAAGATTATTTGATGCAAAAAGAAATACATACTATTGT
TATATTTATCATCTTCCTTAAGTTTCTCTTTGTGCAAAGTATAACAATGACATGATCAAATTAAAGATATGACATAAATCACCTTTATGCT
GAAAGAGTGTCCTTTTTTTATGGTGTTTTTCATCATAATATATAGTACCAATTTCTATAATAAGTTTAGTTACTATAATTTTACTACTTGT
AAGAGTTATAAGTAGTTAGTTTCTCTAGAACATAACAAGCAGTCACTTTTATGACATCTAGTAAGTGTTGTATAAATTAAAATTGATAAAT
AAATATATTTGCTCAAAAATTGGGTAACATTAGTTAAAGGTTTTGGCTTTCTTTTTTAAACATTTCAATGTAAGAAATTAAAATTAAAAAC
TCGTGTAAATGCACGATTATAGAGTTTTGAGAAGTCGTCCATCCTGACATTGCTCAAGACCACTACTAATACAACTTTATATACAATATAA
TTCTTCAAAAAAAATATTTTTTTAAAAAAAAGG

>VfSat16_TA
AAGAAGGAAAGGAAAAATTTCGAATAAAACCCACAAACAAAGGATAAGATGGTCTTCGAGACCAAAGAGAGGGTACATGAGTCGGTTATGC
AAGGGGAAGGTATTAGCACCCCTCACATTCATCGTACTCGATGGGAACCATTTGGTTCGTGTGTGTGTGTTCGAGTGGTAGTGTGATAGTT
TGCAATCTTCTACTTATTAATCTTGAAAGGAGAAAGAAGTAGGCTTTTTGTTTTTTAGTTTGTTGAGTTCGACAAGATTCGCATCTTGTGT
CTACGTACTCCCTCGTGCAATGGGAAAGTCAGAACTCCGTAGTTCTTCTAAAAAAGACCAACGGTGTATTGCTTGATTTTAGAAGAATGAT
GAGTTAGACATTTCAAACGTTTGAACTTCGACTTGTTTTGCTCGTTCGCGGAAACTAAGTCTTTGTGTTTGTTTTCCTATTAAAATGGCTA
AAACACATTCCTTTTATGAAAAGGTTTTTGATGTCGCGCAAGGGCGAAAAAACAAGTTTGATGAGTTGAAGTTGTTTTTATGTGGGTGATG
AGTACCGAAAAATCGGACTAACATCCTACGACTCAAATACTAGAAATTAGAGGATAAATGAAGCTCAAAGAGTAGTCTCTCAACCCCAAAA
GTTATTTTCTTATGAAAAAGATGAAGTGAAACAAGGTTCACGCTTATTAGGTTTTTTGCACGAGTCTTAGACAATTGGTCTAACAACCAAC
GATCTAAATACTCAAAATTTATTATGAAAATGTTTTTGAAAATATAAGTCGACGTTGGATCGAGGATTTGAAACTTTATTATGAAAGTGCT
TTAGATAGAGAATGGGAGAAAGTCGATTGCGGTTGCAAAAGCAAACTCGACTTATCAAATTAGTTATGAGCTTCGTATGTGGACCAAGAGT
GCACGAGCTAGATTGATTCAATTAAGTGTTCAATAGTGAATAAGCGATTAAAAGCAAAATAAAACTATTAAGCTATTACACGTCCAAAATG
CAGGGATACACTTGTTTAATGATGATTGACAAGTAAATCACACAAGCCCATACAAGGTGGCCCACACAAATGAAAAGGATAAAGCAAAAGA
GATGAATTATAATCTCTAAGTGCACTAGCAAAGTGGGCTAGTCTTCTAGACGACTAATTAGAATAATTAACCGACTAGGTCTCACTGTAGG
AAAGCCCAGGATTAAGCTATGAATGTGTGCAAGTGTGCATGTGTTTCGGTGTGGAGGGGTCCAAGAGGAAGCCATGCGAGGTCGTTTGCGA
TGCTTA
GAAATAAATTGACTTATGGTGGAAAATTGGCTCGCATTCGAGCATTAGTTCTTTGAAAGGTTCGATGCATGATGATGATGAATGATTAAGG
TAAAGCAATAAAGTACAATTAAATGATTATTACATCACACGGGGATTGGGTACAACCTTTTGAATGGGGATGGAACCAACCAAATCAACCA
CACACGAAAGCCTAATTAAAATCAAACAATACAGACCAATTGAACCAAATAATTAATTAATTAAACTAAATAATTTTATTTAATATAATTT
ATTATTAAATTAATTAATATTATTAAATAAAATTGCTTAAGATAATTAATTATCTAAACACACGTTTTTTGTATTTTTATGATATAAAATA
AAAAATAAAAAGGAAAGGGACATGGAAAATATAAAACAATGTAGTCCCCGCCGAGATTTATTTATTTC

>VfSat17_TA
GGTAATTGTATCACACGCTTTTATTTTATTAATAATAAAAAAGTAAGTTTTTTCTACTTAACTTAAATTCTTTTTTAAGAAGGAAAGCATA
CATTTGGTATGTTTGGCAGATGATGCAGGGATGAATATCTTCACGTTGAACAATTAATAATGGATTGATTTGTTTTTCTTCTAATAACTTC
TATGTTGTAATTTTTTTGAAATGTTTCTTATATAAAGAATGAGGATAATATTAAGCTTTTTTTATTAATCATTGACATGTCACACATTGTAG
ATAGAGTAACACACAATTAAATGTGAAATGCATATTTCATTCGAAACCAAATATCTTTTCTATAATATAGGTATGAGAATGGTGTATCTTT
TATCAAATGTATCCATTTAAAAACACATTAATGATTATATCTCATTTGTTGTAAACCTTATCTTTTTTGTTCAAGAATCATACTAATTATC
TAAGAAAAACACTAAGAAGTGGTGGTATCAACAAGTTTCTTAGTGTCCATGTCAATGAAAAAGCACATAAAATGTTATTCTGTATTGATTT
CAACCTTCACTCGCTTATGTTTAGTAATAGTCTCATTTTTATACTATTTAATTATTTTTTTAAGTATTCTGGCATCGAATATGTGAAAATC
TTACAATTTGTACTTTAAAACAATGTGACTAAAAATTAATATTTATTAAATGAACATGCTTTTAATATTGATAGAGTATGCATCCTCTCAT
CTTTGCCTAAAAGAAATATCAAGAATGAAAGGTTACTATATGAATAAATTAAA

>VfSat18_TA
AAGAATAAAATTTAATAAGAATATAAAAAAGAGTTTTTATTTTTAAGACACAATAGGTCACACAATAAGATTTTTTTTTTAAAAAGTCATG
CATATAGAATATATTTTAACTTTTATGTCAAATTTATTGAAGTTATTGATTTTTAGGTATGATAAAAACAAATTTTATATATGATAACAAT
TTTTAGCTCTTATTTATATGATAAAAACTATATATGCTTATAATAATTTTTTGGAATAACTAAAAATTTATTTTTCTTGACCATTTAAGTT
TACTAAAAAAAAATAACACAATTTTCTTTACAATTTAAGGACAAAAAACACTAATATTTCTAAAGACACTAAATGAATTTCGACTCCTGCCT
```

```
ATCAATCATTTCTTCCTTTTTTTTAACTGCATTTTGTTTCTTACCTTTCATTTTCATTTACTCATAAATTTCTCTGCAATTCAAACAAATA
TCCAAATTCTTTACAAAGTGATTTCTCAAAATTTTTAATTCTCATTCATTAAATCTTATCTTAAATTGTGAATCTAAGATTCATCGGAAGT
ATTCCTTTTGAAATCACATTTATGGTTTTTAGACATAATTTTGTTATTGAATTCACTGCAATTGCGCTAAGCTACTTTAAATTAGAACAAT
GTGGTTTTGGTTTTGATATGATTAAGGTTGTCAGTTAGATGATGCATTTTAGTTATATCTTAAATGTAATGATATTGAATCTTTGAAGATA
TTGTAAATAACTATTATTACCTTACGGGAAAAAGTGACATAAAATTTTATAAATAAATGATAACCACTCAAAATAGGATTGATATTAATTA
TGTGTGTGATTAGTTGAAAAAAAGTGGTTTATTAATTCAACAATATTTCTTTTCTTCTTTACTATATTACTTGTTTTTAAAATATGTAGTT
TATTAGAAAGTTTATATAAATTAATTTAGTGAGTATGCACATAAATTATTAATATGTTTAATTTTTAAATATTATAAATGTTCAAACTATA
TGTTATGTTTGTTATGAAACTTGGGATCCAAAGAAAAATAGAAAAAGTTAAATAAAATATTGTATTTTTGTTGATAAATATAATTTTATTT
ACTATATTTTTCCTTATTACATAACTAAGTAGTAAAAATGATATATATTGAACAAAACTTTAAAAAAATTATTATTTTTC

>VfSat19_TA
TTTTAATATAGATTTTATTACCAAAATGACATGTGACCATAATTATTGCCAAAAAATAACATGTGGCTACGGTTGTCATCATGACTTCTTT
TAAATTTATGATAAATAGATGACAATTGTAAAATCTTATAATGTAGTAATATTTTTTACATATATTTTAAAGTTGTAGAAACGCACCATAT
TTGTTTATGCAAAAAAAAAACATTTCTTGAATATTACCTAACAGATCTTATTTATTGTTACATATTTAATATTGTCAATGTATCTTATATA
TGGTTACATACTTAATATTGTAAATGTGTGAAAGTGAGATGTGGGCCGATGAGGGAGAATTCTCACAACACATTGTTATTATAGGGATTAG
GAATATATAAGTTATCCTATTGGGCAGTGTTATTGAAAAATTGAACATCGTTAGATTATAAATAAATAAATATTAATATGTTATACTTTTA
TGATGACTTTCTATAAGATTTATATTATAATTATTGTTGTCAAATTTTGTCATTTCTAAATATTTCCTAATTATTTAATTGAGTGATTTTA
AATATTATTTAATAGAAATTTTTGTAAATAACATATAATTATGTTTTTAAATTTTTAAGAGTTGTAATTTGCATATTTAAGTTTAAATAAA
AATCATTTAAATTCTTTTAAAAGAGTTCAAAAACCTATTTATAAATCATTTAGGCTTCTATATAATAGACGACTTTATATTTGTTTATAAT
AAATTAGTCGATTAAATTGTATATTGAAAGAAAAGAAAAATATAAAGTTGTTGATTTGTTTTATTAAAAAAAGCTATAAAATACAATTATA
AACTTTTTTGGGTATAAAATAACAAATAACCATTCATTTACTTGTGCAGAATATTTACAAGTTGATTGAATTTGAAGTGTATGATAATTGT
AATGGTTGAAGCCACTTATAAATCTCAAATGACATAAAAAAATAAGTTTAATTAAATTTAATATTTTAATTAAGATTCTTGGAATTAAGCT
GAGGAAAAAAACTATAAGTTACTTGATTTAGTTTATTAAAATAAACTATTAATCAATAAAAATAAATTATGAACTCTTAAGAAAATGATAA
TTATTAAATATATTAATTTTTCCAATTATAACGCAAAATCTATTAACTATAAATATAACTTTGTCAAAAGAGAAATAAACACAACTGTCTT
TTGCCAACAAAAACCTATAAGTTTGATATATAGAATAGCTGATAAACCATATTATTAATAACAAATAAATTAGTTTATTGAATTTGTTGTA
TTTAATAAAATTAATGGTTAAGTTAACTTGTATTTTTAAAATAGTTTTATTACTATTATTATTATTATTTA

>VfSat20_TA
AAACTATATCTTTCTTTAATTGTGTTGGCTAAAATACACTAATCAATTTGTCATTCATTTTTAATTTGTAATTTAATATTTTATTATCTGA
AATCTTTTAATTAACTAATTATTAACTAATTTAATATTTATCAAAAATATTTTAATCAATTAATTTTTAACTAACTTAACTCATCATCCAT
TGGAAGATTAAAAGACATATATCACATTTTTTATAAAAAGATTCAAATATGTAAAAAAAAGAATATGAAATAAATTAAGGAGAAAACATGC
CTTCGATATCTTGCATCTGATTTAAGAATGATTGAACCATATTCATATAGATGTTTCTTTTATTCTTCTATTGCATCCGTTTTTGTTCACA
TTTTTCTTCAAGTCCTCTTTCAACCAATAGATTTCATACTGTTAATTCCTCTTCATCAAGTACTAAATCAACTCTTGAAGATGGTCGTTTA
TAGATAACTAACTCCCTTTATTTTTTGGCATCTAACTTATTTATCTTTTGTTGTTTTTTCTTCTTTATATAAGATAGTTTATTTATTATAA
GTCTCATTTTTCTGCCCAGTCTAGAACTTCCTTCTACATGTAAAATTTCAAAATATTTTCTCTTTTTCAACAAAAGACTAGATCTTTTCAA
CTTTGCAAAAGGGGTTAGCACTTTTATAAAACACATTATTTGTATTATAATTCTTTTCTTCTGATCACAACTAGAGACAAATATGACAATG
GAGTAATCTTCTGATAACTTTGCAGAGAATTATATTCAATGGCATGTTACCATGATCAATGGTTCAAAAATTGATATTTTCTAAAAATGGT
AGAGATATGAAAAGTCCATTATCATCAAATAATGGTCGCAGACACTTATATGCTCTTTTGACATATGGTTTTATTTGCTTCGGAGTTGAAG
GTTTTGGAGTGAAA

>VfSat21_TA
CTTTTAGATTTTTAAACGTCAATGATAGCATTAAAAAAACAAGAATTTGAGAAAAAGAAAATAAACTTTTGTAATTTGAAAGTAGACGATG
ATTCATGGTGGTGGTAGGAAGTCCATGATATGAGAAAAAAACTAAAAAACATTTATGTGTGAGTGGTACATTATATCATTTAGTACAATCC
TTTAATCTAATGGATCTTTTCCATGGTTTATCATGAACCATTTAGGTGACGACTAAAATTTTAATGTCTTTTATCATTATTTTGAAATAAA
TATTTAAAATTATAAATTTTTGACTATCCCTTATAAAGTTGAAGATAATTCATGTCAAATGAGTATCAATAAATTTATTATTATTTTACAA
ATTATTTATGCTAACTTTTTGGATGAGGGTTCAATTGAATCTCTCATGTCATCAGTGAACAATTCAACTTTACCATAAAATTAAACAAGTC
TTCATTATATATTTATATATACTGTATTAGATTAGATCTTTCCATGCTACCCCATTTTGTATTTCTGTAGTGACAAATATGATCACACTAT
TGTAATCTCCTTGCAAATTGATTATTATTGTTTTAGAATATCAATCCCATCAGAACAAAGCTAAGGCTCTGTTAGTTCCACCACCACATAT
TATTATTTTTATTTTACTTTAGACAACAAGTACACGACCTATGATTTTTTTTTGGAATCTAAATCTTTCCTTTTATGTTTTTTCAACTATA
CATTCATGCACGCTTCTATAAAATACCAAAAAACATATACACTAGTATATTTATTTATTTTTTCTGTTATCTAACTTGCCTCCTCTATTTT
CATTTGATGTAGTTTTTCACCCAAAGCATTAAAATATAAAAAGAAAATAGATGCCATAATATCAAATTCTAAACGTATTCATGATAACAAA
AATGTCAATTTTTTAAAAACAATATTCAAAAATCATTCAAGATGGATAATGAAATATAAGAGTGACACATTATGAAGTACTTAACATCATT
CAACCACTCACAAAGCTTGATATAATTTCCCAATCCAAA
ATTTTTATCAACAACTT

>VfSat22_TA
GATTATTAAGTACATCCACCTATCTCCTTTTAATGCTTTCTGAGTGGGGGAGATTACTAAGTAGACCCACATATCTGTCATTTGATGTTTT
CTCATATTACAAGTTCAATACGAATGATGTTAAAATATTATATGTTTGCCATGTTAACACTGCAAACATACTTCATACTGTATATTTCCAG
TTCATTTATACTAAAAAAAAAATCAATTAAATAGTCAAACTATTTTTTCTCTGTTTATATATTCAGTAGACGGAATTTATTATTCAATTTAA
TTTAAGATTTCGCTTTACTTCCTTATCTAAAAATCATTAAATTGTTCTGTTAGCATTATTGGTCCCTCCGTTAACTTTTCTGAAAATTATT
AACTTTTGCTAACATGGCATGACAACTAGGATACTCAGTCGGTAAATATATTAACAAAATTGATATTTTATATAAACTTGAGAGTATTGTT
TTCAGCGGTTAAAAACCCCCACAAATTGTTATTAATGTGTAAAAAAAAAGAGACAATTATTCTCCTTTCATTAATAATATAATTATTCTCC
TTTCAAATTTCAACGCTCCTTCAAGTATAACTGCATGTAAATATATTCATACATTAATACAATTCTTCTCCTTTCAAATTAAAAAACACTT
GGAGCTTTAAATACACCAAATACATAATTCAAAACAACACATGACAATAATGAAATTATTCAATTGGTAAATTAATAAAAATTGTCATCAA
GCAGAACAAGCTTGAGCAGGAGCCTCAACACTTTCATAATTAATGATTTCTTGTTACAAACACTTCATTATATAACCAGATTGACATGTTT
CTTTTCCTAAAACAAACACAAAATTACATTTAAGTAAACAAATAGCTACTATTCGCTTGGATTAGTGACAAATAGTGTTTCATATTTACCG
AGCATCCAACATTGACAACACTGCATTAAAAATCAAATACGACAACGTATTTAAGCAATTAAACCTCTGTATTACAAATTTGGTAATATGA
AATTAACTCATAATATATCTAATTGCTGAAACAACTTAATGCAAATAACAGGGTTGCAACCACAACTTGAACATCTGGTTGAAGGAGAACT
TTTTTAAACACATCAAACTTAAATTCCGATTTCTTTTCATTCTCATCTTGTGTTGTTTGCAACCTATAATCTTCTGCAAACAACCCAAAAC
AACATAAAACTCATAATAAATCAATAAAAGAATTTTTTTCTTCGATTTGATTCAGGAACCACAAACCCCAAAATTTGCATCAATTTAAAAT
TATCGGAAAAAAAACCCTTAACCTTAACCTTTTTTTGGAAAATTAAATAAAACTTAAAATTGAAAGTTAAAAGAAGATCCAAAATAACTAAT
```

```
CCAAACAGCGTTACCAACCACATCAAGGTAAGTAAAAGCAGGAGGAAGATGTAAAAAAGAAAACAAATGAAGATGAAGGATGGGAAGCAGG
TGAAGAGTTTATGGATGAGAATGAAGAGAAAGAGGGTGACGGCTGCTAGGGTTCTTGGAAGAAAAGAGAAAGAAGGTTTTCATAAATATCA
AAAAAACCACGTGATACGTTAAAAATCTATCACTGCCACATATGCAATAATTAACGTTGATTCTTGAAATTTAACAGAAATGACTAAAATA
ACTAACGGAAAAACATAAAGAAATAAAATAATAAATTAAATTAAGAAATTAAAAGATGAATTAAGTCGAATAAGAACAATCAATTTTATAG
AATAAGAATACGTGTTGATTTTTACATTAAAAAAACTTGGGGAATTTTTCTATTTCAAAAAAAGTATTATTAATCTTCATGCTATATTGGC
TAGAATGGTGGAGA

>VfSat23_TA
TTGAGAGATTTTCATGAAGGTTTGCAATTATTTAGAATGATAATGGTTCATAAAAAGACATAATTTTGAATTATGTACATGTCTTATTAAG
TCGGTTAAACGGCCACTTGATTTAACAAGTTTGCAACATTTTTTAAATTTCAAATAATTTTTCAAATTTCAAATAGCATCAGGTTTTATTA
AATCGGTTAATCAGGCCTGGCCTAAAAAGTGAAAACGACCTAACTTAAATATTTCTGGCAGCTTTTTAAAAATCAACTATTTTTATGCCTT
TATTTCAAAGATTTTTATGCATGACAATTTGAGAACTTTATATATGAAAATGTATGAACTACTTTGAGCATCTAAAGTTTATACATAATGA
ATTATAATTGTTTACCTTAGCATCAAGATCAATTTCCTTTACTTCCAAAAATATCCATAAGCTTATGCATTCTTGATTCCAAGCTGTTTCT
TCTTTGACATTTCTTTTTTGATAAAATTTTGTCTTCCTCAAAACTAAATTAAAATAGATGATGAACATCTTCTTCACAATGATTTGATGTC
AAATTTGATATTAAAGTTGATGCTTCTTCAAATCATCTTTCATTCAATGATATTTTTTGGAACTACTATATCATATGGTGTTGTATTTTGA
AGGATCAATCATGATTTGACAATATTAGCACATGATCTTCAAGTACCTACAAAAAATTAGTAATACTTAGGTACCCAATTTTCATTGGGTC
CTCTTGGGTTAGATACCTTCCTCACCCATACATACTCACCATAAGGAATACCATAATTTCTTACGTTGCAGACATTATAAGTATGGTCTTT
TGCATTACAATAAAAACAAATAGGTCTAAAAACATGATTCTTTTTATTGACTTAAAAGTTATTTCTATCATAAGGCTTTTTATGATGATTT
ACAACACGTTTTTTCTTTGAATCCTAATTGTTGAATTTCTTAGTAGCCTTTAGGAAGATAGTTGGACTTGTACTAGGTTTGTCAAAATTAA
AAAATTCAAGTCCAAATTTATTATTTGAAAATATTTGACTACATGATACATTATCCAAGCTAATTTCCCCTTTTTCATATTTTTCAACGAC
TTGGTTTAATTCCATAATTTTTTATTCAAAATAAGAACAATTATTGCATGCTGAATTTTTTATTAAGTCTAACTCCACTTTTGTGATATTA
ACCTCACTTTCTAGATTTAAAATAGTTTTATTTCGATTTAAAAATTTTCTAAAAAGTTTAAAAAAATCATCATGTAATTTATTAAAAGCAC
TTTGTAATTCATCATATGAAGGACTATCATCACTTGAGTTGATGTAAAAGT
```

Chapter II

Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe *Fabeae*.

Extraordinary Sequence Diversity and Promiscuity of Centromeric Satellites in the Legume Tribe *Fabeae*

Laura Ávila Robledillo,[1,2] Pavel Neumann,[1] Andrea Koblížková,[1] Petr Novák,[1] Iva Vrbová,[1] and Jiří Macas*,[1]
[1]Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic
[2]Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

***Corresponding author:** E-mail: macas@umbr.cas.cz.

Associate editor: Juliette de Meaux

Abstract

Satellite repeats are major sequence constituents of centromeres in many plant and animal species. Within a species, a single family of satellite sequences typically occupies centromeres of all chromosomes and is absent from other parts of the genome. Due to their common origin, sequence similarities exist among the centromere-specific satellites in related species. Here, we report a remarkably different pattern of centromere evolution in the plant tribe *Fabeae*, which includes genera *Pisum*, *Lathyrus*, *Vicia*, and *Lens*. By immunoprecipitation of centromeric chromatin with CENH3 antibodies, we identified and characterized a large and diverse set of 64 families of centromeric satellites in 14 species. These families differed in their nucleotide sequence, monomer length (33–2,979 bp), and abundance in individual species. Most families were species-specific, and most species possessed multiple (2–12) satellites in their centromeres. Some of the repeats that were shared by several species exhibited promiscuous patterns of centromere association, being located within CENH3 chromatin in some species, but apart from the centromeres in others. Moreover, FISH experiments revealed that the same family could assume centromeric and noncentromeric positions even within a single species. Taken together, these findings suggest that *Fabeae* centromeres are not shaped by the coevolution of a single centromeric satellite with its interacting CENH3 proteins, as proposed by the centromere drive model. This conclusion is also supported by the absence of pervasive adaptive evolution of CENH3 sequences retrieved from *Fabeae* species.

***Key words:* centromere evolution, satellite DNA, CENH3, ChIP-seq, plant chromosomes.**

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustav Molekularni Biologie Rostlin, Ceske Budejovice user on 09 February 2021

Introduction

Satellite DNA (satDNA) is a class of eukaryotic repetitive DNA characterized by its genomic organization into arrays of tandemly arranged units called monomers. It is most clearly distinguished from other tandemly repeated sequences by its formation of much longer arrays spanning up to megabases in length. Although monomer sizes of tens to a few hundred base pairs are predominant (Macas et al. 2002), satellite monomers can range from lengths typical for microsatellites (2–10 bp) (Heckmann et al. 2013; Talbert et al. 2018) to over 5 kb (Gong et al. 2012). Owing to its rapid sequence turnover, satDNA is the most evolutionarily dynamic component of the genome, as demonstrated by the dramatic variation in its abundance among species and the frequent emergence of species-specific repeat families (Garrido-Ramos 2017). In higher plants, satellite repeats may occur at subtelomeric or interstitial chromosomal regions, but they are preferentially located in, and often confined to, centromeres, especially in species with small genomes (Garrido-Ramos 2015; Oliveira and Torres 2018). Preferential association of satDNA with centromeric loci has also been reported for other lineages of eukaryotes (Plohl et al. 2014; Hartley and O'Neill 2019). However, the significance of this association for centromere maintenance and function, as well as the underlying mechanisms of satDNA accumulation in centromeres, remains incompletely understood.

Centromeres are chromosome regions that facilitate faithful chromosome segregation during cell division. This is achieved by providing an anchor point for assembly of the kinetochore, a protein complex connecting centromeric chromatin to the spindle microtubules (Cheeseman 2014). Consequently, centromeres have a number of features that distinguish them from other parts of the chromosomes. They are marked by the presence of the centromere-specific histone variant CENH3 and other proteins of the constitutive centromere-associated network (Hara and Fukagawa 2017). In addition, centromeres are regions of suppressed meiotic recombination and exhibit characteristic profiles of epigenetic chromatin modifications (Fuchs and Schubert 2012; Zhang, Dong, et al. 2014). It remains controversial whether and how these features drive the evolution of underlying centromeric sequences, especially the satellite repeats. Diverse hypotheses have been proposed on this issue, ranging from the idea that satDNA is a passive hitchhiker to the claim that it is a key determinant of centromere identity.

Perhaps the most influential concept regarding centromere evolution is the centromere drive hypothesis (Henikoff et al. 2001). Centromere drive is proposed to occur

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Mol. Biol. Evol. 37(8):2341–2356 doi:10.1093/molbev/msaa090 Advance Access publication April 7, 2020

in species with asymmetric female meiosis, in which homologous chromosomes compete for inclusion into the egg cell. Although the observed interspecific variation in repeat composition rules out the existence of a universal sequence determinant of centromere identity, this hypothesis still presumes that CENH3 or other kinetochore proteins interact with the centromeric satellites in a sequence-specific manner. Allelic expansion of the satellite array in one of the homologs then results in a stronger centromere, which binds more kinetochore proteins, thus facilitating its preferential transmission to the egg. On the other hand, such asymmetry leads to defects in male meiosis and reduced fertility, which is compensated for by changes in the CENH3 sequence that affect its DNA-binding preferences, resulting in restoration of meiotic parity. This evolutionary arms race between selfish centromeric DNA and its associated kinetochore proteins is predicted to result in diversification of centromeric repeats between species, as well as adaptive evolution of CENH3 or other kinetochore proteins that directly interact with the centromeric sequences (Henikoff et al. 2001; Malik 2009).

In line with the predictions of the centromere drive model, a single centromeric satellite whose sequence has diverged between related species has been reported in *Oryza* (Lee et al. 2005), *Medicago* (Yu et al. 2017), and some *Brassicaceae* species (Lermontova et al. 2014). Adaptive evolution of CENH3 proteins was detected in some of these species (Cooper and Henikoff 2004; Hirsch et al. 2009) as well as in several other taxa with asymmetric female meiosis (Zedek and Bureš 2016). Direct evidence for centromere drive was obtained in the plant genus *Mimulus* (Finseth et al. 2015), the fly *Drosophila melanogaster* (Wei et al. 2017), and mouse (Iwata-Otsubo et al. 2017) in which the molecular mechanisms underlying centromere drive have also been elucidated (Akera et al. 2017).

On the other hand, considering the widespread occurrence of centromeric satellites in plant and animal genomes, it is surprising that so few examples of centromere drive have been reported so far. Moreover, some observations are not consistent with the presumed evolutionary arms race between CENH3 and its underlying centromeric satellite(s) (Kawabe et al. 2006; Masonbrink et al. 2014). In addition, maize lines carrying homologous chromosomes with different centromere sizes exhibit no significant distortions in their meiotic segregation (Han et al. 2018), and CENH3 proteins from the phylogenetically distant species *Lepidium oleraceum* and *Zea mays* exhibit binding patterns on *Arabidopsis thaliana* centromeres that were indistinguishable from native CENH3 (Maheshwari et al. 2017). This may indicate that the process is not as common as expected, or that it is active only during limited periods of centromere evolution.

Although the considerations described earlier are mainly based on the presumed sequence-specific interactions of kinetochore proteins with their underlying sequences, it has recently become evident that features other than primary sequence may also be important for the coevolution of satDNA and centromeres. Specifically, it has been proposed that the repeated structure itself is advantageous, as homologous recombination between identical repeat copies generates DNA loops that are required for efficient centromere function (McFarlane and Humphrey 2010). In addition, centromere propagation and function seems to depend upon transcription of its sequences (Duda et al. 2017; Perea-Resa and Blower 2018); thus, the ability of centromeric satellites to produce transcripts at optimal levels may determine their fate in these regions. Finally, in domesticated maize inbred lines, the sequence composition of centromeres can be shaped by inbreeding and selection for centromere-linked genes, a process that may also act during speciation in natural systems (Schneider et al. 2016). Therefore, it is likely that the structure and sequence composition of centromeres in a particular species reflects an interplay of various structural features and evolutionary forces, the nature and importance of which are yet to be determined.

The questions outlined above could be answered by gathering information on centromeric sequences and kinetochore proteins from a wide range of species and examining them in the phylogenetic context. Because it is important to discriminate sequences that are truly associated with centromeric chromatin from surrounding repeats, it would be necessary to perform chromatin immunoprecipitation using antibodies against centromeric proteins coupled with sequencing of retrieved DNA (ChIP-seq). To date, however, relatively few such studies have been conducted in plants, and most have focused on one (Gong et al. 2012; Zhang, Kobližkova, et al. 2014; Kowar et al. 2016) or a small group of species (Gent et al. 2017).

In our previous work, we analyzed centromeric repeats in garden pea (*Pisum sativum*), a species with peculiar centromere organization consisting of multiple separated domains of CENH3 chromatin arranged along extended primary constrictions of metaphase chromosomes (Neumann et al. 2012). This “meta-polycentric” chromosome organization was later reported in *Lathyrus*, but not in genera *Vicia* and *Lens*, which are phylogenetically related members of the same legume tribe, *Fabeae* (Neumann et al. 2015). In *P. sativum*, ChIP-seq experiments with CENH3 antibody revealed unprecedented diversity of centromeric satellites consisting of 13 repeats with different distribution patterns among chromosomes (Neumann et al. 2012). Relative to *Vicia* and *Lens*, *Pisum* and *Lathyrus* species possess an additional copy of the *CENH3* gene, which was speculated to serve as a possible trigger for the expansion of centromeres and the emergence of diverse centromeric satellites. However, subsequent study of *Vicia faba*, a species with simple centromeres and only one copy of *CENH3*, also revealed multiple centromeric satellites, three of which are present in the same centromere, whereas the other four are chromosome-specific (Ávila Robledillo et al. 2018). Therefore, *Fabeae* is a taxon with unusual distribution patterns and possibly highly dynamic turnover of centromeric repeats.

Prompted by these results, in this study, we focused on characterization of centromeric satellites across the whole *Fabeae* tribe, investigating 15 species in addition to the two analyzed previously. In these experiments, we employed a set of CENH3 antibodies (Neumann et al. 2012, 2015) to perform ChIP-seq in these species and identified centromeric satellites

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

using repetitive sequences characterized by graph-based clustering of genomic reads (Macas et al. 2015) as the reference. As previously demonstrated (Zhang, Koblížkova, et al. 2014; Ávila Robledillo et al. 2018), this approach provides comprehensive information about centromere-associated repeats without the need for an assembled reference genome and as such is suitable for nonmodel species. The identified repeats were further investigated by a combination of fluorescence *in situ* hybridization (FISH) with immunodetection of CENH3 proteins to confirm their centromeric localization and map their distribution among chromosomes. The experiments revealed an extraordinary diversity of centromeric satellites in *Fabeae*, their irregular distributions among species, and unexpected localization patterns of some of these repeats on the chromosomes. Finally, we analyzed these findings with respect to the sequence diversity and evolution of *CENH3* genes in *Fabeae*.

Results

ChIP-seq Analysis Reveals Unprecedented Diversity of Centromeric Satellites in *Fabeae*

To investigate the repeat composition of *Fabeae* centromeres, we sequenced and analyzed DNA fragments retrieved from centromeric chromatin immunoprecipitated with CENH3 antibody. These experiments were performed with a set of 15 species selected to represent all major evolutionary lineages of *Fabeae*, as described by Schaefer et al. (2012). To verify the species phylogeny, we calculated a maximum likelihood (ML) tree based on *matK–rbcL* sequences for the selected set, supplemented with seven additional *Fabeae* species in which centromeric satellites and/or *CENH3* gene sequences have been characterized previously (Neumann et al. 2012, 2015; Ávila Robledillo et al. 2018). The resulting tree topology (fig. 1*A*) was in general agreement with a previously reported tree (Schaefer et al. 2012), confirming that *Pisum* and *Lathyrus* are closely related and form a separate lineage, whereas the other major lineage consists of most *Vicia* species along with *Lens culinaris*; in addition, a separate group of two *Vicia* species (*V. ervilia* and *V. hirsuta*) is basal to all *Fabeae*.

The ChIP-seq experiments were performed using antibodies raised against CENH3 proteins from *V. faba*, *P. sativum*, *Lathyrus sativus*, and *Le. culinaris* (supplementary table 1, Supplementary Material online), previously shown to specifically label centromeric chromatin in a number of *Fabeae* species (Neumann et al. 2012, 2015). The immunoprecipitated DNA fragments were sequenced on the Illumina platform along with control DNA samples extracted from chromatin preparations prior to ChIP (input control). Centromeric repeats were then identified as sequences enriched in the ChIP sample relative to the input. Enrichment of all repeats representing at least 0.01% of the genome was evaluated by similarity-based mapping of ChIP and input reads onto the reference repeat sequences previously generated for individual *Fabeae* species using the RepeatExplorer pipeline (Macas et al. 2015). In the two species for which a reference was not available (*L. niger* and *L. clymenum*), the ChIP and input reads were used directly for comparative RepeatExplorer analysis (Novák et al. 2013), and the enrichment was calculated as a ratio of ChIP to input reads in individual repeat clusters.

Centromeric satellites were identified in 12 out of the 15 investigated species as sequences with ChIP/input ratios between 3 and 333 (table 1). Such enrichment in the ChIPed fraction was revealed for up to eight satellites per species, whereas the majority of investigated repeats exhibited no enrichment (supplementary fig. 1, Supplementary Material online). In five species, one to five additional nontandem sequences, mostly classified as putative LTR-retrotransposons or unknown repeats, were ChIP-enriched (supplementary table 2, Supplementary Material online), whereas in the rest of the species, all enriched repeats represented satellites. In three *Vicia* species, *V. narbonensis*, *V. ervilia*, and *V. hirsuta*, we identified no ChIP-enriched repeats. Hence, we performed additional experiments to verify that the antibodies used for the ChIP recognize centromeric chromatin. In all three cases, we observed relatively weak but specific immunostaining of primary constrictions of isolated metaphase chromosomes (supplementary fig. 2, Supplementary Material online). These results may reflect a lack of centromere-enriched repetitive sequences in these three species; however, we cannot rule out the possibility that the antibodies failed specifically in the ChIP reaction because the conditions differed from those used for chromosome immunostaining.

Including the previously reported centromeric satellites from *P. sativum* (Neumann et al. 2012) and *V. faba* (Ávila Robledillo et al. 2018), we identified a total of 64 centromeric satDNA families in *Fabeae*. In most species, we detected multiple centromeric satellites, and none of these repeats was shared across all species. The basic characteristics of centromeric satellites are summarized in table 1 and their consensus monomer sequences are provided in supplementary file 1, Supplementary Material online. Monomer lengths varied considerably (33–2,979 bp), as did their nucleotide composition (50–79% AT). To evaluate sequence similarities that could point to a common origin of centromeric satellites from different species, we compared the monomer sequences using alignment-free similarity measures as defined by D_2^* statistics (Reinert et al. 2009). We also performed these analyses on the complete set of 430 putative satellite repeats predicted previously for the investigated species (Macas et al. 2015) to detect similarities between centromeric and noncentromeric satellites. The results revealed that most satellite repeat families, regardless of their association with centromeres, were species-specific (supplementary fig. 3*A*, Supplementary Material online). A subset of the repeats exhibited sequence similarities that led to the definition of 13 superfamilies that included centromeric satellites and consisted of the families present in two to five species (fig. 1*B*). Although satellites assigned to the same superfamily exhibited significant similarities, some families had sequence variations, especially with respect to monomer size (supplementary fig. 3*B*–*D*, Supplementary Material online and table 1). Moreover, some of the centromeric superfamilies

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

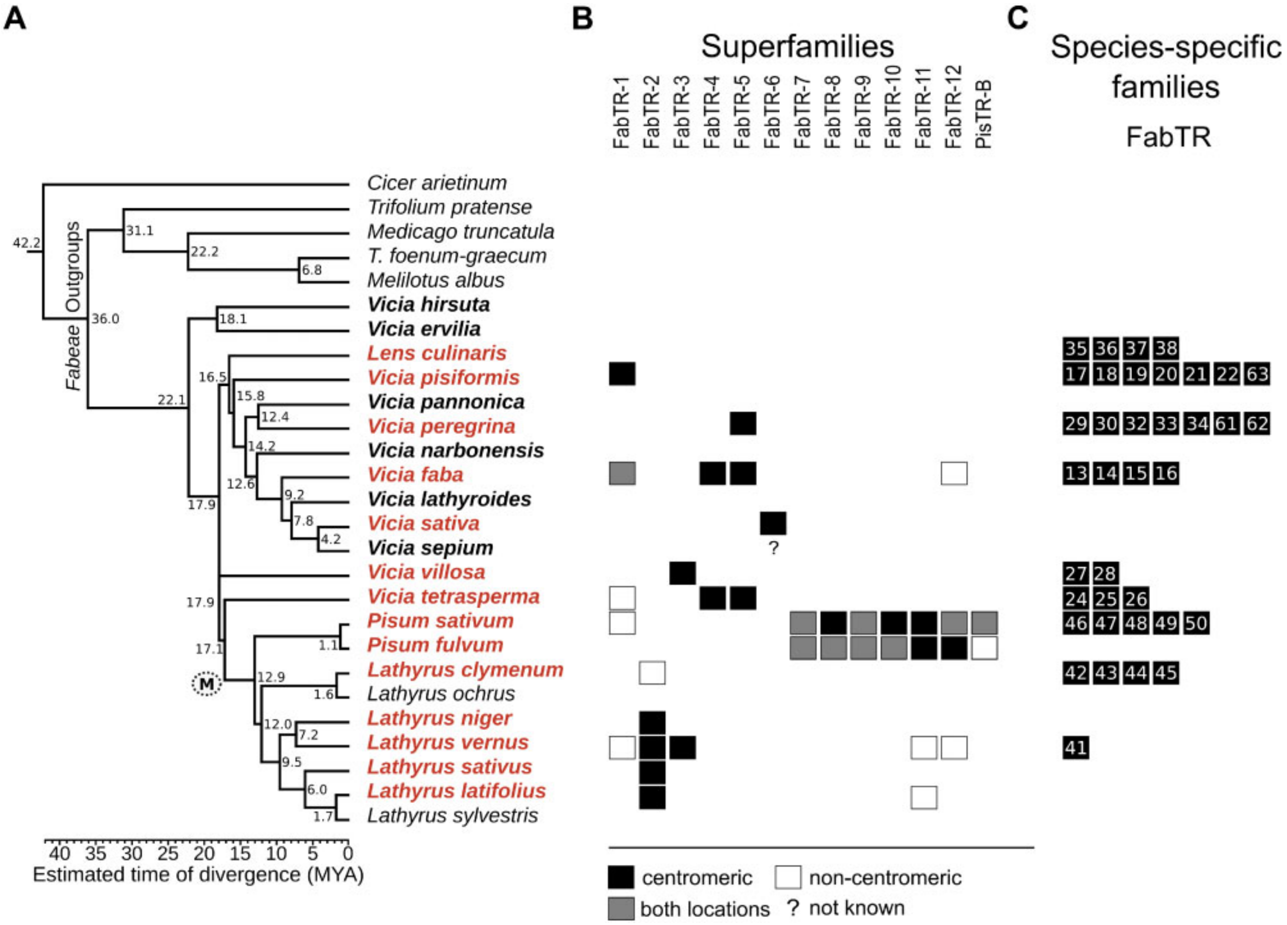


Fig. 1. Overview of centromeric satellite families identified in *Fabeae*. Species are arranged based on their phylogenetic distances inferred from a comparison of *matK–rbcL* sequences using the maximum likelihood algorithm (*A*). The tree was rooted using five species representing related legume genera as outgroups. Numbers represent estimated node ages in million years ago (MYA), and correspond to the divergence time scale below the tree. The branch leading to the species with meta-polycentric chromosomes is marked with (M). Names of *Fabeae* species in which satellite repeats were identified using CENH3 ChIP-seq are printed in red, whereas species not analyzed by ChIP but included in the similarity searches are printed in bold. (*B* and *C*) The presence of individual satellite families in analyzed species is indicated by squares. Black squares indicate families associated with centromeric chromatin, as revealed by their enrichment in the CENH3 ChIP-seq experiments. The centromeric satellites that simultaneously occur in the genome as additional, noncentromeric loci (revealed by FISH) are marked with gray squares, whereas those present in the respective species but not enriched in ChIP-seq experiments are marked with empty squares. The question mark in FabTR-6 column indicates that this repeat is present in *Vicia sepium* genome but was not investigated by ChIP-seq in this species. (*B*) The satellite families from different species displaying sequence similarities are grouped into superfamilies and arranged in columns labeled with the superfamily name. (*C*) Numbers of species-specific families are symbolized by squares in each row, ranging from one in *Lathyrus vernus* to seven in *V. pisiformis* and *V. peregrina*. Numbers within the squares refer to the family names (FabTR-numbers) listed in table 1.

included satellites that were ChIP-enriched in only some species, but were not centromeric in the rest (supplementary fig. 3*B* and *C*, Supplementary Material online and fig. 1*B*).

Most Species Possess Multiple Centromeric Satellites That Are Often Species-Specific

The families of centromeric satellites that we identified were unevenly distributed among *Fabeae* species. A large fraction of families (37 of 64; 58%) were species-specific (fig. 1*C*), whereas the remaining repeats belonged to satellite superfamilies shared by several species (fig. 1*B*). In the phylogenetic lineages including *Vicia* spp. and *Le. culinaris*, all species but one (*V. sativa*) possessed multiple (two to eight) centromeric satellites, with up to seven species-specific satellites in *V. pisiformis* and *V. peregrina* (fig. 1*C*). The largest number of centromeric satellites, 12, occurred in the *P. sativum* genome; however, only six of these satellites were shared with its sister species *P. fulvum* (fig. 1*B* and *C*).

Three of the investigated *Lathyrus* species, *L. sativus*, *L. latifolius*, and *L. niger*, possessed single centromeric satellites that were classified as members of the same superfamily, FabTR-2. The same centromeric repeat was also identified in closely related *L. vernus*; however, this species possessed two additional, albeit far less abundant centromeric satellites (table 1). The existence of a single-dominant centromeric satellite in these three *Lathyrus* species contrasted with the situation in the remaining species, *L. clymenum*, in which FabTR-2 sequences were also present but were not associated with centromeric chromatin. Instead, four species-specific centromeric satellites were identified in this species (fig. 1).

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

Table 1. Satellite Repeats Associated with Centromeric Chromatin.

Species	Family	Superfamily	Monomer (bp)	AT (%)	Genome (%)	ChIP Enrichment	Previous Code
Lens culinaris	FabTR-35-LNS		579	73.7	0.061	52.01	
	FabTR-36-LNS		1,086	76.5	0.058	40.13	
	FabTR-37-LNS		967	74.3	0.032	34.55	
	FabTR-38-LNS		1,315	74.9	0.019	40.19	
Vicia pisiformis	FabTR-17-VPF		580	71.6	0.0317	5.0	
	FabTR-18-VPF		2,087	63.5	0.199	20	
	FabTR-63-VPF		61	60.7	0.158	57.13	
	FabTR-19-VPF		84	69	0.101	53.78	
	FabTR-20-VPF		763	64	0.06	79.4	
	FabTR-1-VPF	FabTR-1	72	62.9	0.043	3.0	
	FabTR-21-VPF		778	74.3	0.039	16	
	FabTR-22-VPF		1,793	74.3	0.017	6.7	
Vicia peregrina	FabTR-5-VPR	FabTR-5	114	74.1	0.344	60	
	FabTR-29-VPR		718	73.8	0.29	67.48	
	FabTR-30-VPR		898	73.8	0.267	91.59	
	FabTR-32-VPR		1,189	63.6	0.115	333.2	
	FabTR-33-VPR		30	56.7	0.087	22.35	
	FabTR-34-VPR		324	76	0.083	55.74	
	FabTR-61-VPR		569	68.5	0.031	130.8	
	FabTR-62-VPR		1,244	71.9	0.013	10.58	
Vicia faba	FabTR-1-VFB	FabTR-1	50	64	0.132	103.6	VfSat6
	FabTR-5-VFB	FabTR-5	44	70.5	0.102	103.2	VfSat7
	FabTR-4-VFB	FabTR-4	2,033	71.7	0.061	91.3	VfSat8
	FabTR-13-VFB		1,762	74.7	0.042	41.2	VfSat10
	FabTR-14-VFB		47	68.1	0.036	149.2	VfSat13
	FabTR-15-VFB		1,712	65.1	0.038	109.9	VfSat16
	FabTR-16-VFB		1,325	73.3	0.008	81.9	VfSat23
Vicia sativa	FabTR-6-VSA	FabTR-6	624	67.3	0.101	62.3	
Vicia villosa	FabTR-27-VVL		156	64.7	2.063	9.7	
	FabTR-3-VVL	FabTR-3	602	75.9	0.226	80.9	
	FabTR-28-VVL		1,792	67.3	0.053	92.9	
Vicia tetrasperma	FabTR-24-VTS		959	66.9	0.361	59.01	
	FabTR-4-VTS	FabTR-4	1,614	69.8	0.254	69.62	
	FabTR-25-VTS		33	69.7	0.069	100	
	FabTR-26-VTS		470	64.5	0.055	72.11	
	FabTR-5-VTS	FabTR-5	44	63.6	0.045	103.21	
Pisum sativum	FabTR-7-PST	FabTR-7	867	77	0.02	51.7	TR1
	FabTR-8-PST	FabTR-8	244	76.6	0.01	59.3	TR6
	FabTR-46-PST		164	72.6	0.124	49.7	TR7
	FabTR-9-PST	FabTR-9	658	74.5	0.01	76.3	TR10
	FabTR-10-PST-A	FabTR-10	459	75.4	0.127	65.9–74.9	TR11-TR19
	FabTR-10-PST-B	FabTR-10	1,975	76.6	0.127	65.9–74.9	TR11-TR19
	FabTR-47-PST		105	69.5	0.013	5.4	TR12
	FabTR-11-PST	FabTR-11	1,637	74.3	0.012	82.5	TR18
	FabTR-12-PST	FabTR-12	844	78	0.179	50.7	TR20
	FabTR-48-PST		613	71.6	0.013	44	TR21
	FabTR-49-PST		882	76.2	0.003	102.9	TR22
	FabTR-50-PST		1,812	69.9	0.087	10.7	TR23
	PisTR-B	PisTR-B	50	72	1.26	20.5	
Pisum fulvum	FabTR-7-PFL	FabTR-7	864	77	0.059	42.1	TR1
	FabTR-8-PFL	FabTR-8	242	77.3	0.033	45.9	TR6
	FabTR-9-PFL	FabTR-9	659	74.7	0.044	73.5	TR10
	FabTR-10-PFL-A	FabTR-10	502	76.9	0.236	53.4	TR11-TR19
	FabTR-10-PFL-B	FabTR-10	2,170	76.7	0.236	53.4	TR11-TR19
	FabTR-11-PFL	FabTR-11	2,979	74.2	0.009	98.3	TR18
	FabTR-12-PFL	FabTR-12	864	73.6	0.01	52.8	TR20
Lathyrus clymenum	FabTR-42-LACLM		36	61.1	4.119	78.31	
	FabTR-43-LACLM		30/60/70	60	0.805	60.29–98.18	
	FabTR-44-LACLM		60	50	0.977	57.21	
	FabTR-45-LACLM		102	57.4	0.162	76.39	
Lathyrus niger	FabTR-2-LNGER	FabTR-2	49/50/100	75.5	0.069	86.6–98.27	
Lathyrus vernus	FabTR-2-LAV	FabTR-2	49	77.6	0.584	62.1	
	FabTR-3-LAV	FabTR-3	972	77.8	0.022	3.08	
	FabTR-41-LAV		54	79.2	0.017	8.34	
Lathyrus sativus	FabTR-2-LAS	FabTR-2	49	73.5	1.679	38.53	
Lathyrus latifolius	FabTR-2-LAL	FabTR-2	49	73.5	1.228	46.54	

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

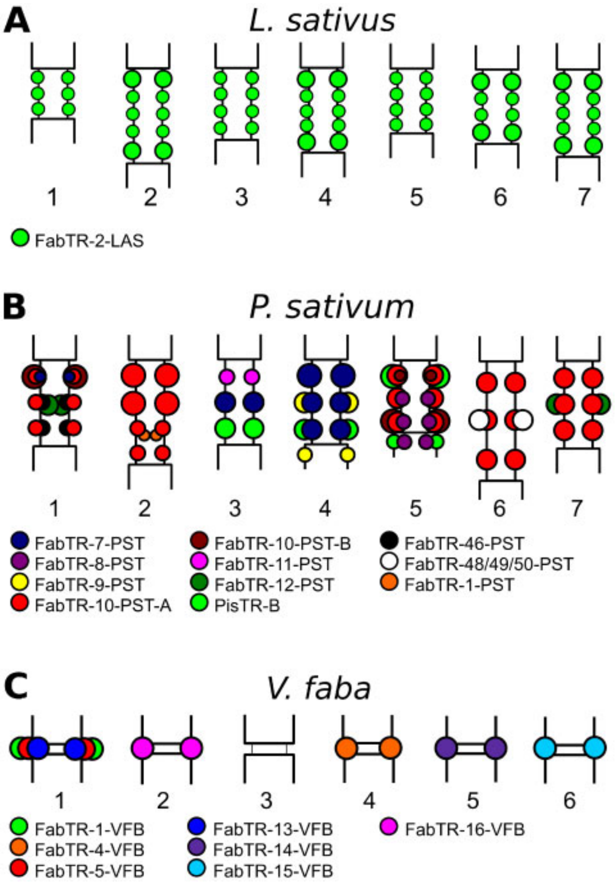


FIG. 2. Schematic representation of the satellite repeat distribution in centromeric regions of (*A*) *Lathyrus sativus* ($n = 7$), (*B*) *Pisum sativum* ($n = 7$), and (*C*) *Vicia faba* ($n = 6$) chromosomes. Different families of satellite repeats are distinguished by colors according to the legend provided for each species. In meta-polycentric chromosomes (*A* and *B*), the satellite loci associated with CENH3 chromatin are located at the outer periphery of the primary constrictions, whereas those located within the inner regions of *P. sativum* constrictions lack CENH3.

Next, we used FISH combined with immunodetection of CENH3 proteins to confirm ChIP-seq results and investigate the genome distribution of the selected satellite sequences. Contrasting patterns of centromeric satellite distributions were revealed, some of which are schematically depicted on figure 2. When applied to *L. sativus*, a species containing FabTR-2 as the single centromeric satellite, these experiments confirmed the location of this repeat in all domains of centromeric chromatin distributed along the primary constrictions of the chromosomes (fig. 2*A* and supplementary fig. 4*A*–*D*, Supplementary Material online). In *L. vernus*, the experiment revealed identical patterns of FabTR-2 colocalization with CENH3 chromatin (supplementary fig. 4*E*–*H*, Supplementary Material online), whereas the two additional ChIP-enriched satellites identified for this species were detected as minor loci overlapping with FabTR-2 signals (data not shown). In species with large numbers of centromeric satellite families, these families were unevenly distributed between the chromosomes, as shown in *P. sativum* (this work and Neumann et al. [2012]) (fig. 2*B*). The same pattern was also observed for *P. fulvum*, as none of its six centromeric satellites occurred on all chromosomes (data not shown). Similar types of distribution patterns are also likely shared by *Vicia* species with high diversity of centromeric satellites. For example, all seven centromeric satellites in *V. faba* were chromosome-specific (fig. 2*C*), and FISH localization of two randomly chosen centromeric satellites in *V. peregrina* revealed their presence in centromeres of four (FabTR-30) and one (FabTR-32) of the seven pairs of chromosomes (supplementary fig. 4*I* and *J*, Supplementary Material online).

Association of Some Satellites with Centromeric Chromatin Differs between Species or Even between Chromosomes of the Same Species

A striking feature of some satellite superfamilies was their association with centromeric chromatin in some species, but no enrichment in CENH3 ChIP-seq experiments in the others, suggesting that they were absent from centromeres in these genomes. This pattern was found for five superfamilies: FabTR-1, 2, 11, 12, and PisTR-B (fig. 1). To obtain better insight into their genomic distribution, we performed FISH on metaphase chromosomes, as shown for FabTR-1 repeats in figure 3. FabTR-1 was ChIP-enriched in *V. pisiformis* and *V. faba*, and corresponding FISH signals were detected in centromeres of two chromosome pairs in *V. pisiformis* (fig. 3*A*) and in the centromere of chromosome 1 of *V. faba*. An additional minor noncentromeric signal was present within the long arm of

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

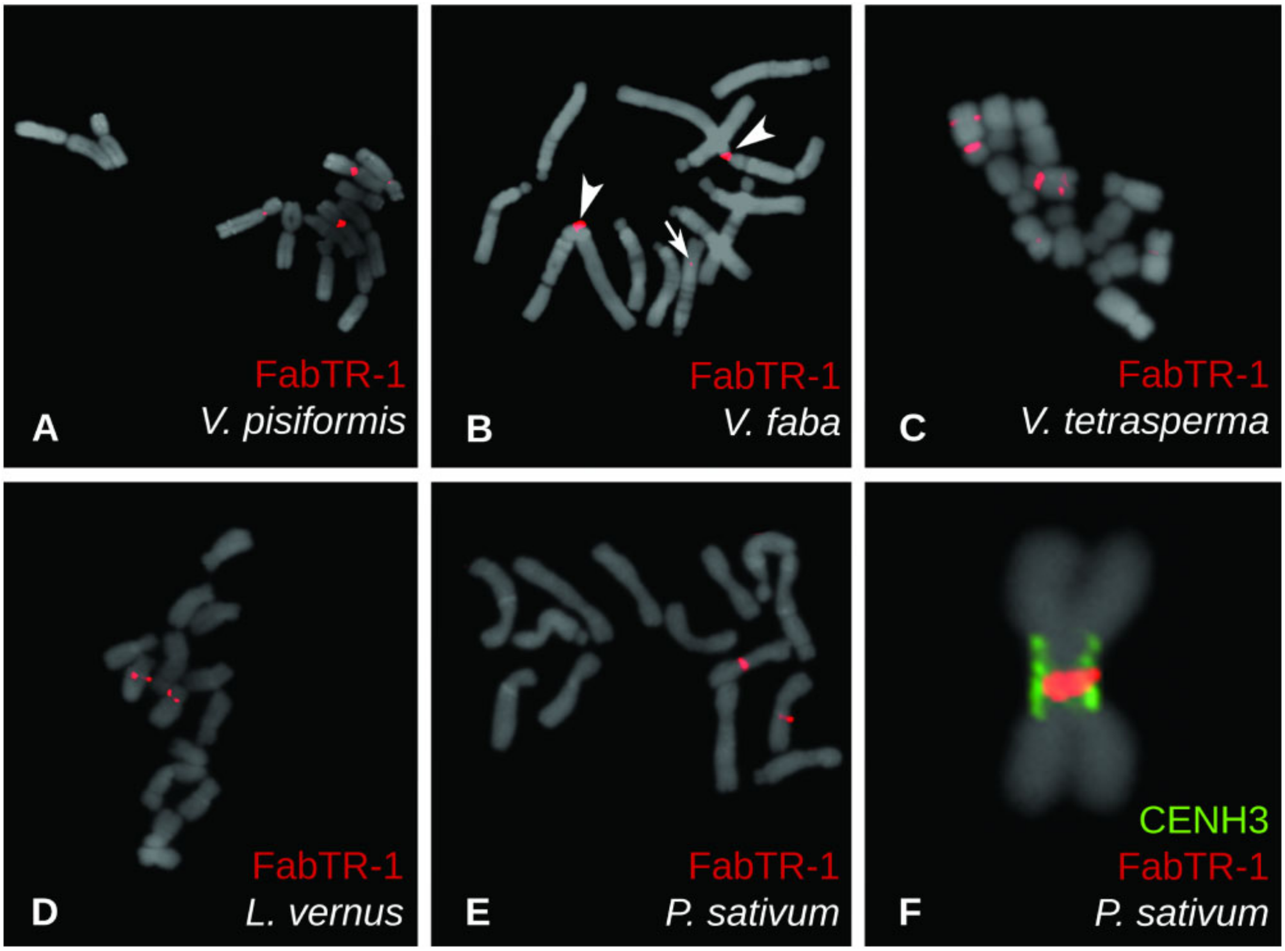


Fig. 3. Localization of FabTR-1 repeats on metaphase chromosomes of five *Fabeae* species. Repeats were detected using FISH (red signals), showing signals within centromeres of two chromosome pairs in *Vicia pisiformis* (*A*) and one pair in *V. faba* (*B*). A minor noncentromeric signal on *V. faba* chromosome 6 is marked with an arrow. Two pericentromeric and one interstitial signal were detected in *V. tetrasperma* (*C*), whereas *Lathyrus vernus* (*D*) and *Pisum sativum* (*E*) exhibited signals adjacent to or within primary constrictions of one pair of chromosomes. Closer examination of *P. sativum* chromosomes using a combination of FISH (red) with immunolabeling of CENH3 proteins (green) revealed that FabTR-1 is located within the inner part of the primary constriction, apart from the CENH3 chromatin located along the constriction periphery (*F*). Chromosomes counterstained with DAPI are shown in gray.

chromosome 6 of *V. faba* (fig. 3*B*). In the remaining three species, *V. tetrasperma*, *L. vernus*, and *P. sativum*, the repeat is not associated with CENH3 chromatin; in all of them, however, it was found to be located close to the centromeres. In *V. tetrasperma*, FabTR-1 signals almost entirely overlapped with the primary constrictions on two chromosome pairs, and an additional repeat locus was revealed within the long arm of one of these chromosomes (fig. 3*C*). One FabTR-1 locus close to the centromere of one chromosome pair was identified in *L. vernus* (fig. 3*D*). In *P. sativum*, the signal was located directly within the extended primary constriction of chromosome 2 (fig. 3*E*). Detailed examination of metaphase chromosomes employing simultaneous immunodetection of CENH3 revealed that FabTR-1 is located within the inner part of the constriction close to the chromosome axis, whereas the CENH3 chromatin is located on the periphery of the constriction (fig. 3*F*). These findings confirmed that despite its presence in the centromeric region, the repeat is not associated with centromeric chromatin, consistent with the results of the ChIP-seq experiments.

The existence of additional noncentromeric loci containing centromeric satellites was confirmed for most superfamilies shared by the two *Pisum* species (fig. 1*B*). In some cases, this pattern was combined with the presence of the repeat in additional species. For example, FabTR-12 was centromeric in both *Pisum* species, but noncentromeric in *V. faba* and *L. vernus*. In all four species, the repeat was located on two pairs of chromosomes, but was fully associated with CENH3 chromatin only in *P. fulvum* (fig. 4*A* and *B*). In *P. sativum*, FabTR-12 signals overlapped with CENH3 chromatin only on chromosome 7 (fig. 4*D*), whereas the FISH signals on chromosome 1 were located within the inner part of the constriction (fig. 4*C*), similar to FabTR-1 on chromosome 2 (fig. 3*F*). In *V. faba*, the repeat was located within long arms of two chromosome pairs (fig. 4*E*), whereas in *L. vernus* it was adjacent to primary constrictions (fig. 4*F*). Yet another interesting

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

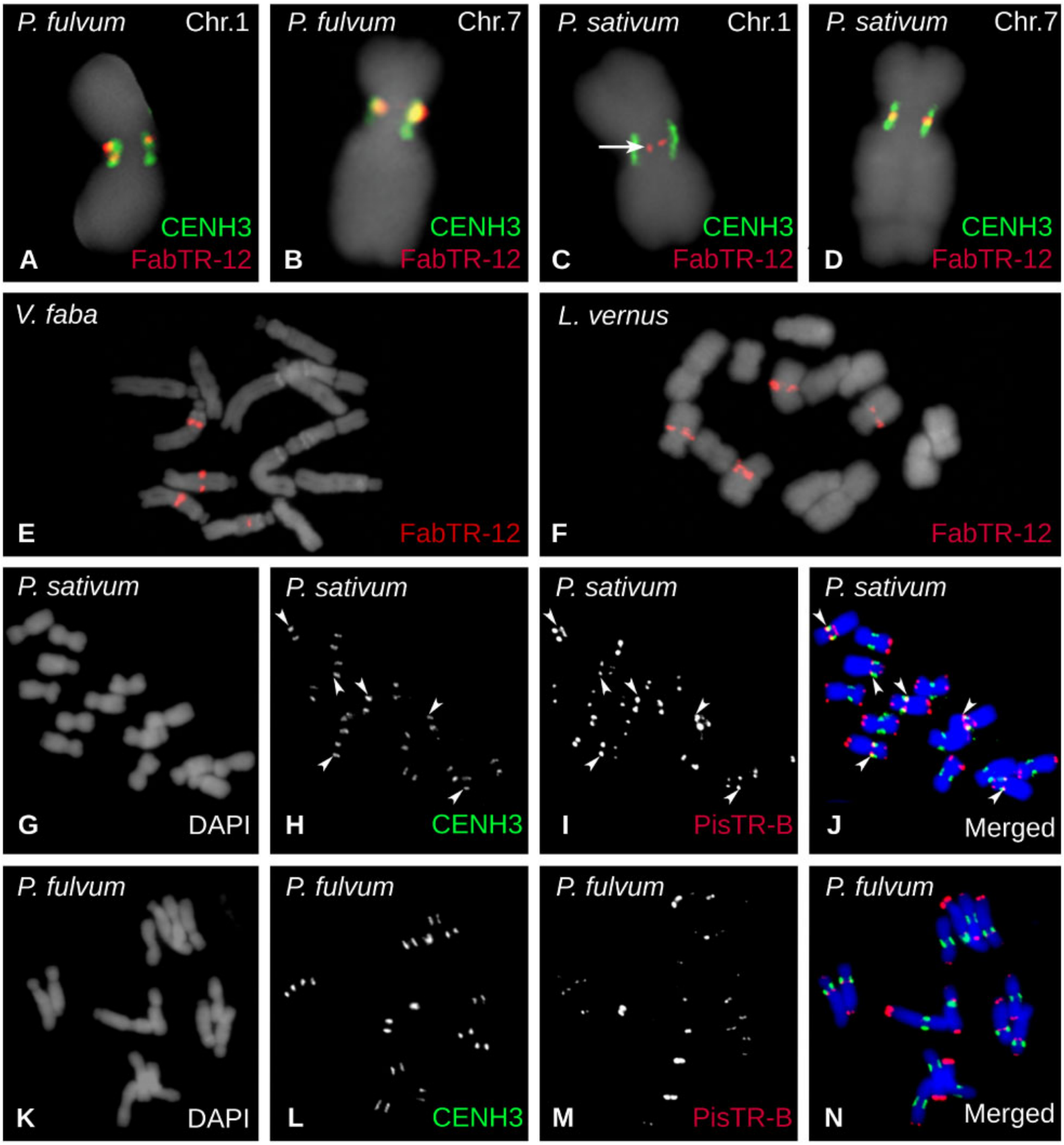


Fig. 4. Localization of FabTR-12 and PisTR-B repeats on metaphase chromosomes. Repeats were detected using FISH (red) alone or in combination with immunolabeling of CENH3 (green signals). (*A*–*D*) FISH detection of FabTR-12 showing signals overlapping with CENH3 loci on chromosomes 1 and 7 of *Pisum fulvum* and on chromosome 7 of *P. sativum*. On the contrary, FabTR-12 signals were located apart from the CENH3 chromatin on *P. sativum* chromosome 1 (arrow). In *Vicia faba* (*E*) and *Lathyrus vernus* (*F*), the repeat was also present on two chromosome pairs, but the signals were not centromeric and were instead located within the long chromosome arms. (*G*–*N*) Distribution of PisTR-B repeats on chromosomes of the two *Pisum* species. There are three centromeric PisTR-B loci (arrowheads) that colocalize with CENH3 in *P. sativum* (*G*–*J*); however, this satellite is not associated with the centromeric chromatin in *P. fulvum* (*K*–*N*).

example of such distribution is the major *Pisum* satellite PisTR-B (Neumann et al. 2001) which in *P. sativum* is associated with centromeric chromatin on chromosomes 3, 4, and 5, whereas most of its loci are distributed in pericentric and subtelomeric regions (fig. 4*G*–*J*). Although of similar genomic abundance and chromosomal distribution, it is not associated with centromeres in *P. fulvum* (figs. 1*B* and 4*K*–*N*).

CENH3 Genes Evolved Mainly under Purifying Selection

To determine whether centromeric repeat composition is correlated with the mode of evolution of *CENH3* genes, we performed a phylogenetic analysis of CENH3 coding sequences. In our previous study (Neumann et al. 2015), we found two *CENH3* variants in *Fabeae* that differed significantly,

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

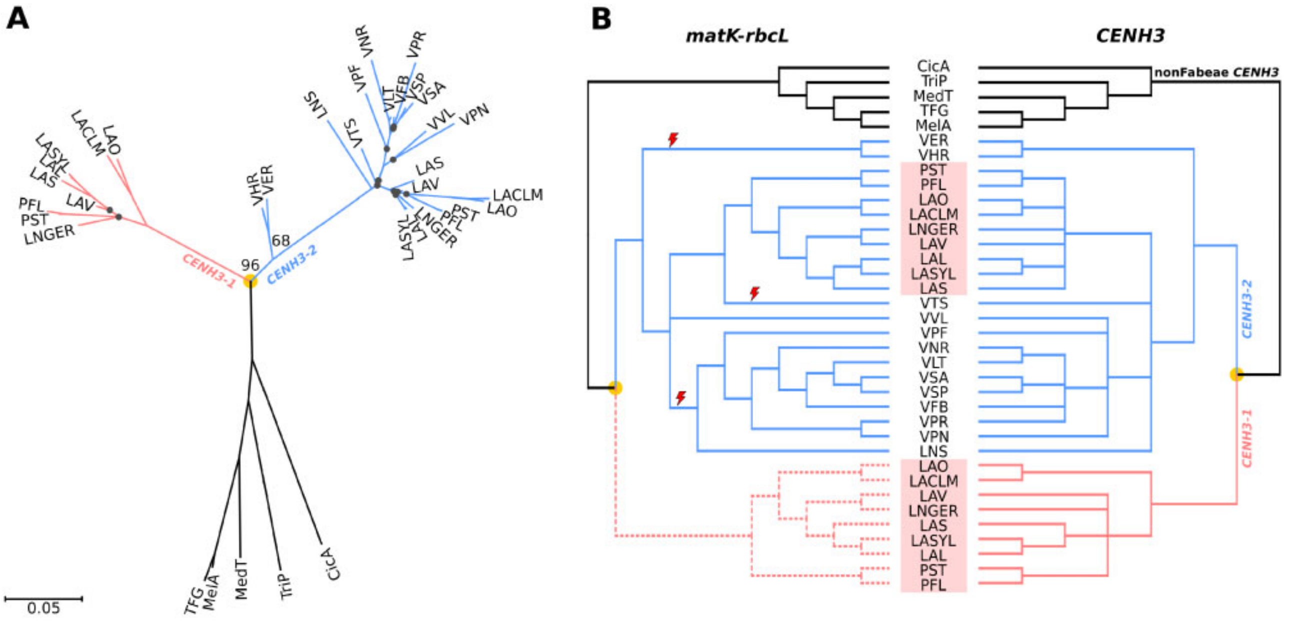


Fig. 5. Phylogenetic trees of CENH3 sequences. (*A*) Phylogenetic tree inferred from the alignment of CENH3-coding sequences using the maximum likelihood method, excluding the INDEL region near the 5′ end (see supplementary fig. 5, Supplementary Material online). Bootstrap values are shown only for key nodes. Black dots indicate nodes with low bootstrap support (<50). The scale bar shows genetic distance. (*B*) Tanglegram showing comparison of the CENH3 tree from the panel (*A*) with the species tree inferred from *matK–rbcL* shown in figure 1*A*. Nodes with low bootstrap support (<50) were collapsed in both trees. The part of the *matK–rbcL* tree depicted by dashed lines was manually added to the tree to show comparison of phylogenies inferred from *matK–rbcL* and CENH3-1, and to allow the use of the *matK–rbcL* tree for analysis of positive selection in CENH3 genes. Red lightning symbols mark three independent losses of CENH3-1 genes. *Pisum* and *Lathyrus* species are highlighted by red rectangles. Orange dots indicate CENH3 duplication events.

particularly in their N-terminal regions. Although the *CENH3-2* variant is shared by all species within the tribe, *CENH3-1* occurs as an additional gene only in the *Pisum*/*Lathyrus* lineage. To better date the *CENH3* duplication event, we identified *CENH3* sequences in four additional species representing the basal group (*V. ervilia*, *V. hirsuta*) or less-represented parts of the *Fabeae* phylogenetic tree (*V. pisiformis*, *V. tetrasperma*) and analyzed them in combination with 32 CENH3-coding sequences identified previously. The results revealed that all four new sequences belong to *CENH3-2*, and that *CENH3-1* and *CENH3-2* diverged before radiation of the *Fabeae* species included in this study (fig. 5). Because *V. ervilia* and *V. hirsuta* represent the clade that split earliest from all other *Fabeae* (Schaefer et al. 2012; fig. 1*A*), it is likely that the duplication occurred in an ancestor of all *Fabeae*. Our analysis further suggested that following the duplication, the *CENH3-1* gene was lost independently at least three times in: 1) an ancestor of *V. hirsuta* and *V. ervilia*, 2) an ancestor of most other *Vicia* species and *Le. culinaris*, and 3) in *V. tetrasperma* or its ancestor (fig. 5). To confirm that *CENH3-1* is indeed absent in *Vicia* species, we sequenced genomic DNA of *V. ervilia* and *V. tetrasperma* at 17× and 26× coverage, respectively. *CENH3* sequences were either selectively assembled using GRAbB (Brankovics et al. 2016) or identified in super-reads assembled by *MaSuRCA* (Zimin et al. 2013). Both approaches revealed only a single functional *CENH3-2* gene in each species, confirming the absence of *CENH3-1*. In *V. tetrasperma*, we detected fragments of an additional *CENH3* gene with partial similarity to exon 2, intron 2, exon 3, and intron 3. It was not possible to identify the *CENH3* variant from these recovered sequences, but it is likely that they represent remnants of a nonfunctional gene copy (data not shown).

Protein sequences of CENH3 histones in *Fabeae* are 119–123 aa in length, share 70.6–100% similarity, and are invariant at only 60 sites (supplementary fig. 5*A*, Supplementary Material online). To determine whether their divergence was due to positive selection, we analyzed the sequences using BUSTED (Murrell et al. 2015) to detect gene-wide positive selection, FEL (Kosakovsky Pond and Frost 2005) to detect sites under pervasive positive selection, and MEME (Murrell et al. 2012) to detect sites under episodic positive selection. BUSTED found no evidence of gene-wide positive selection of CENH3 genes in *Fabeae* (table 2). Estimates of ω (ω = Ka/Ks) calculated for the CENH3-1 and CENH3-2 branches were, depending on the tree, 0.374 or 0.375 and 0.254 or 0.269, respectively, suggesting that both CENH3 variants evolved mainly under purifying selective pressure (table 2). FEL and MEME predicted ($P < 0.05$) a total of eight and two sites that may have evolved under positive selection in CENH3-1 and CENH3-2, respectively, indicating that positive selection explains very little of the variability observed among CENH3 protein sequences in *Fabeae* (supplementary fig. 5*A*, Supplementary Material online). We also performed FEL and MEME analyses focusing specifically on CENH3 sequences from the four *Lathyrus* species possessing FabTR-2 as a single-dominant centromeric satellite, but differing considerably in their centromere sizes (fig. 1*B* and supplementary fig. 4,

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

Table 2. Tests for Positive Selection.

Tested Branches	BUSTED		FEL		MEME	
	CENH3	*matk–rbcL*	*CENH3*	*matk–rbcL*	*CENH3*	*matk–rbcL*
All branches	Evidence of positive selection ($P = 0.0434$) $\omega = 0.321$	Evidence of positive selection ($P = 0.0068$) $\omega = 0.3132$	—	—	7 (0.0000) 33 (0.0166) 69 (0.0088) 128 (0.0279)	— 33 (0.0210) 69 (0.0093) 128 (0.0315)
All CENH3-1 branches	No evidence $\omega = 0.3741$	No evidence $\omega = 0.3749$	9 (0.0339) 24 (0.0218) — 33 (0.0191) 94 (0.0392) 105 (0.0025)	9 (0.0304) 24 (0.0234) 27 (0.0443) 33 (0.0209) — 105 (0.0017)	9 (0.0489) 24 (0.0330) — 33 (0.0293) — 105 (0.0045)	9 (0.0443) 24 (0.0351) — 33 (0.0317) — 105 (0.0032)
All CENH3-2 branches	No evidence $\omega = 0.269$	No evidence $\omega = 0.254$	—	—	23 (0.0000)	—
All non-Fabeae branches (outgroup)	Evidence of positive selection ($P = 0.0015$) $\omega = 0.3927$	Evidence of positive selection ($P = 0.0020$) $\omega = 0.3987$	25 (0.0128) 28 (0.0191) — — — 99 (0.0069) —	25 (0.0128) 28 (0.0166) — — — 99 (0.0059) —	25 (0.0201) 28 (0.0293) 33 (0.0421) 45 (0.0073) 69 (0.0080) 99 (0.0115) 128 (0.0122)	25 (0.0200) 28 (0.0259) 33 (0.0442) 45 (0.0074) 69 (0.0083) 99 (0.0101) 128 (0.0129)
Single CENH3-1 branch after duplication	Not tested	Not tested	Not tested	Not tested	7 (0.0151) 24 (0.0434) 33 (0.0083) 49 (0.0148) 94 (0.0378)	7 (0.0186) — 33 (0.0117) 49 (0.0156) —
Single CENH3-2 branch after duplication	Not tested	Not tested	Not tested	Not tested	7 (0.0207)	7 (0.0265)
CENH3-1 in LAS, LAL, LASYL, LNGER, LAV	NA	No evidence $\omega = 0.1955$	NA	105 (0.0145)	NA	105 (0.0228)
CENH3-2 in LAS, LAL, LASYL, LNGER, LAV	NA	No evidence $\omega = 0.1977$	NA	—	NA	—

P values (*P*) are shown in parenthesis. Sites with $P < 0.01$ are underlined.

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

Supplementary Material online). The analyses revealed only one positively evolving site in CENH3-1 and none in CENH3-2 (table 2). Pairwise comparison of CENH3 sequences from these species showed one to eight and zero to four amino acid substitutions in CENH3-1 and CENH3-2, respectively (supplementary fig. 5*B* and *C*, Supplementary Material online). Of these variable sites, one to three and zero to one appeared to have been predicted ($P < 0.05$) as positively evolving in the tests performed on the entire branches of CENH3-1 and CENH3-2 or single branches immediately following the CENH3 duplication event (table 2 and supplementary figure 5*B* and *C*, Supplementary Material online). These results indicated that the expansion of centromeres in the *Lathyrus* species was accompanied by very few changes in CENH3 protein sequences and that the positive selection had almost no impact on CENH3 diversification.

Discussion

In this study, we identified and characterized centromeric satellites in 14 *Fabeae* species and investigated their distribution with respect to the species phylogeny and the evolution of their *CENH3* genes. In terms of the number of included species and newly described centromeric repeats, this is the largest study to date to be conducted on a group of related plants. The methodology employed for the centromeric repeat identification has been proven to be efficient and accurate in a number of studies (Gong et al. 2012; Zhang, Kobližkova, et al. 2014; Kowar et al. 2016; Yang et al. 2018). Compared with an alternative setup in which centromeric sequences are identified by mapping ChIPed and input reads to the genome assemblies (Park 2009), our approach is limited with respect to identification of a single- or low-copy centromeric sequences. However, this limitation is not relevant for repeat-focused studies, as in this case. Moreover, the use of repeated sequences identified by clustering analysis of low-pass Illumina reads as a reference provides several benefits, including unbiased repeat representation and significant reductions in cost and labor relative to building the reference assembly.

Bioinformatic analysis of all highly and moderately repeated sequences revealed CENH3 ChIP-enriched centromeric repeats in all but three species. Except for a small number of retrotransposon and unclassified sequences, all identified centromeric repeats corresponded to families of satDNA (table 1), showing that this class of repeats is dominant in *Fabeae* centromeres. In three species, no ChIP-enriched sequences were identified, suggesting the absence of abundant repeats in the centromeres of these species. By contrast, most *Fabeae* species harbor numerous and abundant centromeric repeats, although the FISH mapping of seven centromeric satellites identified in *V. faba* revealed their absence in the centromere of one chromosome pair (Ávila Robledillo et al. 2018). For practical reasons, our analysis was limited to repeats representing at least 0.01% of the genome, and thus was not exhaustive; hence, an additional analysis targeting individual species with larger volumes of sequencing data would be needed to determine whether their centromeres are truly repeat-free. On the other hand, the negative result of the ChIP-seq analysis obtained for these three species should be interpreted with caution, as it could also have arisen due to the technical issues. This is especially true in *V. ervilia* and *V. hirsuta*, which represent ancient phylogenetic lineage of *Fabeae* and have CENH3 proteins that are relatively divergent from those used to raise the ChIP antibodies (supplementary table 1, Supplementary Material online).

The major finding of this study is the large number and sequence diversity of centromeric satellites within and between *Fabeae* species which is unique among eukaryotic taxa investigated so far. In many organisms, a single satellite repeat family dominates all centromeres, although it may partially differentiate into chromosome-specific variants or higher order repeats. Although these satellites evolve relatively rapidly, similarities are still detectable between sequences retrieved from related species. Examples of such centromeric satellite superfamilies include the primate alpha satellites (McNulty and Sullivan 2018; Hartley and O'Neill 2019), CentO/CentC in *Oryza* and *Zea* (Lee et al. 2005; Bilinski et al. 2015) and cen180 in *Arabidopsis* and other *Brassicaceae* (Lermontova et al. 2014). Sequence diversification of such shared centromere-specific superfamilies along with the adaptive evolution of CENH3 proteins found in some taxa, led to the formulation of the centromere drive model (Henikoff et al. 2001; Malik 2009). The model proposes that specific interactions of CENH3 or other inner kinetochore proteins with their underlying centromeric satellites result in stronger centromeres on homologs with expanded satellite arrays that are consequently preferentially transmitted to the germ cells during asymmetric female meiosis. This process is then compensated for by the adaptive evolution of the interacting protein(s), leading to the evolutionary race of arms between selfish centromeric DNA and its associated kinetochore proteins. However, it is unlikely that centromere drive is at work in *Fabeae*, as the presence of multiple centromeric satellites with different sequences precludes any sequence-dependent coevolution with CENH3 or other kinetochore proteins. This is further supported by the observed lack of pervasive adaptive evolution of *Fabeae* CENH3 proteins, which was not detected even in the set of *Lathyrus* species possessing a single centromeric satellite (table 2). Another argument against sequence-dependent deposition of CENH3 to *Fabeae* centromeres comes from CENH3–YFP fusion protein expression experiments showing that CENH3-2 from *V. faba* is efficiently deposited onto *P. sativum* centromeres, and conversely, that CENH3-1 from *P. sativum* targets centromeres in *V. faba* (Neumann et al. 2015). Similar results were reported by Maheshwari et al. (2017), demonstrating that CENH3 from evolutionary distant species can replace the native CENH3 in *Arabidopsis thaliana*.

Another factor to consider when seeking an explanation for the observed diversity of centromeric satellites is the duplication and partial diversification of the two *CENH3* gene copies in *Pisum* and *Lathyrus*. Coincidentally, species from these genera also exhibit a distinctive type of centromere morphology characterized by extended primary constrictions

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

and occurrence of multiple CENH3 loci (supplementary fig. 4, Supplementary Material online and Neumann et al. 2015). However, neither the *CENH3* duplication nor the centromere morphology can be directly linked to the diversity of centromeric satellites, as this group of species includes both of the observed extremes: the single centromeric satellite associated with all CENH3 loci in *L. sativus*, as well as the most diverse population of centromeric satellites with uneven distribution on *P. sativum* chromosomes (fig. 2).

Satellite DNA is not necessary or sufficient for centromere establishment and propagation (Piras et al. 2010; Logsdon et al. 2019). In plants, satellite-free centromeres are present on five of the 12 chromosomes of potato (*Solanum tuberosum*). The remaining seven potato centromeres contain mostly chromosome-specific satellites with exceptionally long monomers originating from recombination of LTR-retrotransposon fragments with other genomic sequences (Gong et al. 2012). This type of centromeric satellites is also present in the closely related *S. verrucosum*; however, the repeats are mostly species-specific, suggesting their recent and independent origin (Zhang, Kobližkova, et al. 2014). Based on these findings, along with the presence of partially homogenized centromeric satellites in switchgrass species (Yang et al. 2018), it was hypothesized that evolutionarily young centromeres may be repeat-free and only later accumulate random satellites that are subsequently homogenized across different chromosomes, resulting in the selection of a single, structurally favorable repeat to dominate all centromeres (Gong et al. 2012; Zhang, Kobližkova, et al. 2014; Yang et al. 2018). Considering our results in light of this hypothesis, we can see a number of differences suggesting that the diversity of centromeric satellites in *Fabeae* is not due to their origin in newly formed or relocated centromeres. First, some satellite superfamilies occur in species from different phylogenetic lineages, indicating that their origin dates back to the diversification of *Fabeae* (FabTR-1 and FabTR-12, fig. 1). Moreover, FISH mapping revealed that some centromeric satellites also occur at additional, noncentromeric loci, suggesting that they might have originated elsewhere in the genome and subsequently invade the centromeres. In addition, we have no evidence of frequent neocentromere formation or chromosome rearrangements in *Fabeae*, which have relatively stable karyotypes (Badr 2006).

Compared with other plant taxa, most *Fabeae* species are exceptional in terms of their high diversity of satellite repeats in general (Neumann et al. 2012; Macas et al. 2015; Ávila Robledillo et al. 2018), which might also be reflected in their large numbers of centromeric repeats. This diversity of satellites contrasts even with the closest relatives of *Fabeae*, genera *Trifolium*, *Medicago*, and *Cicer*, whose species possess one or (rarely) two centromeric satellites (Zatloukalová et al. 2011; Yu et al. 2017; Dluhošová et al. 2018). The molecular or evolutionary processes that made *Fabeae* so rich in satDNA remain to be fully elucidated, but one possible mechanism was revealed in our recent investigation of *L. sativus* repeats using ultralong nanopore reads. Most noncentromeric satellites in this species originated relatively recently by amplification of short tandem repeat arrays present in LTR-retrotransposons (Vondrak et al. 2020). The same mechanism was previously proposed for the origin of PisTR-A satellite in *P. sativum* (Macas et al. 2009); thus, it is likely to contribute to the emergence of species-specific satellites and their high turnover across *Fabeae*. It is worth noting that the LTR-retrotransposons providing these short array templates belong to the lineage of Ty3/gypsy Ogre elements (Neumann et al. 2019) which represent dominant repeats in the *Fabeae* genomes (Macas et al. 2015) but are comparably less abundant in the related legume taxa (Macas et al. 2007; Dluhošová et al. 2018), potentially resulting in smaller numbers of Ogre-derived satellite repeats.

Taken together, the results presented in this work, along with the recent data from other species, suggest that the patterns of association and eventual coevolution of satellite repeats with plant centromeres may be far more complex than previously envisioned. It is possible that the mechanisms leading to the centromere drive act only episodically, or in specific cases in which only a single repeat with properties favorable for supporting centromeric chromatin is available. However, should multiple such satellites occur in a genome, they might be co-opted simultaneously or alternatively during centromere evolution, and this seems to have occurred in *Fabeae*. Several features are thought to be important for "centromere competence" of satellite repeats, including the presence of dyad symmetries (Kasinathan and Henikoff 2018) or WW dinucleotide periodicities in their sequences (Zhang et al. 2013; Yang et al. 2018), as well as a proper level of transcription (Duda et al. 2017; Perea-Resa and Blower 2018). The sequence data acquired in this study will be instrumental in future research of these properties, as it includes diverse satellite sequences and allows for their comparative analysis in species with different modes of association between individual satellite families and centromeric chromatin.

Materials and Methods

Plant Material

Seeds of most *Vicia* species were obtained from the seed bank of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany. Their accessions in the seed bank are as follows: *V. hirsuta* (L.) S.F.Gray, VIC728; *V. ervilia* (L.) Willd., ERV52; *V. pisiformis* L., VIC36; *V. peregrina* L., VIC765; *V. villosa*, VIC876; and *V. tetrasperma* (L.) Schreb., VIC726. Commercial varieties of *V. pannonica* "Dětěnická panonská," *V. faba* "Merkur," and *P. sativum* "Terno" were obtained from Osiva Boršov, Czech Republic; *V. sativa* "Ebena" from the Agricultural Research Institute Kroměříž, Czech Republic; and *Le. culinaris* "Eston" from the Nohel garden, Dobříš, Czech Republic. *Vicia narbonensis* (ICARDA 14) was provided by A. M. Torres (IFAPA Cordoba, Spain). *Lathyrus sativus*, *L. latifolius*, and *L. niger* were purchased from Fratelli Ingegnoli S.p.A., Milano, Italy (Cat. No.: 455), SEMO Smržice, Czech Republic (accession number 1-0040-68867-01), and Arboretum Paseka Makču Pikču, Paseka, Czech Republic, respectively. *Lathyrus vernus* was collected from a wild population at Vidov, Czech Republic (GPS

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

48°55′17.401″N, 14°29′44.158″E). *Pisum fulvum* accession (ICARDA IG64207) was provided by Petr Smýkal, Palacký University, Olomouc, Czech Republic.

Genomic DNA Isolation and Phylogenetic Analysis of *Fabeae* Species

Genomic DNA was extracted from leaf tissues according to Dellaporta et al. (1983). Sequences of the chloroplast loci (*matK* and *rbcL*) used for phylogenetic reconstructions were obtained by PCR amplification of the corresponding DNA fragments from total genomic DNA preparations using the primers MatK-L-F (5′-ATG AAG GAM TAT HMA GTA TAT TTA G-3′) and Matk-L-R (5′-TCA TTC ATC ATG GAC CAG ATC-3′), and rbcL-L_F2 (5′-ATG TCA CCA CAA ACA GAA ACT AAA-3′) and rbcL-L_R2 (5′-TTA CAA AGT ATC CAT TGC TGG G-3′). Alternatively, the *matK* and *rbcL* sequences were assembled from previously published NGS data sets (Macas et al. 2015) or retrieved from GenBank, as specified in supplementary table 3, Supplementary Material online. Nucleotide sequences were aligned using Muscle (Edgar 2004). ML phylogenies were estimated using PhyML 3.0 (Guindon et al. 2010) with automatic model selection by SMS (Lefort et al. 2017). Starting trees for ML analysis were calculated using neighbor-joining (NJ) algorithm implemented in SeaView (Gouy et al. 2010). The branch support was evaluated using bootstrap analysis (≥10,000 replicates). Divergence times were estimated using RelTime method implemented in MEGA X (Mello 2018), taking into account that the most recent common ancestor of *P. sativum* and *V. sativa* existed 12.9–22.8 Ma (Lavin et al. 2005). Phylogenetic trees were edited using ITOL (Letunic and Bork 2019). Alignment of the concatenated sequences of *matK* and *rbcL* used to infer the species tree (fig. 1A) is provided in supplementary file 2, Supplementary Material online.

Identification of Centromeric Repeats Using Chromatin Immunoprecipitation

Chromatin immunoprecipitation was performed on nuclei isolated from fresh leaves as described (Neumann et al. 2012) using custom-made antibodies raised against peptides designed according to the previously identified *Fabeae* CENH3 protein sequences (Neumann et al. 2012, 2015). A single antibody was always used for ChIP experiments, and it was selected based on 1) the similarity between peptide antigen and the CENH3 sequence in particular species and 2) its performance in *in situ* immunodetection experiments. Information about the antibodies and their use in individual species is provided in the supplementary table 1, Supplementary Material online and references cited therein. Rabbit polyclonal antibody to CENH3-2 of *L. sativus* (ID: P60) was produced in the course of this study by Genscript (Piscataway, NJ) using "complete affinity-purified peptide polyclonal package" (Cat. No.: SC1031). ChIPed DNA and input DNA control were sequenced on the Illumina platform in a single-end, 101 nt read mode. The resultant reads were trimmed to 100 nt by removing the first base and quality filtered to exceed the cutoff quality score of 10 over at least 95 nucleotides. Quality-filtered reads were mapped to reference contigs assembled from clusters of genome shotgun sequencing reads representing repetitive sequences of the corresponding species produced and characterized in our previous work (Macas et al. 2015). Similarity-based mapping of reads to repeat contigs was performed using BlastN (Altschul et al. 1997) with the parameters "-m 8 -b 1 -e 1e-20 -W 9 -r 2 -q -3 -G 5 -E 2 -F F," and was followed by output parsing to ensure that each read was mapped to a maximum of one repeat cluster with the highest similarity score. The proportion of ChIP and input reads mapped to individual clusters was evaluated to identify repeats with a ChIP/input ratio ≥3, which were considered to represent repeats enriched in the ChIP sample. In the two species for which reference contigs were not available (*L. niger* and *L. clymenum*), the ChIP and input reads were used directly for comparative RepeatExplorer analysis (Novák et al. 2013) and enrichment was calculated as a ratio of ChIP to input reads in individual repeat clusters.

Sequence Analysis of Satellite Repeats

Putative satellite repeats were identified in the course of our previous study (Macas et al. 2015) via graph-based clustering of genomic shotgun reads using the RepeatExplorer pipeline (Novák et al. 2013). Reconstruction of monomer sequences of selected satellites was performed using TAREAN (Novák et al. 2017). Similarities between satellite sequences were evaluated using alignment-free sequence comparison using D_2^*distance (Reinert et al. 2009) as implemented in d2-tools (https://code.google.com/archive/p/d2-tools/; last accessed September 23, 2019). Dissimilarity measurement matrices were calculated using shotgun reads from individual satellite clusters for k-mer lengths *k* from three to nine nucleotides under the zero- to third-order Markov model *M*. The resultant distance matrix was calculated as arithmetical average of all 27 dissimilarity matrices. The similarity threshold used for visualization was defined as:$\sum_{k=3}^{9}\sum_{M=0}^{4}\frac{D_2^*}{27} < 0.33$. This threshold was selected based on the empirical exploration of multiple satellite sequences using dotplot comparisons (Sonnhammer and Durbin 1995).

Alternatively, similarities between centromeric satellites and all other repetitive sequences were detected using BlastN search with default parameters. Contigs assembled from clusters representing repetitive sequences of the corresponding species produced and characterized in our previous work (Macas et al. 2015) were searched against the database of TAREAN-reconstructed satellite centromeric sequences. The top percentile of similarity hits was manually explored using dotplot.

FISH and Immunolabeling

Mitotic chromosomes used for cytogenetic experiments were prepared from root apical meristems synchronized as described previously (Neumann et al. 2015) to increase the proportion of simultaneously dividing cells. The synchronized meristems were processed using different protocols depending on their intended use. For FISH experiments, they were fixed in a 3:1 v/v solution of methanol:glacial acetic acid for

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

2 days at 4 °C, washed in ice-cold water, and digested in a solution of 4% cellulase (Onozuka R10, Serva Electrophoresis, Heidelberg, Germany), 2% pectinase, and 0.4% pectolyase Y23 (both MP Biomedicals, Santa Ana, CA) in 0.01 M citrate buffer (pH 4.5) for 90 min at 37 °C. One to three digested meristems were transferred to a drop of freshly made 3:1 fixation solution on a glass slide and further macerated using a forceps. The slide was then placed over an alcohol flame to induce chromosome spreading as described by Dong et al. (2000). Following air-drying, the slides were stored at −20 °C. FISH was performed using either oligonucleotide probes that were 5′-labeled with biotin or Rhodamine Red-X during their synthesis (Integrated DNA Technologies, Leuven, Belgium), or using cloned fragments of satellite sequences labeled with biotin using nick-translation (Kato et al. 2006). Nucleotide sequences of the probes are provided in supplementary file 3, Supplementary Material online. FISH was performed as described (Macas et al. 2007) with hybridization and washing temperatures adjusted to account for AT/GC content and hybridization stringency allowing for 10–20% mismatches.

Immunolabeling of CENH3 proteins was performed with chromosomes isolated from the meristems fixed using 4% formaldehyde for 25 min at 23 °C Following fixation, suspensions of purified metaphase chromosomes were prepared as described (Neumann et al. 2002). Alternatively, the meristems were digested in a solution of 2% cellulase and 2% pectinase in phosphate-buffered saline (PBS) for 80–120 min at 28 °C, transferred to a glass slide, and squashed under the coverslip. Immunodetection was performed as follows, slides with chromosome suspensions and squash preparations were treated identically, and all incubations were performed at room temperature unless stated otherwise. Slides were washed in PBS for 5 min, PBS-T1 buffer (1× PBS, 0.5% Triton, pH 7.4) for 25 min, and twice in PBS for 5 min and once in PBS-T2 buffer (1× PBS, 0.1% Tween 20, pH 7.4) for 30 min. The slides were then incubated with the primary CENH3 antibody diluted 1:1,000 in PBS-T2 at 4 °C overnight, and then washed twice in PBS for 5 min and once in PBS-T2 for 5 min. The primary antibodies were detected with anti-rabbit-Rhodamine Red-X-AffiniPure (1:500, Jackson ImmunoResearch, Suffolk, UK; catalog number 111-295-144) or anti-chicken-DyLight488 (1:500, Jackson ImmunoResearch; catalog number 103-485-155) diluted in PBS-T2 buffer for 1 h. After two washes in PBS for 5 min and one wash in PBS-T2 for 5 min, the slides were mounted for observation or processed further if combined detection of DNA sequences by FISH was needed. In such cases, the slides were immediately postfixed in 4% formaldehyde in PBS for 10 min at RT and dehydrated in a series of 70% and 96% ethanol at RT for 5 min each. Chromosomes were denatured by incubation in 1× PCR buffer (Promega, Madison, WI) supplemented with 4 mM $MgCl_2$ for 2 min at 94 °C and used for FISH as described earlier. The slides were counterstained with 4′,6-diamidino-2-phenylindole (DAPI), mounted in Vectashield mounting medium (Vector Laboratories, Burlingame, CA), and examined using a Zeiss AxioImager.Z2 microscope with an AxioCam 506 mono camera. Images were captured and processed using the ZEN pro 2012 software (Carl Zeiss GmbH).

Identification and Analysis of *CENH3* Genes

Partial CENH3-coding sequences of *V. ervilia*, *V. hirsuta*, *V. pisiformis*, and *V. tetrasperma* were identified in Illumina sequence data by BlastN using a query containing all *CENH3* sequences identified in *Fabeae* species previously (Neumann et al. 2012, 2015). Primers designed based on these sequences were then used for RT-PCR and RACE amplification of fragments of *CENH3* transcripts, as described by Neumann et al. (2015). Finally, fragments surrounding the 5′ and 3′ end of the coding sequences were used to design primers for amplification of full-length CENH3-coding sequences. Sequences of these primers and details of the amplification conditions are provided in supplementary table 4, Supplementary Material online.

Entire *CENH3* genes in *V. ervilia* and *V. tetrasperma* were selectively assembled using GRABb (Brankovics et al. 2016) using as input Illumina paired-end reads (2 × 151 nt) and a bait file containing all CENH3-coding sequences available in *Fabeae*. The CENH3 sequences were also identified in superreads assembled from the Illumina paired-end reads by *MaSuRCA* (Zimin et al. 2013). Illumina sequence data used for assembly were custom-produced at Admera Health, LLC (South Plainfield, NJ), and deposited into the SRA database under accessions ERR3523145 and ERR3523144, respectively. Exon/intron structure of the genes and their translation products were predicted using est2genome (Rice et al. 2000) and GeneWise (Birney et al. 2004).

CENH3 sequences were aligned using Muscle (Edgar 2004). Pairwise similarities between CENH3 sequences were inferred from the proportions of variable sites (p-distances) calculated from CENH3 alignment in MEGA (Kumar et al. 2018). Phylogenetic analyses were performed using NJ and ML algorithms implemented in SeaView (Gouy et al. 2010) and PhyML 3.0 (Guindon et al. 2010), respectively. Bootstrap values were calculated from at least 1,000 replications. Phylogenetic trees were drawn and edited using the FigTree program (http://tree.bio.ed.ac.uk/software/figtree/; last accessed May 15, 2017). Tests for positive selection were carried out using the BUSTED (Murrell et al. 2015), FEL (Kosakovsky Pond and Frost 2005), and MEME (Murrell et al. 2012) tools implemented in the software package HyPhy (Kosakovsky Pond et al. 2005).

Availability of Sequence Data

Illumina reads from the ChIPed and control input samples are available in the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under run accession numbers ERR3063140–ERR3063141, ERR3063378–ERR3063383, ERR3063416–ERR3063425, and ERR3063493–ERR3063500. The runs are associated with the study "Repeat characterization in *Fabeae* genomes" (PRJEB5241) which also includes the corresponding genomic NGS data. Newly identified CENH3 gene sequences are available from GenBank (https://www.ncbi.nlm.nih.gov/genbank/) under accession numbers MK415838–MK415841.

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Ms Vlasta Tetourová and Ms Jana Látalová for their excellent technical assistance. This work was supported by the grants from the Czech Science Foundation (Grant No. 17-09750S), the Czech Academy of Sciences (Grant No. RVO:60077344), and the ELIXIR CZ Research Infrastructure Project (Czech Ministry of Education, Youth and Sports; Grant No. LM2015047), including access to computing and storage facilities.

References

Akera T, Chmátal L, Trimm E, Yang K, Aonbangkhen C, Chenoweth DM, Janke C, Schultz RM, Lampson MA. 2017. Spindle asymmetry drives non-Mendelian chromosome segregation. *Science* 358(6363):668–672.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–3402.

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. *Sci Rep.* 8(1):5838.

Badr SF. 2006. Karyotype analysis and chromosome evolution in species of *Lathyrus* (Fabaceae). *Cytologia* 71(4):447–455.

Bilinski P, Distor K, Gutierrez-Lopez J, Mendoza Mendoza G, Shi J, Dawe RK, Ross-Ibarra J. 2015. Diversity and evolution of centromere repeats in the maize genome. *Chromosoma* 124(1):57–65.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res.* 14(5):988–995.

Brankovics B, Zhang H, van Diepeningen AD, van der Lee TAJ, Waalwijk C, de Hoog GS. 2016. GRAbB: selective assembly of genomic regions, a new niche for genomic research. *PLoS Comput Biol.* 12(6):e1004753.

Cheeseman I. 2014. The kinetochore. *Cold Spring Harb Perspect Biol.* 6(7):a015826.

Cooper JL, Henikoff S. 2004. Adaptive evolution of the histone fold domain in centromeric histones. *Mol Biol Evol.* 21(9):1712–1718.

Dellaporta SL, Wood J, Hicks JB. 1983. A plant DNA minipreparation: version II. *Plant Mol Biol Rep.* 1(4):19–21.

Dluhošová J, Ištvánek J, Nedělník J, Řepková J. 2018. Red clover (*Trifolium pratense*) and zigzag clover (*T. medium*) – a picture of genomic similarities and differences. *Front Plant Sci.* 9:724.

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

Duda Z, Trusiak S, O'Neill R. 2017. Centromere transcription: means and motive. In: Black BE, editor. Centromeres and kinetochores, progress in molecular and subcellular biology. Cham (Switzerland): Springer. p. 257–281.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5):1792–1797.

Finseth FR, Dong Y, Saunders A, Fishman L. 2015. Duplication and adaptive evolution of a key centromeric protein in *Mimulus*, a genus with female meiotic drive. *Mol Biol Evol.* 32(10):2694–2706.

Fuchs J, Schubert I. 2012. Chromosomal distribution and functional interpretation of epigenetic histone marks in plants. In: Bass HW, Birchler JA, editors. Plant Cytogenetics. New York: Springer. p. 232–246.

Garrido-Ramos MA. 2015. Satellite DNA in plants: more than just rubbish. *Cytogenet Genome Res.* 146(2):153–170.

Garrido-Ramos MA. 2017. Satellite DNA: an evolving topic. *Genes.* 8(9):230.

Gent JI, Wang N, Dawe RK. 2017. Stable centromere positioning in diverse sequence contexts of complex and satellite centromeres of maize and wild relatives. *Genome Biol.* 18(1):121.

Gong Z, Wu Y, Koblížková 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. *Plant Cell* 24(9):3559–3574.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27(2):221–224.

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

Han F, Lamb JC, Mccaw ME, Gao Z, Zhang B, Swyers NC, Birchler JA, Anderson L. 2018. Meiotic studies on combinations of chromosomes with different sized centromeres in maize. *Front Plant Sci.* 9:785.

Hara M, Fukagawa T. 2017. Critical foundation of the kinetochore: the Constitutive Centromere-Associated Network (CCAN). In: Black BE, editor. Centromeres and kinetochores, progress in molecular and subcellular biology. Vol. 56. Cham (Switzerland): Springer. p. 29–57.

Hartley G, O'Neill RJ. 2019. Centromere repeats: hidden gems of the genome. *Genes* 10(3):223.

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. *Plant J.* 73(4):555–565.

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

Hirsch CD, Wu Y, Yan H, Jiang J. 2009. Lineage-specific adaptive evolution of the centromeric protein CENH3 in diploid and allotetraploid *Oryza* species. *Mol Biol Evol.* 26(12):2877–2885.

Iwata-Otsubo A, Dawicki-McKenna JM, Akera T, Falk SJ, Chmátal L, Yang K, Sullivan BA, Schultz RM, Lampson MA, Black BE. 2017. Expanded satellite repeats amplify a discrete CENP-A nucleosome assembly site on chromosomes that drive in female meiosis. *Curr Biol.* 27(15):2365–2373.

Kasinathan S, Henikoff S. 2018. Non-B-form DNA is enriched at centromeres. *Mol Biol Evol.* 35(4):949–962.

Kato A, Kato A, Albert PS, Vega JM, Kato A, Albert PS, Vega JM, Birchler JA. 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(2–3):71–78.

Kawabe A, Nasuda S, Charlesworth D. 2006. Duplication of centromeric histone H3 (HTR12) gene in *Arabidopsis halleri* and *A. lyrata*, plant species with multiple centromeric satellite sequences. *Genetics* 174(4):2021–2032.

Kosakovsky Pond SL, Frost S. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol.* 22(5):1208–1222.

Kowar T, Zakrzewski F, Macas J, Koblížková A, Viehoever P, Weisshaar B, Schmidt T. 2016. Repeat composition of CenH3-chromatin and H3K9me2-marked heterochromatin in sugar beet (*Beta vulgaris*). *BMC Plant Biol.* 16(1):120.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol.* 35(6):1547–1549.

Lavin M, Herendeen PS, Wojciechowski MF. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Syst Biol.* 54(4):575–594.

Lee H-R, Zhang W, Langdon T, Jin W, Yan H, Cheng Z, Jiang J. 2005. Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. *Proc Natl Acad Sci U S A.* 102(33):11793–11798.

Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

Lefort V, Longueville J-E, Gascuel O. 2017. SMS: smart model selection in PhyML. *Mol Biol Evol*. 34(9):2422–2424.

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

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 47(W1):W256–W259.

Logsdon GA, Gambogi CW, Liskovykh MA, Barrey EJ, Larionov V, Miga KH, Heun P, Black BE. 2019. Human artificial chromosomes that bypass centromeric DNA. *Cell* 178(3):624–639.

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(2):198–206.

Macas J, Mészáros T, Nouzová M. 2002. PlantSat: a specialized database for plant satellite repeats. *Bioinformatics* 18(1):28–35.

Macas J, Neumann P, 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(1):427.

Macas J, Novák P, Pellicer J, Čížková J, Koblížková A, Neumann P, Fuková I, Doležel J, Kelly LJ, Leitch IJ. 2015. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe *Fabeae*. *PLoS One* 10(11):e0143424.

Maheshwari S, Ishii T, Brown CT, Houben A, Comai L. 2017. Centromere location in *Arabidopsis* is unaltered by extreme divergence in CENH3 protein sequence. *Genome Res*. 27(3):471–478.

Malik HS. 2009. The centromere-drive hypothesis: a simple basis for centromere complexity. *Prog Mol Subcell Biol*. 48:33–52.

Masonbrink RE, Gallagher JP, Jareczek JJ, Renny-Byfield S, Grover CE, Gong L, Wendel JF. 2014. CenH3 evolution in diploids and polyploids of three angiosperm genera. *BMC Plant Biol*. 14(1):383.

McFarlane RJ, Humphrey TC. 2010. A role for recombination in centromere function. *Trends Genet*. 26(5):209–213.

McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res*. 26(3):115–138.

Mello B. 2018. Estimating TimeTrees with MEGA and the TimeTree Resource. *Mol Biol Evol*. 35(9):2334–2342.

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. *Mol Biol Evol*. 32(5):1365–1371.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*. 8(7):e1002764.

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 Genet*. 8(6):e1002777.

Neumann P, Nouzová M, Macas J. 2001. Molecular and cytogenetic analysis of repetitive DNA in pea (*Pisum sativum* L.). *Génome* 44(4):716–728.

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

Neumann P, Pavlíková Z, Koblížková A, Fuková I, Jedličková V, Novák P, Macas J. 2015. Centromeres off the hook: massive changes in centromere size and structure following duplication of CENH3 gene in *Fabeae* species. *Mol Biol Evol*. 32(7):1862–1879.

Neumann P, Požárková D, Vrána J, Doležel J, Macas J. 2002. Chromosome sorting and PCR-based physical mapping in pea (*Pisum sativum* L.). *Chromosome Res*. 10(1):63–71.

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

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

Oliveira LC, Torres GA. 2018. Plant centromeres: genetics, epigenetics and evolution. *Mol Biol Rep*. 45(5):1491–1497.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 10(10):669–680.

Perea-Resa C, Blower MD. 2018. Centromere biology: transcription goes on stage. *Mol Cell Biol*. 38:e00263–e00318.

Piras FM, Nergadze SG, Magnani E, Bertoni L, Attolini C, Khoriauli L, Raimondi E, Giulotto E. 2010. Uncoupling of satellite DNA and centromeric function in the genus *Equus*. *PLoS Genet*. 6(2): e1000845.

Plohl M, Meštrović N, Mravinac B. 2014. Centromere identity from the DNA point of view. *Chromosoma* 123(4):313–325.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21(5):676–679.

Reinert G, Chew D, Sun F, Waterman MS. 2009. Alignment-free sequence comparison (I): statistics and power. *J Comput Biol*. 16(12):1615–1634.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 16(6):276–277.

Schaefer H, Hechenleitner P, Santos-Guerra A, de Sequeira MM, Pennington RT, Kenicer G, Carine MA. 2012. Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-Atlantic island lineages. *BMC Evol Biol*. 12(1):250.

Schneider KL, Xie Z, Wolfgruber TK, Presting GG. 2016. Inbreeding drives maize centromere evolution. *Proc Natl Acad Sci U S A*. 113(8):E987–E996.

Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167(1–2):GC1–GC10.

Talbert P, Kasinathan S, Henikoff S. 2018. Simple and complex centromeric satellites in *Drosophila* sibling species. *Genetics* 208(3):977–990.

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. *Plant J*. 101(2):484–500.

Wei K-C, Reddy HM, Rathnam C, Lee J, Lin D, Ji S, Mason JM, Clark AG, Barbash DA. 2017. A pooled sequencing approach identifies a candidate meiotic driver in *Drosophila*. *Genetics* 206(1):451–465.

Yang X, Zhao H, Zhang T, Zeng Z, Zhang P, Zhu B, Han Y, Braz GT, Casler MD, Schmutz J, et al. 2018. Amplification and adaptation of centromeric repeats in polyploid switchgrass species. *New Phytol*. 218(4):1645–1657.

Yu F, Dou Q, Liu R, Wang H. 2017. A conserved repetitive DNA element located in the centromeres of chromosomes in *Medicago* genus. *Genes Genom*. 39(8):903–911.

Zatloukalová P, Hřibová E, Kubaláková M, Suchánková P, Šimková H, Adoración C, Kahl G, Millán T, Doležel J. 2011. Integration of genetic and physical maps of the chickpea (*Cicer arietinum* L.) genome using flow-sorted chromosomes. *Chromosome Res*. 19(6):729–739.

Zedek F, Bureš P. 2016. CenH3 evolution reflects meiotic symmetry as predicted by the centromere drive model. *Sci Rep*. 6:33308.

Zhang B, Dong Q, Su H, Birchler JA, Han F. 2014. Histone phosphorylation: its role during cell cycle and centromere identity in plants. *Cytogenet Genome Res*. 143(1–3):144–149.

Zhang H, Koblížková A, Wang K, Gong Z, Oliveira L, Torres GA, Wu Y, Zhang W, Novák P, Buell CR, et al. 2014. Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* species: rapid evolution of DNA sequences associated with centromeres. *Plant Cell* 26(4):1436–1447.

Zhang T, Talbert PB, Zhang W, Wu Y, Yang Z, Henikoff JG, Henikoff S, Jiang J. 2013. The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. *Proc Natl Acad Sci U S A*. 110(50):E4875–E4883.

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

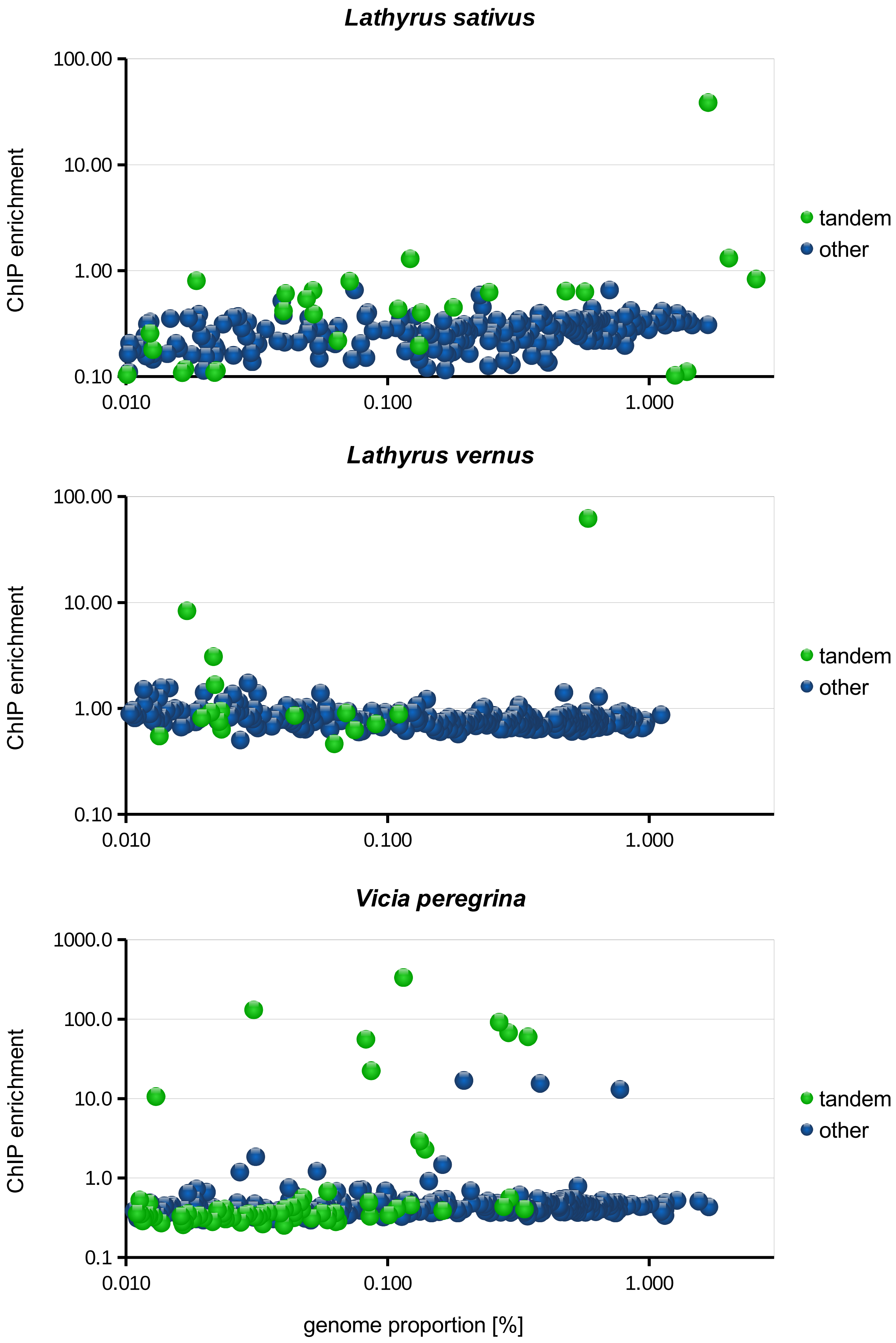
Downloaded from https://academic.oup.com/mbe/article/37/8/2341/5817320 by Ustavy AV - Ceske Budejovice user on 09 February 2021

Supplementary Information

Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe *Fabeae*.

Laura Ávila Robledillo, Pavel Neumann, Andrea Koblížková, Petr Novák, Iva Vrbová, and Jiří Macas.

Supplementary fig. 1



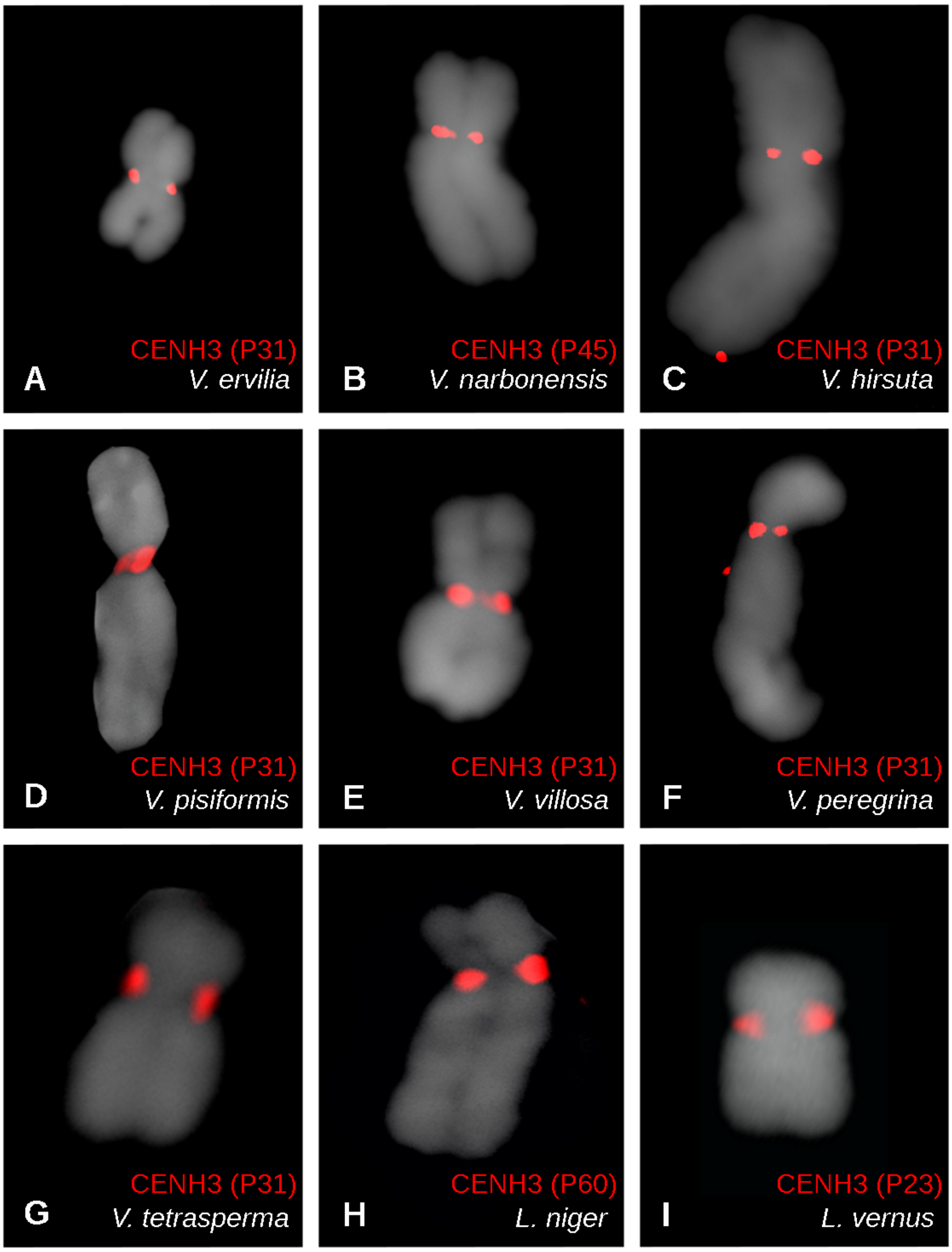
Lathyrus sativus

Lathyrus vernus

Vicia peregrina

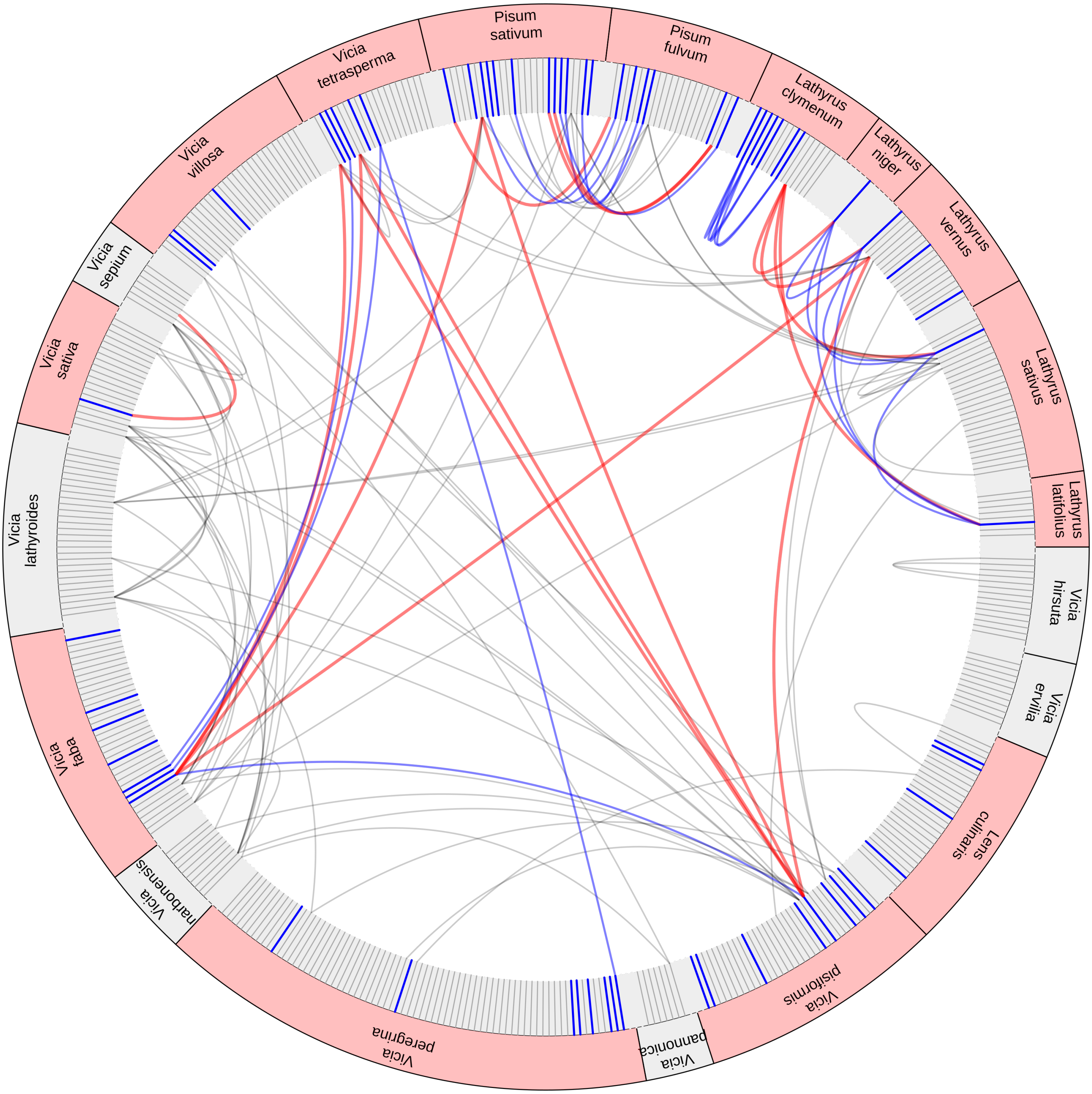
Supplementary fig. 1. Identification of repeat clusters associated with CENH3 chromatin. Repeat clusters are represented by dots and their positions reflect genomic abundance of corresponding repeats (x axis) and their ChIP enrichment (ChIP/input ratio; y axis). Only repeats with genomic proportions of at least 0.01% were analyzed. Tandemly organized repeats are highlighted as green dots whereas all remaining repeats are blue. The plots show only three of the analyzed species as examples of genomes with different numbers of centromeric satellites.

Supplementary fig. 2



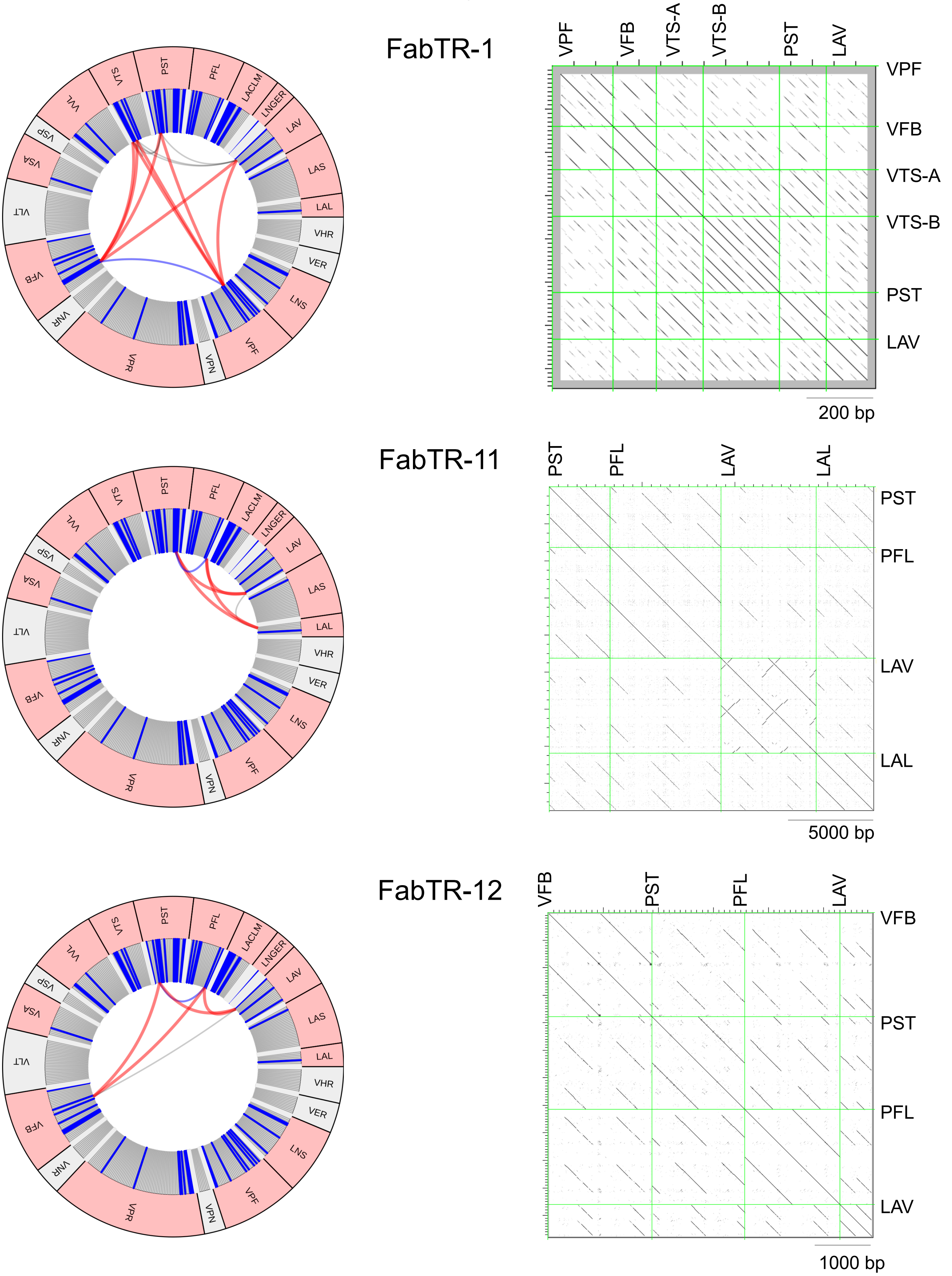
Supplementary fig. 2. Immunodetection of CENH3 proteins on isolated chromosomes to confirm antibody specificity. The CENH3 antibodies are shown as red signals, the chromosomes counterstained with DAPI are gray. Results are shown for all nine *Fabeae* species where the centromere labeling have not been demonstrated previously. Antibody IDs provided in parenthesis correspond to those listed in supplementary table 1.

Supplementary fig. 3A



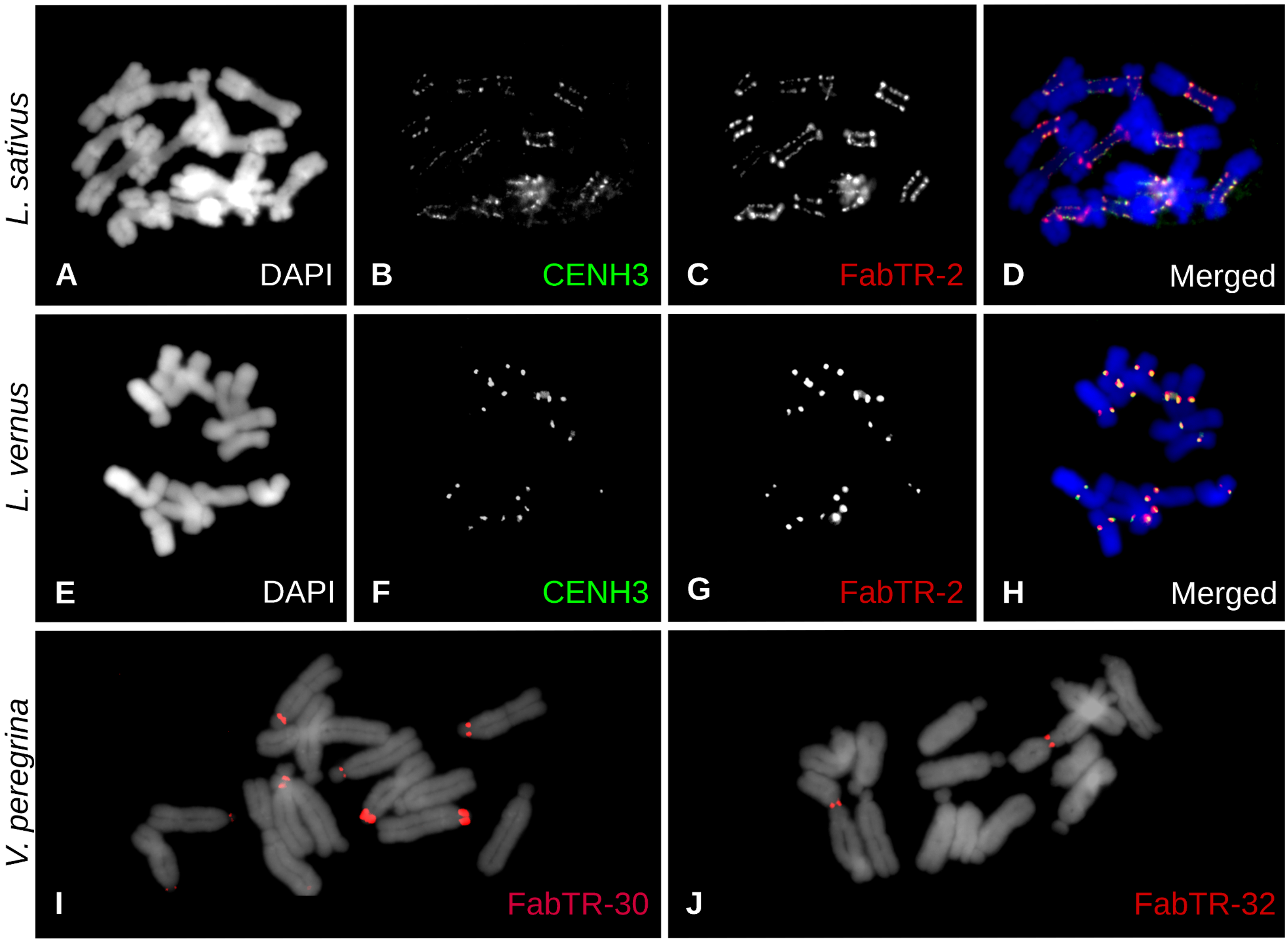
Supplementary fig. 3A. Overview of sequence similarities between *Fabeae* satellites. (A) Each species is represented by a segment of the outer circle which is pink in the case of species analyzed by ChIP-seq. Radial lines within the inner circle represent different satellite repeats in each species. Blue radial lines distinguish centromeric (ChIP-enriched) satellites from those that either were not enriched or were not analyzed by ChIP (gray radial lines). Connecting lines indicate sequence similarities as follows: gray connection lines represent similarities between non-centromeric satellites or satellites not analyzed by ChIP, blue lines show similarities between two centromeric satellites, and red lines connect satellites that are centromeric in some and non-cetromeric in other species.

Supplementary fig. 3B-D



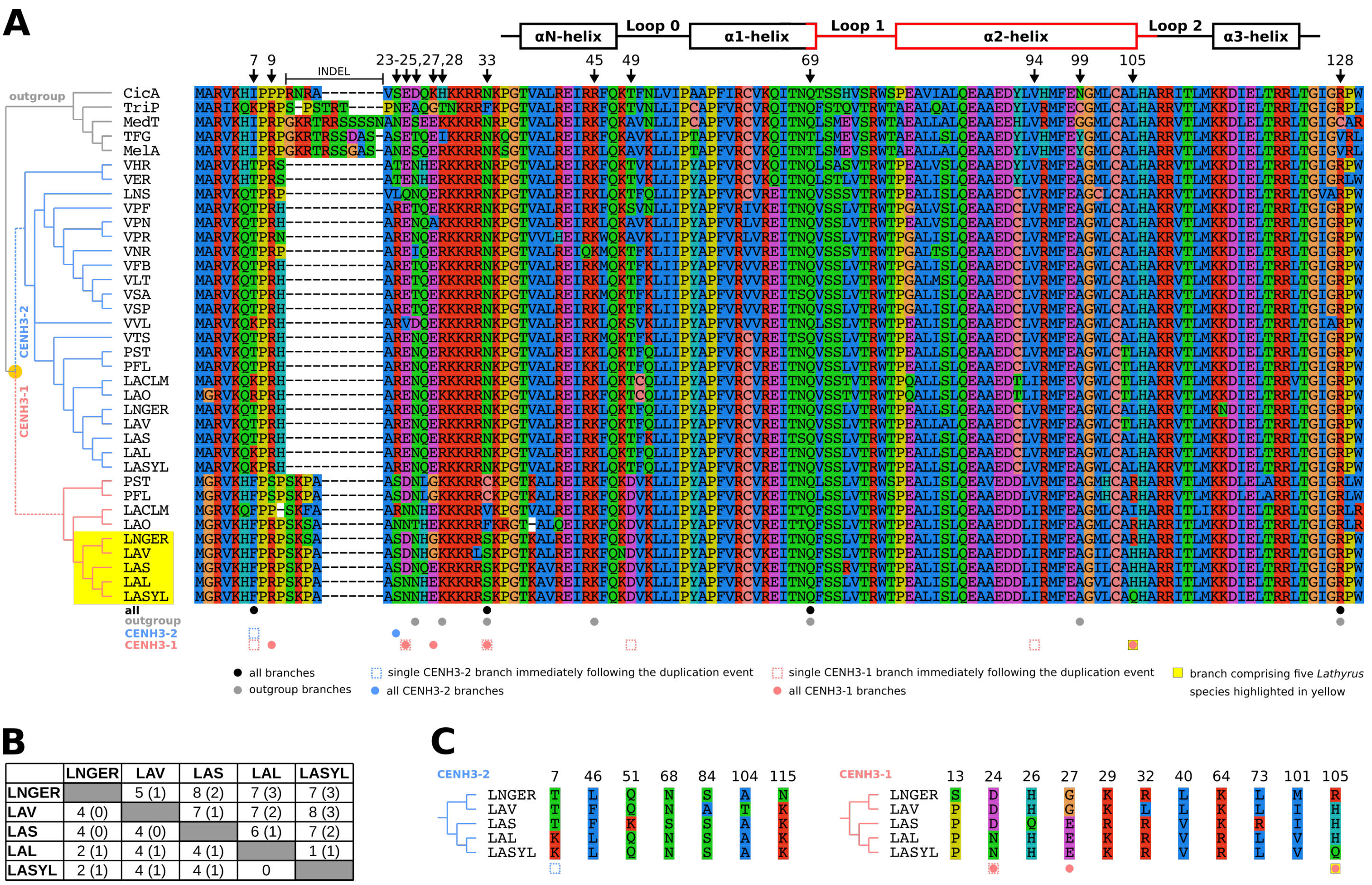
Supplementary fig. 3B-D. Circos plots constructed as in the panel 3A but selectively showing only three individual superfamilies. Sequence similarity dotplots on the right show comparisons of corresponding satellites from different species. Dimers of the consensus monomer sequences were used for the dotplots.

Supplementary fig. 4



Supplementary fig. 4. Localization of CENH3 (green) and FabTR-2 (red) on metaphase chromosomes of *Lathyrus sativus* (A-D) and *L. vernus* (E-H) detected by immunostaining followed by FISH. In both species FabTR-2 co-localized with CENH3 as predicted by ChIP-seq analysis. (I-J) Localization of two species-specific centromeric satellites (FabTR-30 and FabTR-32; red) on metaphase chromosomes of *V. peregrina*. Chromosomes counterstained with DAPI are shown in blue (D,H) or gray (I,J).

Supplementary fig. 5



	LNGER	LAV	LAS	LAL	LASYL
LNGER		5 (1)	8 (2)	7 (3)	7 (3)
LAV	4 (0)		7 (1)	7 (2)	8 (3)
LAS	4 (0)	4 (0)		6 (1)	7 (2)
LAL	2 (1)	4 (1)	4 (1)		1 (1)
LASYL	2 (1)	4 (1)	4 (1)	0	

Supplementary fig. 5. Comparison of CENH3 sequences. (**A**) Alignment of CENH3 protein sequences. The CENH3 sequences are ordered according to species tree inferred from *matK-rbcL* sequences which is shown on the left. CENH3-1 and CENH3-2 branches are highlighted in blue and red, respectively. Dashed lines mark the CENH3-1 and CENH3-2 branches immediately following the duplication event. Sites that were predicted using FEL and MEME to evolve under pervasive or episodic diversifying positive selection are marked with vertical black arrows. Numbers above the arrows show positions in the alignment. Symbols below the alignment indicate branches in which the sites under diversifying positive selection were predicted. The secondary structure of histone fold domain is shown above the alignment, as adopted from (Tachiwana *et al.* 2011). The putative centromere targeting domain is shown in red. (**B**) A table showing numbers of substitutions among CENH3-1 (above the diagonal) and CENH3-2 (below the diagonal) protein sequences from five *Lathyrus* species that possess the same centromeric satellite (FabTR-2). Numbers in brackets show how many of the substitutions were found at sites predicted to evolve under positive selection in any of the tests described in the table 2. Note, however, that the tests carried out specifically for this lineage predicted only one site in CENH3-1 and none in CENH3-2. (**C**) Distributions of variable sites in the alignments of CENH3-1 and CENH3-2 protein sequences of the five *Lathyrus* species compared in panel B. Note that positions of the variable sites are different between CENH3-1 and CENH3-2.

Supplementary table 1. Antibodies used for ChIP.

Antibody ID	Based on	Source	Peptide sequence	Reference	Species	ChIP result
P22	CENH3-1, *P. sativum*	Rabbit	GRVKHFPSPSKPAASDNLGKKKRRCKPGTKC	Neumann et al. 2012	*Pisum fulvum* [1]	ok
					Pisum sativum [1]	ok
P23	CENH3-2, *P. sativum*	Chicken	TPRHARENQERKKRRNKPGC	Neumann et al. 2012	*Lathyrus clymenum* [1]	ok
					Lathyrus vernus	ok
					Lathyrus latifolius [1]	ok
P31	CENH3-2, *V. faba*	Rabbit	CQTPRHARETQEKKKRRNKPG	Neumann et al. 2015	*Vicia faba* [1]	ok
					Vicia peregrina	ok
					Vicia pisiformis	ok
					Vicia sativa [1]	ok
					Vicia narbonensis [2]	No enrichment
					Vicia hirsuta	No enrichment
					Vicia ervilia	No enrichment
					Vicia tetrasperma	ok
					Vicia villosa	ok
					Lathyrus sativus [1]	ok
P45	CENH3-2, *L. culinaris*	Rabbit	PRPVLQNQERKKRRNKPGC	Neumann et al. 2015	*Lens culinaris* [1]	ok
					Vicia narbonensis [2]	No enrichment
P60	CENH3-2, *L. sativus*	Rabbit	QTPRHARENQERKKRRNKC	this study	*Lathyrus niger*	ok

[1] The centromeric specificity of the antibodies in these species has previously been demonstrated using in *situ* immunodetection by Neumann et al. (2012, 2015).

[2] Two ChIP-seq experiments were performed in *Vicia narbonensis*.

Supplementary table 2. Other (non-satellite) repetitive elements enriched in ChIP. RepeatExplorer contigs representing consensus sequences of these repeats are provided in supplementray file 4.

Species	ID	Repeat	Genome %	ChIP enrichment
V. pisiformis	VPF_CL121	Unclassified	0.186	66.8
	VPF_CL150	LTR	0.107	103.5
	VPF_CL240	LTR/gypsy/chromo	0.017	3.17
V. peregrina	VPR_CL14	LTR	0.772	13
	VPR_CL81	LTR	0.382	15.5
	VPR_CL127	LTR	0.196	16.9
V. sativa	VSA_CL16	LTR	0.215	77
V. villosa	VVL_CL77	LTR/gypsy/chromo	0.282	12.6
	VVL_CL228	LTR/gypsy/chromo	0.025	84.1
	VVL_CL255	LTR/copia/Angela	0.019	85.1
V. tetrasperma	VTS_CL168	LTR/copia/Maximus	0.118	64.96

Supplementary table 3. Sources of *matK* and *rbcL* sequences

Species	**Code**	***matK***	***rbcL***
Cicer arietinum	CicA	Neumann et al., 2015	EU835853.1
Trifolium pratense	TriP	KX538847.1	KF241982.1
Medicago trucatula	MedT	Neumann et al., 2015	KF241982.1
Trigonella foenum-graecum	TFG	AF522147.2	MG946901.1
Melilotus albus	MelA	Neumann et al., 2015	KP126850.1
Vicia hirsuta	VHR	NGS data	NGS data
Vicia ervilia	VER	NGS data	NGS data
Lens culinaris	LNS	Neumann et al., 2015	JN661189.1
Vicia pisiformis	VPF	JX505896.1	JX505517.1
Vicia pannoica	VPN	Neumann et al., 2015	NGS data
Vicia peregrina	VPR	Neumann et al., 2015	NGS data
Vicia narbonensis	VNR	Neumann et al., 2015	NGS data
Vicia lathyroides	VLT	Neumann et al., 2015	NGS data
Vicia sativa	VSA	Neumann et al., 2015	JN661204.1
Vicia sepium	VSP	Neumann et al., 2015	NGS data
Vicia villosa	VVL	Neumann et al., 2015	JN661208.1
Vicia tetrasperma	VTS	NGS data	NGS data
Pisum sativum	PST	Neumann et al., 2015	JN661190.1
Pisum fulvum	PFL	Neumann et al., 2015	NGS data
Lathyrus clymenum	LACLM	JX505793.1	KJ850235.1
Lathyrus ochrus	LAO	PCR amplification from genomic DNA	JN661184.1
Lathyrus niger	LNG	PCR amplification from genomic DNA	HE963532.1
Lathyrus vernus	LAV	Neumann et al., 2015	NGS data
Lathyrus sativus	LAS	Neumann et al., 2015	NC014063
Lathyrus latifolius	LAL	Neumann et al., 2015	HM029364.1
Lathyrus sylvestris	LASYL	PCR amplification from genomic DNA	PCR amplification from genomic DNA

Supplementary table 4. Primers used for *CENH3* amplification.

Species	Reverse transcription primer	Reaction type	Forward PCR primer	Reverse PCR primer	Reaction profile
V. tetrasperma	GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTTTTTTV	RT-PCR	CGGTTGCTCCAAGTTCAT	TCAGCAACAATGGTTTTCAC	94°C - 50 sec; 35 cycles of 94°C - 30 sec, 55°C - 50 sec, 72 °C - 1 min; 72 °C - 10 min
V. ervilia	GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTTTTTTV	RT-PCR	CGTTGCTCCAAGTTCATTTAG	GGCTTTCACTACAGGTGCC	94°C - 50 sec; 35 cycles of 94°C - 30 sec, 55°C - 50 sec, 72 °C - 1 min; 72 °C - 10 min
V. pisiformis	GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTTTTTTV	RT-PCR	CAGAATCAAATGGCGAGAG	ATGTTGTGTCGGTTCTCTCA	94°C - 50 sec; 35 cycles of 94°C - 30 sec, 55°C - 50 sec, 72 °C - 1 min; 72 °C - 10 min
V. hirsuta	GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTTTTTTV	3'RACE	CTTTGACTGCACATAATCAAATG	CACGCGTCGACTAGTACTTTT	94°C - 50 sec; 35 cycles of 94°C - 30 sec, 55°C - 50 sec, 72 °C - 1 min; 72 °C - 10 min

Chapter III:

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

the plant journal

S E B 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.

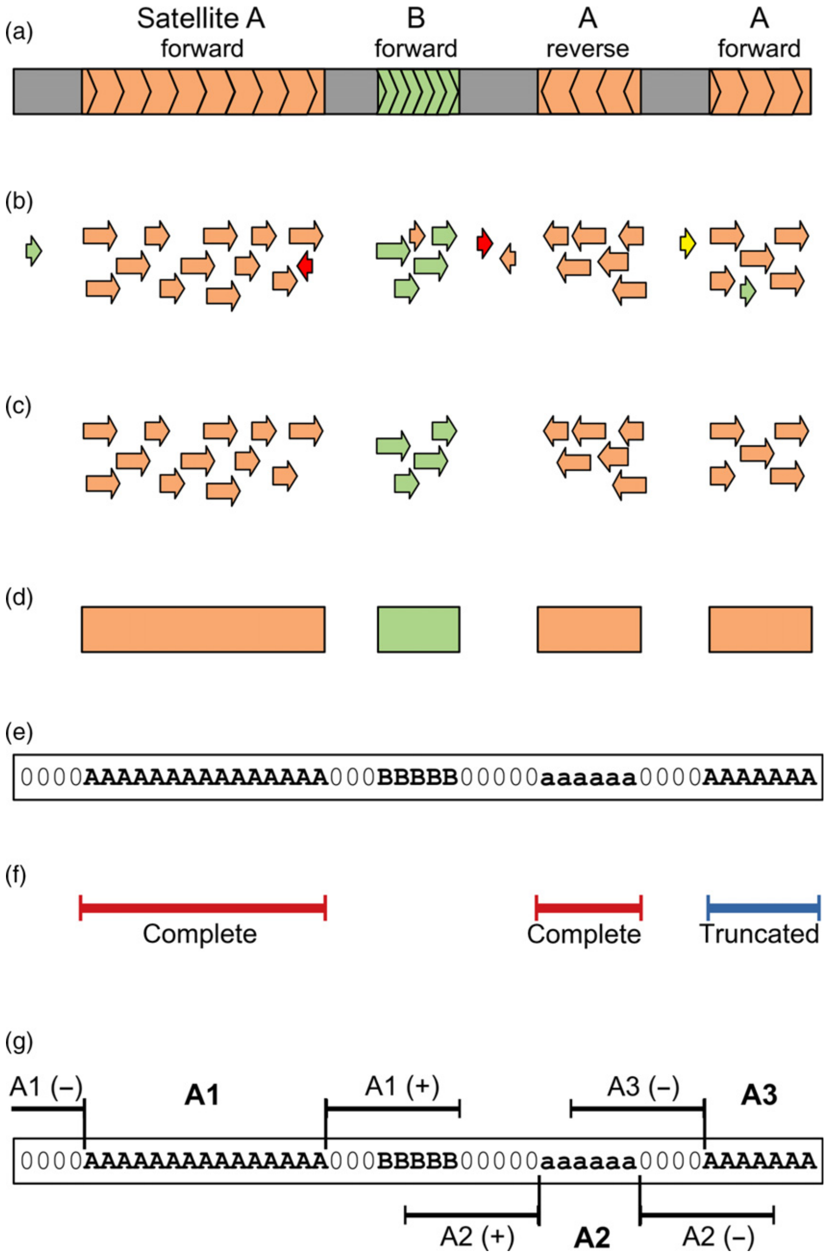


Figure 1. Schematic representation of the analysis strategy. (a) Nanopore read (grey bar) containing arrays of satellites A (orange) and B (green). The orientations of the arrays with respect to sequences in the reference database are indicated. (b) LASTZ search against the reference database results 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

© 2019 The Authors.

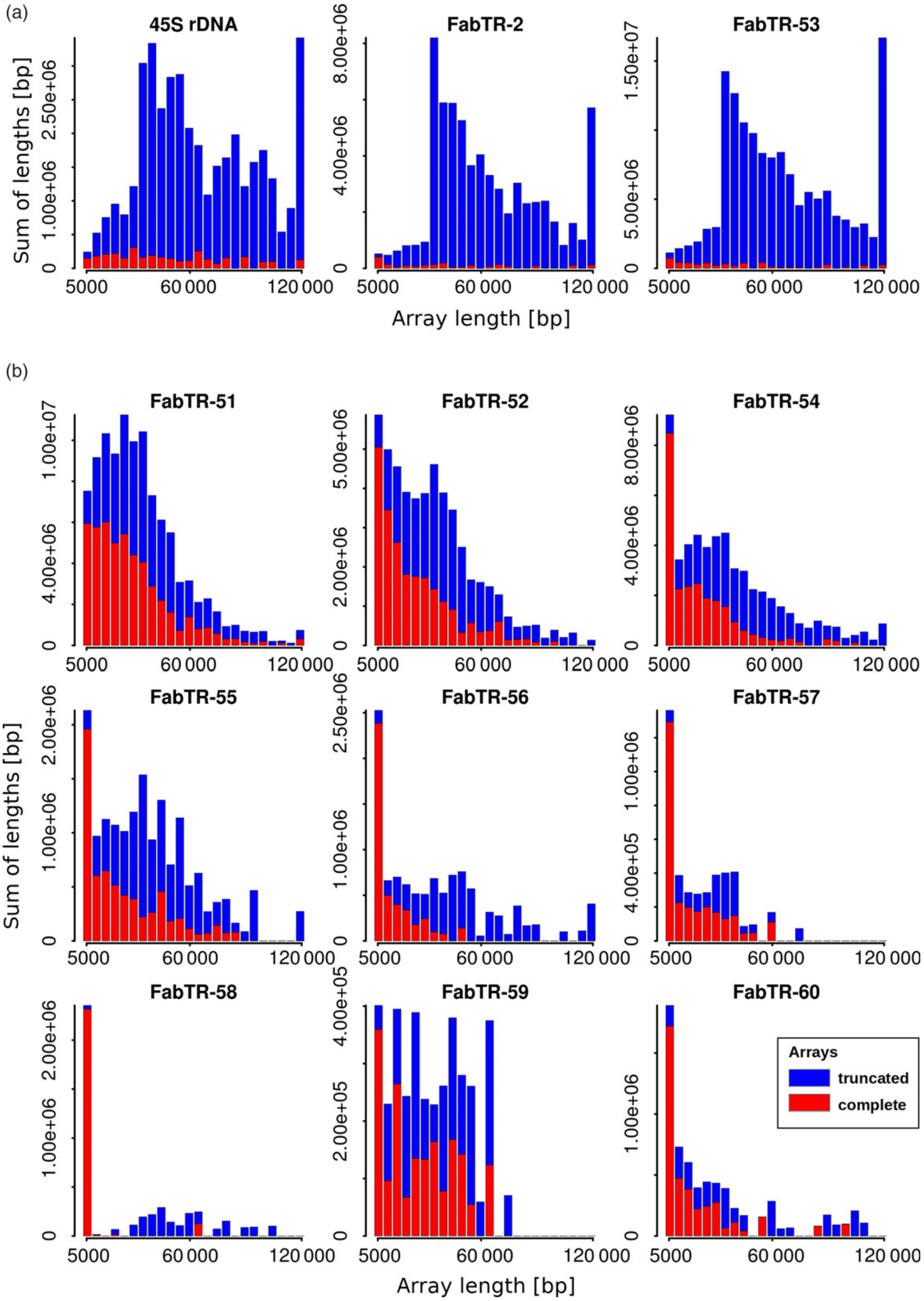


Figure 2. Length distributions of the satellite repeat arrays. The lengths of the arrays detected in the nanopore reads are displayed as weighted histograms with a bin size of 5 kb; the last bin includes all arrays longer than 120 kb. The arrays that were completely embedded within the reads (red bars) 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).

© 2019 The Authors.

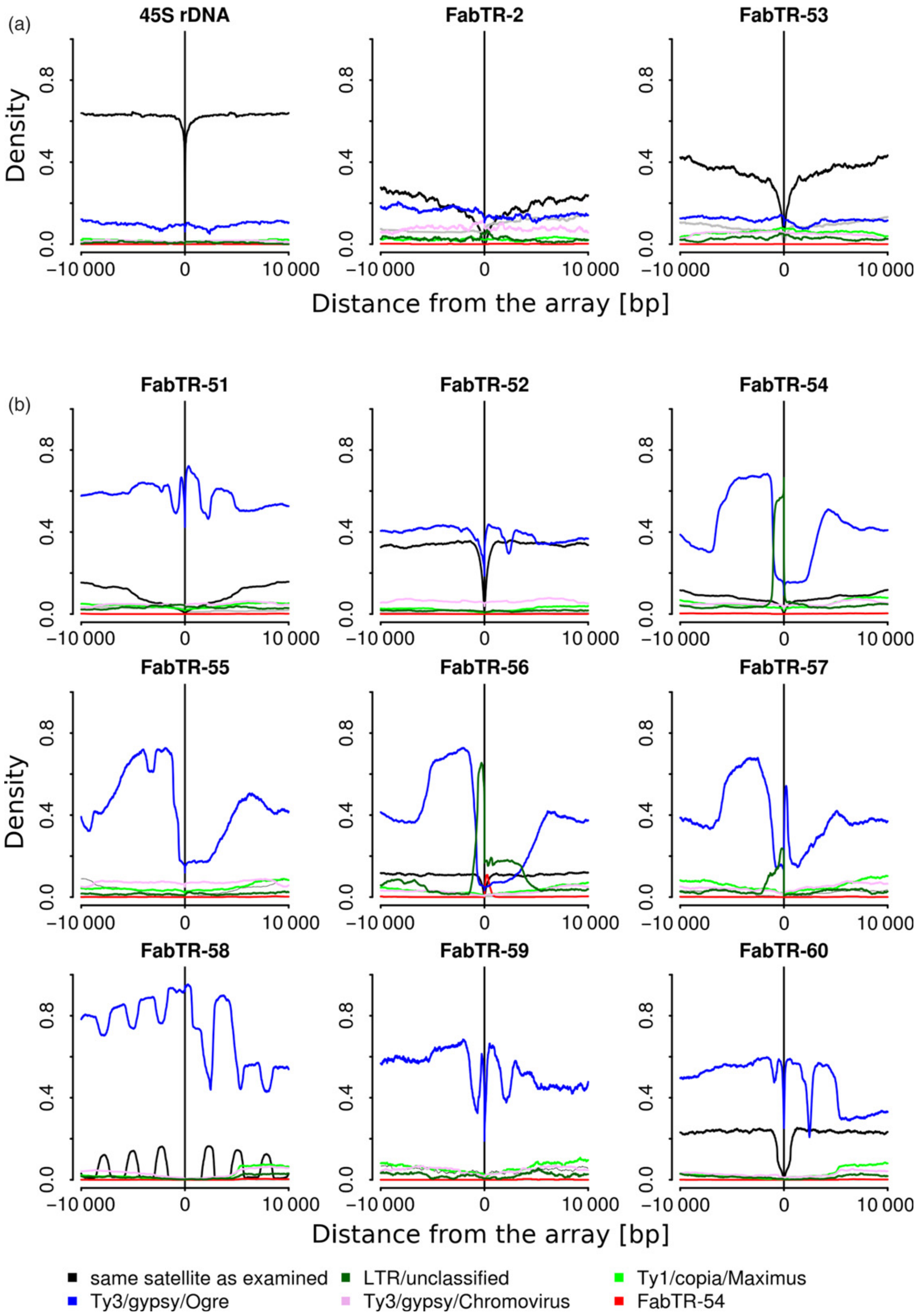


Figure 3. Sequence composition of the genomic regions adjacent to the satellite repeat arrays. The plots show the proportions of repetitive sequences identified within 10 kb regions upstream (positions −1 to −10 000) and downstream (1 to 10 000) of the arrays of individual satellites (the array positions 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

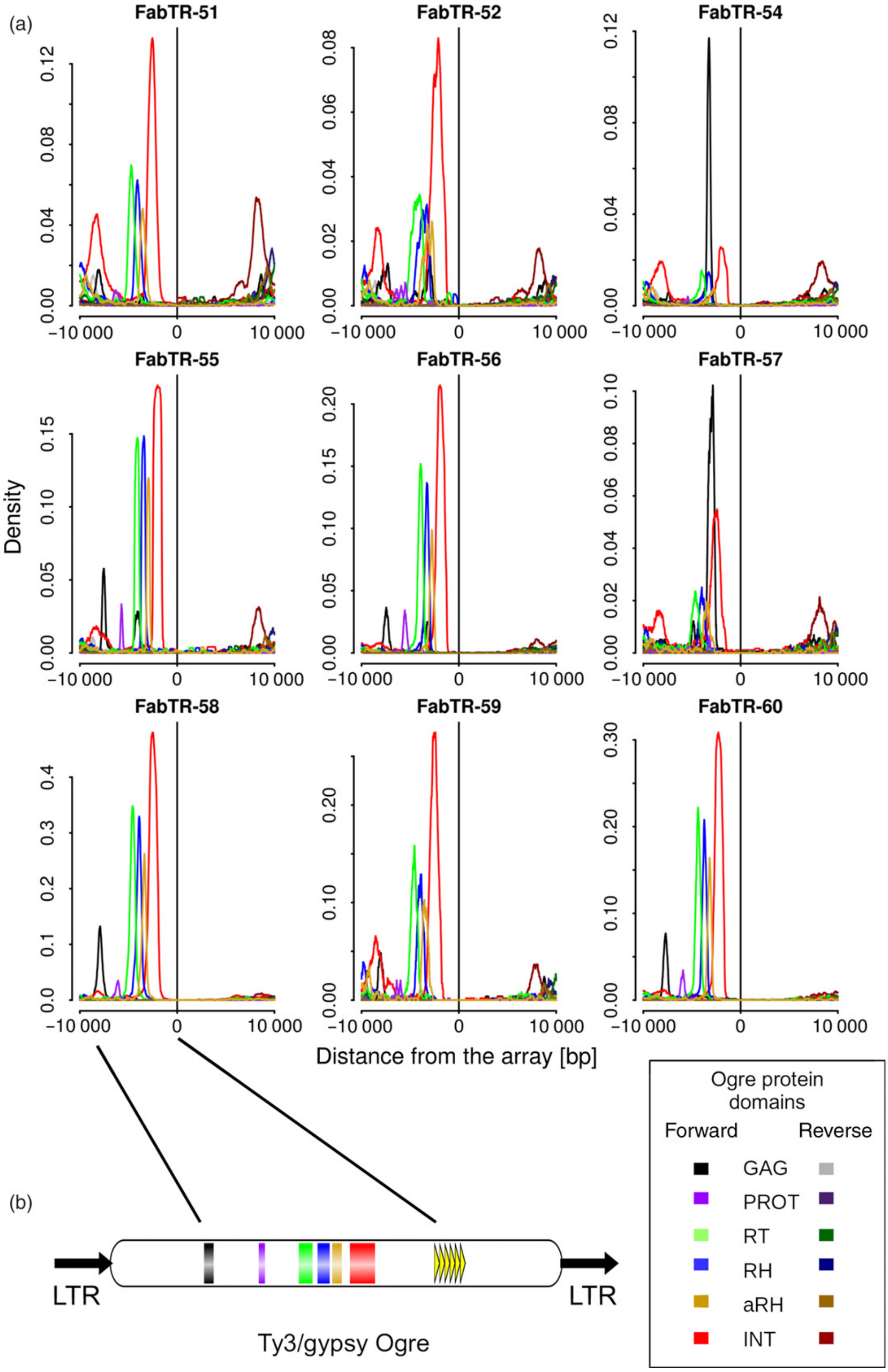
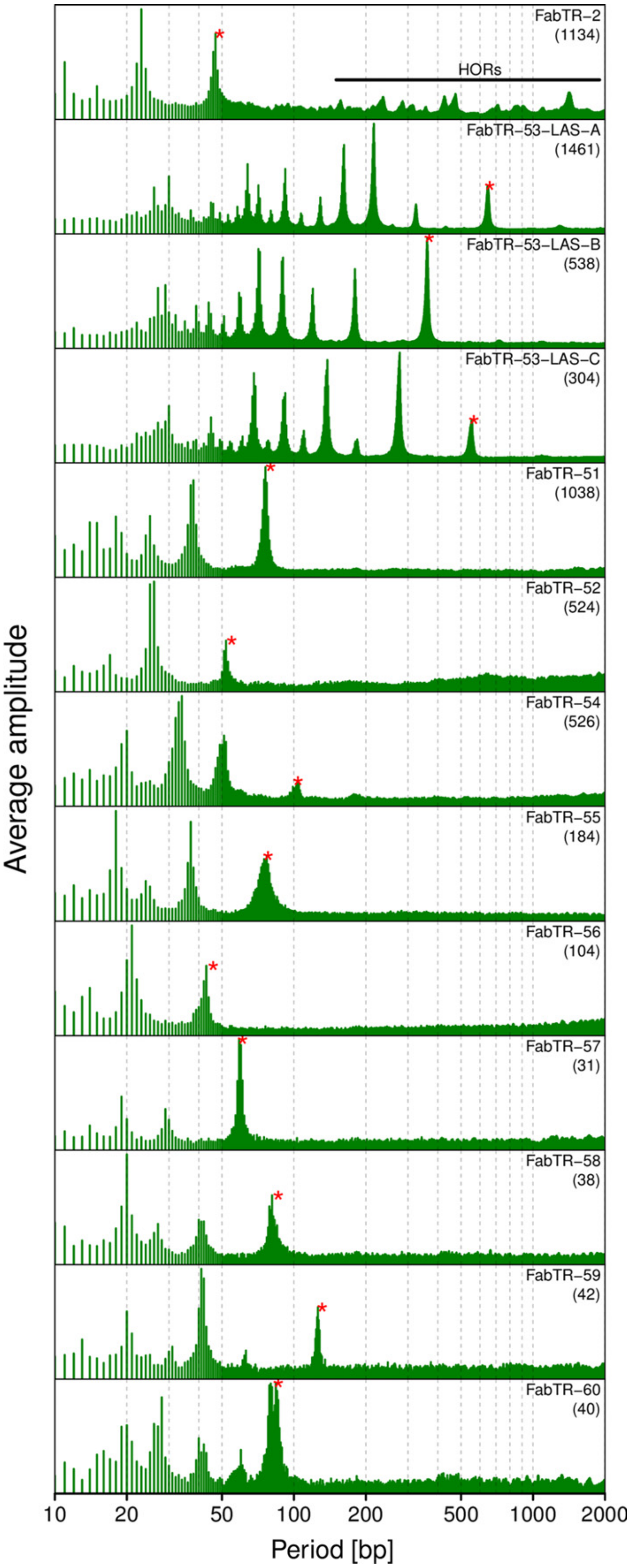


Figure 4. Detection of the Ogre sequences coding for the retrotransposon conserved protein domains in the genomic regions adjacent to the satellite repeat arrays. (a) The plots show the proportions of similarity hits from the individual domains and their orientation with respect to the forward-oriented 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.

© 2019 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd,

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

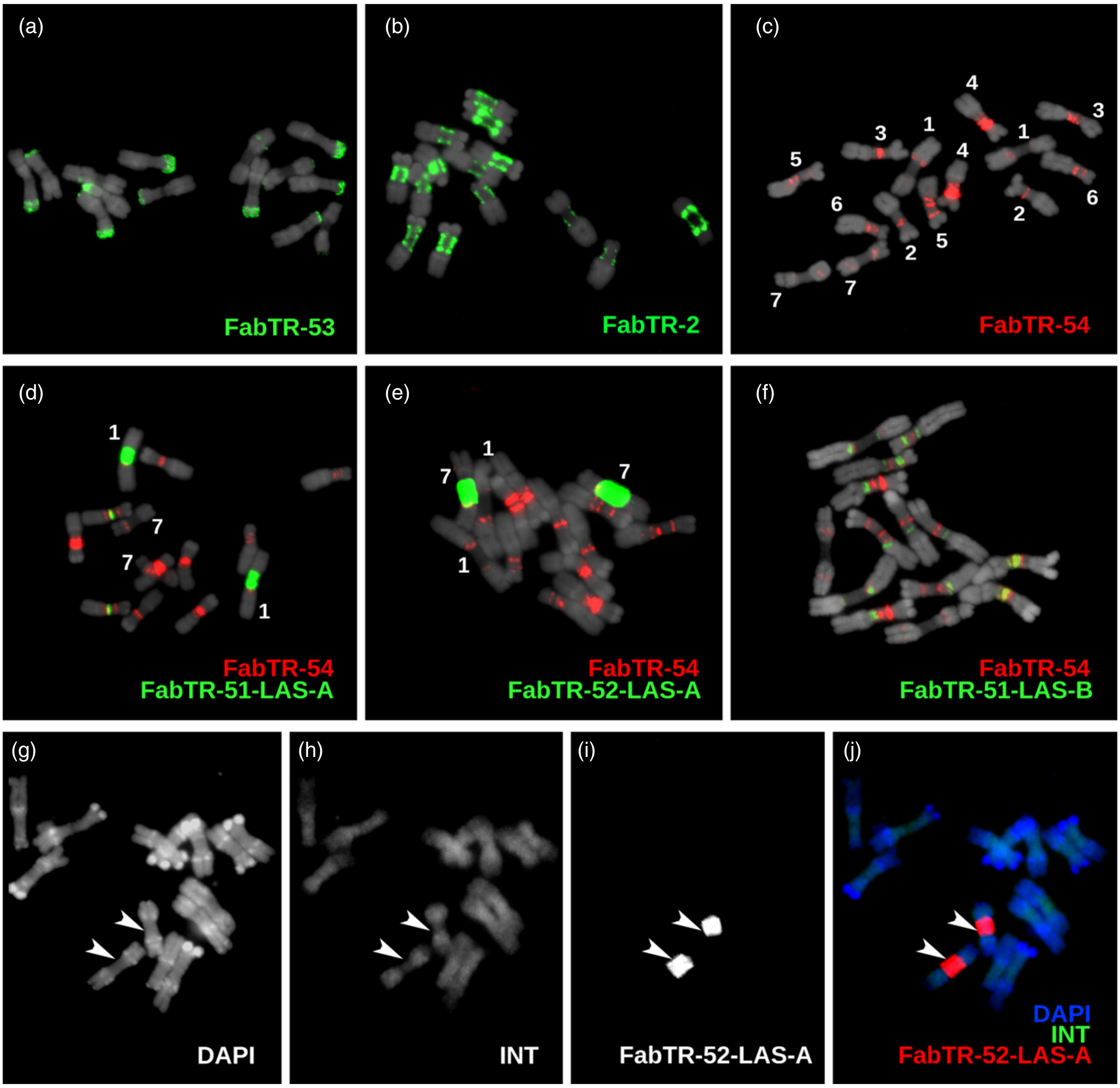


Figure 6. Distribution of the satellite repeats on the metaphase chromosomes of *Lathyrus sativus* ($2n = 14$). (a–f) The satellites were visualized using multi-colour FISH, with individual probes labelled as indicated by the colour-coded descriptions. The chromosomes counterstained with DAPI are shown 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′-CCTTC RTARTTGGGAGTCCA-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.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd,

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.

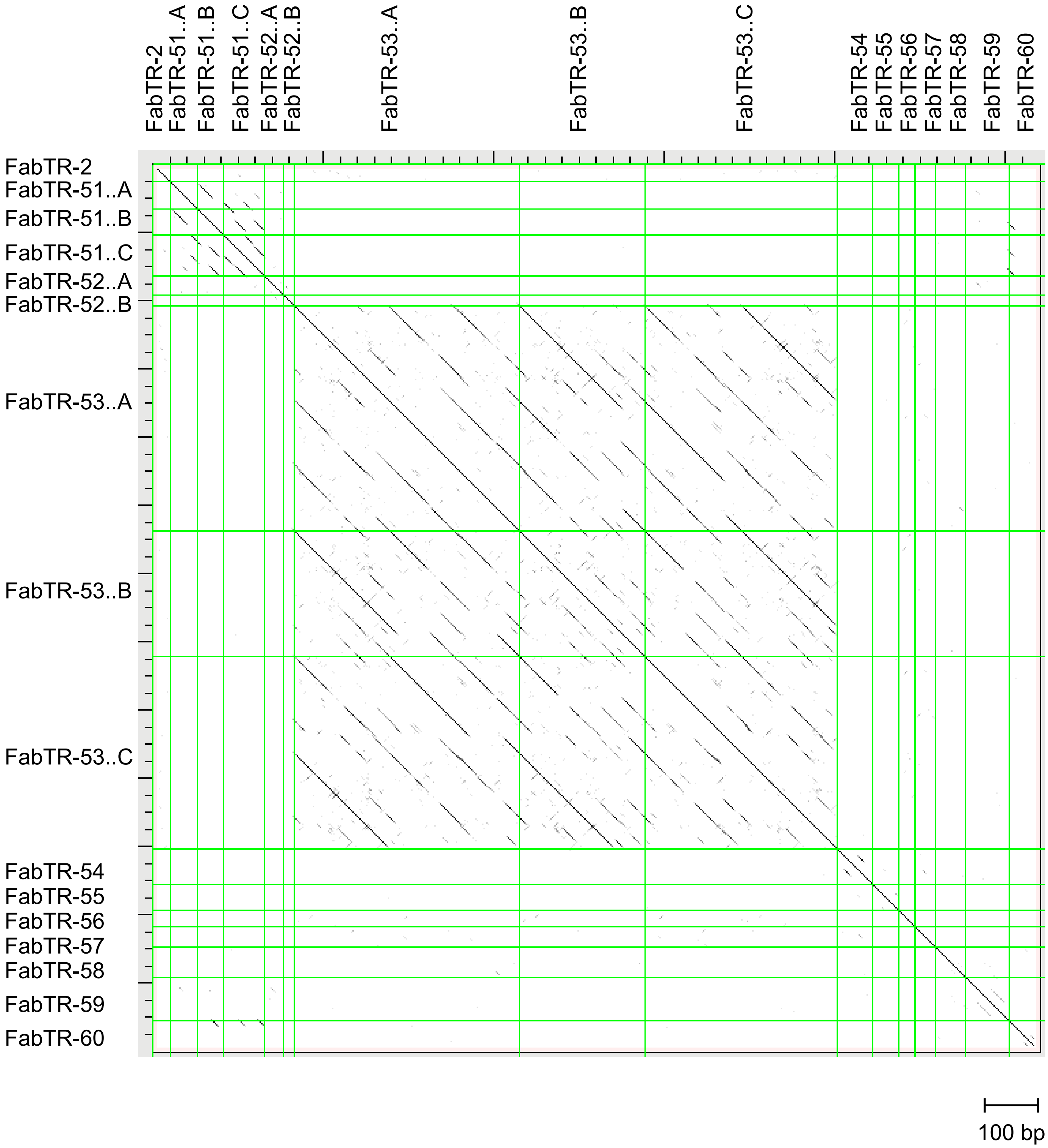
© 2019 The Authors.

Supplementary Information

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

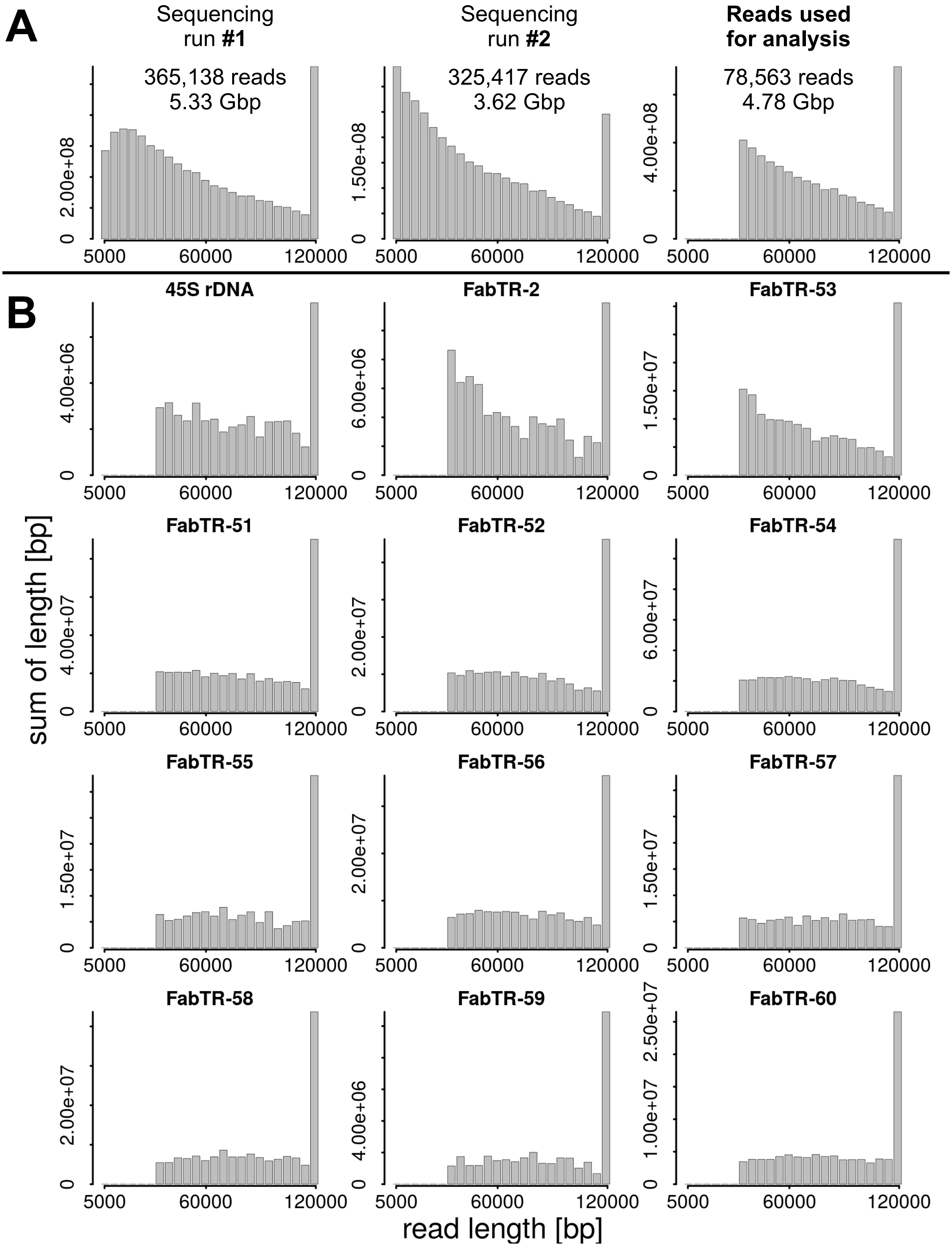
Tihana Vondrak, Laura Ávila Robledillo, Petr Novák, Andrea Koblížková, Pavel Neumann and Jiří Macas.

Supplementary Fig. S1



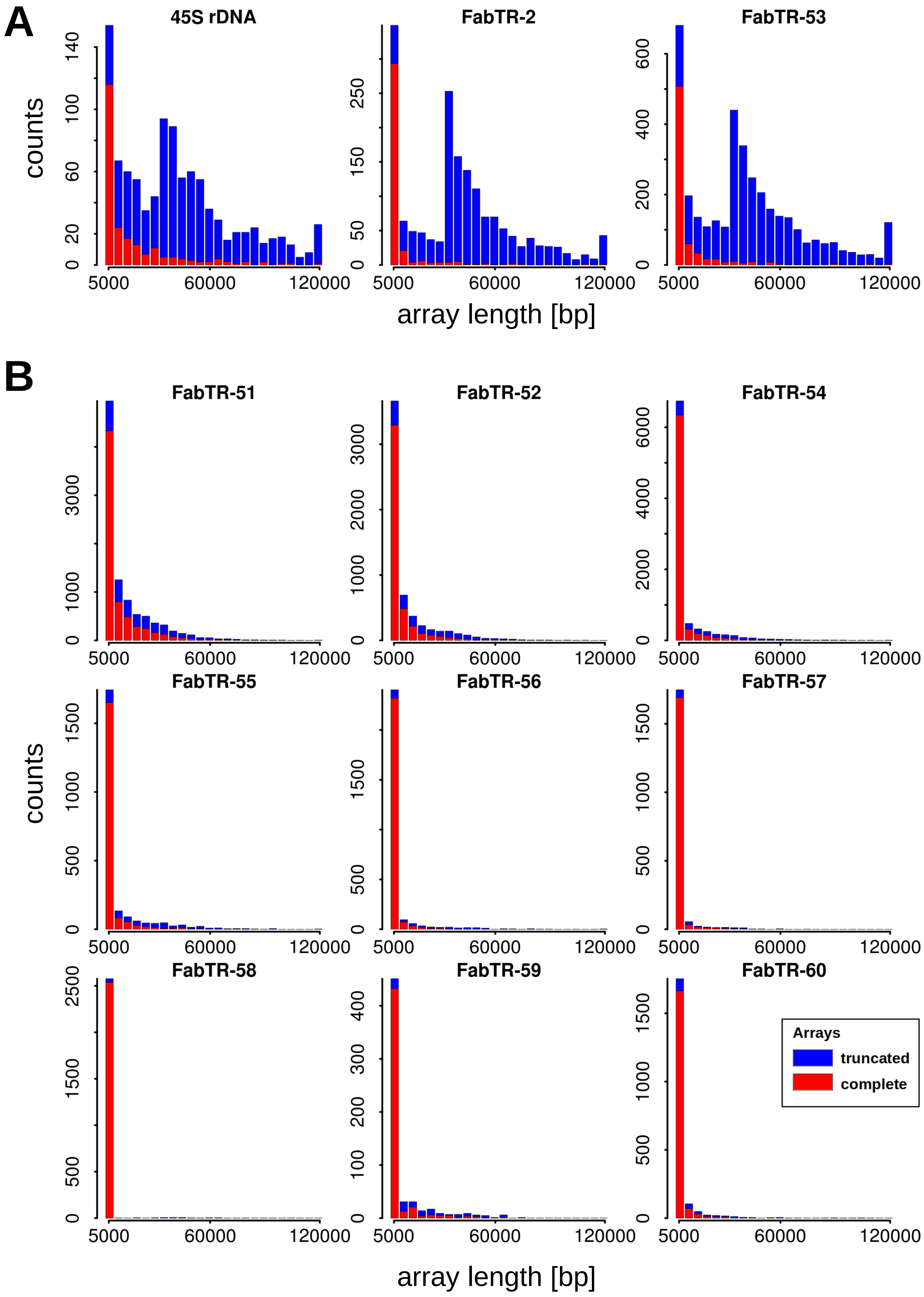
Supplementary Fig. S1. Dot-plot sequence similarity comparison of consensus monomer sequences. The sequences are separated by green lines and their similarities exceeding 40% over a 100 bp sliding window are displayed as black dots or diagonal lines.

Supplementary Fig. S2



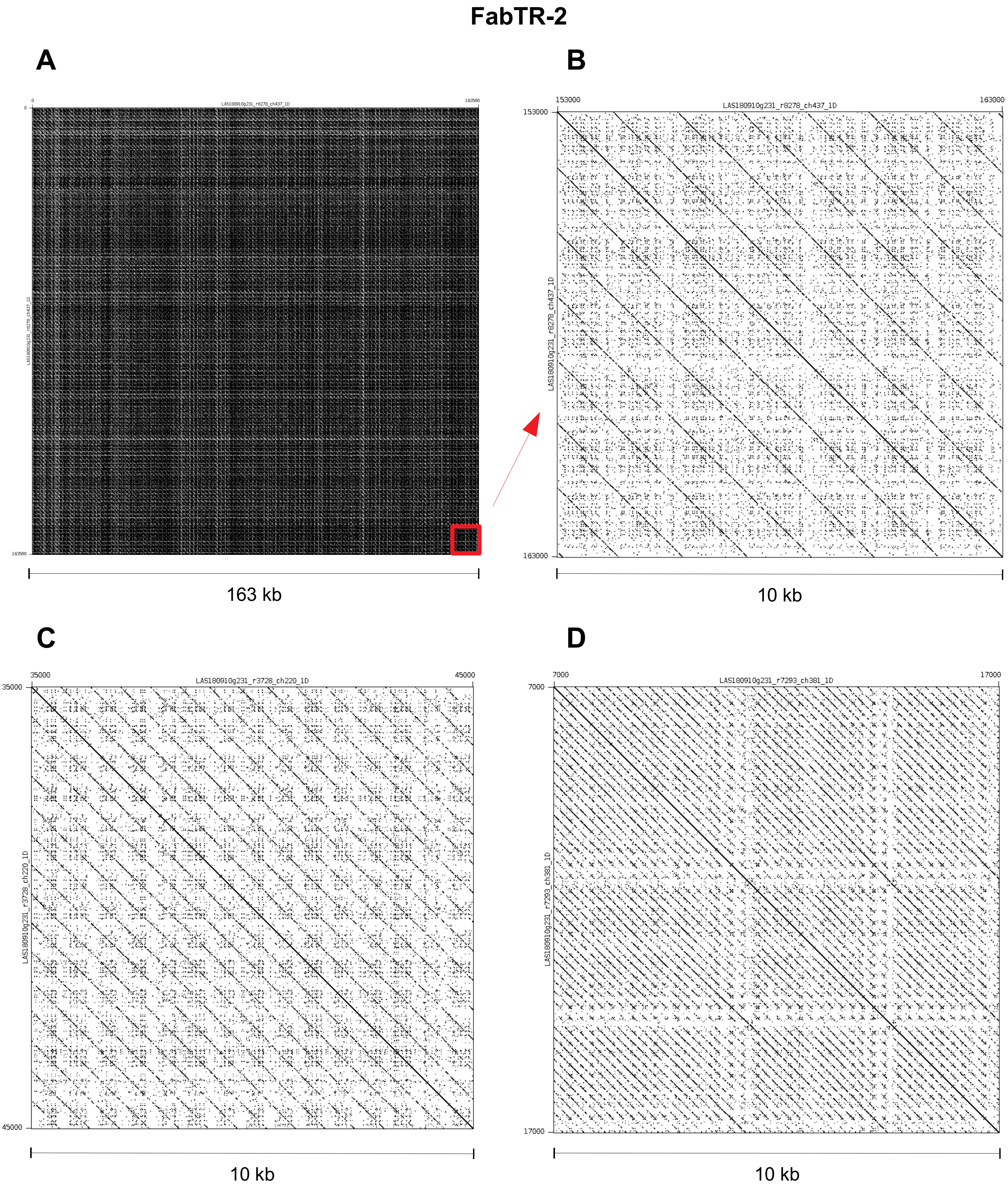
Supplementary Fig. S2. Length distributions of nanopore reads displayed as weighted histograms with bin size of 5 kb, with the last bin including all reads longer than 120 kb. (**A**) Length distributions of raw reads from two sequencing runs and the final set of quality-filtered and size-selected (>30kb)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

Supplementary Fig. S3



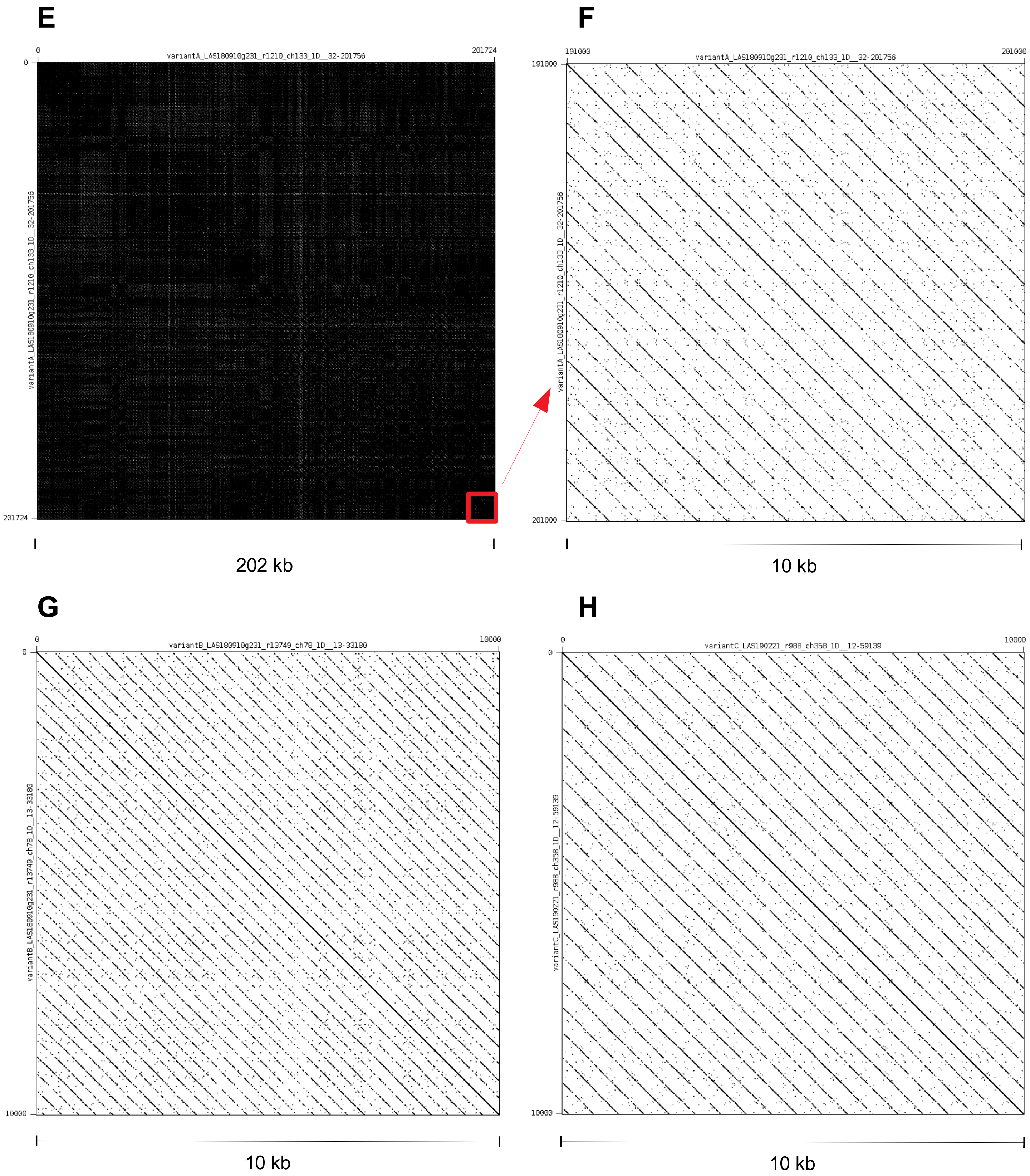
Supplementary Fig. S3. Length distributions of satellite repeat arrays displayed as histograms with bin size of 5 kb, with the last bin including all arrays longer than 120 kb. Arrays which were completely embedded within the reads (red bars) are distinguished from those truncated due to their positions 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

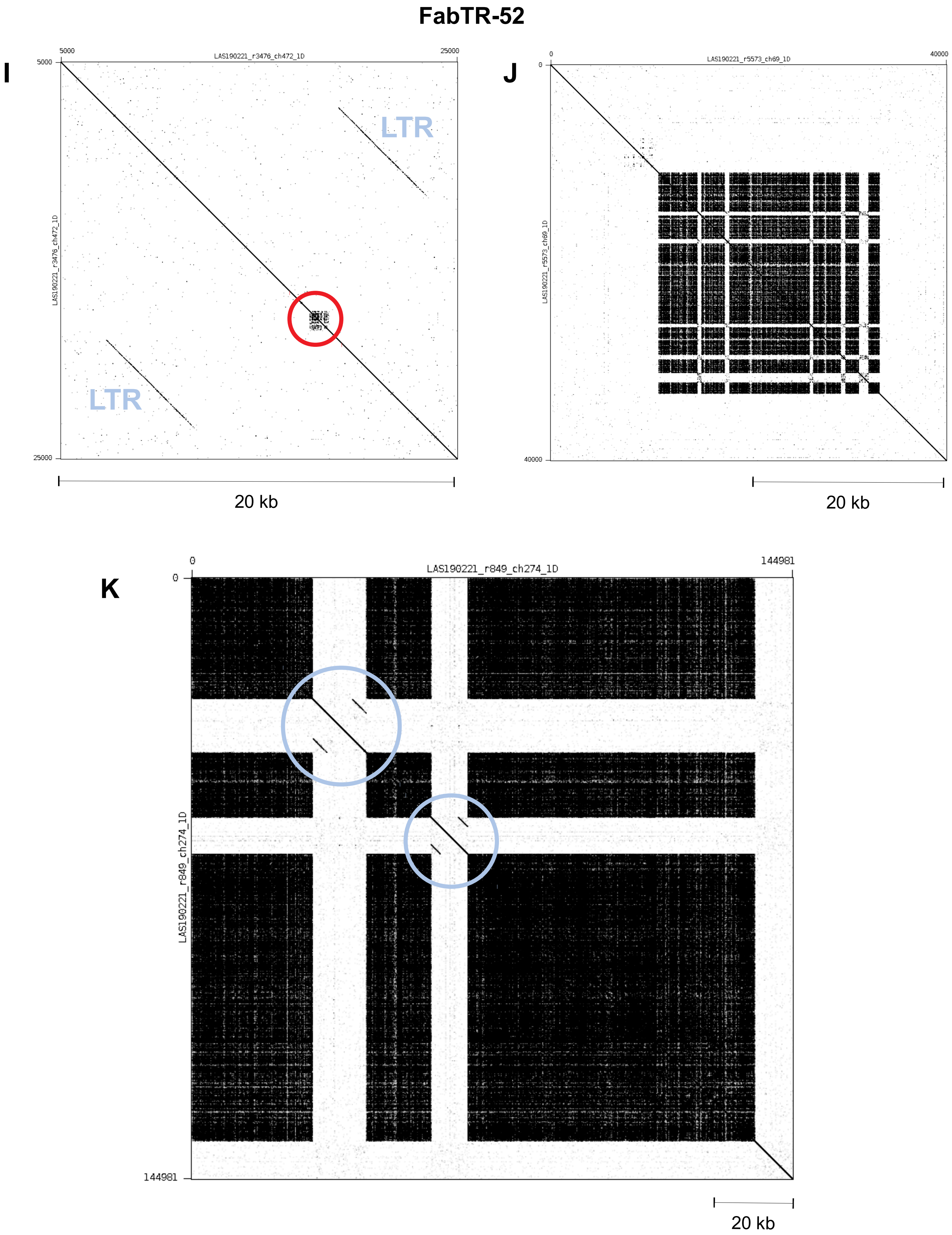


Supplementary Fig. S4 A-D. Self-similarity dot-plot visualization of FabTR-2 arrays. Tandem repeats are revealed as diagonal lines with spacing corresponding to monomer length. (**A**) Example of a 163 kb read completely made of FabTR-2 array (the periodicity pattern is obscured by the high density of 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).

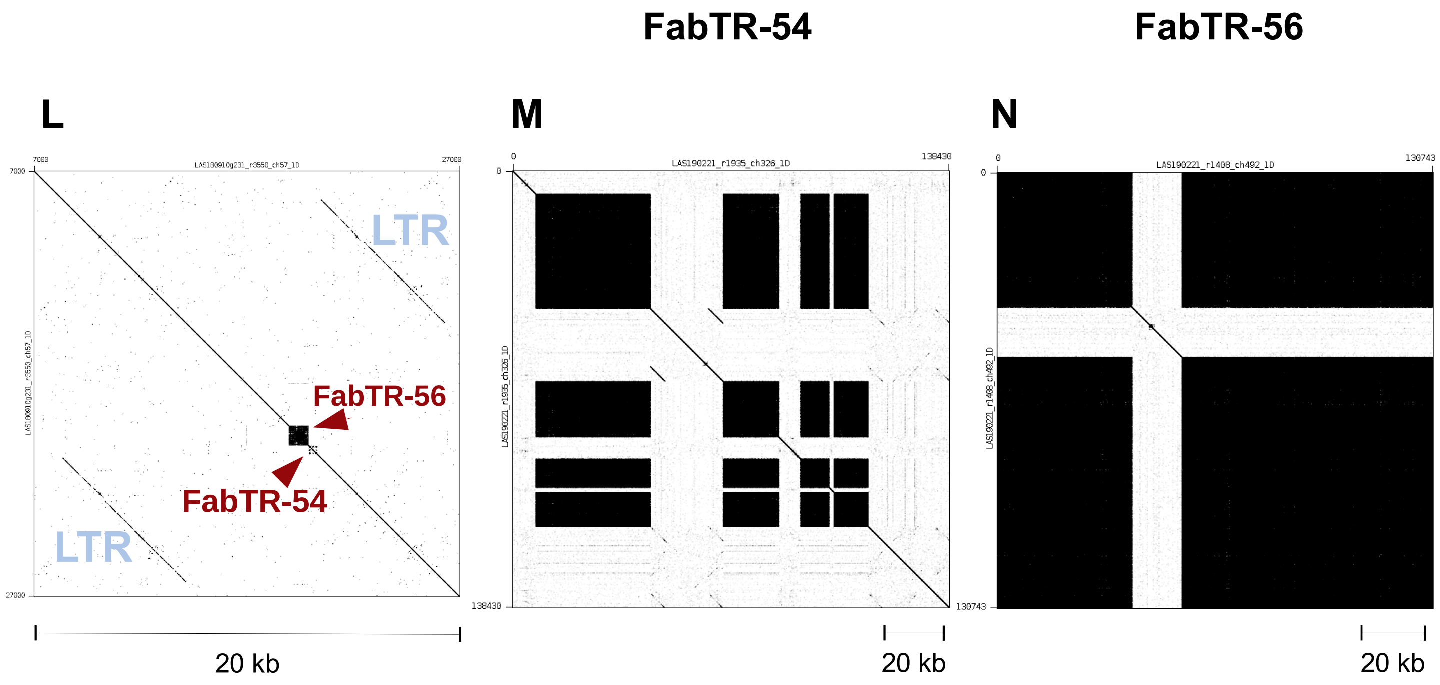
FabTR-53



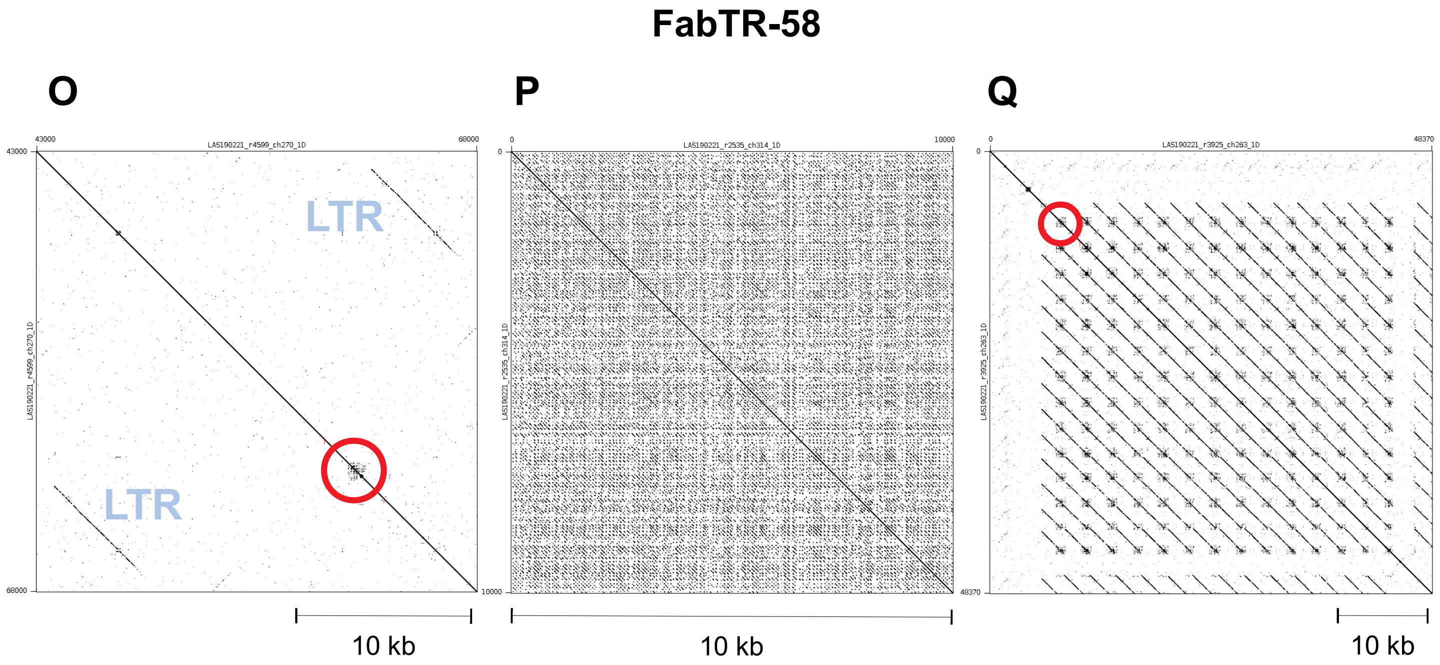
Supplementary Fig. S4 E-H. Self-similarity dot-plot visualization of FabTR-53 arrays. (**E**) Example of a 202 kb read completely made of FabTR-2 array (the periodicity pattern is obscured by the high density of lines). (**F**) Magnification of the 10 kb region highlighted by a red square on panel A. (**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).



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**).

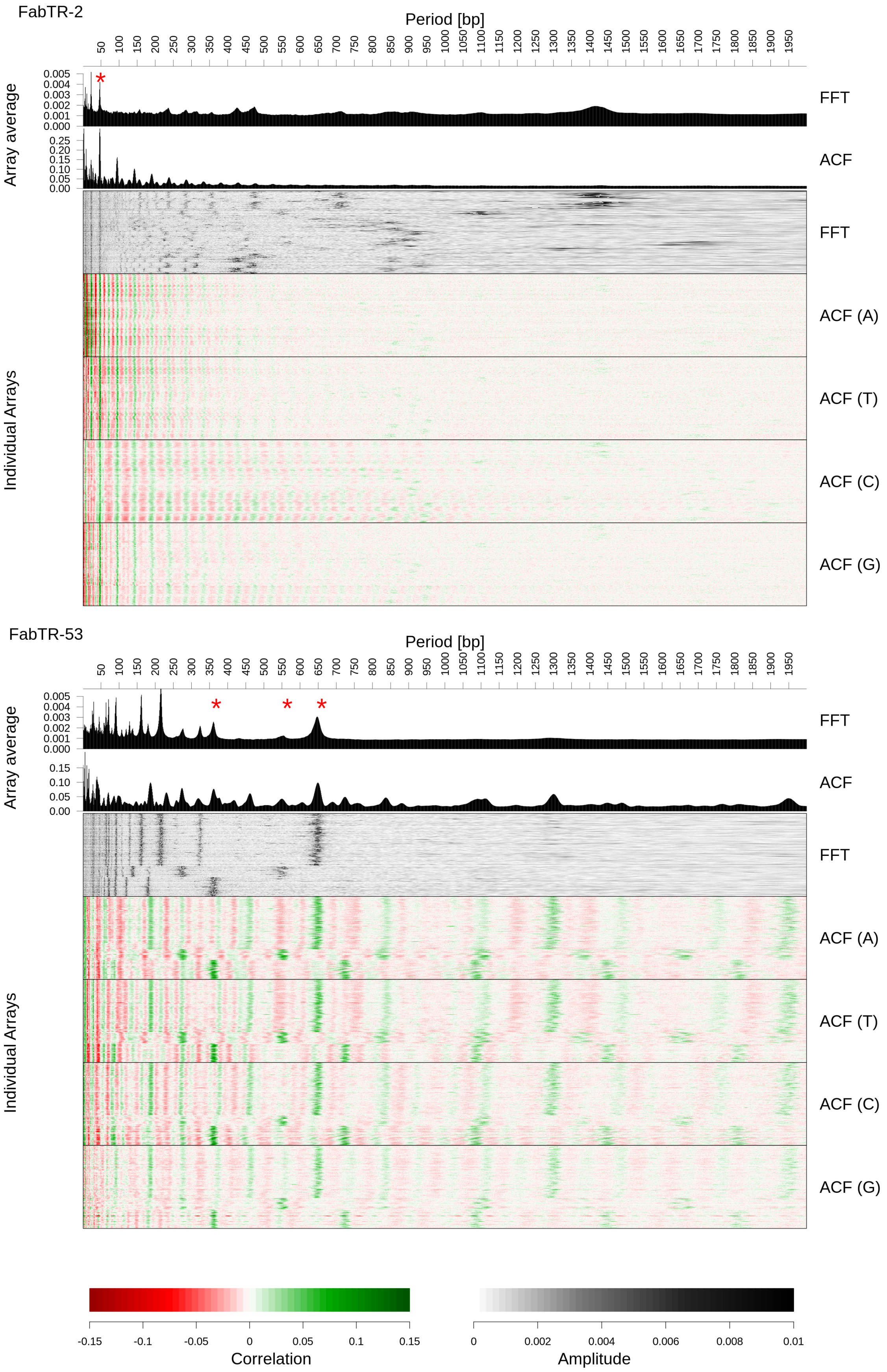


Supplementary Fig. S4 L-N. (**L**) Example of LTR-retrotransposon carrying short FabTR-54 and FabTR-56 arrays. Reads with those tandem repeats expanded to long arrays are shown on panels **M** (FabTR-54) and **N** (FabTR-56). The expanded tandem arrays appear as black squares on the dot-plots due to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



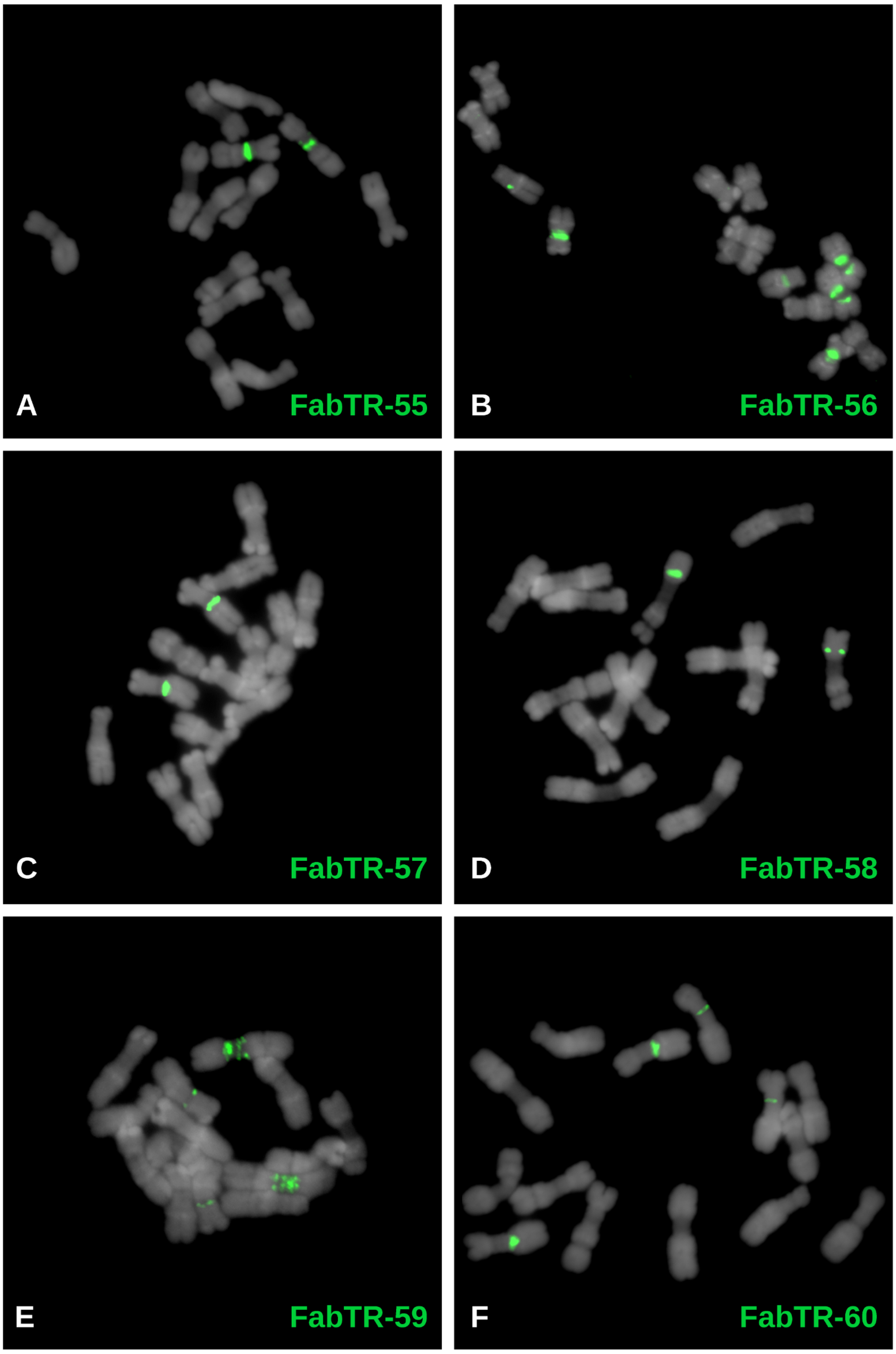
Supplementary Fig. S4 O-Q. Three types of genome organization of FabTR-58 repeats: (O) short array (marked by red circle) within LTR-retrotransposon, (P) expanded array, (Q) short arrays embedded within a longer tandem repeat monomer.

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

SUMMARY

This thesis demonstrates the advantages of new sequencing approaches, combined with specific bioinformatic tools, to *de novo* characterize satDNA in plant genomes. From this we have been able to collect precise information about the satDNA sequence features among different species of the tribe *Fabeae*. A complex pattern of the arrangement of different satDNA families was experimentally confirmed by FISH experiments in *Vicia faba* chromosomes, where different repeats often cluster around pericentromeres or interstitially. Although most of the families detected in *V. faba* show a preference for A-T sequences, there are few other similarities. The satDNA families identified for *V. faba* together with those identified in another 13 species of the tribe *Fabeae* were further evaluated using alignment-free sequence comparison. It was shown that most satellites sequences are species-specific, reflecting either their independent origin or rapid sequence diversification. Furthermore, this work demonstrated that long-monomer satDNAs are more common than previously envisioned using traditional techniques. Whether they represent early stages of satDNA evolution, or are the result of the homogenization and sequence diversification of shorter monomers, is yet to be determined.

In addition, this work is the most extensive study on centromeric proteins and associated satDNA families in a group of related plant species. It was shown that most *Fabeae* species carry several different satellites that are often species-specific. Among the satellites detected by our approach, those associated with centromeric chromatin often differ between species or even between chromosomes of the same species. Consequently, the data provided in this thesis challenges the centromere drive hypothesis. Thus, the arm race scenario proposed for the evolution of centromeric repeats is unlikely to occur in the *Fabeae* tribe since the presence of multiple centromeric satellites with different sequences rules out the possibility of any sequence-dependent co-evolution with the kinetochore proteins. However, since a number of features are thought to be important for centromere function, the data provided in this thesis allows for future investigation of these features, including dyad symmetries and WW dinucleotide periodicity in satellite repeat sequences or detecting centromere-derived transcripts.

This thesis also reports on the investigation of satDNA origin in *Lathyrus sativus,* by using ultra long reads derived from nanopore sequencing. The detection of the satellite arrays in nanopore reads revealed repeats with contrasting array length distributions. Furthermore, analysis of genomic sequences adjacent to the satellite arrays identified a group of satellites whose origin is linked to LTR retrotransposons. Our study demonstrate that the majority of satDNA originated from short tandem repeats located in the 3’UTR of Ogre elements. The array expansion of the retrotransposon-derived satellites occurred preferentially in the pericentromeric regions of *L. sativus* chromosomes.

CURRICULUM VITAE

Laura Ávila Robledillo

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

l.avila.robledillo@gmail.com

Born: Elche, Alicante (Spain), 20th November 1992

Nationality: Spanish

Languages: Spanish, English

Education:

September 2015 – Present: Ph.D. candidate at the University of South Bohemia, České Budějovice, Czech Republic.

September 2014 – July 2015: Master in Genetics and Evolution, Granada University, Granada (Spain)

September 2010 – September 2014: Bachelor degree in Biology, Granada University, Granada (Spain)

Publications:

Novák, P., **Ávila Robledillo, L.**, Koblížková, A., Vrbová, I., Neumann, P., & Macas, J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic acids research*, *45*(12), e111-e111. https://doi.org/10.1093/nar/gkx257

Ávila Robledillo, L., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., ... & Macas, J. (2018). Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Scientific Reports*, 8(1), 1-11. https://doi.org/10.1038/s41598-018-24196-3

Ávila Robledillo, L., Neumann, P., Koblížková, A., Novák, P., Vrbová, I., & Macas, J. (2020). Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe *Fabeae*. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msaa090

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. *Plant Journal*, 101(2), 484. https://doi.org/10.1111/tpj.14546

Conferences:

2015: Poster presentation at the "XL Congreso de la Socieadad Espanola de Genetica", Córdoba, Spain.

2018: Poster presentation at the "22nd International Chromosome Conference", Prague, Czech Republic.

© for non-published parts Laura Ávila Robledillo

l.avila.robledillo@gmail.com

Evolutyionary dynamics of satellite DNA in plant genomes

Ph.D. Thesis

All rights reserved

For non-commercial use only

University of South Bohemia in České Budějovice

Faculty of Science

Branišovská 1760

CZ-37005 České Budějovice, Czech Republic

Phone: +420 777 6540 76

www.prf.jcu.cz, e-mail: sekret-fpr@prf.jcu.cz