Progress in Molecular and Subcellular Biology

Đurđica Ugarković *Editor*

Satellite DNAs in Physiology and Evolution

Springer

Progress in Molecular and Subcellular Biology

Volume 60

This series gives an insight into the most current, cutting edge topics in molecular biology, including applications in biotechnology and molecular medicine. In the recent years, the progress of research in the frontier area of molecular and cell biology has resulted in an overwhelming amount of data on the structural components and molecular machineries of the cell and its organelles and the complexity of intra- and intercellular communication. The molecular basis of hereditary and acquired diseases is beginning to be unravelled, and profound new insights into development and evolutionary biology, as well as the genetically driven formation of 3D biological architectures, have been gained from molecular approaches. Topical volumes, written and edited by acknowledged experts in the field, present the most recent findings and their implications for future research. This series is indexed in PubMed.

More information about this series at http://www.springer.com/series/388

Đurđica Ugarković
Editor

Satellite DNAs in Physiology and Evolution

Springer

Editor
Đurđica Ugarković
Rudjer Boskovic Institute
Zagreb, Croatia

ISSN 0079-6484 ISSN 2197-8484 (electronic)
Progress in Molecular and Subcellular Biology
ISBN 978-3-030-74888-3 ISBN 978-3-030-74889-0 (eBook)
https://doi.org/10.1007/978-3-030-74889-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Non-coding repetitive DNAs constitute a considerable portion of most eukaryotic genomes, and their function is being intensively investigated. Among the most investigated non-coding repetitive DNAs are mobile transposable elements which represent an important source of regulatory sequences. The functional significance of another abundant class of non-coding repetitive elements such as satellite DNA is also now beginning to be discerned. Satellite DNAs are tandemly repeated sequences assembled within constitutive heterochromatin at the (peri)centromeric and subtelomeric regions. However, many satellite DNAs are not only clustered within centromeres or pericentromeric heterochromatin but are dispersed as short arrays within euchromatin, in the vicinity of genes. Besides playing a role in the modulation of global heterochromatin structure and centromere function, recent results reveal a role for satellite DNAs in gene expression regulation during different processes such as cell cycle progression, development, differentiation and stress responses. Here, we review the rapidly advancing field of satellite DNAs describing their structure, origin, organization and function in diverse eukaryotic systems. In addition, the evolutionary aspect of activation of satellite DNAs in terms of transcription and proliferation is highlighted, revealing the role of satellite DNAs in the process of adaptation to changing environment and in the speciation process. This book also deals with satellite DNA activation during pathological transformation and the mechanisms by which they affect disease progression.

Since the discovery of satellite DNAs more than 50 years ago, species from the *Drosophila* genus have continuously been used as models to study several aspects of satellite DNA biology. Chapter 1 written by Maggie P. Lauria Sneideman and Victoria H. Meller focuses on the functions of satellite repeats in *Drosophila* with particular attention to the properties that make satellites a versatile and powerful force in nuclear organization, gene regulation and evolution. The involvement of *Drosophila* satellite DNAs in dosage compensation, meiotic drive and hybrid incompatibilities is presented and discussed. Most satellite DNA studies in the *Drosophila* genus have been largely focused on *Drosophila melanogaster* and closely related species from the *Sophophora* subgenus, although the vast majority of all *Drosophila* species belong to the *Drosophila* subgenus. Chapter 2 by Gustavo C.S. Kuhn and collaborators deals with studies on satellite DNA structure, organization and evolution in two species groups from the *Drosophila* subgenus: the *repleta* and *virilis* groups. The authors highlight the centromeric satellite DNAs in these species groups, their common structural features and association of satellite DNAs with transposable elements.

Eva Šatović and Miroslav Plohl in the first part of Chap. 3 summarize the approaches that have contributed to the development of conceptual views and added new levels to the understanding of the biology of satellite DNAs. Continuing on, the topic of satellite DNA outside of heterochromatin and their association with mobile elements is discussed. Following these aspects, in the third part they present the current state of knowledge on satellite DNAs and heterochromatin in bivalve molluscs, the group of species with rapidly accumulating genome data and with certain peculiarities in abundance, ancestry, connection to transposable elements, conserved sequence boxes, methylation patterns and evolutionary aspects of satellite DNAs that bring into question the classical form of the "library model" within this group of organisms.

In Chap. 4, Juan Pedro M. Camacho and co-authors review the current state of knowledge related to the satellite DNA of the B chromosome. Since B chromosomes often contain a large amount of satellite DNA, the question arises whether satellite DNA has a functional significance or is it simply a consequence of B chromosome properties such as dispensability and late replication. The authors discuss the origin, evolution and possible function of B chromosome satellite DNA in different eukaryotes, in particular its role in B chromosome drive.

Satellite DNAs may comprise a significant portion of plant genomes and are often responsible for the genome size differences between related species. Chapter 5 written by Manuel A. Garrido-Ramos collects some of the most important advances and the main lessons that were learnt about plant satellite DNAs with respect to several aspects related to their origin, evolution and organization. In addition, the role of satellite DNAs in plant centromere and telomere function is discussed.

The recent findings on the role of satellite DNA in gene expression regulation/modulation and on the molecular mechanisms by which satellite DNA affects genes are presented in Chap. 6 written by Đurđica Ugarković and co-authors. The chapter is particularly focused on the impact of euchromatic satellite DNA repeats dispersed outside of (peri)centromeric regions as well as of satellite transcripts on gene expression modulation. Also discussed is the implication of satellite DNA-mediated gene regulation on the evolution of gene-regulatory networks and on the process of environmental adaptation as well as the effects and possible consequences on different physiological processes.

Centromeres are chromosomal domains specialized for the faithful segregation of the genetic material between daughter cells at each cell division, and in most higher eukaryotes, they are made up of satellite DNA. In Chap. 7, Claire Francastel and collaborators discuss the various roles proposed for centromere transcription and their transcripts and the potential molecular mechanisms involved. In addition, evidence is presented on the unscheduled transcription of centromeric repeats or aberrant accumulation of their transcripts and their association with various diseases.

Chapter 8 written by Vladimir Paar and collaborators focuses on higher order repeats (HORs) which are characteristic of many satellite DNAs, in particular of the major primate alpha satellite DNA. While HORs were so far largely investigated only within the centromeric region, here more attention is turned to the cases of HORs in human genes. The HOR-searching and monomer-searching methods are explained and discussed, and novel human HORs discovered using the HOR-searching method with GRM algorithm are presented. In addition, evolution of the HORs among different primate species is analysed and their potential functional significance is discussed.

In conclusion, the book gives a comprehensive overview of unique roles that satellite DNAs play in different physiological and evolutionary processes.

Zagreb, Croatia Đurðica Ugarković

Contents

List of Contributors

Branka Bruvo Mađarić Department of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

Josefa Cabrero Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain

Juan Pedro M. Camacho Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain

Guilherme Borges Dias Department of Genetics and Institute of Bioinformatics, University of Georgia, Athens, GA, USA

Isidoro Feliciello Department of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia
Dipartimento di Medicina Clinica e Chirurgia, Universita' degli Studi di Napoli Federico II, Naples, Italy

Claire Francastel Epigenetics and Cell Fate, CNRS, UMR7216 - Epigénétique et Destin Cellulaire, Université Paris 7 Diderot, Paris Cedex 13, France

Manuel A. Garrido-Ramos Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain

Matko Glunčić Faculty of Science, University of Zagreb, Zagreb, Croatia

Sabrine Hédouin Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Pedro Heringer Department of Genetics, Ecology and Evolution, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG, Brazil

Gustavo C. S. Kuhn Department of Genetics, Ecology and Evolution, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG, Brazil

Sven Ljubić Department of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

María Dolores López-León Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain

Victoria H. Meller Department of Biological Sciences, Wayne State University, Detroit, MI, USA

Pia Mihìc Epigenetics and Cell Fate, CNRS, UMR7216 - Epigénétique et Destin Cellulaire, Université Paris 7 Diderot, Paris Cedex 13, France

Vladimir Paar Croatian Academy of Sciences and Arts, Zagreb, Croatia
Faculty of Science, University of Zagreb, Zagreb, Croatia

Željka Pezer Department of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

Miroslav Plohl Ruđer Bošković Institute, Zagreb, Croatia

Marija Rosandić Croatian Academy of Sciences and Arts, Zagreb, Croatia
University Hospital Centre Zagreb, Zagreb, Croatia

Francisco J. Ruiz-Ruano Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain

Antonio Sermek Department of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

Maggie P. Lauria Sneideman Department of Biological Sciences, Wayne State University, Detroit, MI, USA

Đurðica Ugarković Department of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

Ines Vlahović Algebra University College, Zagreb, Croatia

Eva Šatović Vukšić Ruđer Bošković Institute, Zagreb, Croatia

Chapter 1
Drosophila Satellite Repeats at the Intersection of Chromatin, Gene Regulation and Evolution

Maggie P. Lauria Sneideman and Victoria H. Meller

Abstract Satellite repeats make up a large fraction of the genomes of many higher eukaryotes. Until recently these sequences were viewed as molecular parasites with few functions. *Drosophila melanogaster* and related species have a wealth of diverse satellite repeats. Comparative studies of Drosophilids have been instrumental in understanding how these rapidly evolving sequences change and move. Remarkably, satellite repeats have been found to modulate gene expression and mediate genetic conflicts between chromosomes and between closely related fly species. This suggests that satellites play a key role in speciation. We have taken advantage of the depth of research on satellite repeats in flies to review the known functions of these sequences and consider their central role in evolution and gene expression.

Keywords Satellite DNA · Heterochromatin · Dosage compensation · Meiotic drive

1.1 Introduction

Repetitive DNA makes up a large portion of the genomes of higher eukaryotes. Satellite DNA, composed of tandem repeats that assemble into constitutive heterochromatin, was first described when mouse DNA was subjected to density gradient centrifugation and "satellite bands" of different densities formed above or below the bulk of the genome (Kit 1961; reviewed in Garrido-Ramos 2017). Fifty years ago the pioneering technique of in situ hybridization to mitotic chromosomes demonstrated that mouse satellite DNA was strikingly localized around the centromere (Pardue and Gall 1970). Although satellites propagate and may be mobile, expansion and movement are passive, relying on processes such as replication slippage, unequal crossing-over, or gene conversion to expand and move. This distinguishes satellites from transposable elements that typically insert as monomers and encode genes

M. P. Lauria Sneideman · V. H. Meller (✉)
Department of Biological Sciences, Wayne State University, Detroit, MI, USA
e-mail: Victoria.Meller@wayne.edu

Đ. Ugarković (ed.), *Satellite DNAs in Physiology and Evolution*, Progress in Molecular and Subcellular Biology 60,
https://doi.org/10.1007/978-3-030-74889-0_1

necessary for mobilization. However, this dichotomy is not clean. Some satellite repeats may be derived from transposable elements (Dias et al. 2015; Meštrović et al. 2015). Both transposable elements and satellite repeats are enriched in heterochromatic regions, are subject to silencing by heterochromatin formation, and are often grouped with transposable elements for the purpose of analysis and discussion.

In spite of their abundance, satellite repeats are typically thought of as having few cellular functions besides contributing to the formation of heterochromatin, centromeres, and telomeres. But satellite DNA and RNA participate in a number of diverse processes, including gene regulation, stress response, and nuclear organization in *Drosophila melanogaster* and many other organisms. The mutability of satellites makes them prominent actors in the evolution of genomes. In accord with this, satellite repeats are a potent and adaptable weapon in genomic conflicts between species, and between chromosomes within a species. This review will focus on the functions of satellite repeats in *Drosophila* with particular attention to the properties that make satellites a versatile and powerful force in nuclear organization, gene regulation and evolution.

1.2 Seeing the Dark Matter of the Genome

Much of eukaryotic genomes are comprised of vast, uncharted blocks of heterochromatin surrounding centromeres and telomeres. These regions, made up of satellite repeats and transposable elements, resist cloning and have posed an insurmountable barrier to traditional methods of genome sequencing and assembly. But advances in long-read sequencing of unamplified DNA have allowed the most challenging regions of genomes to be assembled. Nanopore sequencing of high molecular weight DNA was used to complete human centromeres (Miga et al. 2020). PacBio sequencing has similarly enabled *Drosophila* centromeres to be assembled (Chang et al. 2019). These methods avoid bias in library preparation but have high error rates, making assembly of repetitive regions challenging. The performance of correction and assembly methods must consequently be validated before being used to reconstruct repetitive regions (Khost et al. 2017). At present, sequencing and assembly of major repetitive regions remains labor intensive and technically challenging. In contrast, the diversity and abundance of different types satellite repeats in the genome can be determined by sequencing of unamplified genomic DNA (Lower et al. 2018). This approach revealed differences in satellite composition between strains of *Drosophila melanogaster*, supporting the idea that satellites are a rapidly evolving portion of the genome (Wei et al. 2014). Interestingly, the satellite composition of different chromosomes is often distinct. This is observed in humans, where the variants of the α-satellite arrays that make up centromeres are chromosome specific (Rudd et al. 2006). It is also the rule in flies, where distinctive combinations of satellite repeats make up the pericentric heterochromatin of different chromosomes (Lohe et al. 1993; Blattes et al. 2006; Jagannathan et al. 2017; Chang et al. 2019).

Table 1.1 Major *D. melanogaster* satellite repeats

Sequence	%AT	Notable features	Citation
AATAT	100	Binds D1	1,8,12
AATAAAC	86		1,8
AATAG	80		1,8
AATAC	80		1,8
AATAACATAG	80	Prodsat, binds Prod	1,8,9
AATAGAC	71		1,8
AAGAC	60		1,8
AAGAG	60		1,8
TCAT	75		8
AAAAC	80		8
AACAC	60	Binds Lhr	8,11
AACAAAC	71		8
Hsrω	68	Stress induced, sequesters RNA processing factors	13
AAGAGAG	57		1,8
CCCGTACTCGGT	33	Dodeca, 11 and 12 bp variants, binds DP1	2,8,10
359 bp	69	X heterochromatin, binds D1 and topoisomerase 2, produces siRNA and lncRNA	1,3,4,5,8
1.688^X	65–72	359 bp variant in X euchromatin, guides dosage compensation	14
Rsp	71	Expansion on 2R, target of *Segregation distorter*	6,7
260 bp	71	2L heterochromatin	6,8

(1) Lohe et al. (1993), (2) Abad et al. (1992), (3) Dibartolomeis et al. (1992), (4) Waring and Pollack (1987), (5) Kuhn et al. (2012), (6) Khost et al. (2017), (7) Wu et al. (1988), (8) Jagannathan et al. (2017), (9) Török et al. (2000), (10) Huertas et al. (2017), (11) Satyaki et al. (2014), (12) Jagannathan et al. (2019), (13) Jolly and Lakhotia (2006), (14) Menon et al. (2014)

1.3 Biophysical Properties of Satellites

Many satellites have interesting biophysical properties that are often commented on. How these contribute to function remains unclear in most instances. With the exception of *Dodeca*, *D. melanogaster* satellites are notably AT rich (Table 1.1). Indeed, AT richness is common in satellite DNA and may contribute to a curving of the duplex that enhances nucleosome stability (Fitzgerald et al. 1994; reviewed in Palomeque and Lorite 2008). Also suggestive is the observation that monomers of longer and more complex satellites often approximate the length of mono-, di-, or tri-nucleosomes, suggesting the potential for nucleosome phasing (Henikoff et al. 2001). For example, α-satellite repeats of human centromeres (171 bp) and the 359 bp satellite family of *D. melanogaster* suggest mono- and di-nucleosomes, respectively (Table 1.1). Nucleosomes are phased over the centromeric satellites of multiple species (reviewed in Heslop-Harrison and Schwarzacher 2013). The human centromere protein CENP-B enforces phasing by binding a 17 bp sequence in α-satellite repeats (Ando et al. 2002). The *Responder* (*Rsp*) repeats of

D. melanogaster, composed of two similar 120 bp units, also enforce phasing as demonstrated by an extended nucleosome periodicity of 240 bp (Doshi et al. 1991). The *Rsp* locus is known for its role as the target of meiotic drive, but the potential role of nucleosome phasing in this process is unknown. Taken together, these observations suggest that satellites display intrinsic features that are expected to contribute to nucleosome stability and influence the biophysical properties of chromatin.

1.4 How Do Satellites Expand, Move and Change?

D. melanogaster is rich in satellites, having approximately twice the diversity as humans (Shatskikh et al. 2020). The most abundant of these are 12 bp or less. As satellites generally lack coding potential their movement is nonautonomous. In spite of this limitation, they have been extraordinarily successful. Expansion and mobilization of satellite DNA reflects the propensities of replication and repair systems. For example, short tandem repeats are intrinsically unstable as they expand and contract by replication slippage (Fig. 1.1a) (Tautz et al. 1986; Bzymek and Lovett 2001; reviewed in Richards and Sutherland 1994; Levinson and Gutman 1987). Satellites also expand and contract by unequal crossing over during replication or repair (Fig. 1.1b). As longer, more complex monomers pose less of a challenge to the replication machinery, unequal crossing over is presumed to be a major factor variation of long repeats (Cabot et al. 1993; Southern 1975). A relevant question is why the expansion of noncoding sequence is tolerated. Some organisms have a considerably lower accumulation of satellite DNA, suggesting differences in susceptibility to slippage and unequal crossing over or tolerance of repetitive sequence. A comparison of related organisms with dramatic differences in genome size supports the idea that tolerance of additional genetic material is species specific (Petrov et al. 2000; Hartl 2000).

The movement of satellites may also occur by the formation of extrachromosomal loops that occur by recombination within an array. rDNA and noncoding tandem repeats are recovered as extrachromosomal loops in *Drosophila* and mammalian cells (Kiyama et al. 1986, 1987; Pont et al. 1987; Cohen et al. 2003, 2006; reviewed in Cohen and Segal 2009). This suggests a simple mechanism for movement to new sites. Extrachromosomal loops could undergo recombination with similar sequences or insert at random (Fig. 1.1c). The risk of extrachromosomal loops to genome integrity is moderated by the assembly of satellite repeats into heterochromatin. In accord with this idea, the loss of heterochromatin factors elevates the level of extrachromosomal loops and increases genomic damage (Larson et al. 2012; Peng and Karpen 2007). The erosion of heterochromatic silencing that is observed in aging and cancer is presumed to lead to the increase in extrachromosomal loops and

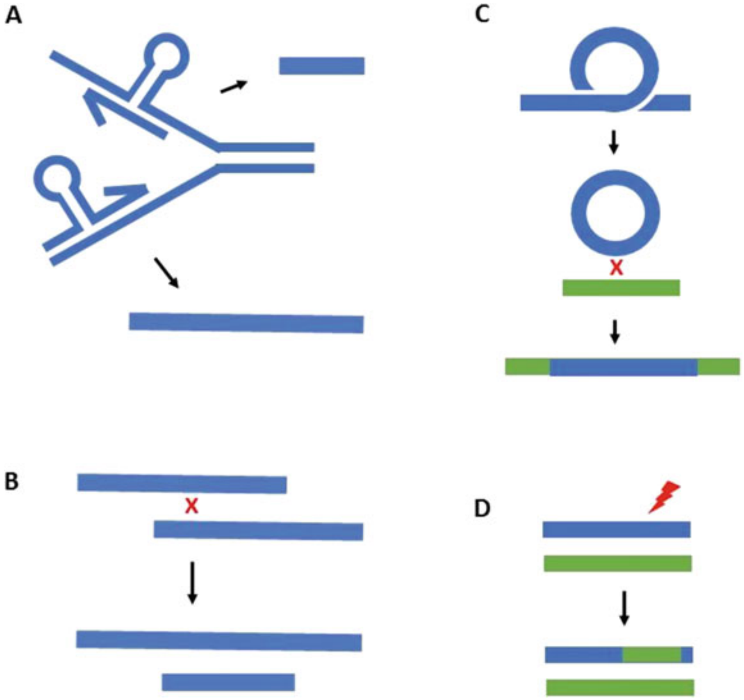


Fig. 1.1 The repetitive structure of satellite repeats facilitates movement and change. (**a**) Replication stalling at repeats allows mispairing and template slippage. This produces contraction (top) or expansion (bottom) of tandem repeats. (**b**) Unequal crossing over leads to the expansion and contraction of tandem arrays. Cycles of unequal crossing over homogenize repeats at the center of an array. (**c**) Extrachromosomal loops generated by recombination within a tandem array can insert at a new site. (**d**) Gene conversion occurs when a related sequence serves as a template for recombination or repair

contribute to genome instability in these cells (Sinclair and Guarente 1997; Larson et al. 2012; Turner et al. 2017; deCarvalho et al. 2018; Kim et al. 2020).

Tandem arrays are also subject to gene conversion by recombination within a cluster or with similar sequences from elsewhere in the genome (Fig. 1.1d). This general process is of interest for its role in the evolutionary divergence of duplicated genes (Osada and Innan 2008). The extreme abundance of satellite repeats that could be used as templates favors this process but raises the potential for large, damaging genome rearrangements. The idea that protection of repetitive DNA from inappropriate recombination is one of the functions of heterochromatin is supported by the behavior of repair foci in heterochromatic regions (Caridi et al. 2018). These foci move out of the nuclear territory occupied by heterochromatin before completion of the repair, supporting the idea that recombination and repair of repetitive DNA are potentially dangerous and under tight control.

1.5 Evolution of Satellite Repeats Is Rapid and Driven

Closely related species often display striking variations in satellite repeat composition and abundance (Bosco et al. 2007; Jagannathan et al. 2017; Lohe and Brutlag 1987). A comparative study of the satellite composition of four closely related species, *D. melanogaster, D. sechellia, D. simulans,* and *D. mauritiana* used hybridization to mitotic chromosomes to compare satellite composition and localization (Jagannathan et al. 2017). Some classes of satellites undergo complete replacement in closely related species. For example, Prodsat (AATAACATAG, Table 1.1) makes up 2% of the *D. melanogaster* genome but is not detected in the other three species (Török et al. 2000; Jagannathan et al. 2017). One caveat of this approach is that sites with low copy numbers of repeats are below the detection limit on mitotic chromosomes. For example, several megabases of 359 bp satellites in pericentromeric heterochromatin on the *D. melanogaster* X are detected by this method, but hundreds of closely related satellites dispersed throughout X euchromatin are not.

Dispersed satellites in euchromatin have also been subject to rapid, widespread changes in sequence, position, and abundance (Sproul et al. 2020). Although striking, this wholesale replacement is the natural outcome of two mutagenic processes with vastly different speeds. Point mutations diversify sequence, and the accumulation of mutations will eventually destroy the identity between sequences derived from the same progenitor. In contrast, gene conversion occurs orders of magnitude much more rapidly and acts to homogenize repeats within an array, and between arrays at different sites in the genome (Ohta and Dover 1984). The outcome of these competing mutational processes is the replacement and homogenization of satellites throughout the genome, termed molecular drive (Dover 1982). This is particularly dramatic when comparing closely related species, such as the Drosophilids (Jagannathan et al. 2017; Sproul et al. 2020; de Lima et al. 2020; Larracuente 2014).

1.6 Satellites, Silencing, and Organization of Chromatin

One of the most prominent features of satellite repeats is their role in heterochromatin formation. Large arrays of tandem repeats trigger silencing through heterochromatin formation that is largely sequence independent (Henikoff 1998). This has bedeviled mouse genetic studies because random transgene insertions produce tandem arrays subject to silencing. Using Cre/LoxP to excise extra copies from a mouse transgene array, Garrick et al. (1998) demonstrated that chromatin compaction and transgene silencing was not an intrinsic feature of the insertion site or transgene sequence, but was instead induced by multicopy arrays. Silencing of tandem transgenes is also observed in flies and plants, indicative of a common strategy for inactivating repetitive DNA (Dorer and Henikoff 1994). As most repetitive sequences are potential threats to genome integrity, the recognition and silencing of repeats represent a triumph of genome defense.

RNA derived from transposable elements and satellites direct the chromatin modifications that initiate heterochromatin formation in fission yeast and this serves as a useful model for the process (reviewed by Grewal and Elgin 2007). Transcription through repeats generates RNAs that are processed into siRNAs and loaded onto Argonaut effector complexes (Höck and Meister 2008). Nascent RNA from cognate regions of the genome is bound by these complexes, which recruit a histone methyltransferase that places the H3K9me mark. The heterochromatin protein Swi6, and a number of small RNA processing factors, are recruited by H3K9me to ensure maintenance of silencing (Zhang et al. 2008). While *Drosophila* heterochromatin is heterogeneous, evidence suggests that small RNA pathways also contribute to chromatin regulation in flies (Swenson et al. 2016; Cernilogar et al. 2011). Mislocalization of heterochromatin proteins and break down of silencing have been observed when Argonaut effectors or the genes necessary to produce small RNAs are inactivated (Fagegaltier et al. 2009). A genetically distinct silencing system in the germ line controls transposable elements by message destruction and transcriptional silencing (Khurana et al. 2010). This is directed by Piwi RNAs (piRNAs), generated from transposon sequences archived in piRNA clusters and expressed in the germ line (Brennecke et al. 2007). The resulting piRNAs enable Piwi, a germ line-specific Argonaut protein, to identify and bind nascent transcripts from mobile elements. Piwi recruits an H3K9 methyltransferase through an adapter protein to establish silencing (Sienski et al. 2015). Components of the Piwi system are also involved in chromatin compaction and silencing at later developmental stages. Maternal depletion of Piwi impairs heterochromatic silencing in the adult, a long-lasting effect that is observed by reduction of Position Effect Variegation (PEV) (Gu and Elgin 2013). PEV occurs when transgenes in repressive environments are silenced in some cells (Elgin and Reuter 2013). The majority of piRNAs have the identity to transposons, consistent with their vital role in the repression of mobile elements (Brennecke et al. 2007). But satellite piRNA are also present and may direct heterochromatin compaction of some repeats in the early embryo. Maternally deposited cues, possibly small RNA, direct the formation of zygotic heterochromatin over a cluster of 359 bp satellites on the X chromosome (Ferree and Barbash 2009; Yuan and O'Farrell 2016). The 359 bp satellites are notable for their role in hybrid incompatibility between closely related species, discussed in a following section.

Heterochromatin itself displays remarkable biophysical properties. Visualization of *D. melanogaster* heterochromatin reveals a subnuclear compartment that is distinct from euchromatin and which may consolidate the major heterochromatic regions of all chromosomes (see Caridi et al. 2018). The discovery that fly and human HP1 phase separate in vitro, and that heterochromatin itself displays the properties of phase separation in cells, suggested a biophysical explanation for how segregation is achieved (Strom et al. 2017; Larson et al. 2017). Phase separation of subcellular bodies occurs by the self-association of disordered proteins (reviewed in Hall et al. 2019). Separation is favored by multivalent interactions, protein crowding, and assembly with a polymer, such as RNA or chromatin. HP1 has disordered domains, interacts with a large number of proteins and also binds RNA (Alekseyenko et al. 2014; Muchardt et al. 2002; Roach et al. 2020). A functional role for RNA in HP1 localization is suggested by the finding that HP1a is released from mouse nuclei by RNase and the association of fly HP1a with chromatin is also RNA dependent (Maison et al. 2002; Piacentini et al. 2003).

Many chromatin proteins in addition to HP1 have RNA-binding domains or interact with RNA-binding proteins, in a manner that suggests a structural role for RNA in chromatin organization. One of these, Decondensation factor 31 (Df31), a small, hydrophobic, and highly disordered RNA binding protein, also boasts a large protein–protein interaction network. The general distribution of Df31 in the nucleus suggests a role in maintaining chromosome territories (Rohrbaugh et al. 2013). In cultured *Drosophila* cells association of Df31 with RNA is necessary for accessible chromatin (Schubert et al. 2012). In vitro assays found that Df31 association with chromatin was RNA dependent and RNase treatment collapsed chromatin into a nuclease-resistant state. While Df31 shows hallmarks of a protein involved in phase separation, it is enriched in euchromatic regions.

Scaffold Attachment Factor A (SAF-A, HNRNPU in humans) and SAF-B have DNA binding domains that recognize AT-rich matrix or scaffold attachment sites (Fackelmayer et al. 1994; Göhring and Fackelmayer 1997; Nozawa et al. 2017; Fan et al. 2018). These similar proteins also have RNA binding domains and large disordered regions. Loss of these proteins disrupts chromatin structure and DNA accessibility, as does RNase digestion (Nickerson et al. 1989; Nozawa et al. 2017; Fan et al. 2018). Mouse SAF-B binds a variety of long noncoding RNAs, but transcripts from pericentric satellite repeats are its predominant partners (Huo et al. 2020). Depletion of mouse SAF-B allowed heterochromatin bodies in the nucleus to expand and make interchromosomal contacts. Imaging reveals that SAF-B coats the exterior of H3K9me3-rich heterochromatin bodies, suggesting a SAF-B shell that prevents inappropriate mingling of phase-separated heterochromatin domains from different chromosomes (Huo et al. 2020). *Drosophila* SAF-B binds chromatin and is also visualized as an extrachromosomal network (Alfonso-Parra and Maggert 2010). The Association of fly SAF-B with chromatin responds to transcription and is differentially affected by mutation of its DNA binding domain and RNase treatment, but whether or not fly SAF-B interacts with specific RNAs is unknown.

Responses to heat shock and stress suggest that satellite RNA is situated in an interconnected web of RNA-binding proteins that organize chromatin and coordinate mRNA processing. When mammalian cells are subjected to stress, transcription of Sat III RNA is dramatically upregulated (Rizzi et al. 2004). SAF-B, and many RNA-binding factors involved in message processing, are recruited to nuclear stress bodies that form at sites of Sat III transcription (Valgardsdottir et al. 2008). Knockdown of Sat III RNA partially reversed the transcriptional repression induced by heat shock, suggesting a mechanism for rapidly restructuring chromatin and RNA processing pathways during stress (Goenka et al. 2016). In flies the *Heat shock RNA omega* (*Hsrω*) RNA serves a similar function. This noncoding transcript orchestrates stress response by sequestering splicing and RNA processing factors (reviewed by Jolly and Lakhotia 2006). *D. melanogaster Hsrω* includes 20 kb of AT-rich, 280 bp tandem repeats, thus conforming to the pattern of AT-rich satellites with a repeat length corresponding to multiples of nucleosome length.

1.7 Tandem Repeats in Euchromatin Modulate Nearby Genes

The role of satellite repeats in nucleating heterochromatin formation is well known, but tandem repeats of all types, including satellites, play interesting and surprising roles in gene regulation in euchromatin. A portion of satellite DNA is distributed throughout the euchromatic genome in tandem arrays. Changes in the number of repeats have created a wealth of genetic variation that has been exploited in forensic analysis, population genetics, and conservation. Although often considered neutral, microsatellites are highly represented in the promoters of human genes (Sawaya et al. 2013; Tomilin 2008). Dinucleotide repeats are enriched in fly enhancers, where they contribute to normal expression levels (Yanez-Cuna et al. 2014). These authors concluded that the association of short tandem repeats with regulatory regions is broadly conserved. Roughly 25% of the promoters in baker's yeast, *Saccharomyces cerevisiae*, also contain tandem repeats (Vinces et al. 2009). These increase gene expression as the length of the repeats expanded. Tandem repeats also mediate repression. In the beetle *Tribolium castaneum* a major satellite DNA family near euchromatic genes maintains repression after heat stress (Feliciello et al. 2015). The mutability of short tandem repeats suggests a potential source of phenotypic variation. In accord with this idea, variation in repetitive DNA has been linked to expression differences in plants and insects (Ranathunge et al. 2018; Brajković et al. 2012). Repeat length variations in developmental genes, coupled with selection by breeders, are responsible for rapid phenotypic evolution in dogs (Fondon and Garner 2004). Social behavior in voles is influenced by satellite polymorphisms in a vasopressin receptor and length variants of repeats in the *period* (*per*) gene of *D. melanogaster* determine the male courtship song rhythm (Yu et al. 1987; Hammock and Young 2005). In addition to providing a source of genetic diversity, satellite repeats have been recruited to wage genomic conflicts and enable sex chromosome dosage compensation, described in the following sections. Their usefulness in these contexts owes to the properties described above: mobility, rapid evolution, and multifaceted roles in the structure and regulation of chromatin.

1.8 Chromosome Identification During Dosage Compensation

Organisms with highly differentiated sex chromosomes, such as humans and *Drosophila*, must address the problem of sex chromosome gene dosage. Males are functionally hemizygous for X-linked genes. Flies meet this challenge by increasing expression from virtually every gene on the single male X chromosome to match that of the two female X chromosomes. The Male-Specific Lethal (MSL) complex, composed of five proteins and one of two redundant RNAs, is essential for this process (reviewed in Kuroda et al. 2016). The MSL complex is selectively recruited to actively expressed X-linked genes (Alekseyenko et al. 2006; Bell et al. 2008; Sural et al. 2008). One of the MSL proteins, Males absent on the first (Mof), is a histone acetyltransferase that deposits the H4K16ac mark within the gene body (Kind et al. 2008; Copur et al. 2018). Histone acetylation increases the likelihood that initiated transcripts will be completed, raising the level of transcripts approximately twofold (Larschan et al. 2011). A long noncoding RNA, *roX1* or *roX2*, must be part of the complex for proper X localization (Meller and Rattner 2002). Severe *roX1 roX2* mutants are male lethal, the expression of X-linked genes is reduced and MSL proteins localize to ectopic autosomal sites (Deng and Meller 2006). How the MSL complex identifies the X chromosome with the required selectivity is still unknown. Studies in a number of laboratories characterized Chromatin Entry Sites (CES) on the X chromosome that bind an adapter protein and recruit the MSL complex directly (Alekseyenko et al. 2008; Straub et al. 2008; Soruco et al. 2013). However, the adapter protein binds related sites on all chromosome arms but only recruits the MSL complex in the context of X-linked sites. This suggests the presence of additional X identity elements.

The striking enrichment of a clade of 359 bp repeats, termed the 1.688^{X} repeats (Table 1.1) in X euchromatin pointed to a potential role in an X chromosome-specific process such as dosage compensation (Hsieh and Brutlag 1979; Waring and Pollack 1987; Dibartolomeis et al. 1992). The 1.688^{X} repeats are enriched near genes, including promoters and introns, leading to the suggestion that they could modulate expression (Kuhn et al. 2012). Autosomal insertions of short clusters of these repeats induced recruitment of the MSL complex and partial compensation of genes as much as 140 kb away (Joshi and Meller 2017; Deshpande and Meller 2018). A clue to how these repeats function came from the discovery that mutations in the siRNA pathway enhanced the lethality of males with partial loss of function *roX1* and *roX2* chromosomes (Menon and Meller 2012). Furthermore, ectopic expression of siRNA from one 1.688^{X} repeat partially restored MSL localization and rescued *roX1 roX2* males (Menon et al. 2014). Taken together, these studies reveal that the 1.688^{X} satellite repeats are X identify elements and suggest that the siRNA pathway mediates their function. Interestingly, other Drosophilid X chromosomes are highly enriched for chromosome-specific repeats, although the sequence of these repeats is not highly conserved (Gallach 2014). Particularly striking is the rapid acquisition of repeats by neo-X chromosomes that arise by the fusion of an autosome to a sex chromosome. These fusions also produce a neo-Y, fated to degenerate as it passes exclusively through males without recombination (reviewed by Wei and Barbash 2015). Degeneration of the neo-Y necessitates compensation of genes on the neo-X. The relative mobility of satellite repeats makes them well suited for marking a young X chromosome to enable it to capture the dosage compensation machinery.

The manner in which the 1.688^{X} satellites identify the X chromosome remains unknown, but there are clues that the mechanism is very different than that of the CES. The 1.688^{X} satellites on the X chromosome are not generally enriched for the MSL proteins, suggesting that they do not recruit directly but mark the X chromosome in some fashion (Deshpande and Meller 2018). Ectopic expression of 1.688^{X} siRNA increased the repressive H3K9me2 mark on autosomal 1.688^{X} satellite insertions (Deshpande and Meller 2018). Contrary to conventional expectations for a repressive mark, this increased the expression of nearby genes in males. The finding that HP1 is modestly enriched on the male X chromosome, and that mutations in several heterochromatin factors selectively disrupt the structure of the polytenized male X, support the idea that repressive marks are in some way linked to fly dosage compensation (Spierer et al. 2005, 2008; De Wit et al. 2005). How satellite repeats and repressive marks might accomplish this is speculative, but one possibility is through influencing chromatin organization in the nucleus. Nuclear organization is a factor in X chromosome compensation in other organisms. For example, the single male X chromosome of *C. elegans* is located at the periphery of the nucleus, but the two female X chromosomes are centrally located (Sharma et al. 2014). Interaction with nuclear pore proteins may elevate the expression of X-linked genes in this species. A role for nuclear pore proteins in MSL loading and activation has also been proposed in flies (Mendjan et al. 2006). More generally, the ability of X-linked genes to acquire dosage compensation during development is attributed to the three-dimensional organization of X chromosomes in the nucleus of mammals and flies (Engreitz et al. 2013; Schauer et al. 2017; Ramírez et al. 2015). A role of the location or organization of the X chromosome is one way that 1.688^X satellites might promote X recognition.

1.9 Satellites and Centromeres

The most widely appreciated function of satellite DNA is at centromeres. Human centromeres contain α-satellite arrays harboring a motif that interacts with the centromeric H3 variant CENP-A, suggesting determination by sequence (Masumoto et al. 1989; reviewed by McNulty et al. 2017; Willard 1985; Schueler et al. 2001). However, human centromeres occupy only part of the array and satellites are absent from some neo-centromeres, challenging the idea that sequence is the primary centromere determinant. Fly centromeres also form within extensive arrays of satellite repeats, but the centromere itself assembles at “islands” of transposons embedded in this sea of satellites (Chang et al. 2019). Fly centromeres are defined by the incorporation of an H3 variant called Cid. The importance of epigenetic information in specifying centromeres in flies is demonstrated by the fact that the transposons at the fly centromere are by no means limited to the centromere, appearing in both heterochromatic and euchromatic contexts throughout the genome (Chang et al. 2019). It is also reflected in the persistence of a functional centromere following transient anchoring of the centromere-specific chaperone CAL-1, which is capable of loading Cid at ectopic sites (Chen et al. 2014). These findings indicate that both fly and humans centromeres are specified by a combination of DNA sequence, genomic context, and epigenetic marking. Centromeres are surrounded by heterochromatin that contributes to their function. A large deletion of heterochromatin flanking a fly centromere produced mitotic instability and premature sister chromatid separation (Wines and Henikoff 1992). This is consistent with the enrichment of cohesin in heterochromatin and suggests that the mitotic machinery is tuned to a certain arrangement of heterochromatin surrounding the centromere (Bernard and Allshire 2002).

RNA from satellites has also been found to localize to centromeres. Transcripts from a large block of pericentric 359 bp satellites on the *D. melanogaster* X chromosome bind *in cis* to the centromeric region (Bobkov et al. 2018). RNA from the mammalian α-satellite also binds to centromeric proteins and localizes *in cis* at centromeres (Wong et al. 2007; reviewed in Ideue and Tani 2020; McNulty et al. 2017). This RNA is also necessary for characteristic localization of centromeric proteins CENPC1 and INCENP, and so may function to recruit or stabilize components of the centromere.

1.10 Satellites Are the Ammunition of Genomic Conflicts

Chromosomes are fundamental units of inheritance and take an active role in biasing their own transmission to the next generation. Meiotic drive, when a genetic element manipulates reproduction to favor its own transmission and overthrow Mendel's rules, is the outcome. A selfish chromosome able to accomplish this will increase in the population. Evolutionary theory posits that systems of meiotic drive emerge frequently and sweep through the population, driving enrichment of one chromosome and limiting genetic variation. In addition, unfavorable, genetically linked alleles are allowed to proliferate (Courret et al. 2019). This extracts a cost in fitness that enables suppressors of drive to emerge and restore Mendelian segregation. A history of recurring cycles of drive and suppression is revealed when wild-caught flies are outcrossed and suppressed drivers emerge (Hartl and Hartung 1975). These conflicts are often mediated through satellite repeats and the outcome shapes the genome.

Sex differences in meiosis ensure that a strategy for biasing chromosome transmission can only function in one sex. While all products of male meiosis have the potential to develop into sperm, only one of the four products of female meiosis will become the egg. To gain an advantage in female meiosis the critical point is the alignment of homologs on the spindle at the first division (Fig. 1.2a). A centromere that attaches to the egg pole will escape elimination in the polar body (Rosin and Mellone 2017; Kursel and Malik 2018). To take advantage of this requires asymmetry in the meiotic spindle and a centromere able to exploit the asymmetry. The extraordinary reproductive advantage that a stronger centromere holds is thought to fuel an evolutionary race that drives rapid changes in centromeric DNA and proteins (Malik 2009). Predictions of this model are fulfilled in mice where an expansion of satellite repeats has produced a large centromere with an advantage over a homolog with fewer satellites (Iwata-Otsubo et al. 2017). Interestingly, the kinetochores of the larger centromere detach more frequently from cortical spindle fibers, providing an opportunity to reattach to the egg pole (Akera et al. 2019). Suppressors of centromere drive would benefit a population in which a chromosome had begun to cheat. The observation that the centromeric variants Cid and CEN-A are remarkably fast-evolving and divergent from other histones suggests their involvement (Black et al.

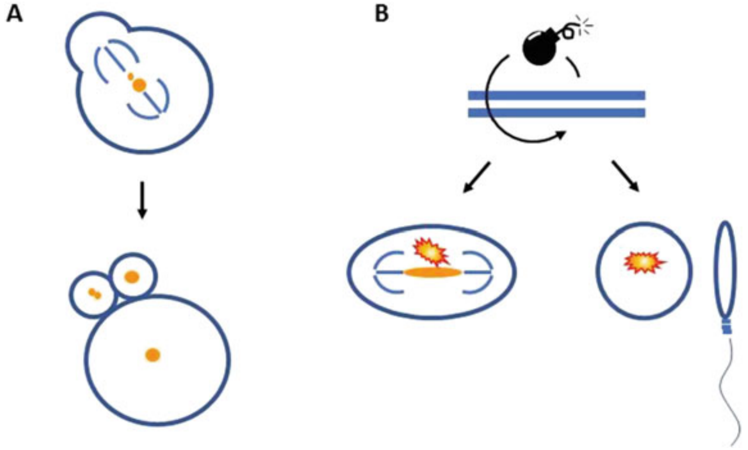


Fig. 1.2 Meiotic drive in female and male germlines. (**a**) Stronger centromeres (larger dot) gain an advantage in the female germline by avoiding the cortical spindle and becoming an egg nucleus more than 50% of the time. (**b**) Chromosomes that achieve drive in males do so by sabotaging their homolog. This may be direct, as implied in the cartoon, or indirect by the establishment of an environment that is toxic to cells inheriting the susceptible homolog. Left) The Paris sex ratio X chromosome produces a factor that blocks segregation of the Y at the second meiotic division. Failure to form Y-bearing sperm ensures predominantly female broods. Right) The *Sd* and Winters drivers sabotage the maturation of sperm carrying susceptible homologs. In both systems, arrest occurs before chromatin compaction and these malformed cells are eliminated (center). Sperm carrying the driver (right) develop normally

2004). Amino acid changes in Cid are concentrated in a region that interacts with H4 and an extended loop that contacts DNA (Vermaak et al. 2002). The rapid evolution of centromeres and the proteins that bind them seems at odds with the very conservative function of centromeres but is in accord with the idea that these structures are the site of an evolutionary battle (Malik 2009).

All products of male meiosis have the opportunity to become sperm. To gain an advantage in the male germ line a chromosome must exert a negative effect on its homolog (Fig. 1.2b). Several examples of this are well known, including *Segregation distorter/Responder* (*Sd/Rsp*) in *D. melanogaster*. *Rsp*, the target of drive, is an array of two very similar 120 bp repeats present in dozens to thousands of copies (Khost et al. 2017). Larger arrays confer increased sensitivity to *Sd* (Larracuente and Presgraves 2012; Moschetti et al. 1996). *Sd* is a truncated but enzymatically active duplication of RanGAP that mislocalizes to the interior of the nucleus (Kusano et al. 2001). When a male has *Sd* on one homolog and a sensitive *Rsp* allele (*Rsp*^{S}) on the other, maturation of sperm carrying *Rsp*^{S} is arrested and these cells are eliminated. The sperm carrying *Sd* develop normally and are responsible for most or all fertilizations. The *Sd* phenotype is enhanced by a number of modifiers, all genetically linked to *Sd* on the second chromosome (reviewed by Larracuente and Presgraves 2012). Of course, *Sd* chromosomes must themselves carry insensitive *Rsp* arrays in order to escape elimination. Although the precise molecular defect that causes arrest is unclear, abnormal localization of mutant RanGAP is thought to disrupt the RanGTP/GDP gradient across the nuclear envelope and this may interfere with transport in and out of the nucleus (Kusano et al. 2003). In this environment, the expanded repeats on *Rsp*^{S} chromosome precipitate failure of sperm maturation. The process that is affected must be unique to sperm development as *Sd/Rsp*^{S} females produce normal offspring ratios. During sperm maturation chromatin is remodeled by the replacement of histones, a process requiring the import of protamine. This step is male-limited and appropriate to the stage of arrest, but it is unclear how expanded *Rsp*^{S} arrays and defects in protamine levels would induce arrest. Disruption of small RNA import leading to a defect in repackaging *Rsp* chromatin is also possible. Small RNAs from *Rsp* have been identified in the germ line and mutations in *aubegine* (*aub*), an argonaut family protein that participates in piRNA production, enhances distortion by *Sd* (Gell and Reenan 2013; Nagao et al. 2010). This finding suggests that an additional role of germ line small RNA systems is to defend against meiotic drive.

Rsp provides an excellent example of satellite turnover. Two families of repeats, the *Rsp* and *Rsp-like* family and the 359 bp family, which includes an extensive array of pericentromeric 359 bp repeats in X heterochromatin and the euchromatic 1.688^{X} satellites, were found to occupy overlapping sites in related species (Sproul et al. 2020). Both the *Rsp* and 359 families are AT-rich, but the 359 bp repeats are widespread, older, and more diversified in related species. Examination of satellite repeats in several *Drosophila* species revealed that 1.688^{X} and *Rsp-like* satellites occupy many of the same euchromatic sites (Sproul et al. 2020). Sites in which *Rsp-like* repeats have been inserted in an existing 1.688^{X} array, possibly in the process of replacing it, were identified in *D. simulans* and *D. mauritiana*. This suggests a model in which young *Rsp-like* repeats use homology with existing 1.688^{X} repeats to enter these sites, a process that may be facilitated by long-range interactions in the nucleus. Extrachromosomal circular DNAs are also a potential mechanism for movement. The correlation between the abundance of one of the repetitive elements and extrachromosomal DNA also suggests a role in *Rsp-like* invasion (Sproul et al. 2020).

D. simulans has at least three meiotic drive systems that bias sex chromosome inheritance and thus distort the sex ratio. All of these involve drivers on the X chromosome. In the Winters system, named for the location where the flies were collected, the X chromosome distorter prevents Y-bearing gametes from completing maturation. Failure occurs during condensation of the haploid nucleus, a timing that is similar to that observed in the *D. melanogaster Sd/Rsp* system (Tao et al. 2007b). The driver, *Distorter on the X chromosome* (*Dox*), is a partial duplication that produces an RNA with limited coding potential (Tao et al. 2007a). The mechanism of *Dox* action is unknown but suppressors of *Dox* on the second chromosome generate siRNAs that reduce levels of the *Dox* transcript (Lin et al. 2018). As the Y chromosome is primarily composed of satellite repeats and transposons, it is likely that the toxic effect of *Dox* depends on the unique sequence and chromatin composition of this chromosome. A second *D. simulans* sex ratio distortion system, Paris, induces anaphase bridges and failure of Y chromosome disjunction during the second meiotic division (Fig. 1.2b, Cazemajor et al. 2000). One component of the X-linked driver was discovered to be a loss of function mutation in a rapidly evolving member of the *HP1* family, *HP1D2* (Helleu et al. 2016). Intriguingly, the HP1D2 protein is specifically enriched on the Y chromosome, suggesting that a defect in the organization or compaction of this chromosome prevents segregation. Sex ratio distortion leads to populations with unbalanced ratios of males and females and creates a strong selective advantage for an individual with a novel suppressor of drive. In accord with this, Y chromosomes that are resistant to the Paris or Winters driver have been discovered (Branco et al. 2013; Helleu et al. 2019). As the coding potential on the Y is limited, it is quite possible that these suppressors are changes in satellite or transposon content that make them insensitive to the X-linked driver.

1.11 Satellite Repeats Mediate Conflict Between Species

Hybrid incompatibilities enforce the reproductive isolation that defines species (Castillo and Barbash 2017). The rapid evolution of heterochromatin DNA and proteins is a potential source of incompatibilities that produce lethality or infertility upon hybridization of closely related Drosophilids (Presgraves 2010; Ferree and Barbash 2009; Gatti et al. 1976; Yunis and Yasmineh 1971; reviewed in Ferree and Prasad 2012). For example, when *D. melanogaster* males are mated to *D. simulans* females, male offspring emerge as sterile adults but females die as embryos. Early female lethality is attributable to the *D. melanogaster* X chromosome in *D. simulans* cytoplasm. Specifically, the large array of 359 bp repeats at the base of the *D. melanogaster* X chromosome fails to compact and X chromatids become entangled in anaphase bridges (Ferree and Barbash 2009). But when *D. melanogaster* males transmitted a *Zygotic hybrid rescue* (*Zhr*) chromosome that was deleted for the pericentric 359 bp repeats, female offspring survived (Sawamura et al. 1993). Small RNAs from the 359 bp satellite are present in oocytes from *D. melanogaster* females, and it is plausible that these direct heterochromatin formation over the 359 bp satellites in fertilized embryos (Ferree and Barbash 2009). The X chromosome of *D. simulans* lacks 359 bp repeats and the relevant class of siRNA is not present in *D. simulans* eggs. These ideas are supported by a study demonstrating that heterochromatin formation at the 359 bp satellites occurred with different timing than that of another large satellite array and required maternal factors missing from *D. simulans* ooplasm (Yuan and O'Farrell 2016).

The reciprocal mating, *D. melanogaster* females mated to *D. simulans* males, produced sterile female adults but no male offspring. The toxic interaction producing male lethality can be traced to heterochromatin proteins, *Lethal hybrid rescue (Lhr, D. simulans)* and *Hybrid male rescue* (*Hmr*, *D. melanogaster*) (Maheshwari and Barbash 2012). Higher expression of *Lhr* from the *D. simulans* chromosome is the basis of hybrid lethality. Loss of *D. simulans Lhr* rescues hybrid lethality, but loss of *D. melanogaster Lhr*, which is expressed at only half the rate as *D. simulans*, does not achieve rescue. *Lhr* encodes a rapidly evolving heterochromatin protein that interacts with HP1 (Brideau et al. 2006; Brideau and Barbash 2011; Thomae et al. 2013). Loss of *Lhr* results in bloated polytene chromosomes in *D. simulans*, a phenotype associated with a loss of chromosome structure (Pal Bhadra et al. 2006). The *Hmr* gene is also rapidly evolving and encodes a DNA binding protein that localizes to heterochromatin in a complex with Lhr and HP1a (Satyaki et al. 2014; Alekseyenko et al. 2014). Hybrid males, which die as larvae, display poorly condensed chromosomes and anaphase bridges between sister chromatids, consistent with the idea that heterochromatin assembly and compaction is the primary defect (Blum et al. 2017).

The conflicts between genetic elements within a species that produce meiotic drive, and between species that lead to hybrid incompatibility, rely on an overlapping cast of characters. In accord with this, it has been suggested that meiotic drive contributes to the genetic divergence that produces hybrid incompatibility (McDermott and Noor 2010). This notion is supported by the discovery that a single gene, *Overdrive* (*Ovd*), appears responsible for both meiotic drive and hybrid incompatibility in *D. pseudoobscura* (Phadnis and Orr 2009). Males from a mating between subspecies are sterile when young but become weakly fertile and produce almost exclusively daughters when aged. Although the molecular mechanisms at play are currently unknown, the finding that one gene is involved in both phenomena supports the idea that meiotic drive and hybrid incompatibility are produced by similar genetic conflicts.

Summary

Satellite repeats appeared both troublesome and singularly unpromising at the dawn of the genomics era. The typical concentration of satellites in vast, unclonable blocks of heterochromatin was an additional deterrent to their study. But the ability of satellites to move, expand, and undergo relatively rapid genome-wide replacement enables them to shape genomes and respond to evolutionary pressures. Satellite DNA, and small RNA pathways capable of directing modifications to chromatin, are a powerful combination that can be adapted to novel roles. This can be appreciated by the dispersed, euchromatic 1.688^X satellites that recruit dosage compensation while very similar heterochromatic 359 bp satellites mediate hybrid incompatibility at a different life stage and in different sex. In spite of the stark differences in the roles of these satellites, it is likely that small RNA normally directs chromatin modifications to both, and that these modifications are essential for normal function. When heterochromatin is compromised satellites become unstable and devastating disruptions of nuclear organization result. This intrinsic risk can be appreciated by the destruction unleased by 359 bp satellites in a hybrid environment. The remarkable ability of heterochromatin to assemble satellite DNA into a nondestructive form can be credited with enabling satellite repeats to expand to their current position of prominence in higher eukaryotes. All of the properties described above, including the mutability of satellites and their intrinsic danger, put repetitive sequences at the leading edge of evolution.

References

Abad JP, Carmena M, Baars S, Saunders RD, Glover DM, Ludeña P, Sentis C, Tyler-Smith C, Villasante A (1992) Dodeca satellite: a conserved G+C-rich satellite from the centromeric heterochromatin of *Drosophila melanogaster*. PNAS 89:4663–4667. https://doi.org/10.1073/pnas.89.10.4663

Akera T, Trimm E, Lampson MA (2019) Molecular strategies of meiotic cheating by selfish centromeres. Cell 178(5):1132–1144.e10. https://doi.org/10.1016/j.cell.2019.07.001

Alekseyenko AA, Larschan E, Lai WR, Park PJ, Kuroda MI (2006) High-resolution ChIP-chip analysis reveals that the Drosophila MSL complex selectively identifies active genes on the male X chromosome. Genes Dev 20(7):848–857. https://doi.org/10.1101/gad.1400206

Alekseyenko AA, Peng S, Larschan E, Gorchakov AA, Lee OK, Kharchenko P, McGrath SD, Wang CI, Mardis ER, Park PJ, Kuroda MI (2008) A sequence Motif within chromatin entry sites directs MSL establishment on the drosophila X chromosome. Cell 134(4):599–609. https://doi.org/10.1016/j.cell.2008.06.033

Alekseyenko AA, Gorchakov AA, Zee BM, Fuchs SM, Kharchenko PV, Kuroda MI (2014) Heterochromatin-associated interactions of Drosophila HP1a with dADD1, HIPP1, and repetitive RNAs. Genes Dev 28(13):1445–1460. https://doi.org/10.1101/gad.241950.114

Alfonso-Parra C, Maggert KA (2010) Drosophila SAF-B links the nuclear matrix, chromosomes, and transcriptional activity. PLoS One 5(4). https://doi.org/10.1371/journal.pone.0010248

Ando S, Yang H, Nozaki N, Okazaki T, Yoda K (2002) CENP-A, -B, and -C chromatin complex that contains the I-type α-satellite array constitutes the prekinetochore in HeLa cells. Mol Cell Biol 22(7):2229–2241. https://doi.org/10.1128/mcb.22.7.2229-2241.2002

Bell O, Conrad T, Kind J, Wirbelauer C, Akhtar A, Schübeler D (2008) Transcription-coupled methylation of histone H3 at Lysine 36 regulates dosage compensation by enhancing recruitment of the MSL complex in *Drosophila melanogaster*. Mol Cell Biol 28(10):3401–3409. https://doi.org/10.1128/mcb.00006-08

Bernard P, Allshire RC (2002) Centromeres become unstuck without heterochromatin. Trends Cell Biol 12(9):419–424. https://doi.org/10.1016/S0962-8924(02)02344-9

Black BE, Foliz DR, Chakravarthy S, Luger K, Woods VL, Cleveland DW (2004) Structural determinants for generating centromeric chromatin. Nature 430(6999):578–582. https://doi.org/10.1038/nature02766

Blattes R, Monod C, Susbielle G, Cuvier O, Wu JH, Hsieh TS, Laemmli UK, Käs E (2006) Displacement of D1, HP1 and topoisomerase II from satellite heterochromatin by a specific polyamide. EMBO J 25(11):2397–2408. https://doi.org/10.1038/sj.emboj.7601125

Blum JA, Bonaccorsi S, Marzullo M, Palumbo V, Yamashita YM, Barbash DA, Gatti M (2017) The hybrid incompatibility genes Lhr and Hmr are required for sister chromatid detachment during anaphase but not for centromere function. Genetics 207(4):1457–1472. https://doi.org/10.1534/genetics.117.300390

Bobkov GOM, Gilbert N, Heun P (2018) Centromere transcription allows CENP-A to transit from chromatin association to stable incorporation. J Cell Biol 217:1957–1972

Bosco G, Campbell P, Leiva-Neto JT, Markow TA (2007) Analysis of drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177(3):1277–1290. https://doi.org/10.1534/genetics.107.075069

Brajković J, Feliciello I, Bruvo-Madarić B, Ugarković D (2012) Satellite DNA-like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. G3 Genes Genomes Genetics 2(8):931–941. https://doi.org/10.1534/g3.112.003467

Branco AT, Tao Y, Hartl DL, Lemos B (2013) Natural variation of the Y chromosome suppresses sex ratio distortion and modulates testis-specific gene expression in *Drosophila simulans*. Heredity 111(1):8–15. https://doi.org/10.1038/hdy.2013.5

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128 (6):1089–1103. https://doi.org/10.1016/j.cell.2007.01.043

Brideau NJ, Barbash DA (2011) Functional conservation of the Drosophila hybrid incompatibility gene Lhr. BMC Evol Biol 11(1):57. https://doi.org/10.1186/1471-2148-11-57

Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA (2006) Two Dobzhansky-Muller genes interact to cause hybrid lethality in drosophila. Science 314(5803):1292–1295. https://doi.org/10.1126/science.1133953

Bzymek M, Lovett ST (2001) Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. Proc Natl Acad Sci USA 98(15):8319–8325. https://doi.org/10.1073/pnas.111008398

Cabot EL, Doshi P, Wu ML, Wu CI (1993) Population genetics of tandem repeats in centromeric heterochromatin: unequal crossing over and chromosomal divergence at the responder locus of *Drosophila melanogaster*. Genetics 135(2):477–487

Caridi CP, D'agostino C, Ryu T, Zapotoczny G, Delabaere L, Li X, Khodaverdian VY, Amaral N, Lin E, Rau AR, Chiolo I (2018) Nuclear F-actin and myosins drive relocalization of heterochromatic breaks. Nature 559(7712):54–60. https://doi.org/10.1038/s41586-018-0242-8

Castillo DM, Barbash DA (2017) Moving speciation genetics forward: modern techniques build on foundational studies in drosophila. Genetics 207(3):825–842. https://doi.org/10.1534/genetics.116.187120

Cazemajor M, Joly D, Montchamp-Moreau C (2000) Sex-ratio meiotic drive in *Drosophila simulans* is related to equational nondisjunction of the Y chromosome. Genetics 154 (1):229–236

Cernilogar FM, Onorati MC, Kothe GO, Burroughs AM, Parsi KM, Breiling A, Sardo FL, Saxena A, Miyoshi K, Siomi H, Siomi MC, Carninci P, Gilmour DS, Corona DFV, Orlando V (2011) Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. Nature 480(7377):391–395. https://doi.org/10.1038/nature10492

Chang CH, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, Chen CC, Erceg J, Beliveau BJ, Wu CT, Larracuente AM, Mellone BG (2019) Islands of retroelements are major components of Drosophila centromeres. PLoS Biol 17(5). https://doi.org/10.1371/journal.pbio.3000241

Chen CC, Dechassa ML, Bettini E, Ledoux MB, Belisario C, Heun P, Luger K, Mellone BG (2014) CAL1 is the Drosophila CENP-A assembly factor. J Cell Biol 204(3):313–329. https://doi.org/10.1083/jcb.201305036

Cohen S, Segal D (2009) Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. Cytogenet Genome Res 124(3–4):327–338. https://doi.org/10.1159/000218136

Cohen S, Yacobi K, Segal D (2003) Extrachromosomal circular DNA of tandemly repeated genomic sequences in drosophila. Genome Res 13(6):1133–1145. https://doi.org/10.1101/gr.907603

Cohen Z, Bacharach E, Lavi S (2006) Mouse major satellite DNA is prone to eccDNA formation via DNA Ligase IV-dependent pathway. Oncogene 25(33):4515–4524. https://doi.org/10.1038/sj.onc.1209485

Copur Ö, Gorchakov A, Finkl K, Kuroda MI, Müller J (2018) Sex-specific phenotypes of histone H4 point mutants establish dosage compensation as the critical function of H4K16 acetylation in Drosophila. Proc Natl Acad Sci USA 115(52):13336–13341. https://doi.org/10.1073/pnas.1817274115

Courret C, Chang CH, Wei KHC, Montchamp-Moreau C, Larracuente AM (2019) Meiotic drive mechanisms: lessons from Drosophila. Proc R Soc B Biol Sci 286(1913). https://doi.org/10.1098/rspb.2019.1430

de Lima LG, Hanlon SL, Gerton JL (2020) Origins and evolutionary patterns of the 1.688 satellite DNA family in drosophila phylogeny. G3 Genes Genomes Genetics 10(11):4129–4146. https://doi.org/10.1534/g3.120.401727

De Wit E, Greil F, Van Steensel B (2005) Genome-wide HP1 binding in Drosophila: developmental plasticity and genomic targeting signals. Genome Res 15(9):1265–1273. https://doi.org/10.1101/gr.3198905 deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D, Koeman J, Seth S, Protopopov A, Felicella M, Zheng S, Multani A, Jiang Y, Zhang J, Nam D-H, Petricoin EF, Chin L, Mikkelsen T, Verhaak RGW (2018) Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat Genet 50(5):708–717. https://doi.org/10.1038/s41588-018-0105-0

Deng X, Meller VH (2006) roX RNAs are required for increased expression of X-linked genes in *Drosophila melanogaster* males. Genetics 174(4):1859–1866. https://doi.org/10.1534/genetics.106.064568

Deshpande N, Meller VH (2018) Chromatin that guides dosage compensation is modulated by the siRNA pathway in *Drosophila melanogaster*. Genetics 209(4):1085–1097. https://doi.org/10.1534/genetics.118.301173

Dias GB, Heringer P, Svartman M, Kuhn GCS (2015) Helitrons shaping the genomic architecture of Drosophila: enrichment of DINE-TR1 in α- and β-heterochromatin, satellite DNA emergence, and piRNA expression. Chromosom Res 23(3):597–613. https://doi.org/10.1007/s10577-015-9480-x

Dibartolomeis SM, Tartof KD, Rob Jackson F (1992) A superfamily of Drosophila satellite related (SR) DNA repeats restricted to the X chromosome euchromatin. Nucleic Acids Res 20 (5):1113–1116. https://doi.org/10.1093/nar/20.5.1113

Dorer DR, Henikoff S (1994) Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. Cell 77(7):993–1002. https://doi.org/10.1016/0092-8674(94)90439-1

Doshi P, Kaushal S, Benyajati C, Wu C-I (1991) Molecular analysis of the responder satellite DNA in *Drosophila melanogaster*: DNA bending, nucleosome stucture, and Rsp-binding proteins. Mol Biol Evol 8(5):721–741

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Elgin SCR, Reuter G (2013) Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harb Perspect Biol 5(8):1–26. https://doi.org/10.1101/cshperspect.a017780

Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES, Plath K, Guttman M (2013) The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341(6147):1–13. https://doi.org/10.1126/science.1237973

Fackelmayer FO, Dahm K, Renz A, Ramsperger U, Richter A (1994) Nucleic-acid-binding properties of hnRNP-U/SAF-A, a nuclear-matrix protein which binds DNA and RNA in vivo and in vitro. Eur J Biochem 221(2):749–757. https://doi.org/10.1111/j.1432-1033.1994.tb18788.x

Fagegaltier D, Bougé AL, Berry B, Poisot É, Sismeiro O, Coppée JY, Théodore L, Voinnet O, Antoniewski C (2009) The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. Proc Natl Acad Sci USA 106(50):21258–21263. https://doi.org/10.1073/pnas.0809208105

Fan H, Lv P, Huo X, Wu J, Wang Q, Cheng L, Liu Y, Tang Q-Q, Zhang L, Zhang F, Zheng X, Wu H, Wen B (2018) The nuclear matrix protein HNRNPU maintains 3D genome architecture globally in mouse hepatocytes. Genome Res 28(2):192–202. https://doi.org/10.1101/gr.224576.117

Feliciello I, Akrap I, Ugarković Ð (2015) Satellite DNA modulates gene expression in the Beetle *Tribolium castaneum* after heat stress. PLoS Genet 11(8):1–18. https://doi.org/10.1371/journal.pgen.1005466

Ferree PM, Barbash DA (2009) Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. PLoS Biol 7(10). https://doi.org/10.1371/journal.pbio.1000234

Ferree PM, Prasad S (2012) How can satellite DNA divergence cause reproductive isolation? Let us count the chromosomal ways. Genetics Res Int 2012:1–11. https://doi.org/10.1155/2012/430136

Fitzgerald DJ, Dryden GL, Bronson EC, Williams JS, Anderson JN (1994) Conserved patterns of bending in satellite and nucleosome positioning DNA. J Biol Chem 269(33):21303–21314. https://doi.org/10.1016/S0021-9258(17)31963-4

Fondon JW, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci USA 101(52):18058–18063. https://doi.org/10.1073/pnas.0408118101

Gallach M (2014) Recurrent turnover of chromosome-specific satellites in Drosophila. Genome Biol Evol 6(6):1279–1286. https://doi.org/10.1093/gbe/evu104

Garrick D, Fiering S, Martin DIK, Whitelaw E (1998) Repeat-induced gene silencing in mammals. Nat Genet 18(1):56–59. https://doi.org/10.1038/ng0198-56

Garrido-Ramos MA (2017) Satellite DNA: an evolving topic. Genes 8(9). https://doi.org/10.3390/genes8090230

Gatti M, Pimpinelli S, Santini G (1976) Characterization of Drosophila heterochromatin. Chromosoma 57(4):351–375. https://doi.org/10.1007/BF00332160

Gell SL, Reenan RA (2013) Mutations to the piRNA pathway component aubergine enhance meiotic drive of segregation distorter in *Drosophila melanogaster*. Genetics 193(3):771–784. https://doi.org/10.1534/genetics.112.147561

Goenka A, Sengupta S, Pandey R, Parihar R, Mohanta GC, Mukerji M, Ganesh S (2016) Human satellite-III non-coding RNAs modulate heat-shockinduced transcriptional repression. J Cell Sci 129(19):3541–3552. https://doi.org/10.1242/jcs.189803

Göhring F, Fackelmayer FO (1997) The Scaffold/Matrix attachment region binding protein hnRNP-U (SAF-A) is directly bound to chromosomal DNA in vivo: a chemical cross-linking study. Biochemistry 36(27):8276–8283. https://doi.org/10.1021/bi970480f

Grewal SIS, Elgin SCR (2007) Transcription and RNA interference in the formation of heterochromatin. Nature 447(7143):399–406. https://doi.org/10.1038/nature05914

Gu T, Elgin SCR (2013) Maternal depletion of Piwi, a component of the RNAi system, impacts heterochromatin formation in drosophila. PLoS Genet 9(9). https://doi.org/10.1371/journal.pgen.1003780

Hall AC, Ostrowski LA, Mekhail K (2019) Phase separation as a melting pot for DNA repeats. Trends Genet 35(8):589–600. https://doi.org/10.1016/j.tig.2019.05.001

Hammock EAD, Young LJ (2005) Genetics: microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308(5728):1630–1634. https://doi.org/10.1126/science.1111427

Hartl DL (2000) Molecular melodies in high and low C. Nat Rev Genet 1(2):145–149. https://doi.org/10.1038/35038580

Hartl DL, Hartung N (1975) High frequency of one element of segregation distorter in natural populations of *Drosophila melanogaster*. Evolution 29(3):512–518

Helleu Q, Gérard PR, Dubruille R, Ogereau D, Prud'homme B, Loppin B, Montchamp-Moreau C (2016) Rapid evolution of a Y-chromosome heterochromatin protein underlies sex chromosome meiotic drive. Proc Natl Acad Sci USA 113(15):4110–4115. https://doi.org/10.1073/pnas.1519332113

Helleu Q, Courret C, Ogereau D, Burnham KL, Chaminade N, Chakir M, Aulard S, Montchamp-Moreau C (2019) Sex-ratio meiotic drive shapes the evolution of the y chromosome in *Drosophila simulans*. Mol Biol Evol 36(12):2668–2681. https://doi.org/10.1093/molbev/msz160

Henikoff S (1998) Conspiracy of silence among repeated transgenes. BioEssays 20(7):532–535. https://doi.org/10.1002/(SICI)1521-1878(199807)20:7<532::AID-BIES3>3.0.CO;2-M

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

Heslop-Harrison JS, Schwarzacher T (2013) Nucleosomes and centromeric DNA packaging. Proc Natl Acad Sci USA 110(50):19974–19975. https://doi.org/10.1073/pnas.1319945110

Höck J, Meister G (2008) The Argonaute protein family. Genome Biol 9(2). https://doi.org/10.1186/gb-2008-9-2-210

Hsieh T-S, Brutlag D (1979) Sequence and sequence variation within the 1.688 g/cm^3 satellite DNA of *Drosophila melanogaster*. J Mol Biol 135(2):465–481. https://doi.org/10.1016/0022-2836(79)90447-9

Huertas D, Cortes A, Casanova J, Azorin F (2004) *Drosophila DDP1*, a multi-KH-domain protein, contributes to centromeric silencing and chromosome segregation. Curr Biol 14:1611–1620

Huo X, Ji L, Zhang Y, Lv P, Cao X, Wang Q, Yan Z, Dong S, Du D, Zhang F, Wei G, Liu Y, Wen B (2020) The nuclear matrix protein SAFB cooperates with major satellite RNAs to stabilize heterochromatin architecture partially through phase separation. Mol Cell 77(2):368–383.e7. https://doi.org/10.1016/j.molcel.2019.10.001

Ideue T, Tani T (2020) Centromeric non-coding RNAs: conservation and diversity in function. Non-coding RNA 6(1). https://doi.org/10.3390/ncrna6010004

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.e8. https://doi.org/10.1016/j.cub.2017.06.069

Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM (2017) Comparative analysis of satellite DNA in the *Drosophila melanogaster* species complex. G3 Genes Genomes Genetics 7 (2):693–704. https://doi.org/10.1534/g3.116.035352

Jagannathan M, Cummings R, Yamashita YM (2019) The modular mechanism of chromocenter formation in Drosophila. BioRxiv:1–16. https://doi.org/10.1101/481820

Jolly C, Lakhotia SC (2006) Human sat III and Drosophila hsrω transcripts: a common paradigm for regulation of nuclear RNA processing in stressed cells. Nucleic Acids Res 34(19):5508–5514. https://doi.org/10.1093/nar/gkl711

Joshi S, Meller VH (2017) Satellite repeats idenitfy X chromatin for dosage compesation in *Drosophila melanogaster* males. Curr Biol 176(1):139–148. https://doi.org/10.1016/j.physbeh.2017.03.040

Khost DE, Eickbush DG, Larracuente AM (2017) Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila. Genome Res 27(5):709–721. https://doi.org/10.1101/gr.213512.116

Khurana JS, Xu J, Weng Z, Theurkauf WE (2010) Distinct functions for the Drosophila piRNA pathway in genome maintenance and telomere protection. PLoS Genet 6:e1001246

Kim H, Nguyen N-P, Turner K, Wu S, Gujar AD, Luebeck J, Liu J, Deshpande V, Rajkumar U, Namburi S, Amin SB, Yi E, Menghi F, Schulte JH, Henssen AG, Chang HY, Beck CR, Mischel PS, Bafna V, Verhaak RGW (2020) Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat Genet 52(9):891–897. https://doi.org/10.1038/s41588-020-0678-2

Kind J, Vaquerizas JM, Gebhardt P, Gentzel M, Luscombe NM, Bertone P, Akhtar A (2008) Genome-wide analysis reveals MOF as a key regulator of dosage compensation and gene expression in drosophila. Cell 133(5):813–828. https://doi.org/10.1016/j.cell.2008.04.036

Kit S (1961) Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. J Mol Biol 3(6):711–716. https://doi.org/10.1016/S0022-2836(61)80075-2

Kiyama R, Matsui H, Oishi M (1986) A repetitive DNA family (Sau3A family) in human chromosomes: extrachromosomal DNA and DNA polymorphism. Proc Natl Acad Sci 83 (13):4665–4669. https://doi.org/10.1073/pnas.83.13.4665

Kiyama R, Matsui H, Okumura K, Oishi M (1987) A group of repetitive human DNA families that is characterized by extrachromosomal oligomers and restriction-fragment length polymorphism. J Mol Biol 193(4):591–597. https://doi.org/10.1016/0022-2836(87)90342-1

Kuhn GCS, Küttler H, Moreira-Filho O, Heslop-Harrison JS (2012) The 1.688 repetitive DNA of drosophila: concerted evolution at different genomic scales and association with genes. Mol Biol Evol 29(1):7–11. https://doi.org/10.1093/molbev/msr173

Kuroda MI, Hilfiker A, Lucchesi JC (2016) Dosage compensation in Drosophila—a model for the coordinate regulation of transcription. Genetics 204(2):435–450. https://doi.org/10.1534/genetics.115.185108

Kursel LE, Malik HS (2018) The cellular mechanisms and consequences of centromere drive. Curr Opin Cell Biol 52(206):58–65. https://doi.org/10.1016/j.ceb.2018.01.011

Kusano A, Staber C, Ganetzky B (2001) Nuclear mislocalization of enzymatically active RanGAP causes segregation distortion in Drosophila. Dev Cell 1(3):351–361. https://doi.org/10.1016/S1534-5807(01)00042-9

Kusano A, Staber C, Chan HYE, Ganetzky B (2003) Closing the (Ran)GAP on segregation distortion in Drosophila. BioEssays 25(2):108–115. https://doi.org/10.1002/bies.10222

Larracuente AM (2014) The organization and evolution of the Responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. BMC Evol Biol 14(1):1–12. https://doi.org/10.1186/s12862-014-0233-9

Larracuente AM, Presgraves DC (2012) The selfish segregation distorter gene complex of *Drosophila melanogaster*. Genetics 192(1):33–53. https://doi.org/10.1534/genetics.112.141390

Larschan E, Bishop EP, Kharchenko PV, Core LJ, Lis JT, Park PJ, Kuroda MI (2011) X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. Nature 471 (7336):115–118. https://doi.org/10.1038/nature09757

Larson K, Yan S-J, Tsurumi A, Liu J, Zhou J, Gaur K, Guo D, Eickbush TH, Li WX (2012) Heterochromatin formation promotes longevity and represses ribosomal RNA synthesis. PLoS Genet 8(1):e1002473. https://doi.org/10.1371/journal.pgen.1002473

Larson AG, Elnatan D, Keenen MM, Trnka MJ, Johnston JB, Burlingame AL, Agard DA, Redding S, Narlikar GJ (2017) Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature 547(7662):236–240. https://doi.org/10.1038/nature22822

Levinson G, Gutman G (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. https://doi.org/10.1093/oxfordjournals.molbev.a040442

Lin CJ, Hu F, Dubruille R, Vedanayagam J, Wen J, Smibert P, Loppin B, Lai EC (2018) The hpRNA/RNAi pathway is essential to resolve intragenomic conflict in the drosophila male germline. Dev Cell 46(3):316–326.e5. https://doi.org/10.1016/j.devcel.2018.07.004

Lohe AR, Brutlag DL (1987) Identical satellite DNA sequences in sibling species of drosophila. J Mol Biol 194:161–170

Lohe AR, Hilliker AJ, Roberts PA (1993) Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. Genetics 134(4):1149–1174. https://doi.org/10.1016/0168-9525(93)90135-5

Lower SS, McGurk MP, Clark A, Barbash DA (2018) Satellite DNA evolution: old ideas, new approaches. Curr Opin Genet Dev 49:70–78. https://doi.org/10.1016/j.gde.2018.03.003

Maheshwari S, Barbash DA (2012) Cis-by-trans regulatory divergence causes the asymmetric lethal effects of an ancestral hybrid incompatibility gene. PLoS Genet 8(3):e1002597. https://doi.org/10.1371/journal.pgen.1002597

Maison C, Bailly D, Peters AHFM, Quivy J-P, Roche D, Taddei A, Lachner M, Jenuwein T, Almouzni G (2002) Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. Nat Genet 30(3):329–334. https://doi.org/10.1038/ng843

Malik HS (2009) The centromere-drive hypothesis: a simple basis for centromere complexity. Prog Mol Subcell Biol 48(June):33–52. https://doi.org/10.1007/978-3-642-00182-6_2

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T (1989) A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol 109(5):1963–1973. https://doi.org/10.1083/jcb.109.5.1963

McDermott SR, Noor MAF (2010) The role of meiotic drive in hybrid male sterility. Philos Trans R Soc B Biol Sci 365(1544):1265–1272. https://doi.org/10.1098/rstb.2009.0264

McNulty SM, Sullivan LL, Sullivan BA (2017) Human centromeres produce chromosome-specific and array-specific alpha satellite transcripts that are complexed with CENP-A and CENP-C. Dev Cell 42(3):226–240.e6. https://doi.org/10.1016/j.devcel.2017.07.001

Meller VH, Rattner BP (2002) The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. EMBO J 21(5):1084–1091. https://doi.org/10.1093/emboj/21.5.1084

Mendjan S, Taipale M, Kind J, Holz H, Gebhardt P, Schelder M, Vermeulen M, Buscaino A, Duncan K, Mueller J, Wilm M, Stunnenberg HG, Saumweber H, Akhtar A (2006) Nuclear pore components are involved in the transcriptional regulation of dosage compensation in Drosophila. Mol Cell 21(6):811–823. https://doi.org/10.1016/j.molcel.2006.02.007

Menon DU, Meller VH (2012) A role for siRNA in X-chromosome dosage compensation in *Drosophila melanogaster*. Genetics 191(3):1023–1028. https://doi.org/10.1534/genetics.112.140236

Menon DU, Coarfa C, Xiao W, Gunaratne PH, Meller VH (2014) siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. Proc Natl Acad Sci USA 111(46):16460–16465. https://doi.org/10.1073/pnas.1410534111

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M (2015) Structural and functional liaisons between transposable elements and satellite DNAs. Chromosom Res 23 (3):583–596. https://doi.org/10.1007/s10577-015-9483-7

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, Schneider VA, Potapova T, Wood J, Chow W, Armstrong J, Fredrickson J, Pak E, Tigyi K, Kremitzki M et al (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

Moschetti R, Caizzi R, Pimpinellit S (1996) Segregation distortion in drosophila rnelanogaster: genomic organization of responder sequences. Genetics 144:1665–1671

Muchardt C, Guillemé M, Seeler J, Trouche D, Dejean A, Yaniv M (2002) Coordinated methyl and RNA binding is required for heterochromatin localization of mammalian HP1 α. EMBO Rep 3 (10):975–981. https://doi.org/10.1093/embo-reports/kvf194

Nagao A, Mituyama T, Huang H, Chen D, Siomi MC, Siomi H (2010) Biogenesis pathways of piRNAs loaded onto AGO3 in the Drosophila testis. RNA 16(12):2503–2515. https://doi.org/10.1261/rna.2270710

Nickerson JA, Krochmalnic G, Wan KM, Penman S (1989) Chromatin architecture and nuclear RNA. Proc Natl Acad Sci 86(1):177–181. https://doi.org/10.1073/pnas.86.1.177

Nozawa R-S, Boteva L, Soares DC, Naughton C, Dun AR, Buckle A, Ramsahoye B, Bruton PC, Saleeb RS, Arnedo M, Hill B, Duncan RR, Maciver SK, Gilbert N (2017) SAF-A regulates interphase chromosome structure through oligomerization with chromatin-associated RNAs. Cell 169(7):1214–1227.e18. https://doi.org/10.1016/j.cell.2017.05.029

Ohta T, Dover GA (1984) The cohesive population genetics of molecular drive. Genetics 108 (2):501–521

Osada N, Innan H (2008) Duplication and gene conversion in the *Drosophila melanogaster* genome. PLoS Genet 4(12). https://doi.org/10.1371/journal.pgen.1000305

Pal Bhadra M, Bhadra U, Birchler JA (2006) Misregulation of sex-lethal and disruption of male-specific lethal complex localization in drosophila species hybrids. Genetics 174(3):1151–1159. https://doi.org/10.1534/genetics.106.060541

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

Pardue ML, Gall JG (1970) Chromosomal localization of mouse satellite DNA. Science 168 (3937):1356–1358. https://doi.org/10.1126/science.168.3937.1356

Peng JC, Karpen GH (2007) H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. Nat Cell Biol 9(1):25–35. https://doi.org/10.1038/ncb1514

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL (2000) Evidence for DNA loss as a determinant of genome size. Science 287(5455):1060–1062. https://doi.org/10.1126/science.287.5455.1060

Phadnis N, Orr HA (2009) Sterility and segregation distortion in Drosophila hybrids. Science 323 (January):376–379

Piacentini L, Fanti L, Berloco M, Perrini B, Pimpinelli S (2003) Heterochromatin protein 1 (HP1) is associated with induced gene expression in Drosophila euchromatin. J Cell Biol 161 (4):707–714. https://doi.org/10.1083/jcb.200303012

Pont G, Degroote F, Picard G (1987) Some extrachromosomal circular DNAs from Drosophila embryos are homologous to tandemly repeated genes. J Mol Biol 195(2):447–451. https://doi.org/10.1016/0022-2836(87)90665-6

Presgraves DC (2010) The molecular evolutionary basis of species formation. Nat Rev Genet 11 (3):175–180. https://doi.org/10.1038/nrg2718

Ramírez F, Lingg T, Toscano S, Lam KC, Georgiev P, Chung HR, Lajoie BR, de Wit E, Zhan Y, de Laat W, Dekker J, Manke T, Akhtar A (2015) High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in drosophila. Mol Cell 60 (1):146–162. https://doi.org/10.1016/j.molcel.2015.08.024

Ranathunge C, Wheeler GL, Chimahusky ME, Kennedy MM, Morrison JI, Baldwin BS, Perkins AD, Welch ME (2018) Transcriptome profiles of sunflower reveal the potential role of microsatellites in gene expression divergence. Mol Ecol 27(5):1188–1199. https://doi.org/10.1111/mec.14522

Richards RI, Sutherland GR (1994) Simple repeat DNA is not replicated simply. Nat Genet 6 (2):114–116. https://doi.org/10.1038/ng0294-114

Rizzi N, Denegri M, Chiodi I, Corioni M, Valgardsdottir R, Cobianchi F, Riva S, Biamonti G (2004) Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. Mol Biol Cell 15(2):543–551. https://doi.org/10.1091/mbc.e03-07-0487

Roach RJ, Garavís M, González C, Jameson GB, Filichev VV, Hale TK (2020) Heterochromatin protein 1α interacts with parallel RNA and DNA G-quadruplexes. Nucleic Acids Res 48 (2):682–693. https://doi.org/10.1093/nar/gkz1138

Rohrbaugh M, Clore A, Davis J, Johnson S, Jones B, Jones K, Kim J, Kithuka B, Lunsford K, Mitchell J, Mott B, Ramos E, Tchedou MR, Acosta G, Araujo M, Cushing S, Duffy G, Graves F, Griffin K et al (2013) Identification and characterization of proteins involved in nuclear organization using drosophila GFP protein trap lines. PLoS One 8(1). https://doi.org/10.1371/journal.pone.0053091

Rosin LF, Mellone BG (2017) Centromeres drive a hard bargain. Trends Genet 33(2):101–117. https://doi.org/10.1016/j.tig.2016.12.001

Rudd MK, Wray GA, Willard HF (2006) The evolutionary dynamics of α-satellite. Genome Res 16 (1):88–96. https://doi.org/10.1101/gr.3810906

Satyaki PRV, Cuykendall TN, Wei KH-C, Brideau NJ, Kwak H, Aruna S, Ferree PM, Ji S, Barbash DA (2014) The Hmr and Lhr hybrid incompatibility genes suppress a broad range of heterochromatic repeats. PLoS Genet 10(3):e1004240. https://doi.org/10.1371/journal.pgen.1004240

Sawamura K, Yamamoto MT, Watanabe TK (1993) Hybrid lethal systems in the *Drosophila melanogaster* species complex. II. The Zygotic hybrid rescue (Zhr) gene of *D. melanogaster*. Genetics 133(2):307–313

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N (2013) Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One 8(2). https://doi.org/10.1371/journal.pone.0054710

Schauer T, Ghavi-Helm Y, Sexton T, Albig C, Regnard C, Cavalli G, Furlong EE, Becker PB (2017) Chromosome topology guides the drosophila dosage compensation complex for target gene activation. EMBO Rep 18(10):1854–1868. https://doi.org/10.15252/embr.201744292

Schubert T, Pusch MC, Diermeier S, Benes V, Kremmer E, Imhof A, Längst G (2012) Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. Mol Cell 48 (3):434–444. https://doi.org/10.1016/j.molcel.2012.08.021

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional human centromere. Science 294(5540):109–115. https://doi.org/10.1126/science.1065042

Sharma R, Jost D, Kind J, Gómez-Saldivar G, van Steensel B, Askjaer P, Vaillant C, Meister P (2014) Differential spatial and structural organization of the X chromosome underlies dosage compensation in *C. elegans*. Genes Dev 28(23):2591–2596. https://doi.org/10.1101/gad.248864.114

Shatskikh AS, Kotov AA, Adashev VE, Bazylev SS, Olenina LV (2020) Functional significance of satellite DNAs: insights from drosophila. Front Cell Dev Biol 8(May):1–19. https://doi.org/10.3389/fcell.2020.00312

Sienski G, Batki J, Senti KA, Dönertas D, Tirian L, Meixner K, Brennecke J (2015) Silencio/CG9754 connects the piwi-piRNA complex to the cellular heterochromatin machinery. Genes Dev 29(21):2258–2271. https://doi.org/10.1101/gad.271908.115

Sinclair DA, Guarente L (1997) Extrachromosomal rDNA Circles—a cause of aging in yeast. Cell 91(7):1033–1042. https://doi.org/10.1016/S0092-8674(00)80493-6

Soruco MML, Chery J, Bishop EP, Siggers T, Tolstorukov MY, Leydon AR, Sugden AU, Goebel K, Feng J, Xia P, Vedenko A, Bulyk ML, Park PJ, Larschan E (2013) The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. Genes Dev 27(14):1551–1556. https://doi.org/10.1101/gad.214585.113

Southern EM (1975) Long range periodicities in mouse satellite DNA. J Mol Biol 94(1):51–69. https://doi.org/10.1016/0022-2836(75)90404-0

Spierer A, Seum C, Delattre M, Spierer P (2005) Loss of the modifiers of variegation Su(var)3-7 or HP1 impacts male X polytene chromosome morphology and dosage compensation. J Cell Sci 118(21):5047–5057. https://doi.org/10.1242/jcs.02623

Spierer A, Begeot F, Spierer P, Delattre M (2008) SU(VAR)3-7 links heterochromatin and dosage compensation in drosophila. PLoS Genet 4(5):3–7. https://doi.org/10.1371/journal.pgen.1000066

Sproul JS, Khost DE, Eickbush DG, Negm S, Wei X, Wong I, Larracuente AM (2020) Dynamic evolution of euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the simulans Clade. Mol Biol Evol 37(8):2241–2256. https://doi.org/10.1093/molbev/msaa078

Straub T, Grimaud C, Gilfillan GD, Mitterweger A, Becker PB (2008) The chromosomal high-affinity binding sites for the Drosophila dosage compensation complex. PLoS Genet 4 (12):1–14. https://doi.org/10.1371/journal.pgen.1000302

Strom AR, Emelyanov AV, Mir M, Fyodorov DV, Darzacq X, Karpen GH (2017) Phase separation drives heterochromatin domain formation. Nature 547(7662):241–245. https://doi.org/10.1038/nature22989

Sural TH, Peng S, Li B, Workman JL, Park PJ, Kuroda MI (2008) The MSL3 chromodomain directs a key targeting step for dosage compensation of the *Drosophila melanogaster* X chromosome. Nat Struct Mol Biol 15(12):1318–1325. https://doi.org/10.1038/nsmb.1520

Swenson JM, Colmenares SU, Strom AR, Costes SV, Karpen GH (2016) The composition and organization of Drosophila heterochromatin are heterogeneous and dynamic. eLife 5:1–37. https://doi.org/10.7554/elife.16096

Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, Hartl DL (2007a) A sex-ratio meiotic drive system in *Drosophila simulans*. II: an X-linked distorter. PLoS Biol 5(11):2576–2588. https://doi.org/10.1371/journal.pbio.0050293

Tao Y, Masly JP, Araripe L, Ke Y, Hartl DL (2007b) A sex-ratio meiotic drive system in *Drosophila simulans*. I: an autosomal suppressor. PLoS Biol 5(11):2560–2575. https://doi.org/10.1371/journal.pbio.0050292

Tautz D, Trick M, Dover G (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Thomae AW, Schade GOM, Padeken J, Borath M, Vetter I, Kremmer E, Heun P, Imhof A (2013) A pair of centromeric proteins mediates reproductive isolation in drosophila species. Dev Cell 27 (4):412–424. https://doi.org/10.1016/j.devcel.2013.10.001

Tomilin NV (2008) Regulation of mammalian gene expression by retroelements and non-coding tandem repeats. BioEssays 30(4):338–348. https://doi.org/10.1002/bies.20741

Török T, Gorjánácz M, Bryant PJ, Kiss I (2000) Prod is a novel DNA-binding protein that binds to the 1.686 g/cm^3 10 bp satellite repeat of *Drosophila melanogaster*. Nucleic Acids Res 28 (18):3551–3557. https://doi.org/10.1093/nar/28.18.3551

Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, Li B, Arden K, Ren B, Nathanson DA, Kornblum HI, Taylor MD, Kaushal S, Cavenee WK, Wechsler-Reya R, Furnari FB, Vandenberg SR, Rao PN, Wahl GM et al (2017) Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature 543(7643):122–125. https://doi.org/10.1038/nature21356

Valgardsdottir R, Chiodi I, Giordano M, Rossi A, Bazzini S, Ghigna C, Riva S, Biamonti G (2008) Transcription of satellite III non-coding RNAs is a general stress response in human cells. Nucleic Acids Res 36(2):423–434. https://doi.org/10.1093/nar/gkm1056

Vermaak D, Hayden HS, Henikoff S (2002) Centromere targeting element within the histone fold domain of Cid. Mol Cell Biol 22(21):7553–7561. https://doi.org/10.1128/mcb.22.21.7553-7561.2002

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ (2009) Unstable tandem repeats in promoters confer transcriptional evolvability. Science 324(5931):1213–1216. https://doi.org/10.1126/science.1170097

Waring GL, Pollack JC (1987) Cloning and characterization of a dispersed, multicopy, X chromosome sequence in *Drosophila melanogaster*. Proc Natl Acad Sci USA 84(9):2843–2847. https://doi.org/10.1073/pnas.84.9.2843

Wei KHC, Barbash DA (2015) Never settling down: frequent changes in sex chromosomes. PLoS Biol 13(4):1–6. https://doi.org/10.1371/journal.pbio.1002077

Wei KHC, Grenier JK, Barbash DA, Clark AG (2014) Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. Proc Natl Acad Sci USA 111(52):18793–18798. https://doi.org/10.1073/pnas.1421951112

Willard HF (1985) Chromosome-specific organization of human alpha satellite DNA. Am J Hum Genet 37(3):524–532

Wines DR, Henikoff S (1992) Somatic instability of a drosophila chromosome. Genetics 131 (3):683–691

Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL, Hannan R, Saffery R, Shaw ML, Williams E, Choo KHA (2007) Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res 17(8):1146–1160. https://doi.org/10.1101/gr.6022807

Wu CI, Lyttle TW, Wu ML, Lin GF (1988) Association between a satellite DNA sequence and the *responder of segregation distorter* in *D. melanogaster*. Cell 54:179–189. https://doi.org/10.1016/0092-8674(88)90550-8

Yanez-Cuna JO, Arnold CD, Stampfel G, Boryń LM, Gerlach D, Rath M, Stark A (2014) Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Res 24(7):1147–1156. https://doi.org/10.1101/gr.169243.113

Yu Q, Colot HV, Kyriacou CP, Hall JC, Rosbash M (1987) Behavior modification by in vitro mutagenesis of a variable region within the period gene of Drosophila. Nature 326:765–769

Yuan K, O'Farrell PH (2016) TALE-light imaging reveals maternally guided, H3K9me2/3-independent emergence of functional heterochromatin in Drosophila embryos. Genes Dev 30 (5):579–593. https://doi.org/10.1101/gad.272237.115

Yunis JJ, Yasmineh WG (1971) Heterochromatin, satellite DNA, and cell function. Science 174 (4015):1200–1209. https://doi.org/10.1126/science.174.4015.1200

Zhang K, Mosch K, Fischle W, Grewal SIS (2008) Roles of the Clr4 methyltransferase complex in nucleation, spreading and maintenance of heterochromatin. Nat Struct Mol Biol 15(4):381–388. https://doi.org/10.1038/nsmb.1406

Chapter 2
Structure, Organization, and Evolution of Satellite DNAs: Insights from the *Drosophila repleta and D. virilis* Species Groups

Gustavo C. S. Kuhn, Pedro Heringer, and Guilherme Borges Dias

Abstract The fact that satellite DNAs (satDNAs) in eukaryotes are abundant genomic components, can perform functional roles, but can also change rapidly across species while being homogenous within a species, makes them an intriguing and fascinating genomic component to study. It is also becoming clear that satDNAs represent an important piece in genome architecture and that changes in their structure, organization, and abundance can affect the evolution of genomes and species in many ways. Since the discovery of satDNAs more than 50 years ago, species from the *Drosophila* genus have continuously been used as models to study several aspects of satDNA biology. These studies have been largely concentrated in *D. melanogaster* and closely related species from the *Sophophora* subgenus, even though the vast majority of all *Drosophila* species belong to the *Drosophila* subgenus. This chapter highlights some studies on the satDNA structure, organization, and evolution in two species groups from the *Drosophila* subgenus: the *repleta* and *virilis* groups. We also discuss and review the classification of other abundant tandem repeats found in these species in the light of the current information available.

Keywords Satellite DNA · Heterochromatin · Drosophila · Tandem repeats

G. C. S. Kuhn (✉) · P. Heringer
Departamento de Genética, Ecologia e Evolução, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brazil
e-mail: gcskuhn@ufmg.br

G. B. Dias
Department of Genetics and Institute of Bioinformatics, University of Georgia, Athens, GA, USA

Đ. Ugarković (ed.), *Satellite DNAs in Physiology and Evolution*, Progress in Molecular and Subcellular Biology 60,
https://doi.org/10.1007/978-3-030-74889-0_2

2.1 Introduction

Satellite DNAs are sequences typically found in large arrays containing more than one thousand repeats (extending up to megabase-Mb size arrays) that are mainly concentrated in heterochromatin-rich regions of the chromosomes, such as centromeres and subtelomeric regions (Tautz 1993; Charlesworth et al. 1994; López-Flores and Garrido-Ramos 2012; Plohl et al. 2012; Garrido-Ramos 2017), but sometimes also showing dispersed distribution along the euchromatin in the form of small arrays usually containing 1–20 tandem repeats (between full and partial repeats) (Fig. 2.1; Kuhn et al. 2012; Brajković et al. 2018; Sproul et al. 2020).

In the *Drosophila* genus, satDNAs account for more than 20% of the total DNA in several species including *D. melanogaster* (Bosco et al. 2007) and can reach up to 70%, as in the Hawaiian *D. cyrtoloma* (Craddock et al. 2016). Several satDNAs are usually found in the genome of a single *Drosophila* species. For example, there are at least 15 satDNAs in the genome of *D. melanogaster* (Lohe et al. 1993). The satDNA repeat length described in *Drosophila* typically ranges from a few bp (≤10 bp) up to ~400 bp (Palomeque and Lorite 2008; Melters et al. 2013). Several studies in *Drosophila* support the assumption that satDNAs are among the fastest evolving components of eukaryotic genomes, both in abundance and at the nucleotide sequence level. Accordingly, a single satDNA family can be found restricted to one species, as found in *D. guanche* (Bachmann et al. 1989), or shared by a group of closely related species, as found in some species from the *Drosophila obscura* group (Bachmann and Sperlich 1993).

Despite the general lack of evolutionary conservation, an increasing number of studies in *Drosophila* and other organisms has been showing the participation of satDNAs in diverse functional roles, such as in spatial chromosome organization (Pathak et al. 2013; Jagannathan et al. 2018, 2019), centromeric architecture (Rošić

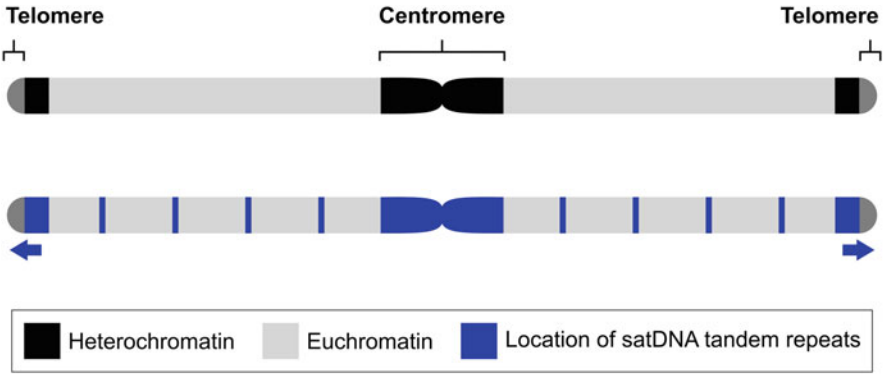


Fig. 2.1 Schematic representation of a eukaryotic chromosome and the distribution of satDNA repeats. While the main bulk of satDNA repeats resides in the heterochromatic regions of the chromosomes (here illustrated at the centromere and subtelomeric regions), arrays typically containing few repeats (less than 20) may be found dispersed along the euchromatin. Arrows indicate that repeats may expand towards the telomeres

et al. 2014), male fertility (Mills et al. 2019), and gene regulation (Menon et al. 2014; Joshi and Meller 2017).

In *Drosophila*, satDNAs studies have been mainly focused on *D. melanogaster* and other species from the *Sophophora* subgenus, despite the fact that more than 80% of all *Drosophila* species belong to the *Drosophila* subgenus (O'Grady and DeSalle 2018). In the subgenus *Drosophila*, the *virilis-repleta* radiation (Throckmorton 1975) with over 200 species is one of the most numerous and includes several species groups, among them the *repleta* and *virilis* groups (O'Grady and DeSalle 2018).

The New World *repleta* group includes more than 100 described species, most of which use cactus decaying tissues as breeding and feeding sites (Oliveira et al. 2012). The *virilis* group comprises 13 Palearctic and Nearctic tree sap-feeding species (Spicer and Bell 2002; Morales-Hojas et al. 2011). Our group and others have been using species from the *repleta* and *virilis* groups as models to address several aspects of satDNA structure, organization, and evolution. While initial studies involved satDNAs isolated by gradient centrifugation and restriction enzyme digestion (e.g., Gall et al. 1971; Kuhn and Sene 2005), more recent analyses have been conducted directly from whole sequenced genomes (e.g., Dias et al. 2015; de Lima et al. 2017; Silva et al. 2019; Flynn et al. 2020). Here we summarize some of the main findings.

2.2 Contrasting Patterns in the *repleta* and *virilis* Groups

In the *repleta* group, we have been mainly studying satDNAs in seven species from the *buzzatii* cluster (*D. buzzatii*, *D. koepferae*, *D. serido*, *D. antonietae*, *D. seriema*, *D. gouveai,* and *D. borborema*), that belong to the *buzzatii* complex, and in *D. mojavensis*, from the more distant *mulleri* complex (Fig. 2.2). Divergence times between these species range from 11 My to less than 1 My, which provides us with interesting time frames to study satDNA evolution at both long and short evolutionary terms. All these species share similar karyotypes consisting of 4 pairs of telocentric autosomes, 1 pair of microchromosomes (or chromosome 6), and 1 pair of sex chromosomes (X and Y). Heterochromatic blocks are present in the proximal region of the chromosomes, while the microchromosomes and Y chromosome are almost entirely heterochromatic (Kuhn et al. 2007).

The pBuM satDNA, initially found in *D. buzzatii*, was the first satDNA described in species from the *repleta* group (Kuhn et al. 1999). This satDNA has been found in species from both *mulleri* and *buzzatii* complexes, so that its age could be estimated as at least 11 million years (My), the oldest satDNA found in the *repleta* group to date (Fig. 2.2) (Kuhn and Sene 2005; Kuhn et al. 2008; de Lima et al. 2017).

Within the *buzzatii* complex, the pBuM satDNA exists as two main repeat variants: in the form of 190 bp repeat units, called pBuM-1 *alpha* repeats, and as ~370 bp repeats, called pBuM-2 *alpha/beta* repeats, made by an *alpha* 190 bp sequence plus a *beta* 180 bp sequence. The pBuM-1 repeats are present in species

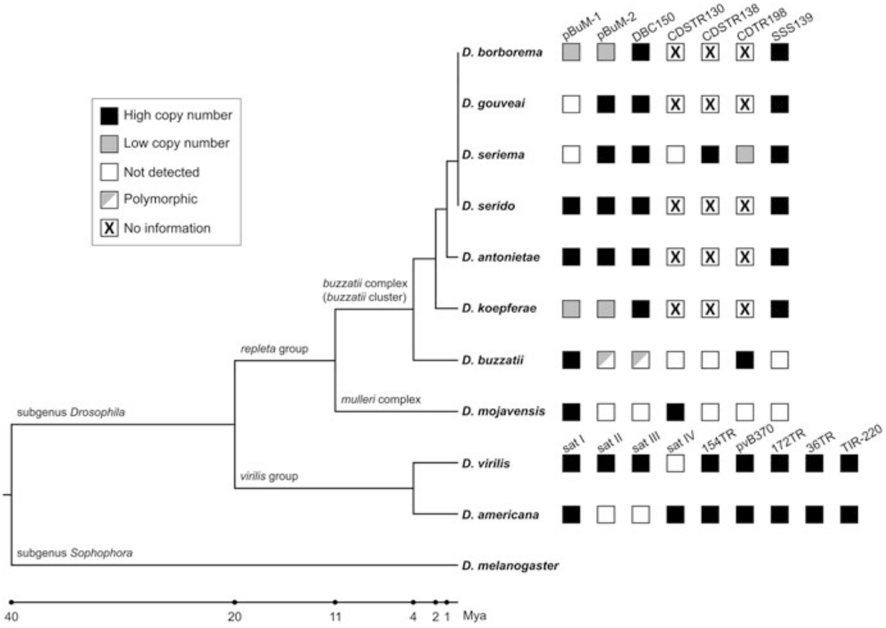


Fig. 2.2 Phylogenetic tree containing *Drosophila* species from the *repleta* and *virilis* groups mentioned in the text and the distribution of the most abundant tandem repeat families found in each species. The species *D. melanogaster* is only shown here as a reference. The scale under the phylogeny represents distances between taxa millions of years ago (Mya). Data from Kuhn et al. (2007, 2008) Franco et al. (2008); de Lima et al. (2017), Gall et al. (1971, 1974); Gall and Atherton (1974); Cohen and Bowman (1979); Zelentsova et al. (1986); Vashakidze et al. (1989); Heikkinen et al. (1995); Biessmann et al. (2000); Abdurashitov et al. (2013); Dias et al. (2014, 2015); Silva et al. (2019); Flynn et al. (2020)

from both *mulleri* and *buzzatii* complexes, but pBuM-2 repeats are restricted to the *buzzatii* complex (Fig. 2.2). Therefore, pBuM-1 repeats most likely represent the primitive state of this satellite, while the origin of pBuM-2 could be explained by the insertion of a *beta* sequence into an *alpha* array, creating an *alpha/beta* repeat that subsequently underwent amplification. The *beta* sequence has no significant sequence identity with any other described genetic element, so that *beta* could have been derived from a previously single-copy noncoding DNA sequence. The pBuM satDNA repeats are mainly located in the centromeric regions, although in *D. seriema*, it is also present in the subtelomeric regions of the microchromosomes (Kuhn et al. 2009). The pBuM chromosomal distribution varies across species, from being present in only one chromosome (the microchromosome) in *D. mojavensis* to all chromosomes in *D. buzzatii* except the X (Kuhn et al. 2008).

There is an interesting pattern of evolutionary turnover of pBuM-1 and pBuM-2 repeats across species from the *buzzatii* complex. While pBuM-1 is the main variant present in *D. buzzatii*, both pBuM-1 and pBuM-2 were found in *D. serido* and *D. antonietae*, only pBuM-2 were found in *D. seriema* and *D. gouveai* and both

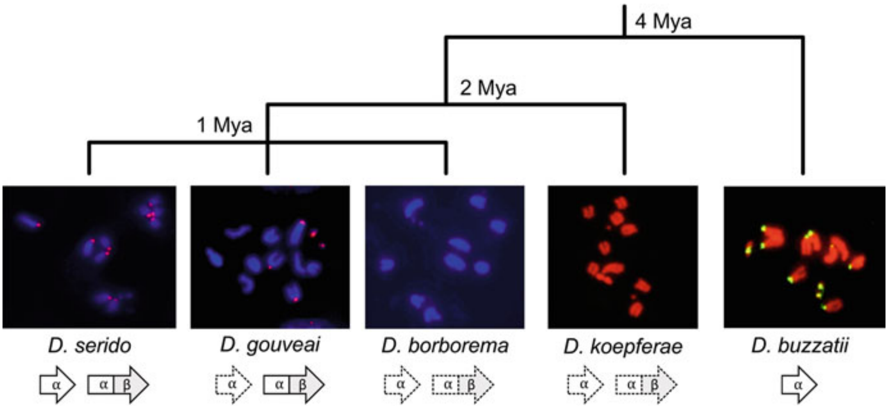


Fig. 2.3 Rapid evolutionary turnover of pBuM across species from the *Drosophila buzzatii* cluster (*repleta group*). The species shown here may contain only pBuM-1 (*alpha*), both pBuM-1 and pBuM-2 (*alpha/beta*), only pBuM-2, or not detectable levels of pBuM-1 and pBuM-2 based on fluorescence in situ hybridization (FISH) experiments. The pBuM variants are depicted as arrows and undetected repeats are represented with doted lines. Chromosomes were counterstained with DAPI (blue), or PI (red). The pBuM location is seen in red for DAPI or yellow for PI. Adapted from Kuhn et al. (2008)

pBuM-1 and pBuM-2 are almost absent from the genomes of *D. borborema* and *D. koepferae*. Given the phylogenetic position of the last two species, it can be assumed that the events resulting in pBuM loss occurred independently twice. Such pattern of turnover of pBuM variants has happened remarkably fast, considering that most forementioned species diverged from each other less than 2 Mya (Figs. 2.2 and 2.3).

The DBC-150 satDNA, with repeat units around 150 bp, has been found in species from the *buzzatii* complex, but it is absent in the more distantly related *D. mojavensis*. Therefore, this satDNA originated at least around 4 million years ago (Ma) (Fig. 2.2). In almost all species from the *buzzatii* cluster, DBC-150 satDNA repeats are abundantly located (and likely restricted) to the microchromosomes (Kuhn et al. 2007). In *D. buzzatii*, however, DBC-150 repeats are very scarce and detected only by the sensitive method of PCR, suggesting that DBC-150 likely underwent amplification after the split of *D. buzzatii* from the remaining species from the *buzzatii* cluster. Interestingly, while DBC-150 repeats could be detected in a *D. buzzatii* strain from Ibotirama (Brazil), they seem to be absent in the *D. buzzatii* strain used for genome sequencing, which was founded by flies from Carboneras (Spain) (Guillén et al. 2015). Such pattern of satDNA polymorphism has also been reported in a study comparing satDNA abundance among populations of *D. melanogaster* (Wei et al. 2014) and it is likely a common phenomenon, especially for species containing low-copy-number satDNAs.

The CDSTR138 satDNA, with repeat units 138 bp long was found in *D. seriema* and could not be detected in the sequenced genomes of *D. buzzatii* or *D. mojavensis*

(Fig. 2.2). This satDNA co-localizes with pBuM-2 in the centromeric region of four autosomes. Finally, the CDSTR130 satDNA, with repeat units 130 bp long, was found in *D. mojavensis* and could not be detected in the genomes of *D. buzzatii* and *D. seriema* (Fig. 2.2). This satDNA is located in the centromeric region of all *D. mojavensis* chromosomes, except the Y (de Lima et al. 2017).

All the four satDNAs presented above (pBuM, DBC-150, CDSTR130, and CDSTR138) are abundant in at least one species, with an estimated average number of repeats per locus >1000, except for CDSTR138 (~720 repeats). In addition, they all map to the heterochromatin, as expected for satDNAs. A fifth putative satDNA, called SSS139, has been described in all species from the *buzzatii* cluster except *D. buzzatii*, but its chromosome location has not been determined (Fig. 2.2) (Franco et al. 2008).

The whole set of satDNAs from *D. buzzatii*, *D. mojavensis,* and *D. seriema* account for 1.5%, 2.5%, and 2.9% of the total genomic DNA, respectively (de Lima et al. 2017). These values are surprisingly low compared to the satDNA content in other *Drosophila* species and eukaryotes in general, where satDNAs typically comprise more than 20% of the total genomic DNA (Bosco et al. 2007; Craddock et al. 2016; Garrido-Ramos 2017). Interestingly, the estimated heterochromatin content in *D. mojavensis* is also low (2%) compared to most *Drosophila* species, where heterochromatic regions typically comprise more than 15% of the genome (e.g., 24% in *D. melanogaster* or 44% in *D. virilis*) (Bosco et al. 2007). Finally, genome sizes have been estimated for *D. mojavensis* and *D. buzzatii* and they are also in a lower range (~150 Mb) compared to several other *Drosophila* species including *D. melanogaster*, where genome sizes are typically above 180 Mb (Bosco et al. 2007; Gregory and Johnston 2008). Interestingly, genome sizes of more distant *Drosophila* species from the *repleta* group, such as *D. mercatorum* (*mercatorum* subgroup) and *D. hydei* (*hydei* subgroup), are also small (Bosco et al. 2007), suggesting that perhaps small genomes could be a widespread characteristic in the *repleta* group.

Moving to the *D. virilis* group, we have been mainly studying satDNAs and other tandem repeats in *D. virilis* and *D. americana*, species that have diverged from each other between 4.1 and 4.5 Mya (Fig. 2.2; Caletka and McAllister 2004; Morales-Hojas et al. 2011). *D. virilis* has a karyotype consisting of 4 pairs of acrocentric autosomes, 1 pair of microchromosomes (or chromosome 6), and 1 pair of sex chromosomes (X and Y). The *D. americana* karyotype differs from the one found in *D. virilis* by centromeric fusions between chromosomes 2 and 3, and chromosomes X and 4 (Caletka and McAllister 2004). In these species, heterochromatic blocks are found in the proximal region of all autosomes and the X chromosome, with the Y chromosome appearing entirely heterochromatic (Mahan and Beck 1986).

The studies of *D. virilis* satDNAs were among the first in *Drosophila*, and revealed that most of the (peri)centromeric heterochromatin in this species is composed of three related satDNAs, all with 7 bp repeat units: sat I (AAACTAC), sat II (AAACTAT), and sat III (AAATTAC) (Gall et al. 1971, 1974; Gall and Atherton 1974). A fourth satDNA from this family, sat IV (AAACAAC), was identified in

D. virilis and other species from the *virilis* group. Based on its phylogenetic distribution, the 7 bp satDNA family likely arose at least 4.5 Mya (Fig. 2.2) (Flynn et al. 2020). The *pvB370* family, with 370 bp repeat units, has also been described as a satDNA in species from the *virilis* group (Heikkinen et al. 1995). However, given its relationship with transposable elements (TEs) and genomic distribution, this tandem repeat family will be discussed in separate topics in this chapter.

In *D. virilis*, the 7 bp satDNAs are located in the centromeric and pericentromeric regions from all chromosomes, while in *D. americana*, only the Y chromosome seems to lack them (Silva et al. 2019). Consistent with a fast evolutionary turnover, each one of the 7 bp satDNAs differ in their chromosomal distribution among species from the *virilis* group, especially in the centromeric region (Flynn et al. 2020).

There are marked differences in the 7 bp satDNAs genomic abundance between *D. virilis* and *D. americana*: while in *D. virilis* these satDNAs account for 40% of the genomic DNA (Gall et al. 1971; Gall and Atherton 1974), in *D. americana* they account for less than 20% (Flynn et al. 2020), similar to the amount of satDNAs found in other *Drosophila* species. Therefore, the amount of satDNAs in *D. virilis* can be regarded as unusually high among *Drosophila*, suggesting that satDNAs experienced a large expansion in the genome of this species. Accordingly, the heterochromatin content in *D. virilis* is also high (~40%) (Gall et al. 1971; Bosco et al. 2007) compared to other *Drosophila* species, and its genome size, around ~390 Mb, is also among the largest found in *Drosophila* (Bosco et al. 2007). In this context, it is important to mention that *D. americana* has not only a smaller satDNA content but also a smaller genome (~240 Mb) compared to *D. virilis* (Bosco et al. 2007).

In summary, in this topic, we reviewed the satDNA data in some selected species from the *repleta* and *virilis* groups. The satDNAs in the *repleta* group have repeat lengths ranging from 130 to 370 bp. In contrast, the most abundant satDNAs in the *D. virilis* and *D. americana* have repeat lengths of only 7 bp. Despite these differences, some important structural features are shared among some of them (see Sect. 2.5). While satDNAs account for up to 2% of the total DNA in the studied species from the *repleta* group, satDNAs can account for more than 20% in the *virilis* group and remarkably reaching 40% in *D. virilis* (Gall et al. 1971; Gall and Atherton 1974). Accordingly, species from the *repleta* group have small genomes (around 150 Mb) and low heterochromatin content (e.g., 2% in *D. mojavensis*), while species from the *virilis* group have much larger genomes (>240 Mb) and higher heterochromatin content (e.g., 40% in *D. virilis*). These figures are in accordance with the known positive correlation between satDNA content, heterochromatin content, and genome sizes, found in many *Drosophila* species (Bosco et al. 2007; Gregory and Johnston 2008).

In *Drosophila* variation in genome sizes might affect phenotypic features such as cell sizes, body size, sperm length, and development time (Gregory and Johnston 2008). On the heterochromatin level, changes in its content are expected to affect gene expression, centromere function, and genome stability, being therefore also considered of biological significance (reviewed in Allshire and Madhani 2018). Since changes in satDNA affect both heterochromatin content and genome sizes, it remains to be investigated whether the large discrepancies in satDNA content between species from the *repleta* and *virilis* groups have any adaptive explanation(s).

2.3 Testing Concerted Evolution

Despite the fact that satDNAs are among the most rapidly evolving components of the genome, repeats from the same satDNA family within a species usually exhibit very low levels of sequence variability. This means that within a species, individual repeats are not evolving and diverging independently from each other but that somehow, they are evolving "together." This pattern is known as "concerted evolution" and can also be seen in non-satellite multigene families (Dover 1982; Ganley and Kobayashi 2007; Goebel et al. 2017).

It is generally accepted that the evolution of satDNA is affected by mutations that create repeat variants, and by the "molecular drive" mechanisms of unequal exchanges such as unequal crossing over and gene conversion, that may increase the frequency of some particular variants in the array, leading to concerted evolution (Dover 1982). For centromeric satellites, homogenization can be further accelerated if the expanded variants confer an advantage for the chromosome to be transmitted to the egg during female meiosis and consequently to the next generation, as proposed in the "Centromere Drive" model (Henikoff et al. 2001; Malik 2009).

Early studies in *Drosophila* have been pivotal to show that concerted evolution is essentially the result of a gradual process (Strachan et al. 1985). However, there are also cases where satDNA changes happen too rapidly in evolution so that transitional stages of satDNA turnover cannot be seen between species. An extreme example of such rapid evolution is given by the existence of species-specific satDNAs, where satDNA repeats present in one species cannot be detected even in a closely related species (Bachmann et al. 1989). Such dramatic changes in satDNAs (both quantitatively and at the sequence level) occurring rapidly between populations likely contribute to the speciation process. This is particularly interesting in the case of centromeric satDNAs, where centromeric binding proteins also appear to be coevolving with satDNAs (Malik 2009). In fact, it has already been shown in *Drosophila* that changes in satDNA abundance and nucleotide sequence are related to post-zygotic lethality in hybrids between *D. melanogaster* and *D. simulans*, possibly because of incompatibilities between centromeric proteins from one species and satDNA from the other (Ferree and Barbash 2009).

We tested the prediction of concerted evolution in two satDNAs from the *repleta* group, the pBuM, and the DBC-150, given that these satellites are present in several species and feature a reasonable number of nucleotide sites for phylogenetic analysis (between 150 and 370 bp).

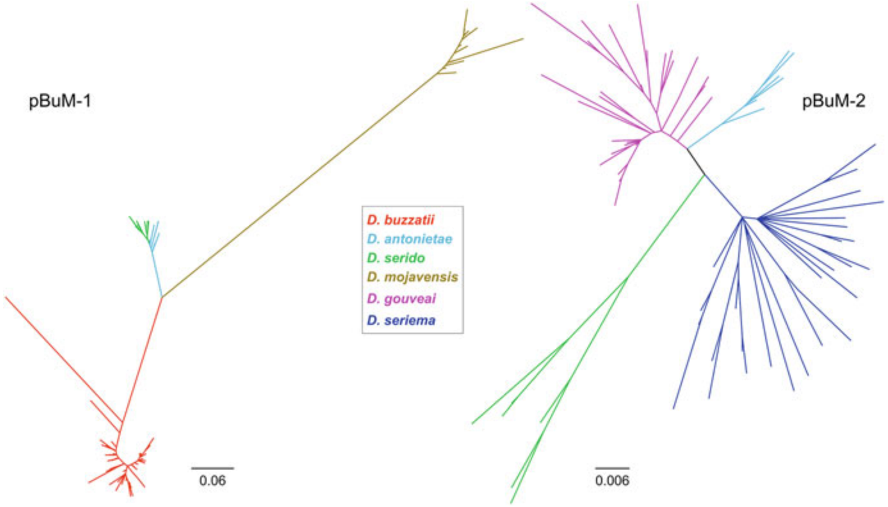


Fig. 2.4 Maximum likelihood (ML) trees containing pBuM-1 or pBuM-2 repeats sampled from species belonging to the repleta group. Concerted evolution is seen here because repeats from the same species (same color) are allocated in species-specific branches. Scale bar represents the number of substitutions per site

For the pBuM satDNA, we found a clear pattern of concerted evolution, with repeats forming species-specific branches on phylogenetic trees (Fig. 2.4; Kuhn et al. 2007; de Lima et al. 2017). However, for species presenting very low amounts of pBuM (e.g., only detectable by PCR amplification), these repeats do not form species-specific branches (Kuhn et al. 2008). It is not clear why these low copy number repeats did not undergo concerted evolution. Among possible hypotheses, these low copy number repeats could belong to a pool of ancestral variants or a "library" (Plohl et al. 2008) that independently underwent homogenization in some species. Alternatively, these low copy number repeats could have been brought to the genome through hybridization with other species during the early stages of speciation. In fact, there is evidence for introgression of genetic material (both mitochondrial and nuclear) among species from the *buzzatii* complex (Franco et al. 2010; Moreyra et al. 2019). Given the lack of evidence for the existence of a pool of ancestral repeats in the sequenced genomes of *D. buzzatii* or *D. mojavensis*, the introgression hypothesis seems more likely.

In *D. buzzatii*, we further detected an interesting case of concerted evolution among chromosomes. In this species, pBuM-1 repeats are found in all autosomes and the Y, but a particular group of divergent pBuM-1 variants, first described in Kuhn et al. (2003), is predominantly located in the Y chromosome (de Lima et al. 2017). These pBuM-1 variants linked to the Y chromosome illustrate how the presence of a satDNA on a non-recombining chromosome may lead to efficient local homogenization and chromosome-specific arrays.

The evolution of the DBC-150 in species from the *buzzatii* complex revealed a more complex pattern. For this satellite, the individual repeats isolated from each species were not more similar to each other compared to repeats from different species. Consequently, they were not grouped in species-specific branches on phylogenetic trees (Kuhn et al. 2007). Moreover, the within-species inter-repeat variability is also significantly higher in DBC-150 compared to pBuM. This was an unexpected finding, considering that pBuM has a multi-chromosomal distribution and DBC-150 is located on a single chromosome, a situation that in theory could facilitate homogenization and concerted evolution (Strachan et al. 1985). The microchromosomes in *Drosophila* do not undergo recombination during meiosis, which led us to first hypothesize that the low homogenization of DBC-150 could be somehow related to such lack of meiotic recombination (Kuhn et al. 2007).

In subsequent studies, we were able to study a larger sample of DBC-150 adjacent monomers and this analysis revealed that DBC-150 can be organized in the form of higher-order-repeats (HOR) (Kuhn et al. 2009). For example, in *D. serido*, we found that DBC-150 is organized in the form of 3 variant monomers (3mer) that are tandemly repeated. Although the variability between monomers within the HOR is very similar to the variability between monomers obtained individually (~9%), when the whole 3mers are compared to each other, the variability drops to 1.2% on average.

The finding that DBC-150 can be organized as HOR showed that intraspecific homogenization did take place in this satellite, but at the level of the HOR. Accordingly, it is possible that as more DBC-150 repeats from the other species become available, the concerted evolution of this satellite may become more evident.

2.4 SatDNAs as Major (but Likely Not Exclusive) Components of Centromeres

Although centromeric and pericentromeric satDNAs are seemingly ubiquitous in most eukaryotes (Plohl et al. 2012; Garrido-Ramos 2017; Hartley and O'Neill 2019), the presence of long tandemly repeated arrays is not a sine qua non feature of functional centromeric loci, especially in neocentromeres (Talbert and Henikoff 2020). Aside from a few exceptions, centromeric and pericentromeric satDNAs appear to provide stabilization and characterize mature centromeres (Kalitsis and Choo 2012; Nergadze et al. 2018).

In the three species from the *repleta* group with sequenced genomes, *D. buzzatii*, *D. seriema,* and *D. mojavensis*, we were able to identify likely all the most abundant centromeric satellites (Kuhn et al. 2008, 2009; de Lima et al. 2017). In *D. buzzatii*, we found that the pBuM-1 is the satellite associated with the centromere on chromosomes 2, 3, 4, 5, 6, and Y. In *D. seriema*, we found three satellites at the centromeric region: (i) the pBuM-2 on chromosomes 2, 3, 4, 5, and 6; (ii) the CDSTR138 satellite on the chromosomes 2, 3, 4, and 5, and (iii) the DBC-150 on chromosome 6 (the microchromosome). In *D. mojavensis*, we found the CDSTR130 in the centromeres of chromosomes 2, 3, 4, 5, 6, and the X.

None of the above satDNAs were mapped to the X chromosome of *D. buzzatii*, X and Y chromosomes of *D. seriema* or the Y chromosome of *D. mojavensis,* suggesting that these sex chromosomes may have satellite-free centromeres. In this context, a TE called PERI, related to DINEs (a family of *Helitrons* found in *Drosophila*; Locke et al. 1999; Yang and Barbash 2008), have been found enriched at (or near) the centromeric regions of chromosomes X and Y in species from the *buzzatii* cluster (Kuhn and Heslop-Harrison 2011; Rius et al. 2016) and are likely candidates to fulfill the centromeric DNA from the forementioned sex chromosomes. For *D. seriema,* it is also possible that the SSS139 tandem family (Fig. 2.2) (Franco et al. 2008) might be present in some centromeres, including from sex chromosomes, but unfortunately, the chromosome distribution of this family has not been determined yet.

In species from the *virilis* group, the 7 bp satDNAs (sat I, II, III, and IV) dominate the centromeric regions from most chromosomes (Gall et al. 1971; Silva et al. 2019; Flynn et al. 2020). However, the specific satDNA linked to each centromere varies within and between species (Flynn et al. 2020). For instance, in *D. virilis*, the major centromeric satDNAs are the sat II and sat III, while in *D. americana*, sat IV is the major centromeric satDNA. In *D. novamexicana*, a species more closely related to *D. americana*, sat IV also dominates the centromere of all chromosomes, except the microchromosomes (Flynn et al. 2020).

In addition to the 7 bp satDNAs, another DINE TE, called *DINE-TR1* (see Sects 2.7 and 2.11), has been found enriched in the centromeres of chromosome 5 and Y in *D. virilis* and the centromere of chromosome Y in *D. americana* (Dias et al. 2015). As both *DINE-TR1* and 7 bp satDNAs are found in the centromeric region of chromosomes 5 and Y from *D. virilis*, it is not possible to determine which sequence likely corresponds to the functional centromeric DNA. In contrast, *DINE-TR1* is the only sequence known to cover the Y centromeric region from *D. americana*.

Therefore, both *repleta* a and *virilis* groups may contain centromeres with DINE-related transposable elements. Interestingly, both DINE-PERI and DINE-TR1 feature internal tandem repeats as part of their structure. While PERI contains two blocks with internal tandem repeats, one with 97–153 bp repeats and another with 383 bp repeats (Kuhn and Heslop-Harrison 2011), DINE-TR1 contains one block with 154 bp repeats, called 154TR (Dias et al. 2015). Furthermore, we found that in both DINEs, these internal tandem repeats underwent independent expansions in some *Drosophila* species. For example, the 154TR is one of the most abundant tandem repeats found in *D. virilis* (Melters et al. 2013; Dias et al. 2015). Therefore, there might be an interesting link between the presence DINEs in the centromeres and the expansion of their internal tandem repeats, possibly forming satDNA-like arrays in the centromeres (see also Sects. 2.7 and 2.11).

In summary, the centromeres of the *Drosophila* species from the *repleta* and *virilis* groups are composed of one or more satDNAs, and perhaps by different types of DINE *Helitrons*. Interestingly, non-LTR retroelements have recently been found to participate in centromere function in *D. melanogaster* and *D. simulans* (Chang et al. 2019). It remains to be investigated whether DINEs also perform a centromeric function in species from the *repleta* and *virilis* groups.

The centromeres of *Drosophila* species seem to have been evolving by (i) turnover of different variants from the same satDNA family, for example, the pBuM-1 in *D. buzzatii* and pBuM-2 in *D. seriema*; (ii) turnover of nonhomologous satDNA families, for example, pBuM in several species from the *buzzatii* cluster but CTR130 in *D. mojavensis*; and (iii) turnover of satDNAs and possibly non-satellite sequences, as seen in the presence of *DINE-1s* in the centromeres of some chromosomes in species from both *repleta* and *virilis* groups. In most cases, a complete turnover of centromeric satDNAs between species took only 4 My to happen, while in other cases, it took even less than 1 My (Fig. 2.3). At the protein side, Cid (*Drosophila* CenH3) and Cenp-C, which are centromeric proteins known to bind to the centromeric DNA and essential for centromere function, also showed rapid evolution in species from the *repleta* and *virilis* groups, including with some instances of positive selection (Kursel and Malik 2017; Teixeira et al. 2018). Therefore, it is possible that the rapid evolution of centromeric satDNAs, possibly associated with deleterious effects associate with it, might be the driving force behind the rapid evolution of these centromeric proteins (Malik 2009).

2.5 Common Structural Features Among Centromeric satDNAs

Despite satDNA rapid evolution, dyad symmetries in stretches of short (<10 bp) palindromic sequences are common features of eukaryotic centromeric satDNAs, with the notable exception of the ones found in great apes and mice (Talbert and Henikoff 2020). In centromeric regions, these palindromic sequences are expected to adopt non-B-form DNA structures, such as stem-loops, that are thought to recruit specific proteins which, in turn, work as centromere identifiers (Kasinathan and Henikoff 2018; Talbert and Henikoff 2020). AT-rich DNA is another characteristic of centromeric sequences, which may also facilitate non-B DNA forms by its tendency to melt more easily in comparison to GC-rich sequences (Talbert and Henikoff 2020).

All centromeric satDNA sequences found in species from the *repleta* group contain short dyad symmetries covering a large portion of their length. In the *virilis* group, only sat III and the 154TR repeats (the latter are expanded tandem repeats initially present inside DINE-TR1 TEs), have dyad symmetries spanning most of their length. In accordance with these observations, all these tandem repeats were predicted to form stable secondary structures (also more stable than what would be expected by chance) (Fig. 2.5), except for individual repeats from the DBC-150 satDNA found in the *repleta* group. However, DBC-150 was also found to be organized as HOR in the form of 3mers (see Sect. 2.3), and taking into consideration this longer repeated structure, DBC-150-HOR forms more stable secondary structures than DBC-150 individual repeats Fig. 2.5).

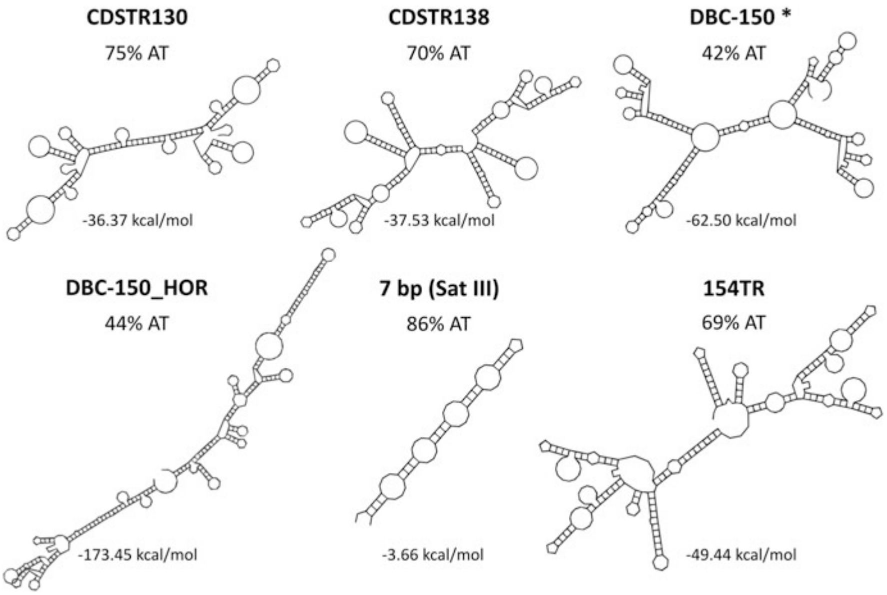


Fig. 2.5 Predicted secondary structures of centromeric satDNAs found in species from the *repleta* (CDSTR130, CDSTR138, DBC-150) and *virilis* (Sat III, 154 TR) groups. AT content (in %) is shown below each satDNA. Minimum free energy (MFE) is shown below each predicted structure. The asterisk in DBC-150 indicates that the secondary structure predicted for this satDNA is not more stable than expected by chance

In addition, all centromeric satDNAs found in the *repleta* and *virilis* group species, and 154TR, have a high AT content (~63-86%), except for DBC-150 and the related sequence DBC-150_HOR, that have a 42% and 44% AT content, respectively. Although DBC-150 and DBC-150_HOR diverge from the general trend observed in the other centromeric satDNAs, they are still enriched in dyad symmetries and expected to form secondary structures (Fig. 2.5).

The pBuM satDNA from the *repleta* group deserves some special comments. As discussed previously, this satDNA is found as two main variants. The pBuM-1 variant consists of *alpha* repeats approximately 190 bp long. The pBuM-2 variant consists of 370 bp repeats composed of a 190 bp *alpha* sequence plus a 180 bp *beta* sequence. There is no homology between *alpha* and *beta* sequences. These pBuM variants are found in high copy numbers in *D. mojavensis* (pBuM-1), *D. buzzatii* (pBuM-1), *D. antonietae* (pBuM-1 and pBuM-2), *D. serido* (pBuM-1 and pBuM-2), *D. seriema* (pBuM-2), and *D. gouveai* (pBuM-2). As these pBuM variants are potentially associated with the centromeres in these species, we further verified if both contain structural features typically found in centromeric DNAs.

Our analyses suggest that both pBuM-1 and pBuM-2 variants are enriched in dyad symmetries and are predicted to form stable secondary structures (Fig. 2.6).

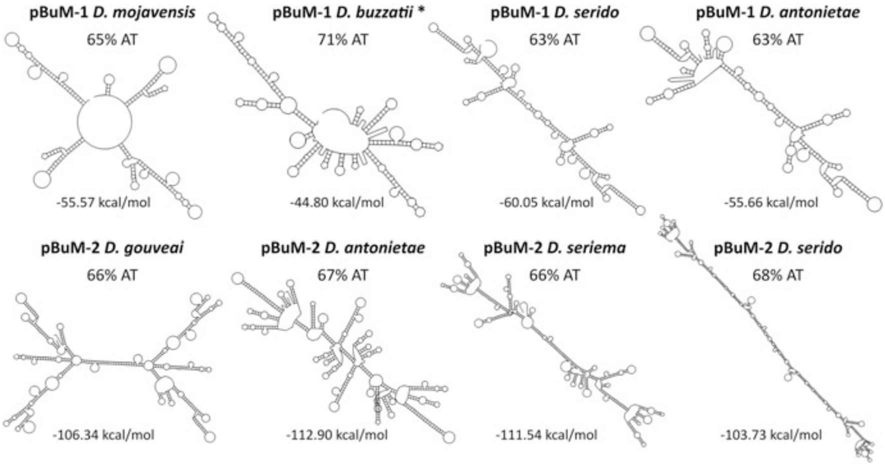


Fig. 2.6 Predicted secondary structures of pBuM satDNA variants found in the centromeric regions of species from the *buzzatii* cluster (*repleta* group). AT content (in %) is shown below each satDNA. Minimum free energy (MFE) is shown below each predicted structure. The asterisk in pBuM-1 from *D. buzzatii* indicates that the secondary structure predicted for this satDNA is not more stable than expected by chance

Moreover, both pBuM-1 and pBuM-2 have a high AT content (~63%-71%), which is a common feature for centromeric satDNAs. That is interesting, considering that pBuM-1 and pBuM-2 differ in their repeat lengths (~190 and ~370 bp) and the fact that almost half of the pBuM-2 length is nonhomologous to pBuM-1.

We have previously noted that both *alpha* and *beta* sequences share similar lengths (~190 and 180 bp) and AT content (65% and 70%) (Kuhn and Sene 2005). Therefore, it is possible that the origin and subsequent centromeric expansion of the most derived *alpha/beta* repeats were only possible because the *beta* sequence shared similar size and structural features with *alpha* sequences, including propensity to adopt non-B-form DNA structures.

In summary, even though the satDNA sequences from the *virilis* and *repleta* group species may be very different concerning their repeat lengths and nucleotide composition, they are generally enriched for dyad symmetries, they are expected to form stable secondary structures and they are AT rich. Importantly, these features appear to be essential for centromeric functions and are conserved between species, despite the high turnover rates of centromeric satDNAs. It is also tempting to speculate that the presence of structural features of centromeric satDNAs in the 154TR may have enabled this tandem repeat, initially restricted to the DINE-TR1 TE at non-centromeric regions, to successfully colonize the centromeres of some chromosomes in *D. virilis* (Fig. 2.9).

2.6 Not Just Homogeneous Arrays

While satDNAs arrays have been often described as long and homogeneous, in some species from the *repleta* group we found instances where two nonhomologous satDNAs not only co-localize in the same chromosome regions, such as pBuM-2 and DBC-150 or pBuM-1 and CDSTR130, but are also highly interspersed with each other in the same arrays. All the cases of satDNA interspersion were confirmed using cytology and DNA sequence data (Kuhn et al. 2009; de Lima et al. 2017). The analyses of sequence junctions between nonhomologous satDNAs indicate that these high levels of interspersion arose multiple times through illegitimate recombination, and possibly with subsequent rounds of unequal crossing-over expanding the copy number of some of the junctions (Kuhn et al. 2009).

The interspersion between DBC-150 and pBuM, and between pBuM and CDSTR130, take place in the microchromosomes. These highly heterochromatic chromosomes can be found in several *Drosophila* species (including *D. melanogaster*) and are typically referred to as the “dot” chromosomes. Due to their small size, it is not possible to determine which satDNA is present in the centromere core using cytology. However, the microchromosome present in *D. seriema* is unusually large (Kuhn et al. 1996), which allowed us to verify that in this chromosome DBC-150 is located in the centromere and pericentromeric regions, while pBuM is distributed from more distal to terminal regions (Fig. 2.7). For other species, like *D. antonietae*, looking at the distribution of pBuM and DBC-150 repeats in the much less condensed chromosomes in the interphase nuclei, it is possible to note a large compartment of DBC-150 repeats alone (that are probably located in the centromere core) in contrast to a region where the distribution of pBuM and DBC-150 overlaps (Fig. 2.7).

Our data showing repeats mainly from one satellite in the centromere, but interspersion of repeats from different satellites at more distal positions, is in accordance with what has been found in other well-studied centromeres, where

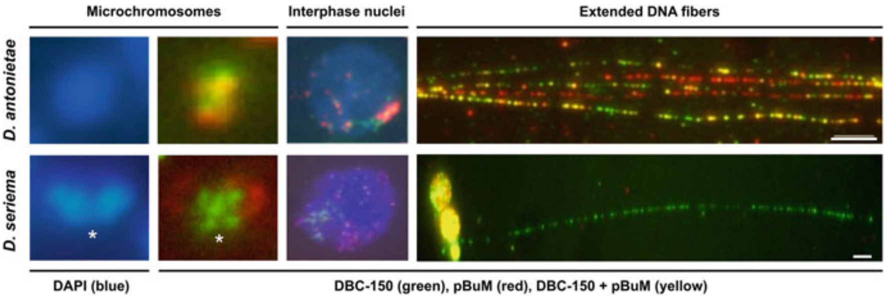


Fig. 2.7 Colocalization and interspersion of pBuM and DBC-150 satDNA repeats in the microchromosomes in two species from the *repleta* group. Different levels of interspersion can be seen in all species where both satellites are present, but in the centromere, DBC-150 arrays are most likely homogeneous (as seen here in *D. seriema*). Asterisk marks the centromere. Scale bar = 10 kb. Adapted from Kuhn et al. (2009)

repeat arrays at the core of the centromere are more contiguous and homogeneous, while arrays at more distal positions, usually present higher levels of inter-repeat variability and transposon insertions (Schueler et al. 2001; Khost et al. 2017). This pattern of variation suggests that homogenization mechanisms, such as unequal crossing over and gene conversion, act more efficiently at the core of satDNA arrays compared to their edges. However, it is important to point out that different sets of centromeric repeats may be homogenized in different species. In fact, this chapter provides several examples showing a remarkably fast rate of evolution for centromeric satDNAs across species.

2.7 TE-Tandem Repeat Associations

Multiple studies have reported on the evolutionary relationships between tandem repeats, including satDNAs, and transposable elements (TEs). These reports include data from both plant and animal species and demonstrate many possible routes by which tandem repeats and TEs can interact in eukaryotic genomes (Meštrović et al. 2015). Our group and others have described aspects of these associations in *Drosophila* species, as well as explored the possible consequences of these interactions to genome structure and function.

The first description of a TE-tandem repeat association in the *virilis* group came from Heikkinen et al. (1995), where the authors detected sequence homology between the *pvB370* satDNA repeats and the termini of pDv transposable elements (Fig. 2.8; Heikkinen et al. 1995). Because the known phylogenetic distribution of pDv elements was restricted to the *virilis* subgroup and the *pvB370* was more broadly distributed, being found in all species of the *virilis* group, the authors suggested that pDv TEs were derived from *pvB370* repeats through sequence rearrangements. This is the only report in *Drosophila* where a tandem repeat has potentially contributed to the origin of a TE, although Heikkinen et al. (1995) has cautioned that the evidence for pDv mobility is limited to interspecific hybrid studies and no target site duplications were found.

Interestingly, in addition to their association with *pvB370*, pDv elements harbor an internal array of 36 bp tandem repeats (later named 36TR) (Fig. 2.8; Zelentsova et al. 1986; Silva et al. 2019). In *D. virilis*, these 36TR sequences are distributed in ~200 loci across chromosome arms and in the telomeric region. In *D. lummei* a similar distribution was observed, albeit with a significantly lower abundance and number of loci (~20 loci) (Vashakidze et al. 1989). This pattern likely replicates the overall observed distribution of pDv itself (Zelentsova et al. 1986) and *pvB370* sequences (Biessmann et al. 2000).

The second TE-tandem repeat association in the *virilis* group involves a Terminal Inverted Repeat (TIR) transposable element named Tetris, a foldback DNA transposon that includes 220 bp tandem repeats (TIR-220) as part of its large TIRs (Dias et al. 2014; Fig. 2.8). TIR-220 can also be found forming long satDNA-like arrays in the genomes of both *D. virilis* and *D. americana*, suggesting this tandem repeat

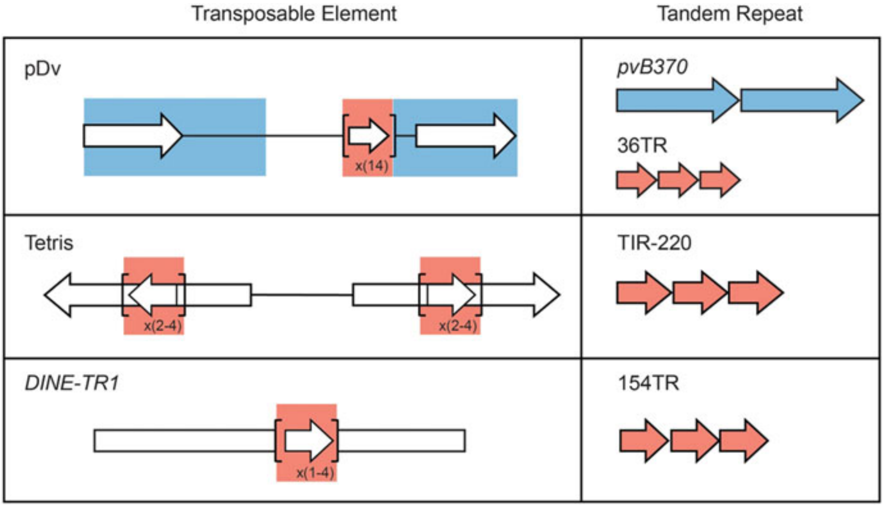


Fig. 2.8 Transposable elements (TEs), and their associated tandem repeats (TRs) found in species from the *virilis* group. Colored blocks indicate regions within the TE structure that are highly similar to an existing TR. Numbers below brackets indicate the typical copy number of TRs within structurally complete TE insertions, as determined from genome assembly data. Not drawn to scale

expansion event predates the *virilis* subgroup radiation (>4.1 Mya). Tetris and TIR-220 are highly enriched in the β-heterochromatin of *D. virilis* and *D. americana* chromosomes, and the molecular characteristics of this genomic compartment were suggested as being conducive for the formation of long tandem repeat arrays (see Sect. 2.8.) (Dias et al. 2014).

The third and latest description of a TE-tandem repeat association in the *virilis* group involves the *DINE* (*Drosophila* Interspersed DNA Elements) TE family (Locke et al. 1999). These elements comprise one of the most abundant TE families in *Drosophila* and are currently classified as Helentrons, a group of nonautonomous endonuclease-encoding *Helitrons* (Kapitonov and Jurka 2007; Yang and Barbash 2008; Thomas et al. 2014). Based on previous work that identified an abundant *Helitron*-associated tandem repeat in the genome of *D. virilis* (Abdurashitov et al. 2013), we have described a new group of *DINEs* in Acalyptratae (Diptera) called *DINE-TR1* that contains central tandem repeats (CTRs) with ~150 bp monomers (Fig. 2.8; Dias et al. 2015). *DINE-TR1* CTRs (later named 154TR) were also detected forming multi-kb sized satDNA-like arrays in the genomes of both *D. virilis* and *D. americana*, suggesting an expansion event that took place before the divergence of the *virilis* subgroup (>4.1 Mya). Although *DINE-TR1* can be detected in multiple euchromatic sites, an overabundance of 154TR was observed in the β-heterochromatin of multiple chromosomes in *D. virilis* and *D. americana* (Fig. 2.9). Additionally, *DINE-TR1* and 154TR are highly enriched in the heterochromatic Y chromosome in *D. virilis* and *D. americana*, covering most of its length (Dias et al. 2015). A similar enrichment in the Y chromosome was also observed for

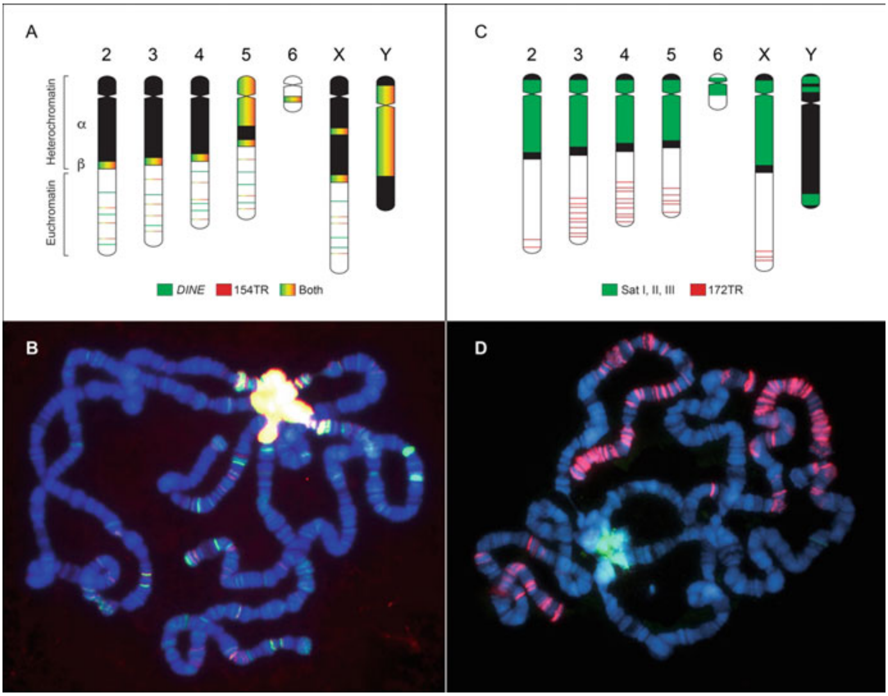


Fig. 2.9 Distribution of euchromatic and heterochromatic tandem repeats (TR) in the *Drosophila virilis* genome. Ideograms representing the six acrocentric chromosomes of *D. virilis* are depicted with black and white regions representing the heterochromatin and euchromatin, respectively. Heterochromatin is further subdivided into α and β, the latter being the most distal to the centromere. (**a**, **b**) Distribution of DINE and 154TR in metaphasic and polytenic chromosomes. (**c**, **d**) Distribution of Sat I, II, and III, and the putative minisatellite 172TR in mitotic and metaphasic and polytenic chromosomes. Metaphasic chromosomal distribution of repeats was compiled from fluorescence in situ hybridization (FISH) experiments (Dias et al. 2015; Silva et al. 2019; Flynn et al. 2020). Depiction of euchromatic repeats indicates approximate abundance rather than actual locations. The morphology of chromosomes Y and 6 can be hard to determine in DAPI-stained preparations and, as a result, the inferred position of probes relative to the centromeres of Y and 6 might be inaccurate

a related element named *PERI* in *D. serido* from the *D. buzzatii* cluster (Marin et al. 1992; Kuhn and Heslop-Harrison 2011), indicating that *DINE*-derived sequences might be important players in sex chromosome differentiation in *Drosophila*. An independent tandem repeat expansion event from a homologous *DINE-TR1* element was also suggested in *D. biarmipes*, which shares a common ancestor with *D. virilis* > 40 Mya. This suggests that *DINE-TR1* might be a recurrent source of abundant tandem repeats, and maybe satDNAs, in *Drosophila* (Dias et al. 2015).

The four satDNAs we studied in the *repleta* group (pBuM, DBC-150, CDSTR130, and CDSTR138), are nonhomologous and showed no significant sequence identity to any known transposable element, intron sequence, or annotated coding sequence. Thus, they likely originated from single copy noncoding DNA sequences.

The fact that all TE-tandem repeat associations described in the *virilis*-*repleta* radiation so far are restricted to the *virilis* group is intriguing. One possibility is that the frequency of tandem repeat expansion events correlates with the diversity and abundance of repeats in any given genome, ultimately reflecting the known correlation between repeat abundance and genome size (Elliott and Gregory 2015). Indeed, the *virilis* and *repleta* groups contain species with some of the highest and lowest genome sizes and repeat contents in *Drosophila*, respectively. Specifically, *D. virilis* has the highest and *D. buzzatii* the lowest amount of satDNAs reported in the *Drosophila* genus (see Sect. 2.2).

Another aspect of identifying TE–tandem repeat relationships is that such analyses heavily depend on the availability and quality of the DNA sequencing data and the genome assemblies. In this sense, one thing to note is the high number of species in the *repleta* group for which no genome assembly or DNA sequencing data is available. Only ~4% of the 106 described species in the *repleta* group have a genome assembly available in GenBank as of October 2020. In contrast, nearly two-thirds of *virilis* group species have an assembly available. Importantly, the contiguity of genome assemblies greatly impacts the ability to analyze repetitive DNAs, and it has been discussed that more fragmented short-read assemblies result in an overall underestimation of repeat content (Treangen and Salzberg 2011; Rius et al. 2016).

Additional DNA sequencing and analysis of *repleta* group species genomes, especially with long-read technologies, will provide a clearer picture of the differences in repetitive DNA abundance and composition between the *repleta* and *virilis* species groups. Importantly, assembly-free methods based on short-read sequencing such as RepeatExplorer and TAREAN (Novák et al. 2013, 2017) have been shown to enable initial detection of TE–tandem repeat associations, which should provide a valuable method for conducting large surveys of these connections across taxa (Silva et al. 2019).

2.8 β-Heterochromatin: Origin of New Tandem Repeats and the Chromatin Sink Hypothesis

In *Drosophila* chromosomes, the transition zone between the highly compacted α-heterochromatin, which includes the centromere and pericentromeric regions, and the more lightly packed euchromatin, which includes most protein-coding genes, is called β-heterochromatin (Heitz 1934). This region does not develop a precise banding pattern during salivary gland chromosome polytenization and instead forms a mesh-like mass around the chromocenter. We have identified two independent events of TE–internal tandem repeat expansions in the *Drosophila virilis* subgroup involving TEs with radically different structures and copy numbers. Despite these differences, in both cases, we have detected a significant enrichment of

TEs and their derived tandem repeats in the β-heterochromatin (Dias et al. 2014, 2015).

Tetris and TIR-220 are enriched in the β-heterochromatin of chromosome 2 in *D. virilis*, and chromosome 2;3 (fused) in *D. americana* (Dias et al. 2014). *DINE-TR1* and the 154TR satDNA are enriched in the β-heterochromatin of chromosomes 2, 3, 4, 5, and X in *D. virilis*, and chromosomes 2;3 (fused), 5, and X;4 (fused) in *D. americana*. This recurrent localization of recently formed abundant tandem repeats in the β-heterochromatin indicates a possible role for this genomic compartment as a "nursery" of new satDNA-like sequences (Dias et al. 2014, 2015). Specifically, the lower gene density compared to euchromatin and the higher recombination rates compared to α-heterochromatin could be argued to make β-heterochromatin a suitable environment for the frequent insertion and rearrangement of TE sequences that could lead to the generation and expansion of tandem repeats. Earlier work in *D. melanogaster* has already highlighted the complex repetitive nature of β-heterochromatin (Miklos et al. 1988; Vaury et al. 1989). This region has been dubbed a "graveyard" of TEs given its "clustered-scrambled" organization and density of repetitive sequences (Wensink et al. 1979).

More recent genomic data has revealed that the β-heterochromatin contains most piRNA clusters in *Drosophila* (Brennecke et al. 2007). These loci are transcribed regions of the genome which act like a catalog of repetitive sequences to be silenced transcriptionally (through histone modifications and heterochromatinization) or post-transcriptionally (through RNA degradation) by the *PIWI*-piRNA machinery (Brennecke et al. 2007; Thomas et al. 2013). We have found *DINE-TR1* copies inserted in multiple piRNA clusters in *D. virilis*, as well as short RNAs derived from *DINE-TR1* that match the piRNA sequence profile (Dias et al. 2015). This indicates that *DINE-TR1,* as well as 154TR, could be targeted for piRNA silencing, which led us to speculate more broadly that TE-derived tandem repeat arrays could be transcriptionally silenced when their parental TE sequence inserts within an active piRNA cluster. Since piRNA-mediated transcriptional silencing in *Drosophila* involves the deposition of repressive chromatin marks such as tri-methylation of lysine 9 in histone H3 (H3K9me3), insertion of a TE containing internal tandem repeats could result in heterochromatinization of both the TE loci across the genome, as well as the tandem repeat loci derived from that TE (Dias et al. 2016). In cases where TE-derived repeat arrays are very abundant (or satDNA-like), this silencing could interfere with chromatin dynamics in the whole genome by introducing sudden shifts in the amount of heterochromatin, i.e., acting as "chromatin sinks" (Dimitri and Pisano 1989; Francisco and Lemos 2014; Berloco et al. 2014). Additionally, tandem repeat array length fluctuations inside piRNA-targeted TEs in the euchromatin could act as tuning knobs of gene expression by altering the amount of chromatin repressors covering the nearby regions (King et al. 1997; Lee 2015). In this sense, the apparent tendency of tandem repeat expansions from TEs within β-heterochromatin could have manifold consequences for genome structure.

2.9 Alternative Scenarios for Tandem Repeat Origin from TEs

Although the number of cases describing TE–tandem repeat relationships in eukaryotes has been mounting, no clear themes have emerged so far indicating recurrent features such as TE family, tandem repeat length, chromosome distribution, etc. One exception to this lack of a pattern might be the tendency of TEs containing preexisting tandem repeats to serve as the substrate and vehicle for the formation of satDNA-like repeats. This phenomenon has been more extensively studied in plant genomes where multiple long terminal repeat (LTR) retrotransposons with preexisting tandem repeats were found to be a recurrent source of novel satDNAs as well as microsatellite repeats in a wide range of species (Macas et al. 2009; Smýkal et al. 2009). This pathway of repeat-containing TEs acting as seeds for satDNA formation is somewhat expected since replicative transposition provides an immediate path for copy number expansion, with ectopic recombination enabling further amplification of tandem arrays even if transposition stops.

An alternative scenario for the origin of abundant tandem repeats from TEs has been proposed, in which tandem insertions of entire TEs could kickstart satDNA formation (McGurk and Barbash 2018). This model suggests insertion site preference during transposition as the main cause of tandem insertions of TEs, as in the case of the 16 tandem copies of *hobo* found in a population of *D. melanogaster*. According to this hypothesis, TE tandem repeats could be a frequent by-product of transposition, and thus TE activity itself could be seen as a substrate for satDNA emergence (McGurk and Barbash 2018). A related scenario has been explored by us and others involving tandem insertions of *Helitrons*. While *Helitrons* appear to have rather unspecific target sites (5′-A 3′-T) (Kapitonov and Jurka 2007) we have previously detected up to 11 tandem insertions of *DINE-TR1 Helitrons* in *D. virilis* (Dias et al. 2015). We predicted that such arrays could only be formed by rolling-circle replication if *Helitron* transposition involved a double-stranded extrachromosomal circular DNA intermediate (Dias et al. 2016). The reconstruction of *Hellraiser*, a modified *Helitron* from the bat *Myotis lucifugus*, has offered direct evidence for the generation of double-strand DNA circles during *Helitron* mobilization. This provided the first experimental data that could explain *Helitron* tandem insertions in eukaryotic genomes (Grabundzija et al. 2016, 2018). Together, these data indicate that *Helitrons* and other TEs might contribute to abundant tandem repeat and satDNA formation in more than one way in eukaryotes.

2.10 Euchromatic satDNAs: It Depends

Besides satDNAs, other types of non-protein-coding tandemly repeated DNAs are typically found in the genome of eukaryotes, most notably the micro- and minisatellites (Tautz 1993; Charlesworth et al. 1994). However, both are found in smaller arrays (from two repeats up to a few hundred repeats) compared to satDNAs arrays (typically more than 100 repeats) and are dispersed in euchromatic regions. While the repeat sizes of microsatellites are small (1–10 bp), minisatellite repeats are in the range of 10–100 bp and can also be found enriched at subtelomeric regions. This repeat length range for minisatellites was largely compiled from loci used as markers in human individual identification and population genetics, and there is no reason to assume that minisatellite repeat length should be constrained to this size variation across eukaryotes.

The main bulk of repeats from satDNAs resides in heterochromatin, but some homologous repeats might exist in the euchromatin, usually in the form of small arrays (less than 20 repeats). Examples of satDNAs showing this kind of organization include the 1.688 (Kuhn et al. 2012) and the Responder (Larracuente 2014) satellites of *D. melanogaster*. Abundant and dispersed tandem repeats exclusively located in the euchromatin have also been found in *Drosophila*, such as the 175–200 bp repeats of *D. ananassae* (Nozawa et al. 2006), but these are usually present in small arrays compatible with the minisatellite DNA definition. Below, we show examples of euchromatic repeats found in species from the *repleta* and *virilis* and review their classification based on their current known features (Table 2.1).

In the *repleta* group, we found two examples of euchromatic tandem repeats. The first is illustrated by the pBuM-1 satDNA of *D. buzzatii*. As already mentioned, this satellite is the main component of the centromeric heterochromatin in *D. buzzatii* species cluster and there is no doubt pBuM-1 is a typical satDNA. But in addition, we found in the euchromatic assembled genome of *D. buzzatii* four arrays with less than two tandem repeats on chromosomes X, 2, and 4.

The second example of euchromatic tandem repeats in the *repleta* group is the CDTR198 family, with repeat lengths around 198 bp and making up 0.23% and 0.02% of the genomic DNA of *D. buzzatii* and *D. seriema*, respectively. The CDTR198 repeats have been found in a highly dispersed pattern along euchromatin and subtelomeric regions of some chromosomes (de Lima et al. 2017). In the euchromatic assembled genome of *D. buzzatii*, we identified around 150 arrays containing CDTR198 repeats. Based on its abundance, exclusive euchromatic distribution, and low estimated number of repeats per array (less than 20 repeats on average), the CDTR198 shares more features with minisatellite DNAs.

In the *virilis* group**,** we found five examples of euchromatic tandem repeats**.** Short arrays containing the 7 bp satDNAs have been detected in a few euchromatic loci in *D. americana*, *D. texana*, *D. novamexicana* (Cohen and Bowman 1979), and *D. virilis* (Silva et al. 2019). Similar to pBuM-1 in *D. buzzatii*, the 7 bp satDNA repeats are the main components of the centromeric heterochromatin in species from the *virilis* group (Flynn et al. 2020), and there is no doubt they represent typical satDNAs.

A second example is the *pvB370* family (370 bp long repeats), which is present in all species from the *virilis* group and it is associated with the pDv transposon (Heikkinen et al. 1995) (see Sect. 2.7). This abundant tandem repeat localizes in the subtelomeric region of all chromosomes and in many euchromatic loci in *D. virilis*, *D. americana*, *D. novamexicana*, *D. lummei*, and *D. montana*. In

Table 2.1 Main features of the most abundant tandem repeat families found in *Drosophila* species from the *repleta* and *virilis* groups and proposed classification

Tandem repeat family	Aprox. Repeat length	A + T content	Chromosome location	Association with known TEs	Classification
***repleta* group**[a]					
pBuM-1	190 bp	>60%	Centromeric heterochromatin, occasionally at euchromatin	No	Satellite DNA
pBuM-2	370 bp	>60%	Centromeric heterochromatin, occasionally at subtelomeric regions	No	Satellite DNA
DBC-150	150 bp	43–50%	Centromeric heterochromatin	No	Satellite DNA
CDSTR138	138 bp	>60%	Centromeric heterochromatin	No	Satellite DNA
CDSTR130	130 bp	>60%	Centromeric heterochromatin	No	Satellite DNA
CDSTR198	198 bp	>60%	Euchromatin and subtelomeric regions	No	Minisatellite DNA
SSS139	139 bp	>60%	ND	ND	ND
***virilis* group**[b]					
Sat I	7 bp	71%	Centromeric heterochromatin, occasionally at euchromatin	No	Satellite DNA
Sat II	7 bp	86%	Centromeric heterochromatin, occasionally at euchromatin	No	Satellite DNA
Sat III	7 bp	86%	Centromeric heterochromatin, occasionally at euchromatin	No	Satellite DNA
Sat IV	7 bp	71%	Centromeric heterochromatin	No	Satellite DNA
pvB370	370 bp	67%	Euchromatin and subtelomeric regions	Yes (pDv)	Minisatellite DNA
36TR	36 bp	50%	Euchromatin and subtelomeric regions	Yes (pDv)	TE-internal tandem repeat
172TR	172 bp	>61%	Euchromatin and subtelomeric regions	No	Minisatellite DNA
TIR-220	220 bp	~70%	Euchromatin and β-heterochromatin	Yes (Tetris)	TE-internal tandem repeat
154TR	154 bp	69%	Mainly euchromatin and β-heterochromatin, but occasionally at centromeric heterochromatin	Yes (DINE-TR1)	Transitional satDNA

ND = not determined based on available data

Data from: [a]Kuhn et al. (2007, 2008, 2009), Franco et al. (2008), de Lima et al. (2017). [b]Gall et al. (1971, 1974), Gall and Atherton (1974), Cohen and Bowman (1979), Zelentsova et al. (1986), Vashakidze et al. (1989), Heikkinen et al. (1995), Biessmann et al. (2000), Abdurashitov et al. (2013), Dias et al. (2014, 2015), Silva et al. (2019), Flynn et al. (2020)

D. littoralis and *D. ezoana*, only a small number of *pvB370* arrays were detected in the euchromatin, with no evidence for *pvB370* sequences in the subtelomeric region (Biessmann et al. 2000). In *D. virilis* and *D. americana*, *pvB370* covers ~1.7% of their genomes (Silva et al. 2019), and are found in array sizes with up to ~110 copies in their assembled genomes. Although *pvB370* has been originally referred to as a satDNA (Heikkinen et al. 1995; Biessmann et al. 2000), its predominant distribution in euchromatic and subtelomeric loci, organized in likely small arrays, indicates that pvB370 should be regarded as a minisatellite DNA.

The 36TR family (36 bp long repeats) is also associated with the pDv transposon (Fig. 2.8) and was found distributed across ~200 loci in the euchromatin of *D. virilis* (Vashakidze et al. 1989). Recently, we found that 36TR cover 0.7% and 0.4% of the *D. virilis* and *D. americana* genomes, respectively (Silva et al. 2019). In a different study, 36TR was estimated to cover ~0.2% (~800 kb) of the *D. virilis* assembled genome (Flynn et al. 2020). Altogether, the available data indicate that 36TR are distributed in arrays with a few hundred copies. However, there is no current evidence that 36TR exists independently from pDv. Hence, 36TR should be classified as a "TE-internal tandem repeat," with a distribution that reflects the putative mobile nature of pDV sequences.

The fourth example of euchromatic repeats in the *virilis* group is illustrated by the 172TR family (172 bp long repeats) (Abdurashitov et al. 2013; Silva et al. 2019), which is located across multiple euchromatic and subtelomeric regions in all chromosomes from *D. virilis* and *D. americana* (Fig. 2.9; Silva et al. 2019). The 172TR repeats cover ~1% and ~ 4% of *D. virilis* and *D. americana* genomes, respectively (Silva et al. 2019). In *D. virilis*, we found that some 172TR arrays can reach more than 200 copies, with an average of 27 tandem repeats per array. Therefore, 172TR displays several attributes of a minisatellite DNA.

Finally, TIR-220 (220 bp long repeats), which are part of the terminal inverted repeats from Tetris transposons (see Sect. 2.7), are found in a few euchromatic loci from *D. virilis*. However, because there is no evidence of TIR-220 existing independently from Tetris in the euchromatin, and only a few arrays have more than 50 copies (average of 9 copies per array) (Dias et al. 2014), we propose that TIR-220 should be classified as a "TE-internal tandem repeat," similarly to the 36TR case.

The existence of euchromatic arrays containing repeats homologous to satDNAs that are not associated with TEs raises the question about what mechanisms lead to their dispersion. One possibility is that satDNA movement can be mediated by TEs during their transposition process. For example, *Helitrons* are particularly abundant near centromeric regions in *D. buzzatii* (Rius et al. 2016) and we found copies of *Helitrons* in the vicinity of pBuM euchromatic arrays in the same species. Given the fact that during transposition *Helitrons* can capture downstream DNA sequences (reviewed in Thomas and Pritham 2015), it is possible that these pBuM repeats could have been brought from heterochromatin to euchromatin together with the transposition of their neighbor *Helitron* sequences. Another possibility involves the deletion of a few tandem repeats from heterochromatin through ectopic recombination between repeats, leading to the formation of satDNA-containing circular DNAs followed by re-integration in euchromatin through illegitimate recombination

(Walsh 1987). Circular DNAs made by tandem repeats have been reported in *Drosophila* and several organisms (Cohen et al. 2003). Whatever the mechanisms of dispersion are, there is an increasing number of studies reporting euchromatic arrays containing satellite or minisatellite repeats (Pita et al. 2017; Brajković et al. 2018). Some of these euchromatic families are relatively abundant and their distribution closely resembles that of TEs, which are known to affect genome evolution in many ways. In fact, there are data from *Drosophila* and other organisms showing the participation of euchromatic satDNA repeats in gene regulation (Menon et al. 2014; Feliciello et al. 2015; Joshi and Meller 2017), revealing that at least some of these euchromatic arrays may have an important functional role in the genome.

2.11 The Special Case of 154 TR: A Transitional satDNA?

The TE-internal tandem repeats discussed in this chapter are almost exclusively distributed in short- or medium-sized arrays within euchromatic and β-heterochromatic regions, thus not fitting classical definitions for satDNAs. In contrast, 154TR, which is an internal tandem repeat from *DINE-TR1* TE (Fig. 2.8), is found not only dispersed in euchromatic and β-heterochromatic regions, but also in large heterochromatic blocks within some chromosomes from *virilis* group species (Fig. 2.9). For instance, 154TR covers a large portion of *D. virilis* and *D. americana* Y chromosomes, and the centromeric heterochromatin from chromosome 5 in *D. virilis* (Fig. 2.9; Dias et al. 2015). Hence, 154TR shares features from both TE-internal tandem repeats and satDNAs. Considering these unique intermediate features, we propose that 154TR should be classified as a "transitional satDNA." In this case, transitional satDNAs would encompass tandem repeats that, despite being part of TEs and displaying features of TE-internal tandem repeats, are also found as large expanded arrays within constitutive heterochromatic loci, similarly to classical satDNAs.

Acknowledgments The authors would like to thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) (fellowship Capes-Print/UFMG Proc. 88887.371254/2019-00 to GK; PhD fellowship to PH), Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq (Proc. 308386/2018-3 to GK) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais FAPEMIG for funding support.

References

Abdurashitov MA, Gonchar DA, Chernukhin VA et al (2013) Medium-sized tandem repeats represent an abundant component of the *Drosophila virilis* genome. BMC Genomics 14:771

Allshire RC, Madhani HD (2018) Ten principles of heterochromatin formation and function. Nat Rev Mol Cell Biol 19:229–244

Bachmann L, Sperlich D (1993) Gradual evolution of a specific satellite DNA family in *Drosophila ambigua*, *D. tristis*, and *D. obscura*. Mol Biol Evol 10:647–659

Bachmann L, Raab M, Sperlich D (1989) Satellite DNA and speciation: a species specific satellite DNA of *Drosophila guanche*. J Zool Syst Evol Res 27:84–93

Berloco M, Palumbo G, Piacentini L et al (2014) Position effect variegation and viability are both sensitive to dosage of constitutive heterochromatin in *Drosophila*. G3 Genes Genomes Genetics 4:1709–1716

Biessmann H, Zurovcova M, Yao JG et al (2000) A telomeric satellite in *Drosophila virilis* and its sibling species. Chromosoma 109:372–380

Bosco G, Campbell P, Leiva-Neto JT, Markow TA (2007) Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177:1277–1290

Brajković J, Pezer Ž, Bruvo-Mađarić B et al (2018) Dispersion profiles and gene associations of repetitive DNAs in the euchromatin of the beetle *Tribolium castaneum*. G3 Genes Genomes Genetics 8:875–886

Brennecke J, Aravin AA, Stark A et al (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell 128:1089–1103

Caletka BC, McAllister BF (2004) A genealogical view of chromosomal evolution and species delimitation in the *Drosophila virilis* species subgroup. Mol Phylogenet Evol 33:664–670

Chang C-H, Chavan A, Palladino J et al (2019) Islands of retroelements are major components of *Drosophila* centromeres. PLoS Biol 17:e3000241

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

Cohen EH, Bowman SC (1979) Detection and location of three simple sequence DNAs in polytene chromosomes from *virilis* group species of *Drosophila*. Chromosoma 73:327–355

Cohen S, Yacobi K, Segal D (2003) Extrachromosomal circular DNA of tandemly repeated genomic sequences in *Drosophila*. Genome Res 13:1133–1145

Craddock EM, Gall JG, Jonas M (2016) Hawaiian *Drosophila* genomes: size variation and evolutionary expansions. Genetica 144:107–124

de Lima LG, Svartman M, Kuhn GCS (2017) Dissecting the satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes. G3 Genes Genomes Genetics 7:2831–2843

Dias GB, Svartman M, Delprat A et al (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

Dias GB, Heringer P, Svartman M, Kuhn GCS (2015) Helitrons shaping the genomic architecture of *Drosophila*: enrichment of DINE-TR1 in α- and β-heterochromatin, satellite DNA emergence, and piRNA expression. Chromosom Res 23:597–613

Dias GB, Heringer P, Kuhn GCS (2016) Helitrons in *Drosophila*: chromatin modulation and tandem insertions. Mob Genet Elements 6:e1154638

Dimitri P, Pisano C (1989) Position effect variegation in *Drosophila melanogaster*: relationship between suppression effect and the amount of Y chromosome. Genetics 122:793–800

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Elliott TA, Gregory TR (2015) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc B 370. https://doi.org/10.1098/rstb.2014.0331

Feliciello I, Akrap I, Ugarković Đ (2015) Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. PLoS Genet 11:e1005466

Ferree PM, Barbash DA (2009) Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. PLoS Biol 7:e1000234

Flynn JM, Long M, Wing RA, Clark AG (2020) Evolutionary dynamics of abundant 7-bp satellites in the genome of *Drosophila virilis*. Mol Biol Evol 37:1362–1375

Francisco FO, Lemos B (2014) How do Y-chromosomes modulate genome-wide epigenetic states: genome folding, chromatin sinks, and gene expression. J Genomics 2:94–103

Franco FF, Sene FM, Manfrin MH (2008) Molecular characterization of SSS139, a new satellite DNA family in sibling species of the *Drosophila buzzatii* cluster. Genet Mol Biol 31:155–159

Franco FF, Silva-Bernardi ECC, Sene FM et al (2010) Intra- and interspecific divergence in the nuclear sequences of the clock gene period in species of the *Drosophila buzzatii* cluster. J Zool Syst Evol Res 48:322–331

Gall JG, Atherton DD (1974) Satellite DNA sequences in *Drosophila virilis*. J Mol Biol 85:633–664

Gall JG, Cohen EH, Polan ML (1971) Repetitive DNA sequences in *Drosophila*. Chromosoma 33:319–344

Gall JG, Cohen EH, Atherton DD (1974) The satellite DNAs of *Drosophila virilis*. Cold Spring Harb Symp Quant Biol 38:417–421

Ganley ARD, Kobayashi T (2007) Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res 17:184–191

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

Goebel J, Promerová M, Bonadonna F et al (2017) 100 million years of multigene family evolution: origin and evolution of the avian MHC class IIB. BMC Genomics 18:460

Grabundzija I, Messing SA, Thomas J et al (2016) A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat Commun 7:10716

Grabundzija I, Hickman AB, Dyda F (2018) Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition. Nat Commun 9:1278

Gregory TR, Johnston JS (2008) Genome size diversity in the family Drosophilidae. Heredity 101:228–238

Guillén Y, Rius N, Delprat A et al (2015) Genomics of ecological adaptation in cactophilic *Drosophila*. Genome Biol Evol 7:349–366

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

Heikkinen E, Launonen V, Müller E, Bachmann L (1995) The pvB370 BamHI satellite DNA family of the *Drosophila virilis* group and its evolutionary relation to mobile dispersed genetic pDv elements. J Mol Evol 41:604–614

Heitz E (1934) Über α-und β-heterochromatin sowie konstanz und bau der chromomeren bei *Drosophila*. Biol Zbl 54:588–609

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

Jagannathan M, Cummings R, Yamashita YM (2018) A conserved function for pericentromeric satellite DNA. eLife 7:e34122

Jagannathan M, Cummings R, Yamashita YM (2019) The modular mechanism of chromocenter formation in *Drosophila*. eLife 8:e43938

Joshi SS, Meller VH (2017) Satellite repeats identify X chromatin for dosage compensation in *Drosophila melanogaster* males. Curr Biol 27:1393–1402

Kalitsis P, Choo KHA (2012) The evolutionary life cycle of the resilient centromere. Chromosoma 121:327–340

Kapitonov VV, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 23:521–529

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

Khost DE, Eickbush DG, Larracuente AM (2017) Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome Res 27:709–721

King DG, Soller M, Kashi Y (1997) Evolutionary tuning knobs. Endeavour 21:36–40

Kuhn GCS, Heslop-Harrison JS (2011) Characterization and genomic organization of PERI, a repetitive DNA in the *Drosophila buzzatii* cluster related to DINE-1 transposable elements and highly abundant in the sex chromosomes. Cytogenet Genome Res 132:79–88

Kuhn GCS, Sene FM (2005) Evolutionary turnover of two pBuM satellite DNA subfamilies in the *Drosophila buzzatii* species cluster (*repleta* group): from alpha to alpha/beta arrays. Gene 349:77–85

Kuhn GCS, Ruiz A, Alves MAR, Sene FM (1996) The metaphase and polytene chromosomes of *Drosophila seriema* (*repleta* group; *mulleri* subgroup). Brazil J Genet 19:209–216

Kuhn GCS, Bollgonn S, Sperlich D, Bachmann L (1999) Characterization of a species-specific satellite DNA of *Drosophila buzzatii*. J Zool Syst Evol Res 37:109–112

Kuhn GCS, Franco FF, Silva WA Jr et al (2003) On the pBuM189 satellite DNA variability among South American populations of *Drosophila buzzatii*. Hereditas 139:161–166

Kuhn GCS, Franco FF, Manfrin MH et al (2007) Low rates of homogenization of the DBC-150 satellite DNA family restricted to a single pair of microchromosomes in species from the *Drosophila buzzatii* cluster. Chromosom Res 15:457–469

Kuhn GCS, Sene FM, Moreira-Filho O et al (2008) Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. Chromosom Res 16:307–324

Kuhn GCS, Teo CH, Schwarzacher T, Heslop-Harrison JS (2009) Evolutionary dynamics and sites of illegitimate recombination revealed in the interspersion and sequence junctions of two nonhomologous satellite DNAs in cactophilic *Drosophila* species. Heredity 102:453–464

Kuhn GCS, Küttler H, Moreira-Filho O, Heslop-Harrison JS (2012) The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. Mol Biol Evol 29:7–11

Kursel LE, Malik HS (2017) Recurrent gene duplication leads to diverse repertoires of centromeric histones in *Drosophila* species. Mol Biol Evol 34:1445–1462

Larracuente AM (2014) The organization and evolution of the responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. BMC Evol Biol 14:233

Lee YCG (2015) The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in *Drosophila melanogaster*. PLoS Genet 11:e1005269

Locke J, Howard LT, Aippersbach N et al (1999) The characterization of DINE-1, a short, interspersed repetitive element present on chromosome and in the centric heterochromatin of *Drosophila melanogaster*. Chromosoma 108:356–366

Lohe AR, Hilliker AJ, Roberts PA (1993) Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. Genetics 134:1149–1174

López-Flores I, Garrido-Ramos MA (2012) The repetitive DNA content of eukaryotic genomes. In: Garrido-Ramos MA (ed) Repetitive DNA. Karger, Basel, pp 1–28

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

Mahan JT, Beck ML (1986) Heterochromatin in mitotic chromosomes of the *virilis* species group of *Drosophila*. Genetica 68:113–118

Malik HS (2009) The centromere-drive hypothesis: a simple basis for centromere complexity. In: Ugarkovic D (ed) Centromere. Springer, Berlin, Heidelberg, pp 33–52

Marin I, Labrador M, Fontdevila A (1992) The evolutionary history of *Drosophila buzzatii*. XXIII High content of nonsatellite repetitive DNA in *D buzzatii* and in its sibling *D koepferae*. Genome 35:967–974

McGurk MP, Barbash DA (2018) Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. Genome Res 28:714–725

Melters DP, Bradnam KR, Young HA et al (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10

Menon DU, Coarfa C, Xiao W et al (2014) siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. Proc Natl Acad Sci USA 111:16460–16465

Meštrović N, Mravinac B, Pavlek M et al (2015) Structural and functional liaisons between transposable elements and satellite DNAs. Chromosom Res 23:583–596

Miklos GL, Yamamoto MT, Davies J, Pirrotta V (1988) Microcloning reveals a high frequency of repetitive sequences characteristic of chromosome 4 and the beta-heterochromatin of *Drosophila melanogaster*. Proc Natl Acad Sci USA 85:2051–2055

Mills WK, Lee YCG, Kochendoerfer AM et al (2019) RNA from a simple-tandem repeat is required for sperm maturation and male fertility in *Drosophila melanogaster*. eLife 8:e48940

Morales-Hojas R, Reis M, Vieira CP, Vieira J (2011) Resolving the phylogenetic relationships and evolutionary history of the *Drosophila virilis* group using multilocus data. Mol Phylogenet Evol 60:249–258

Moreyra NN, Mensch J, Hurtado J et al (2019) What does mitogenomics tell us about the evolutionary history of the *Drosophila buzzatii* cluster (*repleta* group)? PLoS One 14:e0220676

Nergadze SG, Piras FM, Gamba R et al (2018) Birth, evolution, and transmission of satellite-free mammalian centromeric domains. Genome Res 28:789–799

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

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

Nozawa M, Kumagai M, Aotsuka T, Tamura K (2006) Unusual evolution of interspersed repeat sequences in the *Drosophila ananassae* subgroup. Mol Biol Evol 23:981–987

O'Grady PM, DeSalle R (2018) Phylogeny of the genus *Drosophila*. Genetics 209:1–25

Oliveira DCSG, Almeida FC, O'Grady PM et al (2012) Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila repleta* species group. Mol Phylogenet Evol 64:533–544

Palomeque T, Lorite P (2008) Satellite DNA in insects: a review. Heredity 100:564–573

Pathak RU, Mamillapalli A, Rangaraj N et al (2013) AAGAG repeat RNA is an essential component of nuclear matrix in *Drosophila*. RNA Biol 10:564–571

Pita S, Panzera F, Mora P et al (2017 Jul 19) Comparative repeatome analysis on Triatoma infestans Andean and non-Andean lineages, main vector of Chagas disease. PLoS One 12(7):e0181635. https://doi.org/10.1371/journal.pone.0181635

Plohl M, Luchetti A, Mestrović N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)-chromatin. Gene 409:72–82

Plohl M, Meštrović N, Mravinac B (2012) Satellite DNA evolution. In: Garrido-Ramos MA (ed) Repetitive DNA. Karger, Basel, pp 126–152

Rius N, Guillén Y, Delprat A et al (2016) Exploration of the *Drosophila buzzatii* transposable element content suggests underestimation of repeats in *Drosophila* genomes. BMC Genomics 17:344

Rošić S, Köhler F, Erhardt S (2014) Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. J Cell Biol 207:335–349

Schueler MG, Higgins AW, Rudd MK et al (2001) Genomic and genetic definition of a functional human centromere. Science 294:109–115

Silva BSML, Heringer P, Dias GB et al (2019) De novo identification of satellite DNAs in the sequenced genomes of *Drosophila virilis* and *D. americana* using the RepeatExplorer and TAREAN pipelines. PLoS One 14:e0223466

Smýkal P, Kalendar R, Ford R et al (2009) Evolutionary conserved lineage of Angela-family retrotransposons as a genome-wide microsatellite repeat dispersal agent. Heredity 103:157–167

Spicer GS, Bell CD (2002) Molecular phylogeny of the *Drosophila virilis* species group (Diptera: Drosophilidae) inferred from mitochondrial 12S and 16S ribosomal RNA genes. Ann Entomol Soc Am 95:156–161

Sproul JS, Khost DE, Eickbush DG et al (2020) Dynamic evolution of euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the *simulans* clade. Mol Biol Evol 37:2241–2256

Strachan T, Webb D, Dover GA (1985) Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. EMBO J 4:1701–1708

Talbert PB, Henikoff S (2020) What makes a centromere? Exp Cell Res 389:111895

Tautz D (1993) Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In: Pena SDJ, Chakraborty R, Epplen JT, Jeffreys AJ (eds) DNA fingerprinting: state of the science. Birkhäuser, Basel, pp 21–28

Teixeira JR, Dias GB, Svartman M, Ruiz A, Kuhn GCS (2018) Concurrent duplication of *Drosophila* Cid and Cenp-C genes resulted in accelerated evolution and male germline-biased expression of the new copies. J Mol Evol 86:353–364

Thomas J, Pritham EJ (2015) Helitrons, the eukaryotic rolling-circle transposable elements. Mobile DNA III 2015:891–924

Thomas AL, Le Thomas A, Rogers AK et al (2013) Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev 27:390–399

Thomas J, Vadnagara K, Pritham EJ (2014) DINE-1, the highest copy number repeats in *Drosophila melanogaster* are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). Mob DNA 5:18

Throckmorton LH (1975) The phylogeny, ecology, and geography of *Drosophila*. In: King RC (ed) Handbook of genetics, vol 3. Plenum Publishing Corporation, New York, pp 421–469

Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36–46

Vashakidze R, Zelentsova H, Korochkin L, Evgen'ev M (1989) Expression of dispersed 36 bp sequences in *Drosophila virilis*. Chromosoma 97:374–380

Vaury C, Bucheton A, Pelisson A (1989) The β heterochromatic sequences flanking the I elements are themselves defective transposable elements. Chromosoma 98:215–224

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

Wei KH-C, Grenier JK, Barbash DA, Clark AG (2014) Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. Proc Natl Acad Sci USA 111:18793–18798

Wensink PC, Tabata S, Pachl C (1979) The clustered and scrambled arrangement of moderately repetitive elements in *Drosophila* DNA. Cell 18:1231–1246

Yang H-P, Barbash DA (2008) Abundant and species-specific DINE-1 transposable elements in 12 *Drosophila* genomes. Genome Biol 9:R39

Zelentsova ES, Vashakidze RP, Krayev AS, Evgen'ev MB (1986) Dispersed repeats in *Drosophila virilis*: elements mobilized by interspecific hybridization. Chromosoma 93:469–476

Chapter 3
Exploring Satellite DNAs: Specificities of Bivalve Mollusks Genomes

Eva Šatović Vukšić and Miroslav Plohl

Abstract Noncoding DNA sequences repeated in tandem or satellite DNAs make an integral part of every eukaryotic genome. Development and application of new methodological approaches through time enabled gradual improvement in understanding of structural and functional roles of these sequences, early misconsidered as "junk DNA". Advancing approaches started adding novel insights into details of their existence on the genomic scale, traditionally hard to access due to difficulties in analyzing long arrays of nearly identical tandem repeats of a satellite DNA. In turn, broadened views opened space for the development of new concepts on satellite DNA biology, highlighting also specificities coming from different groups of organisms. Observed diversities in different aspects and in organizational forms of these sequences proclaimed a need for a versatile pool of model organisms. Peculiarities of satellite DNAs populating genomes of bivalve mollusks, an important group of marine and fresh-water organisms, add to the diversity of organizational principles and associated roles in which tandemly repeated sequences contribute to the genomes.

Keywords Satellite DNA · Mobile elements · Satellitome · Heterochromatin · Genome evolution · Bivalve mollusks

3.1 Introduction

The ubiquitous and at the same time still the least understood DNA components of every eukaryotic genome are repetitive DNA sequences. The totality of repetitive DNAs in a genome, the repeatome, defines and is responsible for significant variations in the genome size among species, regardless of the genome complexity. Repetitive DNA sequences are traditionally subdivided into two groups, one

E. Šatović Vukšić · M. Plohl (✉)
Ruđer Bošković Institute, Zagreb, Croatia
e-mail: Plohl@irb.hr

Đ. Ugarković (ed.), *Satellite DNAs in Physiology and Evolution*, Progress in Molecular and Subcellular Biology 60,
https://doi.org/10.1007/978-3-030-74889-0_3

composed of arrays formed by sequences repeated in tandem, and the other consisting of repeats distributed in the genome in an interspersed manner, as a result of transposition processes (Charlesworth et al. 1994; Schmidt and Heslop-Harrison 1998; Jurka et al. 2007; López-Flores and Garrido-Ramos 2012; Biscotti et al. 2015).

Noncoding sequences repeated in tandem, satellite DNAs (satDNAs), are predominantly associated with tightly packed heterochromatic chromosomal regions. Except for this feature, and the capability to build megabase-long arrays of head-to-tail repeated satDNA monomers, they represent an extremely heterogeneous group of sequences (Charlesworth et al. 1994; Ugarković and Plohl 2002; Plohl et al. 2008, 2012; Garrido-Ramos 2017; Hartley and O'Neill 2019).

Sequences repeated in tandem evolve in a nonindependent manner, known as concerted evolution. In this process, divergences among monomers in arrays of a satDNA are homogenized by mechanisms of nonreciprocal sequence transfer (unequal crossover, for example), keeping monomer sequence variability within the genome low, and fixed at the species level (usually <5%). This mode of evolution is called molecular drive and assumes rapid divergent evolution of a satDNA sequence in reproductively isolated organisms (Elder and Turner 1995; Dover 1982, 1986; Plohl et al. 2008, 2012). Despite that, some satDNA families can remain preserved in diverged taxa through unexpectedly long evolutionary periods, probably because of non-stochastic preferences in the process of concerted evolution and/or putative constraints on the satDNA monomer sequence (Strachan et al. 1985; Plohl et al. 2012). Many satDNA subfamilies or unrelated families exist in a genome, and they differ strikingly in genomic abundance (from <0.5% to >50%), which is also subject to rapid alterations in short evolutionary periods (Ugarković and Plohl 2002). In this regard, even if nucleotide sequences of satDNAs are preserved, changes in copy number are sufficient to define species-specific profiles, even between very closely related genomes (Fry and Salser 1977; Meštrović et al. 1998; Ugarković and Plohl 2002).

Recent methodological advances changed dramatically views on satDNAs and brought into the focus their essential roles, such as in functional architecture and evolution of chromosomes, chromatin modulation, reproductive isolation, genome stability, and evolution (Henikoff et al. 2001; Slamovits and Rossi 2002; Pezer and Ugarković 2008; Adega et al. 2009; Ferree and Barbash 2009; Garrido-Ramos 2017; Lower et al. 2018; Louzada et al. 2020; Shatskikh et al. 2020). SatDNAs represent the most common form of DNA sequences in functional centromeres, where they associate with the centromere protein determinant, histone variant CenH3 (Henikoff et al. 2001; Plohl et al. 2014; Garrido-Ramos 2017; Hartley and O'Neill 2019; Talbert and Henikoff 2020). In addition, significant roles in chromatin organization and gene expression have transcripts of satDNAs, exampled in insects (Pezer et al. 2011; Feliciello et al. 2015). Data are also accumulating showing that disturbances in some of satDNA features are associated with diseases, including cancer (Miga 2019; Louzada et al. 2020).

Because satDNAs are, together with transposable elements (TEs), considered to be the main determinants of genome architecture and drivers of its evolution, it is important to understand different patterns of organization, mutual links, and functional roles of repetitive sequences in different groups of species, distinct in the biology of repetitive DNAs. In the first part of this review, we summarize approaches that have built conceptual views and added new levels to our understanding of the biology of satDNAs. In continuation, the topic of satDNA outside of heterochromatin and their association to mobile elements is attended. Following these aspects, in the third part we present the current knowledge about satDNAs and heterochromatin in bivalve mollusks, the group of species with rapidly accumulating genome data, and with certain peculiarities in abundance, ancestry, connection to TEs, conserved sequence boxes, methylation patterns and evolutionary aspects of satDNAs that even bring into question the classical form of the "library model" within this group of organisms.

3.2 Chronology of Key Advancements in satDNA Research

3.2.1 Detection and Characterization of satDNAs in the Pre-genomic Era

The term "satellite DNA" has been coined about 60 years ago. It was originally used to describe DNA fraction contained in the additional band, separated from the bulk of mouse genomic DNA in experiments of density gradient centrifugation because of differences in nucleotide composition (Kit 1961; Sueoka 1961; Szybalski 1968). Such fractions turned out to be enriched in highly repetitive DNAs (Waring and Britten 1966). Since then, the name "satellite DNA" has been commonly used for all noncoding DNA sequences repeated in tandem. SatDNAs were localized in situ as dominant DNA components of constitutive heterochromatin (Pardue and Gall 1970), known to accumulate around centromeres and telomeres, structures indispensable for division and stability of chromosomes. Based on these observations, some early works anticipated the structural importance of repetitive sequences at the chromosomal and the nuclear organization level, including in speciation, supporting the proposed roles of heterochromatin (Yunis and Yasmineh 1971). Nevertheless, an opposing view presented satDNAs (as well as other repetitive sequences) as useless genomic ballast, accumulated just because of DNA sequence dynamics in heterochromatic regions and tolerated until overloading (Ohno 1972; Orgel and Crick 1980). This opinion had been based on the monotony of highly similar monomers repeated one after the other, lack of the coding capacity, and, as it was thought in that time, of transcription, as well as by rapid evolution resulting in a high diversity of satDNAs among species.

Breakthroughs in methodologies and introduction of new strategies suitable for satDNA research (Fig. 3.1) generated results that gradually broadened our views on this class of genomic sequences and significantly altered general perception of (Csink and Henikoff 1998; Schmidt and Heslop-Harrison 1998; Garrido-Ramos 2017; Louzada et al. 2020). Briefly, after the initial gradient centrifugation era, a

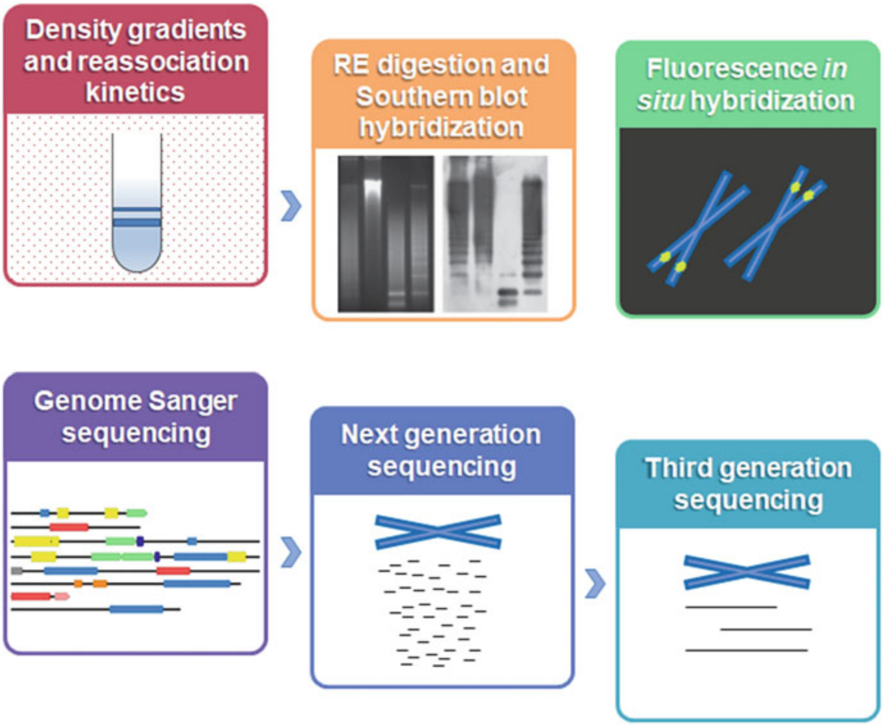

Density gradients and reassociation kinetics
RE digestion and Southern blot hybridization
Fluorescence *in situ* hybridization
Genome Sanger sequencing
Next generation sequencing
Third generation sequencing

Fig. 3.1 Advancements in methodologies and strategies employed in satDNA research throughout time. Density gradient centrifugation that enabled the initial satDNA detection was followed by restriction enzyme digestions and Southern blot-based methods, complemented with fluorescence in situ hybridization localization on chromosomes (upper panel). Sanger sequencing of genomic fragments started revealing close proximity of different types of repetitive sequences and enabled studying transition patterns among them while next generation and third-generation sequencing enabled complete satellitome analyses and detailed sequential order of repetitive sequences in large genomic segments (lower panel)

large number of studies were done using electrophoretic separation of genomic DNA fragments obtained after the restriction endonuclease digestion. The method is based on the low sequence variability in the satDNA family: if appropriate restriction endonuclease is used, tandemly repeated satDNA monomers form a characteristic ladder pattern on the gel. A sample of satDNA monomers and short multimeres could be subsequently cloned and sequenced and relatively easily mapped on chromosomes after the introduction of fluorescence in situ hybridization (FISH) methodology (Singer 1982; Garrido-Ramos 2017). Although this strategy enabled efficient detection and analysis, satDNAs remained limited to one or a few sequence families per genome, discovered if the appropriate endonuclease could be selected and if repeats were abundant enough to be detected on the gel. In an alternative approach, satDNAs (as well as other repetitive sequences) can be identified by analyzing clones of interest from the library of cloned genomic fragments selected after the colony-lift hybridization with labeled fragmented total genomic DNA. In this way, the strongest signals give clones bearing fragments that contain the most abundant repetitive sequences (Sainz et al. 1992; Biscotti et al. 2007). Although usually short (<1 kb), such fragments can be hybrids of more than one sequence type and can be of particular interest in studying transition patterns among them (Šatović and Plohl 2013; Šatović et al. 2016).

Despite limitations, these studies forwarded significantly our understanding of the diversity of satDNA families, their patterns of evolution, repeat unit organization, life-cycle, and genomic distribution (Dover 1986; Willard and Waye 1987; Plohl et al. 2012; Garrido-Ramos 2017). An important milestone in addressing questions about the total number of satDNA families in a genome and their possible intergenomic distribution was raised by the idea about satDNA library (Fry and Salser 1977). This hypothesis proposes that closely related species share a common collection of satDNA families. SatDNAs in the library differ in each species in abundance because of extensive expansions and contractions of arrays, which alters dramatically the profile of these sequences in the particular genome. In this regard, the most dominant satDNA family (or families) can be falsely considered "species specific," just because their low-copy variants remained undetected in related taxa. Introduction of PCR methodology enabled precise identification of these low-copies and proved the library hypothesis (Meštrović et al. 1998; Ugarković and Plohl 2002; Plohl et al. 2008; Garrido-Ramos 2017).

3.2.2 SatDNA Studies in the Genomic Era

Employment of Sanger sequencing on genomic fragments started bringing the information on the close proximity of different types of repetitive sequences and the complex web formed thanks to their vicinity (Fig. 3.1). However, in the era of sequenced genomes, assemblies based on Sanger sequencing are generally smaller than the estimated genome size, with gaps occurring mostly in segments enriched in repetitive DNAs (Miga 2015; Peona et al. 2018; Tørresen et al. 2019). The main reasons are difficulties in overlapping nearly identical repeats in long arrays of satDNAs, due to which they remain miss-presented or left out from genome outputs, appearing mostly in unplaced scaffolds and singletons. For example, one abundant subfamily of pericentromerically located satDNA detected by restriction endonuclease was estimated to build 17% of the beetle *Tribolium castaneum* (Ugarković et al. 1996) but only about 0.3% of the genome assembly (Wang et al. 2008). An upgraded assembly of *T. castaneum* genome retained unmapped gaps of ~20% of the estimated genome size, primarily in the heterochromatic and centromeric areas (Herndon et al. 2020). Even in the genomes with low levels of heterochromatin and satDNAs, such as the Pacific oyster *Crassostrea gigas,* the situation is similar. There, a complex approach combining fosmid pooling, next-generation sequencing (NGS), and hierarchical assembly (Zhang et al. 2012) also could not assemble arrays of its most abundant satDNA family, detected experimentally to populate 1–4% of the genome (Clabby et al. 1996; Wang et al. 2001).

Implementation of NGS methodologies accompanied by the development of specialized bioinformatics tools provide a powerful strategy for comprehensive analysis of repetitive DNA content on the genome-wide scale (Fig. 3.1). For example, a widely used RepeatExplorer computational platform (Novák et al. 2013) detects repetitive sequences by clustering highly similar short-read unassembled genomic datasets representing low genome coverage (up to 0.5x), to reduce the "noise" of single-copy genomic segments. This approach enables the determination of consensus sequences of repetitive DNA families in any species, without the need for the reference genome or for the reference database of repetitive sequences. Specially focused on satDNAs is Tandem Repeat Analyzer (TAREAN; Novák et al. 2017), implemented in the upgraded RepeatExplorer2 protocol. To facilitate the detection of low-copy satDNA families in large genomes, already detected satDNAs can be filtered out in each cycle of the repeated clustering procedure (Ruiz-Ruano et al. 2016). Other approaches have been also developed, such as repeatConnector that screens NGS data for specific satDNAs in different species (Smalec et al. 2019). Overall, the number of available programs and the program improvements is increasing constantly (Garrido-Ramos 2017; Lower et al. 2018; Smalec et al. 2019; Šatović et al. 2020).

An important outcome of NGS-based bioinformatics is a complete (or almost complete) inventory of repetitive DNA sequences in the genome, the repeatome (Kim et al. 2014), or if only satDNAs are considered, the satellitome (Ruiz-Ruano et al. 2016). These analyses also highlighted enormous diversities in organizational patterns and distribution of repetitive sequences on chromosomes and in genomes. For instance, the content of repetitive sequences can be shifted in favor of some particular groups, as in the repeatome of the common oat. There, repetitive DNAs build ~70% of the genome, with the dominance of a relatively small number of retroelement families, while satDNA families compose only the modest 2% of genomic DNA (Liu et al. 2019).

Combined with experimental methods, especially FISH mapping, these pipelines provoked a large number of studies and a burst of new information about content, structure, evolution, and chromosomal distribution of repetitive sequences, within and between genomes, for instance: Macas et al. (2007, 2015), Klemme et al. (2013), Palacios-Gimenez et al. (2017), Utsunomia et al. (2019), Belyayev et al. (2019). An earlier assumption about a large number of satDNAs populating the genome has been confirmed, for example, a collection of 62 satDNA families was revealed in the migratory locust (Ruiz-Ruano et al. 2016), 129 satDNAs in the morabine grasshoppers (Palacios-Gimenez et al. 2020b) and 164 satDNA families were characterized in the fish *Megaleporinus microcephalus* (Utsunomia et al. 2019).

Interspecies comparisons of satellitomes enabled detailed characterization of shared sets, and confirmed postulates of the satDNA library, explained above, on a genome-wide scale (Macas et al. 2015; Ruiz-Ruano et al. 2016; de Silva et al. 2017; Utsunomia et al. 2017; Pita et al. 2017; Palacios-Gimenez et al. 2018, 2020a, b). In this regard, satDNAs forming the library shared by related species can be equivalent to the concept of interspecifically comparable satellitomes. However, the content of the satDNA library and that of the satellitome may not be identical, because the satellitome may also incorporate species-specific satDNAs, not distributed in other species. In other words, the library of satDNAs is a subset of sequences detected in the satellitome.

3.2.3 SatDNAs and the Third-Generation Sequencing

Despite recent significant advances in understanding repetitive DNA genomics, our comprehension of the detailed sequential order of repetitive sequences within large genomic segments continued to be an elusive goal if addressed by Sanger sequencing and/or NGS (Alkan et al. 2011). To solve this problem, an important step forward in accurate reading of the long segments composed of satDNAs is brought by the third-generation sequencing that uses a single-molecule long-read methodology, such as PacBio and Oxford Nanopore Technologies, combined with new mapping protocols and bioinformatics tools (Van Dijk et al. 2018; Sedlazeck et al. 2018). These technologies are able to produce continuous reads between 10 and 100 kb long, and the longest could be over 1 Mb. Assembly of long reads can therefore enable filling the gaps that were left in earlier genome outputs because of the problems caused by repetitive-rich genome segments, as explained above. Finally, incorporating the third-generation methodology into sequencing projects enable high-quality end-to-end assembly of human chromosomes, for example: Y (Kuderna et al. 2019), X (Miga et al. 2020), and 8 (Logsdon et al. 2021). A combination of long and short reads was also used for de novo sequencing and assembly of mollusk genomes, such as of a variety of *Crassostrea gigas*, the black-shelled oyster (Wang et al. 2019), and to obtain the referent chromosome-level genome assembly of the Pacific oyster *C. gigas* (Peñaloza et al. 2021). In addition to the sequencing projects, this methodology can be specifically used to focus on satDNA loci (Khost et al. 2017), or on the detailed composition of repeat-rich centromeres (Jain et al. 2018; Chang et al. 2019). If a high-quality reference genome is available, long sequence reads of satDNAs can be used to study individual array length variations which, for instance, can be linked with some pathogenic states in humans (Mitsuhashi et al. 2019). In addition, general characterization of satDNAs in the genome oriented at array length and their surrounding can be revealed by statistical analysis of individual nanopore reads obtained at low sequence coverage and without the need for the reference genome (Vondrak et al. 2020).

3.3 SatDNA Outside of the Heterochromatin and Their Association to Mobile Elements

Thanks to the rapid advancements in bioinformatics tools, availability of genomic datasets and genome assemblies, tandemly repeated noncoding DNA sequences that are not exclusively associated with heterochromatin started coming into the focus. Sequences repeated in tandem and related to the classical heterochromatin-associated satDNAs can be dispersed along the chromosomal arms in different forms and abundancies (Fig. 3.2). Several organizational patterns are distinctive, although the present knowledge is based on analysis of a small number of species, mostly insects. Classical heterochromatin-associated satDNAs can exist in euchromatin as short arrays, isolated monomers or monomer fragments, located also near the coding regions (Paar et al. 2011; Brajković et al. 2012, 2018; Kuhn et al. 2012; Ruiz-Ruano et al. 2016; de Lima et al. 2017; Chaves et al. 2017; Sproul et al. 2020). The opposite example makes relatively long arrays similar to typical satDNAs but located exclusively or almost exclusively in euchromatin (Pavlek et al. 2015; Pita et al. 2017).

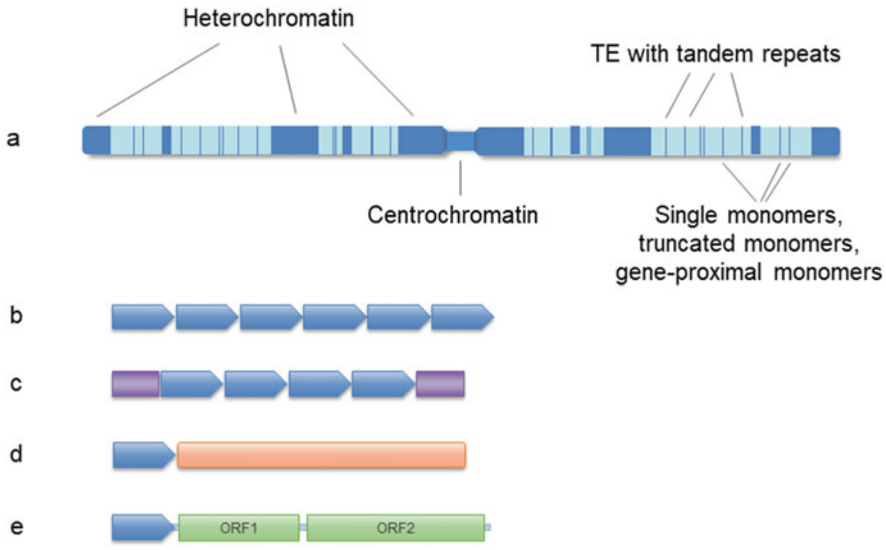


Fig. 3.2 Chromosomal locations and different structural forms built from and/or occupied by satDNA sequences. (**a**) SatDNAs are predominantly localized in heterochromatic chromosomal regions and are the most frequent DNA sequences underlying the centromeres. In addition, these sequences can be found in the interspersed forms of single monomers, truncated monomers, and short arrays also along chromosome arms. (**b**) One of the most common organizational forms of satDNA, long array of monomers repeated in tandem. (**c**) Short arrays of tandem repeats are often incorporated into the central part of mobile elements as their structural component. (**d**) SatDNA monomers can frequently be found in the close proximity of other repetitive sequences (TE or other satDNAs). (**e**) SatDNA monomers in gene-proximal regions

Dynamic evolution of euchromatic satDNA segments defines the landscape of chromosomal regions, as in *Drosophila melanogaster* where they affect the process of a meiotic drive (Larracuente 2014). Dispersed euchromatic copies of a dominant pericentromeric satDNA in the beetle *T. castaneum* exert their effect on the whole-genome scale by initializing chromatin condensation under heat-shock conditions and silencing the expression of nearby genes (Feliciello et al. 2015). Relatively long arrays of *T. castaneum* euchromatin-only satDNAs show features typical for the evolution of a classical (heterochromatic) satDNA, such as low variability of monomers within chromosome-specific arrays, although, at the same time, the most dispersed of them indicate putative links with mobile elements (Pavlek et al. 2015). In this regard, detailed studies of euchromatic satDNAs on the X chromosome of *D. melanogaster* showed similarities with the expansion and diversification of mobile elements in their evolution (Sproul et al. 2020).

There are many ways in which TEs and classical satDNAs are interlinked (reviewed in Meštrović et al. 2015). For instance, satDNA arrays can be interrupted with TEs (Palomeque et al. 2006) or satDNA repeats can be formed by tandem amplification of a TE or its part (Macas et al. 2009; Sharma et al. 2013). Short satDNA-like arrays of tandem repeats are often incorporated into mobile elements as their central structural component (Fig. 3.2), and the same repeats can also appear as builders of classical satDNAs (for example, Gaffney et al. 2003; Dias et al. 2014, 2015; Luchetti 2015). Frequently mentioned in that context are Helitron/Helentron mobile elements, known to be holding arrays of tandem repeats and using a rolling circle mechanism for their propagation (Thomas and Pritham 2015). In that respect, it is not surprising that, with the progress of the sequencing techniques, short forms of satellite DNA arrays are being detected outside of the heterochromatin more and more, existing in different organizational forms (Dias et al. 2015; Šatović et al. 2016; Vojvoda Zeljko et al. 2020; Feliciello et al. 2020; Vondrak et al. 2020). However, information related to origin, function, distribution, and organizational patterns of short arrays of tandem repeats found in the genome and their relation to classical satDNA arrays is still limited.

All aforementioned studies on the satellite DNA sequences so far resulted in a need for a versatile pool of model systems, as it was shown that different organisms seem to follow different rules with respect to the abundance, distribution, organization, function, and evolution of these sequences. In continuation, we provide an overview of satellite DNAs in bivalve mollusks and their specificities within this group of organisms.

3.4 The Importance of Bivalve Mollusks in Genome Research

Bivalve mollusks are organisms that populate marine and fresh-water habitats throughout the world, playing important roles in ecosystems. Their impact has been registered in many different processes, including biofiltration, turnover, and storage of nutrients, participating in the transfer of organic substances and minerals, stimulation of primary and secondary production, creation and modification of natural habitats, and biogeochemical transformations (Vaughn and Hoellein 2018). Bivalves have also been recognized as organisms important for environmental monitoring (Gosling 2003). Their ecological significance is especially accented in cases when invasive bivalve species start to occupy new habitats where they can attain very high abundance, causing significant side effects on the food webs of the affected area (Vaughn and Hoellein 2018). In accordance with their high nutritional value, they represent a food source around the world and hold great importance in aquaculture. The employment in the farming industry results in million-ton production and well portraits such large commercial significance. Previous research on these organisms was mostly focused on finding the genetic basis for traits of interest: metabolism and growth, susceptibility to diseases, resilience to environmental stressors—all applicable in the farming industry (Saavedra and Bachère 2006). In continuation, their potential started to be noticed and reflected in many other research areas, e.g., in stem cells differentiation, the ability to fight pathogens in the absence of adaptive immunity, as a source of alternative drugs, in mucosal immunity, toxicology, and even in cancer resistance (Robledo et al. 2018). Recent studies have been moving toward genome-wide analyses (Gomes-dos-Santos et al. 2020) with a number of sequenced genomes growing rapidly. Nowadays, data from 27 sequenced bivalve genomes are available, assembled to the level of scaffolds, contigs, or chromosomes (PubMed, December 2020). In these genome projects, many different taxonomical groups have been encompassed (Table 3.1) and bivalve mollusks are being forwarded toward the model organisms (Robledo et al. 2018).

3.5 Satellite DNAs and Heterochromatin in Bivalve Mollusks

3.5.1 Heterochromatin in Bivalves

Heterochromatin represents a compacted, transcriptionally repressed form of chromatin within the genomes of higher eukaryotes, and it holds an important function in the silencing of repetitive elements and genome stability maintenance. It is frequently defined by the presence of a specific posttranslational histone H3 modification, H3K9me3 (reviewed by Nicetto and Zaret 2019). Genomic regions belonging to constitutive heterochromatin in bivalves were mostly accessed cytogenetically,

Table 3.1 Classification of bivalve species with currently available genome sequencing data

Kingdom	Phylum	Class	Infraclass	Order	Superfamily	Family	Genus	Species
Animalia	Mollusca	Bivalvia	Pteriomorphia	Pectinida	Pectinoidea	Pectinidae	Argopecten	*Argopecten irradians concentricus*
								Argopecten irradians irradians
				Mytilida	Mytiloidea	Mytilidae	Mytilus	*Mytilus galloprovincialis*
								Mytilus coruscus
							Bathymodiolus	*Bathymodiolus platifrons*
				Ostreida	Ostreoidea	Ostreidae	Crassostrea	*Crassostrea gigas*
								Crassostrea virginica
								Crassostrea hongkongensis
							Ostrea	*Ostrea lurida*
							Saccostrea	*Saccostrea glomerata*
					Pterioidea	Margaritidae	Pinctada	*Pinctada imbricata*
				Mytilida	Mytiloidea	Mytilidae	Limnoperna	*Limnoperna fortunei*
							Modiolus	*Modiolus philippinarum*
				Pectinida	Pectinoidea	Pectinidae	Mizuhopecten	*Mizuhopecten yessoensis*
							Pecten	*Pecten maximus*
				Arcida	Arcoidea	Arcidae	Tegillarca	*Tegillarca granosa*
			Heteroconchia	Myida	Pholadoidea	Teredinidae	Bankia	*Bankia setacea*
				Venerida	Veneroidea	Veneridae	Cyclina	*Cyclina sinensis*
							Ruditapes	*Ruditapes phillipinarum*
							Mercenaria	*Mercenaria mercenaria*
					Glossoidea	Vesicomyidae	Archivesica	*Archivesica marissinica*
					Mactroidea	Mactridae	Lutraria	*Lutraria rhynchaena*
					Corbiculoidea	Corbiculidae	Corbicula	*Corbicula fluminea*
				Myida	Dreissenoidea	Dreissenidae	Dreissena	*Dreissena rostriformis*
				Adapedonta	Hiatelloidea	Hiatellidae	Panopea	*Panopea generosa*

(continued)

Table 3.1 (continued)

Kingdom	Phylum	Class	Infraclass	Order	Superfamily	Family	Genus	Species
					Solenoidea	Pharidae	Sinonovacula	*Sinonovacula constricta*
				Unionoida	Unionoidea	Unionidae	Venustaconcha	*Venustaconcha ellipsiformis*

using the C-banding method (reviewed by Leitão and Chaves 2008). The results have shown that the abundance and localization of heterochromatin vary significantly among bivalve species, even between those belonging to the same genus. Its distribution can be extremely scarce, e.g., in *Crassostrea gigas,* where it is limited only to the centromeric region of one chromosome pair and the telomeric region of another pair (Bouilly et al. 2008). On the contrary, sister-species with the possibility of cross-fertilization, *C. angulata*, shows an abundant presence of heterochromatin in most of the ten chromosome pairs, located at pericentric, telomeric, and intercalary positions (Cross et al. 2005). In *Sphaerium* species (Petkevičiūtė et al. 2018) constitutive heterochromatin is limited exclusively to (peri)centromeres while in *Donax trunculus* (Petrović et al. 2009) those areas are completely heterochromatin-devoid, across all chromosomal pairs. (Peri)centromeric localization of constitutive heterochromatin was frequent in oysters, but not in mussels or scallops (Leitão and Chaves 2008). While species from the genus *Mytilus* contained sets of C-bands that were common to the three tested taxa, one of the species harbored also several additional heterochromatic loci (Martínez-Lage et al. 1995). Pacific oyster, *C. gigas*, is the first bivalve species in which heterochromatin was explored also on the molecular level. For that purpose, chromatin immunoprecipitation was employed, followed by high-throughput next-generation sequencing of the H3K9me3-associated sequences, revealing that the heterochromatin of this species is predominantly constituted of DNA transposons (Tunjić Cvitanić et al. 2020). Immunofluorescent detection of H3K9me3 histone mark performed by the same authors confirmed also general paucity and limited localization of heterochromatin in this organism, previously attended by Bouilly et al. (2008) using the C-banding method. Overall information available so far speak in favor of great heterogeneity in contribution and localization of constitutive heterochromatin in bivalve mollusks.

3.5.2 Genome Sequencing and Repetitive DNA Characterization in Bivalve Mollusks

One of the main challenges in bivalve genome assembly lies in the high heterozygosity and amount of repetitive elements these organisms contain. The mussels *Limnoperna fortunei, Modiolus philippinarum,* and the oyster *Crassostrea gigas* genomes were estimated to have heterozygosity rates of 2.3%, 2.02%, and 1.95%, respectively (Uliano-Silva et al. 2018), significantly exceeding many other animal genomes (Zhang et al. 2018). Repetitive sequences comprise about 35% of the genomes of bivalve species studied so far (Murgarella et al. 2016; Zhang et al. 2012; Takeuchi et al. 2012; Sun et al. 2017; Wang et al. 2017; Mun et al. 2017; Du et al. 2017), with the exception of *Modiolus philippinarum* where this number is doubled (Renaut et al. 2018). Among them, satDNAs comprise only small parts of the sequenced bivalve genomes, for example: 1.85% in *Pinctada fucata* (Takeuchi et al. 2012), 0.08% in *Ruditapes philippinarum* (Mun et al. 2017), 1.2% in *C. gigas* (Zhang et al. 2012). At the same time, a large amount of repetitive DNA sequences (>70%) consistently remained unclassified in all inspected bivalve genomes (Murgarella et al. 2016). The same authors propose that comparisons of the abundance of repetitive sequence between species should be performed with restraint, as the large portion of unclassified repeats might contain species-specific variants of certain types, and that may therefore change the relative contribution of each category on the total. Another potential reason that could explain this deficient classification of repetitive sequences in bivalves is the high contribution of mobile elements holding tandem repeats in their structure (Tunjić Cvitanić et al. 2020). Mobile elements of the Helitron superfamily are hybrid structures, as they usually contain arrays of tandem repeats in their central part, flanked with left and right conserved sequence segments (Thomas and Pritham 2015). A problem in the classification of such elements was observed in the Pacific oyster, using RepeatExplorer pipeline (Tunjić Cvitanić et al. 2020). There, in certain cases, tandem repeats from central parts of previously described Helitrons were placed in one cluster and classified as a satellite DNA, while sequences surrounding central repeats were allocated to separate clusters, without clear classification. To conclude, repetitive DNA sequences in bivalve mollusk genomes still pose a challenge and need to be both quantified and classified in more detail.

Although currently available data from bivalve genome projects bring very little information regarding satDNA sequences in bivalves, many information exist based on conventional methods for their detection and cover a significant number of 48 species (Šatović et al. 2018). Conventional methods yielded satDNAs from the families Ostreidae (Clabby et al. 1996; López-Flores et al. 2004), Donacidae (Plohl and Cornudella 1997; Petrović et al. 2009), Pectinidae (Canapa et al. 2000; Petraccioli et al. 2015), Mactridae (García-Souto et al. 2017), and other. The method based on the construction of partial genomic libraries followed by colony lift hybridization using fragmented total genomic DNA, yielded repeat-enriched DNA sequences, employed by Biscotti et al. (2007), Šatović et al. (2016), Šatović and Plohl (2018). Monomer size of satDNAs detected by such conventional methods can vary, and usually ranging between 40 and 400 bp, and predominance of 150–210 bp monomers can be noticed (reviewed in Šatović et al. 2018). The correspondence to the mononucleosomal length (Henikoff et al. 2001) is considered to be evolutionarily favored for chromatin packing (Heslop-Harrison and Schwarzacher 2013). Novel methods, based on short-read NGS data, are just starting to be employed on bivalves, bringing the information that monomer size in these species can increase to about 2000 bp, as observed for the Pacific oyster *C. gigas* (unpublished data).

Transcription of satDNA in this group of organisms is still very poorly attended. The transcriptional activity was reported for Ac4p3 satDNA and the CvA transposon holding tandem repeats in the species *Adamussium colbecki* (Biscotti et al. 2018). In addition, DTHS3 satDNA of several bivalve species was found to be present in NCBI EST (Expressed Sequence Tags) database (Šatović and Plohl 2018). However, before putative biological implications of such observations are brought, furthering this area of research in these organisms is necessary.

3.5.3 Extremely Long Ancestry of Bivalve satDNAs

Satellite DNA detection by conventional methods was frequently directed by the library hypothesis (Fry and Salser 1977), predicting that related species share a series of satDNAs derived from a common ancestor, as described above. Thereby, the choice of bivalve species used for screening was frequently limited to (closely) related sets of species, e.g., Veneridae (Passamonti et al. 1998), Mytilidae (Martínez-Lage et al. 2002), Ostreidae (López-Flores et al. 2004). As the number of inspected species started to broaden and the availability of data from different bivalve species increased in NCBI GenBank database, wider distribution of a specific satDNA sequence started to be noticed. An example is PjHhaI satDNA, which is found to be present even outside the class Bivalvia, and very likely have an extraordinary long evolutionary ancestry (Petraccioli et al. 2015). In continuation, two satDNAs were inspected across a distant set of bivalves and have also shown remarkable age. The first one, DTHS3 satDNA, was detected in 12 bivalve species belonging to subclasses Heterodonta and Pteriomorphia (Šatović and Plohl 2018), and its minimal age, based on the separation of these two lineages (Bieler et al. 2014), was estimated to be 516 MY. The second one, BIV160 satDNA, is especially interesting. It was originally detected in nine species distributed across all of the main bivalve subclasses: Protobranchia, Pteriomorphia, and Heteroconchia. In accordance with such dispersal, it was estimated to be at least 540 MY old (Plohl et al. 2010). An emerging amount of genomic data from different species could be of use to determine if many other satDNAs are also widely preserved across different taxa within the class Bivalvia, or even wider. Nonetheless, despite the occasional occurrence of species-specific variants, these organisms exhibit impressive long-term satDNA sequence preservation throughout evolutionary history.

3.5.4 Close Connection of Bivalve satDNAs to Mobile Elements

Although observed in other species, the connection of satDNA sequences and mobile elements is particularly evident in bivalves. In *Crassostrea virginica* the *pearl* element incorporates short arrays of ~160 bp long monomers (Gaffney et al. 2003) which are related to several satDNAs found in bivalve species, distant from *C. virginica* and from each other. Related satDNAs are HindIII from oysters (López-Flores et al. 2004), DTE of *Donax trunculus* (Plohl and Cornudella 1996), and BIV160, broadly distributed among bivalve species (Plohl et al. 2010). In the clam *D. trunculus* another structurally equivalent element incorporating an array of tandem repeats has been characterized, DTC84 (Šatović and Plohl 2013). Mg1 satellite DNA of *Mytilus galloprovincialis* was found to be also the core repetitive sequence of a putative TE named MgE (Kourtidis et al. 2006). The Mg1 repeats and their flanking regions were noticed to exhibit sequence and structural homology to the respective regions of CvE, a member of the *pearl* family of mobile elements (Gaffney et al. 2003).

A survey of the Pacific oyster *C. gigas* revealed that its genome is replete with satDNA-like tandem repeats incorporated into Helitron/Helentron elements (Šatović et al. 2016; Vojvoda Zeljko et al. 2020). Illustrative are 11 nonautonomous elements named Cg_HINE, where each of the described elements is formed by a unique combination of flanking sequences and satDNA-like central repeats. In addition, some of the satDNA-like arrays of Cg_HINE (Vojvoda Zeljko et al. 2020) were related to the most abundant classical Cg170/HindIII satDNA of *C. gigas* (Clabby et al. 1996; López-Flores et al. 2004). PjHhaI satellite DNA isolated from *Pecten jacobaeus* shows high sequence similarity among mollusks, and was found surrounded with structures belonging to TEs in species *C. gigas* and *Capitella teleta* (Petraccioli et al. 2015). Another long-ancestry satDNA, widely distributed among bivalve species, DTHS3, was found to be embedded in structures that could be responsible for its mobility in *Mercenaria mercenaria, Spisula solidissima, C. virginica*, and *M. galloprovincialis* (Šatović and Plohl 2018).

Similarity to SINE elements has also been observed for several bivalve satDNAs. Examples are ApaI-repeats found in *Mytilus* species, holding sequence segments that show similarity to RNA Pol III A and B boxes (Martínez-Lage et al. 2005). A 150-bp-long BclI repeats of oysters also contain sequence blocks, in this case degenerate ones, with similarity to the boxes A and B of SINE elements (López-Flores et al. 2010). In addition, satDNAs of bivalve mollusks frequently exhibit distribution patterns that are most probably connected to their association to mobile elements, providing the ability to build a large number of (short) dispersed arrays on many locations (Fig. 3.3). Such an organization is often observed by FISH-detection on the chromosomes (Wang et al. 2001; Biscotti et al. 2007; López-Flores et al. 2010, etc.), or by an in silico analysis on pseudochromosomes (Tunjić Cvitanić et al. 2020; Vojvoda Zeljko et al. 2020).

3.5.5 *Conserved Boxes in satDNA Monomers of Bivalve Species*

Conserved boxes that connect bivalve satDNAs with mobile elements have already been mentioned, yet other different sequence motifs exist. Already mentioned *pearl* element CvA and three satDNAs (HindIII, DTE, and BIV160) share two conserved monomer segments, which are putatively important motifs from the functional aspect (Plohl et al. 2010). Such sequence segments of reduced variability in comparison to the rest of the monomer sequence are proposed to be a result of functional constraints imposed. One of them could be a role in a DNA–protein interaction, well-exampled by a CENP-B box. CENP-B protein, upon recognizing and binding to the conserved CENP-B box, present within the human alpha satellite DNA, participates in the formation and assembly of the centromere (Masumoto et al. 1989). Related to

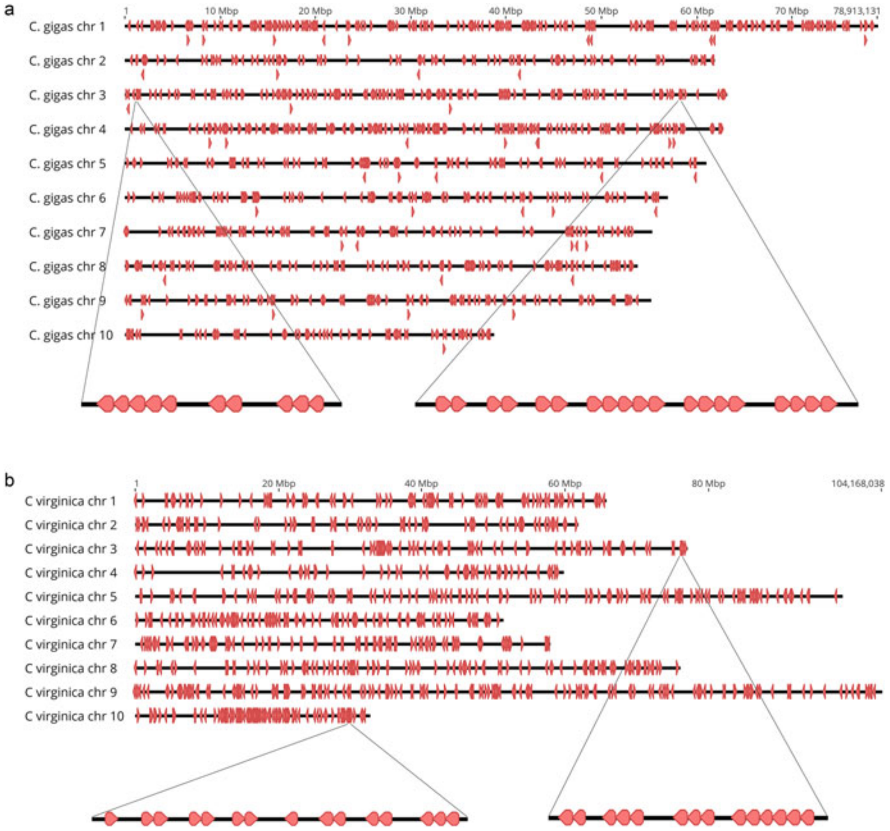


Fig. 3.3 SatDNA with monomer size of 437 bp localized in silico on the chromosomes of the pacific oyster *C. gigas* (**a**) and eastern oyster *C. virginica* (**b**). A large number of highly interspersed short arrays and single monomers can be noticed on many locations along the chromosomes, potentially related to their association to mobile elements in these species

that, pACS satellite DNA found in the antarctic scallop *Adamussium colbecki* was found to contain sequence segments showing similarity to the CENP-B box of higher primates and centromeric DNA element III of yeast (Canapa et al. 2000). Respectively, the localization of this satDNA was shown to be centromeric on the chromosomes of this species (Odierna et al. 2006). Although sequence motifs exist in these species, functional studies yet need to be performed to confirm true involvement in functional interactions. On the other hand, similarities to conserved boxes of mobile elements, especially SINE (Martínez-Lage et al. 2005; López-Flores et al. 2010), would suggest that such sequence segments in those repeats point to the sequence origin and not its functional involvements.

3.5.6 Low Abundance and a Large Palette of satDNAs in Bivalve Genomes Questioning the satDNA Library Concept

As already mentioned, available information from genomic projects brings an extremely low abundance of satDNAs in bivalve genomes, <2% (Takeuchi et al. 2012; Zhang et al. 2012; Murgarella et al. 2016; Mun et al. 2017), especially in comparison with their presence in other organisms (Garrido-Ramos 2017). In accordance, classical detection methods revealed many satDNAs in different bivalve species, each of them constituting only 0.008–2% of the respective genome, with very few exceptions (reviewed by Šatović et al. 2018). Interestingly, such low genomic abundance does not presume a reduced number of different satDNAs present in the genome. In the wedge clam *Donax trunculus* eight different satDNAs have been detected so far (Plohl and Cornudella 1996, 1997; Petrović and Plohl 2005; Petrović et al. 2009; Plohl et al. 2010), while in the scallop *Pecten maximus* at least 10 satDNAs are present (Martínez-Lage et al. 2002; Biscotti et al. 2007; Petraccioli et al. 2015). Novel sequencing methods and specialized bioinformatics programs start to reveal that even a significantly larger number of different satDNA can exist in a bivalve genome, increasing up to about 40 in the Pacific oyster *Crassostrea gigas* (unpublished data). Although the number of satDNAs coexisting in the genome can be significant and all of them show very low genomic occupancy, certain satDNA severalfoldly take lead in the respective genome compared to the rest of the satDNA inventory, e.g., PjHhaI in *P. maximus* (Petraccioli et al. 2015), DTF2 in *D. trunculus* (Petrović et al. 2009), Cg170/HindIII satDNA in *C. gigas* (Clabby et al. 1996; López-Flores et al. 2004). On the other hand, several satDNA were found not to be limited only to closely related species but to be widely distributed among bivalves (Plohl et al. 2010; Šatović and Plohl 2018) or even to transcend the taxonomical level of the class Bivalvia (Petraccioli et al. 2015). This would potentially broaden the library concept even to distantly related species and presume long-term preservation of a spectrum of satDNA sequences, derived from the common ancestor, across different taxa. On the other hand, the close connection of some of these sequences with mobile elements opens the possibility of their horizontal transfer, affecting the conclusions related to their ancestry, based solely on vertical inheritance. Constantly generated new data will hopefully provide sufficient and adequate information in order to provide answers to this evolutionary question.

3.5.7 Methylation Patterns of satDNA Repeats

Repetitive sequences account for the majority of methylated sites of the genome and their methylated state is necessary for proper genome function and maintenance of its integrity. DNA methylation has been associated with most of the biological processes, in addition to the well-known transcriptional repression (reviewed in Francastel and Magdinier 2019). Information related to methylation profiles of repetitive DNA sequences in bivalve species is very limited. However, Wang et al. (2014) have noticed that in the Pacific oyster *C. gigas* DNA transposons, Helitrons, satDNAs, simple repeats, and other tandem repeats all displayed methylation levels that were on average 2 times higher than the genome background. On the contrary, the overall methylation level of repetitive elements in the species *Pinctada fucata martensii* was lower than the genome background (Zhang et al. 2020). In the latter case, no information was brought for satDNAs, but Helitron elements exhibit average methylation levels when compared to other mobile elements of that species. Although methylation patterns of these sequences are still very perplexing, some specific features have been noticed. Wang et al. (2014) have divided the repeats into methylated and unmethylated ones and noticed that the divergence rate within the methylated group was significantly lower when compared to the unmethylated one. By the employment of methylation-sensitive and methylation-insensitive restriction endonucleases in wedge clam *Donax trunculus,* DTF2 satDNA was also shown to be methylated. This satDNA displays high and uniform sequence conservation throughout the entire monomer sequence (Petrović et al. 2009). Similar situation was observed during the investigation of SSU satDNA of cut trough shell *Spisula subtruncata* (García-Souto et al. 2017). There, more detailed analysis had shown that segments of monomer sequence show differences in nucleotide diversity which is inversely correlated with DNA methylation. The general level of methylation of SSU satellite is quite high and triplicates the mean of the *S. subtruncata* genome (García-Souto et al. 2017). Specificities in the methylation pattern of repetitive sequences in bivalves suggest functions that go beyond the mere DNA silencing, and the nonrandom nature of DNA methylation profiles implies that the methylation machinery must be guided to specific genomic locations (Francastel and Magdinier 2019). In accordance with that is the existence of a single chromosome pair in *S. subtruncata* where the presence of distinctly under-methylated SSU satDNA monomers can be observed (García-Souto et al. 2017). It can be concluded that the methylation processes that shape repetitive genome compartments of bivalve mollusks are nonrandom, quite complex, and not necessarily uniform, and the implication of these patterns remains to be revealed.

3.6 Future Perspectives

The extreme biological, ecological, and commercial importance of bivalve mollusks resulted in increased interest for every aspect of bivalve genomics, as well as in rapidly accumulating sequenced genomes. Observed specificities of sequences repeated in tandem, their association with mobile elements, and the peculiarities of heterochromatin distribution, promote bivalves as a promising group of organisms in studies of satDNAs. It can be expected that the advantages of novel approaches in satDNA research, such as using NGS and third-generation sequencing, will reveal even more interesting details and will broaden existing concepts and add new data to the area of satDNA biology. This could be achieved on the two levels, performing detailed studies of high-quality sequenced referent genomes, and by increasing the number of explored species for more comprehensive comparative analyses. In particular, studies of satDNAs should be directed toward function-oriented analyses, including the putative roles of satDNA transcripts in bivalves, so far *terra incognita* in these species.

References

Adega F, Guedes-Pinto H, Chaves R (2009) Satellite DNA in the karyotype evolution of domestic animals - clinical considerations. Cytogenet Genome Res 126:12–20. https://doi.org/10.1159/000245903

Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8:61–65. https://doi.org/10.1038/nmeth.1527

Belyayev A, Josefiová J, Jandová M et al (2019) Natural history of a satellite DNA family: from the ancestral genome component to species-specific sequences, concerted and non-concerted evolution. Int J Mol Sci 20. https://doi.org/10.3390/ijms20051201

Bieler R, Mikkelsen PM, Collins T et al (2014) Investigating the bivalve tree of life – an exemplar-based approach combining molecular and novel morphological characters. Invertebr Syst 28:32–115. https://doi.org/10.1071/IS13010

Biscotti MA, Canapa A, Olmo E et al (2007) Repetitive DNA, molecular cytogenetics and genome organization in the King scallop (*Pecten maximus*). Gene 406:91–98. https://doi.org/10.1016/j.gene.2007.06.027

Biscotti MA, Olmo E, Heslop-Harrison JS (2015) Repetitive DNA in eukaryotic genomes. Chromosom Res 23:415–420. https://doi.org/10.1007/s10577-015-9499-z

Biscotti MA, Barucca M, Canapa A (2018) New insights into the genome repetitive fraction of the Antarctic bivalve *Adamussium colbecki*. PLoS One 13:1–17. https://doi.org/10.1371/journal.pone.0194502

Bouilly K, Chaves R, Leitao A et al (2008) Chromosomal organization of simple sequence repeats in chromosome patterns. J Genet 87:119–125

Brajković J, Feliciello I, Bruvo-Mađaric B, Ugarković D (2012) Satellite DNA-like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. G3 Genes Genomes Genetics 2:931–941. https://doi.org/10.1534/g3.112.003467

Brajković J, Pezer Ž, Bruvo-Mađarić B et al (2018) Dispersion profiles and gene associations of repetitive DNAs in the euchromatin of the beetle *Tribolium castaneum*. G3 Genes Genomes Genetics 8:875–886. https://doi.org/10.1534/g3.117.300267

Canapa A, Barucca M, Cerioni PN, Olmo E (2000) A satellite DNA containing CENP-B box-like motifs is present in the Antarctic scallop *Adamussium colbecki*. Gene 247:175–180

Chang CH, Chavan A, Palladino J et al (2019) Islands of retroelements are major components of *Drosophila* centromeres. PLoS Biol. https://doi.org/10.1371/journal.pbio.3000241

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

Chaves R, Ferreira D, Mendes-Da-Silva A et al (2017) FA-SAT is an old satellite DNA frozen in several bilateria genomes. Genome Biol Evol 9:3073–3087. https://doi.org/10.1093/gbe/evx212

Clabby C, Goswami U, Flavin F et al (1996) Cloning, characterization and chromosomal location of a satellite DNA from the Pacific oyster, *Crassostrea gigas*. Gene 168:205–209

Cross I, Díaz E, Sánchez I, Rebordinos L (2005) Molecular and cytogenetic characterization of *Crassostrea angulata* chromosomes. Aquaculture 247:135–144. https://doi.org/10.1016/j.aquaculture.2005.02.039

Csink AK, Henikoff S (1998) Something from nothing: the evolution and utility of satellite repeats. Trends Genet 14:200–204. https://doi.org/10.1016/S0168-9525(98)01444-9

de Lima LG, Svartman M, Kuhn GCS (2017) Dissecting the satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes. G3 Genes Genomes Genetics 7:2831–2843. https://doi.org/10.1534/g3.117.042093

de Silva DMZA, Utsunomia R, Ruiz-Ruano FJ et al (2017) High-throughput analysis unveils a highly shared satellite DNA library among three species of fish genus *Astyanax*. Sci Rep 7:12726. https://doi.org/10.1038/s41598-017-12939-7

Dias GB, Svartman M, Delprat A et al (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. https://doi.org/10.1093/gbe/evu108

Dias GB, Heringer P, Svartman M, Kuhn GCSS (2015) Helitrons shaping the genomic architecture of *Drosophila*: enrichment of DINE-TR1 in α- and β-heterochromatin, satellite DNA emergence, and piRNA expression. Chromosom Res 23:597–613. https://doi.org/10.1007/s10577-015-9480-x

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117. https://doi.org/10.1038/299111a0

Dover GA (1986) Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet 2:159–165. https://doi.org/10.1016/0168-9525(86)90211-8

Du X, Fan G, Jiao Y et al (2017) The pearl oyster *Pinctada fucata martensii* genome and multi-omic analyses provide insights into biomineralization. GigaScience 6:1–12. https://doi.org/10.1093/gigascience/gix059

Elder JF, Turner BJ (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. Q Rev Biol 70:297–320. https://doi.org/10.1086/419073

Feliciello I, Akrap I, Ugarković Ð (2015) Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. PLoS Genet 11:e1005466. https://doi.org/10.1371/journal.pgen.1005466

Feliciello I, Pezer Ž, Kordiš D et al (2020) Evolutionary history of alpha satellite DNA repeats dispersed within human genome euchromatin. Genome Biol Evol 12:2125–2138. https://doi.org/10.1093/gbe/evaa224

Ferree PM, Barbash DA (2009) Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. PLoS Biol 7. https://doi.org/10.1371/journal.pbio.1000234

Francastel C, Magdinier F (2019) DNA methylation in satellite repeats disorders. Essays Biochem 63:757–771. https://doi.org/10.1042/EBC20190028

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

Gaffney PM, Pierce JC, Mackinley AG et al (2003) *Pearl*, a novel family of putative transposable elements in bivalve mollusks. J Mol Evol 56:308–316. https://doi.org/10.1007/s00239-002-2402-5

García-Souto D, Mravinac B, Šatović E et al (2017) Methylation profile of a satellite DNA constituting the intercalary G+C-rich heterochromatin of the cut trough shell *Spisula subtruncata* (Bivalvia, Mactridae). Sci Rep 7:6930. https://doi.org/10.1038/s41598-017-07231-7

Garrido-Ramos MA (2017) Satellite DNA: an evolving topic. Genes (Basel) 8:1–41. https://doi.org/10.3390/genes8090230

Gomes-dos-Santos A, Lopes-Lima M, Castro LFC, Froufe E (2020) Molluscan genomics: the road so far and the way forward. Hydrobiologia 847:1705–1726. https://doi.org/10.1007/s10750-019-04111-1

Gosling E (ed) (2003) Bivalve Molluscs: biology, ecology and culture. Blackwell Science, Oxford and Malden

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

Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293:1098–1102. https://doi.org/10.1126/science.1062939

Herndon N, Shelton J, Gerischer L et al (2020) Enhanced genome assembly and a new official gene set for *Tribolium castaneum*. BMC Genomics 21:1–13. https://doi.org/10.1186/s12864-019-6394-6

Heslop-Harrison JSP, Schwarzacher T (2013) Nucleosomes and centromeric DNA packaging. Proc Natl Acad Sci USA 110:19974–19975. https://doi.org/10.1073/pnas.1319945110

Jain M, Olsen HE, Turner DJ et al (2018) Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol 36:321–323. https://doi.org/10.1038/nbt.4109

Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 8:241–259. https://doi.org/10.1146/annurev.genom.8.080706.092416

Khost D, Eickbush D, Larracuente A (2017) Single molecule long read sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome Res 27:1–13. https://doi.org/10.1101/gr.213512.116.Freely

Kim YB, Oh JH, McIver LJ et al (2014) Divergence of *Drosophila melanogaster* repeatomes in response to a sharp microclimate contrast in evolution canyon, Israel. Proc Natl Acad Sci 111:10630–10635. https://doi.org/10.1073/pnas.1410372111

Kit S (1961) Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. J Mol Biol 3:711–716. https://doi.org/10.1016/S0022-2836(61)80075-2

Klemme S, Banaei-Moghaddam AM, Macas J et al (2013) High-copy sequences reveal distinct evolution of the rye B chromosome. New Phytol 199:550–558. https://doi.org/10.1111/nph.12289

Kourtidis A, Drosopoulou E, Pantzartzi CN et al (2006) Three new satellite sequences and a mobile element found inside HSP70 introns of the Mediterranean mussel (*Mytilus galloprovincialis*). Genome 49:1451–1458. https://doi.org/10.1139/g06-111

Kuderna LFK, Lizano E, Julià E et al (2019) Selective single molecule sequencing and assembly of a human Y chromosome of African origin. Nat Commun 10. https://doi.org/10.1038/s41467-018-07885-5

Kuhn GCS, Küttler H, Moreira-Filho O, Heslop-Harrison JS (2012) The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. Mol Biol Evol 29:7–11. https://doi.org/10.1093/molbev/msr173

Larracuente AM (2014) The organization and evolution of the responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. BMC Evol Biol 14:1–12. https://doi.org/10.1186/s12862-014-0233-9

Leitão A, Chaves R (2008) Banding for chromosomal identification in bivalves: a 20-year history. In: Dynamic biochemistry, process biotechnology and molecular biology. Global Science Books, Ikenobe, Japan, pp 44–49

Liu Q, Li X, Zhou X et al (2019) The repetitive DNA landscape in Avena (Poaceae): chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads. BMC Plant Biol 19:1–17. https://doi.org/10.1186/s12870-019-1769-z

Logsdon GA, Vollger MR, Hsieh P et al (2021) The structure, function, and evolution of a complete human chromosome 8. Nature 593:101-107. https://doi.org/10.1038/s41586-021-03420-7

López-Flores I, Garrido-Ramos MA (2012) The repetitive DNA content of eukaryotic genomes. Genome Dyn 7:1–28

López-Flores I, de la Herrán R, Garrido-Ramos MA et al (2004) The molecular phylogeny of oysters based on a satellite DNA related to transposons. Gene 339:181–188. https://doi.org/10.1016/j.gene.2004.06.049

López-Flores I, Ruiz-Rejón C, Cross I et al (2010) Molecular characterization and evolution of an interspersed repetitive DNA family of oysters. Genetica 138:1211–1219. https://doi.org/10.1007/s10709-010-9517-1

Louzada S, Lopes M, Ferreira D et al (2020) Decoding the role of satellite DNA in genome architecture and plasticity—an evolutionary and clinical affair. Genes (Basel) 11. https://doi.org/10.3390/genes11010072

Lower SS, McGurk MP, Clark AG, Barbash DA (2018) Satellite DNA evolution: old ideas, new approaches. Curr Opin Genet Dev 49:70–78. https://doi.org/10.1016/j.gde.2018.03.003

Luchetti A (2015) terMITEs: miniature inverted-repeat transposable elements (MITEs) in the termite genome (Blattodea: Termitoidae). Mol Gen Genomics 290:1499–1509. https://doi.org/10.1007/s00438-015-1010-1

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 8:1–16. https://doi.org/10.1186/1471-2164-8-427

Macas J, Koblízková A, Navrátilová A, Neumann P (2009) Hypervariable 3' UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448:198–206. https://doi.org/10.1016/j.gene.2009.06.014

Macas J, Novak 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 10:1–23. https://doi.org/10.1371/journal.pone.0143424

Martínez-Lage A, González-Tizón A, Méndez J (1995) Chromosomal markers in three species of the genus *Mytilus* (Mollusca: Bivalvia). Heredity (Edinb) 74:369–375. https://doi.org/10.1038/hdy.1995.55

Martínez-Lage A, Rodríguez F, González-Tizón A et al (2002) Comparative analysis of different satellite DNAs in four *Mytilus* species. Genome 45:922–929

Martínez-Lage A, Rodríguez-Fariña F, González-Tizón A, Méndez J (2005) Origin and evolution of *Mytilus* mussel satellite DNAs. Genome 48:247–256. https://doi.org/10.1139/G04-115

Masumoto H, Masukata H, Muro Y et al (1989) A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol 109:1963–1973

Meštrović N, Plohl M, Mravinac B, Ugarković D (1998) Evolution of satellite DNAs from the genus *Palorus*-experimental evidence for the "library" hypothesis. Mol Biol Evol 15:1062–1068

Meštrović N, Mravinac B, Pavlek M et al (2015) Structural and functional liaisons between transposable elements and satellite DNAs. Chromosom Res 23:583–596. https://doi.org/10.1007/s10577-015-9483-7

Miga KH (2015) Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosom Res 23:421–426. https://doi.org/10.1007/s10577-015-9488-2

Miga KH (2019) Centromeric satellite DNAs: hidden sequence variation in the human population. Genes (Basel) 10. https://doi.org/10.3390/genes10050352

Miga KH, Koren S, Rhie A et al (2020) Telomere-to-telomere assembly of a complete human X chromosome. Nature 585:79–84. https://doi.org/10.1038/s41586-020-2547-7

Mitsuhashi S, Frith MC, Mizuguchi T et al (2019) Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol 20:1–17. https://doi.org/10.1186/s13059-019-1667-6

Mun S, Kim YJ, Markkandan K et al (2017) The whole-genome and transcriptome of the Manila clam (*Ruditapes philippinarum*). Genome Biol Evol 9:1487–1498. https://doi.org/10.1093/gbe/evx096

Murgarella M, Puiu D, Novoa B et al (2016) A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. PLoS One 11:1–22. https://doi.org/10.1371/journal.pone.0151561

Nicetto D, Zaret KS (2019) Role of H3K9me3 heterochromatin in cell identity establishment and maintenance. Curr Opin Genet Dev 55:1–10. https://doi.org/10.1016/j.gde.2019.04.013

Novák P, Neumann P, Pech J et al (2013) Repeat explorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29:792–793. https://doi.org/10.1093/bioinformatics/btt054

Novák P, Robledillo LÁ, Koblížková A et al (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res 45:e111. https://doi.org/10.1093/nar/gkx257

Odierna G, Aprea G, Barucca M et al (2006) Karyology of the Antarctic scallop *Adamussium colbecki*, with some comments on the karyological evolution of pectinids. Genetica 127:341–349. https://doi.org/10.1007/s10709-005-5366-8

Ohno S (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23:366–370

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607

Paar V, Glunčić M, Rosandić M et al (2011) Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. Mol Biol Evol 28:1877–1892. https://doi.org/10.1093/molbev/msr009

Palacios-Gimenez OM, Dias GB, De Lima LG et al (2017) High-throughput analysis of the satellitome revealed enormous diversity of satellite DNAs in the neo-Y chromosome of the cricket *Eneoptera surinamensis*. Sci Rep 7:1–11. https://doi.org/10.1038/s41598-017-06822-8

Palacios-Gimenez OM, Bardella VB, Lemos B, Cabral-De-Mello DC (2018) Satellite DNAs are conserved and differentially transcribed among *Gryllus* cricket species. DNA Res 25:137–147. https://doi.org/10.1093/dnares/dsx044

Palacios-Gimenez OM, Milani D, Song H et al (2020a) Eight million years of satellite DNA evolution in grasshoppers of the genus *Schistocerca* illuminate the ins and outs of the library hypothesis. Genome Biol Evol 12:88–102. https://doi.org/10.1093/gbe/evaa018

Palacios-Gimenez OM, Koelman J, Flores MP et al (2020b) Comparative analysis of morabine grasshopper genomes reveals highly abundant transposable elements and rapidly proliferating satellite DNA repeats. BMC Biology 18:199. https://doi.org/10.1186/s12915-020-00925-x

Palomeque T, Antonio Carrillo J, Muñoz-López M, Lorite P (2006) Detection of a mariner-like element and a miniature inverted-repeat transposable element (MITE) associated with the heterochromatin from ants of the genus *Messor* and their possible involvement for satellite DNA evolution. Gene 371:194–205. https://doi.org/10.1016/j.gene.2005.11.032

Pardue ML, Gall JG (1970) Chromosomal localization of mouse satellite DNA. Science 168:1356–1358. https://doi.org/10.1126/science.168.3937.1356

Passamonti M, Mantovani B, Scali V (1998) Characterization of a highly repeated DNA family in Tapetinae species (Mollusca Bivalvia: Veneridae). Zool Sci 15:599–605. https://doi.org/10.2108/0289-0003(1998)15[599,COAHRD]2.0.CO;2

Pavlek M, Gelfand Y, Plohl M, Meštrović N (2015) Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. DNA Res 22:387–401. https://doi.org/10.1093/dnares/dsv021

Peñaloza C, Gutierrez AP, Eory L et al (2021) A chromosome-level genome assembly for the Pacific oyster *Crassostrea gigas*. GigaScience 10:1–9. https://doi.org/10.1093/gigascience/giab020

Peona V, Weissensteiner MH, Suh A (2018) How complete are "complete" genome assemblies?-An avian perspective. Mol Ecol Resour 18:1188–1195. https://doi.org/10.1111/1755-0998.12933

Petkevičiūtė R, Stunžėnas V, Stanevičiūtė G (2018) Comments on species divergence in the genus *Sphaerium* (Bivalvia) and phylogenetic affinities of Sphaerium nucleus and *S. corneum var. mamillanum* based on karyotypes and sequences of 16S and ITS1 rDNA. PLoS One 13:1–17. https://doi.org/10.1371/journal.pone.0191427

Petraccioli A, Odierna G, Capriglione T et al (2015) A novel satellite DNA isolated in *Pecten jacobaeus* shows high sequence similarity among molluscs. Mol Gen Genomics 290:1717–1725. https://doi.org/10.1007/s00438-015-1036-4

Petrović V, Plohl M (2005) Sequence divergence and conservation in organizationally distinct subfamilies of *Donax trunculus* satellite DNA. Gene 362:37–43. https://doi.org/10.1016/j.gene.2005.06.044

Petrović V, Pérez-García C, Pasantes JJ et al (2009) A GC-rich satellite DNA and karyology of the bivalve mollusk *Donax trunculus*: a dominance of GC-rich heterochromatin. Cytogenet Genome Res 124:63–71. https://doi.org/10.1159/000200089

Pezer Ž, Ugarković Đ (2008) Role of non-coding RNA and heterochromatin in aneuploidy and cancer. Semin Cancer Biol 18:123–130. https://doi.org/10.1016/j.semcancer.2008.01.003

Pezer Z, Brajković J, Feliciello I, Ugarković D (2011) Transcription of satellite DNAs in insects. Prog Mol Subcell Biol 51:161–178

Pita S, Panzera F, Mora P et al (2017) Comparative repeatome analysis on Triatoma infestans Andean and non-Andean lineages, main vector of Chagas disease. PLoS One 12:e0181635. https://doi.org/10.1371/journal.pone.0181635

Plohl M, Cornudella L (1996) Characterization of a complex satellite DNA in the mollusc *Donax trunculus*: analysis of sequence variations and divergence. Gene 169:157–164

Plohl M, Cornudella L (1997) Characterization of interrelated sequence motifs in four satellite DNAs and their distribution in the genome of the mollusc *Donax trunculus*. J Mol Evol 44:189–198

Plohl M, Luchetti A, Mestrović N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)-chromatin. Gene 409:72–82. https://doi.org/10.1016/j.gene.2007.11.013

Plohl M, Petrović V, Luchetti A et al (2010) Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. Heredity (Edinb) 104:543–551. https://doi.org/10.1038/hdy.2009.141

Plohl M, Meštrović N, Mravinac B (2012) Satellite DNA evolution. In: Garrido-Ramos M (ed) Genome dynamics. Karger AG, Basel, pp 126–152

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

Renaut S, Guerra D, Hoeh WR et al (2018) Genome survey of the freshwater mussel *Venustaconcha ellipsiformis* (Bivalvia: Unionida) using a hybrid de novo assembly approach. Genome Biol Evol 10:1637–1646. https://doi.org/10.1093/gbe/evy117

Robledo JAF, Yadavalli R, Allam B et al (2018) From the raw bar to the bench: bivalves as models for human health. Dev Comp Immunol 92:260–282. https://doi.org/10.1016/j.dci.2018.11.020

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM (2016) High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep 6:28333. https://doi.org/10.1038/srep28333

Saavedra C, Bachère E (2006) Bivalve genomics. Aquaculture 256:1–14

Sainz J, Prats E, Ruiz S, Cornudella L (1992) Organization of repetitive DNA sequences in the genome of the echinoderm *Holothuria tubulosa*. Biochimie 74:1067–1074

Šatović E, Plohl M (2013) Tandem repeat-containing MITE elements in the clam *Donax trunculus*. Genome Biol Evol 5:2549–2559. https://doi.org/10.1093/gbe/evt202

Šatović E, Plohl M (2018) Distribution of DTHS3 satellite DNA across 12 bivalve species. J Genet 97:575–580. https://doi.org/10.1007/s12041-018-0940-x

Šatović E, Vojvoda Zeljko T, Luchetti A et al (2016) Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. BMC Genomics 17:997. https://doi.org/10.1186/s12864-016-3347-1

Šatović E, Vojvoda Zeljko T, Plohl M (2018) Characteristics and evolution of satellite DNA sequences in bivalve mollusks. Eur Zool J 85:95–104. https://doi.org/10.1080/24750263.2018.1443164

Šatović E, Tunjić Cvitanić M, Plohl M (2020) Tools and databases for solving problems in detection and identification of repetitive DNA sequences. Period Biol 121–122:7–14. https://doi.org/10.18054/pb.v121-122i1-2.10571

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

Sedlazeck FJ, Lee H, Darby CA, Schatz MC (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet 19:329–346. https://doi.org/10.1038/s41576-018-0003-4

Sharma A, Wolfgruber TK, Presting GG (2013) Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14:142. https://doi.org/10.1186/1471-2164-14-142

Shatskikh AS, Kotov AA, Adashev VE et al (2020) Functional significance of satellite DNAs: insights from *Drosophila*. Front Cell Dev Biol 8:1–19. https://doi.org/10.3389/fcell.2020.00312

Singer MF (1982) Highly repeated sequences in mammalian genomes. Int Rev Cytol 76:67–122

Slamovits C, Rossi M (2002) Satellite DNA: agent of chromosomal evolution in mammals. A review. Mastozoología Neotrop 9:297–308

Smalec BM, Heider TN, Flynn BL, O'Neill RJ (2019) A centromere satellite concomitant with extensive karyotypic diversity across the *Peromyscus* genus defies predictions of molecular drive. Chromosom Res. https://doi.org/10.1007/s10577-019-09605-1

Sproul JS, Khost DE, Eickbush DG et al (2020) Dynamic evolution of Euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the simulans clade. Mol Biol Evol 37:2241–2256. https://doi.org/10.1093/molbev/msaa078

Strachan T, Webb D, Dover GA (1985) Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. EMBO J 4:1701–1708

Sueoka N (1961) Variation and heterogeneity of base composition of deoxyribonucleic acids: a compilation of old and new data. J Mol Biol 3:31–40. https://doi.org/10.1016/S0022-2836(61)80005-3

Sun J, Zhang Y, Xu T et al (2017) Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nat Ecol Evol 1:1–7. https://doi.org/10.1038/s41559-017-0121

Szybalski W (1968) Use of cesium sulfate for equilibrium density gradient centrifugation. In: Methods in enzymology. Academic Press, New York, pp 330–360

Takeuchi T, Kawashima T, Koyanagi R et al (2012) Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA Res 19:117–130. https://doi.org/10.1093/dnares/dss005

Talbert PB, Henikoff S (2020) What makes a centromere? Exp Cell Res 111895. https://doi.org/10.1016/j.yexcr.2020.111895

Thomas J, Pritham EJ (2015) Helitrons, the eukaryotic rolling-circle transposable elements. Microbiol Spectr 3:1–32. https://doi.org/10.1128/microbiolspec.MDNA3-0049-2014

Tørresen OK, Star B, Mier P et al (2019) Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res 47:10994–11006. https://doi.org/10.1093/nar/gkz841

Tunjić Cvitanić M, Vojvoda Zeljko T, Pasantes JJ et al (2020) Sequence composition underlying Centromeric and heterochromatic genome compartments of the Pacific oyster *Crassostrea gigas*. Genes (Basel) 11:695. https://doi.org/10.3390/genes11060695

Ugarković Đ, Plohl M (2002) Variation in satellite DNA profiles-causes and effects. EMBO J 21:5955–5959

Ugarković D, Podnar M, Plohl M (1996) Satellite DNA of the red flour beetle *Tribolium castaneum*-comparative study of satellites from the genus *Tribolium*. Mol Biol Evol 13:1059–1066

Uliano-Silva M, Dondero F, Dan Otto T et al (2018) A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel, *Limnoperna fortunei*. Gigascience 7:1–10. https://doi.org/10.1093/GIGASCIENCE/GIX128

Utsunomia R, Ruiz-Ruano FJ, Silva DMZA et al (2017) A glimpse into the satellite DNA library in characidae fish (Teleostei, Characiformes). Front Genet 8:1–11. https://doi.org/10.3389/fgene.2017.00103

Utsunomia R, de Silva DMZA, Ruiz-Ruano FJ et al (2019) Satellitome landscape analysis of *Megaleporinus macrocephalus* (Teleostei, Anostomidae) reveals intense accumulation of satellite sequences on the heteromorphic sex chromosome. Sci Rep 9:1–10. https://doi.org/10.1038/s41598-019-42383-8

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018) The third revolution in sequencing technology. Trends Genet 34:666–681. https://doi.org/10.1016/j.tig.2018.05.008

Vaughn CC, Hoellein TJ (2018) Bivalve impacts in freshwater and marine ecosystems. Annu Rev Ecol Evol Syst 49:183–208. https://doi.org/10.1146/annurev-ecolsys-110617-062703

Vojvoda Zeljko T, Pavlek M, Meštrović N, Plohl M (2020) Satellite DNA-like repeats are dispersed throughout the genome of the Pacific oyster *Crassostrea gigas* carried by Helentron non-autonomous mobile elements. Sci Rep 10:1–12. https://doi.org/10.1038/s41598-020-71886-y

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

Wang Y, Xu Z, Guo X (2001) A centromeric satellite sequence in the Pacific oyster (*Crassostrea gigas* Thunberg) identified by fluorescence *in situ* hybridization. Mar Biotechnol 3:486–492. https://doi.org/10.1007/s10126-001-0063-3

Wang S, Lorenzen MD, Beeman RW, Brown SJ (2008) Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. Genome Biol 9:1–14. https://doi.org/10.1186/gb-2008-9-3-r61

Wang X, Li Q, Lian J et al (2014) Genome-wide and single-base resolution DNA methylomes of the Pacific oyster *Crassostrea gigas* provide insight into the evolution of invertebrate CpG methylation. BMC Genomics 15:1119. https://doi.org/10.1186/1471-2164-15-1119

Wang S, Zhang J, Jiao W et al (2017) Scallop genome provides insights into evolution of bilaterian karyotype and development. Nat Ecol Evol 1:1–12. https://doi.org/10.1038/s41559-017-0120

Wang X, Xu W, Wei L et al (2019) Nanopore sequencing and *De novo* assembly of a black-shelled Pacific oyster (*Crassostrea gigas*) genome. Front Genet 10:1211. https://doi.org/10.3389/fgene.2019.01211

Waring M, Britten RJ (1966) Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. Science 154:791–794. https://doi.org/10.1126/science.154.3750.791

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

Yunis JJ, Yasmineh WG (1971) Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation. Science 174:1200–1209. https://doi.org/10.1126/science.174.4015.1200

Zhang G, Fang X, Guo X et al (2012) The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490:49–54. https://doi.org/10.1038/nature11413

Zhang M, Peng WF, Hu XJ et al (2018) Global genomic diversity and conservation priorities for domestic animals are associated with the economies of their regions of origin. Sci Rep 8:1–12. https://doi.org/10.1038/s41598-018-30061-0

Zhang J, Luo S, Gu Z et al (2020) Genome-wide DNA methylation analysis of mantle edge and mantle central from pearl oyster *Pinctada fucata martensii*. Mar Biotechnol 22:380–390. https://doi.org/10.1007/s10126-020-09957-4

Chapter 4
Satellite DNA Is an Inseparable Fellow Traveler of B Chromosomes

Juan Pedro M. Camacho, Francisco J. Ruiz-Ruano, María Dolores López-León, and Josefa Cabrero

Abstract Next-Generation Sequencing (NGS) has revealed that B chromosomes in several species are enriched in repetitive DNA, mostly satellite DNA (satDNA). This raises the question of whether satDNA is important to B chromosomes for functional reasons or else its abundance on Bs is simply a consequence of properties of B chromosomes such as their dispensability and late replication. Here we review current knowledge in this respect and contextualize it within the frame of practical difficulties to perform this kind of research, the most important being the absence of good full genome sequencing for B-carrying species, which is an essential requisite to ascertain the intragenomic origin of B chromosomes. Our review analysis on 16 species revealed that 38% of them showed B-specific satDNAs whereas only one of them (6%) carried an inter-specifically originated B chromosome. This shows that B-specific satDNA families can eventually evolve in intraspecifically arisen B chromosomes. Finally, the possibility of satDNA accumulation on B chromosomes for functional reasons is exemplified by B chromosomes in rye, as they contain B-specific satDNAs which are transcribed and occupy chromosome locations where they might facilitate the kind of drive shown by this B chromosome during pollen grain mitosis.

Keywords Satellite DNA · B chromosome · Heterochromatin · Repeatome

4.1 Introduction

Since satellite DNA (satDNA) was uncovered by Kit (1961) through cesium chloride (CsCl) sedimentation analysis, and its repetitive nature was shown by Waring and Britten (1966), its possible presence on supernumerary (B) chromosomes was soon claimed in the grasshopper *Myrmeleotettix maculatus* (Gibson and Hewitt 1970,

J. P. M. Camacho (✉) · F. J. Ruiz-Ruano · M. D. López-León · J. Cabrero
Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain
e-mail: jpmcamac@ugr.es

Đ. Ugarković (ed.), *Satellite DNAs in Physiology and Evolution*, Progress in Molecular and Subcellular Biology 60,
https://doi.org/10.1007/978-3-030-74889-0_4

1972). However, the fact that Chilton and McCarthy (1973) did not find differences in the buoyant density distribution in CsCl gradients between 0B and 5B maize plants, and Timmis et al. (1975) reached similar results on B chromosomes in rye, appeared to support the conclusion that DNA of B chromosomes is roughly similar to that of standard (A) chromosomes. Likewise, Klein and Eckhardt (1976) found no significant differences between the buoyant densities or thermal denaturation profiles of B-carrying and B-lacking DNA in the mealy bug. Similarly, Dover (1975) did not find any new highly repetitious DNA family related to the presence of B chromosomes in *Aegilops*, by using different approaches such as comparisons of the percentage of heterologous associations in DNA/DNA hybridization experiments. Finally, the fact that Dover and Henderson (1976) reanalyzed the *M. maculatus* case and did not find any satellite to the main band in grasshoppers with 0, 1, or 2 B chromosomes made prevailing the conclusion that A and B chromosomes show considerable similarity in base composition.

Remarkably, 5 years later, G. Dover himself, in collaboration with A. Amos, showed the presence of satellite DNAs shared between A and B chromosomes in tsetse flies (*Glossina austeni* and *G. morsitans morsitans*) by means of CsCl gradient density centrifugation and radioactive in situ hybridization. They suggested the first molecular mode of origin for a B chromosome, with important involvement of the Y sex chromosome and satDNA amplification (Amos and Dover 1981). In maize, Peacock et al. (1981) isolated a 185-bp satellite DNA by CsCl gradient centrifugation and cloned it into a plasmid for its chromosomal mapping by in situ hybridization, concluding that this satDNA is present on A chromosomes knob heterochromatin and on the long arm proximal knob of the B chromosome, but at lower copy number. However, the first demonstration of the existence of B-specific satDNA should have to wait until Nur et al. (1988) who showed that the paternal sex ratio B chromosome (PSR) in the wasp *Nasonia vitripennis* contains at least three B-specific tandem repetitive DNAs (see also Eickbush et al. 1992). B-specific satDNAs were soon found in rye (*Secale cereale*) (Sandery et al. 1990; Blunden et al. 1993), maize (*Zea mays*) (Alfenito and Birchler 1993), and the Australian daisy (*Brachycome dichromosomatica)* (John et al. 1991; Houben et al. 1997), at the same time as satDNAs shared between A and B chromosomes were found in other species, e.g., the grasshopper *Eyprepocnemis plorans* (López-León et al. 1994, 1995), the fly *Drosophila subsilvestris*, where the pSsP216 satDNA might have arisen from the dot chromosome (Gutknecht et al. 1995), and the fish *Prochilodus lineatus* (de Jesus et al. 2003) (see other examples in Camacho 2005).

These observations revealed that B chromosomes at different species may contain a heterogeneous sample of satDNAs, some of them being also located on A chromosomes and others apparently being B-specific. The first hypothesis on the origin of these B-specific satDNAs was suggested by Langdon et al. (2000), who described the de novo creation of satellite repeats, from complex euchromatic sequences, on the rye B chromosome. Therefore, by the end of the twentieth century, it was already clear that satDNA may actually be a major constituent of B chromosomes, in some cases being satDNAs shared with the A chromosomes and, in others, being B-specific satDNAs which might have arisen de novo on the B chromosome itself or else reflecting B chromosome origin through hybridization, as was the case of *N. vitripennis* (see below).

4.2 SatDNA Is the Prevalent Repetitive DNA on B Chromosomes in Some Species

The arrival of the twenty-first century witnessed the advent of powerful DNA sequencing methods able to yield huge amounts of DNA sequences with low input of time and cost, i.e., the so-called Next-Generation Sequencing (NGS) (Mardis 2008). These NGS technologies were soon applied to the analysis of repetitive DNA thanks to the development of software, such as RepeatExplorer (RE), being able to assemble repetitive DNA from short Illumina sequences (Novák et al. 2013).

The first report on satDNA content for a B chromosome, by means of RE analysis, was performed in rye (Martis et al. 2012) and revealed that these B chromosomes contain a similar proportion of repeats as A chromosomes, but Bs showed an additional massive accumulation of B-specific satDNAs, which were characterized by exceptionally long monomers (0.9–4.0 kb), some of them suggesting chimeric origins. In addition, rye B chromosomes showed accumulation of Bianka Ty1/copia elements and plastid (NUPT)- and mitochondrial (NUMT)-derived sequences. Banaei-Moghaddam et al. (2013) later suggested that the accumulation of some of these repetitive sequences (i.e., mobile elements and satDNA) promoted the pseudogenization of many genes in the B chromosome. Likewise, Klemme et al. (2013) observed that B-enriched satellites were mostly accumulated in the nondisjunction control region of the rye B chromosome.

The finding of B chromosomes in *Drosophila melanogaster* (Bauerly et al. 2014) opened the possibility to analyze B chromosomes with the help of all kinds of tools amenable in this model species. These B chromosomes were primarily composed of the AATAT satellite sequence which is also characteristic of autosome 4. Later, Hanlon et al. (2018) reported that this microsatellite showed FISH (fluorescent in situ hybridization) bands on the B, autosome 4, X and Y chromosomes. In addition, these authors found another microsatellite (AAGAT) which only hybridized with the B chromosomes and autosome 4, and suggested the possible origin of B chromosomes from this A chromosome. They also found that *D. melanogaster* B chromosomes did not carry any known euchromatic sequence and that they are rich in transposable elements and long arrays of short nucleotide repeats, the most abundant being the AAGAT microsatellite. Likewise, Milani and Cabral-de-Mello (2014) suggested that microsatellites are important components of the B chromosome in the grasshopper *Abracris flavolineata*. These authors later showed, through RE analysis, that this B chromosome shows a 137 bp satDNA (AflaSAT-1) which is shared with A chromosomes (Milani et al. 2017).

The development of the satMiner protocol, based on reiterative searches for repetitive elements on Illumina reads, by means of RepeatExplorer, allowed us to

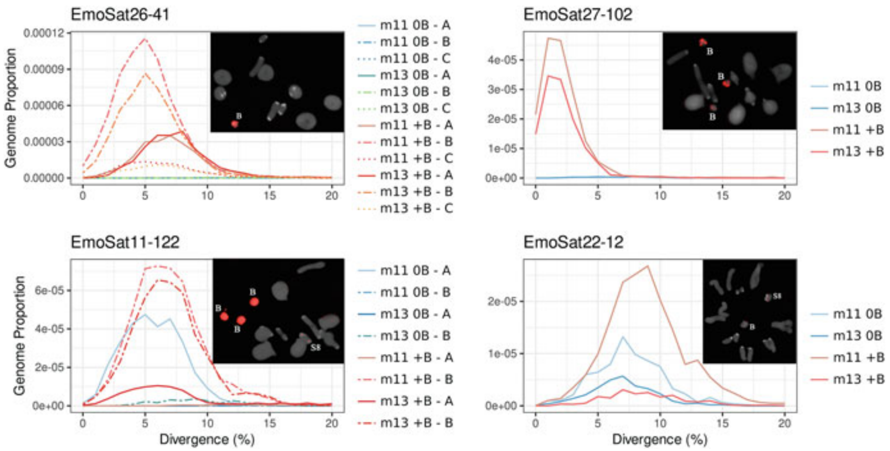


Fig. 4.1 Presence of SatDNA on B chromosomes of the grasshopper *Eumigus monticola*. Two B-lacking (m11_0B and m13_0B, blue lines) and two B-carrying (m11_+B and m13_+B, red lines) males were analyzed by Illumina sequencing (satMiner protocol) and FISH. Two satDNA families (EmoSat26-41 and EmoSat27-102) (upper row) appeared to be B-specific because they yielded large FISH signals on the B chromosomes but no signal on A chromosomes. However, bioinformatic analysis on Illumina reads showed that they are also present on A chromosomes although at very low abundance (note the blue lines, close to the *X*-axis, in the repeat landscapes). Note that EmoSat26-41 showed diversification into three subfamilies (**A**–**C**) amplified on the B chromosome (red lines). The lower row shows two satDNA families showing FISH signals only on the B and S8 chromosomes, thus suggesting the possible B origin from this A chromosome. Note that EmoSat11-122 shows FISH signals across the whole B chromosome area, and its B subfamily is highly abundant on the two B-carrying individuals (most likely on the B chromosome) whereas the A subfamily was abundant on the S8 autosome of the m11_0B male. On the other hand, EmoSat22-12 showed small FISH bands on the B and S8 chromosomes and differential amplification between the two B-carrying males

build the first satellitome in the migratory locust (Ruiz-Ruano et al. 2016). This led to the characterization of tens of different satDNA families constituting a broad satDNA catalog. SatMiner application to Illumina reads obtained from B-carrying and B-lacking genomes in the grasshopper *Eumigus monticola* uncovered 27 satDNA families whose FISH analysis showed the presence of eight of them on the B chromosome (Ruiz-Ruano et al. 2017). In fact, two of them (EmoSat26-41 and EmoSat27-102) showed FISH bands only on the B chromosome, thus appearing to be B-specific (Fig. 4.1). However, the bioinformatic analysis of abundance indicated that, although extremely scarce, they were also present in the 0B genome. Therefore, the most parsimonious explanation for the differential abundance of these two satDNAs on the B chromosomes is their major amplification on the B chromosome. Another satDNA family (EmoSat11-122) was extremely abundant on the B chromosome whereas it showed a single small FISH band on autosome S8 (Fig. 4.1). Remarkably, this satDNA family was composed of two different subfamilies (named A and B), and a repeat landscape (RL), quantifying genomic abundance at 1% divergence intervals, showed that subfamily B was highly amplified on the B chromosomes whereas subfamily A was highly amplified on the S8 autosome of only one of the two 0B males analyzed (Fig. 4.1). Finally, EmoSat22-12 also showed FISH bands on the B and S8 chromosomes, revealing its amplification on B chromosomes in only one of the two B-carrying males analyzed by satMiner (Fig. 4.1).

The same year, Kumke et al. (2016) showed that the 5S rDNA-derived PLsatB satellite DNA found in *Plantago lagopus* makes up 3.3% of the 1B genotype but only 0.09% of the 0B genotype (see also Dhar et al. 2019). In the mice *Apodemus flavicollis* and *A. peninsulae*, Makunin et al. (2018) analyzed B chromosome content by means of low-pass single-chromosome sequencing and found accumulation of repetitive DNA, mainly satDNA and endogenous retroelements. In the plant *Aegilops speltoides*, however, Wu et al. (2019) have shown that the repetitive fraction of the genome "is mostly composed of LTR-retrotransposons, transposons and satellite repeats, with overall proportions of individual repeat types being similar in the 0B/+B genotypes."

In the migratory locust, *Locusta migratoria*, Ruiz-Ruano et al. (2018) performed the quantification of repetitive DNA content of B chromosomes. They found that about 64.1% of the B-lacking genome consists of repetitive DNA, whereas this figure was higher in B-carrying genomes (64.6%) due to B chromosome enrichment in repetitive DNA. Using a subtractive approach, we found that 94.9% of the B chromosome DNA was repetitive. Specifically, 65.2% was satDNA, whereas the most abundant TEs only reached 7.9% for DNA transposons and 7% for LINEs. In addition, several gene families were found on this B chromosome, such as histone genes (2.7%), as reported previously by Teruel et al. (2010), 45S (0.25%) and 5S (0.78%) rRNA genes, snRNA genes (1.3%), especially U2 (1.1%), and tRNA genes (0.7%). Remarkably, about half of the DNA content of the B chromosome corresponded to a single satDNA (LmiSat02-176) whereas all remaining repetitive elements showed abundances lower than 4%. In fact, FISH for this satDNA family yielded a signal occupying the whole B chromosome area (Fig. 4.2). In addition, five other satDNAs showed FISH bands on the B chromosome of different sizes (Fig. 4.2), including the telomeric DNA (LmiSat07-5-tel) and the most abundant satDNA in the *L. migratoria* genome (LmiSat01-185) which is likely involved in the centromeric function since it is the only satDNA being located on all A chromosomes. On the B chromosome, however, this satDNA family is actually scarce (see Fig. S1b in Ruiz-Ruano et al. 2018). The bioinformatic analysis of satDNA abundance on two 0B and four B-carrying males, displayed two main patterns for their RLs (Fig. 4.3), with five satDNA families showing overabundance in all four B-carrying males (LmiSat02-176, LmiSat04-18, LmiSat09-181, LmiSat10-9, and LmiSat16-278) whereas three other showed overabundance in only two B-carrying males (LmiSat06-185, LmiSat18-210, and LmiSat53-47), suggesting the existence of polymorphic B chromosomes in the population analyzed for an abundance of some satDNA families. In addition, about half of these satDNAs showed RLs with a maximum peak of about 4–5% divergence (likewise LmiSat01-185) whereas the remaining families showed their maximum peak at 0% divergence (in resemblance to LmiSat07-5-tel) (Fig. 4.3). As the telomeric DNA is actively homogenized by the

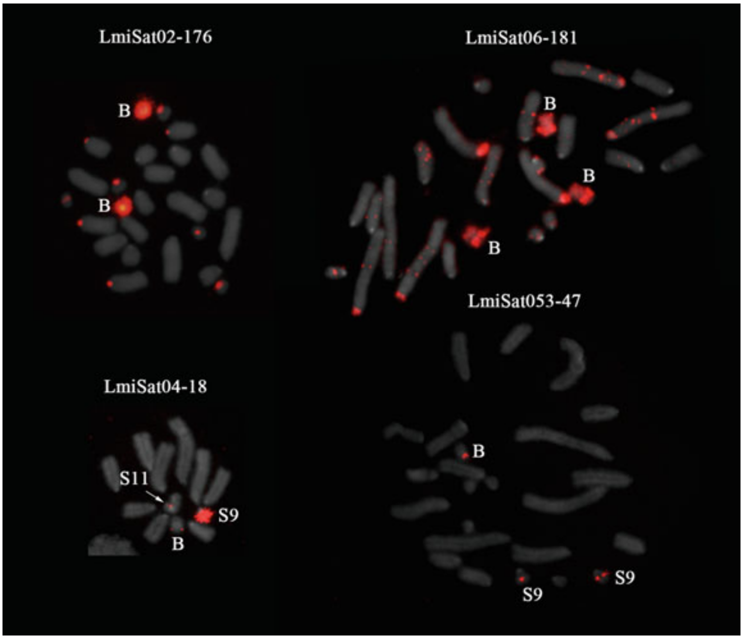


Fig. 4.2 FISH pattern for four SatDNA families in B-carrying embryos of the migratory locust (*Locusta migratoria*). Two of them (upper row) showed large bands which, in LmiSat02-176, occupy the whole B chromosome area, and this satDNA represented about half of the whole B chromosome DNA. No B-specific satDNAs were found in this species, as they were also present on A chromosomes, on exclusively pericentromeric locations for LmiSat02-176 but pericentromeric and interstitial locations for LmiSat06-181. The lower row shows two satDNA families yielding FISH bands on the S9 and B chromosomes, and also on S11 in the case of LmiSat04-18 (note that this cell is haploid)

telomerase during each DNA replication, we can infer that the satDNAs displaying their maximum peak at 0% divergence have undergone recent amplification by the addition of many identical repeat units. On the other hand, those families showing their maximum peak at 4–5% showed their last major amplification some time ago so that point mutations have increased divergence for many repeat units added during that amplification. Ruiz-Ruano et al. (2018) called attention to the contrasting difference found between A and B chromosomes in *L. migratoria*, as TEs comprise 54% of total DNA in A chromosomes but only 19.1% of the B chromosomes. However, satDNA comprises 65.2% of the B chromosome but only 2.4% of A chromosomes. If these B chromosomes are derived from the A chromosomes, it is clear that they have followed a different molecular evolutionary pattern through a marked enrichment in satDNA.

In the same vein, Benetta et al. (2020) have recently shown that 89.80% of the PSR chromosome in *Nasonia vitripennis* is composed of repetitive DNA, the most abundant being complex satellites belonging to four main families (70.32%), three of which were B-specific (PSR2, PSR18, and PSR22) and the other was shared with the

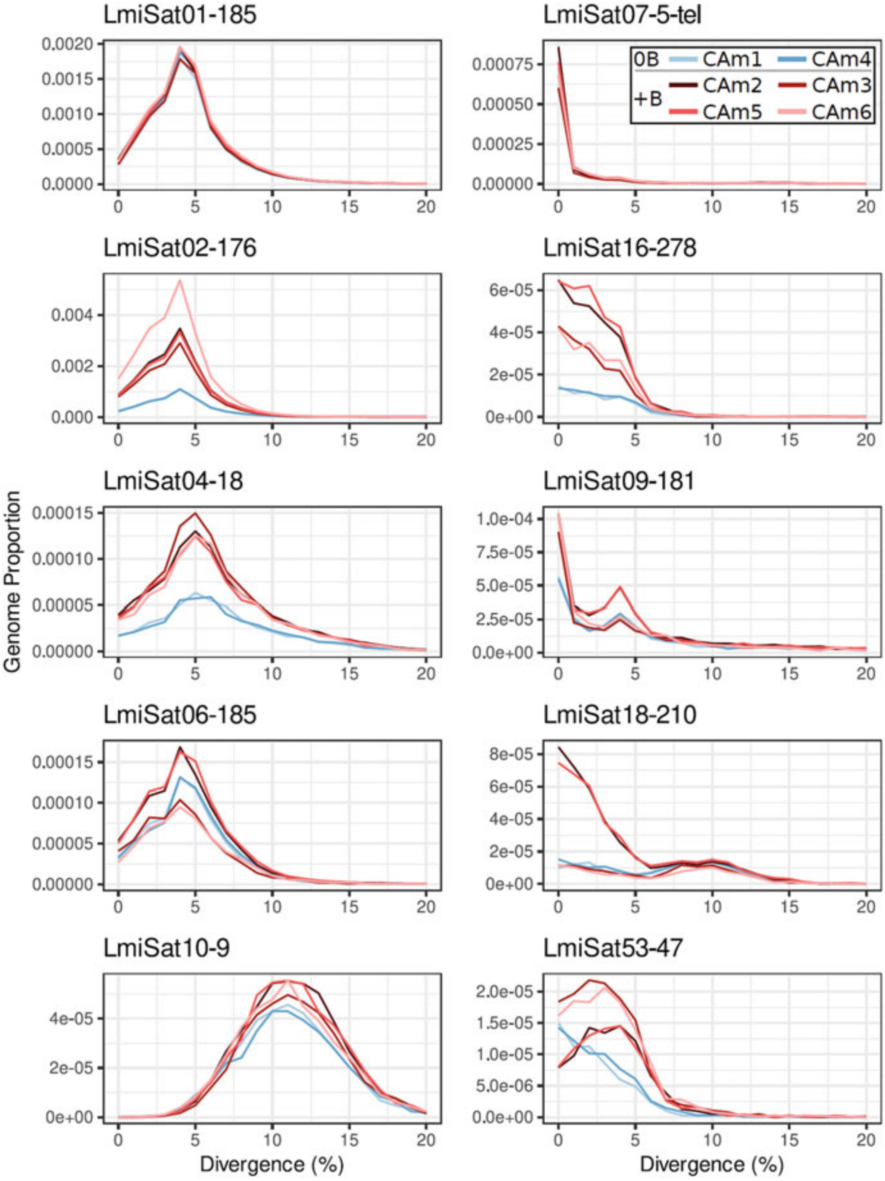


Fig. 4.3 Repeat landscapes for eight satDNA families showing overabundance on the B chromosome of *Locusta migratoria*, compared with the patterns shown by the centromeric (LmiSat01-185) and telomeric (LmiSat07-5-tel) satDNAs (upper row). Note that the left column shows satDNA families showing their maximum peak of abundance at 4–5% divergence, in resemblance to the centromeric satDNA, suggesting that they have shown a certain degree of degeneration since their last major amplification. The only exception was LmiSat10-9 whose peak was placed at 10% divergence, presumably due to faster degeneration of its extremely short repeat units (9 bp). Within the left column, note that LmiSat02-176, LmiSat04-18, and LmiSat10-9 showed overabundance (in respect to B-lacking males) in the four B-carrying males (red lines), whereas LmiSat06-185 showed this fact only in two out of the four B-carrying males analyzed, suggesting the existence of

Table 4.1 The presence of satDNA families on B chromosomes either being shared with A chromosomes or else being B-specific

Type	Species	A-B_shared	B-specific	References
Animals	*Glossina austeni*	Yes		Amos and Dover (1981)
	G. morsitans morsitans	Yes		Amos and Dover (1981)
	Nasonia vitripennis	Yes	Yes	Nur et al. (1988), Eickbush et al. (1992), Benetta et al. (2020)
	Eyprepocnemis plorans	Yes		López-León et al. (1994, 1995)
	Drosophila subsilvestris	Yes		Gutknecht et al. (1995)
	Prochilodus lineatus	Yes		de Jesus et al. (2003)
	Drosophila melanogaster	Yes		Bauerly et al. (2014), Hanlon et al. (2018)
	Moenkhausia sanctafilomenae	Yes		Utsunomia et al. (2016)
	Eumigus monticola	Yes	Yes	Ruiz-Ruano et al. (2017)
	Abracris flavolineata	Yes		Milani et al. (2017)
	Astyanax paranae	Yes	Yes	Silva et al. (2017)
	Apodemus			Makunin et al. (2018)
	Locusta migratoria	Yes		Ruiz-Ruano et al. (2018)
	Characidium gomesi	Yes		Serrano-Freitas et al. (2020)
Plants	*Secale cereale*	Yes	Yes	Sandery et al. (1990), Blunden et al. (1993), Klemme et al. (2013)
	Brachycome dichromosomatica	Yes	Yes	John et al. (1991), Houben et al. (1997)
	Plantago lagopus	Yes	Yes	Kumke et al. (2016)
	Aegilops speltoides			Wu et al. (2019)

A chromosomes (NV79). However, only 13.97% of B-sequences corresponded to TEs. In summary, the information on the species mentioned in this section, reveals that six out of the 16 species where chromosome location was analyzed (Table 4.1),

Fig. 4.3 (continued) B chromosome polymorphism for the abundance of some satDNAs. The right column, however, shows satDNA families showing curves resembling the telomeric DNA pattern, i.e., with the maximum peak at 0% divergence. The only exception was LmiSat53-47 where the curves for B-carrying males showed a peak at about 3% divergence, indicating some degeneration on the B chromosome copies. These 0% peaks suggest recent amplification of these satDNA families on the B chromosome showed B-specific satDNA families (i.e., 38%). In the case of *N. vitripennis*, the presence of B-specific satDNA is explained by the interspecific origin of its B chromosome (McAllister and Werren 1997), but the B chromosomes in the other five species appeared to have originated intraspecifically, so that B-specific satDNAs were most likely originated in them by means of differential amplification on the B chromosomes, as explained above for EmoSat26-41 and EmoSat27-102.

4.3 SatDNA as Marker of B Chromosome Origin

During the pre-NGS times, B chromosome origin was mainly delimited between intra- or interspecific origins (see some examples in Table S1 in Ruiz-Ruano et al. 2017). Assuming that B chromosomes are most likely derived from A chromosomes, the intragenomic origin of B chromosomes is more amenable for analysis in the case of intraspecifically arisen B chromosomes, although not without serious difficulty. In the grasshopper *E. plorans*, we found evidence for contradictory hypotheses on the intragenomic origin of B chromosomes, as we first inferred that B chromosomes in Spanish populations most likely derived from the X chromosome because the order of a satDNA and rDNA in respect to the centromere was only coincident on the pericentromeric region of B and X chromosomes (López-León et al. 1994). We later observed that B chromosomes from Moroccan populations supported the former conclusion, but those found in populations from Daghestan (North Caucasus, Russia) were most likely derived from the smallest autosome, which was the only A chromosome carrying the three markers (5S and 45S rDNA, and the 180-bp satDNA) found on Caucasian B chromosomes (Cabrero et al. 2003). Later, sequence comparison for ITS rDNA sequences obtained from B chromosomes and several A chromosomes (X, M8, S9, S10, and S11), and for the 180-bp satDNA obtained from the X, B, and S11 chromosomes, through chromosome microdissection in spermatocytes from males collected at the Torrox population (Spain), indicated that B chromosome sequences showed higher similarity with those coming from the smallest autosome (S11) than with those from the X chromosome. This gave support to the hypothesis of B origin from the S11 autosome also in Spanish populations (Teruel et al. 2014).

In *Locusta migratoria*, the exclusive presence of genes for H3 and H4 histones on the B chromosome and the M8 autosome, indicated by FISH analysis, provided evidence for B chromosome origin from this A chromosome (Teruel et al. 2010). However, our FISH analysis of 58 satDNA families, previously found in this species (Ruiz-Ruano et al. 2017), on A and B chromosomes revealed that autosome S9 was the only A chromosome carrying all six satDNAs visualized on the B chromosome (Ruiz-Ruano et al. 2018), with one of them being exclusive of S9 and B (Fig. 4.2), on which basis we concluded that both M8 and S9 chromosomes could have contributed to B chromosome origin in this species.

In rye, NGS analysis allowed the identification of several B-specific repeats, mostly being satDNA (Martis et al. 2012; Klemme et al. 2013). This revealed that rye Bs showed higher ancestry from the 3RS and 7R standard (A) chromosomes, with subsequent accumulation of repeats and genic fragments from other A chromosomes and insertions of organellar DNA (Martis et al. 2012). Remarkably, accumulation of B chromosome-enriched tandem repeats was mainly found in the nondisjunction control region of the B (Klemme et al. 2013), involved in B chromosome drive (Langdon et al. 2000).

In the fish *Moenkhausia sanctafilomenae*, Utsunomia et al. (2016) found two types of B chromosomes both containing the same tandem repeat DNA sequences (18S rDNA, H3 histone genes, and the MS3 and MS7 satDNAs), all of which were together only in the paracentromeric region of autosome pair no. 6, suggesting that the B chromosomes derived from this A chromosome.

The FISH analysis of the full satellitome in the grasshopper *Eumigus monticola* (Ruiz-Ruano et al. 2017) revealed the presence of two satDNA families being informative for B origin in this species. These were EmoSat11-122 and EmoSat22-12, which only showed FISH bands on the S8 autosome and the B chromosome (Fig. 4.1). As S8 carries an interstitial FISH band for H3 histone genes which is not apparent on the B chromosome, Ruiz-Ruano et al. (2017) suggested the possible origin of this B chromosome from the proximal third of the S8 autosome, thus excluding the H3 cluster. In addition, two other satDNA families (EmoSat26-41 and EmoSat27-102) showed FISH bands only on the B chromosome, and sequence analysis provided evidence that these two satDNAs were actually present, at very low abundance, in the 0B libraries, suggesting that intraspecifically arisen B chromosomes can harbor satDNAs apparently being B-specific at cytogenetic level but not at the genomic level. In contrast, we did not find any B-specific satDNA in *L. migratoria* (Ruiz-Ruano et al. 2018), and Milani et al. (2018) also failed to find B-specific satDNAs in three other grasshopper species, although satDNA location on A and B chromosomes suggested that B chromosomes might have arisen, in all three cases, from one of the three shortest autosomes.

In the fish *Characidium gomesi*, satellitome analysis was also useful to get insights on the intragenomic origin of B chromosomes. Chromosome painting analysis suggested that B chromosomes in this species most likely derived from sex chromosomes (Pansonato-Alves et al. 2014), and subsequent FISH analysis of 18 satDNA families and the comparison of DNA sequences, obtained through chromosome microdissection, supported this hypothesis (Serrano-Freitas et al. 2020).

All these cases point to specific A chromosomes that could have been ancestors of B chromosomes that arisen intraspecifically. However, a recent analysis of B chromosome content for protein-coding genes is revealing that B chromosome origin appears to be multichromosomal, as claimed, for instance, in rye (Martis et al. 2012), the fish *Astatotilapia latifasciata* (Valente et al. 2014), and the plant *Aegilops speltoides* (Ruban et al. 2020). On this basis, the results obtained only from satDNA location should be interpreted with caution. Of course, we agree with Ruban et al. (2017) that conclusions from satDNA location remain provisional, given the dynamic behavior of satDNA repeats, as it is expected that synteny similarity for protein-coding genes would be more reliable for inferring A chromosome contribution to B chromosome origin. Unfortunately, none of the three cases above are exempt from problems, as the multichromosomal origin was not concluded using the same species genome as a reference, which dilutes the synteny advantage, and they used NGS sequences obtained either from microdissected B chromosomes, whose reliability is low for single-copy DNA, or from flow-sorted B chromosomes where minimal contamination with A chromosomes would unavoidably yield the multichromosomal pattern of B chromosome origin.

In fact, very few B-carrying species have their standard genome sequenced. One of these species is *Locusta migratoria*, but its genome has two serious problems since it has not yet reached the chromosome level (Wang et al. 2014) and was obtained from a B-carrying individual with the consequent interference of B-chromosome sequences for genome assembly (see Ruiz-Ruano et al. 2018, 2019). The B-carrying species with the best-sequenced genome is *N. vitripennis*, but a recent analysis of DNA content of its PSR chromosome has shown that it consists primarily of three complex repeats (70.32%) and other sequences that are undetectable in the standard genome and, in some cases, have strong similarity with genes from other organisms (Benetta et al. 2020), a logical expectation for B chromosomes of interspecific origin.

The origin of small supernumerary marker chromosomes (sSMCs) in humans, however, is easier to infer due to the high quality of the human genome and their youth as extra chromosomes, compared with B chromosomes, as this makes it easier inferring their A chromosome ancestry because they still have not had time for undergoing many changes in sequence or structure (Fuster et al. 2004). Recently, Makunin et al. (2018) performed the low-pass sequencing of a human sSMC derived from the pericentromeric region of chromosome 15 long arm. Interestingly, they could map the two breakpoints that yielded this extra chromosome and both were located within alphoid satDNA. They suggested that “sSMCs might correspond to an early stage of B chromosome evolution, and acquisition of drive mechanism for more efficient transmission could in principle transform those elements into true B chromosomes.” However, in our opinion, sSMCs (and any other kind of extra chromosome, included the so-called proto-Bs) lacking drive from their very origin would most likely be eliminated before they could gain drive and thus reach the polymorphism status in natural populations, and this kind of evolutionary models assuming that drive can be obtained a posteriori are all flawed (for instance, see Martis et al. 2012). In the case of sSMCs in humans, their deleterious effects make it even more unlikely their conversion into B chromosomes. In fact, there are no examples of polymorphic sSMCs in human populations beyond spontaneous cases of independently arisen ones (Fuster et al. 2004).

Evidence for possible differential geographical patterns of chromosomal location for B chromosome content of repetitive DNA in *E. plorans* was found by López-León et al. (2008), whose FISH analysis revealed that B chromosomes from Daghestan, Armenia, Turkey, and Greece were mostly composed of rDNA, whereas those from Spain and Morocco contained about similar amounts of rDNA and a 180-bp satDNA, the latter being actually scarce in eastern Bs. The development of a sequence characterized amplified region (SCAR) marker located at intergenic regions of 45S rDNA provided additional support to the intraspecific origin of B chromosomes in *E. plorans*, as its DNA sequence was identical in B chromosome variants from several localities from Spain and Morocco, and it was highly similar in B chromosome variants from Greece and Armenia. The scarce sequence variation observed between such distant populations suggested either a functional constraint or, most likely, a recent and unique origin for B chromosomes in this species (Muñoz-Pajares et al. 2011). The later finding that the widespread geographical distribution of the B1 variant makes it the best candidate for being the ancestor B chromosome in the whole western Mediterranean region (Cabrero et al. 2014) is consistent with a recent origin of B chromosomes in this species.

In some cases, B chromosomes showing highly similar size and morphology have been found in closely related species, and the high-throughput analysis of satDNA has revealed interesting insights on B chromosome origin. This was the case of several fish species of the genus *Astyanax*, where Silva et al. (2016) analyzed the large metacentric B chromosomes in *A. bockmanni*, *A. paranae*, and *A. fasciatus*, by means of (1) chromosome painting, (2) FISH for 18S rDNA, the H1 histone genes, the As51 satDNA and the $(AC)_{15}$ microsatellite, and (3) ITS rDNA sequence comparison between genomic DNA from B-lacking individuals and DNA obtained from the metacentric B chromosomes in the two latter species. Whereas approaches 1 and 3 suggested the common origin of B chromosomes at different species, approach 2 failed to do it. Subsequent analysis of the satellitome in *A. paranae* revealed the presence of 45 satDNA families, 35 of which were analyzed by FISH in *A. paranae*, *A. fasciatus*, and *A. bockmanni*, showing that most satDNA families were shared between the three species and showed highly similar patterns on their B chromosomes (Silva et al. 2017). The exceptions were two B-specific satDNAs in *A. paranae* (ApaSat44-21 and ApaSat20-18), the former not being observed on A or B chromosomes of the two other species and the latter showing FISH bands on them. The symmetric location of many satDNAs on both B chromosome arms demonstrated the isochromosome nature of these large metacentric B chromosomes, and their high similarity in satDNA content and location gave additional support to the common origin hypothesis for these B chromosomes. Recently, the analysis of gene content in the large metacentric B chromosomes of these three species plus *A. scabripinnis*, by means of the genome and transcriptome sequencing and qPCR, has revealed that the Bs in the four species showed such high similarity in gene content that cannot be explained by chance, thus giving stronger support to the common origin hypothesis (Silva et al. 2021).

4.4 Function of satDNA for B Chromosomes

If satDNA accumulates into B chromosomes becoming the most abundant DNA type in them, a pertinent question is whether it plays an important function for B chromosomes. In principle, we should not expect that a possible function would have nothing to do with being transcribed, as satDNA typically belong to the noncoding fraction of repetitive. Alternatively, satDNA might accumulate in B chromosomes because they are dispensable and thus tolerate the burden of carrying high amounts of useless DNA, as long as this burden can be faced by the host genome which makes the machinery for DNA replication. It is conceivable that the late replication which characterizes B chromosomes (Fox et al. 1974) might facilitate replication errors leading to satDNA accumulation on B chromosomes. In addition, the dispensability of B chromosomes may make them be prone to these failures in DNA replication, e.g., unequal crossovers, leading to the accumulation of satDNA (and other tandem repeats) on them.

A possible functional role of satDNA was suggested for rye B chromosomes, after finding that the E3900 and D1100 B-specific satDNAs are transcriptionally active in the subterminal domain of the B chromosome, which acts as the nondisjunction control region, with the B-transcripts possibly functioning as structural or catalytic RNA (Carchilan et al. 2007). Recently, Gómez-Aldecoa (2021) has found another B-specific satDNA (ScCL11-1), which is interspersed with E3900 in the nondisjunction control region and is the only satDNA being also located on the pericentromeric region of the B chromosome, where persistent cohesion maintains the two B chromatids together for migration to the generative pole during pollen grain mitosis. Banaei-Moghaddam et al. (2012) suggested the possibility that B-derived RNAs could act as guide molecules to direct protein complexes to specific genomic loci, such as the B pericentromeric regions. However, these authors noticed that it is not known whether the B transcripts act directly or indirectly on B nondisjunction, and suggested the possibility that some protein-coding genes located in the rye B chromosome (Martis et al. 2012) might also play a role in nondisjunction control.

The case of B rye is thus suggestive for a possible function of a specific satDNA to increase B chromosome viability in natural populations through transcription to yield noncoding RNAs facilitating B drive, with or without interaction with proteins also coded by B chromosome themselves, thus interfering with the normal course of cell division regulation. In *N. vitripennis*, the PSR chromosome expresses a unique set of small RNAs derived from several satDNAs (Li et al. 2017) and contains a gene named *haploidizer*, which appears to be involved in the sex conversion which this B chromosome drive is based on (Benetta et al. 2020). This B chromosome system is thus the best positioned in the race of demonstrating the molecular basis of B chromosome drive, even though many details are still unknown.

4.5 Future Directions

The extreme scarcity of quality reference genomes of B-carrying species, at the chromosome level, is a serious handicap to investigate the A chromosome ancestry of B chromosomes. Meanwhile, satellitome analysis constitutes an excellent tool to get some insights in this respect in non-model species. In the case of intraspecifically arisen B chromosomes, the best tool would be gene content and the syntenic resemblance between A and B chromosomes, but it needs previous obtaining of high-quality sequenced genomes of B-lacking individuals in the same B-carrying species, a goal that has not yet been reached for any intraspecifically originated B chromosome.

Regarding a possible function of satDNA for B chromosomes, a first indication could be its transcription, as recent research has shown that both B chromosomes (Huang et al. 2016; Ma et al. 2016; Navarro-Domínguez et al. 2017; Kinsella et al. 2019) and satDNA (Menon and Meller 2012; Ugarkovic 2005; Usakin et al. 2007) are not transcriptionally inert. In fact, transcription of B-specific satDNAs in rye and jewel wasp B chromosomes is suggestive of their possible implication in interfering cell division in favor of the B chromosome (see above). However, for an excellent discussion on the possible functional role of B chromosome transcripts, see Benetta et al. (2019).

Highly interesting prospects for B chromosome research have resulted from recent results in mice, where Akera et al. (2017) found that oocyte spindle asymmetry depends on CDC42 signaling inducing microtubule tyrosination, and thus selfish meiotic drivers could exploit this asymmetry to bias their transmission. Likewise, Iwata-Otsubo et al. (2017) showed that "centromeres with more satellite repeats house more nucleosomes that confer centromere identity, containing the histone H3 variant CENP-A, and bias their segregation to the egg relative to centromeres with fewer repeats," and suggested that "amplified repetitive sequences act as selfish elements by promoting expansion of CENP-A chromatin and increased transmission through the female germline." Finally, Akera et al. (2019) showed that drive depends on slowing meiotic progression, and suggested that "selfish centromeres can be suppressed by regulating meiotic timing." These findings suggest the possibility that satDNA accumulation on B chromosome centromere, along with the transcription of some protein-coding genes harbored by B chromosomes with putative functions to slow meiosis progression, could play a role in B chromosome drive. The possibility to focus B chromosome research on these molecular aspects is thus served, at least in some species.

References

Akera T, Chmátal L, Trimm E et al (2017) Spindle asymmetry drives non-Mendelian chromosome segregation. Science 358(6363):668–672. https://doi.org/10.1126/science.aan0092

Akera T, Trimm E, Lampson MA (2019) Molecular strategies of meiotic cheating by selfish centromeres. Cell 178(5):1132–1144. https://doi.org/10.1016/j.cell.2019.07.001

Alfenito MR, Birchler JA (1993) Molecular characterization of a maize B chromosome centric sequence. Genetics 135(2):589–597. https://www.genetics.org/content/135/2/589.long

Amos A, Dover G (1981) The distribution of repetitive DNAs between regular and supernumerary chromosomes in species of Glossina (tsetse): a two-step process in the origin of supernumeraries. Chromosoma 81(5):673–690. https://doi.org/10.1007/BF00329579

Banaei-Moghaddam AM, Schubert V, Kumke K et al (2012) Nondisjunction in favor of a chromosome: the mechanism of rye B chromosome drive during pollen mitosis. Plant Cell 24 (10):4124–4134. https://doi.org/10.1105/tpc.112.105270

Banaei-Moghaddam AM, Meier K, Karimi-Ashtiyani R, Houben A (2013) Formation and expression of pseudogenes on the B chromosome of rye. Plant Cell 25(7):2536–2544. https://doi.org/10.1105/tpc.113.111856

Bauerly E, Hughes SE, Vietti DR et al (2014) Discovery of supernumerary B chromosomes in *Drosophila melanogaster*. Genetics 196(4):1007–1016. https://doi.org/10.1534/genetics.113.160556

Benetta DE, Akbari OS, Ferree PM (2019) Sequence expression of supernumerary B chromosomes: function or fluff? Genes 10(2):123. https://doi.org/10.3390/genes10020123

Benetta DE, Antoshechkin I, Yang T et al (2020) Genome elimination mediated by gene expression from a selfish chromosome. Sci Adv 6(14):eaaz9808. https://doi.org/10.1126/sciadv.aaz9808

Blunden R, Wilkes TJ, Forster JW et al (1993) Identification of the E3900 family, a second family of rye B chromosome specific repeated sequences. Genome 36(4):706–711. https://doi.org/10.1139/g93-095

Cabrero J, Bakkali M, Bugrov A et al (2003) Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. Chromosoma 112(4):207–211. https://doi.org/10.1007/s00412-003-0264-2

Cabrero J, López-León MD, Ruíz-Estévez M et al (2014) B1 was the ancestor B chromosome variant in the western Mediterranean area in the grasshopper *Eyprepocnemis plorans*. Cytogenet Genome Res 142(1):54–58. https://doi.org/10.1159/000356052

Camacho JPM (2005) B chromosomes. In: Gregory TR (ed) The evolution of the genome. Elsevier, San Diego, pp 223–286. https://doi.org/10.1016/B978-012301463-4/50006-1

Carchilan M, Delgado M, Ribeiro T et al (2007) Transcriptionally active heterochromatin in rye B chromosomes. Plant Cell 19(6):1738–1749. https://doi.org/10.1105/tpc.106.046946

Chilton MD, McCarthy BJ (1973) DNA from maize with and without B chromosomes: a comparative study. Genetics 74(4):605–614. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1212976/

de Jesus CM, Galetti PM, Valentini SR, Moreira-Filho O (2003) Molecular characterization and chromosomal localization of two families of satellite DNA in *Prochilodus lineatus* (Pisces, Prochilodontidae), a species with B chromosomes. Genetica 118(1):25–32. https://doi.org/10.1023/A:1022986816648

Dhar M, Kour J, Kaul S (2019) Origin, behaviour, and transmission of B chromosome with special reference to *Plantago lagopus*. Genes 10(2):152. https://doi.org/10.3390/genes10020152

Dover GA (1975) The heterogeneity of B-chromosome DNA: no evidence for a B-chromosome specific repetitive DNA correlated with B-chromosome effects on meiotic pairing in the Triticinae. Chromosoma 53(2):153–173

Dover GA, Henderson SA (1976) No detectable satellite DNA in supernumerary chromosomes of the grasshopper *Myrmeleotettix*. Nature 259(5538):57–59. https://doi.org/10.1038/260170a0

Eickbush DG, Eickbush TH, Werren JH (1992) Molecular characterization of repetitive DNA sequences from a B chromosome. Chromosoma 101(9):575–583. https://doi.org/10.1007/bf00660317

Fox DP, Hewitt GM, Hall DJ (1974) DNA replication and RNA transcription of euchromatic and heterochromatic chromosome regions during grasshopper meiosis. Chromosoma 45(1):43–62. https://doi.org/10.1007/BF00283829

Fuster C, Rigola MA, Egozcue J (2004) Human supernumeraries: are they B chromosomes? Cytogenet Genome Res 106(2–4):165–172. https://doi.org/10.1159/000079283

Gibson I, Hewitt G (1970) Isolation of DNA from B chromosomes in grasshoppers. Nature 225(5227):67–68. https://doi.org/10.1038/225067a0

Gibson I, Hewitt G (1972) Interpopulation variation in the satellite DNA from grasshoppers with B-chromosomes. Chromosoma 38(2):121–138. https://doi.org/10.1007/BF00326190

Gómez-Aldecoa F (2021) Análisis estructural y funcional del cromosoma B de centeno. Doctoral dissertation, Universidad Complutense de Madrid

Gutknecht L, Sperlich D, Bachmann L (1995) A species specific satellite DNA family of *Drosophila subsilvestris* appearing predominantly in B chromosomes. Chromosoma 103(8):539–544. https://doi.org/10.1007/BF00355318

Hanlon S, Miller DE, Eche S, Hawley RS (2018) Origin, composition and structure of the supernumerary B chromosome of *Drosophila*. Genetics 210(4):1197–1212. https://doi.org/10.1534/genetics.118.300904

Houben A, Leach CR, Verlin D et al (1997) A repetitive DNA sequence common to the different B chromosomes of the genus *Brachycome*. Chromosoma 106(8):513–519. https://doi.org/10.1007/s004120050273

Huang W, Du Y, Zhao X, Jin W (2016) B chromosome contains active genes and impacts the transcription of A chromosomes in maize (*Zea mays L.*). BMC Plant Biol 16(1):88. https://doi.org/10.1186/s12870-016-0775-7

Iwata-Otsubo A, Dawicki-McKenna JM, Akera T et al (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. https://doi.org/10.1016/j.cub.2017.06.069

John UP, Leach CR, Timmis JN (1991) A sequence specific to B chromosomes of *Brachycome dichrosomatica*. Genome 34:739–744. https://doi.org/10.1159/000444873

Kinsella CM, Ruiz-Ruano FJ, Dion-Côté AM et al (2019) Programmed DNA elimination of germline development genes in songbirds. Nat Commun 10:5468. https://doi.org/10.1038/s41467-019-13427-4

Kit S (1961) Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. J Mol Biol 3(6):711–716. https://doi.org/10.1016/S0022-2836(61)80075-2

Klein AS, Eckhardt RA (1976) The DNAs of the a and B chromosomes of the mealy bug, *Pseudococcus obscurus*. Chromosoma 57(4):333–340

Klemme S, Banaei-Moghaddam AM, Macas J et al (2013) High-copy sequences reveal distinct evolution of the rye B chromosome. New Phytol 199(2):550–558. https://doi.org/10.1111/nph.12289

Kumke K, Macas J, Fuchs J et al (2016) Plantago lagopus B chromosome is enriched in 5S rDNA-derived satellite DNA. Cytogenet Genome Res 148(1):68–73. https://doi.org/10.1159/000444873

Langdon T, Seago C, Jones RN et al (2000) De novo evolution of satellite DNA on the rye B chromosome. Genetics 154(2):869–884. https://doi.org/10.1093/genetics/154.2.869

Li Y, Jing XA, Aldrich JC et al (2017) Unique sequence organization and small RNA expression of a “selfish” B chromosome. Chromosoma 126(6):753–768. https://doi.org/10.1007/s00412-017-0641-x

López-León M, Neves N, Schwarzacher T et al (1994) Possible origin of a B chromosome deduced from its DNA composition using double FISH technique. Chromosom Res 2(2):87–92. https://doi.org/10.1007/BF01553487

López-León MD, Vázquez P, Hewitt G, Camacho JPM (1995) Cloning and sequence analysis of an extremely homogeneous tandemly repeated DNA in the grasshopper *Eyprepocnemis plorans*. Heredity 75(4):370–375. https://doi.org/10.1038/hdy.1995.148

López-León MD, Cabrero J, Dzyubenko VV et al (2008) Differences in ribosomal DNA distribution on A and B chromosomes between eastern and western populations of the grasshopper *Eyprepocnemis plorans plorans*. Cytogenet Genome Res 121(3–4):260–265. https://doi.org/10.1159/000138894

Ma W, Gabriel TS, Martis MM et al (2016) Rye B chromosomes encode a functional Argonaute-like protein with in vitro slicer activities similar to its A chromosome paralog. New Phytol 213(2):916–928. https://doi.org/10.1111/nph.14110

Makunin AI, Rajičić M, Karamysheva TV et al (2018) Low-pass single-chromosome sequencing of human small supernumerary marker chromosomes (sSMCs) and *Apodemus* B chromosomes. Chromosoma 127(3):301–311. https://doi.org/10.1007/s00412-018-0662-0

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359ç

Martis MM, Klemme S, Banaei-Moghaddam AM et al (2012) Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. Proc Natl Acad Sci USA 109(33):13343–13346. https://doi.org/10.1073/pnas.1204237109

McAllister BF, Werren JH (1997) Hybrid origin of a B chromosome (PSR) in the parasitic wasp *Nasonia vitripennis*. Chromosoma 106(4):243–253. https://doi.org/10.1007/s004120050245

Menon DU, Meller VH (2012) A role for siRNA in X-chromosome dosage compensation in *Drosophila melanogaster*. Genetics 191(3):1023–1028

Milani D, Cabral-de-Mello DC (2014) Microsatellite organization in the grasshopper *Abracris flavolineata* (Orthoptera: Acrididae) revealed by FISH mapping: remarkable spreading in the A and B chromosomes. PLoS One 9(5):e97956. https://doi.org/10.1371/journal.pone.0097956

Milani D, Ramos É, Loreto V et al (2017) The satellite DNA AflaSAT-1 in the A and B chromosomes of the grasshopper *Abracris flavolineata*. BMC Genet 18:81. https://doi.org/10.1186/s12863-017-0548-9

Milani D, Bardella V, Ferretti A et al (2018) Satellite DNAs unveil clues about the ancestry and composition of B chromosomes in three grasshopper species. Genes 9(11):523. https://doi.org/10.3390/genes9110523

Muñoz-Pajares AJ, Martínez-Rodríguez L, Teruel M et al (2011) A single, recent origin of the accessory B chromosome of the grasshopper *Eyprepocnemis plorans*. Genetics 187 (3):853–863. https://doi.org/10.1534/genetics.110.122713

Navarro-Domínguez B, Ruiz-Ruano FJ, Cabrero J et al (2017) Protein-coding genes in B chromosomes of the grasshopper *Eyprepocnemis plorans*. Sci Rep 7:45200. https://doi.org/10.1038/srep45200

Novák P, Neumann P, Pech J et al (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

Nur U, Werren JH, Eickbush DG et al (1988) A “selfish” B chromosome that enhances its transmission by eliminating the paternal genome. Science 240(4851):512–514. https://doi.org/10.1126/science.3358129

Pansonato-Alves JC, Serrano ÉA, Utsunomia R et al (2014) Single origin of sex chromosomes and multiple origins of B chromosomes in fish genus *Characidium*. PLoS One 9(9):e107169. https://doi.org/10.1371/journal.pone.0107169

Peacock WJ, Dennis ES, Rhoades M, Pryor AJ (1981) Highly repeated DNA sequence limited to knob heterochromatin in maize. Proc Natl Acad Sci USA 78(7):4490–4494. https://doi.org/10.1073/pnas.78.7.4490

Ruban A, Schmutzer T, Scholz U, Houben A (2017) How next-generation sequencing has aided our understanding of the sequence composition and origin of B chromosomes. Genes 8(11):294. https://doi.org/10.3390/genes8110294

Ruban A, Schmutzer T, Wu DD et al (2020) Supernumerary B chromosomes of *Aegilops speltoides* undergo precise elimination in roots early in embryo development. Nat Commun 11:2764. https://doi.org/10.1038/s41467-020-16594-x

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM (2016) High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep 6:28333. https://doi.org/10.1038/srep28333

Ruiz-Ruano FJ, Cabrero J, López-León MD, Camacho JPM (2017) Satellite DNA content illuminates the ancestry of a supernumerary (B) chromosome. Chromosoma 26(4):487–500. https://doi.org/10.1007/s00412-016-0611-8

Ruiz-Ruano FJ, Cabrero J, López-León MD et al (2018) Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. Chromosoma 127(1):45–57. https://doi.org/10.1007/s00412-017-0644-7

Ruiz-Ruano FJ, Navarro-Domínguez B, López-León MD et al (2019) Evolutionary success of a parasitic B chromosome rests on gene content. bioRxiv 683417. https://doi.org/10.1101/683417

Sandery MJ, Forster JW, Blunden R, Jones RN (1990) Identification of a family of repeated sequences on the rye B chromosome. Genome 33(1985):908–913. https://doi.org/10.1139/g90-137

Serrano-Freitas ÉA, Silva DMZA, Ruiz-Ruano FJ et al (2020) Satellite DNA content of B chromosomes in the characid fish *Characidium gomesi* supports their origin from sex chromosomes. Mol Genet Genomics 295(1):195–207. https://doi.org/10.1007/s00438-019-01615-2

Silva DMZA, Daniel SN, Camacho JPM et al (2016) Origin of B chromosomes in the genus *Astyanax* (Characiformes, Characidae) and the limits of chromosome painting. Mol Genet Genomics 291(3):1407–1418. https://doi.org/10.1007/s00438-016-1195-y

Silva DMZA, Utsunomia R, Ruiz-Ruano FJ et al (2017) High-throughput analysis unveils a highly shared satellite DNA library among three species of fish genus *Astyanax*. Sci Rep 7:12726. https://doi.org/10.1038/s41598-017-12939-7

Silva DMZA, Ruiz-Ruano FJ, Utsunomia R et al (2021) Long-term persistence of supernumerary B chromosomes in multiple species of Astyanax fish. BMC Biol 19:52. https://doi.org/10.1186/s12915-021-00991-9

Teruel M, Cabrero J, Perfectti F, Camacho JPM (2010) B chromosome ancestry revealed by histone genes in the migratory locust. Chromosoma 119(2):217–225. https://doi.org/10.1007/s00412-009-0251-3

Teruel M, Ruiz-Ruano FJ, Marchal JA et al (2014) Disparate molecular evolution of two types of repetitive DNAs in the genome of the grasshopper *Eyprepocnemis plorans*. Heredity 112 (5):531–542. https://doi.org/10.1038/hdy.2013.135

Timmis JN, Ingle J, Sinclair J, Jones N (1975) The genomic quality of Rye B chromosomes. J Exp Bot 26(3):367–378. https://doi.org/10.1093/jxb/26.3.367

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

Usakin L, Abad J, Vagin VV et al (2007) Transcription of the 1.688 satellite DNA family is under the control of RNA interference machinery in *Drosophila melanogaster* ovaries. Genetics 176:1343–1349. https://doi.org/10.1534/genetics.107.071720

Utsunomia R, Silva DMZA, Ruiz-Ruano FJ et al (2016) Uncovering the ancestry of B chromosomes in *Moenkhausia sanctaefilomenae* (Teleostei, Characidae). PLoS One 11(3):e0150573. https://doi.org/10.1371/journal.pone.0150573

Valente GT, Conte MA, Fantinatti BEA et al (2014) Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. Mol Biol Evol 31 (8):2061–2072. https://doi.org/10.1093/molbev/msu148

Wang X, Fang X, Yang P et al (2014) The locust genome provides insight into swarm formation and long-distance flight. Nat Commun 5:2957. https://doi.org/10.1038/ncomms3957

Waring M, Britten RJ (1966) Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. Science 154(3750):791–794. https://doi.org/10.1126/science.154.3750.791

Wu DD, Ruban A, Fuchs J et al (2019) Nondisjunction and unequal spindle organization accompany the drive of *Aegilops speltoides* B chromosomes. New Phytol 223(3):1340–1352. https://doi.org/10.1111/nph.15875

Chapter 5
The Genomics of Plant Satellite DNA

Manuel A. Garrido-Ramos

Abstract The twenty-first century began with a certain indifference to the research of satellite DNA (satDNA). Neither genome sequencing projects were able to accurately encompass the study of satDNA nor classic methodologies were able to go further in undertaking a better comprehensive study of the whole set of satDNA sequences of a genome. Nonetheless, knowledge of satDNA has progressively advanced during this century with the advent of new analytical techniques. The enormous advantages that genome-wide approaches have brought to its analysis have now stimulated a renewed interest in the study of satDNA. At this point, we can look back and try to assess more accurately many of the key questions that were left unsolved in the past about this enigmatic and important component of the genome. I review here the understanding gathered on plant satDNAs over the last few decades with an eye on the near future.

Keywords Satellite DNA · Heterochromatin · Centromere · Satellitome

5.1 Plant Satellite DNA

In its simplest definition, satellite DNAs (satDNAs) can be described as noncoding repetitive DNA sequences organized in tandem arrays. satDNAs tandem arrays are distributed throughout the genome of eukaryotic species. Regardless, these tandem arrays are commonly concentrated in specific parts of the chromosomes such as the centromeres, the pericentromeric regions, and the subtelomeric regions. In addition, specific chromosomes can bear interstitial satDNAs. Typically, but not always, satDNA loci are organized as heterochromatic blocks. In fact, satDNAs are the main, but not the only, components of heterochromatin. As a whole, these sequences constitute a significant part of the repetitive DNA content of plant genomes.

M. A. Garrido-Ramos (✉)
Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain
e-mail: mgarrido@ugr.es

Đ. Ugarković (ed.), *Satellite DNAs in Physiology and Evolution*, Progress in
Molecular and Subcellular Biology 60,
https://doi.org/10.1007/978-3-030-74889-0_5

Genome size differs enormously between plant species. There is a 2440-fold range of genome size differences (61–148.791 Mbp) among land plants (Pellicer et al. 2018). Notably, genome size differences may be not only important between phylogenetically distant species but also between closely related species (Pellicer et al. 2018). However, land plant species have a broadly similar number of genes, whereas repetitive DNA is the major contributor to genome size disparity (Zhang et al. 2020; Bowles et al. 2020; Novák et al. 2020a). Novák et al. (2020a) have found that there is a repetitive DNA increase proportional to the size of the genome, reaching up to proportions of around 80% of repetitive DNA in large genomes. Curiously, the repetitive DNA increase does not correlate with genome size increase in genomes larger than 10 Gb. In genomes larger than 10 Gb there is a shift in that trend and the largest genomes have about 55% of repetitive DNA, probably by the slow degradation of repeats over time (Novák et al. 2020a).

Interspersed repeats such as transposable elements (TEs) are the major component of the repetitive DNA fraction of the genome in most species (Piegu et al. 2006; Schnable et al. 2009; Hu et al. 2011; Piednoël et al. 2012; López-Flores and Garrido-Ramos 2012). However, satDNA may comprise between 0.1% and 36% of a plant genome and is often responsible for genome size differences between related species (Macas et al. 2000, 2002; De la Herrán et al. 2001; Hribová et al. 2010; Ambrožová et al. 2011; Čížková et al. 2013; Emadzade et al. 2014; Barghini et al. 2014; Kelly et al. 2015; Yang et al. 2019; Pinosio et al. 2020; Neumann et al. 2020). Furthermore, satellite repeats can be exceptionally the most dominating repetitive elements in a species, as occur in radish (He et al. 2015).

In this context, a major question is whether so many noncoding sequences repeated so many times play a role in the plant genome. Indeed, satDNA has long been considered "junk" DNA, if not "garbage" DNA (Ohno 1972; Brenner 1998; Doolittle 2013; Graur et al. 2013, 2015; Garrido-Ramos 2015). satDNA sequences are among the faster-evolving parts of the eukaryotic genomes, and most satDNAs are species-specific or shared circumstantially by a few related species (Garrido-Ramos 2015, 2017). Consequently, they have often been labeled as useless DNA (Garrido-Ramos 2015). However, satDNA repeats occupy the functional centromere locus and, although these sequences are generally not conserved among species, they constitute part of the scaffolding on which the functional centromere/kinetochore complex is built in most plant species (Melters et al. 2013; Oliveira and Torres 2018). Additionally, there is growing evidence supporting fundamental functional roles of satDNAs in chromosome organization, cell division, and genome regulation (Pezer et al. 2012; Garrido-Ramos 2017).

Actually, satDNA has long been one of the most enigmatic parts of the eukaryotic genome. Earliest studies of plant satDNAs date back to the 1970s and early 1980s. These studies were focused on crop plants such as rye, wheat, barley, maize, mustard, faba bean, or radish (Bedbrook et al. 1980; Dennis et al. 1980; Peacock et al. 1981; Viotti et al. 1985; Capesius 1983; Kato et al. 1984; Grellet et al. 1986), mainly because the economic interest of the species itself, but also because most times these sequences comprised an important part of their genomes. Later, many different satDNAs from a great variety of plants were isolated and analyzed during these and subsequent decades. Genomic DNA digestion using restriction enzymes and visualization of satDNAs repeats in agarose gels after electrophoresis of digested genomic DNA popularized the study of satDNAs (reviewed in Garrido-Ramos 2017). satDNA analysis was benefited of the expanding use of two additional and fundamental laboratory techniques, satDNA-repeats amplification using polymerase chain reaction (PCR) and fluorescent in situ hybridization (FISH), as well as the development of increasingly sophisticated computational tools for the analysis of sequence-specific features or for evolutionary analyses (reviewed in Garrido-Ramos 2017). Thus, conventional techniques (restriction enzyme digestion, PCR amplification and FISH, but also Southern-blot hybridization and others) allowed us to obtain approximate interpretations about satDNA nature. However, a high percentage of this part of the genome remained inaccessible to these techniques. In addition, the study of satDNAs was largely omitted from genomic approaches (Henikoff 2002; Kapustova et al. 2019; Pinosio et al. 2020). It has been during the last years that we have made enormous progress in understanding satDNAs, thanks to its study from a genomic point of view. Nowadays, next-generation sequencing (NGS) and computational approaches have revolutionized the study of satDNAs. The next sections collect some of the most important advances obtained during both periods and the main lessons that we have learned about plant satDNAs on several aspects related to their origin, evolution, organization, and functional roles.

5.2 satDNA Origin and Evolution

SatDNA families can emerge by unequal crossing-over from any random sequence of the genome (Smith 1976). Replication slippage may also generate tandem duplications of any sequence (Dover 1982; Tautz et al. 1986; Stephan 1989; Walsh 1987). These types of events might generate the seed for the formation of a satDNA family through amplification events. Amplification is a rather diffuse term that includes copy number increase in the same locus where the repetitive sequence was generated and its spread to other parts of the genome. A variety of mechanisms have been proposed to explain how such amplification would occur. A major amplification mechanism is unequal crossing-over itself. But two other mechanisms could explain better the spread of the novel satDNA array to other parts of the genome: transposition (Dover 1982; Cooper and Henikoff 2004) and genomic reinsertion of repeats generated by the rolling-circle replication of satDNA extrachromosomal circular molecules (Walsh 1987; Cohen et al. 2005, 2010; Navrátilová et al. 2008).

The very origin of satDNAs, from any random sequence, entails the absence of sequence relationship between the repeats of different satDNA families, both intra- and interspecifically. Another consequence of the way in which these sequences originate is the ease for a genome to accumulate a “library” of satDNAs, i.e., a certain number of different unrelated satDNA families. However, one or a few satDNAs from the library are predominant (i.e., more abundant) in each species (Macas et al. 2015; Avila Robledillo et al. 2018). Usually, the most abundant family of repeats is located in centromeric and pericentromeric regions (Melters et al. 2013), but not always (Peacock et al. 1981; Ananiev et al. 1998a; De la Herrán et al. 2001; Bilinski et al. 2015; Gent et al. 2017, 2018; Avila Robledillo et al. 2018; Finke et al. 2019). Phylogenetically related species might share the ancestor satDNA "library". At times, these species may also share the same predominant satDNA family (Macas et al. 2015; Bolsheva et al. 2019; McCann et al. 2020; Avila Robledillo et al. 2018, 2020). However, differential amplification of each member of the "library" in each species usually originates species-specific abundance profiles for each satDNA family (Fry and Salser 1977). In addition, the loss of old satDNA families or the gain of newly emerged ones can also increase differences in species-specific profiles. In that, unequal crossing-over can be responsible both for the explosive and frequent emergence of new species-specific satDNA families and for their rapid disappearance (Smith 1976; Dover 1982). Therefore, satDNA is the evolutionarily most dynamic component of the genome, with a high turnover rate, by continuous replacements of satDNA families for other in different species (Plohl et al. 2012).

In a neutral scenario, satDNA repeats will diverge over time, generating dozens of variants of the original sequence of a particular satDNA family. New cycles of amplification and homogenization of any of these variants will generate the progressive replacement of the others by the amplified variant. This phenomenon leads to a renewed intraspecific homogenization of the satDNA family (Dover 1982). Concurrently, the differential homogenization of alternative variants in different species will lead to satDNA interspecific divergence. As a consequence, repeats of one species would be more similar than repeats of different species, a pattern known as concerted evolution (Brown et al. 1972; Zimmer et al. 1980; Dover 1982). The intraspecific process would first require intra-chromosome homogenization and then inter-chromosome homogenization (Kawabe and Nasuda 2005; Iwata et al. 2013; Bilinski et al. 2015; McCann et al. 2020) and the intervention of unequal crossing-over as well as other mechanisms such as transposition or genomic reinsertion of replicated extrachromosomal satDNA (Cooper and Henikoff 2004; Grellet et al. 1986; Hall et al. 2003; Cohen et al. 2005, 2010; Navrátilová et al. 2008). In addition to unequal crossing-over or any other amplification mechanism, gene conversion would be also an agent involved in both the intra- and the inter-chromosome homogenization of satDNA (Dover 1982; Grellet et al. 1986; Hall et al. 2003). When homogenization of satDNA repeats occurs mostly within individual chromosomes and there are low rates of inter-chromosomal spread, an inter-chromosomal divergence of satellites is observed (Heslop-Harrison et al. 1999; Macas et al. 2010; McCann et al. 2020), which could lead to the emergence of satDNA subfamilies (see below). On the other hand, extreme cases of interspecific divergence can eventually lead to the appearance of apparently different satDNA families in different species. For example, several distinct species-specific centromeric satDNA families were found in different *Arabidopsis* species, but all of them were derived from a common ancestor (Hallden et al. 1987; Berr et al. 2006; Lermontova et al. 2014) and all share homology with the centromeric satDNAs of other Brassicaceae genera (Martinez-Zapater et al. 1986). Similarly, CentO and CentC centromeric satDNAs in *Oryza* and *Zea*, respectively, appear to share a common origin (Lee et al. 2005; Bilinski et al. 2015).

Independently of its random nature, it would be expected that the process of interspecific divergence would depend on time (Pérez-Gutiérrez et al. 2012). However, many other factors besides time can alter the concerted evolution pattern. Thus, the rate of concerted evolution of satDNAs might be slowed or accelerated by the effect of location, organization, and repeat-copy number of tandem arrays (Navajas-Pérez et al. 2009a). Also, population and evolutionary (Suárez-Santiago et al. 2007; Quesada del Bosque et al. 2013, 2014) or biological factors (Luchetti et al. 2003, 2006; Plohl et al. 2008; Lorite et al. 2017; Ruiz-Ruano et al. 2019) can influence the rate of concerted evolution. In addition, selective constraints may also affect concerted evolution since a particular functional sequence variant could be preserved over any other variant (Mravinac et al. 2005).

As indicated before, the differential homogenization of different divergent repeat variants of a satDNA family within a genome can lead to the appearance of several distinct subfamilies (Grellet et al. 1986; Kawabe and Nasuda 2005; Macas et al. 2006; Kazama et al. 2006; Navajas-Pérez et al. 2005a, 2006; Suárez-Santiago et al. 2007; Torres et al. 2011; Quesada del Bosque et al. 2011, 2013, 2014). These subfamilies can be studied as paralogues since they evolve independently of each other and, consequently, repeats of the same subfamily are more similar among them than when compared with repeats of the other subfamilies. Therefore, sequence similarity within subfamilies is higher than between subfamilies in interspecific comparisons. Consequently, phylogenetic trees group the repeats by subfamily provenance instead of doing it by taxonomic affinity (Suárez-Santiago et al. 2007; Quesada del Bosque et al. 2011, 2013, 2014). Indeed, a group of related species can share a “library” of subfamilies of the same satDNA family too. Differential amplification of each subfamily in different species would also lead to species-specific abundance profiles for each subfamily (Quesada del Bosque et al. 2011, 2013, 2014). As mentioned above for satDNA families, satDNA subfamilies divergence can lead eventually to the split in two different families that shared a common origin as occur between RAE730 and RAYSI satDNAs in *Rumex* (Navajas-Pérez et al. 2005a) or between the 500 bp repeat element and pAL1 satDNAs in *Arabidopsis* centromeres (Simoens et al. 1988; Bauwens et al. 1991; Brandes et al. 1997) or among seven centromeric satDNA families in switchgrass (Yang et al. 2018).

5.3 satDNA Location, Organization, and Function

The principle of the equilocal distribution of heterochromatin proposes that heterochromatic blocks are located in equivalent regions in each chromosome of the karyotype of one species (John et al. 1985). That is to say, there is heterochromatin in the pericentromeric and the subtelomeric regions of all its chromosomes. In addition, there may be interstitial blocks of heterochromatin in equivalent positions on each of the chromosomes of the genome. The equilocality principle would also apply to satDNA since this type of repetitive sequence is the main component of heterochromatin. Sugar beet heterochromatin is organized in large interstitial blocks, in the pericentromeric and centromeric regions and in the subtelomeric region (Kowar et al. 2016). Correspondingly, three different satDNA families, in addition to block-specific LTR retrotransposons, populate each heterochromatin section (Kowar et al. 2016). But also there are small interstitially dispersed heterochromatic spots that are composed of a variety of different satDNAs, DNA transposons, and LTR and non-LTR retrotransposons (Kowar et al. 2016). Indeed, although there are examples that fulfill the equilocality principle (Guerra 2000), there are many exceptions too, especially concerning the interstitial blocks of heterochromatin (Guerra 2000; see below).

5.3.1 Centromeres and Pericentromeric Heterochromatin

There is a huge diversity for satDNA families that populate the centromere and pericentromeric heterochromatin in each species. Conservation is not common. On the contrary, each species presents a specific profile of centromeric/pericentromeric sequences that, at most, can be shared by a small group of related species (Lee et al. 2005; Ma et al. 2007; Wang et al. 2009; Lermontova et al. 2014, 2015; Yu et al. 2017; Avila Robledillo et al. 2018, 2020). Adding complexity, some shared satDNAs may be centromeric in one or a few species whereas they are not centromeric in the rest of the species (Avila Robledillo et al. 2020). A pattern of satDNA replacement in centromeres that we have also found in grasshoppers (Camacho et al. in prep.) and is suggestive to explain the observations made in the genus *Medicago* by Yu et al. (2017). Many species possess only one centromeric satDNA family that is common to all centromeres of all their chromosomes (Martinez-Zapater et al. 1986; Kamm et al. 1995; Zhu et al. 1996; Nagaki et al. 2003b; Ansari et al. 2004; Lee et al. 2005; Bilinski et al. 2015; He et al. 2015; Gent et al. 2017, 2018; Yu et al. 2017). However, different satDNAs may populate different centromeres with a chromosome-specific distribution pattern, as occur in *Arabidopsis lyrata* and *A. halleri* (Kawabe and Nasuda 2005), in several species of Fabaceae (Neumann et al. 2012; Iwata et al. 2013; Avila Robledillo et al. 2018, 2020) or in switchgrass (Yang et al. 2018). In the common bean, two different satDNA families evolving independently are predominantly located at two distinct subsets of centromeres (Iwata et al. 2013). Furthermore, different satDNAs can coexist in the same centromeric locus and/or its pericentromeric region (Bauwens et al. 1991; Brandes et al. 1997; Kawabe and Nasuda 2005; Berr et al. 2006; Macas et al. 2007, 2010; Dluhošová et al. 2018; Avila Robledillo et al. 2018, 2020; Yang et al. 2018). In *Cucumis melo* three different satDNA families are present in the centromeric chromatin of all the chromosomes of the genome (Setiawan et al. 2020). The distribution of satDNA families may be compartmentalized as in *Medicago trunculata* where the *Mt*R3 satDNA is the component of every centromere whereas *Mt*R1 and *Mt*R2 satDNAs are the component of the pericentromeric heterochromatin (Kulikova et al. 2004). Moreover, the centromeric locus as well as the pericentromeric region may bear TEs in addition to satDNA sequences (Neumann et al. 2011). Centromeric-specific retrotransposons (CR) form a clade of chromoviruses, a lineage of Ty3/gypsy retrotransposons (Neumann et al. 2011) that have been found in centromeres of banana (Čížková et al. 2013) and grasses such as barley (Presting et al. 1998; Hudakova et al. 2001), maize (Ananiev et al. 1998b; Zhong et al. 2002; Nagaki et al. 2003a; Sharma and Presting 2014), sorghum (Jiang et al. 1996; Miller et al. 1998; Presting et al. 1998), wheat (Liu et al. 2008; Li et al. 2013), goatgrass (Li et al. 2013), rice (Dong et al. 1998; Miller et al. 1998; Bao et al. 2006; Cheng et al. 2002; Nagaki et al. 2005), rye (Francki 2001), and sugarcane (Nagaki and Murata 2005) among others (Neumann et al. 2011; Sharma and Presting 2014). They have also been found in dicotyledonous species such as *Arabidopsis* (Brandes et al. 1997; Fransz et al. 1998), radish (He et al. 2015), and *Beta* (Gindullis et al. 2001; Weber and Schmidt 2009) among others (Neumann et al. 2011). For example, maize centromeres are composed of CentC tandem repeats (156 bp) and interspersed centromeric-specific retrotransposons (CRM) (Ananiev et al. 1998b; Zhong et al. 2002; Nagaki et al. 2003a; Jiang et al. 2003; Jin et al. 2004). The maize inbred B73 has seven of these complex centromeres and three centromeres composed only of CentC sequences, but numbers and locations of each chromosome type vary widely among maize varieties and in wild relatives (Albert et al. 2010; Gent et al. 2015, 2017, 2018; Schneider et al. 2016; Zhao et al. 2017). It should also be taken into account that the abundance of CentC has decreased after domestication (Albert et al. 2010; Bilinski et al. 2015; Schneider et al. 2016). Similarly, rice centromeres are composed mainly of 155-bp CentO tandem repeats (Dong et al. 1998; Nonomura and Kurata 2001; Cheng et al. 2002) and interspersed centromere-specific CRR retrotransposons (Dong et al. 1998; Miller et al. 1998; Cheng et al. 2002; Nagaki et al. 2005). CentC and CentO repeats are homologous and have certain regions of high sequence identity though these species have diverged during more than 50 my (Lee et al. 2005; Cheng et al. 2002). However, CentC sequences are absent from *Sorghum* and *Miscanthus*, both closer to maize than rice (Gent et al. 2018). In the same way, CentO repeats and CRR-related sequences are absent from functional centromeres of the wild rice species *Oryza brachyantha* (Lee et al. 2005; Dawe 2009). In fact, there exists a rapid evolutionary diversification pattern of centromeric DNA among rice and wild related species (Bao et al. 2006). Beyond the diversity described, there are also plant centromeres composed of single-copy DNA sequences. For example, in *Solanum tuberosum* (potato), the centromeres are composed of either satDNA or single and low copy sequences (Gong et al. 2012; Zhang et al. 2014; Wang et al. 2014). There is great satDNA diversity in potatoes. There are several chromosome-specific satellites composed of very long repeats that have been amplified from retrotransposon-related sequences (Gong et al. 2012; Zhang et al. 2014). A comparison between *S. tuberosum* and *S. verrucosum*, which also have that type of centromeric satDNAs, indicated rapid evolution of centromeric sequences in the genus *Solanum* (Gong et al. 2012; Zhang et al. 2014). A connection has been established between TEs and newly emerged centromeric satDNAs (Meštrović et al. 2015; Gong et al. 2012; Zhang et al. 2014; Vondrak et al. 2020).

5.3.2 Subtelomeric Heterochromatin

Subtelomeric heterochromatin shows high diversity (Garrido-Ramos 2015). Intimately associated with telomeric sequences, the subtelomeric region is a highly dynamic region and one of the faster-evolving regions in eukaryotic genomes (Torres et al. 2011; Richard et al. 2013; Mlinarec et al. 2019; Aguilar and Prieto 2020). The great diversity found among subtelomeric satDNAs does not only affect to sequence but also to repeat length and repeat organization too and includes complex compositions of satDNA repeats (Cuadrado and Jouve 1994, 1995; Vershinin and Heslop-Harrison 1998; Contento et al. 2005; Richard et al. 2013). Furthermore, Mlinarec et al. (2019) have demonstrated the highly polymorphic nature of subtelomeric satDNAs in *Tanacetum cinerariifolium*. These authors detected up to 22 polymorphic loci analyzing the location of two subtelomeric satDNAs in different individuals of different populations of *T. cinerariifolium*.

Intact (5′-TTTAGGG-3′) and degenerated telomeric repeats are usually intermingled among subtelomeric satDNA repeats and, sometimes, they are part of the very satDNA monomer (Fajkus et al. 1995; Buzek et al. 1997; Garrido-Ramos et al. 1999; Sýkorová et al. 2003a; Contento et al. 2005; Navajas-Pérez et al. 2009b; Emadzade et al. 2014; Finke et al. 2019; Mlinarec et al. 2019).

Different types of rearrangements and nonhomologous interchanges occurring between chromosome ends may favor sequence diversification of preexisting satDNAs, including the formation of different subfamilies, as well as the formation and amplification of new satDNA families in this region (Macas et al. 2006; Torres et al. 2011). Thus, several satDNA families and/or subfamilies can coexist in the same species, even in the same chromosome (Bedbrook et al. 1980; Vershinin et al. 1996; Vershinin and Heslop-Harrison 1998; Cuadrado and Jouve 1994, 1995; Heacock et al. 2004; Contento et al. 2005; Kazama et al. 2006; Torres et al. 2011; Mlinarec et al. 2019). Although conservation is not the norm and many subtelomeric satDNA families are restricted to one or a few species, even to one or a few chromosomes, some subtelomeric satellites are more conserved and they are present in the genomes of a group of related genera (Kishii et al. 1999; Quesada del Bosque et al. 2013).

5.3.3 Interstitial Heterochromatin

Many interstitial blocks of heterochromatin are made of satDNA families coming from the subtelomeric area. This process of interstitialization assumes the transference of satDNA sequences from the chromosomal ends toward interstitial sites in accordance with the model proposed by Schweizer and Loidl (1987). There are several examples of interstitial satDNAs that are derived from subtelomeric satDNAs (Cuadrado and Jouve 2002; Lim et al. 2006; Carmona et al. 2013b). Telomere association in bouquet configuration during the first meiotic prophase may favor interstitialization processes (John et al. 1985; Schweizer and Loidl 1987). Mechanisms of interstitialization also include chromosome reorganizations such as inversions and/or transpositions as well as Robertsonian translocation. Additionally, these mechanisms could explain the presence of subtelomeric and telomeric satDNA sequences within the (peri)centromeric area of the chromosomes of some species (Tek and Jiang 2004; Bao et al. 2006; Lim et al. 2006; Navajas-Pérez et al. 2009b; Emadzade et al. 2014; Mlinarec et al. 2019; Finke et al. 2019).

Interestingly, the same satDNA family (SatA) is found in pericentromeric, subtelomeric, and interstitial chromosome regions in two species of the genus *Paphiopedilum* (Lee et al. 2018), which constitutes a major exception to the general view that each of these regions should be composed of different satDNAs. Similarly, CmSat189 satDNA of *Cucumis melo* is located not only on centromeric regions but also on chromosome-specific pericentromeric, interstitial, or subtelomeric regions, allowing the characterization of individual chromosomes of melon (Setiawan et al. 2020). Another five species of the genus *Paphiopedilum* also bear SatA in pericentromeric and subtelomeric regions but not interstitially (Lee et al. 2018). This latter pattern was also observed in *Rumex induratus* (Navajas-Pérez et al. 2009b) and in *Oryza rhyzomatis* (Lee et al. 2005).

Furthermore, interstitial satDNA represents one of the major exceptions to the equilocal principle of heterochromatin distribution in many cases. Examples are satDNAs amplified in specific chromosomes such as sex chromosomes (Shibata et al. 1999, 2000a; Mariotti et al. 2006, 2009; Navajas-Pérez et al. 2005a, 2006, 2009a, c; Cuñado et al. 2007; Steflova et al. 2013; Garrido-Ramos 2015; Jesionek et al. 2020), but also in given autosomes (De la Herrán et al. 2001), as well as supernumerary chromosomes (Alfenito and Birchler 1993; Klemme et al. 2013; Banaei-Moghaddama et al. 2015) and supernumerary chromosome segments (Shibata et al. 2000b; Navajas-Pérez et al. 2005a, 2009c; Finke et al. 2019).

Dioecy has independently emerged in about 6% of plant genera. Sex chromosomes have recently evolved (between 15 and 0.6 mya) independently in only a few of those dioecious plant lineages (Guttman and Charlesworth 1998; Filatov et al. 2000; Navajas-Pérez et al. 2005b, Cuñado et al. 2007; Quesada del Bosque et al. 2011; Kubat et al. 2014; Vyskot and Hobza 2015; Charlesworth 2016; Li et al. 2019). One major feature of sex chromosome evolution is the progressive genetic divergence between X and Y(s) chromosomes, including gene degeneration and accumulation of TEs and satDNAs in Y chromosomes (Charlesworth 2002, 2016; Kejnovský et al. 2009; Hobza et al. 2017). The genus *Rumex* is an excellent study case among young sex-chromosome systems. This genus is composed of several dioecious species that differ for the phase of the evolution of their sex-chromosomes, from earliest steps of sex-chromosome differentiation to more advanced phases that include Y-chromosome degeneration (Cuñado et al. 2007; Navajas-Pérez et al. 2005a, 2006, 2009a, c). There are species like *R. acetosella* or *R. suffruticosus* with an XX/XY sex-chromosome system and little X-Y differentiation, and species like *R. acetosa* and *R. papillaris* with a complex XX/XY$_1$Y$_2$ sex-chromosome system and highly diverged sex chromosomes. In this latter case, Y chromosomes have gathered different Y-specific (seven) and Y-preferentially accumulated (three) satDNA families as well as Y-preferentially accumulated active TEs (Cuñado et al. 2007; Navajas-Pérez et al. 2006, 2009a, c; Mariotti et al. 2006, 2009; Steflova et al. 2013; Jesionek et al. 2020). X chromosome also have X-preferentially accumulated active TEs and two satDNAs found also in the Y chromosomes, but the expansion of satellites in X chromosome is not so high (Jesionek et al. 2020). So far, only one of the multiple families of satDNA found in *R. acetosa* has been detected in species like *R. acetosella* and *R. suffruticosus*. This family, RAE180, massively amplified in the Y chromosomes of *R. acetosa* and *R. papillaris*, is scarcely represented in the genomes of *R. acetosella* and *R. suffruticosus* and, furthermore, RAE180 sequences are located in autosomes (Cuñado et al. 2007; Navajas-Pérez et al. 2009a, c). As occur in *R. acetosella* and *R. suffruticosus*, evolutionarily young *S. latifolia* sex chromosomes are not heterochromatinized and do not contain large amounts of chromosome-specific repeats (Macas et al. 2011). However, the process of accumulation and differentiation of repeats is already evident in the case of some satDNA repeats (Hobza et al. 2006; Cermak et al. 2008; Macas et al. 2011). It has been speculated that the seabuckthorn (*Hippophae rhamnoides*) sex-chromosome system could represent a rare example of evolutionarily old plant sex chromosomes, although the date of their origin could not be established (Puterova et al. 2017). This species has an XX/XY sex-chromosome system that resembles the mammalian sex-chromosome system, with a small Y chromosome that contains several satDNAs and a large X chromosome. Younger systems, like those of *Rumex* and *Silene,* are characterized by larger Y chromosomes since they are in earlier expanding stages of sex chromosome evolution accompanied by accumulation of repetitive sequences (Puterova et al. 2017). The small size of the seabuckthorn Y chromosome appears to be caused by the loss of DNA, which may indicate that the Y chromosome could be in a more advanced evolutionarily shrinkage phase (Puterova et al. 2017). Interestingly, satDNA represents 25% of the genome of this species. Some satellites accumulate in the X chromosome, others are specific to the Y chromosome and others are present on both chromosomes, but most satellites were found on autosomes (Puterova et al. 2017).

5.3.4 Monomers

Researchers have paid much attention to features such as sequence composition, length, and internal organization of satDNA repeats. Here too, high diversity is the main conclusion after the analysis of these characteristics. For example, for monomer size, each satDNA family is characterized by its own repeat length. satDNA monomer size varies between a few tens of base pair (Fominaya et al. 1995; Macas et al. 2006) and several thousand base pairs (Gong et al. 2012). In *Fabeae*, for example, the monomer length of satellite repeats range from 33 to 2979 bp (Avila Robledillo et al. 2020). In *Melampodium*, satDNA monomer length range from 4 to 1200 bp, although the most frequently occurring monomer length is around 180 pb (McCann et al. 2020). In fact, monomer lengths of most satDNAs, specially centromeric ones, vary between 135-195 and 315-375 bp (Macas et al. 2002; Mehrotra and Goyal 2014).

Tandem repetitive DNAs have traditionally been classified in three categories according to the length of their repeats. Short tandem repeats between 2 and 10 bp long are known as microsatellites or simple sequence repeats (SSRs) or single tandem repeats (STRs), whereas repeats longer than 10 bp but shorter than a few tens of base pairs are considered minisatellites (López-Flores and Garrido-Ramos 2012). The longest tandem repeats would compose then the "classic" satellites known as satellite DNAs, forming much longer arrays (several kilobases up to megabases) than micro- and minisatellites (Plohl et al. 2012; Garrido-Ramos 2017). Therefore, satDNAs composed of repeats shorter than 100 bp would be outside the range of what is considered "classic" satDNAs according to repeat length. However, this classification was somewhat arbitrary since there are no precise limits for each category. In fact, some "classic" satDNAs were shorter than 100 bp, since tandem repetitive DNA types have been additionally defined according to their location. Thus, all tandem repeats that populate heterochromatin were considered traditionally "classic" satDNA, independently of monomer sizes, even when they were shorter than 10 bp (Pedersen et al. 1996; Hudakova et al. 2001; Ananiev et al. 2005; Heckmann et al. 2013; Talbert et al. 2018). For example, barley centromeric satDNA is composed of short 6-bp monomers (Hudakova et al. 2001; Nasuda et al. 2005). On the contrary, satDNAs were not found in euchromatin, where usually "inhabit" micro- and minisatellites in the form of shorter tandems scattered throughout the genome (López-Flores and Garrido-Ramos 2012). Adding complexity to the concept, (classic) micro- or minisatellites were reported within heterochromatin together the (classic) satDNAs (Hudakova et al. 2001; Cuadrado and Jouve 2007; Carmona et al. 2013a; Cuadrado et al. 2013; Kejnovský et al. 2013). New data coming from the study of the satellitome (see below) have indeed revealed that repeats of different lengths may be organized in arrays of different sizes in both heterochromatin and euchromatin. Furthermore, a "classical" satDNA family can exist in two forms in a genome, organized in long tandem arrays in heterochromatin and organized in short tandems dispersed throughout euchromatin. Therefore, a terminology based on repeat length should be dismissed nowadays since, in addition, all types of tandem repeats show similarities at the genomic and cytological levels (Ruiz-Ruano et al. 2016).

On the opposite side, there are satDNAs composed of monomers with lengths of several hundred or several thousand base pairs (Navajas-Pérez et al. 2005a; Gong et al. 2012; Zhang et al. 2014; Mehrotra and Goyal 2014; Avila Robledillo et al. 2018). Interestingly, *Vicia faba*, one of the species that have a satDNA with the shortest repeats (Macas et al. 2006), also contains other satDNAs with the longest monomers (1.7–2 Kb) (Avila Robledillo et al. 2018). *Solanum* genomes are characterized by the presence of satDNAs composed of longer monomers (Tek et al. 2005; Gong et al. 2012; Zhang et al. 2014). In *S. tuberosum* and *S. verrucosum* there is a diverse group of centromeric satDNAs with monomers up to 5.4 Kb that display similarities to retrotransposons (Gong et al. 2012; Zhang et al. 2014). In fact, satDNAs composed of long monomers are frequently derived from transposable elements (see below).

Many other times, longest satDNAs repeats are usually the result of several rounds of duplication and divergence of ancient shorter monomers. For example, there are two satDNA families in *Rumex*, RAE730 composed of ~730-bp repeats and RAYSI composed of ~920-bp repeats, which are comprised of subrepeats of 120 bp (Navajas-Pérez et al. 2005a). Not only for the case of very long monomers, but it has also been demonstrated that certain shorter satDNA monomers were built from the duplication and divergence of smaller repeats. For example, the 177-bp repeat sequence of the radish centromeric satDNA probably arose by two duplications and the divergence of an ancestral shorter 60-bp monomer (Grellet et al. 1986).

Higher-order repeat (HOR) units could represent an intermediate stage prior to the establishment of an enlarged monomer. In addition to the conventional monomers, some satDNAs are composed of these HORs which result from homogenization cycles of units composed of two or more adjacent repetitions (Nouzová et al. 1999; Grebenstein et al. 1996; Macas et al. 2006; Vondrak et al. 2020; Belyayev et al. 2019). These complex repetitive units show striking sequence identity between them (i.e., the high similarity between counterpart monomers located at the same position in two units), but monomers within the HOR can show remarkable sequence divergence. In addition to regular HORs composed of repeats units of the same satDNA family, there also exist complex HORs composed of sequences of different origins. Thus, for example, the Nazca HOR of *Phaseolus vulgaris* is composed of four consecutive monomers of the CentPv1 satDNA (99-bp repeats) plus an unrelated sequence of 159 bp (Iwata et al. 2013).

5.3.5 Monomer Signatures

The presence of specific hallmarks in repeats has been associated with the functional significance of satDNA in the eukaryotic genome. However, there are no definitive proofs confirming the validity of these observations. For example, there are many AT-rich satDNAs. This biased composition may favor the presence of AT tracts. It has been proved that AT tracts periodically distributed may induce local sequence-dependent bents that could provoke DNA curvatures. It has been proposed that DNA curvature may be involved in specific recognition of DNA-binding proteins of the heterochromatin and facilitate the tight packing of heterochromatic DNA (Fitzgerald et al. 1994; Lee et al. 2005; Pezer et al. 2012; Yang et al. 2018).

A second common signature is the presence of short inverted sequence repeats within the satDNA monomers. These short dyad symmetries can adopt non-B-form thermodynamically stable secondary structures such as stem-loops or cruciform structures. Stem-loops and cruciform structures might have a role in centromere assembly and function (Hall et al. 2003; Luchetti et al. 2003; Plohl et al. 2012; Pezer et al. 2012; Koch 2000; Kasinathan and Henikoff 2018). On the other hand, inverse short repeats might be important for the own dispersal of satDNAs. These structures could be recognized by the machinery of transposition mechanisms and might help the spread and preservation of satDNA in a process intimately linked to processes of transposition (Plohl et al. 2010, 2012; Šatović and Plohl 2013; Pavlek et al. 2015; Meštrović et al. 2015).

A third remarkable feature within monomeric satDNA sequences is the existence of putatively functional short conserved motifs. Sometimes, the conservation of these motifs can be spurious since they might be the remnants of shorter ancestral repeat monomers. Alternatively, the conservation of short motifs within the repeat unit of a satDNA might be the consequence of a selection-driven action. Up to the present, the only example that supports with certainty this second alternative is the CENP-B box, a short motif that is conserved in the disparate primate and mouse centromeric satDNAs (Masumoto et al. 1989, 2004; Muro et al. 1992; Haaf et al. 1995). The centromere protein B (CENP-B) is a DNA-binding protein that specifically recognizes and binds the CENP-B box facilitating the centromere assembly and stabilization, as well as correct chromosome segregation (Fachinetti et al. 2015). There are several reports suggesting the existence of somewhat similar motifs in other species, including plants (Aragon-Alcaide et al. 1996; Nonomura and Kurata 1999; Nagaki et al. 1998; Gindullis et al. 2001). However, no empirical validation of the functionality of these motifs has been done up to now. The centromeric satDNAs of maize and rice, CentC (156 bp repeat length) and CentO (155 bp repeat length), respectively, share an 80-bp motif (Lee et al. 2005). The same motif was found in the CentO-C1 satDNA (126 bp repeat length) of *Oryza rhizomatis* as well as in the 150-bp centromeric satDNA of pearl millet (*Pennisetum glaucum*) (Lee et al. 2005). In addition, seven centromeric satDNAs (between 166 and 187 bp repeat lengths) of switchgrass (*Panicum virgatum*) also share the same 80-bp motif as reported by Yang et al. (2018), who proposed that this motif could have been preserved in all these Poaceae species because of sequence-specific properties that favor centromere assembly. However, this motif has not been identified in other centromeric satDNAs of other Poaceae such as *Oryza brachyantha* (Lee et al. 2005) or *Sorghum* and *Miscanthus* (Gent et al. 2017), whose centromeric satDNAs are unrelated to those of maize and rice and the rest of species sharing the 80-bp motif.

5.3.6 satDNA Function

As we have discussed before, satDNAs are among the fastest evolving sequences in the eukaryotic genome. Therefore, most satDNA families are species-specific or genus-specific. Intriguingly, there are several other satDNAs that are shared by a wide group of species. Whether this is yet another consequence of the random evolutionary dynamics of satDNA or it is due to the fact that those satDNAs have been preserved for millions of years by selective constraints is still the subject of debate. Despite the absence of sequence conservation, centromeres and the adjacent pericentromeric region, as well as the subtelomeric region associated with the telomere, carry out essential functions in preserving and transmitting the genetic material throughout the generations, i.e., they are essential for the maintenance of life. Whether the role carried out by satDNA in these regions is sequence-independent or whether that role depends on specific features of this type of DNA, are questions that have been under debate. Whatever the case may be, thanks to numerous studies conducted in diverse groups of organisms during the last two decades, we have favorably changed our view on the role of satDNA in the regulation and evolution of the genome. These studies will be analyzed in the next section.

5.4 What Has Genomics Contributed to the Study of satDNAs?

5.4.1 On the Satellitome

Satellitome is a recent term, proposed to encompass the whole set of tandem repetitive sequences found in one genome, independently of their repeat length, copy number, and location (Ruiz-Ruano et al. 2016). Conventional techniques of satDNA isolation allowed researchers to isolate one or a few, the most abundant, satDNA families per genome. However, less-represented satDNA families went unnoticed by these methods. Moreover, there are species that contain very low amounts of satDNAs. Most times, isolation of tandem repetitive DNAs in those species has been made inaccessible by such methods (Ruiz-Ruano et al. 2019). However genomic approaches of satDNA analysis have unveiled the existence of several, in some cases tens, satDNA families per genome. Even the study of the satellitome of species with large genome sizes and little satDNA amounts is no longer unreachable (Ruiz-Ruano et al. 2019). Satellitome analysis has revealed a high diversity of satDNAs within plant genomes and important conclusions on their organization, which are set forth below.

satDNA Diversity Satellitome analysis has revealed a surprising diversity of satDNAs both in animals (Ruiz-Ruano et al. 2016, 2017, 2018; Camacho et al. in prep.) and plants. Notwithstanding, usually just a single or a few satellite families are dominant in terms of their genomic abundance (see for example: Macas et al. 2015; Avila Robledillo et al. 2018, 2020). Most plant satellitomes analyzed have a considerably higher quantity of satDNA families per genome than previously found. For example, although all of them were elusive for conventional techniques, we detected up to 11 satDNA families in the genome of the fern *Vandenboschia speciosa* analyzing NGS reads (Ruiz-Ruano et al. 2019). Thirty-four satDNA families included in 21 superfamilies were found in the grass genus *Deschampsia* (González et al. 2020). Some satDNAs are species-specific but most of them were shared by the two *Deschampsia* species analyzed. Some of these satDNAs are shared with other grasses (González et al. 2020). Avila Robledillo et al. (2020) have analyzed 14 species of the legume tribe *Fabeae* and have identified up to 64 highly diverse families of centromeric satDNAs. Most species have centromeres composed of multiple satDNAs. One of these species is *Vicia faba*. The satellitome of this species is composed of 30 satDNA families (26 fully analyzed) whereas only 4 were previously identified (Avila Robledillo et al. 2018). Among these, seven are centromeric satDNAs that have a chromosome-specific distribution (Avila Robledillo et al. 2018). The majority of *V. faba* satDNAs do not show sequence similarities to those from other legume species, which suggested their species-specific origin or rapid satDNA sequence diversification (Avila Robledillo et al. 2018). Interestingly, polymorphic or supernumerary loci of three satDNAs were also found (Avila Robledillo et al. 2018). Thirteen satDNA families have been uncovered in *Pisum sativum* in addition to the two previously described (Macas et al. 2007). Bolsheva et al. (2019) conducted a satellitome analysis in 5 species of *Linum* and found 44 satDNA families. Content and diversity of satDNAs among these species were in agreement with the library hypothesis postulates. The genome of *Lathyrus sativus* contains 23 different satDNA families, summing 10.7% of the genome (Macas et al. 2015). Only 2 out of 12 main satDNA families detected in the dioecious species seabuckthorn (*Hippophae rhamnoides*) had been previously identified (Puterova et al. 2017). Up to five satDNA families were identified by conventional cloning in species of *Rumex* with a multiple sex-chromosome system (Shibata et al. 1999, 2000a, b; Navajas-Pérez et al. 2005a, 2006, 2009a, c; Mariotti et al. 2009). Four of these satDNAs have been massively amplified in the Y chromosomes and one in an autosomic supernumerary chromosome segment of *R. acetosa*. A genomic approach unveiled two more satDNAs in *R. acetosa*, one dominating on the X chromosome and the other localized mostly on the Y_1 chromosome (Steflova et al. 2013). Interestingly, the recent analysis of the satellitome of *R. acetosa* conducted by Jesionek et al. (2020) has revealed the existence of almost 40 satDNA, from which there are 13 major satellites that accumulate in the Y chromosomes of this species, some of which are also found in less amount in the X chromosome as well as in some autosomes (Jesionek et al. 2020). Only two of the six major satDNAs detected in *Olea europea* were identified for the first time in a genomic approximation (Barghini et al. 2014). This species has high satDNA content (more than 30% of the genome) and the four previously known satDNAs represented 85% of tandem repeats of the olive genome (Barghini et al. 2014). All these examples teach us that the conventional cloning of satDNA was very useful for the detection of the most abundant families of each genome, but that it was insufficient to detect all other less abundant families. On the other hand, not all genomes are characterized by a large number of different satDNAs. There are also species relatively poorly diverse for distinct satDNAs. For example, some *Vicia* species (Macas et al. 2015), *Prospero autumnalis* (Emadzade et al. 2014), or *Cucumis melo* (Setiawan et al. 2020).

Promiscuity Avila Robledillo et al. (2020) used the term promiscuity to define the pattern observed in several satDNAs shared by different related species of *Fabeae*. Some satDNAs are located at centromeres in one species while having a non-centromeric location in other species. Furthermore, there is also intragenomic promiscuity since there are some centromeric satDNAs that are simultaneously located in additional non-centromeric loci in the same genome. We have found a similar pattern in grasshoppers (Camacho et al. in prep.). Two grasshopper species that diverged for more than 23 my have a species-specific satellitome profile but share some satDNA families. However, a shared satDNA may be centromeric in one species but not in the other species. Moreover, a satDNA family may be the most abundant in one species whereas is one of the less abundant in the other species, both displaying different organizational patterns. In addition, some satDNAs have different simultaneous locations in the same species (Ruiz-Ruano et al. 2016), intragenomic promiscuity also revealed in other grasshopper species (Ruiz-Ruano et al. 2016, 2017, 2018). As mentioned above, the multi-locus distribution pattern of a satDNA family was also observed in *Cucumis melo* (Setiawan et al. 2020) and in *Paphiopedilum* (Lee et al. 2018). Although the expansion of satellites took place mainly on the Y chromosomes of *R. acetosa*, many short satellite arrays of most of these satDNAs are ubiquitous in the autosomes and/or the X chromosomes (Jesionek et al. 2020). Interestingly, in *Lathyrus sativus*, only 2 out of 11 satDNA families are predominantly organized in long arrays typical for satDNA, one associated with centromeric regions and the other with subtelomeric regions (Vondrak et al. 2020). The remaining 9 tandem repeat families are organized both as prominent pericentromeric bands and as short tandem arrays dispersed throughout the genome (Vondrak et al. 2020). The second pattern was also found in other *Lathyrus* species. This is a dispersed pattern that is consistent with the fact that these short tandem arrays are embedded within the sequence of retrotransposons of the Ogre lineage (Tat family) of LTR/Gypsy retrotransposons (Neumann et al. 2019). Specifically, they are embedded in the 3′-end untranslated region (UTR) of the Tat/Ogre elements (Vondrak et al. 2020). It has been proposed that the longer satellite arrays in centromeres might have been originated by an expansion of tandem sequences originally present only within Tat/Ogre elements and that centromeric regions would be favorable for satDNA accumulation (Vondrak et al. 2020). This type of expansion could be responsible for the emergence of many different satDNAs within a species (see below). On the other hand, a dual pattern of the genomic distribution of satDNAs (long arrays typical of satDNAs and single repeats or short arrays dispersed throughout the genome) was demonstrated previously in several insect species (Ruiz-Ruano et al. 2016; Feliciello et al. 2011, 2015; Kuhn et al. 2012; Brajković et al. 2012; Larracuente 2014; Pavlek et al. 2015; Pita et al. 2017; De Lima et al. 2017). The importance of some of these shorter arrays dispersed throughout the genome as regulators of the expression of nearby genes has been demonstrated (Menon et al. 2014; Feliciello et al. 2015; Joshi and Meller 2017).

5.4.2 On the Origin of satDNAs

Earlier studies demonstrated that many satDNAs monomers evolved from shorter ones by means of alternative cycles of duplication and divergence (Grellet et al. 1986; Navajas-Pérez et al. 2005a; Macas et al. 2006; Emadzade et al. 2014). We have recently confirmed it after an analysis of the satellitome of the fern *V. speciosa* in which we found a marked relationship between several satDNA families grouped into superfamilies. Interestingly, longer (and older) satellites in *V. speciosa* evolved from shorter ones (Ruiz-Ruano et al. 2019). Besides, in some cases, microsatellites were a source of new satDNAs, which would imply the involvement of both replication slippage and unequal crossing-over in the initial monomer formation (Ruiz-Ruano et al. 2019). The existence of superfamilies of related satDNA families within a genome has also been found in the genus *Melampodium*. In this genus, seven satDNA families shared by several species are likely descendants of one common repeat (McCann et al. 2020). Superfamilies composed of different satDNA families were also observed in *Fabeae* (Avila Robledillo et al. 2018, 2020).

In theory, any sequence can act as a seed that gives rise to a repeating monomer. In addition to any single sequence, different kinds of tandem repeats or dispersed repeats can act as a substrate for satDNA emergence. For example, some satDNAs in tomato, potato, tobacco, common bean, and faba bean genomes derive from tandem duplications of a part of the intergenic spacers of the ribosomal RNA (rRNA) genes (Stupar et al. 2002; Macas et al. 2003; Lim et al. 2006; Jo et al. 2009; Almeida et al. 2012). Therefore, it has been proposed that rDNA intergenic spacers dispersion may be one of the processes leading to the formation of novel satDNAs. Conversely, new rDNA loci may also arise by the amplification of orphaned or low copy number rDNA (Matyášek et al. 2012). In addition to the major rDNA locus, the 5S rDNA locus can also be involved in satDNA origin. B chromosomes of *Plantago lagopus* are enriched by a new satDNA family derived from 5S rDNA units (Kumke et al. 2016). One of the eight centromeric satDNAs identified in switchgrass was found to be identical to the 5S rDNA. Yang et al. (2018) demonstrated that 5S rRNA genes were recruited as centromeric DNA in that species.

Regarding dispersed repeats, there is evidence of the involvement of TEs in satDNA origin (reviewed in Meštrović et al. 2015). Macas et al. (2009) found that PisTR-A satDNA in *Pisum sativum* was present both as short dispersed repeats as well as long arrays of tandemly arranged satDNA. Intriguingly, the dispersed repeats occurred in the genome embedded within 3′-end UTR of Tat/Ogre retrotransposons. 3′-end UTR is highly variable among Tat/Ogre elements, including several other tandem repeats along with or instead of PisTR-A (Macas et al. 2009). These authors documented several other cases of satDNAs that likely originated by the amplification of 3′-end UTR tandem repeats. As mentioned above, the majority of *Lathyrus sativu*s satDNAs originated from short tandem repeats present in the 3′-end UTRs of Tat/Ogre retrotransposons (Vondrak et al. 2020). Thus, it has been proposed that dispersed tandem repeats embedded within TEs might populate new larger tandems that would emerge and accumulate in favorable regions such as centromeres (Vondrak et al. 2020) or non-recombining sex chromosomes (Jesionek et al. 2020). According to Vondrak et al. (2020), Tat/Ogre elements may play a general significant role in satDNA evolution by providing a source for novel satellites that would emerge by the expansion of their short tandem repeats arrays. This proposal is based on the widespread occurrence and high copy numbers of Tat/Ogre elements in many plant taxa (Neumann et al. 2006; Macas and Neumann 2007; Kubat et al.

2014; Macas et al. 2015). In *R. acetosa*, the RAE93 satDNA family has also been derived from the 3′-end UTR of Tat/Ogre elements. Tat/Ogre elements are highly amplified in the Y chromosomes, with minor additional signals dispersed through the rest of the genome. These elements contain arrays of five RAE93 monomers in the 3′-end UTR and disperse RAE93 sequences in the genome along with the element (Jesionek et al. 2020). In addition, RAE93 has been expanded into typical long arrays of satDNA mainly on X and Y chromosomes (Jesionek et al. 2020). In addition to Ogre elements, other TEs can serve as a source for new satDNAs. For example, a second satDNA in *R. acetosa* is derived from an LTR/Copia retrotransposon, an AleII element (Jesionek et al. 2020). The AleII satellite monomer contains a full-length non-autonomous copy of the AleII retrotransposon that has been duplicated in tandem, giving rise to a single satDNA locus in a putative pseudoautosomal region mediating recombination between the X and Y_1 chromosomes (Jesionek et al. 2020). Additional examples of retrotransposon-derived satDNAs were documented in wheat (Cheng and Murata 2003), maize (Sharma et al. 2013), and potato (Tek et al. 2005; Gong et al. 2012; Zhang et al. 2014). *Jozin*, an EnSpm/CACTA-like DNA transposon is involved in the generation of monomers of the most abundant satDNA family of the *Chenopodium album* satellitome (Belyayev et al. 2020a). A ~ 40 bp fragment of the transposase gene served as the start monomer of the satDNA array (Belyayev et al. 2020a). In *Arabidopsis*, tandem repeat arrays were also generated from internal parts of an EnSpm-like DNA transposon (Kapitonov and Jurka 1999). Therefore, TEs may significantly contribute to satDNA evolution by generating a "library" of short repeat arrays that can subsequently be dispersed through the genome and eventually further amplified and homogenized into novel satellite repeats (Macas et al. 2009; Vondrak et al. 2020). Something that, in the opinion of Belyayev et al. (2020b), would refute the "library" hypothesis, since newly emerged satDNA families may have a similar sequence in different species given that they can be independently originated from the same fragment of the same TE type in each species. This similarity may create a false perception of conservation even in the event that novel satDNAs would arise repeatedly and independently in different lineages (Belyayev et al. 2020b). Furthermore, these authors have used the idea of "the library of the mechanisms of origin" to refer to the variety of ways in which a satDNA can originate (Belyayev et al. 2020b).

5.4.3 On satDNA Function

satDNA repeats are noncoding sequences but they are sequences that are transcribed. satDNA transcription was seen as a rarity caused by a failure of transcription termination in oocyte lampbrush chromosomes of newts (Varley et al. 1980a, b; Diaz et al. 1981; Gall et al. 1981). However, regulated satDNA transcription has been revealed as a major discovery in very recent years and, we now know that satDNA transcripts are important players in different satDNA-performed functions. These transcripts play important roles in centromere organization and function, in pericentromeric and telomeric assembling as well as in the regulation of heterochromatin formation and maintenance. In addition, satDNA transcripts may have a regulatory role in gene expression.

The "black hole" of the genome (Henikoff 2002) is "less dark" today given that we have now abundant information on how plant and animal centromeres are organized and make their function. The centromere is the assembly site of the kinetochore complex in active centromeres and responsible for the correct chromatid and chromosome segregation. In addition, the pericentromeric region appears essential in maintaining the heterochromatin architecture, sustaining kinetochore formation, maintaining sister-chromatids cohesion, and driving chromosomal segregation during cell divisions (reviewed in Garrido-Ramos 2017). Centromeric repeats (satDNA or satDNA plus CRs) are the main component of both centromeres and pericentromeric regions of plant chromosomes and contribute to the centromere and kinetochore assembly and to the formation of flanking heterochromatin (reviewed in Garrido-Ramos 2017). As indicated before, centromeric sequences are not conserved among plant species, either among eukaryotes in general. Each particular lineage experiences rapid evolutionary diversification patterns of centromeric DNAs. Therefore, centromeric DNA repeat sequences alone appear insufficient to determine centromere identity. Indeed, these repeats may be absent in functional centromeres (Gong et al. 2012) and functional neocentromeres lack centromeric specific repeats (Nasuda et al. 2005; Zhang et al. 2013a, b; Fu et al. 2013; Liu et al. 2015). In addition, only one centromere is active in stable dicentric chromosomes, though both centromeres are composed of centromeric repeats (Zhang et al. 2010; Gao et al. 2011; Fu et al. 2012). Furthermore, there is no obvious delimitation between the DNA forming the functional centromeric locus and the neighboring pericentromeric DNA (Jin et al. 2004; Lamb et al. 2005; Houben et al. 2007; Gao et al. 2011; Wang et al. 2014; Bilinski et al. 2015). All these observations have brought to light the importance of the epigenetic regulation of the functional centromere (Wang et al. 2009; Lermontova et al. 2015). The association of the CenH3 protein, a centromeric-specific histone H3 variant, marks the functional centromere locus. In the functional centromere, the histone H3 is replaced by CenH3. Faced with this scenario, immunological detection (Talbert et al. 2002) and, especially, chromatin immunoprecipitation (ChIP) (Zhong et al. 2002) assays have been largely used as techniques for the delimitation of the centromeric sequences. These assays have demonstrated that only a part of the centromeric repeats interacts with CenH3 in the functional centromere; i.e., only a fraction of the centromeric repeats constitutes the centromere locus (Jin et al. 2004; Houben et al. 2007; Zhang et al. 2013b; Macas et al. 2010), as occur in the centromeric region of human chromosomes (Schueler et al. 2001). Correspondingly, different chromosomes may have similar CenH3 domain sizes but may differ for the sizes of the centromeric/pericentromeric repeat arrays (Zhang et al. 2013b).

The centromere locus is thus a domain composed of CenH3-containing nucleosomes (Schubert et al. 2020). Species-specific CenH3 proteins are key elements in the epigenetic control of centromere function. CenH3 proteins have a conserved core part but differ among species in the N terminal tail (Henikoff et al. 2000; Talbert et al. 2002; Zhong et al. 2002) and could act as the necessary linker between the highly diverse centromeric repeats and the conserved kinetochore proteins (Malik and Henikoff 2003; Maheshwari et al. 2015).

Moreover, centromeres and pericentromeric chromatin are regions that exhibit other characteristic epigenetic modifications. For example, centromeres contain one cluster of CenH3 surrounded by pericentromeric chromatin marked by cell cycle-dependent histone modifications, such as the phosphorylation of the histone H2A at threonine 120 (H2AT120ph) (Dong and Han 2012; Demidov et al. 2014; Neumann et al. 2016; Schubert et al. 2020) or the phosphorylation of the histone H3 at serine 10 or serine 28 (H3S10ph, H3S28ph) (Houben et al. 1999; Kaszás and Cande 2000; Manzanero et al. 2000; Gernand et al. 2003; Zhang et al. 2005; Kurihara et al. 2006; Han et al. 2006; Fu et al. 2012; Neumann et al. 2016; Schubert et al. 2020). Indeed, H2AT120ph is thought to be a universal centromeric marker in plants (Demidov et al. 2014). It appears that there are not differential patterns of histone H3 methylation in plant centromeres and pericentromeres (Zhao et al. 2016; Neumann et al. 2016; Gent et al. 2018), as in animal centromeres (Sullivan and Karpen 2004; Talbert and Henikoff 2018). On the other hand, it was proposed that centromeric repeats are hypomethylated with respect to repeats in the surrounding heterochromatin (Zhang et al. 2008; Yan et al. 2010; Koo et al. 2011; Zakrzewski et al. 2011, 2014). However, this difference has been questioned (Zakrzewski et al. 2011, 2014; Schmidt et al. 2014; Gent et al. 2012, 2014, 2018; Su et al. 2016).

Therefore, it has become evident that the role of centromeric repeats in centromeric function does not depend on their primary sequence. Instead, sequence-specific features of satDNA repeats may be important for the function of the centromere. The repetitive structure of satDNA itself (Dawe and Henikoff 2006; McFarlane and Humphrey 2010) or the ability of the repeats to acquire thermodynamically stable secondary structures (Koch 2000; Zhang et al. 2013b; Kasinathan and Henikoff 2018; Talbert and Henikoff 2018; Yang et al. 2018) might be advantageous for centromere function. The regular positioning of CenH3 nucleosomes could be advantageous for centromere formation, and satDNA repeats might have a crucial role in the stabilization of CenH3-containing nucleosomes (Zhang et al. 2013b; Zhao et al. 2016; Yang et al. 2018). Accordingly, epigenetically defined centromeres and neocentromeres composed of single-copy sequences can evolve to repetitive centromeres, which could have a selective advantage for CenH3 stabilization (Zhang et al. 2013b; Fukagawa and Earnshaw 2014). Such young centromeric repeats could emerge by tandem duplication of a sequence with a selective advantage for CENH3 stabilization (Gong et al. 2012; Zhang et al. 2013b, 2014). In such recombination event could be involved retrotransposon-related sequences (Gong et al. 2012; Zhang et al. 2013b, 2014). Different centromeres starting with different tandem repeats would become homogenized over time (Gong et al. 2012; Zhang et al. 2013b). Alternatively, transposition of preexisting satDNAs could also populate an epigenetically defined centromere or neocentromere and have the same effect on CenH3 stabilization, suggesting that they might originate elsewhere in the genome and subsequently invade the centromeres (Gong et al. 2012; Zhang et al. 2013b, 2014; Yang et al. 2018; Avila Robledillo et al. 2020; Vondrak et al. 2020).

On the other hand, both centromeric satDNA repeats and retrotransposons are transcriptionally active (Topp et al. 2004). Repetitive centromeric satDNA transcripts have been found essential for centromere and kinetochore assembly (Hall et al. 2012; Biscotti et al. 2015; Ferreira et al. 2015; Garrido-Ramos 2017; Talbert and Henikoff 2018). Neumann et al. (2011) proposed that CR transcripts may help to promote a genomic environment that contributes to the establishment of centromeric chromatin. satDNA transcripts might take part also in the establishment of the centromeric chromatin (Lee et al. 2006; Rošić et al. 2014). Centromeric transcripts may affect the stability or activity of several kinetochore components, and centromeric transcription may be required for CenH3 deposition (Talbert and Henikoff 2018, 2020; Gent et al. 2018). It has been proposed that centromeres may be specified by cruciform structures formed by dyad symmetries or induced by DNA-bending proteins and that these non-B form DNA configurations in centromeres may facilitate transcription, enabling CenH3 incorporation during nucleosome remodeling (Kasinathan and Henikoff 2018; Talbert and Henikoff 2018). Alternatively, transcription would facilitate spontaneous or protein-induced cruciform formation at the centromere and, subsequently, cruciform DNA would facilitate CenH3 deposition (Talbert and Henikoff 2018, 2020).

Holocentric chromosomes are characterized by the presence of holocentromeres. That is, instead of the monocentric regional centromere described above, some plant species have chromosomes that lack primary constrictions and their kinetochores are spread along the entire chromosome (Melters et al. 2012; Cuacos et al. 2015). Examples are found in some species of some genera of the families Juncaceae, Cyperaceae, Melanthiaceae Droseraceae, and Convulvulaceae (Melters et al. 2012; Cuacos et al. 2015). *Luzula elegans* and *L. luzuloides,* and mitotic chromosomes of *Rhynchospora pubera* and *R. tenuis*, show line-like holocentromeres (reviewed in Schubert et al. 2020). This type of holocentromeres is characterized by many CenH3-containing chromatin domains forming a contiguous line along the whole chromosome where spindle fibers attach at CenH3 chromatin (Heckmann et al. 2011, 2014a, b; Marques et al. 2015, 2016; Schubert et al. 2020). Holocentromeres of *R. pubera* reorganizes into clusters during meiosis. This organization consists of many evenly dispersed CenH3 clusters where spindle fibers attach along the whole chromosome (Marques et al. 2016; Schubert et al. 2020). In *R. pubera*, centromere domains are composed of a specific satDNA that associates with CenH3 proteins (Marques et al. 2015; Ribeiro et al. 2017). On the contrary, no centromere-specific repeats have been found in *Luzula elegans*, whose genome contains 30 satDNA families (Heckmann et al. 2013). *Cuscuta europaea* is characterized by a third different holocentromeric organization (Schubert et al. 2020). In *Cuscuta europaea*, the spindle fibers attach along the entire chromosome length both to CenH3 and CenH3-free chromatin (Oliveira et al. 2020). CenH3 chromatin is associated mainly with satDNA in this species. However, satDNAs do not constitute the CenH3-lacking holocentromeres in *C. europaea* (Oliveira et al. 2020). The genus *Cuscuta* includes species with monocentric and species with holocentric chromosomes. Transition to holocentricity in *Cuscuta* has been accompanied by significant changes in epigenetic marks, with the loss of histone H2A phosphorylation (H2AT120ph), and in repetitive DNA sequence composition, with the elimination of centromeric retrotransposons (Neumann et al. 2020). Moreover, none of the satDNA families identified in holocentric *Cuscuta* species were distributed along the chromosomes (Neumann et al. 2020).

Finally, meta-polycentric chromosomes of *Lathyrus* and *Pisum* are monocentric chromosomes with multiple centromeric domains. Meta-polycentromeres represent an elongated primary constriction and appear to be an intermediate stage between monocentric-regional and holocentric chromosomes (Neumann et al. 2012, 2015, 2016; Schubert et al. 2020). Centromere domains of the meta-polycentric chromosomes of *Pisum sativum* are composed of 12 different satDNA families (Neumann et al. 2012; Avila Robledillo et al. 2020). Conversely, three *Lathyrus* species possess single-dominant centromeric satellites classified as members of the same FabTR-2 superfamily (Avila Robledillo et al. 2020). Another species has two additional less-abundant satDNA families, while a fifth species have four species-specific satDNAs in addition to FabTR-2, which in this case has a noncentromeric location (Avila Robledillo et al. 2020). Meta-polycentric centromeres show specific patterns of histone phosphorylation and methylation (Neumann et al. 2016).

Telomeres are composed of the repetition of the 5′-TTTAGGG-3′ sequence in most plant species analyzed up to the present (Richards and Ausubel 1988; Lamb et al. 2007). However, there are some exceptions caused by sequence variations (5′-TTAGGG-3′ or 5′-TTTTTTAGGG-3′, for example) (Mizuno et al. 2008; Sýkorová et al. 2003b, c; de la Herrán et al. 2005; Fulnečková et al. 2013; Peška et al. 2015). A deeper change has been produced in *Allium* species in which the telomeric sequence has switched to the 12-bp repeat 5’-CTCGGTTATGGG-3′ (Fajkus et al. 2016). Telomeres protect chromosome ends from exonuclease attack. Telomeres also protect chromosomes from illegitimate fusions between chromosome ends and between these and artificial ends caused by chromosome breaks. Furthermore, they protect chromosomes from progressive shortening after each chromosome replication round.

Under telomeres, the subtelomeric sequences might perform a set of quite important functions in chromosome preservation and segregation (Biscotti et al. 2015; Kwapisz and Morillon 2020; Saint-Leandre and Levine 2020; Mlinarec et al. 2019). For example, the subtelomeric region is important for the process of homologous chromosome recognition and pairing during meiosis (Calderón et al. 2014, 2018). Subtelomeric DNA (and RNA) would be involved in telomeric/subtelomeric chromatin assembly and maintenance (Biscotti et al. 2015; Kwapisz and Morillon 2020). Subtelomere chromatin packaging and subtelomeric transcriptional regulation have profound effects on adjacent telomere function (Saint-Leandre and Levine 2020). Subtelomeres are involved in the regulation of TERRA transcription. Telomeric repeat-containing RNAs (TERRA) are long noncoding RNAs (lncRNA) transcribed from telomeric repeats that initiate transcription in subtelomeric regions and are therefore composed of subtelomeric sequences and telomeric repeats (Azzalin and Lingner 2015). TERRA lncRNAs are evolutionarily conserved (Cusanelli and Chartrand 2015). TERRA transcripts regulate telomere length and stability (Kwapisz and Morillon 2020; Azzalin and Lingner 2015). It has also been proposed that these transcripts might be involved in telomeric/subtelomeric heterochromatin formation and maintenance through a mechanism based on small interfering RNAs (siRNAs) (Vrbsky et al. 2010; Biscotti et al. 2015; Kwapisz and Morillon 2020). In addition to TERRA, ARRET and αARRET (two complementary subtelomeric lncRNAs) and ARIA (composed mostly or exclusively of telomeric repeats) constitute the telomeric transcriptome (Kwapisz and Morillon 2020), although there is little information about these latter lncRNA types. Subtelomeric sequences might have an active role in telomere maintenance and genome stability (van Emden et al. 2019). In addition, Saint-Leandre and Levine (2020) have hypothesized that subtelomeric sequence evolution would shape the recurrent innovation of telomere proteins.

In plants, heterochromatin is mainly characterized by two hallmarks: dimethylation of lysine 9 of histone H3 (H3K9me2) and DNA methylation (reviewed in Kowar et al. 2016). satDNA and satDNA transcripts perform an important role in the regulation of heterochromatin assembly. Small interfering RNAs (siRNAs) of 21-24 nt derived from longer satDNAs transcripts are involved in this regulation and in the regulation of heterochromatin maintenance (May et al. 2005; Lee et al. 2006; Zakrzewski et al. 2011). According to the model of Volpe et al. (2002) for heterochromatin assembly, siRNAs and specific proteins such as Argonauta are associated in RNA-induced transcriptional silencing (RITS) complexes that recruit histone methyltransferases, which promote histone H3 methylation at lysine 9 (H3K9me2) favoring the subsequent recruitment of heterochromatin protein 1 (HP1) and heterochromatic silencing (Volpe and Martienssen 2011; Martienssen and Moazed 2015; Holoch and Moazed 2015; Johnson and Straight 2017). The heterochromatic status is maintained through self-reinforcing positive feedback in which more siRNAs are generated, which favor H3K9 methylation and HP1 recruitment (Martienssen and Moazed 2015; Holoch and Moazed 2015; Johnson and Straight 2017). The process begins with the synthesis of RNA by Pol IV, a plant-specific RNA polymerase. This RNA acts as a template for an RNA-dependent RNA polymerase that synthesizes a complementary RNA strand. The resulting double-stranded RNA is cleaved by Dicer to produce the siRNAs (Martienssen and Moazed 2015; Holoch and Moazed 2015). It appears that, at least in plants, siRNAs also direct the methylation of the DNA from which they are derived (Zakrzewski et al. 2011; Feng and Michaels 2015; Martienssen and Moazed 2015; Holoch and Moazed 2015). Intriguingly, Finke et al. (2019) have found that satDNAs might adopt euchromatic-like features in plant nuclear genomes. Non-pericentromeric heterochromatic segments in Australian crucifer *Ballantinia antipoda* consist of a mixture of unique sequences and a satDNA family (BaSAT1), whose 174-pb repeats are hypomethylated and devoid of heterochromatic H3K9me2. Moreover, these authors have found that individual BaSAT1 repeats may carry either heterochromatin or euchromatin features.

satDNAs have another important, recently discovered, role in eukaryotic genomes. A few studies have revealed that satDNA transcripts might regulate the expression of some genes. satDNA repeats may influence the expression of neighboring genes by siRNA-mediated silencing mechanisms (Menon et al. 2014;

Feliciello et al. 2015; Ferree 2017; Joshi and Meller 2017; Hall et al. 2017). In addition to typical satDNAs long arrays, repeats from a satDNA family may be dispersed throughout the genome, close to genes on which can exert their regulatory influence (Menon et al. 2014; Feliciello et al. 2015; Joshi and Meller 2017). This regulatory control of gene expression has been demonstrated in insects but, up to the present, there are no similar studies on plants. Notwithstanding, the abovementioned studies showing similar patterns of satDNA repeats distribution open the hypothesis that plant dispersed satDNA repeats might play also a role in gene expression regulation.

Taking together, all studies carried out during the last two decades in diverse groups of organisms have made a part of the “dark matter” of the genome less obscure (Kapustova et al. 2019). The initial view of satDNA as “junk” DNA (Ohno 1972; Doolittle 2013; Graur et al. 2013, 2015; Garrido-Ramos 2015) has shifted today to a new view of satDNA as a fraction of the genome with meaningful roles in chromosome organization, replication, chromosome pairing, segregation, and gene expression regulation. Perhaps it is time now that we consider the “junk DNA” descriptor overcome and, as defended by Hartley and O’Neill (2019), we stop contextualizing under that false premise all the knowledge that supports today the role that satDNA has in biology and in the evolution of the eukaryotic genome.

5.5 Concluding Remarks and Perspectives

Satellite DNA is still a fascinating and intriguing part of the genome. Since its discovery more than 50 years ago, it has been of considerable interest to many researchers. Lamentably, the twenty-first century started out being somewhat frustrating with respect to its study. Regions comprised of satDNAs, especially the centromeric region, represented the so-called “black holes” of the genome. Consequently, a significant part of the genome of many plant species has not been incorporated in published genome sequences. However, the initial frustration has changed to progressive optimism. The combined use of short NGS reads, such as Roche/454 or Illumina reads, and computer programs capable of identifying and quantifying all kinds of repeating sequences of the genome have represented a great advance in recent years (Novák et al. 2010, 2013, 2017, 2020b; Weiss-Schneeweiss et al. 2015; Ruiz-Ruano et al. 2016). In addition, chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) of genomic regions associated with specific types of chromatin, such as CenH3 chromatin, has been revealed as a powerful approach to investigate DNA sequence composition of specific parts of the genome, such as the centromere (Henikoff et al. 2015; Kowar et al. 2016). Therefore, the possibility of using genomic tools for satDNA analysis has opened a wide range of possibilities for better understanding the origin, evolution, and organization of satDNAs and we have managed to better understand the relationships that exist between the different families that make up the satellitome of eukaryotic genomes.

Furthermore, we have now a better understanding of the functions these sequences perform.

However, it has been proved that these approaches are insufficient to provide insight into satDNA large-scale arrangement in the genome (see Kapustova et al. 2019; Vondrak et al. 2020). Therefore, the incorporation of long-read DNA sequencing technologies such as those of the Pacific Biosciences (PacBio) and Nanopore platforms has become essential for this purpose. Specifically, nanopore sequencing reads can reach lengths of up to one megabase (van Dijk et al. 2018). A nanopore sequencing strategy combined with short-read variant validation has resulted efficiently used for the assembling and characterization of the centromeric region of a human Y chromosome (Jain et al. 2018). Vondrak et al. (2020) successfully characterized the organization of different satDNAs of *L. sativus* following a genome-wide study by employing bioinformatic analyses of long nanopore reads.

Taken together, all these technological advances are becoming promising tools that open the door to the definitive approach to satellite DNA that we hope to see developed in the next and promising new decade.

Acknowledgments The research in our laboratory is currently financed by Agencia Estatal de Investigación of the Ministerio de Ciencia e Innovación (Research Project PID2019-103996RB-I00) and by Grupos de Investigación de la Junta de Andalucía (Group number BIO109).

References

Aguilar M, Prieto P (2020) Sequence analysis of wheat subtelomeres reveals a high polymorphism among homoeologous chromosomes. Plant Genome. https://doi.org/10.1002/tpg2.20065

Albert PS, Gao Z, Danilova TV, Birchler JA (2010) Diversity of chromosomal karyotypes in maize and its relatives. Cytogenet Genome Res 129:6–16

Alfenito MR, Birchler JA (1993) Molecular characterization of a maize B chromosome centric sequence. Genetics 135:589–597

Almeida C, Fonsêca A, Bezerra dos Santos KG, Mosiolek M, Pedrosa-Harand A (2012) Contrasting evolution of a satellite DNA and its ancestral IGS rDNA in *Phaseolus* (Fabaceae). Genome 55:683–689

Ambrožová K, Mandáková T, Bureš P, Neumann P, Leitch IJ, Koblížková A, Macas J, Lysak MA (2011) Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. Ann Bot 107:255–268

Ananiev EV, Phillips RL, Rines HW (1998a) A knob-associated tandem repeat in maize capable of forming fold-back DNA segments: are chromosome knobs megatransposons? Proc Natl Acad Sci USA 95:10785–10790

Ananiev EV, Phillips RL, Rines HW (1998b) Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. Proc Natl Acad Sci USA 95:13073–13078

Ananiev EV, Chamberlin MA, Klaiber J, Svitashev S (2005) Microsatellite megatracts in the maize (*Zea mays* L.) genome. Genome 48:1061–1069

Ansari HA, Ellison NW, Griffiths AG, Williams WM (2004) A lineage-specific centromeric satellite sequence in the genus *Trifolium*. Chromosom Res 12:357–367

Aragon-Alcaide L, Miller T, Schwarzacher T, Reader S, Moore G (1996) A cereal centromeric sequence. Chromosoma 105:261–268

Avila Robledillo L, Koblizkova A, Novak P, Böttinger K, Vrbova 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 8:5838

Avila Robledillo L, Neumann P, Koblizkova A, Novak P, Vrbova I, Macas J (2020) Extraordinary sequence diversity and promiscuity of centromeric satellites in the legume tribe *Fabeae*. Mol Biol Evol 37:2341–2356

Azzalin CM, Lingner J (2015) Telomere functions grounding on *TERRA firma*. Trends Cell Biol 25:29–36

Banaei-Moghaddama AM, Martis MM, Macas J, Gundlach H, Himmelbach A et al (2015) Genes on B chromosomes: old questions revisited with new tools. Biochim Biophys Acta 1849:64–70

Bao W, Zhang W, Yang Q, Zhang Y, Han B, Gu M, Xue Y, Cheng Z (2006) Diversity of centromeric repeats in two closely related wild rice species, *Oryza officinalis* and *Oryza rhizomatis*. Mol Gen Genomics 275:421–430

Barghini E, Natali L, Cossu RM, Giordani T, Pindo M, Cattonaro F, Scalabrin S, Velasco R, Morgante M, Cavallini A (2014) The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome. Genome Biol Evol 6:776–791

Bauwens S, Van Oostveldt P, Engler G, Van Montagu M (1991) Distribution of the rDNA and three classes of highly repetitive DNA in the chromatin of interphase nuclei of *Arabidopsis thaliana*. Chromosoma 101:41–48

Bedbrook JR, Jones J, O'Dell M, Thompson RD, Flavell RB (1980) A molecular description of telomeric heterochromatin in *Secale* species. Cell 19:545–560

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. Int J Mol Sci 20:5

Belyayev A, Josefiová J, Jandová M, Mahelka V, Krak K, Mandák B (2020a) Transposons and satellite DNA: on the origin of the major satellite DNA family in the *Chenopodium* genome. Mob DNA 11:20

Belyayev A, Jandová M, Josefiová J, Kalendar R, Mahelka V, Mandák B, Krak K (2020b) The major satellite DNA families of the diploid *Chenopodium album* aggregate species: arguments for and against the "library hypothesis". PLoS One 15:e0241206

Berr A, Pecinka A, Meister A, Kreth G, Fuchs J, Blattner FR, Lysak MA, Schubert I (2006) Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Plant J 48:771–783

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

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

Bolsheva NL, Melnikova NV, Kirov IV, Dmitriev AA, Krasnov GS, Amosova AV, Samatadze TE, Yurkevich OY, Zoshchuk SA, Kudryavtseva AV, Muravenko OV (2019) Characterization of repeated DNA sequences in genomes of blue-flowered flax. BMC Evol Biol 19:49

Bowles AMC, Bechtold U, Paps J (2020) The origin of land plants is rooted in two bursts of genomic novelty. Curr Biol 30:530–536

Brajković J, Feliciello I, Bruvo-Mađarić B, Ugarković Đ (2012) Satellite DNA-like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. G3 Genes Genomes Genet 2:931

Brandes A, Thompson H, Dean C, Heslop-Harrison JS (1997) Multiple repetitive DNA sequences in the paracentromeric regions of *Arabidopsis thaliana* L. Chromosom Res 5:238–246

Brenner S (1998) Refuge of spandrels. Curr Biol 8:R669

Brown DD, Wensink PC, Jordan E (1972) A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J Mol Biol 63:57–73

Buzek J, Koutníková H, Houben A, Ríha K, Janousek B, Siroky J, Grant S, Vyskot B (1997) Isolation and characterization of X chromosome-derived DNA sequences from a dioecious plant *Melandrium album*. Chromosome Res 5:57–65

Calderón MC, Rey MD, Cabrera A, Prieto P (2014) The subtelomeric region is important for chromosome recognition and pairing during meiosis. Sci Rep 4:6488

Calderón MC, Rey MD, Martín A, Prieto P (2018) Homoeologous chromosomes from two *Hordeum* species can recognize and associate during meiosis in wheat in the presence of the *Ph1* locus. Front Plant Sci 9:585

Camacho JPM, Cabrero J, López-León MD, Martín-Peciña M, Perfectti F, Garrido-Ramos MA, Ruiz-Ruano FJ (In Preparation) On the contingent nature of satellite DNA evolution

Capesius I (1983) Sequence of the cryptic satellite DNA from the plant *Sinapis alba*. Biochim Biophys Acta 739:276–280

Carmona A, Friero E, de Bustos A, Jouve N, Cuadrado A (2013a) Cytogenetic diversity of SSR motifs within and between *Hordeum* species carrying the H genome: *H. vulgare* L. and *H. bulbosum* L. Theor Appl Genet 126:949–961

Carmona A, Friero E, de Bustos A, Jouve N, Cuadrado A (2013b) The evolutionary history of sea barley (*Hordeum marinum*) revealed by comparative physical mapping of repetitive DNA. Ann Bot 112:1845–1855

Cermak T, Kubat Z, Hobza R, Koblizkova A, Widmer A, Macas J, Vyskot B, Kejnovsky E (2008) Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes. Chromosome Res 16:961–976

Charlesworth D (2002) Plant sex determination and sex chromosomes. Heredity 88:94–101

Charlesworth D (2016) Plant sex chromosomes. Annu Rev Plant Biol 67:397–420

Cheng ZJ, Murata M (2003) A centromeric tandem repeat family originating from a part of Ty3/*gypsy*-retroelement in wheat and its relatives. Genetics 164:665–672

Cheng ZK, Dong F, Langdon T, Ouyang S, Buell CB, Gu MH, Blattner FR, Jiang J (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell 14:1691–1704

Čížková J, Hřibová A, Humplíková L, Christelová P, Suchánková P, Doležel J (2013) Molecular analysis and genomic organization of major DNA satellites in banana (*Musa* spp.). PLoS One 8: e54808

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

Cohen S, Agmon N, Sobol O, Segal D (2010) Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. Mob DNA 1:11

Contento A, Heslop-Harrison JS, Schwarzacher T (2005) Diversity of a major repetitive DNA sequence in diploid and polyploid Triticeae. Cytogenet Genome Res 109:34–42

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

Cuacos M, Franklin FCH, Heckmann S (2015) Atypical centromeres in plants—what they can tell us. Front Plant Sci 6:913

Cuadrado A, Jouve N (1994) Mapping and organization of highly-repeated DNA sequences by means of simultaneous and sequential FISH and C-banding in 6x-triticale. Chromosom Res 2:331–338

Cuadrado A, Jouve N (1995) Fluorescent in situ hybridization and C-banding analyses of highly repetitive DNA sequences in the heterochromatin of rye (*Secale montanum* Guss.) and wheat incorporating *S. montanum* chromosome segments. Genome 38:795–802

Cuadrado A, Jouve N (2002) Evolutionary trends of different repetitive DNA sequences during speciation in the genus *Secale*. J Hered 93:339–345

Cuadrado A, Jouve N (2007) The nonrandom distribution of long clusters of all possible classes of trinucleotide repeats in barley chromosomes. Chromosome Res 15:711–720

Cuadrado A, Carmona A, Jouve N (2013) Chromosomal characterization of the three subgenomes in the polyploids of *Hordeum murinum* L.: new insight into the evolution of this complex. PLoS One 8:e81385

Cuñado N, Navajas-Pérez R, de la Herrán R, Ruiz Rejón C, Ruiz Rejón M, Santos JL, Garrido-Ramos MA (2007) The evolution of sex chromosomes in the genus *Rumex* (Polygonaceae):

identification of a new species with heteromorphic sex chromosomes. Chromosome Res 15:825–832

Cusanelli E, Chartrand P (2015) Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity. Front Genet 6:143

Dawe RK (2009) Maize centromeres and knobs (neocentromeres). In: Bennetzen JL, Hake S (eds) Handbook of maize. Springer, New York, NY

Dawe RK, Henikoff S (2006) Centromeres put epigenetics in the driver's seat. Trends Biochem Sci 31:662–669

de la Herrán R, Robles F, Cuñado N, Santos JL, Ruiz Rejón M, Garrido-Ramos MA, Ruiz Rejón C (2001) A heterochromatic satellite DNA is highly amplified in a single chromosome of *Muscari* (Hyacinthaceae). Chromosoma 110:197–202

de la Herrán R, Cuñado N, Navajas-Pérez R, Santos JL, Ruiz Rejón C, Garrido-Ramos MA, Ruiz Rejón M (2005) The controversial telomeres of lily plants. Cytogenet Genome Res 109:144–147

De Lima LG, Svartman M, Kuhn GCS (2017) Dissecting the satellite DNA landscape in three cactophilic Drosophila sequenced genomes. G3 Genes Genomes Genet 7:2831–2843

Demidov D, Schubert V, Kumke K, Weiss O, Karimi-Ashtiyani R, Buttlar J et al (2014) Anti-phosphorylated histone H2AThr120: a universal microscopic marker for centromeric chromatin of mono- and holocentric plant species. Cytogenet Genome Res 143:150–156

Dennis ES, Gerlach WL, Peacock WJ (1980) Identical polypyrimidine-polypurine satellite DNAs in wheat and barley. Heredity 44:349–366

Diaz MO, Barsacchi-Pilone G, Mahon KA, Gall JG (1981) Transcripts from both DNA strands of a satellite DNA occur on lampbrush chromosome loops of the newt *Notophthalmus*. Cell 24:649–659

Dluhošová J, Išvánek 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 Q, Han F (2012) Phosphorylation of histone H2A is associated with centromere function and maintenance in meiosis. Plant J 71:800–809

Dong F, Miller JT, Jackson SA, Wang GL, Ronald PC, Jiang J (1998) Rice (*Oryza sativa*) centromeric regions consist of complex DNA. Proc Natl Acad Sci USA 95:8135–8140

Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA 110:5294–5300

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Emadzade K, Jang T-S, Macas J, Kovařík A, Novák P, Parker J, Weiss-Schneeweiss H (2014) Differential amplification of satellite PaB6 in chromosomally hypervariable *Prospero autumnale* complex (Hyacinthaceae). Ann Bot 114:1597–1608

Fachinetti D, Han JS, McMahon MA, Ly P, Abdullah A, Wong AJ, Cleveland DW (2015) DNA sequence-specific binding of CENP-B enhances the fidelity of human centromere function. Dev Cell 33:314–327

Fajkus J, Kovarík A, Královics R, Bezděk M (1995) Organization of telomeric and subtelomeric chromatin in the higher plant *Nicotiana tabacum*. Mol Gen Genet 247:633–638

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

Feliciello I, Chinali G, Ugarković Đ (2011) Structure and evolutionary dynamics of the major satellite in the red flour beetle *Tribolium castaneum*. Genetica 139:999–1008

Feliciello I, Akrap I, Ugarković Đ (2015) Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. PLoS Genet 11:e1005466

Feng W, Michaels SD (2015) Accessing the inaccessible: the organization, transcription, replication, and repair of heterochromatin in plants. Annu Rev Genet 49:439–459

Ferree PM (2017) Sex differences: satellite DNA directs male-specific gene expression. Curr Biol 27:1866

Ferreira D, Meles S, Escudeiro A, Mendes-da-Silva A, Adega F, Chaves R (2015) Satellite non-coding RNAs: the emerging players in cells, cellular pathways and cancer. Chromosom Res 23:479–493

Filatov DA, Moneger F, Negrutiu I, Charlesworth D (2000) Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution. Nature 404:388–390

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

Fitzgerald DJ, Dryden GL, Bronson EC, Williams JS, Anderson JN (1994) Conserved patterns of bending in satellite and nucleosome positioning DNA. J Biol Chem 269:21303–21314

Fominaya A, Hueros G, Loarce Y, Ferrer E (1995) Chromosomal distribution of a repeated DNA sequence from C-genome heterochromatin and the identification of a new ribosomal DNA locos in the *Avena* genus. Genome 38:548–557

Francki MG (2001) Identification of Bilby, a diverged centromeric Ty1-copia retrotransposon family from cereal rye (*Secale cereale* L.). Genome 44:266–274

Fransz P, Armstrong S, Alonso-Blanco C, Fischer TC, Torres-Ruiz RA, Jones G (1998) Cytogenetics for the model system *Arabidopsis thaliana*. Plant J Cell Mol Biol 13:867–876

Fry K, Salser W (1977) Nucleotide sequences of HS-alpha satellite DNA from kangaroo rat *Dipodomys ordii* and characterization of similar sequences in other rodents. Cell 12:1069–1084

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

Fu S, Lv Z, Gao Z, Wu H, Pang J, Zhang B, Dong Q, Guo X, Wang X, Birchler JA (2013) *De novo* centromere formation on a chromosome fragment in maize. Proc Natl Acad Sci USA 110:6033–6036

Fukagawa T, Earnshaw WC (2014) The centromere: chromatin Foundation for the Kinetochore Machinery. Dev Cell 30:496–508

Fulnečková J, Sěvčíková T, Fajkus J, Lukešová A, Lukeš M, Vlček C, Lang BF, Kim E, Eliáš M, Sýkorova E (2013) A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. Genome Biol Evol 5:468–483

Gall JG, Stephenson EC, Erba HP, Diaz MO, Barsacchi-Pilone G (1981) Histone genes are located at the sphere loci of newt lampbrush chromosomes. Chromosoma 84:159–171

Gao Z, Fu S, Dong Q, Han F, Birchler JA (2011) Inactivation of a centromere during the formation of a translocation in maize. Chromosom Res 19:755–761

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

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

Garrido-Ramos MA, de la Herrán R, Ruiz Rejón M, Ruiz Rejón C (1999) A subtelomeric satellite DNA family isolated from the genome of the dioecious plant *Silene latifolia*. Genome 42:442–446

Gent JI, Dong Y, Jiang J, Dawe RK (2012) Strong epigenetic similarity between maize centromeric and pericentromeric regions at the level of small RNAs, DNA methylation and H3 chromatin modifications. Nucl Acids Res 40:1550–1560

Gent JI, Madzima TF, Bader R, Kent MR, Zhang X et al (2014) Accessible DNA and relative depletion of H3K9me2 at maize loci undergoing RNA-directed DNA methylation. Plant Cell 26:4903–4917

Gent JI, Wang K, Jiang J, Dawe RK (2015) Stable patterns of CENH3 occupancy through maize lineages containing genetically similar centromeres. Genetics 200:1105–1116

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 18:121

Gent JI, Nannas NJ, Liu Y, Su H, Zhao H, Gao Z, Dawe RK, Jiang J, Han F, Birchler JA (2018) Genomics of maize centromeres. In: Bennetzen J, Flint-Garcia S, Hirsch C, Tuberosa R (eds) The maize genome. Compendium of plant genomes. Springer, Cham

Gernand D, Demidov D, Houben A (2003) The temporal and spatial pattern of histone H3 phosphorylation at serine 28 and serine 10 is similar in plants but differs between mono-and polycentric chromosomes. Cytogenet Genome Res 101:172–176

Gindullis F, Desel C, Galasso I, Schmidt T (2001) The large-scale organization of the centromeric region in *Beta* species. Genome Res 11:253–265

Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, Iovene M, Neumann P, Zhang W, Novák P, Buell CR, Macas J, Jiang J (2012) Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell 24:3559–3574

González ML, Chiapella J, Topalian J, Urdampilleta JD (2020) Genomic differentiation of *Deschampsia antarctica* and *D. cespitosa* (Poaceae) based on satellite DNA. Bot J Linn Soc 194:326–341

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5:578–590

Graur D, Zheng Y, Azevedo RBR (2015) An evolutionary classification of genomic function. Genome Biol Evol 7:642–645

Grebenstein B, Grebenstein O, Sauer W, Hemleben V (1996) Distribution and complex organization of satellite DNA sequences in Aveneae species. Genome 39:1045–1050

Grellet F, Delcasso D, Panabieres F, Delseny M (1986) Organization and evolution of a higher plant alphoid-like satellite DNA sequence. J Mol Biol 187:495–507

Guerra M (2000) Patterns of heterochromatin distribution in plant chromosomes. Genet Mol Biol 23:1029–1041

Guttman DS, Charlesworth D (1998) An X-linked gene with a degenerate Y-linked homologue in a dioecious plant. Nature 393:1009–1014

Haaf T, Mater AG, Wienberg J, Ward DC (1995) Presence and abundance of CENP-B box sequences in great ape subsets of primate-specific alpha-satellite DNA. J Mol Evol 41:487–491

Hall SE, Kettler G, Preuss D (2003) Centromere satellites from *Arabidopsis* populations: maintenance of conserved and variable domains. Genome Res 13:195–205

Hall LE, Mitchell SE, O'Neill RJ (2012) Pericentric and centromeric transcription: a perfect balance required. Chromosom Res 20:535–546

Hall LL, Byron M, Carone DM, Whitfield TW, Pouliot GP, Fischer A, Jones P, Lawrence JB (2017) Demethylated HSATII DNA and HSATII RNA foci sequester PRC1 and MeCP2 into cancer-specific nuclear bodies. Cell Rep 18:2943–2956

Hallden C, Bryngelsson T, Sall T, Gustafsson M (1987) Distribution and evolution of a tandemly repeated DNA sequence in the family Brassicaceae. J Mol Evol 25:318–323

Han F, Lamb JC, Birchler JA (2006) High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize. Proc Natl Acad Sci USA 103:3238–3243

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

He Q, Cai Z, Hu T, Liu H, Bao C, Mao W, Jin W (2015) Repetitive sequence analysis and karyotyping reveals centromere-associated DNA sequences in radish (*Raphanus sativus* L.). BMC Plant Biol 15:105

Heacock M, Spangler E, Riha K, Puizina J, Shippen DE (2004) Molecular analysis of telomere fusions in *Arabidopsis*: multiple pathways for chromosome endjoining. EMBO J 23:2304–2313

Heckmann S, Schroeder-Reiter E, Kumke K, Ma L, Nagaki K, Murata M, Wanner G, Houben A (2011) Holocentric chromosomes of *Luzula elegans* are characterized by a longitudinal centromere groove, chromosome bending, and a terminal nucleolus organizer region. Cytogenet Genome Res 134:220–228

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

Heckmann S, Jankowska M, Schubert V, Kumke K, Ma W, Houben A (2014a) Alternative meiotic chromatid segregation in the holocentric plant *Luzula elegans*. Nat Commun 5:4979

Heckmann S, Schubert V, Houben A (2014b) Holocentric plant meiosis: first sisters, then homologues. Cell Cycle 13:3623–3624

Henikoff S (2002) Near the edge of a chromosome's 'black hole'. Trends Genet 18:165–167

Henikoff S, Ahmad K, Platero JS, van Steensel B (2000) Heterochromatic deposition of centromeric histone H3-like proteins. Proc Natl Acad Sci USA 97:716–721

Henikoff JG, Thakur J, Kasinathan S, Henikoff S (2015) A unique chromatin complex occupies young α-satellite arrays of human centromeres. Sci Adv 1:e1400234

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

Hobza R, Lengerova M, Svoboda J, Kubekova H, Kejnovsky E, Vyskot B (2006) An accumulation of tandem DNA repeats on the Y chromosome in *Silene latifolia* during early stages of sex chromosome evolution. Chromosoma 115:376–382

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 8:302

Holoch D, Moazed D (2015) RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet 16:71–84

Houben A, Wako T, Furushima-Shimogawara R, Presting G, Künzel G, Schubert I et al (1999) The cell cycle dependent phosphorylation of histone H3 is correlated with the condensation of plant mitotic chromosomes. Plant J 18:675–679

Houben A, Demidov D, Caperta AD, Karimi R, Agueci F, Vlasenko L (2007) Phosphorylation of histone H3 in plants: a dynamic affair. Biochim Biophys Acta 1769:308–315

Hribová E, Neumann P, Matsumoto T, Roux N, Macas J, Dolezel J (2010) Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. BMC Plant Biol 10:204

Hu TT et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481

Hudakova S, Michalek W, Presting GG, ten Hoopen R, dos Santos K, Jasencakova Z, Schubert I (2001) Sequence organization of barley centromeres. Nucl Acid Res 29:5029–5035

Iwata A, Tek AL, Richard MMS, Abernathy B, Fonseca A, Schmutz J, Chen NWG, Thareau V, Magdelenat G, Li Y, Murata M, Pedrosa-Harand A, Geffroy V, Nagaki K, Jackson SA (2013) Identification and characterization of functional centromeres of the common bean. Plant J 76:47–60

Jain M, Olsen HE, Turner DJ et al (2018) Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol 36:321–323

Jesionek W, Bodláková M, Kubát Z, Čegan R, Vyskot B, Vrána J, Šafář J, Puterova J, Hobza R (2020) Fundamentally different repetitive element composition of sex chromosomes in *Rumex acetosa*. Ann Bot. https://doi.org/10.1093/aob/mcaa160

Jiang J, Nasuda A, Dong F, Scherrer CW, Woo SS et al (1996) A conserved repetitive DNA element located in the centromeres of cereal chromosomes. Proc Natl Acad Sci USA 93:14210–14213

Jiang J, Birchler JA, Parrott WA, Dawe RK (2003) A molecular view of plant centromeres. Trends Plant Sci 8:570–575

Jin W, Melo JR, Nagaki K, Talbert PB, Henikoff S et al (2004) Maize centromeres: organization and functional adaptation in the genetic background of oat. Plant Cell 16:571–581

Jo S-H, Koo D-H, Kim JF, Hur C-G, Lee S, Yang T-J, Kwon S-Y, Choi D (2009) Evolution of ribosomal DNA-derived satellite repeat in tomato genome. BMC Plant Biol 9:42

John B, King M, Schweizer D, Mendelak M (1985) Equilocality of heterochromatin distribution and heterochromatin heterogeneity in acridid grasshoppers. Chromosoma 91:185–200

Johnson WL, Straight AF (2017) RNA-mediated regulation of heterochromatin. Curr Opin Cell Biol 46:102–109

Joshi SS, Meller VH (2017) Satellite repeats identify x chromatin for dosage compensation in *Drosophila melanogaster* males. Curr Biol 27:1393–1402

Kamm A, Galasso I, Schmidt T, Heslop-Harrison JS (1995) Analysis of a repetitive DNA family from Arabidopsis arenosa and relationships between Arabidopsis species. Plant Mol Biol 27:853–862

Kapitonov VV, Jurka J (1999) Molecular paleontology of transposable elements from *Arabidopsis thaliana*. Genetica 107:27–37

Kapustova V, Tulpova Z, Toegelova H, Novak P, Macas J, Karafiatova M, Hribova E, Dolezel J, Simkova H (2019) The dark matter of large cereal genomes: long tandem repeats. Int J Mol Sci 20:2483

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

Kaszás E, Cande WZ (2000) Phosphorylation of histone H3 is correlated with changes in the maintenance of sister chromatid cohesion during meiosis in maize, rather than the condensation of the chromatin. J Cell Biochem 113:3217–3226

Kato A, Yakura K, Tanifuji S (1984) Sequence analysis of *Vicia faba* repeated DNA, the FokI repeat element. Nucl Acids Res 16:6415–6426

Kawabe A, Nasuda S (2005) Structure and genomic organization of centromeric repeats in *Arabidopsis* species. Mol Gen Genomics 272:593–602

Kazama Y, Sugiyama R, Suto Y, Uchida W, Kawano S (2006) The clustering of four subfamilies of satellite DNA at individual chromosome ends in *Silene latifolia*. Genome 49:520–530

Kejnovský E, Hobza R, Kubat Z, Cermak T, Vyskot B (2009) The role of repetitive DNA in structure and evolution of sex chromosomes in plants. Heredity 102:533–541

Kejnovský E, Michalovova M, Steflova P, Kejnovska I, Manzano S, Hobza R, Kubat Z, Kovarik J, Jamilena M, Vyskot B (2013) Expansion of microsatellites on evolutionary young Y chromosome. PLoS One 8:e45519

Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novák P, Neumann P, Lysak MA, Day PD, Berger M, Fay MF, Nichols RA, Leitch AR, Leitch IJ (2015) Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol 208:596–607

Kishii M, Nagaki K, Tsujimoto H, Sasakuma T (1999) Exclusive localization of tandem repetitive sequences in subtelomeric heterochromatin regions of *Leymus racemosus* (Poaceae, Triticeae). Chromosom Res 7:519–529

Klemme S, Banaei-Moghaddam AM, Macas J, Wicker T, Novák P, Houben A (2013) High-copy sequences reveal distinct evolution of the rye B chromosome. New Phytol 199:550–558

Koo DH, Han FP, Birchler JA, Jiang JM (2011) Distinct DNA methylation patterns associated with active and inactive centromeres of the maize B chromosome. Genome Res 21:908–914

Koch J (2000) Neocentromeres and alpha satellite: a proposed structural code for functional human centromere DNA. Hum Mol Genet 92:149–154

Kowar T, Zakrzewski F, Macas J, Koblizkova 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 16:120

Kubat Z, Zluvova J, Vogel I et al (2014) Possible mechanisms responsible for absence of a retrotransposon family on a plant Y chromosome. New Phytol 202:662–678

Kuhn GSC, Küttler H, Moreira-Filho O, Heslop-Harrison JS (2012) The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. Mol Biol Evol 29:7–11

Kulikova O, Geurts R, Lamine M, Kim DJ, Cook DR, Leunissen J, de Jong H, Roe BA, Bisseling T (2004) Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. Chromosoma 113:276–283

Kumke K, Macas J, Fuchs J, Altschmied L, Kour J, Dhar MK, Houben A (2016) *Plantago lagopus* B chromosome is enriched in 5S rDNA-derived satellite DNA. Cytogenet Genome Res 148:68–73

Kurihara D, Matsunaga S, Kawabe A, Fujimoto S, Noda M, Uchiyama S et al (2006) Aurora kinase is required for chromosome segregation in tobacco BY-2 cells. Plant J 48:572–580

Kwapisz M, Morillon A (2020) Subtelomeric transcription and its regulation. J Mol Biol 432:4199–4219

Lamb JC, Kato A, Birchler JA (2005) Sequences associated with a chromosome centromeres are present throughout the maize B chromosome. Chromosoma 113:337–349

Lamb JC, Meyer JM, Birchler JA (2007) A hemicentric inversion in the maize line knobless Tama flint created two sites of centromeric elements and moved the kinetochore-forming region. Chromosoma 116:237–247

Larracuente AM (2014) The organization and evolution of the responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. BMC Evol Biol 14:233

Lee HR, 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 USA 102:11793–11798

Lee H-R, Neumann P, Macas J, Jiang J (2006) Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice. Mol Biol Evol 23:2505–2520

Lee YI, Yap JW, Izan S, Leitch IJ, Fay MF, Lee YC, Hidalgo O, Dodsworth S, Smulders MJM, Gravendeel B, Leitch AR (2018) Satellite DNA in *Paphiopedilum* subgenus *Parvisepalum* as revealed by high throughput sequencing and fluorescent in situ hybridization. BMC Genomics 19:578

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

Lermontova I, Sandmann M, Mascher M, Schmit AC, Chabout ME (2015) Centromeric chromatin and its dynamics in plants. Plant J 83:4–17

Li B, Choulet F, Heng Y, Hao W, Paux P, Liu Z, Yue W, Jin W, Feuillet C, Zhang X (2013) Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. Plant J 73:952–965

Li SF, Guo YJ, Li JR, Zhang DX, Wang BX, Li N, Deng CL, Gao WJ (2019) The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (*Spinacia oleracea* L.). Mobile DNA 10:3

Lim KY, Kovarik A, Matyášek R, Chase MW, Knapp S, McCarthy E, Clarkson JJ, Leitch AR (2006) Comparative genomics and repetitive sequence divergence in the species of diploid *Nicotiana* section Alatae. Plant J 48:907–919

Liu Z, Yue W, Li D, Wang R, Kong X, Lu K, Wang G, Dong Y, Jin W, Zhang X (2008) Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. Chromosoma 117:445–456

Liu Y, Su H, Pang J, Gao Z, Wang XJ et al (2015) Sequential de novo centromere formation and inactivation on a chromosomal fragment in maize. Proc Natl Acad Sci USA 112:E1263–E1271

López-Flores I, Garrido-Ramos MA (2012) The repetitive DNA content of eukaryotic genomes. In: Garrido-Ramos MA (ed) Repetitive DNA. Genome dynamics, vol 7. Karger, Basel, pp 126–152

Lorite P, Muñoz-López M, Carrillo JA, Sanllorente O, Vela J, Mora P, Tinaut A, Torres MI, Palomeque T (2017) Concerted evolution, a slow process for ant satellite DNA: study of the satellite DNA in the *Aphaenogaster* genus (Hymenoptera, Formicidae). Org Divers Evol 17:595–606

Luchetti A, Cesari M, Carrara G, Cavicchi S, Passamonti M, Scali V, Mantovani B (2003) Unisexuality and molecular drive: Bag320 sequence diversity in *Bacillus* taxa (Insecta Phasmatodea). J Mol Evol 56:587–596

Luchetti A, Marini M, Mantovani B (2006) Non-concerted evolution of RET76 satellite DNA family in Reticulitermes taxa (Insecta, Isoptera). Genetica 128:123–132

Ma J, Wing RA, Bennetzen JL, Jackson SA (2007) Plant centromere organization: a dynamic structure with conserved functions. Trends Genet 23:135–139

Macas J, Neumann P (2007) Ogre elements: a distinct group of plant Ty3/gypsy-like retrotransposons. Gene 390:108–116

Macas J, Pozarkova D, Navratilova A, Nouzova M, Neumann P (2000) Two new families of tandem repeats isolated from genus *Vicia* using genomic self-priming PCR. Mol Gen Genet 263:741–751

Macas J, Meszaros T, Nouzova M (2002) PlantSat: a specialized database for plant satellite repeats. Bioinformatics 18:28–35

Macas J, Navratilova A, Meszaros 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, Navratilova A, Koblizkova 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, Navratilova 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, Koblizkova A, Navratilova A, Neumann P (2009) Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448:198–206

Macas J, Neumann P, Novak P, Jiang J (2010) Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data. Bioinformatics 26:2101–2108

Macas J, Kejnovsky E, Neumann P, Novak P, Koblizkova A, Vyskot B (2011) Next generation sequencing-based analysis of repetitive DNA in the model dioecious plant *Silene latifolia*. PLoS One 6:e27335

Macas J, Novak P, Pellicer J, Cizkova J, Koblizkova A, Neumann P, Fukova I, Dolezel J, Kelly L, Leitch I (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

Maheshwari S, Tan EH, West A, Franklin FC, Comai L, Chan SW (2015) Naturally occurring differences in CENH3 affect chromosome segregation in zygotic mitosis of hybrids. PLoS Genet 11:e1004970

Malik HS, Henikoff S (2003) Phylogenomics of the nucleosome. Nat Struct Mol Biol 10:882–891

Manzanero S, Arana P, Puertas MJ, Houben A (2000) The chromosomal distribution of phosphorylated histone H3 differs between plants and animals at meiosis. Chromosoma 109:308–317

Mariotti B, Navajas-Pérez R, Lozano R, Parker JS, de la Herrán R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos M, Jamilena M (2006) Cloning and characterization of dispersed repetitive DNA derived from microdissected sex chromosomes of *Rumex acetosa*. Genome 49:114–121

Mariotti B, Manzano S, Kejnovský E, Vyskot B, Jamilena M (2009) Accumulation of Y-specific satellite DNAs during the evolution of *Rumex acetosa* sex chromosomes. Mol Gen Genomics 281:249–259

Marques A, Ribeiro T, Neumann P, Macas J, Novák P, Schubert V, Pellino M, Fuchs J, Ma W, Kuhlmann M et al (2015) Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. Proc Natl Acad Sci USA 112:13633–13638

Marques A, Schubert V, Houben A, Pedrosa-Harand A (2016) Restructuring of holocentric centromeres during meiosis in the plant *Rhynchospora pubera*. Genetics 204:555–568

Martienssen R, Moazed D (2015) RNAi and heterochromatin assembly. Cold Spring Harb Perspect Biol 7:a019323

Martinez-Zapater JM, Estelle MA, Somerville CR (1986) A highly repeated DNA sequence in *Arabidopsis thaliana*. Mol Gen Genet 204:417–423

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T (1989) A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol 109:1963–1973

Masumoto H, Nakano M, Ohzeki J (2004) The role of CENP-B and alpha-satellite DNA: De novo assembly and epigenetic maintenance of human centromeres. Chromosom Res 12:543–556

Matyášek R, Renny-Byfield S, Fulnecek J, Macas J, Grandbastien MA, Nichols RA, Leitch AR, Kovarik A (2012) Next generation sequencing analysis reveals a relationship between rDNA unit diversity and locus number in *Nicotiana* diploids. BMC Genomics 13:722

May BP, Lippman ZB, Fang Y, Spector DL, Martienssen RA (2005) Differential regulation of strandspecific transcripts from Arabidopsis centromeric satellite repeat. PLoS Genet 1:e79

McCann J, Macas J, Novák P, Stuessy TF, Villasenor JL, Weiss-Schneeweiss H (2020) Differential genome size and repetitive DNA evolution in diploid species of *Melampodium* sect. *Melampodium* (Asteraceae). Front Plant Sci 11:362

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

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

Melters DP, Paliulis LV, Korf I, Chan SW (2012) Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. Chromosom Res 20:579–593

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, Garcia JF, DeRisi JL, Smith T, Tobias C, Ross-Ibarra J, Korf I, Chan SW (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10

Menon DU, Coarfa C, Xiao W, Gunaratne PH, Meller VH (2014) siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. Proc Natl Acad Sci USA 111:16460–16465

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

Miller JT, Dong F, Jackson SA, Song J, Jiang J (1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. Genetics 150:1615–1623

Mizuno H, Wu J, Katayose Y, Kanamori H, Sasaki T, Matsumoto T (2008) Chromosome-specific distribution of nucleotide substitutions in telomeric repeats of rice (*Oryza sativa* L.). Mol Biol Evol 25:62–68

Mlinarec J, Skuhala A, Jurkovic A, Malenica N, McCann J, Weiss-Schneeweiss H, Bohanec B, Besendorfer V (2019) The repetitive DNA composition in the natural pesticide producer *Tanacetum cinerariifolium*: interindividual variation of subtelomeric tandem repeats. Front Plant Sci 10:613

Mravinac B, Plohl M, Ugarkovic D (2005) Preservation and high sequence conservation of satellite DNAs indicate functional constraints. J Mol Evol 61:542–550

Muro Y, Masumoto H, Yoda K, Nozaki N, Ohashi M, Okazaki T (1992) Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box. J Cell Biol 116:585–596

Nagaki K, Murata M (2005) Characterization of CENH3 and centromere-associated DNA sequences in sugarcane. Chromosom Res 13:195–203

Nagaki K, Tsujimoto H, Sasakuma T (1998) A novel repetitive sequence of sugar cane, SCEN family, locating on centromeric regions. Chromosom Res 6:295–302

Nagaki K, Song J, Stupar SM, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones KM, Dawe RK, Buell CR, Jiang J (2003a) Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics 163:759–770

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

Nagaki K, Neumann P, Zhang DF, Ouyang S, Buell CR, Cheng ZK, Jiang JM (2005) Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. Mol Biol Evol 22:845–855

Nasuda S, Hudakova S, Schubert I, Houben A, Endo TR (2005) Stable barley chromosomes without centromeric repeats. Proc Natl Acad Sci USA 102:9842–9847

Navajas-Pérez R, de la Herrán R, Jamilena M, Lozano R, Rejón CR, Ruiz Rejón MR, Garrido-Ramos MA (2005a) Reduced rates of sequence evolution of Y-linked satellite DNA in *Rumex* (Polygonaceae). J Mol Evol 60:391–399

Navajas-Pérez R, de la Herrán R, López González G, Jamilena M, Lozano R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA (2005b) The evolution of reproductive systems and sex-determining mechanisms within *Rumex* (Polygonaceae) inferred from nuclear and chloroplastidial sequence data. Mol Biol Evol 22:1929–1939

Navajas-Pérez R, Schwarzacher T, de la Herrán R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA (2006) The origin and evolution of the variability in a Y-specific satellite-DNA of *Rumex acetosa* and its relatives. Gene 368:61–71

Navajas-Pérez R, Quesada del Bosque ME, Garrido-Ramos MA (2009a) Effect of location, organization and repeat-copy number in satellite-DNA evolution. Mol Genet Genomics 282:395–406

Navajas-Pérez R, Schwarzacher T, Ruiz Rejón M, Garrido-Ramos MA (2009b) Characterization of RUSI, a telomere-associated satellite DNA, in the genus *Rumex* (Polygonaceae). Cytogenet Genome Res 124:81–89

Navajas-Pérez R, Schwarzacher T, Ruiz Rejón M, Garrido-Ramos MA (2009c) Molecular cytogenetic characterization of *Rumex papillaris*, a dioecious plant with an XX/XY1Y2 sex chromosome system. Genetica 135:87–93

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

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

Neumann P, Navrátilová A, Koblizkova A, Kejnovsky E, Hribova E, Hobza R, Widmer A, Dolezel J, Macas J (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA 2:4

Neumann P, Navrátilová A, Schroeder-Reiter E, Koblizkova A, Steinbauerova V, Chocholova E, Novak P, Wanner G, Macas J (2012) Stretching the rules: monocentric chromosomes with multiple centromere domains. PLoS Genet 8:e1002777

Neumann P, Pavlikova Z, Koblizkova A, Fukova I, Jedlickova V, Novak 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 32:1862–1879

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

Neumann P, Novak P, Hostakova 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, Oliveira L, Cizkova J, Jang TS, Klemme S, Novak P, Stelmach K, Koblizkova A, Dolezel J, Macas J (2020) Impact of parasitic lifestyle and different types of centromere organization on chromosome and genome evolution in the plant genus *Cuscuta*. New Phytol. https://doi.org/10.1111/nph.17003

Nonomura KI, Kurata N (1999) Organization of the 1.9-kb repeat unit RCE1 in the centromeric region of rice chromosomes. Mol Gen Genet 261:1–10

Nonomura KI, Kurata N (2001) The centromere composition of multiple repetitive sequences on rice chromosome 5. Chromosoma 110:284–291

Nouzová M, Kubaláková M, Doleželová M, Koblížková A, Neumann P, Doleel J, Macas J (1999) Cloning and characterization of new repetitive sequences in field bean (*Vicia faba* L.). Ann Bot 83:535–541

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 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 29:792–793

Novák P, Avila Robledillo L, Koblizkova A, Vrbova 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:e111

Novák P, Guignard MS, Neumann P, Kelly LJ, Mlinarec J, Koblizkova A, Dodsworth S, Kovarik A, Pellicer J, Wang W, Macas J, Leitch IJ, Leitch AR (2020a) Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat Plants 6:1325–1329

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

Ohno S (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23:366–370

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

Oliveira L, Neumann P, Jang TS, Klemme S, Schubert V, Koblizkova A, Houben A, Macas J (2020) Mitotic spindle attachment to the holocentric chromosomes of *Cuscuta europaea* does not correlate with the distribution of CENH3 chromatin. Front Plant Sci 10:1799

Pavlek M, Gelfand Y, Plohl M, Meštrović N (2015) Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. DNA Res 22:387–401

Peacock WJ, Dennis ES, Rhoades MM, Pryor AJ (1981) Highly repeated DNA sequence limited to knob heterochromatin in maize. Proc Natl Acad Sci USA 78:4490–4494

Pedersen C, Rasmussen SK, Linde-Laursen I (1996) Genome and chromosome identification in cultivated barley and related species of the Triticeae (Poaceae) by in situ hybridization with the GAA-satellite sequence. Genome 39:93–104

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ (2018) Genome size diversity and its impact on the evolution of land plants. Genes 9:88

Pérez-Gutiérrez MA, Suárez-Santiago VN, López-Flores I, Romero AT, Garrido-Ramos MA (2012) Concerted evolution of satellite DNA in *Sarcocapnos*: a matter of time. Plant Mol Biol 78:19–29

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

Pezer Z, Brajković J, Feliciello I, Ugarković Đ (2012) Satellite DNA-mediated effects on genome regulation. In: Garrido-Ramos MA (ed) Repetitive DNA. Genome dynamics, vol 7. Karger, Basel, pp 153–169

Piednoël M, Aberer AJ, Schneeweiss GM, Macas J, Novak P, Gundlach H, Temsch EM, Renner SS (2012) Next-generation sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. Mol Biol Evol 29:3601–3611

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16:1262–1269

Pinosio S, Marroni F, Zuccolo A, Vitulo N, Mariette S, Sonnante G, Aravanopoulos FA, Ganopoulos I, Palasciano M, Vidotto M, Magris G, Iezzoni A, Vendramin GG, Morgante M (2020) A draft genome of sweet cherry (*Prunus avium* L.) reveals genome-wide and local effects of domestication. Plant J 103:1420–1432

Pita S, Panzera F, Mora P, Vela J, Cuadrado Á, Sánchez A, Palomeque T, Lorite P (2017) Comparative repeatome analysis on *Triatoma infestans* Andean and non-Andean lineages, main vector of Chagas disease. PLoS One 12:e0181635

Plohl M, Luchetti A, Meštrović N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)-chromatin. Gene 409:72–82

Plohl M, Petrović V, Luchetti A, Ricci A, Satović E, Passamonti M, Mantovani B (2010) Long-term conservation vs high sequence divergent: the case of an extraordinarily old satellite DNA in bivalve mollusks. Heredity 104:543–551

Plohl M, Meštrović N, Mravinac B (2012) Satellite DNA evolution. In: Garrido-Ramos MA (ed) Repetitive DNA. Genome dynamics, vol 7. Karger, Basel, pp 126–152

Presting GG, Malysheva L, Fuchs J, Schubert I (1998) A Ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. Plant J 16:721–728

Puterova J, Razumova O, Martinek T, Alexandrov O, Divashuk M, Kubat Z, Hobza R, Karlov G, Kejnovsky E (2017) Satellite DNA and transposable elements in seabuckthorn (*Hippophae rhamnoides*), a dioecious plant with small Y and large x chromosomes. Genome Biol Evol. https://doi.org/10.1093/gbe/evw303

Quesada del Bosque ME, Navajas-Pérez R, Panero JL, Fernández-González A, Garrido-Ramos MA (2011) A satellite DNA evolutionary analysis in the North American endemic dioecious plant *Rumex hastatulus* (Polygonaceae). Genome 54:253–260

Quesada del Bosque ME, López-Flores I, Suárez-Santiago VN, Garrido-Ramos MA (2013) Differential spreading of HinfI satellite DNA variants during radiation in Centaureinae. Ann Bot 112:1793–1802

Quesada del Bosque ME, López-Flores I, Suárez-Santiago VN, Garrido-Ramos MA (2014) Satellite-DNA diversification and the evolution of major lineages in Cardueae (Carduoideae, Asteraceae). J Plant Res 127:575–583

Ribeiro T, Marques A, Novák P, Schubert V, Vanzela ALL, Macas J, Houben A, Pedrosa-Harand A (2017) Centromeric and non-centromeric satellite DNA organization differs in holocentric Rhynchospora species. Chromosoma 126:325–335

Richard MMS, Chen NWG, Thareau V, Pflieger E, Blanchet S, Pedrosa-Harand A, Iwata A, Chavarro C, Jackson SA, Geffroy V (2013) The subtelomeric khipu satellite repeat from *Phaseolus vulgaris*: lessons learned from the genome analysis of the Andean genotype G19833. Front Plant Sci 4:1–14

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

Rošić S, Köhler F, Erhardt S (2014) Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. J Cell Biol 207:335–349

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM (2016) High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep 6:28333

Ruiz-Ruano FJ, Cabrero J, López-León MD, Camacho JPM (2017) Satellite DNA content illuminates the ancestry of a supernumerary (B) chromosome. Chromosoma 126:487–500

Ruiz-Ruano FJ, Castillo-Martínez J, Cabrero J, Gómez R, Camacho JPM, López-León MD (2018) High-throughput analysis of satellite DNA in the grasshopper *Pyrgomorpha conica* reveals abundance of homologous and heterologous higher-order repeats. Chromosoma 127:323–340

Ruiz-Ruano FJ, Navarro-Domínguez B, Camacho JPM, Garrido-Ramos MA (2019) Characterization of the satellitome in lower vascular plants: the case of the endangered fern *Vandenboschia speciosa*. Ann Bot 123:587–599

Saint-Leandre B, Levine MT (2020) The telomere paradox: stable genome preservation with rapidly evolving proteins. Trends Genet 36:4232–4242

Šatović E, Plohl M (2013) Tandem repeat-containing MITEs in the clam *Donax trunculus*. Genome Biol Evol 5:2549–2559

Schmidt M, Hense S, Minoche AE, Dohm JC, Himmelbauer H, Schmidt T, Zakrzewski F (2014) Cytosine methylation of an ancient satellite family in the wild beet *Beta procumbens*. Cytogenet Genome Res 143:157–167

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schneider KL, Xie ZD, Wolfgruber TK, Presting GG (2016) Inbreeding drives maize centromere evolution. Proc Natl Acad Sci USA 113:E987–E996

Schubert V, Neumann P, Marques A, Heckmann S, Macas J, Pedrosa-Harand A, Schubert I, Jang TS, Houben A (2020) Super-resolution microscopy reveals diversity of plant centromere architecture. Int J Mol Sci 21:3488

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional human centromere. Science 294:109–115

Schweizer D, Loidl J (1987) A model for heterochromatin dispersion and the evolution of C-band patterns. In: Hayman DL, Rofe RH, Sharp PJ (eds) Chromosomes today, vol 9. Allen & Unwin, London, pp 61–74

Setiawan AB, Teo CH, Kikuchi S, Sassa H, Kato K, Koba T (2020) Centromeres of *Cucumis melo* L. comprise *Cmcent* and two novel repeats, *CmSat162* and *CmSat189*. PLoS One 15:e0227578

Sharma A, Presting GG (2014) Evolution of centromeric retrotransposons in grasses. Genome Biol Evol 6:1335–1352

Sharma A, Wolfgruber TK, Presting GG (2013) Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14:142

Shibata F, Hizume M, Kurori Y (1999) Chromosome painting of Y chromosomes and isolation of a Y chromosome-specific repetitive sequence in the dioecious plant *Rumex acetosa*. Chromosoma 108:266–270

Shibata F, Hizume M, Kurori Y (2000a) Differentiation and the polymorphic nature of the Y chromosomes revealed by repetitive sequences in the dioecious plant, *Rumex acetosa*. Chromosome Res 8:229–236

Shibata F, Hizume M, Kurori Y (2000b) Molecular cytogenetic analysis of supernumerary heterochromatic segments in *Rumex acetosa*. Genome 43:391–397

Simoens CR, Gielen J, Van Montagu M, Inze D (1988) Characterization of highly repetitive sequences of *Arabidopsis thaliana*. Nucleic Acids Res 16:6753–6766

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Steflova P, Tokan V, Vogel I, Lexa M, Macas J, Novák P, Hobza R, Vyskot B, Kejnovsky E (2013) Contrasting patterns of transposable element and satellite distribution on sex chromosomes (XY_1Y_2) in the dioecious plant *Rumex acetosa*. Genome Biol Evol 5:769–782

Stephan W (1989) Tandem-repetitive noncoding DNA: forms and forces. Mol Biol Evol 6:198–212

Stupar RM, Song J, Tek AL, Cheng Z, Dong F, Jiang J (2002) Highly condensed potato pericentromeric heterochromatin contains rDNA-related tandem repeats. Genetics 162:1435–1444

Su H, Liu Y, Liu YX, Lv Z, Li H, Xie S et al (2016) Dynamic chromatin changes associated with de novo centromere formation in maize euchromatin. Plant J 88:854–866

Suárez-Santiago VN, Blanca G, Ruiz-Rejón M, Garrido-Ramos MA (2007) Satellite-DNA evolutionary patterns under a complex evolutionay scenario: the case of *Acrolophus* subgroup (*Centaurea* L., Compositae) from the western Mediterranean. Gene 404:80–92

Sullivan BA, Karpen GH (2004) Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nat Struct Mol Biol 11:1076–1083

Sýkorová E, Lim KY, Kunická Z, Chase MW, Bennett MD, Fajkus J, Leitch AR (2003a) Telomere variability in the monocotyledonous plant order Asparagales. Proc R Soc Lond B 270:1893–1904

Sýkorová E, Cartagena J, Horáková M, Fukui K, Fajkus J (2003b) Characterization of telomere-subtelomere junctions in *Silene latifolia*. Mol Gen Genomics 269:13–20

Sýkorová E, Lim KY, Chase MW, Knapp S, Leitch IJ, Leitch AR, Fajkus J (2003c) The absence of Arabidopsis-type telomeres in *Cestrum* and closely related genera *Vestia* and *Sessea* (Solanaceae); first evidence from eudicots. Plant J 34:283–291

Talbert PB, Henikoff S (2018) Transcribing centromeres: noncoding RNAs and kinetochore assembly. Trends Genet 34:587–599

Talbert PB, Henikoff S (2020) What makes a centromere? Exp Cell Res 389:111895

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

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

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Tek AL, Jiang J (2004) The centromeric regions of potato chromosomes contain megabase-sized tandem arrays of telomere-similar sequence. Chromosoma 113:77–83

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

Topp CN, Zhong CX, Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. Proc Natl Acad Sci USA 101:15986–15991

Torres GA, Gong Z, Iovene M, Hirsch CD, Buell CR, Bryan GJ, Novák P, Macas J, Jiang (2011) Organization and evolution of subtelomeric satellite repeats in the potato genome. G3 Genes Genomes Genetics 1:85–92

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018) The third revolution in sequencing technology. Trends Genet 34:666–681 van Emden TS, Forn M, Forné I, Sarkadi Z, Capella M, Martín Caballero L, Fischer-Burkart S, Brönner C, Simonetta M, Toczyski D, Halic M, Imhof A, Braun S (2019) Shelterin and subtelomeric DNA sequences control nucleosome maintenance and genome stability. EMBO Rep 20:e47181

Varley JM, Macgregor HC, Erba HP (1980a) Satellite DNA is transcribed on lampbrush chromosomes. Nature 283:686–688

Varley JM, Macgregor HC, Nardi I, Andrews C, Erba HP (1980b) Cytological evidence of transcription of highly repeated DNA sequences during the lampbrush stage in *Triturus cristatus carnifex*. Chromosoma 80:289–307

Vershinin AV, Heslop-Harrison JS (1998) Comparative analysis of the nucleosomal structure of rye, wheat and their relatives. Plant Mol Biol 36:149–161

Vershinin AV, Alkhimova EG, Heslop-Harrison JS (1996) Molecular diversification of tandemly organized DNA sequences and heterochromatic chromosome regions in some Triticeae species. Chromosom Res 4:517–525

Viotti A, Privitera E, Sala E, Pogna N (1985) Distribution and clustering of two highly repeated sequences in the A and B chromosomes of maize. Theor Appl Genet 70:234–239

Volpe T, Martienssen RA (2011) RNA interference and heterochromatin assembly. Cold Spring Harb Perspect Biol 3:a003731

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SIS, Martienssen R (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science 297:1833–1837

Vondrak T, Avila Robledillo L, Novak P, Koblizkova 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 101:484–500

Vrbsky J, Akimcheva S, Watson JM, Turner TL, Daxinger L, Vyskot B, Aufsatz W, Riha K (2010) siRNA-mediated methylation of *Arabidopsis* telomeres. PLoS Genet 6:e1000986

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

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

Wang G, Zhang X, Jin W (2009) An overview of plant centromeres. J Genet Genomics 36:529–537

Wang K, Wu Y, Zhang W, Dawe RK, Jiang J (2014) Maize centromeres expand and adopt a uniform size in the genetic background of oat. Genome Res 24:107–116

Weber B, Schmidt T (2009) Nested Ty3-gypsy retrotransposons of a single *Beta procumbens* centromere contain a putative chromodomain. Chromosom Res 17:379–396

Weiss-Schneeweiss H, Leitch AR, McCann J, Jang TS, Macas J (2015) Employing next generation sequencing to explore the repeat landscape of the plant genome. In: Hörandl E, Appelhans M (eds) Next generation sequencing in plant systematics regnum Vegetabile. Koeltz Scientific Books, Königstein, Germany, pp 155–179

Yan H, Kikuchi S, Neumann P, Zhang W, Wu Y, Chen F, Jiang J (2010) Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice. Plant J 63:353–365

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

Yang S, Chenga C, Qina X, Yua X, Loua Q, Lia J, Qian C, Chen J (2019) Comparative cytomolecular analysis of repetitive DNA provides insights into the differential genome structure and evolution of five *Cucumis* species. Hort Plant J 5:192–204

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: 903-911

Zakrzewski F, Weisshaar B, Fuchs J, Bannack E, Minoche AE, Dohm JC, Himmelbauer H, Schmidt T (2011) Epigenetic profiling of heterochromatic satellite DNA. Chromosoma 120:409–422

Zakrzewski F, Schubert V, Viehoever P, Minoche AE, Dohm JC, Himmelbauer H, Weisshaar B, Schmidt T (2014) The CHH motif in sugar beet satellite DNA: a modulator for cytosine methylation. Plant J 78:937–950

Zhang X, Li X, Marshall JB, Zhong CX, Dawe RK (2005) Phosphoserines on maize CENTROMERIC HISTONE H3 and histone H3 demarcate the centromere and pericentromere during chromosome segregation. Plant Cell 17:572–583

Zhang W, Lee HR, Koo DH, Jiang J (2008) Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the CENH3-associated chromatin in Arabidopsis thaliana and maize. Plant Cell 20:25–34

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

Zhang B, Lv Z, Pang J, Liu Y, Guo X, Fu S, Li J, Dong Q, Wu H-J, Gao Z, Wang X-J, Hana F (2013a) Formation of a functional maize centromere after loss of centromeric sequences and gain of ectopic sequences. Plant Cell 25:1979–1989

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

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

Zhang J, Fu XX, Li RQ et al (2020) The hornwort genome and early land plant evolution. Nat Plants 6:107–118

Zhao H, Zhu X, Wang K, Gent JI, Zhang WL, Dawe RK, Jiang JM (2016) Gene expression and chromatin modifications associated with maize centromeres. G3 Genes Genomes Genet 6:183–192

Zhao HN, Zeng ZX, Koo D-H, Gill BS, Birchler JA, Jiang JM (2017) Recurrent establishment of de novo centromeres in the pericentromeric region of maize chromosome 3. Chromosome Res 25:299–311

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

Zhu JM, Ellison NW, Rowland RE (1996) Chromosomal localization of a tandemly repeated DNA sequence in *Trifolium repens* L. Cell Res 6:39–46

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC (1980) Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc Natl Acad Sci USA 77:2158–2162

Chapter 6
Satellite DNA-Mediated Gene Expression Regulation: Physiological and Evolutionary Implication

Isidoro Feliciello, Željka Pezer, Antonio Sermek, Branka Bruvo Mađarić, Sven Ljubić, and Đurđica Ugarković

Abstract Satellite DNAs are tandemly repeated sequences organized in large clusters within (peri)centromeric and/or subtelomeric heterochromatin. However, in many species, satellite DNAs are not restricted to heterochromatin but are also dispersed as short arrays within euchromatin. Such genomic organization together with transcriptional activity seems to be a prerequisite for the gene-modulatory effect of satellite DNAs which was first demonstrated in the beetle *Tribolium castaneum* upon heat stress. Namely, enrichment of a silent histone mark at euchromatic repeats of a major beetle satellite DNA results in epigenetic silencing of neighboring genes. In addition, human satellite III transcripts induced by heat shock contribute to genome-wide gene silencing, providing protection against stress-induced cell death. Gene silencing mediated by satellite RNA was also shown to be fundamental for the early embryonic development of the mosquito *Aedes aegypti*. Apart from a physiological role during embryogenesis and heat stress response, activation of satellite DNAs in terms of transcription and proliferation can have an evolutionary impact. Spreading of satellite repeats throughout euchromatin promotes the variation of epigenetic landscapes and gene expression diversity, contributing to the evolution of gene regulatory networks and to genome adaptation in fluctuating environmental conditions.

Keywords Satellite DNA · Heterochromatin · Euchromatin · Embryogenesis · Heat stress · Gene expression

I. Feliciello
Department of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

Dipartimento di Medicina Clinica e Chirurgia, Universita' degli Studi di Napoli Federico II, Naples, Italy

Ž. Pezer · A. Sermek · B. Bruvo Mađarić · S. Ljubić · Đ. Ugarković (✉)
Department of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia
e-mail: ugarkov@irb.hr

Đ. Ugarković (ed.), *Satellite DNAs in Physiology and Evolution*, Progress in Molecular and Subcellular Biology 60,
https://doi.org/10.1007/978-3-030-74889-0_6

6.1 Introduction

Noncoding repetitive DNAs comprise a considerable portion of most eukaryotic genomes and their function has been studied in diverse model organisms. Among the most intensively investigated noncoding repetitive DNAs are mobile transposable elements (TE) which represent an important source of regulatory sequences (Faulkner et al. 2009; Kapusta et al. 2013). By mediating the distribution of regulatory elements throughout the genome transposons are known to influence the evolution of gene-regulatory networks (Chuong et al. 2017). Recent evidence suggests that TEs can also have potent "epigenetic" effects on the regulation of gene expression and genome evolution (Choi and Lee 2020). The functional significance of another abundant class of noncoding repetitive elements such as satellite DNA, regarding gene expression regulation was also proposed (Ugarković 2005; Pezer et al. 2012) and has been recently experimentally confirmed in different studies (Feliciello et al. 2015a, 2020a; Menon et al. 2014; Joshi and Meller 2017; Halbach et al. 2020). The aim of this review is to present recent findings on the role of satellite DNA in gene expression regulation/modulation and to explain the molecular mechanisms by which satellite DNA affects genes. We focus particularly on the impact of euchromatic satellite DNA repeats dispersed outside of (peri)centromeric regions on adjacent gene expression, as well as the role of satellite transcripts since their importance for gene expression modulation has been reported in different studies. A physiological role of satellite DNAs and their transcripts in the remodeling of global heterochromatin structure and in the modulation of gene expression during development, stress response, and pathological transformation is presented. We also discuss the implication of satellite DNA-mediated gene regulation in the evolution of gene-regulatory networks and on the processes of environmental adaptation.

6.2 Proliferation and Dispersion of Satellite DNA Within Euchromatin

Satellite DNAs are preferentially organized as tandemly repeated sequences assembled in large arrays within gene-poor constitutive heterochromatin in (peri)-centromeric and/or telomeric regions. Longer arrays of tandem satellite repeats within euchromatin are generally rare, probably due to the instability caused by intrastrand homologous recombination, although blocks of tandem repeats are found in euchromatin of *D. melanogaster* (Kuhn et al. 2012) and *Triatoma infestans (*Pita et al. 2017), while in the beetle *Tribolium castaneum* some euchromatic satellite DNAs arrays are even composed of higher-order repeats (Vlahović et al. 2017; Pavlek et al. 2015). Bioinformatic analyses of sequenced genomes however have revealed many single repeats or short arrays of satellite DNAs dispersed within euchromatin, in the vicinity of genes, in different insects such as *Tribolium castaneum*, *Drosophila melanogaster,* and *Locusta migratoria* (Ruiz-Ruano et al.

2016; Brajković et al. 2012, 2018; Kuhn et al. 2012). In mammals, single repeats of a major human alpha satellite DNA (Feliciello et al. 2020a, b) as well as of a major mouse satellite DNA (Bulut-Karslioglu et al. 2012) are also found interspersed among genes or within introns. It seems therefore that such mixed organization of satellite DNAs with a majority of the repeats clustered within pericentromeric constitutive heterochromatin combined with single repeats or short multimers dispersed within euchromatin is common for many species. Heterochromatic and euchromatic repeats of the same satellite show different evolutionary dynamics as revealed for *Tribolium castaneum* satellites, *Drosophila* 1.688, and human alpha satellite DNAs (Brajković et al. 2012, 2018; de Lima et al. 2020; Feliciello et al. 2020b), suggesting that chromatin domains may influence the evolution of these sequences. While heterochromatic satellite repeats display concerted evolution and a species-specific pattern, euchromatic repeats display high intra- and interspecific divergence. On the other hand, heterochromatic satellites coexisting in different species of the insect genus *Pimelia* evolve in parallel with fairly similar rates (Bruvo et al. 2003), indicating also the effect of chromatin state on satellite sequence evolution. Human euchromatic alpha satellite repeats have sequence characteristics of (peri)centromeric alpha repeats suggesting heterochromatin as their source but do not exhibit the concerted evolution pattern (Feliciello et al. 2020b). Alpha satellite repeats were continuously inserted within euchromatin throughout primate evolutionary history and stably transmitted to the descendant species, while their sequence divergence generally follows the primate species phylogeny (Feliciello et al. 2020b).

The pattern of dispersion of satellite DNA repeats within euchromatin can be very dynamic, differing significantly among related species as shown for *Drosophila* X chromosome euchromatin satellite DNAs (Sproul et al. 2020). This suggests that similar to transposable elements, euchromatic satellite repeats can be subjected to cycles of proliferation. Insertional polymorphism of euchromatic satellite repeats detected among populations of the same species or even among individuals within the same population suggests an ongoing movement of these elements within euchromatin and demonstrates the mutational potency of satellite DNAs (Feliciello et al. 2015a, b). Although a novel insertion of satellite repeat within euchromatin in many cases probably does not have an effect on genes, under particular conditions such as heat stress it can modulate the expression of nearby genes by a novel epigenetic mechanism (Feliciello et al. 2015a) which is described in the next sections. Some insertions however can affect proper gene function or even cause disease as demonstrated for the human beta satellite repeats inserted within the splice-acceptor site of a transmembrane serine protease gene which causes childhood-onset deafness (Scott et al. 2001). Mobilization of some transposons in somatic cells can also induce a pathogenic state, e.g., insertion of human L1 retroelement within the *APC* tumor suppressor gene initiates colorectal cancer (Scott et al. 2016).

The activation of repetitive elements such as transposons, in terms of transcription and transposition, was intensively studied and was shown to be stress-induced, particularly by heat shock (Ratner et al. 1992; Piacentini et al. 2014; Cavrak et al. 2014; Makarevitch et al. 2015; Ito et al. 2016). In addition, environmental stress is

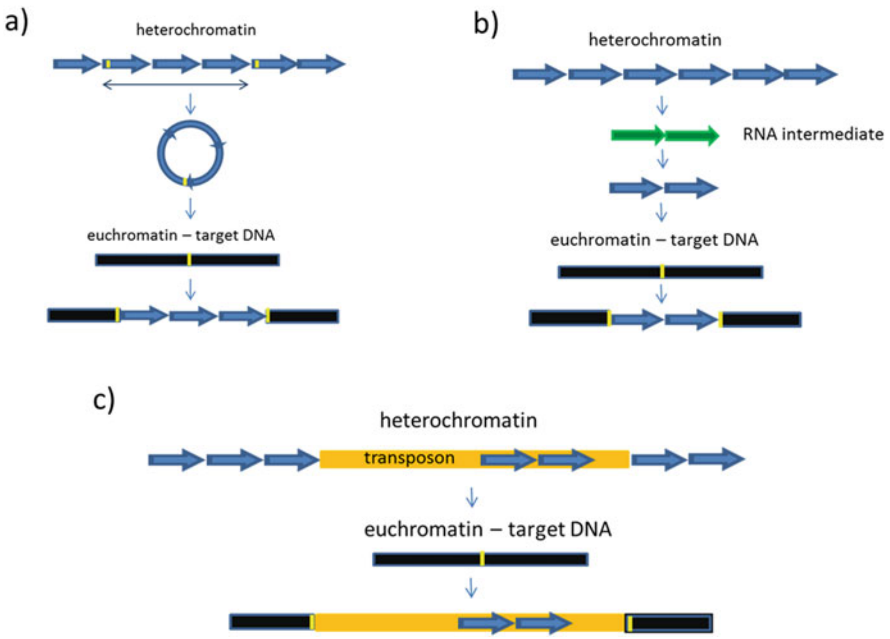


Fig. 6.1 Models of spreading of satellite DNA repeats in euchromatin. (**a**) Intra-chromatid recombination of satellite repeats within heterochromatin gives rise to extrachromosomal circular satellite DNAs. Short segments of homology, indicated in yellow, between circularized repeats and target regions in euchromatin are necessary for the insertion by site-specific homologous recombination. (**b**) Satellite transcripts can be reverse transcribed and by the activity of endonuclease/integrase cDNA is inserted within euchromatin, (**c**) satellite DNAs can be spread throughout the genome as an integral part of DNA transposons

responsible for a significant change in copy number of transposons, as shown in wild barley and in *Drosophila* (Kalendar et al. 2000; Kim et al. 2014). It has been proposed that heat stress induces modulation of heterochromatin structure which is accompanied by the rearrangement of repeats present therein, in particular of tandem satellite repeats, which are prone to homologous recombination (Fig. 6.1a; Brajković et al. 2012). Intra-chromatid recombination events can give rise to extrachromosomal circular satellite DNAs that are common for diverse eukaryotic organisms including insects, plants, and mammals (Cohen et al. 2006; Navratilova et al. 2008; Cohen and Segal 2009; Paulsen et al. 2018; Sproul et al. 2020). Extrachromosomal satellite DNA circles are proposed to be amplified by rolling circle replication and can be reintegrated within the genome by a random process of site-specific recombination which occurs between short sequence motifs within circularized satellite repeats and homologous motifs at different chromosomal sites, either within euchromatin or heterochromatin (Fig. 6.1a; Feliciello et al. 2006; Brajković et al. 2012). This process can lead to a relatively rapid change in a copy number of particular satellite DNA which can be detected at the population level (Wei et al. 2014; Feliciello et al. 2015b) or even at the individual level (Cardone et al. 1997). In addition, the same process of proliferation can spread satellite repeats to new loci and change the dispersion profiles of satellite DNA within euchromatin. These processes lead to an increase in the genetic variability among individuals within a population as well as between populations (Feliciello et al. 2015a, b).

Some satellite DNAs that are preferentially expressed in cancer such as human satellite II have the ability to reverse-transcribe in cancer cells and through RNA-derived DNA intermediates can expand locally and genome-wide (Bersani et al. 2015). This example of human satellite II shows that similar to retrotransposons, some satellite DNAs can proliferate through RNA intermediates and indicates coupling of satellite DNA transcription and proliferation (Fig. 6.1b). DNA transposons belonging to the *Helitron* superfamily have a propensity to capture and mobilize flanking DNA sequences (Thomas et al. 2014). Since some satellite repeats are found as integral parts of DNA transposons while some satellite arrays are flanked by *Helitron* transposons, it was proposed that the spread of satellite repeats throughout the genome can be linked to the process of transposition (Fig. 6.1c; Brajković et al. 2012; Satović et al. 2016; Vojvoda Zeljko et al. 2020). In the human genome, single repeats of a major alpha satellite DNA dispersed within euchromatin are often embedded within abundant retroelements such as *Alu*, L1, or ERVL-LRTs; however, there is no evidence for such elements playing a role in the spreading of alpha repeats throughout euchromatin (Feliciello et al. 2020b). While segmental duplication can be associated with the dispersion of some alpha repeats, the prevalent mechanism of spreading seems to be mediated by extrachromosomal circles of alpha satellite DNA whose insertion is facilitated by short sequence homology between alpha repeats and their target sequences (Feliciello et al. 2020b). Extrachromosomal circular DNAs (eccDNA) composed of alpha satellite repeats ranging in size from less than 2 kb to over 20 kb are detected in human cells (Cohen et al. 2010), revealing the propensity of tandemly arranged alpha repeats to generate eccDNA. The main mechanisms proposed to be responsible for the proliferation of satellite repeats and their dispersion within euchromatin are shown in Fig. 6.1.

6.3 Satellite DNA Transcription: Heat Stress Activation

Apart from a specific genomic organization of satellite DNA which is characterized by their partial dispersion within euchromatin, transcripts of satellite DNAs have also been proposed to have gene-regulatory potential (Ugarković 2005). Although satellite DNAs are preferentially embedded in constitutive heterochromatin which is considered transcriptionally inert, their transcription was reported in many species belonging to vertebrates, invertebrates, and plants (Ugarković 2005). Transcription of satellite DNAs is often bidirectional and proceeds usually by RNA polymerase II (RNAP II) from internal promoters as shown in mice (Lu and Gilbert 2007), humans (Bury et al. 2020) as well as in insects (Pezer and Ugarković 2008, 2009, 2012). The satellite transcripts fall into two main categories: long noncoding RNAs (>200 nt) and small RNAs (<200 nt; reviewed in Arunkumar and Melters 2020). Among small RNAs, the most represented are small interfering RNAs (siRNAs) which, through an RNA interference mechanism (RNAi) are involved in the epigenetic process of heterochromatin formation in fission yeast, insects as well as in plants and nematodes (Volpe et al. 2002; Pal-Bhadra et al. 2004; Grewal and Elgin 2007; Fagegaltier et al. 2009). In mammals, however, long satellite transcripts play a role in heterochromatin formation, maintenance, and regulation (Saksouk et al. 2015; Johnson et al. 2017). During mitosis, the level of mouse major and minor satellite RNA and of human alpha satellite RNA is regulated by Dicer-mediated cleavage (Fukagawa et al. 2004; Huang et al. 2015) while in meiosis the MIWI protein guided by PIWI-interacting RNAs (piRNAs) together with the endoribonuclease Dicer controls satellite RNA level (Hsieh et al. 2020). In diverse species, from plants, insects to mammals, centromeric satellite transcripts are involved in the recruitment and loading of centromere-specific histone H3 variant CENP-A as well as of CENP-B and CENP-C proteins, which are necessary for centromere organization, maintenance, and function (Bouzinba-Segard et al. 2006; Wong et al. 2007; Rosic et al. 2014; Arunkumar and Melters 2020; Chap. 7 of this book). Controlled expression of (peri)centromeric satellite RNAs, therefore, seems to be essential for ensuring proper kinetochore assembly and faithful chromosome segregation.

Constitutive heterochromatin is sensitive to temperature fluctuations and is dynamically regulated in response to environmental stimuli (Ayoub et al. 1999; Wang et al. 2016). Possible mechanism of temperature-mediated heterochromatin modulation includes stress-response transcription factors involved in heterochromatin assembly. In human cells, heat stress activates heat shock transcription factor 1 (HSF1) which recruits major cellular acetyltransferases to pericentric heterochromatin leading to targeted hyperacetylation (Col et al. 2017), facilitating particularly the transcription of satellite III DNA (Jolly et al. 2004; Rizzi et al. 2004) and satellite II (Tilman et al. 2012) but also, to a lower extent, transcription of a major alpha satellite DNA (Feliciello et al. 2020a). The human alpha satellite transcription seems to be controlled by centromere-nucleolar contacts and when the nucleolus is disrupted alpha satellite transcript levels increase substantially (Bury et al. 2020). The possible damage of nucleolus structure upon heat stress might therefore also influence the activation of alpha satellite transcription. Although human pericentromeric satellite DNAs such as alpha, satellites II and III are heavily methylated no change in methylation was detected upon heat stress (Eymery et al. 2009), confirming that transcription activation is not related to DNA methylation status. In *Drosophila*, under standard conditions, transcription factor dATF-2 which regulates expression of stress response genes recruits heterochromatin protein 1 (HP1) to pericentromeric heterochromatin regions that contain dATF-2 binding sites. Under stress conditions activated MAP kinase such as p38 phosphorylates dATF-2 which is released from heterochromatin, leading to the abolishment of HP1 and disruption of heterochromatin (Seong et al. 2011). In vivo studies on insect *T. castaneum* revealed heat-stress induced transcription of a major satellite DNA TCAST1 located within pericentromeric heterochromatic and in centromeric regions, followed by the processing of long satellite transcripts into siRNAs (Pezer and Ugarković 2012; Sermek et al. 2021). Induced satellite DNA transcription is coupled with the almost complete demethylation of satellite DNA suggesting a possible role of DNA methylation in the control of satellite DNA transcription activation upon heat stress (Feliciello et al. 2013). In *Arabidopsis* specific transcription factors HIT4, MED14, and UVH6 are required for transcriptional activation of heterochromatic DNA. Transposons in particular respond to heat stress and this process is accompanied by heterochromatin decondensation and 3D genome reorganization (Bourguet et al. 2018; Wang et al. 2013; Sun et al. 2020).

6.4 Satellite RNA and Euchromatic Satellite Repeats in Gene Expression Regulation

Since expression of heterochromatic satellite DNAs is induced upon heat stress in different model organisms it was investigated whether this could be linked to modulation of expression of genes located in the vicinity of satellite repeats. In the beetle *T. castaneum* enhanced heat stress-induced transcription of a major TCAST1 satellite DNA correlates with an increased level of repressive heterochromatin marks H3K9me2/3 on satellite repeats in constitutive heterochromatin as well as on dispersed TCAST1 satellite elements within euchromatin and their proximal regions up to 6 kb from the insertion site (Feliciello et al. 2015a). TCAST1 satellite DNA repeats dispersed within euchromatin, therefore, seem to serve as nucleation sites for transient heterochromatin formation which results in partial suppression of nearby genes upon heat stress, representing the first experimental proof for the gene-modulatory role of a satellite DNA (Feliciello et al. 2015a). In addition, the role of TCAST1-derived siRNAs in transient H3K9me2/3 enrichment at euchromatic and heterochromatic TCAST1 repeats upon heat stress is proposed (Fig. 6.2a). This proposal is consistent with the fact that small RNAs initiate the epigenetic silencing of repetitive DNAs such as satellite DNAs or transposons (TE), and the strength of these epigenetic effects was shown to be positively correlated with the amount of small RNAs targeting some TE families (Lee 2015; Lee and Karpen 2017; Choi and Lee 2020). Since this novel mode of gene expression regulation does not seem to be unique to a specific satellite DNA it is hypothesized that different satellites which are partially dispersed in the vicinity of genes and whose transcription is induced upon heat stress, could influence the expression of associated genes by the same mechanism of temporary "heterochromatinization." Furthermore, in plants, the strength of epigenetic silencing of a TE family positively correlates with the family copy number (Cheng et al. 2006; Noreen et al. 2007), while in *Drosophila*, the proportion of TEs with *cis*-spreading of repressive marks also increases with family copy number (Lee and Karpen 2017). In addition, it could be proposed that the copy number of satellite DNAs which would be related to the level of satellite transcripts might also influence the strength of their epigenetic effects. Apart from copy number, the influence of chromatin state on the expression of transposon-derived

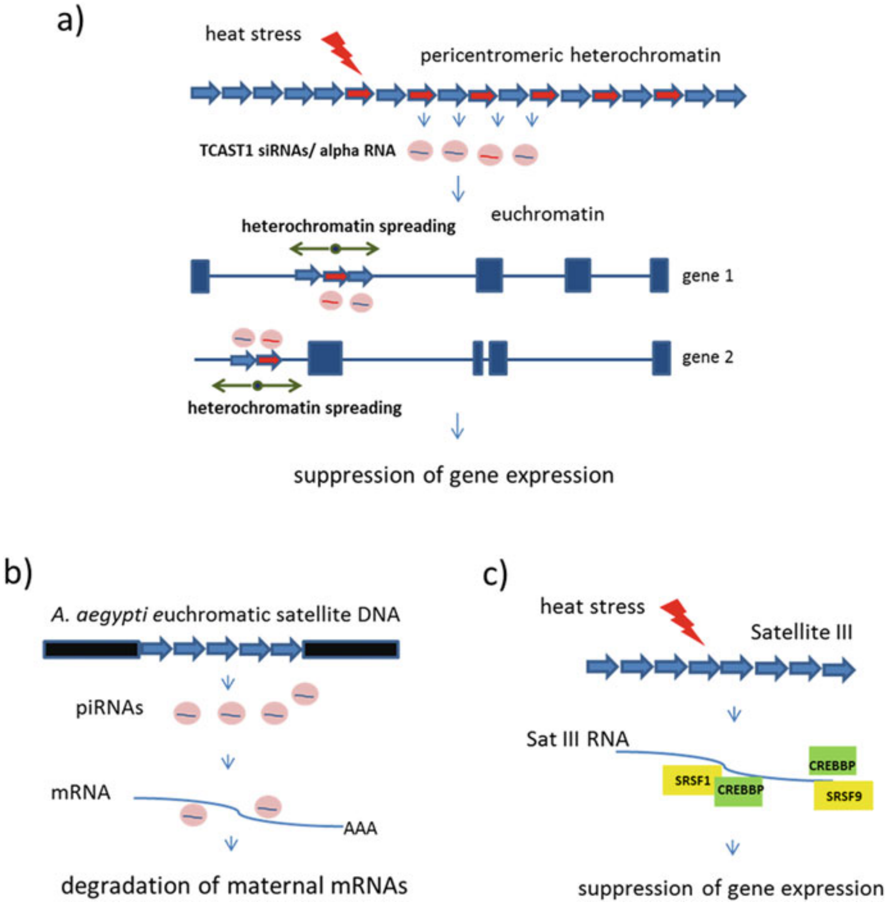

a)
heat stress
pericentromeric heterochromatin
TCAST1 siRNAs/ alpha RNA
heterochromatin spreading
euchromatin
gene 1
gene 2
heterochromatin spreading
suppression of gene expression
b)
A. aegypti euchromatic satellite DNA
piRNAs
mRNA
AAA
degradation of maternal mRNAs
c)
heat stress
Satellite III
Sat III RNA
SRSF1
CREBBP
CREBBP
SRSF9
suppression of gene expression

Fig. 6.2 Mechanisms of satellite DNA-mediated gene expression regulation. (**a**) Heat stress promotes transcription of abundant pericentromeric satellite DNAs: TCAST1 in beetle *T. castaneum* and alpha satellite DNA in human cells. This is accompanied by increased H3K9me3 levels at euchromatic TCAST1 and alpha satellite repeats, respectively, resulting in partial suppression of nearby genes (Feliciello et al. 2015a, 2020a). Genes associated with satellite repeats are schematically shown: exons are represented by rectangles, satellite elements by arrows, and complex containing satellite RNAs by a circle. (**b**) In the mosquito *Aedes aegypti* satellite repeats located at a single euchromatic locus promote sequence-specific gene silencing in *trans* via the expression of piRNAs which participate in the degradation of maternally inherited transcripts during early embryonal development (Halbach et al. 2020). (**c**) The transcription of human Satellite III (sat III) loci is induced upon heat stress and satellite 3 transcripts sequester transcription factor CREBBP and splicing regulatory proteins SRSFs. As a consequence, there is a suppression of gene expression (Goenka et al. 2016; Ninomiya et al. 2020)

small RNAs during embryogenesis was reported in plants (Papareddy et al. 2020). Chromatin organization is also proposed to be responsible for distinct transcription regulation of satellite DNAs in the beetle *T. castaneum* during embryogenesis and heat stress (Sermek et al. 2021). Namely, transcription of a major TCAST1 satellite DNA which proceeds from heterochromatic loci is specifically induced during these processes. In contrast, the transcription of a minor TCAST2 satellite which proceeds predominantly from euchromatic clusters remains unchanged. Consequently, the levels of the silent histone mark H3K9me3 at minor TCAST2 repeats as well as the expression of nearby genes are not influenced by heat stress (Sermek et al. 2021).

Recently it was revealed that repeats of a major human alpha satellite DNA located both in heterochromatin and euchromatin have increased H3K9me3 levels upon heat stress (Feliciello et al. 2020a). H3K9me3 enrichment at alpha repeats upon heat stress correlates with the dynamics of alpha satellite DNA transcription activation while spreading of H3K9me3 up to 1–2 kb from the insertion sites reveals that euchromatic alpha repeats act as modulators of local chromatin structure. Aside from satellite DNAs, some transposons in plants (Eichten et al. 2012) and mammals (Rebollo et al. 2011; Liu et al. 2018) reduce expression of neighboring genes by spreading heterochromatin marks, DNA methylation, and/or H3K9me2/3 from the insertion sites. A widespread influence of transposons on H3K9me3 spreading and expression of neighboring genes was also observed in *Drosophila* (Sienski et al. 2012; Lee and Karpen 2017). All these results suggest that epigenetic effects, in particular H3K9me3 enrichment mediated by siRNAs and piRNAs, respectively, are common for some satellite DNAs and transposons, becoming pronounced upon stress and may affect neighboring gene expression.

While in the beetle *T. castaneum* and in human cells the major satellites' repeats dispersed within euchromatin modulate the local chromatin environment *in cis* inducing neighboring gene silencing, in the mosquito *Aedes aegypti* evolutionary old and conserved satellite repeats located at a single euchromatic locus promote sequence-specific gene silencing *in trans* via the expression of abundant PIWI-interacting RNAs (piRNAs). The satellite-derived piRNAs participate in the degradation of maternally inherited transcripts during the maternal-to-zygotic transition and are fundamental to early embryonic development (Halbach et al. 2020; Fig. 6.2b). Satellite DNA-derived siRNAs also play a specific role in gene expression regulation in *Drosophila*. Namely, short, tandem clusters of 1.688 satellite DNA in the X chromosome euchromatin of *D. melanogaster* males guide the dosage compensation complex MLS which increases expression of nearby genes and the 1.688 siRNAs play a role in this process (Menon et al. 2014; Joshi and Meller 2017, Chap. 1 of this book). The short euchromatic array of 1.688 satellite on the X chromosome is also shown to promote specific targeting of POF protein which is involved in the global regulation of genes on *D. melanogaster* chromosome 4 (Kim et al. 2018), while depletion of a large block of pericentromeric 1.688 satellite seems to affect eggshell formation (Ekhteraei-Tousi et al. 2020). Human alpha satellite DNA repeats in addition to primates have been detected as rare, highly conserved elements in evolutionarily distant species such as chicken and zebrafish (Li and Kirby 2003). The presence of several coding mRNAs in human and chick embryos that contain alpha-like satellite repeats as a part of their 5′ or 3′ untranslated regions indicates that their expression could be controlled *in trans* by alpha satellite RNA (Li and Kirby 2003).

Some satellite DNA repeats are located within introns of particular genes affecting their expression under specific conditions or developmental stages. One such example is the tandem repeats found within the intron of the major histocompatibility complex gene (*MHIIβ*) in the fish *Salvelinus fontinalis* which are involved in temperature-dependant modulation of expression of this gene (Croisetière et al. 2010). The minisatellite was proposed to play a role in the regulation of the adaptive immune response but the molecular mechanism behind its gene-modulatory effect was not investigated. In plants such as *Arabidopsis thaliana* and particularly in rice, introns of many genes contain heterochromatin associated with repetitive elements, mostly transposons (Duan et al. 2017; Espinas et al. 2020). The establishment and maintenance of heterochromatin within introns seem to be critical for transcriptional control of the associated genes which are predominantly required for environmental responses and development (Le et al. 2015; Khan et al. 2013). The transcription of genes with intronic heterochromatin is regulated by an epigenetic mechanism that involves the conserved nuclear protein complex, mutation of which results in severe developmental defects (Duan et al. 2017; Espinas et al. 2020). Introns containing long arrays of satellite DNAs are characteristic for *Drosophila* Y chromosome genes which are solely expressed during spermatogenesis (Hardy et al. 1981). The gigantic introns of these genes are transcribed in line with their exons; however, their expression requires a unique gene expression program, which acts on both transcription and posttranscriptional processing (Fingerhut et al. 2019). It is proposed that satellite DNA-containing gigantic introns could act in a manner similar to enhancers, recruiting transcriptional machinery to the Y-loop genes, while intron size can play a critical role in the regulation of gene expression (Shaul 2017; Fingerhut et al. 2019).

6.5 "Macroheterochromatin" in Gene Expression Regulation

A "micro-heterochromatin" is formed on some satellite repeats or short arrays dispersed within euchromatin and can affect the expression of genes located in the vicinity (Feliciello et al. 2015a, b, 2020a). On the other hand, a "macro-heterochromatin" is composed of megabase stretches of satellite DNA such as those on the *Drosophila* Y chromosome and polymorphism in heterochromatic Y chromosomes results in genome-wide gene expression variation (Lemos et al. 2010). It seems that Y chromosome heterochromatin serves as a source of epigenetic variation in natural populations that interacts with chromatin components to modulate the expression of biologically relevant phenotypic variation. Increasing the amount of repeats on the X or Y *D. melanogaster* chromosome results in a decrease of H3K9me2/3 levels at repeat-rich regions at pericentromeres and the Y chromosome, implying a role for satellite DNA in global chromatin dynamics and redistribution of chromatin regulators across the genome (Brown et al. 2020). Since satellite DNAs are characterized by a rapid copy number change often observed at the intraspecific level (Cardone et al. 1997; Wei et al. 2014; Feliciello et al. 2015b), a significant difference in their amount may contribute to the diversity of expression of genes and repetitive elements among populations and individuals. Analyses of 3D genome structures reveal that pericentromeric heterochromatin spatially contacts distant euchromatin regions enriched for repressive epigenetic marks, such as regions associated with epigenetically silenced transposable elements or other repeats, as shown in *D. melanogaster* (Lee et al. 2020). It can be proposed that due to such interactions, pericentromeric heterochromatin could impact the expression of distant euchromatic genes which are associated with "H3K9me2/3 islands." This also indicates that an interplay between satellite DNA repeats located within heterochromatin and euchromatin might be involved in genome-wide gene expression regulation.

6.6 Satellite DNA Role in Stress Response and Environmental Adaptation

Numerous in vitro studies on human cell lines have shown a strong increase of pericentromeric satellite III expression induced by a large number of stressing agents including heat shock, DNA damaging agents, and hyperosmotic stress (reviewed in Vourc'h and Biamonti 2011). While most of these stressing agents act through heat shock transcription factor 1 (HSF1), transcription of satellite III in response to hyperosmotic stress depends on Tonicity Enhancer Binding Protein (TonEBP) which controls genes responsible for the survival of cells subjected to high osmotic pressure (Valgardsdottir et al. 2008). It was proposed that stress-induced activation of satellite III occurred through at least two independent pathways which both lead to the formation of nuclear stress bodies, and is considered to be a part of a general cellular response to stress. Namely, satellite III transcripts recruit critical factors involved in the transcriptional process, contributing to heat-induced transcriptional silencing and seem to be required to provide protection against heat-shock-induced cell death (Goenka et al. 2016; Fig. 6.2c). Satellite III RNA also mediates in the recruitment of a number of RNA binding proteins involved in pre-mRNA processing and participates in the control of gene expression upon heat stress at the level of splicing regulation (Ninomiya et al. 2020). The alteration of the splicing profile is mainly characterized by an increase in intron retention events during the recovery from heat shock. Intron retention prevents the export of the pre-mRNAs from the nucleus resulting in suppression of gene expression at the posttranscriptional level. Expression of centromeric satellites is also strongly induced by genotoxic stress as shown for mouse minor satellite DNA and their accumulation under stress conditions seems to be a conserved feature of the cellular stress response (Hédouin et al. 2017).

In the insect *Tribolium castaneum* and in human cells activation of (peri)-centromeric satellite DNA transcription during heat stress response reinforces "heterochromatinization" and helps heterochromatin recovery (Pezer and Ugarković 2012; Feliciello et al. 2020a). Since heterochromatin is important for genome stability and integrity, satellite DNA transcripts might have a protective effect in stressed cells/organisms. In addition, induced "heterochromatinization" leads to transient suppression of genes located in the vicinity of dispersed TCAST1 satellite elements (Feliciello et al. 2015a), as described previously in this chapter. However, what is the physiological consequence of such transient gene suppression upon heat stress? It is known that after strong heat stress genomes undergo a substantial transcriptional silencing and the role of human satellite III in this process was demonstrated (Goenka et al. 2016). It could be hypothesized that other satellite DNAs contribute to the same process of gene repression which is necessary to protect the cell from stress-induced damage. While human satellite III RNA affects gene expression genome-wide, in the case of TCAST1 satellite expression of genes located in the vicinity of euchromatic TCAST1 repeats is affected. Within genes associated with euchromatic TCAST1 satellite repeats, there is a significant over-representation of immunoglobulin-like genes (Brajković et al. 2012). Stress and the immune response are tightly connected in insects and mild physical or thermal stress leads to short-term immune memory (Altincicek et al. 2009; Freitak et al. 2012; Marshall and Sinclair 2012; Eggert et al. 2015). In mammals, genes involved in immunity and stress are more likely to contain transposon sequences within UTRs than other genes (van de Lagemaat et al. 2003). In plants, genes required for environmental response and development are enriched with heterochromatic introns associated with repetitive elements (Duan et al. 2017; Espinas et al. 2020). These data indicate that repetitive elements, either transposons or satellite repeats, seem to be preferentially associated with environment susceptible genes such as stress or immune response genes and might affect their expression under specific conditions. In addition, the high evolutionary dynamics of repetitive elements can promote expression variation and the evolution of associated genes. Differential transcription activation of satellite DNA families by heat stress and clustering of their repeats near some genes, as observed in *T. castaneum* (Brajković et al. 2018), may facilitate satellite-mediated gene modulatory effects and increase the complexity of the transcriptional response to the environment. Satellite DNA-induced changes of the transcriptome might create a modified gene interaction network with a strong adaptive potential on which natural selection can act (Fig. 6.3). In ectothermic organisms in particular, whose body temperatures conform to ambient temperature, the temperature is one of the principal environmental variables that drive adaptive evolution. It is also important to mention that satellite DNAs are subjected to a high evolutionary turnover, resulting in a rapid change of their copy number (Meštrović et al. 1998) as well as in the emergence of new satellites which could sometimes contribute to the evolution of a novel feature (Ugarković and Plohl 2002). In the New World Monkey genus *Aotus* the newly amplified satellite DNA builds a centrally located heterochromatin block in the nucleus of the rod cells which is responsible for the evolution of night vision characteristic for species of this genus (Koga et al. 2017). This represents an example of how a newly acquired satellite

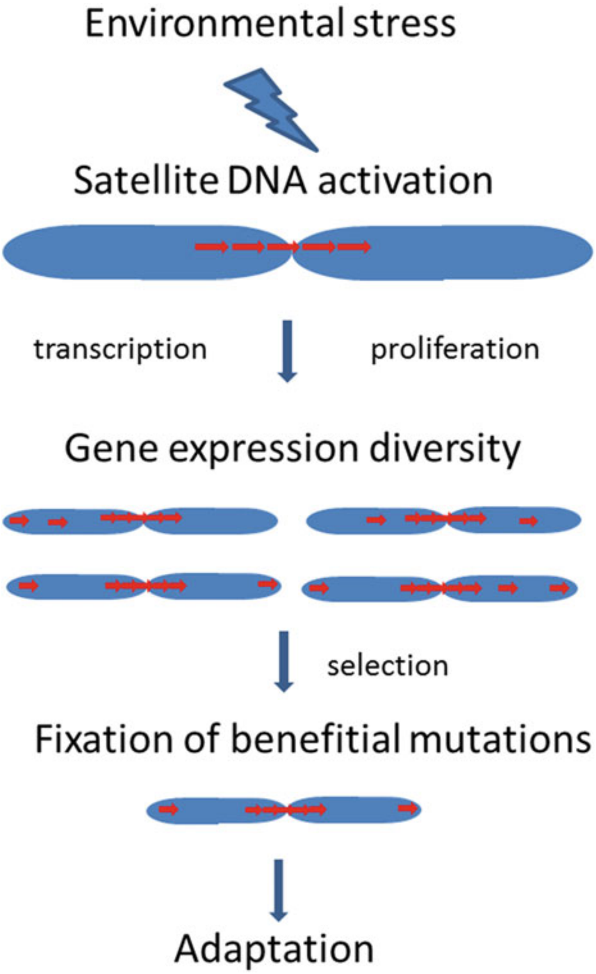


Fig. 6.3 Model explaining the role of satellite DNAs in the adaptation to environmental conditions. Heat stress induces activation of satellite DNAs in terms of transcription and proliferation which can lead to the insertion polymorphism of satellite repeats within euchromatin. This may cause gene expression diversity among individuals upon heat stress, on which natural selection can act, fixing the beneficial mutations and contributing to the adaptation process

DNA contributes to the adaptation of its host organism to exploit an ecological niche.

6.7 Satellite DNA in Pathological Transformation and Development

Satellite DNA transcription is activated not only by environmental stress but also upon pathological conditions. In epithelial cancers increased satellite DNA transcription is observed (Ting et al. 2011) and it is often associated with a deficiency of tumor suppressor proteins, in particular p53 which restrains the movement of repetitive elements (Wylie et al. 2016). Besides p53, deficiency of the tumor suppressor BRCA1 impairs the integrity of constitutive heterochromatin and induces abnormal transcription of satellite DNA repeats (Zhu et al. 2011). Overexpressed heterochromatic satellite RNAs sequester BRCA1-associated proteins causing destabilization of DNA replication forks, and promote breast cancer formation in mice (Zhu et al. 2018). In mouse K-ras-mutated pancreatic precancerous tissues, transcripts of a major pericentromeric satellite DNA inhibit the DNA-damage repair function of YBX1 protein and accelerate tumor formation, and so act as "intrinsic mutagens" (Kishikawa et al. 2016, 2018). Human satellite II transcripts which are preferentially expressed in cancer cells are immunogenic, able to directly activate the innate immune system to produce cytokines, modulating in this way the immune response against tumor cells (Tanne et al. 2015). In addition, demethylated human satellite II and its transcripts act as molecular sponges and sequester chromatin regulatory proteins into abnormal nuclear bodies in cancer (Hall et al. 2017). Expression of human satellite II is also strongly induced in herpesvirus infected cells by viral proteins, while satellite II transcripts modulate viral protein expression and release of infectious particles, having functionally important consequences for viral replication (Nogalski et al. 2019). Hypomethylation of pericentromeric sequences and subsequent derepression of associated satellite transcripts triggers an interferon response in zebrafish (Rajshekar et al. 2018).

Apart from stress and pathological states, activation of satellite DNA transcription in many organisms is associated with cell cycle progression, development, and differentiation (Probst et al. 2010; Kishi et al. 2012; Park et al. 2018; Ferreira et al. 2020). The bidirectional transcription of murine minor satellite DNA occurs during mitosis and transcripts stabilize the overall kinetochore structure in the G2/M phase (Ferri et al. 2009), while in meiosis transcription mostly occurs during the early-pachytene stage (Hecht 1986). During early mouse embryogenesis, the major pericentromeric satellite RNA modulates the activity of histone methyltransferase SUV39H2 and reduces H3K9me3 levels in zygotes (Burton et al. 2020), while during the midblastula stage a burst of strand-specific transcription of a major pericentromeric satellite DNA is essential for heterochromatin formation and early development progression (Probst et al. 2010; Casanova et al. 2013). In addition, the same satellite is differentially expressed in cells of the developing central nervous system (Rudert et al. 1995) and this satellite RNA whose level is significantly increased during neuronal differentiation (Kishi et al. 2012) is necessary for the correct higher-order organization of pericentromeric heterochromatin (Fioriniello et al. 2020). It is interesting that although the transcription of the major satellite proceeds from both DNA strands only the satellite forward RNA is involved in the initial heterochromatin formation during embryogenesis (Maison et al. 2011) and in pericentromeric heterochromation organization in neurons (Fioriniello et al. 2020). Major and minor mouse satellite RNAs are also involved in the large-scale reorganization of constitutive heterochromatin during muscle differentiation (Park et al. 2018). In chicken and zebrafish, transcription of alphoid repeat sequences displays a specific temporal and spatial expression pattern during embryogenesis (Li and Kirby 2003). In insects, transcription of satellite DNAs is also developmentally regulated being increased during specific stages of embryogenesis as revealed for the major TCAST1 satellite DNA of *T. castaneum*, and transcripts in the form of TCAST1 piRNAs and siRNAs are proposed to be necessary for initial heterochromatin formation (Pezer and Ugarković 2012; Sermek et al. 2021). In the mosquito *Aedes aegypti*, piRNAs which derive from conserved euchromatic satellite DNA are necessary for embryonic development (Halbach et al. 2020), while RNA from a simple satellite DNA of *D. melanogaster* is required for sperm maturation and male fertility (Mills et al. 2019). Examples in other species also indicate that transcription activation of satellite DNAs as well as of some other repetitive families such as LINE1 might be a part of normal developmental and differentiation processes. While pericentromeric satellite DNA transcripts are important for regulation of heterochromatin establishment during early mouse embryogenesis and for heterochromatin remodeling during differentiation (Park et al. 2018; Burton et al. 2020), LINE-1 transcripts are relevant for regulation of global chromatin dynamics: de- and recondensation (Jachowicz et al. 2017), acting as a nuclear scaffold to direct gene expression programs essential for embryo development (Percharde et al. 2018). Transcripts of some satellite DNAs such as FA-SAT DNA are proposed to be important for cell proliferation (Ferreira et al. 2020). FA-SAT DNA is highly conserved in mammals being primarily located at the telomeres and FA-SAT RNA forms a nuclear complex with Pyruvate Kinase Muscle Isozyme protein (PKM2) which seems to participate in cell-cycle progression.

In conclusion, satellite DNA transcripts activated either by environmental stress or during pathological transformation and viral infection, are implicated in immune system modulation and stress responsiveness, although they act through different molecular pathways and mechanisms. On the one hand, satellite transcripts can promote oncogenic processes by inducing mutations (Kishikawa et al. 2016), affecting epigenetic regulators (Hall et al. 2017), enhancing tumor cell proliferation (Nogalski and Shenk 2020), or compromising replication fork stability and genome integrity (Zeller and Gasser 2017; Zhu et al. 2018). Satellite RNAs also provide protection against heat-shock-induced cell death (Goenka et al. 2016) and are a prerequisite for early embryonic development (Probst et al. 2010; Halbach et al. 2020) and cell differentiation (Park et al. 2018). On the other hand, besides playing a physiological role in the modulation of global heterochromatin structure as well as of the gene expression program during development and stress response, activation of satellite DNAs in terms of transcription and proliferation (mobilization) has an evolutionary impact. It generates spreading and insertion polymorphism of euchromatic satellite repeats, causing variation of epigenetic landscapes and gene expression diversity within species (Feliciello et al. 2015a; Fig. 6.3). This variation is additionally enhanced by the propensity of satellites to change copy numbers and to form new satellite families. This gene expression diversity contributes to the evolution of gene regulatory networks, increases the evolvability of species, and could represent a powerful adaptive response of the genome to changing environmental conditions.

Acknowledgments This work was supported by Croatian Science Foundation grants IP-2014-09-3733 and IP-2019-04-6915, by donation of "Adris" Foundation to Đ. Ugarković, and by the Italian Ministry of Education, University and Research (MIUR), fund for Investments on Basic Research (FIRB) and the International Staff Mobility Program of University of Naples Federico II to I. Feliciello. We thank Mary Sopta for the critical reading of the manuscript.

References

Altincicek B, Knorr E, Vilcinskas A (2009) Beetle immunity: identification of immune-inducible genes from the model insect *Tribolium castaneum*. Dev Comp Immunol 32:585–595

Arunkumar G, Melters DP (2020) Centromeric transcription: a conserved Swiss-Army knife. Genes (Basel) 11:E911

Ayoub N, Goldshmidt I, Cohen A (1999) Position effect variegation at the mating-type locus of fission yeast: a cis-acting element inhibits covariegated expression of genes in the silent and expressed domains. Genetics 152(2):495–508

Bersani F, Lee E, Kharchenko PV, Xu AW, Liu M, Xega K, MacKenzie OC, Brannigan BW, Wittner BS, Jung H, Ramaswamy S, Park PJ, Maheswaran S, Ting DT, Haber DA (2015) Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. Proc Natl Acad Sci USA 112:15148–15153

Bourguet P, de Bossoreille S, López-González Pouch-Pélissier MN, Gómez-Zambrano Á, Devert A, Pélissier T, Pogorelcnik R, Vaillant I, Mathieu O (2018) A role for MED14 and UVH6 in heterochromatin transcription upon destabilization of silencing. Life Sci Alliance 1: e201800197

Bouzinba-Segard H, Guais A, Francastel C (2006) Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. Proc Natl Acad Sci USA 103:8709–8714

Brajković J, Feliciello I, Bruvo-Mađarić B, Ugarković Đ (2012) Satellite DNA-like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. G3 (Bethesda) 2:931–941

Brajković J, Pezer Ž, Bruvo-Mađarić B, Sermek A, Feliciello I, Ugarković Đ (2018) Dispersion profiles and gene associations of repetitive DNAs in the euchromatin of the beetle *Tribolium castaneum*. G3 (Bethesda) 8:875–886

Brown EJ, Nguyen AH, Bachtrog D (2020) The Drosophila Y chromosome affects heterochromatin integrity genome-wide. Mol Biol Evol 37:2808–2824

Bruvo B, Pons J, Ugarković D, Juan C, Petitpierre E, Plohl M (2003) Evolution of low-copy number and major satellite DNA sequences coexisting in two Pimelia species-groups (Coleoptera). Gene 17(312):85–94

Bulut-Karslioglu A, Perrera V, Scaranaro M, de la Rosa-Velazquez IA, van de Nobelen S, Shukeir N, Popow J, Gerle B, Opravil S, Pagani M, Meidhof S, Brabletz T, Manke T, Lachner M, Jenuwein T (2012) A transcription factor-based mechanism for mouse heterochromatin formation. Nat Struct Mol Biol 19:1023–1030

Burton A, Brochard V, Galan C, Ruiz-Morales ER, Rovira Q, Rodriguez-Terrones D, Kruse K, Le Gras S, Udayakumar VS, Chin HG, Eid A, Liu X, Wang C, Gao S, Pradhan S, Vaquerizas JM, Beaujean N, Jenuwein T, Torres-Padilla ME (2020) Heterochromatin establishment during early mammalian development is regulated by pericentromeric RNA and characterized by non-repressive H3K9me3. Nat Cell Biol 22:767–778

Bury L, Moodie B, Ly J, McKay LS, Miga KH, Cheeseman IM (2020) Alpha-satellite RNA transcripts are repressed by centromere-nucleolus associations. eLife 9:e59770

Cardone DE, Feliciello I, Marotta M, Rosati C, Chinali G (1997) A family of centromeric satellite DNAs from the European brown frog *Rana graeca italica*. Genome 40:774–781

Casanova M, Pasternak M, El Marjou F, Le Baccon P, Probst AV, Almouzni G (2013) Heterochromatin reorganization during early mouse development requires a single-stranded noncoding transcript. Cell Rep 4:1156–1167

Cavrak VV, Lettner N, Jamge S, Kosarewicz A, Bayer LM, Mittelsten Scheid O (2014) How a retrotransposon exploits the Plant's heat stress response for its activation. PLoS Genet 10: e1004115

Cheng C, Daigen M, Hirochika H (2006) Epigenetic regulation of the rice retrotransposon Tos17. Mol Gen Genomics 276:378–390

Choi JY, Lee YCG (2020) Double-edged sword: the evolutionary consequences of the epigenetic silencing of transposable elements. PLoS Genet 16:e1008872

Chuong EB, Elde NC, Feschotte C (2017) Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18:71–86

Cohen S, Segal D (2009) Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. Cytogenet Genome Res 124:327–338

Cohen Z, Bacharach E, Lavi S (2006) Mouse major satellite DNA is prone to eccDNA formation via DNA ligase IV-dependent pathway. Oncogene 25:4515–4524

Cohen S, Agmon N, Sobol O, Segal D (2010) Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. Mob DNA 1:11

Col E, Hoghoughi N, Dufour S, Penin J, Koskas S, Faure V, Ouzounova M, Hernandez-Vargash H, Reynoird N, Daujat S, Folco E, Vigneron M, Schneider R, Verdel A, Khochbin S, Herceg Z, Caron C, Vourc'h C (2017) Bromodomain factors of BET family are new essential actors of pericentric heterochromatin transcriptional activation in response to heat shock. Sci Rep 7:5418

Croisetière S, Bernatchez L, Belhumeur P (2010) Temperature and length-dependent modulation of the MH class II beta gene expression in brook charr (*Salvelinus fontinalis*) by a cis-acting minisatellite. Mol Immunol 47:1817–1829

de Lima LG, Hanlon SL, Gerton JL (2020) Origins and evolutionary patterns of the 1.688 satellite DNA family in Drosophila phylogeny. G3 (Bethesda) 10:4129–4146

Duan CG, Wang X, Zhang L, Xiong X, Zhang Z, Tang K, Pan L, Hsu CC, Xu H, Tao WA, Zhang H, Zhu JK (2017) A protein complex regulates RNA processing of intronic heterochromatin-containing genes in Arabidopsis. Proc Natl Acad Sci USA 114:E7377–E7E84

Eggert H, Diddens-de Buhr MF, Kurtz J (2015) A temperature shock can lead to trans-generational immune priming in the red flour beetle, *Tribolium castaneum*. Ecol Evol 5:1318–1326

Eichten SR, Ellis NA, Makarevitch I, Yeh CT, Gent JI, Guo L, McGinnis KM, Zhang X, Schnable PS, Vaughn MW, Dawe RK, Springer NM (2012) Spreading of heterochromatin is limited to specific families of maize retrotransposons. PLoS Genet 8:e1003127

Ekhteraei-Tousi S, Lewerentz J, Larsson J (2020) Painting of fourth and the X-linked 1.688 satellite in *D. melanogaster* is involved in chromosome-wide gene regulation. Cells 9:323

Espinas NA, Tu LN, Furci L, Shimajiri Y, Harukawa Y, Miura S, Takuno S, Saze H (2020) Transcriptional regulation of genes bearing intronic heterochromatin in the rice genome. PLoS Genet 16:e1008637

Eymery A, Horard B, El Atifi-Borel M, Fourel G, Berger F, Vitte AL, Van den Broeck A, Brambilla E, Fournier A, Callanan M, Gazzeri S, Khochbin S, Rousseaux S, Gilson E, Vourc'h C (2009) A transcriptomic analysis of human centromeric and pericentric sequences in normal and tumor cells. Nucleic Acids Res 37:6340–6354

Fagegaltier D, Bougé AL, Berry B, Poisot E, Sismeiro O, Coppée JY, Théodore L, Voinnet O, Antoniewski C (2009) The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. Proc Natl Acad Sci USA 106:21258–21263

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest AR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41:563–571

Feliciello I, Picariell O, Chinali G (2006) Intra-specific variability and unusual organization of the repetitive units in a satellite DNA from *Rana dalmatina*: molecular evidence of a new mechanism of DNA repair acting on satellite DNA. Gene 383:81–92

Feliciello I, Parazajder J, Akrap I, Ugarković Đ (2013) First evidence of DNA methylation in insect *Tribolium castaneum* - environmental regulation of DNA methylation within heterochromatin. Epigenetics 8:534–541

Feliciello I, Akrap I, Ugarković Đ (2015a) Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. PLoS Genet 11:e1005466

Feliciello I, Akrap I, Brajković J, Zlatar I, Ugarković Đ (2015b) Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. Genome Biol Evol 7:228–239

Feliciello I, Sermek A, Pezer Ž, Matulić M, Ugarković Đ (2020a) Heat stress affects H3K9me3 level at human alpha satellite DNA repeats. Genes 11:663

Feliciello I, Pezer Z, Kordiš D, Bruvo-Mađarić B, Ugarković Đ (2020b) Evolutionary history of alpha satellite DNA repeats dispersed within human genome euchromatin. Genome Biol Evol 12:2125–2138

Ferreira D, Escudeiro A, Adega F, Anjo SI, Manadas B, Chaves R (2020) FA-SAT ncRNA interacts with PKM2 protein: depletion of this complex induces a switch from cell proliferation to apoptosis. Cell Mol Life Sci 77:1371–1386

Ferri F, Bouzinba-Segard H, Velasco G, Hubé F, Francastel C (2009) Noncoding murine centromeric transcripts associate with and potentiate Aurora B kinase. Nucleic Acids Res 37:5071–5080

Fingerhut JM, Moran JV, Yamashita YM (2019) Satellite DNA-containing gigantic introns in a unique gene expression program during Drosophila spermatogenesis. PLoS Genet 15:e1008028

Fioriniello S, Csukonyi E, Marano D, Brancaccio A, Madonna M, Zarrillo C, Romano A, Marracino F, Matarazzo MR, D'Esposito M, Della Ragione F (2020) MeCP2 and major satellite forward RNA cooperate for Pericentric heterochromatin organization. Stem Cell Rep 15:1317–1332

Freitak D, Knorr E, Vogel H, Vilcinskas A (2012) Gender- and stressor-specific microRNA expression in *Tribolium castaneum*. Biol Lett 8:860–863

Fukagawa T, Nogami M, Yoshikawa M, Ikeno M, Okazaki T, Takami Y, Nakayama T, Oshimura M (2004) Dicer is essential for formation of the heterochromatin structure in vertebrate cells. Nat Cell Biol 6:784–791

Goenka A, Sengupta S, Pandey R, Parihar R, Mohanta GC, Mukerji M, Ganesh S (2016) Human satellite-III non-coding RNAs modulate heat-shock-induced transcriptional repression. J Cell Sci 129:3541–3552

Grewal SI, Elgin SC (2007) Transcription and RNA interference in the formation of heterochromatin. Nature 447:399–406

Halbach R, Miesen P, Joosten J, Taşköprü E, Rondeel I, Pennings B, Vogels CBF, Merkling SH, Koenraadt CJ, Lambrechts L, van Rij RP (2020) A satellite repeat-derived piRNA controls embryonic development of Aedes. Nature 580:274–277

Hall LL, Byron M, Carone DM, Whitfield TW, Pouliot GP, Fischer A, Jones P, Lawrence JB (2017) HSATII DNA and HSATII RNA foci sequester PRC1 and MeCP2 into cancer-specific nuclear bodies. Cell Rep 18:2943–2956

Hardy RW, Tokuyasu KT, Lindsley DL (1981) Analysis of spermatogenesis in *Drosophila melanogaster* bearing deletions for Y-chromosome fertility genes. Chromosoma 83:593–617

Hecht NB (1986) Regulation of gene expression during mammalian spermatogenesis. In: Experimental approaches to mammalian embryonic development. Cambridge University Press, Cambridge

Hédouin S, Grillo G, Ivkovic I, Velasco G, Francastel C (2017) CENP-A chromatin disassembly in stressed and senescent murine cells. Sci Rep 7:42520

Hsieh CL, Xia J, Lin H (2020) MIWI prevents aneuploidy during meiosis by cleaving excess satellite RNA. EMBO J 17:e103614

Huang C, Wang X, Liu X, Cao S, Shan G (2015) RNAi pathway participates in chromosome segregation in mammalian cells. Cell Discov 1:15029

Ito H, Kim JM, Matsunaga W, Saze H, Matsui A, Endo TA, Harukawa Y, Takagi H, Yaegashi H, Masuta Y, Masuda S, Ishida J, Tanaka M, Takahashi S, Morosawa T, Toyoda T, Kakutani T, Kato A, Seki M (2016) Stress-activated transposon in *Arabidopsis* induces transgenerational abscisic acid insensitivity. Sci Rep 6:23181

Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla ME (2017) LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat Genet 49:1502–1510

Johnson WL, Yewdell WT, Bell JC, McNulty SM, Duda Z, O'Neill RJ, Sullivan BA, Straight AF (2017) RNA-dependent stabilization of SUV39H1 at constitutive heterochromatin. eLife 6: e25299

Jolly C, Metz A, Govin J, Vigneron M, Turner BM, Khochbin S, Vourc'h C (2004) Stress-induced transcription of satellite III repeats. J Cell Biol 164:25–33

Joshi SS, Meller VH (2017) Satellite repeats identify X chromatin for dosage compensation in *Drosophila melanogaster* males. Curr Biol 27:1393–1402

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci USA 97:6603–6607

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9:e1003470

Khan AR, Enjalbert J, Marsollier AC, Rousselet A, Goldringer I, Vitte C (2013) Vernalization treatment induces site-specific DNA hypermethylation at the VERNALIZATION-A1 (VRN-A1) locus in hexaploid winter wheat. BMC Plant Biol 13:209

Kim YB, Oh JH, McIver LJ, Rashkovetsky E, Michalak K, Garner HR, Kang L, Nevo E, Korol AB, Michalak P (2014) Divergence of *Drosophila melanogaster* repeatomes in response to a sharp microclimate contrast in evolution canyon, Israel. Proc Natl Acad Sci USA 111:10630–10635

Kim M, Ekhteraei-Tousi S, Lewerentz J, Larsson J (2018) The X-linked 1.688 satellite in *Drosophila melanogaster* promotes specific targeting by painting of fourth. Genetics 208:623–632

Kishi Y, Kondo S, Gotoh Y (2012) Transcriptional activation of mouse major satellite regions during neuronal differentiation. Cell Struct Funct 37:101–110

Kishikawa T, Otsuka M, Yoshikawa T, Ohno M, Ijichi H, Koike K (2016) Satellite RNAs promote pancreatic oncogenic processes via the dysfunction of YBX1. Nat Commun 7:13006

Kishikawa T, Otsuka M, Suzuki T, Seimiya T, Sekiba K, Ishibashi R, Tanaka E, Ohno M, Yamagami M, Koike K (2018) Satellite RNA increases DNA damage and accelerates tumor formation in mouse models of pancreatic Cancer. Mol Cancer Res 16:1255–1262

Koga A, Tanabe H, Hirai Y, Imai H, Imamura M, Oishi T, Stanyon R, Hirai H (2017) Co-opted megasatellite DNA drives evolution of secondary night vision in Azara's owl monkey. Genome Biol Evol 9:1963–1970

Kuhn GC, Küttler H, Moreira-Filho O, Heslop-Harrison JS (2012) The 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales and association with genes. Mol Biol Evol 29:7–11

Le TN, Miyazaki Y, Takuno S, Saze H (2015) Epigenetic regulation of intragenic transposable elements impacts gene transcription in *Arabidopsis thaliana*. Nucleic Acids Res 43:3911–3921

Lee YCG (2015) The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in *Drosophila melanogaster*. PLoS Genet 11:e1005269

Lee YCG, Karpen GH (2017) Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. elife 6:e25762

Lee YCG, Ogiyama Y, Martins NMC, Beliveau BJ, Acevedo D, Wu CT, Cavalli G, Karpen GH (2020) Pericentromeric heterochromatin is hierarchically organized and spatially contacts H3K9me2 islands in euchromatin. PLoS Genet 16:e1008673

Lemos B, Branco AT, Hartl DL (2010) Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. Proc Natl Acad Sci USA 107:15826–15831

Li YX, Kirby ML (2003) Coordinated and conserved expression of alphoid repeat and alphoid repeat-tagged coding sequences. Dev Dyn 228:72–81

Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, Wysocka J (2018) Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature 553:228–232

Lu J, Gilbert DM (2007) Proliferation- dependent and cell cycle- regulated transcription of mouse pericentromeric heterochromatin. J Cell Biol 179:411–421

Maison C, Bailly D, Roche D, Montes de Oca R, Probst AV, Vassias I, Dingli F, Lombard B, Loew D, Quivy JP, Almouzni G (2011) SUMOylation promotes de novo targeting of HP1α to pericentric heterochromatin. Nat Genet 43:220–227

Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, Springer NM (2015) Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet 11:e1004915

Marshall KE, Sinclair BJ (2012) The impacts of repeated cold exposure on insects. J Exp Biol 215:1607–1613

Menon DU, Coarfa C, Xiao W, Gunaratne PH, Meller VH (2014) siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. Proc Natl Acad Sci USA 111:16460–16465

Meštrović N, Plohl M, Mravinac B, Ugarković Đ (1998) Evolution of satellite DNAs from the genus Palorus—experimental evidence for the "library" hypothesis. Mol Biol Evol 15:1062–1068

Mills WK, Lee YCG, Kochendoerfer AM, Dunleavy EM, Karpen GH (2019) RNA from a simple-tandem repeat is required for sperm maturation and male fertility in *Drosophila melanogaster*. eLife 8:e48940

Navratilova A, Koblizkova A, Macas J (2008) Survey of extrachromosomal circular DNA derived from plant satellite repeats. BMC Plant Biol 8:90

Ninomiya K, Adachi S, Natsume T, Iwakiri J, Terai G, Asai K, Hirose T (2020) LncRNA-dependent nuclear stress bodies promote intron retention through SR protein phosphorylation. EMBO J 39:e102729

Nogalski MT, Shenk T (2020) HSATII RNA is induced via a noncanonical ATM-regulated DNA damage response pathway and promotes tumor cell proliferation and movement. Proc Natl Acad Sci USA 117:31891–31901

Nogalski MT, Solovyov A, Kulkarni AS, Desai N, Oberstein A, Levine AJ, Ting DT, Shenk T, Greenbaum BD (2019) A tumor-specific endogenous repetitive element is induced by herpesviruses. Nat Commun 10:90

Noreen F, Akbergenov R, Hohn T, Richert-Poggeler KR (2007) Distinct expression of endogenous Petunia vein clearing virus and the DNA transposon dTph1 in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. Plant J 50:219–229

Pal-Bhadra M, Leibovitch BA, Gandhi SG, Chikka MR, Bhadra U, Birchler JA, Elgin SC (2004) Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. Science 303:669–672

Papareddy RK, Páldi K, Paulraj S, Kao P, Lutzmayer S, Nodine MD (2020) Chromatin regulates expression of small RNAs to help maintain transposon methylome homeostasis in *Arabidopsis*. Genome Biol 21:251

Park J, Lee H, Han N et al (2018) Long non-coding RNA ChRO1 facilitates ATRX/DAXX-dependent H3.3 deposition for transcription-associated heterochromatin reorganization. Nucleic Acids Res 46:11759–11775

Paulsen T, Kumar P, Koseoglu MM, Dutta A (2018) Discoveries of extrachromosomal circles of DNA in normal and tumor cells. Trends Genet 34:270–278

Pavlek M, Gelfand Y, Plohl M, Meštrović N (2015) Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. DNA Res 22:387–401

Percharde M, Lin CJ, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, Ramalho-Santos M (2018) A LINE1-nucleolin partnership regulates early development and ESC identity. Cell 174:391–405

Pezer Ž, Ugarković Đ (2008) RNA Pol II promotes transcription of centromeric satellite DNA in beetles. PLoS One 3:e1594

Pezer Z, Ugarković Đ (2009) Transcription of pericentromeric heterochromatin in beetles – satellite DNAs as active regulatory elements. Cytogenet Genome Res 124:268–276

Pezer Ž, Ugarković Đ (2012) Satellite DNA-associated siRNAs as mediators of heat shock response in insects. RNA Biol 9:587–595

Pezer Ž, Brajković J, Feliciello I, Ugarković Đ (2012) Satellite DNA-mediated effects on genome regulation. Genome Dyn 7:153–169

Piacentini L, Fanti L, Specchia V, Bozzetti MP, Berloco M, Palumbo G, Pimpinelli S (2014) Transposons, environmental changes, and heritable induced phenotypic variability. Chromosoma 123:345–354

Pita S, Panzera F, Mora P, Vela J, Cuadrado Á, Sánchez A, Palomeque T, Lorite P (2017) Comparative repeatome analysis on Triatoma infestans Andean and non-Andean lineages, main vector of Chagas disease. PLoS One 12:e0181635

Probst AV, Okamoto I, Casanova M, El Marjou F, Le Baccon P, Almouzni G (2010) A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. Dev Cell 19:625–638

Rajshekar S, Yao J, Arnold PK, Payne SG, Zhang Y, Bowman TV, Schmitz RJ, Edwards JR, Goll M (2018) Pericentromeric hypomethylation elicits an interferon response in an animal model of ICF syndrome. eLife 7:e39658

Ratner VA, Zabanov SA, Kolesnikova OV, Vasilyeva LA (1992) Induction of the mobile genetic element Dm-412 transpositions in the Drosophila genome by heat shock treatment. Proc Natl Acad Sci USA 89:5650–5654

Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, Goyal P, Keane TM, Jones S, Hirst M, Lorincz MC, Mager DL (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet 7:e1002301

Rizzi N, Denegri M, Chiodi I, Corioni M, Valgardsdottir R, Cobianchi F, Riva S, Biamonti G (2004) Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. Mol Biol Cell 15:543–551

Rosic S, Kohler F, Erhardt S (2014) Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. J Cell Biol 207:335–349

Rudert F, Bronner S, Garnier JM, Dollé P (1995) Transcripts from opposite strands of gamma satellite DNA are differentially expressed during mouse development. Mamm Genome 6:76–83

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JP (2016) High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep 6:28333

Saksouk N, Simboeck E, Déjardin J (2015) Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin 8:3

Satović E, Vojvoda Zeljko T, Luchetti A, Mantovani B, Plohl M (2016) Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. BMC Genomics 17:997

Scott HS, Kudoh J, Wattenhofer M, Shibuya K, Berry A, Chrast R, Guipponi M, Wang J, Kawasaki K, Asakawa S, Minoshima S, Younus F, Mehdi SQ, Radhakrishna U, Papasavvas MP, Gehrig C, Rossier C, Korostishevsky M, Gal A, Shimizu N, Bonne-Tamir B, Antonarakis SE (2001) Insertion of beta-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness. Nat Genet 27:59–63

Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE (2016) A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res 26:745–755

Seong KH, Li D, Shimizu H, Nakamura R, Ishii S (2011) Inheritance of stress-induced, ATF-2-dependent epigenetic change. Cell 145:1049–1061

Sermek A, Feliciello I, Ugarković Ð (2021) Distinct regulation of the expression of satellite DNAs in the beetle *Tribolium castaneum*. Int J Mol Sci 22:296

Shaul O (2017) How introns enhance gene expression. Int J Biochem Cell Biol 91:145–155

Sienski G, Dönertas D, Brennecke J (2012) Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell 151:964–980

Sproul JS, Khost DE, Eickbush DG, Negm S, Wei X, Wong I, Larracuente AM (2020) Dynamic evolution of Euchromatic satellites on the X chromosome in Drosophila melanogaster and the simulans clade. Mol Biol Evol 37:2241–2256

Sun L, Jing Y, Liu X, Li Q, Xue Z, Cheng Z, Wang D, He H, Qian W (2020) Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. Nat Commun 11:1886

Tanne A, Muniz LR, Puzio-Kuter A, Leonova KI, Gudkov AV, Ting DT, Monasson R, Cocco S, Levine AJ, Bhardwaj N, Greenbaum BD (2015) Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells. Proc Natl Acad Sci USA 112:15154–15159

Thomas J, Phillips CD, Baker RJ, Pritham EJ (2014) Rolling-circle transposons catalyze genomic innovation in a mammalian lineage. Genome Biol Evol 6:2595–2610

Tilman G, Arnoult N, Lenglez S, Van Beneden A, Loriot A, De Smet C, Decottignies A (2012) Cancer-linked satellite 2 DNA hypomethylation does not regulate Sat2 non-coding RNA expression and is initiated by heat shock pathway activation. Epigenetics 7:903–913

Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G, Deshpande V, Iafrate AJ, Letovsky S, Rivera MN, Bardeesy N, Maheswaran S, Haber DA (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science 331:593–596

Ugarković Ð (2005) Functional elements residing within satellite DNAs. EMBO Rep 6:1035–1103

Ugarković Ð, Plohl M (2002) Variation in satellite DNA profiles – causes and effects. EMBO J 21:5955–5959

Valgardsdottir R, Chiodi I, Giordano M, Rossi A, Bazzini S, Ghigna C, Riva S, Biamonti G (2008) Transcription of satellite III non-coding RNAs is a general stress response in human cells. Nucleic Acids Res 36:423–434

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530–536

Vlahović I, Glunčić M, Rosandić M, Ugarković Ð, Paar V (2017) Regular higher order repeat structures in beetle *Tribolium castaneum* genome. Genome Biol Evol 9:2668–2680

Vojvoda Zeljko T, Pavlek M, Meštrović N, Plohl M (2020) Satellite DNA-like repeats are dispersed throughout the genome of the Pacific oyster *Crassostrea gigas* carried by *Helentron* non-autonomous mobile elements. Sci Rep 10:15107

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science 297:1833–1837

Vourc'h C, Biamonti G (2011) Transcription of satellite DNAs in mammals. Prog Mol Subcell Biol 51:95–118

Wang LC, Wu JR, Chang WL, Yeh CH, Ke YT, Lu CA, Wu SJ (2013) Arabidopsis HIT4 encodes a novel chromocentre-localized protein involved in the heat reactivation of transcriptionally silent loci and is essential for heat tolerance in plants. J Exp Bot 64:1689–1701

Wang J, Jia ST, Jia S (2016) New insight into the regulation of heterochromatin. Trends Genet 32:284–294

Wei KH, Grenier JK, Barbash DA, Clark AG (2014) Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. Proc Natl Acad Sci USA 111:18793–18798

Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL, Hannan R, Saffery R, Shaw ML, Williams E, Choo KH (2007) Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res 17:1146–1160

Wylie A, Jones AE, D'Brot A, Lu WJ, Kurtz P, Moran JV, Rakheja D, Chen KS, Hammer RE, Comerford SA, Amatruda JF, Abrams JM (2016) p53 genes function to restrain mobile elements. Genes Dev 30:64–77

Zeller P, Gasser SM (2017) The importance of satellite sequence repression for genome stability. Cold Spring Harb Symp Quant Biol 82:15–24

Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, Nederlof PM, Gage FH, Verma IM (2011) BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. Nature 477:179–184

Zhu Q, Hoong N, Aslanian A, Hara T, Benner C, Heinz S, Miga KH, Ke E, Verma S, Soroczynski J, Yates JR 3rd, Hunter T, Verma IM (2018) Heterochromatin-encoded satellite RNAs induce breast cancer. Mol Cell 70:842–853

Chapter 7
Centromeres Transcription and Transcripts for Better and for Worse

Pia Mihìc, Sabrine Hédouin, and Claire Francastel

Abstract Centromeres are chromosomal regions that are essential for the faithful transmission of genetic material through each cell division. They represent the chromosomal platform on which assembles a protein complex, the kinetochore, which mediates attachment to the mitotic spindle. In most organisms, centromeres assemble on large arrays of tandem satellite repeats, although their DNA sequences and organization are highly divergent among species. It has become evident that centromeres are not defined by underlying DNA sequences, but are instead epigenetically defined by the deposition of the centromere-specific histone H3 variant, CENP-A. In addition, and although long regarded as silent chromosomal loci, centromeres are in fact transcriptionally competent in most species, yet at low levels in normal somatic cells, but where the resulting transcripts participate in centromere architecture, identity, and function. In this chapter, we discuss the various roles proposed for centromere transcription and their transcripts, and the potential molecular mechanisms involved. We also discuss pathological cases in which unscheduled transcription of centromeric repeats or aberrant accumulation of their transcripts are pathological signatures of chromosomal instability diseases. In sum, tight regulation of centromeric satellite repeats transcription is critical for healthy development and tissue homeostasis, and thus prevents the emergence of disease states.

Keywords Centromere · Transcription · Satellite DNA · Cancer · ICF syndrome

Pia Mihìc and Sabrine Hédouin contributed equally with all other contributors.

P. Mihìc · C. Francastel (✉)
Université De Paris, Epigenetics and Cell Fate, CNRS UMR7216, Paris, France
e-mail: claire.francastel@univ-paris-diderot.fr; http://parisepigenetics.com/dmrhd/

S. Hédouin
Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Đ. Ugarković (ed.), *Satellite DNAs in Physiology and Evolution*, Progress in Molecular and Subcellular Biology 60,
https://doi.org/10.1007/978-3-030-74889-0_7

7.1 Satellite DNA Underlies (Peri)centromeric Chromosomal Regions

Centromeres are chromosomal domains specialized in the faithful segregation of the genetic material between daughter cells at each cell division. In most higher eukaryotes, centromeres are made up of satellite DNA, i.e., large arrays of short, mostly A/T-rich, DNA sequences repeated in tandem over chromosomal domains that can range from hundreds of kilobases (Kb) to tens of megabases (Mb) on each centromere. Different families of satellite DNA define two distinct functional chromosomal domains (Choo 2001) with distinct chromatin landmarks (Karpen and Allshire 1997), i.e., the centromere per se and the juxta- or pericentromeric domains. The centromere is defined by the presence of unique nucleosomes, in which histone H3 is replaced by its variant called CENP-A in humans (Cse4 in budding yeast, Cnp1 in fission yeast, and CID/CenH3 in fruit flies) (Earnshaw and Rothfield 1985), interspersed with canonical nucleosomes. The CENP-A nucleosomal domain creates a platform for the assembly of a proteinaceous structure known as the kinetochore, that links chromosomes to mitotic spindle microtubules (Palmer et al. 1991; Fukagawa and Earnshaw 2014; Gambogi et al. 2020). CENP-A deposition is tightly regulated and mediated by a specific histone chaperone Holliday Junction Recognition Protein (HJURP) (Dunleavy et al. 2009; Foltz et al. 2009), although it remains unclear how HJURP directs CENP-A to its default location (Hoffmann and Fachinetti 2017). In juxta- or pericentromeric position, satellite repeats are enriched in repressive epigenetic marks which makes up the bulk of constitutive heterochromatin in mammals (Schueler and Sullivan 2006; Eymery et al. 2009b; Saksouk et al. 2015), and to which several functions have been assigned, including in sister chromatid cohesion at centromeres (Pidoux and Allshire 2005), maintenance of genome stability (Peters et al. 2001), and functional organization of the interphase nucleus (Wijchers et al. 2015; Muller et al. 2019; Francastel et al. 2000).

In contrast to the presence of CENP-A nucleosomes being the main and evolutionary conserved determinant of centromere identity, with very few exceptions in some insect lineages and kinetoplastids (Akiyoshi and Gull 2014; Drinnenberg et al. 2014; Navarro-Mendoza et al. 2019), the evolution of the underlying DNA sequences has been quite dynamic and gave rise to highly divergent centromere organization (Malik and Henikoff 2009). Phylogenetic analysis also found little evidence for satellite sequence conservation, which contrasts with the ancestral conserved structure of tandem repeats at telomeres (Meyne et al. 1989). Centromeres in different species display a wide variety of sequences, repeats organization, and chromosomal positions (Melters et al. 2013; Plohl et al. 2014), that can even diverge between chromosomes of the same species like in human and *Drosophila* (Bracewell et al. 2019; Sullivan and Sullivan 2020). Tandem repeats are highly prevalent at (peri)centromeres of most animal and plant genomes, and monocentric centromeres are the fundamental unit for chromosome inheritance in most species. However, chromosomes in certain insects lack a primary constriction and have adopted holocentric centromeres, i.e., in which the activity of the kinetochore extends over the whole chromosome arms, and therefore lacks a satellite repeats signature (Drinnenberg et al. 2014). There are also examples of atypical, yet functional, centromeres that spontaneously form on unique non-satellite sequences. Originally described in humans, the so-called neocentromeres are functionally and structurally similar to endogenous centromeres but lack the underlying repetitive sequences (Scott and Sullivan 2014). In the domestic horse, the discovery of satellite-less centromeres with variable positions among individuals (Giulotto et al. 2017) reinforced the idea that centromeres are defined, at least in part, by a centromere-specific histone variant, and not by the underlying DNA sequences.

Besides the abundance of repeats and association of centromeres with pericentromeric heterochromatin domains in most eukaryotes, a 17 base-pair (bp) motif is also highly conserved throughout species and called the CENP-B-box (B-Box). This sequence is the consensus binding site for the CENP-B protein, the only centromeric protein with sequence-specific DNA-binding activity (Masumoto et al. 1989). CENP-B protein is itself highly conserved, although its essential nature is debatable since certain centromeres lack a B-box in their satellite repeats, like on the human Y chromosome (Jain et al. 2018), and because CENP-B seems dispensable in the mouse (Kapoor et al. 1998). The development of human artificial chromosomes (HAC) has been instrumental in the determination of the minimal requirements for a functional centromere in terms of protein factors and DNA sequences (Bergmann et al. 2012). Many reports highlighted that both satellite DNA and CENP-B were necessary for de novo assembly of centromeres and HAC formation (Ohzeki et al. 2002; Masumoto et al. 2004), although alternative ways to establish a centromere during HAC formation have been reported (Logsdon et al. 2019), still questioning CENP-B requirement for establishment or maintenance of centromeres.

7.1.1 Examples of Centromere Organization Across Species

The budding yeast *Saccharomyces cerevisiae* and some relatives represent some sort of exceptions in the centromere world. In these organisms, a so-called “point centromere” is defined by short and unique sequences on which the centromere-specific nucleosome is positioned (Lechner and Ortiz 1996; Furuyama and Biggins 2007). This is in contrast with other eukaryotic organisms that feature regional centromeres assembled on kilo- to megabase-scale arrays of tandem repeats at the primary constriction site of the chromosome, and at which CENP-A nucleosomes are interspersed with canonical ones (Blower et al. 2002). The yeast *Schizosaccharomyces pombe* centromeres are composed of a 4–7 Kb-long central core element (ctr) flanked by centromere-specific innermost repeats (imr) sequences and pericentric outer repeats (otr), with an overall size range of 30–110 Kbs depending on the chromosome (Polizzi and Clarke 1991). In maize, centromeres are composed of CentC repeat of 156 bp, which form tandem arrays that span 180 Kb, separated by one or more copies of centromeric retrotransposable

(CR) elements. Similar organization at centromeres is shared by the fruit fly *Drosophila melanogaster*, where the centromere is primarily composed of AATAT and TTCTC satellites, interspersed with complex A/T-rich repeats and mobile elements (Sun et al. 2003; Chang et al. 2019).

A model of choice for the study of centromere organization is the laboratory mouse *Mus musculus* due to its fairly homogeneous centromeres across all chromosomes (Kalitsis et al. 2006). The basic repeat unit in murine centromeres, called a minor satellite, is 120 bp-long and repeated in tandem over about 600 Kbs, which represents around 0.45% of the mouse genome. Murine chromosomes are telocentric, meaning that the centromere is nearly adjacent to the telomere of the short chromosome arm. On this short arm, minor satellite repeats are flanked by a retrotransposable DNA element, the truncated Long Interspersed Nucleotide Element 1 (tL1) and clusters of telocentric tandem (TLC) repeats, which share between 74% and 77% of homology with minor satellites, but lay in the opposite orientation (Kalitsis et al. 2006). On the long chromosome arm, the flanking pericentromeric domains are made up of tandem repeats of 234 bp-long major satellite repeat units over around 6 Mb and representing up to 3% of the mouse genome (Choo 1997; Kalitsis et al. 2006).

As opposed to homogeneous murine satellite repeats, each human centromere shows distinct polymorphisms in the number and sequence of α-satellite repeats (Aldrup-Macdonald and Sullivan 2014). Centromeric regions contain 171 bp-long α-satellite repeat units arranged in a tandem head-to-tail fashion, into higher-order repeat (HORs) units, themselves repeated in a largely uninterrupted fashion up to 5 Mb. The re-iteration of the HOR forms the centromeric α-satellite array, with occasional interruptions by transposable elements (She et al. 2004). In a given HOR, individual α-satellite repeats may only share 50–70% sequence similarity, whereas different HORs from the same chromosome share up to 98% of homology (Miga 2019). Neighboring pericentromeric regions account for about 4% to 5% of the human genome. They can be made of three types of satellite repeats: type I, which are formed by an alternation of 17 and 25 bp monomers and are restricted to chromosomes 2, 3, and acrocentric chromosomes. Type II and type III satellites are made of a 5 bp-long GGAAT repeat unit, found on all chromosomes, although unevenly distributed over several Mb. Notably, large blocks of heterochromatin in juxtacentromeric position on the long arm of chromosomes 1, and 16, or the long arm of chromosome 9, are composed of satellites type II or III, respectively (Vourc'h and Biamonti 2011).

All in all, the few unifying features in centromere organization across eukaryote kingdoms pose somewhat of a paradox given the essential nature of centromeres for the maintenance of genome integrity and the conserved functions and dynamics of kinetochores. Perhaps the most common characteristic of DNA sequences underlying centromeres is that they are transcriptionally competent in most of the species studied.

7.2 Transcription of Centromeric Repeats and Their Transcripts

At a time when recognition of the relatively pervasive aspect of genome transcription was in its infancy, the findings that (peri)centromeric satellite repeats could be transcribed have almost gone unnoticed. However, transcription at centromeric repeats was hinted at by the existence of centromeric transcripts in murine cells (Harel et al. 1968; Cohen et al. 1973) and in lung cells of the newt *Taricha granulosa* (Rieder 1978).

Nowadays, both the transcription of centromeric repeats and its products, the centromeric RNAs (cenRNA), are viewed as a conserved feature of centromeres in a broad range of organisms (Table 7.1), including yeast (Volpe et al. 2002; Choi et al. 2011; Ohkuni and Kitagawa 2011), plants (Topp et al. 2004; Du et al. 2010), beetles (Pezer and Ugarković 2008), *Drosophila* (Grewal and Elgin 2007; Rošić et al. 2014), amphibians (Varley et al. 1980; Diaz et al. 1981; Blower 2016), mouse (Rudert et al. 1995; Bouzinba-Segard et al. 2006; Ferri et al. 2009), and humans (Chan et al. 2012; Quénet and Dalal 2014; McNulty et al. 2017). Interestingly, work using a structurally dicentric chromosome that contains two α-satellite arrays demonstrated that RNA Polymerase II (RNA Pol II) localizes at active centromeres, i.e., at which the kinetochore assembles, but not at the inactive one (Chan et al. 2012), although another study reported that inactive arrays can also produce cenRNAs but just less stable than those originating from active arrays (McNulty et al. 2017).

In essence, in the absence of conserved centromeric DNA sequences across species, it is tempting to speculate that transcription through centromeric repeats or their derived transcripts may be functionally relevant to centromeres identity or function.

7.2.1 Regulation of Centromere Transcription/Transcripts Levels

Most of our knowledge of centromeric repeats transcription has been inferred from the existence of transcripts with centromeric sequences. Yet, the consequence of their highly repetitive and near-identical nature in some species is that they are mostly absent from reference genomes and specifically excluded from high-throughput sequencing analysis. Yet, transcripts with sequences of the identified centromeric repeat units can be found in various databases, including Expressed Sequence Tags (EST) databases. In addition, dedicated experimental testing of the levels of cenRNAs in a given tissue or at a specific developmental stage argued against the idea of simple transcription noise and even provided evidence for some level of transcriptional regulation. The challenge is rather to determine whether the observed differences in transcript abundance are the result of transcriptional or posttranscriptional control mechanisms.

Table 7.1 Satellite transcripts observed in normal and pathological conditions from various model organisms

Organism	Origin	Condition	Size	Pol II	Characteristics	References
S. pombe	Centromeric **Cnt** and **imr**	Exosome mutants	~0.5 Kb	Yes	Not detected in WT cells, transcribed from both strands	Choi et al. (2011)
	Pericentromeric **dg** and **dh**	WT cells	21–25 nt	Yes	RNAi machinery-dependent siRNAs	Volpe et al. (2002)
		RNAi depletion	1.4–2.4 Kb	Yes	Accumulation of cenRNAs longer than siRNAs	
S. cerevisiae	Centromere	Exosome mutants	1.3 Kb	Yes	Not detected in WT cells	Ohkuni and Kitagawa (2012)
Plants	Centromeric **retroelement** Satellite **CentC** repeat	WT cells	40–900 bp	N. D.	Single-stranded, polyadenylated RNA	Du et al. (2010), Topp et al. (2004)
Beetles	Centromeric repeats	WT cells	0.5–5 Kb	Yes	Transcribed from both strands	Pezer and Ugarković (2008)
Drosophila	Centromeric repeats	WT cells	1.3 Kb	Yes	Transcribed from the X chromosome	Rošić et al. (2014)
Amphibians	Pericentromeric **Sat I**	WT cells	N.D.	Yes	Transcribed from both strands	Diaz et al. (1981)
Mouse	Centromeric **Minor satellite**	WT somatic cells	2–4 Kb	Yes	Transcribed from both strands Peak in G2/M	Bouzinba-Segard et al. (2006), Ferri et al. (2009), Hédouin et al. (2017)
		5AZA-treated or differentiated cells	120 nt	N. D.	Transcribed from both strands	
	Pericentromeric **Major satellite**	WT mESCs	25–30 nt	Yes	Transcribed from both strands	Kanellopoulou et al. (2005)
		Dicer-KO mESCs	>1 Kb	Yes	Accumulation of cenRNAs longer than siRNAs	Kanellopoulou et al. (2005)
		Murine embryo, 2-cell stage	N.D.	N. D.	Transcribed first from the Forward then from the Reverse strand	Probst et al. (2010)

Human	Centromeric **α-satellite**	WT cells	1.3 Kb	Yes	Transcribed from both strands Peak in G1	Quénet and Dalal (2014), Bury et al. (2020)
	Pericentromeric **Satellite II**	Pancreatic adenocarcinomas	N.D.	Yes	Transcribed from both strands	Ting et al. (2011)
	Pericentromeric **Satellite III**	Heat shock treated cells	2–5 Kb	Yes	Transcribed from the reverse strand	Rizzi et al. (2004)

Pol II transcribed by RNA polymerase II; *5AZA* 5-aza-2′-deoxycytidine; *KO* knock-out; *mESCs* mouse embryonic stem cells; *N.D.* not determined; *RNAi* RNA interference; *WT* wild-type

In most species, the levels of cenRNAs appear to vary with particular developmental stages and with cell types, tissues, or organs. They have been detected in coleopteran insect species at all three developmental stages: larvae, pupae, and adults (Pezer and Ugarković 2008). In chicken and zebrafish, transcripts from an α-like satellite repeat are detected during early embryogenesis but are limited to the cardiac neural crest, the head, and the heart (Li and Kirby 2003). During mouse early development, pericentromeric major satellite RNAs (pericenRNAs) start being detected at the 2-cells stage and are required for the major reorganization of the nucleus that occurs at this stage, most notably characterized by the assembly of pericentromeric heterochromatin nuclear compartments, concomitantly with zygotic gene activation (Probst et al. 2010). In somatic cells, murine cen- and pericenRNAs accumulate with terminal differentiation, a process also accompanied by major spatial reorganization of constitutive heterochromatin compartments and changes in gene expression programs (Terranova et al. 2005; Bouzinba-Segard et al. 2006).

Transcription of centromeric repeats seems to be also regulated during the cell cycle. In cycling murine cells, cenRNAs begin to accumulate at the end of S phase and peak in the G2/M phase, just before the onset of mitosis (Ferri et al. 2009). This accumulation coincides with the late S phase when murine centromeres are being replicated (Müller and Almouzni 2017), although no formal demonstration that this would facilitate active transcription has been established. In human cells, the levels of cenRNAs do not change throughout the cell cycle (McNulty et al. 2017), although a recent study showed that their levels could fluctuate and peak in G2/M (Bury et al. 2020). In contrast, active RNA Pol II has been detected at human centromeric repeats in G1 (Quénet and Dalal 2014), when cenRNAs levels are low (Bury et al. 2020). More strikingly, elongating RNA Pol II was detected at mitotic centromeres in humans and mice (Chan et al. 2012). This is paradoxical since mitosis is regarded as a phase during which the bulk of the genome is transcriptionally silent (Christova and Oelgeschläger 2002), and this RNA Pol II localization could represent storage or bookmarking for further transcriptional activation when cells reenter the cell cycle. However, incorporation of fluorescent uridine-5′-triphosphate nucleotides (UTP) at the mitotic centromere suggested that human and murine centromeres are indeed actively transcribed during mitosis (Chan et al. 2012).

7.2.2 *Mechanisms of Transcriptional Regulation*

7.2.2.1 Transcriptional Machinery at Centromeric Repeats

The characteristic organization of most satellite DNA sequences is based on tandem repeats devoid of canonical promoter sequences, which led to the proposal that they could be transcribed by read-through from upstream genes or promoters of transposable elements, which is the case in maize (Topp et al. 2004). It should be noted that not all centromeres, like the human Y chromosome, may contain transposon sequences (Miga et al. 2014). In addition, a candidate TATA-box has been identified within human α-satellite sequences, as well as an SV40 enhancer-core sequence with spacing and orientation characteristic of RNA Pol II-transcribed genes (Vissel et al. 1992). The hypothesis that cryptic promoter elements are present within repeat sequences would not be surprising since, like any genomic sequence, centromeric repeats are stuffed with consensus binding sites for regulatory proteins. Some of these binding sites have long been known to serve as entry sites for the basal transcriptional machinery, like GATA sites, in place of a canonical TATA-box (Aird et al. 1994). Of note, consensus binding sites for GATA factors (WGATAR in which W indicates A/T and R indicates A/G) are frequently occurring in mammalian genomes, including at murine and human centromeric repeats, although occupancy by GATA-family members has not been described.

Whether cells have adapted to the fortuitous binding of various transcription factors to genomic loci essential for their survival is not known. Nevertheless, it may explain why centromere transcription appears to be regulated depending on cellular contexts, and hence, may rely on context-specific transcription factors. In addition, accumulation of cenRNAs of the size of a repeat unit in the mouse suggested that each repeat unit might contain a transcription start site (Bouzinba-Segard et al. 2006), although we cannot exclude that posttranscriptional processing of longer transcripts occurs, which is discussed below.

The question of the RNA polymerase(s) involved, and the means employed to regulate transcription of centromeric repeats, is also important. Centromeres of both budding and fission yeast are transcribed by RNA Pol II (Ohkuni and Kitagawa 2011; Sadeghi et al. 2014). In beetles, the presence of a cap structure and poly (A) tails in a subset of cenRNAs, termed PRAT, is also indicative of RNA Pol II-dependent transcription (Pezer and Ugarković 2008). Differential inhibition of RNA Pol I, II, and III and detection of active RNA Pol II at humans (Quénet and Dalal 2014; McNulty et al. 2017) and murine (Chan et al. 2012) centromeres indicated that RNA Pol II orchestrates the transcription of their centromeres.

In sum, RNA Pol II seems to be responsible for most part of the transcription of both point and regional centromeres, suggesting that transcription of this essential chromosomal domain has been conserved throughout evolution.

7.2.2.2 Transcription Factors

Compatible with the plethora of putative consensus binding sites for transcription factors in centromeric sequences, activators and repressors have been identified to control transcription of satellite repeats in various systems in a similar mode to the regulation of gene promoters.

In *S. cerevisiae*, the transcription factor Centromere-binding protein 1 (Cbf1) has been implicated as an activator of centromere transcription in an RNA Pol II-dependent manner (Ohkuni and Kitagawa 2011), although this is still debated. Indeed, other studies reported the upregulation of cenRNAs in cells lacking Cbf1, in association with chromosomal instability through the downregulation of the protein levels and mislocalization of CENP-A, HJURP, and components of the

Chromosome Passenger Complex (CPC) (Ling and Yuen 2019; Chen et al. 2019). Along the same line, Htz1 (human homolog of H2A.Z) was identified as a transcriptional repressor since its deletion resulted in an upregulation of cenRNAs levels (Ling and Yuen 2019). Interestingly, the double invalidation of Cbf1 and Htz1 resulted in an additive effect on the upregulation of cenRNAs, suggesting that these two proteins operate in distinct pathways to repress centromere transcription (Ling and Yuen 2019). Noteworthy, Cbf1 is conserved among species with point centromeres, but not in eukaryotic species that have regional centromeres.

The Daxx-like motif-containing GATA factor Ams2 was actually one of the first transcription factors shown to be required for centromere function in *S. pombe* (Chen et al. 2003). Ams2 is a cell cycle-regulated factor that occupies centromeric chromatin in S phase where it is required for SpCENP-A deposition, although if this occurs through promoting transcription of cenRNAs has not been established. Again, whether a role for GATA factors is conserved among species has not yet been tested.

The only reported transcriptional regulator for transcriptional activation at murine and human centromeric repeats is the Zinc Finger and AT-Hook Domain Containing (ZFAT) protein, through binding to the ZFAT box, a short sequence present at centromeres of all chromosomes in mouse and human (Ishikura et al. 2020). In mammals, more is known about the transcription of the neighboring pericentromeric satellite repeats. In the mouse, transcriptional repressors YY1 (Shestakova et al. 2004), C/EBPα (Liu et al. 2007), and Ikaros (Brown et al. 1997; Cobb et al. 2000) appear to bind directly to major satellite sequences, although their link with transcriptional repression of these repeats has not been tested. In contrast, heat-shock transcription factor 1 (HSF1) has been shown to promote transcriptional activation of Sat III repeats in response to cellular stress in human cells (Jolly et al. 2004; Rizzi et al. 2004).

7.2.2.3 Histone Marks and DNA Methylation

In addition to occupancy by transcription factors, epigenetic modifications are likely to participate in the control of the transcription of centromeric repeats. In contrast to nearby pericentromeric heterochromatin, and in addition to CENP-A-containing nucleosomes, centromeres exhibit marks of euchromatin such as dimethylation of lysine 4 of histone H3 (H3K4me2) and lack of heterochromatin marks such as di- and tri-methylation of lysine 9 at histone H3 (H3K9me2 or me3) (Sullivan and Karpen 2004). In addition, a hypoacetylated state at centromeres seems to be conserved across eukaryotes (Wako et al. 2003; Sullivan and Karpen 2004; Choy et al. 2011). More specifically, hypoacetylated lysine 16 of histone H4 (H4K16) was shown to be required for kinetochore function and accurate chromosome segregation, although the link with transcriptional competence of centromeric repeats was not assessed (Choy et al. 2011). More recently, acetylated lysine 4 of histone H3 (H3K4ac) was described at centromeres, at which it is required for the centromeric localization of the chromatin reader Bromodomain-containing protein 4 (BRD4), which in turn recruits the RNA Pol II (Ishikura et al. 2020). At pericentromeric satellite repeats, SIRT6, a member of the Sirtuin family of deacetylases, has been implicated in the maintenance of their silent state through deacetylation of histone H3 at lysine 18 (H3K18ac) (Tasselli et al. 2016). Knockdown of SIRT6 caused an aberrant accumulation of pericenRNAs associated with mitotic errors through desilencing of pericentromeric repeats, which also lent support to the importance of heterochromatin maintenance in centromere function.

Additional important insights into the causal roles of histone modifications for transcription at centromeres came from the study of human artificial chromosomes (HACs) in which a CENP-B box was replaced by a tetracycline operator (tetO). Demethylation of H3K4me2, through targeting of the Lysine-Specific histone Demethylase 1A (LSD1) to the tetO sequence, induced a strong decrease in centromere transcription associated with impaired recruitment of HJURP, the CENP-A chaperone (Bergmann et al. 2011). This study provided a nice demonstration of a causal link between a specific activating histone mark, transcription of centromeric repeats, maintenance of CENP-A, and kinetochore functions.

Ubiquitination of histones is also important for the transcription of satellite repeats, with activating or repressive roles for ubiquitination of histone H2B or H2A (H2A-Ub; H2B-Ub), respectively (Zhu et al. 2011; Sadeghi et al. 2014). Loss of function of the Breast Cancer type 1 susceptibility protein (BRCA1) in cancer cells led to reduced H2A-Ub at pericentromeric satellite repeats, accompanied by their transcriptional derepression and loss of heterochromatin integrity (Zhu et al. 2011). Knockdown of Ring Finger Protein 20 (RNF20), the ubiquitin ligase responsible for H2B-Ub in *S. pombe,* led to reduced levels of transcription and nucleosome turnover at centromeres, associated with impaired centromere function. These data suggested that H2B-ub is essential for the maintenance of active centromeric chromatin (Sadeghi et al. 2014).

In contrast to their divergent sequence and structure, a common feature of mammalian satellite sequences is their methylated state at CpG dinucleotides, the main context for DNA methylation at least in mammals, which is also a major epigenetic mechanism to consider. The density of methylatable CpGs is higher at pericentromeres, which is consistent with their heterochromatin status, whereas centromeric repeat units in mice and humans contain only 2 to 3 CpGs *per* repeat unit. At repetitive elements, DNA methylation has been implicated in the inhibition of transposition and mitotic recombination between repeats, in part through maintaining these elements in a silent state (reviewed in Saksouk et al. 2014; Francastel and Magdinier 2019; Scelfo and Fachinetti 2019). However, our vision of the direct relationship between DNA methylation and transcriptional states of these regions may only be partial. For example, treatment of cells with the DNA demethylating agent 5-aza-2′-deoxycytidine (5AZA) led to increased levels of pericenRNAs in human cells (Eymery et al. 2009a) and of cenRNAs in murine cells (Bouzinba-Segard et al. 2006). However, it remains to be determined whether is it directly through demethylation of satellite repeats or through more indirect effects caused by 5AZA treatment, e.g., increased DNA damage also known to cause accumulation of Sat III pericenRNAs in humans (Valgardsdottir et al. 2008) or cenRNAs in murine cells (Hédouin et al. 2017). The identification of the DNA methyltransferases (DNMTs) responsible for de novo methylation (DNMT3A and DNMT3B) (Okano et al. 1999) or maintenance of methylation (DNMT1) (Bestor et al. 1988) was decisive in the dissection of more direct links. Notably, mouse minor satellites and human pericentromeric Sat II and Sat III repeats were identified as specific targets of the de novo methyltransferase DNMT3B (Okano et al. 1999; Xu et al. 1999). Human cells deficient for DNMT1 and DNMT3B are hypomethylated on Sat III repeats but do not accumulate pericenRNAs compared to wild-type cells (Eymery et al. 2009a). Similarly, hypomethylation of (peri)centromeric regions in murine embryonic stem cells (mESCs) deficient for Dnmt1 and Dnmt3a/b is not sufficient to promote their transcriptional activation (Lehnertz et al. 2003; Martens et al. 2005). However, in physiopathological cellular contexts further discussed below, loss of DNA methylation at (peri)centromeres has been associated with their transcriptional derepression, although it remains to be determined whether this is a direct consequence or a mere byproduct of the disease states.

It is possible that the methylated state at satellite repeats could influence the binding of transcriptional activators or repressors. It is important to note that two of the methylatable CpGs in human and murine centromeric repeat units reside within the CENP-B-box. Yet, the impact of CpG methylation on CENP-B binding and centromere architecture is still debated (Scelfo and Fachinetti 2019). DNA methylation was shown to prevent CENP-B binding to the CENP-B-box in vitro (Tanaka et al. 2005). Conversely, global demethylation using 5AZA treatment in cultured cells led to CENP-B spreading over demethylated repeats (Mitchell et al. 1996), although nearby pericentromeric repeats without a CENP-B-box are also demethylated in these conditions, making it difficult to conclude. Nevertheless, DNA methylation could be directly linked to the correct assembly of centromere architecture independently of the transcription of the repeats.

In sum, hypomethylation may create a favorable environment for transcriptional activation of satellite repeats, although it might not be sufficient. Hence, one has to consider that specific cellular contexts and their associated tissue- or context-specific transcription factors would be a prerequisite.

7.3 Functional Relevance of Centromeric Transcripts/Transcription

7.3.1 Centromeres Transcription and Chromatin Remodeling Processes

It appeared that low transcriptional activity is a characteristic of centromeric repeats in normal somatic cells. Several studies suggested that active transcription at centromeres would not only be compatible with but even required for centromere function. Indirect evidence came from the global inhibition of transcription by

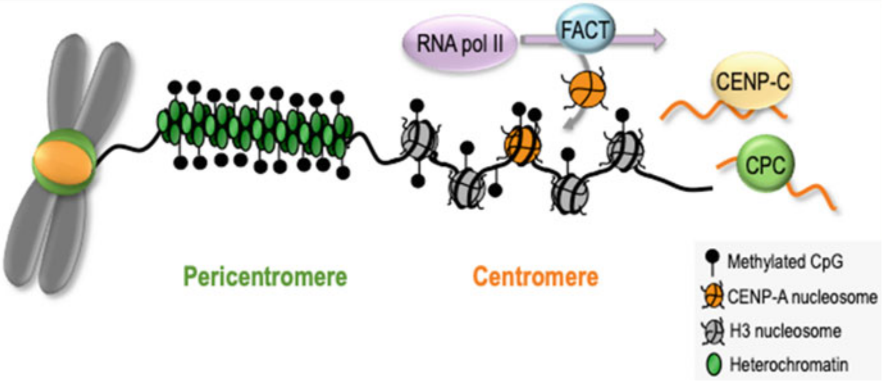


Fig. 7.1 Role of centromere transcription and transcripts in physiological conditions. Centromeric chromosomal domains are marked by a combination of nucleosomes containing histone H3 (gray) or CENP-A (orange) and are flanked by pericentromeric heterochromatin domains (green). RNA pol II-mediated transcription of centromeric repeats, together with chromatin remodelers such as the FACT complex, contribute to the dynamics of chromatin at centromeres and to the deposition of CENP-A. The derived centromeric transcripts serve as scaffolds or guides for the correct localization of centromeric proteins (e.g., CENP-C) and their associated complexes such as the chromosomal passenger complex (CPC), which includes Survivin, INCENP, and Aurora B kinase

RNA Pol II inhibitors which led to compromised centromere function (Chan et al. 2012; Quénet and Dalal 2014). Transcription or nucleosome turnover at centromeres may be important for the dynamics of nucleosomes at centromeric repeats and deposition of CENP-A (Fig. 7.1). Notably, the complex Facilitates Chromatin Transcription (FACT) is localized to centromeres (Foltz et al. 2006) and is involved in CENP-A deposition via the recruitment of the chromatin remodeler Chromodomain Helicase DNA Binding Protein 1 (CHD1) in chickens (Okada et al. 2009). In mammals, a subunit of FACT, the Structure Specific Recognition Protein 1 (SSRP1), co-localized with RNA Pol II at centromeres during mitosis and was necessary for the efficient deposition of CENP-A in early G1 (Chan et al. 2012). Similarly, FACT-mediated transcription was also shown to be required for the de novo incorporation of CENP-A in *Drosophila* S cells (Chen et al. 2015). In this study, knockdown of FACT led to the loss of transcription at centromeres and reduced CENP-A loading. In recent years, a two-step process for *Drosophila* CENP-A loading was proposed (Bobkov et al. 2018). The first step was transcription-independent, during which CENP-A localized to centromeres through the *Drosophila*-specific chaperone Chromosome Alignment defect 1 (CAL1) (Chen et al. 2014), but a second step of active transcription was necessary for its stable incorporation into chromatin. In *S. pombe* mutants that are unable to restart stalled RNA Pol II at centromeres, CENP-A was still efficiently deposited, suggesting that halting RNA Pol II at centromeres may promote local chromatin remodeling events sufficient for CENP-A deposition (Catania et al. 2015). In a context where centromeric transcription-dependent chromatin remodeling is required for stable incorporation of its epigenetic determinant CENP-A, which also relies on the eviction of previously deposited H3/H3.3-placeholder nucleosomes (Dunleavy et al. 2011), the replication-independent histone chaperone and transcription elongation factor Spt6 (SUPT6H in human) was identified as a conserved CENP-A maintenance factor (Bobkov et al. 2020). Spt6 was shown to prevent loss of centromere identity in a transcription-dependent manner, through promoting the recycling of CENP-A and maintenance of parental CENP-A nucleosomes in both *Drosophila* and human cells (Bobkov et al. 2020).

Importantly, targeting of a strong trans-activation domain from Herpes simplex virus (VP16) to centromeres of HACs impaired the incorporation of newly synthesized CENP-A and led to the eviction of the parental one (Bergmann et al. 2011). Hence, even though transcriptional activation of centromeric repeats is necessary for centromere identity, increased transcription at this locus is incompatible with centromere function since it leads to loss of CENP-A at HAC centromeres (Bergmann et al. 2011).

To date, our knowledge of the immediate contribution of centromere transcription on centromere identity in different species still remains incomplete. Since the direct output from transcription at centromeres is the production of cenRNAs, whether long or short-lived, a question that arises is whether cenRNAs themselves could be implicated in the maintenance of centromere identity and function.

7.3.2 Functional Relevance of Centromeric Transcripts themselves

A growing body of evidence suggests that cenRNAs themselves may contribute to proper kinetochore assembly (Chen et al. 2003; Nakano et al. 2003; Topp et al. 2004; Ferri et al. 2009). Notably, cenRNAs are an integral part of the centromeric fraction (Ferri et al. 2009; McNulty et al. 2017; Kabeche et al. 2018), and coprecipitate with CENP-A in maize (Topp et al. 2004), mouse (Ferri et al. 2009), and humans (Chueh et al. 2009). Importantly, knockdown of human cenRNAs, without affecting transcription of the locus per se, led to impaired CENP-A deposition (Quénet and Dalal 2014). This finding suggested that correct loading of CENP-A does not only depend on active transcription at centromeres (see above), but also requires the transcripts themselves. This was further evidenced by the knockdown of cenRNAs in extracts of *Xenopus* oocytes, which led to decreased occupancy of CENP-A at centromeres (Grenfell et al. 2016).

Besides CENP-A deposition, several studies showed a direct association between cenRNAs and CENP-C in gel shift or immunoprecipitation experiments, probably through the CENP-C RNA binding domain. This was the case for example in maize (Du et al. 2010) and *Drosophila* (Rošić et al. 2014). RNase treatment of human cells induced the delocalization of CENP-C but not that of CENP-A (Wong et al. 2007). Similarly, inhibition of transcription in mitosis, a stage at which centromeres are transcribed by RNA Pol II, impaired the localization of CENP-C at centromeres (Chan et al. 2012).

Beyond their association with the constitutive components of the centromere, cenRNAs also interact with components of the CPC: Aurora B, INCENP, and Survivin in the G2/M phase in murine cells (Ferri et al. 2009). More specifically, the association of cenRNAs with the mitotic kinase Aurora B was necessary for Aurora B interaction with its partners Survivin and CENP-A and potentiated its kinase activity (Fig. 7.1). Since cenRNAs peak in G2/M, these data therefore suggested a role for cenRNAs in the timely recruitment or stabilization of the CPC components specifically at centromeres before the onset of mitosis. In turn, unscheduled accumulation of murine cenRNAs throughout the cell cycle led to ectopic localization of CPC proteins, mitotic abnormalities, and loss of cohesion between sister chromatids (Bouzinba-Segard et al. 2006). This interaction between cenRNAs and CPC proteins is conserved in *Xenopus* and humans, as seen with the knockdown of cenRNAs which impaired the recruitment of CPC at centromeres (Ideue et al. 2014; Blower 2016). Strikingly, the association of cenRNAs with Aurora B was also shown to be required for both telomerase activity and maintenance of telomere length in mESCs (Mallm and Rippe 2015). In the mouse, centromeres and telomeres being in close proximity, it is possible that cenRNAs may favor a local concentration of the kinase for shared functions on two essential chromosomal domains.

Together, these data emphasize that the fine-tuning of centromeric transcription/ transcript levels is absolutely required, since too low or too high transcription or cenRNAs levels have deleterious consequences for centromere identity and function, and hence for normal cell growth and survival, and are emerging as new kinds of players in the development of disease as discussed in the following chapters.

7.3.3 Regulation of the Levels of (Peri)centromeric Transcripts Themselves

As mentioned above, the dynamic levels of cenRNAs according to cellular contexts may not only result from regulatory processes at the level of transcription but also at the level of the transcripts themselves, through fine-tuning of their stability or their processing. In *S. cerevisiae*, low levels of cenRNAs might be insured by their degradation by the exosome (Houseley et al. 2007; Ling and Yuen 2019). In many species, centromeric repeats are transcribed in both sense and antisense orientations (Topp et al. 2004; Li et al. 2008; Carone et al. 2009; Ideue et al. 2014), which would favor the formation of double-stranded RNAs (dsRNAs) that are substrates for further processing by the RNA interference (RNAi) machinery. The first example of the possibility that cenRNAs could be processed into smaller species was provided by the discovery of endogenous small interfering RNA (siRNAs) of centromeric origin in *S. pombe* (Reinhart and Bartel 2002). In addition, deletion of factors of the *S. pombe* RNAi machinery such as Dicer, the RNA-binding protein Argonaute (AGO), or the RNA-dependent RNA polymerase (RdRP), led to aberrant accumulation of cenRNAs and chromosome missegregation due to defective pericentromeric heterochromatin formation (Volpe et al. 2002). Similarly, in human–chicken hybrid cells (chicken DT40 cells carrying a human chromosome), ablation of Dicer led to mitotic defects and premature sister chromatid separation that was attributed to the loss of HP1 at pericentromeric heterochromatin and mislocalization of the cohesin complex (Fukagawa et al. 2004). In mouse embryonic stem cells (mESCs), Dicer deficiency also caused an accumulation of pericenRNAs, ranging from 40 nt to over 200 nt in size, i.e., not in the size range of siRNAs (Kanellopoulou et al. 2005; Murchison et al. 2005). Dicer has been involved in the repression of pericentromeric repeats in many species, but its depletion did not lead to mitotic defects in mESCs, although they exhibited differentiation and proliferation defects. In human cells, knockdown of Dicer or AGO2 resulted in chromosome lagging and increased levels of cenRNAs (Huang et al. 2015).

In fact, outside of the well-characterized *S. pombe* model, the literature is punctuated by opposing views of the role of Dicer/RNAi pathway in the regulation of the levels of satellite transcripts, with consequences for heterochromatin assembly and chromosomal stability. This is probably related to the failure to detect 25–30 nt RNA species, at least in the mouse (Kanellopoulou et al. 2005; Bouzinba-Segard et al. 2006). However, in mESCs, 150 nt-long cenRNAs, and smaller species but longer than siRNAs, have been detected and shown to rely on Dicer for their biogenesis (Kanellopoulou et al. 2005). Whether these transcripts play similar roles as (peri)centromeric siRNAs found in *S. pombe* is an interesting possibility. Of note, whereas exponentially growing somatic murine cells exhibit low levels of 2–4 Kb cenRNAs, the accumulation of 120 nt-long cenRNAs in physiopathological conditions recapitulated the same phenotypic defects observed in Dicer-deficient *S. pombe* that failed to produce centromeric siRNAs, including loss of sister chromatid cohesion, impaired centromere architecture and heterochromatin organization, associated with mitotic defects (Bouzinba-Segard et al. 2006). Although it is still not known whether these shorter cenRNAs result from cleavage through unknown mechanisms or are produced by multiple transcription initiation events, these data suggested that the absence of mature siRNAs or the accumulation of unprocessed longer RNA species have the same impact on the integrity of centromeric regions (Bouzinba-Segard et al. 2006).

The low levels of cenRNAs in normal conditions could indicate a short half-life caused by the rapid degradation of the transcripts through mechanisms exposed above or through posttranscriptional modifications shown to regulate the stability of transcripts (Nachtergaele and He 2017; Boo and Kim 2020). Among the myriad of known posttranscriptional RNA modifications, the adenosine to inosine (A to I) modification mediated by the adenosine deaminase acting on RNA (ADAR) machinery was shown to edit structured dsRNAs, with selectivity for certain internal loops and bulges rather than for a consensus sequence (Levanon et al. 2005). Just like the excitement of a connection between the RNAi machinery in keeping (peri)-centromeric transcription in check in a broad range of eukaryotes in the early years of 2000 (Lippman and Martienssen 2004), the involvement of ADAR in the silencing of repetitive sequences in heterochromatin also attracted much interest a few years later (Fernandez et al. 2005). The A/T-rich and potentially dsRNAs produced from murine pericentromeric satellite repeats (Kanellopoulou et al. 2005) would make excellent substrates for the deamination of adenosine to inosine residues by ADAR. However, the search for such editing and for an immunolocalization of factors of the ADAR complex at (peri)centromeric domains remained unsuccessful (Lu and Gilbert 2008). Yet, it is interesting to note that RNA editing by ADAR is incompatible with the RNAi machinery (Scadden and Smith 2001). This is exemplified by the A to I conversion on microRNAs (miRNAs) derived from repetitive Long interspersed nuclear element 2 (LINE2), which blocks their cleavage by Dicer in human and mouse (Kawahara et al. 2007).

Thus, there is an exciting possibility that the dynamic size range of cenRNAs, with potentially distinct roles, could result from the fine-tuning of a balance between RNA modifications and RNA processing depending on phases of the cell cycle or cellular contexts.

7.4 Centromeric Transcripts: Cause or Consequence of Disease?

7.4.1 Accumulation of Satellite Transcripts in Various Types of Cellular Stress

Cells are constantly exposed to various environmental or endogenous stresses that may jeopardize their identity and viability. Exogenous sources of cellular stress include high temperatures [heat shock (HS)], DNA damaging chemicals, UV and ionizing radiation, hyperosmotic and oxidative stresses, whereas endogenous stress can originate from cellular metabolism or replication defects. In this context, a major challenge for the cells is therefore to safeguard their genome integrity. This is fundamental for their survival but also for the normal functioning of the whole organism and protection from the emergence of disease states. In that respect, cells have evolved sophisticated mechanisms that can trigger a rapid and adapted cellular response, including cell-cycle arrest, to allow time to repair DNA lesions or activate sets of genes to recover from stress, and therefore maintain their genomic integrity. There is now a large body of evidence showing that transcription of (peri)-centromeric satellite repeats is rapidly induced in cells under various stress conditions, through the activity of transcription factors belonging to different stress-response pathways, thereby increasing the levels of resulting satellite transcripts as integral components of the stress response.

The accumulation of human pericenRNAs from Sat III of chromosome 9 was the first and best-characterized example of the accumulation of satellite transcripts in stress conditions. Originally described in response to HS (Jolly et al. 2004; Rizzi et al. 2004), it occurs in all of the above-cited types of stress conditions (Valgardsdottir et al. 2008). In response to HS, transient subnuclear organelles, called nuclear stress bodies (nSB), assemble on blocks of Sat III DNA repeats at which the Heat Shock Factor 1 (HSF1) binds, which in turn recruits the RNA Pol II and promotes transcriptional activation of the repeats (Biamonti and Vourc'h 2010). In stressed cells, nSBs are thought to contribute to the rapid and transient shutdown or reprogramming of gene expression programs required for the cells to recover from stress, through a Sat III transcripts-mediated trapping of a subset of splicing and transcription factors away from their site of action (Eymery et al. 2010). Pericentromeric heterochromatin is also central to the functional organization of the cell nucleus and to the maintenance of gene silencing through their positioning in the vicinity of subnuclear heterochromatin compartments (Francastel et al. 2000; Fisher and Merkenschlager 2002). Hence, transcriptional activation of pericentromeric Sat III sequences may also have a long-range impact on gene expression programs through the disorganization of these repressive nuclear compartments. It is interesting to note that activated transcription of satellite repeats in response to thermal perturbations appears to be a shared feature as it also occurs in beetles (Pezer and Ugarkovic 2012), *Arabidopsis thaliana* (Pecinka et al. 2010; Tittel-Elmer et al. 2010), and mouse (Hédouin et al. 2017). However, whether it triggers similar mechanisms in these organisms is not known. In fact, in the mouse, HS stimulates only a modest increase in cenRNAs levels but not that from pericentromeric repeats, whereas nSBs described in human cells do not form in murine cells. Hence, cellular responses to stress may vary between organisms although satellite transcripts appear to be central to the stress response.

In contrast to HS in murine cells, genotoxic stress led to a strong and rapid transcriptional activation of centromeric repeats, followed by local accumulation of cenRNAs at their site of transcription (Hédouin et al. 2017). Transcriptional activation, and not the transcripts themselves, has been causally linked to the loss of centromere identity characterized by the delocalization of CENP-A away from its default location. CENP-A delocalization, or eviction of CENP-A nucleosomes, was dependent on the chromatin remodeler FACT, pinpointing another function of FACT at centromeres in nucleosome destabilization at centromeres, and on the DNA Damage Response (DDR) effector ATM (Hédouin et al. 2017). Importantly, genotoxic stress-induced transcriptional activation of centromeric repeats had distinct functional consequences for cellular phenotypes depending on the integrity of the p53 checkpoint. Whereas immortalized cells continued to cycle, while accumulating micronuclei indicative of mitotic errors and centromere dysfunction, primary cells with normal p53 entered premature cell-cycle arrest and senescence (Hédouin et al. 2017). Hence, in the mouse, activated transcription at centromeric repeats provides a safeguard mechanism to prevent genomic instability in the context of persistent DNA damage signaling, through the disassembly of the core components of centromere identity and function. Whether this mechanisms is related to the acrocentric structure of murine chromosomes, i.e., with telomeres close to centromeric repeats, or to the mouse having very long telomeres so they enter senescence through mechanisms distinct than telomere shortening, is not known. However, together with the example of increased levels of Sat III pericenRNAs in heat-shocked cells, this illustrated the functional relevance for satellite transcripts in the stress response and protection of genome integrity.

7.4.2 Deregulation of Satellite Transcription/Transcripts in Cancer

Consistent with the above-mentioned links between transcription of satellite repeats and the triggering of safeguard mechanisms, it is not surprising that aberrant accumulation of transcripts from satellite repeats characterizes disease states with chromosomal instability like cancer, and that they actually represent good biomarkers of cancerous lesions (Eymery et al. 2009a; Ting et al. 2011; Zhu et al. 2011; Bersani et al. 2015; Tasselli et al. 2016; Hall et al. 2017). Yet, whether they are mere byproducts of disease phenotypes or act as drivers of disease, and through which mechanisms, are still open questions.

Macroarray-based approaches, designed to assess levels of transcripts from various repeated elements including satellite sequences, showed that the levels of satellite transcripts are higher in a variety of cancer cells compared to their normal healthy counterparts (Eymery et al. 2009a). Aberrant accumulation of satellite transcripts has also been reported in a wide range of primary epithelial tumors, both in humans and mice (Eymery et al. 2009a; Ting et al. 2011; Zhu et al. 2011). High-throughput sequencing has further highlighted that satellite transcripts actually represent up to 50% of transcriptional output in these tumors, which is in stark contrast to the low levels of these RNAs in normal cells (Ting et al. 2011).

An important open question remains as to the mechanisms that lead to the pathological transcriptional derepression of satellite sequences. Abnormal levels of satellite transcripts are often associated with the global hypomethylation that characterizes cancer cells, which in fact reflects reduced DNA methylation at repeated sequences owing to the large fraction of the genome they represent and their heavily methylated state in normal cells (Ross et al. 2010). The derepression of satellite transcripts was indeed shown to correlate with reduced DNA methylation of the underlying repeats (Ting et al. 2011; Unoki et al. 2020).

More hints into causal links between the aberrant accumulation of satellite transcripts and chromosomal instability came from gain-of-function experiments where cenRNAs were ectopically transcribed from expression vectors. In murine cells, ectopic expression of minor satellites from one repeat unit was sufficient to promote mitotic defects and alterations of nuclear organization typical of cancer cells (Bouzinba-Segard et al. 2006). Unscheduled accumulation of cenRNAs led to mitotic errors and disorganized centromere architecture through the trapping of centromeric protein complexes away from their default location (Bouzinba-Segard et al. 2006), providing a direct link between high levels of cenRNAs and centromere dysfunction. Likewise, ectopic expression or injection of satellite transcripts in cultured human or murine cells also led to mitotic errors (Zhu et al. 2011, 2018; Kishikawa et al. 2016, 2018), in correlation with the accumulation of foci of phosphorylated histone H2A.X (γ-H2A) that marks DNA double-strand breaks (DSBs) (Zhu et al. 2011). These data suggested that increased levels of satellite transcripts, and not transcriptional activation per se, is deleterious to the cells and leads to increased DNA mutation rates. They also put forward an interesting interplay between pathological high levels of satellite transcripts and DNA damage. In support of this hypothesis, there is the finding that these satellite transcripts accumulate strongly in breast cancer cells deficient for the *BRCA1* gene (Zhu et al. 2011). Likewise, BRCA1 depletion has been causally linked to elevated levels of cenRNAs associated with impaired centromere architecture and chromosome missegregation (Di Paolo et al. 2014). BRCA1 is an important repair factor which, in normal conditions, occupies centromeric chromatin in interphase and throughout mitosis in normal cells (Pageau and Lawrence 2006; Di Paolo et al. 2014; Gupta et al. 2018), and may serve as a guardian of centromere integrity. Aberrant levels of cenRNAs, just like it has been described for kinetochore proteins (Bouzinba-Segard et al. 2006), were proposed to lead to BRCA1 delocalization away from centromeres and further exposure of this locus to the accumulation of unrepaired genotoxic insults (Zhu et al. 2018). Yet, the molecular mechanisms may not be that straightforward since BRCA1, besides its role in DNA damage repair, operates pleiotropic functions linked to the maintenance of chromosomal stability including at the replication fork, control of the cell cycle, and many other regulatory functions (Savage and Harkin 2015). BRCA1 was also recently shown to be an important determinant of the epigenetic states of centromeric and pericentromeric chromatin, through its ubiquitin ligase activity (Zhu et al. 2011). H2A ubiquitination at Lys 19 by BRCA1 provides a repressive mark at centromeric repeats important for their transcriptional repression. Hence, pathogenic variants of BRCA1 would promote DNA damage through the derepression of centromeric repeats in addition to, or instead of, promoting the accumulation of satellite transcripts. Along the same lines, centromeric targeting of VP16 for transcriptional activation of the underlying repeats in murine cells also promoted chromosomal instability (Zhu et al. 2018). Conversely, genotoxic stress using DSB inducers triggered the rapid transcriptional activation of murine centromeric repeats in a p53- and ATM-dependent dependent manner (Hédouin et al. 2017). In that case, transcriptional activation preceded, and was required for, eviction of CENP-A from centromeric chromatin, suggesting a direct link between activated transcription at centromeres and loss of centromere function and identity.

All these data suggested that our vision of the functional consequences of activated transcription of satellite repeats or accumulation of their related transcripts, and the actors at play, on centromere function and chromosomal stability is still only partial and may depend on organisms and cellular contexts. In the mouse, activated transcription of centromeric repeats and delocalization of CENP-A is associated with premature senescence in primary cells, whereas immortalized cells with impaired p53 checkpoint continue to cycle while accumulating mitotic errors and micronuclei, indicative of chromosomal instability (Hédouin et al. 2017). Thus, at least in the mouse, a functional p53 pathway is an important surveillance mechanism for centromere integrity, although the mechanisms remain unknown. Interestingly, p53-deficient mice ectopically expressing either human cenRNAs or murine pericenRNAs, were susceptible to tumor formation in mammary glands (Zhu et al. 2018). Hence, alterations to centromeric transcription may cooperate with oncogenic events or loss of tumor-suppressor function to promote oncogenesis.

Many questions remain unanswered as to the direct and reciprocal links between pathological hypomethylation of satellite repeats, their transcriptional derepression, the accumulation of DNA damage, and chromosomal instability. As cancer is a complex multifactorial disease, the current challenge is to dissect further and order these events.

7.4.3 Deregulation of Satellite Transcripts in the ICF Syndrome

A major breakthrough in the medical field came from the identification of inherited disorders of the epigenetic machinery, which provided interesting monogenic contexts and unsuspected players in a number of biological processes (Velasco and Francastel 2019). The first example of such developmental rare diseases was the Immunodeficiency with Centromeric instability and Facial anomalies (ICF) syndrome, a rare autosomal recessive immunological/neurological disorder with typical centromeric instability, including the presence of unusual multiradial chromosomal figures, decondensation, and rearrangement of (peri)centromeric regions (Ehrlich et al. 2006; Francastel and Magdinier 2019). At the molecular level, it is a remarkable case where these chromosomal alterations are caused by constitutive defects in DNA methylation, especially visible at heterochromatin blocks in juxtacentromeric position regions of chromosomes 1, 9, and 16 (Satellites type II and III) in all patients (Ehrlich et al. 2006). In a subset of patients, additional hypomethylation of centromeric α-satellite repeats suggested the genetic heterogeneity of the disease (Jiang et al. 2005; Toubiana et al. 2018).

Studies of the etiology of this rare disease have been instrumental in the identification of essential factors for the methylated state of (peri)centromeric repeats and the maintenance of their integrity. Hypomorphic mutations in the DNMT3B gene were the first identified genetic cause in about half of the patients (Xu et al. 1999), concomitantly implicated in de novo DNA methylation at centromeres in the mouse (Okano et al. 1999). The disease gained renewed interest when, in the reminder of patients, under the same diagnosis but with additional DNA methylation loss at centromeric repeats, exome sequencing identified mutations in factors with very few known functions and strikingly devoid of DNA methyltransferase (DNMT) activity (de Greef et al. 2011; Thijssen et al. 2015). These factors are transcription factors (ZBTB24, CDCA7) or chromatin remodeler of the SWI/SNF2 family (HELLS), the latter having already been shown to play a role in DNA methylation at murine centromeric repeats (Zhu et al. 2006). RNA interference performed in somatic cells, where DNA methylation profiles are already established, further demonstrated the requirement for ZBTB24, CDCA7, and HELLS in DNA methylation maintenance at murine centromeric repeats (Thijssen et al. 2015). These findings represented a major breakthrough in the knowledge of the determinants of DNA methylation at centromeric repeats. Yet, they raised again the question of the mechanisms that link hypomethylation of centromeric repeats to centromere loss of integrity, and the question of the contribution of non-DNMT ICF factors in DNA methylation and integrity of centromeres.

Independently of a putative role in DNA methylation pathways, a role for CDCA7 and HELLS in DNA repair pathways has been recently reported, reinforcing the idea of a link between DNA damage and centromere integrity (Burrage et al. 2012; Unoki et al. 2019). Notably, human embryonic kidney HEK-293T cells engineered to reproduce CDCA7 and HELLS mutations found in ICF patients exhibited a compromised nonhomologous end joining (NHEJ) DNA repair pathway (Unoki et al. 2019). Consistent with an aberrant accumulation of defects in DNA repair at centromeres, these cells accumulated micronuclei and suffered from abnormal chromosome segregation, while satellite repeats retained their methylated status. These engineered ICF cells, as well as cells from ICF patients, also exhibited increased transcription of (peri)centromeric repeats (Unoki et al. 2020). Given that genotoxic stress promotes a rapid transcriptional activation at centromeres (Hédouin et al. 2017), and along with the findings that satellite transcripts accumulate in breast cancer cells deficient for the DNA repair factor BRCA1 (Zhu et al. 2011), these data, therefore, suggested that DNA damage may trigger transcriptional activation at satellite repeats. An alternative, or concomitant, the mechanism could be that factors like CDCA7, HELLS, or BRCA1 may protect transcribed centromeric repeats from the accumulation of deleterious DNA:RNA hybrids (R-loops), just like BRCA1 does at transcriptional termination pause sites of actively transcribed genes (Hatchi et al. 2015). R-loops are dynamic and abundant structures that have been involved in a variety of physiological processes including chromosome segregation (Kabeche et al. 2018), whereas their unscheduled accumulation is also a source of DNA damage and genome instability (Costantino and Koshland 2018; Mishra et al. 2021). R-loops have been observed to accumulate at (peri)centromeres in engineered ICF cells and cells from ICF patients (Unoki et al. 2020). However, whether the transcriptional derepression and subsequent R-loop formation arise directly through DNA methylation loss or loss of function of ICF factors acting as "guardians" or transcriptional repressors of (peri)centromeric repeats, remains to be determined.

Like in cancer cells, transcriptional derepression or accumulation of the related satellite transcripts may represent intermediate steps between pathological hypomethylation of satellite repeats and chromosomal instability (Fig. 7.2). Importantly, the ICF syndrome leads to the premature death of the patients in early childhood from repeated infections, and despite a few reported cases where patients developed cancer, it is not clear whether pathological hypomethylation of centromeric repeats would favor later complications and further emergence of cancer. It

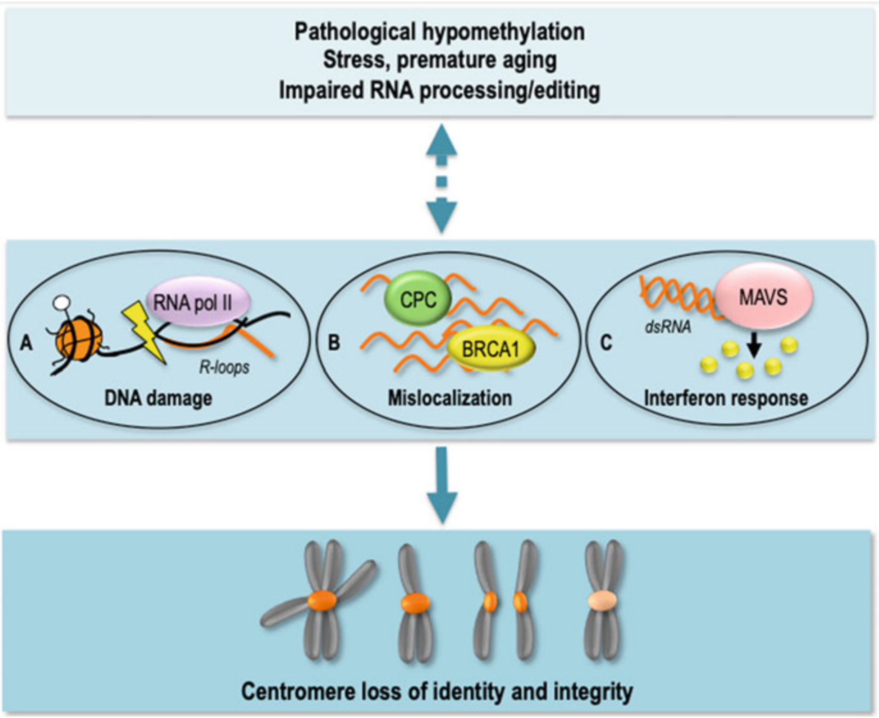


Fig. 7.2 Increased centromere transcription/transcripts levels: missing links between physiopathological DNA hypomethylation/DNA damage at centromeres and loss of centromere identity and function. Pathological hypomethylation, DNA damage at centromeres, or potentially impaired RNA processing or editing could promote: (**a**) unscheduled activated transcription of centromeric repeats, which in turn would lead to the formation of genotoxic R-loops; (**b**) aberrant accumulation of cenRNAs and trapping of kinetochore and DNA repair proteins away from centromeres; (**c**) the formation of double-stranded cenRNAs, known to trigger inflammatory responses. Although direct links between all these events remain to be formally dissected, the deregulation of centromeric transcription/transcripts ultimately lead to loss of centromere identity and integrity, as exemplified by multiradial chromosome figures, loss of sister chromatid cohesion, or recombination events between satellite repeats

also pointed out again that the impact of higher levels of satellite repeats transcripts on cellular phenotypes depends on the context in which it occurs.

A possibility that has not been evoked yet is that loss of DNA methylation at (peri)centromeric repeats and its associated abnormal levels of satellite RNAs may trigger surveillance mechanisms, which *in fine* would activate an interferon inflammatory response (Rajshekar et al. 2018). This has been nicely shown in a Zebrafish ICF model where one of the earliest in vivo consequences of ZBTB24 loss of function is a progressive loss of DNA methylation at pericentromeric regions associated with the derepression of sense and antisense pericenRNAs. This in turn triggered an interferon-dependent immune response mediated by the Melanoma Differentiation-Associated gene 5 (MDA5) and Mitochondrial AntiViral Signaling

(MAVS) machinery, an antiviral surveillance mechanism that senses dsRNAs (Berke et al. 2013). Injection of sense and antisense pericenRNAs in Zebrafish embryos was also sufficient to stimulate the innate immunity (Rajshekar et al. 2018), implicating the accumulation of pericenRNAs as an important trigger of autoimmunity in a variety of diseases.

7.5 Conclusion

All of the data exposed in this chapter lend support to the essential nature of temporal control of the act of transcription through centromeric satellite repeats for the determination and correct functioning of this chromosomal region in most of the species studied so far. Transcription per se would facilitate the dynamic exchange of nucleosomes for the deposition of the key determinant of centromere identity, CENP-A, but would also favor a local concentration of the transcripts themselves for the timely recruitment of other centromere components. Yet, it is still unclear which molecular mechanisms and regulatory pathways are involved for a timely control in normal conditions, although we mentioned transcription factors acting in defined chromatin environments and RNA-based mechanisms for the regulation of the transcripts levels.

In turn, unscheduled transcription or aberrant levels of the transcripts have profound consequences for both centromere function and cell fate. We have seen that activated transcription of (peri)centromeric repeats or ectopic accumulation of the transcripts, i.e. elsewhere than at centromeres, under stress conditions is a mechanism adopted by many organisms to trigger rapid cellular responses for cells to recover from stress. This type of response is possible through the trapping of various regulatory factors away from their site of action and impairment of their associated functions, to favor genome repair or remodeling of gene expression programs. In turn, unscheduled transcription or accumulation of the transcripts coincides with disease states. In that case, they are not seen as safeguard mechanisms, which implicitly infer collaborative effects with disease conditions like defective checkpoints, oncogenic events, or even an inflammatory environment that ultimately alter cellular phenotypes. Pathological hypomethylation of satellite repeats like in cancer or ICF syndrome is a good candidate for uncontrolled transcription of satellite repeats, although we have seen that it is not necessarily sufficient and that opportunistic tissue- or context-specific factors may come into play. This might explain why all cancer cells do not necessarily exhibit increased transcription of satellite repeats, and why ICF patients do not have widespread alterations in all their tissues. Alternatively, or in addition to, defective RNA-based surveillance mechanisms might also contribute to the abnormal elevation of the levels of satellite transcripts.

In sum, the use of a wide range of model organisms and artificial centromeres allowed to identify a large number of centromere and kinetochore proteins, to address the relevance of DNA sequences for centromere identity, and to tackle the functional relevance of centromeres transcription/transcripts for centromere identity and function. Studies of the etiology of complex or monogenic human diseases further identified key determinants for centromere integrity and function, among which we can cite factors with DNA repair or chromatin remodeling activities, many of which could not be suspected before their implication in centromeric instability diseases. Yet, our vision of the intricate contribution of all the actors and mechanisms mentioned throughout this chapter still remains fragmentary and will require the development of targeted approaches, many of which are still missing in mammalian systems.

Acknowledgments We are grateful to Ekaterina Boyarchuk for the critical reading of the manuscript and Florent Hubé for the help with figures.

We thank our tutelles University de Paris, Centre National de la Recherche Scientifique (CNRS) and Institut National de la Recherche Médicale (INSERM). Work in C.F.'s lab is supported by research grants from Agence Nationale pour la Recherche (grant ANR-19-CE12-0022), Ligue Nationale Contre le Cancer, Fondation ARC and Fondation Maladies Rares. P.M. is a recipient of a fellowship from the French Ministry of Research through the Ecole Universitaire Génétique et Epigénétique Nouvelle Ecole (EUR GENE).

References

Aird WC, Parvin JD, Sharp PA, Rosenberg RD (1994) The interaction of GATA-binding proteins and basal transcription factors with GATA box-containing core promoters A model of tissue-specific gene expression. J Biol Chem 269:883–889

Akiyoshi B, Gull K (2014) Discovery of unconventional kinetochores in kinetoplastids. Cell 156:1247–1258

Aldrup-Macdonald ME, Sullivan BA (2014) The past, present, and future of human centromere genomics. Genes (Basel) 5:33–50

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

Bergmann JH, Martins NMC, Larionov V, Masumoto H, Earnshaw WC (2012) HACking the centromere chromatin code: insights from human artificial chromosomes. Chromosom Res 20:505–519

Berke IC, Li Y, Modis Y (2013) Structural basis of innate immune recognition of viral RNA. Cell Microbiol 15:386–394

Bersani F, Lee E, Kharchenko PV, Xu AW, Liu M, Xega K, MacKenzie OC, Brannigan BW, Wittner BS, Jung H et al (2015) Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. Proc Natl Acad Sci USA 112:15148–15153

Bestor T, Laudano A, Mattaliano R, Ingram V (1988) Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. J Mol Biol 203:971–983

Biamonti G, Vourc'h C (2010) Nuclear stress bodies. Cold Spring Harb Perspect Biol 2:a000695

Blower MD (2016) Centromeric transcription regulates Aurora-B localization and activation. Cell Rep 15:1624–1633

Blower MD, Sullivan BA, Karpen GH (2002) Conserved organization of centromeric chromatin in flies and humans. Dev Cell 2:319–330

Bobkov GOM, Gilbert N, Heun P (2018) Centromere transcription allows CENP-A to transit from chromatin association to stable incorporation. J Cell Biol 217:1957–1972

Bobkov GOM, Huang A, van den Berg SJW, Mitra S, Anselm E, Lazou V, Schunter S, Feederle R, Imhof A, Lusser A et al (2020) Spt6 is a maintenance factor for centromeric CENP-A. Nat Commun 11:2919

Boo SH, Kim YK (2020) The emerging role of RNA modifications in the regulation of mRNA stability. Exp Mol Med 52:400–408

Bouzinba-Segard H, Guais A, Francastel C (2006) Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. Proc Natl Acad Sci 103:8709–8714

Bracewell R, Chatla K, Nalley MJ, Bachtrog D (2019) Dynamic turnover of centromeres drives karyotype evolution in Drosophila. eLife 8

Brown KE, Guest SS, Smale ST, Hahm K, Merkenschlager M, Fisher AG (1997) Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. Cell 91:845–854

Burrage J, Termanis A, Geissner A, Myant K, Gordon K, Stancheva I (2012) The SNF2 family ATPase LSH promotes phosphorylation of H2AX and efficient repair of DNA double-strand breaks in mammalian cells. J Cell Sci 125:5524–5534

Bury L, Moodie B, Ly J, McKay LS, Miga KH, Cheeseman IM (2020) Alpha-satellite RNA transcripts are repressed by centromere-nucleolus associations. eLife 9:e59770

Carone DM, Longo MS, Ferreri GC, Hall L, Harris M, Shook N, Bulazel KV, Carone BR, Obergfell C, O'Neill MJ et al (2009) A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. Chromosoma 118:113–125

Catania S, Pidoux AL, Allshire RC (2015) Sequence features and transcriptional stalling within centromere DNA promote establishment of CENP-A chromatin. PLoS Genet 11:e1004986

Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KHA, Wong LH (2012) Active transcription and essential role of RNA polymerase II at the centromere during mitosis. Proc Natl Acad Sci USA 109:1979–1984

Chang C-H, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, Chen C-C, Erceg J, Beliveau BJ, Wu C-T et al (2019) Islands of retroelements are major components of Drosophila centromeres. PLoS Biol 17:e3000241

Chen ES, Saitoh S, Yanagida M, Takahashi K (2003) A cell cycle-regulated GATA factor promotes centromeric localization of CENP-A in fission yeast. Mol Cell 11:175–187

Chen C-C, Dechassa ML, Bettini E, Ledoux MB, Belisario C, Heun P, Luger K, Mellone BG (2014) CAL1 is the Drosophila CENP-A assembly factor. J Cell Biol 204:313–329

Chen C-C, Bowers S, Lipinszki Z, Palladino J, Trusiak S, Bettini E, Rosin L, Przewloka MR, Glover DM, O'Neill RJ et al (2015) Establishment of Centromeric chromatin by the CENP-A assembly factor CAL1 requires FACT-mediated transcription. Dev Cell 34:73–84

Chen C-F, Pohl TJ, Chan A, Slocum JS, Zakian VA (2019) *Saccharomyces cerevisiae* centromere RNA is negatively regulated by Cbf1 and its unscheduled synthesis impacts CenH3 binding. Genetics 213:465–479

Choi ES, Strålfors A, Castillo AG, Durand-Dubief M, Ekwall K, Allshire RC (2011) Identification of noncoding transcripts from within CENP-A chromatin at fission yeast centromeres. J Biol Chem 286:23600–23607

Choo KH (1997) Centromere DNA dynamics: latent centromeres and neocentromere formation. Am J Hum Genet 61:1225–1233

Choo KHA (2001) Domain organization at the centromere and neocentromere. Dev Cell 1:165–177

Choy JS, Acuña R, Au W-C, Basrai MA (2011) A role for histone H4K16 hypoacetylation in *Saccharomyces cerevisiae* kinetochore function. Genetics 1:11–21

Christova R, Oelgeschläger T (2002) Association of human TFIID-promoter complexes with silenced mitotic chromatin in vivo. Nat Cell Biol 4:79–82

Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KHA, Wong LH (2009) LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. PLoS Genet 5:e1000354

Cobb BS, Morales-Alcelay S, Kleiger G, Brown KE, Fisher AG, Smale ST (2000) Targeting of Ikaros to pericentromeric heterochromatin by direct DNA binding. Genes Dev 14:2146–2160

Cohen AK, Huh TY, Helleiner CW (1973) Transcription of satellite DNA in mouse L-cells. Can J Biochem 51:529–532

Costantino L, Koshland D (2018) Genome-wide map of R-loop-induced damage reveals how a subset of R-loops contributes to genomic instability. Mol Cell 71:487–497e3

de Greef JC, Wang J, Balog J, den Dunnen JT, Frants RR, Straasheijm KR, Aytekin C, van der Burg M, Duprez L, Ferster A et al (2011) Mutations in ZBTB24 are associated with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2. Am J Hum Genet 88:796–804

Di Paolo A, Racca C, Calsou P, Larminat F (2014) Loss of BRCA1 impairs centromeric cohesion and triggers chromosomal instability. FASEB J 28:5250–5261

Diaz MO, Barsacchi-Pilone G, Mahon KA, Gall JG (1981) Transcripts from both strands of a satellite DNA occur on lampbrush chromosome loops of the newt Notophthalmus. Cell 24:649–659

Drinnenberg IA, de Young D, Henikoff S, Malik HS (2014) Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. eLife 3:e03676

Du Y, Topp CN, Dawe RK (2010) DNA binding of centromere protein C (CENPC) is stabilized by single-stranded RNA. PLoS Genet 6:e1000835

Dunleavy EM, Roche D, Tagami H, Lacoste N, Ray-Gallet D, Nakamura Y, Daigo Y, Nakatani Y, Almouzni-Pettinotti G (2009) HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. Cell 137:485–497

Dunleavy EM, Almouzni G, Karpen GH (2011) H33 is deposited at centromeres in S phase as a placeholder for newly assembled CENP-A in G1 phase. Nucleus 2:146–157

Earnshaw WC, Rothfield N (1985) Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma 91:313–321

Ehrlich M, Jackson K, Weemaes C (2006) Immunodeficiency, centromeric region instability, facial anomalies syndrome (ICF). Orphanet J Rare Dis 1:1

Eymery HB, El Atifi-Borel M, Fourel G, Berger F, Vitte A-L, Van den Broeck A, Brambilla E, Fournier A, Callanan M et al (2009a) A transcriptomic analysis of human centromeric and pericentric sequences in normal and tumor cells. Nucleic Acids Res 37:6340–6354

Eymery A, Callanan M, Vourc'h C (2009b) The secret message of heterochromatin: new insights into the mechanisms and function of centromeric and pericentric repeat sequence transcription. Int J Dev Biol 53:259–268

Eymery A, Souchier C, Vourc'h C, Jolly C (2010) Heat shock factor 1 binds to and transcribes satellite II and III sequences at several pericentromeric regions in heat-shocked cells. Exp Cell Res 316:1845–1855

Fernandez HR, Kavi HH, Xie W, Birchler JA (2005) Heterochromatin: on the ADAR radar? Curr Biol 15:R132–R134

Ferri F, Bouzinba-Segard H, Velasco G, Hubé F, Francastel C (2009) Non-coding murine centromeric transcripts associate with and potentiate Aurora B kinase. Nucleic Acids Res 37:5071–5080

Fisher AG, Merkenschlager M (2002) Gene silencing, cell fate and nuclear organisation. Curr Opin Genet Dev 12:193–197

Foltz DR, Jansen LE, Black BE, Bailey AO, Yates JR 3rd, Cleveland DW (2006) The human CENP-A centromeric nucleosome-associated complex. Nat Cell Biol 8:458–469

Foltz DR, Jansen LET, Bailey AO, Yates JR, Bassett EA, Wood S, Black BE, Cleveland DW (2009) Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. Cell 137:472–484

Francastel C, Magdinier F (2019) DNA methylation in satellite repeats disorders. Essays Biochem 63:757–771

Francastel C, Schübeler D, Martin DIK, Groudine M (2000) Nuclear compartmentalization and gene activity. Nat Rev Mol Cell Biol 1:137–143

Fukagawa T, Earnshaw WC (2014) The centromere: chromatin foundation for the kinetochore machinery. Dev Cell 30:496–508

Fukagawa T, Nogami M, Yoshikawa M, Ikeno M, Okazaki T, Takami Y, Nakayama T, Oshimura M (2004) Dicer is essential for formation of the heterochromatin structure in vertebrate cells. Nat Cell Biol 6:784–791

Furuyama S, Biggins S (2007) Centromere identity is specified by a single centromeric nucleosome in budding yeast. Proc Natl Acad Sci USA 104:14706–14711

Gambogi CW, Dawicki-McKenna JM, Logsdon GA, Black BE (2020) The unique kind of human artificial chromosome: bypassing the requirement for repetitive centromere DNA. Exp Cell Res 391:111978

Giulotto E, Raimondi E, Sullivan KF (2017) The unique DNA sequences underlying equine centromeres. Prog Mol Subcell Biol 56:337–354

Grenfell AW, Heald R, Strzelecka M (2016) Mitotic noncoding RNA processing promotes kinetochore and spindle assembly in Xenopus. J Cell Biol 214:133–141

Grewal SIS, Elgin SCR (2007) Transcription and RNA interference in the formation of heterochromatin. Nature 447:399–406

Gupta R, Somyajit K, Narita T, Maskey E, Stanlie A, Kremer M, Typas D, Lammers M, Mailand N, Nussenzweig A et al (2018) DNA repair network analysis reveals Shieldin as a key regulator of NHEJ and PARP inhibitor sensitivity. Cell 173:972–988e23

Hall LL, Byron M, Carone DM, Whitfield TW, Pouliot GP, Fischer A, Jones P, Lawrence JB (2017) Demethylated HSATII DNA and HSATII RNA foci sequester PRC1 and MeCP2 into cancer-specific nuclear bodies. Cell Rep 18:2943–2956

Harel J, Hanania N, Tapiero H, Harel L (1968) RNA replication by nuclear satellite DNA in different mouse cells. Biochem Biophys Res Commun 33:696–701

Hatchi E, Skourti-Stathaki K, Ventz S, Pinello L, Yen A, Kamieniarz-Gdula K, Dimitrov S, Pathania S, McKinney KM, Eaton ML et al (2015) BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. Mol Cell 57:636–647

Hédouin S, Grillo G, Ivkovic I, Velasco G, Francastel C (2017) CENP-A chromatin disassembly in stressed and senescent murine cells. Sci Rep 7:42520

Hoffmann S, Fachinetti D (2017) A time out for CENP-A. Mol Cell Oncol 4:e1293596

Houseley J, Kotovic K, El Hage A, Tollervey D (2007) Trf4 targets ncRNAs from telomeric and rDNA spacer regions and functions in rDNA copy number control. EMBO J 26:4996–5006

Huang C, Wang X, Liu X, Cao S, Shan G (2015) RNAi pathway participates in chromosome segregation in mammalian cells. Cell Discov 1:15029

Ideue T, Cho Y, Nishimura K, Tani T (2014) Involvement of satellite I noncoding RNA in regulation of chromosome segregation. Genes Cells 19:528–538

Ishikura S, Nakabayashi K, Nagai M, Tsunoda T, Shirasawa S (2020) ZFAT binds to centromeres to control noncoding RNA transcription through the KAT2B-H4K8ac-BRD4 axis. Nucleic Acids Res 48:10848–10866

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH (2018) Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol 36:321–323

Jiang YL, Rigolet M, Bourc'his D, Nigon F, Bokesoy I, Fryns JP, Hultén M, Jonveaux P, Maraschio P, Mégarbané A, Moncla A, Viegas-Péquignot E (2005) DNMT3B mutations and DNA methylation defect define two types of ICF syndrome. Hum Mutat 25:56–63

Jolly C, Metz A, Govin J, Vigneron M, Turner BM, Khochbin S, Vourc'h C (2004) Stress-induced transcription of satellite III repeats. J Cell Biol 164:25–33

Kabeche L, Nguyen HD, Buisson R, Zou L (2018) A mitosis-specific and R loop-driven ATR pathway promotes faithful chromosome segregation. Science 359:108–114

Kalitsis P, Griffiths B, Choo KHA (2006) Mouse telocentric sequences reveal a high rate of homogenization and possible role in Robertsonian translocation. Proc Natl Acad Sci USA 103:8786–8791

Kanellopoulou C, Muljo SA, Kung AL, Ganesan S, Drapkin R, Jenuwein T, Livingston DM, Rajewsky K (2005) Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev 19:489–501

Kapoor M, Montes de Oca Luna R, Liu G, Lozano G, Cummings C, Mancini M, Ouspenski I, Brinkley BR, May GS (1998) The cenpB gene is not essential in mice. Chromosoma 107:570–576

Karpen GH, Allshire RC (1997) The case for epigenetic effects on centromere identity and function. Trends Genet 13:489–496

Kawahara Y, Zinshteyn B, Chendrimada TP, Shiekhattar R, Nishikura K (2007) RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. EMBO Rep 8:763–769

Kishikawa T, Otsuka M, Yoshikawa T, Ohno M, Yamamoto K, Yamamoto N, Kotani A, Koike K (2016) Quantitation of circulating satellite RNAs in pancreatic cancer patients. JCI Insight 1: e86646

Kishikawa T, Otsuka M, Suzuki T, Seimiya T, Sekiba K, Ishibashi R, Tanaka E, Ohno M, Yamagami M, Koike K (2018) Satellite RNA increases DNA damage and accelerates tumor formation in mouse models of pancreatic cancer. Mol Cancer Res 16:1255–1262

Lechner J, Ortiz J (1996) The *Saccharomyces cerevisiae* kinetochore. FEBS Lett 389:70–74

Lehnertz B, Ueda Y, Derijck AA, Braunschweig U, Perez-Burgos L, Kubicek S, Chen T, Li E, Jenuwein T, Peters AHFM (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol 13:1192–1200. https://doi.org/10.1016/s0960-9822(03)00432-9

Levanon EY, Hallegger M, Kinar Y, Shemesh R, Djinovic-Carugo K, Rechavi G, Jantsch MF, Eisenberg E (2005) Evolutionarily conserved human targets of adenosine to inosine RNA editing. Nucleic Acids Res 33:1162–1168

Li Y-X, Kirby ML (2003) Coordinated and conserved expression of alphoid repeat and alphoid repeat-tagged coding sequences. Dev Dyn 228:72–81

Li F, Sonbuchner L, Kyes SA, Epp C, Deitsch KW (2008) Nuclear non-coding RNAs are transcribed from the centromeres of plasmodium falciparum and are associated with centromeric chromatin. J Biol Chem 283:5692–5698

Ling YH, Yuen KWY (2019) Centromeric non-coding RNA as a hidden epigenetic factor of the point centromere. Curr Genet 65:1165–1171

Lippman Z, Martienssen R (2004) The role of RNA interference in heterochromatic silencing. Nature 431:364–370

Liu X, Wu B, Szary J, Kofoed EM, Schaufele F (2007) Functional sequestration of transcription factor activity by repetitive DNA. J Biol Chem 282:20868–20876

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 178:624–639e19

Lu J, Gilbert DM (2008) Cell cycle regulated transcription of heterochromatin in mammals vs fission yeast: functional conservation or coincidence? Cell Cycle 7:1907–1910

Malik HS, Henikoff S (2009) Major evolutionary transitions in centromere complexity. Cell 138:1067–1082

Mallm J-P, Rippe K (2015) Aurora kinase B regulates telomerase activity via a centromeric RNA in stem cells. Cell Rep 11:1667–1678

Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J 24:800–812

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T (1989) A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol 109:1963–1973

Masumoto H, Nakano M, Ohzeki J-I (2004) The role of CENP-B and alpha-satellite DNA: de novo assembly and epigenetic maintenance of human centromeres. Chromosom Res 12:543–556

McNulty SM, Sullivan LL, Sullivan BA (2017) Human centromeres produce chromosome-specific and Array-specific alpha satellite transcripts that are complexed with CENP-A and CENP-C. Dev Cell 42:226–240e6

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby J, Sebra R, Peluso P, Eid J, Rank D et al (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10

Meyne J, Ratliff RL, MoYzIs RK (1989) Conservation of the human telomere sequence (TTAGGG) among vertebrates. Proc Natl Acad Sci USA 5

Miga KH (2019) Centromeric satellite DNAs: hidden sequence variation in the human population. Genes (Basel) 10:352

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ (2014) Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res 24:697–707

Mishra PK, Chakraborty A, Yeh E, Feng W, Bloom KS, Basrai MA (2021) R-loops at centromeric chromatin contribute to defects in kinetochore integrity and chromosomal instability in budding yeast. MBoC 32:74–89

Mitchell AR, Jeppesen P, Nicol L, Morrison H, Kipling D (1996) Epigenetic control of mammalian centromere protein binding: does DNA methylation have a role? J Cell Sci 109:2199–2206

Müller S, Almouzni G (2017) Chromatin dynamics during the cell cycle at centromeres. Nat Rev Genet 18:192–208

Muller H, Gil J, Drinnenberg IA (2019) The impact of centromeres on spatial genome architecture. Trends Genet 35:565–578

Murchison EP, Partridge JF, Tam OH, Cheloufi S, Hannon GJ (2005) Characterization of dicer-deficient murine embryonic stem cells. Proc Natl Acad Sci USA 102:12135–12140

Nachtergaele S, He C (2017) The emerging biology of RNA post-transcriptional modifications. RNA Biol 14:156–163

Nakano M, Okamoto Y, Ohzeki J, Masumoto H (2003) Epigenetic assembly of centromeric chromatin at ectopic alpha-satellite sites on human chromosomes. J Cell Sci 116:4021–4034

Navarro-Mendoza MI, Pérez-Arques C, Panchal S, Nicolás FE, Mondo SJ, Ganguly P, Pangilinan J, Grigoriev IV, Heitman J, Sanyal K et al (2019) Early diverging fungus Mucor circinelloides lacks centromeric histone CENP-A and displays a mosaic of point and regional centromeres. Curr Biol 29:3791–3802e6

Ohkuni K, Kitagawa K (2011) Endogenous transcription at the centromere facilitates centromere activity in budding yeast. Curr Biol 21:1695–1703

Ohkuni K, Kitagawa K (2012) Role of transcription at centromeres in budding yeast. Transcription 3:193–197

Ohzeki J, Nakano M, Okada T, Masumoto H (2002) CENP-B box is required for de novo centromere chromatin assembly on human alphoid. DNA J Cell Biol 159:765–775

Okada M, Okawa K, Isobe T, Fukagawa T (2009) CENP-H-containing complex facilitates centromere deposition of CENP-A in cooperation with FACT and CHD1. Mol Biol Cell 20:3986–3995

Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99:247–257

Pageau GJ, Lawrence JB (2006) BRCA1 foci in normal S-phase nuclei are linked to interphase centromeres and replication of pericentric heterochromatin. J Cell Biol 175:693–701

Palmer DK, O’Day K, Trong HL, Charbonneau H, Margolis RL (1991) Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. Proc Natl Acad Sci USA 88:3734–3738

Pecinka A, Dinh HQ, Baubec T, Rosa M, Lettner N, Mittelsten Scheid O (2010) Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. Plant Cell 22:3118–3129

Peters AH, O'Carroll D, Scherthan H, Mechtler K, Sauer S, Schöfer C, Weipoltshammer K, Pagani M, Lachner M, Kohlmaier A et al (2001) Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. Cell 107:323–337

Pezer Z, Ugarković D (2008) RNA Pol II promotes transcription of centromeric satellite DNA in beetles. PLoS One 3:e1594

Pezer Z, Ugarkovic D (2012) Satellite DNA-associated siRNAs as mediators of heat shock response in insects. RNA Biol 9:587–595

Pidoux AL, Allshire RC (2005) The role of heterochromatin in centromere function. Philos Trans R Soc Lond Ser B Biol Sci 360:569–579

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

Polizzi C, Clarke L (1991) The chromatin structure of centromeres from fission yeast: differentiation of the central core that correlates with function. J Cell Biol 112:191–201

Probst AV, Okamoto I, Casanova M, El Marjou F, Le Baccon P, Almouzni G (2010) A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. Dev Cell 19:625–638

Quénet D, Dalal Y (2014) A long non-coding RNA is required for targeting centromeric protein A to the human centromere. eLife 3:e03254

Rajshekar S, Yao J, Arnold PK, Payne SG, Zhang Y, Bowman TV, Schmitz RJ, Edwards JR, Goll M (2018) Pericentromeric hypomethylation elicits an interferon response in an animal model of ICF syndrome. eLife 7

Reinhart BJ, Bartel DP (2002) Small RNAs correspond to centromere heterochromatic repeats. Science 297:1831

Rieder CL (1978) Effect of elevated temperatures on spindle microtubules and chromosome movements in cultured newt lung cells. Cytobios 18:201–233

Rizzi N, Denegri M, Chiodi I, Corioni M, Valgardsdottir R, Cobianchi F, Riva S, Biamonti G (2004) Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. Mol Biol Cell 15:543–551

Rošić S, Köhler F, Erhardt S (2014) Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. J Cell Biol 207:335–349

Ross JP, Rand KN, Molloy PL (2010) Hypomethylation of repeated DNA sequences in cancer. Epigenomics 2:245–269

Rudert F, Bronner S, Garnier JM, Dollé P (1995) Transcripts from opposite strands of gamma satellite DNA are differentially expressed during mouse development. Mamm Genome 6:76–83

Sadeghi L, Siggens L, Svensson JP, Ekwall K (2014) Centromeric histone H2B monoubiquitination promotes noncoding transcription and chromatin integrity. Nat Struct Mol Biol 21:236–243

Saksouk N, Barth TK, Ziegler-Birling C, Olova N, Nowak A, Rey E, Mateos-Langerak J, Urbach S, Reik W, Torres-Padilla M-E et al (2014) Redundant mechanisms to form silent chromatin at pericentromeric regions rely on BEND3 and DNA methylation. Mol Cell 56:580–594

Saksouk N, Simboeck E, Déjardin J (2015) Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin 8:3

Savage KI, Harkin DP (2015) BRCA1, a "complex" protein involved in the maintenance of genomic stability. FEBS J 282:630–646

Scadden AD, Smith CW (2001) RNAi is antagonized by A-->I hyper-editing. EMBO Rep 2:1107–1111

Scelfo A, Fachinetti D (2019) Keeping the centromere under control: a promising role for DNA methylation. Cell 8

Schueler MG, Sullivan BA (2006) Structural and functional dynamics of human centromeric chromatin. Annu Rev Genomics Hum Genet 7:301–313

Scott KC, Sullivan BA (2014) Neocentromeres: a place for everything and everything in its place. Trends Genet 30:66–74

She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark R, Graves T, Gulden CL, Alkan C et al (2004) The structure and evolution of centromeric transition regions within the human genome. Nature 430:857–864

Shestakova EA, Mansuroglu Z, Mokrani H, Ghinea N, Bonnefoy E (2004) Transcription factor YY1 associates with pericentromeric gamma-satellite DNA in cycling but not in quiescent (G0) cells. Nucleic Acids Res 32:4390–4399

Sullivan BA, Karpen GH (2004) Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nat Struct Mol Biol 11:1076–1083

Sullivan LL, Sullivan BA (2020) Genomic and functional variation of human centromeres. Exp Cell Res 389:111896

Sun X, Le HD, Wahlstrom JM, Karpen GH (2003) Sequence analysis of a functional Drosophila centromere. Genome Res 13:182–194

Tanaka Y, Kurumizaka H, Yokoyama S (2005) CpG methylation of the CENP-B box reduces human CENP-B binding. FEBS J 272:282–289

Tasselli L, Xi Y, Zheng W, Tennen RI, Odrowaz Z, Simeoni F, Li W, Chua KF (2016) SIRT6 deacetylates H3K18ac at pericentric chromatin to prevent mitotic errors and cellular senescence. Nat Struct Mol Biol 23:434–440

Terranova R, Sauer S, Merkenschlager M, Fisher AG (2005) The reorganisation of constitutive heterochromatin in differentiating muscle requires HDAC activity. Exp Cell Res 310:344–356

Thijssen PE, Ito Y, Grillo G, Wang J, Velasco G, Nitta H, Unoki M, Yoshihara M, Suyama M, Sun Y et al (2015) Mutations in CDCA7 and HELLS cause immunodeficiency-centromeric instability-facial anomalies syndrome. Nat Commun 6:7870

Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G, Deshpande V, Iafrate AJ, Letovsky S et al (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science 331:593–596

Tittel-Elmer M, Bucher E, Broger L, Mathieu O, Paszkowski J, Vaillant I (2010) Stress-induced activation of heterochromatic transcription. PLoS Genet 6:e1001175

Topp CN, Zhong CX, Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. Proc Natl Acad Sci USA 101:15986–15991

Toubiana S, Velasco G, Chityat A, Kaindl AM, Hershtig N, Tzur-Gilat A, Francastel C, Selig S (2018) Subtelomeric methylation distinguishes between subtypes of immunodeficiency, centromeric instability and facial anomalies syndrome. Hum Mol Genet 27:3568–3581

Unoki M, Funabiki H, Velasco G, Francastel C, Sasaki H (2019) CDCA7 and HELLS mutations undermine nonhomologous end joining in centromeric instability syndrome. J Clin Invest 129:78–92

Unoki M, Sharif J, Saito Y, Velasco G, Francastel C, Koseki H, Sasaki H (2020) CDCA7 and HELLS suppress DNA:RNA hybrid-associated DNA damage at pericentromeric repeats. Sci Rep 10:17865

Valgardsdottir R, Chiodi I, Giordano M, Rossi A, Bazzini S, Ghigna C, Riva S, Biamonti G (2008) Transcription of satellite III non-coding RNAs is a general stress response in human cells. Nucleic Acids Res 36:423–434

Varley JM, Macgregor HC, Erba HP (1980) Satellite DNA is transcribed on lampbrush chromosomes. Nature 283:686–688

Velasco G, Francastel C (2019) Genetics meets DNA methylation in rare diseases. Clin Genet 95:210–220

Vissel B, Nagy A, Choo KH (1992) A satellite III sequence shared by human chromosomes. 13, 14, and 21 that is contiguous with alpha satellite DNA. Cytogenet Cell Genet 61:81–86

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SIS, Martienssen RA (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science 297:1833–1837

Vourc'h C, Biamonti G (2011) Transcription of satellite DNAs in mammals. Prog Mol Subcell Biol 51:95–118

Wako T, Houben A, Furushima-Shimogawara R, Belyaev ND, Fukui K (2003) Centromere-specific acetylation of histone H4 in barley detected through three-dimensional microscopy. Plant Mol Biol 51:533–541

Wijchers PJ, Geeven G, Eyres M, Bergsma AJ, Janssen M, Verstegen M, Zhu Y, Schell Y, Vermeulen C, de Wit E et al (2015) Characterization and dynamics of pericentromere-associated domains in mice. Genome Res 25:958–969

Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL, Hannan R, Saffery R, Shaw ML, Williams E et al (2007) Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res 17:1146–1160

Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, Hulten M, Qu X, Russo JJ, Viegas-Péquignot E (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature 402:187–191

Zhu H, Geiman TM, Xi S, Jiang Q, Schmidtmann A, Chen T, Li E, Muegge K (2006) Lsh is involved in *de novo* methylation of DNA. EMBO J 25:335–345

Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, Nederlof PM, Gage FH, Verma IM (2011) BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. Nature 477:179–184

Zhu Q, Hoong N, Aslanian A, Hara T, Benner C, Heinz S, Miga KH, Ke E, Verma S, Soroczynski J et al (2018) Heterochromatin-encoded satellite RNAs induce breast Cancer. Mol Cell 70:842–853e7

Chapter 8
Global Repeat Map (GRM): Advantageous Method for Discovery of Largest Higher-Order Repeats (HORs) in Neuroblastoma Breakpoint Family (NBPF) Genes, in Hornerin Exon and in Chromosome 21 Centromere

Vladimir Paar, Ines Vlahović, Marija Rosandić, and Matko Glunčić

Abstract Here we present three interesting novel human Higher-Order Repeats (HORs) discovered using the HOR-searching method with GRM algorithm: (a) The novel Neuroblastoma Breakpoint Family gene (NBPF) 3mer HOR, discovered applying GRM algorithm to human chromosome 1 (Paar et al., Mol Biol Evol 28:1877–1892, 2011). NBPF 3mer HOR is based on previously known ~1.6 kb NBPF primary repeat monomers (known as DUF1220 domain) in human chromosome 1, but the NBPF HOR was not known before its discovery by using GRM. It should be stressed that the NBPF HOR presents a unique human-specific pattern, distinguishing human from nonhuman primates. (b) The novel quartic HOR (2mer ⊃ 2mer ⊃ 9mer) discovered using the GRM algorithm for analysis of hornerin genes in human chromosome 1 (Paar et al., Mol Biol Evol 28:1877–1892, 2011). This quartic HOR is based on 39 bp hornerin primary repeat monomer in human chromosome 1. To our knowledge, this is the first known case of quartic HOR, with four levels of hierarchy of HOR organization. (c) The novel 33mer alpha satellite HOR in human chromosome 21, discovered using the GRM

V. Paar (✉)
Croatian Academy of Sciences and Arts, Zagreb, Croatia

Faculty of Science, University of Zagreb, Zagreb, Croatia
e-mail: vpaar@hazu.hr

I. Vlahović
Algebra University College, Zagreb, Croatia

M. Rosandić
Croatian Academy of Sciences and Arts, Zagreb, Croatia

University Hospital Centre Zagreb, Zagreb, Croatia

M. Glunčić
Faculty of Science, University of Zagreb, Zagreb, Croatia

Đ. Ugarković (ed.), *Satellite DNAs in Physiology and Evolution*, Progress in Molecular and Subcellular Biology 60,
https://doi.org/10.1007/978-3-030-74889-0_8

algorithm (Glunčić et al., Sci Rep 9:12629, 2019). This 33mer HOR in the smallest human chromosome is the largest alpha satellite HOR copy among all 22 somatic human chromosomes. Moreover, the same 33mer HOR is present in the hg38 human genome assembly of four human chromosomes: 21, 22, 13, and 14. We point out that the DUF1220 encoding genomic structures in NBPF genes in human chromosome 1, recently studied and related to the brain evolution and pathologies and cognitive aptitude, can be considered in the framework of the general concept of HORs, already extensively studied in genomics, especially in the centromeric region.

Keywords HOR · NBPF · Hornerin · Human genome · Neanderthal genome · Chromosome 21

8.1 Introduction

The concerted evolution of tandemly repeated DNA families involves many genetic turnover mechanisms and insight into these processes can be obtained by investigation of repeat structure and organization (Brown et al. 1971; Southern 1975; Smith 1976; Dover 1982; Willard 1985, 1991; Dover 1986; Willard and Waye 1987; Choo et al. 1991; Charlesworth et al. 1994; Warburton and Willard 1996; Alexandrov et al. 2001; Rudd et al. 2006; Aldrup-Macdonald and Sullivan 2014; Ruiz-Ruano et al. 2016; Sullivan et al. 2017; Jain et al. 2018; Miga et al. 2020). The alpha satellite DNA, extensively investigated at the human and other centromeres of human and other primates, can be considered as a paradigm for studies of concerted evolution in tandemly repeated DNA families (Willard and Waye 1987; Willard 1991; Choo et al. 1991; Warburton and Willard 1996; Alexandrov et al. 2001; Garrido-Ramos 2017; Sullivan et al. 2017). The primary repeat of alpha satellite DNA is based on ~20–40% diverged monomers of approximately 171 bp. Most alpha satellite monomers are organized into higher-order repeats (HORs) in which monomers (alpha satellite repeat units) are reiterated as a single repeat structure with high sequence identity (more than 95%) (Willard and Waye 1987; Willard 1991; Choo et al. 1991; Warburton and Willard 1996; Alexandrov et al. 2001). The impressive recent progress of sequencing technology (Aldrup-Macdonald and Sullivan 2014; Miga 2015, 2017; Ruiz-Ruano et al. 2016; Turner et al. 2018; Jain et al. 2015, 2018; Lower et al. 2018; Uralsky et al. 2019; Miga et al. 2020; Logsdon et al. 2020) gives a new impetus for HOR studies.

While HORs were so far largely investigated in the centromeric region, here we turn more attention to some cases of HORs in genes. The most intriguing case could be the 3mer HORs in NBPF (neuroblastoma breakpoint family) genes in human chromosome 1, having an important role in human brain evolution and function. They contain a repetitive structure with rather divergent (~20%) repeat units of ~1.6 kb (Vinogradova et al. 2002; Fortna et al. 2004; Vandepoele et al. 2005; Popesco et al. 2006; Dumas et al. 2007; Dumas and Sikela 2009). These repeats encode the protein domain *of unknown function*, DUF1220; these DNA repeats are called DUF1220 repeats or DUF1220 domains, with dramatically increased copy number in the human genome (Vandepoele et al. 2005; Popesco et al. 2006). The DUF1220 copy number was correlated with brain size, cortical neuron number, brain pathologies (autism, schizophrenia, microcephaly, macrocephaly, and neuroblastoma), IQ scores, cognitive aptitude, and evolution (Vandepoele et al. 2005, 2008; Popesco et al. 2006; Andries et al. 2012; Dumas et al. 2012; Davis et al. 2014; Keeney et al. 2014; Quick et al. 2015; Astling et al. 2017; Mitchell and Silver 2018; Fiddes et al. 2019; Heft et al. 2020).

Using the novel robust HOR searching algorithm GRM (Paar et al. 2011), convenient for detecting and analyzing long HOR units, the novel ~4.8 kb NBPF 3mer HORs were discovered (Paar et al. 2011) based on ~0.6 kb primary repeats in human chromosome 1. In general, GRM identifies simultaneously both HORs and their primary repeat monomers in a given sequence without the need for any prior knowledge on primary repeats and HORs. It turns out automatically that these three constituent monomer types coincide with DUF1220 repeats. As pointed out (Andries et al. 2012), before the discovery of NBPF 3mer HOR (Paar et al. 2011), it was not realized that the DUF1220 repeats are of three types, forming a remarkable 3mer HOR organization embedded within the NBPF genes. In fact, the ~4.8 kb NBPF 3mer HOR is a classical HOR pattern.

On the basis of divergence among DUF1220 repeats, this pattern was expressed in the DUF1220 terminology, with the use of the novel name HLS DUF1220 triplet (O'Bleness et al. 2012, 2014), but it was not noted that it corresponds in classical HOR terminology to the previously identified 3mer HOR. Bioinformatically, the HOR-searching method is simpler than the monomer (DUF1220) searching method because of a much smaller divergence between HOR copies than between neighboring DUF1220 domains.

Comparing the NBPF 3mer HORs identified by GRM in human, Neanderthal, and chimpanzee genomes interesting results on a possible evolutionary role of these HORs are emerging.

The second case of pronounced novel intra-gene HOR discovered by using the GRM algorithm is the ~1.41 kb quartic HOR fully embedded within exon in the human hornerin gene. This quartic HOR is characterized by the three-level-hierarchy HOR organization (Paar et al. 2011), based on the 39 bp primary repeat monomer. The fourth case of the long HOR unit presented here is the ~5.6 kb 33mer human alpha satellite HOR (Glunčić et al. 2019) in chromosome 21, based on the ~171 bp alpha satellite primary repeat. Interestingly, this longest alpha satellite HOR unit among somatic chromosomes is located in the smallest of all somatic chromosomes.

8.2 HORs and GRM

8.2.1 Higher-Order Repeats (HORs)

HORs have been extensively studied in human and nonhuman primate chromosomes, with alpha satellite primary repeat units of ~171 bp (Manuelidis 1978;

Willard 1985; Warburton and Willard 1996; Alexandrov et al. 2001; Warburton et al. 2008). The best-known prototypes of HORs are alpha satellite arrays, located in the centromeric region of all human chromosomes (Wu and Manuelidis 1980; Willard 1985; Jorgensen et al. 1986; Waye and Willard 1987; Willard and Waye 1987; Tyler-Smith and Brown 1987; Warburton and Willard 1996; Choo 1997; Alexandrov et al. 2001; Rudd and Willard 2004; Jurka et al. 2005; Rosandić et al. 2003; Paar et al. 2005; Miga 2017; Sullivan et al. 2017; Aldrup-Macdonald and Sullivan 2014; Lower et al. 2018; Uralsky et al. 2019; Glunčić et al. 2019). The term *satellite DNA* defines highly repetitive DNA sequences organized in tandem arrays (Pech et al. 1979; Singer 1982; Garrido-Ramos 2017). Alpha satellite arrays consist of primary repeat units, diverged monomers of approximately 171 bp in length, tandemly arranged in a head-to-tail fashion. Individual alpha satellite monomers diverge in sequence from each other by 20–40%. Some stretches of alpha satellites are hierarchically organized into HORs, secondary repeat units with highly convergent HOR copies (types of monomers based on sequence similarity, the divergence between monomers of the same type less than 5%, in some cases even below 1%). A sequence of a certain number of diverging monomers, forming a secondary repeat unit containing *n* monomers (*n*mer HOR unit), is tandemly repeated, with a much smaller divergence between HOR copies than the divergence between monomers within each HOR copy (Warburton and Willard 1996). In the *n*mer HOR array, the HOR copies containing *n* monomers are referred to as canonical *n*mer HOR copies, while the corresponding HOR copies missing one or more of *n* monomers, or having more than *n* monomers are referred to as noncanonical.

An explanation for generating HORs involves unequal crossing over between misaligned HOR units aligned on the register of homologous monomers. Unequal crossing over, restricted to tandem sequences, explains the generation and local homogenization of HOR units and accounts for large size variation among HORs on homologous chromosomes (Southern 1975; Smith 1976; Willard and Waye 1987; Warburton and Willard 1996; Schueler et al. 2001; Alkan et al. 2004; Rudd et al. 2006). HORs are in particular interesting since they are due to more recent evolution and enable a rapid evolutionary process.

Alpha satellite HORs in human and nonhuman primates have been first identified by hybridization (Willard 1985; Waye and Willard 1987; Wolfe et al. 1985; Jorgensen et al. 1986; Tyler-Smith and Brown 1987; Willard and Waye 1987; Choo et al. 1991; Alexandrov et al. 1991; Ge et al. 1992; Greig et al. 1993; Mashkova et al. 1994; Warburton and Willard 1996) and later by bioinformatics tools, as for example TRF (Benson 1999), BLAST (Altschul et al. 1990), etc., applied to genomic sequences. Willard and coworkers have pointed out that the alpha satellite DNA can be considered as a paradigm for addressing specific questions about the processes of concerted evolution in tandemly repeated DNA families (Willard 1991; Willard and Waye 1987).

HORs have been also identified for a number of other types of repeat monomers, even fully embedded in genes, or even within a single exon (for example, Paar et al. 2011); and surprisingly, even in such evolutionary distant species like insects (Vlahović et al. 2017). It should be stressed that the repeat elements in the genome have been associated with a regulatory role in eukaryotic organisms (King and Wilson 1975; Pennacchio and Rubin 2001; Ugarković 2005; Haygood et al. 2010; Noonan and McCallion 2010; Pezer et al. 2012).

8.2.2 HOR-Searching and Monomer-Searching Methods

The simplest way to identify HORs in a given genomic sequence is to use directly a HOR-searching method: to identify directly the HOR copy repeats in a given sequence. However, with an increase in HOR copy length, the efficacy of computational HOR searching tools decreases. Different computational algorithms are available to identify large tandem arrays. Lower et al. (2018) include the following software for assessing satellite DNA: TRF (Benson 1999), alfa-CENTAURI (Sevim et al. 2016), GRM (Glunčić and Paar 2013), RepeatExplorer (Novak et al. 2013), TAREAN (Novak et al. 2017), RepeatMasker (Smit et al. 2015), Spectral Repeat Finder (Sharma et al. 2004), etc., and for extended analysis BLAST (Altschul et al. 1990). In Ref. Lower et al. (2018), the purpose *HOR discovery* is associated with two novel methods: alfa-CENTAURI (Sevim et al. 2016) and GRM (Glunčić and Paar 2013).

GRM algorithm is a robust method, convenient for the identification and study of long HOR copies, as well as with deviations from regular HOR patterns. The main advantage of HOR-searching methods is due to the characteristic small divergence between neighboring HOR copies. After identification of HOR copies, the constituent monomers are deduced, and inter-monomeric divergence is determined. Simultaneously, GRM identifies all types of pronounced HORs present in a given genomic sequence, without the need for any prior knowledge on constituent monomers.

The other way to identify HORs in a given sequence is the monomer-searching method (monomer denotes a primary repeat sequence). This method starts with the identification of monomers in a given sequence, but the divergence between neighboring monomers can be sizable, much larger than divergence among HOR copies. Once diverged repeat monomers are identified, in the next step a search for low divergence among some equidistant monomers enables a posteriori identification of HOR copies.

Concluding, the HOR-searching methods identify in the first step the repeating HOR copies, with low mutual divergence (less than 5%). In the second step, the monomers that are constituents of HOR copies are deduced from HOR copies identified in the first step. The monomer-searching methods identify repeating monomers in the first step, which have sizable mutual divergence (~20–40%). This can have an impact on the accuracy or resolution of the method. In the second step, HOR copies are obtained by combining monomers resulting from the first step. Because of this divergence, the HOR-searching methods can have a computational advantage for the identification and study of long and/or distorted HOR copies.

8.2.3 *GRM Algorithm: A Robust Tool for Identification of Large Repeats and Higher-Order Repeats in a Given Genomic Sequence*

GRM is a novel efficient and robust method to identify and study large repeats, especially HORs, in a given DNA sequence. The GRM algorithm (Paar et al. 2011; Glunčić and Paar 2013; Glunčić et al. 2019; Vlahović et al. 2020) is an extension of KSA (Key String Algorithm) (Rosandić et al. 2003, 2006; Paar et al. 2005, 2007) and ColorHor algorithm (Paar et al. 2005).

For long DNA sequences, the noise in detecting repeats increases with the increasing length of the HOR repeat unit, which can mask some peaks corresponding to HOR copies. This background noise is significantly reduced in the GRM algorithm. The novelty of the GRM approach is a direct mapping of symbolic DNA sequence into the frequency domain using a complete *K*-string ensemble instead of statistically adjusted individual *K*-strings optimized locally. In this way, GRM provides a straightforward identification of DNA repeats using frequency domain, but avoids mapping of symbolic DNA sequence to numerical sequence, and uses *K*-string matching, but avoids statistical methods and locally optimizing individual *K*-strings. For a given sequence, the GRM algorithm provides in the first step (*identification step*) the corresponding GRM diagram; each significant peak (*fragment length*) presents the length of a repeat unit. In the second step (*analysis step*), for each significant GRM peak, the algorithm determines the corresponding repeat sequences and their positions, the consensus repeat unit and divergence between repeat copies and with respect to consensus (Paar et al. 2011; Glunčić and Paar 2013; Glunčić et al. 2019).

In the case of alpha satellite HORs, when using the hg38 genome assembly, GRM is supplemented by the novel ALPHAsub algorithm, which efficiently recognizes and detects alpha satellite arrays in DNA sequence. As an “ideal key word,” robust 28-bp segment from alpha satellite DNA sequences, TGAGAAACTGCTT TGTGATGTGTGCATT is used (Glunčić et al. 2019). The first step identifies locations of alpha satellite arrays in DNA sequence and the second step performs GRM computation for them. In this way, an ensemble of all alpha satellite HORs is extracted from a given genomic sequence.

In summary, characteristics of the GRM algorithm are robustness with respect to deviations from ideal repeats (substitutions, insertions, deletions), straightforward and parameter-free identification of simple repeats (tandem and dispersed), applicability to very large repeat units—both simple repeats and HORs, straightforward determination of consensus lengths and consensus sequences for simple repeats and HORs. In particular, GRM has no such limitation on the length of HOR copy as the TRF algorithm. The GRM method is a straightforward method to provide a global repeat map in a GRM diagram, identifying all pronounced repeats in a given sequence, without the need for any prior knowledge on repeats. Once the size of the repeat is determined, GRM provides straightforwardly the location of the corresponding repeat arrays. GRM is particularly useful for obtaining precise sequence information since the method does not involve any averaging procedure. It is also useful that the method is rather robust with respect to sizeable substitutions and indels. Once the consensus repeat unit is determined using GRM, in the next step it could be combined with BLAST search for dispersed HOR copies or their constituent monomers. Further information on the GRM code is available on request from the authors.

8.3 Human-Specific NBPF HORs in Human Chromosome 1

8.3.1 The NBPF Gene Family with ~1.6 kb Primary Repeat

The NBPF genes in human chromosome 1 contain a repetitive structure, both in coding and noncoding regions (Vinogradova et al. 2002; Vandepoele et al. 2005), with rather divergent (~20%) repeat units of ~1.6 kb (Fortna et al. 2004; Vandepoele et al. 2005; Popesco et al. 2006). The NBPF monomer repeat encodes the protein domain *of unknown function*, called DUF1220 (Vandepoele et al. 2005), and the NBPF repeat was referred to as NBPF/DUF1220 repeat or DUF1220 repeat or DUF1220 domain. Each ~1.6 kb primary repeat (DUF1220) shows a unique signature of an evenly spaced two exons (Popesco et al. 2006). Recently, Sikela and van Roy (2017) proposed to change the name of the DUF1220 domain to the Olduvai domain.

The DUF1220 copy number is dramatically increased in the human genome with respect to other primates and it was correlated with brain size, cortical neuron number, brain pathologies (autism, schizophrenia, microcephaly, macrocephaly, and neuroblastoma), IQ scores, cognitive aptitude, and evolution (Vandepoele et al. 2005, 2008; Popesco et al. 2006; Dumas et al. 2007, 2012; Dumas and Sikela 2009; Andries et al. 2012; Davis et al. 2014; Keeney et al. 2014; Quick et al. 2015; Astling et al. 2017; Mitchell and Silver 2018; Fiddes et al. 2019; Heft et al. 2020).

DUF1220 domains are the primary repeat units, tandemly repeated in the NBPF genes.

8.3.2 NBPF HORs in Human Chromosome 1 (Build 36.3 Human Genome Assembly) Determined Using HOR-Searching GRM Method

As pointed out in Ref. Andries et al. (2012), before the discovery of NBPF 3mer HOR in Ref. Paar et al. (2011), it was not realized that these diverging NBPF monomer repeats (DUF1220) are of three types, forming a remarkable 3mer HOR organization based on ~1.6 kb primary repeat unit, fully embedded within the NBPF genes.

The GRM diagram for human chromosome 1 was first determined in 2011 for Build 36.3 human genome assembly. In the GRM diagram, the peaks at ~1.6 kb, corresponding to NBPF repeat, and the pronounced GRM peaks at its multiples ~3.2 kb and ~ 4.8 kb, have been identified. These GRM peaks correspond to the 3mer HOR based on the ~1.6 kb primary repeat NBPF monomer (Paar et al. 2011). The largest NBPF 3mer HOR array was found in the contig NT_113799.1 from Build 36.3 (Fig. 8.7 of Ref. Paar et al. (2011), consisting of 17 HOR copies in tandem. The aligned monomer scheme of that tandemly organized HOR array is displayed in Fig. 8.1. We see that out of 17 HOR copies, 14 contain three monomers and the remaining three (from the sixth to eighth row) are missing one of three monomers. Thus, the pattern of the GRM HOR copy array in this contig can be expressed as: 14 × (3 monomer types) + 3 × (2monomer types, 1 monomer type out of 3 missing). On the other hand, using the results of the DUF1220-monomer-searching method (O'Bleness et al. 2012), the corresponding pattern for the NBPF20 gene can be expressed analogously as for the earlier HOR method as 14 × (HLS triplets) + 3 × (2 HLS doublets, one of the three HLS types missing), revealing the congruency of both methods.

In the other three contigs, NT_079497.3, NT_004434.18, and NT_034400.4 from the Build 36.3 assembly, the tandemly organized HOR arrays were also found,

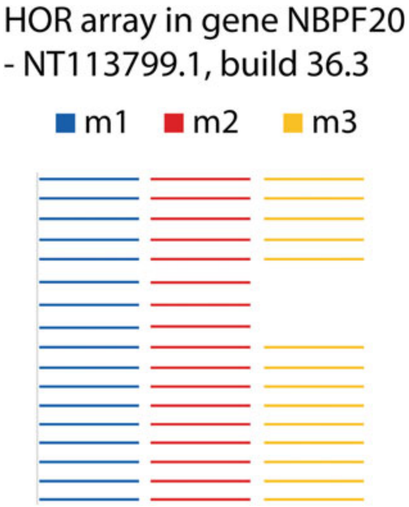


Fig. 8.1 Aligned monomer scheme of 3mer HORs discovered 2011. in NBPF genes from contig NT113799.1, Build 36.3 for human chromosome 1 (Paar et al. 2011). Each row in the scheme represents a HOR copy. The top enumeration of three columns corresponds to three constituent monomer types: m1, m2, and m3. Each monomer in a HOR copy is presented by a horizontal bar in the corresponding column. Monomers of the same type in different HOR copies are presented by bars in the same column. For example, the first HOR copy is presented by three bars: the first, second, and third bar correspond to monomers of types m1, m2, and m3, respectively. Calculated divergence among monomers of the same type was ≤1%. This was the largest NBPF HOR array identified in Build 36.3 (Paar et al. 2011)

containing 16, 12, and 2 HOR copies, respectively (Paar et al. 2011). The total number of identified NBPF HOR copies tandemly organized within the NBPF genes was 47. Additionally, 10 dispersed NBPF HOR copies (not tandemly organized) and 9 dispersed NBPF monomers (outside of HOR copies) were also identified (Paar et al. 2011).

It should be noted that the 3mer NBPF HOR arrays were identified without searching for any specific type of monomers, just applying the GRM algorithm to the available Build 36.3 genome assembly of chromosome 1. From the NBPF HORs, identified by the GRM algorithm, the constituent NBPF monomers were deduced in the second step. It was found that they belong to three monomer types, which were denoted m1, m2, and m3.

The GRM-searching method has the ability to automatically resolve the NBPF primary repeat monomers into three diverging types (m1, m2, m3) that show intragenic copy number increases specific to the human lineage (HLS1, HLS2, HLS3). This was previously considered as a problem in DUF1220-monomer-searching approaches and improved alignment and summarization strategies were recently described (Astling et al. 2017).

Divergence among these three monomer types was ~20%, while the divergence between monomers of the same type in different HOR copies (for example, between the monomer m1 in the first and in the second HOR copy) was small, mostly below 1% (Paar et al. 2011). As a consequence, divergence among 3mer HOR copies was very small. This is a classical HOR pattern, reminiscent of the divergence pattern for alpha satellite HORs (Warburton and Willard 1996).

8.3.3 NBPF HORs in Human Chromosome 1 (hg38 Human Genome Assembly) Using HOR-Searching GRM Method

Using the GRM algorithm, we analyze the genome of human chromosome 1 from recent human genome assembly hg38. Similarly, as GRM results for Build 36.3 assembly (Paar et al. 2011), the GRM peaks for hg38 assembly reveal three HORs with long HOR copies: the NBPF 3mer HOR based on the ~1.6 kb NBPF primary repeat, the hornerin quartic HOR based on 39 bp hornerin primary repeat and the alpha satellite 11mer HOR based on the ~171 bp alpha satellites. GRM diagram for hg38.p2 assembly of the NBPF20 gene shows pronounced peaks at ~1.6, ~3.2, and ~4.8 kb (Fig. 8.2). This is a signature of the NBPF 3mer HOR based on the ~1.6 kb NBPF monomers, similarly as in the case of Build 36.3 genome assembly. It should be noted that the three close-lying peaks around ~1.6 kb correspond to three NBPF monomer types (denoted m1, m2, m3). Using key strings corresponding to the highest frequency of ~1.6 kb NBPF monomers, we identify in the NT_004487.20 contigs the five 3mer HOR arrays: 22-copy, 14-copy, 2-copy, and 11-copy and 14-copy HOR arrays, denoted H1, H2, H3, H4, and H5, respectively.

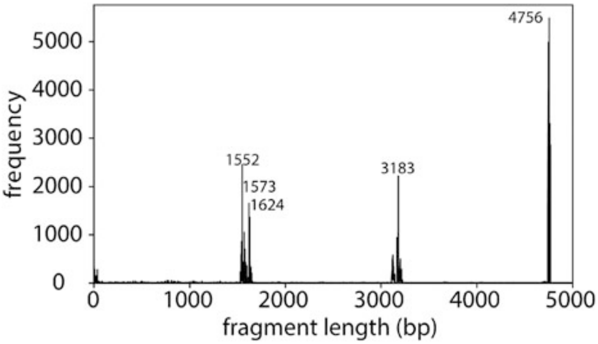


Fig. 8.2 GRM diagram for the NBPF20 gene in human chromosome 1 (hg38.p2 assembly, segment 145,294,660–145,393,360). Three pronounced peaks at ~1.6, ~3.2, and ~4.8 kb correspond to the NBPF 3mer HOR based on the ~1.6 kb NBPF monomers

Their aligned monomer schemes are presented in Fig. 8.3. In comparison to our previous results obtained for Build 36.3 assembly (Paar et al. 2011), the hg38 assembly contains a larger number of NBPF HOR copies. In each HOR array, the HOR copies are denoted by h1, h2, h3, ... Most HOR copies are complete, i.e., m1m2m3, containing all three monomer types and are referred to as canonical HOR copies. In the largest HOR array H1, out of 22 HOR copies 20 are canonical, and two are m1m2, i.e., without the m3 monomer. On the other hand, in HOR array H5 all 14 HOR copies are canonical, i.e., m1m2m3.

HOR array H1 has a tandem of canonical HOR copies h1–h17, and a tandem of canonical HOR copies h20–h22 while between them are two HOR copies h18, h19 which are missing the third monomer m3. Then the whole HOR array h1–h22 is referred to as canonical array H1.

In HOR array H2 the HOR copies h2–h13 are canonical and they are referred to as canonical HOR array H2. The variant copies at lower and upper boundaries, h1 and h4, are noncanonical.

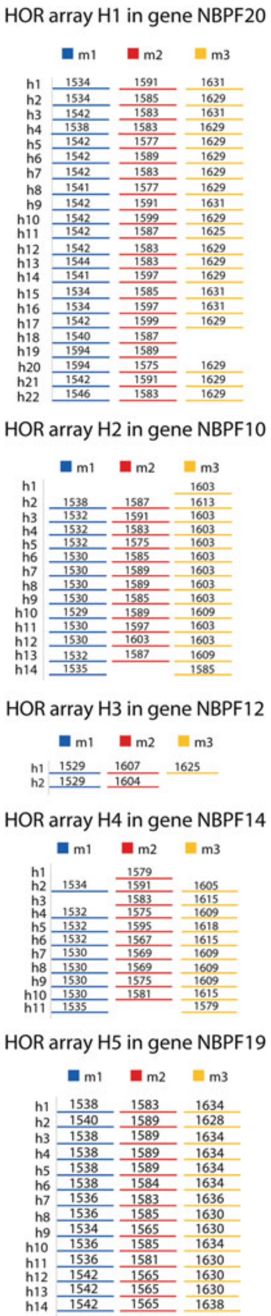
In HOR array H3 only the HOR copy h1 is canonical, while in the h2 copy the monomer m3 is absent from the canonical HOR copy. Exceptionally, we will refer to this two-copy HOR array as canonical.

In HOR array H4 the HOR copies h4–h10 are canonical and therefore are referred to as canonical HOR array H4. At the upper boundary, the copies h1 and h3 are noncanonical and is a 3mer copy isolated from canonical HOR copies and therefore not assigned to canonical HOR array. At the lower boundary, the HOR copy h11 is noncanonical. Thus, the HOR copies h1, h2, h3, h11 do not belong to canonical H4.

In the HOR array H5 all HOR copies are canonical, so the HOR array H5 is canonical.

The pattern of five GRM HOR arrays (Fig. 8.3), excluding isolated HOR copies or segments near the ends of arrays (i.e., h1 and h4 in H2, h2 in H3, h1–h3 and h11 in H4, which lie outside of tandem of canonical HOR copies), the HOR content of NBPF genes in hg38 genome assembly (Fig. 8.3) can be expressed as:

H1: 20 × (canonical 3mer) + 2 × (noncanonical variant, with one monomer absent from canonical 3mer)



HOR array H1 in gene NBPF20

	m1	m2	m3
h1	1534	1591	1631
h2	1534	1585	1629
h3	1542	1583	1631
h4	1538	1583	1629
h5	1542	1577	1629
h6	1542	1589	1629
h7	1542	1583	1629
h8	1541	1577	1629
h9	1542	1591	1631
h10	1542	1599	1629
h11	1542	1587	1625
h12	1542	1583	1629
h13	1544	1583	1629
h14	1541	1597	1629
h15	1534	1585	1631
h16	1534	1597	1631
h17	1542	1599	1629
h18	1540	1587	
h19	1594	1589	
h20	1594	1575	1629
h21	1542	1591	1629
h22	1546	1583	1629

HOR array H2 in gene NBPF10

	m1	m2	m3
h1			1603
h2	1538	1587	1613
h3	1532	1591	1603
h4	1532	1583	1603
h5	1532	1575	1603
h6	1530	1585	1603
h7	1530	1589	1603
h8	1530	1589	1603
h9	1530	1585	1603
h10	1529	1589	1609
h11	1530	1597	1603
h12	1530	1603	1603
h13	1532	1587	1609
h14	1535		1585

HOR array H3 in gene NBPF12

	m1	m2	m3
h1	1529	1607	1625
h2	1529	1604	

HOR array H4 in gene NBPF14

	m1	m2	m3
h1		1579	
h2	1534	1591	1605
h3		1583	1615
h4	1532	1575	1609
h5	1532	1595	1618
h6	1532	1567	1615
h7	1530	1569	1609
h8	1530	1569	1609
h9	1530	1575	1609
h10	1530	1581	1615
h11	1535		1579

HOR array H5 in gene NBPF19

	m1	m2	m3
h1	1538	1583	1634
h2	1540	1589	1628
h3	1538	1589	1634
h4	1538	1589	1634
h5	1538	1589	1634
h6	1538	1584	1634
h7	1536	1583	1636
h8	1536	1585	1630
h9	1534	1565	1630
h10	1536	1585	1634
h11	1536	1581	1630
h12	1542	1565	1630
h13	1542	1565	1630
h14	1542	1565	1638

Fig. 8.3 Aligned monomer scheme of five NBPF HOR arrays in hg38 assembly for human chromosome 1 using GRM algorithm. Top enumeration of three columns corresponds to three constituent monomer types: m1, m2, and m3. Each monomer in a HOR copy is presented by a horizontal bar in the corresponding column. In analogy to Fig. 8.1, each row in the scheme represents a HOR copy. HOR copies are clustered in five distinct HOR arrays. In each HOR array, its HOR copies are denoted h1, h2, h3. Most of HOR copies are complete 3mers—composed of three types of mutually diverging ~1.6 kb monomers, denoted m1 (blue bars), m2 (red bars), and m3 (yellow bars). The HOR arrays are denoted H1 (in the NBPF20 gene), H2 (in the NBPF10 gene), H3 (in the NBPF12 gene), H4 (in the NBPF14 gene), and H5 (in the NBPF19 gene). Each HOR array is composed of HOR copies: H1 of 22 HOR copies (denoted h1–h22 within H1), H2 of 14 (denoted h1–h14 within H2), H3 of 2 (denoted h1–h2 within H3), H4 of 11 (denoted h1–h11 within H4), and H5 of 14 (denoted h1–h14 within H5)

H2: 12 × (canonical 3mer)
H3: 1 × (canonical 3mer)
H4: 7 × (canonical 3mer)
H5: 14 × (canonical 3mer)

The criterion for excluding from counting an isolated canonical HOR copy or its segments is in accordance with the results for triplets obtained by the DUF1220-monomer-searching method (underlined in Fig. 8.2 of Ref. O'Bleness et al. 2014). Due to this criterion, the noncanonical HOR copies h1 and h14 from H2, HOR copy h2 from H3, and HOR copies h1–h3, h11 from H4 are excluded from the comparison. The exception is H3 with only one canonical HOR copy present H1), while the other (h2 in H3) is noncanonical with m3 constituent monomer missing.

On the other hand, using the results of the DUF1220-monomer-searching method, the corresponding pattern for the NBPF20 gene can be expressed analogously as for the HOR-searching method as:

H1: 20 × (HLS triplet) + 2 × (HLS doublet, one of three HLS types is missing)
H2: 12 × (HLS triplet)
H3: 2 × (HLS triplet)
H4: 7 × (HLS triplet)
H5: 14 × (HLS triplet)

Therefore, it is seen that both the HOR-searching and DUF1220-monomer-searching methods are largely congruent in spite of employing different computational procedures.

As seen from divergence, the NBPF HORs have a pattern like classical HORs, in analogy to the alpha satellite HORs (Warburton and Willard 1996). This is clearly seen from the divergence pattern. For example, divergence among canonical HOR copies in HOR array H1 is very small, on the average less than 1% (Table 8.1a); only the divergence between h22 and other HOR copies is somewhat higher, but h22 is near the boundary of the HOR array. On the other hand, divergence among monomers within each HOR copy in H1 is sizable, ~17–19% (Table 8.1b). On average, divergence is ~15–20% among consensus monomers within HOR copies. HOR pattern is clearly seen from divergence matrix among NBPF monomers ordered along HOR copies in HOR array 1 (Table 8.1c). Aligned consensus sequences of three NBPF monomer types, each with two exons, are presented in Fig. 8.4.

8.3.4 NBPF HORs in Neanderthal Genome and Evolution

Recent sequencing of the Neanderthal genome (Kelso and Prüfer 2014) gives opportunities to compare Neanderthal to the modern human genome. It was pointed out that this could contribute to a better understanding of human evolution, especially regarding cognitive aptitude (Noonan and McCallion 2010). Neanderthal had, on average, a larger brain than modern humans did and more DUF1220 copies,

Table 8.1 Divergence pattern of major human NBPF HOR array H1

	(**a**) Divergence matrix (%) among canonical HOR copies (h1–h17, h20–h22) in HOR array H1 (in NBPF 20 gene). Canonical NBPF HOR copies are defined as those which contain full 3mer NBPF monomers. Divergence values are rounded off to closest integers																		
	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11	h12	h13	h14	h15	h16	h17	h20	h21	h22
h1	0	0	0	0	1	0	1	1	0	1	1	1	1	1	0	0	1	2	1	8
h2		0	0	0	1	0	0	1	1	1	0	0	0	1	0	0	1	2	0	7
h3			0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	0	7
h4				0	0	0	0	0	1	1	0	0	0	1	0	1	1	2	0	7
h5					0	0	0	0	1	1	0	0	0	1	1	1	1	2	0	7
h6						0	0	0	0	0	0	0	0	0	1	1	0	2	0	7
h7							0	0	0	0	0	0	0	0	0	1	1	2	0	7
h8								0	1	1	0	0	0	1	1	1	1	2	0	7
h9									0	0	0	0	0	0	1	0	0	3	0	7
h10										0	0	1	1	0	1	0	0	3	0	7
h11											0	0	0	0	1	1	1	2	0	7
h12												0	0	0	0	1	1	2	0	7
h13													0	0	0	1	1	2	0	7
h14														0	1	0	0	3	0	7
h15															0	0	1	2	1	7
h16																0	0	3	1	8
h17																	0	3	0	8
h20																		0	3	9
h21																			0	7
h22																				0

(continued)

Table 8.1 (continued)

	(**a**) Divergence matrix (%) among canonical HOR copies (h1–h17, h20–h22) in HOR array H1 (in NBPF 20 gene). Canonical NBPF HOR copies are defined as those which contain full 3mer NBPF monomers. Divergence values are rounded off to closest integers																		
	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11	h12	h13	h14	h15	h16	h17	h20	h21	h22
	(**b**) Divergence (%) among monomers within canonical HOR copies from HOR array H1																			
	HOR copy																			
m1/m2	19	18	18	18	18	19	18	18	19	19	19	18	18	19	18	19	19	21	19	18
m1/m3	18	17	17	17	17	17	17	17	17	17	17	17	17	17	18	18	17	12	17	26
m2/m3	17	17	17	17	17	17	17	17	17	18	17	17	17	17	17	17	18	17	17	30
	(**c**) Divergence matrix (%) among NBPF monomers along HOR array H1. Monomers are denoted m1 m2 m3 m4 m5 m6...m64, in order of appearance for HOR array H1																			

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23	m24	m25	m26	m27	m28	m29	m30	m31	m32	m33	m34	m35	m36	m37	m38	m39	m40	m41	m42	m43	m44	m45	m46	m47	m48	m49	m50	m51	m52	m53	m54	m55	m56	m57	m58	m59	m60	m61	m62	m63	m64
m1	0	19	18	0	18	17	1	18	17	1	18	17	1	18	17	1	19	17	1	18	17	1	18	17	1	19	18	1	19	18	1	19	17	1	18	17	1	18	17	1	19	17	0	18	18	0	19	18	1	19	17	1	19	6	19	6	18	17	1	19	17	1	18	26
m2		0	17	19	0	17	19	1	17	19	1	17	19	1	17	19	0	17	18	1	17	19	1	17	19	0	17	19	1	17	19	0	17	18	1	17	19	1	17	18	0	17	19	0	17	19	0	17	19	1	17	19	1	21	0	21	1	17	18	0	17	18	1	30
m3			0	18	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	1	17	17	0	17	17	0	17	17	0	18	17	0	18	17	0	17	18	0	17	17	12	17	12	16	0	17	17	0	17	17	20
m4				0	18	17	1	18	17	1	18	17	1	18	17	1	19	17	1	18	17	1	18	17	1	19	18	1	19	18	1	19	17	1	18	17	1	18	17	1	19	17	0	18	18	0	19	18	1	19	17	1	19	6	19	6	18	17	1	19	17	1	18	26
m5					0	17	18	0	17	18	0	17	18	1	17	18	0	17	18	0	17	18	1	17	18	1	17	18	1	17	18	0	17	18	0	17	18	0	17	18	1	17	18	0	17	18	1	17	18	1	17	18	0	21	0	21	1	17	18	1	17	18	0	30
m6						0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	18	0	17	18	0	17	17	12	17	12	17	0	17	17	0	17	17	20
m7							0	18	17	1	18	17	0	18	17	0	19	17	0	18	17	0	18	17	0	19	17	0	19	17	0	19	17	0	18	17	0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	18	6	18	6	18	17	0	19	17	0	18	26
m8								0	17	18	0	17	18	1	17	18	0	17	18	0	17	18	1	17	18	1	17	18	1	17	18	0	17	18	0	17	18	0	17	18	1	17	18	0	17	18	1	17	18	1	17	18	1	21	1	21	1	17	18	1	17	18	0	30
m9									0	17	17	0	17	16	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	1	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	12	17	12	16	0	17	17	0	17	17	20
m10										0	18	17	1	18	17	1	19	17	1	18	17	1	18	17	1	19	17	1	19	17	1	19	17	1	18	17	1	18	17	1	19	17	1	18	17	1	19	17	1	19	17	0	19	6	19	6	18	17	1	19	17	1	18	26
m11											0	17	18	1	17	18	0	17	18	0	17	18	1	17	18	1	17	18	1	17	18	0	17	18	0	17	18	0	17	18	1	17	18	0	17	18	1	17	18	1	17	18	0	21	0	21	1	17	18	1	17	18	0	30
m12												0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	18	0	17	18	0	17	17	12	17	12	17	0	17	17	0	17	17	20
m13													0	18	17	0	19	17	0	18	17	0	18	17	0	19	17	0	19	17	0	19	17	0	18	17	1	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	18	6	18	6	18	17	0	19	17	1	18	26
m14														0	17	18	1	17	18	1	17	18	0	17	18	1	17	18	1	17	18	1	17	18	0	17	18	1	17	18	1	17	18	1	17	18	1	17	18	2	17	18	1	21	1	21	0	17	18	1	17	18	1	30
m15															0	17	17	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	18	0	17	18	0	17	17	12	17	12	17	0	17	17	0	17	17	20
m16																0	19	17	0	18	17	0	18	17	0	19	17	0	19	17	0	19	17	0	18	17	0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	19	6	18	6	18	17	0	19	17	0	18	26
m17																	0	17	19	0	17	19	1	17	19	0	17	19	1	17	19	0	17	18	1	17	19	0	17	19	1	17	19	0	17	19	1	17	19	1	17	19	1	21	0	22	1	17	19	0	17	18	1	30
m18																		0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	18	0	17	18	0	17	17	12	17	12	17	0	17	17	0	17	17	20
m19																			0	18	17	0	18	17	0	19	17	0	19	17	0	18	17	0	18	17	0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	18	6	18	6	18	17	0	19	17	0	18	26
m20																				0	17	18	1	17	18	1	17	18	1	17	18	0	17	18	0	17	18	0	17	18	1	17	18	0	17	18	1	17	18	1	17	18	0	21	0	21	1	17	18	1	17	18	0	30
m21																					0	17	17	0	17	17	0	17	18	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	0	17	18	0	17	18	0	17	17	12	17	12	17	0	17	17	0	17	17	20
m22																						0	18	17	0	19	17	0	19	17	0	19	17	0	18	17	0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	18	6	18	6	18	17	0	19	17	0	18	26
m23																							0	17	18	1	17	18	2	17	18	1	17	18	1	17	18	1	17	18	1	17	18	1	17	18	1	17	18	2	17	18	1	21	1	21	0	17	18	1	17	18	1	30
m24																								0	17	17	0	17	18	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	17	0	17	18	0	17	17	12	17	12	16	0	17	17	0	17	17	20
m25																									0	19	17	0	19	17	0	19	17	0	18	17	0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	19	6	18	6	18	17	0	19	17	0	18	26
m26																										0	17	19	1	17	19	0	17	19	1	17	19	1	17	19	1	17	19	1	17	19	1	17	19	1	17	19	1	21	0	22	1	17	19	0	17	19	1	30
m27																											0	17	18	0	17	17	1	17	17	0	17	17	0	17	17	0	18	17	0	18	17	0	17	18	0	17	17	12	17	12	16	0	17	17	0	17	17	20
m28																												0	19	17	0	19	17	0	18	17	0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	19	6	18	6	18	17	0	19	17	0	18	26
m29																													0	18	19	1	18	19	1	18	19	1	18	19	0	18	19	1	18	19	0	18	19	0	18	19	1	22	1	22	2	18	19	1	18	19	1	30
m30																														0	17	17	0	17	17	0	17	17	0	17	18	0	18	17	0	18	18	0	17	18	0	17	17	12	17	12	17	0	17	17	0	17	17	20
m31																															0	19	17	0	18	17	0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	18	6	18	6	18	17	0	19	17	0	18	26
m32																																0	17	18	1	17	19	0	17	18	1	17	19	0	17	19	1	17	19	1	17	19	0	21	0	21	1	17	18	0	17	18	1	30
m33																																	0	17	17	0	17	17	0	17	18	0	17	17	0	17	18	0	17	18	1	17	17	11	17	12	17	0	17	17	0	17	17	20
m34																																		0	18	17	0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	18	6	18	6	18	17	0	18	17	0	18	26
m35																																			0	17	18	0	17	18	1	17	18	0	17	18	1	17	18	1	17	18	0	21	0	21	1	17	18	1	17	18	0	30
m36																																				0	17	17	0	17	18	0	17	17	0	17	18	0	17	18	0	17	17	12	17	12	17	0	17	17	0	17	17	20
m37																																					0	18	17	0	19	17	1	18	17	1	19	17	0	19	17	0	18	6	18	6	18	17	0	19	17	0	18	26
m38																																						0	17	18	1	17	18	0	17	18	1	17	18	1	17	18	0	21	0	21	1	17	18	1	17	18	0	30
m39																																							0	17	18	0	17	17	0	17	18	0	17	18	0	17	17	12	17	12	17	0	17	17	0	17	17	20
m40																																								0	19	17	1	18	17	1	19	17	0	19	17	0	18	6	18	6	18	17	0	19	17	0	18	26
m41																																									0	17	19	1	17	19	0	17	19	0	18	19	1	22	1	22	1	18	19	1	18	19	1	30
m42																																										0	17	17	0	17	17	0	17	18	0	17	17	12	17	12	16	0	17	17	0	17	17	20
m43																																											0	18	18	0	19	18	1	19	17	1	19	6	19	6	18	17	1	19	17	1	18	26
m44																																												0	17	18	1	17	18	1	17	18	0	21	0	21	1	17	18	1	17	18	0	30
m45																																													0	18	17	0	17	17	0	17	17	12	17	12	16	0	17	17	0	17	17	20
m46																																														0	19	18	1	19	17	1	19	6	19	6	18	17	1	19	17	1	18	26
m47																																															0	17	19	0	18	19	1	22	1	22	1	18	19	1	18	19	1	30
m48																																																0	17	17	0	17	17	12	17	12	16	0	17	17	0	17	17	20
m49																																																	0	19	17	0	19	6	18	6	18	17	0	19	17	0	18	26
m50																																																		0	18	19	1	22	1	22	2	18	19	1	18	19	1	30
m51																																																			0	17	17	12	17	12	17	0	17	17	0	17	17	20
m52																																																				0	19	6	18	6	18	17	0	19	17	0	18	26
m53																																																					0	21	0	21	1	17	18	1	17	18	1	30
m54																																																						0	21	0	21	12	6	21	12	6	21	20
m55																																																							0	21	1	17	18	0	17	18	1	30
m56																																																								0	21	12	6	22	12	6	21	20
m57																																																									0	17	18	1	17	18	1	29
m58																																																										0	17	17	0	17	17	20
m59																																																											0	19	17	0	18	26
m60																																																												0	17	18	1	30
m61																																																													0	17	17	20
m62																																																														0	18	26
m63																																																															0	30
m64																																																																0

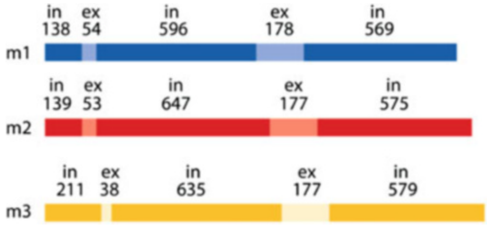


Fig. 8.4 Positions of exons and introns in three types of consensus monomers from human NBPF HOR array H1. Internal intron/exon substructure of consensus monomers m1, m2, m3 from HOR copies in HOR array 1 are shown. Each NBPF monomer contains two exons (dark color) and three introns (light color). Above segments presenting exons (denoted *ex*) and introns (denoted *in*) the corresponding lengths are given

Table 8.2 Comparison of HOR array structure of Neanderthal and human genome

Neanderthal			Human		
HOR array	No. HOR copies		HOR array	No. HOR copies	
	Total	Canonical		Total	Canonical
N1	16	8	H1	22	20
N2	13	11	H2	14	12
N3	4	2	H3	2	1
N4	12	8	H4	11	8
N5	4	2	H5	14	14
N6	18	14			
	67	45		63	55

highly expressed in brain regions associated with higher cognitive function (Holloway 1985; Dumas et al. 2012; O'Bleness et al. 2012; Keeney et al. 2014). But it was argued that, given the limited knowledge, one cannot make conclusions regarding whether the increased number of DUF1220 copies in Neanderthal conferred any evolutionary advantage or disadvantage (Keeney et al. 2014).

To address this intriguing question, we study the Neanderthal orthologues of human NBPF HORs, applying the GRM algorithm. We identify in the Neanderthal genome six NBPF HOR arrays containing canonical ~4.7 kb 3mer HOR copies m1m2m3. The number of NBPF HOR copies in NBPF HOR arrays in the Neanderthal genome is presented in Table 8.2, in comparison with the human genome. We find in Neanderthal NBPF HOR copies that the intra-HOR monomer divergence is an order of magnitude larger than the inter-HOR copy monomer divergence, in accordance with the HOR pattern in the human genome (Table 8.3).

The main difference between the Neanderthal and human genome is a larger number of canonical 3mer NBPF HOR copies in the human than in the Neanderthal genome, in spite of a smaller number of NBPF monomers (i.e., DUF1220 domains) in the human genome. Furthermore, Neanderthal has a symmetry-breaking 4mer variant of NBPF HOR copy, which is not present in human HOR copies.

Table 8.3 Divergence in Neanderthals canonical HOR copies in NBPF HOR array N1

(**a**) intra-HOR copy divergence (%) among consensus monomers.
div (m1 vs. m2) = 20, div (m1 vs. m3) = 14, div (m2 vs. m3) = 17.
(**b**) inter-HOR divergence (%) among canonical HOR copies.

	n1	**n2**	**n6**	**n7**	**n8**	**n9**	**n11**	**n15**
n1	0.0	2.0	2.5	0.9	0.8	0.7	2.3	2.3
n2		0.0	4.0	2.6	2.5	2.3	3.9	4.0
n6			0.0	2.7	2.8	2.4	0.4	0.4
n7				0.0	0.6	0.6	2.8	2.6
n8					0.0	0.8	2.7	2.4
n9						0.0	2.5	2.5
n11							0.0	0.4
n15								0.0

Neanderthal HOR arrays exhibit larger monomer dispersion, at the expense of canonical 3mer HORs. One could hypothesize that a larger number of canonical HOR copies in the human genome may have contributed to the evolutionary advantage of human over Neanderthal, resulting in a more favorable DUF1220 protein distribution within the human brain. The importance of DUF1220 protein distribution in the human brain has been stressed previously (Kochiyama et al. 2018).

The largest human NBPF HOR array H1 contains 20 canonical 3mer NBPF HOR copies m1m2m3, and two 2mer variant HOR copies m1m2 (deletion of m3 monomer), while the largest Neanderthal HOR array N6 contains 14 canonical 3mer HOR copies m1m2m3, three 2mer variant HOR copies m1m2 (deletion of m3 monomer) and one variant HOR copy m3 (deletion of m1m2 monomers). We could hypothesize that the number of canonical 3mer HOR copies could have a role of an additional signature: human HOR arrays contain 55 canonical HOR copies compared to 45 in Neanderthal. There is also a question of a possible role of an additional 4mer variant in the Neanderthal genome, containing the additional fourth monomer m4, sizably divergent with respect to monomers m1, m2, and m3.

It was noted that a reconstructed Neanderthal genome sequence could help reveal the evolutionary genetic events that produced modern humans (Noonan and McCallion 2010). Our results indicate that the cognitive capabilities may depend significantly not only on the number of NBPF monomers (HLS DUF1220 domains in the terminology of the DUF1220-monomer-searching method) but also on the number of canonical HOR copies. We discovered that, for the available Neanderthal genome, significant differences in NBPF HORs were created after the human–Neanderthal split, during a relatively short period of at most about 1 million years. The present Neanderthal study provides additional insight into recent human lineage-specific changes on a finer "tuning." An intriguing question is also whether this violation of the HOR pattern might have contributed to the extinction of Neanderthal, our closest extinct relative.

8.3.5 NBPF HORs in Chimpanzee Genome and Cognitive Evolution

In Ref. Paar et al. (2011) it was shown that the number of NBPF monomer copies and the number of all NBPF HOR copies gradually decrease from human to chimpanzee to orangutan to Rhesus macaque (Table 8.4). However, in the chimpanzee genome assembly Build 2.1 the number of tandemly organized NBPF HOR copies drops to 0 from 47 in Build 36.3 for the human genome (Paar et al. 2011). Tandemly organized NBPF copies were also absent in orangutan and Rhesus macaque genomes. The tandem repeat of the NBPF HOR copies shows a discontinuous jump in the evolutionary step from chimpanzee to the human genome: from total absence of tandem repeats of NBPF HOR copies in chimpanzees to 47 tandem repeat HOR copies in humans. This human accelerated HOR pattern (HAHOR) is one of the factors that distinguish humans from nonhuman primates (Paar et al. 2011). It will be interesting to check whether such a drastic drop of tandemly organized HOR copies persists in more recent genome assemblies of nonhuman primates. Such a pattern would be consistent with our finding of decreasing number of tandemly organized NBPF HOR copies in the Neanderthal genome with respect to the human genome.

8.3.6 HOR-Searching Method Versus DUF1220-Monomer-Searching Method for NBPF Repeats

In the HOR-searching method, HORs have been previously directly identified (Paar et al. 2011) using the GRM algorithm for Build 36.3 human genome assembly of human chromosome 1. GRM identifies directly NBPF HOR copies, which consist of three ~1.6 kb primary repeat sequences—NBPF monomers. Such identification was possible with higher accuracy than in the case of identification of ~1.6 kb repeat

Table 8.4 Comparison of the number of NBPF monomers and NBPF HOR copies in NCBI assemblies of chromosome 1 for humans, chimpanzee, orangutan, and rhesus macaque (Paar et al. 2011) using GRM

	All monomer copies	All HOR copies	Tandemly organized HOR copies
Human[a]	165	57	47
Chimpanzee[b]	48	14	0
Orangutan[c]	17	7	0
Rhesus macaque[d]	7	2	0

[a] Build 36.3
[b] Build 2.1
[c] WUSTL Pongo albelii-2.02
[d] Build 1.1

sequences, because generally, the divergence between HOR copies (<5%) is much smaller than the divergence between constituent ~1.6 kb repeat sequences (~20%). In five human NBPF genes the 3mer HORs, determined by GRM in human genome sequence are lined up in tandem, forming a HOR array (Paar et al. 2011).

On the other hand, in the first step, the DUF1220 repeats (domains) are identified using the DUF1220-monomer-searching method. However, it was noted that there is a problem to determine DUF1220 domains with high accuracy and more recently, a novel method was developed to determine copies of the DUF1220 domain from the whole genome sequence data with optimal resolution (Astling et al. 2017). Using DUF1220 domains identified in the first step, in the second step the repeating HLS DUF1220 triplets are obtained. These triplets consist of three types of DUF1220 domains, called HLS1, HLS2, and HLS3 (O'Bleness et al. 2012). In five human NBPF genes, these triplets are lined up in tandem (O'Bleness et al. 2012, 2014; Dumas et al. 2012). The three ~1.6 kb monomers obtained in the HOR-method correspond to the three ~1.6 kb HLS DUF1220 domains in DUF1220-method (Table 8.5).

In accordance with the classical HOR terminology from Warburton and Willard (1996), in the HOR-searching method, we use the name NBPF *3mer HOR copy*, while in the DUF1220-monomer-searching method the name *HLS DUF1220 triplet* was used for the same structure (O'Bleness et al. 2012, 2014; Dumas et al. 2012; Keeney et al. 2014; Davis et al. 2014; Astling et al. 2017).

The correspondence between the terms used in the HOR-searching method and in the DUF1220 (monomer)-searching method (Table 8.5):

*NBPF monomer i*s named *DUF1220*;
monomer m1 is named *DUF1220 HLS1*;
monomer m2 is named *DUF1220 HLS2*;
monomer m3 is named *DUF1220 HLS3*.;

NBPF 3mer HOR is named *HLS DUF1220 triplet*.

Table 8.5 Correspondence between two different terminologies used for the presentation of the same NBPF higher-order repeat structure. The same higher-order repeat structure was discovered by different computational methods. In Ref. Paar et al. (2011) the higher-order repeat pattern was obtained by HOR-searching method using GRM algorithm, and in Refs. O'Bleness et al. (2012, 2014) by monomer-searching tools. In both cases, the resulting pattern was the same but given different names. The HOR-searching method (Paar et al. 2011) is simpler because divergence between higher-order structures (3mer HORs, i.e., HLS DUF1220 triplets) is much smaller than the divergence between constituting primary repeats (NBPF monomers i.e. DUF1220 HLS domains). The HOR terminology is in accordance with the terminology in extensive studies of alpha satellite HORs (Warburton and Willard 1996)

Ref. Paar et al. (2011)	Refs. O'Bleness et al. (2012, 2014), Astling et al. (2017)
NBPF monomer m1	DUF1220 HLS1 domain
NBPF monomer m2	DUF1220 HLS2 domain
NBPF monomer m3	DUF1220 HLS3 domain
NBPPF 3mer HOR	HLS DUF1220 triplet

Both methods give similar results, and some computational differences may arise because of a larger divergence between repeat monomers than between HOR copies.

8.4 Unique Hornerine Quartic HOR Array Embedded Within One Hornerin Exon

GRM diagram of contig NT_004487.18 (Build 36.3 assembly) for human chromosome 1 shows a pronounced peak at 1410 bp (Fig. 8.5a). We found that it corresponds to an array of five copies of 1410 repeat units (start position 2,676,458 in contig NT_00487.18). We have determined its consensus in GRM (Paar et al. 2011). The average divergence of five repeat copies with respect to consensus is ~4%. In the next step, we computed the GRM diagram for consensus of the 1410 bp repeat unit, in order to reveal its internal repeat structure (Fig. 8.5b). This diagram shows a series of equidistant peaks at multiples of 39 bp ($n \times 39$ bp, $n = 1, 2, 3, ...$). Among the multiples of 39 bp in this GRM diagram, the next pronounced peak is at ~0.35 kb, which is equal to 9 × 39 bp; it corresponds to 9mer secondary HOR based on the 39 bp primary repeat.

In general, for a primary repeat of length $l_{\text{prim.}}$, the multiple peak at the length $L = n \times l_{\text{prim.}}$ corresponds to the *n*mer HOR unit if the peak is much higher than the neighboring primary repeat multiples (peaks corresponding to the lengths of multiples $(n - 1) \times l_{\text{prim.}}$ and $(n + 1) \times l_{\text{prim.}}$. In general, a HOR peak corresponds to a pronounced local maximum among multiple peaks in the GRM diagram. Usually, it is sufficient that the multiple peak is sizably higher than the neighboring multiple peak to the left. In the GRM diagram from Fig. 8.5b the peak at 9 × 39 bp = 0.35 kb is much higher than the neighboring multiple peaks at 8 × 39 bp and 10 × 39 bp and therefore corresponds to the 9mer HOR with respect to the 39 bp primary repeat (Table 8.6).

The next larger length at which there is a pronounced peak much higher than the neighboring multiples of primary repeat ~39 bp is ~0.70 kb (corresponding to $n = 18$). However, simultaneously the ~0.35 kb HOR unit by itself acts as a primary repeat for the ~0.70 kb repeat.

$$2 \times 0.35 \text{ kb} = 0.70 \text{ kb}$$

Since there is no GRM peak at 3 × 0.35 kb = 1.05 kb, as seen from Fig. 8.5a, it follows that the ~0.70 kb peak corresponds to the 2mer tertiary HOR with respect to the ~0.35 kb secondary HOR repeat (Table 8.6).

Analogously, the peak at 0.70 kb acts as a primary repeat for the ~1.41 kb repeat.

$$2 \times 0.70 \text{ kb} \sim 1.41 \text{ kb}$$

and the ~1.41 kb peak corresponds to the 2mer quartic HOR with respect to the ~0.70 kb tertiary HOR repeat (Fig. 8.5a and Table 8.6).

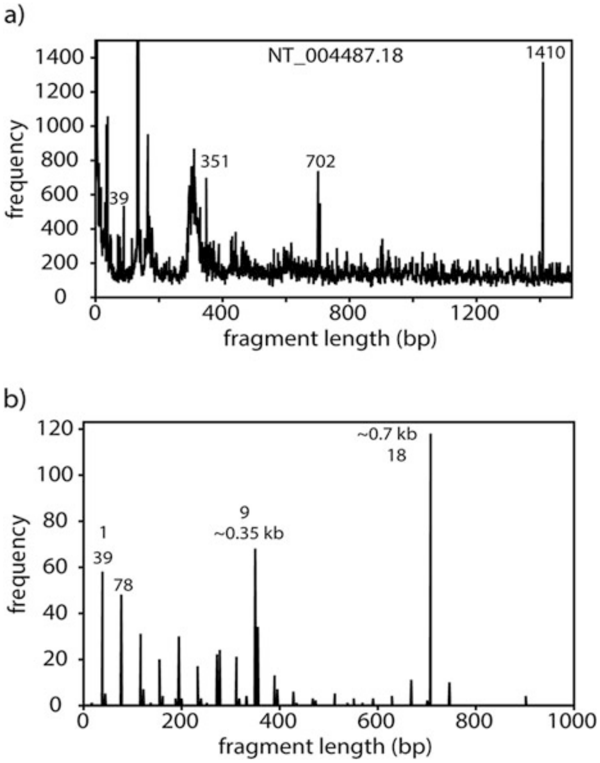


Fig. 8.5 GRM diagrams for hornerin quartic HOR in build 36.3 assembly of human chromosome 1. (**a**) GRM diagram for the 0 - 2.7 Mb segment in human contig NT_113799.1 of chromosome 1 from build 36.3 (Paar et al. 2011). The peaks at 39 bp, 351 bp, 702 bp, and 1410 bp correspond to primary repeat unit, secondary HOR, tertiary HOR, and quartic HOR, respectively. (**b**) GRM diagram for consensus sequence of the 1410 bp quartic HOR copy in contig NT_00487.18

Table 8.6 Three levels of hornerin HOR organization based on the 39 bp hornerin primary repeat. Hornerin primary repeat unit of 39 bp was at first amplified ninefold, and then duplicated, and finally further duplicated

Repeat unit level	Length
PRU (primary)	39 bp
SRU (secondary HOR)	39 × 9 bp = 351 bp ≈ 0.35 kb
TRU (tertiary HOR)	(39 × 9) × 2 bp = 702 bp ≈ 0.70 kb
QRU (quartic HOR)	[(39 × 9) × 2] × 2 bp = 1404 bp ≈ 1.40 kb

Thus, the GRM diagram reveals three levels of hornerin repeat organization of the 39 bp primary repeat unit: (1) nine 39 bp primary repeat units are organized into a ~0.35 kb secondary HOR repeat unit, (2) two ~0.35 kb secondary repeat units are organized into a ~0.70 kb tertiary HOR repeat unit, and (3) the two ~0.70 kb tertiary repeat units are organized into the ~1.4 kb quartic HOR repeat unit (Paar et al. 2011). This quartic HOR repeat scheme is schematically presented in Fig. 8.6. Such higher-order organization was also confirmed by a straightforward analysis of Build 36.3 genomic assembly of human chromosome 1 (Paar et al. 2011, Fig. 8.7).

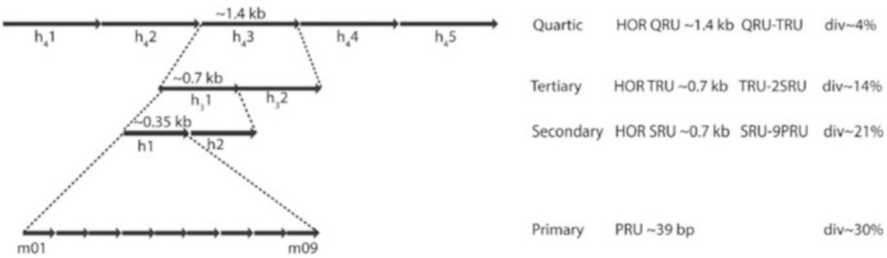

~1.4 kb
$h_4 1$
$h_4 2$
$h_4 3$
$h_4 4$
$h_4 5$
~0.7 kb
$h_3 1$
$h_3 2$
~0.35 kb
h1
h2
m01
m09
Quartic HOR QRU ~1.4 kb QRU-TRU div~4%
Tertiary HOR TRU ~0.7 kb TRU-2SRU div~14%
Secondary HOR SRU ~0.7 kb SRU-9PRU div~21%
Primary PRU ~39 bp div~30%

Fig. 8.6 Schematic presentation of the hierarchical structure of 1.41 kb hornerin quartic HOR array

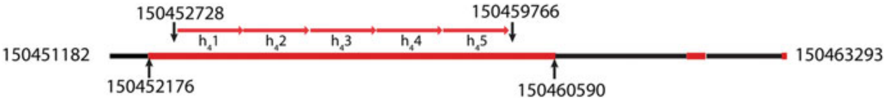


Fig. 8.7 Domain of hornerin quartic HOR array embedded within the large exon in hornerin gene. The hornerin quartic HOR array is positioned from 150,452,728 to 150,459,766 in Build 36.3 assembly of human chromosome 1, embedded completely within the long exon (150,451,182 to 150,460,590) in hornerin gene. This gene contains also the two short exons on the r.h.s of the gene. Exons are presented by red lines and introns by black. The start and end of the hornerin quartic HOR array are marked by vertical arrows pointing to the location within the domain of the long exon. The quartic HOR array covers most of the domain of the long exon

$$\left\{ \left\{ \left[(39\text{ bp})_{\text{primary}} \times 9 \right]_{\text{secondary}} \times 2 \right\}_{\text{tertiary}} \times 2 \right\}_{\text{quartic}} \times 5$$

Nine 39-bp primary repeat units $m1$, $n2$, . . ., $m9$, with mutual divergence ~30% (the first line from below), form the ~0.35 kb secondary HOR unit. Two secondary HOR units, denoted h1 and h2 with mutual divergence ~21% (second line from below), form the ~0.7 kb tertiary HOR unit. Two tertiary HOR units, denoted $h_3 1$ and $h_3 2$, with mutual divergence ~14% (third line from below), form the ~0.7 kb quartic HOR unit, denoted $h_4 3$ (third line from below). Here, the secondary, tertiary, and quartic HOR units are denoted by h, h_3, and h_4, respectively.

By using consensus sequence, it was shown that the average divergence between neighboring copies is gradually decreasing with increasing level of HOR organization, from ~32% between 39 bp copies of primary repeats to ~4% between ~1.41 kb quartic HOR copies. Such hierarchy of divergence is a signature of HOR organization.

In Ref. Paar et al. (2011) the quartic HOR was detected in the human genome only, and no counterpart was found in then-available chimpanzee genome.

Using the structure of human hornerin protein, Takaishi et al. (2005) have deduced the corresponding amino acid sequence. The repetitive region was divided into segments. The smallest units of 39 bp amino acids showed a moderate homology to each other. However, Takaishi et al. (2005) concluded that the analysis was compatible with the notion that the unit of 39 bp was at first amplified fourfold and then triplicated to form three segments in tandem, these being further amplified sixfold, implying a {[(39 bp) × 4] × 3} × 6 = 2.8 kb organization, which differs from the HOR organization {[(39 bp) × 9] × 2} × 2 = 1.4 kb by (Paar et al. 2011). In the corresponding GRM diagram, the peaks at 39, ~0.35, ~0.70, and ~1.41 kb are present (Table 8.5a), in full accordance with our annotation of ~1.4 kb quartic HOR, whereas no peak appears at ~2.8 kb, that was predicted by Takaishi et al. (2005).

It should be noted that a tandem repeat unit of ~1.4 kb was found by the Tandem Repeat Finder algorithm (Warburton et al. 2008), but the HOR pattern was not detected.

In a recent extensive study, Romero et al. (2018) fully confirmed the human hornerin quartic repeat organization given by Paar et al. (2011). Moreover, Romero et al. discovered the same formation also in recent genome assemblies of all primates, except crab-eating macaque (Romero et al. 2018).

8.5 33mer Alpha Satellite HOR in Human Chromosome 21: the Longest HOR Repeat Unit in Human Chromosomes

8.5.1 GRM Diagrams for 33mer, 23mer, 22mer, two 20mers, 16mer, 11mer, and 8mer in Human Chromosome 21

The centromeres in primate genomes contain tandem repeats of ~171 bp alpha satellite DNA, commonly organized into HORs. In spite of their importance, these satellites have been understudied because of still existing gaps in centromere sequencing, genomic "black holes." Using the GRM algorithm we identified in the hg38 assembly of human chromosome 21 complete ensemble of alpha satellite HORs with six long repeat units (≥20mers), five of them novel (Glunčić et al. 2019). The novel 33mer alpha satellite HOR has the longest HOR unit identified so far among all human somatic chromosomes.

In genome assembly hg38 (GCA_000001405.15) a number of human chromosome 21*p* clones have been added and the centromeric gap was filled with "reference models," which are representations of alpha satellite HOR domains. Alpha satellite HOR ideogram obtained for chromosome 21 is shown in Fig. 8.8. To our knowledge, this is the first time that a complete ensemble of $n \geq 8$ alpha satellite HORs of a human chromosome was proposed for a centromeric region.

The GRM diagram was computed for the hg38 DNA sequence of the whole chromosome 21 (Fig. 8.9a). Pronounced peaks at fragment lengths that are approximately equal to 171 × *n* bp, i.e., to multiples of alpha satellite monomer length ~171 bp, are candidates for *n*mer alpha satellite HORs, usually, if a peak at ~171 × *n* bp is sizably higher than the neighboring peak at lower fragment length ~171 × (*n* − 1) bp. For example, for 8mer HOR, the peak at ~171 × 8 bp is sizably stronger than the peak at ~171 × 7 bp; for 11mer HOR, the peak at ~171 × 11 bp

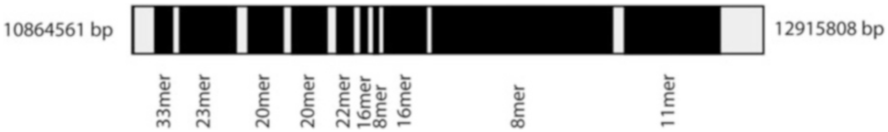


Fig. 8.8 Alpha satellite HOR ideogram for alpha satellite HOR arrays in human chromosome 21. Alpha satellite HOR arrays for long alpha satellite repeat units ($n \geq 8$ monomers), determined by applying GRM algorithm to the hg38 assembly of the centromere of human chromosome 21 (positions 10,864,561–12,915,808 bp). An additional 23mer HOR (with reverse monomers) is obtained by GRM in the long arm of chromosome 21 away from the centromere (start at position 7,970,290, not shown in the figure). Within the centromere, the GRM algorithm identifies ten major alpha satellite arrays ordered in the direction from the long toward the short arm of the chromosome: 33mer, 23mer, 20mer, 20mer, 22mer, 16mer, 8mer, 16mer, 8mer, 11mer

is sizably stronger than at ~171 × 10 bp; for 16mer HOR, the peak at ~171 × 16 bp is sizably stronger than at ~171 × 15 bp; for 20mer HOR the peak at ~171 × 20 bp is sizably stronger than at ~171 × 19 bp, etc. We can directly confirm this attribution by analyzing the corresponding hg38 DNA sequences. For monomeric alpha satellite arrays, the frequencies of peaks at approximately 171 × *n* bp gradually decrease with increasing *n* and a peak sizably above this background is an indication for HOR.

In order to identify a complete ensemble of alpha satellite HORs in a given DNA sequence, we extended the use of the GRM algorithm combined with novel algorithm ALPHAsub (Glunčić et al. 2019; Vlahović et al. 2020) to identify positions of all alpha satellite arrays (regardless whether being of HOR type or not). In this way, we determine contigs in which alpha satellite arrays are located and the GRM algorithm is applied to each of these contigs. The GRM analysis of contigs containing alpha satellite *n*mer HORs provides more peaks at position 171*n* bp, than the GRM analysis of the whole genome, because the noise due to the other repeats is sizably smaller for a contig than for the whole chromosome. In this way, it is straightforward to determine whether the alpha satellite array is an HOR array.

The GRM peak corresponding to 33mer in the GRM diagram for the whole chromosome 21 is small, but visible at 5639 bp in the magnified segment of the HOR diagram. On the other hand, the 5639 bp peak of 33mer HOR is sizeable in the GRM diagram for contig NT_187321.1, in which is the 33mer HOR located (Fig. 8.9b). The length of a 33mer HOR copy is ~33 × 0.171 kb ~ 5.6 kb.

The 33mer HOR array has a very regular structure; its four HOR copies are canonical. Using GRM, the DNA sequences of 33mer HOR copies are determined from hg38 for human chromosome 21 and the consensus sequence was determined (Glunčić et al. 2019; Vlahović et al. 2020). The average divergence among monomers in the 33mer consensus HOR is ~19%, and the divergence between 33mer HOR copies is ~5%.

The computed GRM diagrams for some other *n*mer HORs in the corresponding contigs are presented in Figs. 8.9c–i).

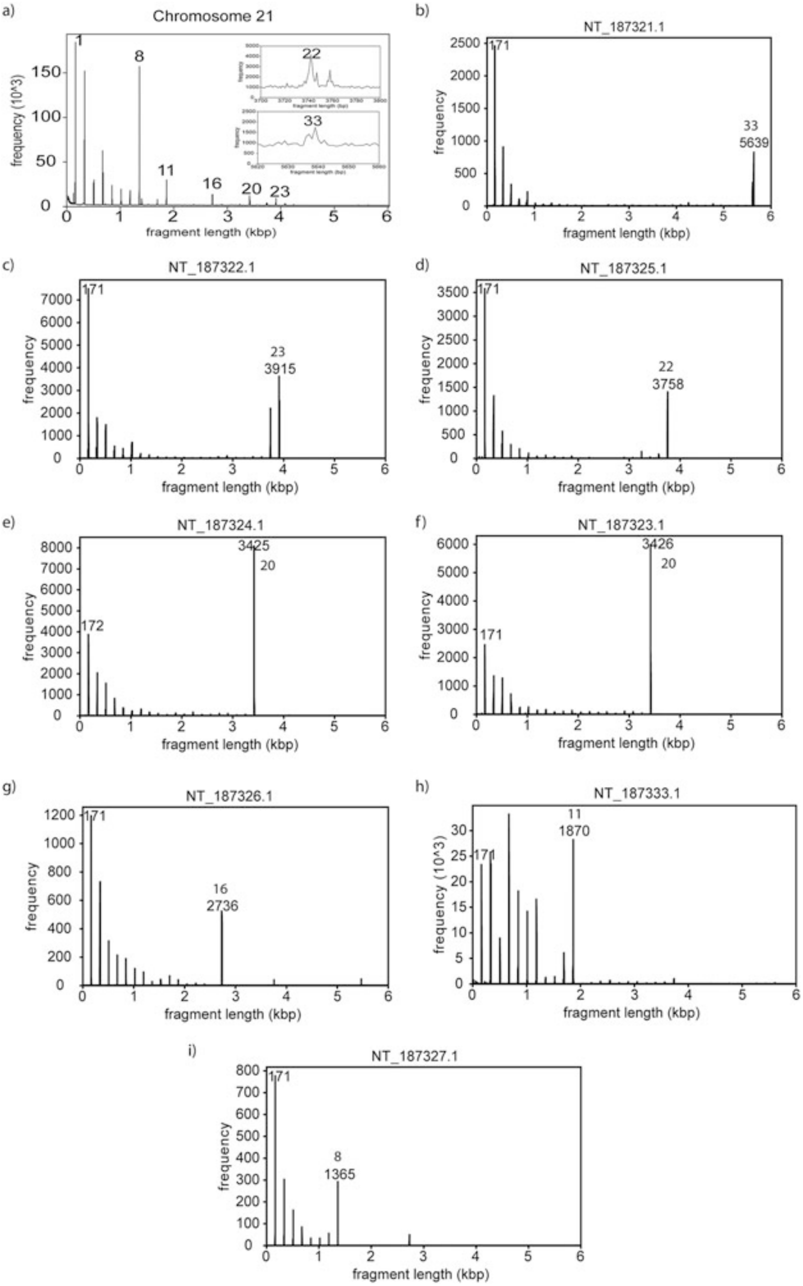


Fig. 8.9 GRM diagrams for alpha satellite HORs in the centromere of human chromosome 21. (**a**) GRM diagram for alpha satellite HORs of the whole chromosome 21. Pronounced peaks that correspond to alpha satellite *n*mer HORs are denoted by number *n* of monomers, given above the fragment length of major peaks. Two inserts give a magnified presentation of weak peaks for 22mer and 33mer, which are sizably screened by the noise of different other repeats. (**b**) GRM diagram for contig NT_187321 in which the 33mer array is located. The pronounced GRM peak at 5539 bp is a

8.5.2 Dot-Matrix Analysis of HORs in Chromosome 21 Identified Using GRM Diagrams

For each alpha satellite array identified by the GRM diagram, the corresponding dot-matrix diagram was computed to confirm its HOR structure. The dot-matrix for 33mer HOR is shown in Fig. 8.10a. The regular HOR pattern is characterized by off-diagonal lines at a spacing equal to the number of monomers in the HOR unit, $n = 33$, parallel to the self-diagonal. We computed regular dot-matrix diagrams also for some other high-multiple *n*mer HORs, 23mer, 22mer, two 20mers, and 8mer (Fig. 8.10b–f).

However, for some more complex HORs, the dot-matrix is not regular. Such an example is the approximately intertwined 17mer + 8mer HOR-like pattern (Fig. 8.11). In this case, the GRM diagram reveals peaks at ~17 × 171 bp and at ~8 × 171 bp (Fig. 8.11a). In such case, a peculiar partial symmetry arises in the corresponding dot-matrix (Fig. 8.11b), where almost all points lie on two sets of off-diagonal straight lines, parallel to the main diagonal and mutually shifted by 8 bp, with an approximate asymmetry described in the caption to Fig. 8.11b).

8.5.3 Four Human Chromosomes, 21, 13, 22, and 14 Share the Same 33mer HOR in hg38 Assembly

In accordance with GRM diagrams (Fig. 8.12a–d) and dot-matrix (Fig. 8.10a), which is for hg38 genome assembly of four chromosomes the same, the identical 33mer HOR appears in hg38 genomes of chromosomes 21, 13, 14, and 22. A careful check of sequencing in this region seems to be required.

8.5.4 Novel GRM Tandem Repeat Database

Details about human alpha satellite arrays (their start position in chromosome, sequence length of array, monomer length, and number of monomers for specific array) are given in novel tandem repeat database at http://genom.hazu.hr/search.html as well as for Neanderthal and chimpanzee genomes. It contains around 3000 records and in the future, we will expand it with data for other tandem repeats

◀

Fig. 8.9 (continued) signature of 33merHOR (5639:171 ≈ 33). (**c**) GRM diagram for 23mer in NT_187322.1. (**d**) GRM diagram for 22mer in NT_187325.1. (**e**) GRM diagram for 20mer in NT_187324.1. (**f**) GRM diagram for 20mer in NT_187323.1. (**g**) GRM diagram for 16mer in NT_187326.1. (**h**) GRM diagram for 11mer in NT_187333.1. (**i**) GRM diagram for 8mer in NT_187327.1

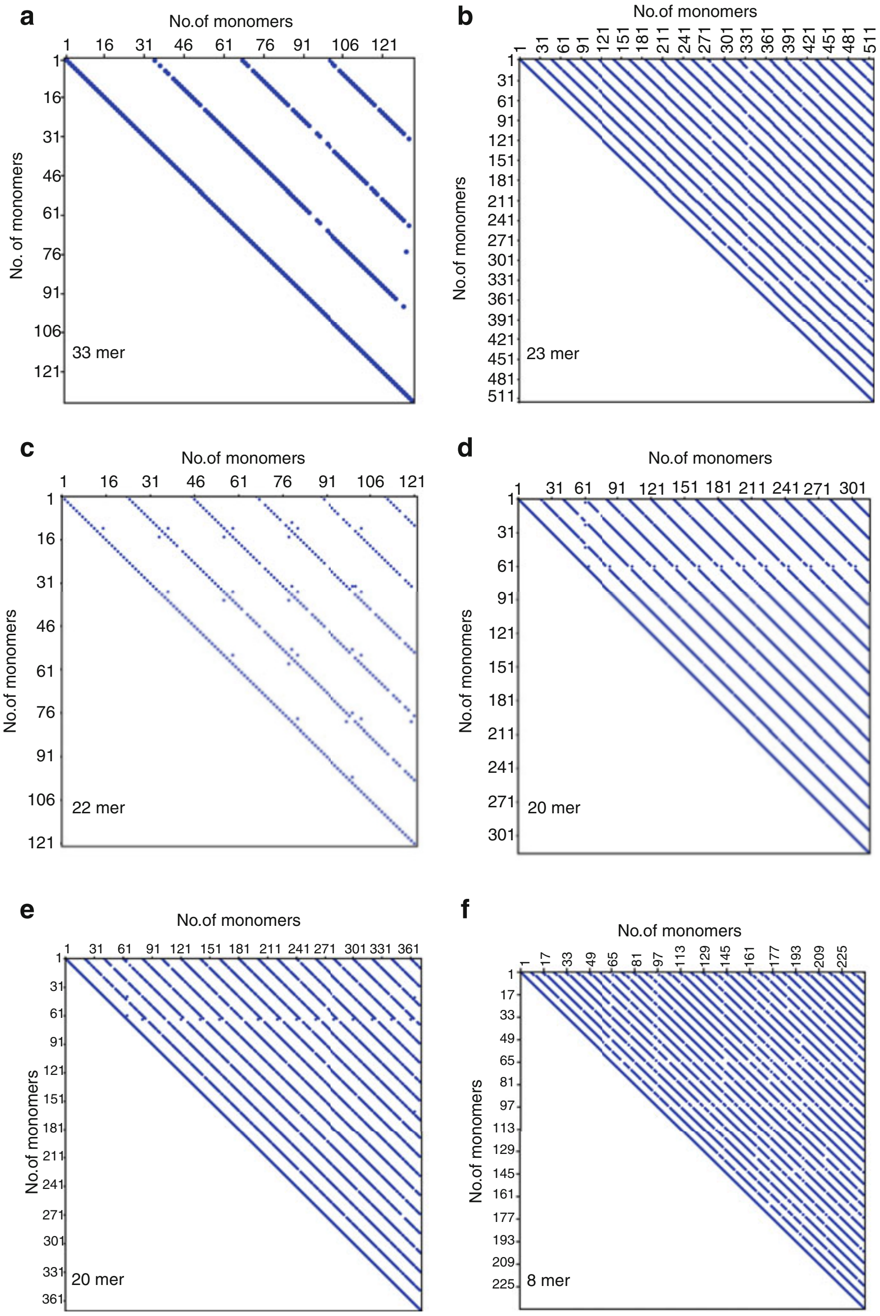


Fig. 8.10 Dot-matrix analysis of alpha satellite HORs identified in human chromosome 21. For each alpha satellite HOR array, identified by the GRM algorithm we computed the corresponding dot-matrix plots (Needleman–Wunsch) to confirm HOR structure. (**a**) 33mer HOR array; (**b**) 23mer HOR array; (**c**) 22mer HOR array; (**d**) 20mer HOR; (**e**) 20mer HOR array; (**f**); 8mer HOR array

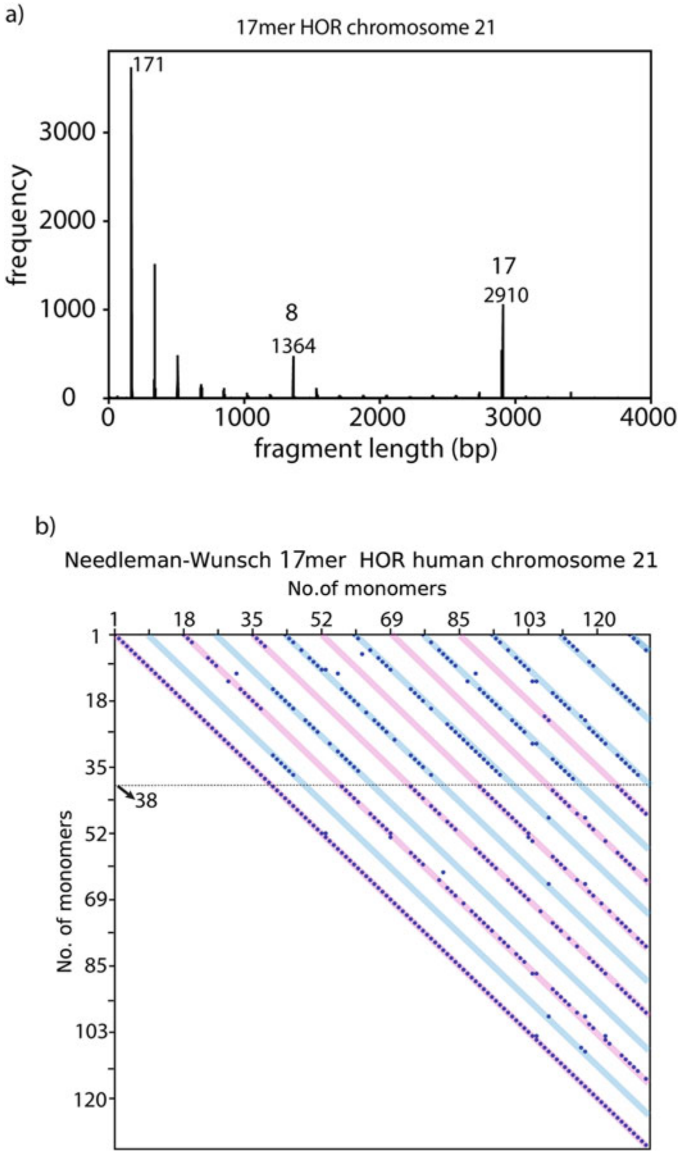


Fig. 8.11 17mer HORs with intertwined 17mer + 8mer HOR pattern in human chromosome 21. (**a**) GRM diagram for alpha satellite HOR in NT_187328.1. There are two pronounced peaks at ~17 × 171 bp and at ~8 × 171 bp. In the case of nonoverlapping HOR sequences, this indicates the presence of 17mer and 8mer HORs. However, if the distances between repeat monomers of the same type have systematically two or more different values, such repetitions lead to the intertwining of repeats, which leads to a more complex pattern. (**b**) However, there is a peculiar underlying partial symmetry seen in the corresponding dot-matrix (both on horizontal and vertical axis monomers are displayed in order of appearance), representing by points the divergence pattern among monomers. Almost all points lie on two sets of off-diagonal straight lines, parallel to the main diagonal. The first set contains the main diagonal and equidistant off-diagonals shifted to the right from the main diagonal by 17, 34, 51 ... (17*k*, *k* = 1, 2, 3 ...) monomers, shown by red straight lines.

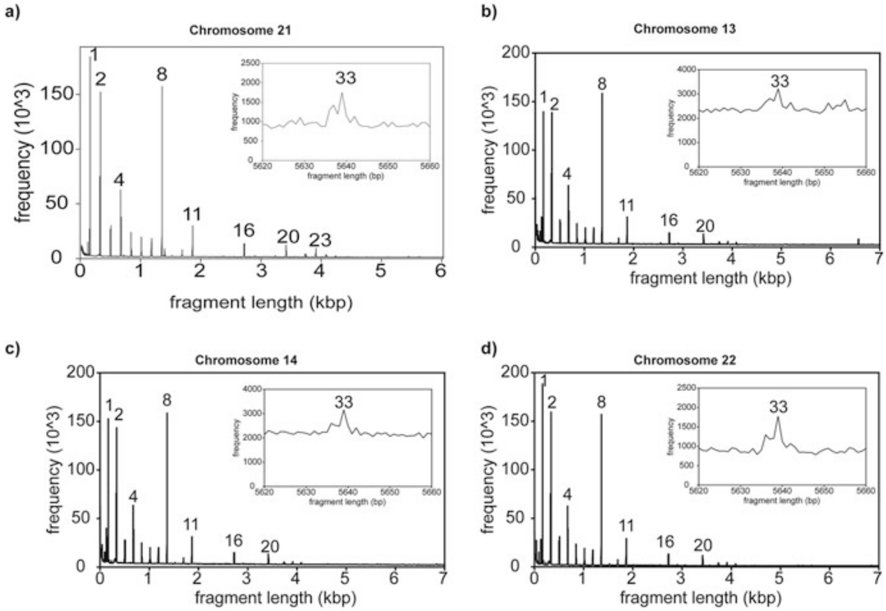


Fig. 8.12 Comparison of GRM diagrams for alpha satellite HORs in centromeres of human chromosomes 21, 22, 13, and 14

from a wide range of species. This expansion of our database will allow us to understand their roles and function and enable us to make conclusions about evolution between closely related and distant species with improvement of technological limitations of assembly of tandem repeats (Vlahović et al. 2020).

Acknowledgments This work was supported by Croatian Science Foundation under the project IP-2019-04-2757.

Fig. 8.11 (continued) The second set of straight lines is obtained from the first set by translating each straight line of the first set to the right by 8 bp. Thus the second set contains off-diagonal straight lines shifted to the right from the main diagonal by 8, 25, 42 ... (8*k*, *k* = 1, 2, 3 ...) monomers shown by blue straight lines. The points representing divergence are to some extent randomly distributed along the two sets of off-diagonals, but with an approximate asymmetry: in the upper part of the matrix ("blue region"), with the first 38 horizontal lines, the majority of points lie on the blue lines, while in the lower part ("red region") a large majority of points lie on the red lines. In the case of full realization of that approximate pattern, the result would be complete 17mer. However, due to the partial violation of this "red-blue region rule" in the upper part ("blue region"), there appears also a characteristic of 8mer

References

Aldrup-Macdonald ME, Sullivan BA (2014) The past, present, and future of human centromere genomics. Genes (Basel) 5:33–50

Alexandrov IA, Mashkova TD, Akopian TA, Medvedev LI, Kisselev LL, Mitkevich SP, Yurov YB (1991) Chromosome specific alpha satellites: two distinct families on human chromosome 18. Genomics 11:15–23

Alexandrov IA, Kazakov A, Tumeneva I, Shepelev V, Yurov Y (2001) Alpha-satellite DNA of primates: old and new families. Chromosoma 110:253–266

Alkan C, Eichler EE, Bailey JA, Sahinalp SC, Tuzun E (2004) The role of unequal crossover in alpha-satellite DNA evolution: a computational analysis. J Comp Biol 11:933–944

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andries V, Vandepoele K, van Roy F (2012) The *NBPF* gene family. In: Shimada H (ed) Neuroblastoma – present and future. InTech, Rijeka, Croatia, pp 185–214. https://doi.org/10.5772/28470

Astling DP, Heft IE, Jones KL, Sikela JM (2017) High resolution measurement of DUF1220 domain copy number from whole genome sequence data. BMC Genomics 18:614

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Brown DD, Wensink PC, Jordan E (1971) A comparison of the ribosomal DNAs of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J Mol Biol 63:57–73

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

Choo KHA (1997) The centromere. Oxford University Press, Oxford

Choo KH, Vissel B, Nagy A, Earle E, Kalitsis P (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. Nucleic Acids Res 19:1179–1182

Davis JM, Searles VB, Anderson N, Keeney J, Dumas L, Sikela JM (2014) DUF1220 dosage is linearly associated with increasing severity of the three primary symptoms of autism. PLoS Genet 10:e1004241

Dover GA (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Dover GA (1986) Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet 2:159–165

Dumas L, Sikela JM (2009) DUF1220 domains, cognitive disease, and human brain evolution. Cold Spring Harb Symp Quant Biol 74:375–382

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM (2007) Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res 17:1266–1277

Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N et al (2012) DUF1220-domain copy number implicated in human brain size pathology and evolution. Am J Hum Genet 91:444–454

Fiddes IT, Pollen AA, Davis JM, Sikela JM (2019) Paired involvement of human-specific Olduvai domains and *NOTCH2NL* genes in human brain evolution. Hum Genet 138:715–721

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T et al (2004) Lineage specific gene duplication and loss in human and great ape evolution. PLoS Biol 2:937–954

Garrido-Ramos MA (2017) Satellite DNA: an evolving topic. Genes 8:230–241

Ge Y, Wagner MJ, Siciliano M, Wells DE (1992) Sequence, higher order repeat structure, and long-range organization of alpha satellite DNA specific to human chromosome 8. Genomics 13:585–593

Glunčić M, Paar V (2013) Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Res 41:e17

Glunčić M, Vlahović I, Paar V (2019) Discovery of 33mer in chromosome 21 – the largest alpha satellite higher order repeat unit among all human somatic chromosomes. Sci Rep 9:12629

Greig GM, Warburton PE, Willard HF (1993) The organization and evolution of an alpha satellite subset shared by chromosomes 13 and 21. J Mol Evol 37:464–475

Haygood R, Babbitt CC, Fedrigo O, Wray GA (2010) Contrasts between adaptive coding and noncoding changes during human evolution. Proc Natl Acad Sci USA 107:7853–7857

Heft IE, Mostovoy Y, Levy-Sakin M, Ma W, Stevens AJ, Pastor S, McCaffrey J, Bofelly D, Martin DI, Xiao M, Kennedy MA, Kwok PY, Sikela M (2020) The driver of extreme human specific Olduvai repeat expansion remains highly active in the human genome. Genetics 214:179–191

Holloway RL (1985) The poor brain of *Homo sapiens* neanderthalensis: see what you please. In: Delson E (ed) Ancestors: the hard evidence. A.R. Liss, Inc, New York, NY, pp 319–324

Jain M, Fiddes JT, Miga KH, Olsen HE, Paten B, Akeson M (2015) Improved data analysis for the MinION nanopore sequencer. Nat Methods 12:351–356

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH (2018) Linear assembly of a human Y chromosome centromere. Nat Biotechnol 36:321–323

Jorgensen AL, Bostock CJ, Bak AL (1986) Chromosome specific subfamilies within human alphoid repetitive DNA. J Mol Biol 187:185–196

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Keeney JG, Dumas L, Sikela JM (2014) The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. Human Neurosci 8:427

Kelso J, Prüfer K (2014) Ancient humans and the origin of modern humans. Curr Opin Genet Dev 29:133–138

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188:107–116

Kochiyama T, Ogihara N, Tanabe HC, Kondo O, Amano H, Hasegawa K, Suzuki H, de Leon MSP, Zollikofer CPE, Bastir M et al (2018) Reconstructing the Neanderthal brain using computational anatomy. Sci Rep 8:6296

Logsdon GA, Vollger MR, Eichler EE (2020) Long-read human genome sequencing and its applications. Nat Rev Genet 21:597–614

Lower SS, McGurk MP, Clark AG, Barbash DA (2018) Satellite DNA evolution: old ideas, new approaches. Curr Opin Genet Dev 49:70–78

Manuelidis L (1978) Complex and simple sequences in human repeated DNAs. Chromosoma 66:1–21

Mashkova TD, Akopian TA, Romanova LY, Mitkevich SP, Yurov YB, Kisselov LL, Alexandrov IA (1994) Genomic organization, sequence and polymorphism of the human chromosome 4 specific alpha satellite DNA. Gene 14:211–217

Miga KH (2015) Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosome Res 23:421–426

Miga KH (2017) The promises and challenges of genomic studies of human centromeres. Prog Mol Subcell Biol 56:285–304

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A et al (2020) Telomere to telomere assembly of a complete human X chromosome. Nature 585:79–84

Mitchell C, Silver DL (2018) Enhancing our brains: genomic mechanisms underlying cortical evolution. Semin Cell Dev Biol 76:23–32

Noonan JP, McCallion AS (2010) Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet 11:1–23

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

Novak P, Avila Robledillo L, Koblizkova A, Vrbova 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:e111

O'Bleness MS, Dickens CM, Dumas LJ, Kehrer-Sawatzki H, Wyckoff GI, Sikela JM (2012) Evolutionary history and genome organization of DUF1220 protein domains. G3 (Bethesda) 2:977–986

O'Bleness MS, Searles VB, Dickens CM, Astling D, Albracht D, Mak ACY, Lai YYY, Lin C, Chu C, Graves T, Kwok PY, Wilson RK, Sikela JM (2014) Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genomics 15:387

Paar V, Pavin N, Rosandić M, Glunčić M, Basar I, Pezer R, Durajlija-Žinić S (2005) ColorHOR – novel graphical algorithm for fast scan of alpha satellite higher-order repeats and HOR annotation for GenBank sequence of human genome. Bioinformatics 21:846–852

Paar V, Basar I, Rosandić M, Glunčić M (2007) Consensus higher order repeats and frequency of string distributions in human genome. Curr Genomics 8:93–111

Paar V, Glunčić M, Rosandić M, Basar I, Vlahović I (2011) Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees. Mol Biol Evol 28:1877–1892

Pech M, Igo-Kemenes T, Zachau HG (1979) Nucleotide sequence of a highly repetitive component of rat DNA. Nucleic Acids Res 7:417–432

Pennacchio LA, Rubin EM (2001) Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2:100–109

Pezer Ž, Brajković J, Feliciello I, Ugarković Đ (2012) Satellite DNA-mediated effects on genome regulation. Genome Dyn 7:153–169

Popesco MC, Maclaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, McGavran L, Wyckoff GJ, Sikela JM (2006) Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. Science 313:1304–1307

Quick VBS, Davis JM, Olincy A, Sikela JM (2015) DUF1220 copy number is associated with schizophrenia risk and severity: implications for understanding autism and schizophrenia as related diseases. Transl Psychiatry 5:e697

Romero V, Nakaoka H, Hosomichi K, Inoue I (2018) High order formation and evolution of hornerin in primates. Genome Biol Evol 10:3167–3175

Rosandić M, Paar V, Basar I (2003) Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7. J Theor Biol 221:29–37

Rosandić M, Paar V, Basar I, Glunčić M, Pavin N, Pilaš I (2006) CENP-B box and pJα sequence distribution in human alpha satellite higher-order repeats (HOR). Chromosom Res 14:735–753

Rudd MK, Willard HF (2004) Analysis of the centromeric regions of the human genome assembly. Trends Genet 20:529–533

Rudd MK, Wray GA, Willard HF (2006) The evolutionary dynamics of alpha-satellite. Genome Res 16:88–96

Ruiz-Ruano FJ, Lopez-Leon MD, Cabrero J, Camacho JPM (2016) High-throughput analysis of the satellitome illuminates satellite DNA revolution. Sci Rep 6:28333

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional human centromere. Science 294:109–115

Sevim V, Bashir A, Chin C-S, Miga KH (2016) Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing. Bioinformatics 32:1921–1924

Sharma D, Isaac B, Raghava GPS, Ramaswamy R (2004) Spectral repeat finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinformatics 20:1405–1412

Sikela JM, van Roy F (2017) Changing the name of the NBPF/DUF1220 domain to the Olduvai domain. F1000Res 6:2185

Singer MF (1982) Highly repeated sequences in mammalian genomes. Int Rev Cytol 76:67–112

Smit A, Hubley R, Green P (2015) RepeatMasker Open-4.0. 2013-2015. Institute for Systems Biology. http://repeatmasker.org

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Southern EM (1975) Long range periodicities in mouse satellite DNA. J Mol Biol 94:51–69

Sullivan LL, Chew K, Sullivan BA (2017) Alpha satellite DNA variation and function of the human centromere. Nucleus 8:331–339

Takaishi M, Makino T, Morohashi M, Huh N (2005) Identification of human hornerin and its expression in regenerating and psoriatic skin. J Biol Chem 280:4696–4703

Turner DJ, Stoddart D, Bulazel KV, Haussler D, Willard HF (2018) Linear assembly of a human Y chromosome centromere. Nat Biotechnol 36:321–323

Tyler-Smith C, Brown WRA (1987) Structure of the major block of alphoid satellite DNA on the human Y chromosome. J Mol Biol 195:457–470

Ugarković D (2005) Functional elements residing within satellite DNAs. EMBO Rep 6:1035–1039

Uralsky LI, Shepelev VA, Alexandrov AA, Yurov YB, Rogaev EI, Alexandrov IA (2019) Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly. Data Brief 24:103708

Vandepoele K, van Roy N, Staes K, Speleman F, Van Roy F (2005) A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. Mol Biol Evol 22:2265–2274

Vandepoele K, Andries V, Van Roy N, Staes K, Vandesompele J, Laureys G, De Smet E, Berx G, Speleman F, Van Roy F (2008) A constitutional translocation t (1; 17) (p36.2;q11.2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes. PLoS One 3:e2207

Vinogradova TV, Zhulidov PA, Illarionova AE, Sverdlov ED (2002) A new family of *KIAA1245* genes with and without *HERV-K LTRs* in their introns. Russ J Bioorg Chem 28:312–315

Vlahović I, Glunčić M, Rosandić M, Ugarković Đ, Paar V (2017) Regular higher order repeat structures in beetle *Tribolium castaneum* genome. Genome Biol Evol 9:2668–2680

Vlahović I, Glunčić M, Dekanić K, Mršić L, Jerković H, Martinjak I, Paar V. (2020) Global repeat map algorithm (GRM) reveals differences in alpha satellite number of tandem and higher order repeats (HORs) in human, Neanderthal and chimpanzee genomes – novel tandem repeat database. In: 2020 43rd international Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 237–242. https://doi.org/10.23919/MIPRO48935.2020.9245278. https://ieeexplore.ieee.org/document/9245278

Warburton PE, Willard HF (1996) Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In: Jackson M, Strachan T, Dover G (eds) Human genome evolution. BIOS Scientific Publishers, Oxford, pp 121–145

Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G (2008) Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics 9:533

Waye JS, Willard HF (1987) Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. Nucleic Acids Res 15:7549–7569

Willard HF (1985) Chromosome-specific organization of human alpha satellite DNA. Am J Hum Genet 37:524–532

Willard HF (1991) Evolution of alpha satellite. Curr Opin Genet Dev 1:509–514

Willard HF, Waye JS (1987) Hierarchical order in chromosome-specific human alpha satellite DNA. Trends Genet 3:192–198

Wolfe J, Darling SM, Erickson RP, Craig IW, Buckle VJ, Rigby PWJ, Willard HF, Goodfellow PN (1985) Isolation and characterization of an alphoid centromeric repeat family from Y chromosome. J Mol Biol 182:477–485

Wu JC, Manuelidis L (1980) Sequence definition and organization of a human repeated DNA. J Mol Biol 142:363–386