UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Exploring nanopore long reads and whole-genome sequencing data to characterize long tandem repeats and chromosome structure in mammalian genomes

Daniel Alexandre Varela Marques Eleutério

Mestrado em Bioquímica e Biomedicina

Dissertação orientada por:
Professora Doutora Margarida Gama Carvalho
Doutor Ricardo Dias

2023

Acknowledgements

First and most notably I want to thank my professor Margarida Gama Carvalho for her great patience and dedication in dealing with me through all these years.

I also want to thank Mariana Lopes, Sandra Louzada and Raquel Chaves from the CytoGenomics Lab (UTAD) for providing me with the opportunity to participate in the project regarding the HSat1A study. This project also led to a scientific publication, indicated in reference: (Lopes et al., 2023).

I want to thank Mariana Nascimento and Ricardo Dias for supplying their time and effort in teaching me to use Nanopore sequencing technology.

Big thanks to all the people in the RNA Systems Biology group that helped me with the little and big questions, shared their compassion and knowledge.

Thank you to some of my friends that always provided me with care and attention: Raquel Teixeira, for sharing some of the wildest stories out there; Noémi Velez, for being my “comadre”; André Salvada, for your compassion and care of both my mental and physical health; Afonso Bravo, for providing a great guidance into the future; Diogo Tomaz, for your compliance with the weirdness and impartial outlook on life; Nuno Ribeiro and João Raposo, for being my great childhood friends, and I hope that I can make you proud.

Special thank you to Larissa Godoi, for loving me through this crazy and long journey.

Lastly and most importantly, I want to thank my father, my mother and sister, that always supported, cared for, provided, and loved me through all these years, and that helped me reach my goal. Thank you so much.

Abstract

Long tandem repeats (LTRs) sequences, namely satellite DNA (satDNA), play a critical role in genome organization and stability. Their detection and characterization still represent a challenge, which long-read Nanopore sequencing is helping to simplify. Thus, there is a growing interest in developing bioinformatic tools for the analysis of LTRs and in understanding their role in genome evolution. The main aim of this work was the characterization of LTRs in mammalian genomes, using WGS data from two model organisms: *Peromyscus* genus and *Homo sapiens*. In the first case, the focus was to characterize PMSat, an evolutionarily conserved satDNA. In the second case, the key objective was to perform a comparative assessment of LTRs between different human long-read genome datasets, to characterize in detail the human satellite HSat1A and its transcripts, and to establish methods to analyze satDNA involvement in Robertsonian translocations (specifically the t(14;21)).

The analysis performed identified PMSat as the most abundant LTR of the *Peromyscus* genus, in accordance with previous cytogenetic studies, with GC-content, monomer and array length exhibiting little variation. Moreover, a tendency in the positioning in *Peromyscus* genus chromosomes was identified. In the study of the three WGS Nanopore human datasets, the LTRs ALR, BSR, HSat2 and HSat1A were identified as the most abundant. It was possible to demonstrate the potential of chromosome sorting to facilitate the analysis of translocated chromosomes. However, low sequencing yield and DNA fragmentation did not allow for characterization of the target region. Finally, LTRs of HSat1A were shown to have a predominance of 9-mer monomers, and to express polyadenylated transcripts of varying lengths, probably resulting from alternative polyadenylation.

This study represents a relevant contribution to understanding the role of LTRs in genome organization and evolution.

Keywords: LTR; satellite DNA; Nanopore sequencing; PMSat; HSat1A.

Resumo Alargado

Sequências repetitivas encontradas em regiões peri/centroméricas têm demonstrado um papel crucial na organização e estabilidade genómica. Estas regiões são compostas por repetições longas em tandem (LTRs). Incluídas neste grupo estão os DNA satélite (satDNA), que são caracterizados por montagens de matrizes de sequências repetitivas de cabeça-a-cauda. PMSat, um satDNA presente no género *Peromyscus*, embora demonstrando variação do número de cópias e localização, preserva a sequência e estrutura, desafiando as previsões de impulso molecular. Portanto, este género demonstrou um grande potencial para o estudo de LTRs em regiões peri/centroméricas, pois as espécies deste género demonstram ter rearranjos cromossómicos enquanto mantêm um número estável de 48 cromossomas.

No genoma humano existem várias famílias de DNA satélite humano, que são compostas por uma grande variedade de populações e qualidades. Embora existam estratégias recorrentes para detecção e caracterização de LTRs, nomeadamente satDNA, estas são baseadas em digestão de DNA genómico com endonucleases de restrição, seguida de análise de sequências de monómeros ou multímeros curtos aleatoriamente clonados. Isto dificulta a análise destes elementos repetitivos devido às suas curtas dimensões. No entanto, novas tecnologias de sequenciação (NGS) têm emergindo, como a tecnologia de sequenciação por nanoporos, que assistirá a esclarecer a tendência e selectividade baseada em sequência dos LTRs. No estudo de LTRs, é também relevante entender o impacto de satDNA a nível transcricional. Embora a família de HSat1A seja a menos abundante dos DNA satélites humanos, esforços iniciais demonstraram que estas sequências são activamente transcritas, enquanto também demonstram várias características que são consideradas interessantes para investigações adicionais relacionadas à actividade e função em regiões peri/centroméricas deste satDNA.

O objectivo principal deste trabalho foi realizar uma caracterização aprofundada de LTRs de mamíferos recorrendo a dados de sequenciação de genomas completos, com foco em dois organismos modelo: género *Peromyscus* e *Homo sapiens*. Isto implica identificar a abundância e diversidade de LTRs nos genomas, determinando as suas dimensões e número de cópias, caracterizando a suas variações em sequências e mapeando a sua localização cromossómica quando possível. Especificamente, este trabalho teve como objectivo avaliar LTRs específicos através de dados de sequenciação disponíveis de quatro espécies do género de *Peromyscus*, com intuito também de determinar as suas posições relativas numa comparação de cariótipo/ideograma entre as espécies. A medição do conteúdo GC, dimensões de monómeros e matrizes do PMSat também foi relevante para uma melhor descrição das matrizes de repetição deste satDNA dentro deste género. Também foi relevante para este trabalho a avaliação do impacto dos diferentes métodos de sequenciação na detecção e quantificação de LTRs. As LTRs têm-se demonstrado difíceis de detector e caracterizar devidamente devido às suas sequências contíguas altamente repetitivas. Com isso em mente, outro objectivo específico deste trabalho foi o aperfeiçoamento da caracterização da diversidade de matrizes repetitivas no genoma humano, tirando proveito de longas e contíguas leituras produzidas por tecnologia de sequenciação por nanoporos. Este estudo teve também como propósito a construção do cromossoma translocado Robertsoniano t(14;21) em células humanas, para uma melhor caracterização de LTRs envolvidos nesta translocação. Foi também importante para este trabalho a medição do número de cópias genómicas e da diversidade de produtos transcriptómicos do satélite humano HSat1A, com o intuito de revelar o panorama de transcrição deste satDNA.

A identificação de LTRs foi realizada com a ferramenta Tandem Repeat Finder, enquanto a caracterização foi desenvolvida com a ferramenta blastclust, um algoritmo de *clustering* baseado em sequência. Os métodos de sequenciação demonstraram algumas variações na detecção e quantificação de LTRs entre espécies. Todavia, ainda foi possível determinar que PMSat é o LTR mais abundante entre o género *Peromyscus*. O conteúdo GC dos monómeros de PMSat apresentaram uma variação entre 40-48% entre as quatro espécies seleccionadas. A dimensão dos monómeros de PMSat variam entre 340-345 bp entre as quatro espécies, enquanto a dimensão das matrizes demonstrou maior variação entre as espécies, no intervalo de 2,000-40,000 bp. Alguns LTRs foram identificados pela primeira vez, no entanto, apenas o LTR_PMAN_1 e o LTR_PMAN_2 apresentaram elevada abundância em comparação com outras novas classes identificadas. Localização relativa das matrizes de LTRs permitiram uma melhor visualização da interligação das classes de LTRs entre os cromossomas das espécies do género *Peromyscus*, permitindo também demonstrar a impulsão evolutiva dos LTRs, como a classe PMSat apresentando selectividade a nível cromossómico e de posição. Na detecção e quantificação de LTRs através de longas e contíguas leituras por nanoporos no genoma humano, dois conjuntos de dados de sequenciação do genoma humano foram usados, o NA12878 e o CHM13. Adicionalmente, foi usado neste estudo a linhagem celular GM03417, uma linhagem humana específica contendo uma translocação Robertsoniana t(14;21), para comparação com os resultados obtidos com esses conjuntos de dados de sequenciação. As quatro primeiras classes de satDNA agrupadas com mais matrizes foram as mesmas em todos os conjuntos de dados de sequenciação, com a ordem descendente ALR>BSR>HSat2>HSat1A. Para além disso, foi possível a identificação de seis novas classes de LTRs nos três conjuntos de dados de sequenciação, as quais estavam contidas nos primeiros 20 grupos de LTRs com mais matrizes. A técnica de *flow sorting* de cromossomas foi usada em conjunto com a sequenciação por nanopore para a construção do cromossoma translocado Robertsoniano t(14;21) em células humanas. No entanto, este método não produziu quantidade suficiente de dados para a sua análise. Para além disso, através dos dados obtidos foi possível verificar uma fragmentação de DNA significativa antes da preparação da biblioteca de sequenciação, o que poderá indicar que este método não seja o mais eficiente para o estudo de sequências repetitivas. Os dados de sequenciação de NA12878 foram usados para analisar o número de cópias genómicas de HSat1A, enquanto uma 5'/3' RACE (rápida amplificação de extremidades de cDNA) juntamente com sequenciação de elevado rendimento foi realizada para o estudo da diversidade dos produtos transcriptómicos também de HSat1A. As matrizes de HSat1A a nível genómico demonstram uma predominância para a dimensão de 42 bp, que é a sua forma de monómero único. No entanto, demonstra também uma predominância para uma organização de uma ordem superior (HOR) de repetições juntas em tandem (NTRs) 9-mer. Os produtos de 3' RACE demonstraram estar poliadenilados, indicando a ocorrência de transcrição, possivelmente pela RNA polimerase II. A montagem dos transcritos de HSat1A apresentou uma distribuição entre 51 e ~400 nucleótidos, demonstrando também picos correspondentes a múltiplos da dimensão do monómero de 42 bp. Elevada variabilidade de transcritos poliadenilados de HSat1A indica uma regulação complexa da poliadenilação alternativa deste satDNA.

Especificamente, no estudo de LTRs no género *Peromyscus* verificou-se padrões e relações que permitiram uma melhor caracterização de prévias e novas classes de LTRs identificadas. O posicionamento relativo da família satDNA mais abundante do género *Peromyscus*, PMSat, juntamente com outras classes de LTRs demonstraram uma preferência posicional e cromossómica destes elementos repetitivos. A tecnologia de sequenciação por nanoporos também demonstrou um grande potencial para a classificação e quantificação de LTRs no genoma humano. No entanto, alguns desafios na combinação desta tecnologia com *flow sorting* de cromossomas foram identificados na detecção e caracterização de LTRs. Algumas

novas classes de LTRs foram identificados no genoma humano. A caracterização da família de HSat1A revelou não só ser activamente transcrita, como também apresentar um elevado grau de variabilidade entre as sequências genómicas e trancriptómicas. Em geral, esta análise contribuiu para uma caracterização melhorada dos LTRs, com implicações na organização e evolução genómica. Contudo, investigação adicional e estudos funcionais são de devera importância para esclarecer funções e mecanismos específicos destes elementos repetitivos.

Palavras-chave: LTR; DNA satélite; Sequenciação Nanopore; PMSat; HSat1A.

Contents

List of Figures

List of Tables

List of Abbreviations/Acronyms

3′ RACE	**3’ R**apid **A**mplification of **c**DNA **E**nds
ACRO1	Human **acro**meric satellite **1** DNA
ALR	Human alpha satellite DNA [**al**pha **R**epetitive DNA]
APA	**A**lternative **P**oly**a**denylation
B1_Mus1/2	**B1** SINE **Mus** musculus **1/2** DNA
BSR	Human beta satellite DNA [**b**eta **S**atellite **R**epeat]
CER	D22Z3 satellite DNA [**Ce**ntromeric **R**epeat]
dsDNA	**d**ouble-**s**tranded **DNA**
GSatII	Gamma Satellite II DNA
HSat	**H**uman **Sat**ellite
HSat1A (SAR)	**H**uman **Sat**ellite **1A** (SAR) DNA
HSat2/3	**H**uman **Sat**ellite 2/3 DNA
HSat4	**H**uman **Sat**ellite 4 DNA
L1_Mur2_orf2	**L1** Retrotransposon **Mur**idae **2 orf2** DNA
L1_Mur3_orf2	**L1** Retrotransposon **Mur**idae **3 orf2** DNA
lncRNA	**L**ong **n**on**c**oding **RNA**
LTR	**L**arge **T**andem **R**epeat
Lx5c_3end	Retrotransposon in murids - 3' UTR DNA
MMSAT4	***Mus Musculus* Sat**ellite **4** DNA
MurSatRep1	**Mur**idae **Sat**ellite **Rep**eat **1** DNA
ncDNA	**n**on-**c**oding **DNA**
NGS	**N**ext-**G**eneration **S**equencing
NTR	**N**ested **T**andem **R**epeat
ONT	**O**xford **N**anopore **T**echnology
PAS	**P**oly**a**denylation **S**ignal
PCAL	***P**eromyscus **cal**ifornicus*
PCPA	**P**remature 3'-end **c**leavage and **p**oly**a**denylation
PERE	***P**eromyscus **ere**micus*
PLEU	***P**eromyscus **leu**copus*

PMAN	***P****eromyscus* ***man****iculatus*
PMSat	***P****eromyscus* ***m****aniculatus* **Sat**ellite
RMER1C	**RMER1C** subfamily of L1-dependent DNA
ROB	**Rob**ertsonian translocation
ssDNA	**s**ingle-**s**tranded **DNA**
SATR1	**Sat**ellite-like **R**epeat **1** DNA
snRNP	**S**mall **n**uclear **r**ibo**n**ucleo**p**roteins
SST1	Large human satellite identified with **SstI** enzyme DNA
T2T	**T**elomere-**2**-**T**elomere
TE	**T**ransposable **E**lement
TRF	**T**andem **R**epeat **F**inder
WGS	**W**hole **G**enome **S**equencing

1. Introduction

This section provides an introductory summary of fundamental concepts regarding peri/centromeric regions and their components, as well as depicting general insights of nanopore sequencing technology and the objectives of this dissertation.

1.1 Centromere complexity and its multifaceted components

Centromeres play an important role in chromosome segregation of genetic material in both mitosis and meiosis. They represent the domain that defines the primary constriction of each eukaryotic metaphase chromosome. Other functions associated with centromeres are the recognition of homologous chromosomes during pairing in meiosis, attachment of sister chromatids before or until mitotic anaphase and during the first meiotic division, nucleation of the kinetochore apparatus, and cell cycle checkpoint control to regulate the metaphase-anaphase transition (Lee et al., 1997).

The peri/centromeric regions are composed of long tandem repeats (LTRs), mainly satellite DNA, which are highly repetitive sequences of DNA (Plohl et al., 2008). The satellite DNA sequences have been fundamental for the study of karyotypic evolution in eukaryotes (Adega et al., 2006; Gong et al., 2012; Kuhn et al., 2008; Langdon et al., 2000; Vondrak et al., 2020), understanding the foundation of abnormal centric fusion (Barra & Fachinetti, 2018; Lin & Davidson, 1974), defining the ancestry and evolution of satellite DNA families (Melters et al., 2013), and the distinction of homologous chromosomes (M. E. Aldrup-MacDonald et al., 2016). Other types of LTRs have also shown to be highly represented in centromeres, such as the case of some heterogeneous or transposable element (TE)-related superfamilies, however, these can be positioned anywhere on the genome (Komissarov et al., 2011; Sulovari et al., 2019). In some circumstances, noncoding LTRs have been associated with disease mechanisms (DeJesus-Hernandez et al., 2011), with evolutionary drive of speciation in eukaryotes (Raskina et al., 2008; Sulovari et al., 2019) and with transcriptomics implications on chromatin organization (Isiktas et al., 2022; Trigiante et al., 2021). Moreover, LTRs are also involved in structural chromosomal rearrangements, such as Robertsonian translocations (ROBs), which specifically involves the fusion of two acrocentric chromosomes at their centromeric regions, resulting in a single, larger chromosome (Poot & Hochstenbach, 2021). Even though some functions from these LTRs have been established, there are still unresolved cases due to constrictions regarding their highly repetitive contiguous sequences (Tørresen et al., 2019).

As stated before, satellite DNA is the main component of peri/centromeric regions and, subsequently, most of the focus will be on these sequences. This element is characterized by highly repeated segments of non-coding DNA (ncDNA) sequences, assembled into large tandem arrays of head-to-tail repeats in the constitutive heterochromatin, the eukaryotic chromosomal regions that stay condensed through the cell cycle (Plohl et al., 2008). The monomers of these sequences, which are the basic repeating units, are mainly composed of A+T and vary from a few base pairs (bp) to up more than 1 kilobases (kb) and can reach up to 100 Megabases (Mb) on array length (Plohl et al., 2008). The propensity of monomer satellite length seems to be around 150-180 bp and 300-360 bp, as it has been verified in satellites from both animals and plants. The explanation for this has been associated with the DNA length requirement of

wrapping of one or two nucleosomes (Henikoff et al., 2001). Contribution of satellite DNA to total genomic content has a significant impact on variability among species, as these sequences can sometime exceed 50% of total DNA. Consequently, eukaryotes acquire vast variation in genome size due to these sequences (Plohl et al., 2008). As a result of this great diversity of satellite DNA in sequence composition, genomic abundance, intricacy, and the existence of several families of satellite DNA that are unrelated, some potential functions from these sequences in the genome are still unknown. To this extent, though new data is still being acquired, connections between mechanisms and functions of theses satellite DNA are still very reliant on deduction rather than to experimental evidence (Plohl et al., 2008).

Human satellite DNA families comprise a wide range of sequence types and functions. The most abundant group is the centromeric alpha (α) satellite DNA family, influencing genome stability and recruitment of centromere and kinetochore proteins (McNulty & Sullivan, 2018). Other members are the beta (β) and gamma (γ) satellite DNA families, with the first family having shown to potentially originate from horizontal transfers in primates (Yang et al., 2020), while the latter being considered to control heterochromatin expansion in chromosomal arms (Lee et al., 1997). Additionally, human satellite DNAs HSat1 and HSat2/3 are also very important families represented in peri/centromeric regions (Prosser et al., 1986). HSat3 has shown to be located in the Robertsonian translocation (ROB) breakpoint of chromosomes 14 and 21 (Earle et al., 1992). Although HSat1 is the least abundant family of human satellite DNAs, early efforts demonstrate these sequences are transcriptionally active, while also showing several characteristics that are deemed to be interesting for further investigation regarding satellite DNA activity and function on peri/centromeric regions (Lopes et al., 2023). As aforementioned, satellite DNA transcription has been confirmed to have major implications on heterochromatin structure and centromeric protein recruitment, however, regarding this HSat1 family there are still many mechanisms and functions that are unknown, which is mainly due to their highly repetitive contiguous sequences.

Current strategies for LTR detection and characterization, mainly for satellite DNA, are based on genomic DNA digestion with restriction endonucleases, followed by sequence analysis of randomly cloned monomers or short multimers (Garrido-Ramos, 2017). However, new next-generation sequencing (NGS) technologies have been emerging, such as nanopore sequencing, which will assist in further elucidating the sequence-based arrangements and positioning of LTRs in chromosomes.

1.2 Advancing Long Tandem Repeat (LTR) analysis through Nanopore Sequencing Technology

Oxford Nanopore Technology (ONT) is an emerging sequencing method with promising potential to solve some of the current issues with assembly of large repetitive sequences, such as satellite DNA. This technology is based on sequencers using nanopore biosensors. In general, these biosensors can be classified into two main categories: solid-state nanopores that can be fabricated from several materials by usage of semiconductor production processes, that permit a wide variety of experimental conditions; and biological nanopores, consisting of genetically engineered transmembrane protein channels, that are embedded in a matrix. Currently, the biosensor used in ONT is a biological nanopore based on mutants of the Curlin sigma S-dependent growth (CsgG) nanopore from *Escherichia coli* (Magi et al., 2017; Y. Wang et al., 2021). CsgG is a secretion channel, implicated in curli formation. Curli are functional amyloid fibers that can be found

in the extracellular matrix of biofilms formed by some bacteria, such as α-*Proteobacteria* and γ-*Proteobacteria* (Goyal et al., 2014).

Library preparation strategies for nanopore sequencing can either be: 1D, where each strand is ligated with an adapter and pre-loaded motor proteins on both ends of the molecule, and sequenced independently; 2D, where the leader adapter guides fragments of double-stranded DNA (dsDNA) to the nanopore vicinity, and it starts the sequencing process by unzipping the dsDNA, enabling the template strand to pass through the nanopore, while at the end at the template strand, a hairpin adapter with pre-motor proteins is also sequenced with the complement strand; or $1D^2$, where each strand is ligated at both ends with a special adapter, increasing the probability that one strand will immediately be captured by the same nanopore following sequencing of the other strand of dsDNA (**Figure 1A**) (Goodwin et al., 2016; Magi et al., 2017; Y. Wang et al., 2021). Concomitantly, as the molecule of single-stranded DNA passes through the nanoscopic pore, a sensor measures ionic current shifts with a constant sampling frequency, with the additional shift between template and complement strand being recognized by the pore via the specific signal generated by an apurinic/apyrimidinic site located in the hairpin in the case of the 2D sequencing strategy (**Figure 1B**) (Magi et al., 2017). Data captured from raw current is then subjected to base calling with a machine learning approach (such as recurrent neural network, RNN, or hidden Markov Model) (**Figure 1C**) to obtain a consensus sequence for the template only (1D) or both template and complement strands separately (2D) (**Figure 1D**) (Magi et al., 2017). 2D and $1D^2$ sequencing strategies have an increased accuracy detection compared to 1D sequencing strategy, however, with the improvement of new base-calling algorithms, 1D read accuracy has also improved (Y. Wang et al., 2021).

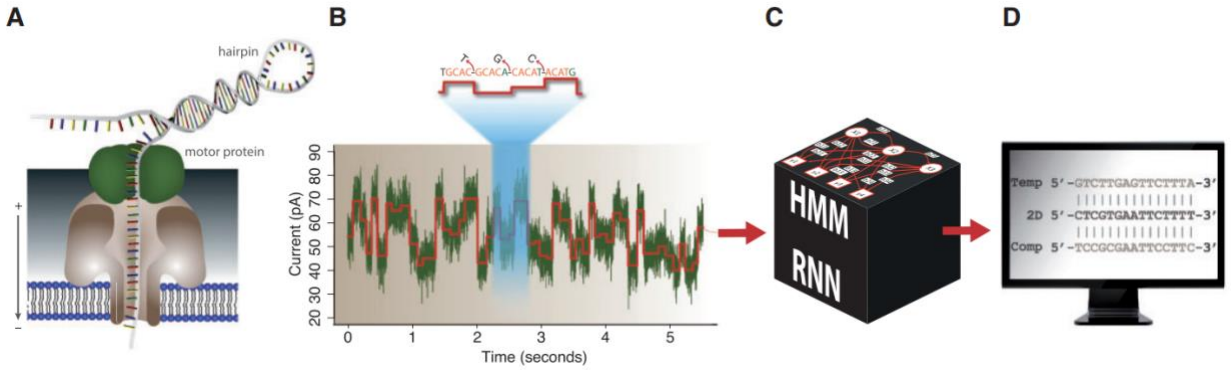


Figure 1 – The Oxford Nanopore sequencing process and performance. The motor protein unwinds the dsDNA allowing ssDNA to pass through the pore, while a sensor measures ionic current shifts (**A**) with a constant sampling frequency (at present 5000 Hz). Raw ionic current signals are segmented into discrete 'events' summarized by mean, SD and length (**B**). Segmented events are then analyzed with machine learning approach (black box, **C**) that outputs the sequence of template and complement signals (**D**) (Magi et al., 2017).

Application of these library preparation strategies using nanopore sequencing technology have been used in emerging projects, such as the Telomere-2-Telomere (T2T) consortium, which has solved major gaps in the human genome is the (Nurk et al., 2022). Their efforts have managed to correctly identify and quantify the sequences in the major peri/centromeric gaps of the human genome, while further driving the study of these regions (Altemose, 2022; Logsdon & Eichler, 2023; Lopes et al., 2023).

1.3 Satellite DNA: Exploring its role in centromere structure, chromosomal rearrangements, and evolution

The structure of the centromere is complex and involves a variety of molecular components, with satellite DNA playing a prominent role in defining its structure and function (Malik & Henikoff, 2009). Satellite DNA is involved in the formation of kinetochores, which are essential for the proper segregation of chromosomes during cell division (Shatskikh et al., 2020). It also has been associated with regulatory intron-mediated mechanisms that are important for gene expression (Fingerhut & Yamashita, 2022), and determines chromatin organization (Brändle et al., 2022; Thakur et al., 2021). Furthermore, satellite DNA strongly influences equilocality of centromeres in chromosomes, which refers to the tendency of centromeres to occupy fixed positions relative to other genetic elements on chromosomes (John et al., 1985; Ruiz-Ruano et al., 2016) and is thought to play a role in the evolution of karyotypes in eukaryotes (Adega et al., 2009; Smalec et al., 2019; K. Wang et al., 2022; Zattera et al., 2020). These sequences also play a crucial role in maintaining the structural and functional integrity of chromosomes (Plohl et al., 2008). The movement of a chromosomal locus as a result of rearrangements into the satellite DNA region can lead to a change in the activity of genes located in that locus because of the "position effect" (Zhimulev & Belyaeva, 2003). Moreover, rearrangement of satellite DNA can lead to genome instability, which can result in altered gene expression patterns, due to changes in the position of genes relative to regulatory elements (Fournier et al., 2010). The repetitive nature of the satellite DNA sequences makes it susceptible to recombination during cell division, which can lead to structural abnormalities in chromosomes (Giunta & Funabiki, 2017). Rearrangement of these sequences may also confer susceptibility to ROBs, which could lead to chromosomal disorders such as Down syndrome (Bandyopadhyay et al., 2001), although the mechanisms for this are still unknown. Additionally, satellite DNA structures can suffer rearrangements due to transposable elements (TEs) and other repetitive sequences being inserted into adjacent regions (Klein & O'Neill, 2018). Moreover, due to their tandemly repetitive nature, satellite DNA can undergo rapid evolution and contribute to chromosomal rearrangements, with the occurrence of events such as translocations and inversions (Louzada et al., 2020). Concerning specifically ROBs, these are a common type of chromosomal rearrangement that occurs between acrocentric chromosomes that involve the fusion of two chromosomes at their centromeric regions, resulting in a single, larger chromosome. Satellite DNA is often found at the breakpoints of these translocations, which could be an indicator of a crucial role of these sequences regarding these events (Altemose et al., 2014; Puppo et al., 2020).

Chromosome rearrangements in peri/centromeric regions may drive speciation events by contributing to reproductive isolation, but the relationship between the evolution of heterochromatic sequences and the karyotypic dynamics of these regions is poorly understood. In the order Rodentia of the genus *Peromyscus*, a single conserved satellite DNA sequence (PMSat) is located in recurrent sites of chromosome rearrangements and heterochromatic amplifications (Louzada et al., 2015). Despite wide variation in the copy number and location of repeat blocks among different species, PMSat maintains a conserved sequence and homogenized tandem repeat structure, defying predictions of molecular drive (Smalec et al., 2019). The conservation of this satellite monomer results in common, abundant, and large blocks of chromatin that are homologous among chromosomes within one species and among diverged species. The study of other LTRs in these species and including PMSat, can provide insights into molecular drives behind their wide range of ecological adaptations and on chromosomal evolution.

1.4 Strategies for tackling repetitive sequences in genome assembly

Repetitive sequences have always represented a great challenge for correct assembly and copy number assessment, especially in LTRs, which can reach massive array lengths (Plohl et al., 2008). In accordance with this, several methodologies have been fundamental for tandem repeat analysis. The most common tool employed for this analysis is the Tandem Repeats Finder (TRF) (Benson, 1999). This tool allows for the detection, quantification and establishing monomer consensus of tandem repeats located in provided DNA sequences. Additionally, RepeatMasker (developed by A.F.A. Smit, R. Hubley, and P. Green; see http://www.repeatmasker.org/) was also designed with the intent to identify and annotate repetitive elements in DNA sequences, and similarly contributed immensely for tandem repeat analysis. Nevertheless, research studies to detect and classify these tandem repeats are highly dependent on specific parameters, such as selecting for minimum monomer length and non-overlapping LTRs, in order to retrieve the most probable LTR correspondent to its detected location (Easterling et al., 2020; Ummat & Bashir, 2014).

Some concepts are of great importance to understand the process of LTR detection and characterization. As mentioned before, sequencing is the first step of the analysis, which produces reads. The reads are the primary unit of sequencing, as they constitute the raw sequences retrieved from the organism DNA. To improve the study of DNA sequences, reads are then assembled at high confidence into a contiguous fragment, the contig. Assembly of contigs is performed by overlapping these reads obtained from sequencing and then merging them together to retrieve an elongated and/or a more accurate consensus sequence. When forming these contigs, there are several regions that may contain highly repetitive or difficult sections between them. As such, when linking several contigs together these then contain gaps between them. These linked contigs are known as scaffold. There are some methods that provide enough evidence to support that a set of contigs are positioned at a certain distance from each other, such as paired-end or mate-pair reads that bridge two contigs (McCartney et al., 2019). Reads obtained from sequencing allow assessment of the sequencing coverage of the technology used. Sequencing coverage was described as LN/G, where L is the average read length, N is the number of reads, and G is the haploid genome length (e.g., an average read length of 6,200 bp, from $3x10^7$ reads for the human genome of 3.1Gb would give an estimated coverage of 60x) (Sims et al., 2014). This sequencing coverage is assessed for each genome assembly. A genome assembly is the reconstruction of a complete or near-complete representation of the genome of an organism using DNA sequencing data. It involves piecing together shorter DNA fragments (reads) obtained through sequencing into longer contiguous sequences (contigs/scaffolds) or complete chromosomes. An assembly can be composed of only contigs, only unplaced scaffolds, only chromosomes, or a combination of the three (Baker, 2012).

Even though downstream computational analysis has shown great improvement, there are still some challenges to consider. The study of translocated chromosomes focusing on repetitive sequences cannot be performed in an efficient manner through whole genome sequencing approaches, as this would require the generation of a huge amount of data that is irrelevant for the question to be addressed to ensure appropriate coverage of the target region. Therefore, other methods need to be deployed in order to facilitate sufficient sequencing data collection for specific translocated chromosomes. A relevant approach in this context is the flow-sorting technique, that can be used to isolate specific chromosomes from a population of cells, and then subject them to sequencing (Doležel et al., 2021; Kuderna et al., 2019). By separating individual chromosomes, flow-sorting can greatly reduce the complexity of the DNA sample and increase the

sequencing coverage of the target chromosome. The process of flow-sorting involves labeling chromosomes with fluorescent dyes and passing them through a flow cytometer, which sorts the chromosomes based on their scatter properties and fluorescence intensity. Overall, flow-sorting offers a valuable tool to study specific chromosomes, including those involved in translocations.

1.5 Objectives

The main objective of this work was to perform an in-depth characterization of mammalian LTRs resorting to whole genome sequencing data, with focus on two model organisms: *Peromyscus* genus and *Homo sapiens*. This involves identifying the abundance and diversity of LTRs within the genome, determining their length and number of copies, characterizing the variation in their nucleotide sequence, and mapping their chromosomal location when possible.

Specifically, the study of the LTR class diversity present in *Peromyscus* genus genomes was of interest, centering on the satellite PMSat family, in an effort to evaluate their species-specific LTRs, while also determining their relative position between the species for karyotype/ideogram comparison. Assessment of the impact from distinct sequencing methods in LTR detection and quantification was an additional aim regarding this specific analysis.

Additionally, another aim of this work was to improve the characterization of the diversity of repetitive arrays in the human genome, taking advantage of long contiguous reads produced by nanopore sequencing technology. Furthermore, this study aims to re-construct the Robertsonian translocation region in chromosome t(14;21) in human cells, for a better characterization of the long tandem repeats involved in this translocation. It was a final aim of this work to characterize human satellite HSat1A genomic copy number, sequence variation, and transcription product diversity with the purpose of unraveling the transcriptional landscape of this human satellite.

2. Methods

2.1 Whole genome datasets

The genomes selected for the LTR family diversity study on the *Peromyscus* genus were from the *Peromyscus maniculatus* (PMAN 1.0, PMAN 2.1 CHR and PMAN 2.1 CONTIGS), the *Peromyscus californicus* (PCAL), the *Peromyscus eremicus* (PERE) and the *Peromyscus leucopus* (PLEU) species. These genomes are reported in GenBank under accession numbers GCA_003704035.1 (PMAN 1.0), GCA_003704035.3 (PMAN 2.1 CHR), RCWR01000001-RCWR01155965 (PMAN 2.1 CONTIGS), GCA_007827085.3 (PCAL), GCA_902702925.1 (PERE) and GCA_004664715.2 (PLEU). For the study of LTRs on long contiguous reads from nanopore sequencing in human genome two Whole Human Genome Sequencing Projects were selected, the NA12878 (https://github.com/nanopore-wgs-consortium/NA12878) and CHM13 (https://github.com/marbl/CHM13). The header of each read was renamed by numerical order. The length and abundance of raw reads from each assembly or dataset was determined for the whole dataset and different sequence categories with the use of custom in-house R (R version 4.2.2) scripts (https://github.com/GamaPintoLab/DanielEleuterio-MSc-Thesis).

2.2 Cell culture, genomic DNA isolation and Oxford Nanopore Technology (ONT) sequencing

In this work the human fibroblast GM03417 cell line (Coriell Institute) was used, which presents mosaicism with 32% of cells being a balanced 45,XX, t(14;21) (ISCN: 46,XX,der(14;21)(14qter> 14q10::21q10>21qter),+21[34]/45,XX,der(14;21)(14qter>14q10::21q10>21qter)[16]). The GM03417 cell line was grown according to the protocol from Coriell Institute. $2x10^7$ freshly cultured fibroblasts from the GM03417 cell line were used to extract genomic DNA with the Qiagen Genomic Tips 100/G extraction kit, described to yield high quality purified DNA-fragments in the range between 20-150kb (Jain et al., 2018). Using this method, it was possible to isolate ~100µg of high-quality DNA (as quantified by Qbit). Genomic DNA from the GM03417 cell line was prepared for ligation genomic DNA 1D R9.4 chemistry sequencing and subjected to one round of library preparation using the SQK-LSK109 Kit from ONT. This was followed by two runs of nanopore sequencing on a FLO-MIN106 cell, performed on a GridION sequencer. Basecalling was performed with Guppy, which is only available via their community site (https://community.nanoporetech.com) and the results were compiled into FASTQ format. These FASTQ files were then converted into FASTA format for analysis. The sample with the flow sorted chromosomes from the GM03417 cell line was prepared for ligation genomic DNA 1D R9.4 chemistry sequencing in-house and subjected to one round of library preparation using the SQK-LSK109 Kit from ONT. This was followed by one run of nanopore sequencing on a FLO-MIN106 cell, performed on a GridION sequencer. Basecalling was performed with Guppy, which is only available via their community site (https://community.nanoporetech.com) and the results were compiled into FASTQ format. These FASTQ files were then converted into FASTA format for analysis. The sequencing was stopped at an early stage

due to a low sequencing data yield. To optimize the reading experience of this work, the sequencing data obtained from the GM03417 cell line will be referred to as the GM03417 dataset.

Samples were generated by the host lab and libraries were synthesized and sequenced at the BioISI Genomics facility.

2.3 Flow-sorting of chromosomes in GM03417 cell line

Purification of flow-sorted chromosomes was performed by the CytoGenomics Lab (UTAD), following Kuderna et al. 2019 protocol with modifications: the sorted chromosomes were treated overnight with 10µl of proteinase K (20mg/ml) at 50ºC; after treatment the solution was transferred to a Pur-A-Lyzer Maxi Dialysis column (Sigma) for dialysis against 500ml 1xTE buffer, in order to remove the proteinase K as well as chromamycin-A3 and Hoeschst 33258 (intercalating dyes used for the chromosome flow-sorting); dialysis was carried for 48h exchanging the buffer every 10-16h; after the dialysis the DNA was transferred to a 1.5 ml tube and quantified with Qubit using dsDNA HS Kit, with a Qubit reading of 59ng/ml and total sample volume of ~500µl. This sample contained 330,000 chromosomes isolated and was stored at 4ºC.

Raw reads from sequencing performed on the flow-sorted chromosomes were aligned using the minimap2 tool (Li, 2018) (https://github.com/lh3/minimap2) to the 24 chromosomes of the human genome assembly GRCh38.p13, individually. From this alignment, quantification of mapped reads was performed.

2.4 LTR detection and clustering analysis

The detection of tandem repeats was performed using the Tandem Repeat Finder (TRF) algorithm (Benson, 1999), with the following parameters: *match*, *mismatch* and *delta* 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 in R (R version 4.2.2) (https://github.com/GamaPintoLab/DanielEleuterio-MSc-Thesis). Redundant entries from the TRF output were eliminated, considering all embedded LTR arrays and for those that possessed the same sequence coordinates, the LTR with the larger consensus length was discarded.

A clustering algorithm, blastclust (Altschul et al., 1990) was considered LTR characterization. This was used for the clustering of sequence-based similarity within LTRs detected assembly and dataset. This clustering analysis was performed with parameters of a minimum length coverage of 51% and a similarity threshold of 50%. Based on this, LTRs within these thresholds were clustered to the same group.

Quantification of alignments of satDNA arrays (ALR, BSR, HSat1A and HSat2) from GM03417, NA12878 and CHM13 with chromosomes 14, 21, and both together from the human genome assembly GRCh38.p13, was performed using the BLAST tool (Z. Zhang et al., 2000).

2.5 Annotation and classification of LTR classes

The monomer sequences corresponding to each group that were clustered together were then compared to a repetitive sequence database Dfam (https://dfam.org/home) (Hubley et al., 2016), and assigned their name matching with those repetitive elements in the database. The LTR groups from the *Peromyscus* genus analysis that had not been identified in Dfam had their names assigned according to their assembly, numbered from largest to smallest group (e.g., PMAN: LTR_PMAN_1 to LTR_PMAN_11). For the nanopore sequencing datasets from the human genome, a similar process was performed for LTR groups that had not been identified in Dfam, with their names also numbered from largest to smallest group (e.g., H_LTR_1 to H_LTR_6). The BLAST tool (Z. Zhang et al., 2000) was used to identify similar LTR groups across assemblies in the *Peromyscus* genus LTR analysis, considering the first species in which the LTRs were originally identified, following the order PMAN>PCAL>PERE>PLEU. This process was also performed on the nanopore sequencing datasets from the human genome; however, these maintained a unique identification letter representative of human (H) and numbered following the order GM03417>NA12878>CHM13. Groups that contained a query cover >80% and identity score >80% in other assemblies were aggregated into a class. Clusters with values below the cut-offs maintained their primary classification. A representative workflow of this process can be found in **Figure 3A**.

Comparative analysis based on GC-content, monomer and array length was performed on PMSat arrays detected in the *Peromyscus* genus assemblies selected, using custom scripts in R (R version 4.2.2) (https://github.com/GamaPintoLab/DanielEleuterio-MSc-Thesis).

2.6 Distribution and statistical analysis of PMSat characteristics across the *Peromyscus* genus

Selection of PMSat arrays for a distribution and statistical test analysis across the *Peromyscus* genus assemblies was performed. The ANOVA One-way statistical test was selected for this analysis, with the additional use of a Tukey Post-Hoc test for a pairwise comparison of PMSat monomer length, monomer GC-content, array length, and array GC-content among the *Peromyscus* genomes. The analysis was performed using custom scripts in R (R version 4.2.2) (https://github.com/GamaPintoLab/DanielEleuterio-MSc-Thesis).

2.7 Relative positioning and orientation of LTR arrays

Selection of LTR arrays for relative positioning and orientation on chromosomes from PMAN, PCAL, PERE and PLEU assemblies was based on their presence in at least two species, such as described in **Figure 3**. The output information from the TRF analysis performed on these four *Peromyscus* spp. provided the coordinates of the LTR arrays on the chromosomes, which was used for relative positioning on each chromosome correspondent with each species. Additionally, orientation of each LTR array was

defined based on a randomly selected monomer corresponding to each LTR class, for a baseline comparison between LTR arrays on each species.

2.8 Quantification and periodicity study of HSat1A

The HSat1A repetitive sequence was extracted from the Dfam database (https://dfam.org/), with the name SAR, accession DF0001062.4. Quantification was performed using RepeatMasker (Smit et al., 2013) to detect the presence of HSat1A (SAR) on reads from NA12878 sequencing data. For the periodicity studies of HSat1A repetitive sequences, reads with a number higher than 400 Kbp containing only these sequences were selected. Then, the NTRprism (https://github.com/-altemose/NTRprism) scripts were used on the selected reads.

2.9 3' RACE, RACE-seq, and sequencing analysis

HeLa cDNA was prepared using the 5′/3′ RACE (rapid amplification of cDNA ends) Kit, 2nd Generation (Roche), and subjected to PCR using the provided PCR anchor primer (5'-GACCACGCGTATCGATGTCGAC-3') and the HSat1A forward primer (5'-TGTGCGGTACATAAGATATCAAAG-3') at the CytoGenomics Lab (UTAD). 3′ RACE was coupled with high-throughput sequencing, performed by STAB VIDA NGS sequencing service. The analysis of the generated sequence raw data was carried out using CLC Genomics Workbench 12.0.3 (https://www.qiagenbioinfor-matics.com/). The following data processing is detailed in **Figure 10**. RACE-PCR was followed by paired-end 300-nt Illumina sequencing. To evaluate the quality of the reads throughout the workflow, FASTQC (Andrews, 2010) was used. CutAdapt (Martin, 2011) was used to remove the Universal Adaptors on the extremities of the reads. The paired reads were merged using the PEAR (J. Zhang et al., 2014) tool. To guarantee the quality of the reads, a score Phred of equal or higher than 30 was applied with the tool Seqkit (Shen et al., 2016). Furthermore, to also guarantee that at least one monomer of HSat1A was present on every read, selection for the size of the sequences to be at least 85 bp or higher was applied, considering the downstream removal of the Oligo Anchor primer of the reads. For the next step, all the reads were set in the same orientation, to ensure that we could proceed with a Multiple Sequence Alignment analysis downstream. This was also done with the tool Seqkit (Shen et al., 2016) and the GNU grep. Then, the identification of reads containing the HSat1A (SAR) satellite sequence was performed with RepeatMasker (Smit et al., 2013). The count and removal of duplicates within these reads containing the HSat1A (SAR) satellite sequence were performed using the Seqkit (Shen et al., 2016) tool. Subsequently, clustering of these sequences was achieved with the MESHClust v3.0 program (Girgis, 2022), selecting for a threshold identity score of 90% to determine the cluster membership. Multiple sequence alignment was performed with an R (R version 4.2.2) script on cluster center sequences to produce a dendrogram. From the clustering, it was possible to discover motifs based on the sequence in each group using the Improbizer tool (Ao et al., 2004). The scripts and produced data are publicly available at https://github.com/GamaPintoLab/HSAT1A-transcript-analysis.

3. Results

3.1 Study of LTR diversity in *Peromyscus* spp. genome assemblies

This section describes the results obtained in the frame of the first objective of this thesis, to characterize LTR diversity in the *Peromyscus* genus. The interest in this topic stemmed from the previous identification in *Peromyscus eremicus* of the PMSat LTR by the CytoGenomics Lab (UTAD) (Louzada et al., 2015), leading to the exploration of tandem repeats within selected *Peromyscus* genus available genome assemblies, to further improve our understanding of the complexity of these sequences within these genomes.

3.1.1 Overview of publicly available *Peromyscus* genomes datasets

The study began with a survey of genomes of the *Peromyscus* genus in public repositories. *Peromyscus maniculatus* version 1.0 (PMAN 1.0), the contigs of *Peromyscus maniculatus* version 2.1 (PMAN 2.1 CONTIGS) and the chromosomes with unplaced scaffolds of the same species (PMAN 2.1 CHR), the *Peromyscus californicus* (PCAL), the *Peromyscus eremicus* (PERE) and the *Peromyscus leucopus* (PLEU) species, were identified and downloaded for analysis (**Table 1)**. The *Peromyscus maniculatus* genome had two assemblies available (PMAN 1.0 and PMAN 2.1), increasing the information available to study repetitive sequences. For a better comparison in terms of similar contig/scaffold length between PMAN 1.0 and PMAN 2.1, contigs from the PMAN 2.1 (PMAN 2.1 CONTIGS) assembly were also selected for this analysis. The PMAN 1.0 and PMAN 2.1 CONTIGS assemblies do not possess assembled chromosomes. All other genome data only had one assembly available.

Firstly, through the retrieval of length differences and assembly methods used, dataset discrepancies were analyzed for any implication in the subsequent LTR analysis. Sequencing data from all assemblies was collected and metrics based on unplaced scaffold length were calculated for this purpose (**Table 1**). The number of contigs from the PMAN 2.1 CONTIGS assembly is much higher than the total number of scaffolds and contigs from the PMAN 1.0 assembly. Moreover, both of these assemblies have a minimum scaffold or contig length of 201 bp, although the contigs from the PMAN 2.1 CONTIGS assembly have a much lower average and maximum length than the scaffolds and contigs from the PMAN 1.0 assembly. The unplaced scaffolds from the PMAN 2.1 CHR assembly had similar average length with the contigs from the PMAN 2.1 CONTIGS assembly. Additionally, the maximum length was considerably higher in the unplaced scaffolds from the PMAN 2.1 CHR assembly, however, still much lower than that of the PMAN 1.0 assembly. These differences between the reconstructed sequences of the three PMAN assemblies suggest there is some length distribution variability in each one that needs to be assessed before the LTR analysis. Moreover, it will be important to assess that variability across the assemblies from other *Peromyscus* genomes.

Table 1 – Overview of publicly available *Peromyscus* genus genomes datasets, sequencing, and assembly methods. Quantification and distribution of contigs/scaffolds belonging to each assembly. *The nucleotide sum value is equal to the genome size because these assemblies do not have their sequences assembled into chromosomes. **The custom method is referenced in (Long et al., 2019).

	Genome Assemblies					
	PMAN 1.0	PMAN 2.1 CONTIGS	PMAN 2.1 CHR	PCAL	PERE	PLEU
Genome size (GS) (bp)	2,630,541,020	2,387,876,961	2,512,423,440	2,561,163,842	2,713,541,378	2,475,180,836
Chromosomes (Haploid)	-	-	24	24	24	24
Scaffolds/Contigs	30,921	155,965	8,499	34,100	6,784	1,832
Scaffolds nucleotides sum (bp)	2,630,541,020*	2,387,876,961*	75,979,398	180,923,057	172,598,049	69,829,826
Scaffolds min length (bp) (Q0)	201	201	1,000	692	199	1,000
Scaffolds first quartile length (bp) (Q1)	630	2,794	1,376	1,300	10,451	6,000
Scaffolds median length (bp) (Q2)	2,048	8,682	2,935	2,001	14,452	17,349
Scaffolds third quartile length (bp) (Q3)	7,056	20,406	6,687	4,171	25,000	36,053
Scaffolds average length (bp)	85,073.00	15,310.30	8,939.80	5,305.70	25,441.90	38,116.70
Scaffolds max length (bp) (Q4)	18,898,765	256,906	803,378	9,464,350	1,387,000	861,000

Table 1 (cont.) – Overview of publicly available *Peromyscus* genus genomes datasets, sequencing, and assembly methods. Quantification and distribution of contigs/scaffolds belonging to each assembly. *The nucleotide sum value is equal to the genome size because these assemblies do not have their sequences assembled into chromosomes. **The custom method is referenced in (Long et al., 2019).

	Genome Assemblies					
	PMAN 1.0	PMAN 2.1 CONTIGS	PMAN 2.1 CHR	PCAL	PERE	PLEU
Sequencing Technology	FLX 454; Illumina HiSeq	Illumina HiSeq	Illumina HiSeq	Illumina NovaSeq	Whole Genome Shotgun	Illumina; PacBio; HiC-Illumina
Genome coverage	110x	115x	115x	37x	50x	60x
Assembly method	Newbler v. 2.3 and 2.5; AllPaths v. 41070; ATLAS-gapfill v. 2.2; ATLAS-link v. 1.0	AllPaths v. 2016-03	AllPaths v. 2016-03	SuperNova v. 2.0.0; HiRise v. APRIL-2019	-	Custom Method**
Assembly accession	GCA_000500345.1	RCWR01	GCA_003704035.3	GCA_007827085.2	GCA_902702925.1	GCA_004664715.2
Assembly level	Scaffold	Contigs	Chromosome	Scaffold (Chromosome based on scaffold)	Scaffold (Chromosome based on scaffold)	Chromosome
Publish Date	03/12/2013	13/10/2018	17/09/2020	25/03/2020	26/11/2019	02/10/2020

For a better representation of the differences between the assemblies from the *Peromyscus* genomes, a distribution of unplaced scaffold/contig length for all assemblies was generated (**Figure 2**).

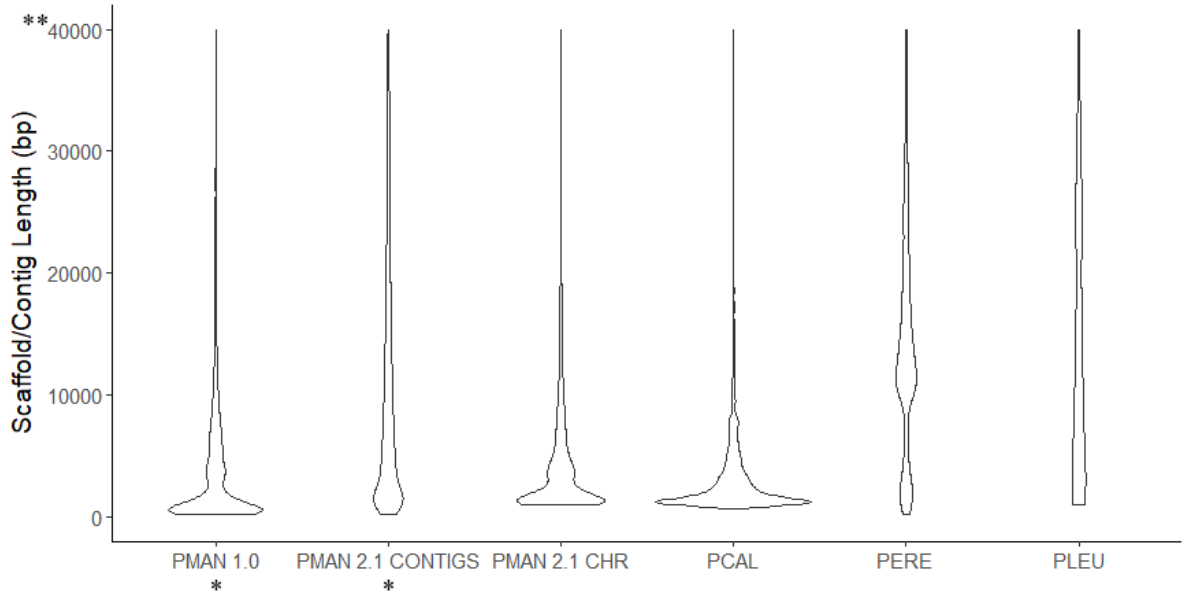


Figure 2 – Comparison of unplaced scaffolds length variation of all assemblies. A violin plot for length distribution of contig/scaffolds, or unplaced scaffolds in each assembly. The range length was restricted between 0 and 40,000 bp. *PMAN 1.0 and PMAN 2.1 CONTIGS contain all scaffolds/contigs from their assemblies, while the length variation of other assemblies only represents unplaced scaffolds. **The maximum length presented was selected based on the third quartile length (bp) (Q3) and for better visualization of the distribution.

The length distribution between the scaffolds from the PMAN 1.0 assembly and the unplaced scaffolds from the PMAN 2.1 CHR assembly could indicate similar sequencing data for repetitive elements analysis, as both revealed a similar length distribution (**Figure 2**). Nevertheless, the scaffolds and contigs from the PMAN 1.0 assembly and the contigs from the PMAN 2.1 CONTIGS assembly have similar nucleotide sum (bp) and also a higher number in comparison with the PMAN 2.1 CHR assembly, which might provide a more representative landscape of the LTRs present in the genome of this species (**Table 1**).

The PCAL assembly appears to have a similar sequencing data length distribution of unplaced scaffolds relative to that of the PMAN 1.0 and the PMAN 2.1 CHR assemblies, although the unplaced scaffolds nucleotide sum is double that of the PMAN 2.1 CHR assembly (**Figure 2**; **Table 1**). Moreover, the distribution of unplaced scaffolds in the PCAL assembly shows a higher quantity of shorter length sequences in comparison to the PMAN 1.0 and the PMAN 2.1 CHR assemblies (**Figure 2**).

The PERE assembly has a lower number of unplaced scaffolds when compared to the PMAN 2.1 CHR assembly (**Table 1**). The unplaced scaffold nucleotide sum is also double that of the PMAN 2.1 CHR assembly. However, the length distribution of the unplaced scaffolds for the PERE assembly is concentrated at a higher length, with an average of ~12,000 bp (**Figure 2**).

The unplaced scaffolds from the PLEU assembly have a length distribution more similar to those from the PERE assembly (**Table 1**; **Figure 2**). Moreover, the unplaced scaffold nucleotide sum is similar to that of the PMAN 2.1 CHR assembly (**Table 1**). This might have implications in the quantification of LTRs.

Based on this analysis, it can be concluded that the publicly available genome assemblies for the *Peromyscus* genus can support the comparative analysis of LTR sequences present in different species, although with some limitations.

3.1.2 Exploring LTR array detection and PMSat satellite distribution in *Peromyscus* genomes

In this section, detection of LTR arrays was performed in the selected genomes of PMAN (PMAN 1.0; PMAN 2.1 CONTIGS; PMAN 2.1 CHR), PCAL, PERE and PLEU. Initially, quantification of LTR arrays and nucleotide sum were evaluated to assess LTR characteristics across the *Peromyscus* genus. Subsequently, a clustering algorithm was used to discriminate specific LTR classes. From this, in agreement with previous data (Louzada et al., 2015; Smalec et al., 2019), a predominant LTR class was evident on first examination, the PMSat satellite family. Considering this, PMSat was also selected for quantification of its arrays and nucleotide sum evaluation, with the aim of comparing these values with the overall LTR population (**Table 2)**.

The first assembly, PMAN 1.0, showed to have the highest number of LTR arrays detected of all PMAN assemblies, in addition to the highest array nucleotide sum. However, regarding the PMSat arrays, there are some slight variations in comparison with the overall LTR detection. The PMAN 1.0 assembly also had the highest number of PMSat arrays detected and subsequent nucleotide sum, however, their relative ratio to the number of LTR arrays and LTR nucleotide sum (52.03% and 62.37%, respectively) is much lower in comparison to the other assemblies of PMAN 2.1 CHR (70.86% and 78.26%, respectively) and of PMAN 2.1 CONTIGS (89.92% and 91.55%, respectively) (**Table 2**).

The vast number of contigs contained in the PMAN 2.1 CONTIGS genome does not seem to improve LTR array detection, with 784 arrays being detected, against the 1672 arrays in the PMAN 1.0 assembly, which had a much lower number of scaffolds in its assembly (**Table 1-2**). The discrepancy in scaffold and contig length distribution between the PMAN 1.0 assembly and the PMAN 2.1 CONTIGS assembly could account for this difference, as the former exhibits smaller scaffolds compared to the distribution of contigs in the latter (**Figure 2)**. Moreover, despite the PMAN 2.1 CONTIGS assembly having a lower count of detected LTR arrays compared to the PMAN 1.0 assembly, this does not seem to be connected with the number of contigs or the overall higher median length value in the PMAN 2.1 CONTIGS assembly (**Table 1-2**). However, it is possible that the presence of a substantial number of long scaffolds in the PMAN 1.0 assembly could have influenced the detection of LTR arrays. This might explain the observed half-term value of 937 LTR arrays detected in the PMAN 2.1 CHR assembly (**Table 1-2**).

To further explore this topic, the percentage of LTR and PMSat arrays detected in the chromosomes and unplaced scaffolds was assessed (**Table 3**). The ratio of PMSat/LTR arrays in the chromosomes of PMAN 2.1 CHR and the scaffolds of PMAN 1.0 exhibits a notable similarity. Additionally, the ratio of PMSat/LTR arrays in unplaced scaffolds in the PMAN 2.1 CHR assembly is similar to the ratio found in

Table 2 – Overview of LTR and PMSat detection in assemblies from *Peromyscus* genus genomes. Quantification of LTR and PMSat arrays, and nucleotide sum was performed, with calculation of their relative ratios, and with the genome size.

	Genome Assemblies					
	PMAN 1.0	PMAN 2.1 CONTIGS	PMAN 2.1 CHR	PCAL	PERE	PLEU
Genome size (GS) (bp)	2,630,541,020	2,387,876,961	2,512,423,440	2,561,163,842	2,713,541,378	2,475,180,836
Large Tandem Repeats (LTR) arrays	1,672	784	937	1,163	18,233	334
LTR nucleotides sum (bp)	6,940,254	3,488,457	4,371,890	5,642,312	54,425,850	1,831,918
% (LTR/GS arrays length sum) (bp)	0.26%	0.15%	0.17%	0.22%	2.01%	0.07%
PMSat arrays	870	705	664	494	988	96
PMSat nucleotides sum (bp)	4,328,851	3,193,550	3,421,368	2,351,453	7,235,801	734,407
% (PMSat nucleotides sum / GS)	0.16%	0.13%	0.14%	0.09%	0.27%	0.03%
% (PMSat/LTR nucleotides sum)	62.37%	91.55%	78.26%	41.68%	13.29%	40.09%
% (PMSat/LTR arrays)	52.03%	89.92%	70.86%	42.48%	5.42%	28.74%

Table 3 – Assessment of LTR and PMSat arrays distributed between chromosomes, scaffolds, contigs and unplaced scaffolds for all *Peromyscus* genus selected assemblies. Quantification of LTR and PMSat arrays across all six *Peromyscus* genus assemblies, with calculation of their relative ratio. *This assembly is scaffold based. **This assembly is contig based.

	Genome Assemblies					
	PMAN 1.0	PMAN 2.1 CONTIGS	PMAN 2.1 CHR	PCAL	PERE	PLEU
LTR arrays in chromosomes	-	-	441	234	16,666	253
LTR arrays in unplaced scaffolds	1672	784	496	929	1,567	81
% (LTR arrays in chromosomes/all LTR arrays)	-	-	47.07%	20.12%	91.41%	75.75%
% (LTR arrays in unplaced scaffolds/all LTR arrays)	-	-	52.93%	79.88%	8.59%	24.25%
PMSat arrays in chromosomes	-	-	226	155	76	53
PMSat arrays in unplaced scaffolds	870	705	438	339	912	43
% (PMSat arrays in chromosomes/all PMSat arrays)	-	-	34.04%	31.38%	7.69%	55.21%
% (PMSat arrays in unplaced scaffolds/all PMSat arrays)	-	-	65.96%	68.62%	92.31%	44.79%
% (PMSat/LTR arrays) in chromosomes	-	-	51.25%	66.24%	0.46%	20.95%
% (PMSat/LTR arrays) in unplaced scaffolds	52.03%	89.92%	88.31%	36.49%	58.20%	53.09%

contigs from the PMAN 2.1 CONTIGS assembly. As these scaffolds were not placed in any region of the chromosomes, this might suggest that PMSat arrays detected in unplaced scaffolds from the PMAN 2.1 CHR assembly mainly derive from the PMAN 2.1 CONTIGS PMSat arrays that overlapped in the assembly process. The number of LTR arrays in the PMAN 2.1 CHR assembly showed a similar distribution between the chromosomes and unplaced scaffolds. However, for the detected PMSat arrays, the distribution appears slightly different, favoring the unplaced scaffolds in comparison to the chromosomes.

The other *Peromyscus* genomes (PCAL, PERE and PLEU) also have many variations between the number and cumulative length of detected arrays. The PCAL assembly showed a slightly higher number of LTR arrays detected in comparison with the PMAN 2.1 CHR assembly (**Table 2)**, although the observed length distribution of their respective reconstructed sequences is similar in both (**Figure 2**), which might be due to the maximum unplaced scaffold length from the PCAL assembly being much larger (**Table 1**). Moreover, the PCAL assembly has a slightly lower number of PMSat arrays detected in comparison with the PMAN 2.1 CHR assembly (**Table 2)**, which might be due to it having a higher percentage of LTR arrays detected in unplaced scaffolds (**Table 3**), considering these reconstructed sequences have a higher distribution of shorter lengths (**Figure 2**). Additionally, this may mean that the PCAL assembly has more of other (non-PMSat) LTR arrays.

The PERE genome appears to have the highest number of detected LTR arrays and corresponding nucleotide sum of all *Peromyscus* genomes in this study (**Table 2**). Even so, considering the PERE assembly has a lower number of unplaced scaffolds than the PMAN 2.1 CHR and PCAL assemblies (**Table 1**), this large number of LTR arrays detected could be due to the length of the unplaced scaffolds of the PERE assembly being at least two to three times larger than the length of unplaced scaffolds of PMAN 2.1 CHR and PCAL (**Figure 2**). The analysis also showed that the LTR arrays from the PERE assembly are mostly detected in chromosomes (**Table 3**). Concerning the PMSat arrays detected in the PERE assembly, it is possible to distinguish that there is a partial increase of PMSat arrays and subsequent nucleotide sum, however, their ratio in relation to the LTRs are the lowest of all the assemblies (**Table 2**). Moreover, most of the PMSat arrays in this genome are found in the unplaced scaffolds, while the PMSat arrays detected in the chromosomes only account for 7.69% (**Table 3**).

The PLEU assembly has the less LTR arrays and LTR nucleotide sum in comparison to all other assemblies (**Table 2**), which could be associated with having the lowest number of unplaced scaffolds across all *Peromyscus* genomes. The length distribution of the unplaced scaffolds from the PLEU assembly are more similar to that of the PERE assembly (**Figure 2**), also corresponding to the pattern observed of LTR arrays detected in chromosomes in comparison to unplaced scaffolds (**Table 3**). However, the ratio of PMSat/LTR arrays detected in chromosomes is not as low as in the PERE assembly, while the ratio of PMSat/LTR arrays in unplaced scaffolds is also similar to that of the PERE assembly (**Table 3**). The impact on overall LTR detection in the PLEU assembly in comparison to the other assemblies may be due to the custom method used for assembly of the reconstructed sequences.

It was possible to conclude that PMSat detection was in similar proportion across all the genomes, even with some variations regarding LTR detection, namely in the PERE and PLEU assemblies.

3.1.3 Clustering analysis of LTR arrays from *Peromyscus* genomes

To assess the similarities between LTR arrays detected within each assembly, a clustering analysis was performed to identify LTR groups (**Table 4**; see also **Supplementary Table 1-6**). LTR groups that were not described in Dfam (Storer et al., 2021), were named according to their assembly, numbered from largest to smallest (Ex: LTR_PMAN_1 to LTR_PMAN_11). The BLAST tool (Z. Zhang et al., 2000) was used to identify similar LTR groups across assemblies, considering the first species in which the LTRs were originally identified, following the order PMAN>PCAL> PERE>PLEU. Groups containing a query cover >50% and identity score >50% in other assemblies were aggregated into a class, while those within the same class that had a query cover >80%, and identity

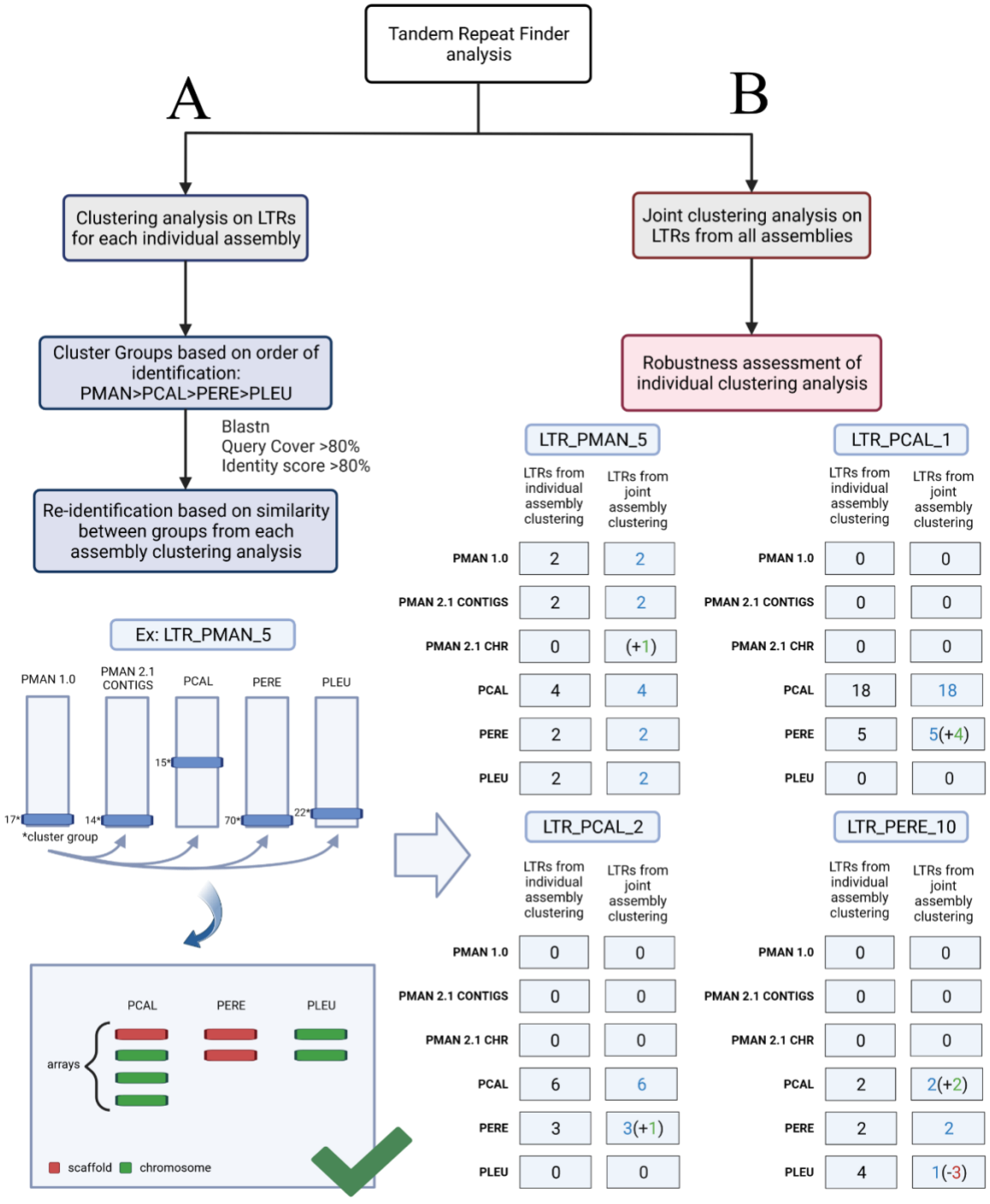


Figure 3 – Overview of the LTR clustering workflow. A: Assembly-centered LTR group identification. Clustering of LTR arrays was performed on each assembly followed by cross-species mapping using Blast in the order PMAN >PCAL >PERE >PLEU; LTRs with a query cover >80% and identity score >80% were defined as belonging to the same class. Bottom: selection criteria for LTR group validation against the global clustering approach (one array located in chromosomes from at least two different species assemblies). **B**: Global clustering classification of LTR classes and example of comparative analysis with assembly-based classification for four selected classes.

score >80% in other assemblies were designated as a sub-class from that class. Clusters with values below the cut-offs maintained their primary classification (**Figure 3A**).

The clustering analysis was able to generate groups based on previous and newly identified LTR classes (**Table 4**). The assemblies that showed a higher percentage of clustered LTRs were the PMAN 2.1 CONTIGS and PCAL. The PERE assembly has the smallest percentage of clustered LTR groups, while the PMAN 1.0, PMAN 2.1 CHR and PLEU assemblies have at least half or more of their LTRs detected clustered in groups. Furthermore, the clustered groups of all assemblies were mostly composed of PMSat arrays. Interestingly, the PCAL assembly appears to contain a higher percentage of the newly identified LTR classes, LTR_PMAN_1 and LTR_PMAN_2, in comparison with the other assemblies.

Table 4 – Clustering overview of different assemblies of *Peromyscus* genus genomes, based on similarity. Count of arrays, % of arrays, number of classes and sub-classes for each group of clustered LTRs and for orphan LTRs, corresponding to each genome. Only LTR classes that are present in 2 or more assemblies were selected.

	PMAN 1.0				**PMAN 2.1 CONTIGS**				**PMAN 2.1 CHR**			
	Arrays	% Arrays	Class	Sub-class	Arrays	% Arrays	Class	Sub-class	Arrays	% Arrays	Class	Sub-class
Orphan LTRs	**752**	**44.92**			**45**	**5.74**			**232**	**24.76**		
Grouped LTRs	**922**	**55.08**	**13**	**8**	**739**	**94.26**	**9**	**7**	**705**	**75.24**	**10**	**10**
Satellite	**880**	**52.57**	**3**	**4**	**713**	**90.94**	**3**	**5**	**673**	**71.82**	**3**	**5**
PMSat	870	51.97		1	705	89.92		2	664	70.86		2
MurSatRep1	8	0.48		2	6	0.77		2	7	0.75		2
MMSAT4	2	0.12		1	2	0.26		1	2	0.21		1
TE	**12**	**0.72**	**3**	**4**	**4**	**0.51**	**1**	**2**	**16**	**1.71**	**4**	**5**
L1_Mur2_orf2	7	0.42		2	0	0.00		0	8	0.85		1
L1_Mur3_orf2	0	0.00		0	0	0.00		0	2	0.21		1
B1_Mus1/2	3	0.18		1	4	0.51		2	4	0.43		2
RMER1C	2	0.12		1	0	0.00		0	2	0.21		1
Lx5c_3end	0	0.00		0	0	0.00		0	0	0.00		0
PMAN Class	**25**	**1.49**	**7**	**4**	**17**	**2.17**	**5**	**5**	**13**	**1.39**	**3**	**3**
LTR_PMAN_1	11	0.66		1	8	1.02		1	8	0.85		1
LTR_PMAN_2	4	0.24		1	3	0.38		1	3	0.32		1
LTR_PMAN_4	2	0.12		1	2	0.26		1	2	0.21		1
LTR_PMAN_5	2	0.12		1	2	0.26		1	0	0.00		0
LTR_PMAN_7	0	0.00		0	2	0.26		1	0	0.00		0
LTR_PMAN_9	2	0.12		0	0	0.00		0	0	0.00		0
LTR_PMAN_10	2	0.12		0	0	0.00		0	0	0.00		0
LTR_PMAN_11	2	0.12		0	0	0.00		0	0	0.00		0
PCAL Class	**0**	**0.00**	**0**	**0**	**0**	**0.00**	**0**	**0**	**0**	**0.00**	**0**	**0**
LTR_PCAL_1	0	0.00		0	0	0.00		0	0	0.00		0
LTR_PCAL_2	0	0.00		0	0	0.00		0	0	0.00		0
LTR_PCAL_6	0	0.00		0	0	0.00		0	0	0.00		0
PERE Class	**0**	**0.00**	**0**	**0**	**0**	**0.00**	**0**	**0**	**0**	**0.00**	**0**	**0**
LTR_PERE_10	0	0.00		0	0	0.00		0	0	0.00		0

Table 4 (cont.) – Clustering overview of different assemblies of *Peromyscus* genus genomes, based on similarity. Count of arrays, % of arrays, number of classes and sub-classes for each group of clustered LTRs and for orphan LTRs, corresponding to each genome. Only LTR classes that are present in 2 or more assemblies were selected.

	PCAL				PERE				PLEU			
	Arrays	% Arrays	Class	Sub-class	Arrays	% Arrays	Class	Sub-class	Arrays	% Arrays	Class	Sub-class
Orphan LTRs	**72**	**6.19**			**17025**	**93.37**			**162**	**48.50**		
Grouped LTRs	**1091**	**93.81**	**10**	**5**	**1208**	**6.63**	**13**	**12**	**172**	**51.50**	**13**	**8**
Satellite	**503**	**43.25**	**2**	**4**	**998**	**5.47**	**2**	**4**	**113**	**33.83**	**3**	**3**
PMSat	494	42.48		2	990	5.43		3	96	28.74		1
MurSatRep1	9	0.77		2	8	0.04		1	15	4.49		1
MMSAT4	0	0.00		0	0	0.00		0	2	0.60		1
TE	**0**	**0.00**	**0**	**0**	**17**	**0.09**	**3**	**6**	**7**	**2.10**	**3**	**3**
L1_Mur2_orf2	0	0.00		0	5	0.03		2	0	0.00		0
L1_Mur3_orf2	0	0.00		0	6	0.03		3	0	0.00		0
B1_Mus1/2	0	0.00		0	0	0.00		0	3	0.90		1
RMER1C	0	0.00		0	0	0.00		0	2	0.60		1
Lx5c_3end	0	0.00		0	6	0.03		1	2	0.60		1
PMAN Class	**521**	**44.80**	**5**	**8**	**21**	**0.12**	**4**	**5**	**21**	**6.29**	**6**	**7**
LTR_PMAN_1	315	27.09		2	8	0.04		1	2	0.60		1
LTR_PMAN_2	190	16.34		3	7	0.04		2	0	0.00		0
LTR_PMAN_4	9	0.77		1	4	0.02		1	11	3.29		2
LTR_PMAN_5	4	0.34		1	2	0.01		1	2	0.60		1
LTR_PMAN_7	3	0.26		1	0	0.00		0	0	0.00		0
LTR_PMAN_9	0	0.00		0	0	0.00		0	2	0.60		1
LTR_PMAN_10	0	0.00		0	0	0.00		0	2	0.60		1
LTR_PMAN_11	0	0.00		0	0	0.00		0	2	0.60		1
PCAL Class	**28**	**2.41**	**3**	**3**	**10**	**0.05**	**3**	**4**	**0**	**0.00**	**0**	**0**
LTR_PCAL_1	18	1.55		1	5	0.03		2	0	0.00		0
LTR_PCAL_2	6	0.52		1	3	0.02		1	0	0.00		0
LTR_PCAL_6	4	0.34		1	2	0.01		1	0	0.00		0
PERE Class	**0**	**0.00**	**0**	**0**	**2**	**0.01**	**1**	**1**	**4**	**1.20**	**1**	**2**
LTR_PERE_10	0	0.00		0	2	0.01		1	4	1.20		2

In addition to the PMSat satellite class detected in all assemblies, there were also other two satellite classes that were detected in clustered LTRs, MurSatRep1 and MMSAT4, although in much lower count. Nevertheless, the number of arrays from these LTR classes was found to be similar across assemblies.

Transposable elements (TE) arrays were also detected within clustered groups, pertaining to the L1_Mur2_orf2, L1_Mur3_orf2, B1_Mus1/2, RMER1C and Lx5c_3end classes. The distribution of these classes was significantly different across assemblies (**Table 4**). All TE classes were found in the PMAN 2.1 CHR assembly, except for the Lx5c_3end, which was only found in the PERE and PLEU assemblies. Distinctively, in the PCAL assembly no TE arrays were found, while in the PMAN 2.1 CONTIGS only B1_Mus1/2 arrays were found, which may be due to the sequencing and assembly methods, considering all other assemblies had several TE classes detected.

Some of the newly identified LTRs classes did not show significant variation in the number of arrays detected between assemblies (**Table 4**). However, the LTR_PMAN_1 and LTR_PMAN_2 classes had a very high number of arrays detected in the PCAL assembly, in comparison with the other assemblies.

Considering the identification of several highly similar LTR classes across different species described above, it was important to perform a global clustering analysis of LTRs from all assemblies. This alternative method aimed to support the robustness of the assembly-focused approach described above and to confirm the classification of orphan LTRs (**Table 5**).

Table 5 – Clustering overview of joint clustering of LTRs from all assemblies of *Peromyscus* genus genomes and the total from individual clustering of LTRs from assemblies, based on their similarity. Count of arrays, % of arrays, number of classes and sub-classes for each group of clustered LTRs and for orphan LTRs, corresponding to joint and total individual clustering. Only LTR classes that are present in 2 or more assemblies were selected.

	JOINT CLUSTERING				TOTAL INDIVIDUAL CLUSTERING	
	Arrays	% Arrays	Class	Sub-class	Arrays	% Arrays
Orphan LTRs	**18384**	**79.51**			**18288**	**79.08**
Grouped LTRs	**4739**	**20.49**	**34**	**45**	**4837**	**20.92**
Satellite	**3765**	**16.28**	**3**	**8**	**3880**	**16.78**
PMSat	3686	15.94		2	3819	16.51
MurSatRep1	65	0.28		3	53	0.23
MMSAT4	14	0.06		3	8	0.03
TE	**85**	**0.37**	**4**	**7**	**56**	**0.24**
L1_Mur2_orf2	69	0.30		4	20	0.09
L1_Mur3_orf2	0	0.00		0	8	0.03
B1_Mus1/2	9	0.04		1	14	0.06
Lx5c_3end	0	0.00		0	8	0.03
RMER1C	5	0.02		1	6	0.03
ERVB4_1-I_MM	2	0.01		1	0	0.00
PMAN Class	**659**	**2.85**	**13**	**15**	**618**	**2.67**
LTR_PMAN_1	359	1.55		1	352	1.52
LTR_PMAN_2	213	0.92		2	207	0.90
LTR_PMAN_4	31	0.13		2	30	0.13
LTR_PMAN_5	13	0.06		1	12	0.05
LTR_PMAN_7	15	0.06		1	5	0.02
LTR_PMAN_9	3	0.01		1	4	0.02
LTR_PMAN_10	3	0.01		1	4	0.02
LTR_PMAN_11	0	0.00		0	4	0.02
LTR_PMAN_12	6	0.03		1	0	0.00
LTR_PMAN_13	5	0.02		1	0	0.00
LTR_PMAN_14	4	0.02		1	0	0.00
LTR_PMAN_15	3	0.01		1	0	0.00
LTR_PMAN_16	2	0.01		1	0	0.00
LTR_PMAN_17	2	0.01		1	0	0.00
PCAL Class	**88**	**0.38**	**11**	**12**	**38**	**0.16**
LTR_PCAL_1	28	0.12		1	23	0.10
LTR_PCAL_2	10	0.04		1	9	0.04
LTR_PCAL_4	6	0.03		1	0	0.00
LTR_PCAL_5	7	0.03		1	0	0.00
LTR_PCAL_6	11	0.05		1	**6**	0.03
LTR_PCAL_9	4	0.02		1	0	0.00
LTR_PCAL_16	8	0.03		2	0	0.00
LTR_PCAL_17	4	0.02		1	0	0.00
LTR_PCAL_18	4	0.02		1	0	0.00
LTR_PCAL_19	3	0.01		1	0	0.00
LTR_PCAL_20	3	0.01		1	0	0.00
PERE Class	**19**	**0.08**	**3**	**3**	**6**	**0.03**
LTR_PERE_2	11	0.05		1	0	0.00
LTR_PERE_10	5	0.02		1	6	0.03
LTR_PERE_14	3	0.01		1	0	0.00
PLEU Class	**4**	**0.02**	**1**	**1**	**0**	**0.00**
LTR_PLEU_4	4	0.02		1	0	0.00

Overall, results appear concordant with the initial grouping based on BLAST. Four classes from three different assemblies containing arrays found in chromosomes were selected for a detailed comparison between the two approaches, confirming the robustness of the proposed classification (**Figure 3B**).

LTR classes identified by the two approaches were essentially overlapping, with some orphan LTRs from the assembly-centered analysis being clustered with other groups in the global approach (**Supplementary Table 1-7**). A few LTRs did not cluster to their original group regarding the LTR_PERE_10 class. The global clustering analysis further identified ten new classes that had not been established previously (**Supplementary Table 7**).

The clustering analysis successfully identified LTR groups, including those not previously described in the Dfam database, which were named according to the aforementioned methods. Furthermore, it was possible to conclude that PMSat was the most predominant LTR class across all assemblies. The results also demonstrated the presence of highly similar LTR arrays across species, further confirming the widespread distribution of specific LTR classes in the *Peromyscus* genomes.

3.1.4 PMSat array variations between different *Peromyscus* species

After establishing the major presence of PMSat arrays across all *Peromyscus* species, it was relevant to assess similarities and differences in array sequence and structure PMSat across assemblies. For that purpose, monomer length, GC-content, and array length were determined (**Figure 4-6**). The monomers of PMSat arrays are most predominantly with 345 bp across assemblies, however, the GC-content seems to be different (**Figure 4**). To analyze these differences, a ANOVA One-way statistical test was performed to detect significant differences of monomer length and GC-content of PMSat monomers across the four species. A statistically significant difference was found across the four Peromyscus species in average monomer length ($F(3)=12.1$, $p < 0.001$). A Tukey post-hoc test was also performed to assess the significant pairwise differences between genomes, regarding PMSat monomer length. This test revealed significant pairwise differences between assemblies, except between the PMAN and PLEU assemblies, and between the PCAL and PERE assemblies (**Supplementary Table 1A**). Furthermore, the mean of monomer length across all assemblies is ~345 bp (**Figure 5A**). Regarding the average monomer GC-content, a statistically significant difference was found between the four species ($F(3) = 271.5$, $p < 0.001$). The Tukey post-hoc test revealed significant pairwise differences between assemblies, except between the PERE and PLEU assemblies (**Supplementary Table 1B**). Furthermore, the mean of GC-content of PMSat monomers from the PCAL assembly is closer to 45%, from the PERE and PLEU assemblies closer to 44%, and from the PMAN assembly closer to 43% (**Figure 5B**).

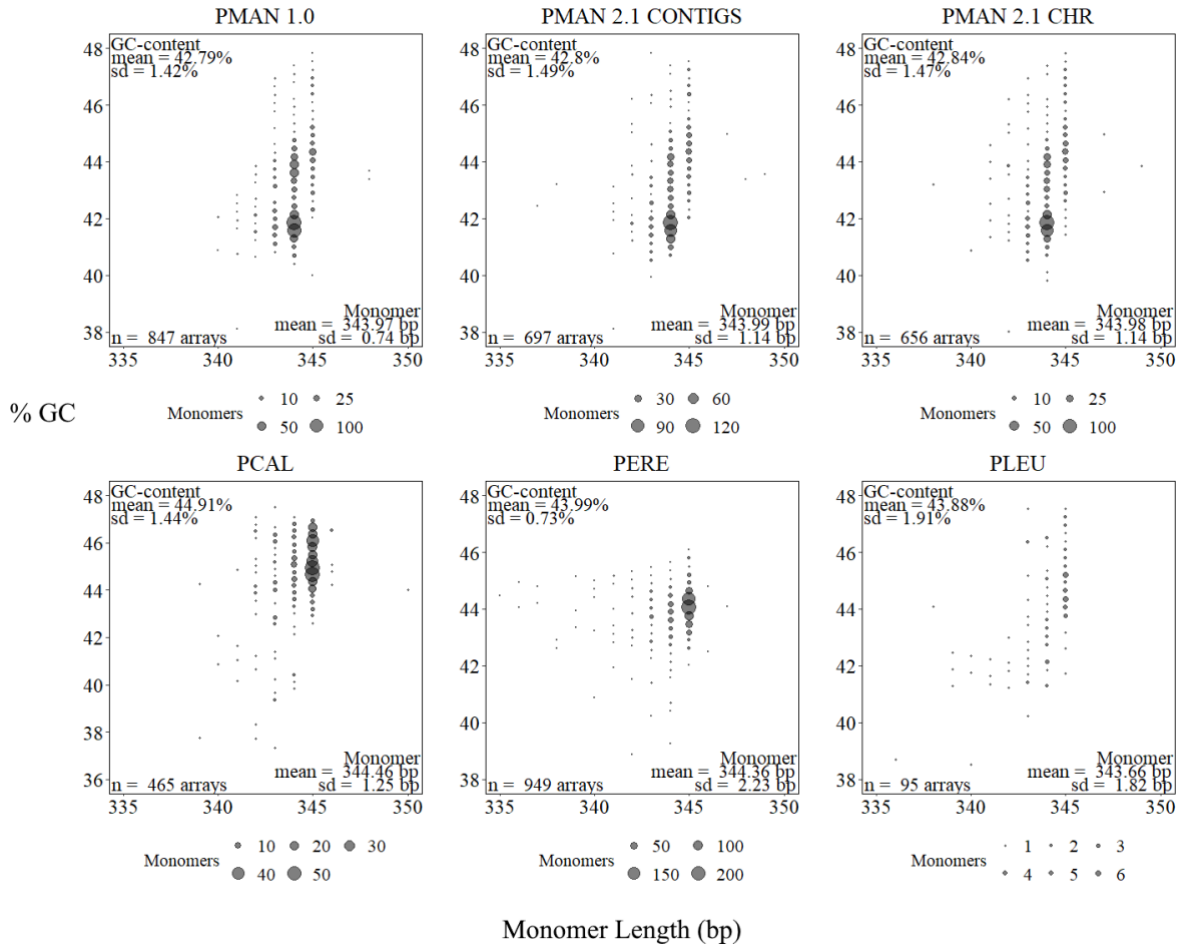


Figure 4 – Distribution of % monomer GC-content and monomer length (bp) in *Peromyscus* PMSat arrays.

The PMSat arrays have varied lengths and GC-content across all *Peromyscus* assemblies (**Figure 6**). A statistically significant difference was found across the four Peromyscus species in average array length ($F(3)=32.03$, $p < 0.001$). The Tukey post-hoc test revealed significant pairwise differences between assemblies, except between the PMAN and PCAL assemblies, and between the PERE and PLEU assemblies (**Supplementary Table 1C**). Furthermore, the mean array length of PMSat arrays from the PERE and PLEU assemblies is higher than the PMAN and PCAL assemblies (**Figure 5C**). Regarding the average array GC-content, a statistically significant difference was found between the four species ($F(3)=283.8$, $p < 0.001$). The Tukey post-hoc test revealed significant pairwise differences between assemblies, except between the PERE and PLEU assemblies, similar to the differences in mean levels of PMSat monomer GC-content (**Supplementary Table 1D**). Furthermore, the mean GC-content of PMSat arrays from the PCAL assembly is closer to 43%, from the PERE and PLEU assemblies closer to 42%, and the PMAN assembly closer to 41% (**Figure 5D**).

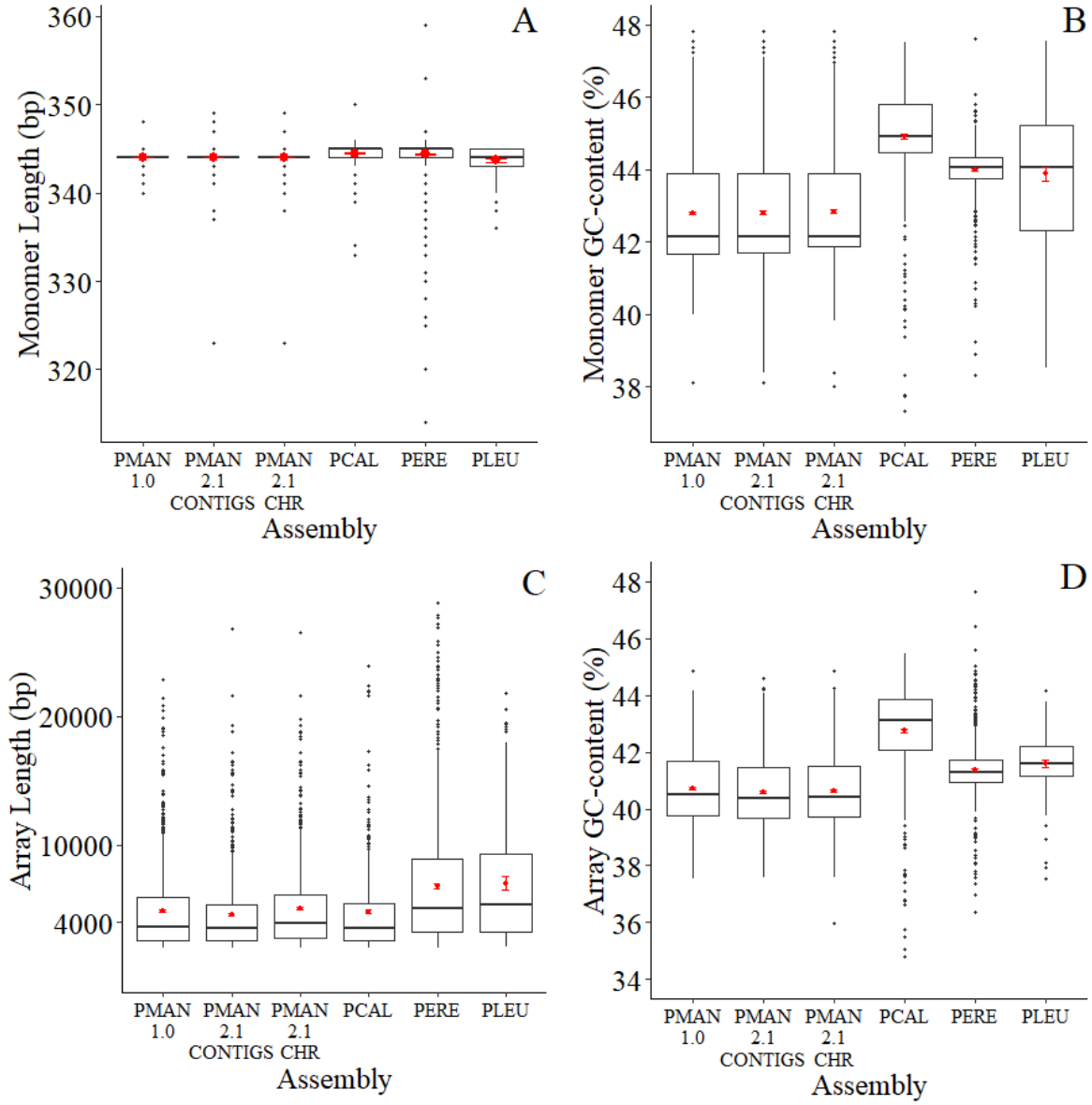


Figure 5 – Boxplot of PMSat monomer length (A), monomer GC-content (B), array length (C) and array GC-content (D) across the assemblies. Red dot represents the mean value, while the red error bars represent the standard error of the mean.

The analysis of PMSat arrays in the four *Peromyscus* species revealed that the mean monomer length predominantly remains consistent at 345 bp across assemblies, while slight differences exist in monomer GC-content among the four species. Additionally, the PMSat array mean GC-content also slightly varied among the four species, while PMSat array length showed significant differences among them, with PERE and PLEU having a higher mean array length in comparison with PMAN and PCAL. This might provide further insight into the relative positioning of PMSat arrays across the *Peromyscus* genomes.

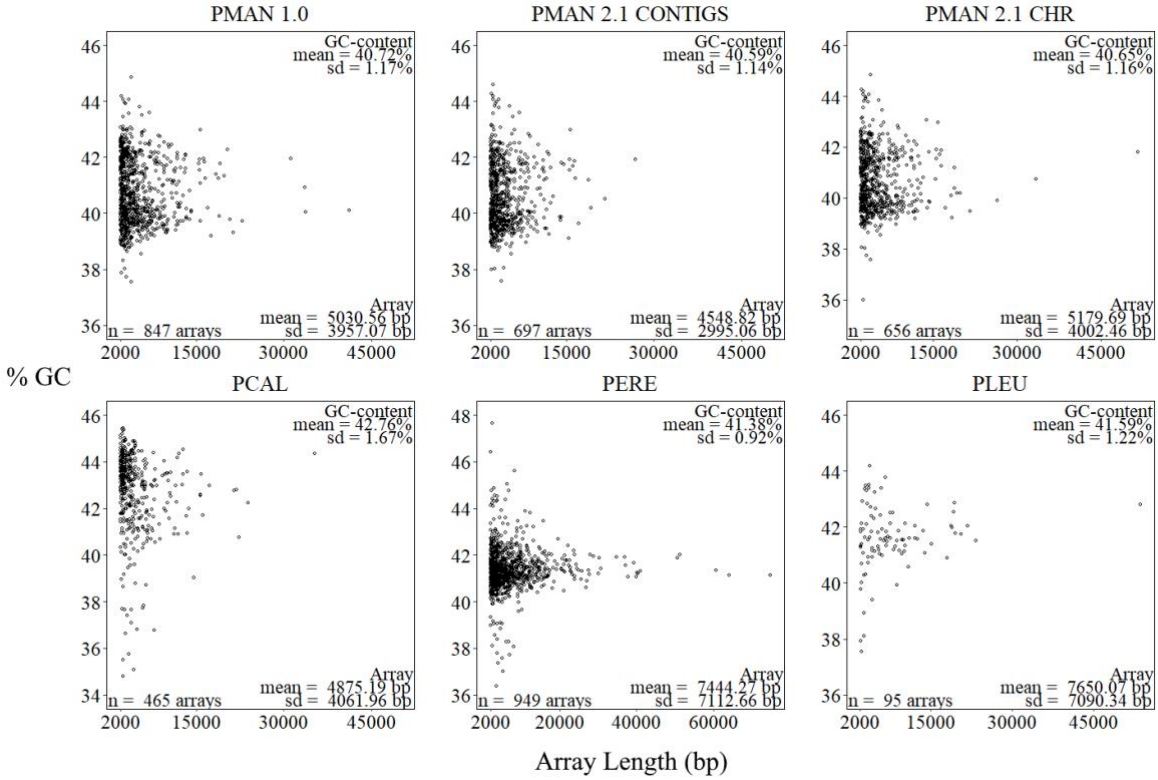


Figure 6 – Distribution of % array GC-content and array length (bp) in *Peromyscus* PMSat arrays.

3.1.5 Relative positioning of PMSat and other LTR arrays in *Peromyscus* spp. chromosomes

To further improve the annotation of LTR classes in the *Peromyscus* genus, the chromosome length of the four *Peromyscus* genomes (**Supplementary Table 8**) was used for LTR mapping on those chromosomes based on their coordinates of detection (**Supplementary Table 9**). Furthermore, by selecting a monomer of a LTR from each class and comparing its relative orientation to other LTRs (LTR in bold – **Supplementary Table 9**), it was possible to accomplish a global arrangement of LTR relative position and orientation. Additionally, cluster groups were added to this table based on the order in which class in the joint clustering table was established (**Supplementary Table 7**) (for example, as seen for the MMSAT4 class, which has 3 sub-classes, has also 3 cluster groups, 18, 33, and 47). The results obtained are described in the following subsections by LTR class.

3.1.5.1 Arrangement of PMSat arrays relative to their position on chromosomes in *Peromyscus* spp.

Chromosome 1 contains the majority of PMSat arrays detected in the PCAL and PERE assemblies, while chromosome 22 of the PMAN and PLEU assemblies contains the most PMSat arrays detected (**Supplementary Table 10**). The results of the positioning are presented in **Figure 7**. Some patterns appear to be similar between assemblies, such as the configuration of PMSat arrays in chromosome 13 for PMAN and PCAL, containing similar number of arrays and the same orientation,

although with distinct array lengths in comparison (**Supplementary Table 9**). Other patterns are difficult to outline, considering the different positioning and orientation of these PMSat arrays.

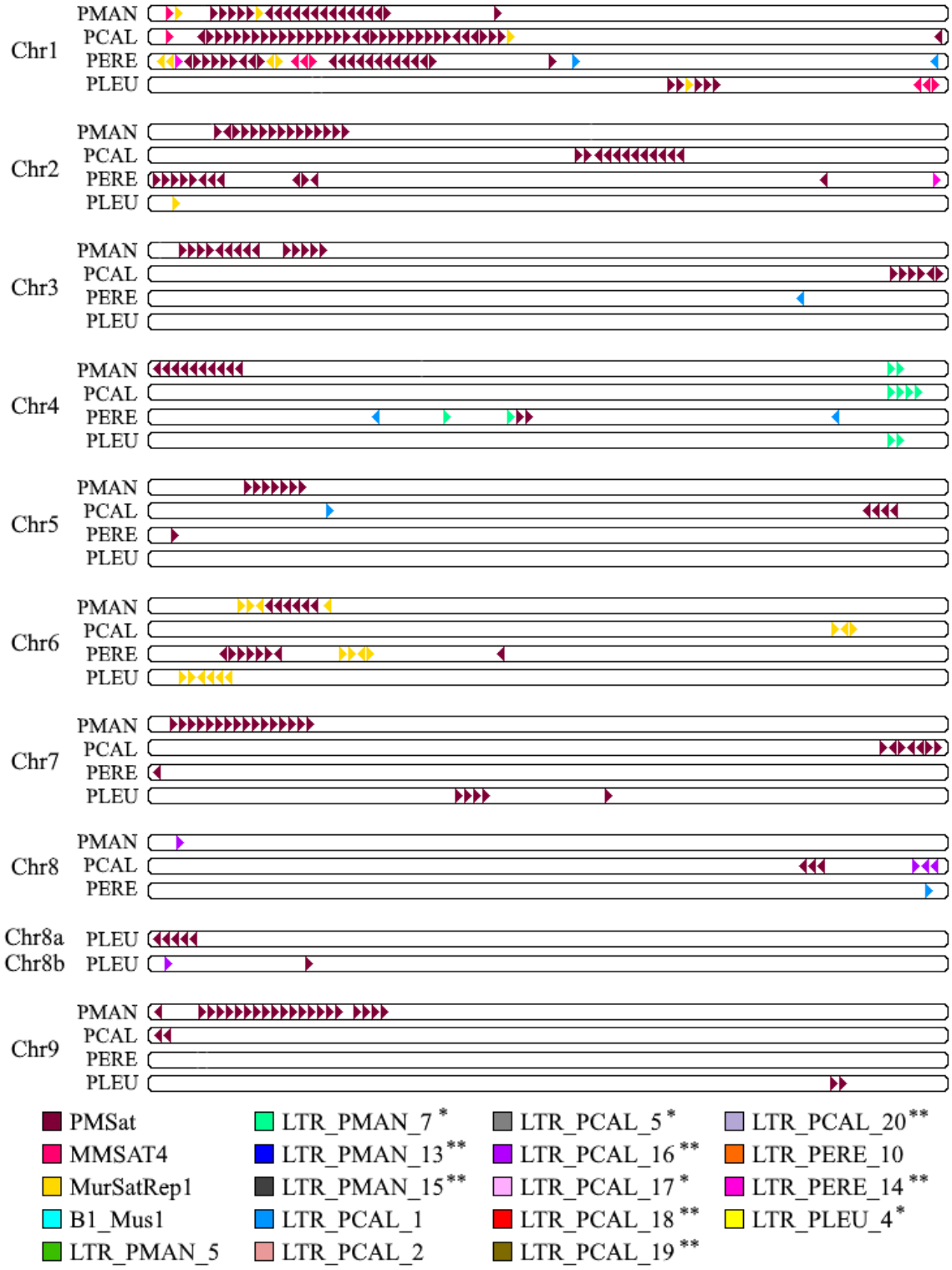


Figure 7 – Orientation and relative positioning of clustered LTR classes from *Peromyscus* genus. LTR arrays are represented on chromosomes matching each *Peromyscus* genome assembly. Each LTR array was placed based on the relative position of the coordinates of where it was detected and designated based on its LTR class. * These LTRs were not previously detected on chromosomes from at least 2 different species. ** These were new LTR groups clustered from joint clustering and followed the conditions established for relative positioning in **Figure 3**.

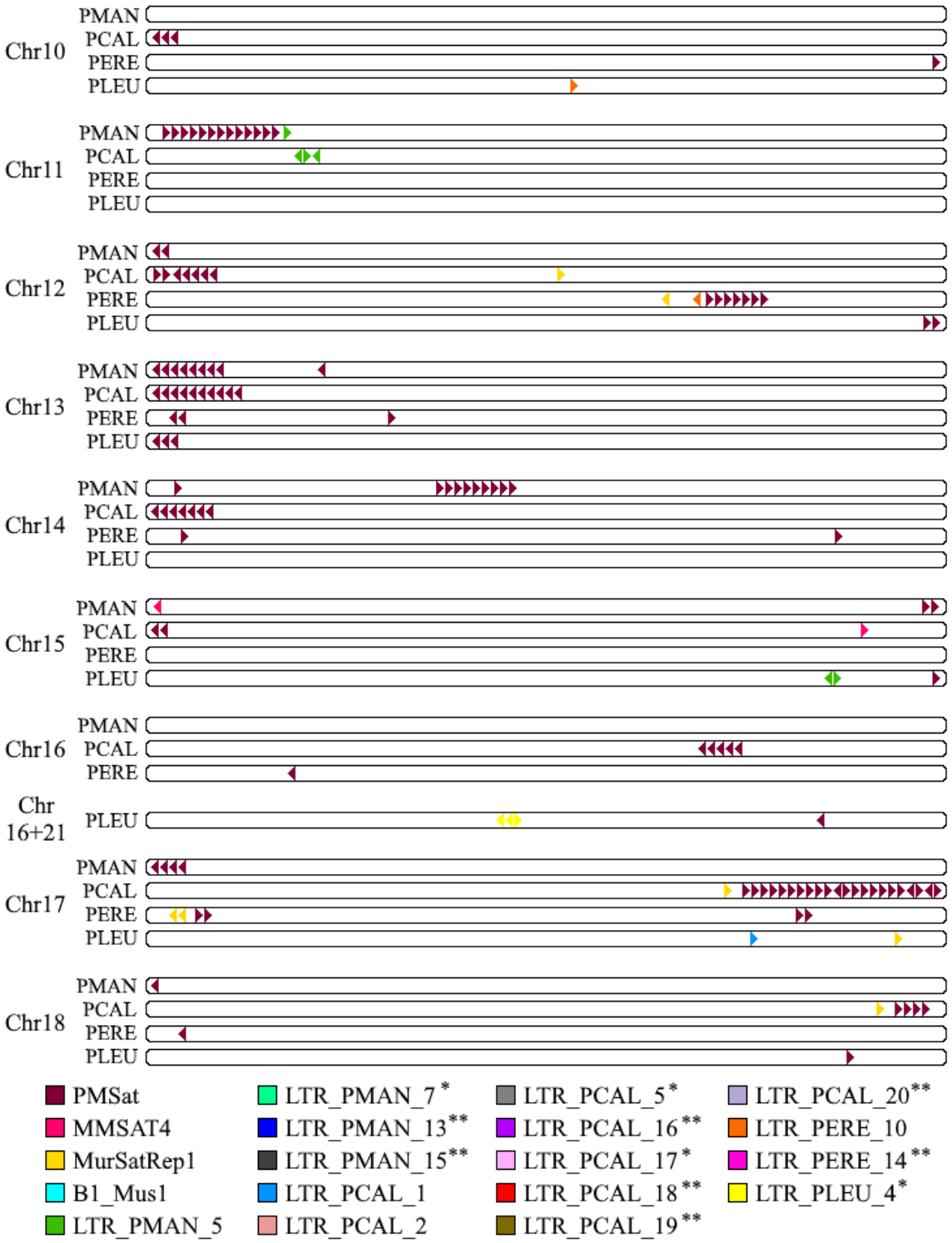


Figure 7 (cont.) – Orientation and relative positioning of clustered LTR classes from *Peromyscus* genus. LTR arrays are represented on chromosomes matching each *Peromyscus* genome assembly. Each LTR array was placed based on the relative position of the coordinates of where it was detected and designated based on its LTR class. * These LTRs were not previously detected on chromosomes from at least 2 different species. ** These were new LTR groups clustered from joint clustering and followed the conditions established for relative positioning in **Figure 3**.

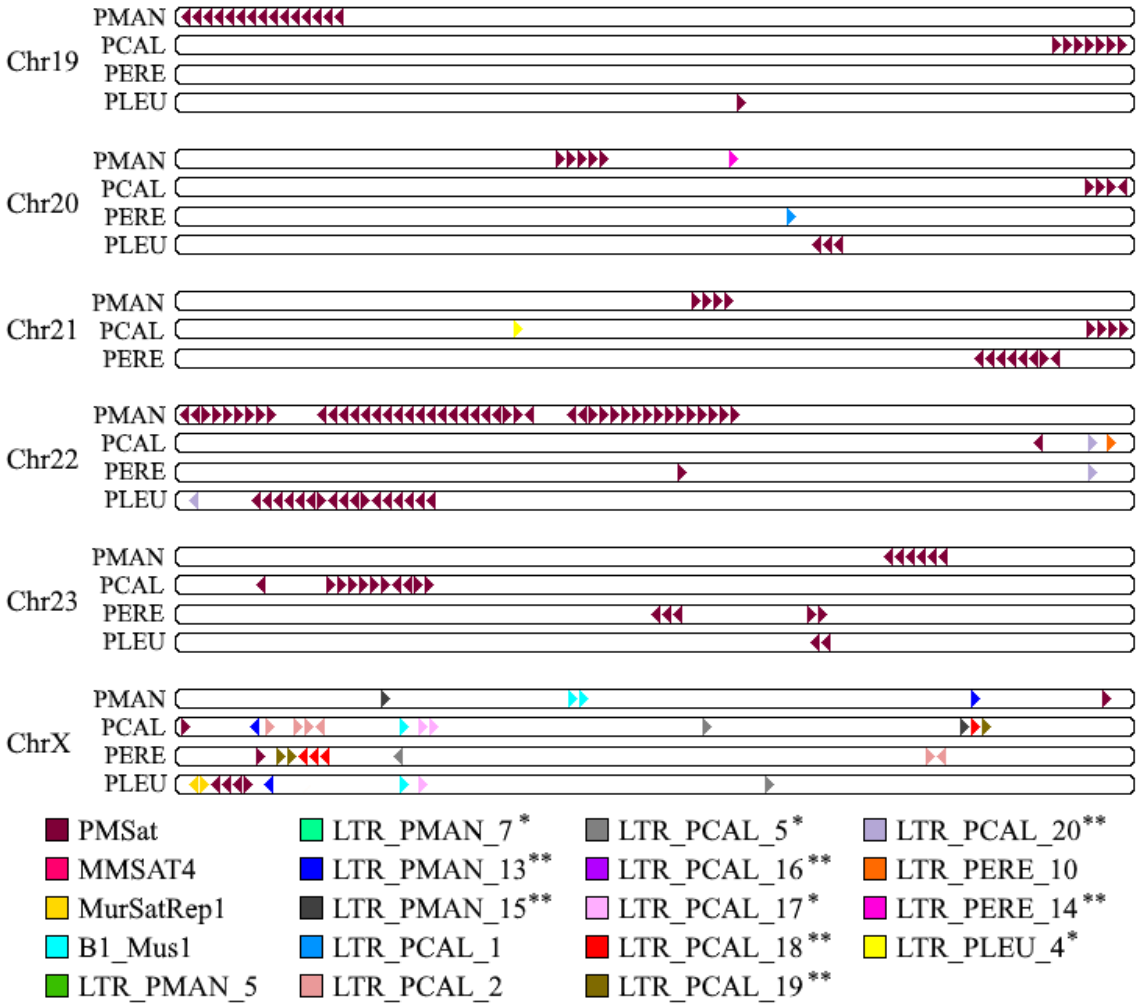


Figure 7 (cont.) – Orientation and relative positioning of clustered LTR classes from *Peromyscus* genus. LTR arrays are represented on chromosomes matching each *Peromyscus* genome assembly. Each LTR array was placed based on the relative position of the coordinates of where it was detected and designated based on its LTR class. * These LTRs were not previously detected on chromosomes from at least 2 different species. ** These were new LTR groups clustered from joint clustering and followed the conditions established for relative positioning in **Figure 3**.

Considering this, another clustering analysis was performed with the MeshClust tool (Girgis, 2022), in order to understand the connection between each PMSat monomer. This clustering method is based on identity scores without alignment, which could help in this case, considering these are highly repetitive sequences. A group of PMSat monomers clustered was selected and then their arrays represented in their relative localization on their respective chromosomes from all assemblies (**Figure 8**). PMSat arrays from this clustered group are located on two chromosomes in the PMAN assembly, on four chromosomes in the PCAL assembly, and on three chromosomes in the PERE assembly. There seems to be a pattern of the arrays being orientated on the minus strand, for the monomers in this selected clustered group. Additionally, there seems to be a prevalent pattern of two arrays on chromosome 12 on the PMAN and PCAL assemblies, however, they have slightly distinct array lengths, with the upstream from both PMAN and PCAL being ~3200 bp and ~3900 bp, respectively, and the downstream being ~3100 bp and 6000 bp, respectively (**Supplementary Table 9**; **11**). Furthermore, three of the arrays within this group seem to be positioned relatively closer to the end of the chromosome, contrary to the more common position closer to the start of chromosomes. One array is located on chromosome 5 in the PCAL assembly, and two arrays are located on chromosome 21 in the PERE assembly.

Figure 8 – Orientation and relative localization of cluster group 1 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).

Other PMSat clustered groups also showed interesting arrangement connection on chromosomes from the four *Peromyscus* genus assemblies (**Supplementary Figure 2-16**). Some PMSat clustered groups seem to be distributed across all four assemblies (**Supplementary Figure 2; 8**), with some groups showing a higher connection between the PCAL and PERE assemblies (**Supplementary Figure 3-4; 6**), another group showing connection between the PCAL, PERE and PLEU assemblies (**Supplementary Figure 5**), some groups showing a higher connection between the PMAN and PLEU assemblies (**Supplementary Figure 10-11; 13-15**), and some only showing presence mostly in the PMAN assembly (**Supplementary Figure 12; 16**). Some groups do not show significant connection due to their low number of arrays present within their respective group (**Supplementary Figure 7; 9**). Some PMSat arrays within each group have a varied relative positioning on chromosomes within and outside each assembly (**Supplementary Figure 2-7; 10; 14-15**), while others appear to be predominantly located on a specific region of the chromosomes within and outside each assembly (**Supplementary Figure 8-9; 11-13; 16**).

From these observations, it was possible to conclude that PMSat arrays have variable arrangements across all genomes, although some still retain similar arrangements in specific chromosomes. This might provide further insight into the connection of other LTR classes across the *Peromyscus* genomes.

3.1.5.2 Arrangement of novel LTR arrays relative to their position on chromosomes in *Peromyscus* spp.

To further increase the knowledge of the connections regarding other LTR classes in chromosomes of these *Peromyscus* species, the classes LTR_PMAN_5, LTR_PCAL_1, LTR_PCAL_2 and LTR_PERE_10 were initially selected for relative positioning analysis in chromosomes, due to the LTRs detected in their groups being present in the individual LTR detection analysis in chromosomes from different assemblies (see **Figure 3**). However, upon performing joint clustering analysis of LTRs from all assemblies additional extra LTR arrays from previous defined classes, and new LTR classes mentioned in section **3.2** were also considered for relative positioning mapping (**Figure 7**). The LTR arrays from the initial identified four classes within **Figure 3** conditions, LTR_PMAN_5,

LTR_PCAL_1, LTR_PCAL_2 and LTR_PCAL_10 classes demonstrate distinguishing chromosomic configurations within their classes.

The LTR_PMAN_5 arrays shown on chromosome 11 of the PMAN and PCAL assemblies are closer to the start of chromosome, while on chromosome 15 of the PLEU assembly they are shown to be closer to the end of chromosome (**Figure 7**). Interestingly, the monomer of this class is of smaller length (~63 bp), similar to the size of satellite sequences (**Supplementary Table 7**). Moreover, one the arrays in the PMAN, PCAL and PLEU assemblies has similar lengths (~15,000-16,000 bp), though it is on the plus strand in chromosome 11 of the PMAN assembly and in chromosome 15 of the PLEU assembly, while in chromosome 11 of the PCAL assembly it is on the minus strand (**Figure 7**). Additionally, another array of ~55,000 bp is also common in chromosome 11 of the PCAL assembly and chromosome 15 of the PLEU assembly, with the same orientation on the minus strand. Distinctively, while the distance between the arrays with ~55,000 bp and ~16,000 bp in the PCAL assembly is ~115 kbp, the distance between the similar arrays in the PLEU assembly is ~1.3 Mbp (**Supplementary Table 9**).

Concerning the LTR_PCAL_1 class, their lengths and orientations are different from the ones detected in the chromosomes of the PCAL assembly. The array in the PCAL assembly has a copy number of ~18, while in the PERE and PLEU assemblies the arrays have a copy number of ~2-3. Moreover, the arrays are localized in distinct chromosomes in the PERE assembly (1, 3, 4, 8 and 20), while in the PCAL assembly only being detected on chromosome 5 and in the PLEU assembly on chromosome 17. The monomer length of this class ranges from 1025-1865 bp (**Supplementary Table 7**). It does not seem to show a preferable orientation or localization (**Figure 7**).

Regarding the LTR_PCAL_2 class, it was only detected in chromosome X in two of the *Peromyscus* genomes (**Figure 7**). There are four arrays closer to the start of chromosome X in the PCAL assembly, while only two arrays are closer to the end of chromosome X in the PERE assembly. The monomer length of this class ranges from 1707-1784 bp (**Supplementary Table 7**). Only one pattern concerning the orientation is prevalent in both assemblies, being that of the two arrays closely together and on opposite strands, the one upstream being on the plus strand, while the other downstream being on the minus strand (**Figure 7; Supplementary Table 9**). However, their lengths are distinct between them, whereas in the PCAL assembly, the one upstream and on the plus strand has ~5100 bp, while in the PERE assembly, the one upstream and also on the plus strand has ~7600 bp. The one array downstream and on the minus strand in the PCAL assembly has ~8700 bp, while in the PERE assembly, the most downstream array and on the minus strand has ~5000 bp (**Figure 7; Supplementary Table 9**).

LTR_PERE_10 class seems to also have monomers of higher length ranging with 1398-1980 bp, however, there is only one array detected on distinct chromosomes of the PCAL, PERE and PLEU assemblies (**Figure 7; Supplementary Table 9**). In chromosome 22 of the PCAL assembly, its array is closer to the end of the chromosome and on the plus strand (**Figure 7; Supplementary Table 9**). In chromosome 12 of the PERE assembly, its array seems to be located on the middle part of the end of the chromosome and on the minus strand (**Figure 7; Supplementary Table 9**). In the PLEU assembly was located around the middle region in the chromosome 10 and on the plus strand (**Figure 7; Supplementary Table 9**). Additionally, all of these arrays have different lengths, with ~4100 bp, ~2800 bp and ~6000 bp in the PCAL, PERE and PLEU assemblies, respectively (**Supplementary Table 9**).

There were also more classes that previously did not show any LTRs being detected on chromosomes that after joint clustering revealed to be present. One of these was the LTR_PMAN_7 class. With a monomer length of 69 bp, it was detected only in chromosome 4 of all four genomes. Moreover, it was on the plus strand of all four genomes (**Figure 7**). Although all of the arrays varied in

length, ranging with ~3400-16,000 bp, there was a similar array in the PMAN and PLEU assemblies, with ~16,000 bp (**Supplementary Table 9**).

LTR_PCAL_5 class also had new arrays added to its cluster group that were detected in chromosomes. Specifically in this one, the arrays were detected in chromosome X of the PCAL, PERE and PLEU assemblies (**Figure 7**). The monomer length of this class is ~310 bp (**Supplementary Table 7**). Although on different strands, there is one array in the PCAL assembly similar in length to that of the one in the PERE assembly, with ~3300-3800 bp, and also with similar localization in the chromosome (**Figure 7**; **Supplementary Table 9**). The array in the PCAL assembly is much larger in length, with ~29,000 bp and closer to the start of the chromosome.

LTR_PCAL_17 class also had additional arrays detected in other assemblies. This class also only had arrays detected in chromosome X, however, in this case exclusive to the PCAL and PLEU genomes (**Figure 7**). The monomer length is 51 bp (**Supplementary Table 7**). These arrays were only on the plus strand and in similar localization in the chromosome X of these genomes (**Figure 7**). The length of the two arrays detected in the PCAL assembly are of smaller size (~2500 bp and 4000 bp) compared to the one array detected in the PLEU assembly (~12,000 bp) (**Supplementary Table 9**).

LTR_PLEU_4 class also had more arrays detected after joint clustering of all assemblies. The monomer length ranged from 1554-1601 bp (**Supplementary Table 7**). All of the arrays detected were located near the middle of chromosome 21 in the PCAL assembly and chromosome 16+21 in the PLEU assembly (**Figure 7**). Additionally, the array detected in the PCAL assembly is on plus strand, while the three arrays detected in the PLEU assembly are on the minus>minus>plus order (**Figure 7; Supplementary Table 9**).

Regarding the new classes of LTRs identified based on the joint clustering analysis, there were the LTR_PMAN_13, LTR_PMAN_15, LTR_PCAL_18 and LTR_PCAL_19 classes that showed to only have arrays in chromosome X. The monomer length of these classes is ~72 bp, ~72 bp, 77 bp and 126-168 bp, respectively (**Supplementary Table 7**). The LTR_PMAN_13 class detected in the PMAN assembly is on the plus strand and closer to the end of the chromosome X, while the arrays from this class detected in the PCAL and PLEU assemblies are on the minus strand and closer to the start of the chromosome X (**Figure 7**). The LTR_PMAN_15 class has arrays detected on opposite sides of the chromosome X in the PMAN and PCAL assemblies, while maintaining their orientation on the plus strand but with different array length (**Figure 7**; **Supplementary Table 9**). The LTR_PCAL_18 class was also detected on opposite sides of the chromosome X in the PCAL and PERE assemblies, having arrays with similar length (~2300-3400 bp), with the exception of one array in the PERE assembly (~28,000 bp) (**Supplementary Table 9**). The array in the PCAL assembly was on the plus strand and the three arrays in the PERE assembly were on the minus strand (**Figure 7**; **Supplementary Table 9**). The arrays from the LTR_PCAL_19 class had similar length (2700-3300 bp) (**Supplementary Table 9**) and were shown to be on the plus strand in chromosome X of the PCAL assembly, while the two arrays detected in the PERE assembly had one array on the minus strand and another on the plus strand (**Figure 7**; **Supplementary Table 9**).

Chromosome 8 of the PMAN, PCAL and PLEU assemblies also had exclusive detection of the LTR_PCAL_16 class, with one array located at the start of chromosome 8 in the PMAN assembly and chromosome 8b in the PLEU assembly, while three arrays were located at the end of chromosome 8 in the PCAL assembly (**Figure 7**). The monomer length of this class is 56 bp (**Supplementary Table 7**). Moreover, there is one array that had similar length (~2000 bp) in both the PCAL and PMAN assemblies, while only array in the PLEU assembly has a larger copy number of the monomer in its array (**Supplementary Table 9**). A similar arrangement was observed regarding the LTR_PCAL_20 class,

where it only had arrays detected on chromosome 22 (**Figure 7**). It had an array on the plus strand closer to the end of chromosome 22 in the PCAL and PERE assemblies, while the array in the PLEU assembly was on the minus strand and closer to the start of chromosome 22 (**Figure 7**). The monomer length of this class was 99 bp (**Supplementary Table 7**). Furthermore, the length of these arrays is different in PCAL, PERE and PLEU assemblies (~2400 bp, ~4000 bp and ~3000 bp, respectively) (**Supplementary Table 9**). The LTR_PERE_14 class had arrays located in different chromosomes, with one array located on the minus strand closer to the start of chromosome 1 in the PERE assembly, and another array also on the minus strand of chromosome 2 in the PERE assembly, though closer to end of the chromosome (**Figure 7**). Another array that was detected in the PMAN assembly was on the plus strand closer to the middle section of chromosome 20 (**Figure 7**). The monomer length of this class ranges with 1845-1947 bp (**Supplementary Table 7**), with the length of arrays being similar (~3700-3900 bp) (**Supplementary Table 9**).

3.1.5.3 Arrangement of MMSAT4 arrays relative to their position on chromosomes in *Peromyscus* spp.

The MMSAT4 family showed 3 sub-classes detected in chromosomes across all the assemblies. The monomer length of two of the sub-classes is ~84 bp, while the other sub-class had a length of ~168 bp (**Supplementary Table 7**). One of the sub-classes had an array on the plus strand closer to the start of chromosome 1 in the PMAN assembly, two arrays closely together on the minus strand closer to the start of chromosome 1 in the PERE assembly, and two arrays closely together on the minus strand closer to the end of the chromosome 1 in the PLEU assembly (**Figure 7**). One of the arrays on the three assemblies (PMAN, PERE and PLEU) regarding the sub-class from cluster group 18, has similar length (2100-2200 bp), while another array from the same sub-class that was upstream to the previous array in the PERE and PLEU assemblies has a larger length (3000 bp and 2500 bp, respectively) (**Figure 7; Supplementary Table 9**). The arrays of the sub-class from the cluster group 47 are only located in chromosome 1 of the PCAL, PERE and PLEU assemblies (**Figure 7; Supplementary Table 9**). In the PCAL and PERE assemblies the arrays were on the plus strand closer to the start of chromosome 1, while in the PLEU assembly, though also being on the plus strand, the array was closer to the end of chromosome 1 (**Figure 7**). The length of these arrays is similar, with 2100 bp (**Supplementary Table 9**). There seems to be an arrangement of three arrays of the MMSAT4 class on chromosome 1 in the PERE and PLEU assembly, with the two most upstream arrays belonging to the sub-class from the cluster group 18, while the array downstream of those two arrays belonging to the sub-class from the cluster group 47, and though these maintain their orientation, they are located on opposite sides of the chromosome (**Figure 7**; **Supplementary Table 9**). Moreover, the distance between them is also similar, with an interval of ~500 kbp and ~407 kbp between the first two arrays in the PERE and PLEU assemblies, respectively, and with an interval of ~362 kbp and ~285 kbp between the last two arrays of this set in the PERE and PLEU assemblies (**Supplementary Table 9**). In chromosome 15 of the PMAN and PCAL assemblies was also detected another sub-class of MMSAT4 from the cluster group 33. Although the arrays having similar length (~2200 bp), the one detected in the PMAN assembly was on the minus strand and closer to the start of the chromosome, the one detected in the PCAL assembly was on the plus strand and closer to the end of the chromosome (**Figure 7**; **Supplementary Table 9**).

3.1.5.4 Arrangement of MurSatRep1 arrays relative to their position on chromosomes in *Peromyscus* spp.

The MurSatRep1 class also had three sub-classes in all four genomes. The monomer length of two sub-classes is ~84 bp, while one of the sub-classes had varying length, with 168-269 bp (**Supplementary Table 7**). For the most abundant sub-class (from cluster group 4), there were arrays detected in chromosome 1 and 6 of the PMAN assembly, in chromosomes 6, 17 and 18 of the PCAL assembly, in chromosomes 1, 6, 12 and 17 of the PERE assembly, and in chromosomes 2, 6, 17 and X of the PLEU assembly (**Figure 7**). Particularly regarding this sub-class, there seems to be a specific arrangement in chromosome 6 of all assemblies, with one array on the minus strand with ~2100 bp, followed by an array on the plus strand with ~2900 bp (**Supplementary Table 9**). However, in the PMAN assembly this set of arrays may be inverted, as the array with ~2900 bp is upstream of the array with ~2100 bp, while also being closer to the end of the chromosome 6 (**Figure 7**; **Supplementary Table 9**). Another sub-class of MurSatRep1 was only detected in chromosome 1 of all *Peromyscus* genomes. Interestingly, the length of the arrays in this sub-class (from cluster group 17) in all the genomes was similar (~3000 bp), with exception of one array in the PERE assembly (~2100 bp) (**Supplementary Table 9**). However, these arrays have different orientation and location in the chromosomes. In the PMAN and PCAL assemblies the arrays were on the plus strand and around the middle section of the start of chromosome 1, while the array in the PERE assembly was on the minus strand and closer to the start of chromosome 1, and the array in the PLEU assembly was on the plus strand and closer to the end of chromosome 1 (**Figure 7**; **Supplementary Table 9**). There was also one sub-class that only had one array located in chromosome 12 of the PCAL assembly, on the plus strand and in the middle section of the chromosome (cluster group 46) (**Figure 7**; **Supplementary Table 7; 9**).

From the relative positioning and orientation of both PMSat and other LTR classes in the chromosomes across all genomes it could be concluded that, although they have diverse arrangements, it still exists some arrays that have similarities pertaining to these characteristics across all *Peromyscus* genomes. Furthermore, the analysis of the PMSat satellite family found in the *Peromyscus* genus allowed for a detailed portrayal of its arrangements, showing that arrays with a high degree of similarity had preferential chromosome positioning across all four genomes. This analysis also revealed that these methodologies had relevant results, which might help with the exploration of LTRs in the human genome, mainly with the characterization of human satDNA families.

3.2. Improving the characterization of LTR arrays and satDNA in the human genome

This section describes the results obtained in the frame of the second objective of this project: to take advantage of the recently available Nanopore long read sequencing methodology to gain more in-depth knowledge on the organization and function of LTRs and satDNA in the human genome. The analysis of LTRs in the human genome was focused on satellite DNA, with a particular interest in assessing peri/centromeric region arrangements in chromosomes. These regions have proven difficult to characterize, namely due to their highly contiguous repetitive sequence composition (M. Aldrup-MacDonald & Sullivan, 2014; Miga, 2015; Miga et al., 2014). Consequently, highly contiguous sequencing technologies such as nanopore sequencing technology (Jain et al., 2018) have emerged and shown to be important in the characterization of long tandem repeats arrays (Kinkar et al., 2021; Vondrak et al., 2020). We took advantage of two whole genome datasets that are publicly available (NA12878 and CHM13) and one additional dataset was generated in-house by low-depth sequencing of total DNA and sorted chromosomes of the human cell line GM03417 with a ROB t(14;21) translocation to investigate the impact of unusual chromosomal rearrangement. Furthermore, an in-depth study of a specific human satellite family, HSat1A, was performed, involving genomic copy number and transcription product diversity analysis.

3.2.1 Detection and analysis of human LTRs in nanopore sequencing datasets

Concurrently with this study, nanopore sequencing was performed in-house on total DNA from the GM03417 cell line, using a total of two flowcells. This yielded a dataset of 20.3 Gbps that was analyzed in the frame of this dissertation project, corresponding to a total of 6.5x the estimated genome coverage. Sequencing metrics for the publicly available NA12878 and CHM13 and this in-house dataset were calculated for comparative purposes (**Table 6**).

Cell culture and genomic DNA isolation of cell line GM03417 containing ROB t(14;21) for Oxford Nanopore Technology (ONT) sequencing showed overall improvement on the percentage of long reads when compared to the publicly available NA12878 dataset for the same cell line and technology, with 6.27% of total reads being above 30 kbps, while the NA12878 dataset showed only 2.70% of total reads to be above 30 kbps. However, the NA12878 dataset produced a higher percentage of ultra long reads, with 0.48% of total reads being above 100 kbps, while the GM03417 dataset produced 0.03% of total reads being above 100 kbps. The CHM13 dataset showed better overall results (**Table 6**).

Table 6 – Overview analysis of the GM03417 cell line and the NA12878 and CHM13 datasets. Quantification and distribution of raw reads belonging to each dataset**.** Additional ratios were analyzed based on each relevant category (30 kbp, 100 kbp) relative to the total reads. Sequencing coverage was also performed relative to the total bases and relative to the > 30 kbp reads. Gpb per flow cell was measured based on the total Gbp detected versus the number of runs (flow cells) performed.

	GM03417	NA12878	CHM13
Total Gbp	20.3	132.9	357.4
Total reads	2,326,357	15,666,888	28,449,385
Mean read length (bp)	8,727	8,485	12,563
Reads > 800bp	2,075,550	14,349,948	13,560,254
Mean read length > 800bp (bp)	9,748	9,220	26,138
Longest read (bp)	255,987	2,974,128	7,213,890
Long reads (>30 kbp)	145,927	423,418	3,933,689
% 30 kbp/total (reads)	6.27	2.70	13.83
Ultra long reads (>100 kbp)	708	75,353	603,417
% 100 kbp/total (reads)	0.03	0.48	2.12
Estimated human genome coverage	6.5	42.9	115.3
Nucleotide - Total (bp)	20,302,191,845	132,931,102,331	357,422,186,747
Nucleotides > 800 bp (bp)	20,232,377,516	132,314,118,262	354,431,747,269
Nucleotides > 30 kbp (bp)	6,495,465,843	30,446,578,605	266,749,095,583
Coverage (reads > 30 kbp)	2	9.5	83.4
Nucleotides > 100 kbp (bp)	82,360,206	13,343,989,706	97,481,032,422
% 30 kbp/total (nt)	31.99	22.90	74.63
Runs (flow cells)	2	52	481
Gbp per flow cell	10.15	2.56	0.74

Detection of tandem repeats in raw reads above 30 kbps was the selected method for this analysis with the Tandem Repeat Finder tool (Benson, 1999), in an effort to assess the LTR detection performance after sequencing on long contiguous reads. Moreover, investigation of LTR diversity largely found in peri-centromeric and centromeric regions was the focus of this analysis, comparing between the sequencing data of the GM03417 cell line and the other two cell lines without chromosomal abnormalities (GM12878 and CHM13). When comparing the detection of LTRs in the GM03417, NA12878 and CHM13 datasets, the latter showed a much higher number of detected arrays (**Table 7**).

Table 7 – Quantification of nucleotide sum from LTR arrays detected on GM03417, NA12878 and CHM13 datasets.

	GM03417	NA12878	CHM13
Total nt from reads (bp)	20,302,191,845	132,931,102,331	357,422,186,747
LTR arrays nt sum (bp)	105,263,128	527,018,833	7,535,000,448
% LTR arrays nt sum/Total nt from reads	0.52	0.40	2.11

Following this, a clustering analysis with blastclust (Altschul et al., 1990) was performed on the LTRs detected (**Supplementary Table 11-13**). The top 20 clustered groups regarding number of detected arrays from each dataset and corresponding to > 93% of all LTRs in clustered groups, were selected to simplify this analysis while maintaining a broad number of LTR groups.

A sequence search was performed using Dfam (Hubley et al., 2016; Storer et al., 2021) to identify known repetitive sequences, corresponding to each group. The results from this analysis are presented in **Table 8** and are discussed in detail next.

Table 8 – Clustering overview of the top first 20 clustered groups of LTRs from GM03417, NA12878 and CHM13 datasets. Count of arrays, % of arrays, number of classes and sub-classes for each group of clustered LTRs and for orphan LTRs, corresponding to each dataset, from the top first 20 clustered groups (**Supplementary Table 11-13)**. TE – Transposable Element. RE – Repetitive Element.

	GM03417				**NA12878**				**CHM13**			
	Arrays	% Ar-rays	Class	Sub-class	Arrays	% Ar-rays	Class	Sub-class	Arrays	% Ar-rays	Class	Sub-class
Orphan LTRs	**1,143**	**15.19**			**14,266**	**35.90**			**42,248**	**12.04**		
Grouped LTRs	**6,383**	**84.81**	**16**	**20**	**25,475**	**64.10**	**15**	**20**	**308,743**	**87.96**	**19**	**20**
Satellite	**6,149**	**81.70**	**7**	**11**	**21,709**	**54.63**	**7**	**9**	**281,749**	**80.27**	**10**	**10**
ALR	4,858	64.55		2	15,812	39.79		1	207,000	58.98		1
BSR	467	6.21		1	2,165	5.45		1	31,999	9.12		1
CER	49	0.65		1	211	0.53		1	3,212	0.92		1
HSAT1A (SAR)	271	3.60		1	1,623	4.08		1	4,634	1.32		1
HSAT2	485	6.44		4	1,762	4.43		3	28,925	8.24		1
HSAT4	0	0.00		0	0	0.00		0	381	0.11		1
ACRO1	11	0.15		1	80	0.20		1	1,355	0.39		1
SST1	8	0.11		1	56	0.14		1	577	0.16		1
GSATII	0	0.00		0	0	0.00		0	461	0.13		1
SATR1	0	0.00		0	0	0.00		0	3,205	0.91		1
TE	**8**	**0.11**	**2**	**2**	**40**	**0.10**	**1**	**1**	**1,302**	**0.37**	**1**	**2**
L1 retrotransposon	4	0.05		1	40	0.10		1	1,302	0.37		2
MER20	4	0.05		1	0	0.00		0	0	0.00		0
RE	**11**	**0.15**	**2**	**2**	**1,820**	**4.58**	**2**	**5**	**2,957**	**0.84**	**3**	**3**
MER5A1r	5	0.07		1	55	0.14		1	820	0.23		1
teucerv2_3edge	6	0.08		1	0	0.00		0	0	0.00		0
Walusat	0	0.00		0	0	0.00		0	621	0.18		1
Simple Repetition	0	0.00		0	1,765	4.44		4	1,516	0.43		1
H_LTR Class	**50**	**0.66**	**5**	**5**	**351**	**0.88**	**5**	**5**	**3,918**	**1.12**	**5**	**5**
H_LTR_1	16	0.21		1	137	0.34		1	1,365	0.39		1
H_LTR_2	11	0.15		1	62	0.16		1	739	0.21		1
H_LTR_3	10	0.13		1	0	0.00		0	790	0.23		1
H_LTR_4	9	0.12		1	43	0.11		1	601	0.17		1
H_LTR_5	4	0.05		1	35	0.09		1	423	0.12		1
H_LTR_6	0	0.00		0	74	0.19		1	0	0.00		0

In all datasets, the majority of LTRs could be grouped in clusters, with the most abundant groups corresponding to satellite sequences. The ALR (Human alpha satellite) was the most abundant LTR detected, as expected, reaching above 50% of all LTRs detected in the GM03417 and CHM13 dataset,

and almost 40% for NA12878. The BSR (Beta satellite repeat) was the second most abundant LTR detected in the NA12878 and CHM13 dataset, however, slightly below HSAT2 percentage in the GM03417 dataset. HSAT1A (SAR) and HSAT2 were slightly less abundant, with a major distinction in CHM13, that had larger abundance of HSAT2 detected in comparison to HSAT1A (~7%). Interestingly, the HSAT2 family had four and three sub-classes detected in the GM03417 and NA12878 datasets, respectively, while the CHM13 dataset only had one.

Other satellites, such as CER, ACRO1 and SST1, were detected in all datasets, but in lower count in comparison (**Table 8**). The HSAT4, GSATII and SATR1 satellites were not detected in the GM03417 and NA12878 datasets, only in the CHM13 dataset. Some transposable elements were detected, with the L1 retrotransposon family being the major represented family in this category. Moreover, some repetitive elements were detected in all 3 datasets (MER5A1r), while some were only detected in the GM03417 dataset (teucerv2_3edge) and others only in the CHM13 dataset (Walusat). NA12878 and CHM13 were the only datasets to show arrays of simple repetition elements within the first 20 groups, containing four and one sub-classes, respectively.

Un-identified LTR groups (i.e., not described in the Dfam database) were compared using the BLAST tool (Z. Zhang et al., 2000) to identify similar groups across datasets and named in decreasing order of group size (H_LTR_1 to H_LTR_6). The H_LTR_1, H_LTR_2, H_LTR_4 and H_LTR_5 groups were detected in all the three datasets. However, the H_LTR_3 group was only detected in the GM03417 and CHM13 datasets, while the H_LTR_6 group was only present in the NA12878 dataset.

Additionally, it was of interest to assess if there were any differences in regards to satDNA detection on chromosomes 14 and 21 from the GM03417 cell line, which has the translocated chromosome t(14;21), with the other two cell lines without chromosomal abnormalities (GM12878 and CHM13 datasets). For this, alignment of reads containing at least one of the top four satDNA clustered groups (ALR, BSR, HSat2 and HSat1A) with chromosome 14, 21 or both from the human genome assembly GRCh38.p13 was performed (**Table 9**). There are no apparent significant percentual distinctions between the three datasets, which could indicate insufficient data from the GM03417 cell line in order to assess the translocation t(14;21).

This analysis allowed to show the capacity of nanopore sequencing for the detection and characterization of long tandem repeats (LTRs), as with fewer flowcells used, it was still possible to produce similar results in percentage. LTR analysis revealed the dominance of satellite sequences present in each dataset, albeit with some differences in percentage across the datasets. Furthermore, when assessing satDNA detection on chromosomes 14 and 21, no apparent differences were observed among the datasets, which a flow-sorting of chromosomes might prove beneficial for the analysis of translocation t(14;21).

Table 9 – Alignment of reads from each dataset to chromosomes 14 and 21 from reference human genome GRCh38.p13. The first row of each dataset indicates the number of reads that aligned with each chromosome or the combination of both (chromosome 14 and 21). The percentage calculation is based on the number reads that aligned with each chromosome or both. Each satDNA selected for this analysis was based on the top 4 clustered groups from each dataset (**Supplementary Table 11-13**).

		Chr14		Chr21		Chr14 + Chr21	
		Reads	**%**	**Reads**	**%**	**Reads**	**%**
GM03417	-	86,074	-	86,798	-	84,893	-
	HSAT1A	231	0.27	110	0.13	110	0.13
	HSAT2	373	0.43	373	0.43	373	0.44
	ALR	2,684	3.12	2,684	3.09	2,684	3.16
	BSR	96	0.11	123	0.14	89	0.1
NA12878	-	338,663	-	337,628	-	322,940	-
	HSAT1A	1,062	0.31	413	0.12	413	0.13
	HSAT2	1,136	0.34	1,136	0.34	1,136	0.35
	ALR	6,937	2.05	6,936	2.05	6,935	2.15
	BSR	364	0.11	460	0.14	332	0.1
CHM13	-	3,780,312	-	3,799,642	-	3,774,591	-
	HSAT1A	3,937	0.1	2,930	0.08	2,930	0.08
	HSAT2	25,358	0.67	25,358	0.67	25,358	0.67
	ALR	105,574	2.79	105,574	2.78	105,574	2.8
	BSR	6,428	0.17	7,955	0.21	6,153	0.16

3.2.2 Assessment of sequencing data from flow sorted chromosomes of the ROB t(14;21) GM03417 cell line

A different approach was taken to improve on the coverage limitation of the sequencing performed to study the ROB t(14;21) on the GM03417 cell line. Flow Cytometry (flow-sorting) of chromosomes has shown great progress in their isolation and enrichment to simplify assembly and study of individual chromosomes (Cápal et al., 2016; Doležel et al., 2021; Kuderna et al., 2019). This technique relies on chromosome size for discrimination. As a consequence, some groups of chromosomes will be sorted together, with translocated chromosome rob(14:21) being captured along with chromosomes 9, 10, 11, and 12.

Flow-sorting of chromosomes to capture rob(14:21) was performed on the GM03417 cell line, following Kuderna et al. protocol. Subsequently, sequencing was performed using ONT on these flow-sorted chromosomes, with the same intent as before. In this pilot assay, 330,000 chromosomes were isolated, which is 10x less than the referenced number by Kuderna et al. Notwithstanding, the flow-sorting and sequencing performed was successful, as 182 of the 226 reads that were obtained showed

alignment with the intended flow-sorted chromosomes (**Figure 9**). Moreover, the mean length of the reads was ~1790 bp, with read length ranging from 124 bp to 16,777 bp. Unfortunately, due to the low sequencing data yield, further analysis to characterize the translocated region was not possible.

Although the low sequencing data yield did not allow for the study of the ROB t(14;21) in the GM03417 cell line, it was still possible to conclude that these methods are able to provide an improvement to chromosome coverage, albeit with the sufficient quantity of flow-sorted chromosomes.

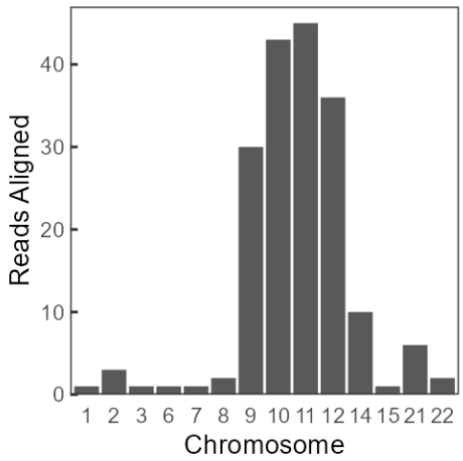


Figure 9 – Alignment of reads from flow-sorted chromosomes of nanopore sequencing on GM03417 cell line. The reads were aligned to each 24 individual chromosomes from the GRCh38.p13 genome assembly.

3.2.3 HSat1A genomic copy number and transcript analysis

As satDNA genomic occurrence and preservation seem to be associated with long-range organization and structure of peri/centromeric regions (Plohl et al., 2014), the HSAT1A satellite, which is highly present in these regions (Altemose et al., 2022), is of particular interest for our study. The human satellites have shown to be transcriptionally active (Ugarkovic, 2005), however, the role of this is still not fully understood. To further increase knowledge on these repetitive sequences, HSat1A was selected for a focused analysis to inspect its long-array structure and transcriptional activity. A previously done analysis on the CHM13 dataset with the NTRprism software tool, which was specifically developed for this purpose, reported that HSAT1A displays a higher-order organization (HOR) with the predominance of 9-mer NTRs (nested tandem repeats) (Altemose et al., 2022). To expand on this information and assess the recurrency of HSAT1A organization in an additional reference cell line, the same approach was applied to the NA12878 dataset. It was also confirmed with this analysis, that the most frequent periodicity identified corresponds to the HSat1A 42-bp monomer, followed by the 378-bp 9-mer array (**Figure 10**). Although these are the two most predominant monomers, there is still some relevant representation of other n-mer (**Figure 10**).

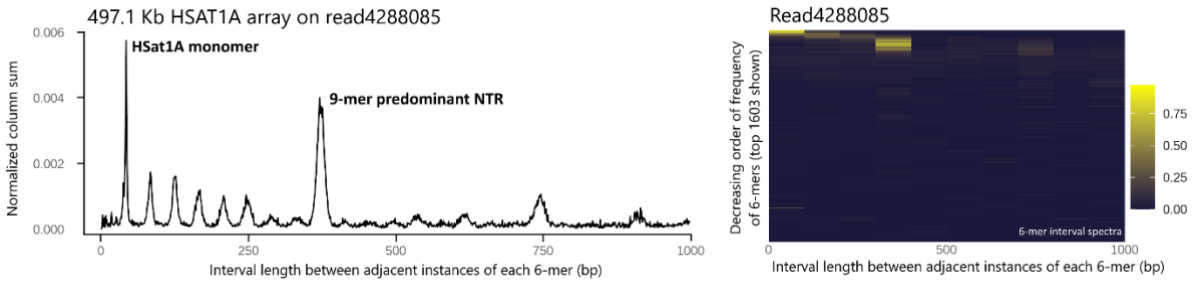


Figure 10 – HSat1A periodicity spectrum and heatmap in a selected read from the NA12878 dataset. NTRprism reveals two predominant peaks: one corresponding to HSat1A monomer and the second to a 9-mer higher repeat. (NTR – Nested Tandem Repeat).

Following this, HSAT1A transcription activity was assessed (**Figure 11**). For this, 3′ RACE (Yeku & Frohman, 2011) in HeLa RNA using an oligo-dT anchor primer for reverse-transcription and a forward PCR primer targeting the HSat1A sequence was performed by the CytoGenomics Lab. Analysis of 3′ RACE products by agarose gel electrophoresis revealed a ladder of products, with the most intense band around 550 bp (**Figure 12A**). This result suggests that HSAT1A transcripts are polyadenylated and thus likely transcribed by RNA polymerase II (Hirose & Manley, 1998). To characterize these amplified products, a 300-nt paired-end high-throughput sequencing using the Illumina MiSeq platform was performed. A total of ~2×10^5 reads **(Supplementary Table 15**) were obtained (for R1 and R2), which were quality and size filtered and assembled into ~3.5×10^4 complete 3′ RACE transcripts. Approximately 70% of these assembled transcripts presented the HSat1A motif and were distributed across a size range of 51 to ~400 nucleotides, with peaks corresponding to multiples of the 42-monomer size (**Figure 12A**). Although PCR size-amplification bias cannot be ignored, the most prevalent sequence size was ~170 nt. Within this universe, 16,332 sequences were found to be unique, attesting to the high complexity of the HSAT1A transcriptome. By analyzing sequence reads in a window of 200 nucleotides, a progressive reconstruction of longer reads by piecing together smaller ones was observed (**Figure 12B**), possibly suggesting mechanisms of alternative polyadenylation (APA). To test this theory, the structure of a representative read was examined, having found multiple alternative and non-canonical polyadenylation signals (PASs) (Tian et al., 2005), organized in a known poly(A) signal structure (Proudfoot, 2011), and cleavage sites often corresponding to the actual read lengths (**Figure 12C**). To get a better perspective of the degree of sequence variability, this dataset was clustered into groups with a minimum sequence identity of 90%, identifying a total of 257 clusters, 50 of which had more than 50 elements (**Figure 12D-E**). Subsequently, unbiased search for sequence motifs within each of the 50 mentioned clusters was performed. The obtained repeated motifs invariably compose, or were composed of, HSat1A 42 nt monomers (**Supplementary Figure 17**).

Based on the findings of this analysis, the HSat1A human satellite family revealed to have NTR arrangements in the human genome. Moreover, there is no indication that transcription of HSat1A arises from preferable *loci*, as the transcripts showed a great sequence diversity. The size variability observed within each cluster of sequences seems to be the consequence of the frequent presence of sequence elements similar to PolyA signals within the HSat1A monomer, creating multiple alternative polyadenylation sites within arrays derived transcripts.

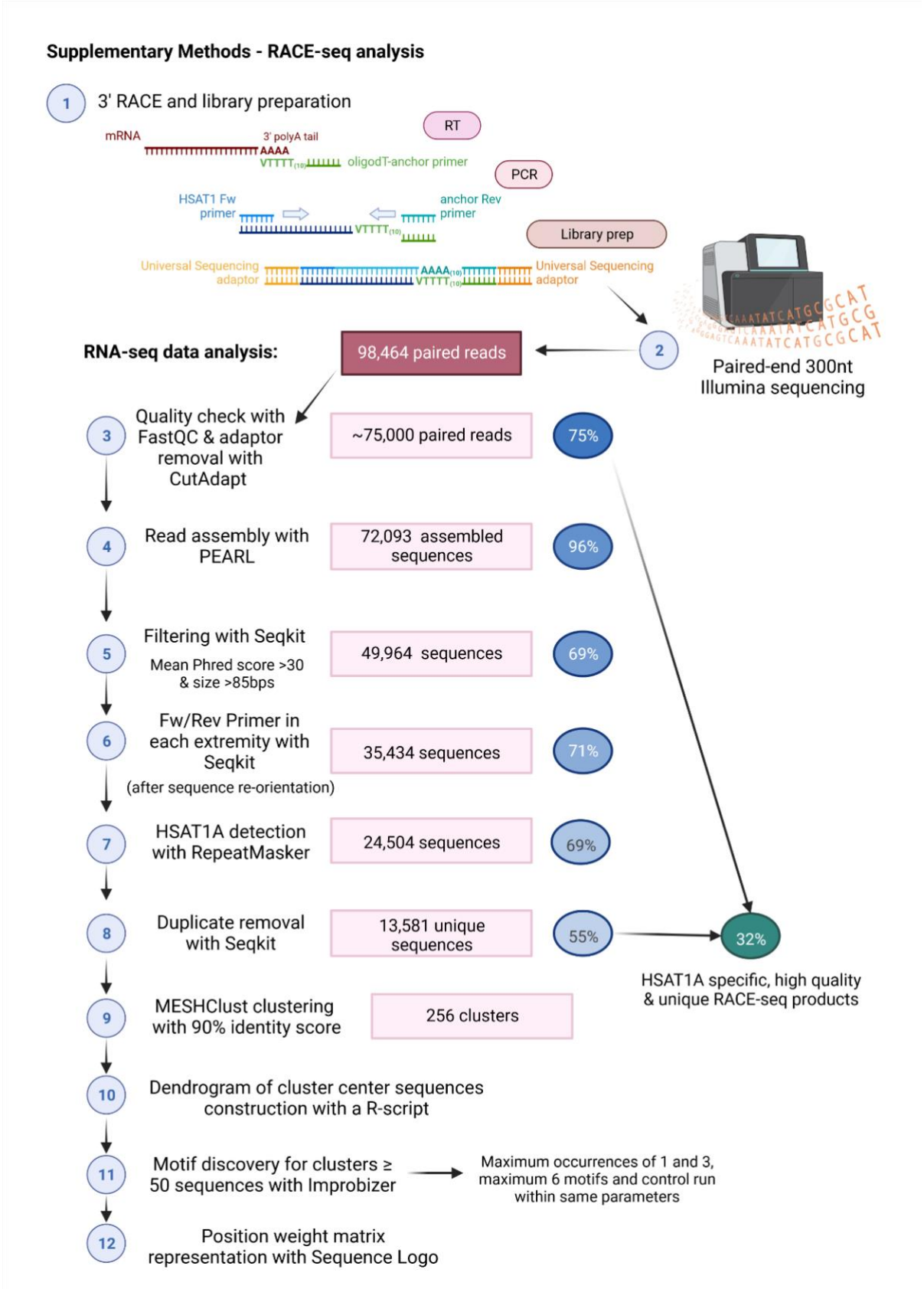


Figure 11 – Workflow for the analysis of HSat1A RACE-Seq.

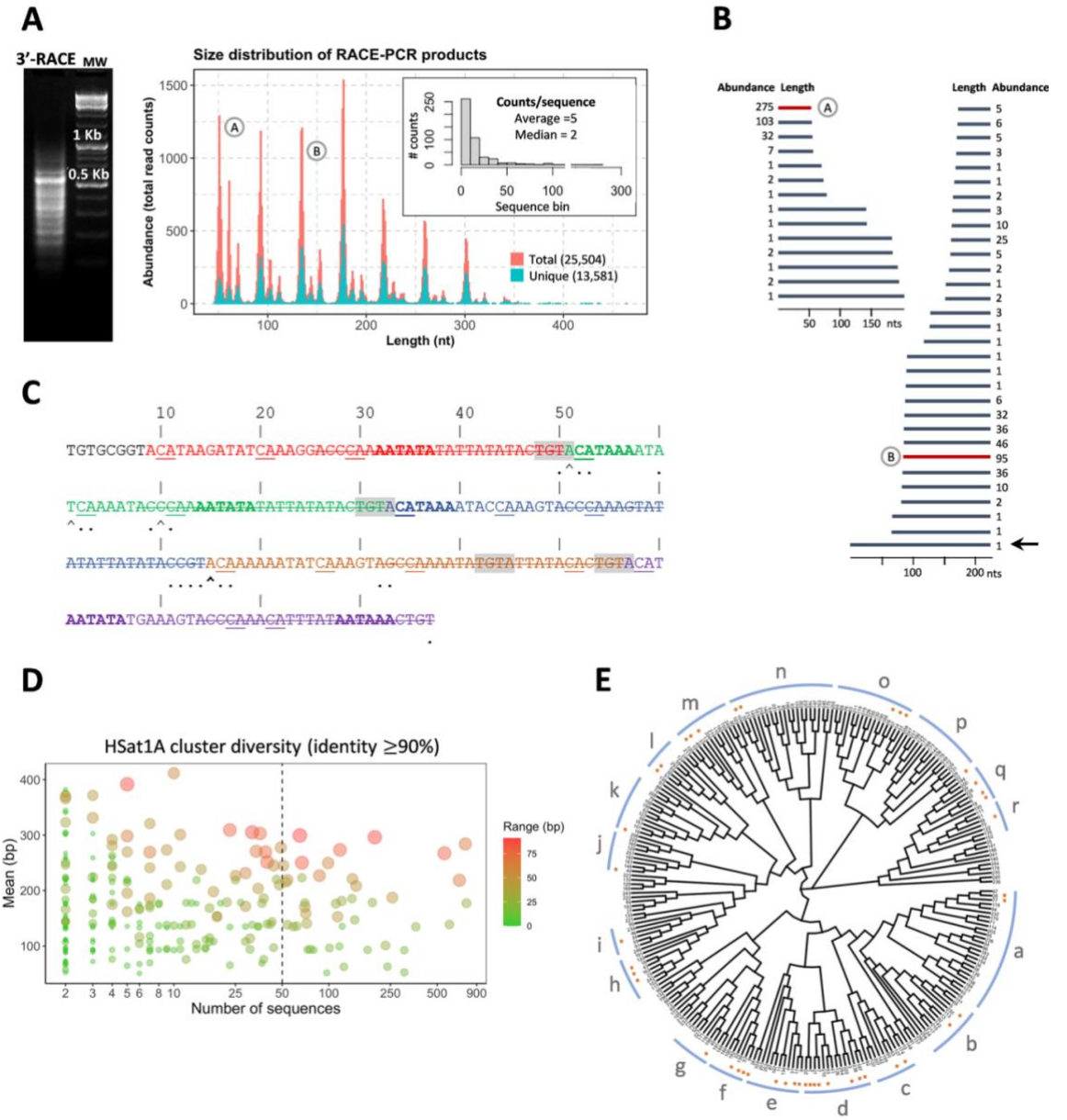


Figure 12 - HSat1A 3′ RACE analysis. A: Agarose gel corresponding to HSat1A 3′ RACE-amplified transcripts (left); Molecular weight (right). A size distribution plot is presented for the graphical representation of HSat1A reads. Assembled transcripts contained HSat1A peaks corresponding to multiples of the 42-monomer. From the total of HSat1A sequences, 16,332 sequences were found to be unique (blue in plot). The bar chart (top right corner) shows the high representation of unique sequences, visible in the distribution of counts/sequence. A and B (round) sequences are representative of the identified peaks and are displayed in B. **B**: HSat1A tandem transcript organization. In a universe of 200 nucleotides, it is possible to reconstruct transcripts of longer lengths with smaller sequences. The black arrow points to the longer represented read (structure explored in C). **C**: HSat1A read structure analyzed in the light of the consensus mammalian poly(A) signal. Different colors display HSat1A monomers: HSat1A are organized in alternative A (17 nt) and B (25 nt); strikethrough nucleotides in the figure. Sequences that may function as poly(A) signal hexamers (Hirose & Manley, 1998) are highlighted in bold. Shades of gray correspond to the sequence that functions as the recognition of the poly(A) signal in the absence of the canonical hexamer element [A(A/U)UAAA]. Nucleotides located at the site of optimal 3′ cleavage, named the poly(A) site, are underlined. Arrows point to the largest number of duplicates that are cleaved at that nucleotide position (bold for the largest most abundant). Dots represent the cleavage location of duplicates that contain a diference≥1 nucleotide from the previous sequence. The cleavage positions address the possible occurrence of alternative polyadenylation, resulting in the observable variation of transcript length. **D**: HSat1A transcript cluster membership. Colors determine the range (bp) between sequences of the same cluster. **E**: Phylogenetic tree depicting transcript variability, constructed from the multiple alignment between the center sequences of each cluster. Clusters can be grouped according to their distance (groups a–r). Orange dots represent clusters with more than 50 elements.

4. Discussion

4.1 PMSat as the most abundant satellite DNA of *Peromyscus genus*

This analysis showed that PMSat is the most abundant satDNA in the *Peromyscus* genus, consistent with previous studies (Louzada et al., 2015; Smalec et al., 2019). The abundance of this satDNA in the *Peromyscus genus* indicates a great importance of this sequence, that might be relevant in chromosome rearrangements that may drive speciation events, as it been shown on previous studies (Smalec et al., 2019). In accordance with Smalec et al. study it is possible to assert some similarities between the PMSat hybridization positioning and the sequence-based analysis performed in this work with subsequent relative positioning of PMSat arrays on chromosomes in four *Peromyscus* genus assemblies (PMAN, PCAL, PERE and PLEU). Some exceptions occur specifically in the PCAL assembly, as some PMSat arrays appear assembled near the end of multiple chromosomes in the sequence-based relative positioning, while in the PMSat hybridization this is not the case (Smalec et al., 2019). This may be derived by the presence of translocations within these regions of the chromosomes, or from misassembling of the chromosomes corresponding to this species. Notwithstanding, phylogenetic relation between the PMSat clustered groups from these four *Peromyscus* spp. seem to be in accordance with the phylogenetic tree assembled with previous studies performed with mitochondrial cytochrome-*b* sequences (Bradley et al., 2007) and PMSat hybridization data from *Peromyscus* spp. (Smalec et al., 2019). An important characteristic regarding the relative positioning of the PMSat arrays from each clustered group is the representativity on many different chromosomes across different assemblies, which could imply that translocations events between non-homologous chromosomes are frequent within species in this genus. This PMSat sequence clustering representation might also be an indication of molecular drive for speciation in this genus.

PMSat seemed to demonstrate common features across all genomes with similar monomer length, however, monomer and array GC-content, and array length showed variations among the genomes. A major common feature is that the monomer length seems to be ~345 bp, and that GC-content ranges from mainly 40-48%, which is correspondent with satellite DNA characteristics. The GC-content of both PMSat monomer and arrays showed slight variations among the genomes. The feature that showed a significant variation was the array length, which in some species showed a low range, such as the PMAN and PCAL, while the PERE and PLEU genomes showed a propensity for PMSat with higher array length.

Localization of LTRs within the four genomes selected showed important information regarding not only selective orientation, but also that series of ordered arrays may be maintained across their genomes. A good example of this would be in chromosome 13, that showed PMSat arrays detected across all genomes to be ordered in the minus strand, located mainly closest to the start of the chromosome. This type of behavior from LTR arrays could be relevant, given the instability associated with repetitive DNA sequences (Bzymek & Lovett, 2001; Vondrak et al., 2020). Furthermore, inspection of clustered groups of PMSat may also show important connections maintained based on similarity across all genomes. Such as the case of a specific cluster group that showed 2 similar arrays ordered together in chromosome 12 and in similar vicinity, detected in both the PMAN and PCAL assembly. This might indicate a strong connection of these

PMSat arrays between these two species, although their phylogenetic relationship showing greater distance when compared to other selected species, such as PLEU (Smalec et al., 2019).

4.2 LTR classes from the *Peromyscus* genus may have an important role across different genomes

The clustering analysis of LTR arrays within various *Peromyscus* genome assemblies has shed light on the diversity and distribution of these repetitive elements across species. The presence of highly similar LTR arrays across different species highlights the conservation of specific LTR classes within the *Peromyscus* genus. These findings might indicate a significant role of these LTRs in these genomes.

It is worth noting that the differences in the distribution of LTR classes across assemblies, as observed in this study, may reflect variations in genome assembly quality and sequencing methods, as these factors can impact the detection and annotation of repetitive elements (Treangen & Salzberg, 2012). Nevertheless, these differences might still be attributed to the evolutionary history of the *Peromyscus* genus (Harringmeyer & Hoekstra, 2022; Smalec et al., 2019).

Newly identified LTRs were found present across *Peromyscus* genus genomes. Of these new LTRs, the LTR_PMAN_1 and LTR_PMAN_2 classes have a higher count detection in comparison with other newly identified LTRs in the PCAL assembly. These classes also have a high copy number of monomers in their arrays, which might suggest a significant role of these newly identified LTRs in this species. However, both of these classes had no representativity shown on the chromosomes for relative positioning analysis, because they were only detected in scaffolds.

LTR classes LTR_PCAL_1, LTR_PCAL_2, LTR_PERE_10, LTR_PERE_14, and LTR_PLEU_4 although showing representativity on at least two assemblies from different species, only had between two or three copy number in their arrays. This may mean that these classes are correlated with transposable elements inserted on chromosomes in the ancestral of *Peromyscus* spp. Additionally, as shown in another study, these classes can also represent an important factor in the unique landscape of TEs of this genus in comparison to other mammalian genomes (Gozashti et al., 2023).

Newly identified LTR classes that showed similar characteristics but had more than three copy numbers in their arrays also had important characteristics to denote. The LTR_PMAN_5 class had arrays detected on chromosomes 11 and 15 in different species, which might be an indication of a translocation event. Furthermore, other newly identified LTR classes with these characteristics, such as LTR_PMAN_7 and LTR_PCAL_16, were only detected in the same chromosome between species (chromosomes 4 and 8, respectively). This might indicate a relevant role of these LTRs in their respective chromosomes. Some of these newly identified LTR classes were also located on distinct regions of chromosomes in different species, which could indicate the occurrence of crossing-over events of these LTRs.

From other satDNA families outside of PMSat it was also possible to determine some important interactions. The MMSAT4 family showed that a crossing-over event in chromosome 1 might have driven to a divergence between species seen between PERE and PLEU, which have been shown to be further apart phylogenetically (Bradley et al., 2007; Smalec et al., 2019). However, MurSatRep1 showed indication of a

crossover event of representative arrays on chromosome 6 that might separate PCAL from PMAN, PERE and PLEU, which goes against the phylogenetic distance between these four species.

Studying these LTR classes taking advantage of long contiguous reads, such as produced by nanopore sequencing technology, could provide further insight into specific arrangements from crossing-over events within the *Peromyscus* genus. Furthermore, it would also be of interest to study these LTR classes in additional *Peromyscus* spp., in order to obtain information regarding the chromosomal evolution and speciation events within this genus.

4.3 Diversity of human LTRs and HSAT1A transcription activity

Identifying and characterizing LTR arrays is difficult because due to their variable length and sequence they require long sequencing reads for their proper characterization. Although nanopore sequencing technology fixes this problem, this sequencing method was developed very recently and is still under constant improvement (Jain et al., 2018; Nurk et al., 2022). Several research efforts are underway to improve the identification and characterization of LTRs using a combination of sequencing technologies, developing new bioinformatics tools and algorithms, and making the technology more accessible and affordable, which can advance our understanding of LTRs (Altemose, 2022; Lopes et al., 2021; Miga, 2021). The results presented in this thesis can be seen as a part of this endeavor. For example, the DNA isolation and sequencing protocol used in this study showed improvement on the overall percentage of long reads compared to the publicly available NA12878 dataset for the same nanopore sequencing technology used. Nevertheless, there are still some challenges in the production of ultra-long reads.

The results obtained in this thesis suggest that there are significant differences in the NA12878 dataset in comparison with the GM03417 and CHM13 datasets, as it demonstrated a much lower percentage of clustered LTR groups. Additionally, detection of the most abundant satellite DNA, ALR, also showed a lower percentage of all the in the NA12878 sequencing data, although other satellite DNA did not show significant variation in percentages. This might be an indication of a higher fragmentation in the isolation of DNA for library preparation in this dataset. Nevertheless, it was possible to confirm the presence of the most common human satellite DNA in the top 20 clustered groups of LTRs on all assemblies, corresponding with previous characterization of satellite studies in the human genome (Altemose, 2022; Lopes et al., 2021).

Of note, the clustering analysis performed in this thesis identified some LTRs that had not been previously described in other studies, which were also present in the top 20 most abundant groups, specifically in the CHM13 Telomore-2-Telomere Consortium dataset (Altemose et al., 2022; Nurk et al., 2022). Although the number of these new arrays is lower than those of human satellite DNA, they could still represent significant repetitive element families with important functions.

Regarding the use of flow cytometry-based isolation for sequencing of individual chromosomes, it seems to be a valuable approach for studying chromosomal abnormalities such as translocations, as it was successful in isolating the translocated chromosome rob(14:21) along with other chromosomes. This is in line with a previous study using this approach (Kuderna et al., 2019). However, the results indicate that this method might not be efficient for the study of repetitive sequences, as the length of the reads produced suggested significant DNA fragmentation before library preparation.

Finally, with the use of 3' RACE to characterize HSat1A transcripts, it was possible to verify that the size variability of HSat1A transcripts is most likely associated to alternative polyadenylation (APA), which can be regulated by various factors, including a fundamental co-transcriptional gene regulation process that relies on U1 small nuclear ribonucleoprotein (snRNP) to suppress premature 3'-end cleavage and polyadenylation (PCPA) in RNA polymerase II transcripts, which is necessary for full-length transcription of thousands of protein-coding (pre-mRNAs) and long noncoding (lncRNA) genes (a process called U1 telescripting), and non-canonical polyadenylation signals (Cugusi et al., 2022). The accumulation of HSat1A transcripts is thought to be dependent on a complex interplay between transcriptional and post-transcriptional processes (Eymery et al., 2009). It is also of note that abnormal transcription of satellite sequences is a common feature of tumor cells (Eymery et al., 2009; Puppo et al., 2020), suggesting that further functional studies are needed to determine the role of HSat1A in genome architecture, gene expression regulation, and cellular pathways involved in stress, development, and pathology. To better understand the function of HSat1A, performing experiments to knock down the transcript and evaluate the resulting cellular phenotypes is recommended. Additionally, investigating the pathways that regulate the expression of HSat1A transcripts is crucial for understanding their role in tumor development. Exploring HSat1A transcripts could provide valuable information about the mechanisms underlying abnormal transcription in tumor cells and their involvement in disease development.

4.4 Concluding remarks

This work demonstrated the ability to detect and assess LTRs using different sequencing methods. The study of LTRs on *Peromyscus* spp. showed interesting patterns and relationships that allowed for a better characterization of these previously and newly identified LTR classes. The relative positioning of the most abundant satDNA family on *Peromyscus* genus, PMSat, with other LTR classes also allowed for a better depiction of these repetitive elements.

Nanopore sequencing technology showed great potential for a better quantification and characterization of LTRs in the human genome. However, some challenges in identifying and characterizing LTRs using this sequencing technology simultaneously with flow-sorting of chromosomes were acknowledged, and resolutions to improve these methods were highlighted. Some novel LTRs were also identified in the human genome. Furthermore, the characterization of the HSat1A family revealed not only to be transcriptionally active, but also that a high degree of variability between genomic and transcriptomic sequences.

Overall, this analysis contributed to a better characterization of LTRs, with implications in genomic organization and evolution. Notwithstanding, further investigation and functional studies are of great importance to elucidate the specific functions and mechanisms of these repetitive elements.

5. References

Adega, F., Chaves, R., Guedes-Pinto, H., & Heslop-Harrison, J. S. (2006). Physical organization of the 1.709 satellite IV DNA family in Bovini and Tragelaphini tribes of the Bovidae: Sequence and chromosomal evolution. *Cytogenetic and Genome Research*, *114*(2), 140–146. https://doi.org/10.1159/000093330

Adega, F., Guedes-Pinto, H., & Chaves, R. (2009). Satellite DNA in the karyotype evolution of domestic animals - Clinical considerations. In *Cytogenetic and Genome Research* (Vol. 126, Issues 1–2, pp. 12–20). https://doi.org/10.1159/000245903

Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K., & Sullivan, B. A. (2016). Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Research*, *26*(10), 1301–1311. https://doi.org/10.1101/gr.206706.116

Aldrup-MacDonald, M., & Sullivan, B. (2014). The Past, Present, and Future of Human Centromere Genomics. *Genes*, *5*(1), 33–50. https://doi.org/10.3390/genes5010033

Altemose, N. (2022). A classical revival: Human satellite DNAs enter the genomics era. In *Seminars in Cell and Developmental Biology* (Vol. 128, pp. 2–14). Elsevier Ltd. https://doi.org/10.1016/j.semcdb.2022.04.012

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., Hoyt, S. J., Uralsky, L., Ryabov, F. D., Shew, C. J., Sauria, M. E. G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V. A., Dvorkina, T., Kunyavskaya, O., Vollger, M. R., Rhie, A., … Miga, K. H. (2022). Complete genomic and epigenetic maps of human centromeres. *Science*, *376*(6588). https://doi.org/10.1126/science.abl4178

Altemose, N., Miga, K. H., Maggioni, M., & Willard, H. F. (2014). Genomic Characterization of Large Heterochromatic Gaps in the Human Genome Assembly. *PLoS Computational Biology*, *10*(5), e1003628. https://doi.org/10.1371/journal.pcbi.1003628

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andrews, S. (2010). *FASTQC. A quality control tool for high throughput sequence data*.

Ao, W., Gaudet, J., Kent, W. J., Muttumu, S., & Mango, S. E. (2004). Environmentally Induced Foregut Remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, *305*(5691), 1743–1746. https://doi.org/10.1126/science.1102216

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature Methods*, *9*(4),

333–337. https://doi.org/10.1038/nmeth.1935

Bandyopadhyay, R., Berend, S. A., Page, S. L., Choo, K. H., & Shaffer, L. G. (2001). Satellite III sequences on 14p and their relevance to Robertsonian translocation formation. *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *9*(3), 235–242. https://doi.org/10.1023/a:1016652621226

Barra, V., & Fachinetti, D. (2018). The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. In *Nature Communications* (Vol. 9, Issue 1). Nature Publishing Group. https://doi.org/10.1038/s41467-018-06545-y

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. In *Nucleic Acids Research* (Vol. 27, Issue 2). https://academic.oup.com/nar/article/27/2/573/1061099

Bradley, R. D., Durish, N. D., Rogers, D. S., Miller, J. R., Engstrom, M. D., & Kilpatrick, C. W. (2007). Toward a Molecular Phylogeny for Peromyscus: Evidence from Mitochondrial Cytochrome- b Sequences. *Journal of Mammalogy*, *88*(5), 1146–1159. https://doi.org/10.1644/06-MAMM-A-342R.1

Brändle, F., Frühbauer, B., & Jagannathan, M. (2022). Principles and functions of pericentromeric satellite DNA clustering into chromocenters. *Seminars in Cell and Developmental Biology*, *128*(January), 26–39. https://doi.org/10.1016/j.semcdb.2022.02.005

Bzymek, M., & Lovett, S. T. (2001). Instability of repetitive DNA sequences: The role of replication in multiple mechanisms. *Proceedings of the National Academy of Sciences*, *98*(15), 8319–8325. https://doi.org/10.1073/pnas.111008398

Cápal, P., Endo, T. R., Vrána, J., Kubaláková, M., Karafiátová, M., Komínková, E., Mora-Ramírez, I., Weschke, W., & Doležel, J. (2016). The utility of flow sorting to identify chromosomes carrying a single copy transgene in wheat. *Plant Methods*, *12*(1), 24. https://doi.org/10.1186/s13007-016-0124-8

Cugusi, S., Mitter, R., Kelly, G. P., Walker, J., Han, Z., Pisano, P., Wierer, M., Stewart, A., & Svejstrup, J. Q. (2022). Heat shock induces premature transcript termination and reconfigures the human transcriptome. *Molecular Cell*, *82*(8), 1573-1588.e10. https://doi.org/10.1016/j.molcel.2022.01.007

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., Finch, N. C. A., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G. Y. R., Karydas, A., Seeley, W. W., Josephs, K. A., Coppola, G., Geschwind, D. H., … Rademakers, R. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*, *72*(2), 245–256. https://doi.org/10.1016/j.neuron.2011.09.011

Doležel, J., Lucretti, S., Molnár, I., Cápal, P., & Giorgi, D. (2021). Chromosome analysis and sorting. In *Cytometry Part A* (Vol. 99, Issue 4, pp. 328–342). John Wiley and Sons Inc. https://doi.org/10.1002/cyto.a.24324

Earle, E., Shaffer, L. G., Kalitsis, P., McQuillan, C., Dale, S., & Choo, K. H. A. (1992). Identification of DNA sequences flanking the breakpoint of human t(14q21q) Robertsonian translocations. *American Journal of Human Genetics*, *50*(4), 717–724.

Easterling, K. A., Pitra, N. J., Morcol, T. B., Aquino, J. R., Lopes, L. G., Bussey, K. C., Matthews, P. D., & Bass, H. W. (2020). Identification of tandem repeat families from long-read sequences of Humulus lupulus. *PLOS ONE*, *15*(6), e0233971. https://doi.org/10.1371/journal.pone.0233971

Eymery, A., Callanan, M., & Vourc'h, C. (2009). The secret message of heterochromatin: New insights into the mechanisms and function of centromeric and pericentric repeat sequence transcription. In *International Journal of Developmental Biology* (Vol. 53, Issues 2–3, pp. 259–268). https://doi.org/10.1387/ijdb.082673ae

Fingerhut, J. M., & Yamashita, Y. M. (2022). The regulation and potential functions of intronic satellite DNA. *Seminars in Cell and Developmental Biology*, *128*(April), 69–77. https://doi.org/10.1016/j.semcdb.2022.04.010

Fournier, A., Mcleer-florin, A., Lefebvre, C., Duley, S., Debernardi, A., Rousseaux, S., Fraipont, F. De, Figeac, M., Kerckaert, J., Vos, J. De, Usson, Y., Delaval, K., Grichine, A., Vourc, C., Khochbin, S., Feil, R., Leroux, D., & Callanan, M. B. (2010). *1q12 chromosome translocations form aberrant heterochromatic foci associated with changes in nuclear architecture and gene expression in B cell lymphoma*. 159–171. https://doi.org/10.1002/emmm.201000067

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

Girgis, H. Z. (2022). MeShClust v3.0: high-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. *BMC Genomics*, *23*(1), 423. https://doi.org/10.1186/s12864-022-08619-0

Giunta, S., & Funabiki, H. (2017). Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. *Proceedings of the National Academy of Sciences*, *114*(8), 1928–1933. https://doi.org/10.1073/pnas.1615133114

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

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Goyal, P., Krasteva, P. V., Van Gerven, N., Gubellini, F., Van Den Broeck, I., Troupiotis-Tsaïlaki, A., Jonckheere, W., Péhau-Arnaudet, G., Pinkner, J. S., Chapman, M. R., Hultgren, S. J., Howorka, S., Fronzes, R., & Remaut, H. (2014). Structural and mechanistic insights into the bacterial amyloid

secretion channel CsgG. *Nature*, *516*(7530), 250–253. https://doi.org/10.1038/nature13768

Gozashti, L., Feschotte, C., & Hoekstra, H. E. (2023). Transposable Element Interactions Shape the Ecology of the Deer Mouse Genome. *Molecular Biology and Evolution*, *40*(4), 1–17. https://doi.org/10.1093/molbev/msad069

Harringmeyer, O. S., & Hoekstra, H. E. (2022). Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nature Ecology & Evolution*, *6*(12), 1965–1979. https://doi.org/10.1038/s41559-022-01890-0

Henikoff, S., Ahmad, K., & Malik, H. S. (2001). The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. *Science*, *293*(5532), 1098–1102. https://doi.org/10.1126/science.1062939

Hirose, Y., & Manley, J. L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. *Nature*, *395*(6697), 93–96. https://doi.org/10.1038/25786

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89. https://doi.org/10.1093/nar/gkv1272

Isiktas, A. U., Eshov, A., Yang, S., & Guo, J. U. (2022). Systematic generation and imaging of tandem repeats reveal base-pairing properties that promote RNA aggregation. *Cell Reports Methods*, *2*(11). https://doi.org/10.1016/j.crmeth.2022.100334

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

John, B., King, M., Schweizer, D., & Mendelak, M. (1985). Equilocality of heterochromatin distribution and heterochromatin heterogeneity in acridid grasshoppers. *Chromosoma*, *91*(3–4), 185–200. https://doi.org/10.1007/BF00328216

Kinkar, L., Gasser, R. B., Webster, B. L., Rollinson, D., Littlewood, D. T. J., Chang, B. C. H., Stroehlein, A. J., Korhonen, P. K., & Young, N. D. (2021). Nanopore sequencing resolves elusive long tandem-repeat regions in mitochondrial genomes. *International Journal of Molecular Sciences*, *22*(4), 1–12. https://doi.org/10.3390/ijms22041811

Klein, S. J., & O'Neill, R. J. (2018). Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Research*, *26*(1–2), 5–23. https://doi.org/10.1007/s10577-017-9569-5

Komissarov, A. S., Gavrilova, E. V., Demin, S. J., Ishov, A. M., & Podgornaya, O. I. (2011). Tandemly repeated DNA families in the mouse genome. *BMC Genomics*, *12*. https://doi.org/10.1186/1471-2164-

12-531

Kuderna, L. F. K., Lizano, E., Julià, E., Gomez-Garrido, J., Serres-Armero, A., Kuhlwilm, M., Alandes, R. A., Alvarez-Estape, M., Juan, D., Simon, H., Alioto, T., Gut, M., Gut, I., Schierup, M. H., Fornas, O., & Marques-Bonet, T. (2019). Selective single molecule sequencing and assembly of a human Y chromosome of African origin. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-07885-5

Kuhn, G. C. S., Sene, F. M., Moreira-Filho, O., Schwarzacher, T., & Heslop-Harrison, J. S. (2008). Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the Drosophila buzzatii cluster. *Chromosome Research*, *16*(2), 307–324. https://doi.org/10.1007/s10577-007-1195-1

Langdon, T., Seago, C., Jones, R. N., Ougham, H., Thomas, H., Forster, J. W., & Jenkins, G. (2000). *De Novo Evolution of Satellite DNA on the Rye B Chromosome*. https://academic.oup.com/genetics/article/154/2/869/6047885

Lee, C., Wevrick, R., Fisher, R. B., Ferguson-Smith, M. A., & Lin, C. C. (1997). Human centromeric DNAs. *Human Genetics*, *100*(3–4), 291–304. https://doi.org/10.1007/s004390050508

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lin, M. S., & Davidson, R. L. (1974). Centric Fusion, Satellite DNA, and DNA Polarity in Mouse Chromosomes. *Science*, *185*(4157), 1179–1181. https://doi.org/10.1126/science.185.4157.1179

Logsdon, G. A., & Eichler, E. E. (2023). The Dynamic Structure and Rapid Evolution of Human Centromeric Satellite DNA. In *Genes* (Vol. 14, Issue 1). MDPI. https://doi.org/10.3390/genes14010092

Long, A. D., Baldwin-Brown, J., Tao, Y., Cook, V. J., Balderrama-Gutierrez, G., Corbett-Detig, R., Mortazavi, A., & Barbour, A. G. (2019). The genome of Peromyscus leucopus , natural host for Lyme disease and other emerging infections. *Science Advances*, *5*(7), 1–10. https://doi.org/10.1126/sciadv.aaw6441

Lopes, M., Louzada, S., Ferreira, D., Veríssimo, G., Eleutério, D., Gama-Carvalho, M., & Chaves, R. (2023). Human Satellite 1A analysis provides evidence of pericentromeric transcription. *BMC Biology*, *21*(1). https://doi.org/10.1186/s12915-023-01521-5

Lopes, M., Louzada, S., Gama-Carvalho, M., & Chaves, R. (2021). Genomic tackling of human satellite dna: Breaking barriers through time. In *International Journal of Molecular Sciences* (Vol. 22, Issue 9). MDPI. https://doi.org/10.3390/ijms22094707

Louzada, S., Lopes, M., Ferreira, D., Adega, F., Escudeiro, A., Gama-carvalho, M., & Chaves, R. (2020).

Decoding the role of satellite DNA in genome architecture and plasticity—an evolutionary and clinical affair. In *Genes* (Vol. 11, Issue 1). MDPI AG. https://doi.org/10.3390/genes11010072

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. https://doi.org/10.1016/j.ympev.2015.06.008

Magi, A., Semeraro, R., Mingrino, A., Giusti, B., & D'Aurizio, R. (2017). Nanopore sequencing data analysis: State of the art, applications and challenges. *Briefings in Bioinformatics*, *19*(6), 1256–1272. https://doi.org/10.1093/bib/bbx062

Malik, H. S., & Henikoff, S. (2009). Major Evolutionary Transitions in Centromere Complexity. In *Cell* (Vol. 138, Issue 6, pp. 1067–1082). Elsevier B.V. https://doi.org/10.1016/j.cell.2009.08.036

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

McCartney, M. A., Mallez, S., & Gohl, D. M. (2019). Genome projects in invasion biology. In *Conservation Genetics* (Vol. 20, Issue 6, pp. 1201–1222). Springer Netherlands. https://doi.org/10.1007/s10592-019-01224-x

McNulty, S. M., & Sullivan, B. A. (2018). Alpha satellite DNA biology: finding function in the recesses of the genome. In *Chromosome Research* (Vol. 26, Issue 3, pp. 115–138). Springer Netherlands. https://doi.org/10.1007/s10577-018-9582-3

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

Miga, K. H. (2015). Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research*, *23*(3), 421–426. https://doi.org/10.1007/s10577-015-9488-2

Miga, K. H. (2021). Breaking through the unknowns of the human reference genome. *Nature*, *590*(7845), 217–218. https://doi.org/10.1038/d41586-021-00293-8

Miga, K. H., Newton, Y., Jain, M., Altemose, N., Willard, H. F., & Kent, W. J. (2014). Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research*, *24*(4), 697–707. https://doi.org/10.1101/gr.159624.113

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53.

https://doi.org/10.1126/science.abj6987

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. https://doi.org/10.1016/j.gene.2007.11.013

Plohl, M., Meštrović, N., & Mravinac, B. (2014). Centromere identity from the DNA point of view. In *Chromosoma* (Vol. 123, Issue 4, pp. 313–325). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00412-014-0462-0

Poot, M., & Hochstenbach, R. (2021). Prevalence and Phenotypic Impact of Robertsonian Translocations. *Molecular Syndromology*, *12*(1), 1–11. https://doi.org/10.1159/000512676

Prosser, J., Frommert, M., Paul, C., & Vincent, P. C. (1986). Sequence Relationships of Three Human Satellite DNAs. In *J. Mol. Riol* (Vol. 187).

Proudfoot, N. J. (2011). Ending the message: poly(A) signals then and now. *Genes & Development*, *25*(17), 1770–1782. https://doi.org/10.1101/gad.17268411

Puppo, I. L., Saifitdinova, A. F., & Tonyan, Z. N. (2020). The Role of Satellite DNA in Causing Structural Rearrangements in Human Karyotype. In *Russian Journal of Genetics* (Vol. 56, Issue 1, pp. 41–47). Pleiades Publishing. https://doi.org/10.1134/S1022795419080155

Raskina, O., Barber, J. C., Nevo, E., & Belyayev, A. (2008). Repetitive DNA and chromosomal rearrangements: Speciation-related events in plant genomes. In *Cytogenetic and Genome Research* (Vol. 120, Issues 3–4, pp. 351–357). https://doi.org/10.1159/000121084

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

Shatskikh, A. S., Kotov, A. A., Adashev, V. E., Bazylev, S. S., & Olenina, L. V. (2020). Functional Significance of Satellite DNAs: Insights From Drosophila. *Frontiers in Cell and Developmental Biology*, *8*(May), 1–19. https://doi.org/10.3389/fcell.2020.00312

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

Smalec, B. M., Heider, T. N., Flynn, B. L., & O'Neill, R. J. (2019). A centromere satellite concomitant with extensive karyotypic diversity across the Peromyscus genus defies predictions of molecular drive. *Chromosome Research*, *27*(3), 237–252. https://doi.org/10.1007/s10577-019-09605-1

Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>. (n.d.).

Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., & Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, *12*(1), 2. https://doi.org/10.1186/s13100-020-00230-y

Sulovari, A., Li, R., Audano, P. A., Porubsky, D., Vollger, M. R., Logsdon, G. A., Warren, W. C., Pollen, A. A., Chaisson, M. J. P., & Eichler, E. E. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(46), 23243–23253. https://doi.org/10.1073/pnas.1912175116

Thakur, J., Packiaraj, J., & Henikoff, S. (2021). Sequence, Chromatin and Evolution of Satellite DNA. *International Journal of Molecular Sciences*, *22*(9), 4309. https://doi.org/10.3390/ijms22094309

Tian, B., Hu, J., Zhang, H., & Lutz, C. S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, *33*(1), 201–212. https://doi.org/10.1093/nar/gki158

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, *47*(21), 10994–11006. https://doi.org/10.1093/nar/gkz841

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

Trigiante, G., Blanes Ruiz, N., & Cerase, A. (2021). Emerging Roles of Repetitive and Repeat-Containing RNA in Nuclear and Chromatin Organization and Gene Expression. In *Frontiers in Cell and Developmental Biology* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fcell.2021.735527

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

Ummat, A., & Bashir, A. (2014). Resolving complex tandem repeats with long reads. *Bioinformatics*, *30*(24), 3491–3498. https://doi.org/10.1093/bioinformatics/btu437

Vondrak, T., Ávila Robledillo, L., Novák, P., Koblížková, A., Neumann, P., & Macas, J. (2020). Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *Plant Journal*, *101*(2), 484–500. https://doi.org/10.1111/tpj.14546

Wang, K., Xiang, D., Xia, K., Sun, B., Khurshid, H., & Esh, A. M. H. (2022). *Characterization of Repetitive DNA in Saccharum officinarum and Saccharum spontaneum by Genome Sequencing and Cytological Assays*. *13*(February), 1–13. https://doi.org/10.3389/fpls.2022.814620

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Yang, J., Yuan, B., Wu, Y., Li, M., Li, J., Xu, D., Gao, Z. hong, Ma, G., Zhou, Y., Zuo, Y., Wang, J., & Guo, Y. (2020). The wide distribution and horizontal transfers of beta satellite DNA in eukaryotes. *Genomics*, *112*(6), 5295–5304. https://doi.org/10.1016/j.ygeno.2020.10.006

Yeku, O., & Frohman, M. A. (2011). Rapid Amplification of cDNA Ends (RACE). In H. Nielsen (Ed.), *Methods in molecular biology (Clifton, N.J.)* (Vol. 703, pp. 107–122). Humana Press. https://doi.org/10.1007/978-1-59745-248-9_8

Zattera, M. L., Gazolla, C. B., Soares, A. de A., Gazoni, T., Pollet, N., Recco-Pimentel, S. M., & Bruschi, D. P. (2020). Evolutionary Dynamics of the Repetitive DNA in the Karyotypes of Pipa carvalhoi and Xenopus tropicalis (Anura, Pipidae). *Frontiers in Genetics*, *11*(July), 1–10. https://doi.org/10.3389/fgene.2020.00637

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology*, *7*(1–2), 203–214. https://doi.org/10.1089/10665270050081478

Zhimulev, I. F., & Belyaeva, E. S. (2003). Intercalary heterochromatin and genetic silencing. *BioEssays*, *25*(11), 1040–1051. https://doi.org/10.1002/bies.10343

6. Supplementary Data

Supplementary Table 1 – Clustering of LTRs detected in the PMAN 1.0 assembly.

Cluster Group	Arrays	min size (bp)	mean size (bp)	max size (bp)	Type	Repeat
1	870	340	357.6	1436	Satellite	PMSAT
2	11	69	70.1	73	Unclassified	LTR_PMAN_1
3	4	1648	1702	1781	TE	L1_Mur2_orf2
4	4	93	140	234	Unclassified	LTR_PMAN_2
5	4	83	84	85	Satellite	MurSatRep1
6	4	84	84	84	Satellite	MurSatRep1
7	3	1201	1528.7	1860	TE	L1_Mur2_orf2
8	3	70	105	141	TE	B1_Mus1/2
9	3	77	77	77	Unclassified	LTR_PMAN_3
10	2	1399	1408.5	1418	Unclassified	LTR_PMAN_8
11	2	1031	1111	1191	Unclassified	LTR_PMAN_9
12	2	803	937.5	1072	TE	RMER1C
13	2	993	999.5	1006	Unclassified	LTR_PMAN_10
14	2	499	500	501	Unclassified	LTR_PMAN_11
15	2	84	84	84	Satellite	MMSAT4
16	2	72	72	72	Unclassified	LTR_PMAN_4
17	2	63	63	63	Unclassified	LTR_PMAN_5

Supplementary Table 2 – Clustering of LTRs detected in the PMAN 2.1 CONTIGS assembly.

Cluster Group	Arrays	min size (bp)	mean size (bp)	max size (bp)	Type	Repeat
1	698	323	344	349	Satellite	PMSAT
2	8	68	70.1	73	Unclassified	LTR_PMAN_1
3	7	682	760.9	1223	Satellite	PMSAT
4	3	93	108.3	139	Unclassified	LTR_PMAN_2
5	3	84	84.3	85	Satellite	MurSatRep1
6	3	84	84	84	Satellite	MurSatRep1
7	3	77	77	77	Unclassified	LTR_PMAN_3
8	2	243	243	243	TE	B1_Mus1
9	2	70	87	104	TE	B1_Mus1
10	2	95	95	95	Unclassified	LTR_PMAN_6
11	2	84	84	84	Satellite	MMSAT4
12	2	72	72	72	Unclassified	LTR_PMAN_4
13	2	69	69	69	Unclassified	LTR_PMAN_7

14	2	63	63	63	Unclassified	LTR_PMAN_5

Supplementary Table 3 – Clustering of LTRs detected in the PMAN 2.1 CHR assembly.

Cluster Group	Arrays	min size (bp)	mean size (bp)	max size (bp)	Type	Repeat
1	657	323	344	349	Satellite	PMSAT
2	8	1340	1702.9	1969	TE	L1_Mur2_orf2
3	8	68	70.1	73	Unclassified	LTR_PMAN_1
4	7	682	743.1	1095	Satellite	PMSAT
5	4	84	84.2	85	Satellite	MurSatRep1
6	3	93	108.3	139	Unclassified	LTR_PMAN_2
7	3	84	84	84	Satellite	MurSatRep1
8	3	77	77	77	Unclassified	LTR_PMAN_3
9	2	1706	1828	1950	TE	L1_Mur3_orf2
10	2	790	802.5	815	TE	RMER1C
11	2	243	243	243	TE	B1_Mus1
12	2	70	87	104	TE	B1_Mus1
13	2	84	84	84	Satellite	MMSAT4
14	2	72	72	72	Unclassified	LTR_PMAN_4

Supplementary Table 4 – Clustering of LTRs detected in the PCAL assembly.

Cluster Group	Arrays	min size (bp)	mean size (bp)	max size (bp)	Type	Repeat
1	467	333	344.5	350	Satellite	PMSAT
2	310	67	75.9	206	Unclassified	LTR_PMAN_1
3	185	74	119.5	562	Unclassified	LTR_PMAN_2.1
4	27	672	943.2	1418	Satellite	PMSAT
5	18	1025	1365.4	1541	Unclassified	LTR_PCAL_1
6	9	66	71.3	72	Unclassified	LTR_PMAN_4
7	7	83	83.9	84	Satellite	MurSatRep1
8	6	1707	1745.5	1783	Unclassified	LTR_PCAL_2
9	6	685	686.3	687	Unclassified	LTR_PCAL_3
10	5	1273	1515	1678	Unclassified	LTR_PCAL_4
11	5	281	459	563	Unclassified	LTR_PMAN_1.1
12	4	309	309.8	310	Unclassified	LTR_PCAL_5
13	4	79	99.8	120	Unclassified	LTR_PCAL_6
14	4	114	114.8	115	Unclassified	LTR_PCAL_7
15	4	63	63	63	Unclassified	LTR_PMAN_5
16	3	93	108.7	140	Unclassified	LTR_PMAN_2

17	3	69	69	69	Unclassified	LTR_PMAN_7
18	2	1603	1660	1717	Unclassified	LTR_PCAL_8
19	2	1425	1426.5	1428	Unclassified	LTR_PCAL_9
20	2	1363	1363	1363	Unclassified	LTR_PCAL_10
21	2	503	531.5	560	Unclassified	LTR_PCAL_11
22	2	168	218.5	269	Satellite	MurSatRep1
23	2	235	235.5	236	Unclassified	LTR_PCAL_12
24	2	185	207.5	230	Unclassified	LTR_PMAN_2
25	2	78	102.5	127	Unclassified	LTR_PCAL_13
26	2	75	87.5	100	Unclassified	LTR_PCAL_14
27	2	72	72	72	Unclassified	LTR_PCAL_15
28	2	56	56	56	Unclassified	LTR_PCAL_16
29	2	51	51	51	Unclassified	LTR_PCAL_17

Supplementary Table 5 – Clustering of LTRs detected in the PERE assembly.

Cluster Group	Arrays	min size (bp)	mean size (bp)	max size (bp)	Type	Repeat
1	961	314	344.4	359	Satellite	PMSAT
2	43	1071	1414.3	1882	Unclassified	LTR_PERE_1
3	27	638	879	1719	Satellite	PMSAT
4	9	1015	1356.1	1742	Unclassified	LTR_PERE_2
5	8	84	84	84	Satellite	MurSatRep1
6	8	68	70.9	73	Unclassified	LTR_PMAN_1
7	6	1010	1128.3	1361	TE	Lx5c_3end
8	5	1122	1400.2	1641	Unclassified	LTR_PERE_3
9	5	79	79.8	80	Unclassified	LTR_PERE_4
10	4	1136	1223	1310	Unclassified	LTR_PERE_5
11	4	76	95	114	Unclassified	LTR_PMAN_2.1
12	4	72	72	72	Unclassified	LTR_PMAN_4
13	3	1055	1477.3	1894	Unclassified	LTR_PERE_6
14	3	1093	1574.7	1832	Unclassified	LTR_PERE_7
15	3	1038	1493.3	1792	TE	L1_Mur2_orf2
16	3	1778	1782	1784	Unclassified	LTR_PCAL_2
17	3	1033	1223.3	1491	Unclassified	LTR_PERE_8
18	3	1093	1281.7	1406	Unclassified	LTR_PCAL_1
19	3	1091	1191.3	1247	Unclassified	LTR_PERE_9
20	3	93	109	140	Unclassified	LTR_PMAN_2
21	2	1398	1689	1980	Unclassified	LTR_PERE_10
22	2	1214	1596.5	1979	Unclassified	LTR_PERE_11
23	2	1049	1512.5	1976	Unclassified	LTR_PERE_12
24	2	1960	1960	1960	Unclassified	LTR_PERE_13

25	2	1924	1935.5	1947	Unclassified	LTR_PERE_14
26	2	1915	1915	1915	Unclassified	LTR_PERE_15
27	2	1899	1899	1899	Unclassified	LTR_PERE_16
28	2	1860	1862.5	1865	Unclassified	LTR_PERE_17
29	2	1244	1536	1828	Unclassified	LTR_PERE_18
30	2	1589	1646.5	1704	Unclassified	LTR_PERE_19
31	2	1701	1701	1701	Unclassified	LTR_PERE_20
32	2	1320	1495	1670	Unclassified	LTR_PERE_21
33	2	1326	1483.5	1641	Unclassified	LTR_PERE_22
34	2	1457	1524	1591	Unclassified	LTR_PERE_23
35	2	1335	1462.5	1590	TE	MYSERV
36	2	1537	1537	1537	Unclassified	LTR_PERE_24
37	2	1266	1375.5	1485	Unclassified	LTR_PERE_25
38	2	1259	1363	1467	TE	L1_Mur3_orf2
39	2	1102	1258	1414	Unclassified	LTR_PERE_26
40	2	1102	1248.5	1395	Unclassified	LTR_PCAL_1
41	2	1359	1359	1359	Unclassified	LTR_PERE_27
42	2	1222	1287	1352	Unclassified	LTR_PERE_28
43	2	1283	1314	1345	Unclassified	LTR_PERE_29
44	2	1079	1211	1343	Unclassified	LTR_PERE_30
45	2	1342	1342	1342	Unclassified	LTR_PERE_31
46	2	1310	1310	1310	Unclassified	LTR_PERE_32
47	2	1235	1272	1309	Unclassified	LTR_PERE_33
48	2	1081	1178.5	1276	Unclassified	LTR_PERE_34
49	2	1145	1210	1275	Unclassified	LTR_PERE_35
50	2	1270	1270	1270	Unclassified	LTR_PERE_36
51	2	1065	1153	1241	Unclassified	LTR_PERE_37
52	2	1053	1135	1217	Unclassified	LTR_PERE_38
53	2	1216	1216	1216	Unclassified	LTR_PERE_39
54	2	1007	1097.5	1188	Unclassified	LTR_PERE_40
55	2	1174	1174	1174	Unclassified	LTR_PERE_41
56	2	1067	1111	1155	Unclassified	LTR_PERE_42
57	2	1144	1144	1144	Unclassified	LTR_PERE_43
58	2	1023	1067	1111	Unclassified	LTR_PERE_44
59	2	1094	1094.5	1095	Unclassified	LTR_PERE_45
60	2	1076	1076	1076	Unclassified	LTR_PERE_46
61	2	1014	1039.5	1065	TE	L1_Mur3_orf2
62	2	1015	1037.5	1060	TE	L1_Mur3_orf2
63	2	1044	1045.5	1047	Unclassified	LTR_PERE_47
64	2	1043	1043.5	1044	TE	L1_Mur2_orf2
65	2	1022	1022	1022	Unclassified	LTR_PERE_48

66	2	1015	1015.5	1016	Unclassified	LTR_PERE_49
67	2	343	344	345	Satellite	PMSAT
68	2	159	160.5	162	Unclassified	LTR_PCAL_6
69	2	63	63	63	Unclassified	LTR_PERE_50
70	2	63	63	63	Unclassified	LTR_PMAN_5

Supplementary Table 6 – Clustering of LTRs detected in the PLEU assembly.

Cluster Group	Arrays	min size (bp)	mean size (bp)	max size (bp)	Type	Repeat
1	96	178	341.9	345	Satellite	PMSAT
2	15	83	83.9	84	Satellite	MurSatRep1
3	9	72	72	72	Unclassified	LTR_PMAN_4
4	7	159	159	159	Unclassified	LTR_PLEU_1
5	4	1110	1225.5	1428	Unclassified	LTR_PLEU_2
6	3	70	93.3	105	TE	B1_Mus1
7	2	1844	1857.5	1871	Unclassified	LTR_PERE_10
8	2	1840	1841.5	1843	Unclassified	LTR_PLEU_3
9	2	1193	1506	1819	TE	Lx5c_3end
10	2	1782	1784	1786	Unclassified	LTR_PERE_10
11	2	1557	1579	1601	Unclassified	LTR_PLEU_4
12	2	1031	1111	1191	Unclassified	LTR_PMAN_9
13	2	814	950.5	1087	TE	RMER1C
14	2	1036	1040.5	1045	Unclassified	LTR_PLEU_5
15	2	691	691.5	692	Unclassified	LTR_PLEU_6
16	2	498	499.5	501	Unclassified	LTR_PMAN_11
17	2	383	383.5	384	Unclassified	LTR_PMAN_10
18	2	144	144	144	Unclassified	LTR_PMAN_4
19	2	72	103	134	Unclassified	LTR_PMAN_1
20	2	77	82.5	88	Unclassified	LTR_PLEU_7
21	2	83	83.5	84	Satellite	MMSAT4
22	2	62	62.5	63	Unclassified	LTR_PMAN_5
23	2	58	58	58	Unclassified	LTR_PLEU_8
24	2	54	54	54	Unclassified	LTR_PLEU_9
25	2	51	51	51	Unclassified	LTR_PLEU_10

Supplementary Table 7 – Global clustering of LTRs. LTR classes are ordered by number of arrays in group. Column "Class" refers to the classification from the analysis performed on section **3.1.3**.

Cluster Group	Arrays	min size (bp)	mean size (bp)	max size (bp)	Type	Repeat
1	3684	314	355.7	1719	Satellite	PMSAT
2	359	67	81.7	563	Unclassified	LTR_PMAN_1
3	190	74	119.4	562	Unclassified	LTR_PMAN_2
4	55	83	84	85	Satellite	MurSatRep1
5	54	1010	1375.9	1882	TE	L1_Mur2_orf2
6	28	1025	1361.1	1865	Unclassified	LTR_PCAL_1
7	28	66	71.8	72	Unclassified	LTR_PMAN_4
8	23	93	151.7	327	Unclassified	LTR_PMAN_2-2
9	15	69	69	69	Unclassified	LTR_PMAN_7
10	13	62	62.9	63	Unclassified	LTR_PMAN_5
11	11	1015	1365.5	1742	Unclassified	LTR_PERE_2
12	11	79	101.7	162	Unclassified	LTR_PCAL_6
13	10	1707	1755.5	1784	Unclassified	LTR_PCAL_2
14	9	104	147.2	243	TE	B1_Mus1
15	8	1038	1657	1865	TE	L1_Mur2_orf2
16	7	309	309.9	310	Unclassified	LTR_PCAL_5
17	7	83	83.9	84	Satellite	MurSatRep1
18	7	83	83.9	84	Satellite	MMSAT4
19	6	1273	1543.7	1687	Unclassified	LTR_PCAL_4
20	6	685	686.3	687	Unclassified	LTR_PCAL_3
21	6	72	72	72	Unclassified	LTR_PMAN_12
22	5	1398	1772.6	1980	Unclassified	LTR_PERE_10
23	5	1122	1400.2	1641	Unclassified	LTR_PERE_3
24	5	803	1086.8	1361	TE	RMER1C
25	5	71	72	73	Unclassified	LTR_PMAN_13
26	5	55	55.8	56	Unclassified	LTR_PCAL_16-2
27	4	1241	1710	1948	Unclassified	LTR_PMAN_14
28	4	1259	1508	1894	TE	L1_Mur2_orf2
29	4	1554	1576.5	1601	Unclassified	LTR_PLEU_4
30	4	1197	1378.8	1465	Unclassified	LTR_PCAL_9
31	4	1136	1223	1310	Unclassified	LTR_PERE_5
32	4	114	114.8	115	Unclassified	LTR_PCAL_7
33	4	84	84	84	Satellite	MMSAT4
34	4	77	77	77	Unclassified	LTR_PCAL_18
35	4	51	51	51	Unclassified	LTR_PCAL_17
36	3	1845	1905.3	1947	Unclassified	LTR_PERE_14
37	3	1055	1477.3	1894	Unclassified	LTR_PERE_6

38	3	1093	1574.7	1832	Unclassified	LTR_PERE_7
39	3	1448	1469.3	1506	Unclassified	LTR_PERE_1
40	3	1033	1223.3	1491	Unclassified	LTR_PERE_8
41	3	1043	1142.3	1340	TE	L1_Mur2_orf2
42	3	1112	1165	1260	Unclassified	LTR_PERE_1
43	3	1091	1191.3	1247	Unclassified	LTR_PERE_9
44	3	1031	1084.3	1191	Unclassified	LTR_PMAN_9
45	3	779	926	1006	Unclassified	LTR_PMAN_10
46	3	168	201.7	269	Satellite	MurSatRep1
47	3	167	167.7	168	Satellite	MMSAT4
48	3	126	140	168	Unclassified	LTR_PCAL_19
49	3	144	144	144	Unclassified	LTR_PMAN_4
50	3	99	99	99	Unclassified	LTR_PCAL_20
51	3	71	71.7	72	Unclassified	LTR_PMAN_15
52	3	56	56	56	Unclassified	LTR_PCAL_16
53	2	1214	1596.5	1979	Unclassified	LTR_PERE_11
54	2	1049	1512.5	1976	Unclassified	LTR_PERE_12
55	2	1960	1960	1960	Unclassified	LTR_PERE_13
56	2	1915	1915	1915	Unclassified	LTR_PERE_15
57	2	1899	1899	1899	Unclassified	LTR_PERE_16
58	2	1860	1862.5	1865	Unclassified	LTR_PERE_17
59	2	1244	1536	1828	Unclassified	LTR_PERE_18
60	2	1603	1660	1717	Unclassified	LTR_PCAL_8
61	2	1701	1701	1701	Unclassified	LTR_PERE_20
62	2	1320	1495	1670	Unclassified	LTR_PERE_21
63	2	1326	1483.5	1641	Unclassified	LTR_PERE_22
64	2	1335	1462.5	1590	TE	MYSERV
65	2	1537	1537	1537	Unclassified	LTR_PERE_24
66	2	1266	1375.5	1485	Unclassified	LTR_PERE_25
67	2	1102	1258	1414	Unclassified	LTR_PERE_26
68	2	1363	1363	1363	Unclassified	LTR_PCAL_10
69	2	1359	1359	1359	Unclassified	LTR_PERE_27
70	2	1222	1287	1352	Unclassified	LTR_PERE_28
71	2	1283	1314	1345	Unclassified	LTR_PERE_29
72	2	1079	1211	1343	Unclassified	LTR_PERE_30
73	2	1342	1342	1342	Unclassified	LTR_PERE_31
74	2	1091	1215	1339	TE	ERVB4_1-I_MM
75	2	1310	1310	1310	Unclassified	LTR_PERE_32
76	2	1081	1178.5	1276	Unclassified	LTR_PERE_34
77	2	1145	1210	1275	Unclassified	LTR_PERE_35
78	2	1270	1270	1270	Unclassified	LTR_PERE_36

79	2	1065	1153	1241	Unclassified	LTR_PERE_37
80	2	1053	1135	1217	Unclassified	LTR_PERE_38
81	2	1216	1216	1216	Unclassified	LTR_PERE_39
82	2	1007	1097.5	1188	Unclassified	LTR_PERE_40
83	2	1174	1174	1174	Unclassified	LTR_PERE_41
84	2	1144	1144	1144	Unclassified	LTR_PERE_43
85	2	1023	1067	1111	Unclassified	LTR_PERE_44
86	2	1076	1076	1076	Unclassified	LTR_PERE_46
87	2	1044	1045.5	1047	Unclassified	LTR_PERE_47
88	2	1022	1022	1022	Unclassified	LTR_PERE_48
89	2	1015	1015.5	1016	Unclassified	LTR_PERE_49
90	2	674	675.5	677	Satellite	PMSAT
91	2	503	531.5	560	Unclassified	LTR_PCAL_11
92	2	235	235.5	236	Unclassified	LTR_PCAL_12
93	2	171	174	177	Unclassified	LTR_PMAN_16
94	2	78	102.5	127	Unclassified	LTR_PCAL_13
95	2	75	87.5	100	Unclassified	LTR_PCAL_14
96	2	75	75	75	Unclassified	LTR_PMAN_17
97	2	63	63	63	Unclassified	LTR_PERE_50

Supplementary Table 8 – Chromosomes length of the four Peromyscus spp. for relative positioning analysis of LTRs.

PMAN		PCAL		PERE		PLEU	
Chr	Length (bp)	Chr	Length (bp)	Chr	Length (bp)	Chr	Length (bp)
1	193,310,054	1	190,877,991	1	203,417,028	1	193,658,164
2	168,715,211	2	166,545,015	2	176,795,481	2	154,649,009
3	161,151,335	3	160,280,158	3	171,116,754	3	159,738,685
4	154,712,973	4	153,108,158	4	153,290,812	4	151,869,327
5	139,359,418	5	135,108,212	5	147,472,828	5	145,700,154
6	134,808,455	6	132,295,648	6	140,708,079	6	133,087,326
7	119,402,526	7	116,922,554	7	125,590,875	7	118,845,208
8	108,536,456	8	97,107,874	8	114,385,551	8a	58,348,646
9	115,033,041	9	113,663,737	9	119,957,392	8b	106,604,048
10	98,965,349	10	97,435,531	10	104,037,021	9	114,273,790
11	94,517,625	11	92,652,276	11	98,732,353	10	93,642,674
12	83,173,863	12	80,356,967	12	85,538,547	11	69,407,300
13	65,685,024	13	64,292,899	13	70,523,061	12	83,119,451
14	88,525,187	14	85,699,326	14	91,988,365	13	66,178,483
15	78,974,444	15	77,065,821	15	80,037,048	14	91,293,109
16	63,928,082	16	62,407,110	16	68,911,995	15	90,487,634
17	63,635,831	17	61,658,153	17	67,543,381	16+21	79,846,141
18	46,762,208	18	45,679,219	18	50,732,679	17	56,153,843
19	79,940,924	19	77,901,386	19	83,049,736	18	47,548,717
20	70,296,724	20	68,216,057	20	73,759,366	19	84,080,322
21	70,257,074	21	68,451,442	21	74,320,920	20	65,585,891
22	54,709,200	22	54,699,646	22	58,711,749	22	56,493,492
23	47,673,962	23	45,237,853	23	48,037,570	23	46,490,564
X	134,369,076	X	132,577,752	X	132,284,738	X	138,232,706

Supplementary Table 9 – Coordinates, orientation, and array length of LTRs detected on chromosomes from *Peromyscus* genus assemblies. Orientation baseline is positive based on the monomer from the LTR represented in bold, corresponding to its cluster group.

Species	Chromosome	start	end	LTR	strand	length	cluster group
PCAL	1	6,257,712	6,259,835	MMSAT4	plus	2,124	47
PCAL	1	36,221,457	36,229,297	PMSAT	minus	7,841	1
PCAL	1	36,288,558	36,293,164	PMSAT	plus	4,607	1
PCAL	1	36,297,146	36,299,816	PMSAT	plus	2,671	1
PCAL	1	36,300,957	36,305,349	PMSAT	plus	4,393	1
PCAL	1	36,312,220	36,326,817	PMSAT	plus	14,598	1
PCAL	1	36,351,232	36,364,716	PMSAT	plus	13,485	1
PCAL	1	36,364,712	36,369,922	PMSAT	plus	5,211	1
PCAL	1	36,369,952	36,375,783	PMSAT	plus	5,832	1

PCAL	1	36,398,334	36,401,849	PMSAT	plus	3,516	1
PCAL	1	36,402,017	36,404,879	PMSAT	plus	2,863	1
PCAL	1	36,404,882	36,407,057	PMSAT	plus	2,176	1
PCAL	1	36,414,251	36,416,948	PMSAT	plus	2,698	1
PCAL	1	36,418,178	36,420,429	PMSAT	plus	2,252	1
PCAL	1	36,420,452	36,424,666	PMSAT	plus	4,215	1
PCAL	1	36,428,611	36,432,520	PMSAT	plus	3,910	1
PCAL	1	36,434,246	36,445,748	PMSAT	plus	11,503	1
PCAL	1	36,918,392	36,921,154	PMSAT	plus	2,763	1
PCAL	1	36,921,256	36,924,927	PMSAT	minus	3,672	1
PCAL	1	36,926,231	36,929,437	PMSAT	minus	3,207	1
PCAL	1	36,932,859	36,935,570	PMSAT	plus	2,712	1
PCAL	1	36,946,498	36,949,452	PMSAT	plus	2,955	1
PCAL	1	36,989,817	36,992,588	PMSAT	plus	2,772	1
PCAL	1	36,993,030	36,998,088	PMSAT	plus	5,059	1
PCAL	1	37,314,775	37,321,253	PMSAT	plus	6,479	1
PCAL	1	37,323,548	37,327,216	PMSAT	plus	3,669	1
PCAL	1	37,419,242	37,423,009	PMSAT	plus	3,768	1
PCAL	1	37,455,481	37,458,792	PMSAT	plus	3,312	1
PCAL	1	37,461,040	37,463,102	PMSAT	plus	2,063	1
PCAL	1	37,499,487	37,501,750	PMSAT	minus	2,264	1
PCAL	1	37,509,483	37,511,869	PMSAT	minus	2,387	1
PCAL	1	37,518,965	37,523,242	PMSAT	minus	4,278	1
PCAL	1	37,542,593	37,545,027	PMSAT	plus	2,435	1
PCAL	**1**	**37,546,172**	**37,548,237**	**PMSAT**	**plus**	2,066	**90**
PCAL	1	37,561,342	37,563,799	PMSAT	plus	2,458	1
PCAL	1	37,935,149	37,938,200	MurSatRep1	plus	3,052	17
PCAL	1	190,872,036	190,875,191	PMSAT	minus	3,156	1
PCAL	2	131,040,614	131,048,452	PMSAT	plus	7,839	1
PCAL	2	131,048,553	131,050,967	PMSAT	plus	2,415	1
PCAL	2	131,056,406	131,059,862	PMSAT	minus	3,457	1
PCAL	2	131,059,862	131,077,165	PMSAT	minus	17,304	1
PCAL	2	131,085,752	131,098,344	PMSAT	minus	12,593	1
PCAL	2	131,105,306	131,109,146	PMSAT	minus	3,841	1
PCAL	2	131,115,894	131,127,484	PMSAT	minus	11,591	1
PCAL	2	131,127,635	131,131,054	PMSAT	minus	3,420	1
PCAL	2	131,133,984	131,139,738	PMSAT	minus	5,755	1
PCAL	2	131,143,224	131,150,585	PMSAT	minus	7,362	1
PCAL	2	131,166,283	131,172,979	PMSAT	minus	6,697	1
PCAL	2	131,172,985	131,178,904	PMSAT	minus	5,920	1
PCAL	3	160,237,787	160,244,137	PMSAT	plus	6,351	1

PCAL	3	160,244,142	160,246,664	PMSAT	plus	2,523	1
PCAL	3	160,246,657	160,251,279	PMSAT	plus	4,623	1
PCAL	3	160,258,616	160,262,942	PMSAT	plus	4,327	1
PCAL	3	160,270,657	160,273,564	PMSAT	minus	2,908	1
PCAL	3	160,274,377	160,280,158	PMSAT	plus	5,782	1
PCAL	4	148,837,226	148,840,657	LTR_PMAN_7	plus	3,432	9
PCAL	4	148,842,294	148,846,499	LTR_PMAN_7	plus	4,206	9
PCAL	4	148,848,479	148,853,642	LTR_PMAN_7	plus	5,164	9
PCAL	4	148,858,917	148,867,521	LTR_PMAN_7	plus	8,605	9
PCAL	**5**	**46,235,244**	**46,257,187**	**LTR_PCAL_1**	plus	21,944	**6**
PCAL	5	126,893,837	126,897,668	PMSAT	minus	3,832	1
PCAL	5	126,900,788	126,903,196	PMSAT	minus	2,409	1
PCAL	5	126,904,328	126,906,788	PMSAT	minus	2,461	1
PCAL	5	126,909,170	126,911,675	PMSAT	minus	2,506	1
PCAL	6	115,261,381	115,264,699	MurSatRep1	plus	3,319	4
PCAL	6	115,454,851	115,456,899	MurSatRep1	minus	2,049	4
PCAL	6	115,595,265	115,598,232	MurSatRep1	plus	2,968	4
PCAL	7	116,829,908	116,832,755	PMSAT	plus	2,848	1
PCAL	7	116,857,674	116,860,190	PMSAT	minus	2,517	1
PCAL	7	116,872,720	116,876,295	PMSAT	plus	3,576	1
PCAL	7	116,890,652	116,896,796	PMSAT	minus	6,145	1
PCAL	7	116,896,912	116,899,688	PMSAT	minus	2,777	1
PCAL	7	116,899,826	116,905,148	PMSAT	plus	5,323	1
PCAL	7	116,918,216	116,921,728	PMSAT	plus	3,513	1
PCAL	8	85,006,204	85,009,019	PMSAT	minus	2,816	1
PCAL	8	85,015,401	85,019,542	PMSAT	minus	4,142	1
PCAL	8	85,020,057	85,023,642	PMSAT	minus	3,586	1
PCAL	8	94,208,343	94,211,140	LTR_PCAL_16-2	plus	2,798	26
PCAL	8	94,232,258	94,234,312	LTR_PCAL_16	minus	2,055	52
PCAL	8	94,254,354	94,256,465	LTR_PCAL_16	minus	2,112	52
PCAL	9	36,311	40,134	PMSAT	minus	3,824	1
PCAL	9	85,874	90,035	PMSAT	minus	4,162	1
PCAL	10	2	3,793	PMSAT	minus	3,792	1
PCAL	10	41,325	46,991	PMSAT	minus	5,667	1
PCAL	10	67,086	73,657	PMSAT	minus	6,572	1
PCAL	11	17,724,187	17,777,210	LTR_PMAN_5	minus	53,024	10
PCAL	11	17,777,317	17,789,374	LTR_PMAN_5	plus	12,058	10
PCAL	11	17,904,819	17,921,440	LTR_PMAN_5	minus	16,622	10
PCAL	12	17,020	28,795	PMSAT	plus	11,776	1
PCAL	12	30,457	34,041	PMSAT	plus	3,585	1
PCAL	12	42,479	46,884	PMSAT	minus	4,406	1

PCAL	12	59,471	61,866	PMSAT	minus	2,396	1
PCAL	12	73,273	77,129	PMSAT	minus	3,857	1
PCAL	12	85,194	89,055	PMSAT	minus	3,862	1
PCAL	12	90,894	96,890	PMSAT	minus	5,997	1
PCAL	**12**	**58,334,439**	**58,336,877**	**MurSatRep1**	plus	2,439	**46**
PCAL	13	2,709	5,075	PMSAT	minus	2,367	1
PCAL	13	5,086	11,443	PMSAT	minus	6,358	1
PCAL	13	20,967	27,520	PMSAT	minus	6,554	1
PCAL	13	37,451	42,177	PMSAT	minus	4,727	1
PCAL	13	60,412	63,254	PMSAT	minus	2,843	1
PCAL	13	65,007	67,324	PMSAT	minus	2,318	1
PCAL	13	67,533	74,107	PMSAT	minus	6,575	1
PCAL	13	83,232	88,309	PMSAT	minus	5,078	1
PCAL	13	88,320	93,738	PMSAT	minus	5,419	1
PCAL	13	96,195	99,394	PMSAT	minus	3,200	1
PCAL	14	53,079	66,788	PMSAT	minus	13,710	1
PCAL	14	76,440	79,369	PMSAT	minus	2,930	1
PCAL	14	87,065	93,070	PMSAT	minus	6,006	1
PCAL	14	93,082	98,951	PMSAT	minus	5,870	1
PCAL	14	112,948	136,882	PMSAT	minus	23,935	1
PCAL	14	150,482	159,250	PMSAT	minus	8,769	1
PCAL	14	170,937	174,186	PMSAT	minus	3,250	1
PCAL	15	427	7,785	PMSAT	minus	7,359	1
PCAL	15	7,785	12,252	PMSAT	minus	4,468	1
PCAL	15	70,594,319	70,596,516	MMSAT4	plus	2,198	33
PCAL	16	48,243,625	48,245,639	PMSAT	minus	2,015	1
PCAL	16	48,245,795	48,252,418	PMSAT	minus	6,624	1
PCAL	16	48,255,394	48,257,681	PMSAT	minus	2,288	1
PCAL	16	48,259,470	48,266,686	PMSAT	minus	7,217	1
PCAL	16	48,275,139	48,279,054	PMSAT	minus	3,916	1
PCAL	17	58,095,270	58,097,554	MurSatRep1	plus	2,285	4
PCAL	17	61,402,799	61,405,557	PMSAT	plus	2,759	1
PCAL	17	61,407,387	61,410,081	PMSAT	plus	2,695	1
PCAL	17	61,417,883	61,421,133	PMSAT	plus	3,251	1
PCAL	17	61,429,946	61,441,174	PMSAT	plus	11,229	1
PCAL	17	61,452,510	61,455,680	PMSAT	plus	3,171	1
PCAL	17	61,473,551	61,476,034	PMSAT	plus	2,484	1
PCAL	17	61,478,588	61,500,930	PMSAT	plus	22,343	1
PCAL	17	61,508,569	61,510,823	PMSAT	plus	2,255	1
PCAL	17	61,510,826	61,515,096	PMSAT	plus	4,271	1
PCAL	17	61,515,361	61,525,802	PMSAT	plus	10,442	1

PCAL	17	61,547,115	61,549,159	PMSAT	minus	2,045	1
PCAL	17	61,555,714	61,567,705	PMSAT	plus	11,992	1
PCAL	17	61,573,457	61,575,828	PMSAT	plus	2,372	1
PCAL	17	61,578,513	61,585,414	PMSAT	plus	6,902	1
PCAL	17	61,585,432	61,588,097	PMSAT	plus	2,666	1
PCAL	17	61,588,104	61,594,666	PMSAT	plus	6,563	1
PCAL	17	61,602,480	61,606,766	PMSAT	plus	4,287	1
PCAL	17	61,607,751	61,611,954	PMSAT	plus	4,204	1
PCAL	17	61,617,051	61,620,023	PMSAT	minus	2,973	1
PCAL	17	61,620,211	61,623,172	PMSAT	plus	2,962	1
PCAL	17	61,628,358	61,631,496	PMSAT	minus	3,139	1
PCAL	17	61,636,465	61,640,924	PMSAT	plus	4,460	1
PCAL	18	42,152,253	42,154,484	MurSatRep1	plus	2,232	4
PCAL	18	45,443,752	45,448,572	PMSAT	plus	4,821	1
PCAL	18	45,646,670	45,654,659	PMSAT	plus	7,990	1
PCAL	18	45,654,818	45,659,908	PMSAT	plus	5,091	1
PCAL	18	45,664,026	45,669,847	PMSAT	plus	5,822	1
PCAL	19	77,844,149	77,848,030	PMSAT	plus	3,882	1
PCAL	19	77,848,028	77,852,322	PMSAT	plus	4,295	1
PCAL	19	77,852,347	77,854,496	PMSAT	plus	2,150	1
PCAL	19	77,863,624	77,874,984	PMSAT	plus	11,361	1
PCAL	19	77,875,339	77,880,006	PMSAT	plus	4,668	1
PCAL	19	77,881,302	77,883,557	PMSAT	plus	2,256	1
PCAL	19	77,885,323	77,898,096	PMSAT	plus	12,774	1
PCAL	20	68,165,772	68,168,385	PMSAT	plus	2,614	1
PCAL	20	68,176,462	68,179,690	PMSAT	plus	3,229	1
PCAL	20	68,179,933	68,190,660	PMSAT	plus	10,728	1
PCAL	20	68,206,282	68,212,894	PMSAT	minus	6,613	1
PCAL	**21**	**28,237,747**	**28,242,474**	**LTR_PLEU_4**	plus	4,728	**29**
PCAL	21	68,375,195	68,378,768	PMSAT	plus	3,574	1
PCAL	21	68,378,768	68,385,513	PMSAT	plus	6,746	1
PCAL	21	68,393,155	68,400,888	PMSAT	plus	7,734	1
PCAL	21	68,418,341	68,425,778	PMSAT	plus	7,438	1
PCAL	22	43,704,648	43,720,837	PMSAT	minus	16,190	1
PCAL	**22**	**49,760,816**	**49,763,210**	**LTR_PCAL_20**	plus	2,395	**50**
PCAL	**22**	**52,028,939**	**52,033,037**	**LTR_PERE_10**	plus	4,099	**22**
PCAL	23	10,244,616	10,247,832	PMSAT	minus	3,217	1
PCAL	23	15,618,129	15,622,026	PMSAT	plus	3,898	1
PCAL	23	15,622,083	15,624,649	PMSAT	plus	2,567	1
PCAL	23	15,625,295	15,627,418	PMSAT	plus	2,124	1
PCAL	23	15,630,697	15,633,634	PMSAT	plus	2,938	1

PCAL	23	15,636,300	15,642,620	PMSAT	plus	6,321	1
PCAL	23	15,648,693	15,657,429	PMSAT	plus	8,737	1
PCAL	23	15,659,063	15,661,258	PMSAT	minus	2,196	1
PCAL	23	15,672,501	15,678,666	PMSAT	minus	6,166	1
PCAL	23	15,678,668	15,682,274	PMSAT	plus	3,607	1
PCAL	23	15,682,272	15,692,779	PMSAT	plus	10,508	1
PCAL	X	1,392	7,860	PMSAT	plus	6,469	1
PCAL	X	10,813,649	10,817,777	LTR_PMAN_13	minus	4,129	25
PCAL	**X**	**11,073,596**	**11,077,917**	**LTR_PCAL_2**	plus	4,322	**13**
PCAL	X	18,359,975	18,365,008	LTR_PCAL_2	plus	5,034	13
PCAL	X	18,365,747	18,370,883	LTR_PCAL_2	plus	5,137	13
PCAL	X	18,597,450	18,606,219	LTR_PCAL_2	minus	8,770	13
PCAL	X	47,116,506	47,122,978	B1_Mus1	plus	6,473	14
PCAL	**X**	**49,589,455**	**49,592,025**	**LTR_PCAL_17**	plus	2,571	**35**
PCAL	X	49,595,244	49,599,267	LTR_PCAL_17	plus	4,024	35
PCAL	**X**	**84,425,742**	**84,429,083**	**LTR_PCAL_5**	plus	3,342	**16**
PCAL	X	112,375,252	112,379,318	LTR_PMAN_15	plus	4,067	51
PCAL	**X**	**112,388,936**	**112,391,234**	**LTR_PCAL_18**	plus	2,299	**34**
PCAL	**X**	**112,552,239**	**112,555,583**	**LTR_PCAL_19**	plus	3,345	**48**
PERE	1	1,617,943	1,620,071	MurSatRep1	minus	2,129	17
PERE	1	1,643,971	1,646,947	MurSatRep1	minus	2,977	17
PERE	1	1,787,828	1,791,667	LTR_PERE_14	minus	3,840	36
PERE	1	2,036,721	2,039,055	PMSAT	minus	2,335	90
PERE	1	2,053,553	2,055,816	PMSAT	plus	2,264	1
PERE	1	2,094,327	2,102,468	PMSAT	plus	8,142	1
PERE	1	2,165,847	2,167,970	PMSAT	plus	2,124	1
PERE	1	2,169,949	2,173,777	PMSAT	plus	3,829	1
PERE	1	2,175,384	2,180,123	PMSAT	plus	4,740	1
PERE	1	2,189,704	2,192,746	PMSAT	minus	3,043	1
PERE	1	2,230,508	2,234,871	PMSAT	minus	4,364	1
PERE	1	2,273,318	2,278,486	PMSAT	plus	5,169	1
PERE	1	2,592,195	2,594,893	MurSatRep1	plus	2,699	4
PERE	1	2,731,441	2,733,568	MurSatRep1	minus	2,128	4
PERE	1	8,542,475	8,545,553	MMSAT4	minus	3,079	18
PERE	1	9,045,931	9,048,146	MMSAT4	minus	2,216	18
PERE	1	9,410,117	9,412,237	MMSAT4	plus	2,121	47
PERE	1	18,919,872	18,926,404	PMSAT	minus	6,533	1
PERE	1	18,929,083	18,932,837	PMSAT	minus	3,755	1
PERE	1	18,937,253	18,940,027	PMSAT	minus	2,775	1
PERE	1	18,940,372	18,945,965	PMSAT	minus	5,594	1
PERE	1	18,959,947	18,962,254	PMSAT	minus	2,308	1

PERE	1	18,962,427	18,968,559	PMSAT	minus	6,133	1
PERE	1	18,974,219	18,978,301	PMSAT	minus	4,083	1
PERE	1	18,978,309	18,980,728	PMSAT	minus	2,420	1
PERE	1	18,982,741	18,994,324	PMSAT	minus	11,584	1
PERE	1	19,054,383	19,057,287	PMSAT	minus	2,905	1
PERE	1	19,058,384	19,062,324	PMSAT	minus	3,941	1
PERE	1	19,079,199	19,085,945	PMSAT	plus	6,747	1
PERE	1	81,209,745	81,223,131	PMSAT	plus	13,387	1
PERE	1	108,135,755	108,138,548	LTR_PCAL_1	minus	2,794	6
PERE	1	187,806,777	187,810,678	LTR_PCAL_1	plus	3,902	6
PERE	2	27,343	30,059	PMSAT	plus	2,717	1
PERE	2	32,030	39,434	PMSAT	plus	7,405	1
PERE	2	41,360	43,735	PMSAT	plus	2,376	1
PERE	2	46,394	49,092	PMSAT	plus	2,699	1
PERE	2	49,104	53,222	PMSAT	plus	4,119	1
PERE	2	60,475	69,391	PMSAT	minus	8,917	1
PERE	2	76,510	87,800	PMSAT	minus	11,291	1
PERE	2	87,810	97,562	PMSAT	minus	9,753	1
PERE	2	45,596,657	45,600,721	PMSAT	minus	4,065	1
PERE	2	45,607,828	45,611,209	PMSAT	plus	3,382	1
PERE	2	45,616,013	45,619,051	PMSAT	minus	3,039	1
PERE	2	158,246,481	158,249,284	PMSAT	minus	2,804	1
PERE	2	176,439,630	176,443,523	LTR_PERE_14	minus	3,894	36
PERE	3	143,524,050	143,526,556	LTR_PCAL_1	minus	2,507	6
PERE	4	54,789,970	54,792,799	LTR_PCAL_1	minus	2,830	6
PERE	4	68,526,831	68,533,084	LTR_PMAN_7	plus	6,254	9
PERE	4	72,738,127	72,750,703	LTR_PMAN_7	plus	12,577	9
PERE	4	72,784,644	72,792,064	PMSAT	plus	7,421	1
PERE	4	72,792,652	72,799,584	PMSAT	plus	6,933	1
PERE	4	131,031,678	131,033,854	LTR_PCAL_1	minus	2,177	6
PERE	5	14,469,975	14,472,427	PMSAT	plus	2,453	1
PERE	6	26,749,772	26,753,579	PMSAT	minus	3,808	1
PERE	6	26,754,753	26,776,696	PMSAT	plus	21,944	1
PERE	6	26,776,915	26,780,910	PMSAT	plus	3,996	1
PERE	6	26,789,110	26,791,137	PMSAT	plus	2,028	1
PERE	6	26,793,811	26,803,338	PMSAT	plus	9,528	1
PERE	6	26,815,426	26,821,309	PMSAT	plus	5,884	1
PERE	6	26,823,764	26,827,337	PMSAT	minus	3,574	1
PERE	6	43,685,445	43,687,845	MurSatRep1	plus	2,401	4
PERE	6	43,743,826	43,745,893	MurSatRep1	plus	2,068	4
PERE	6	43,751,908	43,754,023	MurSatRep1	minus	2,116	4

PERE	6	44,127,724	44,131,042	MurSatRep1	plus	3,319	4
PERE	6	73,020,011	73,022,171	PMSAT	minus	2,161	1
PERE	7	1,451,043	1,456,291	PMSAT	minus	5,249	1
PERE	8	113,961,582	113,963,773	LTR_PCAL_1	minus	2,192	6
PERE	10	103,969,214	103,971,366	PMSAT	plus	2,153	1
PERE	12	70,009,421	70,011,549	MurSatRep1	minus	2,129	4
PERE	12	78,089,088	78,091,853	LTR_PERE_10	minus	2,766	22
PERE	12	78,323,014	78,329,069	PMSAT	plus	6,056	1
PERE	12	78,340,316	78,343,589	PMSAT	plus	3,274	1
PERE	12	78,344,110	78,350,723	PMSAT	plus	6,614	1
PERE	12	78,350,886	78,356,801	PMSAT	plus	5,916	1
PERE	12	78,359,801	78,362,400	PMSAT	plus	2,600	1
PERE	12	78,363,676	78,365,819	PMSAT	plus	2,144	1
PERE	12	78,365,821	78,368,982	PMSAT	plus	3,162	1
PERE	13	5,280,120	5,286,168	PMSAT	minus	6,049	1
PERE	13	5,286,340	5,326,163	PMSAT	minus	39,824	1
PERE	13	31,946,095	31,948,906	PMSAT	plus	2,812	1
PERE	14	8,292,490	8,295,152	PMSAT	plus	2,663	1
PERE	14	78,631,334	78,634,548	PMSAT	plus	3,215	1
PERE	16	19,399,899	19,415,623	PMSAT	minus	15,725	1
PERE	17	2,916,367	2,918,522	MurSatRep1	minus	2,156	4
PERE	17	3,100,913	3,103,020	MurSatRep1	minus	2,108	4
PERE	17	7,994,202	7,996,372	PMSAT	plus	2,171	1
PERE	17	8,004,067	8,007,924	PMSAT	plus	3,858	1
PERE	17	51,396,259	51,407,879	PMSAT	plus	11,621	1
PERE	17	51,409,053	51,413,110	PMSAT	plus	4,058	1
PERE	18	3,594,736	3,598,316	PMSAT	minus	3,581	1
PERE	20	58,870,007	58,872,808	LTR_PCAL_1	minus	2,802	6
PERE	21	62,785,523	62,790,369	PMSAT	minus	4,847	1
PERE	21	62,797,913	62,800,608	PMSAT	minus	2,696	1
PERE	21	62,802,442	62,807,520	PMSAT	minus	5,079	1
PERE	21	62,812,213	62,815,370	PMSAT	minus	3,158	1
PERE	21	62,827,336	62,830,031	PMSAT	minus	2,696	1
PERE	21	62,831,991	62,838,981	PMSAT	minus	6,991	1
PERE	21	63,174,566	63,180,030	PMSAT	plus	5,465	1
PERE	21	63,180,125	63,185,842	PMSAT	minus	5,718	1
PERE	22	33,910,245	33,924,073	PMSAT	plus	13,829	1
PERE	22	56,562,045	56,566,128	LTR_PCAL_20	plus	4,084	50
PERE	23	22,236,625	22,239,706	PMSAT	minus	3,082	1
PERE	23	22,247,797	22,252,816	PMSAT	minus	5,020	1
PERE	23	22,401,568	22,410,826	PMSAT	minus	9,259	1

PERE	23	33,976,349	33,978,752	PMSAT	plus	2,404	1
PERE	23	37,313,250	37,320,232	PMSAT	plus	6,983	1
PERE	X	16,114,181	16,116,679	PMSAT	plus	2,499	1
PERE	X	18,722,671	18,725,390	LTR_PCAL_19	minus	2,720	48
PERE	X	18,899,867	18,902,899	LTR_PCAL_19	plus	3,033	48
PERE	X	19,063,502	19,091,550	LTR_PCAL_18	minus	28,049	34
PERE	X	19,091,556	19,093,874	LTR_PCAL_18	minus	2,319	34
PERE	X	19,093,814	19,097,216	LTR_PCAL_18	minus	3,403	34
PERE	X	47,083,808	47,087,705	LTR_PCAL_5	minus	3,898	16
PERE	X	112,474,817	112,482,493	LTR_PCAL_2	plus	7,677	13
PERE	X	116,863,306	116,868,355	LTR_PCAL_2	minus	5,050	13
PLEU	1	150,397,527	150,408,431	PMSAT	plus	10,905	1
PLEU	1	150,418,438	150,436,405	PMSAT	plus	17,968	1
PLEU	1	150,595,973	150,599,023	MurSatRep1	plus	3,051	17
PLEU	1	152,363,548	152,366,300	PMSAT	plus	2,753	1
PLEU	1	152,366,300	152,369,073	PMSAT	plus	2,774	1
PLEU	1	152,376,774	152,386,927	PMSAT	plus	10,154	1
PLEU	1	188,612,207	188,614,733	MMSAT4	minus	2,527	18
PLEU	1	189,022,583	189,024,734	MMSAT4	minus	2,152	18
PLEU	1	189,310,145	189,312,268	MMSAT4	plus	2,124	47
PLEU	2	11,203,726	11,205,922	MurSatRep1	plus	2,197	4
PLEU	4	142,740,732	142,756,942	LTR_PMAN_7	plus	16,211	9
PLEU	4	142,772,266	142,775,926	LTR_PMAN_7	plus	3,661	9
PLEU	6	16,785,343	16,787,744	MurSatRep1	plus	2,402	4
PLEU	6	17,008,268	17,010,726	MurSatRep1	plus	2,459	4
PLEU	6	17,155,826	17,157,952	MurSatRep1	minus	2,127	4
PLEU	6	17,270,928	17,273,054	MurSatRep1	minus	2,127	4
PLEU	6	17,502,786	17,505,686	MurSatRep1	plus	2,901	4
PLEU	6	17,597,993	17,600,551	MurSatRep1	plus	2,559	4
PLEU	7	66,507,849	66,509,995	PMSAT	plus	2,147	1
PLEU	7	66,510,338	66,517,943	PMSAT	plus	7,606	1
PLEU	7	66,526,480	66,528,694	PMSAT	plus	2,215	1
PLEU	7	66,531,253	66,535,470	PMSAT	plus	4,218	1
PLEU	7	87,345,661	87,351,417	PMSAT	plus	5,757	1
PLEU	9	85,063,855	85,068,511	PMSAT	plus	4,657	1
PLEU	9	85,068,505	85,072,381	PMSAT	plus	3,877	1
PLEU	10	55,937,537	55,943,582	LTR_PERE_10	minus	6,046	22
PLEU	12	83,057,656	83,060,416	PMSAT	plus	2,761	1
PLEU	12	83,116,048	83,119,451	PMSAT	plus	3,404	1
PLEU	13	34	4,055	PMSAT	minus	4,022	1
PLEU	13	10,995	23,533	PMSAT	minus	12,539	1

PLEU	13	36,027	44,873	PMSAT	minus	8,847	1
PLEU	15	74,746,832	74,802,605	LTR_PMAN_5	minus	55,774	10
PLEU	15	76,181,875	76,198,144	LTR_PMAN_5	plus	16,270	10
PLEU	15	89,496,061	89,502,777	PMSAT	plus	6,717	1
PLEU	17	42,760,572	42,763,589	LTR_PCAL_1	plus	3,018	6
PLEU	17	52,409,116	52,411,443	MurSatRep1	plus	2,328	4
PLEU	18	44,531,181	44,534,915	PMSAT	plus	3,735	1
PLEU	19	53,977,317	53,982,629	PMSAT	plus	5,313	1
PLEU	20	40,010,224	40,017,791	PMSAT	minus	7,568	1
PLEU	20	40,022,919	40,035,075	PMSAT	minus	12,157	1
PLEU	20	40,035,483	40,049,394	PMSAT	minus	13,912	1
PLEU	22	2,475,907	2,478,978	LTR_PCAL_20	minus	3,072	50
PLEU	22	11,189,272	11,195,666	PMSAT	minus	6,395	1
PLEU	22	11,196,241	11,206,166	PMSAT	minus	9,926	1
PLEU	22	11,208,215	11,212,297	PMSAT	minus	4,083	1
PLEU	22	11,219,104	11,221,531	PMSAT	minus	2,428	1
PLEU	22	11,221,559	11,225,113	PMSAT	minus	3,555	1
PLEU	22	11,225,111	11,227,932	PMSAT	minus	2,822	1
PLEU	22	11,437,963	11,441,507	PMSAT	plus	3,545	1
PLEU	22	11,442,785	11,446,275	PMSAT	minus	3,491	1
PLEU	22	11,453,137	11,455,446	PMSAT	minus	2,310	1
PLEU	22	11,456,925	11,459,935	PMSAT	minus	3,011	1
PLEU	22	11,459,963	11,463,175	PMSAT	plus	3,213	1
PLEU	22	12,435,517	12,441,563	PMSAT	minus	6,047	1
PLEU	22	12,441,732	12,449,329	PMSAT	minus	7,598	1
PLEU	22	12,449,966	12,458,756	PMSAT	minus	8,791	1
PLEU	22	12,467,222	12,475,290	PMSAT	minus	8,069	1
PLEU	22	12,480,727	12,487,379	PMSAT	minus	6,653	1
PLEU	22	12,487,392	12,491,735	PMSAT	minus	4,344	1
PLEU	23	30,486,697	30,506,114	PMSAT	minus	19,418	1
PLEU	23	30,509,004	30,512,257	PMSAT	minus	3,254	1
PLEU	16+21	32,668,219	32,671,962	LTR_PLEU_4	minus	3,744	29
PLEU	16+21	32,672,534	32,675,636	LTR_PLEU_4	minus	3,103	29
PLEU	16+21	32,840,626	32,844,695	LTR_PLEU_4	plus	4,070	29
PLEU	16+21	69,258,342	69,311,997	PMSAT	minus	53,656	1
PLEU	8a	3	5,464	PMSAT	minus	5,462	1
PLEU	8a	15,208	24,827	PMSAT	minus	9,620	1
PLEU	8a	54,717	57,004	PMSAT	minus	2,288	1
PLEU	8a	57,905	60,456	PMSAT	minus	2,552	1
PLEU	8a	106,265	108,486	PMSAT	minus	2,222	1
PLEU	8b	3,468,791	3,472,363	LTR_PCAL_16-2	plus	3,573	26

PLEU	8b	29,911,725	29,914,532	PMSAT	plus	2,808	1
PLEU	X	3,970,894	3,976,065	MurSatRep1	minus	5,172	4
PLEU	X	4,312,612	4,314,741	MurSatRep1	plus	2,130	4
PLEU	X	4,353,863	4,360,999	PMSAT	minus	7,137	1
PLEU	X	4,363,518	4,368,199	PMSAT	minus	4,682	1
PLEU	X	4,368,196	4,371,770	PMSAT	minus	3,575	1
PLEU	X	4,387,348	4,391,538	PMSAT	plus	4,191	1
PLEU	X	15,005,714	15,008,491	LTR_PMAN_13	minus	2,778	25
PLEU	X	52,245,979	52,271,381	B1_Mus1	plus	25,403	14
PLEU	X	54,651,107	54,663,580	LTR_PCAL_17	plus	12,474	35
PLEU	X	90,333,419	90,362,828	LTR_PCAL_5	plus	29,410	16
PMAN	**1**	**9,689,031**	**9,691,180**	**MMSAT4**	plus	2,150	**18**
PMAN	**1**	**12,959,095**	**12,961,270**	**MurSatRep1**	plus	2,176	**4**
PMAN	**1**	**30,912,716**	**30,915,120**	**PMSAT**	plus	2,405	**1**
PMAN	1	30,915,161	30,921,226	PMSAT	plus	6,066	1
PMAN	1	30,921,224	30,924,012	PMSAT	plus	2,789	1
PMAN	1	30,934,094	30,944,410	PMSAT	plus	10,317	1
PMAN	1	30,946,355	30,954,452	PMSAT	plus	8,098	1
PMAN	**1**	**31,137,865**	**31,140,916**	**MurSatRep1**	plus	3,052	**17**
PMAN	1	32,980,188	32,987,177	PMSAT	minus	6,990	1
PMAN	1	32,989,784	32,996,975	PMSAT	minus	7,192	1
PMAN	1	32,997,647	33,000,336	PMSAT	minus	2,690	1
PMAN	1	33,008,531	33,015,485	PMSAT	minus	6,955	1
PMAN	1	33,016,187	33,018,491	PMSAT	minus	2,305	1
PMAN	1	33,018,490	33,020,660	PMSAT	minus	2,171	1
PMAN	1	33,028,065	33,030,516	PMSAT	minus	2,452	1
PMAN	1	33,038,381	33,048,103	PMSAT	minus	9,723	1
PMAN	1	33,066,302	33,071,475	PMSAT	minus	5,174	1
PMAN	1	33,079,023	33,081,658	PMSAT	minus	2,636	1
PMAN	1	33,081,658	33,098,184	PMSAT	minus	16,527	1
PMAN	1	33,107,737	33,110,530	PMSAT	minus	2,794	1
PMAN	1	33,110,530	33,114,128	PMSAT	minus	3,599	1
PMAN	1	33,200,699	33,203,154	PMSAT	plus	2,456	1
PMAN	1	74,282,880	74,286,322	PMSAT	plus	3,443	1
PMAN	2	42,015,998	42,021,275	PMSAT	plus	5,278	1
PMAN	2	42,021,288	42,025,744	PMSAT	minus	4,457	1
PMAN	2	42,029,764	42,032,256	PMSAT	plus	2,493	1
PMAN	2	42,049,829	42,063,045	PMSAT	plus	13,217	1
PMAN	2	42,091,407	42,093,672	PMSAT	plus	2,266	1
PMAN	2	42,095,730	42,101,305	PMSAT	plus	5,576	1
PMAN	2	42,103,790	42,107,709	PMSAT	plus	3,920	1

PMAN	2	42,107,708	42,111,424	PMSAT	plus	3,717	1
PMAN	2	42,118,168	42,121,077	PMSAT	plus	2,910	1
PMAN	2	42,125,888	42,131,392	PMSAT	plus	5,505	1
PMAN	2	42,134,414	42,137,716	PMSAT	plus	3,303	1
PMAN	2	42,138,103	42,149,757	PMSAT	plus	11,655	1
PMAN	2	42,149,769	42,154,307	PMSAT	plus	4,539	1
PMAN	2	42,160,059	42,178,836	PMSAT	plus	18,778	1
PMAN	2	42,179,784	42,193,615	PMSAT	plus	13,832	1
PMAN	3	28,273,920	28,282,536	PMSAT	plus	8,617	1
PMAN	3	28,284,292	28,287,131	PMSAT	plus	2,840	1
PMAN	3	28,291,290	28,294,355	PMSAT	plus	3,066	1
PMAN	3	28,296,051	28,302,793	PMSAT	plus	6,743	1
PMAN	3	29,567,993	29,576,190	PMSAT	minus	8,198	1
PMAN	3	29,611,000	29,613,001	PMSAT	minus	2,002	1
PMAN	3	29,613,052	29,620,760	PMSAT	minus	7,709	1
PMAN	3	29,621,582	29,628,804	PMSAT	minus	7,223	1
PMAN	3	29,658,700	29,671,030	PMSAT	minus	12,331	1
PMAN	3	41,237,289	41,242,493	PMSAT	plus	5,205	1
PMAN	3	41,279,240	41,290,341	PMSAT	plus	11,102	1
PMAN	3	41,298,685	41,314,244	PMSAT	plus	15,560	1
PMAN	3	41,321,136	41,333,144	PMSAT	plus	12,009	1
PMAN	3	41,350,347	41,357,663	PMSAT	plus	7,317	1
PMAN	4	53,980	60,788	PMSAT	minus	6,809	1
PMAN	4	63,538	65,718	PMSAT	minus	2,181	1
PMAN	4	68,573	73,086	PMSAT	minus	4,514	1
PMAN	4	96,778	105,194	PMSAT	minus	8,417	1
PMAN	4	110,121	128,398	PMSAT	minus	18,278	1
PMAN	4	128,404	136,887	PMSAT	minus	8,484	1
PMAN	4	141,894	144,987	PMSAT	minus	3,094	1
PMAN	4	145,461	152,975	PMSAT	minus	7,515	1
PMAN	4	152,985	168,626	PMSAT	minus	15,642	1
PMAN	4	168,646	175,593	PMSAT	minus	6,948	1
PMAN	**4**	**150,553,946**	**150,570,566**	**LTR_PMAN_7**	plus	16,621	**9**
PMAN	4	150,579,699	150,587,469	LTR_PMAN_7	plus	7,771	9
PMAN	5	32,694,927	32,746,417	PMSAT	plus	51,491	1
PMAN	5	32,748,031	32,752,361	PMSAT	plus	4,331	1
PMAN	5	32,775,115	32,782,339	PMSAT	plus	7,225	1
PMAN	5	32,791,375	32,794,645	PMSAT	plus	3,271	1
PMAN	5	32,794,643	32,800,758	PMSAT	plus	6,116	1
PMAN	5	32,825,720	32,829,480	PMSAT	plus	3,761	1
PMAN	5	32,844,206	32,850,509	PMSAT	plus	6,304	1

PMAN	6	20,471,665	20,474,066	MurSatRep1	plus	2,402	4
PMAN	6	20,904,721	20,907,631	MurSatRep1	plus	2,911	4
PMAN	6	21,040,340	21,042,413	MurSatRep1	minus	2,074	4
PMAN	6	21,351,929	21,378,391	PMSAT	minus	26,463	1
PMAN	6	21,380,331	21,383,520	PMSAT	minus	3,190	1
PMAN	6	21,392,710	21,397,349	PMSAT	minus	4,640	1
PMAN	6	21,399,618	21,403,975	PMSAT	minus	4,358	1
PMAN	6	21,403,975	21,412,725	PMSAT	minus	8,751	1
PMAN	6	21,412,741	21,415,456	PMSAT	minus	2,716	1
PMAN	6	25,255,367	25,258,138	MurSatRep1	minus	2,772	4
PMAN	7	10,686,051	10,688,242	PMSAT	plus	2,192	1
PMAN	7	10,688,516	10,694,199	PMSAT	plus	5,684	1
PMAN	7	10,694,370	10,697,833	PMSAT	plus	3,464	1
PMAN	7	10,706,941	10,711,285	PMSAT	plus	4,345	1
PMAN	7	10,711,351	10,714,239	PMSAT	plus	2,889	1
PMAN	7	10,722,328	10,738,511	PMSAT	plus	16,184	1
PMAN	7	10,749,705	10,753,909	PMSAT	plus	4,205	1
PMAN	7	10,754,064	10,756,489	PMSAT	plus	2,426	1
PMAN	7	10,756,645	10,759,465	PMSAT	plus	2,821	1
PMAN	7	10,759,462	10,763,031	PMSAT	plus	3,570	1
PMAN	7	10,763,754	10,768,252	PMSAT	plus	4,499	1
PMAN	7	10,768,257	10,775,302	PMSAT	plus	7,046	1
PMAN	7	10,775,605	10,778,777	PMSAT	plus	3,173	1
PMAN	7	10,779,415	10,789,404	PMSAT	plus	9,990	1
PMAN	7	10,789,404	10,794,489	PMSAT	plus	5,086	1
PMAN	7	10,799,156	10,804,601	PMSAT	plus	5,446	1
PMAN	**8**	**9,776,467**	**9,778,502**	**LTR_PCAL_16-2**	plus	2,036	**26**
PMAN	9	58,330	63,937	PMSAT	minus	5,608	1
PMAN	9	23,462,024	23,464,503	PMSAT	plus	2,480	1
PMAN	9	23,464,505	23,466,709	PMSAT	plus	2,205	1
PMAN	9	23,467,275	23,470,772	PMSAT	plus	3,498	1
PMAN	9	23,471,151	23,479,944	PMSAT	plus	8,794	1
PMAN	9	23,484,904	23,491,031	PMSAT	plus	6,128	1
PMAN	9	23,496,204	23,498,653	PMSAT	plus	2,450	1
PMAN	9	23,505,335	23,508,298	PMSAT	plus	2,964	1
PMAN	9	23,508,312	23,511,388	PMSAT	plus	3,077	1
PMAN	9	23,541,599	23,544,785	PMSAT	plus	3,187	1
PMAN	9	23,546,719	23,548,756	PMSAT	plus	2,038	1
PMAN	9	23,549,472	23,554,366	PMSAT	plus	4,895	1
PMAN	9	23,554,357	23,565,343	PMSAT	plus	10,987	1
PMAN	9	23,565,921	23,568,195	PMSAT	plus	2,275	1

PMAN	9	23,568,176	23,570,824	PMSAT	plus	2,649	1
PMAN	9	23,576,564	23,580,692	PMSAT	plus	4,129	1
PMAN	9	23,583,943	23,592,398	PMSAT	plus	8,456	1
PMAN	9	30,137,723	30,143,358	PMSAT	plus	5,636	1
PMAN	9	30,143,354	30,159,738	PMSAT	plus	16,385	1
PMAN	9	30,160,784	30,163,790	PMSAT	plus	3,007	1
PMAN	9	30,167,305	30,173,164	PMSAT	plus	5,860	1
PMAN	11	11,249,220	11,251,937	PMSAT	plus	2,718	1
PMAN	11	11,255,517	11,261,648	PMSAT	plus	6,132	1
PMAN	11	11,261,845	11,276,091	PMSAT	plus	14,247	1
PMAN	11	11,288,882	11,290,957	PMSAT	plus	2,076	1
PMAN	11	11,291,127	11,298,703	PMSAT	plus	7,577	1
PMAN	11	11,299,395	11,302,409	PMSAT	plus	3,015	1
PMAN	11	11,302,408	11,306,676	PMSAT	plus	4,269	1
PMAN	11	11,310,205	11,318,048	PMSAT	plus	7,844	1
PMAN	11	11,320,141	11,329,414	PMSAT	plus	9,274	1
PMAN	11	11,329,412	11,332,166	PMSAT	plus	2,755	1
PMAN	11	11,332,827	11,335,835	PMSAT	plus	3,009	1
PMAN	11	11,335,833	11,338,587	PMSAT	plus	2,755	1
PMAN	11	11,338,637	11,341,223	PMSAT	plus	2,587	1
PMAN	**11**	**13,143,449**	**13,158,518**	**LTR_PMAN_5**	plus	15,070	**10**
PMAN	12	64,196	67,400	PMSAT	minus	3,205	1
PMAN	12	89,843	92,955	PMSAT	minus	3,113	1
PMAN	13	5,394,759	5,398,467	PMSAT	minus	3,709	1
PMAN	13	5,400,150	5,408,606	PMSAT	minus	8,457	1
PMAN	13	5,408,630	5,414,781	PMSAT	minus	6,152	1
PMAN	13	5,414,945	5,420,369	PMSAT	minus	5,425	1
PMAN	13	5,420,369	5,427,641	PMSAT	minus	7,273	1
PMAN	13	5,431,234	5,434,741	PMSAT	minus	3,508	1
PMAN	13	5,436,242	5,439,485	PMSAT	minus	3,244	1
PMAN	13	5,440,792	5,443,015	PMSAT	minus	2,224	1
PMAN	13	28,791,687	28,794,352	PMSAT	minus	2,666	1
PMAN	14	7,325,814	7,330,126	PMSAT	plus	4,313	1
PMAN	14	40,280,884	40,283,454	PMSAT	plus	2,571	1
PMAN	14	40,283,770	40,285,952	PMSAT	plus	2,183	1
PMAN	14	40,286,611	40,288,674	PMSAT	plus	2,064	1
PMAN	14	40,293,530	40,298,437	PMSAT	plus	4,908	1
PMAN	14	40,302,285	40,305,398	PMSAT	plus	3,114	1
PMAN	14	40,320,455	40,324,610	PMSAT	plus	4,156	1
PMAN	14	40,328,131	40,330,338	PMSAT	plus	2,208	1
PMAN	14	40,330,339	40,335,239	PMSAT	plus	4,901	1

PMAN	14	40,343,568	40,349,377	PMSAT	plus	5,810	1
PMAN	15	6,809,273	6,811,448	MMSAT4	minus	2,176	33
PMAN	15	78,959,084	78,962,137	PMSAT	plus	3,054	1
PMAN	15	78,963,788	78,967,819	PMSAT	plus	4,032	1
PMAN	17	204	7,195	PMSAT	minus	6,992	1
PMAN	17	17,620	20,991	PMSAT	minus	3,372	1
PMAN	17	21,407	25,142	PMSAT	minus	3,736	1
PMAN	17	25,688	31,073	PMSAT	minus	5,386	1
PMAN	18	161,703	163,916	PMSAT	minus	2,214	1
PMAN	19	47,587	54,029	PMSAT	minus	6,443	1
PMAN	19	86,978	89,421	PMSAT	minus	2,444	1
PMAN	19	91,746	97,985	PMSAT	minus	6,240	1
PMAN	19	100,508	103,567	PMSAT	minus	3,060	1
PMAN	19	105,002	110,219	PMSAT	minus	5,218	1
PMAN	19	110,220	115,613	PMSAT	minus	5,394	1
PMAN	19	122,795	131,218	PMSAT	minus	8,424	1
PMAN	19	131,934	143,559	PMSAT	minus	11,626	1
PMAN	19	143,723	147,787	PMSAT	minus	4,065	1
PMAN	19	147,802	149,860	PMSAT	minus	2,059	1
PMAN	19	149,886	152,342	PMSAT	minus	2,457	1
PMAN	19	153,474	155,742	PMSAT	minus	2,269	1
PMAN	19	155,736	159,056	PMSAT	minus	3,321	1
PMAN	19	160,868	163,432	PMSAT	minus	2,565	1
PMAN	19	183,124	188,640	PMSAT	minus	5,517	1
PMAN	20	30,194,600	30,197,366	PMSAT	plus	2,767	1
PMAN	20	30,211,528	30,222,711	PMSAT	plus	11,184	1
PMAN	20	30,225,421	30,235,020	PMSAT	plus	9,600	1
PMAN	20	30,254,009	30,256,238	PMSAT	plus	2,230	1
PMAN	20	30,261,367	30,263,820	PMSAT	plus	2,454	1
PMAN	**20**	**42,528,065**	**42,531,764**	**LTR_PERE_14**	plus	3,700	**36**
PMAN	21	57,298,381	57,301,015	PMSAT	plus	2,635	1
PMAN	21	57,310,208	57,315,857	PMSAT	plus	5,650	1
PMAN	21	57,319,091	57,321,212	PMSAT	plus	2,122	1
PMAN	21	57,321,239	57,324,034	PMSAT	plus	2,796	1
PMAN	21	57,330,119	57,333,921	PMSAT	plus	3,803	1
PMAN	22	4,593	13,148	PMSAT	minus	8,556	1
PMAN	22	13,159	17,557	PMSAT	minus	4,399	1
PMAN	22	29,116	37,416	PMSAT	plus	8,301	1
PMAN	22	45,448	53,884	PMSAT	plus	8,437	1
PMAN	22	53,882	57,847	PMSAT	plus	3,966	1
PMAN	22	68,897	71,325	PMSAT	plus	2,429	1

PMAN	22	71,355	73,420	PMSAT	plus	2,066	1
PMAN	22	76,757	82,109	PMSAT	plus	5,353	1
PMAN	22	82,296	85,424	PMSAT	plus	3,129	1
PMAN	22	6,835,561	6,843,988	PMSAT	minus	8,428	1
PMAN	22	6,844,335	6,847,178	PMSAT	minus	2,844	1
PMAN	22	6,847,176	6,849,501	PMSAT	minus	2,326	1
PMAN	22	6,852,506	6,856,174	PMSAT	minus	3,669	1
PMAN	22	6,856,177	6,860,282	PMSAT	minus	4,106	1
PMAN	22	6,860,293	6,864,040	PMSAT	minus	3,748	1
PMAN	22	6,864,038	6,874,251	PMSAT	minus	10,214	1
PMAN	22	6,875,930	6,879,352	PMSAT	minus	3,423	1
PMAN	22	6,879,355	6,884,659	PMSAT	minus	5,305	1
PMAN	22	6,888,669	6,891,295	PMSAT	minus	2,627	1
PMAN	22	6,892,553	6,898,665	PMSAT	minus	6,113	1
PMAN	22	6,898,663	6,904,015	PMSAT	minus	5,353	1
PMAN	22	6,905,428	6,921,238	PMSAT	minus	15,811	1
PMAN	22	6,929,690	6,932,113	PMSAT	minus	2,424	1
PMAN	22	6,946,904	6,949,158	PMSAT	minus	2,255	1
PMAN	22	6,951,053	6,953,504	PMSAT	minus	2,452	1
PMAN	22	6,955,326	6,957,856	PMSAT	minus	2,531	1
PMAN	22	6,958,967	6,962,596	PMSAT	plus	3,630	1
PMAN	22	6,963,035	6,965,347	PMSAT	plus	2,313	1
PMAN	22	6,974,820	6,979,651	PMSAT	minus	4,832	1
PMAN	22	10,337,713	10,350,588	PMSAT	minus	12,876	1
PMAN	22	10,350,588	10,354,793	PMSAT	minus	4,206	1
PMAN	22	10,355,266	10,368,040	PMSAT	plus	12,775	1
PMAN	22	10,368,261	10,373,435	PMSAT	plus	5,175	1
PMAN	22	10,373,433	10,381,428	PMSAT	plus	7,996	1
PMAN	22	10,381,425	10,383,521	PMSAT	plus	2,097	1
PMAN	22	10,383,518	10,385,609	PMSAT	plus	2,092	1
PMAN	22	10,385,852	10,390,344	PMSAT	plus	4,493	1
PMAN	22	10,396,311	10,402,082	PMSAT	plus	5,772	1
PMAN	22	10,402,925	10,406,471	PMSAT	plus	3,547	1
PMAN	22	10,409,122	10,411,233	PMSAT	plus	2,112	1
PMAN	22	10,441,597	10,447,045	PMSAT	plus	5,449	1
PMAN	22	10,448,212	10,451,494	PMSAT	plus	3,283	1
PMAN	22	10,451,493	10,455,612	PMSAT	plus	4,120	1
PMAN	22	10,455,920	10,459,048	PMSAT	plus	3,129	1
PMAN	22	10,459,254	10,462,846	PMSAT	plus	3,593	1
PMAN	23	30,282,139	30,288,967	PMSAT	minus	6,829	1
PMAN	23	30,288,978	30,292,529	PMSAT	minus	3,552	1

PMAN	23	30,320,614	30,328,482	PMSAT	minus	7,869	1
PMAN	23	30,329,930	30,333,022	PMSAT	minus	3,093	1
PMAN	23	30,333,035	30,337,917	PMSAT	minus	4,883	1
PMAN	23	30,349,322	30,353,194	PMSAT	minus	3,873	1
PMAN	**X**	**33,084,937**	**33,087,140**	**LTR_PMAN_15**	plus	2,204	**51**
PMAN	**X**	**84,617,938**	**84,621,964**	**B1_Mus1**	plus	4,027	**14**
PMAN	X	84,625,322	84,630,117	B1_Mus1	plus	4,796	14
PMAN	**X**	**114,208,799**	**114,214,221**	**LTR_PMAN_13**	plus	5,423	**25**
PMAN	X	133,180,330	133,183,284	PMSAT	plus	2,955	1

Supplementary Table 10 – Distribution of PMSat arrays on chromosomes across the four *Peromyscus* genomes. * The joint total of arrays calculated for chromosome 8 was done with both chromosomes 8a and 8b together (0+3+0+5+1), and 16 and 21 was done with chromosome 16_21 for each (chr16: 0+5+1+1; chr21: 5+4+8+1).

	PMAN	**PCAL**	**PERE**		**PLEU**		
Chromosome	Arrays	Arrays	Arrays	Chromosome	Arrays	Chromosome	TOTAL
1	20	34	20	1	5	1	79
2	15	12	12	2	0	2	39
3	14	6	0	3	0	3	20
4	10	0	2	4	0	4	12
5	7	4	1	5	0	5	12
6	6	0	8	6	0	6	14
7	16	7	1	7	5	7	29
8	0	3	0	8a	5	8*	9
9	21	2	0	8b	1	9	26
10	0	2	1	9	3	10	3
11	13	0	0	10	0	11	13
12	2	7	7	11	0	12	18
13	8	10	2	12	2	13	23
14	10	7	2	13	3	14	19
15	2	2	0	14	0	15	5
16	0	5	1	15	1	16*	7
17	4	22	4	16_21	1	17	30
18	1	4	1	17	0	18	7
19	15	7	0	18	1	19	23
20	5	4	0	19	1	20	12
21	5	4	8	20	3	21*	18
22	45	1	1	22	16	22	63
23	6	11	4	23	2	23	23
X	1	1	1	X	4	X	7
Total each genome	226	155	76		53	Joint total	511

Supplementary Table 11 – Clustered groups of PMSat based on identity score threshold of 0.9. PMSat monomers from each array were clustered based on their identity score with MeShClust v3.0 (Girgis, 2022). The letters on column 'C/M/E' represent the classification within the clustered group: C – Center (identity score ≈ 1); M – Member (0.9 < identity score < 1); E – Extended Member (identity score < 0.9).

Cluster group	Assembly	Chromosome	Start	End	Identity score	C/M/E	Length (bp)
1	PCAL	1	36221457	36229297	0.8809	E	7,841
1	PCAL	5	126909170	126911675	0.9089	M	2,506
1	PERE	7	1451043	1456291	0.9125	M	5,249
1	PCAL	12	85194	89055	0.8814	E	3,862
1	PCAL	12	90894	96890	0.9272	M	5,997
1	PMAN	12	64196	67400	0.8831	E	3,205
1	PMAN	12	89843	92955	0.9345	M	3,113
1	PCAL	13	96195	99394	0.8905	E	3,200
1	PERE	18	3594736	3598316	0.9118	M	3,581
1	PMAN	19	183124	188640	0.8702	E	5,517
1	PERE	21	62802442	62807520	0.9534	M	5,079
1	PERE	21	62831991	62838981	0.9939	C	6,991
2	PCAL	1	36351232	36364716	0.946	M	13,485
2	PCAL	1	36364712	36369922	0.9222	M	5,211
2	PCAL	1	36369952	36375783	0.8697	E	5,832
2	PCAL	1	36398334	36401849	0.9176	M	3,516
2	PCAL	1	36402017	36404879	0.9352	M	2,863
2	PCAL	1	36404882	36407057	0.8989	E	2,176
2	PCAL	1	36414251	36416948	0.9503	M	2,698
2	PCAL	1	36418178	36420429	0.8773	E	2,252
2	PCAL	1	36420452	36424666	0.898	E	4,215
2	PERE	2	27343	30059	0.9504	M	2,717
2	PERE	2	32030	39434	0.9651	M	7,405
2	PERE	2	41360	43735	0.9288	M	2,376
2	PERE	2	46394	49092	0.9552	M	2,699
2	PERE	2	49104	53222	0.9245	M	4,119
2	PMAN	3	28296051	28302793	0.8791	E	6,743
2	PERE	4	72784644	72792064	0.9373	M	7,421
2	PERE	4	72792652	72799584	0.9413	M	6,933
2	PERE	6	26776915	26780910	0.958	M	3,996
2	PERE	6	26789110	26791137	0.9346	M	2,028
2	PERE	6	26793811	26803338	0.9611	M	9,528
2	PERE	6	26815426	26821309	0.9493	M	5,884
2	PERE	12	78363676	78365819	0.8896	E	2,144
2	PERE	12	78365821	78368982	0.8794	E	3,162
2	PCAL	17	61429946	61441174	0.9362	M	11,229

2	PCAL	17	61452510	61455680	0.9572	M	3,171
2	PCAL	17	61515361	61525802	0.9616	M	10,442
2	PCAL	17	61573457	61575828	0.9369	M	2,372
2	PCAL	17	61578513	61585414	0.9623	M	6,902
2	PCAL	17	61585432	61588097	0.8735	E	2,666
2	PCAL	17	61607751	61611954	0.924	M	4,204
2	PCAL	18	45646670	45654659	0.9685	M	7,990
2	PCAL	18	45654818	45659908	0.9415	M	5,091
2	PCAL	18	45664026	45669847	0.9629	M	5,822
2	PCAL	19	77863624	77874984	0.9477	M	11,361
2	PCAL	20	68176462	68179690	1	C	3,229
2	PMAN	20	30211528	30222711	0.8893	E	11,184
2	PCAL	21	68378768	68385513	0.9135	M	6,746
2	PCAL	21	68393155	68400888	0.9655	M	7,734
2	PMAN	22	6958967	6962596	0.9185	M	3,630
2	PMAN	22	6963035	6965347	0.9111	M	2,313
2	PCAL	23	15618129	15622026	0.8699	E	3,898
2	PCAL	23	15622083	15624649	0.8824	E	2,567
2	PCAL	23	15625295	15627418	0.9235	M	2,124
2	PCAL	23	15636300	15642620	0.8992	E	6,321
2	PCAL	23	15648693	15657429	0.9369	M	8,737
2	PLEU	X	4387348	4391538	0.9254	M	4,191
3	PCAL	1	36428611	36432520	0.8887	E	3,910
3	PCAL	1	36434246	36445748	0.886	E	11,503
3	PCAL	1	36918392	36921154	0.8812	E	2,763
3	PCAL	1	36932859	36935570	0.877	E	2,712
3	PCAL	2	131040614	131048452	0.8877	E	7,839
3	PCAL	2	131048553	131050967	0.8786	E	2,415
3	PCAL	3	160237787	160244137	0.9275	M	6,351
3	PCAL	3	160244142	160246664	0.8886	E	2,523
3	PCAL	3	160246657	160251279	0.9185	M	4,623
3	PCAL	3	160258616	160262942	0.9203	M	4,327
3	PCAL	3	160274377	160280158	0.9286	M	5,782
3	PERE	6	26754753	26776696	0.9124	M	21,944
3	PCAL	7	116872720	116876295	0.8932	E	3,576
3	PCAL	7	116899826	116905148	0.8965	E	5,323
3	PCAL	12	17020	28795	0.9187	M	11,776
3	PCAL	12	30457	34041	0.8805	E	3,585
3	PCAL	17	61473551	61476034	0.9074	M	2,484
3	PCAL	17	61478588	61500930	0.8956	E	22,343
3	PCAL	17	61508569	61510823	0.9124	M	2,255

3	PCAL	17	61510826	61515096	0.9041	M	4,271
3	PCAL	17	61555714	61567705	0.9064	M	11,992
3	PCAL	17	61588104	61594666	0.9266	M	6,563
3	PCAL	17	61602480	61606766	0.8968	E	4,287
3	PCAL	17	61636465	61640924	0.9291	M	4,460
3	PERE	17	7994202	7996372	0.98	M	2,171
3	PERE	17	8004067	8007924	0.9597	M	3,858
3	PERE	17	51396259	51407879	0.9704	M	11,621
3	PERE	17	51409053	51413110	0.9962	C	4,058
3	PCAL	19	77844149	77848030	0.9364	M	3,882
3	PCAL	19	77848028	77852322	0.9281	M	4,295
3	PCAL	19	77875339	77880006	0.9099	M	4,668
3	PCAL	19	77881302	77883557	0.9124	M	2,256
3	PCAL	19	77885323	77898096	0.9319	M	12,774
3	PCAL	20	68179933	68190660	0.9226	M	10,728
3	PCAL	21	68418341	68425778	0.9044	M	7,438
3	PERE	21	63174566	63180030	0.9034	M	5,465
3	PMAN	22	10459254	10462846	0.8774	E	3,593
3	PCAL	23	15678668	15682274	0.8833	E	3,607
3	PCAL	23	15682272	15692779	0.8861	E	10,508
3	PCAL	X	1392	7860	0.8868	E	6,469
4	PERE	1	18919872	18926404	0.8873	E	6,533
4	PERE	1	18937253	18940027	0.8838	E	2,775
4	PERE	1	18940372	18945965	0.8709	E	5,594
4	PERE	1	18962427	18968559	0.8719	E	6,133
4	PERE	1	18974219	18978301	0.8878	E	4,083
4	PCAL	2	131056406	131059862	0.8718	E	3,457
4	PCAL	2	131059862	131077165	0.882	E	17,304
4	PCAL	2	131085752	131098344	0.8797	E	12,593
4	PCAL	2	131115894	131127484	0.8865	E	11,591
4	PCAL	2	131133984	131139738	0.8988	E	5,755
4	PCAL	2	131143224	131150585	0.9026	M	7,362
4	PERE	2	60475	69391	0.8717	E	8,917
4	PERE	2	87810	97562	0.9384	M	9,753
4	PERE	2	45596657	45600721	0.8863	E	4,065
4	PERE	2	45616013	45619051	0.9036	M	3,039
4	PERE	2	158246481	158249284	0.8846	E	2,804
4	PCAL	3	160270657	160273564	0.8693	E	2,908
4	PERE	6	26749772	26753579	0.892	E	3,808
4	PERE	6	26823764	26827337	0.9279	M	3,574
4	PCAL	7	116890652	116896796	0.8814	E	6,145

4	PCAL	7	116896912	116899688	0.9	M	2,777
4	PCAL	8	85006204	85009019	0.9113	M	2,816
4	PCAL	9	36311	40134	0.8798	E	3,824
4	PCAL	9	85874	90035	0.8884	E	4,162
4	PCAL	10	67086	73657	0.9351	M	6,572
4	PCAL	12	42479	46884	0.8754	E	4,406
4	PCAL	12	73273	77129	0.8882	E	3,857
4	PCAL	13	2709	5075	0.8751	E	2,367
4	PCAL	13	5086	11443	0.8979	E	6,358
4	PCAL	13	20967	27520	0.9036	M	6,554
4	PCAL	13	37451	42177	0.8997	E	4,727
4	PCAL	13	65007	67324	0.8716	E	2,318
4	PCAL	13	67533	74107	0.8918	E	6,575
4	PCAL	13	83232	88309	0.8712	E	5,078
4	PCAL	13	88320	93738	0.9585	M	5,419
4	PERE	13	5280120	5286168	0.8933	E	6,049
4	PERE	13	5286340	5326163	0.9132	M	39,824
4	PCAL	14	53079	66788	0.8761	E	13,710
4	PCAL	14	87065	93070	0.878	E	6,006
4	PCAL	14	93082	98951	0.9517	M	5,870
4	PCAL	14	112948	136882	0.8954	E	23,935
4	PCAL	14	150482	159250	0.9973	C	8,769
4	PCAL	15	427	7785	0.901	M	7,359
4	PCAL	15	7785	12252	0.8871	E	4,468
4	PCAL	16	48243625	48245639	0.9134	M	2,015
4	PCAL	16	48245795	48252418	0.9401	M	6,624
4	PCAL	16	48259470	48266686	0.9333	M	7,217
4	PERE	16	19399899	19415623	0.9049	M	15,725
4	PCAL	17	61547115	61549159	0.9613	M	2,045
4	PCAL	17	61628358	61631496	0.926	M	3,139
4	PMAN	19	143723	147787	0.9166	M	4,065
4	PMAN	19	147802	149860	0.9004	M	2,059
4	PMAN	19	149886	152342	0.9162	M	2,457
4	PMAN	19	153474	155742	0.9044	M	2,269
4	PMAN	19	155736	159056	0.9068	M	3,321
4	PLEU	20	40035483	40049394	0.8937	E	13,912
4	PERE	21	63180125	63185842	0.91	M	5,718
4	PCAL	22	43704648	43720837	0.9348	M	16,190
4	PMAN	22	6974820	6979651	0.8933	E	4,832
4	PCAL	23	15659063	15661258	0.8698	E	2,196
5	PCAL	1	37455481	37458792	0.9307	M	3,312

5	PERE	1	2094327	2102468	0.8828	E	8,142
5	PERE	1	2169949	2173777	0.9009	M	3,829
5	PERE	1	2273318	2278486	0.8722	E	5,169
5	PERE	1	81209745	81223131	0.8893	E	13,387
5	PLEU	1	152363548	152366300	0.8854	E	2,753
5	PERE	12	78344110	78350723	0.9097	M	6,614
5	PERE	14	78631334	78634548	0.9321	M	3,215
5	PCAL	17	61407387	61410081	0.9988	C	2,695
5	PCAL	17	61417883	61421133	0.8914	E	3,251
5	PERE	23	37313250	37320232	0.9084	M	6,983
5	PLEU	8b	29911725	29914532	0.8912	E	2,808
6	PERE	1	2230508	2234871	0.9292	M	4,364
6	PCAL	2	131166283	131172979	0.9533	M	6,697
6	PCAL	2	131172985	131178904	0.8819	E	5,920
6	PCAL	8	85020057	85023642	0.91	M	3,586
6	PCAL	14	170937	174186	0.8942	E	3,250
6	PERE	21	62797913	62800608	0.9968	C	2,696
6	PERE	21	62827336	62830031	0.9728	M	2,696
6	PCAL	23	10244616	10247832	0.8987	E	3,217
6	PERE	23	22236625	22239706	0.9262	M	3,082
6	PERE	23	22247797	22252816	0.8756	E	5,020
6	PERE	23	22401568	22410826	0.9119	M	9,259
7	PCAL	16	48275139	48279054	0.9988	C	3,916
7	PLEU	8a	106265	108486	0.8729	E	2,222
8	PCAL	7	116918216	116921728	0.8728	E	3,513
8	PMAN	7	10686051	10688242	0.8693	E	2,192
8	PERE	10	103969214	103971366	0.8716	E	2,153
8	PLEU	12	83057656	83060416	0.906	M	2,761
8	PCAL	17	61402799	61405557	0.9246	M	2,759
8	PCAL	18	45443752	45448572	0.9939	C	4,821
8	PERE	X	16114181	16116679	0.9444	M	2,499
9	PERE	21	62785523	62790369	0.9939	C	4,847
9	PERE	21	62812213	62815370	0.9451	M	3,158
10	PLEU	1	150397527	150408431	0.9359	M	10,905
10	PLEU	1	150418438	150436405	0.9719	M	17,968
10	PLEU	1	152376774	152386927	0.894	E	10,154
10	PMAN	1	30946355	30954452	0.9983	C	8,098
11	PMAN	1	30912716	30915120	0.9181	M	2,405
11	PMAN	1	30915161	30921226	0.9499	M	6,066
11	PMAN	1	30921224	30924012	0.9486	M	2,789
11	PMAN	1	30934094	30944410	0.9468	M	10,317

11	PMAN	1	74282880	74286322	0.9111	M	3,443
11	PMAN	2	42015998	42021275	0.9283	M	5,278
11	PMAN	2	42029764	42032256	0.9004	M	2,493
11	PMAN	2	42049829	42063045	0.908	M	13,217
11	PMAN	2	42091407	42093672	0.8955	E	2,266
11	PMAN	2	42095730	42101305	0.9307	M	5,576
11	PMAN	2	42103790	42107709	0.9268	M	3,920
11	PMAN	2	42107708	42111424	0.9172	M	3,717
11	PMAN	2	42118168	42121077	0.9568	M	2,910
11	PMAN	2	42125888	42131392	1	C	5,505
11	PMAN	2	42134414	42137716	0.9499	M	3,303
11	PMAN	2	42179784	42193615	0.8922	E	13,832
11	PMAN	3	41237289	41242493	0.9526	M	5,205
11	PMAN	3	41321136	41333144	0.8744	E	12,009
11	PMAN	5	32748031	32752361	0.8956	E	4,331
11	PMAN	5	32775115	32782339	0.9041	M	7,225
11	PMAN	5	32791375	32794645	0.8889	E	3,271
11	PMAN	5	32794643	32800758	0.9091	M	6,116
11	PMAN	5	32825720	32829480	0.891	E	3,761
11	PMAN	5	32844206	32850509	0.9204	M	6,304
11	PLEU	7	87345661	87351417	0.8997	E	5,757
11	PMAN	7	10706941	10711285	0.8796	E	4,345
11	PMAN	7	10711351	10714239	0.8903	E	2,889
11	PMAN	7	10722328	10738511	0.885	E	16,184
11	PMAN	7	10749705	10753909	0.8982	E	4,205
11	PMAN	7	10754064	10756489	0.9167	M	2,426
11	PMAN	7	10756645	10759465	0.8814	E	2,821
11	PMAN	7	10759462	10763031	0.9014	M	3,570
11	PMAN	7	10763754	10768252	0.9339	M	4,499
11	PMAN	7	10768257	10775302	0.9305	M	7,046
11	PMAN	7	10775605	10778777	0.9133	M	3,173
11	PMAN	7	10779415	10789404	0.9151	M	9,990
11	PMAN	7	10789404	10794489	0.9276	M	5,086
11	PMAN	7	10799156	10804601	0.9333	M	5,446
11	PMAN	9	23546719	23548756	0.9443	M	2,038
11	PMAN	9	23549472	23554366	0.9182	M	4,895
11	PMAN	9	23554357	23565343	0.9554	M	10,987
11	PMAN	9	23565921	23568195	0.9468	M	2,275
11	PMAN	9	23568176	23570824	0.9665	M	2,649
11	PMAN	9	23576564	23580692	0.9134	M	4,129
11	PMAN	9	23583943	23592398	0.9139	M	8,456

11	PMAN	9	30137723	30143358	0.9202	M	5,636
11	PMAN	9	30143354	30159738	0.8906	E	16,385
11	PMAN	11	11261845	11276091	0.9129	M	14,247
11	PMAN	11	11288882	11290957	0.9133	M	2,076
11	PMAN	11	11291127	11298703	0.905	M	7,577
11	PMAN	11	11299395	11302409	0.9283	M	3,015
11	PMAN	11	11302408	11306676	0.9273	M	4,269
11	PMAN	11	11310205	11318048	0.9273	M	7,844
11	PMAN	11	11320141	11329414	0.9392	M	9,274
11	PMAN	11	11329412	11332166	0.9122	M	2,755
11	PMAN	11	11332827	11335835	0.9248	M	3,009
11	PMAN	11	11335833	11338587	0.9122	M	2,755
11	PMAN	11	11338637	11341223	0.915	M	2,587
11	PLEU	12	83116048	83119451	0.9019	M	3,404
11	PMAN	14	7325814	7330126	0.9136	M	4,313
11	PMAN	14	40293530	40298437	0.9439	M	4,908
11	PMAN	14	40302285	40305398	0.9143	M	3,114
11	PMAN	14	40320455	40324610	0.9208	M	4,156
11	PMAN	14	40328131	40330338	0.9046	M	2,208
11	PMAN	14	40330339	40335239	0.9021	M	4,901
11	PMAN	14	40343568	40349377	0.9271	M	5,810
11	PLEU	19	53977317	53982629	0.8738	E	5,313
11	PMAN	20	30254009	30256238	0.9179	M	2,230
11	PMAN	20	30261367	30263820	0.9101	M	2,454
11	PMAN	21	57298381	57301015	0.9113	M	2,635
11	PMAN	21	57310208	57315857	0.9089	M	5,650
11	PMAN	21	57321239	57324034	0.8977	E	2,796
11	PMAN	21	57330119	57333921	0.9237	M	3,803
11	PMAN	22	29116	37416	0.8901	E	8,301
11	PMAN	22	76757	82109	0.9097	M	5,353
11	PMAN	22	10355266	10368040	0.9286	M	12,775
11	PMAN	22	10368261	10373435	0.9025	M	5,175
11	PMAN	22	10373433	10381428	0.8983	E	7,996
11	PMAN	22	10381425	10383521	0.877	E	2,097
11	PMAN	22	10383518	10385609	0.9101	M	2,092
11	PMAN	22	10385852	10390344	0.9061	M	4,493
11	PMAN	22	10396311	10402082	0.8779	E	5,772
11	PMAN	22	10409122	10411233	0.9088	M	2,112
11	PMAN	22	10441597	10447045	0.8869	E	5,449
11	PMAN	22	10448212	10451494	0.9047	M	3,283
11	PMAN	22	10451493	10455612	0.9047	M	4,120

11	PMAN	22	10455920	10459048	0.9003	M	3,129
11	PMAN	X	133180330	133183284	0.9132	M	2,955
12	PMAN	1	32980188	32987177	0.9263	M	6,990
12	PMAN	1	32989784	32996975	0.9282	M	7,192
12	PMAN	1	32997647	33000336	0.9227	M	2,690
12	PMAN	1	33008531	33015485	0.9309	M	6,955
12	PMAN	1	33016187	33018491	0.9458	M	2,305
12	PMAN	2	42021288	42025744	0.9389	M	4,457
12	PMAN	4	53980	60788	0.9625	M	6,809
12	PMAN	4	63538	65718	0.932	M	2,181
12	PMAN	4	68573	73086	0.9528	M	4,514
12	PMAN	4	96778	105194	0.931	M	8,417
12	PMAN	4	110121	128398	1	M	18,278
12	PMAN	4	128404	136887	0.9918	M	8,484
12	PMAN	4	141894	144987	0.9858	M	3,094
12	PMAN	4	145461	152975	0.9906	M	7,515
12	PMAN	4	152985	168626	0.8949	E	15,642
12	PMAN	6	21351929	21378391	0.9611	M	26,463
12	PMAN	6	21380331	21383520	0.9368	M	3,190
12	PMAN	6	21392710	21397349	0.9416	M	4,640
12	PMAN	6	21399618	21403975	0.9605	M	4,358
12	PMAN	6	21403975	21412725	0.8716	E	8,751
12	PMAN	9	58330	63937	0.9167	M	5,608
12	PMAN	13	5394759	5398467	1	C	3,709
12	PMAN	17	204	7195	0.9509	M	6,992
12	PMAN	17	17620	20991	1	M	3,372
12	PMAN	17	21407	25142	0.974	M	3,736
12	PMAN	18	161703	163916	0.9578	M	2,214
12	PMAN	19	47587	54029	0.9631	M	6,443
12	PMAN	19	86978	89421	0.9529	M	2,444
12	PMAN	19	91746	97985	0.9789	M	6,240
12	PMAN	19	100508	103567	0.924	M	3,060
12	PMAN	19	105002	110219	0.9717	M	5,218
12	PMAN	19	122795	131218	0.8738	E	8,424
12	PMAN	22	4593	13148	0.9271	M	8,556
12	PMAN	22	6835561	6843988	1	M	8,428
12	PMAN	22	6844335	6847178	0.922	M	2,844
12	PMAN	22	10337713	10350588	0.9904	M	12,876
12	PMAN	22	10350588	10354793	0.9758	M	4,206
12	PMAN	23	30282139	30288967	0.9828	M	6,829
12	PMAN	23	30288978	30292529	0.963	M	3,552

12	PMAN	23	30320614	30328482	0.8718	E	7,869
12	PLEU	X	4353863	4360999	0.9534	M	7,137
13	PMAN	3	29567993	29576190	0.9624	M	8,198
13	PMAN	3	29611000	29613001	0.9422	M	2,002
13	PMAN	3	29613052	29620760	0.9545	M	7,709
13	PMAN	3	29621582	29628804	0.9451	M	7,223
13	PMAN	3	29658700	29671030	0.9103	M	12,331
13	PMAN	4	168646	175593	0.9664	M	6,948
13	PMAN	6	21412741	21415456	0.9444	M	2,716
13	PLEU	13	34	4055	0.9666	M	4,022
13	PMAN	13	5408630	5414781	0.9766	M	6,152
13	PMAN	13	5414945	5420369	0.9695	M	5,425
13	PMAN	13	5420369	5427641	0.973	M	7,273
13	PMAN	19	131934	143559	0.9623	M	11,626
13	PLEU	20	40010224	40017791	0.9578	M	7,568
13	PLEU	20	40022919	40035075	1	C	12,157
13	PLEU	22	11208215	11212297	0.8708	E	4,083
13	PLEU	22	12441732	12449329	0.8741	E	7,598
13	PMAN	22	6847176	6849501	0.9229	M	2,326
13	PMAN	22	6879355	6884659	0.8772	E	5,305
13	PMAN	22	6888669	6891295	0.8715	E	2,627
13	PLEU	8a	3	5464	0.9765	M	5,462
14	PMAN	2	42138103	42149757	0.8817	E	11,655
14	PMAN	2	42149769	42154307	0.8825	E	4,539
14	PMAN	2	42160059	42178836	0.8849	E	18,778
14	PMAN	3	28273920	28282536	0.9557	M	8,617
14	PMAN	3	41279240	41290341	0.9259	M	11,102
14	PMAN	3	41298685	41314244	0.9341	M	15,560
14	PMAN	3	41350347	41357663	0.8975	E	7,317
14	PMAN	5	32694927	32746417	0.9266	M	51,491
14	PLEU	7	66510338	66517943	0.9726	M	7,606
14	PMAN	7	10688516	10694199	0.9365	M	5,684
14	PLEU	9	85063855	85068511	0.9983	C	4,657
14	PMAN	9	23462024	23464503	0.9738	M	2,480
14	PMAN	9	23464505	23466709	0.9541	M	2,205
14	PMAN	9	23467275	23470772	0.9145	M	3,498
14	PMAN	9	23471151	23479944	0.8938	E	8,794
14	PMAN	9	30160784	30163790	0.9017	M	3,007
14	PMAN	9	30167305	30173164	0.8971	E	5,860
14	PMAN	11	11249220	11251937	0.969	M	2,718
14	PMAN	11	11255517	11261648	0.9244	M	6,132

14	PMAN	14	40280884	40283454	0.9279	M	2,571
14	PMAN	14	40286611	40288674	0.883	E	2,064
14	PLEU	15	89496061	89502777	0.9688	M	6,717
14	PLEU	18	44531181	44534915	0.9305	M	3,735
14	PMAN	20	30225421	30235020	0.8973	E	9,600
14	PMAN	22	45448	53884	0.9161	M	8,437
14	PMAN	22	53882	57847	0.9022	M	3,966
14	PMAN	22	68897	71325	0.8795	E	2,429
14	PMAN	22	82296	85424	0.9103	M	3,129
15	PMAN	22	6929690	6932113	0.8845	E	2,424
15	PLEU	23	30509004	30512257	0.913	M	3,254
15	PMAN	23	30349322	30353194	1	C	3,873
16	PMAN	1	33018490	33020660	0.879	E	2,171
16	PMAN	1	33028065	33030516	0.9406	M	2,452
16	PMAN	1	33038381	33048103	0.9448	M	9,723
16	PMAN	1	33066302	33071475	0.9988	C	5,174
16	PMAN	1	33079023	33081658	0.9421	M	2,636
16	PMAN	1	33081658	33098184	0.9091	M	16,527

Supplementary Table 12 – Clustering of LTRs detected in the GM03417 nanopore sequencing dataset.

Number	Arrays	min size (bp)	mean size (bp)	max size (bp)	Total nt (bp)	Matching Repeat
1	4,849	155	244	1,905	75,022,553	ALR
2	467	63	68	133	3,334,859	BSR
3	418	50	73	214	8,935,785	HSAT2
4	271	75	114	377	6,009,809	HSAT1A
5	49	90	105	188	438,065	CER
6	35	51	94	220	423,822	HSAT2
7	25	1,293	1,520	1,802	118,799	HSAT2
8	16	764	789	808	47,478	H_LTR_1
9	11	623	922	1,196	35,305	H_LTR_2
10	11	140	143	146	24,386	ACRO1
11	10	81	84	86	30,867	H_LTR_3
12	9	1,835	1,844	1,860	99,408	ALR
13	9	1,503	1,532	1,549	79,747	H_LTR_4
14	8	1,122	1,345	1,402	123,158	SST1
15	7	440	454	464	87,086	HSAT2
16	6	1,683	1,745	1,798	57,223	teucerv2_3edge
17	5	1,836	1,854	1,867	77,271	MER5A1r
18	4	87	88	89	13,153	MER20
19	4	63	63	64	16,653	H_LTR_5

20	4	62	63	63	10,248	L1MD_orf2

Supplementary Table 13 – Clustering of LTRs detected in the NA12878 nanopore sequencing dataset.

Number	Arrays	min size (bp)	mean size (bp)	max size (bp)	Total nt (bp)	Matching Repeat
1	15,812	151	245	2,011	277,734,313	ALR
2	2,165	62	69	335	14,335,309	BSR
3	1,678	50	73	218	32,350,555	HSAT2
4	1,623	73	123	1,474	37,280,688	HSAT1A
5	830	50	577	1,992	6,532,774	(GAG)
6	382	50	105	471	5,609,214	(GGC)
7	341	50	83	271	6,256,219	(TCCAT)
8	212	50	129	708	1,498,520	(TGG)
9	211	51	108	428	2,083,030	CER
10	137	744	786	912	385,451	H_LTR_1
11	80	135	141	147	188,865	ACRO1
12	74	461	584	1,513	288,883	H_LTR_6
13	62	610	1,059	1,958	207,893	H_LTR_2
14	56	1,088	1,341	1,388	1,667,243	SST1
15	55	1,770	1,830	1,889	690,119	MER5A1r
16	44	408	440	468	199,904	HSAT2
17	43	1,470	1,507	1,573	489,848	H_LTR_4
18	40	686	1,078	1,929	94,737	L1 retrotransposon
19	40	1,201	1,400	1,775	168,325	HSAT2
20	35	59	62	64	185,237	H_LTR_5

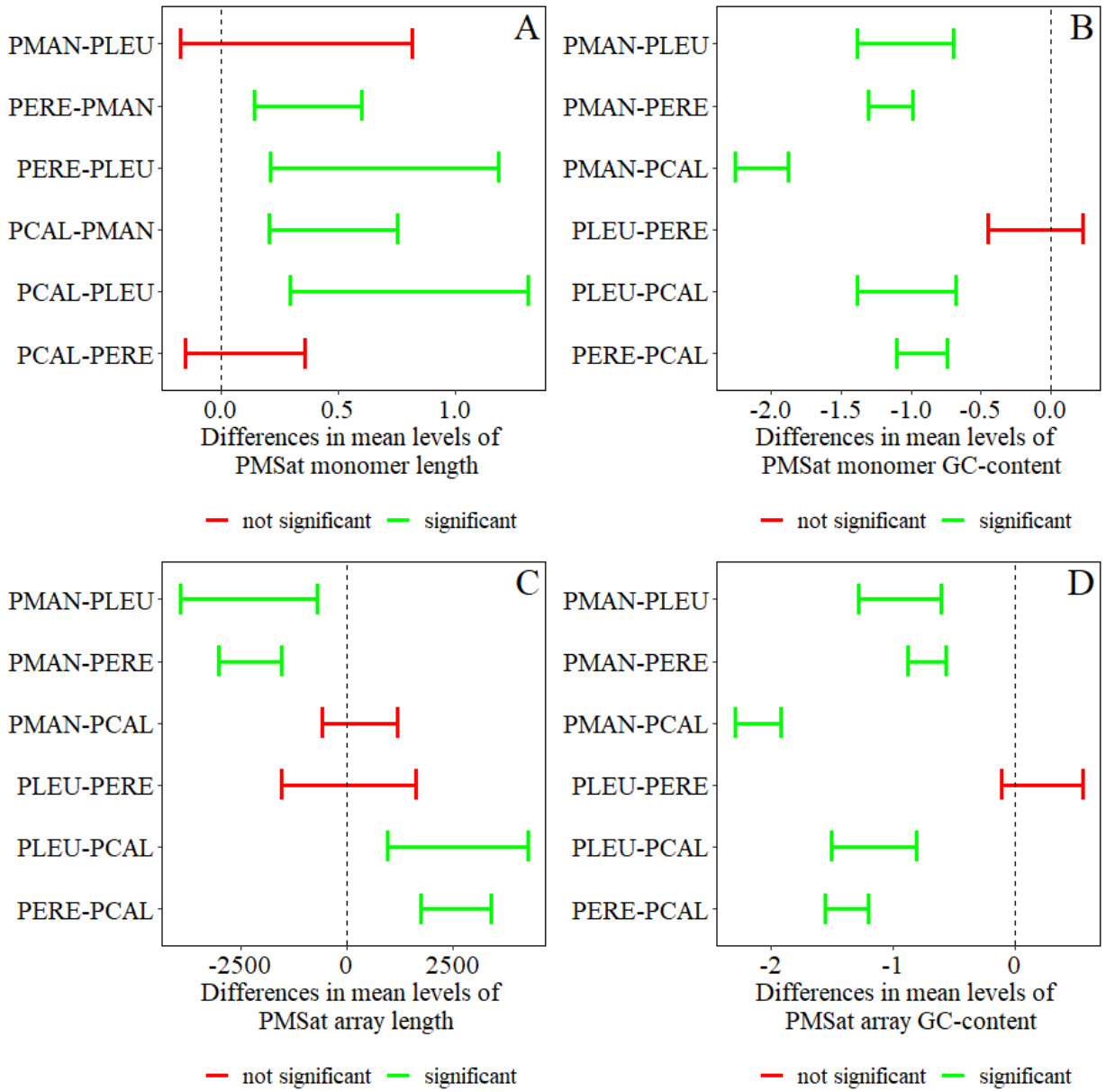
Supplementary Table 14 – Clustering of LTRs detected in the CHM13 nanopore sequencing dataset.

Number	Arrays	min size (bp)	mean size (bp)	max size (bp)	Total nt (bp)	Matching Repeat
1	207,000	149	202	2,027	5,251,475,232	ALR
2	31,999	60	69	605	322,852,114	BSR
3	28,925	50	81	579	1,095,920,854	HSAT2
4	4,634	72	115	581	109,577,492	HSAT1A
5	3,212	50	120	711	38,130,843	CER
6	3,205	60	74	152	10,722,036	SATR1
7	1,516	50	85	234	31,782,441	(GGAAT)
8	1,365	734	798	1,562	3,913,395	H_LTR_1
9	1,355	131	153	432	3,303,077	ACRO1
10	882	746	1,218	1,984	2,259,920	L1 retrotransposon
11	820	1,688	1,850	1,887	11,271,470	MER5A1r
12	790	79	91	334	2,074,995	H_LTR_3

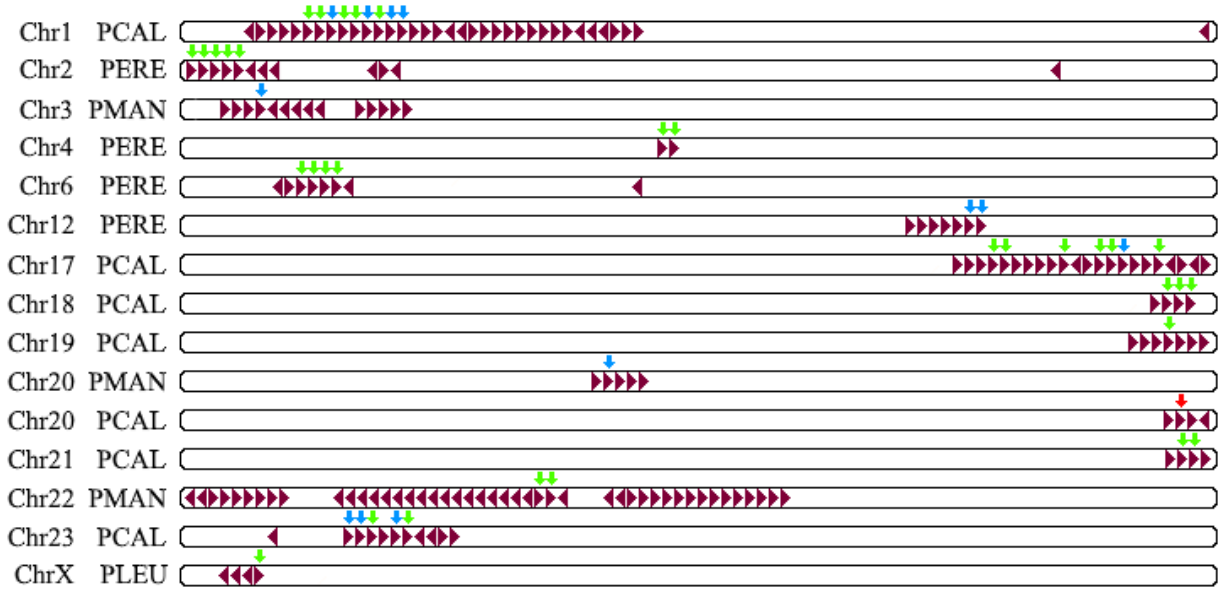
13	739	598	1,061	1,988	2,674,794	H_LTR_2
14	621	54	64	117	23,418,856	Walusat
15	601	1,457	1,524	1,636	6,656,412	H_LTR_4
16	577	1,086	1,351	1,648	15,816,952	SST1
17	461	182	258	1,414	1,282,517	GSATII
18	423	58	63	119	2,087,650	H_LTR_5
19	420	70	76	78	2,221,431	L1PA10_3end
20	381	67	84	279	2,796,755	HSAT4

Supplementary Table 15 – Sequencing run of 3'RACE-Seq. Quantification of number of bases and reads, with mean read length (Whole Genome Library Preparation, Illumina MiSeq plataform, NGS Sequencing service STAB VIDA).

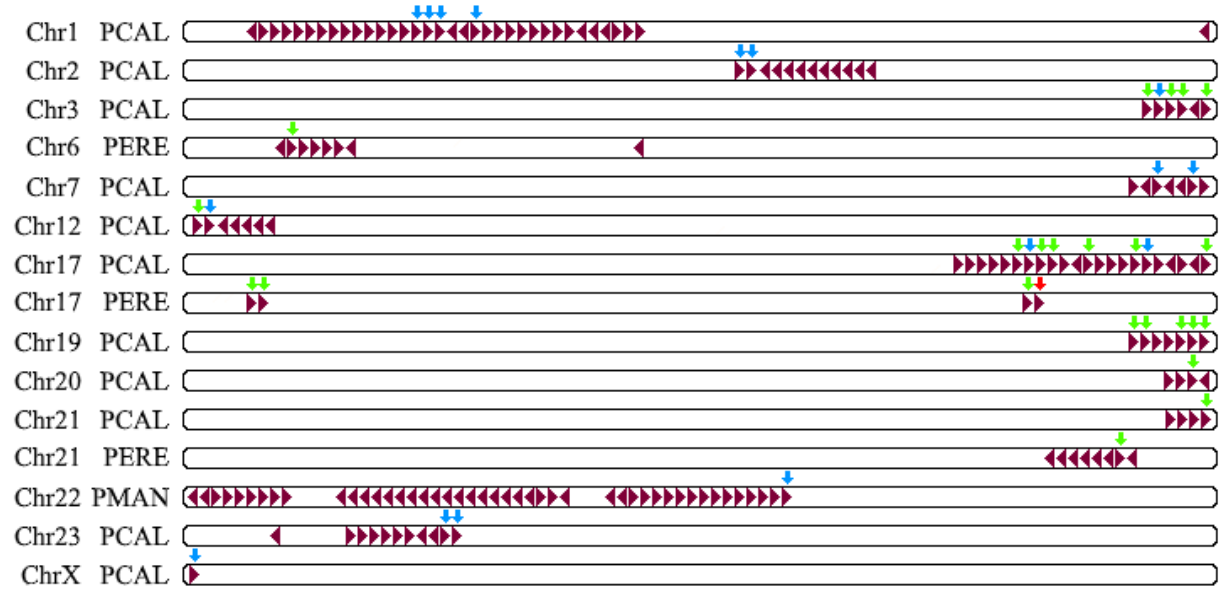
Number Of Bases (Raw Data) [Mbp]	Number Of Reads (Raw Data)	Mean Read Length [bp]
59.27	196,928	300



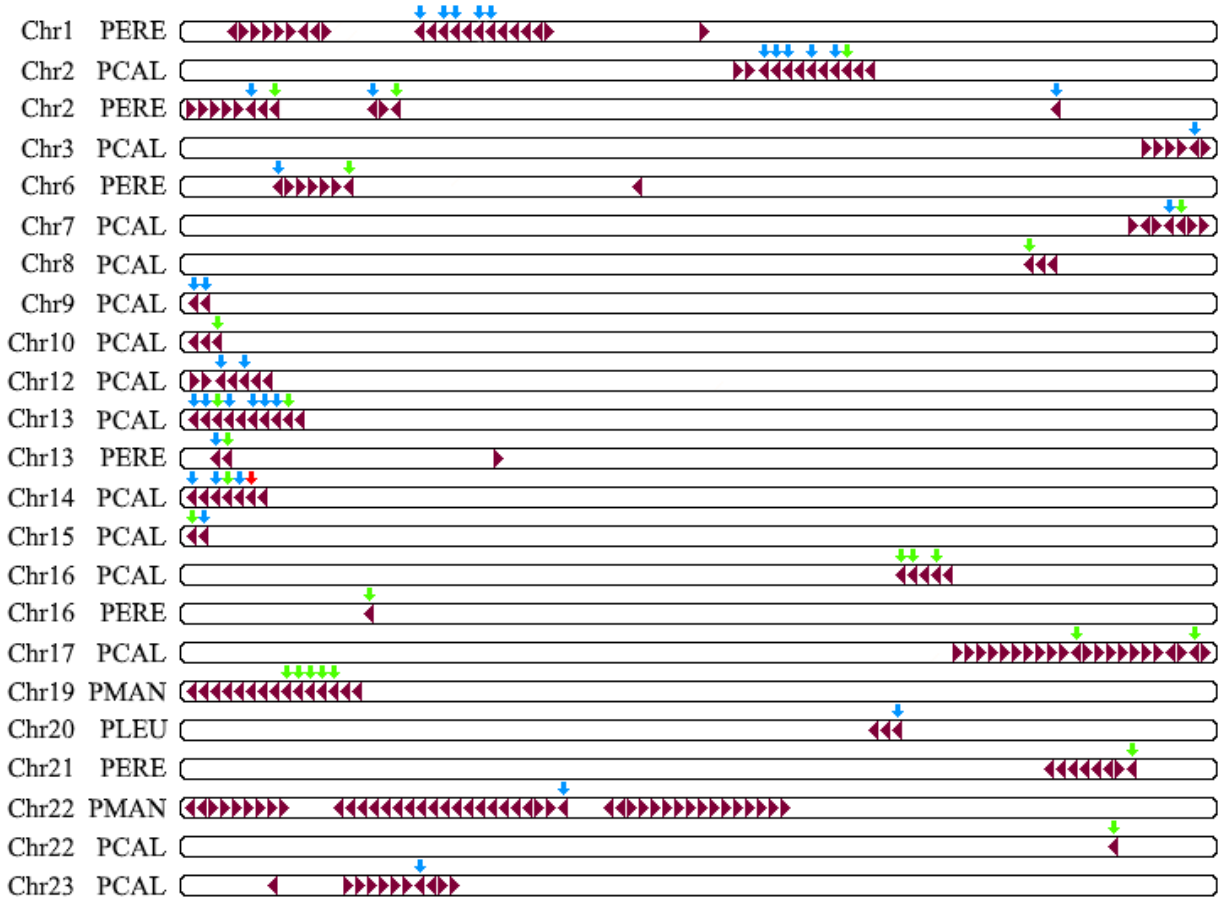
Supplementary Figure 1 – Tukey Post-Hoc test of differences in mean levels of PMSat monomer length (A), monomer GC-content (B), array length (C) and array GC-content (D) across the four Peromyscus species. The significant groupwise differences are anywhere the 95% confidence interval doesn't include zero ($p < 0.05$).



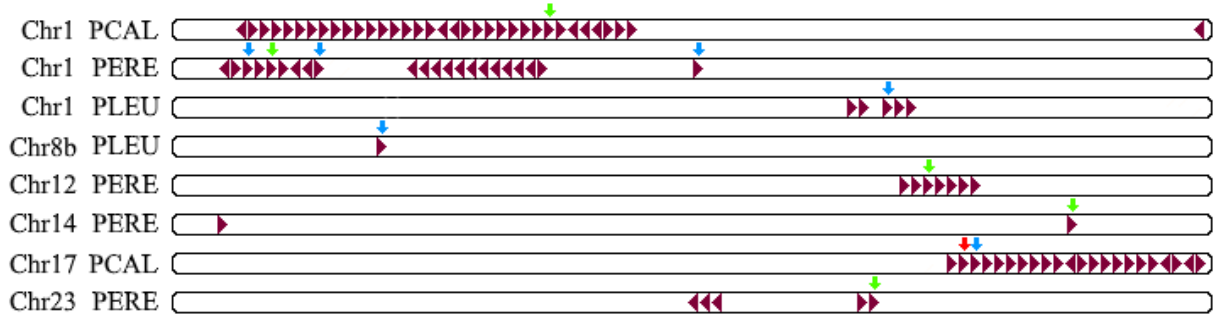
Supplementary Figure 2 - Orientation and relative localization of cluster group 2 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).



Supplementary Figure 3 – Orientation and relative localization of cluster group 3 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).

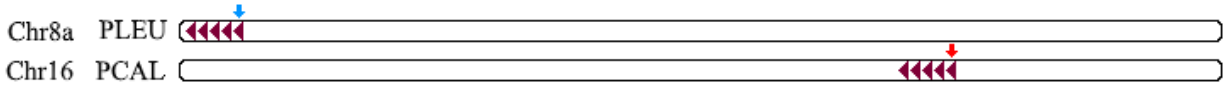


Supplementary Figure 4 – Orientation and relative localization of cluster group 4 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).

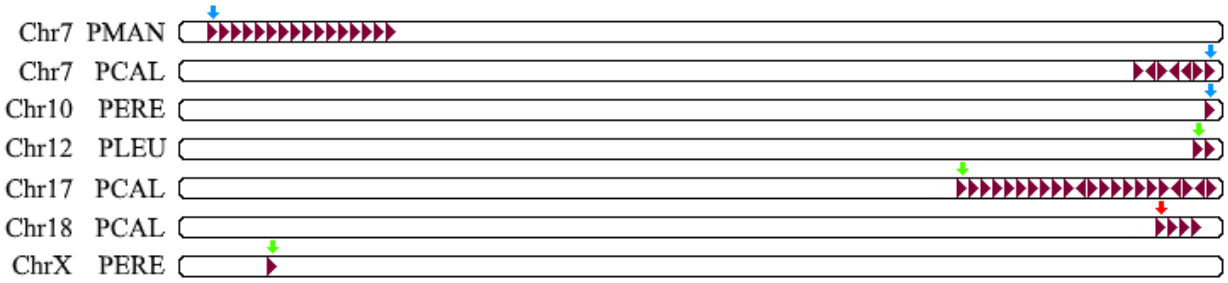


Supplementary Figure 5 – Orientation and relative localization of cluster group 5 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).

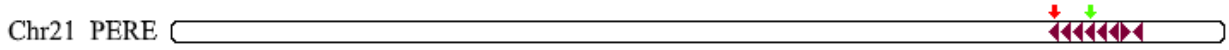
Supplementary Figure 6 – Orientation and relative localization of cluster group 6 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).



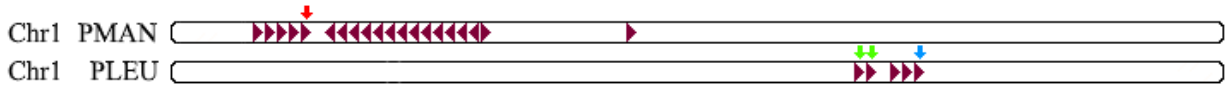
Supplementary Figure 7 – Orientation and relative localization of cluster group 7 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).



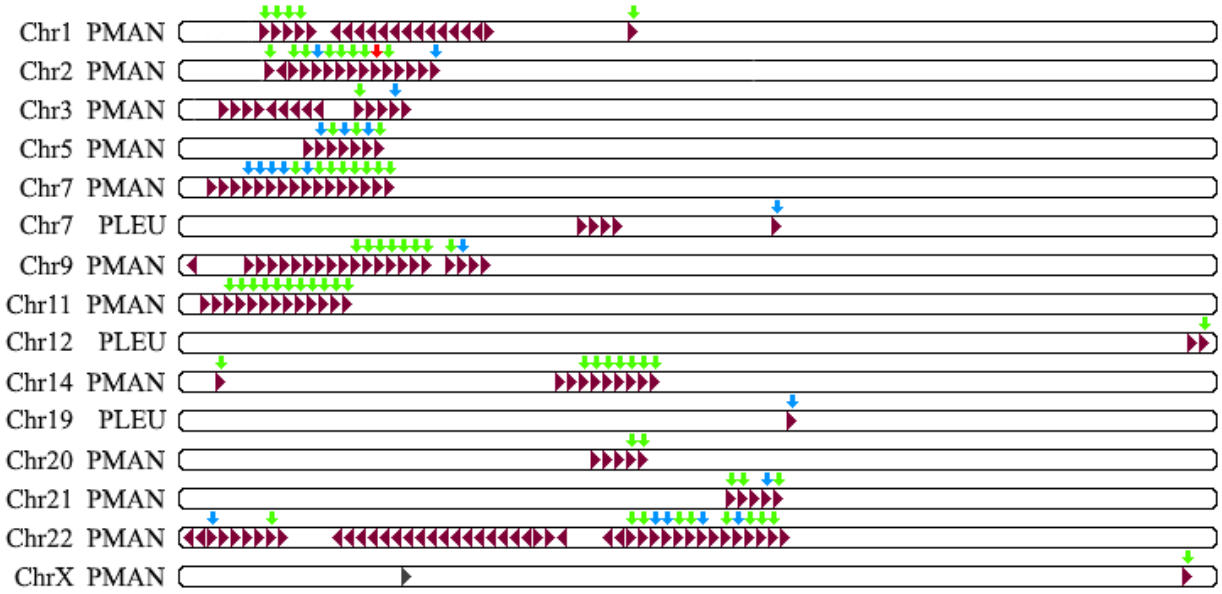
Supplementary Figure 8 – Orientation and relative localization of cluster group 8 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).



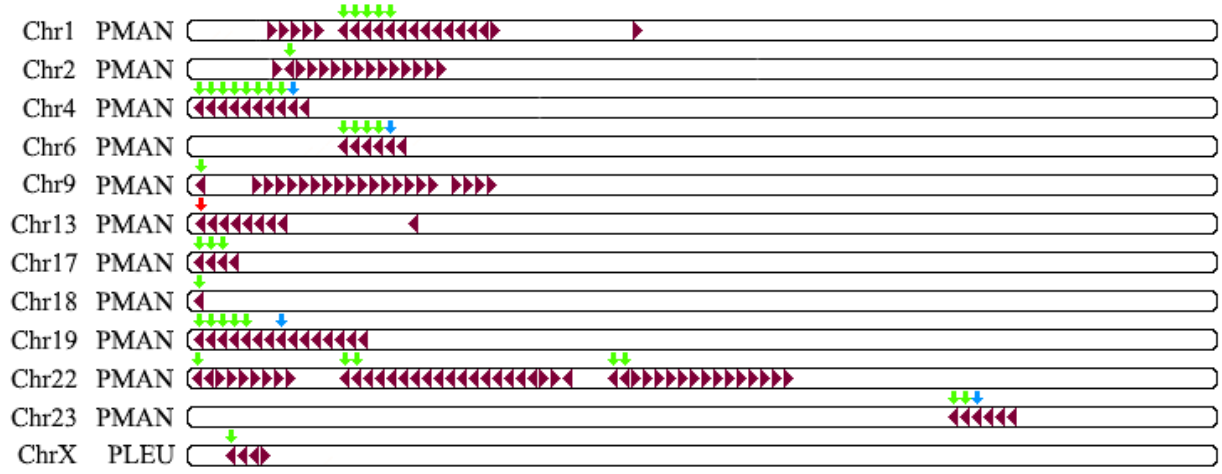
Supplementary Figure 9 – Orientation and relative localization of cluster group 9 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).



Supplementary Figure 10 – Orientation and relative localization of cluster group 10 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).

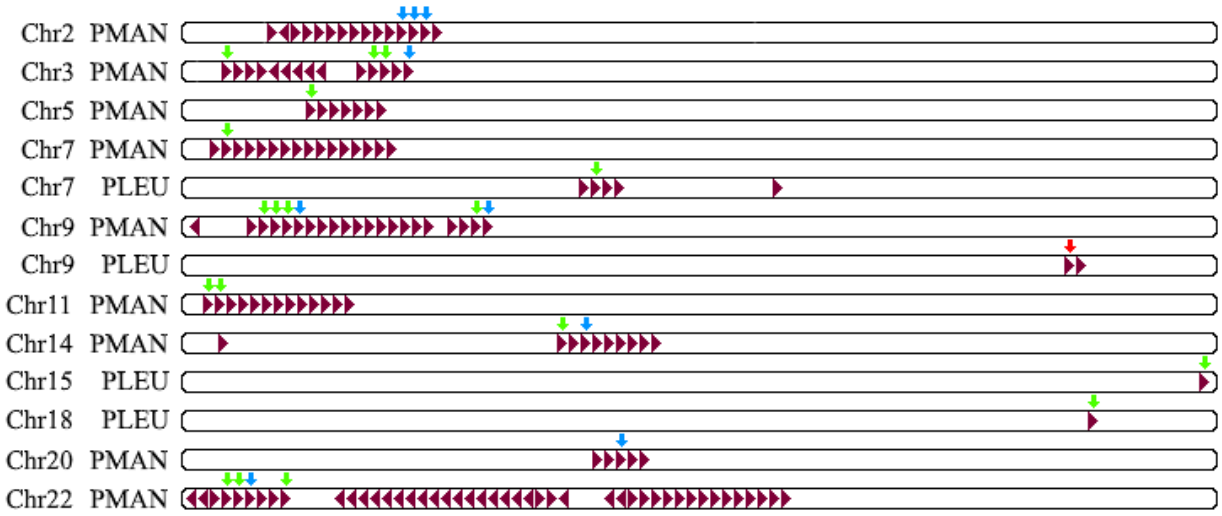


Supplementary Figure 11 – Orientation and relative localization of cluster group 11 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).



Supplementary Figure 12 – Orientation and relative localization of cluster group 12 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).

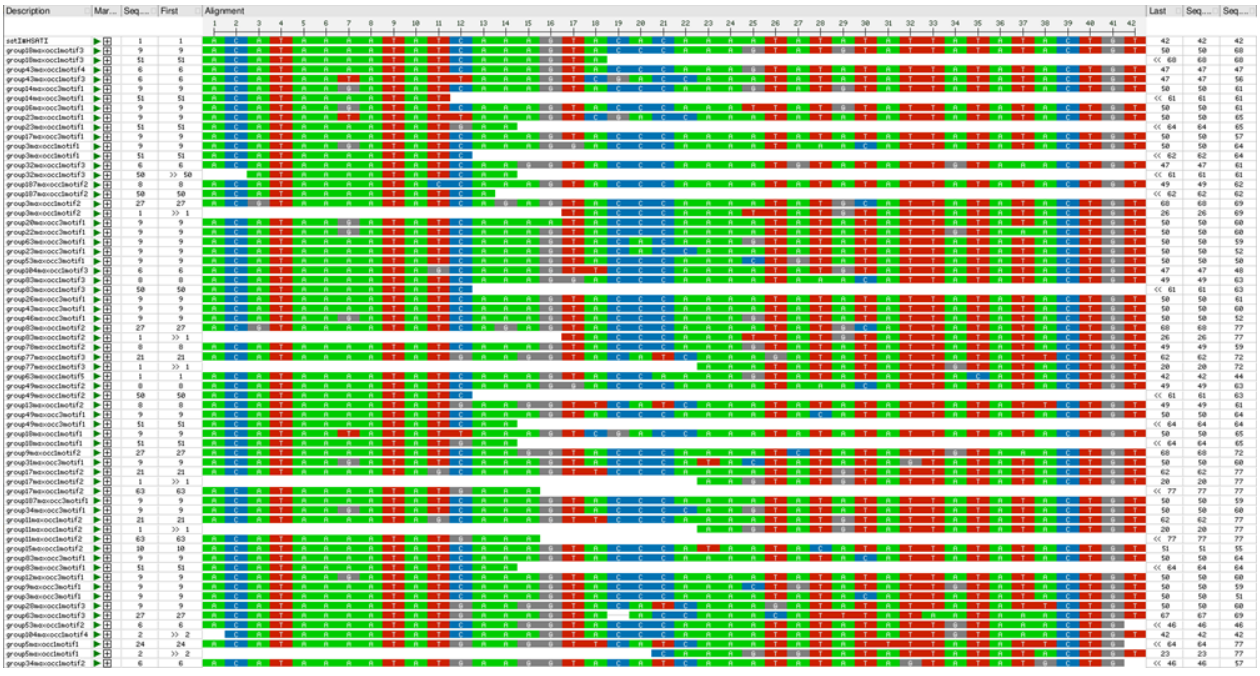
Supplementary Figure 13 – Orientation and relative localization of cluster group 13 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).



Supplementary Figure 14 – Orientation and relative localization of cluster group 14 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).

Supplementary Figure 15 – Orientation and relative localization of cluster group 15 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).

Supplementary Figure 16 – Orientation and relative localization of cluster group 16 of PMSat arrays based on identity score threshold of 0.9. These arrays are represented on the chromosomes matching each *Peromyscus* genome assembly. Each PMSat array was placed based on the relative position of the coordinates of where it was detected. The arrows represent the classification within the clustered group, based on their identity score from MeShClust v3.0 (Girgis, 2022) clustering: Red – Center (identity score ≈ 1); Green – Member (0.9 < identity score < 1); Blue – Extended Member (identity score < 0.9).



Supplementary Figure 17 - Comparison between the HSat1A monomer sequence and a selection of motifs sequences in each cluster group through a merged alignment.