University of Trás-os-Montes and Alto Douro

HOW SATELLITE DNA COMES INTO PLAY IN GENOMES?

A RODENTIA CELL-MODEL APPROACH

PhD Thesis in
Technologic, Comparative and Molecular Genetics

Ana Cristina Mendes da Silva

Supervisors:

Professora Doutora Raquel Maria Garcia dos Santos Chaves
Professora Doutora Maria Filomena Lopes Adega

VILA REAL, 2020

University of Trás-os-Montes and Alto Douro

HOW SATELLITE DNA COMES INTO PLAY IN GENOMES?
A RODENTIA CELL-MODEL APPROACH

PhD Thesis in
Technologic, Comparative and Molecular Genetics

Ana Cristina Mendes da Silva

Supervisors:

Professora Doutora Raquel Maria Garcia dos Santos Chaves
Professora Doutora Maria Filomena Lopes Adega

Jury Composition:

Professor Doutor Jorge Ventura Ferreira Cardoso
Professor Doutor António Manuel Amorim dos Santos
Professora Doutora Raquel Maria Garcia dos Santos Chaves
Professor Doutor Agostinho Antunes Fernandes
Professora Doutora Estela Maria Bastos Martins de Almeida

VILA REAL, 2020

I declare for all due purposes that the PhD thesis meets the technical and scientific standards required by the regulations of the University of Trás-os-Montes and Alto Douro. The presented doctrines are the exclusive responsibility of the author.

This thesis was specifically prepared to obtain the PhD degree in Technologic, Comparative and Molecular Genetics.

Financial support provided by FCT ("Fundação para a Ciência e a Tecnologia") and POPH-QREN/FSE ("Programa Operacional Potencial Humano-Quadro de Referência Estratégico Nacional/Fundo Social Europeu"), PhD Grant SFRH/BD/80808/2011.

“Success is not final, failure is not fatal:
it is the courage to countinue that counts”
Sir Winston Churchill

To my Father

ACKNOWLEDGMENTS

Um projeto de Doutoramento, de nominação singular, é na sua essência um conjunto de contribuições quer físicas quer emocionais de muitas pessoas que permitiram o seu desenvolvimento e concretização. Desta forma, gostaria de exprimir o meu sincero agradecimento a todas as pessoas e instituições essenciais neste percurso.

Institucionalmente, por todos os esforços realizados para providenciar as melhores condições para o desenvolvimento e realização deste trabalho, gostaria de agradecer à Universidade de Trás-os-Montes e Alto Douro, na pessoa do Reitor Professor Doutor António Augusto Fontainhas Fernandes, à Escola de Ciências da Vida e do Ambiente (UTAD) na pessoa do seu Presidente Professor Doutor Artur Agostinho de Abreu e Sá e ao Presidente do Conselho Científico da mesma, Professor Doutor Luís Herculano Melo de Carvalho, ao Departamento de Genética e Biotecnologia, ao BioISI – "Biosystems & Integrative Sciences Institute" (Universidade de Lisboa) e ao laboratório de Citogenómica e Genómica Animal do grupo de "Gene Expression Regulation" (BioISI), na pessoa da Professora Doutora Margarida Gama-Carvalho.

Pelo apoio financeiro através da atribuição de uma bolsa de doutoramento, gostaria de agradecer à "Fundação para a Ciência e a Tecnologia" (FCT) e ao "Programa Operacional Potencial Humano/Fundo Social Europeu" (POPH/FSE).

À Diretora do 3º Ciclo em Genética Molecular Comparativa e Tecnológica, na pessoa da Professora Doutora Paula Filomena Martins Lopes, muito obrigada por todo o seu apoio ao longo do meu percurso acadêmico, científico e pessoal.

À Professora Doutora Raquel Chaves, agradeço por me ter aceite no laboratório e por ser minha orientadora. O meu agradecimento sincero por todo o apoio quer a nível científico quer a nível pessoal, os conhecimentos e saberes transmitidos desde que ingressei no ensino superior, pela disponibilidade, pela competência, e por toda a ajuda na realização deste projeto. Durante este projeto, passei por uma das maiores provas de superação de vida, e agradeço profundamente todo o apoio que me deu e todas as opções e oportunidades que me proporcionou. Muito obrigada por tudo!

À Doutora Filomena Adega, também minha orientadora, agradeço por todo o apoio, disponibilidade, ensinamentos e amizade. Obrigada por todas as cantorias e gargalhadas! Agradeço pela forma que sempre me apoiou e aconselhou ao longo destes anos, pelo seu exemplo e pela constante disponibilidade. Obrigada por todo o apoio nos bons e, sobretudo, nos maus momentos. Obrigada "Meníssima" por tudo!

À Professora Doutora Margarida Gama-Carvalho, agradeço por toda a colaboração e disponibilidade na execução de todo o trabalho de análise *in silico*. Também, pela forma calorosa que me acolheu no laboratório. Agradeço também a toda a equipa, em especial ao Hugo Santos por toda a simpatia, disponibilidade, por todo o tempo que passamos a fazer a análise *in silico* e por todas as (longas!) conversas *skype*.

Ao grupo no qual me integrei quando iniciei a minha bolsa de doutoramento, Ana Luísa Borges, Jorge Pereira, Sara Santos, Sandra Louzada, Ana Neta, Ana Paço, Susana Meles, Daniela Ferreira e Ana Escudeiro, o meu muito obrigada por todo o apoio e pelos bons momentos que vivemos. À Lú agradeço toda a amizade e apoio incondicional durante o primeiro ano de doutoramento. Obrigada por não me fazeres desistir! À Sú, agradeço toda a amizade e todo o apoio que sempre me deu! Obrigada por continuares a fazer parte da minha vida mesmo com um "mundo de Km" entre nós! À Daniela Ferreira e Ana Escudeiro, minhas colegas de laboratório com quem partilhei todos os momentos passados ao longo dos 4 anos de doutoramento, obrigada por toda a ajuda, pelo apoio e todos os momentos que passámos juntas.

À Patrícia Silva, Vera Cardoso e Richard Gonçalves, com quem partilhei os melhores momentos *"out-of-office"* agradeço todas as gargalhadas! À Patrícia Silva, minha amiga desde o primeiro dia de licenciatura, agradeço por toda amizade, companheirismo e apoio. Obrigada por estares sempre presente na minha vida, por seres uma das minhas melhores amigas e por me demonstrares constantemente a tua amizade e amor.

À Leonor Pereira, Márcia Carvalho e Vanessa Ferreira, obrigada por todos os momentos que passamos juntas, todo o companheirismo e apoio ao longo deste meu percurso.

À minha querida Sónia Gomes, minha grande amiga e companheira de todos os momentos. Obrigada por todas as viagens diárias partilhadas, por todas as conversas e ideias malucas que tivemos (temos!) e por todos os momentos de pura felicidade. Obrigada por todos o apoio incondicional ao longo do doutoramento e por me acompanhares nesta minha nova aventura. Contigo partilhei o pior e o melhor que a vida me tem dado nos últimos anos, e tu, sempre abriste os braços quer para não me deixar cair quer para me levantar no ar. Obrigada por seres quem és e por existires na minha vida!

À "La Equipa" Nuno, Diogo, Carol, Henrique, Lili, Cláudio e Ângela agradeço toda a amizade, todo o apoio, companheirismo e todos os momentos que passamos juntos. Estar com vocês é Vida!

Neste último ano, conheci pessoas fabulosas com quem tenho partilhado a paixão pela Medicina. À Sara, Natália, Joana, Ana e Mariana, minhas grandes companheiras nesta viagem, agradeço todos os momentos que temos vivenciado. À Sara e Nats, pelos momentos incríveis que temos vivido, pelas histórias que temos já para contar, e, acima de tudo, por sermos verdadeiras compinchas nesta nossa caminhada!

À minha família, a melhor que poderia ter, obrigada por sermos a família que somos! À minha avó, meus tios e tias, meus primos e primas, obrigada por sermos a grande e unida família que somos. Obrigada por sermos um só, por nos apoiarmos uns aos outros e estarmos sempre presentes para sermos a muleta de quem precisa. Acima de tudo, agradeço todo o apoio que sempre me deram ao longo do doutoramento e, mais recentemente, por me apoiarem na luta pelos meus sonhos! Vocês são os melhores!

Aos meus Pais e irmão, obrigada por estarem sempre comigo e por me ensinarem diariamente o verdadeiro significado do Amor. Não existem palavras suficientes para exprimir o quanto vos amo e o quanto estou grata por todo o apoio e companheirismo em todas as decisões na minha vida. À minha Mãe, uma verdadeira força da Natureza, obrigada por seres minha amiga, minha confidente e acima de tudo, me mostrares o quanto podemos superar na vida contra todas as adversidades. Tu és minha e eu sou tua!

Ao meu Pai, que apesar de não estares connosco presencialmente, enches os nossos dias com as mais maravilhosas memórias. Apoiaste-me, deste-me o ombro e foste sempre o meu mais fiel confidente. Agradeço diariamente o privilégio que tive em acompanhar-te em toda a tua luta e todos os momentos em que estive contigo, para ti. Não ter a tua presença nesta reta final é sem dúvida a parte mais difícil de todo este percurso. Obrigada por me teres dito o quanto orgulho tens em mim. Tu és meu e eu sou tua!

E por último, mas não menos importante, ao Filipe. Não existem palavras que consigam descrever todo o apoio e amor incondicional que sempre me deste. Viste-me crescer, entrar para a universidade, abraçar o mundo da investigação e entrar num doutoramento. Nos entretantos, casámos, iniciámos uma linda família, viste o meu mundo desabar e ergueste-me. Eu quis seguir os meus sonhos e tu vieste comigo, sem nunca olhar para trás e sem nunca me questionares. Agradeço-te por seres o meu pilar e o meu porto de abrigo. Obrigada Meu Amor!

Sou uma verdadeira privilegiada em ter os melhores dos melhores!
Obrigada!

ABSTRACT

Satellite DNA (satDNA) sequences constitute the major component of constitutive heterochromatin (CH) and have been considered one of the most fascinating and intriguing repetitive DNA elements of eukaryotic genomes. For many years, satDNA was considered "junk" and a transcriptional inert fraction of eukaryotic genomes. Today is generally accepted that satDNAs play important structural and functional roles in genomes, such as genome architecture, chromosomal reorganization during evolution, or genome regulation, mainly driven by satellite transcripts or satellite non-coding RNAs (satncRNAs). Centromeric satncRNAs have been highlighted as crucial players in remodeling/CENP-A deposition and correct kinetochore assembly, essential for proper chromosome segregation.

The advent of Next Generation Sequencing (NGS) strategies and the overwhelming advances in genome sequencing technologies have provided a massive amount of sequencing data from hundreds of model and non-model species. Similarly, a growing in bioinformatics tools and strategies have been established towards genome-wide identification and characterization of the repetitive genome elements, namely satDNAs – the Satellitome.

In the last decades, rodents belonging to the *Peromyscus* genus have emerged as model systems across a variety of scientific disciplines, including chromosomal evolution, and the release of the first *Peromyscus* representative genome sequence (from *P. maniculatus*) was the gateway to further dissect the repetitive content of this genome. A bioinformatics pipeline was thus defined in this work, based on the Tandem Repeats Finder algorithm and an integrated analysis of sequence similarity allowed the identification of 21 distinct families of large tandem repeats in *Peromyscus maniculatus* genome (array length larger than 2 kb) being the majority of these satellite- or transposable elements-related families, presenting a tandem organization. Two orthologous satDNA families of the rat and mouse genomes were recognized for the first time in *P. maniculatus* genome: RNSAT1 and MMSAT4, respectively. The most prevalent satDNA family of the *P. maniculatus* satellitome corresponded to the previously described *Peromyscus* satDNA – PMSat –, an AT-rich satDNA displaying a 345 bp monomeric size. Physical mapping of PMSat conducted in four *Peromyscus* species (*P. eremicus*, *P. maniculatus*, *P. leucopus* and *P. californicus*) revealed that PMSat is mainly located at the active centromeres and pericentromeric regions of all chromosomes, and at other constitutive heterochromatin rich regions as telomeres and p-arms of some chromosomes. In all the studied species, PMSat showed a high degree of nucleotide conservation, despite the different number of PMSat copies *per* genome. Our results strongly suggest that the evolution of PMSat was driven by copy number fluctuations and the high similarity among *Peromyscus* and non-*Peromyscus* species may reflect non-concerted evolutionary events. Also, in light of the karyotype differences of these species, as well as many of the chromosome polymorphisms found in *Peromyscus* species, the distinct pattern of CH, and PMSat locations, we hypothesized that PMSat evolutionary molecular events may have promoted *Peromyscus* karyotype variations and genome evolution. Furthermore, the PMSat copy number fluctuations, promoted by molecular mechanisms such as unequal crossing-over and rolling circle amplification are clearly observed in the heterochromatin additions found in some of the genus species, namely on *P. eremicus* genome.

The transcriptional analysis of PMSat in proliferative cells from all the studied *Peromyscus* species uncovered a positive correlation between PMSat expression and DNA copy number in each genome. Despite the pronounced variation levels of transcripts, the analysis of specific cell cycle phases revealed a similar transcriptional cellular profile throughout the cell cycle: PMSat satncRNA accumulates mostly at G2/M transition and at the mitosis onset and are restricted to the nucleus. To gain more insights on the putative function(s) of PMsat transcripts, a functional assay based on PMSat RNA knockdown on *P. eremicus* proliferative cells anticipated its potential role as key players on kinetochore assembly and centromeric function. Moreover, according to the putative transcription factors' binding sites on PMSat monomer sequence found, RNA polymerase II may be the enzyme conducting the transcription of this satellite family in a variety of cell conditions, namely in response to cellular stresses.

The work presented on this thesis uncovered PMSat not only as the trigger of *Peromyscus* karyotype evolution but also as a crucial element of the centromeric function and chromosome segregation fidelity, that seems to be conducted by their derived satncRNAs.

Keywords: *Peromyscus*, satellitome, PMSat, PMSat satncRNAs, karyotype evolution, centromeric function

RESUMO

As sequências de DNA satélite (satDNA) constituem o principal componente da heterocromatina constitutiva (HC) e têm sido consideradas como um dos elementos repetitivos mais fascinantes e intrigantes dos genomas eucariotas. Durante muitos anos, estas sequências de DNA foram consideradas "lixo" e uma fração genómica trancricionalmente inerte. Atualmente, é reconhecida a importante função das sequências de satDNA na arquitetura dos genomas, na reorganização cromossómica durante a evolução, ou na regulação dos genomas, maioritariamente conduzida pelos seus transcritos ou RNAs satélite não codificantes (satncRNAs). Os transcritos centroméricos têm sido destacados como importantes reguladores na remodelação/deposição da CENP-A e no correto "assembly" do cinetocóro, fatores determinantes para a correta segregação cromossómica.

O advento de novas estratégias de sequenciação dos genomas (*"Next Generation Sequencing - NGS"*) e os avanços impressionantes das tecnologias de sequenciação têm fornecido uma enorme quantidade de dados de sequenciação de centenas de espécies. Da mesma forma, um crescente número de ferramentas e estratégias bioinformáticas têm sido desenvolvidas para identificar e caracterizar a fracção repetitiva de todo um genoma, nomeadamente as sequências de satDNAs - o *"*Satellitome*"*.

Nas últimas décadas, os roedores pertencentes ao género *Peromyscus* emergiram como animal modelo em várias áreas científicas, incluindo a evolução cromossómica. A disponibilidade do primeiro genoma sequenciado representativo de uma espécie *Peromyscus* (o *P. maniculatus*), representou a oportunidade para uma perceção global do conteúdo em sequências repetitivas deste genoma. Neste trabalho, foi definido um "pipeline" bioinformático com base no algoritmo "Tandem Repeats Finder" e uma análise integrada baseada na similaridade entre sequências que permitiu a identificação de 21 famílias repetitivas em "tandem" (tamanho de "array" maior que 2 kb), sendo a maioria correspondente a sequências relacionadas com sequências de satDNA ou elementos transponíveis, que se apresentam em "tandem". Duas sequências ortólogas de satDNA de ratazana e ratinho foram identificadas pela primeira vez no genoma de *P.maniculatus*: RNSAT1 e MMSAT4, respetivamente. A família mais predominante encontrada no genoma de *P. maniculatus* foi uma sequência de satDNA previamente descrita no genoma de *P. eremicus* – PMSat –, uma sequência rica em AT com um monómero de 345 bp. Esta sequência foi mapeada fisicamente em quatro espécies de *Peromyscus* (*P. eremicus*, *P. maniculatus*, *P. leucopus* e *P. californicus*) e revelaram que o PMSat está localizado principalmente nos centrómeros e regiões pericentroméricas de todos os cromossomas, para além de outras regiões ricas em HC, como os telómeros e os braços curtos de alguns cromossomas. Em todas as espécies estudadas, o PMSat apresenta uma elevada conservação da sequência nucleotídica, apesar da grande variação no número de cópias encontrado em cada genoma. Assim, os nossos resultados sugerem que a evolução do PMSat foi impulsionada por flutuações no número de cópias, sendo que a elevada similaridade da sequência entre espécies *Peromyscus* e não-*Peromyscus* podem refletir eventos evolutivos que ocorreram de forma não-concertada. Para além disto, e de acordo com as diferenças nos cariótipos, bem como muitos dos polimorfismos cromossómicos encontrados nas espécies de *Peromyscus*, o padrão distinto de HC e a localização cromossómica do PMSat, hipotetizamos que os eventos moleculares evolutivos do PMSat podem ter promovido as variações de cariótipo em *Peromyscus* e a evolução destes genomas. Mais ainda, as flutuações no número de cópias de PMSat, promovidas por mecanismos moleculares como "crossing-over" desigual e amplificação por círculo-rolante, são claramente observadas nas adições de HC encontradas em algumas espécies, principalmente no genoma de *P. eremicus*.

A atividade transcricional do PMSat em células proliferativas de espécies *Peromyscus* revelou uma correlação positiva entre a expressão do PMSat e o número de cópias em cada genoma. Apesar da variação pronunciada no nível de transcritos detetados, a análise transcricional ao longo do ciclo celular revelou um perfil semelhante: o PMSat satncRNA acumula-se preferencialmente na transição G2/M e no início da mitose e está confinado ao núcleo. De forma a obter mais informações sobre a(s) função(ões) putativa(s) dos transcritos de PMSat, foi realizado um ensaio funcional de silenciamento dos transcritos de PMSat em células proliferativas de *P. eremicus* que revelou uma possível intervenção dos transcritos de PMSat no "assembling" do cinetocóro e na função centromérica. De acordo com os possíveis locais de ligação a fatores de transcrição na sequência de PMSat, a transcrição desta família de satDNA parece ser realizada pela RNA polimerase II em diversas condições celulares, nomeadamente em resposta a stress celular.

O trabalho aqui apresentado revelou que o PMSat, não só foi um "motor" na evolução do cariótipo do género *Peromyscus*, como também é um elemento crucial na função centromérica e fidelidade da segregação cromossômica, tarefa que parece ser desempenhada pelos seus transcritos.

Palavras-chave: *Peromyscus*, satellitome, PMSat, PMSat satncRNAs, evolução cariotípica, função centromérica

PUBLICATIONS AND COMMUNICATIONS

This thesis is based on the collection of the following articles throughout the PhD period:

Publications in international scientific journals with Referee

Mendes-da-Silva A, Santos HAF, Adega F, Gama-Carvalho M, Chaves R (2019) The major repetitive element of *Peromyscus*, PMSat, ensures mitotic fidelity and drives chromosome evolution. *In preparation*

Louzada S, Vieira-da-Silva A, **Mendes-da-Silva A**, Kubickova S, Rubes J, Adega F, Chaves R (2015) A novel satellite DNA sequence in the *Peromyscus* genome (PMSat): Evolution via copy number fluctuation. Molecular phylogenetics and evolution. 92: 193-203. doi: 10.1016/j.ympev.2015.06.008

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. Chromosome Research. 23(3), 479-493. doi: 10.1007/s10577-015-9482-8

Poster Communications

Mendes-da-Silva A, Santos AFH, Adega F, Gama-Carvalho M, Chaves R (2016) Genome wide analysis of highly repetitive sequences in the deer mouse genome. XL Portuguese Genetics Conferences, Coimbra, Portugal

Mendes-da-Silva A, Escudeiro C, Adega F, Chaves R (2015) Disclosing the centromeric satDNA function in *Peromyscus eremicus* genome. VII National Conference of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal

Mendes-da-Silva A, Adega F, Chaves R (2015) PMSat - a novel satDNA conserved across time as a starting point to disclose the role of centromeric non-coding RNAs. XXXIX Portuguese Genetics Conferences, Braga, Portugal

GENERAL INDEX

INDEX OF FIGURES

INDEX OF TABLES

ABBREVIATIONS

In this list are presented only the words/sentences abbreviations that are used more than twice in the text. Chemical formulas and symbols contained in the IUPAC (International Union of Pure and Applied Chemistry) are not included in this list as there are internationally recognized. Genes and Proteins abbreviations are not included in this list once they can be consulted in the respective nomenclature committees of each species.

BLAST – Basic Local Alignment Search Tool

CH – Constitutive Heterochromatin

CPC – Chromosome Passenger Complex

CRS – CENP-B box Core Recognition Sequence

CRS*var* – CENP-B box Core Recognition Sequence with 1-2 mismatches

CRS*wt* – wild-type CENP-B box Core Recognition Sequence

CT – Centromere

CTRs – Clustered Tandem Repeats

FISH – Fluorescent *in situ* hybridization

FN – Fundamental Number

HOR – High Order Repeats

IF– Immunofluorescence

LNA – Locked Nucleic Acid

MaSat – Mouse Major Satellite

MiSat – Mouse Minor Satellite

NCBI – National Center for Biotechnology Information

ncRNA – Non-coding RNA

NGS – Next Generation Sequencing

OTRs – Orphan Tandem Repeats

PCR – Polymerase Chain Reaction

PCT – Pericentromere

qPCR – Quantitative Polymerase Chain Reaction

RNAi – RNA interference

RNApolII – RNA Polymerase II

RT – Room temperature

RT-qPCR – Reverse Transcriptase Quantitative Polymerase Chain Reaction satDNA – Satellite DNA

satncRNA – Satellite Non-Coding RNA

SD – Standard Deviation

siRNAs – Short Interfering RNAs

TEs – Transposable Elements

TF – Transcription Factor

TRF – Tandem Repeats Finder

TRs – Tandem Repeats

WGS – Whole Genome Shotgun

Species List

CCR – *Cricetus cricetus*

MAR – *Microtus arvalis*

PCA – *Peromyscus californicus*

PER – *Peromyscus eremicus*

PLE – *Peromyscus leucopus*

PMA – *Peromyscus maniculatus*

PSU – *Phodopus sungorus*

CHAPTER I

GENERAL INTRODUCTION

Historically, constitutive heterochromatin (CH) has been viewed as "dark matter" with an unvarying and static structure, in which only a few regulatory processes occur. However, over the past two decades, the synergy between the "Omic" technologies, such as genomics, epigenomics, proteomics and transcriptomics, has revealed unexpected plasticity. Additionally, transcription of centromeric regions has emerged as a feature of eukaryotic genomes in various biologic contexts (Chapter III). The CH domains formation and regulation has been considered to be more dynamic than anticipated, and understanding the biogenesis and function of repetitive sequences, mainly at centromeres, is raised as of fundamental interest.

CH is a major component of eukaryotic genomes composing about 30% in *Drosophila* and human genomes, 60% in rodents and up to 70-90% in nematodes and plants (Lander et al. 2001; Dimitri et al. 2005; Vicient and Casacuberta 2017). In most organisms, CH occurs as large blocks mainly at (peri)centromeric regions and telomeres, and represents the typical "inert" chromatin structure with a stable structural organization characterized by histone hypoacetylation (important for chromatin compaction), di- and trimethylation of lysine 9 of histone H3 (H3K9me2 and H3K9me3, respectively) and its binding protein, heterochromatin protein 1 (HP1) (Bannister et al. 2001; Yan and Boyd 2006; Janssen et al. 2018). This specialized chromatin plays critical roles in proper chromosomal segregation and genome stability. At centromeric region, CH components allow the recruitment of the cohesion complex that promotes sister chromatid cohesion and the recruitment of kinetochore proteins, like Mis12 complex (reviewed in Grewal and Jia 2007; Fukagawa and Earnshaw 2014). Heterochromatic regions are typically formed by repetitive DNA sequences that, due to their high molecular dynamics, act like "hotspots" for the occurrence of structural chromosome rearrangements and consequently, genome evolution (Chaves et al. 2004; Ruiz-Herrera et al. 2006; Adega et al. 2009; Paço et al. 2014).

I. 1. EUKARYOTIC GENOMES "ARE DOOMED TO REPEAT"

A great proportion of the eukaryotic genomes are built of repetitive DNA sequences. Their abundance varies significantly among species, and is involved in the huge variation of genome size in eukaryotes, performing around 50% of the human genome (Shapiro and von Sternberg 2005; Gregory et al. 2007; Gregory 2018). Repetitive DNA is simply defined as sequence motifs that occur repeated several of times (i.e. hundreds or thousands) in the genome. However, encompasses a huge variety of DNA elements of very diverse structure and origin. Based on distribution modes in genomes, repetitive DNAs are classified into two main classes: dispersed (interspersed) repeats and tandemly repeated DNA (reviewed in Slamovits and Rossi 2002; Richard et al. 2008). In Figure I.1 each of these main classes is shown according to their principal repeat elements and genomic locations across species.

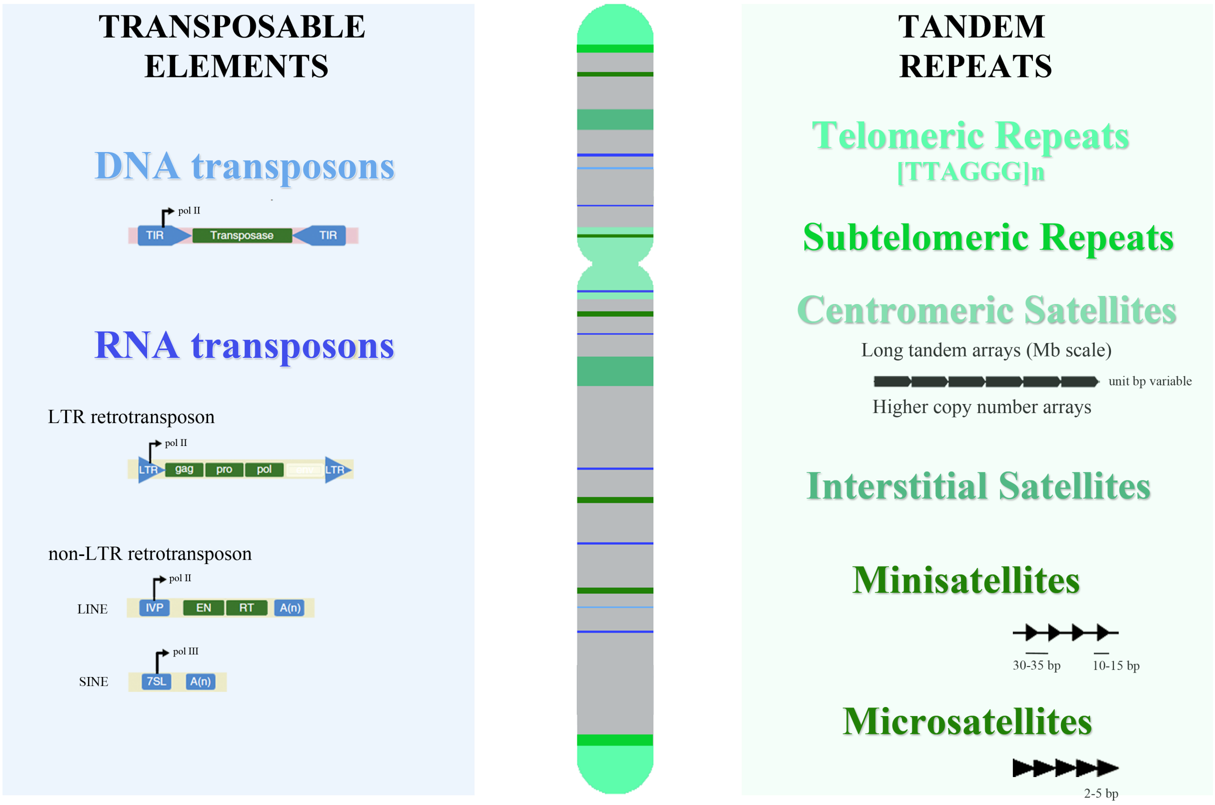


Figure I.1. Repetitive DNA in Eukaryotic genomes. Transposable elements (TEs), interspersed repeats scattered through the genome, are mainly located at heterochromatin, but also can be present in euchromatin (gene-rich regions). The elements are composed by DNA transposons and RNA transposons (with subclasses identified on the figure). Tandem repeats, sequential arrangement of repeat units, are composed by mini-, microsatellites and satellite DNA (satDNA). Mini- and microsatellites are predominantly located in euchromatin, but microsatellite arrays can be often also detected in heterochromatin. SatDNAs are located in heterochomatic regions [(peri)centromere, telomeres, and some interstitial block], but some arrays in euchromatin region was also identified. TEs and satDNAs are the main constitutes of Constitutive Heterochromatin (CH). Figure information as a result of the data compilation (Plohl et al. 2008; Richard et al. 2008; Meštrović et al. 2015; Padeken et al. 2015).

Interspersed repeats refer to sequences scattered through the genome, commonly identified as transposable elements (TEs) due their ability to “jump” (transpose) within distinct genomic locations. According to their transposition mechanism these repetitive elements are subdivided into two major classes: retrotransposons or class I, that transpose by a “copy and paste” mechanism through an RNA intermediate (e.g. L1 elements in mammals); and class II, the DNA transposons, that transpose by excision and integration (“cut and paste”) without an RNA intermediate (reviewed in Meštrović et al. 2015; Padeken et al. 2015;

Kojima 2018). The **tandem repeats (TRs)** are structurally characterized by a sequential arrangement (arrays) of copies/repeat units (monomer) that are normally presented in a head-to-tail fashion (Richard et al. 2008). Within a genome, distinct groups of tandem repeats are found with different properties: telomeric and subtelomeric repeats, microsatellites, minisatellites and satellite DNA. Microsatellites include simple short repeat units, 2-5 bp, with a total length of hundreds of basepairs (bp), while minisatellites have a unit length of 30–35 bp with a conserved core sequence of 10–15 bp, span from 1 to 15 kb (Padeken et al. 2015). Both classes can be found distributed through the genome in euchromatin regions. Further, microsatellite arrays can be often also detected in heterochromatin (Plohl et al. 2008; Padeken et al. 2015). **Satellite DNAs (satDNA)**, that are the repetitive sequences and focused in this thesis, are characterized by noncoding long tandem arrays and it are usually present in several million copies in genomes (Charlesworth et al. 1994). Indeed, their prominent copy number arrays constitute the main feature that allows its differentiation from micro- and minisatellites (Plohl et al. 2008). SatDNAs do not have the ability to transpose by themselves as TEs. However, there are some reported examples showing that TEs may act as a substrate for satDNA emergence and mobility (Dias et al. 2015; Meštrović et al. 2015; Satović et al. 2016; Chaves et al. 2017). Together with TEs, satDNAs are the main constituent of CH (Ugarković and Plohl 2002; Chaves et al. 2004; Meštrović et al. 2015), being preferentially found at the centromeric and pericentromeric heterochromatin, but also at the interstitial and terminal chromosomal positions (reviewed in Adega et al. 2009). Actually, recent findings also revealed the presence of short arrays dispersed along the euchromatin (gene-rich regions) (Brajković et al. 2012; Kuhn et al. 2012; Pavlek et al. 2015). In general, the specific sequence and monomer length, copy number and chromosome distribution, define the different satDNA families (Plohl et al. 2008). These sequences are generally characterized by a high AT content, but GC-rich families have been identified (Kuznetsova et al. 2006). These sequences often have a high linguistic complexity of nucleotides that leads to complex conformational curvatures of the DNA helix axis, resulting in tertiary or quaternary structures important for the heterochromatin identity (Ugarković 2005). Basic features of centromeric and pericentromeric satDNA, as structural and evolutionary concerns, are highlighted in the following sections.

I. 1.1. Satellite DNA and the centromere identity

The faithful inheritance of the genome during cell division is ensured by a region of specialized chromatin found in all eukaryotic chromosomes – the centromere. Two distinct domains are of vital importance for the centromeric function: the centromere core domain and its flanking pericentric heterochromatin (pericentromere), which are epigenetically defined by different sets of proteins that are concomitant with their structure and function in kinetochore formation and sister chromatid cohesion, respectively (Figure I.2a) (Chan and Wong 2012; Plohl et al. 2014). At the centromere, heterochromatin contains the histone H3 variant (CENP-A in mammals) interspersed with histone H3 (H3.1), which is characterized by the methylation of lysine residues of H3.1 tails by a lysine methyltransferase: methylation and di-metylation of lysine 4 (H3K4me1/2) and di- and tri-methylation of lysine 36 (H3K36me2/3) (Hall et al. 2012). Pericentromere is characterized by di- and tri-methylation of H3.3 lysine 9 and 27 (H3K9me2/3 and H3K27me2/3), and also by the presence of non-histonic proteins, which are associated to chromatin by the H3K9me recognition, the HP1 (in mammals) (Chan and Wong 2012; Hall et al. 2012).

While centromeric structure and function is conserved through eukaryotes, both centromere DNA sequences (mainly satDNAs) and protein components are paradoxically variable (Henikoff et al. 2001). In fact, it was assumed that both satDNA and protein evolve in parallel at the centromere, but at the same time provide a stable complex essential for centromere activity (Dawe and Henikoff 2006). Epigenetic pathways also play a crucial role for structural and functionality of centromeres, but the synergy among centromeric components remains unclear (Hayden et al. 2013; Fukagawa and Earnshaw 2014; Plohl et al. 2014). Nevertheless, a more comprehensive contribution of satDNA to centromeric activity was accomplished by the discovery of the transcription of centromeric satDNAs as non-coding RNAs, which are important for a functional centromere/kinetochore complex (a review of satDNA transcription was presented on Chapter III).

Centromere region differ greatly among species, showing variations in nucleotide sequence, monomer length, copy number of repeats and their organization itself. Variations exist even in different chromosomes of the same species. Primate centromeres are made of a 171 bp satDNA (alpha or alphoid satDNA) that acquires distinct organization features (reviewed in Plohl et al. 2012). Human pericentromere occurs in the form of monomeric tandem repeats and centromere core domain is organized into high-order repeats (HOR) structures that consist of multiple (from 2 to 34) head-to-tail basic 171 bp repeat units (Figure

I.2b) (reviewed in Plohl et al. 2012; Garrido-Ramos 2017). Conversely, in the mouse centromeres and pericentromeres, two distinct AT-rich satDNAs have been characterized: the minor satellite (MiSat; 120 bp) and major satellite (MaSat; 234 bp) (Guenatri et al. 2004a). Additionally, two GC-rich satDNAs are present in some mouse centromeres and pericentromeric regions of chromosomes, mouse satellite 3 (MS3, 150 bp) and mouse satellite 4 (MS4, 300 bp), (Kuznetsova et al. 2006). Interestingly, some evidences shows that MaSat can also present HOR structures (Komissarov et al. 2011).

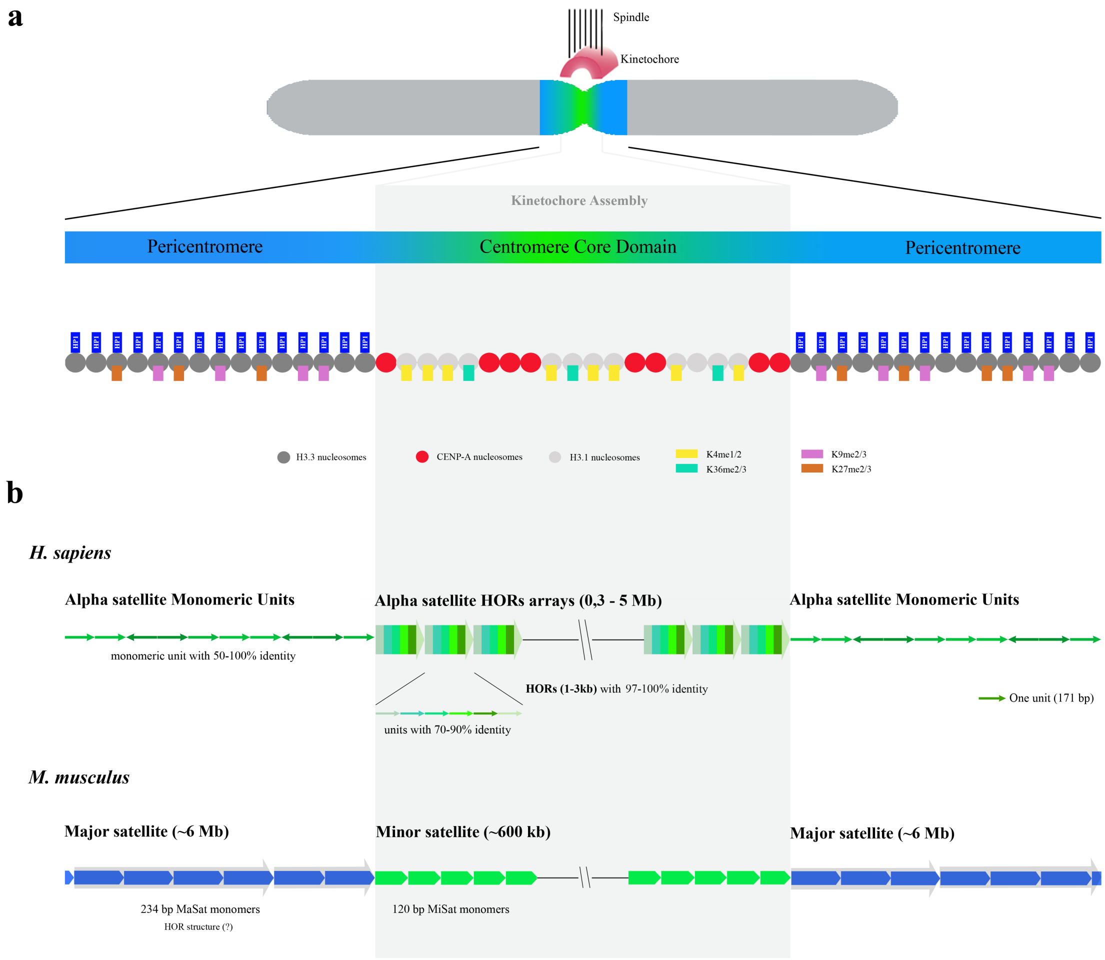


Figure I.2. Centromere organization features in human and mouse chromosomes. (a) The centromere core domain, which specifies kinetochore formation, contains centromere-specific proteins (not shown) and consists of clusters of CENP-A and H3.1 nucleosomes. The pericentromere regions with H3.3 nucleosomes contain typical heterochromatin markers including heterochromatin protein 1 (HP1). The epigenetic features of lysine residues are described in the text. (b) Human centromere was made by long tandem array of high-order repeats (multicolored arrows HOR) made up of a set of 171 bp alpha satellite monomers flanked by monomeric units disorderly arranged; mouse centromere was constituted by Major satellite (234 bp units; possible forming HORs) located at pericentromere and Minor satellite (120 bp units) located at the centromere core domain. The figure information was collected from the studies (Guenatri et al. 2004; Roizès 2006; Chan and Wong 2012; Plohl et al. 2012; Fukagawa and Earnshaw 2014; Biscotti et al. 2015).

Despite satDNAs monomer length variation range from only a few bp up to more than 1kb across species, it seems to exist a preferential monomer length between 150-180 bp and 300-360 bp, which can be explained by the required DNA length to be wrapped around one or two nucleosomes, respectively (Henikoff et al. 2001). Further, some DNA motifs can be preserved across centromeric satDNAs. The presence of a short motif with 17 bp, known as CENP-B box, was found in several mammalian centromeres (Masumoto et al. 2004). These motif represents the binding site for the centromere protein B (CENP-B), an essential component for centromere function (Ugarković 2005; Fachinetti et al. 2015).

I. 1.2. Satellite DNA evolution

A remarkable feature of satDNA is their rapid turnover even among closely related species, in which differences in nucleotide sequence, copy number and/or composition of satDNA families reflect their evolutionary dynamics (reviewed in Ugarković and Plohl 2002; Plohl et al. 2008). Several satDNA families can coexist in a single genome constituting a "library" of satDNAs, that each of them can be independently amplified/deleted in each genome (Fry and Salser 1977; Plohl et al. 2012). According to the "library model" (Fry and Salser 1977), related species share an ancestral hypothetical bulk of distinct satDNA families, and expansions and contractions of satDNA monomers/arrays can lead a species-specific satDNA profile with a copy number variation among related species, or even between distinct chromosomes (Figure I.3a). Thus, resulting in the replacement of one dominant satellite repeat (major satellite) by another less represented (minor satellite) (reviewed in Ugarković and Plohl 2002). This model of satDNA evolution is verified in some satDNA already reported (Mravinac et al. 2002; Bruvo et al. 2003; Plohl et al. 2010). The major satDNA present in the genomes of the *Ctenomys* rodents (RPCS - repetitive PvuII *Ctenomys* sequence) revels copy number fluctuations among these rodents genomes (Slamovits et al. 2001; Ellingsen et al. 2007; Caraballo et al. 2010).

It is generally accepted that satDNAs sequences evolves according to the principles of concerted evolution, that a non-independent evolution of satellite monomers results in a high repeat homogenization of satDNAs within a genome (Elder and Turner 1995; Meštrović et al. 1998, 2013). This evolution mode is promoted by molecular drive (Figure I.3b), a two-level process in which mutations are spread or eliminated throughout members of a repetitive family, and concomitantly to its fixation within a population (reviewed in Plohl 2010). The sequence homogenization is promoted by mechanisms of non-reciprocal transfer, mainly unequal crossing-over, gene conversion, rolling circle replication, and also transposition-related mechanisms (Dover 1986, 2002; Elder and Turner 1995). Generally, these mechanisms act more efficiently within proximal monomers, decreasing their efficiency when occurring between different arrays on the same chromosome, homologous or heterologous chromosomes (Plohl et al. 2008). Thus, adjacent monomers reveal higher degree of sequence similarity and, in some cases, the homogenization process may originate new repeat units compose in HORs. For example, in the human alpha-satellite (mentioned above in Figure I.2b), HORs are typically 97-100% identical while internal subunits are ~70% identical (Roizès 2006; Palomeque and Lorite 2008).

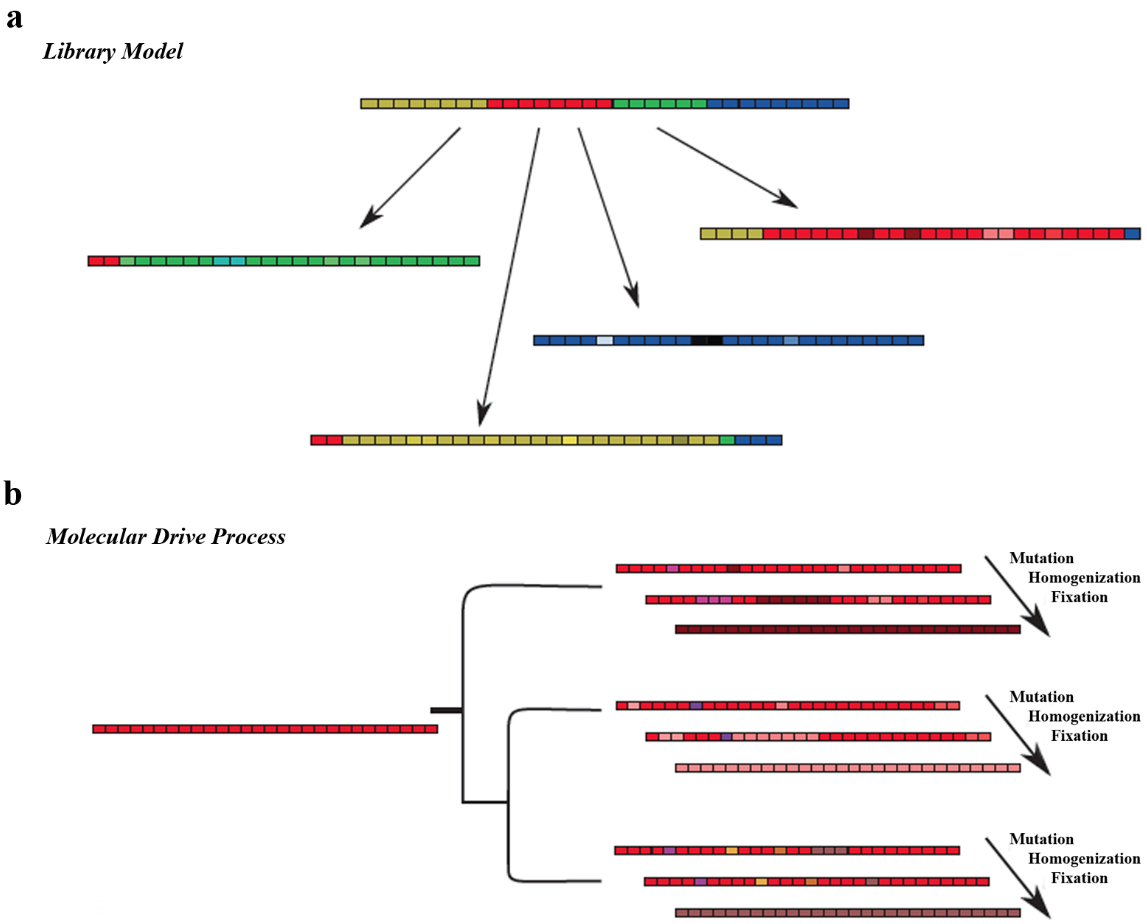


Figure I. 3. The Library Model and Molecular Drive Process. (a) The library model of satDNA evolution predicts that several satDNA families can coexist with different representation among species/chromosomes, which can be differentially amplified resulting in a distinct satellite landscape leading to a species-specific profile. Each rectangle represents a repeat unit. Different colors indicate different satDNA variants in an ancestral species. Subsequent variation in each species is represented by different color gradients. (b) The molecular drive process explains intraspecific homogenization and the gradual divergence of a specific satDNA family (color gradient) between species. Adapted from Garrido-Ramos (2015).

The centromere was traditionally referred as a genomic *locus* of suppressed recombination, but mechanisms as unequal crossing over and gene conversion have been identified as involved is satDNA dynamics on these genomic region (Talbert and Henikoff 2010). Also, segmental duplications has been implicated in large satDNA amplifications of satDNA arrays and rearrangements of (peri)centromeric regions (Catacchio et al. 2015). The study of human centromeres (Schueler et al. 2005) reveals that centromeric satDNA evolves according to "Proximal Progressive Expansion" (Figure I.4). According to this model, new satDNA sequences originated by mutations are consequently added to the core centromere, in a progressive manner that includes both copy number changes and mutation/homogenization mechanisms. Each addition moves previous centromeric DNA outwards, being the older sequences located more distantly. The terminal monomers present a low efficacy of the homogenization mechanisms, and these outwards monomers are more divergent than those located in the centromere core (for instance see Figure I.2) (Schueler et al. 2005; Schueler and Sullivan 2006).

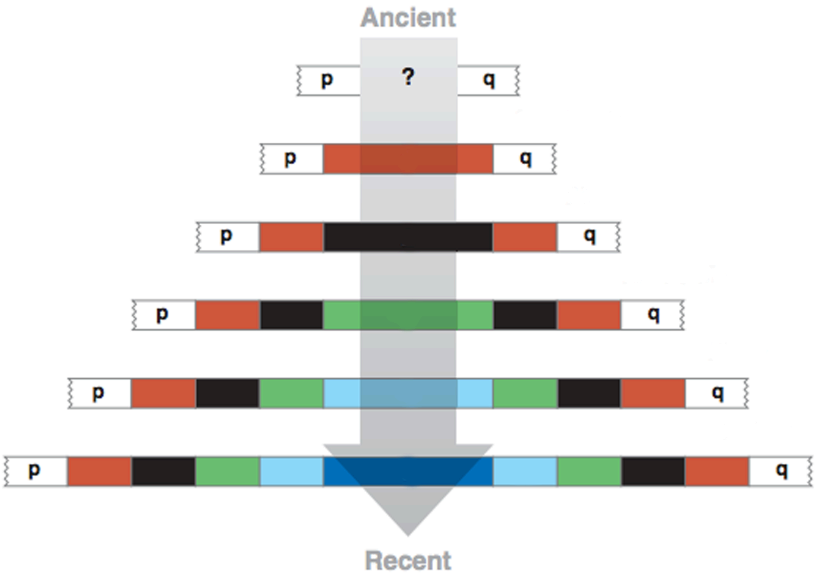


Figure I.4. Centromeric satDNA evolution by Proximal Progressive Expansion. New satDNA sequences (originated by mutations) are successively added (colour bars) to the centromere along evolution. Each addition moves previous centromeric DNA outward. Adapted from Schueler and Sullivan (2006).

Altogether, the dynamic evolution of satDNAs leads to a species-specific satellite profiles that reflects a combinatorial evolve mechanisms include nucleotide sequence, monomer size, copy number fluctuations and/or chromosome locations. In fact, the final outcome of concerted evolution is a highly dynamic molecular behavior of satDNAs that results in their occurrence in a restricted lineage (e.g. taxonomic group, species, chromosomes) (Dover 1986; Rudd et al. 2006; Ellingsen et al. 2007). Nonetheless, some satDNA sequences, designated as "frozen" satDNAs, persist in the genomes over long evolutionary times, even in the form of low-copy number repeats, most likely because the concerted evolution of these sequences is influenced by selective constraints and/or slowing down mutation rates (Mravinac et al. 2002; Mravinac et al. 2005; Plohl et al. 2010; Petraccioli et al. 2015; Chaves et al. 2017). The most striking example is the FA-SAT satellite that was recently characterized by our group (Chaves et al. 2017). This satDNA family is the oldest satDNA already reported being present in several Bilateria species. FA-SAT copy number changes accompanied by low sequence variability, are observed between Carnivora (i.e., cat and genet) and non-Carnivora genomes, and are arranged in tandem arrays at telomeric or centromeric regions, or presented in an interspersed fashion, respectively.

I. 1.3. The link between satDNA and Chromosomal Evolution

The ever-increasing volume of genomic and cytogenetic information focused on chromosomal evolution highlight satDNA sequences as active players in the structural and functional evolution of the genome (Garagna et al. 2001; Slamovits et al. 2001; Louzada et al. 2008). Some authors described CH as hotspots for structural chromosomal rearrangements (Chaves et al. 2004). Actually, repetitive sequences are involved in chromosomal rearrangements and are responsible for significant proportions of the karyotypic variations observed in many taxa (Slamovits and Rossi 2002; Schibler et al. 2006; Chaves et al. 2012; Paço et al. 2015; Vieira-da-Silva et al. 2015; Li et al. 2017). Ruiz-Herrera et al (2006) performed a broad analysis of evolutionary breakpoint regions between several species and disclosed that human chromosomes possess fragile sites (where evolutionary rearrangement events accumulated) characterized by the presence of tandem repeat elements.

Although the specific mechanisms underlying the close relationship between satDNA and chromosomal rearrangements are unclear, several reports suggest satDNA intragenomic movements among non-homologous chromosomes and between different chromosomal regions (centromere, telomere, and short and long arm) (Wichman et al. 1991; Garagna et al. 2001; Slamovits et al. 2001; Louzada et al. 2008). Slamovits and colleagues (2001) confirmed that karyotype reshuffle in the rodent *Ctenomys* was accomplished by localization and copy number variations of RPCS satDNA (referred before). Specifically, expansion and contraction events were accomplished by chromosome fissions or fusions, respectively (Slamovits et al. 2001; Caraballo et al. 2010). Also, the presence of specific sequence motifs on satDNA sequences, like the CENP-B box, seems to play an important role in recombination events promoting, in some cases, translocations involving the centromeric region (e.g. Garagna et al. 2001; Kalitsis et al. 2006; Meštrović et al. 2013). The complex satDNA/CENP-B seems to promote recombination events in a dual manner: 1) the sequence similarity can promote misalignments between monomers/arrays on non-homologous chromosomes, and 2) facilitate recombination due the protein nicking activity that is bound to satDNA itself (Kipling and Warburton 1997; Garagna et al. 2001). In this context, the MiSat-CENP-B protein complex looks to be involved in robertsonian translocations in mouse (Garagna et al. 2001).

Despite the significant findings underlying the link between satDNA and chromosomal rearrangements in the light of comparative cytogenetics, also the clinical cytogenetics, namely cancer cytogenetics, has been crucial to disclose the satDNAs role(s) in eukaryotic genomes (for a comprehensive detail about a practical cancer cytogenetics approach see Mendes-da-Silva et al. 2016). Indeed, carcinogenesis is considered a microevolutionary process, in which similar events such as the ones occurring during chromosomal evolution are observed in tumorigenesis progression. Santos and colleagues (2006) reported that in a cat fibrosarcoma the amplification of the FA-SAT satDNA was linked to the complex patterns of chromosome abnormalities detected. Also, not only FA-SAT satDNA sequence itself was correlated with carcinogenesis, as the FA-SAT non-coding transcripts also seem to play an important role in tumour progression (Ferreira et al. 2015, 2019).

I. 2. THE SATELLITOME – NEW CHALLENGES FROM THE GENOMICS ERA

Although initially called "junk" DNA, an increasing number of studies reinforce the importance of satDNA in genome plasticity and regulation. In addition, the assessing of the whole collection of satDNA families within a genome has been one of the main challenges in the new genomic era. Advances in Next Generation Sequencing (NGS) techniques have improved the quickness, high throughput and reduced cost of whole genome sequencing. However, the read length was replaced by speed, and the average read length of the most widely used NGS platforms is about 100-500 bp (Buermans and den Dunnen 2014). The reduced read lengths, compared with capillary-based approaches (Sanger sequencing), makes the "genome puzzle" an arduous challenging because more overlapping sequence reads (i.e., additional coverage) are necessary to generate a comparable assembly (Schatz et al. 2010; Treangen and Salzberg 2011). However, higher depth of coverage cannot overcome the difficulties of repetitive sequences. In fact, for *de novo* assembly, the assessment of a high throughput of sequences with read length smaller than repetitive monomers, with an additional high similarity between repeat units, resulted in several assembly gaps and produced more fragmented assemblies in recent years than in the pre-NGS era (Schatz et al. 2010; Ye et al. 2011). As an example, the satellite TCAST1 comprises the (peri)centromeric regions of all the chromosomes and comprising 35% of the beetle *Tribolium castaneum* genome; however, only 0.3% of TCAST1 of the assembled genome consists of a major TCAST1 satellite (Wang et al. 2008). Actually, even the human genome reference sequence remains incomplete due to the challenge of assembling long arrays of highly similar repeats at the centromeric regions and the short arms of the acrocentric chromosomes (for review, see Miga 2015). In the last years, the innovation and challenges in the NGS platforms allowed the generation of ever-larger reads (up to over 10.000 bp) strategies, known also as third-generation sequencing methodologies, that are currently provided by several companies such as Pacific Biosciences (PacBio) (Khost et al. 2017) and Oxford Nanopore (MinION) (Jain et al. 2018). Recently Jain et al. (2018) reported the complete assembly and characterization of the centromeric region of human chromosome Y by an implemented nanopore sequencing strategy, which represents a key advance on the understanding of these genomic regions.

NGS-based approaches are providing a growing number of sequenced genomes, while innovative and efficient bioinformatic tools have been specifically developed toward genome-wide identification of repetitive DNAs. Currently, there are new tools and strategies to access the whole collection of satDNAs from a given genome – termed as Satellitome (Ruiz-Ruano et al. 2016). In the last decade several NGS technologies have been developed with distinct sequencing strategies (Table I.1), which ultimately present advantages/disadvantages for the satDNA sequences assessment (reviewed in Lower et al. 2018). Further, many computational methods are design to assess repeats only in assembled data that can be focused to detect novel repeats [e.g. Tandem Repeats Finder (Benson 1999; Warburton et al. 2008; Komissarov et al. 2011; Bose et al. 2014; de Lima et al. 2017; Lang et al. 2019)] or based on similarity to known repetitive sequences [e.g. RepeatMasker (Bose et al. 2014; Ruiz-Ruano et al. 2016)]. Interestingly, new algorithms have been emerged for the repetitive sequences discovery at unassembled sequences, however, the maximum detectable repeat monomer size is constrained by read length [e.g. RepeatExplorer (Novák et al. 2013, 2014; Belyayev et al. 2019) and TOREAN (Novák et al. 2017)].

Table I.1. Next Generation and Third Generation Sequencing platforms for assessing highly conserved and repetitive DNA (satDNA). Adapted from Lower et al. 2018.

Platform	Method	Read length (Up to)	Pros	Cons	Examples
NEXT GENERATION SEQUENCING					
Illumina	Amplicons	300 bp	Inexpensive, low error rate	PCR bias in library preparation (PCR-free libraries reduces bias); short reads	Ruiz-Ruano et al. 2018 Ostromyshenskii et al. 2018
Ion torrent		400 bp	Fast, inexpensive	Lower yield; high error rate in homopolymer tracts	Cacheux et al. 2016
THIRD GENERATION SEQUENCING*					
Pacific Biosciences	Single molecule	50 kb	Long reads; can assemble complex satellite regions	Expensive; high error rate (some exceptions for Circular Consensus Sequencing approach)	Khost et al. 2017
Optical mapping (nanochannel)		220 kb	Long-range positional information; orthogonal method to sequencing	Requires a reference genome; large nicking intervals preclude mapping simple sequences	Weissensteiner et al. 2017
Oxford Nanopore		300 kb**	Longest reads	High error rate; extracting high molecular weight DNA is limiting	Jain et al. 2018

Pros: advantages; Cons: limitations

* Oxford Nanopore is also considered as fourth generation sequencing technique (e.g. Srinivasan and Batra 2014).

** Read length is only limited by the size of the DNA molecules; thus, reads up to 2 Mb can be obtained (https://nanoporetech.com/).

Traditionally, satDNA have been mostly studied from experimental approaches (mainly restriction digestion and/or PCR) with a small sample of cloned repeats that was isolated from a specific genome, or eventually, a few genomes. These expensive and time-consuming experimental strategies are insufficient for the identification of the satellitome from a chosen genome. The synergy between NGS technologies and computational algorithms has allowed the quest for repetitive content from multiple lineages of non-model taxa. Therefore, an increasing number of studies focuses in assessing satellite diversity across a wide range of species, including animals (Warburton et al. 2008; Alkan et al. 2011; Komissarov et al. 2011; García et al. 2015; Cacheux et al. 2016; Silva et al. 2017), insects (Ruiz-Ruano et al. 2016; Palacios-Gimenez et al. 2017; Ruiz-Ruano et al. 2018) and plants (Macas et al. 2011; Novák et al. 2014; Belyayev et al. 2019). Melters et al. (2013) developed a bioinformatic pipeline to identify the most abundant TRs from 282 selected sequenced genomes from animal and plant species. This approach confirmed the rapid evolution of satDNA at the centromere, in which centromeric repeat monomers were highly variable in both nucleotide sequence and length, however, showing similar modes of concerted evolution. On the mouse genome, the comprehensive analysis of tandemly repeated sequences, not only reveled new putative satDNA families (usually defined as tandem repeats - TRs - rather than satDNA in the Genomic era), as revealed new insights about MaSat evolution as HORs (Komissarov et al. 2011). Further, the genome-wide analysis of TRs have allowed the identification of several novel TR families that can reveal a specific pattern of hybridization, which might provide a kind of “bar code” for each chromosome that can be used in cytogenetic analysis (Komissarov et al. 2011; Podgornaya et al. 2013).

Since its discovery, satDNA is still the most enigmatic fraction of eukaryotic genomes. Howsoever, the genomic era opens new perspectives not only to disclose the fundamental features of satDNA (structure, composition, origin and evolution) but also to unveil the universal framework for understanding the roles of repetitive DNAs as a whole.

I. 3. RODENTIA – A SOURCE OF ANIMAL MODELS

Since its divergence (~82 MYA[1]), rodents underwent an impressive adaptive radiation, accounting for over one third of the current mammalian species (Carleton and Musser 2005). Rodentia order (NCBI:txid9989) comprises an evolving taxon with more than 2.200 species distributed by 33 families (Wilson and Reeder, 2005). Currently, new species and genera are being described each year (e.g. Fabre et al., 2018; Pérez et al. 2017). These small to medium-size mammals are ubiquitous and having spread almost over all continents (except Antarctica), where they occupy basically all terrestrial ecosystems, including tropical rainforest and deserts (Carleton and Musser 2005).

Rodents diversity is also reflected karyotypically, due the extreme variation in the diploid chromosome number ranging from 2n=10 to 2n=102 (Romanenko et al. 2012, http://www.bionet.nsc.ru/labs/chromosomes/). Even in the same genera, some species may differ by almost 50 chromosomal rearrangements (Aniskin et al. 2006). Indeed, rodent genomes experienced a rapid chromosomal evolution (Veyrunes et al. 2007). Specifically, the species belonging to the Superfamily Muroidea, were characterized by intense chromosome reshufflings, presenting many complex rearrangements compared to humans and other mammals (Romanenko et al. 2012). As a result, these species were considered a preferential animal model for studying the process of karyotype evolution (Romanenko et al. 2007) that is accompanied by variations at the heterochromatin content and consequently satDNA distribution patterns (Volobouev et al. 2006). Altogether, rodents have been considered a fine candidate to study the dynamic behavior of satDNA sequences and its contribution to genome evolution (e.g. Paço et al. 2014; Paço et al. 2015; Vieira-da-Silva et al. 2015).

Intrinsic characteristics such as small size, easily housing and maintenance, and well adaptation to new environments make rodents, specifically mouse (*Mus musculus*) and rat (*Rattus norvegicus*), the animal model of choice for a broad range of scientific fields like, for instance, physiology, nutrition, pharmacology, toxicology, immunology or cancer, among others (Morse 2007). The advent of the new genomic era with an increasing of whole genome sequencing platforms allows sequencing an increasing number of Rodentia non-model genomes. Currently, in addition to mouse (*Mus musculus*) and rat (*Rattus norvegicus*), about more than 30 Rodentia genomes were sequenced and several others are in the process in sequencing (Accession NCBI:txid9989, NCBI taxonomy, www.ncbi.nlm.nih.gov/taxonomy).

[1] Estimated divergence times between Lagomorpha and Rodentia were derived from 45 molecular and paleontological studies. These inferences were accomplished and available at http://www.timetree.org/.

Thus, as a result of the increasing genomic and comparative studies, different rodent species are emerging as new models. Importantly, due to the rodent high content in CH, we are able to combine different strategies, including whole genome shotgun (WGS) technologies/data and *in silico* analysis conjugated with *in situ* approaches (e.g. Fluorescent *in situ* hybridization approaches; for more details see Mendes-da-Silva et al. 2016) to unveil the secrets that govern satDNA regions.

I. 3.1. *Peromyscus* as an emerging animal model

The *Peromyscus* genus (Cricetidae, Neotominae) constitutes the most abundant and diverse group of North American mammals found from Alaska to Central America. This genera comprises 56 recognized species, being the *Peromyscus maniculatus* (deer mouse) and *Peromyscus leucopus* (white-footed mouse) the two most abundant and widespread (Figure I.5) (Carleton and Musser 2005). As in Rodentia species in general, the phylogenetic relationships inside *Peromyscus* remains a challenge due their complexity (Bradley et al. 2007).

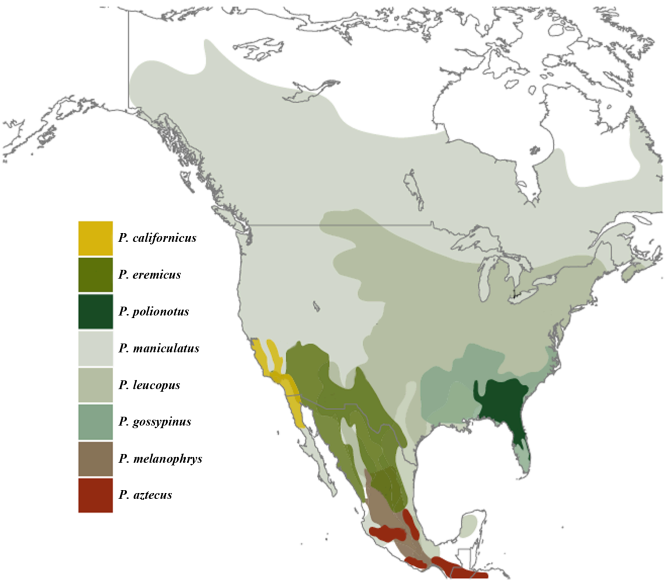


Figure I.5. Geographic distribution of *Peromyscus* species in North American. Only the species currently maintained as laboratory stocks are shown. Adapted from Bedford and Hoekstra (2015).

Despite the morphological similarities between *Peromyscus*, *Mus* and *Rattus*, the two last ones share a more recent common ancestor each other than *Peromyscus* (Figure I.6). The *Peromyscus* not only assisted in the phylogenetic relationships disclosures the *Mus/Rattus* lineage by serving as outgroup but also represents an intermediary species between the two major rodent genetic models and humans (O'Neill et al. 2007). However, comparative genomic analyses suggest the deer mouse genomic organization more closely to rat than mouse (Ramsdell et al. 2008).

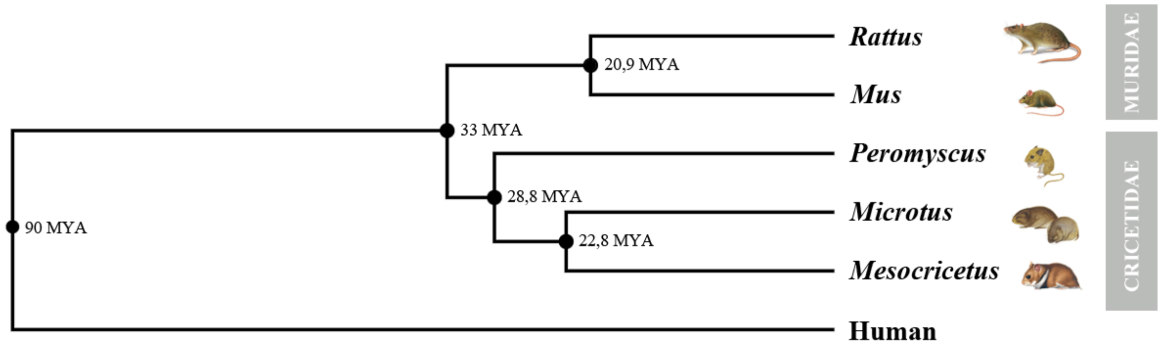


Figure I.6. Phylogeny of some muroid rodent models and relation with human. *Peromyscus* belong to the Cricetidae family, which includes voles (*Microtus*) and hamsters (*Mesocricetus*). The laboratory rat (*Rattus norvegicus*) and mouse (*Mus musculus*) belongs to the Muridae family. Muridae and Cricetidae diverged ~33 million years ago (MYA). The divergence time between rodents and human was estimated in 88 MYA. Schematic phylogeny and divergence time was based on molecular and paleontological studies that were compiled and available on public database "Time Tree of Life" (http://www.timetree.org/).

In addition to mouse and rat, peromyscine species are quickly becoming models in diverse areas of science such as ecology, physiology, chromosomal evolution or reproductive and developmental biology (complied by Bradley et al. 2007). The majority of the studies focus in biomedical research (autism, epilepsy, cancer, diabetes, aging, infectious diseases, toxicology, haematology) and natural variation (behavior, habitat adaptation, etc.) (O'Neill et al. 2007; Shorter et al. 2012; Sun et al. 2014; Bedford and Hoekstra 2015; Havighorst et al. 2017). Indeed, the importance of these rodent species across a variety of scientific disciplines is highlighted by Dewey and Dawson (2001) that refer to *Peromyscus* as "The Drosophila of North American mammalogy". *Peromyscus* acquired special interest when they were recognized as a natural reservoir for infectious diseases, mainly the hantavirus pulmonary syndrome (Netski et al. 1999; Burns et al. 2018), Lyme disease (Schwanz et al. 2011) and hepatitis C (Kapoor et al. 2013; Vandegrift et al. 2017). In contrast to mice, which have a 2- to 3-year life span, *Peromyscus* species have life spans ranging between 5 and 8 years (Sacher and Hart 1978). As such, these species, in special *Peromyscus leucopus*, are being considered good models for aging research (Ungvari et al. 2008; Labinskyy et al. 2009). The frequent occurrence of adenocarcinomas in inbred *Peromyscus leucopus* strains makes them a spontaneous metastasis model in laboratory mammals (Parnell et al. 2005). Also, a recent work of Kaza et al. (2018) report that *Peromyscas californicus* supports the growth of estrogen-dependent breast cancers and, on the contrary to what happens in mice, exogenous supplementation is not necessary.

Currently, diverse research programs are carried out on rodents provided by the *Peromyscus* Genetic Stock Center at the University of South Carolina (PGSC;

http://stkctr.biol.sc.edu/) that provide healthy, uniform, disease-free and parasite-free *Peromyscus* individuals to research and to the educational community, and develop the genetic and molecular resources that will further enhance the research value of the species (e.g. Shorter et al. 2012; Kenney-Hunt et al. 2014; Vrana et al. 2014; Brown et al. 2018). Additionally, the Coriell Institute (https://www.coriell.org/) possesses distinct *Peromyscus* cell lines karyotipically characterized, providing an excellent source for cytogenetic research, as the research presented on this thesis.

Four *Peromyscine* species were studied in this thesis: *Peromyscus maniculatus*, *P. leucopus*, *P. californicus* and *P. eremicus*. Their phylogenetic relationships are shown in Figure I.7, and followed the estimated divergence time inferred from both fossil records and molecular studies compiled and available in the public database "Time Tree of Life" (http://www.timetree.org/).

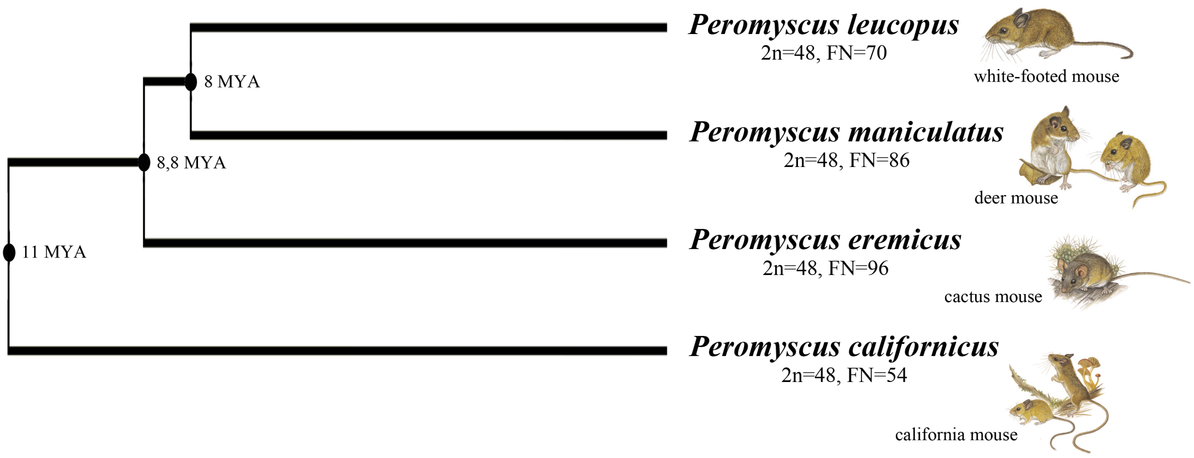


Figure I.7. Phylogenetic tree of the *Peromyscus* species studied on this thesis. The estimated divergence time among species are shown. The resulting tree was obtained from the data compilation available on public database "Time Tree of Life" (http://www.timetree.org/).

I. 3.1.1. Satellite DNA: A neglected element on *Peromyscus* studies

An interesting feature in *Peromyscus* species is the high degree of karyotypic conservation: all species exhibit 2n=48 (Romanenko et al. 2012). The differences between species reside on the number of chromosomal arms (Fundamental Number, FN) that ranges from 52 (*P. boylii*) to 96 (*P. eremicus*) (Robbins and Baker 1981; Rogers et al. 1984). One of the major goals in chromosomal evolution studies is the reconstruction of the putative ancestral karyotype of Muroidea and its subfamilies. Despite the initial purpose that considered the *P. eremicus* karyotype close to the putative ancestral for Muroidea

(Romanenko et al. 2007), some disagreement signatures were found when more *Peromyscus* genomes were analysed, as *P. maniculatus* (Romanenko et al. 2007; Mlynarski et al. 2008). These findings reveal the importance of comparative studies with as many as possible *Peromyscus* species for the reconstruction of the Neotominae ancestral karyotype (Romanenko et al. 2012).

The karyotypic differences among *Peromyscus* species were attributed to pericentric inversions or additions/deletions of large CH blocks present at the short arm of biarmed chromosomes that increase the FN (Deaven et al. 1977; Robbins and Baker 1981). All the *P. eremicus* chromosomes are biarmed (submetacentric), with all the autosomal short arms composed by CH. As a result, *Peromyscus* species greatly differs in the CH content, composing, for instance, in ~36% of *P. eremicus* genome and ~6% of *P. maniculatus* (Deaven et al. 1977).

Firstly referred by Jalal et al. (1974), it seems to exist a close relationship between satDNA sequences and the evolutionary rearrangements that originate the *Peromyscus* karyotype. With a combinatorial method of traditional analysis of CH and satDNA, quinacrine banding and density gradient centrifugation, these authors revealed that all the short arms of *P. eremicus* were heterochromatic and contained a large amount of satDNA. In the early 90's, Hamilton and colleagues (1992) isolated four satDNA clones from *P. leucopus* genome and shown that the majority of CH in several *Peromyscus* species is composed by satDNA. Interestingly, the studied species (e.g. *P. leucopus*, *P.maniculatus* and *P.eremicus*) revealed a conserved satDNA at the centromeric region in all the chromosomes that shared, at least, 70% of similarity (Hamilton et al. 1992). However, the molecular features of this satDNA (e.g. nucleotide sequence and length) were not defined, and the inferences were obtained through the stringency used on *in situ* hybridization experiments.

Despite its initial interest, the repetitive sequences on *Peromyscus* genomes, as also their involvement in the karyotyping reshufflings was neglected for many years. In the last decade, our group highlighted some evolutionary features involving the contribution of CH, and satDNA sequences on *Peromyscus* genome evolution. The complete characterization of CH from *Cricetus cricetus* and *P. eremicus* (Paço et al. 2009) contributed to the construction of the first combined chromosome comparative maps between these two Cricetidae species and the index species *M. musculus* and *R. norvegicus* (Vieira-da-Silva et al. 2015). On both genomes, evolutionary breakpoint regions co-localize with CH, reinforcing the involvement of CH in the karyotype restructuring of these species (Paço et al. 2009; Vieira-da-Silva et al. 2015). Several mammalian species were investigated in our lab and two orthologous satDNAs were reported in *P. eremicus* genome. A centromeric satDNA in *C. cricetus* genome present on chromosomes 4 and 10 (CCR4/10sat) is also present on *P. eremicus* genome displaying an interesting pattern: a scattered distribution in all the chromosomes of the species and localized within CH regions (Louzada et al. 2008). However, no sequence information of CCR4/10sat is available. Figure I.8 summarizes the reports on satellite DNA in *Peromyscus* until the beginning of the work presented in this thesis.

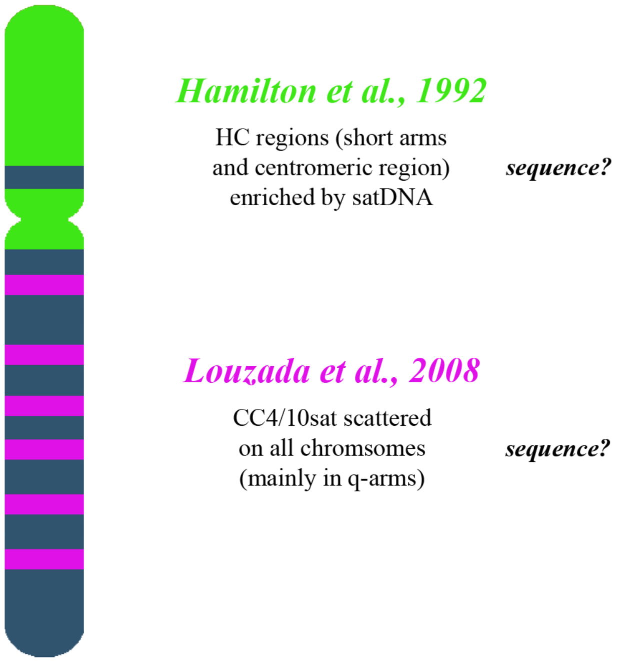


Figure I.8. Schematic representation of *Peromyscus* chromosome related to satDNA reports until the beginning of this work. The scheme is based on reports by Hamilton et al. (1992) and Louzada et al. (2008). In all these studies, *P. eremicus* is one of the analyzed species and, for that, a biarmed (submetacentric) chromosome is represented.

I. 4. AIMS OF THE WORK

The peculiarities of the *Peromyscus* karyotype diversity among the genus, mainly due to the heterochromatic content, highlights the potential of Peromyscine rodents for the study of the dynamic behaviour of satDNA. Despite the few studies regarding the notorious constitutive heterochromatin regions of the *Peromyscus* genome, mainly at the (peri)centromeric region, the release of the representative *Peromyscus* annotated genome assembly (*P. maniculatus bairdii*; Pman_1.0, GenBank assembly accession GCA_000500345.1) and the recent assembly of *P. maniculatus* chromosomes (HU_Pman_2.1, GenBank assembly accession GCA_003704045.1) represents a gateway for genome-wide analysis of the *Peromyscus* satellitome and tandem repeats assessment. Over the last years, an increasing number of studies reinforced the functional role of satDNAs carried out not only by the DNA molecule but also by its derived satellite non-coding RNAs as dynamic elements of mammalian genomes driving the structural and functional evolution of the genome. The great question underlying this work was "What is the role of satellite DNA in the genome?" To address this scientific question, the *Peromyscus* genus (Rodentia) was used as model, and an integrated and comprehensive approach was carried out to disclose the satDNA content of *Peromyscus* genome and its functional role. To approach this goal, specific objectives were achieved:

i) Genome-wide analysis of the repetitive fraction content on *Peromyscus* genome through a bioinformatics pipeline;

ii) Molecular and cytogenetic characterization of the major satDNA family in four distinct *Peromyscus* species, *P. eremicus*, *P. maniculatus*, *P. leucopus* and *P. californicus*;

iii) Evaluation of the transcriptional *status* of this satDNA family and the putative function(s) of the derived satellite non-coding RNAs on *Peromyscus* genomes.

This thesis is divided in two major sections considering the study of satDNA as DNA sequences (Chapter I and II), and as satellite non-coding RNAs (Chapter III and IV).

Following the General Introduction in Chapter I, Chapter II describes the repetitive content of *Peromyscus* genome using a bioinformatic pipeline to unveil the diversity of large tandem repeat families, mainly satDNA. The *in silico* analysis was performed for *P. maniculatus* genome and the major (peri)centromeric satDNA family found – PMSat – was further experimentally characterized in terms of presence, localization, and abundance in the four distinct genomes previously referred. Also the analysis of the dynamic behaviour of this satDNA family on *Peromyscus'* karyotype evolution was also performed and presented in this chapter.

In Chapter III, a general review emphasizes the functional features underlying satellite non-coding RNAs, especially regarding centromeric and pericentromeric transcripts. Chapter IV focus on the main purpose of understanding the role of (peri)centromeric satellite non-coding RNA on *Peromyscus* genomes, where the transcriptional profile of PMSat was performed in two genomes, *P. eremicus* and *P. maniculatus*, both in terms of space (cellular location) and time (cell cycle). The depletion of PMSat transcripts was carried out to disclose the putative cellular function(s) of these (peri)centromeric non-coding RNAs and the possible pathways in which it is involved.

Finally, Chapter V aimed to integrate and discuss the global findings, followed by the main conclusions and future perspectives of the work carried out.

I. 5. REFERENCES

Adega F, Guedes-Pinto H, Chaves R. 2009. Satellite DNA in the Karyotype Evolution of Domestic Animals – Clinical Considerations. Cytogenet Genome Res. 126(1–2):12–20. doi:10.1159/000245903.

Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, Ryder OA, Purgato S, Zoli M, Della Valle G, Eichler EE, et al. 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 21(1):137–145. doi:10.1101/gr.111278.110.

Aniskin VM, Benazzou T, Biltueva L, Dobigny G, Granjon L, Volobouev V. 2006. Unusually extensive karyotype reorganization in four congeneric *Gerbillus* species (Muridae: Gerbillinae). Cytogenet Genome Res. 112(1–2):131–140. doi:10.1159/000087525.

Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T. 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature. 410(6824):120–124. doi:10.1038/35065138.

Bedford NL, Hoekstra HE. 2015. The natural history of model organisms: Peromyscus mice as a model for studying natural variation. Elife. 4:e06813.

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). doi:10.3390/ijms20051201.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580. doi:10.1093/nar/27.2.573.

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

Bose P, Hermetz KE, Conneely KN, Rudd MK. 2014. Tandem Repeats and G-Rich Sequences Are Enriched at Human CNV Breakpoints. Chadwick BP, editor. PLoS ONE. 9(7):e101607. doi:10.1371/journal.pone.0101607.

Bradley RD, Durish ND, Rogers DS, Miller JR, Engstrom MD, Kilpatrick CW. 2007. Toward a Molecular Phylogeny for Peromyscus: Evidence from Mitochondrial Cytochrome- *b* Sequences. J Mammal. 88(5):1146–1159. doi:10.1644/06-MAMM-A-342R.1.

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

Brown J, Crivello J, O'Neill RJ. 2018. An updated genetic map of *Peromyscus* with chromosomal assignment of linkage groups. Mammal Genome. 29(5–6):344–352. doi:10.1007/s00335-018-9754-7.

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. 312:85–94.

Buermans HPJ, den Dunnen JT. 2014. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta BBA - Mol Basis Dis. 1842(10):1932–1941. doi:10.1016/j.bbadis.2014.06.015.

Burns JE, Metzger ME, Messenger S, Fritz CL, Vilcins I-ME, Enge B, Bronson LR, Kramer VL, Hu R. 2018. Novel Focus of Sin Nombre Virus in *Peromyscus eremicus* Mice, Death Valley National Park, California, USA. Emerg Infect Dis. 24(6):1112–1115. doi:10.3201/eid2406.180089.

Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, Escudé C. 2016. Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: *Cercopithecus solatus*. BMC Genomics. 17(1). doi:10.1186/s12864-016-3246-5.

Caraballo DA, Belluscio PM, Rossi MS. 2010. The library model for satellite DNA evolution: a case study with the rodents of the genus *Ctenomys* (Octodontidae) from the Iberá marsh, Argentina. Genetica. 138(11–12):1201–1210. doi:10.1007/s10709-010-9516-2.

Carleton MD, Musser GG. 2005. Order Rodentia. In: Wilson DE, Reeder DM, editors. Mammal Species of the World: A Taxonomic and Geographic Reference. 3rd ed. Baltimore: Johns Hopkins University Press. p. 745–1601.

Catacchio CR, Chiatante G, Anaclerio F, Ventura M. 2015. Segmental Duplications: A Source of Diversity, Evolution and Disease. eLS. doi:10.1002/9780470015902.a0020838.pub2

Chan FL, Wong LH. 2012. Transcription in the maintenance of centromere chromatin identity. Nucleic Acids Res. 40(22):11178–11188. doi:10.1093/nar/gks921.

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

Chaves R, Ferreira D, Mendes-da-Silva A, Meles S, Adega F. 2017. FA-SAT Is an Old Satellite DNA Frozen in Several Bilateria Genomes. Genome Biol Evol. 9(11):3073–3087. doi:10.1093/gbe/evx212.

Chaves R, Frönicke L, Guedes-Pinto H, Wienberg J. 2004. Multidirectional chromosome painting between the Hirola antelope (Damaliscus hunteri, Alcelaphini, Bovidae), sheep and human. Chromosome Res. 12(5):495–503. doi:10.1023/B:CHRO.0000034751.84769.4c.

Chaves R, Louzada S, Meles S, Wienberg J, Adega F. 2012. *Praomys tullbergi* (Muridae, Rodentia) genome architecture decoded by comparative chromosome painting with *Mus* and *Rattus*. Chromosome Res. 20(6):673–683. doi:10.1007/s10577-012-9304-1.

Dawe RK, Henikoff S. 2006. Centromeres put epigenetics in the driver's seat. Trends Biochem Sci. 31(12):662–669. doi:10.1016/j.tibs.2006.10.004.

Deaven LL, Vidal-Rioja L, Jett JH, Hsu TC. 1977. Chromosomes of *Peromyscus* (Rodentia, Cricetidae). Cytogenet Genome Res. 19(5):241–249. doi:10.1159/000130816.

Dewey MJ, Dawson WD. 2001. Deer mice: "The Drosophila of North American mammalogy." Genes. 29(3):105–109.

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. 23(3):597–613. doi:10.1007/s10577-015-9480-x.

Dimitri P, Corradini N, Rossi F, Mei E, Zhimulev IF, Vernì F. 2005. Transposable elements as artisans of the heterochromatic genome in *Drosophila melanogaster*. Cytogenet Genome Res. 110(1–4):165–172. doi:10.1159/000084949.

Dover G. 2002. Molecular drive. Trends Genet. 18(11):587–589. doi:10.1016/S0168-9525(02)02789-0.

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

Elder JF, Turner BJ. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. Q Rev Biol. 70(3):297–320.

Ellingsen A, Slamovits CH, Rossi MS. 2007. Sequence evolution of the major satellite DNA of the genus *Ctenomys* (Octodontidae, Rodentia). Gene. 392(1–2):283–290. doi:10.1016/j.gene.2007.01.013.

Fabre P-H, Reeve AH, Fitriana YS, Aplin KP, Helgen KM. 2018. A new species of Halmaheramys (Rodentia: Muridae) from Bisa and Obi Islands (North Maluku Province, Indonesia). J Mammal. 99(1):187–208. doi:10.1093/jmammal/gyx160.

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(3):314–327. doi:10.1016/j.devcel.2015.03.020.

Ferreira D, Escudeiro A, Adega F, Anjo SI, Manadas B, Chaves R. 2019. FA-SAT ncRNA interacts with PKM2 protein: depletion of this complex induces a switch from cell proliferation to apoptosis. Cell Mol Life Sci. doi:10.1007/s00018-019-03234-x.

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. Chromosome Res. 23(3):479–493. doi:10.1007/s10577-015-9482-8.

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(4):1069–1084.

Fukagawa T, Earnshaw WC. 2014. The Centromere: Chromatin Foundation for the Kinetochore Machinery. Develop Cell. 30(5):496–508. doi:10.1016/j.devcel.2014.08.016.

Garagna S, Marziliano N, Zuccotti M, Searle JB, Capanna E, Redi CA. 2001. Pericentromeric organization at the fusion point of mouse Robertsonian translocation chromosomes. Proc Natl Acad Sci. 98(1):171–175. doi:10.1073/pnas.98.1.171.

García G, Ríos N, Gutiérrez V. 2015. Next-generation sequencing detects repetitive elements expansion in giant genomes of annual killifish genus *Austrolebias* (Cyprinodontiformes, Rivulidae). Genetica. 143(3):353–360. doi:10.1007/s10709-015-9834-5.

Garrido-Ramos M. 2017. Satellite DNA: An Evolving Topic. Genes. 8(9):230. doi:10.3390/genes 8090230.

Garrido-Ramos MA. 2015. Satellite DNA in Plants: More than Just Rubbish. Cytogenet Genome Res. 146(2):153–170. doi:10.1159/000437008.

Gregory T. 2018. Animal genome size database. http://www.genomesize.com.

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2007. Eukaryotic genome size databases. Nucleic Acids Res. 35:D332-338. doi:10.1093/nar/gkl828.

Grewal SIS, Jia S. 2007. Heterochromatin revisited. Nat Rev Genet. 8(1):35–46. doi:10.1038/nrg2008.

Guenatri M, Bailly D, Maison C, Almouzni G. 2004. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. Cell Biol. 166(4):493–505. doi:10.1083/jcb.2004 03109.

Hall LE, Mitchell SE, O'Neill RJ. 2012. Pericentric and centromeric transcription: a perfect balance required. Chromosome Res. 20(5):535–546. doi:10.1007/s10577-012-9297-9.

Hamilton MJ, Hong G, Wichman HA. 1992. Intragenomic movement and concerted evolution of satellite DNA in *Peromyscus:* evidence from *in situ* hybridization. Cytogenet Genome Res. 60(1):40–44. doi:10.1159/000133292.

Havighorst A, Crossland J, Kiaris H. 2017. *Peromyscus* as a model of human disease. Semin Cell Dev Biol. 61:150–155. doi:10.1016/j.semcdb.2016.06.020.

Hayden KE, Strome ED, Merrett SL, Lee H-R, Rudd MK, Willard HF. 2013. Sequences associated with centromere competency in the human genome. Mol Cell Biol. 33(4):763–772. doi:10.1128/MCB.01198-12.

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

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(4):321–323. doi:10.1038/nbt.4109.

Jalal SM, Clark RW, Hsu TC, Pathak S. 1974. Cytological differentiation of constitutive heterochromatin. Chromosoma. 48(4):391–403. doi:10.1007/BF00290995.

Janssen A, Colmenares SU, Karpen GH. 2018. Heterochromatin: Guardian of the Genome. Annu Rev Cell Dev Biol.:24. doi:10.1146/annurev-cellbio-100617-062653.

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. 103(23):8786–8791. doi:10.1073/pnas.0600250103.

Kapoor A, Simmonds P, Scheel TKH, Hjelle B, Cullen JM, Burbelo PD, Chauhan LV, Duraisamy R, Sanchez Leon M, Jain K, et al. 2013. Identification of Rodent Homologs of Hepatitis C Virus and Pegiviruses. MBio. 4(2). doi:10.1128/mBio.00216-13.

Kaza V, Farmaki E, Havighorst A, Crossland J, Chatzistamou I, Kiaris H. 2018. Growth of human breast cancers in *Peromyscus*. Dis Model Mech. 11(1):dmm031302. doi:10.1242/dmm.031302.

Kenney-Hunt J, Lewandowski A, Glenn TC, Glenn JL, Tsyusko OV, O'Neill RJ, Brown J, Ramsdell CM, Nguyen Q, Phan T, et al. 2014. A genetic map of *Peromyscus* with chromosomal assignment of linkage groups (a *Peromyscus* genetic map). Mamm Genome. 25(3–4):160–179. doi:10.1007/s00335-014-9500-8.

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(5):709–721. doi:10.1101/gr.213512.116.

Kipling D, Warburton PE. 1997. Centromeres, CENP-B and Tigger too. Trends Genet. 13(4):141–145.

Kojima KK. 2018. Human transposable elements in Repbase: genomic footprints from fish to humans. Mob DNA. 9(1). doi:10.1186/s13100-017-0107-y.

Komissarov AS, Gavrilova EV, Demin SJ, Ishov AM, Podgornaya OI. 2011. Tandemly repeated DNA families in the mouse genome. BMC Genomics. 12(1):531. doi:10.1186/1471-2164-12-531.

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. doi:10.1093/molbev/msr173.

Kuznetsova I, Podgornaya O, Ferguson-Smith MA. 2006. High-resolution organization of mouse centromeric and pericentromeric DNA. Cytogenet Genome Res. 112(3–4):248–255. doi:10.1159/000089878.

Labinskyy N, Mukhopadhyay P, Toth J, Szalai G, Veres M, Losonczy G, Pinto JT, Pacher P, Ballabh P, Podlutsky A, et al. 2009. Longevity is associated with increased vascular resistance to high glucose-induced oxidative stress and inflammatory gene expression in *Peromyscus leucopus*. Am J Physiol Heart Circ Physiol. 296(4):H946-956. doi:10.1152/ajpheart.00693.2008.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409(6822):860–921. doi:10.1038/35057062.

Lang T, Li G, Yu Z, Ma J, Chen Q, Yang E, Yang Z. 2019. Genome-Wide Distribution of Novel Ta-3A1 Mini-Satellite Repeats and Its Use for Chromosome Identification in Wheat and Related Species. Agronomy. 9(2):60. doi:10.3390/agronomy9020060.

Li S-F, Su T, Cheng G-Q, Wang B-X, Li X, Deng C-L, Gao W-J. 2017. Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. Genes. 8(10). doi:10.3390/genes8100290.

de Lima LG, Svartman M, Kuhn GCS. 2017. Dissecting the Satellite DNA Landscape in Three Cactophilic *Drosophila* Sequenced Genomes. G3: Genes Genomes Genet. 7(8):2831–2843. doi:10.1534/g3.117.042093.

Louzada S, Paço A, Kubickova S, Adega F, Guedes-Pinto H, Rubes J, Chaves R. 2008. Different evolutionary trails in the related genomes *Cricetus cricetus* and *Peromyscus eremicus* (Rodentia, Cricetidae) uncovered by orthologous satellite DNA repositioning. Micron. 39(8):1149–1155. doi:10.1016/j.micron.2008.05.008.

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

Macas J, Kejnovský E, Neumann P, Novák P, Koblížková A, Vyskot B. 2011. Next generation sequencing-based analysis of repetitive DNA in the model dioecious plant Silene latifolia. PloS One. 6(11):e27335. doi:10.1371/journal.pone.0027335.

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. Chromosome Res. 12(6):543–556. doi:10.1023/B:CHRO.0000036593.72788.99.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14(1):R10.

Mendes-da-Silva A, Adega F, Chaves R. 2016. Importance of Fluorescent In Situ Hybridization in Rodent Tumors. In: Aziz SA, Mehta R, editors. Technical Aspects of Toxicological Immunohistochemistry. New York, NY: Springer New York. p. 21–49.

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. 23(3):583–596. doi:10.1007/s10577-015-9483-7.

Meštrović N, Pavlek M, Car A, Castagnone-Sereno P, Abad P, Plohl M. 2013. Conserved DNA Motifs, Including the CENP-B Box-like, Are Possible Promoters of Satellite DNA Array Rearrangements in Nematodes. PloS One. 8(6):e67328. doi:10.1371/journal.pone.0067328.

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(8):1062–1068. doi:10.1093/oxfordjournals.molbev.a026005.

Miga KH. 2015. Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosome Res. 23(3):421–426. doi:10.1007/s10577-015-9488-2.

Mlynarski EE, Obergfell CJ, Rens W, O’Brien PCM, Ramsdell CM, Dewey MJ, O’Neill MJ, O’Neill RJ. 2008. *Peromyscus maniculatus – Mus musculus* chromosome homology map derived from reciprocal cross species chromosome painting. Cytogenet Genome Res. 121(3–4):288–292. doi:10.1159/000138900.

Morse H. 2007. Building a Better Mouse: One Hundred Years of Genetics and Biology. In: The Mouse in Biomedical Research. Vol. I. Elsevier. p. 1–11.

Mravinac B, Plohl M, Mestrović N, Ugarković Đ. 2002. Sequence of PRAT Satellite DNA "Frozen" in Some Coleopteran Species. Mol Evol. 54(6):774–783. doi:10.1007/s0023901-0079-9.

Mravinac B, Plohl M, Ugarković Đ. 2005. Preservation and High Sequence Conservation of Satellite DNAs Suggest Functional Constraints. Mol Evol. 61(4):542–550. doi:10.1007/s00239-004-0342-y.

Netski D, Thran BH, St Jeor SC. 1999. Sin Nombre virus pathogenesis in *Peromyscus maniculatus*. Virol. 73(1):585–591.

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

Novák P, Hřibová E, Neumann P, Koblížková A, Doležel J, Macas J. 2014. Genome-Wide Analysis of Repeat Diversity across the Family Musaceae. PLoS ONE. 9(6):e98918. doi:10.1371/journal.pone.0098918.

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. Bioinforma Oxf Engl. 29(6):792–793. doi:10.1093/bioinformatics/btt054.

O'Neill R, Szalai G, Gibbs R, Weinstock G, Vrana P, Glenn J, Dewey M, Felder M. 2007. White paper proposal for sequencing the genome of *Peromyscus*. Natl Hum Genome Res Inst. http://www.genome.gov/pages/research/sequencing/seqproposals/peromyscus.pdf.

Ostromyshenskii DI, Chernyaeva EN, Kuznetsova IS, Podgornaya OI. 2018. Mouse chromocenters DNA content: sequencing and in silico analysis. BMC Genomics. 19(1). doi:10.1186/s12864-018-4534-z.

Paço A, Adega F, Guedes-Pinto H, Chaves R. 2009. Hidden heterochromatin: characterization in the Rodentia species *Cricetus cricetus*, *Peromyscus eremicus* (Cricetidae) and *Praomys tullbergi* (Muridae). Genet Mol Biol. 32(1):56–68. doi:10.1590/S1415-47572009000100009.

Paço A, Adega F, Meštrović N, Plohl M, Chaves R. 2015. The puzzling character of repetitive DNA in *Phodopus* genomes (Cricetidae, Rodentia). Chromosome Res. 23(3):427–440. doi:10.1007/s10577-015-9481-9.

Paço A, Adega F, Me trovi N, Plohl M, Chaves R. 2014. Evolutionary Story of a Satellite DNA from *Phodopus sungorus* (Rodentia, Cricetidae). Genome Biol Evol. 6(10):2944–2955. doi:10.1093/gbe/evu233.

Padeken J, Zeller P, Gasser SM. 2015. Repeat DNA in genome organization and stability. Curr Opin Genet Dev. 31:12–19. doi:10.1016/j.gde.2015.03.009.

Palacios-Gimenez OM, Dias GB, de Lima LG, Kuhn GC e S, Ramos É, Martins C, Cabral-de-Mello DC. 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). doi:10.1038/s41598-017-06822-8.

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

Parnell PG, Crossland JP, Beattie RM, Dewey MJ. 2005. Frequent Harderian Gland Adenocarcinomas in Inbred White-Footed Mice (*Peromyscus leucopus*). Comp Med. 55(4):5.

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(6):387–401. doi:10.1093/dnares/dsv021.

Pérez ME, Deschamps CM, Vucetich MG. 2017. Diversity, phylogeny and biogeography of the South American 'cardiomyine' rodents (Hystricognathi, Cavioidea) with a description of two new species. Pap Palaeontol. 4(1), 1-19. doi:10.5061/dryad.pj562.

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

Plohl M. 2010. Those mysterious sequences of satellite DNAs. Period Biol. 112(4):403–410.

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(1–2):72–82. doi:10.1016/j.gene.2007.11.013.

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA evolution. In: Repetitive DNA. Vol. 7. Karger Publishers. p. 126–152.

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

Plohl M, Petrović V, Luchetti A, Ricci A, Šatović E, Passamonti M, Mantovani B. 2010. Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. Heredity. 104(6):543–551. doi:10.1038/hdy.2009.141.

Podgornaya O, Gavrilova E, Stephanova V, Demin S, Komissarov A. 2013. Large tandem repeats make up the chromosome bar code: a hypothesis. Adv Protein Chem Struct Biol. 90:1–30. doi:10.1016/B978-0-12-410523-2.00001-8.

Ramsdell CM, Lewandowski AA, Glenn J, Vrana PB, O'Neill RJ, Dewey MJ. 2008. Comparative genome mapping of the deer mouse (*Peromyscus maniculatus*) reveals greater similarity to rat (*Rattus norvegicus*) than to the lab mouse (*Mus musculus*). BMC Evol Biol. 8(1):65. doi:10.1186/1471-2148-8-65.

Richard G-F, Kerrest A, Dujon B. 2008. Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. Microbiol Mol Biol Rev. 72(4):686–727. doi:10.1128/MMBR.00011-08.

Robbins LW, Baker RJ. 1981. An assessment of the nature of chromosomal rearrangements in 18 species of *Peromyscus* (Rodentia: Cricetidae). Cytogenet Genome Res. 31(4):194–202. doi:10.1159/000131649.

Rogers DS, Greenbaum IF, Gunn SJ, Engstrom MD. 1984. Cytosystematic Value of Chromosomal Inversion Data in the Genus *Peromyscus* (Rodentia: Cricetidae). J Mammal. 65(3):457–465. doi:10.2307/1381092.

Roizès G. 2006. Human centromeric alphoid domains are periodically homogenized so that they vary substantially between homologues. Mechanism and implications for centromere functioning. Nucleic Acids Res. 34(6):1912–1924. doi:10.1093/nar/gkl137.

Romanenko SA, Perelman PL, Trifonov VA, Graphodatsky AS. 2012. Chromosomal evolution in Rodentia. Heredity. 108(1):4–16. doi:10.1038/hdy.2011.110.

Romanenko SA, Volobouev VT, Perelman PL, Lebedev VS, Serdukova NA, Trifonov VA, Biltueva LS, Nie W, O'Brien PCM, Bulatova NS, et al. 2007. Karyotype evolution and phylogenetic relationships of hamsters (Cricetidae, Muroidea, Rodentia) inferred from chromosomal painting and banding comparison. Chromosome Res. 15(3):283–297. doi:10.1007/s10577-007-1124-3.

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

Ruiz-Herrera A, Castresana J, Robinson TJ. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? Genome Biol. 7(12)(R115). doi:10.1186/gb-2006-7-12-r115.

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(3):323–340. doi:10.1007/s00412-018-0666-9.

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep. 6(1). doi:10.1038/srep28333.

Sacher GA, Hart RW. 1978. Longevity, aging and comparative cellular and molecular biology of the house mouse, *Mus musculus*, and the white-footed mouse, *Peromyscus leucopus*. Birth Defects Orig Artic Ser. 14(1):71–96.

Santos S, Chaves R, Adega F, Bastos E, Guedes-Pinto H. 2006. Amplification of the major satellite DNA family (FA-SAT) in a cat fibrosarcoma might be related to chromosomal instability. Heredity. 97(2):114–118. doi:10.1093/jhered/esj016.

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(1). doi:10.1186/s12864-016-3347-1.

Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. Genome Res. 20(9):1165–1173. doi:10.1101/gr.101360.109.

Schibler L, Roig A, Mahe M-F, Laurent P, Hayes H, Rodolphe F, Cribiu EP. 2006. High-resolution comparative mapping among man, cattle and mouse suggests a role for repeat sequences in mammalian genome evolution. BMC Genomics. 7:194. doi:10.1186/1471-2164-7-194.

Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiano L, NISC Comparative Sequencing Program, Rocchi M, Willard HF, Green ED. 2005. Progressive proximal expansion of the primate X chromosome centromere. Proc Natl Acad Sci. 102(30):10563–10568. doi:10.1073/pnas.0503346102.

Schueler MG, Sullivan BA. 2006. Structural and Functional Dynamics of Human Centromeric Chromatin. Annu Rev Genomics Hum Genet. 7(1):301–313. doi:10.1146/annurev.genom.7.080505.115613.

Schwanz LE, Voordouw MJ, Brisson D, Ostfeld RS. 2011. Borrelia burgdorferi has minimal impact on the Lyme disease reservoir host *Peromyscus leucopus*. Vector Borne Zoonotic Dis. 11(2):117–124. doi:10.1089/vbz.2009.0215.

Shapiro JA, von Sternberg R. 2005. Why repetitive DNA is essential to genome function. Biol Rev. 80(2):227–250. doi:10.1017/S1464793104006657.

Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, Vrana PB. 2012. *Peromyscus* as a Mammalian Epigenetic Model. Genet Res Int. 2012:1–11. doi:10.1155/2012/179159.

Silva DMZ de A, Utsunomia R, Ruiz-Ruano FJ, Daniel SN, Porto-Foresti F, Hashimoto DT, Oliveira C, Camacho JPM, Foresti F. 2017. High-throughput analysis unveils a highly shared satellite DNA library among three species of fish genus *Astyanax*. Sci Rep. 7(1). doi:10.1038/s41598-017-12939-7.

Slamovits CH, Cook JA, Lessa EP, Susana Rossi M. 2001. Recurrent Amplifications and Deletions of Satellite DNA Accompanied Chromosomal Diversification in South American Tuco-tucos (Genus *Ctenomys*, Rodentia: Octodontidae): A Phylogenetic Approach. Mol Biol Evol. 18(9):1708–1719. doi:10.1093/oxfordjournals.molbev.a003959.

Slamovits CH, Rossi MS. 2002. Satellite DNA: agent of chromosomal evolution in mammals. A review. Mastozool Neotropical. 9:297–308.

Srinivasan S, Batra J. 2014. Four Generations of Sequencing - Is it Ready for the Clinic Yet? Next Gener Seq Appl. 1(107). doi:10.4172/2469-9853.1000107.

Sun Y, Desierto MJ, Ueda Y, Kajigaya S, Chen J, Young NS. 2014. *Peromyscus leucopus* mice: a potential animal model for haematological studies. Exp Pathol. 95(5):342–350. doi:10.1111/iep.12091.

Talbert PB, Henikoff S. 2010. Centromeres Convert but Don't Cross. PLoS Biol. 8(3):e1000326. doi:10.1371/journal.pbio.1000326.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. doi:10.1038/nrg3117.

Ugarković D. 2005. Functional elements residing within satellite DNAs. EMBO Rep. 6(11):1035–1039. doi:10.1038/sj.embor.7400558.

Ugarković Đ, Plohl M. 2002. Variation in satellite DNA profiles - causes and effects. EMBO J. 21(22):5955–5959. doi:10.1093/emboj/cdf612.

Ungvari Z, Krasnikov BF, Csiszar A, Labinskyy N, Mukhopadhyay P, Pacher P, Cooper AJL, Podlutskaya N, Austad SN, Podlutsky A. 2008. Testing hypotheses of aging in long-lived mice of the genus *Peromyscus*: association between longevity and mitochondrial stress resistance, ROS detoxification pathways, and DNA repair efficiency. AGE. 30(2–3):121–133. doi:10.1007/s11357-008-9059-y.

Vandegrift KJ, Critchlow JT, Kapoor A, Friedman DA, Hudson PJ. 2017. *Peromyscus* as a model system for human hepatitis C: An opportunity to advance our understanding of a complex host parasite system. Semin Cell Dev Biol. 61:123–130. doi:10.1016/j.semcdb.2016.07.031.

Veyrunes F, Watson J, Robinson TJ, Britton-Davidian J. 2007. Accumulation of rare sex chromosome rearrangements in the African pygmy mouse, *Mus* (Nannomys) minutoides: a whole-arm reciprocal translocation (WART) involving an X-autosome fusion. Chromosome Res. 15(2):223–230. doi:10.1007/s10577-006-1116-8.

Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. Ann Bot. 120(2):195–207. doi:10.1093/aob/mcx078.

Vieira-da-Silva A, Louzada S, Adega F, Chaves R. 2015. A High-Resolution Comparative Chromosome Map of *Cricetus cricetus* and *Peromyscus eremicus* Reveals the Involvement of Constitutive Heterochromatin in Breakpoint Regions. Cytogenet Genome Res. 145(1):59–67. doi:10.1159/000381840.

Volobouev VT, Gallardo MH, Graphodatsky AS. 2006. Rodents cytogenetics. In: O'Brien SJ, Nash WG, Menninger JC, editors. Atlas of Mammalian Karyotypes. Chichester, UK: Wiley. p. 173–176.

Vrana PB, Shorter KR, Szalai G, Felder MR, Crossland JP, Veres M, Allen JE, Wiley CD, Duselis AR, Dewey MJ, et al. 2014. *Peromyscus* (deer mice) as developmental models: *Peromyscus* as developmental models. Wiley Interdiscip Rev Dev Biol. 3(3):211–230. doi:10.1002/wdev.132.

Wang S, Lorenzen MD, Beeman RW, Brown SJ. 2008. Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. Genome Biol. 9(3):R61. doi:10.1186/gb-2008-9-3-r61.

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. doi:10.1186/1471-2164-9-533.

Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Petterson O, Suh A, Wolf JBW. 2017. Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. Genome Res. 27(5):697–708. doi:10.1101/gr.215095.116.

Wichman HA, Payne CT, Ryder OA, Hamilton MJ, Maltbie M, Baker RJ. 1991. Genomic Distribution of Heterochromatic Sequences in Equids: Implications to Rapid Chromosomal Evolution. Heredity. 82(5):369–377. doi:10.1093/oxfordjournals.jhered.a111106.

Yan C, Boyd DD. 2006. Histone H3 acetylation and H3 K4 methylation define distinct chromatin regions permissive for transgene expression. Mol Cell Biol. 26(17):6357–6371. doi:10.1128/MCB.00311-06.

Ye L, Hillier LW, Minx P, Thane N, Locke DP, Martin JC, Chen L, Mitreva M, Miller JR, Haub KV, et al. 2011. A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol. 12(3):R31.

CHAPTER II

UNVEILING THE SATELLITE DNA LANDSCAPE IN *PEROMYSCUS* GENOMES

II.1. A novel satellite DNA sequence from *Peromyscus* genome (PMSat): Evolution via copy number fluctuation

This subchapter summarizes the main features and findings of the satDNA sequence PMSat on its first report.

BASED ON THE RESEARCH ARTICLE:
Louzada S, Vieira-da-Silva A, **Mendes-da-Silva A**, Kubickova S, Rubes J, Adega F and Chaves R (2015) A Novel Satellite DNA Sequence in the Peromyscus Genome (PMSat): Evolution via Copy Number Fluctuation. *Molecular Phylogenetics and Evolution* 92: 193–203. https://doi.org/10.1016/j.ympev.2015.06.008.

For many years, the repetitive fraction of the *Peromyscus* genome was disregarded from the cytogenetic analyses, and although some studies reported the presence of satellite DNA (satDNA) at the (peri)centromeric region in some chromosomes' short arms (Hamilton et al. 1992), and in an interspersed location in *Peromyscus eremicus* chromosomes (Louzada et al. 2008), no DNA sequence information was available. Later on, the work performed by Louzada and colleagues, in 2015, characterized and provided molecular information on the monomer sequence of the major satDNA in *P. eremicus*, PMSat. The next sections focus on the main features and findings of this satDNA sequence.

II. 1.1. PMSAT - A NOVEL SATELLITE DNA FROM *P. EREMICUS* GENOME

Over the years, the development of effective methodologies for *de novo* isolation of repetitive sequences has been one of the main factors in the success of the knowledge and understanding of the functions of these sequences in genomes. One of these methodologies is laser microdissection procedures (Kubickova et al. 2002) that have allowed the isolation of centromeric repetitive sequences (Li et al. 2005; Pauciullo et al. 2006) in different mammalian genomes such as rodent genomes (Louzada et al. 2008). Isolation of centromeres from *P. eremicus* chromosomes was conducted by Louzada et al. (2015), by laser microdissection and subsequent amplification and cloning produced clones of a repetitive sequence. After sequencing, additional clones were also obtained by PCR amplification with specific primers and by digested genomic DNA with restriction enzymes. Southern blot analysis revealed the tandem genomic organization of the isolated sequence, exhibiting a monomeric size of 345 bp. This novel satDNA was named PMSat – **P**ero**m**yscus **Sat**ellite.

The PMSat clones isolated from *P. eremicus* genome revealed an identity ranging from 89.5 to 100 % (Table II.1.1) with a length variation extending from 343 to 464 bp and an average AT content of 55% (Figure II.1.1 a). In addition to southern blot hybridization pattern, also the dot plot analysis on PMSat clones confirmed a monomer size of 345 bp (Figure II.1.1 b,c).

Table II. 1.1. Matrix of sequence identity of PMSat isolated sequences in *P. eremicus* based in the alignment of Figure II.1.1.

	GQ902036	KC351938	KC351942	KC351943	KC351941	KC351939	KC351940
GQ902036		1.000	1.000	0.997	0.997	0.910	0.921
KC351938	1.000		1.000	0.997	0.997	0.910	0.921
KC351942	1.000	1.000		0.997	0.997	0.910	0.921
KC351943	0.997	0.997	0.997		0.997	0.907	0.918
KC351941	0.997	0.997	0.997	0.997		0.912	0.922
KC351939	0.910	0.910	0.910	0.907	0.912		0.895
KC351940	0.921	0.921	0.921	0.918	0.922	0.895	

The PMSat sequence corresponding to the accession number is available in Table II.1.2.

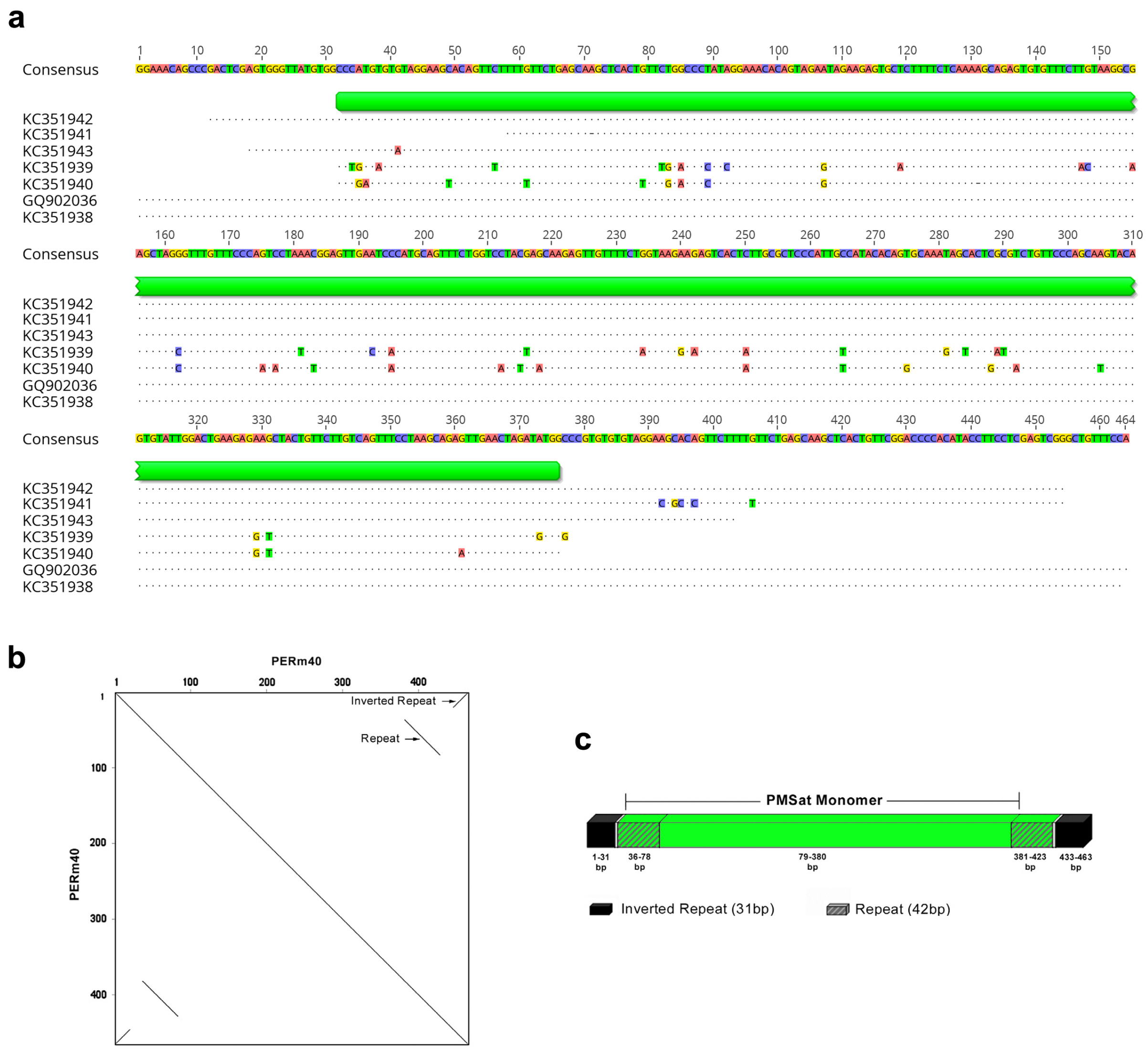


Figure II. 1.1. PMSat monomer sequence. (a) Alignment of *P. eremicus* PMSat isolated clones. The green box corresponds to PMSat monomer. Only differences in sequences are indicated, while the positions of sequence identity are represented by a dot. (b) Dot plot diagram of the clone GQ902036 (PERm40) compared to itself, showing two internal repeats (42 bp length) and two inverted repeats (31 bp length). (c) Schematic representation of the satellite detected features with indication of the monomer. Adapted from Louzada et al. (2015).

The physical mapping of PMSat on *P. eremicus* chromosomes (Figure II.1.2 a,b), conducted by fluorescent *in situ* hybridization (FISH) and sequential C-banding, showed a co-localization with constitutive heterochromatin (CH) (Figure II.1.2c,d), mainly at the (peri)centromeric region in all the autosomes and also in the entire p-arm (e.g. PER2, PER4 and PER9) or at the terminal region of some autosomes (e.g. PER6). On sex chromosomes, this sequence was also presented at the (peri)centromere and interstitially in the X chromosome p-arm (Figure II.1.2b).

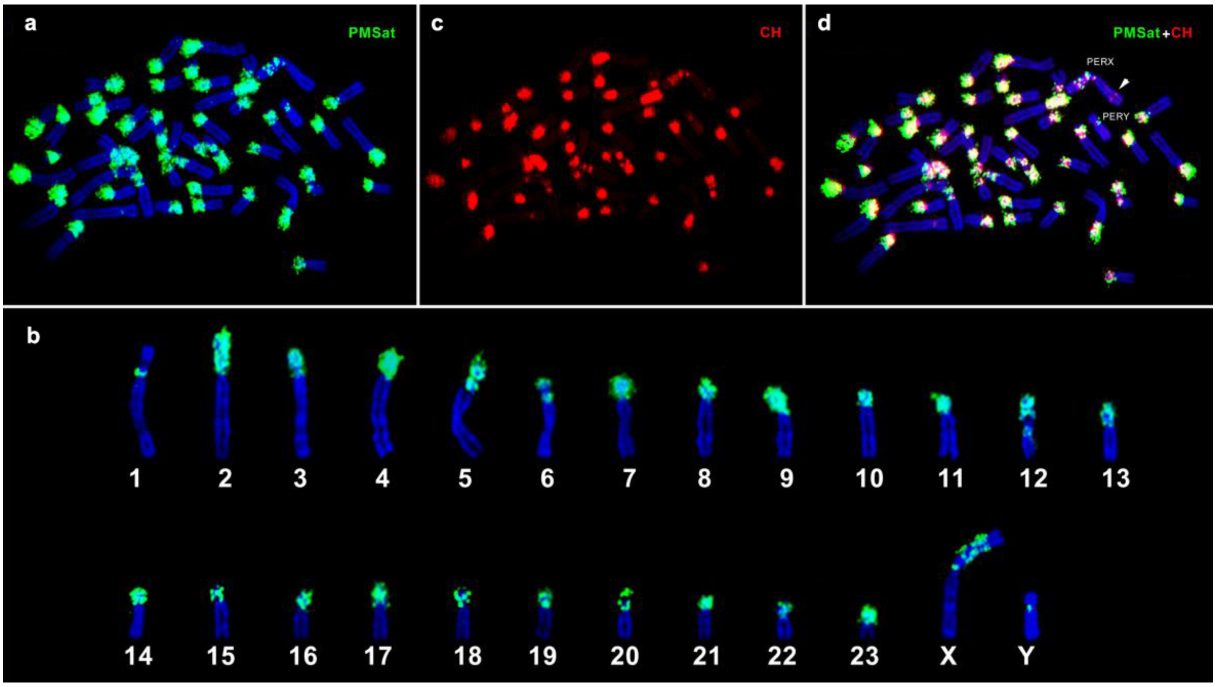


Figure II. 1.2. Physical mapping of PMSsat on *Peromyscus eremicus* chromosomes. (A) Representative in situ hybridization presenting the chromosomal localization of PMSat. (B) Haploid karyotype of *P. eremicus* chromosomes showing PMSat hybridization signal. (C) Same metaphase as in (A) after sequential C-banding. (D) Overlapping of PMSat hybridization signal with C-banding. The arrowhead indicates a chromosomal region containing CH but no PMSat signal. Adapted from Louzada et al. (2015).

II. 1.2. ORTHOLOGOUS PMSAT IN OTHER CRICETIDAE SPECIES

The presence of orthologous PMSat sequences was investigated in three distinct Cricetidae species, *Cricetus cricetus*, *Phodopus sungorus* and *Microtus arvalis*. The sequences were isolated by PCR amplification with specific primers, cloned and sequenced. The comparison between clones and *P. eremicus* consensus sequence revealed similarity of 95% with *C. cricetus*, 94% with *P. sungorus* and 100% with *M. arvalis* (Table II.1.2). Indeed, the PMSat clones showed a very high interspecies similarity with high conservation of the monomer sequence among species.

Table II. 1.2. Summary of the analysis in all PMSat isolated clones and Genbank accession number sequences.

Phylum	Species	%Similarity	Designation	Isolation	Length (bp)	% AT	Access. number
Craniata	*P. eremicus*		PERm40	Microdissection	464	54	GQ902036
			PERm57	Microdissection	463	54	KC351938
			PERp25	PCR	396	55	KC351941
			PERp45	PCR	443	54	KC351942
			PERp62	PCR	386	55	KC351943
			PERHaeIIIA	RE HaeIII	346	56	KC351939
			PERHaeIIIB	RE HaeIII	343	57	KC351940
	C. cricetus	95	CCRpA1	PCR	446	54	KC351944
			CCRpB1	PCR	399	54	KC351945
			CCRpC1	PCR	442	54	KC351946
			CCRpR1	PCR	411	54	KC351947
	M. arvalis	94	MARpA1	PCR	430	54	KC351948
			MARAp15	PCR	431	54	KC351949
			MARpB1	PCR	430	54	KC351950
	P. sungorus	100	PSUpA11	PCR	387	54	KC351951
			PSUpA25	PCR	392	54	KC351952
			PSUpE1	PCR	443	54	KC351953
			PSUpR1	PCR	425	57	KC351954

Similarity percentage refers to the number of identical nucleotide positions compared with PER consensus sequence.

Despite the presence of PMSat orthologous sequences, the physical mapping and southern blot hybridization revealed different results in the other three Cricetidae species from those observed in *P. eremiscus*, namely, no hybridization signal could be detected. This could be a consequence of copy number variations of PMSat content on these genomes. Using a new methodology, based on real-time quantitative PCR allied to TaqMan chemistry, the quantification of satDNA was conducted with a specific assay (two primers and a probe) designed based on the PMSat consensus sequence in a conserved region amongst all the species under study (Figure II.1.3). Although it was observed a high similarity among the PMSat monomers in the different Cricetidae species, PMSat family may also comprise other divergent monomers that were not detected. This approach was also described by our group in Paço et al. (2014) and revealed to be more specific than previous experiments that combine the use of standard primers with SYBR Green I chemistry to access the copy number of repetitive sequences (Navajas-Pérez et al. 2009). Absolute quantification revealed that PMSat comprises, at least, 20% of the *P. eremicus* genome (corresponding to at least 1.73x10^6 copies per genome), and the relative quantification showed, as suspected, a much lower amount of this satellite family in the others studied genomes, approximately 10^6-fold fewer copies compared to *P. eremicus* (Table II.1.3).

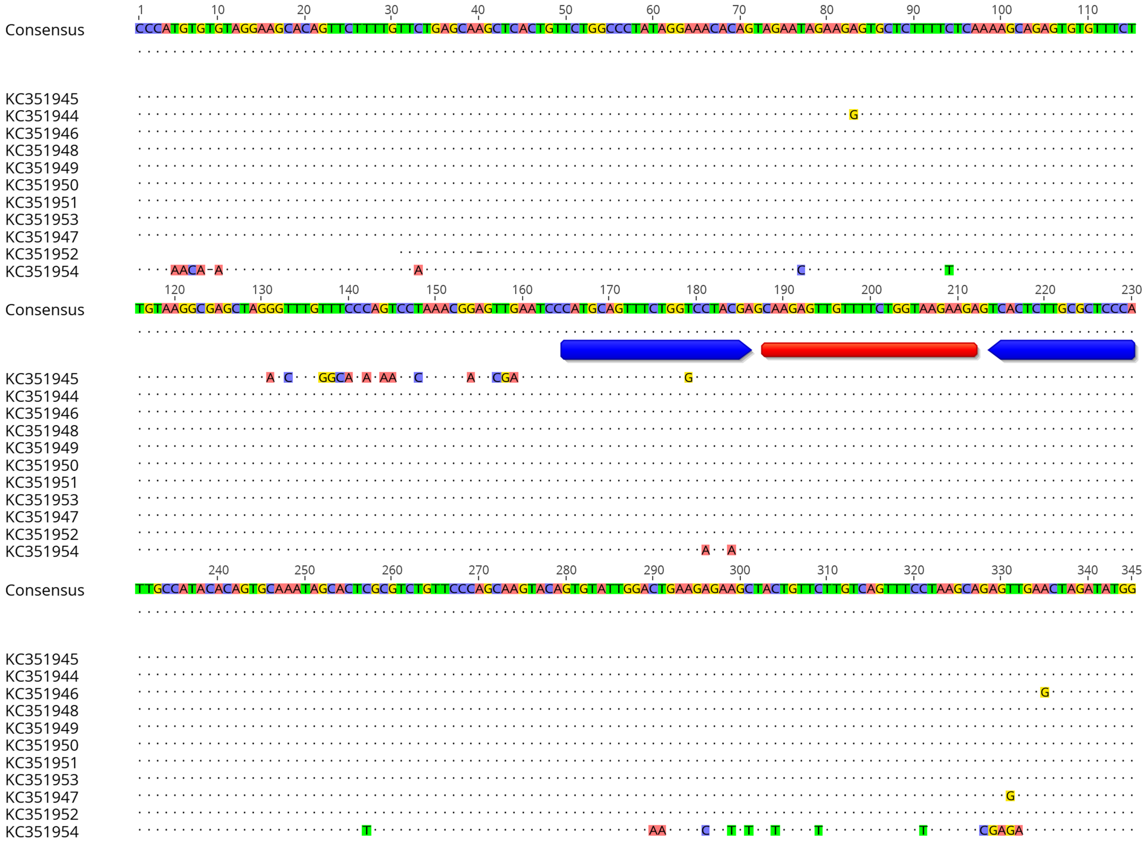


Figure II. 1.3. Alignment of PMSat monomer sequence. The image shows the alignment between PMSat monomer consensus sequence of *P. eremicus* with other species PMSat sequences. Only differences in sequences are indicated, while the positions of sequence identity are represented by a dot. The primer/probe used in the quantification experiments is indicated in blue and red, respectively. Adapted from Louzada et al. (2015).

Table II. 1.3. List of species and number of clones analyzed, percentage of similarity and quantification of PMSat satellite DNA family in the species genomes.

Species	Clones	Similarity (%)	PMSat copy number analysis
P. eremicus	7		≥ 1,73x106 copies/genome
			Relative quantification
C. cricetus	4	95	- $1x10^6$ fold
P. sungorus	4	94	- $6x10^6$ fold
M. arvalis	3	100	- $4x10^6$ fold

Similarity percentage refers to the number of identical nucleotide positions compared with PER consensus sequence (Figure II.1.1). The amount of PMSat in *C. cricetus*, *M. arvalis* and *P. sungorus* is presented as fold change relative to the amount determined for PER.

Altogether, the presence of PMSat orthologous sequences in non-*Peromyscus* species in a much lower copy number (comparatively to *P. eremicus*) suggests that PMSat is only amplified in the *Peromyscus* species and presents a typical satDNA sequence ‘‘behavior’’, i.e., located at the chromosomes' CH rich regions (at the pericentromeric regions in *P. eremicus*), highly amplified and organized in a tandem array fashion.

II. 1.3. EVOLUTION OF PMSAT SATELLITE DNA SEQUENCE

Due to the high interspecies PMSat sequence identity and a wide range of PMSat copy number content on the studied Cricetidae species, it could be hypothesized that PMSat evolves through copy number fluctuations, appearing in *P. eremicus* genome highly repeated and organized in tandem arrays. The amplification of repetitive DNA sequences has been attributed to distinct mechanisms, including unequal crossing-over and rolling circle amplification (Walsh 1987; Dover 2002; Plohl 2010).

On *P. eremicus* genome, the copy number variation of PMSat seems to be associated with the karyotype evolution of this species. As referred on Chapter I, *Peromyscus* genus reveals a high degree of karyotypic conservation, being the variations attributed to CH additions and pericentric inversions (Robbins and Baker 1981; Rogers et al. 1984). As an element of the CH (observed on *P. eremicus* genome), PMSat dynamics through amplification events certainly contributed to CH additions, resulting in the large CH blocks enriched in PMSat observed on *P. eremicus* chromosomes.

It has been postulated that satDNA sequences rapidly evolve even among related species based on the principles of concerted evolution, where a non-independent mode of monomers evolution results in the homogenization of the accumulated mutations of satDNAs within a genome (Dover 1986, 2002; Plohl et al. 2008). Interestingly, in non-*Peromyscus* species, PMSat orthologous sequences exhibit remarkable similarity to *Peromyscus* PMSat sequences but do not present the archetypical features of a satellite DNA sequence (i.e., highly repeated and organized in tandem arrays). These findings leave several questions open: Why is PMSat so conserved in these genomes? What is the function of this sequence? Some satDNAs persist in the genomes in a conserved manner and have been considered as "frozen" satDNAs (Plohl et al. 2010; Petraccioli et al. 2015). In fact, putative functional interactions of some segments inside the monomeric unit of a satDNA family can lead to low mutation rates by substitution (Plohl 2010). Altogether, PMSat findings suggest that this satDNA family is under evolutionary constraints in the studied species and may constitute a functional element in their genomes.

II.1.4. REFERENCES

Dover G. 2002. Molecular drive. Trends Genet. 18(11):587–589. doi:10.1016/S0168-9525(02)02789-0.

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

Hamilton MJ, Hong G, Wichman HA. 1992. Intragenomic movement and concerted evolution of satellite DNA in *Peromyscus:*evidence from in situ hybridization. Cytogenet Genome Res. 60(1):40–44. doi:10.1159/000133292.

Kubickova S, Cernohorska H, Musilova P, Rubes J. 2002. The use of laser microdissection for the preparation of chromosome-specific painting probes in farm animals. Chromosome Res. 10(7):571–577.

Li Y-C, Cheng Y-M, Hsieh L-J, Ryder OA, Yang F, Liao S-J, Hsiao K-M, Tsai F-J, Tsai C-H, Lin CC. 2005. Karyotypic evolution of a novel cervid satellite DNA family isolated by microdissection from the Indian muntjac Y-chromosome. Chromosoma. 114(1):28–38. doi:10.1007/s00412-005-0335-7.

Louzada S, Paço A, Kubickova S, Adega F, Guedes-Pinto H, Rubes J, Chaves R. 2008. Different evolutionary trails in the related genomes *Cricetus cricetus* and *Peromyscus eremicus* (Rodentia, Cricetidae) uncovered by orthologous satellite DNA repositioning. Micron. 39(8):1149–1155. doi:10.1016/j.micron.2008.05.008.

Louzada S, Vieira-da-Silva A, Mendes-da-Silva A, Kubickova S, Rubes J, Adega F, Chaves R. 2015. A novel satellite DNA sequence in the *Peromyscus* genome (PMSat): Evolution via copy number fluctuation. Mol Phyl Evol. 92:193–203. doi:10.1016/j.ympev.2015.06.008.

Navajas-Pérez R, Quesada del Bosque ME, Garrido-Ramos MA. 2009. Effect of location, organization, and repeat-copy number in satellite-DNA evolution. Mol Genet Genomics. 282(4):395–406. doi:10.1007/s00438-009-0472-4.

Paço A, Adega F, Me trovi N, Plohl M, Chaves R. 2014. Evolutionary Story of a Satellite DNA from *Phodopus sungorus* (Rodentia, Cricetidae). Genome Biol Evol. 6(10):2944–2955. doi:10.1093/gbe/evu233.

Pauciullo A, Kubickova S, Cernohorska H, Petrova K, Di Berardino D, Ramunno L, Rubes J. 2006. Isolation and physical localization of new chromosome-specific centromeric repeats in farm animals. Veterinarni Medicina. 51(No. 5):224–231. doi:10.17221/5541-VETMED.

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

Plohl M. 2010. Those mysterious sequences of satellite DNAs. Periodicum biologorum. 112(4):403–410.

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(1–2):72–82. doi:10.1016/j.gene.2007.11.013.

Plohl M, Petrović V, Luchetti A, Ricci A, Šatović E, Passamonti M, Mantovani B. 2010. Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. Heredity. 104(6):543–551. doi:10.1038/hdy.2009.141.

Robbins LW, Baker RJ. 1981. An assessment of the nature of chromosomal rearrangements in 18 species of *Peromyscus* (Rodentia: Cricetidae). Cytogenet Genome Res. 31(4):194–202. doi:10.1159/000131649.

Rogers DS, Greenbaum IF, Gunn SJ, Engstrom MD. 1984. Cytosystematic Value of Chromosomal Inversion Data in the Genus *Peromyscus* (Rodentia: Cricetidae). J Mammal. 65(3):457–465. doi:10.2307/1381092.

Walsh JB. 1987. Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics. 115(3):553–567.

II. 2. Genome-wide analysis of Tandem Repeats on the genome of *Peromyscus*

BASED ON THE RESEARCH ARTICLES:

Louzada S, Vieira-da-Silva A, **Mendes-da-Silva A**, Kubickova S, Rubes J, Adega F and Chaves R (2015) A Novel Satellite DNA Sequence in the Peromyscus Genome (PMSat): Evolution via Copy Number Fluctuation. *Molecular Phylogenetics and Evolution* 92: 193–203. doi:10.1016/j.micron.2008.05.008.

Mendes-da-Silva A, Santos HAF, Adega F, Gama-Carvalho M, Chaves R (2019) The major repetitive element of Peromyscus, PMSat, ensures mitotic fidelity and drives chromosome evolution. *In preparation*

ABSTRACT |

The continuous advances in genome sequencing technologies have provided an enormous amount of genomic data from hundreds of models and non-model species. Further, bioinformatics tools and strategies have been developed toward genome-wide identification of repetitive DNAs, namely satellite DNA (satDNA) – the Satellitome. Here, we describe the first genome-wide identification and analysis of tandem repeats on a Peromyscine species – *Peromyscus maniculatus*. A bioinformatics pipeline based on the Tandem Repeat Finder (TRF) algorithm and an integrated analysis of sequence similarity allowed the identification of 21 distinct families of large tandem repeats (array length longer than 2 kb) being the majority of them satellite- or transposable elements-related families. The major constituent of the *Peromyscus* satellitome corresponded to the PMSat satDNA family originally isolated on *P. eremicus* genome. *In situ* experiments conducted in four *Peromyscus* species (*P. eremicus*, *P. maniculatus*, *P. leucopus* and *P. californicus*) revealed that PMSat is mainly located at the active centromeres and pericentromeric regions of all chromosomes, as also at other constitutive heterochromatin rich regions as telomeres and p-arms of some chromosomes, maintaining a high degree of conservation in all the studied species despite the different number of copies of PMSat *per* genome. Our data strongly suggest that the evolution of PMSat was driven by copy number fluctuations and the high similarity among *Peromyscus* and non-*Peromyscus* species reflect non-concerted evolutionary events. Moreover, in light of the karyotype differences among *Peromyscus* species and the distinct redistribution of constitutive heterochromatin, we hypothesized that PMSat evolutionary molecular events may have promoted *Peromyscus* karyotype variations and genome evolution.

II. 2.1. INTRODUCTION

An important feature of eukaryotic genomes is their richness in repetitive sequences, namely satellite DNA (satDNA) sequences, which are mainly located at the heterochromatic regions of chromosomes (constitutive heterochromatin, CH), especially at the centromeric, pericentromeric and subtelomeric regions, but also at interstitial locations (Henikoff and Dalal 2005; Plohl et al. 2012). SatDNAs are traditionally organized in megabase-scale arrays of tandemly repeated monomers in a head-to-tail fashion (Plohl et al. 2014). Due to their repetitive nature and to the high frequency of variable sequences, in particular at the centromeres, the correct assembly of these sequences remains challenging and the genome projects have expurgated them from the public sequencing data. Consequently, even in the best studied genomes, as the human genome, satDNAs represent the principal unassembled elements (Jain et al. 2018; Lower et al. 2018), thus limiting our understanding of centromere organization and function. Notwithstanding, the synergy between emerging sequencing technologies and increasingly robust analytical algorithms has strengthened our progress towards end-to-end assemblies of entire chromosomes (Alkan et al. 2011; Jain et al. 2018; Lower et al. 2018). Therefore, some efforts have been made to explore sequencing data from recently sequenced genomes in an attempt to find repetitive elements, mainly satDNAs

(Komissarov et al. 2011; Melters et al. 2013; Ruiz-Ruano et al. 2016; de Lima et al. 2017; Palacios-Gimenez et al. 2017). These new strategies, in association with molecular and cytogenetic approaches, not only contribute to unveil these hidden genome components but also to increase the knowledge about their evolution and function.

Traditionally, satDNA families have been referred to as the most variable elements between genomes, as they usually evolve rapidly between species regarding composition, chromosome organization/location and array length (mainly by expansion and/or contraction events) (Plohl et al. 2008). This is a presumed consequence of concerted evolution, where different molecular mechanisms of non-reciprocal transfer lead to a rapid intraspecific homogenization of occurring changes (Plohl et al. 2012). SatDNAs are thus generally characterized by a high evolutionary mutation rate resulting in species-specific repetitive elements in some genomes, as it is the case of the human genome. Curiously, some satDNA families seem to contradict this fact and persist almost intact at the nucleotide sequence level in phylogenetically distant genomes for long evolutionary periods - the so called “frozen” satDNAs - even in low copy numbers (Mravinac et al. 2002; Mravinac et al. 2005; Petraccioli et al. 2015; Chaves et al. 2017). These findings clearly indicate a functional significance for these ubiquitous genome elements (Chaves et al. 2017) that were once considered “junk” DNA. In fact, several lines of evidence suggest that satDNA represents a dynamic component of mammalian genomes, playing important roles in structure and function (Shapiro and von Sternberg 2005; Biémont and Vieira 2006). Some satDNAs were already described as responsible for chromosomal rearrangements leading to karyotype variations and hence, to genome evolution (Slamovits and Rossi 2002).

The Rodentia genus *Peromyscus* (Cricetidae) comprises the most abundant and widely distributed group of North American mammals (Witmer and Moulton 2012). *Peromyscus* species have emerged as model systems for various aspects of human biology, including aging, epigenetics or cancer (Parnell et al. 2005; Ungvari et al. 2008; Shorter et al. 2012; Kaza et al. 2018), in addition to the study of chromosome evolution (Shorter et al. 2012; reviewed in Bedford and Hoekstra, 2015). Due to their abundance, *P. maniculatus* (deer mouse) and *P. leucopus* represent the most studied *Peromyscus* species.

Within *Peromyscus,* a high degree of conservation in chromosome number is observed, with all the 56 species presenting 2n=48. There is however a substantial variation in the number of chromosomal arms, which ranges from 52 to 96 as a result of heterochromatin additions and pericentric inversions (Rogers et al. 1984; Carleton and Musser 2005). Most of the cytogenetic studies performed in these species are in the fields of comparative genomics, phylogeny and chromosome evolution (e.g. Louzada et al. 2008; Paço et al. 2009; Brown et al. 2018). In contrast, studies that focus on the repetitive fraction of the genome are very scarce and the first molecularly characterized satDNA in this genus (PMSat in *P. eremicus)* having only been recently reported by our group (Louzada et al. 2015).

The availability of the primary genome assembly of the deer mouse, *P. maniculatus* (Pman_1.0 by Baylor College of Medicine, 2014), has allowed us to perform the work presented here, which is the first genome-wide identification and analysis of large tandem repeats (TRs) in this genome. For this purpose, we applied a bioinformatics pipeline based on the Tandem Repeat Finder (TRF) algorithm to identify TR families. The repeat sequences clustered into 21 new families, the majority of which correspond to satDNA or Transposable Elements (TE) and related repeats presenting a tandem organization. The largest TR family identified was the previously reported PMSat, the major satDNA of *P. eremicus* genome (Louzada et al. 2015). The molecular and cytogenetic validation studies performed in several *Peromyscus* species (*P. eremicus*, *P. maniculatus*, *P. leucopus* and *P. californicus*) revealed that PMSat is a component of the active centromere in these species, forming large sequence blocks at the (peri)centromeric regions of the chromosomes. The peculiarities of the evolution of this satDNA family seem to correlate with the evolution of the karyotype of *Peromyscus*.

II. 2.2. MATERIAL AND METHODS

Bioinformatics analysis

Peromyscus maniculatus bairdii (Pman_1.0, GenBank assembly accession GCA_000500345.1, BioProject_PRJNA53563) WGS scaffold sequences were obtained from NCBI in FASTA format. The scaffold nomenclature (1 to 30.921) was defined by default output order.

Tandem repeat search was performed using Tandem Repeats Finder (TRF) (Benson 1999), using the following parameters: *match*, *mismatch* and *delta* that were set to 2, 5, 7, respectively; *match_probability* was set to 80; *indel_probability* was set to 10; the *MinScore* (minimum alignment score to report) was set to 50 and *MaxPeriod* (maximum period size to report) was set to 2000. TRF output analysis was performed with custom in house scripts. Redundant entries from TRF output were eliminated (all embedded TR arrays were discarded; for the same sequence coordinates, a TR with a larger unit size was discarded). Each pair of arrays was compared using bl2seq from BLAST+ suite (Camacho et al. 2009). All pair matches with a score less than 90 were discarded to remove false-positive or suspicious alignments. The remaining arrays were separated in Families by Blast score (members in the same family have bl2seq match with score greater than 90). The resulting families were checked manually for errors. Repbase Rodentia database was used to compare all TRs with known repeats and validate the grouping into families. All matches to Repbase repeats with less than 80% similarity were discarded in order to remove false positive matches from Blast *vs* Repbase. Some TRs that were not grouped originally into families (orphan TRs), were grouped in clustered TRs as ‘transposable element (TE)-related’ due to their respective match on Repbase blast results. For blasting repetitive DNA, sequence alignments were performed with several changes in the search parameters: *max_target_seqs* and *num_descriptions* were set to 10,000, *evalue* was set to 10^{-16}, *word_size* was set to 10, *dust* was set to ‘no’ and *soft_masking* parameter was set to ‘false’. All other search parameters were set to default values. The search for CENP-B box motifs (wild type motif YTTCGTTGGAARCGGGA) on PMSat family arrays was performed using the Fuzznuc from EMBOSS on Geneious R9 version 9.1.2 (Biomatters) with a maximum of two mismatches.

Cell culture and isolation of DNA

Cell lines from *Peromyscus maniculatus* (48,XY), *P. californicus* (48,XX) and *P. leucopus* (48,XX) were provided by the *Peromyscus* Genetic Stock Center from the University of South Carolina (now available from the Coriell Institute). The cell line from *P. eremicus* was gently provided by the Department of Systematics and Evolution, Muséum National d'Histoire Naturelle, Paris, France. The first two cell lines were grown in Ham's F12/DMEM, *P. leucopus* and *P. eremicus* cell lines were grown respectively in EMEM and DMEM. All basal media were supplemented with 13% AmnioMax C-100 Basal Medium, 2% AminoMax C-100 supplement, 10% FBS, 100 U/mL/100 μg/mL of Penicillin/Streptomycin antibiotic mixture and 200 mM L-Glutamine (all from Gibco, Thermo Fisher Scientific). Cells were maintained at 37ºC in a humidified atmosphere of 5% CO_2.

Genomic DNA isolation from the different cell lines was carried out using QuickGene DNA Tissue Kit S (Fujifilm Life Science), according to the manufacturer's instructions.

PMSat isolation, cloning, sequencing and analysis

PMSat orthologous sequences from *P. maniculatus*, *P. leucopus* and *P. californicus* genomes were isolated by PCR amplification from the genomic DNA previously obtained from these genomes and using sequence-specific primers, as previously described for *P. eremicus* (Louzada et al. 2015). PCR amplification fragments were extracted from the agarose gel and purified using the QIAquick PCR purification Kit (QIAGEN). Fast DNA End Repair (Thermo Scientific) was performed for blunting and phosphorylation of DNA ends, and subsequently linked into the SmaI site of the plasmid pUC19 (Thermo Scientific) with T4 DNA ligase (Thermo Scientific). Transformation was performed in DH5α competent cells (Invitrogen, Thermo Fisher Scientific). Clones were screened using the β-galactosidase blue-white color system, and the selected ones were labeled with digoxigenin-11-d'UTP (Roche Diagnostics), validated by DNA-FISH (see below) onto *P. eremicus* chromosomes (to confirm orthologous sequences) and sequenced in the forward direction. Clone sequencing chromatograms and sequence alignments were performed using ClustalW cost matrix on Geneious R9 version 9.1.2 (Biomatters) with parameters set to default values.

Chromosome preparations, GTG-banding, DNA Fluorescent in Situ Hybridization (DNA-FISH) and CBP-Banding Sequential to FISH

Fixed chromosome preparations were obtained from the cell lines referred bellow using standard procedures described elsewhere (Chaves et al. 2004). Air-dried chromosome preparations were aged overnight at 65ºC and subsequently submitted to standard procedures of G-banding with Trypsin and revealed with Giemsa (GTG-banding) (Seabright 1971). Chromosome preparations were fixed in 3% formaldehyde and subjected to sequential physical mapping of PMSat by DNA fluorescent *in situ* hybridization (DNA-FISH) by routine procedures (Schwarzacher and Heslop-Harrison 2000). PMSat cloned sequences were labeled with biotin-16-dUTP (Roche Diagnostics) by PCR. The most stringent post hybridization wash was 50% formamide/2xSSC at 42ºC.

Sequentially to FISH, CBP-banding [C-bands by Barium hydroxide using Propidium Iodide (PI)] was performed according to standard procedures (Sumner 1972) with slight modifications (Adega et al. 2007).

The karyotypes for the species under analysis were organized based on the guidelines of the Committee for Standardization of Chromosomes of *Peromyscus* (1977) and Greenbaum et al. (1994) and are presented in Supplementary Figure II.2.1.

Immunofluorescence (IF) and DNA-FISH on unfixed metaphase spreads

Immunofluorescence on unfixed metaphase spreads was performed using a slight modification of the procedure previously described by Terrenoire et al. (2010). Cells in exponential growth were treated for 2 hours with colcemid (KaryoMax, Gibco) at 0.1 µg/ml. The cells were harvested by mitotic shake-off, washed twice with cold phosphate buffered saline by centrifugation at 1800 rpm for 10 minutes at 4°C, re-suspended in 75 mM KCl at a concentration of 2-3×10^5 cells/ml and left at 37ºC for 10 minutes. Cell suspension (200 µl) was cyto-spun (Hettich Rotofix 32A) onto glass slides at 1200 rpm for 10 minutes. Slides were then incubated in KCM buffer (120 mM KCl, 20 mM NaCl, 10 mM TrisHCl pH 8.0, 0.5 mM EDTA, 0.1% Triton X-100) for 10 minutes at room temperature. The CENP-A antibody (Cell Signaling) was diluted 200-fold in KCM supplemented with 1% BSA (Sigma-Aldrich and incubated at 37ºC for 1 hour. Slides were washed twice in KCM (5 minutes at room temperature) and a FITC conjugated secondary antibody diluted as previously described was then added and the slides incubated for a further hour at 37ºC. Slides were washed twice in KCM (5 minutes at room temperature), fixed in 4% (v/v) formaldehyde (10 minutes, room temperature), rinsed in deionized water and mounted in Vectashield mounting medium containing 4'-6-diamidino-2-phenylindole (DAPI) (Vector Laboratories). After image capture the slides were equilibrated in 50%formamide/2xSSC (v/v) for 48 hours and then DNA-FISH procedures were performed on the same slides.

PMSat Copy Number Quantification (absolute and relative) by TaqMan Assay

For PMSat quantification a TaqMan specific assay (primers/probe) previously used and described in Louzada et al., 2015 was performed. Absolute quantification in *Peromyscus* species was performed by the standard curve method, previously used in other satDNA copy number quantifications (Louzada et al. 2015; Chaves et al. 2017). A 10-fold serial dilution series of the plasmid DNA standard, ranging from 2×10^8 to $3.2\text{x}10^5$ copies, was used to construct the standard curve (5 points series dilutions). The concentration of the plasmid was measured using the NanoDrop ND-1000 (NanoDrop Technologies) equipment and the corresponding plasmid copy number was calculated using the following equation: DNA (copy number) = [$6{,}023\times10^{23}$ (copy number/mol) × DNA amount (g)] / [DNA length (bp) × 660 (g/mol/bp)], where Avogadro number is 6.023×10^{23} (copy number) / 1mol and the average molecular weight of a double-stranded DNA molecule is 660 g/mol/bp. In the respective formula the recombinant plasmid DNA length is 4242 bp (pDrive vector 3851 bp and the insert 391 bp).

C_T values in each dilution were measured using real-time qPCR with the TaqMan specific assay described above to generate the standard curve for PMSat. Briefly, the standard curve includes a plot of the C_T values *versus* the log concentration of the plasmid DNA standard. Genomic DNA of all studied species, the unknown total DNA sample (copy number of PMSat on the genome), was obtained by interpolating its C_T value against the standard curve. We used 1 ng genomic DNA in the PCR reaction. These reactions were performed in a total volume of 25 µL with 1.25 µL of the primer/probe assay mixture and 12.5 µL of TaqMan® Genotyping Master Mix (Life Technologies Applied Biosystems). This experiment was carried out in StepOne real-time PCR system (Life Technologies Applied Biosystems), where the samples were subjected to an initial denaturation at 95ºC (10 min), and then to 40 cycles at 95ºC 15 sec followed by 60ºC 1 min. All reactions were performed in triplicate, and negative controls (without DNA) were also run.

StepOne software (version 2.2.2, Life Technologies Applied Biosystems) was used to generate the standard curve and for data analysis. Only standard curves with the following parameters were considered to be typically acceptable: $R^2 > 0.99$ and slopes between −3.1 and −3.6 giving reaction efficiencies between 90 and 110%.

Antibodies

Cell signaling: anti CENP-A polyclonal rabbit (IF: 1:200, #2186). Millipore: anti-rabbit polyclonal FITC antibody (1:200, #AP132F). Sigma Aldrich: antidigoxigenin-50-TAMRA (1:200, #11207750910). Zymed: anti-mouse monoclonal FITC (1:200, #81-6511).

Microscopy and image acquisition

Chromosomes images were obtained using an Axio Imager Z2 microscope (Zeiss) coupled to a JAI Progressive Scan (CV-M4+CL) digital camera and Cytovision software (Genus, version 4.5.2). Digitized photos were prepared using Adobe Photoshop (version 7.0). Image optimization included contrast and color adjustments and affected the whole image equally.

Statistics and reproducibility

All data are presented as mean ± standard deviation (SD). Data were statistically analyzed in GraphPad Prism 7 (GraphPad Software, Inc.) in which statistical significance was determined using two-tailed Student's t-test for the comparison between two independent samples and analysis of variance (ANOVA) tests when more than two groups were under analysis. Fisher’s exact test was used to analyze the cell phenotypes significance. *p*-values: ns $p>0.05$, * $p\leq0.05$, ** $p\leq0.01$, *** $p\leq0.001$, **** $p\leq0.0001$.

II. 2.3. RESULTS

Genome-wide analysis of highly repetitive sequences in the deer mouse genome

We have taken advantage of the data available from the *Peromyscus* genome sequencing project (*Peromyscus maniculatus bairdii,* Pman_1.0, GenBank assembly accession GCA_000500345.1, BioProject_ PRJNA53563) to perform a systematic search for Tandem Repeats (TRs) in this genome. The Tandem Repeats Finder (TRF) algorithm (Benson 1999) was used to identify all the putative TRs presenting at least two tandem copies of a particular sequence, with a maximal unit size of 2 kb. Since only the scaffold assembly level was available on Pman_1.0, the search was performed on each scaffold sequence (Figure II.2.1). Due to the repetitive nature of the analyzed sequences, the initial TRF output included redundant results due to repeats in the same coordinates with different unit sizes. To eliminate these redundant entries, all the embedded arrays (i.e. arrays within other arrays) were removed and in the case of overlapping arrays, the TR with the largest unit size was discarded. Since the TRF algorithm is more precise in searching for large TRs (monomer length >50 bp), short repeats (monomer lengths <50 bp) were also excluded from the subsequent analysis. Additionally, the data obtained were further filtered for finding arrays larger than 2 kb so we could analyze solely large TRs. Using this approach, we found 1651 large TRs, with monomer and array lengths longer than 50 bp and 2 kb, respectively. The complete report of each TR is presented in Supplementary File 1.

Then, the TRs were compared two-by-two in the bl2seq program, and the score value was used as a measure of TR sequence similarity. According to Komissarov et al. (2011), two tandem repeats belong to the same family if they have a bl2seq match with a score greater than 90. This analysis classified the 1651 large TRs into orphan (OTRs, 32%) and clustered (CTRs, 68%), subdividing these last ones (1126) into 21 families (Table II.2.1; Figure II.2.2). The similarity of the identified CTRs with Rodentia known repeats was checked by blast search against the Rodentia Repbase repeat collection and the nucleotide database from the National Center for Biotechnology Information (NCBI) (Supplementary File 1), resulting in the identification of four families of transposable elements (TEs), three families of satDNA and fourteen families of unclassified tandem repeats within our dataset. The 'Unclassified' TR families were named according to the minimum monomer size found in each array, followed by the species abbreviation PM (*Peromyscus*).

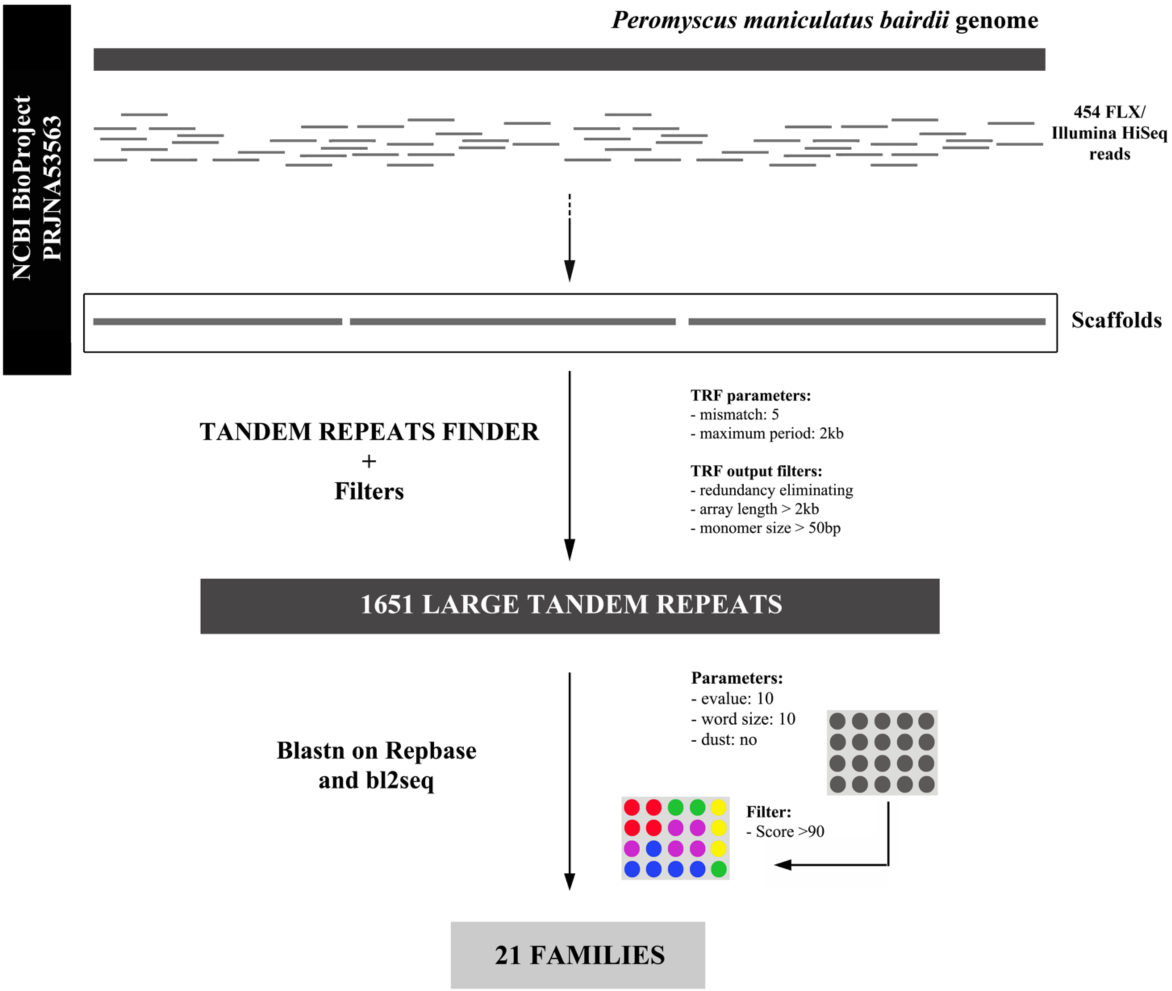


Figure II.2.1. General workflow of the large tandem repeats analysis on *Peromyscus maniculatus* genomic data. The search was performed on the scaffold level sequences and for each program the changed parameters are shown. The complete description of the workflow is described in Methods and Results sections.

Typically, TEs display a scattered distribution throughout the genome due to their ability to "jump" (transpose) to distinct genomic locations. Interestingly, the four families identified as related to disperse transposable elements (TE) are tandemly organized in the *P. maniculatus* genome. The L1-related family seems to be mostly formed by part of the ORF2, and the SINE-related family made by fragments of SINE elements, namely B1, B2 and B3 Rodentia retrotransposons. Moreover, several TRs seem to be composed by fragments of MTA transposons (MTA_related family), mostly MaLR-LTR, endogenous retrovirus (ERV3) and retrovirus-like elements (MYS1), which have structural similarities with MTA transposons. It should be noted that some TRs included in TE related families (based on bl2seq match) do not show any similarities with Repbase repeats.

Table II.2.1. *Peromyscus maniculatus* tandem repeats classification.

			Arrays	% of TR	Families	Subfamilies
ORPHAN TR			**525**	**31.8**	**-**	**-**
CLUSTERED TR			**1126**	**68.2**	**21**	**21**
	Satellite related		***886***	***53.7***	***3***	
		PMSat	875	53		1
		RNSAT1	8	0.5		3
		MMSAT4	3	0.2		1
	TE Related		***198***	***11.9***	***4***	
		L1_related	25	1.5		7
		SINE_related	62	3.7		12
		MTA_related	47	2.8		25
		Uncharacterized	64	3.9		N.A.
	Unclassified		***42***	***2.5***	***14***	
		TR_72B_PM	10	0.6		N.A.
		TR_273_PM	4	0.3		N.A.
		TR_72A_PM	3	0.2		N.A.
		TR_77_PM	3	0.2		N.A.
		TR_1084_PM	3	0.2		N.A.
		TR_1699_PM	3	0.2		N.A.
		TR_56_PM	2	0.1		N.A.
		TR_72C_PM	2	0.1		N.A.
		TR_141_PM	2	0.1		N.A.
		TR_498_PM	2	0.1		N.A.
		TR_1024_PM	2	0.1		N.A.
		TR_1175_PM	2	0.1		N.A.
		TR_1198_PM	2	0.1		N.A.
		TR_1612_PM	2	0.1		N.A.

The tandem repeats (TR) found at *P. maniculatus bairdii* WGS scaffolds (only monomer and array length longer than 50 bp and 2 kb, respectively, were considered) were classified by their bl2seq score. Two TR belongs to the same family if their score was higher than 90. The clustered TR (families composed by at least 2 members) was classified due their similarity (>80%) with known repeats deposited on Repbase Rodentia Database and NCBI. The complete characterization of each family is provided on Supplementary File 1. *TE Related* – Tandem repeats related to transposable elements (TE); L1 *related* – Long interspersed elements related, non-LTR retrotransposon; SINE_related – short interspersed nuclear elements related, non-LTR retrotransposon; MTA_related – mouse transcript retrotransposon related, LTR retrotransposon; PM – *Peromyscus*

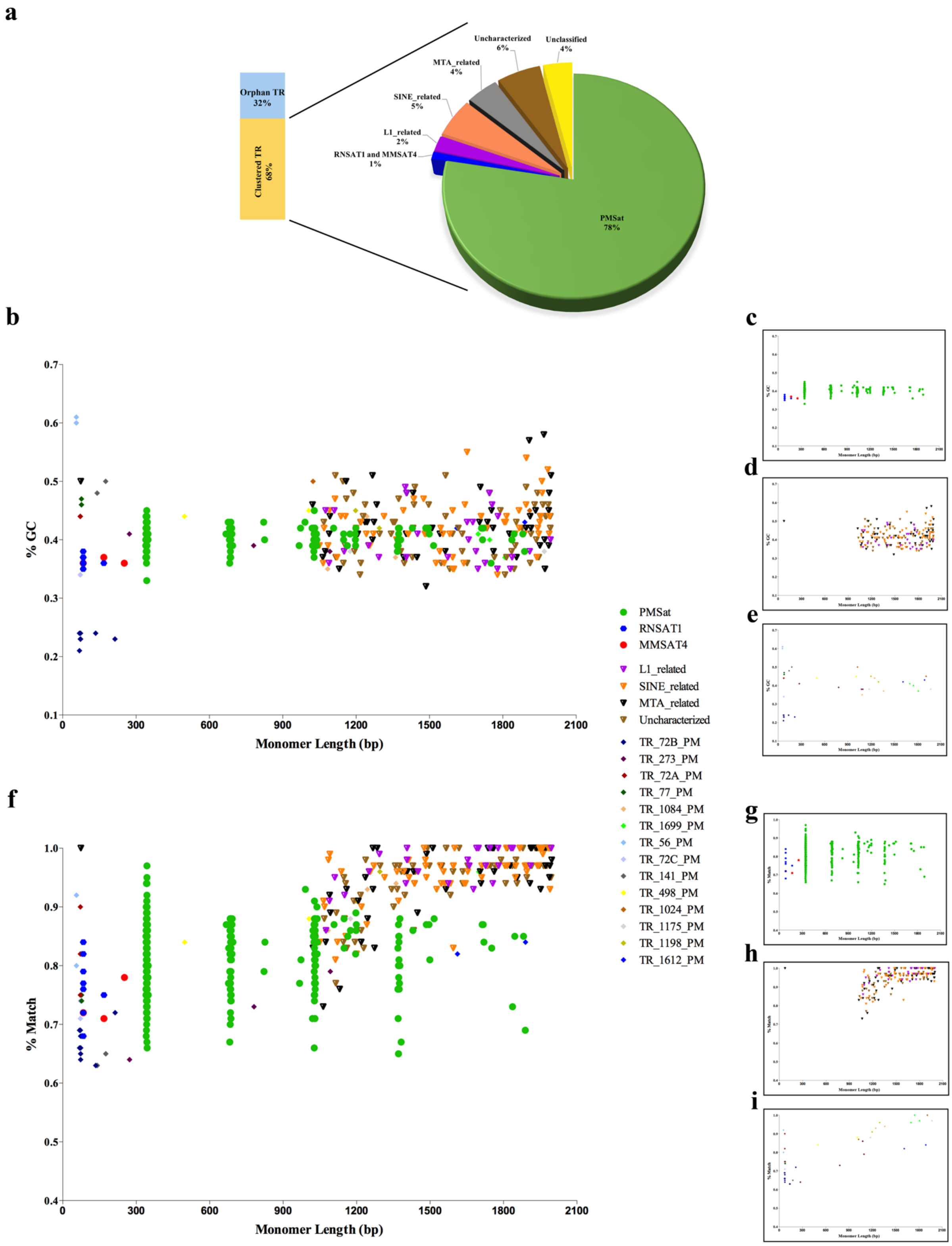


Figure II.2.2. Large tandem repeat families in *Peromyscus maniculatus* genome. (a) Overview of the large TR composition based on TRF output with CTRs highlighted. (b-e) Relationship between all the families depending on monomer length and GC content. Each clustered TR is separately presented: (c) Satellite_related; (d) TE_related; and (e) Unclassified. (f-i) Relationship between all the families depending on monomer length and the degree of repeat unit similarity. Each clustered TR is separately presented: (g) Satellite_related; (h) TE_related; and (i) Unclassified.

Regarding the three satDNA families identified, two were recognized as Rodentia satDNAs: RNSAT1, originally isolated from rat, found in eight arrays; and MMSAT4, a mouse satDNA, found in three arrays. The third and most prevalent family is the previously described *Peromyscus* satDNA (PMSat), which appeared in 875 arrays (Louzada et al. 2015). Thus, our bioinformatics search for *P. maniculatus* TRs identified for the first time the presence of two Rodentia satDNAs in this genome (Table II.2.1; Figure II.2.2). Although presenting a different nucleotide sequence, the monomeric unit of both rodent satDNA families is described as having 168 bp. In *P. maniculatus* genome, in addition to this monomer, we identified repeat size units of 84 bp for both families and 252 bp for MMSAT4. This may indicate the presence of variants/subfamilies of these satDNA families in *Peromyscus*. In fact, the results of the Repbase blastn search recognized three RNSAT1 variants already deposited on Repbase, namely RNSAT1a, b and c (Table II.2.2). All the *P. maniculatus* RNSAT1 and MMSAT4 arrays are AT-rich, with a maximum GC content of 38% and 37%, respectively (Figure II.2.2 b, c). The consensus sequence for each array showed a high similarity with the Repbase consensus sequence for each of the satDNA families (75.4 -89.3% for RNSAT1 and 79.8 – 86.8% for MMSAT4). However, within the array, a low identity was found between monomers (about 76.6% of matches among RNSAT1 arrays and 73.7% among MMSAT4 arrays) (Table II.2.2; Figure II.2.2 f, g).

Finally, all CTR families were compared in terms of monomer length, GC content and inter-repeat units' similarity (Figure II.2.2 b, f). The relationship between these parameters revealed that some of the unclassified families show similar molecular features to the satellite related families: TR_56_PM, TR_72A_PM, TR_72B_PM, TR_72C_PM, TR_141_PM and TR_273_PM (Figure II.2.2 b, e, f, i).

Table II.2.2. RNSAT1 and MMSAT4 satellite related families.

F	SF	N	Scaffold	Coordinates	Length	Unit Size	Copy number	%GC	%match	%id.
RNSAT1	**RNSAT1a**	4	162	6626177--6628249	2074	84	24,7	37	79	89,3
			434	527291--529595	2335	84	27,8	35	72	86,9
			723	1218507--1220682	2192	84	26,1	38	75	85,7
			973	220921--223035	2133	84	25,4	37	76	83,3
	RNSAT1b	2	10	177211--179697	2486	168	14,8	36	82	75,4
			1932	66142--68991	2856	84	34	36	77	82,1
	RNSAT1c	2	596	72004--75055	3066	84	36,5	37	68	77,4
			1058	193388--196605	3217	84	38,3	36	84	78,6
MMSAT4		3	820	148404--150931	2545	252	10,1	36	78	85,7
			1056	101425--103548	2133	84	25,4	36	71	79,8
			1126	190181--192327	2150	168	12,8	37	72	86,8

Subfamilies were ordered by scaffold number. Unit Size in bp; F - Family; SF - Subfamily; N - number of arrays; Coordinates - nucleotide position on the corresponding scaffold; Length – in bp; Copy number – number of monomers in the array; % GC – mean array GC content; % match – mean agreements between monomers in the array; % id – alignment identity between the array *consensus* sequence and the corresponding sequence on Repbase Rodentia Database.

PMSat is the major SatDNA family in Peromyscus maniculatus

As mentioned above, the most prominent TR identified in this study was PMSat satDNA sequence (Table II.2.1; Figure II.2.2), previously experimentally isolated and cloned by our group (Louzada et al. 2015) from *P. eremicus* genome and from other species' genomes belonging to the order Rodentia. The bioinformatics approach now conducted showed that these sequences are present in 231 scaffolds, totalizing 875 independent arrays, with the longest having ~41 kb (Figure II.2.3 a, b; Supplementary File 1). The PMSat family was thus estimated to account for over 0.2% of the total length of the sequenced *P. maniculatus* genome.

The scaffold encompassing the highest PMSat representation contains 34 arrays and a total of 352 PMSat monomers, spanning over ~ 228 kb (scaffold_266). Interestingly, some smaller scaffolds are composed entirely by PMSat (see for instance scaffold_19116 with ~7kb and scaffold_19117 with ~3kb). The PMSat family is AT-rich (55 to 67%) showing high intra-monomeric similarity within each array (~84%) (Figure II.2.2 b, c, f, g). The majority of the arrays exhibit a monomer size varying between 341 and 345 bp (Figure II.2.3). This is in agreement with the experimental characterization performed by Louzada et al. (2015), that identified a common monomer size with approximately 345 bp. The current study also identified the presence of arrays displaying repeat units of dimers (684 to 688 bp) and trimers (1027 to 1031 bp) of PMSat. Furthermore, these units are repeated in multimers, suggesting a possible higher order repeat (HOR) organization structure. However, the dot matrix analysis of scaffolds composed entirely by PMSat monomers shown no signs of HOR structures' presence (Figure II.2.4 a, b). Indeed, the multimers found could be merely the result of sequence divergence originating sequence variants composed by multimers. Furthermore, the pairwise comparisons between PMSat monomers within a scaffold are suggestive of recombination processes amongst PMSat sequences (Figure II.2.4 c).

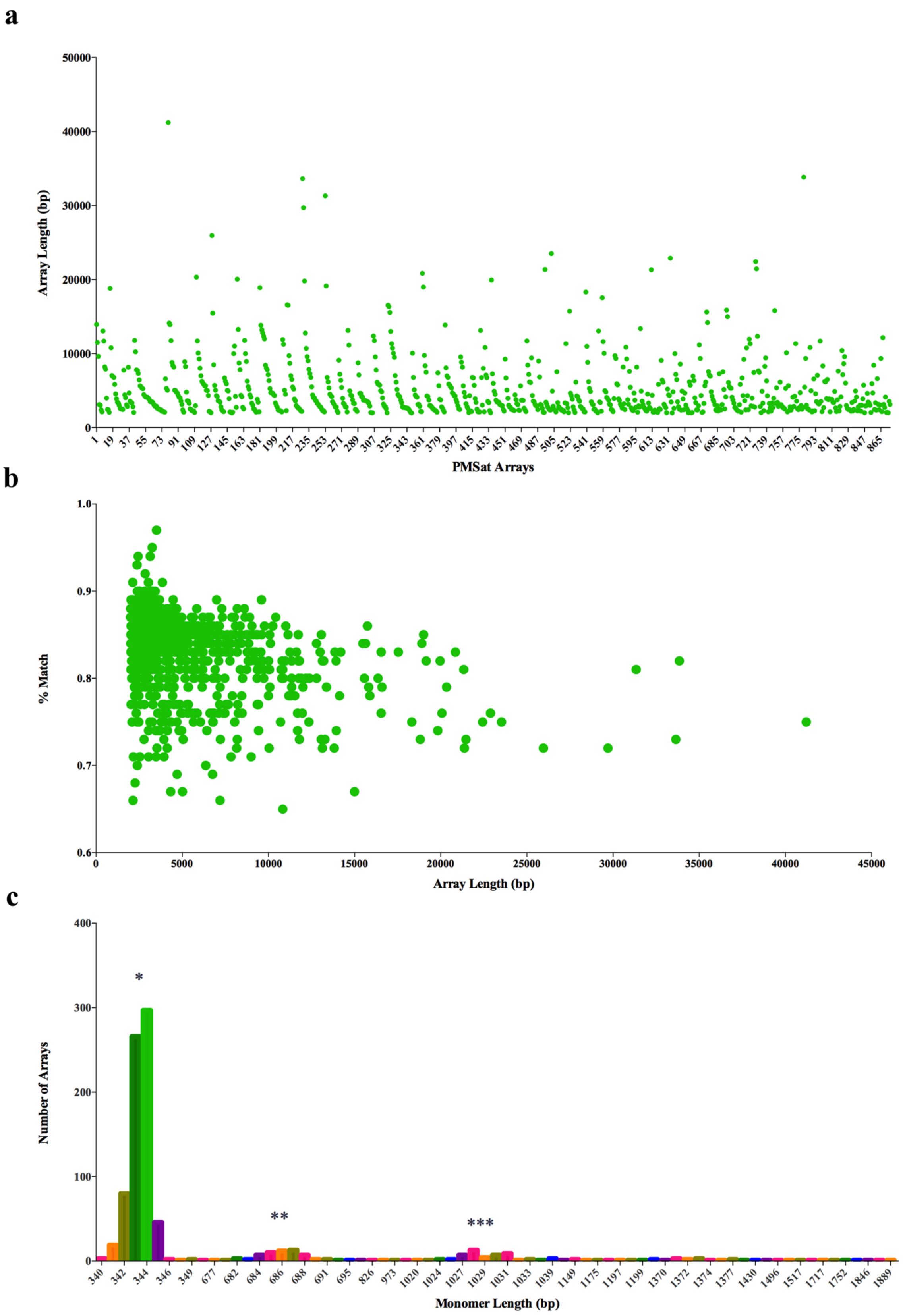


Figure II.2.3. PMSat distribution on *Peromyscus maniculatus* genomic data. (a) Array length distribution among the 875 PMSat arrays. (b) Relationship between the degree of monomer units similarity and array length. (c) PMSat repeat unit distribution shows a predominant PMSat repeat unit with 341 to 345 bp; and the presence of three picks that represent monomers composed by one (*), two (**) or three (***) repeat units.

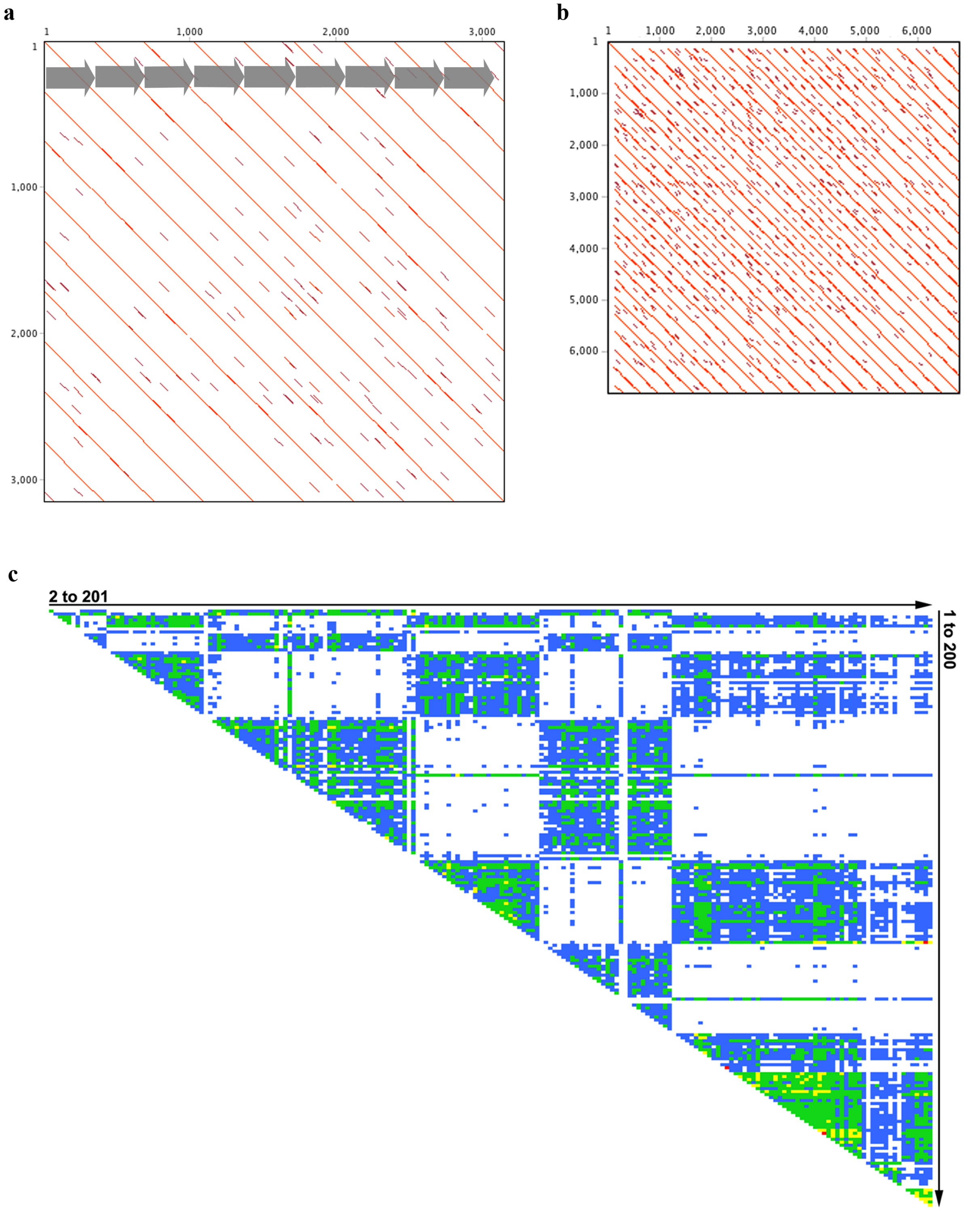


Figure II.2.4. Dot-plot analysis and distance matrix of pairwise alignment. Dot-plot matrix analysis in *P. maniculatus.* PMSat scaffolds 19117 (a) and 19116 (b) for signs of HOR structures. The criterion in the analysis was that a 49- or 50 nucleotide match should exist over a window of 50 nucleotides. If a line spanning more than two repeat units on a dot matrix, it can be interpreted as a putative HOR structure (Sujiwattanarat et al. 2015). However, only the pattern with single repeat units (~ 345 bp) was detected (represented by gray arrows). (c) Distance matrix of pairwise alignments of PMSat repeats in scaffold_2261 (Supplementary Figure II.2.2). The distances were made using the alignment algorithm CLUSTALW. The horizontal and vertical axes of each matrix represent consecutive repeats contained in the scaffold sequence. Cells showing nucleotide identities of ≥90, 85–89, 80–84 and 75–79% are in red, yellow, green and blue respectively. The same matrices containing

the identity values in the cells are shown in Supplementary File 2. All matrixes were generated by Geneious R9 version 9.1.2. (Biomatters) under default settings.

PMSat SatDNA has the hallmarks of a true centromeric sequence

A common feature in most animal and plant centromeres is their abundant content in tandem repeats. The widespread presence of PMSat in a repeated fashion in the centromere region of the *P. eremicus* genome, together with its richness in the genome of *P. maniculatus* determined by our bioinformatics approach, hypothesizes a close relationship between this sequence in both genomes and foresees a centromeric location in the entire *Peromyscus* genus.

Although the divergence of centromeric sequences among species seems to be the rule, these satDNA sequences appear to keep a conserved DNA-binding domain for the centromeric protein CENP-B (CENP-B box), which forms a stable complex with CENP-A nucleosome (Fujita et al. 2015). The CENP-B box is composed by a 17 bp motif (YTTCGTTGGAARCGGGA), in which the underlined nucleotides make the core recognition sequence (CRS), composed by three different binding sites (Tanaka et al., 2001; Masumoto et al., 2004). A search for the CENP-B box motif was thus conducted for all the scaffolds that presented PMSat arrays using the Geneious R9 (Biomatters) software. The CRS was used as the reference motif (CRS*wt*), leading to the identification of a total of 40683 similar motifs scattered across all scaffolds and located within PMSat sequences/arrays (Supplementary File 3), 29% of which were identical to CRS*wt*. The remaining motifs contained 1-2 mismatches and were termed CRS*var* (see for example the scaffold_19.117 on Figure II.2.5. a, b). It is worth noticing that some of the PMSat monomers that presented CRS*wt* motifs are located in scaffolds whose sequences are smaller than 2 kb and flanked by gaps, which escape from our TRF analysis (e.g. scaffold_1.708; Figure II.2.5. c).

To experimentally analyze the genomic context of PMSat, its nucleotide sequence was isolated and the genome organization was analyzed in other *Peromyscus* genomes, including *P. maniculatus*, *P. leucopus* and *P. californicus*. Sequence-specific primers were designed and used for PCR amplification experiments considering *P. eremicus* the control (Louzada et al. 2015). Amplification was successful in all the species under analysis and the PCR fragments were subsequently cloned. Three clones in *P. maniculatus* and two clones both in *P. leucopus* and *P. californicus* were selected for sequencing. All the analyzed sequences shared 99 to 100% identity across all nucleotide positions (Supplementary Tables

II.2.1 and II.2.2). These sequences showed a CRS*var* with two mismatches (Figure II.2.5 d), a motif that was also found in our bioinformatics analysis in distinct scaffolds enriched in PMSat arrays (Supplementary Figure II.2.3).

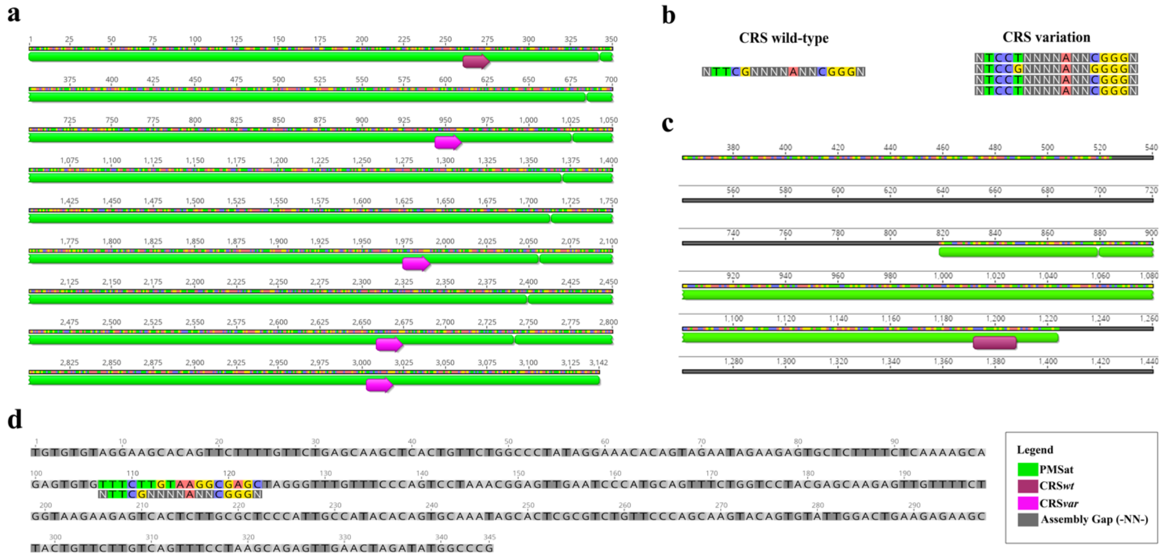

a
b
CRS wild-type
NTTCGNNNNANNCGGGN
CRS variation
NTCCTNNNNANNCGGGN
NTCCGNNNNANNGGGGN
NTCCTNNNNANNCGGGN
NTCCTNNNNANNCGGGN
c
d
TGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCTGGCCCTATAGGAAACACAGTAGAATAGAAGAGTGCTCTTTTCTCAAAAGCA
GAGTGTGTTTCTTGTAAGGCGAGCTAGGGTTTGTTTCCCAGTCCTAAACGGAGTTGAATCCCATGCAGTTTCTGGTCCTACGAGCAAGAGTTGTTTTCT
NTTCGNNNNANNCGGGN
GGTAAGAAGAGTCACTCTTGCGCTCCCATTGCCATACACAGTGCAAATAGCACTCGCGTCTGTTCCCAGCAAGTACAGTGTATTGGACTGAAGAGAAGC
TACTGTTCTTGTCAGTTTCCTAAGCAGAGTTGAACTAGATATGGCCCG
Legend
PMSat
CRSwt
CRSvar
Assembly Gap (-NN-)

Figure II.2.5. CENP-B box motifs on PMSat monomers. (a) The scaffold_19.117 are composed by 9 PMSat monomers, in which the first repeat unit possesses the wild type of the core recognition sequence (CRSwt) of CENP-B box, and the remaining repeat units presents a CRS with 1/2 mismatch or absent. The CRS*wt* and CRS*var* on these scaffold is shown in (b). (c) The initial region of scaffold_1.708 present a unique PMSat repeat that present a CRS*wt* and is flanked by assembly gaps (represented in grey); these PMSat repeats are not detected by the Tandem Repeat Finder (TRF) algorithm due their array length <2 kb. (d) *Consensus* sequence of the isolated clones from *Peromyscus* species shown a CRS*var* with two mismatches.
The corresponding colours for PMsat repeats, CRS*wt*, CRS*var* and assembly gaps are shown in the figure lower right corner. The nucleotides present in (b) and (d) follow the same corresponding code colours: Adenine (A) in red, Thymine (T) in green, Guanine (G) in yellow, and C (cytosine) in purple.

We physically mapped PMSat clones onto *Peromyscus* chromosomes using fluorescent *in situ* hybridization (FISH). The species belonging to the genus *Peromyscus* are characterized by exhibiting the same chromosome number (2n=48), with variation at the fundamental number (FN). In all the species analyzed, *P. maniculatus*, *P. leucopus* and *P. californicus,* PMSat presents a chromosome distribution characteristic of a satellite repeat, organized in large blocks at the (peri)centromeric region of all autosomes and in the sex chromosomes (Figure II.2.6 a). Of note, FISH technique resolution only allows identification of PMSat arrays at the (peri)centromeric region (where these are heavily clustered) and does not allow the physical discrimination between centromeric and pericentromeric domains. In *P. californicus* (PCA, FN =54), the sequence was only found at this (peri)centromeric region. However, in some chromosomes of *P. maniculatus* (PMA, FN= 86) and *P. leucopus* (PLE, FN=70), PMSat was also found at the chromosomes' p-arms, namely PMA14, 17, 18, 22 and

PLE11, 18, 22 and at the telomeric region of PMA11 and PLE23. In *P. eremicus* (PER; FN= 96) chromosomes, PMSat is widely distributed throughout the p-arms (note that all PER chromosomes are submetacentric), as described by Louzada et al. (2015) (Supplementary Figure II.2.4). C-banding sequential to FISH revealed a co-localization of PMSat sequence with constitutive heterochromatin in all the *Peromyscus* species (Figure II.2.6 a; Supplementary Figure II.2.5). FISH was further coupled to an immunostaining against the CENP-A protein on *Peromyscus* chromosomes in order to confirm the centromeric nature of the PMSat sequence, with results showing a perfect co-localization between the two (Figure II.2.6 b).

The analysis of PMSat copy number was performed in the *Peromyscus* genomes using a previously established real-time qPCR approach based on TaqMan chemistry for repetitive sequences (as described in Paço et al. 2014; Louzada et al. 2015). Absolute quantification using the standard curve method (Supplementary Figure II.2.5) revealed significant differences in the copy number of PMSat in *P. eremicus* genome compared with the other *Peromyscus* (Figure II.2.6 c). Our quantification results for PMSat in *P. eremicus* are in agreement with those of Louzada et al. (2015), revealing that, at least, 20% of the genome is composed by PMSat. Since the annotation of the genome size/molecular weight is not available for *P. leucopus* or for *P. californicus* genomes, the results were analyzed in terms of copy number variation between the species, using the genome of *P. maniculatus* as reference. This analysis revealed that *P. eremicus* is the genome presenting the highest number of PMSat copies (Figure II.2.6. c; Supplementary Table II.2.3).

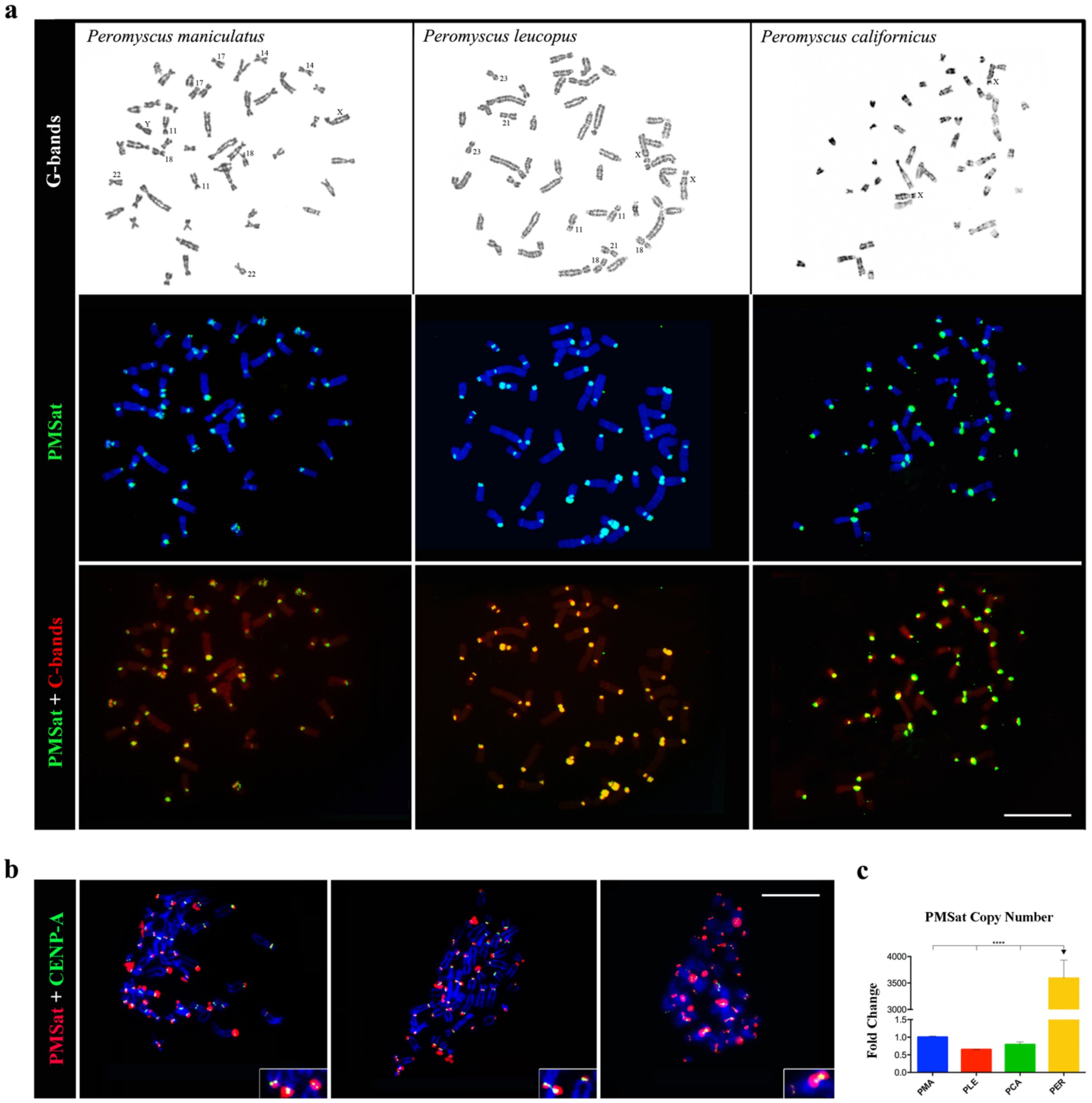


Figure II.2.6. PMSat DNA profile on *Peromyscus* species. (a) Physical mapping of PMSat on *P. maniculatus*, *P. leucopus* and *P. californicus* chromosomes. Representative G-banded metaphases with sequential DNA-FISH presenting the chromosomal localization of PMSat (green signals); chromosomes were counterstained with DAPI (blue). The same metaphases after sequential C-banding (chromosomes counterstained with propidium iodide, red) revels chromosomal regions containing CH, which when overlap with PMSat hybridization results in yellow signals. The karyotypes of the same metaphase are presented in Supplementary Figure II.2.1. (b) Immuno-localization of the centromeric protein CENP-A (green) on *P. maniculatus*, *P. leucopus* and *P. californicus* metaphases with sequential localization of PMSat (red). Some examples of co-localization signals (yellow) are highlighted. Scale bars represent 10 µm in all the panels. (c) PMSat copy number fold change in the different analysed genomes is indicated, considering *P. eremicus* as the reference genome. Values are mean ± SD of three replicates. ****P ≤ 0.0001 as determined by one-way ANOVA.

II. 2.4. DISCUSSION

Since its discovery, satDNA has been the most enigmatic fraction of eukaryotic genomes. An increasing number of studies reinforce the significance of satDNA in genome plasticity and regulation. Assessing the entire collection of satDNA families within a genome has been a major challenge in the new genomic era. Next Generation Sequencing (NGS) technologies have provided an increasing number of sequenced genomes, while new and efficient bioinformatic tools have been specifically developed toward genome-wide identification of repetitive DNAs. Currently, we have new tools and strategies to theoretically access the whole collection of satDNAs from a given genome – termed as the "Satellitome" (Ruiz-Ruano et al. 2016).

The genomic era, with the synergy between *in silico* and *in situ* approaches, opens new perspectives not only for disclosing the fundamental features of satDNA (structure, composition, origin and evolution) but for unveiling the universal framework for understanding the roles of repetitive DNAs as a whole.

The repeatome of the deer mouse genome

Here, we report for the first time, a genome-wide analysis of tandem repeat elements (TRs) on *P. maniculatus bairdii* genome, a *Peromyscus* species whose sequencing genome data assemble (at the scaffold level) is available. For this purpose, we defined a bioinformatics pipeline (Figure II.2.1) that applies the TRF algorithm at a scaffold assembly genome level conjugated with tactical filters for TR discovery. A similar strategy revealed to be effective in the analysis of the repetitive fraction in two mouse whole genome shotgun (WGS) assemblies (Komissarov et al. 2011).

In the present study, we found 1651 large TRs with a monomer and array length larger than 50 bp and 2 kb, respectively, which were clustered into 21 families according to their abundance and similarities with repetitive sequences already reported and deposited on the Repbase and/or NCBI databases. The majority of the families were clustered into satellite- or transposable elements (TE)- related repeats, presenting a tandem repeat organization (Table II.2.1). We identified, for the first time, in *P. maniculatus* genome, two orthologous satDNA families of the rat and mouse genomes: RNSAT1 and MMSAT4, respectively. Mouse and rat shared a common ancestor with the deer mouse lineage ~32.7 MYA (according to the "Time Tree of Life", http://www.timetree.org/). Therefore, these two satDNAs were present on this ancestral genome, and have, at least, 32.7 MYA.

The largest TR family identified is a satDNA already characterized by our group: PMSat (Louzada et al. 2015). This AT-rich satDNA, previously isolated from *P. eremicus* genome, revealed a monomer unit of 345 bp organized in large blocks in the heterochromatin at or near the centromeric region. According to our results, PMSat arrays are constituted not only by monomeric repetitions, but also by repeated units periodically repeated as dimers and trimers. Moreover, our data showed that PMSat arrays are characterized by a high similarity between monomers within an array (~84%) (Figure II.2.3 b). These values may be even underestimated because the more conserved monomers, which form long tracts of nearly identical TRs, represent a major challenge for genome assembly, besides the fact that (peri)centromeric regions are often neglected in genome assembly and annotation. In fact, the smaller arrays of PMSat showed a higher degree of repeat units' similarity (cf. Figure II.2.3. b). The same was observed experimentally between the PMSat clones isolated from the *Peromyscus* species under analysis (above 99-100%; Supplementary Table II.2.1).

PMSat is part of the functional centromere in Peromyscus

Its abundance and high similarity in the genome sequencing data and in previous experimental data (Louzada et al. 2015) place the PMSat as a potential candidate for a centromeric satDNA in *Peromyscus* genomes. Indeed, the high abundance of TRs at the centromeres of eukaryotic genomes drove Melters and co-authors (2013) to apply a bioinformatic pipeline to the available genomic data of hundreds of species in an attempt to identify centromeric candidate sequences. In addition to satDNA sequences, TEs are also present at centromeric regions (Plohl et al., 2014; Jain et al., 2018) and in our analysis we have also found repeats exhibiting homology with partial sequences of endogenous retroviruses (e.g. ERV2-3) and retrotransposons (e.g. MYS1).

The experimental analysis of isolation and mapping of PMSat was carried out on three *Peromyscus* species: *P. maniculatus, P. leucopus* and *P. californicus.* A combined analysis integrating the obtained *in silico* results, with those from the physical mapping (DNA-FISH) and CENP-A identification (immunocytochemistry), clearly demonstrated that PMSat is part of the active centromeres of all the chromosomes in the three species under study (Figure II.2.6 b). In *P. californicus* the sequence is confined to the centromeric region of the entire chromosomal complement, including the sex chromosomes. This was also observed by Smalec and colleagues (2019). Additionally, in *P. leucopus* and *P. maniculatus*, the sequence extends to the p-arm or is present at the telomeric region in some chromosomes, (PMA11, 14, 17, 18, 22 and PLE11, 18, 21 and 23) (Figure II.2.6; Supplementary Figure

II.2.6). Although PMSat physical distribution in some of the chromosomes of these two species is not apparently concordant with that of Smalec et al. (2019), the fact is that *Peromyscus* chromosomes exhibit a considerable number of intraspecific/intraindividual polymorphisms that can be responsible for the differences here observed. In fact, Baker et al. (1983) and Stangl and Baker (1984) refer to the existence of cytotypes or chromosome races when analyzing *P. leucopus* chromosomes.

We also disclosed the presence of the CENP-B box like motif (analyzed by its core recognition sequence, CRS) on PMSat monomers (Figure II.2.5), what foresees PMSat centromeric location and function. Indeed, the *in silico* analysis of *P. maniculatus* revealed the existence of the CENP-B box motif as part of a large amount of PMSat monomers scattered across all the scaffolds analyzed (Figure II.2.5 and Supplementary Figure II.2.3), either the conserved functional motif (CENP-B box motif – CRS*wt*, Masumoto et al., 2004) or presenting one to two mismatches (CRS*var*). We strongly believe that the monomers exhibiting the functional CENP-B box core recognition site (CRS*wt*) found throughout the PMA genome assembly are the ones located at the centromere core and are the ones CENP-A is binding, thus performing a part of its role on the centromeric activity.

PMSat evolution by copy number fluctuation

In *P. eremicus* PMSat shows the highest representativeness amongst the studied species, comprising large (peri)centromeric blocks that extend to the entire p-arm of the majority of the chromosomes, results that corroborate previous findings, that PMSat corresponds to, at least, ~20% of the genome, according to Louzada et al., 2015 (Supplementary Figures II.2.4 and II.2.6). Despite the specificity of the innovative methodology (i.e. TaqMan probe/primers) used for copy number analysis and the high similarity observed between monomers, is not possible to rule out the possibility of the existence of divergent monomers of PMSat that escaped our detection (Louzada et al., 2015). For this reason, the number of copies estimated is considered the minimum present in *P. eremicus* genome. The observed PMSat conservatism seems to widen to other Rodentia genomes also belonging to the Cricetidae family - *Cricetus cricetus*, *Phodopus sungorus* and *Microtus arvalis*, but once again the differentiating feature is the number of copies, that is residual in the latter (Louzada et al. 2015).

Once again, the data here assembled corroborates our previous hypothesis (Louzada et al. 2015) suggesting that copy number fluctuation drove PMSat evolution.

PMSat is a driver of Peromyscus karyotype evolution

As referred, the karyotype differences among *Peromyscus* genus rely in the variation of the fundamental number that ranges from 52 to 96, forming a karyotype of 48 chromosomes in all the studied species. It is established that the karyotypic differences, as well as many of the chromosome polymorphisms found in *Peromyscus* result from heterochromatin additions and pericentric inversions (Rogers et al. 1984; Greenbaum et al. 1994; Romanenko et al. 2012). In fact, PMSat, the major heterochromatin component in *Peromyscus* locates exactly at the supposed target chromosome regions involved in the kayotype evolution of the genus, namely at the (peri)centromeres, p-arms and telomeres. Furthermore, the fluctuations in copy number of this satDNA sequence represent the heterochromatin additions found in some species. In the light of *Peromyscus* karyotype evolution, we believe that PMSat was originally in a strict centric location, as observed in *P. californicus,* mainly composed of acrocentric chromosomes (Supplementary Figure II.2.6). The amplification of the satDNA sequences by mechanisms as unequal crossing-over and rolling circle amplification led to copy number fluctuations, resulting in progressive CH addictions and these consequently to chromosomal rearrangements (Wichman et al. 1991), such as the pericentric inversions clearly verified in some chromosomes of *P. leucopus P. maniculatus* and *P. eremicus*. This effect was quite notorious in *P. eremicus* chromosomes, where all the autosomes are submetacentric presenting very large CH blocks enriched in PMSat in the entire p-arms of some chromosomes (Supplementary Figure II.2.6), what is in accordance with the analysis of PMSat copy number by real-time qPCR (Figure II.2.6 c). It is accepted that repetitive sequences, with emphasis to satDNA, play an important role in mammalian genome evolution as hotspots for the occurrence of chromosomal rearrangements, due to the rapid evolution rates of this genomic fraction (Slamovits and Rossi 2002; Ruiz-Herrera et al. 2006; Adega et al. 2009).

The evolution of specific satDNA families by copy number variations, either by expansion and/or contraction of arrays, have been associated to chromosomal evolution in phylogenetically related species (e.g. Slamovits et al. 2001; Ellingsen et al. 2007; Cazaux et al. 2013; Chaves et al. 2017). Several are the examples where amplification, deletion and intragenomic movements of satDNA sequences seem to have been the engine promoting chromosomal evolution. Some of these include RPCS satellite DNA in rodents belonging to the *Ctenomys* genus (Ellingsen et al. 2007), the TLC satDNA in the subgenus *Mus* (Cazaux et al. 2013) or FA-SAT in a wide range of Bilateria genomes (Chaves et al. 2017), where large fluctuations in the number of copies of this repetitive family have been detected and seem to be causal for the sequence evolution. Also in this work reveals the involvement of a satDNA family – PMSat – in the chromosomal rearrangements that conducted *Peromyscus* karyotype evolution and that is clearly associated with the evolution of the sequence itself.

II. 2.4. CONCLUSION

This work clearly reinforces the potential of genome-wide analysis of newly sequenced genomes for a global characterization of TRs content. An integrated analysis with cellular complementary techniques, allows not only a physical characterization at a chromosomal level, assisting in the subsequent stages of sequencing projects (scaffold/contigs mapping), but also a molecular and functional characterization of satDNAs across genome evolution and function.

The characterization of the repetitive fraction of *Peromyscus maniculatus* assembled genome, revealed the presence of several not yet classified tandem repeats, transposable elements and three satDNA families. Amongst these, PMSat stands out, showing to be the most representative and conserved satDNA family in this genome. These results, together with previous ones from our group points PMSat as the major constituent of the *Peromyscus* satellitome. The molecular and cytogenetic/physical characterization of PMSat in the different analyzed *Peromyscus* species (*P. californicus*, *P. maniculatus*, *P. leucopus* and *P. eremicus*), revealed its presence at the centromeres, constituting large PMSat blocks, that extended to the short arm of the chromosomes in some species, being *P. eremicus* the most evident one, what seems to be the result of an incredible number of sequence amplifications, as verified by the huge number of PMSat DNA copies for us detected in comparison to the other species. In fact, copy number fluctuation seems to have been the evolutionary engine of this satDNA family that consequently resulted in the chromosome differences and rearrangements (promoted by centromeric sequences copy number fluctuations and inversions) behind the karyotype evolution in the genus. Besides its presumable role in the evolution of the *Peromyscus* karyotype, the remarkable sequence similarity found in PMSat orthologous sequences (on *Peromyscus* species and in non-*Peromyscus* species) clearly indicates a functional significance for this repeat. Additionally, the presence of the conserved DNA-binding domain for the centromeric protein CENP-B (CENP-B box) and the co-localization of the CENP-A protein that forms a stable complex with this motif on PMSat monomers proves its centromeric nature and anticipates its involvement in the centromeric function.

II. 2.5. REFERENCES

Adega F, Chaves R, Guedes-Pinto H. 2007. Chromosomal evolution and phylogenetic analyses in *Tayassu pecari* and *Pecari tajacu* (Tayassuidae): tales from constitutive heterochromatin. Genetics. 86(1):19–26. doi:10.1007/s12041-007-0003-1.

Adega F, Guedes-Pinto H, Chaves R. 2009. Satellite DNA in the Karyotype Evolution of Domestic Animals – Clinical Considerations. Cytogenet Genome Res. 126(1–2):12–20. doi:10.1159/000245903.

Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, Ryder OA, Purgato S, Zoli M, Della Valle G, Eichler EE, et al. 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 21(1):137–145. doi:10.1101/gr.111278.110.

Baker RJ, Robbins LW, Stangl FB, Birney EC. 1983. Chromosomal Evidence for a Major Subdivision in *Peromyscus leucopus*. J Mammal. 64(2):356–359. doi:10.2307/1380579.

Bedford NL, Hoekstra HE. 2015. The natural history of model organisms: *Peromyscus* mice as a model for studying natural variation. Elife. 4:e06813.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580. doi:10.1093/nar/27.2.573.

Biémont C, Vieira C. 2006. Junk DNA as an evolutionary force. Nature. 443:4.

Brown J, Crivello J, O'Neill RJ. 2018. An updated genetic map of *Peromyscus* with chromosomal assignment of linkage groups. Mamm Genome. 29(5–6):344–352. doi:10.1007/s00335-018-9754-7.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10(1):421. doi:10.1186/1471-2105-10-421.

Carleton MD, Musser GG. 2005. Order Rodentia. In: Wilson DE, Reeder DM, editors. Mammal Species of the World: A Taxonomic and Geographic Reference. 3rd ed. Baltimore: Johns Hopkins University Press. p. 745–1601.

Cazaux B, Catalan J, Justy F, Escudé C, Desmarais E, Britton-Davidian J. 2013. Evolution of the structure and composition of house mouse satellite DNA sequences in the subgenus *Mus* (Rodentia: Muridea): a cytogenomic approach. Chromosoma. 122(3):209–220. doi:10.1007/s00412-013-0402-4.

Chaves R, Ferreira D, Mendes-da-Silva A, Meles S, Adega F. 2017. FA-SAT Is an Old Satellite DNA Frozen in Several Bilateria Genomes. Genome Biol Evol. 9(11):3073–3087. doi:10.1093/gbe/evx212.

Chaves R, Frönicke L, Guedes-Pinto H, Wienberg J. 2004. Multidirectional chromosome painting between the Hirola antelope (*Damaliscus hunteri*, Alcelaphini, Bovidae), sheep and human. Chromosome Res. 12(5):495–503. doi:10.1023/B:CHRO.0000034751.84769.4c.

Committee for Standardization of Chromosomes of Peromyscus. 1977. Standardized karyotype of deer mice, *Peromyscus* (Rodentia). Cytogenet Genome Res. 19(1):38–43. doi:10.1159/000130792.

Ellingsen A, Slamovits CH, Rossi MS. 2007. Sequence evolution of the major satellite DNA of the genus *Ctenomys* (Octodontidae, Rodentia). Gene. 392(1–2):283–290. doi:10.1016/j.gene.2007.01.013.

Fujita R, Otake K, Arimura Y, Horikoshi N, Miya Y, Shiga T, Osakabe A, Tachiwana H, Ohzeki J, Larionov V, et al. 2015. Stable complex formation of CENP-B with the CENP-A nucleosome. Nucleic Acids Res. 43(10):4909–4922. doi:10.1093/nar/gkv405.

Greenbaum IF, Gunn SJ, Smith SA, McAllister BF, Hale DW, Baker RJ, Engstrom MD, Hamilton MJ, Modi WS, Robbins LW, et al. 1994. Cytogenetic nomenclature of deer mice, *Peromyscus* (Rodentia): revision and review of the standardized karyotype. Cytogenet Genome Res. 66(3):181–195. doi:10.1159/000133696.

Henikoff S, Dalal Y. 2005. Centromeric chromatin: what makes it unique? Curr Opin Genet Dev. 15(2):177–184. doi:10.1016/j.gde.2005.01.004.

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. Nature Biotec. 36(4):321–323. doi:10.1038/nbt.4109.

Kaza V, Farmaki E, Havighorst A, Crossland J, Chatzistamou I, Kiaris H. 2018. Growth of human breast cancers in *Peromyscus*. Disease Models Mec. 11(1):dmm031302. doi:10.1242/dmm.031302.

Komissarov AS, Gavrilova EV, Demin SJ, Ishov AM, Podgornaya OI. 2011. Tandemly repeated DNA families in the mouse genome. BMC genomics. 12(1):531. doi:10.1186/1471-2164-12-531.

de Lima LG, Svartman M, Kuhn GCS. 2017. Dissecting the Satellite DNA Landscape in Three Cactophilic *Drosophila* Sequenced Genomes. 3G: Genes Genomes Genet. 7(8):2831–2843. doi:10.1534/g3.117.042093.

Louzada S, Paço A, Kubickova S, Adega F, Guedes-Pinto H, Rubes J, Chaves R. 2008. Different evolutionary trails in the related genomes *Cricetus cricetus* and *Peromyscus eremicus* (Rodentia, Cricetidae) uncovered by orthologous satellite DNA repositioning. Micron. 39(8):1149–1155. doi:10.1016/j.micron.2008.05.008.

Louzada S, Vieira-da-Silva A, Mendes-da-Silva A, Kubickova S, Rubes J, Adega F, Chaves R. 2015. A novel satellite DNA sequence in the *Peromyscus* genome (PMSat): Evolution via copy number fluctuation. Mol Phyl Evol. 92:193–203. doi:10.1016/j.ympev.2015.06.008.

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

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. Chromosome Res. 12(6):543–556. doi:10.1023/B:CHRO.0000036593.72788.99.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14(1):R10.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14(1):R10.

Mravinac B, Plohl M, Mestrović N, Ugarković Đ. 2002. Sequence of PRAT Satellite DNA "Frozen" in Some Coleopteran Species. J Mol Evol. 54(6):774–783. doi:10.1007/s0023901-0079-9.

Mravinac B, Plohl M, Ugarković Đ. 2005. Preservation and High Sequence Conservation of Satellite DNAs Suggest Functional Constraints. J Mol Evol. 61(4):542–550. doi:10.1007/s00239-004-0342-y.

Paço A, Adega F, Guedes-Pinto H, Chaves R. 2009. Hidden heterochromatin: characterization in the Rodentia species *Cricetus cricetus*, *Peromyscus eremicus* (Cricetidae) and *Praomys tullbergi* (Muridae). Genet Mol Biol. 32(1):56–68. doi:10.1590/S1415-47572009000100009.

Paço A, Adega F, Me trovi N, Plohl M, Chaves R. 2014. Evolutionary Story of a Satellite DNA from *Phodopus sungorus* (Rodentia, Cricetidae). Genome Biol Evol. 6(10):2944–2955. doi:10.1093/gbe/evu233.

Palacios-Gimenez OM, Dias GB, de Lima LG, Kuhn GC e S, Ramos É, Martins C, Cabral-de-Mello DC. 2017. High-throughput analysis of the satellitome revealed enormous diversity of satellite DNAs in the neo-Y chromosome of the cricket Eneoptera surinamensis. Sci Reports. 7(1). doi:10.1038/s41598-017-06822-8.

Parnell PG, Crossland JP, Beattie RM, Dewey MJ. 2005. Frequent Harderian Gland Adenocarcinomas in Inbred White-Footed Mice (*Peromyscus leucopus*). Comparative Medicine. 55(4):5.

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

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(1–2):72–82. doi:10.1016/j.gene.2007.11.013.

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA evolution. In: Repetitive DNA. Vol. 7. Karger Publishers. p. 126–152.

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

Rogers DS, Greenbaum IF, Gunn SJ, Engstrom MD. 1984. Cytosystematic Value of Chromosomal Inversion Data in the Genus *Peromyscus* (Rodentia: Cricetidae). J Mammal. 65(3):457–465. doi:10.2307/1381092.

Romanenko SA, Perelman PL, Trifonov VA, Graphodatsky AS. 2012. Chromosomal evolution in Rodentia. Heredity. 108(1):4–16. doi:10.1038/hdy.2011.110.

Ruiz-Herrera A, Castresana J, Robinson TJ. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? Genome Biol. 7(12)(R115). doi:10.1186/gb-2006-7-12-r115.

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Reports. 6(1). doi:10.1038/srep28333.

Schwarzacher T, Heslop-Harrison P. 2000. Practical in situ hybridization. Oxford: BIOS Scientific Publ [u.a.].

Seabright M. 1971. A rapid banding technique for human chromosomes. Lancet. 2(7731):971–972.

Shapiro JA, von Sternberg R. 2005. Why repetitive DNA is essential to genome function. Biol Reviews. 80(2):227–250. doi:10.1017/S1464793104006657.

Shorter KR, Crossland JP, Webb D, Szalai G, Felder MR, Vrana PB. 2012. *Peromyscus* as a Mammalian Epigenetic Model. Genet Res International. 2012:1–11. doi:10.1155/2012/179159.

Slamovits CH, Cook JA, Lessa EP, Susana Rossi M. 2001. Recurrent Amplifications and Deletions of Satellite DNA Accompanied Chromosomal Diversification in South American Tuco-tucos (Genus *Ctenomys*, Rodentia: Octodontidae): A Phylogenetic Approach. Mol Biol Evolution. 18(9):1708–1719. doi:10.1093/oxfordjournals.molbev.a003959.

Slamovits CH, Rossi MS. 2002. Satellite DNA: agent of chromosomal evolution in mammals. A review. Mastozoología Neotropical. 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. Chromosome Res. 1-16. doi:10.1007/s10577-019-09605-1.

Stangl FB, Baker RJ. 1984. Evolutionary relationships in *Peromyscus*: Congruence in chromosomal, genic, and classical data sets. J Mammal. 65: 668-673.

Sujiwattanarat P, Thapana W, Srikulnath K, Hirai Y, Hirai H, Koga A. 2015. Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. Sci Reports. 5(1). doi:10.1038/srep10315.

Sumner AT. 1972. A simple technique for demonstrating centromeric heterochromatin. Exp Cell Res. 75(1):304–306. doi:10.1016/0014-4827(72)90558-7.

Tanaka Y. 2001. Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. EMBO J. 20(23):6612–6618. doi:10.1093/emboj/20.23.6612.

Terrenoire E, McRonald F, Halsall JA, Page P, Illingworth RS, Taylor AMR, Davison V, O'Neill LP, Turner BM. 2010. Immunostaining of modified histones defines high-level features of the human metaphase epigenome. Genome Biol. 11(11):R110. doi:10.1186/gb-2010-11-11-r110.

Ungvari Z, Krasnikov BF, Csiszar A, Labinskyy N, Mukhopadhyay P, Pacher P, Cooper AJL, Podlutskaya N, Austad SN, Podlutsky A. 2008. Testing hypotheses of aging in long-lived mice of the genus *Peromyscus*: association between longevity and mitochondrial stress resistance, ROS detoxification pathways, and DNA repair efficiency. AGE. 30(2–3):121–133. doi:10.1007/s11357-008-9059-y.

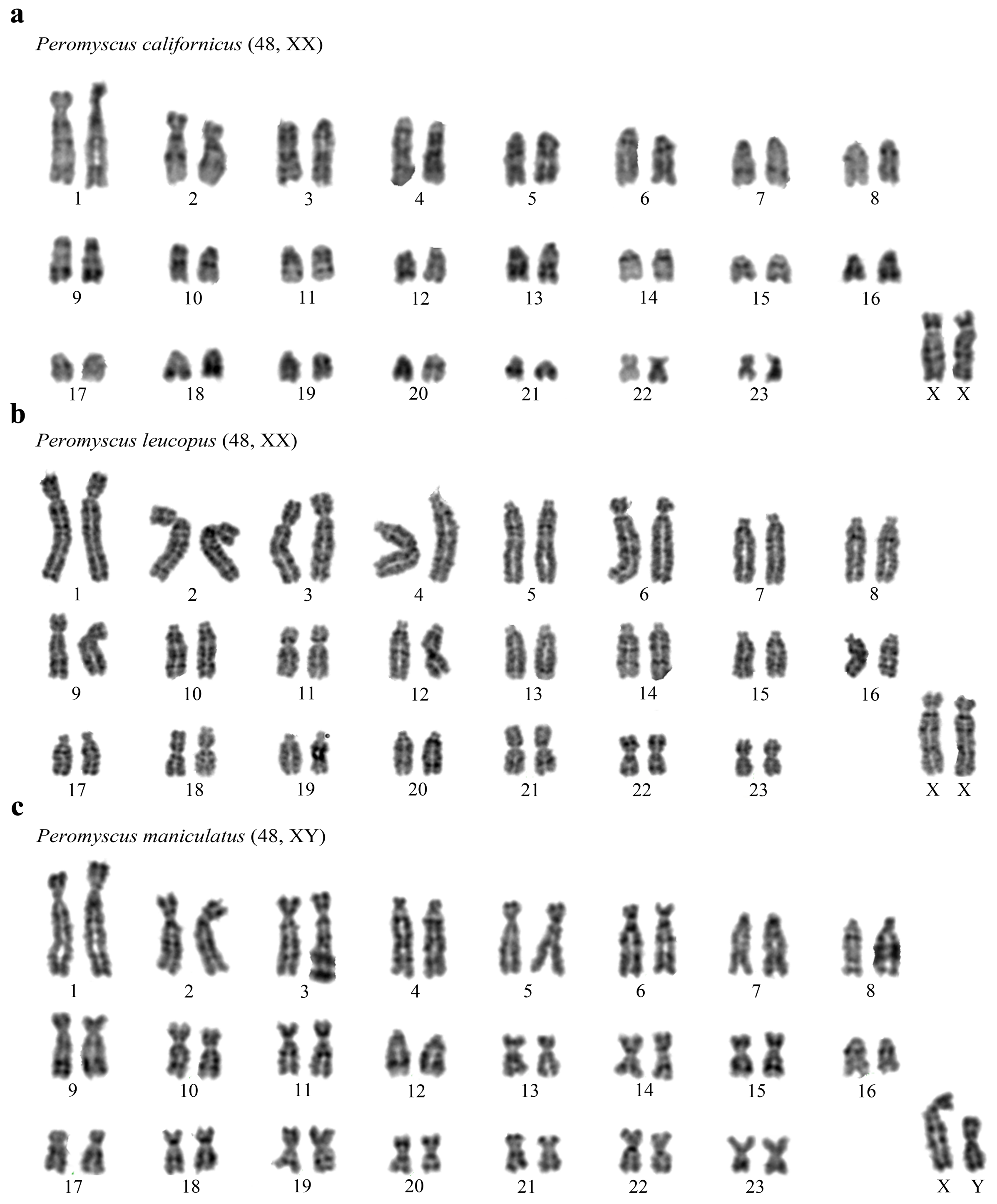
Vittorazzi SE, Lourenço LB, Recco-Pimentel SM. 2014. Long-time evolution and highly dynamic satellite DNA in leptodactylid and hylodid frogs. BMC genetics. 15(1):111.

Wichman HA, Payne CT, Ryder OA, Hamilton MJ, Maltbie M, Baker RJ. 1991. Genomic Distribution of Heterochromatic Sequences in Equids: Implications to Rapid Chromosomal Evolution. Heredity. 82(5):369–377. doi:10.1093/oxfordjournals.jhered.a111106.

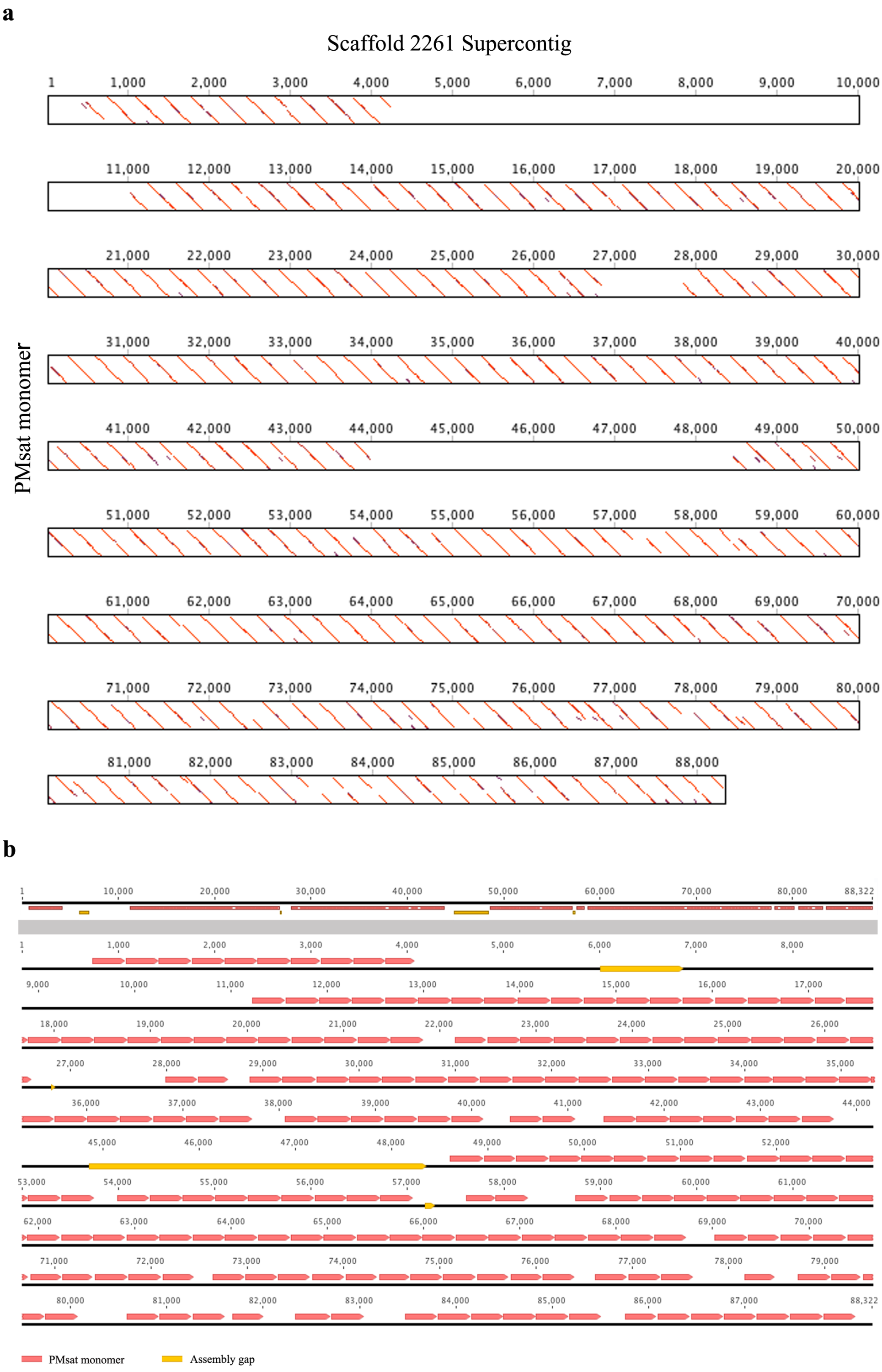
Witmer GW, Moulton RS. 2012. Deer mice (*Peromyscus* spp.) biology, damage, and management: a review. Proceedings of the 25th vertebrate pest conference, University of California, Davis (CA).:213–219.

SUPPLEMENTARY INFORMATION

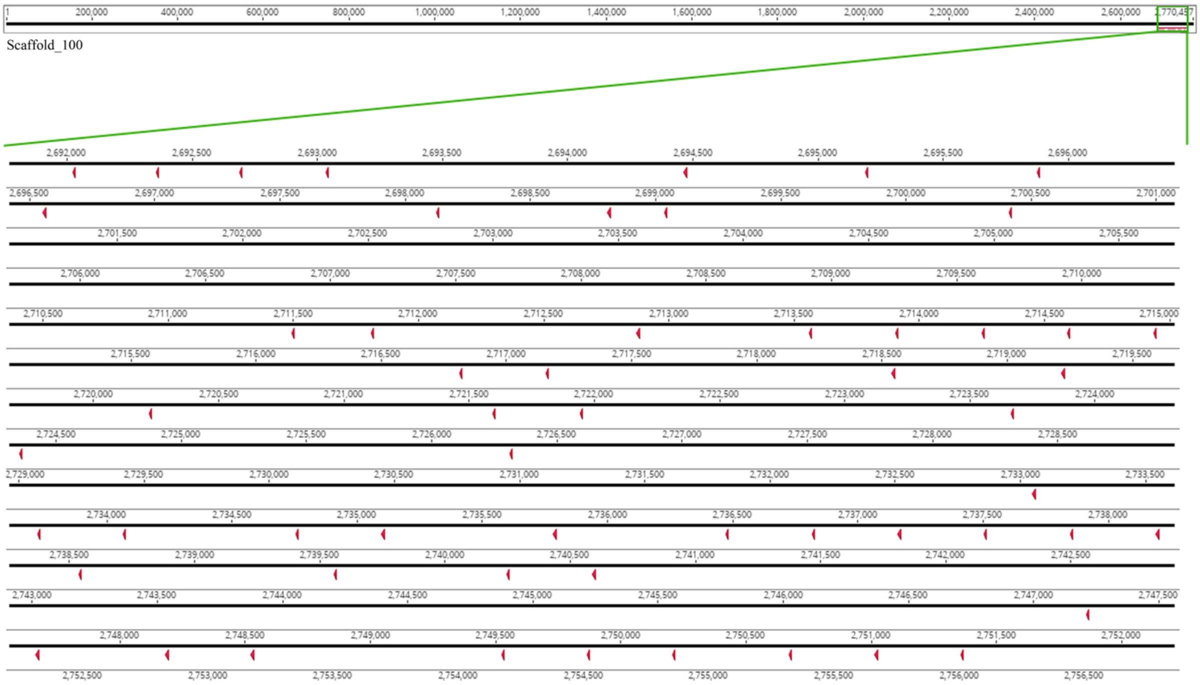
SUPPLEMENTARY FIGURES



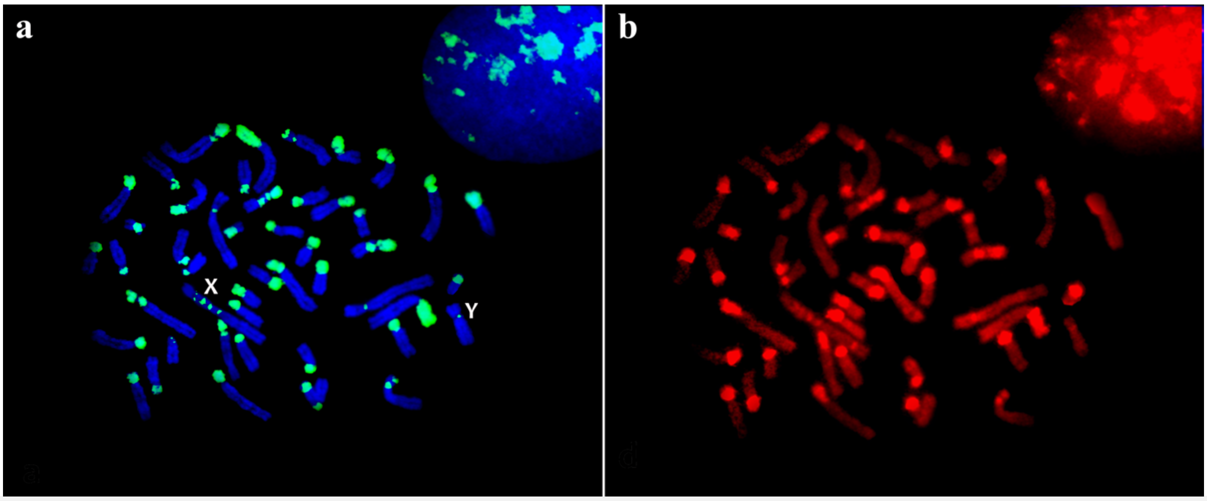
Supplementary Figure II.2.1. Karyotypes of *Peromyscus* species. The representative karyotype of *P. californicus* (a), *P. leucopus* (b) and *P. maniculatus* (c). For each species, 20 metaphases were analysed after of G-banding and after physical mapping of PMSat. The metaphasic chromosomes are the same presented in Figure II.2.4 a. Karyotype were constructed based on chromosomal morphology presented in Committee for Standardization of Chromosomes of *Peromyscus* (1977), and Greenbaum et al. (1994).



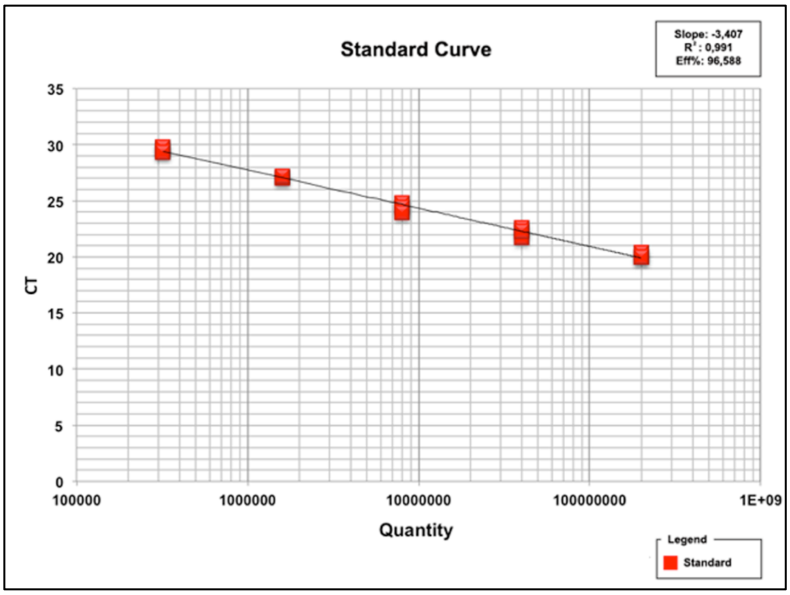
Supplementary Figure II.2.2. PMSat representativeness on Scaffold_2261. (a) Dot-matrix analysis of five contigs on scaffold_2261 (AYHN01158826.1, AYHN01158827.1, AYHN01158828.1, AYHN01158829.1 and AYHN01158830.1) for PMSat repeats. The matrix was performed based on EMBOSS 6.5.7 tool dotmatcher on Geneious R9 version 9.1.2 (Biomatters), with threshold of 50%. (b) Assembly map of PMSat repeats on Scaffold_2261 supercontig. PMSat repeats and assembly gaps were mapped in red and yellow, respectively. Both analyses was carry out on Geneious R9 version 9.1.2. (Biomatters).



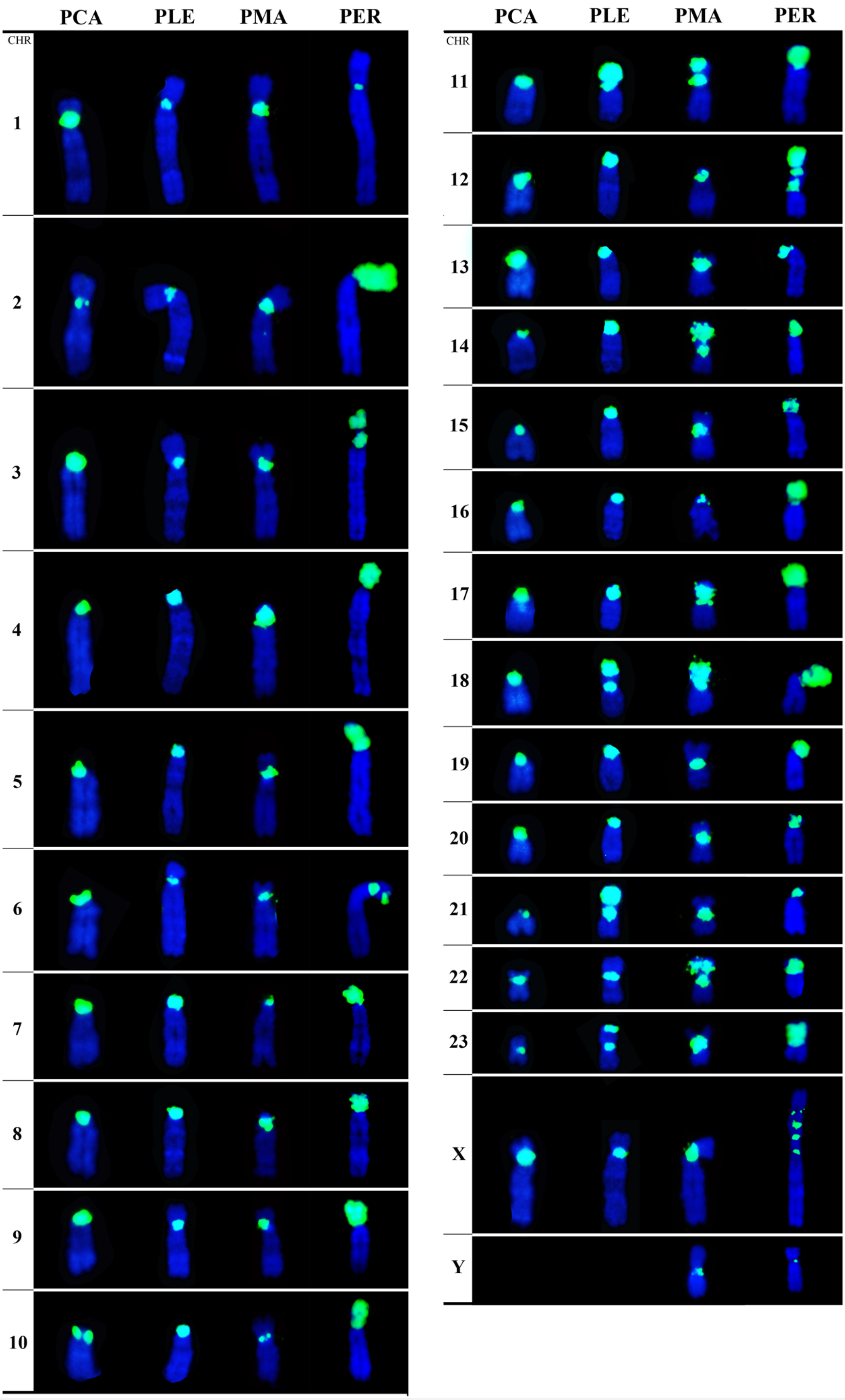
Supplementary Figure II.2.3. CENP-B box motifs mapping on Scaffold_100. The presented scaffold shows a region (highlighted in the figure with corresponding coordinates) with CRS*var* motifs (red arrowheads), which are also found in isolated PMSat clones.



Supplementary Figure II.2.4. Physical mapping of PMSat onto *P. eremicus* chromosomes. (a) Representative metaphase of DNA-FISH presenting the chromosomal localization of PMSat (green signals). Chromosomes were counterstained with DAPI (blue). (b) The same metaphase after sequential C-banding (chromosomes counterstained with propidium iodide, red) revels coincidence of CH at the same PMSat regions (for more detail see in Louzada et al. 2015).



Supplementary Figure II.2.5. Standard curve used in the absolute quantification of PMSat in *Peromyscus* genomes. The parameters follow the acceptable values for the purpose, with R^2 = 0.991, slopes -3.407 and 96.588% of reaction efficiency.



Supplementary Figure II.2.6. Chromosomal mapping of PMSat onto *Peromyscus* chromosomes. The physical location of PMSat (green signals) is shown in terms of comparison for the homologous chromosome in the corresponding *Peromyscus* species: *P. californicus* (PCA, 48,XX), *P. leucopus* (PLE, 48XX), *P. maniculatus* (PMA, 48,XY) and *P. eremicus* (PER, 48,XY). Chromosomes were counterstained with DAPI (blue). The metaphasic chromosomes are the same presented in Figure II.2.6 and Supplementary Figure II.2.4 and the corresponding karyotypes are presented in Supplementary Figure II.2.1.

SUPPLEMENTARY TABLES

Supplementary Table II.2.1. Summary of the analysis in all PMSat isolated clones

Species	% Similarity	Designation	Length (bp)	% GC
P. maniculatus	99.74	PMAp1	387	46.3
	100	PMAp2	388	46.9
	100	PMAp17	405	46.2
P. leucopus	99.73	PLEp2	368	46.7
	99.75	PLEp4	405	45.9
P. californicus	100	PCAp1	391	46.8
	99.74	PCAp2	392	46.9

Similarity percentage refers to the number of identical nucleotide positions compared with PMSat sequence deposited on NCBI (GI: KC351938) isolated from *P. eremicus* (Louzada et al. 2015).

Supplementary Table II.2.2. DNA sequence of PMSat isolated clones from *P. maniculatus*, *P. leucopus* and *P. californicus*.

Designation	Sequence
PMAp1	>TTCTTTTGTTCTGAGCAAGCTCACTGTTCTGGCCCTATAGGAAACACAGTAGAATAGAAGAGTGCTCTTTTCTCAAAAGCAGAGTGTGTTTCTTGTAAGGCGAGCTAGGGTTTGTTTCCCAGTCCTAAACGGAGTTGAATCCCATGCAGTTTCTGGTCCTACGAGCAAGAGTTGTTTTCTGGTAAGAAGAGTCACTCTTGCGCTCCCATTGCCATACACAGTGCAAATAGCACTCGCGTCTGTTCCCGGCAAGTACAGTGTATTGGACTGAAGAGAAGCTACTGTTCTTGTCAGTTTCCTAAGCAGAGTTGAACTAGATATGGCCCGTGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCGGACCCCACATAC
PMAp2	>CGACTCGAGTGGGTTATGTGGCCCATGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCTGGCCCTATAGGAAACACAGTAGAATAGAAGAGTGCTCTTTTCTCAAAAGCAGAGTGTGTTTCTTGTAAGGCGAGCTAGGGTTTGTTTCCCAGTCCTAAACGGAGTTGAATCCCATGCAGTTTCTGGTCCTACGAGCAAGAGTTGTTTTCTGGTAAGAAGAGTCACTCTTGCGCTCCCATTGCCATACACAGTGCAAATAGCACTCGCGTCTGTTCCCAGCAAGTACAGTGTATTGGACTGAAGAGAAGCTACTGTTCTTGTCAGTTTCCTAAGCAGAGTTGAACTAGATATGGCCCGTGTGTGTAGGAAGCACAG
PMAp17	>TGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCTGGCCCTATAGGAAACACAGTAGAATAGAAGAGTGCTCTTTTCTCAAAAGCAGAGTGTGTTTCTTGTAAGGCGAGCTAGGGTTTGTTTCCCAGTCCTAAACGGAGTTGAATCCCATGCAGTTTCTGGTCCTACGAGCAAGAGTTGTTTTCTGGTAAGAAGAGTCACTCTTGCGCTCCCATTGCCATACACAGTGCAAATAGCACTCGCGTCTGTTCCCAGCAAGTACAGTGTATTGGACTGAAGAGAAGCTACTGTTCTTGTCAGTTTCCTAAGCAGAGTTGAACTAGATATGGCCCGTGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCGGACCCCACATAC
PLEp2	>GCCCATGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCTGGCCCTATAGGAAACACAGTAGAATAGAAGAGTGCTCTTTTCTCAAAAGCAGAGTGTGTTTCTTGTAAGGCGAGCTAGGGTTTGTTTCCCAGTCCTAAACGGAGTTGAATCCCATGCAGTTTCTGGTCCTACGAGCAAGAGTTGTTTTCTGGTAAGAAGAGTCACTCTTGCGCTCCCATTGCCATACACAGTGCAAATAGCACTCGCGTCTGTTCCCAGCAAGTACAGTGTATTGGACTGAAGAGAAGCTACTGTTCTTGTCAGTTTCCTAAGCAGAGTTGAACTAGATATGGCCCGTGTGCGTAGGAAGCACAG
PLEp4	>TGTGTGTAGGAAACACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCTGGCCCTATAGGAAACACAGTAGAATAGAAGAGTGCTCTTTTCTCAAAAGCAGAGTGTGTTTCTTGTAAGGCGAGCTAGGGTTTGTTTCCCAGTCCTAAACGGAGTTGAATCCCATGCAGTTTCTGGTCCTACGAGCAAGAGTTGTTTTCTGGTAAGAAGAGTCACTCTTGCGCTCCCATTGCCATACACAGTGCAAATAGCACTCGCGTCTGTTCCCAGCAAGTACAGTGTATTGGACTGAAGAGAAGCTACTGTTCTTGTCAGTTTCCTAAGCAGAGTTGAACTAGATATGGCCCGTGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCGGACCCCACATAC
PCAp1	>CGACTCGAGTGGGTTATGTGGCCCATGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCTGGCCCTATAGGAAACACAGTAGAATAGAAGAGTGCTCTTTTCTCAAAAGCAGAGTGTGTTTCTTGTAAGGCGAGCTAGGGTTTGTTTCCCAGTCCTAAACGGAGTTGAATCCCATGCAGTTTCTGGTCCTACGAGCAAGAGTTGTTTTCTGGTAAGAAGAGTCACTCTTGCGCTCCCATTGCCATACACAGTGCAAATAGCACTCGCGTCTGTTCCCAGCAAGTACAGTGTATTGGACTGAAGAGAAGCTACTGTTCTTGTCAGTTTCCTAAGCAGAGTTGAACTAGATATGGCCCGTGTGTGTAGGAAGCACAGTTC
PCAp2	>CGACTCGAGTGGGTTATGTGGCCCATGTGTGTAGGAAGCACAGTTCTTTTGTTCTGAGCAAGCTCACTGTTCTGGCCCTATAGGAAACACAGTAGAATAGAAGAGTGCTCTTTTCTCAAAAGCAGAGTGTGTTTCTTGTAAGGCGAGCTAGGGTTTGTTTCCCAGTCCTAAACGGAGTTGAATCCCATGCAGTTTCTGGTCCTACGAGCAAGAGTTGTTTTCTGGTAAGAAGAGCCACTCTTGCGCTCCCATTGCCATACACAGTGCAAATAGCACTCGCGTCTGTTCCCAGCAAGTACAGTGTATTGGACTGAAGAGAAGCTACTGTTCTTGTCAGTTTCCTAAGCAGAGTTGAACTAGATATGGCCCGTGTGTGTAGGAAGCACAGTTCT

Supplementary Table II.2.3. Copy number of PMSat in all the species analyzed. The data are presented in comparison to *P. maniculatus* reference genome, resulting in a relative quantification.

Species	Relative quantification
PMA	1.00 (±0.03)
PLE	0.65 (±1.14x10^{-3})
PCA	0.79 (±0.07)
PER	3590 (±343.2)

SUPPLEMENTARY FILES

All the Supplementary Files are available for online view at the link:

https://mega.nz/#F!P8AgDQrJ!n6nD8L4USJa1PDgUNzljYQ

Supplementary File 1

These file (.xml file) comprise 7 distinct sheets with the following information:
1) WGS *Peromyscus maniculatus bairdii* scaffolds. The Scaffold_number used in these work are presented on the first column.
2) Repbase Rodentia Databe Blastn results.
3) Large Tandem Repeats (>50bp period/monomer size; >2000 bp array lenght) after redundancy elimination.
4) PMSat Family
5) RNSAT Family
6) MMSAT4 Family
7) Unclassified TRs

Supplementary File 2

Distance matrix of pairwise alignments of PMSat repeats in five scaffolds of scaffold_2261 (Supplementary Figure II.2.2). The horizontal and vertical axes of each matrix represent consecutive repeats contained in the scaffold sequence. Cells showing nucleotide identities of ≥90, 85–89, 80–84 and 75–79% are red, yellow, green and blue respectively. Also, in each cell the corresponding value was presented (zoom tool for visualization). All the matrixes were generated by Geneious R9 version 9.1.2. (Biomatters) under default settings.

Supplementary File 3

CENP-B box motifs on PMSat scaffolds (.xml file). A total of 875 motifs were analyzed. The sequence and mismatches with the wildtype motif are presented.

CHAPTER III

SATELLITE NON-CODING RNAs: AN EVOLVING TOPIC AROUND CENTROMERE FORMATION, IDENTITY AND FUNCTIONALITY

SOME SECTIONS ON THIS CHAPTER WERE BASED ON THE REVIEW ARTICLE:
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. *Chromosome Research.* 23, 479-493. (doi: 10.1007/s10577-015-9482-8).

The release of the human genome sequence in 2001 unveiled that only about 1-2 % of the genome encodes for proteins, as a huge part of the genome is transcribed as non-coding RNAs (ncRNAs) (Deng and Sui 2013). Over the last years, a growing number of studies focused on ncRNAs point them as crucial elements in fundamental biological processes (Brown et al. 2012; Sana et al. 2012).

Aside with other parts of the genome, satellite DNA (satDNA) was initially considered to be "junk DNA" (Shapiro and von Sternberg 2005; Plohl et al. 2008; Leonova et al. 2013). As referred in Chapter I, satDNA is mainly located at the centromeric region of the chromosomes, besides other heterochromatic regions, and the possibility of its transcription was not considered for several years, as these were considered transcriptionally inert (Plohl et al. 2008). It was in the late 60s and 70s that satDNA transcriptional activity was first reported (Harel et al. 1968; Cohen et al. 1973), however, only in the last years the satellite non-coding RNAs (satncRNAs) started to be characterized and described in different species, including vertebrates, invertebrates and plants (reviewed in Hall et al. 2012; Biscotti et al. 2015; McNulty and Sullivan 2018).

III.1. SATELLITE ncRNAS AND THEIR FUNCTIONAL SIGNIFICANCE

Centromeric transcription has evolved as a fundamental aspect of centromere conserved among eukaryotes. The study of satncRNAs has revealed to be a challenging work, even more because the human centromere, specifically the two main centromeric domains, the core centromere (CT) and the pericentromere (PCT), are both composed by alpha-satellite DNA, being extremely difficult to discriminate the functional role(s) of alpha-satellite transcripts (McNulty and Sullivan 2018). Indeed, transcripts are generated both from the CT and PCT regions; interestingly, processed through different pathways, and exhibiting distinct molecular functions, independently from the DNA sequence. SatncRNAs from CT interact with centromeric proteins and are involved in remodeling/CENP-A deposition and kinetochore assembly (Wong et al. 2007; Ideue et al. 2014; Quénet and Dalal 2014; McNulty et al. 2017). On the contrary, PCT derived transcripts can act as siRNAs to define and maintain PCT heterochromatin in fission yeast, or as long satncRNAs in heterochromatin formation, at least in human and mouse cells (Chan and Wong 2012; Camacho et al. 2017; Johnson et al. 2017).

III.1.1. Pericentromeric transcription

As referred to in chapter I, the heterochromatin is characterized by specific epigenetic marks, and their formation in mammals involves the methylation of histone H3 at lysine 9 (H3K9me) by Suv39h (methyltransferases) and subsequent recruitment of chromodomain proteins such as heterochromatin protein 1 (HP1) (Grewal and Jia 2007). PCT transcripts has been considered as an essential component for heterochromatin formation and maintenance, due to their role in recruiting heterochromatin factors that maintain the heterochromatin modifications, mainly H3K9me2/3 and H3K27me2/3 (Lippman and Martienssen 2004; Chen et al. 2008; Djupedal et al. 2009; Reyes-Turcu et al. 2011). Intriguingly, a paradox was governed at heterochromatic regions: heterochromatin is transcribed to maintain its inactive state.

The mechanism of heterochromatin formation and maintenance was extensively dissected in fission yeast (*Saccharomyces pombe*), in which three distinct mechanisms have been identified (Lippman and Martienssen 2004; Djupedal et al. 2009; Reyes-Turcu et al. 2011). In all the characterized mechanisms, heterochromatin formation involves the transcription of PCT sequences by RNA polymerase II (RNApolII) and subsequent PCT transcripts processing into short interfering RNAs (siRNAs), which can involve the RNA interference (RNAi) pathway (Lippman and Martienssen 2004), an alternate RNAi pathway with secondary stem-loop structures as triggers (Djupedal et al. 2009), or an RNAi-independent mechanism that acts in parallel with the RNAi pathway (Reyes-Turcu et al. 2011).

The heterochromatin establishment at pericentromeres involving similar RNAi machinery has also been identified in other organisms, including plants (e.g. rice, maize and *Arabidopsis*), invertebrates (*Drosophila* and tammar wallaby) and vertebrates (Fukagawa et al. 2004; Lippman and Martienssen 2004; Neumann et al. 2007; Hsieh et al. 2011). Indeed, the involvement of RNAi in heterochromatin formation in vertebrates has been debated in the last years, namely in mouse and human, with some conflicting reports related to transcripts size, cell cycle expression pattern and recognized involvement in heterochromatinization process (reviewed in Chan and Wong 2012). Despite the initial different opinions (e.g. Kanellopoulou et al. 2005 *vs* Murchison et al. 2005), the involvement of Dicer/RNAi analogous pathways has been reported in mouse. The condensation of chromatin might be due to WDHD1 (WD repeat and HMG-box DNA binding protein 1), an acidic nucleoplasmic DNA-binding protein whose activity is coupled to RNApolII transcription, which the association with centromere in mid-to-late S phase plays a role in PCT transcripts processing by a similar pathway to the RNAi pathway Dicer-dependent in yeast (Hsieh et al. 2011). In mouse, in WDHD1 knock-down experiments, the localization of HP1 and epigenetic silencing of (peri)centromeric regions is compromised, leading to an increase in the transcription of both CT and PCT satellites (MiSat and MaSat, respectively and referred in Chapter I) and a decrease in the compaction of centromeric heterochromatin, which in turn result in cell cycle abnormalities due the effects in centromere integrity and subsequently, genomic stability (Hsieh et al. 2011).

The involvement of murine non-siRNA-sized PCT transcripts in establishing heterochromatin has been also reported in several works. In mitotic somatic mouse cells, Lu and Gilbert (2007) reported the presence of both small (~200 nt) and long PCT transcripts (MaSat transcripts with 1 kb to more than 8 kb length) through the cell cycle. The transcription of murine PCT transcripts is cell cycle regulated and long PCT transcripts are present at the G1 phase and are closely located around the chromocenters (nuclear structures formed by the aggregation of heterochromatin from multiple chromosomes), with an increased level in G1/S transition and decrease before the replication of PCT heterochromatin (Lu and Gilbert 2007). Moreover, the accumulation of small PCT transcripts at pericentric regions of condensed chromosomes at the G2/M phase has been observed and reinforces their role in the remodeling and/or maintenance of the pericentric heterochromatin structure during cell division (Lu and Gilbert 2007; Bulut-Karslioglu et al. 2012). Maison et al. (2011) demonstrated that mouse long single-stranded (ss) PCT transcripts associate with HP1 and this complex is guided to the pericentric heterochromatin domain to lead further HP1 localization. A more recent work by Camacho et al. (2017) propose an RNA-mediated process to govern the stable association of the Suv39h enzymes at mouse heterochromatin (Figure III.1), in which MaSat transcripts remain associated with the chromatin and form RNA:DNA hybrids and induce the formation of a higher-order RNA-nucleosome scaffold that would represent the underlying structure of mouse heterochromatin (cf. in Chapter I the MaSat organization into HORs). Also, in humans, alpha satellite ssPCT transcripts in association with chromatin contribute to the localization of SUV39H1 at constitutive heterochromatin (CH) (Johnson et al. 2017). Although, evidence of HP1 localization by alpha satellite RNA binding has not yet been reported in humans.

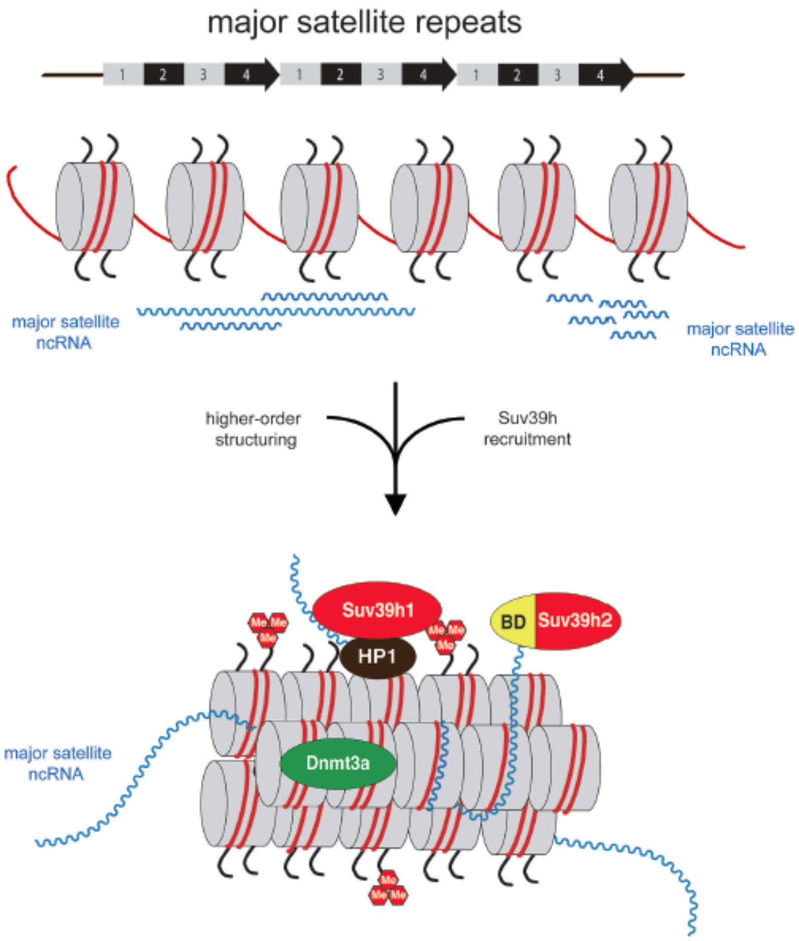


Figure III.1. Model for a higher-order RNA-nucleosome scaffold established by the chromatin association of mouse PCT repeats. In this model, the initial transcriptional activity of the PCT region is needed to build heterochromatin. The intrinsic property of satellite mouse repeats to form RNA:DNA hybrids will facilitate their chromatin retention and most likely occurs in inter-nucleosomal regions. Additional ssPCT transcripts organize the assembly of a higher-order RNA-nucleosome structure and recruit and stabilize Suv39h enzymes to heterochromatin. At the heterochromatin, ssPCT transcripts also provide additional binding affinities such as the basic domain (BD) of Suv39h2 (Camacho et al. 2017), H3K9me3 (Wang et al. 2012) and RNA binding by the chromodomains of both mouse Suv39h1 (Shirai et al. 2017) or human SUV39H1 (Johnson et al. 2017) enzymes and HP1 interaction (Maison et al. 2011). Adapted from Camacho et al. (2017).

Some studies have been clarifying the binding partners for PCT sequences that, ultimately, are involved in heterochromatinization process, which are characterized by a fine regulation at the transcriptional level. As a redundant outcome, the transcription factors Pax3 and Pax9 repress RNA output from major satellite sequences by associating with DNA within PCT heterochromatin (Bulut-Karslioglu et al. 2012). Nevertheless, other transcription factors can be involved in the regulation of PCT transcription, since potential binding sites reside on PCT satellites (e.g. YY1 factor, Shestakova et al. 2004).

Transcription of murine PCT sequences may also be required to regulate growth and development. In early mouse development, the transcription of MaSat is required to establish heterochromatin formation in chromocenters (Probst et al. 2010; Casanova et al. 2013;

Burton and Torres-Padilla 2014) and their disruption causes arrest of the cell cycle, suggesting a PCT transcripts role in the correct progression of the early embryogenesis by heterochromatin establishment (Probst et al. 2010; Burton and Torres-Padilla 2014). Moreover, a strand-specific transcription of MaSat occurs, sense and anti-sense strand of PCT transcripts are differentially expressed throughout the developmental progression, in terms of their expression levels and location within the cell (Probst et al. 2010). In addition to mitotic cells, the roles of PCT sequences in mouse post-mitotic cells were also investigated during neuronal differentiation (Solovei et al. 2004; Kishi et al. 2012). Kishi and colleagues (2012) demonstrate that murine MaSat transcription is significantly increased during neuronal differentiation both *in vitro* and *in vivo* assays. Also, MaSat DNA sequences suffer an increasing of H3K4me3, suggesting both structural and transcriptional roles of MaSat regions in neuronal differentiation (Kishi et al. 2012).

III.1.2. Centromeric transcription

The transcription from de core centromeric domain has been proved more enigmatic than PCT transcription, and much less is known of the molecular mechanisms involving CT transcription. Actually, the overall level of CT sequences transcription was lower than PCT transcription (Ohkuni and Kitagawa 2011), being in some cases, almost undetectable due CT satncRNA rapid turnover (Choi et al. 2011; Ohkuni and Kitagawa 2011; Chan et al. 2012). Notwithstanding, both the act of RNApolII transcription of the centromeric chromatin and the derived nascent satncRNAs are essential for both kinetochore assembly and CT chromatin remodeling/CENP-A deposition and evidenced an extremely regulated process (Perea-Resa and Blower 2018). The initial studies reported the co-localization of human CT transcripts in the nucleolus until their re-localization to the centromere at the onset of mitosis via CENP-C (Wong et al. 2007). However, recent studies reveal that long CT satncRNAs are localized at the centromeres in both interphase and metaphase and co-localized with essential centromeric proteins (Chan et al. 2012; Ideue et al. 2014; Quénet and Dalal 2014; McNulty et al. 2017).

Little is known about the mechanism of RNApolII acting on centromere and the transcription factors and binding domains involved in their recruitment in most organisms are still enigmatic (Talbert and Henikoff 2018; Perea-Resa and Blower 2018). Some molecular players have been revealed in budding yeast (Smurova and De Wulf 2018), but, according to our knowledge, only the general CTDP1 factor (RNA pol II subunit A C-terminal domain phosphatase) has been identified at the human centromere (Chan et al. 2012). Also, additional chromatin remodeling factors and RNApolII associated proteins has been identified at the human centromere, like FACT (facilitates chromatin transcription complex) (Formosa 2012). Indeed, the repetitive nature of CT region difficult the identification of promoter-like regions. Studies in maize (Topp et al. 2004) and tammar wallaby (Carone et al. 2009; O'Neill and Carone 2009) have been suggested that the promoters within transposal elements (e.g. retroviral elements), which also reside on CT region, lead the transcription of neighbouring satDNAs (O'Neill and Carone 2009; Carone et al. 2013). Recently, Kasinathan and Henikoff (2018) hypothesized that non-B-form DNA structures (e.g. cruciforms) on centromeric satellites may facilitate their transcription.

While the specific mechanisms of CT transcription remain elusive, several studies reinforce the link between the CT transcription and nucleosome assembly (Carone et al. 2013; Rošić et al. 2014; Chen et al. 2015; Bobkov et al. 2018). The topological effect of RNApolII transcription is necessary for the stable and specific loading of CENP-A into the CT core domain chromatin (Chen et al. 2015; Molina et al. 2016; Perea-Resa and Blower 2018). Ohzeki et al. (2012) suggested that transcription could promote H3 acetylation at the core domain, in which during mitosis the RNApolII could recruit HAT (Histone acetyltransferase) complexes and generate an acetylated environment that could be favorable for CENP-A loading (Ohzeki et al. 2012). Interestingly, CT transcription seems also to be required for the correct re-localization of Sgo1 from the outer kinetochore to the inner centromere for the correct centromeric cohesion (Liu et al. 2015).

Recent works have been reported that human CT transcripts can be functional in *cis* or in *trans* (Blower 2016; McNulty et al. 2017; Quénet et al. 2017; Kabeche et al. 2018), where array-specific CT satncRNAs act only near the site of transcription as nascent transcripts or can be influence the function of all/others centromere regions. CT transcripts have been identified in pre-assembly histone complexes (i.e. not yet incorporated into DNA to form chromatin) containing CENP-A and the histone chaperone HJURP (Holliday Junction Recognition Protein) prior to association with centromeric chromatin (Quénet and Dalal 2014). In mammalian cells, the CENP-A loading occurs in late telophase/early G1 (Jansen et al. 2007; Dunleavy et al. 2011), and prior to assembly into chromatin long CT transcripts are complexed with chromatin-bound centromere proteins, such as CENP-A and CENP-B (McNulty et al. 2017). In human centromeres, RNApolII is actively elongating during mitosis, before CENP-A deposition (Chan et al. 2012; Liu 2016) and in mouse, MiSat transcripts present a moderate peak in G2/M are barely detectable at G1 (Ferri et al. 2009). Indeed, high activity of RNApolII was verified until kinetochores have achieved stable microtubule attachment (Liu et al. 2015). As a result, the mitotic CT transcripts seem to associate with other centromeric proteins during mitosis.

In addition to CENP-A and HJURP, an increasing number of studies has reported that CT transcripts physically associate with other centromeric proteins, including CENP-B and CENP-C (Wong et al. 2007; Carone et al. 2009; Du et al. 2010; Quénet and Dalal 2014; McNulty et al. 2017), as well as with components of chromosome passenger complex (CPC), including, INCENP, Survivin and Aurora-B (Bouzinba-Segard et al. 2006; Ferri et al. 2009; Ideue et al. 2014). Specifically, Aurora-B is recruited in early mitosis and regulates essential events in chromosome dynamics, as pericentromeric cohesion, chromosome alignment and kinetochore-microtubule attachment (Krenn and Musacchio 2015). Indeed, in mouse and human (Bouzinba-Segard et al. 2006; Ferri et al. 2009; Ideue et al. 2014), in which both centromere transcription and CT transcripts seems to regulate the normal activation and localization of Aurora-B (Ideue et al. 2014). Recently, Kabeche and colleagues (2018) proposed that the presence of R-loops (derived from the transcriptional activity of RNAPolII) in centromere during mitosis is required for normal Aurora-B activation. The inhibition of centromeric transcription or depletion of satncRNAs causes mistargeting of centromeric proteins, namely CENP-A (Quénet and Dalal 2014; Rošić et al. 2014; McNulty et al. 2017), CENP-B (Carone et al. 2009; McNulty et al. 2017) and CENP-C (Wong et al. 2007; Du et al. 2010; Chan et al. 2012), and also displacement of CPC components, leading to abnormal cell shape and error in mitosis (Ideue et al. 2014). Also, their overexpression compromise CPC integrity by mislocalization of Aurora-B (Bouzinba-Segard et al. 2006). Altogether, the recent evidences of centromeric transcription suggest that CT transcripts could act as a scaffold at the mitotic kinetochore to recruit and organize centromeric proteins (Figure III.2).

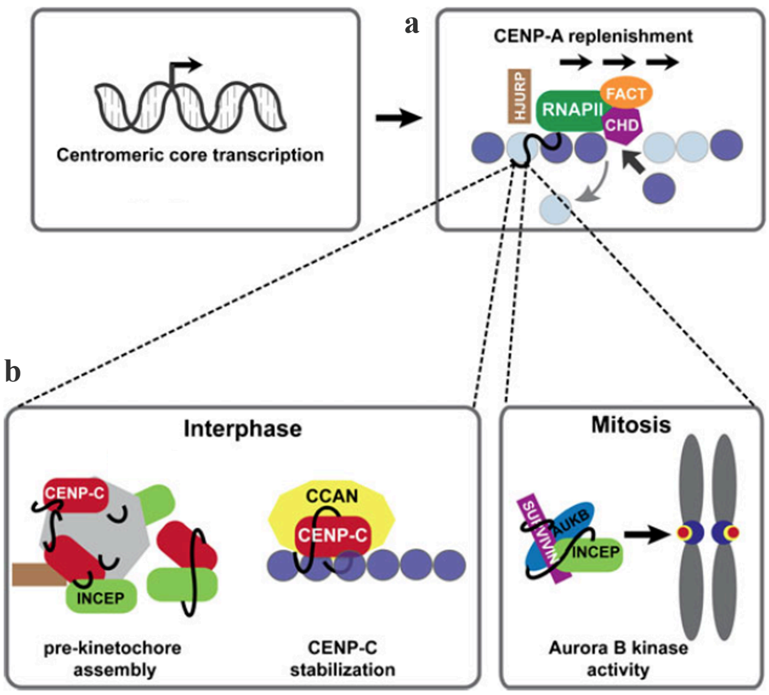


Figure III.2. The dual effect of centromere transcription. a) RNApolII is recruited to the centromeric core (through an unknown mechanism) and the topological effect of transcription results in chromatin remodeling by CENP-A loading. The histone chaperone and chromatin remodeler, HJURP and FACT complex, are important component for remodeling process. (Dark- and light-blue circles represent CENP-A or H3.1 nucleosomes, respectively). b) Centromere satncRNA is an integral component of the kinetochore. During interphase, core-derived satncRNA localized to the nucleolus (left) allows the complex formation or the assembly of pre-kinetochore structures or at the centromere core domain associates with CENP-A and CCAN (Constitutive Centromere Associated Network) complex, namely CENP-C, and stabilize its DNA-binding ability. During mitosis, satncRNAs associates with CPC proteins (INCENP and Survivin) and mediates the kinase activity of other CPC protein, Aurora-B. Adapted from Scott (2013).

III.1.3. Centromeric transcription during cellular stress, disease and cancer

It has been made clear that satncRNAs derived from CT and PCT regions plays important roles in the cell. As a result, it has been expected that their transcription should be maintained through a fine balance, in which some disturbance potentiate cellular disadvantageous consequences (reviewed in Hall et al. 2012). The most studied cellular context that triggers overexpression of satncRNAs is the response to stress, which can be induced by different factors such as high temperature (heat shock), heavy metals, hazardous chemicals, ultraviolet radiation, and hyperosmotic or oxidative conditions (Jolly et al. 2004; Valgardsdottir et al. 2008; Eymery et al. 2010; Goenka et al. 2016).

Besides alpha-satellite, two additional satDNA families are found in some human chromosomes at the PCT region, Satellite II (SATII) and Satellite III (SATIII) (Tagarro et al. 1994), which transcription was verified in specific cell stress conditions such as heat shock. In fact, the association between the heat shock factor 1 (HSF1) transcription factor and

SATIII leads to their transcription by RNApolII, being important for the nuclear stress bodies' (nSBs) assembly (Jolly et al. 2004; Valgardsdottir et al. 2008; Eymery et al. 2010; Goenka et al. 2016). These transcripts are involved in the recruitment of specific splicing factors (e.g. SRSF1, serine/arginine-rich splicing factor 1; also known as SF2) and transcription factors (e.g CREB-binding protein) to nSBs (Goenka et al. 2016) and modulate the expression of stress specific genes (Valgardsdottir et al. 2008; Eymery et al. 2010; Goenka et al. 2016). The transcription of specific PCT satncRNAs seem to contribute to the transitory cell nucleus organization (Eymery et al. 2009a; Eymery et al. 2010), acting as a protective effect against the heat-shock-induced cell death (Enukashvily and Ponomartsev 2013; Goenka et al. 2016). The transcription of SATIII can also be induced by other cellular stresses, regulated by other transcription factors, like the tonicity enhancer-binding protein (TonEBP) during hyperosmotic stress (Valgardsdottir et al. 2008). In contrast, the overexpression of CT satncRNAs (MiSat in mouse) during chemical exposure leads to chromosome abnormalities (Bouzinba-Segard et al. 2006). Indeed, under stress conditions, the transcription of PCTs was globally upregulated and CTs were not, which reinforces that PCT and CT transcripts are under different transcriptional controls during stress (Eymery et al. 2009b).

Interestingly, the accumulation of CT MiSat transcripts was also verified on the induction of apoptosis (Bouzinba-Segard et al. 2006), despite their specific role in the context has not yet been elucidated. In senescent cells, the accumulation of PCT satncRNAs was also been reported (Enukashvily et al. 2007; De Cecco et al. 2013), and their transcripts might be related with heterochromatin structure maintenance (Enukashvily et al. 2007; Eymery et al. 2009a). In the last years, an increasing number of studies associate satncRNAs with genomic instability and tumorigeneses (Bouzinba-Segard et al. 2006; Valgardsdottir et al. 2008; Eymery et al. 2009b; Ting et al. 2011; Deng and Sui 2013; Bersani et al. 2015; Zhu et al. 2018). Due to their functional role on the centromere/kinetochore assembly, satncRNAs may influence the oncogenic process by its dysfunction in mitosis contributing to abnormal chromosome segregation – one of the hallmarks of cancer (Frescas et al. 2008). Indeed, the abnormal expression of satncRNAs might be involved in the cancer genome instability that contributes to cancer cell phenotype (Brown et al. 2012). However, whether if the molecular mechanisms leading to overexpression of satncRNAs in cancer or their expression pattern is a cause or a consequence of genomic instability are yet unclear (Burgess 2011; Plohl et al. 2014). Nevertheless, the involvement of epigenetic mechanisms has been attributed (Ting et al. 2011; Ferreira et al. 2015). Also, satncRNAs have been associated with tumor suppressor intervenients/pathways, like BRCA1 (Kononenko et al. 2014; Zhu et al. 2018), KDM2A (Frescas et al. 2008) and p53 (Leonova et al. 2013).

Along the previous sections, the transcription of satncRNAs in several and distinct biological contexts supports the awareness of a fine-tuning regulatory system of satncRNAs expression. The transcriptional profile of satDNAs can be specific of cell stages: development (Burton and Torres-Padilla 2014), differentiation (Kishi et al. 2012), cell cycle (Lu and Gilbert 2007; Bulut-Karslioglu et al. 2012), cell stresses (Eymery et al. 2010), apoptosis (Bouzinba-Segard et al. 2006), transformation (Eymery et al. 2009a; Eymery et al. 2009b); cell types: stem cells (Probst et al. 2010; Burton and Torres-Padilla 2014), proliferative cells (Lu and Gilbert 2007; Bulut-Karslioglu et al. 2012), senescent cells (Enukashvily et al. 2007), cancer cells (Ting et al. 2011); and tissues (Ugarković 2005; Enukashvily and Ponomartsev 2013). A specific satDNA transcription in a specific biological context may reflect a conserved function of the satncRNAs or of the transcription process itself (Saksouk et al. 2015). Figure III.3 summarizes the cellular processes/functions where satncRNAs seem to be involved in normal, stressed, and cancer cells.

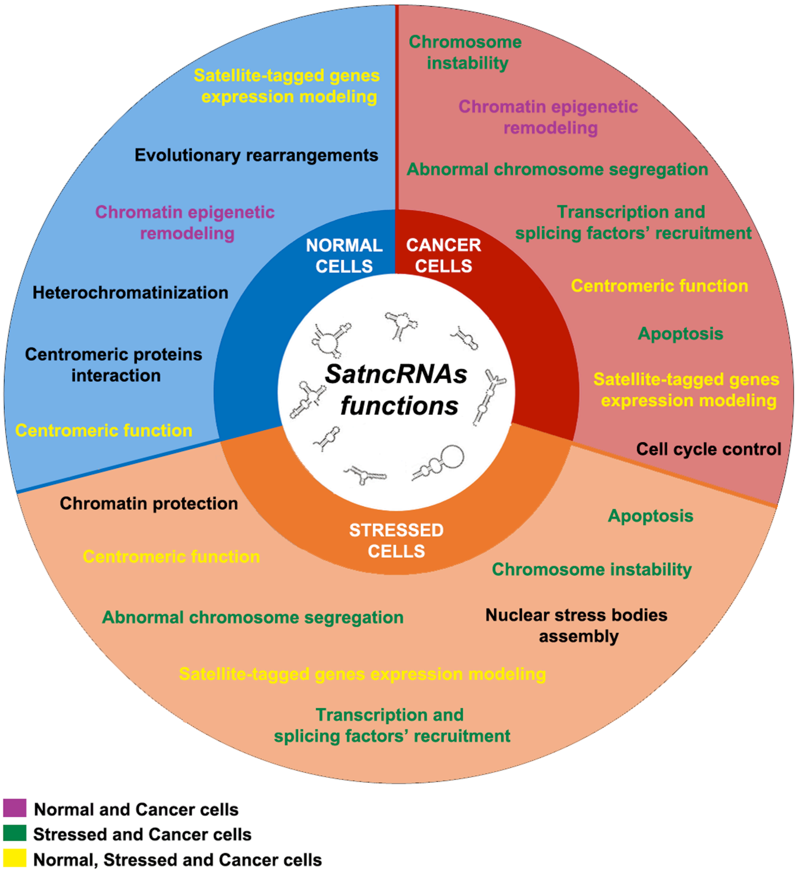

Satellite-tagged genes expression modeling
Evolutionary rearrangements
Chromatin epigenetic remodeling
Heterochromatinization
Centromeric proteins interaction
Centromeric function
Chromosome instability
Chromatin epigenetic remodeling
Abnormal chromosome segregation
Transcription and splicing factors' recruitment
Centromeric function
Apoptosis
Satellite-tagged genes expression modeling
Cell cycle control
NORMAL CELLS
CANCER CELLS
SatncRNAs functions
STRESSED CELLS
Chromatin protection
Centromeric function
Apoptosis
Chromosome instability
Abnormal chromosome segregation
Nuclear stress bodies assembly
Satellite-tagged genes expression modeling
Transcription and splicing factors' recruitment
🟪 Normal and Cancer cells
🟩 Stressed and Cancer cells
🟨 Normal, Stressed and Cancer cells

Figure III.3. Functions attributed to satncRNAs in normal, stressed, and cancer cells. The functions common to all the cells are in yellow, and the functions shared by normal/cancer cells and stressed/cancer cells are in purple and green, respectively. Adapted from Ferreira et al. 2015.

III.1.2. THE CHALLENGES OF SATNCRNAS ANALYSIS

One of the main goals for the scientific community in unveiling the functional role of satDNA transcripts reside on the strategies, methodologies and techniques used. Indeed, for many years, the molecular genetics and genomic approaches are traditionally focused on the study of coding genes, mRNAs and proteins and not repetitive sequences (Ferreira et al. 2015).

The knowledge about satDNA transcripts and their DNA sequences has been delayed by several circumstances, namely a) the restricted number of mammalian studied species (mainly human and mouse) with distinct cell types and cellular conditions among the different works that difficult a comparative analysis; b) the sequences' complexity, the high level of polymorphisms and the repetitive nature of satDNA sequences in the genomes, which is also a major issue in the design of the experimental work as well as in the application of the techniques available that are mainly directed for coding sequences; and c) these are exacerbated by the fact that sequence databases (DNA and RNA) mask the repetitive sequences content that hampers an *in silico* proper analysis (Ferreira et al. 2015).

SatDNA sequences evolve extremely fast in respect to sequence and/or copy number. Indeed, the molecular dynamics and the repetitive nature of satDNAs raises the challenges in their analysis (Eymery et al. 2009a), which requires the improvement and/or adjustments in the available methodologies and techniques. These strategies need to be performed using different approaches that extend from the identification and quantification of satncRNAs in a collection of cells approach to their characterization by single-cell analysis, complemented with other methods and techniques allowing the disclosure of their function(s) and cellular pathway(s) interveners (for approach examples see Ferreira et al. 2015). As mentioned in Chapter I, satDNA sequences are constantly masked in the genome assemblies and, as a consequence, their transcripts are also underrepresented in the ncRNA-specific databases. Despite some currently ncRNA available databases, namely lncRNAdb (Amaral et al. 2011; Quek et al. 2015), NONCODE (Xie et al. 2014) and LncRNADisease (Chen et al. 2012), the information about satellite transcripts is only available in the NONCODE database.

III.1.3. REFERENCES

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. 2011. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 39(suppl_1):D146–D151. doi:10.1093/nar/gkq1138.

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. 112(49):15148–15153. doi:10.1073/pnas.1518008112.

Blower MD. 2016. Centromeric Transcription Regulates Aurora-B Localization and Activation. Cell Rep. 15(8):1624–1633. doi:10.1016/j.celrep.2016.04.054.

Bobkov GOM, Gilbert N, Heun P. 2018. Centromere transcription allows CENP-A to transit from chromatin association to stable incorporation. Cell Biol. 217(6):1957–1972. doi:10.1083/jcb.201611087.

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(23):8709–14.

Brown JD, Mitchell SE, O'Neill RJ. 2012. Making a long story short: noncoding RNAs and chromosome change. Heredity. 108(1):42–49. doi:10.1038/hdy.2011.104.

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, et al. 2012. A transcription factor-based mechanism for mouse heterochromatin formation. Nat Struct Mol Biol. 19(10):1023–1030. doi:10.1038/nsmb.2382.

Burgess DJ. 2011. Chromosome instability: Tumorigenesis via satellite link. Nat Rev Cancer. 11(3):158. doi:10.1038/nrc3031.

Burton A, Torres-Padilla M-E. 2014. Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. Nat Rev Mol Cell Biol. 15(11):723–734. doi:10.1038/nrm3885.

Camacho OV, Galan C, Swist-Rosowska K, Ching R, Gamalinda M, Karabiber F, De La Rosa-Velazquez I, Engist B, Koschorz B, Shukeir N. 2017. Major satellite repeat RNA stabilize heterochromatin retention of Suv39h enzymes by RNA-nucleosome association and RNA: DNA hybrid formation. Elife. 6:e25293.

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(1):113–125. doi:10.1007/s00412-008-0181-5.

Carone DM, Zhang C, Hall LE, Obergfell C, Carone BR, O'Neill MJ, O'Neill RJ. 2013. Hypermorphic expression of centromeric retroelement-encoded small RNAs impairs CENP-A loading. Chromosome Res. 21(1):49–62. doi:10.1007/s10577-013-9337-0.

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(6):1156–1167. doi:10.1016/j.celrep.2013.08.015.

Chan FL, Marshall OJ, Saffery R, Won Kim B, 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. 109(6):1979–1984. doi:10.1073/pnas.1108705109.

Chan FL, Wong LH. 2012. Transcription in the maintenance of centromere chromatin identity. Nucleic Acids Res. 40(22):11178–11188. doi:10.1093/nar/gks921.

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(1):73–84. doi:10.1016/j.devcel.2015.05.012.

Chen ES, Zhang K, Nicolas E, Cam HP, Zofall M, Grewal SIS. 2008. Cell cycle control of centromeric repeat transcription and heterochromatin assembly. Nature. 451(7179):734–737. doi:10.1038/nature06561.

Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. 2012. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 41(D1):D983–D986. doi:10.1093/nar/gks1099.

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. Biol Chemistry. 286(26):23600–23607. doi:10.1074/jbc.M111.228510.

Cohen AK, Huh TY, Helleiner CW. 1973. Transcription of satellite DNA in mouse L-cells. Biochem. 51(5):529–532.

De Cecco M, Criscione SW, Peckham EJ, Hillenmeyer S, Hamm EA, Manivannan J, Peterson AL, Kreiling JA, Neretti N, Sedivy JM. 2013. Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. Aging Cell. 12(2):247–256. doi:10.1111/acel.12047.

Deng G, Sui G. 2013. Noncoding RNA in oncogenesis: a new era of identifying key players. Int J Mol Sci. 14(9):18319–18349. doi:10.3390/ijms140918319.

Djupedal I, Kos-Braun IC, Mosher RA, Söderholm N, Simmer F, Hardcastle TJ, Fender A, Heidrich N, Kagansky A, Bayne E, et al. 2009. Analysis of small RNA in fission yeast; centromeric siRNAs are potentially generated through a structured RNA. EMBO. 28(24):3832–3844. doi:10.1038/emboj.2009.351.

Du Y, Topp CN, Dawe RK. 2010. DNA binding of centromere protein C (CENPC) is stabilized by single-stranded RNA. PLoS Genet. 6(2):e1000835. doi:10.1371/journal.pgen.1000835.

Dunleavy EM, Almouzni G, Karpen GH. 2011. H3.3 is deposited at centromeres in S phase as a placeholder for newly assembled CENP-A in G $_1$ phase. Nucleus. 2(2):146–157. doi:10.4161/nucl.2.2.15211.

Enukashvily NI, Donev R, Waisertreiger IS-R, Podgornaya OI. 2007. Human chromosome 1 satellite 3 DNA is decondensed, demethylated and transcribed in senescent cells and in A431 epithelial carcinoma cells. Cytogenet Genome Res. 118(1):42–54. doi:10.1159/000106440.

Enukashvily NI, Ponomartsev NV. 2013. Mammalian Satellite DNA. In: Advances in Protein Chemistry and Structural Biology. Vol. 90. Elsevier. p. 31–65.

Eymery A, Callanan M, Vourc'h C. 2009a. The secret message of heterochromatin: new insights into the mechanisms and function of centromeric and pericentric repeat sequence transcription. Develop Biol. 53(2–3):259–268. doi:10.1387/ijdb.082673ae.

Eymery A, Horard B, Atifi-Borel ME, Fourel G, Berger F, Vitte A-L, Van den Broeck A, Brambilla E, Fournier A, Callanan M, et al. 2009b. A transcriptomic analysis of human centromeric and pericentric sequences in normal and tumor cells. Nucleic Acids Res. 37(19):6340–6354. doi:10.1093/nar/gkp639.

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(11):1845–1855. doi:10.1016/j.yexcr.2010.02.002.

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. Chromosome Res. 23(3):479–493. doi:10.1007/s10577-015-9482-8.

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(15):5071–5080. doi:10.1093/nar/gkp529.

Formosa T. 2012. The role of FACT in making and breaking nucleosomes. Biochim Biophys Acta. 1819(3–4):247–255. doi:10.1016/j.bbagrm.2011.07.009.

Frescas D, Guardavaccaro D, Kuchay SM, Kato H, Poleshko A, Basrur V, Elenitoba-Johnson KS, Katz RA, Pagano M. 2008. KDM2A represses transcription of centromeric satellite repeats and maintains the heterochromatic state. Cell Cycle. 7(22):3539–3547. doi:10.4161/cc.7.22.7062.

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(8):784–791. doi:10.1038/ncb1155.

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. Cell Sci. 129(19):3541–3552. doi:10.1242/jcs.189803.

Grewal SIS, Jia S. 2007. Heterochromatin revisited. Nat Rev Genet. 8(1):35–46. doi:10.1038/nrg2008.

Hall LE, Mitchell SE, O'Neill RJ. 2012. Pericentric and centromeric transcription: a perfect balance required. Chromosome Res. 20(5):535–546. doi:10.1007/s10577-012-9297-9.

Harel J, Hanania N, Tapiero H, Harel L. 1968. RNA replication by nuclear satellite DNA in different mouse cells. Biochem Biophys Res Commun. 33(4):696–701.

Hsieh C-L, Lin C-L, Liu H, Chang Y-J, Shih C-J, Zhong CZ, Lee S-C, Tan BC-M. 2011. WDHD1 modulates the post-transcriptional step of the centromeric silencing pathway. Nucleic Acids Res. 39(10):4048–4062. doi:10.1093/nar/gkq1338.

Ideue T, Cho Y, Nishimura K, Tani T. 2014. Involvement of satellite I noncoding RNA in regulation of chromosome segregation. Genes to Cells. 19(6):528–538. doi:10.1111/gtc.12149.

Jansen LET, Black BE, Foltz DR, Cleveland DW. 2007. Propagation of centromeric chromatin requires exit from mitosis. Cell Biol. 176(6):795–805. doi:10.1083/jcb.200701066.

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. doi:10.7554/eLife.25299.

Jolly C, Metz A, Govin J, Vigneron M, Turner BM, Khochbin S, Vourc'h C. 2004. Stress-induced transcription of satellite III repeats. Cell Biol. 164(1):25–33. doi:10.1083/jcb.200306104.

Kabeche L, Nguyen HD, Buisson R, Zou L. 2018. A mitosis-specific and R loop-driven ATR pathway promotes faithful chromosome segregation. Science. 359(6371):108–114. doi:10.1126/science.aan6490.

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(4):489–501. doi:10.1101/gad.1248505.

Kasinathan S, Henikoff S. 2018. Non-B-Form DNA Is Enriched at Centromeres. Mol Biol Evol. 35(4):949–962. doi:10.1093/molbev/msy010.

Kishi Y, Kondo S, Gotoh Y. 2012. Transcriptional Activation of Mouse Major Satellite Regions during Neuronal Differentiation. Cell Struct Function. 37(2):101–110. doi:10.1247/csf.12009.

Kononenko AV, Bansal R, Lee NCO, Grimes BR, Masumoto H, Earnshaw WC, Larionov V, Kouprina N. 2014. A portable BRCA1-HAC (human artificial chromosome) module for analysis of BRCA1 tumor suppressor function. Nucleic Acids Res. 42(21). doi:10.1093/nar/gku870.

Krenn V, Musacchio A. 2015. The Aurora B Kinase in Chromosome Bi-Orientation and Spindle Checkpoint Signaling. Front Oncology. 5. doi:10.3389/fonc.2015.00225.

Leonova KI, Brodsky L, Lipchick B, Pal M, Novototskaya L, Chenchik AA, Sen GC, Komarova EA, Gudkov AV. 2013. p53 cooperates with DNA methylation and a suicidal interferon response to maintain epigenetic silencing of repeats and noncoding RNAs. Proc Natl Acad Sci. 110(1):E89-98. doi:10.1073/pnas.1216922110.

Lippman Z, Martienssen R. 2004. The role of RNA interference in heterochromatic silencing. Nature. 431(7006):364–370. doi:10.1038/nature02875.

Liu H. 2016. Insights into centromeric transcription in mitosis. Transcription. 7(1):21–25. doi:10.1080/21541264.2015.1127315.

Liu H, Qu Q, Warrington R, Rice A, Cheng N, Yu H. 2015. Mitotic Transcription Installs Sgo1 at Centromeres to Coordinate Chromosome Segregation. Mol Cell. 59(3):426–436. doi:10.1016/j.molcel.2015.06.018.

Lu J, Gilbert DM. 2007. Proliferation-dependent and cell cycle–regulated transcription of mouse pericentric heterochromatin. Cell Biol. 179(3):411–421. doi:10.1083/jcb.200706176.

Maison C, Bailly D, Roche D, de Oca RM, Probst AV, Vassias I, Dingli F, Lombard B, Loew D, Quivy J-P, et al. 2011. SUMOylation promotes de novo targeting of HP1α to pericentric heterochromatin. Nat Genet. 43(3):220–227. doi:10.1038/ng.765.

McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. Chromosome Res. doi:10.1007/s10577-018-9582-3.

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. Develop Cell. 42(3):226-240.e6. doi:10.1016/j.devcel.2017.07.001.

Molina O, Vargiu G, Abad MA, Zhiteneva A, Jeyaprakash AA, Masumoto H, Kouprina N, Larionov V, Earnshaw WC. 2016. Epigenetic engineering reveals a balance between histone modifications and transcription in kinetochore maintenance. Nat Commun. 7:13334. doi:10.1038/ncomms13334.

Murchison EP, Partridge JF, Tam OH, Cheloufi S, Hannon GJ. 2005. Characterization of Dicer-deficient murine embryonic stem cells. Proc Natl Acad Sci. 102(34):12135–12140. doi:10.1073/pnas.0505479102.

Neumann P, Yan H, Jiang J. 2007. The centromeric retrotransposons of rice are transcribed and differentially processed by RNA interference. Genetics. 176(2):749–761. doi:10.1534/genetics.107.071902.

Ohkuni K, Kitagawa K. 2011. Endogenous transcription at the centromere facilitates centromere activity in budding yeast. Curr Biol. 21(20):1695–1703. doi:10.1016/j.cub.2011.08.056.

Ohzeki J, Bergmann JH, Kouprina N, Noskov VN, Nakano M, Kimura H, Earnshaw WC, Larionov V, Masumoto H. 2012. Breaking the HAC Barrier: histone H3K9 acetyl/methyl balance regulates CENP-A assembly. EMBO. 31(10):2391–2402. doi:10.1038/emboj.2012.82.

O'Neill RJ, Carone DM. 2009. The Role of ncRNA in Centromeres: A Lesson from Marsupials. In: Ugarkovic D, editor. Centromere. Vol. 48. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 77–101.

Perea-Resa C, Blower MD. 2018. Centromere Biology: transcription goes on stage. Molecular and Cellular Biology.:MCB.00263-18. doi:10.1128/MCB.00263-18.

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(1–2):72–82. doi:10.1016/j.gene.2007.11.013.

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

Probst AlineV, 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. Develop Cell. 19(4):625–638. doi:10.1016/ j.devcel.2010.09.002.

Quek XC, Thomson DW, Maag JLV, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. 2015. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res. 43(D1):D168–D173. doi:10.1093/nar/gku988.

Quénet D, Dalal Y. 2014. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. eLife. 3. doi:10.7554/eLife.03254.

Quénet D, Sturgill D, Olson M, Dalal Y. 2017. CENP-A associated lncRNAs influence chromosome segregation in human cells. BioRxiv. 1:097956. doi:10.1101/097956.

Reyes-Turcu FE, Zhang K, Zofall M, Chen E, Grewal SIS. 2011. Defects in RNA quality control factors reveal RNAi-independent nucleation of heterochromatin. Nat Struct Mol Biol. 18(10):1132–1138. doi:10.1038/nsmb.2122.

Rošić S, Köhler F, Erhardt S. 2014. Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. Cell Biol. 207(3):335–349. doi:10.1083/jcb.201404097.

Saksouk N, Simboeck E, Déjardin J. 2015. Constitutive heterochromatin formation and transcription in mammals. Epigenet Chromatin. 8:3. doi:10.1186/1756-8935-8-3.

Sana J, Faltejskova P, Svoboda M, Slaby O. 2012. Novel classes of non-coding RNAs and cancer. Transl Med. 10:103. doi:10.1186/1479-5876-10-103.

Scott KC. 2013. Transcription and ncRNAs: at the cent(rome)re of kinetochore assembly and maintenance. Chromosome Res. 21(6–7):643–651. doi:10.1007/s10577-013-9387-3.

Shapiro JA, von Sternberg R. 2005. Why repetitive DNA is essential to genome function. Biol Reviews. 80(2):227–250. doi:10.1017/S1464793104006657.

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(14):4390–4399. doi:10.1093/nar/gkh737.

Shirai A, Kawaguchi T, Shimojo H, Muramatsu D, Ishida-Yonetani M, Nishimura Y, Kimura H, Nakayama J, Shinkai Y. 2017. Impact of nucleic acid and methylated H3K9 binding activities of Suv39h1 on its heterochromatin assembly. eLife. 6. doi:10.7554/eLife.25317.

Smurova K, De Wulf P. 2018. Centromere and Pericentromere Transcription: Roles and Regulatio in Sickness and in Health. Front Genet. 9:674. doi:10.3389/fgene.2018.00674.

Solovei I, Grandi N, Knoth R, Volk B, Cremer T. 2004. Positional changes of pericentromeric heterochromatin and nucleoli in postmitotic Purkinje cells during murine cerebellum development. Cytogenet Genome Res. 105(2–4):302–310. doi:10.1159/000078202.

Tagarro I, Fernández-Peralta AM, González-Aguilera JJ. 1994. Chromosomal localization of human satellites 2 and 3 by a FISH method using oligonucleotides as probes. Hum Genet. 93(4):383–388.

Talbert PB, Henikoff S. 2018. Transcribing Centromeres: Noncoding RNAs and Kinetochore Assembly. Trends Genet. 34(8):587–599. doi:10.1016/j.tig.2018.05.001.

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(6017):593–596. doi:10.1126/science.1200801.

Topp CN, Zhong CX, Dawe RK. 2004. Centromere-encoded RNAs are integral components of the maize kinetochore. Proc Natl Acad Sci. 101(45):15986–15991. doi:10.1073/pnas.04071 54101.

Ugarković D. 2005. Functional elements residing within satellite DNAs. EMBO rep. 6(11):1035–1039. doi:10.1038/sj.embor.7400558.

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. doi:10.1093/nar/gkm1056.

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(8):1146–1160. doi:10.1101/gr.6022807.

Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y. 2014. NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. 42(D1):D98–D103. doi:10.1093/nar/gkt1222.

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(5):842-853.e7. doi:10.1016/j.molcel.2018.04.023.

CHAPTER IV

DISCLOSING THE ROLE OF PMSAT NON-CODING RNA ON *PEROMYSCUS* GENOMES: A PRELIMINARY STUDY

BASED ON THE RESEARCH ARTICLE:

Mendes-da-Silva A, Santos HAF, Adega F, Gama-Carvalho M, Chaves R (2019) The major repetitive element of *Peromyscus*, PMSat, ensures mitotic fidelity and drives chromosome evolution
In preparation

ABSTRACT |

Satellite DNA (satDNA) constitutes the major component of constitutive heterochromatin and has been considered one of the most intriguing repetitive DNA elements of eukaryotic genomes. Initially considered as "junk DNA", satDNAs play an important role in the occurrence of chromosomal reorganization during chromosomal evolution and the satellite transcripts or satellite non-coding RNAs are key players in genome regulation. Here, we report for the first time the transcription of the *Peromyscus* centromeric satDNA sequence – PMSat. The transcriptional activity in *Peromyscus* proliferative cells reveals a positive correlation between PMSat expression and DNA copy number content in each genome, maintaining, however, a the transcriptional cellular profile throughout the cell cycle. PMSat satncRNAs are nuclear transcripts and accumulate mostly at G2/M transition and in mitosis onset. A preliminary knockdown experiment anticipates the potential involvement of PMSat satncRNAs on kinetochore assembly and function. This work uncovers PMSat transcripts as crucial elements for chromosome segregation fidelity.

IV. 1. INTRODUCTION

As an essential structure in the genome, the centromere is the chromosomal region crucial for chromosome segregation and genome stability across all eukaryotes (Aldrup-Macdonald and Sullivan 2014). During cell division, a dynamic multiprotein kinetochore complex assembles on centromeric DNA and coordinates the attachment to microtubules and the movement of chromosomes along the mitotic spindle (Aldrup-Macdonald and Sullivan 2014; Forer et al. 2015). It is assumed that genomic along with epigenetic pathways are important for the structure and functionality of centromeres. However, the synergy among centromeric components remains unclear (Perpelescu and Fukagawa 2011; Hayden et al. 2013; Plohl et al. 2014). As a vital component of the centromeres of higher eukaryotes, centromeric protein A (CENP-A in mammals), a species-specific histone H3 variant, is considered an epigenetic mark that contributes for specialized chromatin at active centromeres (reviewed by Buscaino et al. 2010; Earnshaw 2015). Long-term established centromeres are frequently formed by repetitive DNA, mainly megabase-scale arrays of tandemly repeated satellite DNAs (satDNAs) (Melters et al. 2013; Plohl et al. 2014). Paradoxically, both centromeric DNA and protein components are characterized by a high degree of divergence in spite of conserved centromeric function (Henikoff et al. 2001).

The days when the centromere was considered a transcriptionally inert region of the chromosome are long gone. This idea has been refuted by the multiple observations showing that the transcription of centromeric sequences is a highly conserved feature of all eukaryote genomes studied so far. In fact, both centromeric transcription and the resultant non-coding satellite RNAs (satncRNAs) have been shown to contribute to centromeric function (Lu and

Gilbert 2007; Ferri et al. 2009; Probst et al. 2010; Hsieh et al. 2011; Kishi et al. 2012; Rošić et al. 2014; McNulty et al. 2017). Studies on the transcription of satncRNAs have revealed the existence of pericentromere (PCT) and centromere (CT) satncRNAs (Chan and Wong 2012; Gent and Dawe 2012; Rošić et al. 2014; Perea-Resa and Blower 2018). PCT and CT satncRNAs appear to be essential for the formation and maintenance of heterochromatin and for kinetochore assembly, respectively; as abnormalities in the transcription of each of these satDNAs adversely affect cellular function, causing centromeric dysfunction, genomic instability and tumorigenesis (Ting et al. 2011; Hall et al. 2012; Zhu et al. 2018).

PMSat, a 345 bp monomeric unit satDNA identified and characterized by our group is the major component of the rodent *Peromyscus eremicus* constitutive heterochromatin (Louzada et al. 2015) and is present at the CT regions of *P. maniculatus*, *P. leucopus*, *P. califonicus* and *P.eremicus* in all the chromosomes of the chromosome complement (Section II.1 and II.2, Louzada et al. 2015). The features of this satDNA led us to investigate its transcriptional potential in *P. eremicus* and *P. maniculatus.* In fact, PMSat is transcribed in these genomes. Depletion of these satncRNAs was conducted to disclose its putative cellular function(s). Altogether, our results strongly suggest that PMSat plays a functionally relevant role in centromeric activity, displaying the characteristic behavior of a centromeric satDNA.

IV. 2. MATERIALS AND METHODS

PMSat in silico analysis on Peromysucs transcriptome and putative transcription factors binding sites assessment

The search for PMSat transcripts was performed on the only available *Peromyscus* species with available FASTA transcriptome sequences (*Peromyscus californicus*, PRJNA350325 BioProjectID: 350325). Sequence analysis was performed using BLAST (Basic Local Alignment Search Tool provided by NCBI). Several search parameters were changed for the analysis of repetitive DNA: max_target_seqs and num_descriptions were set to 10.000, and e-value and word_size were set to 10^{-16} and 11, respectively. All other search parameters were set to default values. Multiple alignments were obtained using CLUSTALW cost matrix on Geneious R9 version 9.1.2 (Biomatters) with parameters set to default values.

The search for transcription factors binding sites on PMSat sequence was carried out on scaffolds that contained PMSat arrays (identified in section II.2). On each of these 231 scaffolds, one random PMSat consensus sequence was selected for further analysis. The search was performed using Vertebrate transcription factor database from Transfac (Wingender et al. 1997) with a minimum limit length of matches to 7 bp, based on the EMBOSS 6.5.7 tool tfscan. The analysis was carried out on Geneious R9 version 9.1.2 (Biomatters) interface and only positive results for Rodentia (mouse/rat) transcription factors were analyzed.

Cell culture and transfection

Cell lines from *Peromyscus maniculatus* (48,XY) and *P. californicus* (48,XX) and *P. leucopus* (48,XX) were provided by the *Peromyscus* Genetic Stock Center from the University of South Carolina (Coriell Institute). The cell line from *P. eremicus* was gently provided by the Department of Systematics and Evolution, Muséum National d'Histoire Naturelle, Paris, France. The first two cell lines were grown in Ham's F12/DMEM, *P. leucopus* and *P. eremicus* cell lines were grown respectively in EMEM and DMEM. All basal media were supplemented with 13% AmnioMax C-100 Basal Medium, 2% AminoMax C-100 supplement, 10% FBS, 100 U/mL/100 µg/mL of Penicillin/Streptomycin antibiotic mixture and 200 mM L-Glutamine (all from Gibco, Thermo Fisher Scientific). Cells were maintained at 37ºC in a humidified atmosphere of 5% CO_2.

To deplete PMSat, 50 nM of a customized Antisense LNA^{TM} GapmeRs was used (5'-FAM TCGTAGGACCAGAAACTGCAT) (Exiqon). PMSat_LNA transient transfection was carried out with Lipofectamine® RNAiMAx Transfection Reagent (Invitrogen, Thermo Fisher Scientific) according to the manufacturer's instructions. A LNA GapmeR negative control (Exiqon) was used to exclude off-target effects. A mock control (with the transfection reagent) was used for all the transfections.

Isolation of RNA

Total and small (< 200 bp) RNA fractions were isolated using the mirVana Isolation Kit (Ambion, Thermo Fisher Scientific) following the manufacturer's recommendations. The total RNA was purified using the TURBO DNA-freeTM Kit (Ambion, Thermo Fisher Scientific). The DNA and RNA obtained were quantified in a NanoDrop 1000 equipment (Thermo Fisher Scientific).

Reverse Transcription Quantitative Real-Time PCR (RT-qPCR Real Time)

Expression experiments were performed using the same TaqMan specific assay (primers/probe) described previously (section II.2.2.) and TaqMan® RNA-to-C_T™ 1-Step Kit. The 20 µL reactions included 80 ng of RNA, 0.5 µL of RT enzyme mix, 10 µL of RT-PCR Mix and 1 µL of TaqMan assay. This experiment was carried out in StepOne real-time PCR (Life Technologies Applied Biosystems), where the samples were subjected to 48ºC for 15 min and 95ºC for 10 min, followed by 40 cycles of 95ºC for 15 sec and 60ºC 1 min. All reactions were performed in triplicate, and negative controls (without template) were run for each master mix. The data were analyzed using the same parameters and the StepOne software (version 2.2.2, Life Technologies Applied Biosystems). The quantification was transformed in fold-change in comparison with a control sample.

RNA-FISH, RNA-FISH/IF and sequential DNA-FISH

The RNA-FISH procedure was performed using cells grown on a Superfrost Excell microscope slides (Thermo Scientific) at a concentration of 100.000 cells/mL. The cells were fixed with 2% (m/v) paraformaldehyde in PBS, for 20 min and permeabilized with 4% (v/v) Triton X-100/100µg/mL digitonin or 4% (v/v) Tween-20 (this last only for PCNA antibody) in PBS supplemented with 200 mM of Ribonucleoside Vanadyl Complex (RVC, Sigma Aldrich) for 15-20 min. Before hybridization, a dehydration was performed in sequential ethanol baths (70%, 90% and 100%). The cells were hybridized with the PMSat probe [PCR amplification of the PMSat cloned sequence and labelled with digoxigenin-11-dUTP (Roche

Applied Science)] overnight at 37°C. The most stringent wash was carried out in 0.1xSSC at 42°C. After this, the cells were incubated with blocking buffer for 30 min. When the RNA-FISH protocol was conjugated with IF, the primary antibody was incubated with the cells for 1 h. In both, RNA-FISH or RNA-FISH/IF, the cells where incubated with the secondary antibody for 1 h. After this, cells were mounted with coverslips and counterstained with Vectashield mounting medium containing DAPI (Vector Laboratories).

After the RNA-FISH procedure and image capture, a sequential DNA-FISH was performed. For that, the slides were equilibrated in 50%formamide/2xSSC (v/v) for 48 hours and then DNA-FISH procedures were performed on the same slides with a denaturation step of 2 minutes instead of 90 sec of the standard procedure followed.

Antibodies

Cell signaling: anti CENP-A polyclonal rabbit (IF: 1:200, #2186). Millipore: anti-cyclin D1 monoclonal mouse (IF 1:50, #05-815), anti-Cdc25 monoclonal mouse (IF 1:100, TC-15 clone, #05-507SP), anti-cyclin A polyclonal rabbit antibody (IF 1:75, #06-138), anti-phospho-histone H3 (Ser10) polyclonal rabbit antibody (IF 1:200, #06-570), anti-rabbit polyclonal FITC antibody (1:200, #AP132F). Sigma Aldrich: antidigoxigenin-50-TAMRA (1:200, #11207750910). Zymed: anti-mouse monoclonal FITC (1:200, #81-6511).

Microscopy and image acquisition

The RNA-FISH and RNA-FISH/IF confocal fluorescence images were acquired on a LSM 510 META with an Axio Imager Z1 microscope (Zeiss) and LSM 510 software (version 4.0 SP2). In order to normalize the results, for all the images it was applied the same microscope settings. The lasers used were: argon (488 nm) set at 12.9%, helium–neon (543 nm) set at 50.8% and Diode (405 nm) set at 9.9%. The pinhole was set to 96 mm (1.02 airy units) for argon laser, 102 mm (0.98 airy units) for helium–neon laser and 112 mm for the Diode laser using a 63x objective. Images were captured at a scan speed of 4 with 1 μm thick Z sections. The images' deconvolution was performed using the AutoQuant X3 software (Media Cybernetics) and processed in TIFF images with ImageJ (1.47v). A Zeiss Axiovert 200 microscope with P.A.L.M. image browser was used for transfected cells live imaging. All the images were prepared at the contrast and colour optimization (at whole image) using Adobe Photoshop (version 7.0).

Statistics and reproducibility

All data are presented as mean ± standard deviation (SD). Data were statistically analyzed in GraphPad Prism 7 (GraphPad Software, Inc.) in which statistical significance was determined using two-tailed Student's t-test for the comparison between two independent samples and analysis of variance (ANOVA) tests when more than two groups were under analysis. Fisher’s exact test was used to analyze the cell phenotypes significance. *p*-values: ns $p>0.05$, * $p\leq0.05$, ** $p\leq0.01$, *** $p\leq0.001$, **** $p\leq0.0001$.

IV. 3. RESULTS

PMSat transcripts are present in Peromyscus transcriptome sequencing projects

The transcriptome sequencing projects on *Peromyscus* genomes are now in progress for several species, including *P. eremicus*, *P. maniculatus*, *P. leucopus* and *P. californicus*. However, until now, RNA FASTA sequences are only available for *P. californicus*. Thus, a blast search was conducted to verify the possible presence of PMSat transcripts. Indeed, our analysis identifies two significant blast hits with a cover PMSat consensus sequence of 77% and 65%, with an identity of 90% and 86%, respectively (Figure IV.1).

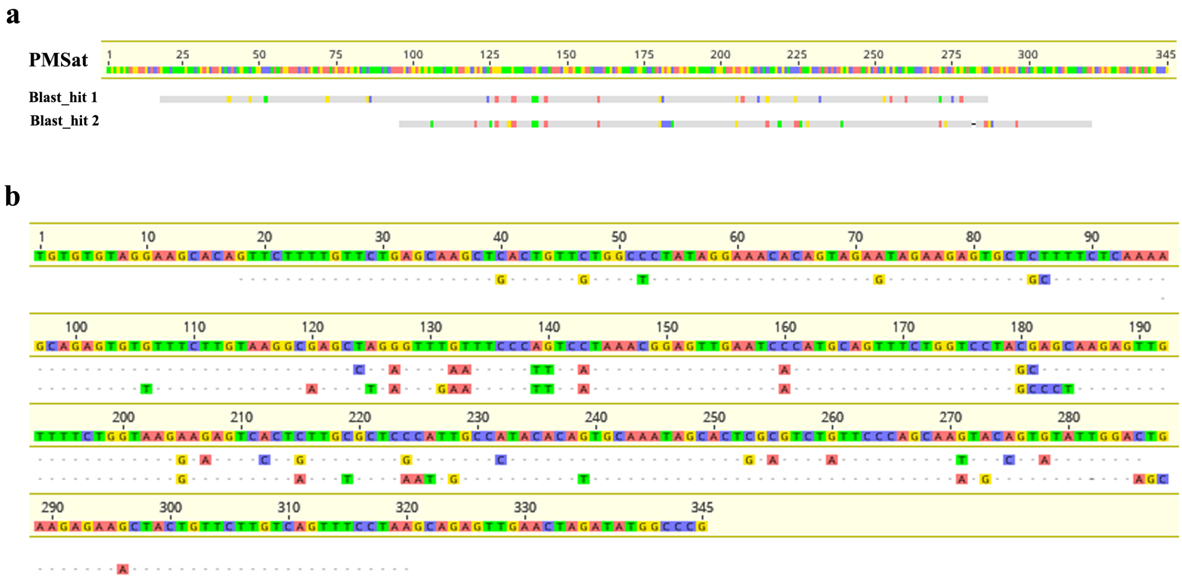


Figure IV.1. PMSat blastn search on *P. californicus* transcriptome project (PRJNA350325 BioProjectID: 350325). (a) Blast hits mapping on PMSat consensus sequence. Hit 1 (GenBank: GFCW01046971.1) and hit 2 (GenBank: GFCW01002455.1) reveals an identity of 90% and 86%, respectively. The detailed mapping with identities is presented in (b). The sequences mapping was carried out on Geneious R9 version 9.1.2. (Biomatters).

PMSat SatDNA is actively transcribed during the G2/M transition and mitosis onset

To get some more insights about the functional significance of PMSat, transcription was analyzed by Real Time RT-qPCR. Total and small RNA fractions were isolated from normal proliferative fibroblast cells from *P. eremicus, P. maniculatus*, *P. leucopus* and *P. californicus*. The transcription of PMSat was verified on both RNA fractions from all the studied species, but only the total fraction showed a significant difference between *P. eremicus* and the other *Peromyscus* species, namely an increased amount of ncsatRNAs of 20.36 times (Figure IV.2 a; Supplementary Table IV.1).

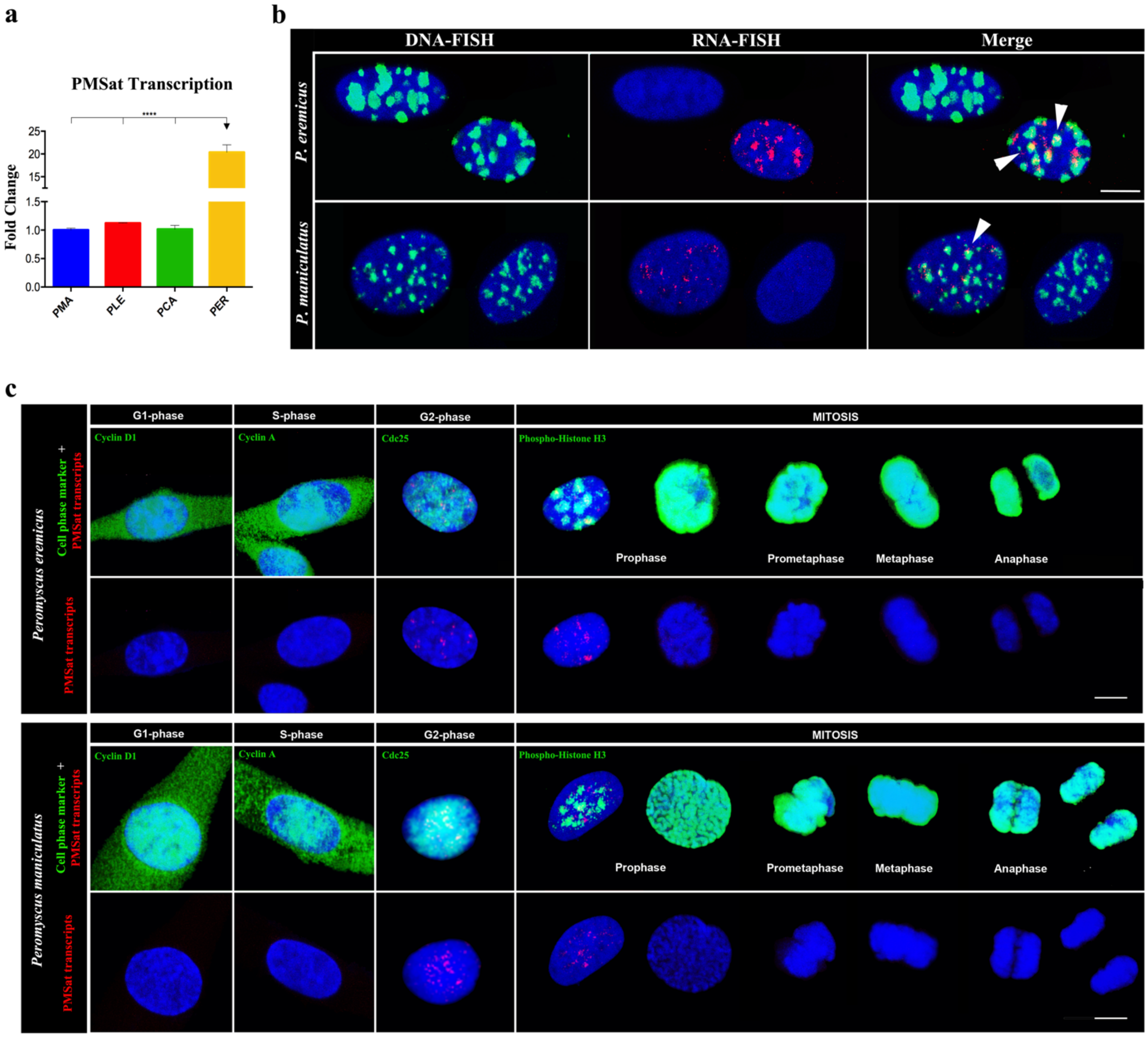


Figure IV.2. PMsat satellite RNAs on *Peromyscus* proliferating cells. (a) PMSat transcripts fold change on *P. maniculatus*, *P. leucopus*, *P. californicus* and *P. eremicus*, considering the last one as the reference genome. (b) Sequential RNA-FISH of PMSat RNA (red) and DNA-FISH of PMSat DNA (green) and merged image of both. Some RNA signals co-localize with DNA signals, resulting in yellow signals (indicated by white arrowheads). (c) Cell cycle distribution of PMSat RNA by RNA-FISH (red) conjugated with IF with cell cycle-specific antibodies; for the G1/S transition, Cyclin D1-positive cells were analysed; Cyclin A-positive cells are in the S phase; cdc25-positive cells are considered in the G2 phase; phospho-histone H3 (ser10) are cells in mitosis. Nuclei are counterstained with DAPI (blue) in all the presented images. All the scale bars represent 10 µm in all the panels. Values are mean ± SD, ****p ≤ 0.0001, as determined by one-way ANOVA.

Due to the expression differences on the analyzed species, a cell imaging analysis was performed only in *P. eremicus* and *P. maniculatus* proliferating cells. The use of RNA-FISH allowed to define the transcripts cellular localization. This analysis revealed that PMSat transcripts are restricted to the nucleus and in a cluster pattern, grouped as spots-like signals in both species (Figure IV.2 b, c). PMSat RNA-FISH followed by DNA-FISH (Figure IV.2 b) exhibits the presence of PMSat DNA sequences and transcripts at the same nuclear regions. The co-localization of specifically RNA and DNA signals suggests that the PMSat transcripts are: 1) nascent transcripts or, alternatively, 2) mature transcripts acting at the PMSat DNA clusters (arrowheads on Figure IV.2 b). RNA-FISH data demonstrated that PMSat ncRNAs are not present in all cycling cells at a specific moment. Both findings may be indicative of a cell cycle-dependent transcription. In order to comprehend this aspect, PMSat transcriptional cellular profile was analyzed throughout the cell cycle using specific cell cycle markers (by immunofluorescence) combined with RNA-FISH (Figure IV.2 c). As can be seen in Figure IV.2 c, PMSat seems to be transcribed at specific phases of the cell cycle, with PMSat ncRNAs accumulating at the G2/M transition (Cdc25 positive cells) and at mitosis onset (phosphoH3-ser10 positive cells) in both species (*P. eremicus* and *P. maniculatus*) . It was not possible to detect PMSat transcripts at the remaining cell cycle phases. This satellite transcription behavior seems to be analogous to that of other cell cycle-dependent centromeric satellite RNAs (Lu and Gilbert, 2007; Chen et al., 2008; Ferri et al., 2009).

An *in silico* analysis was also conducted for the analysis of putative transcription factors (TFs) that could bind to PMSat sequence, in an attempt to disclose the possible role of PMSat ncRNAs. We randomly selected one PMSat consensus sequence *per* scaffold and the search for putative TFs binding sites was carried out using the Rodentia (mouse/rat) TF database from Transfac (Wingender et al. 1997). This analysis identified 208 binding sites for 13 distinct TFs (Supplementary Table IV.2). Table IV.1 summarizes the TFs identified and the biological processes in which these are involved.

Table IV.1. Summary of transcription factors binding sites on PMSat sequences.

Transcription Factor	N	Transfac Accession	Uniprot Accession	Biological processes
Pit-1	96	T00691	Q00286	Development, Differentiation
Myogenin	78	T00528	P12979	Development, Differentiation, Cell cycle progression, Cell response to stimuli
CBF2	11	T00084	P53569	Enhancer of RNA polymerase II
CP2	9	T00152	Q3UNW5	Development, Differentiation Repressor of RNA polymerase II
AP-1	2	T00032	P05627	Development, Differentiation, Cell cycle progression, Aging, Tumorigenesis, Cell response to stimuli
DBP	2	T00183	Q60925	Development, Activator of RNA polymerase II
GR	2	T00333	P06537	Development, Cell response to stimuli
IL-6 RE-BP	2	T01499	P08505	Differentiation, Cell response to stimuli, Cell proliferation
MEP-1	2	T00970	P28825	Transcription repression
CAC-Binding Protein	1	T00076	N.A.	N.A. *

Fra1	1	T01208	P48755	Development, Cell response to stimuli, cell proliferation, Activator of RNA polymerase II
NF-E2	1	T00557	Q07279	Development, Differentiation, cell proliferation, regulation of RNA polymerase II
SRF	1	T00765	Q9JM73	Development, Differentiation, Cell response to stimuli, Cell proliferation and migration, regulation of RNA polymerase II

The biological processes were referred based in Uniprot information for each TF.
N: Number of binding sites among 231 PMSat consensus sequences; N.A. not available;
*any protein that binding to a CACC motif

PMSat RNA depletion appears to lead to mitotic errors during the cell cycle

As a preliminary assay on the role of PMSat transcripts in *Peromyscus* cells, we designed a knockdown experiment on *P. eremicus* proliferating cells, once this species showed the highest expression of PMSat. PMSat transcripts were depleted with locked nucleic acid (LNA) GapmeRs in assays of 24 h (Figure IV.3 a). The knockdown experiments with LNA GapmeR degrade RNA in a RNase H-dependent manner (Kauppinen et al. 2005). A negative control, LNA GapmeR, was also included. The real-time RT-qPCR results showed that the RNA levels of PMSat were reduced by 84% compared with the mock control (Figure IV.3 b). The cells transfected with PMSat_LNA were easily visualized once the LNA GapmeR was FAM labelled (Figure IV.3 a). Immunofluorescence analysis of the transfected cell population (at the 24 h experiment) with an anti-tubulin antibody exhibited nuclear abnormalities, with 10% of the cells containing two or more nuclei (multinucleated cells) and 8% of the cells with large nuclei (Figure IV.3 c, d). These preliminary findings suggest a putative role for PMSat RNAs in chromosome segregation, since the PMSat transcripts demonstrated a propensity for nuclear abnormalities concomitant with aneuploidy.

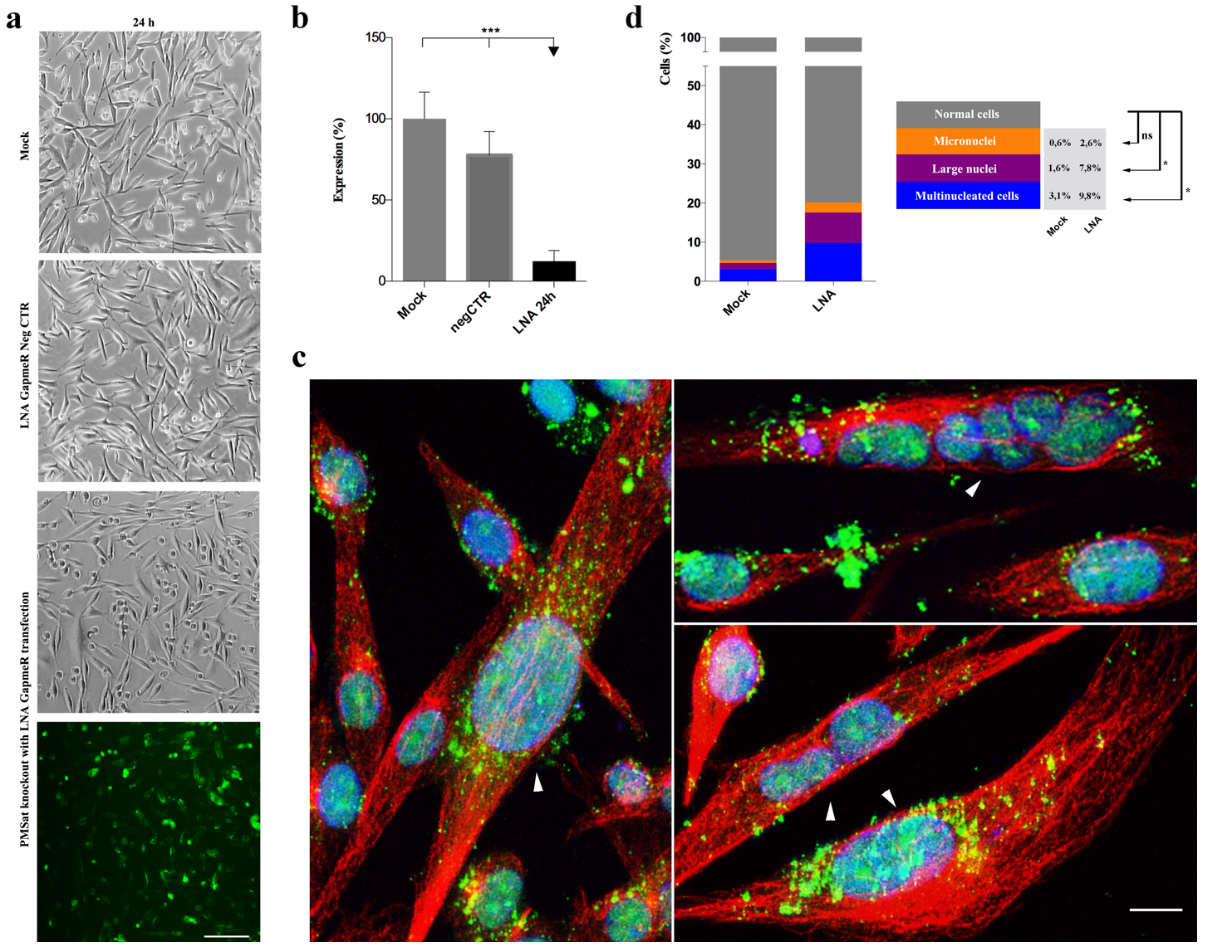


Figure IV.3. PMSat knockout leads to aneuploidy phenotypes. (a) Cell imaging of *P. eremicus* cells at 24 h in mock, negative control and PMSat_LNA. Scale barsrepresent 100 µm. (b) PMSat expression analysis of PMSat knockout in *P. eremicus* cells by real-time RT-qPCR using mock as reference at 24 h shows a high decrease in PMSat RNA. A negative control (negative LNA GapmeR) was used in the experiment. Values are mean ± SD, ***$p \leq 0.001$, as determined by one-way ANOVA. (c) Immunofluorescence of *P. eremicus* cells after PMSat knockout with tubulin antibody (red) reveals several nuclear abnormalities (micronuclei, large nuclei and multinucleated cells; indicated by arrowheads). The nucleus was counterstained with DAPI (blue). (d) Percentage of *P. eremicus* cells exhibiting nuclear abnormalities after PMSat knockout, with statistical significance mock *versus* PMSat_LNA. Was analysed ~1000 cells on mock and ~500 cells on PMSat_LNA. P-value: ns non-significant, * $p \leq 0.05$, as determined by Fisher's exact test.

IV. 4. DISCUSSION

In general, transcripts of satDNAs have been identified in various eukaryotic genomes, however, reports on the cellular location, transcriptional profile, binding partners or function are far more scarce (Wong et al. 2007; Chan et al. 2012; Ideue et al. 2014; McNulty et al. 2017). Our data revealed that all proliferative cells from the different *Peromyscus* species under analysis transcribe PMSat, being *P. eremicus* cells the ones showing the highest transcriptional levels (Figure IV.2 a). This seems to be a reflection of the striking amount of PMSat DNA in the genome of *P. eremicus* largely extended to the chromosomes' p-arms in large (peri)centromeric blocks, comparatively to the other species (Section II.2.). Additionally, comparisons between centromeric and pericentromeric transcripts reveal that the overall level of centromeric satncRNAs is lower than pericentromeric one (Ohkuni and Kitagawa 2011), being in some cases, almost undetectable due to the centromeric ncRNAs rapid turnover (Choi et al. 2011; Ohkuni and Kitagawa 2011; Chan et al. 2012). Additionally, the transcription of centromeric satDNAs is also constrained by the highly condensed state of the centromere (Schalch and Steiner 2017). As an example, the transcriptional level of FA-SAT, a highly abundant satellite DNA family in the Felidae genome, greatly decreased when the sequences were evolutionarily relocated to the centromeric region in some related species (Chaves et al. 2017). These findings indicate that a large amount of copies of satDNA in a particular genome is not the only condition for being highly transcribed, being the chromosome region where it is located also an important factor of this equation. Nevertheless, in *Peromyscus* genomes, it seems that PMSat transcription is highly associated with variation in DNA copy number (cf. Section II.2. Figure II.2.6).

The role of PMSat DNA and ncRNAs in proliferative cells

The transcription analysis by RNA-FISH in *P. eremicus* and *P. maniculatus* demonstrated a similar transcriptional cellular behavior throughout the cell cycle independently from its representativeness in the genomes (Figure IV.2). Depending on the localization of the DNA (i.e. CT or the PCT region), satncRNAs display distinct molecular functions (reviewed in McNulty and Sullivan, 2018; Perea-Resa and Blower, 2018). Long non-coding satellite RNAs derived from PCT repeats directly associate with key players on heterochromatin formation in human and mouse cells, namely with methyltransferases (Suv39h1/2) and heterochromatin protein 1 (HP1) (Maison et al. 2011; Camacho et al. 2017; Johnson et al. 2017; Shirai et al. 2017). However, the transcripts that derive from the central core of the centromere seem to directly contribute to the centromeric function, specifically, in chromatin remodeling/CENP-A deposition and in kinetochore assembly in mitosis (Perea-Resa and Blower 2018). CT transcripts interact with the centromere proteins CENP-A, CENP-B and CENP-C (Wong et al. 2007; Carone et al. 2009; Du et al. 2010; Quénet and Dalal 2014; McNulty et al. 2017), as well as with components of the Chromosome Passenger Complex (CPC), including Aurora-B, INCENP and Survivin (Ferri et al. 2009; Ideue et al. 2014). These fundamental processes in which CT and PCT transcripts are involved occur in different phases of the cell cycle and the variation in the centromere transcript levels throughout the cell cycle has been reported in many works (Lu and Gilbert 2007; Ferri et al. 2009; Probst et al. 2010; Maison et al. 2011). In mouse, long transcripts derived from the PCT region have been detected in G1-phase (CENP-A loading in mammals) (Lu and Gilbert 2007; Maison et al. 2011). Also in mouse, but regarding transcripts from the CT region (mouse minor satellite), these exhibit a different cell cycle profile, with an enrichment at G2/M (Ferri et al. 2009). This profile fits our data, revealing that PMSat transcripts exhibit characteristics of a centromeric satncRNA, a dynamic temporal profile, accumulating in G2 and mitosis onset in both *P. eremicus* and *P. maniculatus* (Figure IV.2 c). Besides, PMSat knockout preliminary experiments highlights this putative major role as it leads to aneuploidy phenotypes (Figure IV.3). Furthermore, our previous *in silico* analysis of *P. maniculatus* (section II.2.) revealed the existence of the CENP-b box motif (in a total of 40683 motifs) as part of a large amount of PMSat monomers scattered across all the scaffolds analyzed, either the conserved functional motif (CENP-B box motif - CRS*wt*, Masumoto et al. 2004) or presenting one to two mismatches (CRS*var*). Finally, the co-localization of the sequence with the centromeric protein CENP-A at the centromeres, proved that PMSat is a component of the active centromere (cf. Section II.2. Figure II.2.6).

The presence of putative binding sites for a variety of TFs involved in distinct biological processes in PMSat highlights the functional significance of the originated satncRNA (Table IV.1 and Supplementary Table IV.2). Indeed, satDNA transcripts have been reported in a variety of cell conditions related to cell proliferation, development and differentiation, cell aging, cell stresses and tumorigenesis (Eymery et al. 2009; Eymery et al. 2010; Biscotti et al. 2015; Ferreira et al. 2015; Zhu et al. 2018). Our data are concordant with the hypothesis that PMSat transcription is performed by RNApolII (cf. Table IV.1, repressors/activators/regulators of RNApolII on PMSat monomer sequence). Furthermore, PMSat ncRNAs seem to participate in a variety of cellular biological conditions in addition to cell proliferation (verified on these study), as cell stresses/stimuli and cell development and differentiation. All these biological roles are in accordance with a satDNA transcribed from a centromeric location.

IV. 5. CONCLUSION

In this preliminary analysis of PMSat satncRNAs we reveal that these sequence are transcribed in a cell cycle dependent mode, accumulating at G2 and in mitosis onset in both *P. eremicus* and *P. maniculatus* proliferative cells. Together with our previous results, specifically the presence of the conserved DNA-binding domain for the centromeric protein CENP-B (CENP-B box); the co-localization of the CENP-A protein that forms a stable complex with this motif on PMSat monomers; the aneuploidy phenotypes resulting from PMSat knockout and the centromeric nature of this satDNA family, anticipates its involvement in the centromeric function, both as a DNA sequence and as a ncRNA. Also, the search for putative TFs binding sites revealed that PMSat could indeed bind to a variety of TFs involved in distinct biological processes, what also anticipates the functional relevance for this satDNA family.

IV. 6. REFERENCES

Aldrup-Macdonald ME, Sullivan BA. 2014. The past, present, and future of human centromere genomics. Genes (Basel). 5(1):33–50.

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

Buscaino A, Allshire R, Pidoux A. 2010. Building centromeres: home sweet home or a nomadic existence? Curr Opin Genet Dev. 20(2):118–126. doi:10.1016/j.gde.2010.01.006.

Camacho OV, Galan C, Swist-Rosowska K, Ching R, Gamalinda M, Karabiber F, De La Rosa-Velazquez I, Engist B, Koschorz B, Shukeir N. 2017. Major satellite repeat RNA stabilize heterochromatin retention of Suv39h enzymes by RNA-nucleosome association and RNA: DNA hybrid formation. Elife. 6:e25293.

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(1):113–125. doi:10.1007/s00412-008-0181-5.

Chan FL, Marshall OJ, Saffery R, Won Kim B, 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. 109(6):1979–1984. doi:10.1073/pnas.1108705109.

Chan FL, Wong LH. 2012. Transcription in the maintenance of centromere chromatin identity. Nucleic Acids Res. 40(22):11178–11188. doi:10.1093/nar/gks921.

Chaves R, Ferreira D, Mendes-da-Silva A, Meles S, Adega F. 2017. FA-SAT Is an Old Satellite DNA Frozen in Several Bilateria Genomes. Genome Biol Evol. 9(11):3073–3087. doi:10.1093/gbe/evx212.

Chen ES, Zhang K, Nicolas E, Cam HP, Zofall M, Grewal SIS. 2008. Cell cycle control of centromeric repeat transcription and heterochromatin assembly. Nature. 451(7179):734–737. doi:10.1038/nature06561.

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. Biol Chemistry. 286(26):23600–23607. doi:10.1074/jbc.M111.228510.

Du Y, Topp CN, Dawe RK. 2010. DNA binding of centromere protein C (CENPC) is stabilized by single-stranded RNA. PLoS Genet. 6(2):e1000835. doi:10.1371/journal.pgen.1000835.

Earnshaw WC. 2015. Discovering centromere proteins: from cold white hands to the A, B, C of CENPs. Nat Rev Mol Cell Biol. 16(7):443–449. doi:10.1038/nrm4001.

Eymery A, Callanan M, Vourc'h C. 2009a. The secret message of heterochromatin: new insights into the mechanisms and function of centromeric and pericentric repeat sequence transcription. Develop Biol. 53(2–3):259–268. doi:10.1387/ijdb.082673ae.

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(11):1845–1855. doi:10.1016/j.yexcr.2010.02.002.

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. Chromosome Res. 23(3):479–493. doi:10.1007/s10577-015-9482-8.

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(15):5071–5080. doi:10.1093/nar/gkp529.

Forer A, Johansen KM, Johansen J. 2015. Movement of chromosomes with severed kinetochore microtubules. Protoplasma. 252(3):775–781. doi:10.1007/s00709-014-0752-7.

Gent JI, Dawe RK. 2012. RNA as a Structural and Regulatory Component of the Centromere. Ann Rev Genet. 46(1):443–453. doi:10.1146/annurev-genet-110711-155419.

Hall LE, Mitchell SE, O'Neill RJ. 2012. Pericentric and centromeric transcription: a perfect balance required. Chromosome Res. 20(5):535–546. doi:10.1007/s10577-012-9297-9.

Hayden KE, Strome ED, Merrett SL, Lee H-R, Rudd MK, Willard HF. 2013. Sequences associated with centromere competency in the human genome. Mol Cell Biol. 33(4):763–772. doi:10.1128/MCB.01198-12.

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

Hsieh C-L, Lin C-L, Liu H, Chang Y-J, Shih C-J, Zhong CZ, Lee S-C, Tan BC-M. 2011. WDHD1 modulates the post-transcriptional step of the centromeric silencing pathway. Nucleic Acids Res. 39(10):4048–4062. doi:10.1093/nar/gkq1338.

Ideue T, Cho Y, Nishimura K, Tani T. 2014. Involvement of satellite I noncoding RNA in regulation of chromosome segregation. Genes to Cells. 19(6):528–538. doi:10.1111/gtc.12149.

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. doi:10.7554/eLife.25299.

Kauppinen S, Vester B, Wengel J. 2005. Locked nucleic acid (LNA): High affinity targeting of RNA for diagnostics and therapeutics. Drug Discovery Today: Technologies. 2(3):287–290. doi:10.1016/j.ddtec.2005.08.012.

Kishi Y, Kondo S, Gotoh Y. 2012. Transcriptional Activation of Mouse Major Satellite Regions during Neuronal Differentiation. Cell Struct Function. 37(2):101–110. doi:10.1247/csf.12009.

Louzada S, Vieira-da-Silva A, Mendes-da-Silva A, Kubickova S, Rubes J, Adega F, Chaves R. 2015. A novel satellite DNA sequence in the Peromyscus genome (PMSat): Evolution via copy number fluctuation. Mol Phyl Evol. 92:193–203. doi:10.1016/j.ympev.2015.06.008.

Lu J, Gilbert DM. 2007. Proliferation-dependent and cell cycle–regulated transcription of mouse pericentric heterochromatin. Cell Biol. 179(3):411–421. doi:10.1083/jcb.200706176.

Maison C, Bailly D, Roche D, de Oca RM, Probst AV, Vassias I, Dingli F, Lombard B, Loew D, Quivy J-P, et al. 2011. SUMOylation promotes de novo targeting of HP1α to pericentric heterochromatin. Nature Genet. 43(3):220–227. doi:10.1038/ng.765.

McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. Chromosome Res. doi:10.1007/s10577-018-9582-3.

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. Develop Cell. 42(3):226-240.e6. doi:10.1016/j.devcel.2017.07.001.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14(1):R10.

Ohkuni K, Kitagawa K. 2011. Endogenous transcription at the centromere facilitates centromere activity in budding yeast. Curr Biol. 21(20):1695–1703. doi:10.1016/j.cub.2011.08.056.

Perea-Resa C, Blower MD. 2018. Centromere Biology: transcription goes on stage. Mol Cel Biol. MCB.00263-18. doi:10.1128/MCB.00263-18.

Perpelescu M, Fukagawa T. 2011. The ABCs of CENPs. Chromosoma. 120(5):425–446. doi:10.1007/s00412-011-0330-0.

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

Probst AlineV, 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. Develop Cell. 19(4):625–638. doi:10.1016/j.devcel.2010.09.002.

Quénet D, Dalal Y. 2014. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. eLife. 3. doi:10.7554/eLife.03254.

Rošić S, Köhler F, Erhardt S. 2014. Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. Cell Biol. 207(3):335–349. doi:10.1083/jcb.201404097.

Schalch T, Steiner FA. 2017. Structure of centromere chromatin: from nucleosome to chromosomal architecture. Chromosoma. 126(4):443–455. doi:10.1007/s00412-016-0620-7.

Shirai A, Kawaguchi T, Shimojo H, Muramatsu D, Ishida-Yonetani M, Nishimura Y, Kimura H, Nakayama J, Shinkai Y. 2017. Impact of nucleic acid and methylated H3K9 binding activities of Suv39h1 on its heterochromatin assembly. eLife. 6. doi:10.7554/eLife.25317.

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(6017):593–596. doi:10.1126/science.1200801.

Wingender E, Kel AE, Kel OV, Karas H, Heinemeyer T, Dietze P, Knüppel R, Romaschenko AG, Kolchanov NA. 1997. TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. Nucleic Acids Res. 25(1):265–268.

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(8):1146–1160. doi:10.1101/gr.6022807.

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(5):842-853.e7. doi:10.1016/j.molcel.2018.04.023.

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY TABLES

Supplementary Table IV.1. Relative quantification of PMSat transcripts in the analyzed species. *P. maniculatus* was considered the reference genome.

Species	Relative quantification
PMA	1.00 (±0.04)
PLE	1.12 (±5.82x10^{-3})
PCA	1.02 (±0.07)
PER	20.36 (±1.66)

Supplementary Table IV.2. Transcription Binding Sites on random PMSat sequences on each 231 scaffold on *P. maniculatus* Genome project (BioProject_PRJNA53563)

Transcription factor	Min. (position on sequence)	Max. (position on sequence)	Length	Species	Transfac Accession	Sequence
AP-1	595	601	7	mouse, Mus musculus.	T00032	random 2168
AP-1	595	601	7	rat, Rattus norvegicus.	T00031	random 2168
CAC-binding protein	625	631	7	mouse, Mus musculus.	T00076	random 1885
CBF (2)	11	18	8	rat, Rattus norvegicus.	T00084	random 90
CBF (2)	11	18	8	rat, Rattus norvegicus.	T00084	random 90
CBF (2)	25	32	8	rat, Rattus norvegicus.	T00084	random 891
CBF (2)	26	33	8	rat, Rattus norvegicus.	T00084	random 807
CBF (2)	318	325	8	rat, Rattus norvegicus.	T00084	random 686
CBF (2)	808	815	8	rat, Rattus norvegicus.	T00084	random 523
CBF (2)	131	138	8	rat, Rattus norvegicus.	T00084	random 523
CBF (2)	78	85	8	rat, Rattus norvegicus.	T00084	random 2262
CBF (2)	185	192	8	rat, Rattus norvegicus.	T00084	random 18949
CBF (2)	185	192	8	rat, Rattus norvegicus.	T00084	random 18944
CBF (2)	298	305	8	rat, Rattus norvegicus.	T00084	random 1833
CP2	10	16	7	mouse, Mus musculus.	T00152	random 90
CP2	24	30	7	mouse, Mus musculus.	T00152	random 891
CP2	25	31	7	mouse, Mus musculus.	T00152	random 807
CP2	317	323	7	mouse, Mus musculus.	T00152	random 686
CP2	807	813	7	mouse, Mus musculus.	T00152	random 523
CP2	77	83	7	mouse, Mus musculus.	T00152	random 2262
CP2	184	190	7	mouse, Mus musculus.	T00152	random 18949
CP2	184	190	7	mouse, Mus musculus.	T00152	random 18944
CP2	297	303	7	mouse, Mus musculus.	T00152	random 1833
DBP	928	934	7	rat, Rattus norvegicus.	T00183	random 19016
DBP	1,02	1,026	7	rat, Rattus norvegicus.	T00183	random 1714
FraI	595	601	7	rat, Rattus norvegicus.	T01208	random 2168
GR	60	68	9	rat, Rattus norvegicus.	T00333	random 775
GR	246	254	9	rat, Rattus norvegicus.	T00333	random 1708
IL-6 RE-BP	282	290	9	rat, Rattus norvegicus.	T01499	random 609
IL-6 RE-BP	363	371	9	rat, Rattus norvegicus.	T01499	random 2168
MEP-1	792	798	7	mouse, Mus musculus.	T00970	random 2575
MEP-1	622	628	7	mouse, Mus musculus.	T00970	random 2504
myogenin	176	183	8	mouse, Mus musculus.	T00528	random 964
myogenin	156	162	7	mouse, Mus musculus.	T00528	random 807
myogenin	41	48	8	mouse, Mus musculus.	T00528	random 736

myogenin	231	238	8	mouse, Mus musculus.	T00528	random 6535
myogenin	132	139	8	mouse, Mus musculus.	T00528	random 639
myogenin	275	282	8	mouse, Mus musculus.	T00528	random 5948
myogenin	22	29	8	mouse, Mus musculus.	T00528	random 5440
myogenin	130	137	8	mouse, Mus musculus.	T00528	random 4932
myogenin	189	196	8	mouse, Mus musculus.	T00528	random 475
myogenin	32	39	8	mouse, Mus musculus.	T00528	random 4642
myogenin	140	147	8	mouse, Mus musculus.	T00528	random 3711
myogenin	129	136	8	mouse, Mus musculus.	T00528	random 2723
myogenin	169	176	8	mouse, Mus musculus.	T00528	random 266
myogenin	246	253	8	mouse, Mus musculus.	T00528	random 2570
myogenin	162	169	8	mouse, Mus musculus.	T00528	random 2261
myogenin	63	70	8	mouse, Mus musculus.	T00528	random 2085
myogenin	117	124	8	mouse, Mus musculus.	T00528	random 1991
myogenin	98	105	8	mouse, Mus musculus.	T00528	random 19260
myogenin	222	229	8	mouse, Mus musculus.	T00528	random 19206
myogenin	129	136	8	mouse, Mus musculus.	T00528	random 19148
myogenin	308	315	8	mouse, Mus musculus.	T00528	random 19130
myogenin	206	213	8	mouse, Mus musculus.	T00528	random 19078
myogenin	121	128	8	mouse, Mus musculus.	T00528	random 19075
myogenin	269	276	8	mouse, Mus musculus.	T00528	random 19074
myogenin	12	19	8	mouse, Mus musculus.	T00528	random 19052
myogenin	334	341	8	mouse, Mus musculus.	T00528	random 19036
myogenin	63	70	8	mouse, Mus musculus.	T00528	random 19034
myogenin	205	212	8	mouse, Mus musculus.	T00528	random 19033
myogenin	870	877	8	mouse, Mus musculus.	T00528	random 19016
myogenin	526	533	8	mouse, Mus musculus.	T00528	random 19016
myogenin	185	192	8	mouse, Mus musculus.	T00528	random 19016
myogenin	648	655	8	mouse, Mus musculus.	T00528	random 19016
myogenin	115	122	8	mouse, Mus musculus.	T00528	random 19015
myogenin	43	50	8	mouse, Mus musculus.	T00528	random 19014
myogenin	134	141	8	mouse, Mus musculus.	T00528	random 19012
myogenin	888	895	8	mouse, Mus musculus.	T00528	random 18998
myogenin	545	552	8	mouse, Mus musculus.	T00528	random 18998
myogenin	202	209	8	mouse, Mus musculus.	T00528	random 18998
myogenin	57	64	8	mouse, Mus musculus.	T00528	random 18997
myogenin	396	403	8	mouse, Mus musculus.	T00528	random 18978
myogenin	52	59	8	mouse, Mus musculus.	T00528	random 18978
myogenin	161	168	8	mouse, Mus musculus.	T00528	random 18951
myogenin	146	153	8	mouse, Mus musculus.	T00528	random 18932
myogenin	30	37	8	mouse, Mus musculus.	T00528	random

						18928
myogenin	171	178	8	mouse, Mus musculus.	T00528	random 18912
myogenin	581	588	8	mouse, Mus musculus.	T00528	random 18911
myogenin	268	275	8	mouse, Mus musculus.	T00528	random 18910
myogenin	298	305	8	mouse, Mus musculus.	T00528	random 18909
myogenin	632	639	8	mouse, Mus musculus.	T00528	random 18908
myogenin	291	298	8	mouse, Mus musculus.	T00528	random 18908
myogenin	318	325	8	mouse, Mus musculus.	T00528	random 18907
myogenin	56	63	8	mouse, Mus musculus.	T00528	random 18906
myogenin	742	749	8	mouse, Mus musculus.	T00528	random 18906
myogenin	121	128	8	mouse, Mus musculus.	T00528	random 18904
myogenin	325	332	8	mouse, Mus musculus.	T00528	random 18895
myogenin	821	827	7	mouse, Mus musculus.	T00528	random 1885
myogenin	236	243	8	mouse, Mus musculus.	T00528	random 18785
myogenin	224	231	8	mouse, Mus musculus.	T00528	random 18711
myogenin	282	289	8	mouse, Mus musculus.	T00528	random 18043
myogenin	36	43	8	mouse, Mus musculus.	T00528	random 18042
myogenin	100	107	8	mouse, Mus musculus.	T00528	random 17833
myogenin	95	102	8	mouse, Mus musculus.	T00528	random 1766
myogenin	783	790	8	mouse, Mus musculus.	T00528	random 1766
myogenin	439	446	8	mouse, Mus musculus.	T00528	random 1766
myogenin	45	52	8	mouse, Mus musculus.	T00528	random 17450
myogenin	36	43	8	mouse, Mus musculus.	T00528	random 1651
myogenin	303	310	8	mouse, Mus musculus.	T00528	random 16503
myogenin	283	290	8	mouse, Mus musculus.	T00528	random 1541
myogenin	112	119	8	mouse, Mus musculus.	T00528	random 1514
myogenin	42	49	8	mouse, Mus musculus.	T00528	random 1513
myogenin	155	162	8	mouse, Mus musculus.	T00528	random 1470
myogenin	206	213	8	mouse, Mus musculus.	T00528	random 1446
myogenin	141	148	8	mouse, Mus musculus.	T00528	random 1416
myogenin	85	92	8	mouse, Mus musculus.	T00528	random 1389
myogenin	114	121	8	mouse, Mus musculus.	T00528	random 13735
myogenin	127	134	8	mouse, Mus musculus.	T00528	random 1355
myogenin	14	21	8	mouse, Mus musculus.	T00528	random 1156
myogenin	39	46	8	mouse, Mus musculus.	T00528	random 1064
NF-E2	595	601	7	mouse, Mus musculus.	T00557	random 2168
Pit-1a	252	258	7	rat, Rattus norvegicus.	T00691	random 964
Pit-1a	339	345	7	rat, Rattus norvegicus.	T00691	random 90
Pit-1a	339	345	7	rat, Rattus norvegicus.	T00691	random 90
Pit-1a	117	123	7	rat, Rattus norvegicus.	T00691	random 736
Pit-1a	895	901	7	rat, Rattus norvegicus.	T00691	random 732
Pit-1a	597	603	7	rat, Rattus norvegicus.	T00691	random 732
Pit-1a	1,237	1,243	7	rat, Rattus norvegicus.	T00691	random 732
Pit-1a	307	313	7	rat, Rattus norvegicus.	T00691	random 6535
Pit-1a	208	214	7	rat, Rattus norvegicus.	T00691	random 639
Pit-1a	98	104	7	rat, Rattus norvegicus.	T00691	random 5440
Pit-1a	206	212	7	rat, Rattus norvegicus.	T00691	random 4932
Pit-1a	108	114	7	rat, Rattus norvegicus.	T00691	random 4642

Pit-1a	216	222	7	rat, Rattus norvegicus.	T00691	random 3711
Pit-1a	205	211	7	rat, Rattus norvegicus.	T00691	random 2723
Pit-1a	245	251	7	rat, Rattus norvegicus.	T00691	random 266
Pit-1a	245	251	7	rat, Rattus norvegicus.	T00691	random 266
Pit-1a	322	328	7	rat, Rattus norvegicus.	T00691	random 2570
Pit-1a	186	192	7	rat, Rattus norvegicus.	T00691	random 2504
Pit-1a	412	418	7	rat, Rattus norvegicus.	T00691	random 2504
Pit-1a	61	67	7	rat, Rattus norvegicus.	T00691	random 2262
Pit-1a	238	244	7	rat, Rattus norvegicus.	T00691	random 2261
Pit-1a	41	47	7	rat, Rattus norvegicus.	T00691	random 2181
Pit-1a	139	145	7	rat, Rattus norvegicus.	T00691	random 2085
Pit-1a	193	199	7	rat, Rattus norvegicus.	T00691	random 1991
Pit-1a	174	180	7	rat, Rattus norvegicus.	T00691	random 19260
Pit-1a	298	304	7	rat, Rattus norvegicus.	T00691	random 19206
Pit-1a	205	211	7	rat, Rattus norvegicus.	T00691	random 19148
Pit-1a	40	46	7	rat, Rattus norvegicus.	T00691	random 19130
Pit-1a	282	288	7	rat, Rattus norvegicus.	T00691	random 19078
Pit-1a	70	76	7	rat, Rattus norvegicus.	T00691	random 19077
Pit-1a	197	203	7	rat, Rattus norvegicus.	T00691	random 19075
Pit-1a	1	7	7	rat, Rattus norvegicus.	T00691	random 19074
Pit-1a	88	94	7	rat, Rattus norvegicus.	T00691	random 19052
Pit-1a	139	145	7	rat, Rattus norvegicus.	T00691	random 19034
Pit-1a	281	287	7	rat, Rattus norvegicus.	T00691	random 19033
Pit-1a	648	655	8	rat, Rattus norvegicus.	T00691	random 19016
Pit-1a	946	952	7	rat, Rattus norvegicus.	T00691	random 19016
Pit-1a	602	608	7	rat, Rattus norvegicus.	T00691	random 19016
Pit-1a	191	197	7	rat, Rattus norvegicus.	T00691	random 19015
Pit-1a	119	125	7	rat, Rattus norvegicus.	T00691	random 19014
Pit-1a	210	216	7	rat, Rattus norvegicus.	T00691	random 19012
Pit-1a	964	970	7	rat, Rattus norvegicus.	T00691	random 18998
Pit-1a	133	139	7	rat, Rattus norvegicus.	T00691	random 18997
Pit-1a	472	478	7	rat, Rattus norvegicus.	T00691	random 18978
Pit-1a	128	134	7	rat, Rattus norvegicus.	T00691	random 18978
Pit-1a	237	243	7	rat, Rattus norvegicus.	T00691	random 18951
Pit-1a	168	174	7	rat, Rattus norvegicus.	T00691	random 18949
Pit-1a	168	174	7	rat, Rattus norvegicus.	T00691	random 18944
Pit-1a	470	476	7	rat, Rattus norvegicus.	T00691	random 18943
Pit-1a	222	228	7	rat, Rattus norvegicus.	T00691	random 18932
Pit-1a	106	112	7	rat, Rattus norvegicus.	T00691	random 18928
Pit-1a	247	253	7	rat, Rattus norvegicus.	T00691	random

						18912
Pit-1a	657	663	7	rat, Rattus norvegicus.	T00691	random 18911
Pit-1a	314	320	7	rat, Rattus norvegicus.	T00691	random 18911
Pit-1a	30	36	7	rat, Rattus norvegicus.	T00691	random 18909
Pit-1a	367	373	7	rat, Rattus norvegicus.	T00691	random 18908
Pit-1a	23	29	7	rat, Rattus norvegicus.	T00691	random 18908
Pit-1a	50	56	7	rat, Rattus norvegicus.	T00691	random 18907
Pit-1a	721	728	8	rat, Rattus norvegicus.	T00691	random 18906
Pit-1a	818	824	7	rat, Rattus norvegicus.	T00691	random 18906
Pit-1a	475	481	7	rat, Rattus norvegicus.	T00691	random 18906
Pit-1a	132	138	7	rat, Rattus norvegicus.	T00691	random 18906
Pit-1a	197	203	7	rat, Rattus norvegicus.	T00691	random 18904
Pit-1a	57	63	7	rat, Rattus norvegicus.	T00691	random 18895
Pit-1a	617	626	10	rat, Rattus norvegicus.	T00691	random 1885
Pit-1a	11	17	7	rat, Rattus norvegicus.	T00691	random 1885
Pit-1a	676	682	7	rat, Rattus norvegicus.	T00691	random 1885
Pit-1a	311	317	7	rat, Rattus norvegicus.	T00691	random 1885
Pit-1a	450	456	7	rat, Rattus norvegicus.	T00691	random 1885
Pit-1a	312	318	7	rat, Rattus norvegicus.	T00691	random 18785
Pit-1a	300	306	7	rat, Rattus norvegicus.	T00691	random 18711
Pit-1a	218	224	7	rat, Rattus norvegicus.	T00691	random 18603
Pit-1a	262	268	7	rat, Rattus norvegicus.	T00691	random 1850
Pit-1a	14	20	7	rat, Rattus norvegicus.	T00691	random 18043
Pit-1a	112	118	7	rat, Rattus norvegicus.	T00691	random 18042
Pit-1a	175	181	7	rat, Rattus norvegicus.	T00691	random 17833
Pit-1a	859	865	7	rat, Rattus norvegicus.	T00691	random 1766
Pit-1a	515	521	7	rat, Rattus norvegicus.	T00691	random 1766
Pit-1a	171	177	7	rat, Rattus norvegicus.	T00691	random 1766
Pit-1a	121	127	7	rat, Rattus norvegicus.	T00691	random 17450
Pit-1a	157	163	7	rat, Rattus norvegicus.	T00691	random 1714
Pit-1a	842	848	7	rat, Rattus norvegicus.	T00691	random 1714
Pit-1a	112	118	7	rat, Rattus norvegicus.	T00691	random 1651
Pit-1a	35	41	7	rat, Rattus norvegicus.	T00691	random 16503
Pit-1a	15	21	7	rat, Rattus norvegicus.	T00691	random 1541
Pit-1a	188	194	7	rat, Rattus norvegicus.	T00691	random 1514
Pit-1a	118	124	7	rat, Rattus norvegicus.	T00691	random 1513
Pit-1a	231	237	7	rat, Rattus norvegicus.	T00691	random 1470
Pit-1a	282	288	7	rat, Rattus norvegicus.	T00691	random 1446
Pit-1a	269	275	7	rat, Rattus norvegicus.	T00691	random 1415
Pit-1a	161	167	7	rat, Rattus norvegicus.	T00691	random 1389
Pit-1a	190	196	7	rat, Rattus norvegicus.	T00691	random 13735
Pit-1a	203	209	7	rat, Rattus norvegicus.	T00691	random 1355
Pit-1a	70	76	7	rat, Rattus norvegicus.	T00691	random 1195
Pit-1a	90	96	7	rat, Rattus norvegicus.	T00691	random 1156
Pit-1a	115	121	7	rat, Rattus norvegicus.	T00691	random 1064
SRF	913	919	7	mouse, Mus musculus.	T00765	random 1885

CHAPTER V

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

Over the years, understanding the evolution, organization, and regulation of genomes has been one of the major research focuses of the scientific community. Since its discovery in the early 1960s, satellite DNA (satDNA) (the major constituent of the repetitive genomes' fraction) has been considered one of the most intriguing elements of eukaryotic genomes. Initially considered as "junk DNA" and a transcriptional inert portion of eukaryotic genomes, the dynamic molecular behaviour of satDNAs plays an important role in the occurrence of chromosomal reorganization during genomes' evolution (e.g. Chaves et al. 2004) and satDNA transcripts are emerging as key players in genome regulation (reviewed in Ferreira et al. 2015).

V. 1. UNVEILING THE *PEROMYSCUS'* SATELLITOME LANDSCAPE

Traditionally, the methods used for the isolation and discovery of satDNA families were time-consuming (e.g. genomic restriction digestion), allowing the identification of only the major or a few satDNA families in each genome (reviewed in Garrido-Ramos 2017). The advances in whole-genome sequencing methodologies and platforms has generated an enormous amount of sequencing data that has been used to decipher the genomes' repetitive fraction (e.g. Alkan et al. 2011; Komissarov et al. 2011; Silva et al. 2017). Due to the easy access of sequencing platforms, more and more non-model species can be studied at the genomic scale and today, the genome-wide analysis of the repetitive content of eukaryotic genomes is evolving as a pivotal step for unveiling the functional roles of these sequences in eukaryotic genomes.

The starting point of the work here presented was to perform, for the first time, a genome-wide analysis of the repetitive DNA sequences on the representative genome sequencing project of the first Peromyscine species whose genomic data was available – the deer mouse, *Peromyscus maniculatus* (Pman_1.0, Genbank assembly accession GCA_000500345.1). A bioinformatics pipeline was thus defined (Chapter II; section II.2.) based on the Tandem Repeat Finder (TRF) algorithm (Benson 1999) and tactical filters application that allowed the identification of more than 1.500 distinct arrays of large tandem repeats (TRs). Additionally, an integrated analysis based on sequence similarity between the identified TRs and repetitive sequences deposited in the Rodentia Repbase and NCBI databases clustered all the sequences into 21 families, being the majority satellite-related or transposable elements (TE)-related families. The major component of the *P. maniculatus* satellitome counts for more than 50% of all the identified TRs and corresponds to the PMSat satDNA family that was firstly physically isolated by laser microdissection from the chromosomes (peri)centromeric region of other *Peromyscus* species, *P. eremicus*, by our group (Chapter II; Section II.1; Louzada et al. 2015). The data of the *in silico* analysis of this AT-rich satDNA were in agreement with our previous report, as in *P. eremicus* genome, also in *P. maniculatus,* PMSat presents a monomer size of approximately 345 bp exhibiting a high intra-monomeric similarity. Despite the divergence of centromeric repetitive sequences, the abundance and repetitive nature of each specific TR (Melters et al. 2013), such as the presence of conserved DNA-binding domains (Fujita et al. 2015) and surrounding TE elements (reviewed in Hartley and O'Neill 2019) have been identified as common features of eukaryotic centromeres. In fact, in addition to the high abundance of PMSat in *P. maniculatus* genome, our data reveals the occurrence of PMSat monomers exhibiting CENP-B box like motifs, as well as PMSat rich regions presenting TE-related elements.

Peromyscus karyotype reorganization driven by PMSat

Experimental approaches focused on the PMSat satDNA family were conducted in four distinct *Peromyscus* species: *P. eremicus*, *P. maniculatus*, *P. leucopus* and *P. californicus* (Chapter II; section II.2.). According to an *in silico* analysis, we proved that PMSat is mainly located at the chromosomes' active centromeres and pericentromeric regions, maintaining a high degree of conservation/similarity in all the studied species despite the different number of copies of PMSat *per* genome. Traditionally, satDNA has been considered one of the most dynamic elements of eukaryotic genomes that rapidly evolve even in close-related species (Plohl et al. 2012). The high variability exhibited in terms of monomer size nucleotide sequence, chromosome organization is mostly promoted by concerted evolution, which culminates in the rapid intraspecific homogenization of occurring changes by a molecular drive process (Plohl 2010; Plohl et al. 2012). Conversely, some satDNA families seem to persist in a conserved fashion in genomes for long evolutionary periods, even if presenting a residual number of copies (e.g. Petraccioli et al. 2015; Chaves et al. 2017). Our results strongly suggest that the evolutionary pathway of PMSat was driven by copy number fluctuations and the high similarity among PMSat on the studied *Peromyscus* and non-*Peromyscus* species (Louzada et al. 2015) can reflect non-concerted evolutionary events (Plohl et al. 2010). Moreover, the genomic location of PMSat at the centromeres can favour this evolutionary form, since there is a suppression of recombination events at this region (Talbert and Henikoff 2010). Besides the centromeric location, PMSat is also present at (peri)centromeric, p-arm and telomeric regions of some chromosomes, and regardless of the high degree of interspecific and intraspecific sequence similarity of PMSat monomers, some nucleotide divergences were identified in specific sequence motifs (i.e. CENP-B box). In fact, the results of the *in silico* results on *P. maniculatus* genome sequencing data, points to the hypothesis that the more conserved PMSat monomers reside on the centromere and, while the monomers displaying sequence modifications locate in other locations, not so constrained in the occurrence of recombination processes.

Copy number changes of satDNA families have been correlated with karyotypic reorganization, where the amplification, deletion and/or intragenomic movement can promote chromosomal alterations and consequently, karyotype evolution (e.g. Ellingsen et al. 2007; Cazaux et al. 2013; Vittorazzi et al. 2014). Despite the initial reports that attribute the karyotype differences among *Peromyscus* species to the distinct repatterning of constitutive heterochromatin (CH) blocks (e.g. Romanenko et al. 2012), until the beginning of this work, no sequence information on the repetitive nature and content of these regions was known. Our results demonstrated that *Peromyscus* CH, mainly at the (peri)centromeric, p-arm and telomeric regions, are enriched in PMSat. It is accepted that chromosomal evolution in the *Peromyscus* genus was driven by progressive CH addictions, deletions and pericentric inversions (e.g. Louzada et al. 2015; Smalec et al. 2019); so, if these are mainly composed of PMSat, we can postulate that it were these satDNA family evolutionary molecular events the responsible for *Peromyscus* genome/or karyotpic evolution. In fact, copy number fluctuations (due to unequal crossing-over and rolling circle replication) can instigate chromosomal rearrangements, such as the pericentric inversions observed in the distinct karyotypes of the genus. In light of the different patterns of PMSat locations on CH regions and the distinct features of karyotype evolution, it seems that PMSat was originally only at the (peri)centromeric regions (as observed on *P. californicus*, that was mainly composed by acrocentric chromosomes) and acquired a more spread distribution during karyotype evolution in the other species (Figure V.1). This effect is notorious on *P. eremicus*, presenting the highest PMSat copy number *per* genome as large CH blocks that form the p-arms of all the autosomes.

Our first report of PMsat on *P. eremicus* genome (Louzada et al. 2015) caught the attention of the scientific community, and in the course of the work here presented, Smalec and colleagues (2019) also mapped and characterized PMSat satDNA in other *Peromyscus* species. As our results, the analysis based on next-generation sequencing reads databases of four distinct *Peromyscus* species (*P. maniculatus*, *P. californicus*, *P. leucopus* and *P. polionotus*) revealed a high degree of nucleotide sequence and size conservation between monomers of PMSat inter- and intra-*Peromyscus* species (Smalec et al. 2019). However, some discrepancies in the PMSat physical distribution findings were verified in some chromosomes of *P. maniculatus*, *P. leucopus* and *P. eremicus*, which can be a result of considerable number of intraspecific/intraindividual chromosome CH polymorphisms resulting in distinct cytotypes or chromosome races. It is also worth mentioning that Smalec and colleagues (2019) do not refer the standard karyotypes used in the organization of the species' karyotypes. Despite these differences, this study corroborates the evolution mode of PMSat by non-concerted evolutionary events and its involvement in *Peromyscus* karyotype variations (Smalec et al. 2019).

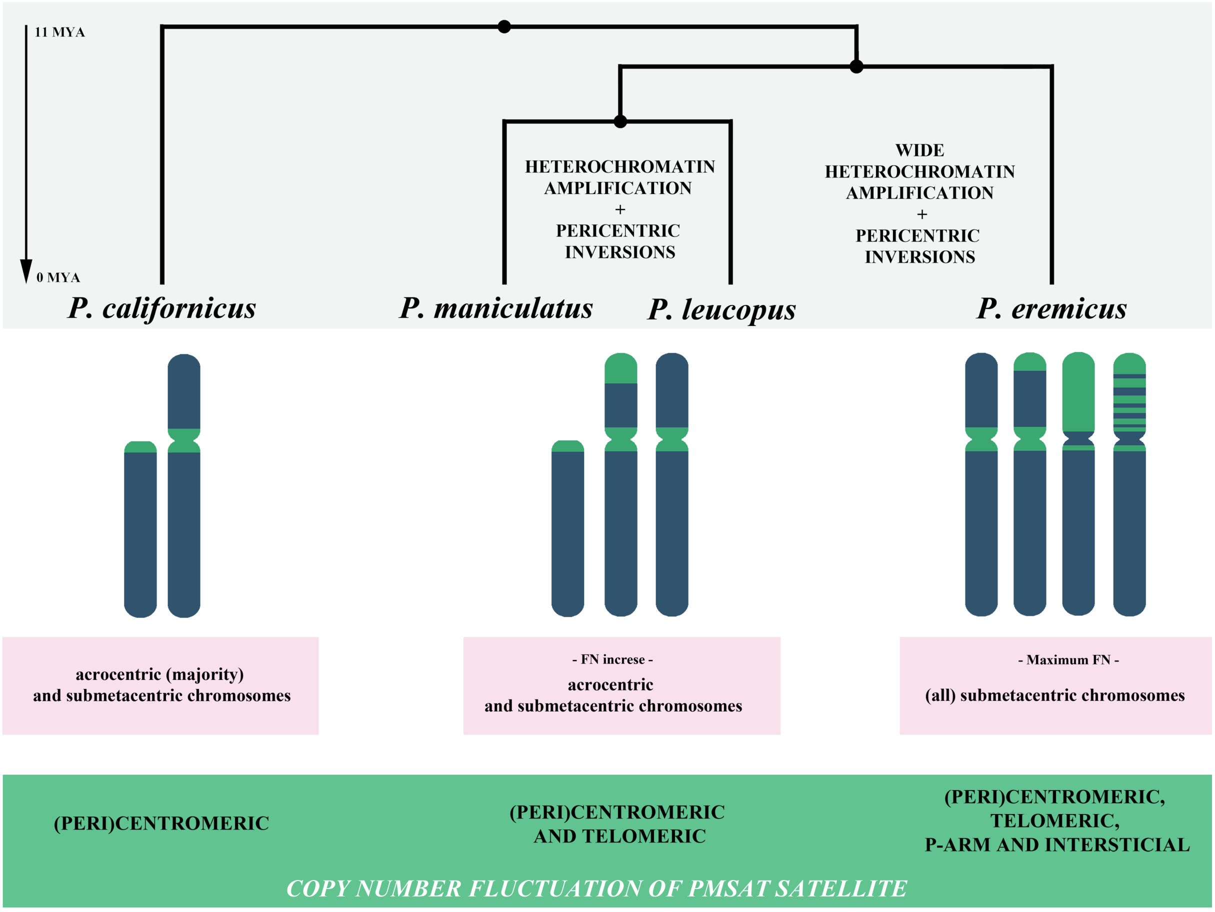

11 MYA
0 MYA
HETEROCHROMATIN AMPLIFICATION + PERICENTRIC INVERSIONS
WIDE HETEROCHROMATIN AMPLIFICATION + PERICENTRIC INVERSIONS
P. californicus
P. maniculatus
P. leucopus
P. eremicus
acrocentric (majority) and submetacentric chromosomes
- FN increse -
acrocentric and submetacentric chromosomes
- Maximum FN -
(all) submetacentric chromosomes
(PERI)CENTROMERIC
(PERI)CENTROMERIC AND TELOMERIC
(PERI)CENTROMERIC, TELOMERIC, P-ARM AND INTERSTICIAL
COPY NUMBER FLUCTUATION OF PMSAT SATELLITE

Figure V.1. Proposed model for PMSat evolution in the studied *Peromyscus* species. *Peromyscus* genus is characterized by a high degree of karyotypic conservation and all the Peromyscine species share a 2n = 48. The interspecific and intraspecific karyotypic variations reside in the number of chromosomal arms (fundamental number (FN) varies from 52 to 92). We propose that chromosomal rearrangements among *Peromyscus* chromosomes were promoted by PMSat and this sequence also evolved in a non-concerted evolutionary fashion. PMSat presents a high degree of interspecific and intraspecific sequence similarity, and the differences among *Peromyscus* species are attributed to the variation of PMSat copy number *per* genome; the amplification/contraction molecular mechanisms (such unequal crossing-over and rolling circle replication) leads to copy number fluctuations among species, resulting in PMSat repatterning from its original location (centromere; as observed in *P. californicus*) to other genomic locations (pericentromere, p-arm and telomere; as observed in other *Peromyscus* species).

In addition to PMSat, two additional satDNA families were identified for the first time on *P. maniculatus* genome, RNSAT1 and MMSAT4, that were previously identified on *Rattus norvegicus* and *Mus musculus* genomes (Ostromyshenskii et al. 2015; Rodentia database reports). This finding together with the lack of molecular and cytogenetic features of these two satDNA families on *Peromyscus* and non-*Peromyscus* genomes represents the gateway for the full characterization of other satDNA families in Peromyscine species. It is important to highlight that the bioinformatics strategy conducted in this work was designed to uncover repetitive sequences exhibiting a classical satDNA behaviour, i.e. with a tandem array organization fashion and high copy number. Recently, our group reported that the cat major satDNA (FA-SAT) constitute one of the most ancient satDNAs described so far, maintaining a high degree of conservation/similarity in several Bilateria species with distinct copy numbers *per* genome and genomic organization, including on *P. eremicus* genome with an interspersed distribution (Chaves et al. 2017). However, FA-SAT escaped our analysis most probably due to its low copy number. Meanwhile, it is also important to mention that *P. maniculatus* genome is not fully sequenced (and annotated) as well as certain repetitive regions may have been masked during the genome sequencing process.

V. 2. DISCLOSING THE ROLE OF PMSAT ncRNA

To get more insights into PMSat on *Peromyscus* genomes, the transcriptional activity of this satDNA family was also investigated (Chapter IV). The experimental achievements unveiled that PMSat is transcribed in proliferative cells of all the studied *Peromyscus* species with a positive correlation between PMSat expression and copy number content on each genome. Since PMSat presents a (peri)centromeric location with a high degree of sequence similarity between the centromeric and pericentromeric regions among all the studied *Peromyscus* genomes, it is difficult to distinguish and analyse specific PMSat variants for each chromosome region or species. Notwithstanding, as already referred, our data suggest that the main difference among species resides on PMSat copy number fluctuations and these variations seem to be the major contributor to PMSat expression variations. Besides the distinct abundance of transcripts between *P. eremicus* and *P. maniculatus* proliferative cells, a similar transcriptional cellular profile was detected throughout the cell cycle in both species' cells. The analysis of specific cell cycle phases revealed that PMSat satncRNA accumulates mostly at G2/M transition and at the mitosis onset. Our findings are in agreement with a cell cycle-dependent manner of transcription, which follows other centromeric satDNAs (Lu and Gilbert 2007; Ferri et al. 2009; Probst et al. 2010; Maison et al. 2011). In accordance with our results, also in mouse cells, the levels of centromeric MiSat RNAs greatly vary throughout the cell cycle presenting a maximum at G2/M phase (Ferri et al. 2009).

In order to unveil the putative function(s) of PMSat transcripts, a knockdown experiment was performed on *P. eremicus* proliferative cells. Interestingly, the depletion of PMSat RNA suggested a tendency for nuclear abnormalities, mainly the occurrence of aneuploidy phenotypes. These results anticipate the potential role of PMSat transcripts as key players on kinetochore assembly and function, in agreement with several others centromeric satncRNAs' reports (e.g. Ideue et al. 2014; Grenfell et al. 2016). In fact, it has been revealed that both the transcription of centromeric satDNAs by RNA polymerase II (RNApolII) process itself and the derived transcripts play important roles in chromatin remodelling/CENP-A deposition and in kinetochore assembly during mitosis (reviewed in Perea-Resa and Blower 2018). Moreover, centromeric RNAs seem to interact with the centromeric proteins CENP-A, CENP-B and CENP-C (e.g. Quénet and Dalal 2014; McNulty et al. 2017), as well as with Aurora-B, INCENP and Survivin, which are elements of the chromosome passenger complex (CPC) that is involved in proper chromosome segregation (e.g. Ferri et al. 2009; Ideue et al. 2014; Grenfell et al. 2016). PMSat monomers seems to keep a conserved motif related to the DNA-binding domain for the centromeric protein CENP-B (CENP-B box) and, crucially, are co-localized with the CENP-A protein that forms a stable complex with this motif on PMSat monomers, which proves its centromeric nature and anticipates its involvement in the centromeric function, both as a DNA sequence and as a satncRNA.

Although satDNA transcripts have been reported in a variety of cell conditions, such as cell proliferation, development and differentiation, cell aging, cell stresses and tumorigenesis (e.g. Ferreira et al. 2015), little is known about the mechanism of RNApolII acting in the centromere, or the transcription factors and binding partners involved in centromere transcription. Some satDNAs present binding sites for diverse transcription factors (Bulut-Karslioglu et al. 2012) and some of these were already identified under specific cell stress conditions, such the human SATIII overexpression after heat shock (Goenka et al. 2016). The *in silico* analysis of putative transcription factors binding sites in PMSat monomer sequences suggests that PMSat transcription can be conducted by RNApolII, as repressors/activators/ regulators of RNApolII were identified in our analysis. Also, the cellular pathways in which these putative transcription factors are involved are in accordance with the PMSat satncRNA involvement in a variety of cell conditions, namely in response to cellular stresses, cell development and differentiation.

V.3. CONCLUDING REMARKS

This work clearly reinforces the potential of genome-wide analysis on newly sequenced genomes for a global characterization of tandem repeats, specifically the satellitome fraction, which, in addition with experimental complementary techniques, allows not only its physical characterization at a chromosome level, assisting in the subsequent stages of sequencing projects (scaffold/contigs mapping), but also at the molecular and functional analysis level of satDNAs across genomes.

The search for repetitive sequences in the representative *P. maniculatus* genome sequencing project, revealed a highly homogenous and conserved satDNA – PMSat - the major constituent of this genome satellitome. The molecular and cytogenetic characterization of PMSat in several *Peromyscus* species (*P. californicus*, *P. maniculatus*, *P. leucopus* and *P. eremicus*) revealed that it must be involved in the centromeric function, as PMSat knockout preliminary experiments lead to aneuploidy phenotypes and in fact PMSat satncRNAs accumulate at G2 and in mitosis onset in both *P. eremicus* and *P. maniculatus.* Moreover, the presence of a conserved CENP-B box-like motif and the co-localization of the CENP-A protein on PMSat monomers emphasize the centromeric role of this satDNA. These findings are also supported by the analysis of PMSat orthologous sequences (found in non-*Peromyscus* species) that share a remarkable sequence similarity, which once again, anticipates its functional significance. Its presence not only at the active centromeres but also at the pericentromeric regions of these species chromosomes, comprising large PMSat blocks at, that further extend to the entire p-arms in some species chromosomes, in addition to its different representativeness in the analysed species, shows that PMSat satDNA family was potentially involved as trigger of the *Peromyscus* karyotype evolution, driven by chromosomal rearrangements promoted by copy number fluctuations of a “simple” satDNA.

Altogether, these results represent the preliminary study of PMSat satncRNAs on *Peromyscus* genomes, whose functions seem to follow the ones postulated for centromeric transcripts: key players on kinetochore assembly and function by association and modulation of centromeric proteins.

V.4. FUTURE PERSPECTIVES

The complex repetitive structure of the centromeric region with long arrays of near-identical satDNA sequences represents one of the major challenges for the correct genome assembly, as the (peri)centromeric regions are usually discarded during genome sequencing. Therefore, the centromere has persisted as the ultimate stronghold in genome annotation. Works as the one presented in this thesis, that discloses the repetitive sequences that "govern" each genome in newly sequenced species, represent a remarkable tool for a widespread understanding of genome organization, regulation, and evolution. The achievements regarding PMSat at the centromeres represent an excellent opportunity to complete the assembly and annotation of *P. maniculatus* centromeric regions. Interestingly, in addition to the *P. maniculatus* genome project that was focused on this work, a novel genome sequencing project (with unplaced contigs and at a chromosome level assembly – Pman_2.1, GenBank assembly accession GCA_003704035.1, BioProject_ PRJNA494228) was recently released by Harvard University/Howard Hughes Medical Institute. Thus, our research group initiates the comparison of the *in silico* genome wide analysis of TRs content between this new one genome sequencing project and the representative genome of *P. maniculatus* (presented in this thesis). Moreover, our results can also assist the ongoing genome sequencing projects of other *Peromyscus* species, namely, *P. californicus*, *P. leucopus* and *P. polionotus* (Baylor College of Medicine, www.hgsc.bcm.edu/peromyscus-genome-project). In the near future, the full characterization of other satDNAs also reported in this work (MMSAT4 and RNSAT1), will certainly contribute to decoding the satellitome of *Peromyscus* genomes.

Understanding the molecular mechanisms underlying satDNAs transcription (i.e., the transcriptional process and machinery and the satncRNAs-binding partners of each satellite DNA) as well as the pathways in which satncRNAs are involved is imperative for the clarification of the roles played by these important DNA sequences, and its transcripts in eukaryotic genomes.

V.4. REFERENCES

Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, Ryder OA, Purgato S, Zoli M, Della Valle G, Eichler EE, et al. 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 21(1):137–145. doi:10.1101/gr.111278.110.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580. doi:10.1093/nar/27.2.573.

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, et al. 2012. A transcription factor-based mechanism for mouse heterochromatin formation. Nat Struct Mol Biol. 19(10):1023–1030. doi:10.1038/nsmb.2382.

Cazaux B, Catalan J, Justy F, Escudé C, Desmarais E, Britton-Davidian J. 2013. Evolution of the structure and composition of house mouse satellite DNA sequences in the subgenus Mus (Rodentia: Muridea): a cytogenomic approach. Chromosoma. 122(3):209–220. doi:10.1007/s00412-013-0402-4.

Chaves R, Ferreira D, Mendes-da-Silva A, Meles S, Adega F. 2017. FA-SAT Is an Old Satellite DNA Frozen in Several Bilateria Genomes. Genome Biol Evol. 9(11):3073–3087. doi:10.1093/gbe/evx212.

Chaves R, Frönicke L, Guedes-Pinto H, Wienberg J. 2004. Multidirectional chromosome painting between the Hirola antelope (Damaliscus hunteri, Alcelaphini, Bovidae), sheep and human. Chromosome Res. 12(5):495–503. doi:10.1023/B:CHRO.0000034751.84769.4c.

Ellingsen A, Slamovits CH, Rossi MS. 2007. Sequence evolution of the major satellite DNA of the genus *Ctenomys* (Octodontidae, Rodentia). Gene. 392(1–2):283–290. doi:10.1016/j.gene.2007.01.013.

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. Chromosome Res. 23(3):479–493. doi:10.1007/s10577-015-9482-8.

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(15):5071–5080. doi:10.1093/nar/gkp529.

Fujita R, Otake K, Arimura Y, Horikoshi N, Miya Y, Shiga T, Osakabe A, Tachiwana H, Ohzeki J, Larionov V, et al. 2015. Stable complex formation of CENP-B with the CENP-A nucleosome. Nucleic Acids Res. 43(10):4909–4922. doi:10.1093/nar/gkv405.

Garrido-Ramos M. 2017. Satellite DNA: An Evolving Topic. Genes. 8(9):230. doi:10.3390/genes8090230.

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. Cell Sci. 129(19):3541–3552. doi:10.1242/jcs.189803.

Grenfell AW, Heald R, Strzelecka M. 2016. Mitotic noncoding RNA processing promotes kinetochore and spindle assembly in *Xenopus*. Cell Biol. 214(2):133–141. doi:10.1083/jcb.201604029.

Hartley G, O'Neill R. 2019. Centromere Repeats: Hidden Gems of the Genome. Genes. 10(3):223. doi:10.3390/genes10030223.

Ideue T, Cho Y, Nishimura K, Tani T. 2014. Involvement of satellite I noncoding RNA in regulation of chromosome segregation. Genes to Cells. 19(6):528–538. doi:10.1111/gtc.12149.

Komissarov AS, Gavrilova EV, Demin SJ, Ishov AM, Podgornaya OI. 2011. Tandemly repeated DNA families in the mouse genome. BMC genomics. 12(1):531. doi:10.1186/1471-2164-12-531.

Louzada S, Vieira-da-Silva A, Mendes-da-Silva A, Kubickova S, Rubes J, Adega F, Chaves R. 2015. A novel satellite DNA sequence in the *Peromyscus* genome (PMSat): Evolution via copy number fluctuation. Mol Phyl Evol. 92:193–203. doi:10.1016/j.ympev.2015.06.008.

Lu J, Gilbert DM. 2007. Proliferation-dependent and cell cycle–regulated transcription of mouse pericentric heterochromatin. Cell Biol. 179(3):411–421. doi:10.1083/jcb.200706176.

Maison C, Bailly D, Roche D, de Oca RM, Probst AV, Vassias I, Dingli F, Lombard B, Loew D, Quivy J-P, et al. 2011. SUMOylation promotes de novo targeting of HP1α to pericentric heterochromatin. Nature Genet. 43(3):220–227. doi:10.1038/ng.765.

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. Develop Cell. 42(3):226-240.e6. doi:10.1016/j.devcel.2017.07.001.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14(1):R10.

Ostromyshenskii DI, Kuznetsova IS, Komissarov AS, Kartavtseva IV, Podgornaya OI. 2015. Tandem repeats in the rodent genome and their mapping. Cell Tiss Biol. 9(3):217–225. doi:10.1134/S1990519X15030116.

Perea-Resa C, Blower MD. 2018. Centromere Biology: transcription goes on stage. Mol Cel Biol. MCB.00263-18. doi:10.1128/MCB.00263-18.

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

Plohl M. 2010. Those mysterious sequences of satellite DNAs. Periodicum biologorum. 112(4):403–410.

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA evolution. In: Repetitive DNA. Vol. 7. Karger Publishers. p. 126–152.

Plohl M, Petrović V, Luchetti A, Ricci A, Šatović E, Passamonti M, Mantovani B. 2010. Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. Heredity. 104(6):543–551. doi:10.1038/hdy.2009.141.

Probst AlineV, 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. Develop Cell. 19(4):625–638. doi:10.1016/j.devcel.2010.09.002.

Quénet D, Dalal Y. 2014. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. eLife. 3. doi:10.7554/eLife.03254.

Romanenko SA, Perelman PL, Trifonov VA, Graphodatsky AS. 2012. Chromosomal evolution in Rodentia. Heredity. 108(1):4–16. doi:10.1038/hdy.2011.110.

Silva DMZ de A, Utsunomia R, Ruiz-Ruano FJ, Daniel SN, Porto-Foresti F, Hashimoto DT, Oliveira C, Camacho JPM, Foresti F. 2017. High-throughput analysis unveils a highly shared satellite DNA library among three species of fish genus Astyanax. Sci Reports. 7(1). doi:10. 1038/s41598-017-12939-7.

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. Chromosome Res. doi:10.1007/s10577-019-09605-1.

Talbert PB, Henikoff S. 2010. Centromeres Convert but Don't Cross. PLoS Biol. 8(3):e1000326. doi:10.1371/journal.pbio.1000326.

Vittorazzi SE, Lourenço LB, Recco-Pimentel SM. 2014. Long-time evolution and highly dynamic satellite DNA in leptodactylid and hylodid frogs. BMC genetics. 15(1):111.