321. Biscotti MA, Olmo E, Heslop-Harrison JS. 2015. Repetitive DNA in eukaryotic genomes.

Chromosome Research DOI: 10.1007/s10577-015-9499-z

Author-prepared/archived version of manuscript.

On-line first 29 October 2015.

Repetitive DNA in Eukaryotic Genomes

Maria Assunta Biscotti[1], Ettore Olmo[1] and JS (Pat) Heslop-Harrison[2]

ADDRESSES

[1] Università Politecnica delle Marche, Dipartimento di Scienze della Vita e dell'Ambiente, Ancona 60131, Italy.

[2] University of Leicester, Department of Genetics, Leicester LE1 7RH, UK. Phh4@le.ac.uk

Abstract

Repetitive DNA – sequence motifs repeated hundreds or thousands of times in the genome – makes up the major proportion of all the nuclear DNA in most eukaryotic genomes, However, the significance of repetitive DNA in the genome is not completely understood, and it has been considered to have both structural and functional roles, or perhaps even no essential role. High throughput DNA sequencing reveals huge numbers of repetitive sequences. Most bioinformatic studies focus on low-copy DNA including genes and hence the analyses collapse repeats in assemblies presenting only one or a few copies, often masking out and ignoring them in both DNA and RNA read data. Chromosomal studies are proving vital to examine the distribution and evolution of sequences because of the challenges of analysis of sequence data. Many questions are open about the origin, evolutionary mode and functions that repetitive sequences might have in the genome. Some, the satellite DNAs, are present in long arrays of similar motifs at a small number of sites, while others, particularly the transposable elements (DNA transposons and retrotranposons),

are dispersed over regions of the genome; in both cases, sequence motifs may be located at relatively specific chromosome domains such as centromeres or sub-telomeric regions. Here, we overview a range of works involving detailed characterization of the nature of all types of repetitive sequences, in particular their organization, abundance, chromosome localization, variation in sequence within and between chromosomes, and, importantly, the investigation of their transcription or expression activity. Comparison of the nature and locations of sequences between more, and less, related species is providing extensive information about their evolution and amplification. Some repetitive sequences are extremely well conserved between species, while others are among the most variable, defining differences between even closely relative species. These data suggest contrasting modes of evolution of repetitive DNA of different types, including selfish sequences that propagate themselves and may even be transferred horizontally between species rather than by descent, through to sequences that have a tendency to amplification because of their sequence motifs, to those that have structural significance because of their bulk rather than precise sequence. Functional consequences of repeats include generation of variability by movement and insertion in the genome (giving useful genetic markers), the definition of centromeres, expression under stress conditions, and regulation of gene expression via RNA moieties. Molecular cytogenetics and bioinformatic studies in a comparative context are now enabling understanding of the nature and behaviour of this major genomic component.

Abbreviation list

GC guanine-cytosine content

miRNA microRNA

ncRNA Non coding RNA

NOR Nucleolar organizing region

rDNA ribosomal DNA

satDNA Satellite DNA

TE Transposable element

Repetitive DNA

Sequence motifs that are repeated hundreds or thousands of times in the nuclear genome are an abundant part of all eukaryotic genomes, in most species representing more than half of the total content of the DNA in the cell nucleus (Fig. 1). For many researchers examining genes or aiming to reconstruct long contiguous stretches of low-copy DNA in genomic sequences, repeats provide a challenge to the assembly and they are removed in the analysis. The contributors to the present volume are investigating the nature, behaviour and consequences for the genome of this major component of the genome (Fig. 2), including sequences that are localized or dispersed in the genome, or are expressed, associated with structural regions of chromosomes, or rapidly evolving.

Long before there was any knowledge of DNA sequences, differentially stained regions of chromosomes were recognized in microscope preparations, sometimes following preparation methods involving DNA extraction to give bands, or visualized with particular stains. Many of these domains were classified as heterochromatin. Nucleolar organizing regions (NORs) were also recognized as having different staining properties, related both to associated proteins (seen with silver staining) and a higher GC content than other parts of chromosomes (see with fluorescent stains such as chromomycin A3). The composition of many of these heterochromatin regions was recognized as including hundreds to thousands of copies of the same sequence motif. In the case of the rDNA at the NORs of secondary constrictions, the motif included the 45S rDNA (a unit typically 9kb long comprising the 18S-5.8S-28S rDNA genes and both transcribed and untranscribed spacer regions). Other major heterochromatin bands were identified around centromeres and telomeres in many organisms, and found to be composed of huge numbers of copies of the same sequence, with a much shorter repeat unit length. Often, the repeat length of such

satellite DNA sequences is related to the folding of the DNA around nucleosomes, 150 base pairs plus the linking DNA, giving a length of 160 to 180 bases. As DNA technologies advanced, DNA fractions including repeats were identified by high-speed density gradient centrifugation of bulk DNA. Because the repetitive DNA was abundant and had a different AT/GC content to the bulk DNA, it formed a satellite band, giving rise to the phrase ‘satellite DNA’ (satDNA), and these bands were extracted and used for the first in situ hybridizations to show their chromosomal locations (Gall and Pardue 1969; John et al. 1969). After the development of restriction enzyme technology, satellite DNA motifs were identified by restriction enzyme digests of DNA followed by size separation by agarose gel electrophoresis. Bulk genomic DNA, with restriction sites randomly distributed through the DNA sequence gives a diffuse smear (or sometimes a largely uncut, high molecular weight band) which has superimposed on it sharper, abundant bands. Some of these bands come from mitochondrial or chloroplast DNA, but many represent highly repeated units of tandemly repetitive sequences defined by the presence of flanking restriction sites, sometimes named “restriction satellite DNA” but mostly shortened to satDNA. DNA sequencing showed that many sequences were highly abundant in the genome, and the combination of membrane and in situ hybridization showed those that were located at discrete sites in the genome as tandemly repeated satDNA. With high-throughput DNA sequencing, there are many bioinformatic approaches to identify abundant satDNA sequences; in particular, the use of programs to find k-mer motifs (DNA sequences of k nucleotides long, where k is typically 16 to 128) where some motifs are far in excess of the expected random number, are finding many repetitive DNA motifs. The graph-based sequence clustering tool RepeatExplorer (Novak et al. 2013) is also proving valuable to identify high and medium copy repeats in genome sequence data. Moreover given the sequence homogeneity of satDNA repeats, the assembly of these sequences into long stretches still represents a challenge; technologies are now becoming available to enable longer reads of several kilobases to be made (eg those from PacBio or Nanopore Minion), but still cannot read continuous stretches of

millions of bases of DNA. The major unassembled component of the reference human genome, one of the most studied, is made up of satDNA. The review by Miga (2015) highlights strengths and weaknesses of using low-copy sequence variants as array markers to support the assembly procedure. Strategies to integrate satellite sequences in genomic studies are also discussed by Miga.

Mutations in sequence of satDNAs within and between arrays of the tandem repeats, and the amplification or contraction of specific arrays and their unit variants can be seen as changes in satDNA profiles; given the abundance of such sequences their abundance, they have been considered to affect the karyotype and hence play a role in the speciation. Data obtained by Paço and colleagues (2015) in two hamster species of the genus *Phodopus* gives evidence that the high molecular dynamics of repetitive sequences are responsible for a chromosomal instability which led to chromosome rearrangements, leading to the extensive divergence of the karyotypes of these rodent species from the ancestral Muroidea. The consequent genetic diversity potentially helps to allow species to survive to environmental changes. In another comparative study, Rojo and colleagues (2015) analysed the intra- and interspecific variability, the genomic organization and the chromosomal distribution of two satDNA families in eight reptilian species of the genus *Iberolacerta*. The evidence obtained showed common and non-common traits in the evolutionary histories of the satDNAs. The different sequence homogenization could be related to different turnover rates which in turn might be due to differences in the karyotypes that may prevent interchromosomal exchanges.

An increasing number of studies are studying the transcriptional activity and dynamics of satDNA, formerly considered an inactive portion of the genome. Biscotti et al. (2015) present an overview of the structural roles of small ncRNAs deriving from satellite DNA: in the formation and maintenance of heterochromatin at centromere and telomere with consequent impact on chromosome integrity and genome stability, in centromere identity, and elongation, capping and replication of telomere. Moreover, the aberrant expression of these sequences has been recorded in

several cellular pathways and biological processes highlights their association with different diseases, with a particular relevance to cancer as reviewed in the work by Ferreira et al. (2015).

Centromeric and pericentromeric satellite ncRNAs are involved also in many aspects of vertebrate embryogenesis but less is known about subtelomeric repeat transcription at this developmental stage. The analyses performed by Trofimova and colleagues (2015) on the distribution of PO41 at different embryonic stages and cell types in *Gallus gallus* revealed an extremely uniform pattern, indicating universal functions of RNA derived from subtelomeric repeats.

Apart from the satDNAs, a second major grouping of repetitive DNA is represented by transposable elements (TEs), elements that amplify and reinsert into the nuclear genome. The DNA transposons, or class II transposable elements, move and amplify through DNA, while class I transposable elements or retrotransposons amplify through an RNA intermediate. Because of their mode of amplification, transposons are often dispersed in the genome, although some may show either a concentration or depletion in particular chromosomal regions such as centromeres or sub-telomeric sites. In various genomes, the nature of the elements present differs: some classes predominate in one species, while another class may be much more abundant in another species. The review by Warren and colleagues (2015) examines the lineage-specific diversity of TEs in terms of composition, content, and age of mobile elements in major vertebrate lineages. They also looked at the role of TE in genome plasticity and the effects on the evolution of the host genome at various levels contributing to the evolution of key innovations responsible for the diversity and success of this taxon.

The computational analysis performed by Scarpato et al. (2015) suggests that TEs in conjunction with miRNA might constitute a gene regulatory network in vertebrates. In particular, they have analysed the structural relationship between V-SINE harboured in the 3’ UTR of several genes and specific miRNAs in *Danio rerio.* Potentially, these mobile elements are targets for

miRNAs. This interaction might explain the preservation over very long evolutionary time of the V-SINE elements in *Anamnia* and might have been responsible for genomic innovations that occurred in this taxon.

Sex chromosomes in both plants and animals have been found to have somewhat different repetitive DNA sequence types to autosomes, presumably at least partially related to their different pairing and recombination behaviour. Chalopin and colleagues (2015) investigate fish, organisms having relatively young sex chromosomes, and evaluated the impact of all the mobile elements – sometimes refered to as the mobilome – on sex chromosome evolution in vertebrates. The recent amplification of specific TEs in sex determination regions in some fish has been related to the suppression of recombination and differentiation of the sex chromosomes. Moreover the accumulation of mobile elements in these regions has been hypothesized to play a role in spreading putative transcription binding sites for sex-specific regulation networks or in dosage compensation processes. Hozba and colleagues (2015) reviewed the evolutionary impact of not only TEs but also tandem repeats on sex chromosome in plants. Expansion and contraction of repetitive DNA could contribute to the formation of the heteromorphic sex chromosomes. Moreover also the mechanisms by which these sequences spread show different intensity if compared to recombining regions of the genome.

TE content is variable also in plants influencing their genome size, karyotypes, and defining differences in the genomes among related species. These genomes may come together in forming new polyploid species. Santos and colleagues (2015) investigated the role of retrotransposon expansions or contractions in genome evolution in the genus *Brachiaria* by identifying characteristic motifs in transcriptome data, and finding the chromosomal distribution and abundance of these sequences on chromosome sets in diploid and polyploid species. Results suggest that amplification of different retroelements has an impact on genome behaviour and on genome divergence leading to speciation.

Although TEs and satDNAs present differences in structure, genomic organization, spreading mechanisms and evolutionary dynamics, several studies are highlighting that they are mutually related. Indeed, as reviewed in the paper by Meštrović et al. (2015) the homology between satDNA and mobile elements led to support the hypothesis that tandemly arranged repeat sequences can be derived from transposons and/or retrotransposons either by ectopic recombination or by the amplification of pre-existing internal repeats. Also, the opposite evolutionary movement may occur, since tandem repeat fragments can be found as part of TEs, more often in DNA transposons than retrotransposons, both in animals and plants. The direction of transition from TE to satDNA might depend on transposition rate which increases until an optimal number of internal repeats is reached, after which it decreases; the opposite transition is linked to recombination rate which decreases gradually with the number of repeats.

The paper by Dias and colleagues (2015) firstly reports the emergence of satDNA from central tandem repeats of a helitron (DINE-TR1) in three *Drosophila* species and this event occurred independently in the lineages examined. In other *Drosophila* species the amplification of the central tandem repeats was not observed. This might indicate that the amplification is at an early stage or other mechanisms/forces are operating to prevent the increase of tandem array length. In situ hybridization shows that the helitron localizes at transitional β-heterochromatin regions. This might be a result of a preferential insertion in these loci or alternatively selection-pressure is less strong against ectopic recombination. Small RNA transcripts originating from DINE-TR1 are expressed in male and female gonads of *D. virilis*, suggesting their involvement in the piRNA pathway, so ensuring a faithful gametogenesis and preventing deleterious transpositions in germ line cells. At local and genome-wide scales, these small RNAs interacting with PIWI proteins might also play regulatory roles affecting the heterochromatic state.

Some transposable element sequences are G-rich, and, as shown in both human and plants, can form structures made of four DNA strands known as G-quadruplexs. Kejnovsky and colleagues (2015)

have discussed the role of these structures in TE life-cycle, on replication, transcription, translation, chromatin status and recombination. The authors proposed that TEs are responsible for the genomic spreading of G-quadruplexes contributing in the evolution of cellular regulatory network.

The establishment and maintenance of nuclear heterochromatin compartments and in particular the peripheral heterochromatin underlying the nuclear envelope, perinucleolar heterochromatin and chromocenters are related to the organization of chromosome territories during interphase (Pombo and Dillon, 2015). Many questions on how tandem repeats contained in these regions affect the organization of chromosome territories in interphase are still open. Maslova and colleagues (2015) analysed the distribution of several types of tandem repeats in relation to large-scale heterochromatin domains, gene-dense and gene-poor chromosomes in interphase nuclei of chicken MDCC-MSB1 cells and somatic tissues. Telomere and subtelomere repeats localize at the nuclear or chromocenters periphery, while the CNM centromeric repeat, localized in regions of gene-dense microchromosomes, forms interchromosome clusters and occupies DAPI-positive chromocenters predominantly present within the nuclear interior. Centromere-specific tandem repeats of the majority of gene-poor macrochromosomes are localized into the peripheral layer of heterochromatin.

Conclusions

Chromosomal studies, and particularly the application of in situ hybridization, have been able to localized repetitive DNA sequences on chromosomes, and show the distribution of the various classes. DNA sequencing has been able to identify repetitive DNA elements, but bioinformatic approaches to assembly have been weak, so there have been relatively few studies compared to whole-genome assemblies of low-copy sequences. As shown in the several articles in this special issue, the chromosome-centric approach is proving vital to examine the distribution and evolution of repeats identified in sequence data. These chromosomal studies are showing the origin,

evolutionary mode and function of repetitive sequences in the genome. The investigation of their transcription or expression activity is giving novel insight into their consequences. Comparison of the nature and locations of sequences between more, and less, related species is providing extensive information about their evolution and amplification. Clearly, contrasting modes of evolution are found for different sequences and different chromosomes (particularly the sex chromosomes), and there may be selfish sequences that propagate themselves and may even be transferred horizontally between species rather than by descent, through to sequences that have a tendency to amplification because of their sequence motifs, to those that have structural significance because of their bulk rather than precise sequence. Molecular cytogenetics studies in a comparative context are now enabling understanding of the nature and behaviour of this major genomic component.

References

Biscotti MA, Canapa A, Forconi M, Olmo E, Barucca M (2015) Transcription of tandemly repetitive DNA: functional roles. Chromosome Res DOI 10.1007/s10577-015-9494-4

Chalopin D, Volff JN, Galiana D, Anderson JL, Schartl M (2015) Transposable elements and early evolution of sex chromosomes in fish. Chromosome Res DOI 10.1007/s10577-015-9490-8

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. Chromosome Res DOI 10.1007/s10577-015-9480-x

Ferreira DP, 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. Chromosome Res DOI 10.1007/s10577-015-9482-8

Gall JG, Pardue ML (1969) Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proc Natl Acad Sci USA 63:378–383

Heslop-Harrison JS, Schmidt T (2012). Plant Nuclear Genome Composition. In: Encyclopedia of Life Sciences. eLS 2012, John Wiley & Sons Ltd: Chichester http://www.els.net/ http://dx.doi.org/10.1002/9780470015902.a0002014.pub2

Hozba R, Kubat Z, Cegan R, Jesioneck W, Vyskot B, Kejnovsky E (2015) Impact of repetitive DNA on sex chromosome evolution in plants. Chromosome Res submitted

John HA, Birnstiel ML, Jones KW (1969) RNA-DNA hybrids at the cytological level. Nature 223:582

Kejnovsky E, Tokan V, Lexa M (2015) Transposable elements and G-quadruplex. Chromosome Res DOI 10.1007/s10577-015-9491-7

Maslova A, Zlotina A, Kosyakova N, Sidorova M, Krasikova A (2015) Three-dimensional architecture of tandem repeats in chicken interphase nucleus. Chromosome Res DOI 10.1007/s10577-015-9485-5

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M (2015) Structural and functional liaisons between transposable elements and satellite DNAs. Chromosome Res DOI 10.1007/s10577-015-9483-7

Miga KH (2015) Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosome Res DOI 10.1007/s10577-015-9488-2

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.

Paço A, Ageda F, Meštrović N, Plohl M, Chaves R (2015) The puzzling character of repetitive DNA in Phodopus genomes (Cricetidae, Rodentia). Chromosome Res DOI 10.1007/s10577-015-9481-9

Pombo A, Dillon N (2015) Three-dimensional genome architecture: players and mechanisms. Nat Rev Mol Cell Biol 16: 245–257.

Rojo V, Martínez-Lage A, Giovannotti M, González-Tizón AM, Nisi Cerioni P, Caputo Barucchi V, Galán P, Olmo E, Naveira H (2015) Evolutionary dynamics of two satellite DNA families in Rock lizards of the *Iberolacerta* (Squamata, Lacertidae): different histories but common traits. Chromosome Res DOI 10.1007/s10577-015-9489-1

Santos Carvalho F, Guyot R, Borges do Valle C, Chiari L, Techio VH, Heslop-Harrison P, Vanzela Laforga AL (2015) Chromosomal distribution a 1 nd evolution of abundant retrotransposons in plants: *gypsy* elements in diploid and polyploid *Brachiaria* forage grasses. Chromosome Res DOI 10.1007/s10577-015-9492-6

Scarpato M, Angelini C, Cocca E, Pallotta MM, Morescalchi MA, Capriglione T (2015) Short interspersed DNA elements and miRNAs: a novel hidden gene regulation layer in zebrafish? Chromosome Res DOI 10.1007/s10577-015-9484-6

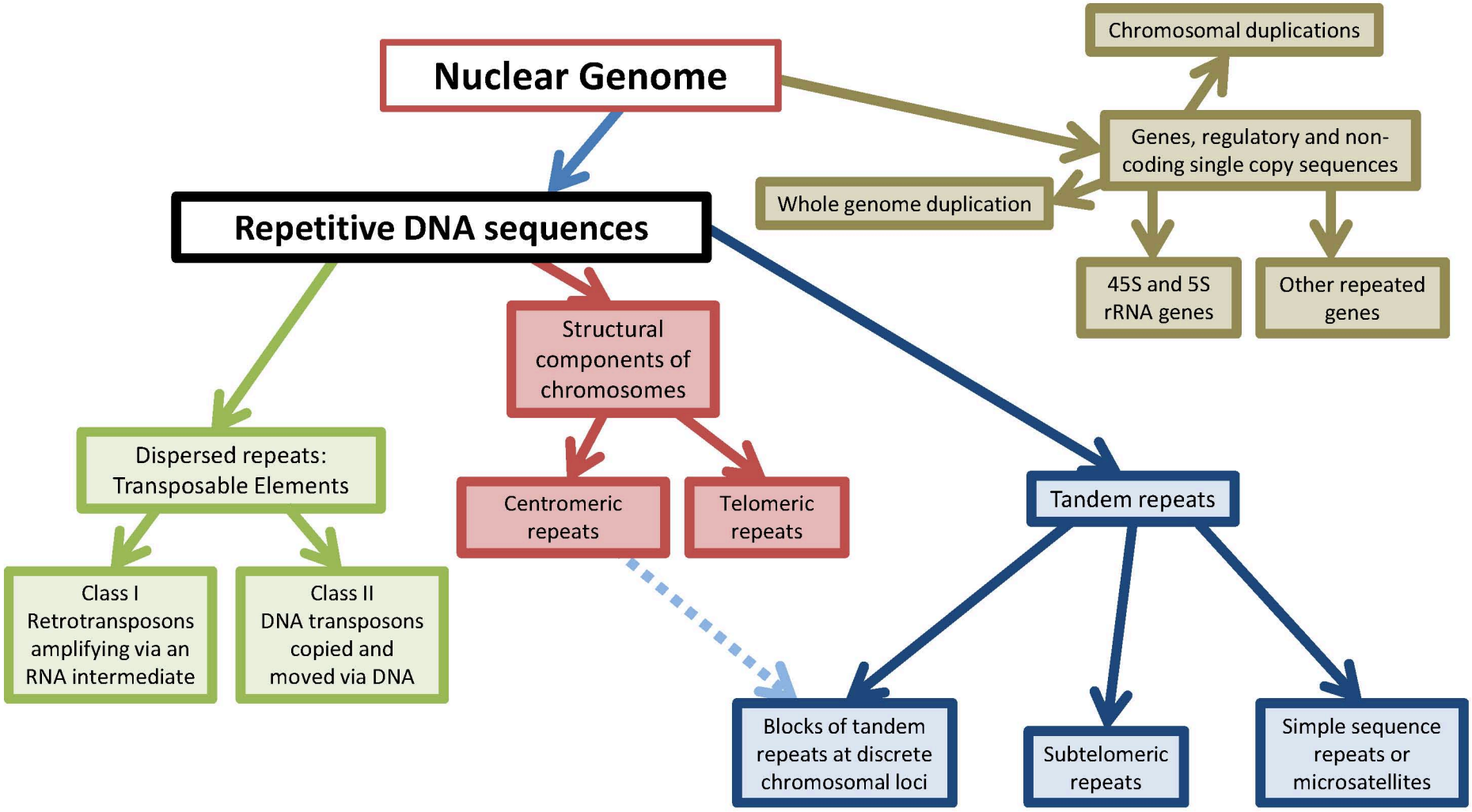
Trofimova I, Chervyakova D, Krasikova A (2015) Transcription of subtelomere tandemly repetitive DNA in chicken embryogenesis. Chromosome Res DOI 10.1007/s10577-015-9487-3

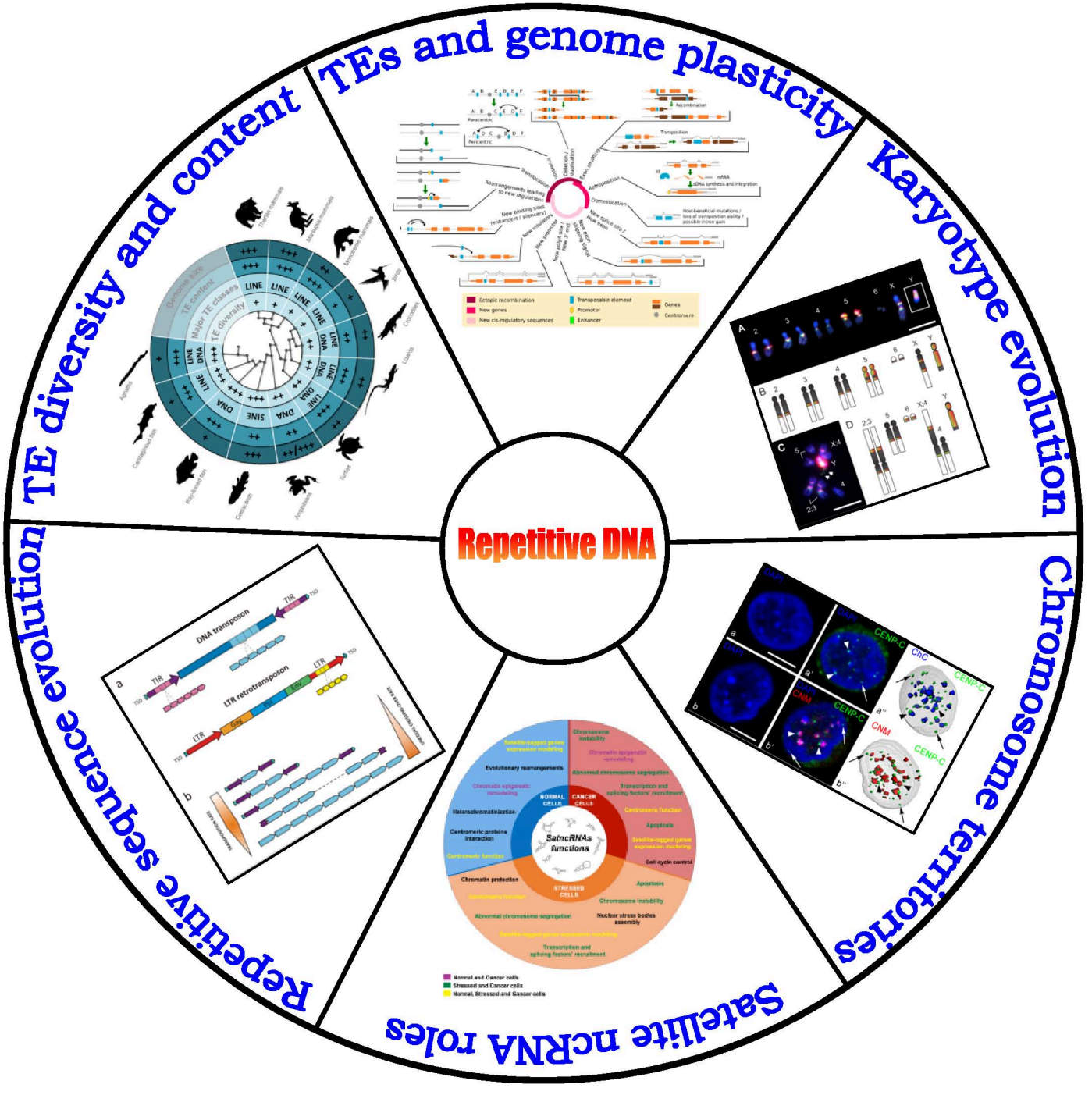
Warren I, Naville M, Chalopin D, Levin P, Berger C, Galiana, D, Volff JN (2015) Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. Chomosome Res DOI 10.1007/s10577-015-9493-5

LEGENDS TO FIGURES

Figure 1. Major divisions of repetitive DNA sequences in the eukaryotic genome. Dispersed repetitive sequences may be widespread over the genome, located over broad regions of one of more chromosomes. Tandem repeats tend to be located in blocks at one or more locations on chromosomes. (After Heslop-Harrison and Schmidr, 2012)

Figure 2. Consequences of repetitive DNA sequences for eukaryotes. The chapters of this Special Issue discuss the significances for genome behaviour and evolution of various classes of satDNA and transposable elements.


Nuclear Genome
Chromosomal duplications
Genes, regulatory and non-coding single copy sequences
Whole genome duplication
45S and 5S rRNA genes
Other repeated genes
Repetitive DNA sequences
Structural components of chromosomes
Dispersed repeats: Transposable Elements
Centromeric repeats
Telomeric repeats
Tandem repeats
Class I Retrotransposons amplifying via an RNA intermediate
Class II DNA transposons copied and moved via DNA
Blocks of tandem repeats at discrete chromosomal loci
Subtelomeric repeats
Simple sequence repeats or microsatellites


TEs and genome plasticity
Karyotype evolution
Chromosome territories
Satellite ncRNA roles
Repetitive sequence evolution
TE diversity and content
Repetitive DNA
Ectopic recombination
New genes
New cis-regulatory sequences
Transposable element
Promoter
Enhancer
Genes
Centromere
DNA transposon
LTR retrotransposon
SatncRNAs functions
NORMAL CELLS
CANCER CELLS
STRESSED CELLS
Normal and Cancer cells
Stressed and Cancer cells
Normal, Stressed and Cancer cells