DNA Research, 2017, 24(4), 377–385
doi: 10.1093/dnares/dsx010
Advance Access Publication Date: 17 July 2017
Full Paper

Full Paper

Identification and characterization of a subtelomeric satellite DNA in Callitrichini monkeys

Naiara Pereira Araújo[1,†], Leonardo Gomes de Lima[1,†], Guilherme Borges Dias[1], Gustavo Campos Silva Kuhn[1], Alan Lane de Melo[2], Yatiyo Yonenaga-Yassuda[3], Roscoe Stanyon[4], and Marta Svartman[1,*]

[1]Universidade Federal de Minas Gerais, Laboratório de Citogenômica Evolutiva, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Avenida Presidente Antônio Carlos, 6627 - Pampulha, 31270-901, Belo Horizonte, Brazil, [2]Universidade Federal de Minas Gerais, Laboratório de Taxonomia e Biologia de Invertebrados, Departamento de Parasitologia, Instituto de Ciências Biológicas, Belo Horizonte, Brazil, [3]Universidade de São Paulo, Laboratório de Citogenética de Vertebrados, Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, São Paulo, Brazil, and [4]University of Florence, Department of Biology, Florence, Italy

*To whom correspondence should be addressed. Tel. +5531 34092965. Email: svartmanm@icb.ufmg.br

[†]These authors contributed equally to this work.

Edited by Dr. Minoru Yoshida

Received 16 September 2016; Editorial decision 15 February 2017; Accepted 2 March 2017

Abstract

Repetitive DNAs are abundant fast-evolving components of eukaryotic genomes, which often possess important structural and functional roles. Despite their ubiquity, repetitive DNAs are poorly studied when compared with the genic fraction of genomes. Here, we took advantage of the availability of the sequenced genome of the common marmoset *Callithrix jacchus* to assess its satellite DNAs (satDNAs) and their distribution in Callitrichini. After clustering analysis of all reads and comparisons by similarity, we identified a satDNA composed by 171 bp motifs, named MarmoSAT, which composes 1.09% of the *C. jacchus* genome. Fluorescent *in situ* hybridization on chromosomes of species from the genera *Callithrix, Mico* and *Callimico* showed that MarmoSAT had a subtelomeric location. In addition to the common monomeric, we found that MarmoSAT was also organized in higher-order repeats of 338 bp in *Callimico goeldii*. Our phylogenetic analyses showed that MarmoSAT repeats from *C. jacchus* lack chromosome-specific features, suggesting exchange events among subterminal regions of non-homologous chromosomes. MarmoSAT is transcribed in several tissues of *C. jacchus*, with the highest transcription levels in spleen, thymus and heart. The transcription profile and subtelomeric location suggest that MarmoSAT may be involved in the regulation of telomerase and modulation of telomeric chromatin.

Key words: heterochromatin, repetitive DNA, Platyrrhini

1. Introduction

New World monkeys (NWM), infraorder Platyrrhini, represent a diverse group of neotropical primates, which are very important in medical, genomics, and evolutionary studies. Among them, the marmosets (tribe Callitrichini, family Cebidae) comprise a group of 21 species of *Callithrix*, endemic in the Atlantic forest, while *Cebuella* and *Mico* are found in the Amazon rainforest.[1] Recent molecular data support a strong relationship between *Callithrix* and *Cebuella* + *Mico*, which are, in turn, sister groups to *Callimico*.[2,3] In addition to the geographical separation, marmosets have different diploid numbers (2n). *Callithrix* species have 2n = 46, whereas both *Cebuella* and *Mico* species have 2n = 44.[4]

Several studies have shown that NWM genomes are rich in repetitive DNAs, most still uncharacterized. Among them, dispersed repetitive sequences, such as transposable elements (TEs), are major components of primate genomes.[5] For instance, the long interspersed element 1 (LINE-1) and the primate-specific Alu element, a short interspersed element (SINE), were considered as the largest contributors to the genome expansion in primates.[6]

Satellite DNA (satDNA) sequences, which are organized as long arrays of head-to-tail tandem repetitions, are also abundant components of primate genomes.[7] SatDNA monomers (repetitive units) form homogeneous arrays, usually enriched in regions of constitutive heterochromatin, and were hypothesized to be related to the maintenance of centromeric function (reviewed by Plohl *et al.*[8]). Multimers of the same satDNA motif may exhibit high similarity to each other, even when the individual monomers show considerable divergence. This organization is referred to as higher order repeats (HORs). Simian centromeres are mainly composed of α-satellite (AS) consisting of units of ∼170 bp in the infraorder Catarrhini and ∼340 or ∼540 bp in NWM.[9–11] It was assumed that the HOR blocks of AS was a unique attribute of hominoids. However, Sujiwattanarat *et al.*[12] have recently reported HORs in the NWM *Aotus azarae* and *C. jacchus*. They suggested that this type of organization probably occurs in the AS of a wide range of simians.

SatDNAs do not code proteins but their transcription has been reported in many organisms, including vertebrates, invertebrates and plants (reviewed by Pezer *et al.*[13]) where they were shown to participate in the formation of heterochromatin,[14] centromeres[15] and in gene regulation.[16] Chan *et al.*[17] showed that AS transcripts are essential for the localization of mitotic centromere proteins including CENP-C, determining the kinetochore structure on centromeric chromatin during mitosis in humans.

In spite of the great biomedical and evolutionary interest of primates, their satDNAs have only been studied in a few groups.[18–21] In the common marmoset *Callithrix jacchus*, the only satDNA reported to date is the AS DNA, located in the centromeres of all chromosomes.[9,11]

The recent availability of genomic data for *C. jacchus* provides an excellent new opportunity to study how satDNAs are organized and influence NWM genome evolution. In this study, we employed an integrated approach, using whole-genome sequence analysis and molecular cytogenetics, to get an in-depth insight into a new satDNA of *Callithrix*, termed MarmoSAT. Our intention was to better understand its evolution by analyzing an array of NWM genomes.

2. Materials and methods

2.1. Identification of satDNA in *Callithrix jacchus* and sequence analysis

Similarity-based clustering, repeat identification, and classification were performed using RepeatExplorer[22] with whole-genome shotgun (WGS) Illumina reads from a male *C. jacchus* (accession number: SRR957684). This pipeline involves an all-to-all comparison of Illumina reads by MEGABLAST and the grouping of similar reads in clusters that represent unique repetitive DNA families. A minimum of 55 nt overlap is required for clustering different reads. A total of 1540214 100 bp reads, representing ∼5% coverage of the *C. jacchus* genome were utilized in the analysis (Supplementary Fig. S1). All clusters with an abundance of at least 0.01% that of the top cluster were analysed in detail (Supplementary Table S1). As the reads utilized represent a random sample of the genome, the abundance of a given repetitive DNA family can be determined by the number of reads present in that specific cluster divided by the total number of reads utilized. The reads from each cluster are further aligned and partially assembled to produce contigs to be used in repeat consensus reconstruction and annotation. All contigs were compared with the mammalian repeat library in Repbase.[23,24] Whenever a significant number of reads from two distinct clusters match the similarity parameters, RepeatExplorer indicates these clusters as 'connected component', pointing to a potential relationship between the repeats.

C. jacchus MarmoSAT repeats were retrieved from this species sequenced genome by BLAST searches on the assembled genome (accession number: ACFV00000000.1) using as query a consensus sequence obtained from the RepeatExplorer analysis. Hits with e-values lower than 1×10^{-5} were considered significant. Furthermore, BLAST searches on *C. jacchus* WGS database present on NCBI were used to retrieve long MarmoSAT arrays on unmapped contigs. In some cases, the Tandem Repeats Finder[25] program was used to help in the delimitation of MarmoSAT monomers. The MarmoSAT arrays analysed in unmapped contigs and in assembled chromosomes files were carefully analysed through dot plots to determine the start and end of each repeat. Dot plots were also used to check for similarity between MarmoSAT and AS. These plots were generated with the Dotlet application with a 15 bp word size and 60% similarity cutoff.[26]

Multiple sequence alignments were performed using Muscle 4.0.[27] The MEGA software version 5.05[28] was used for the calculation of genetic distances and construction of Neighbor-Joining (NJ) trees.

2.2. Samples, DNA extractions, PCR amplifications, cloning and sequencing

Chromosome preparations and genomic DNAs were obtained from fibroblast cultures of one male of each *Callithrix penicillata*, *C. geoffroyi*, *Callimico goeldii* and *Mico argentatus*. Both *Callithrix* specimens are kept by Dr Alan Lane de Melo in animal facilities at Universidade Federal de Minas Gerais (permits 1/31/94/0000-8 and 3106.6995/2012-MG from IBAMA and 167/2006 from CETEA/UFMG, revalidated on 16 March 2012). The *M. argentatus* cells were provided by Dr Yatiyo Yonenaga-Yassuda from the Universidade de São Paulo (Brazil).

AS and MarmoSAT were amplified by polymerase chain reaction (PCR) from genomic DNAs of the three species with the following primer sets: Alpha-F (ACAGGGAAATATCTGCTTCTAAATC) and Alpha-R (GCTTACTGCTGTTTCTTCCATATG); MarmoSAT-F (ACAGAGTAGAATAGGGCATTG) and MarmoSAT-R (CCAACTCAGTATGCTCTCTCATG). The MarmoSAT set of primers were designed from consensus sequences from an unidentified *C. jacchus* satDNA. PCR reactions consisted of an initial denaturation step of 94°C for 3 min, followed by 30 cycles at 94°C for 60 s, 55°C for 60 s and 72°C for 60 s and a final extension at

72°C for 10 min. PCR products were excised from a 1% agarose gel and purified with the Wizard SV Gel and PCR Clean-up System kit (Promega). Selected MarmoSAT repeats were cloned using the pGEM-T-Easy cloning kit (Promega). Recombinant plasmids were sequenced on the ABI3130 platform (Myleus Biotechnology). The sequences generated in this study have GenBank accession numbers KX686899 (MarmoSAT – *Cebuella pygmaea*), KX686900 and KX686901 (MarmoSAT – *Mico argentatus*).

2.3. CBG-banding and fluorescence *in situ* hybridization

CBG-banding was obtained according to Sumner.[29] The *Callithrix*, *Callimico* and *M. argentatus* karyotypes were mounted following Sherlock *et al.*,[30] Neusser *et al.*[31] and Dumas *et al.*,[32] respectively. Fluorescence *in situ* hybridization (FISH) was performed using AS, MarmoSAT and telomeric sequences as probes. The satDNA probes were prepared from PCR purified products labelled by nick translation with digoxigenin-11-dUTP (DIG-Nick Translation mix, Roche Applied Science). A biotinylated telomeric sequence $(TTAGGG)_4$ (Invitrogen) was synthesized and used as probe for FISH. Chromosomes were denatured in 70% formamide/2xSSC at 65% for 1–2 min. The hybridization mix, consisting of 100 ng of labelled probe in 50% formamide/2xSSC, was denatured for 10 min at 98°C and applied to the chromosome preparations. Hybridization was carried out at 37°C for 16–20 hours. Slides were washed in 2xSSC at 37°C for 5 min. Immunodetection was performed with antidigoxigenin conjugated with FITC and neutravidin coupled with rhodamine (Roche Applied Science). The analyses and image acquisition were performed under a Zeiss Axioimager 2 epifluorescence microscope using the AxioVision software (Zeiss).

2.4. Transcription analysis

We investigated the transcription of MarmoSAT in several tissues of *C. jacchus* using the RNA-seq data generated by the Non-Human Primate Reference Transcriptome Resource (NHPRTR; Peng *et al.*[33]). These data are publicly available at NCBI under the BioProject PRJNA271912 and include total Ribo-Zero transcriptomes for bone marrow, the left and right brain hemispheres, the pituitary, colon, heart plus thymus, heart only, kidney, liver, lung, lymph node, muscle and spleen from a female marmoset.

The reads were mapped to a consensus sequence of MarmoSAT using the '–sensitive-local' preset of Bowtie2 implemented on the Galaxy platform (http://usegalaxy.org; Langmead and Salzberg[34]; Giardine *et al.*[35]; Goecks *et al.*[36]). The Neural Network Promoter Prediction tool was used to investigate potential transcription start sites within MarmoSAT (http://www.fruitfly.org/seq_tools/promoter.html; Reese[37]). Transcription of the abundant LINE-1 and Alu elements was also examined as described above and compared with that of MarmoSAT using the Spearman's rank correlation coefficient.

3. Results

3.1. Identification and characterization of MarmoSAT

We identified the most abundant families of repetitive DNAs present in the *C. jacchus* genome using the similarity-based clustering method implemented on RepeatExplorer.[22] We found that highly repetitive elements comprise ~15% of the common marmoset genome, mostly represented by TEs and satDNA families. LINE-like elements represented the most abundant repetitive family, comprising 6.7% of the genome and including 95,112 reads in 21 different clusters. The second most abundant repetitive element, Alu-like/SINE-like, spans 3.6% (54,301 reads) of the common marmoset genome. The AS DNA, the third most abundant repetitive family, represents 1.5% of *C. jacchus* genome (20,308 reads) and is organized on three different clusters inside this species genome. The fourth most abundant cluster of repetitive DNA was a still uncharacterized tandem repeat named herein MarmoSAT, which represents 1.09% of the genome and is composed of 171 bp AT-rich (61%) motifs. Considering an estimated genome size of 3.4 Gb for *C. jacchus*[38] this new satDNA family would account for more than 37 Mb or ~216,000 copies. MarmoSAT arrays were identified in all assembled chromosome files, including the sex chromosomes, and presented an average size of 2,223 bp, ranging from 296 bp on chromosome 13 to 8,188 bp on chromosome 3. The average nucleotide divergence among repeats is 20.18%, ranging from 11.5% (on chromosome 14) to 39.1% (on chromosome 22) (Supplementary Table S2).

The amount of MarmoSAT differed among the assembled chromosome files, but this variation may be related to technical limitations. We also identified arrays present in different loci inside chromosomes 1, 3, 8, 9, 10, 15, 16, 18, 19, and 21, whereas on chromosomes 8, 15, 16 and 21, MarmoSAT is organized as one single array interspersed with several TEs insertions (Supplementary Table S3).

In order to investigate whether MarmoSAT repeats form long-arrays, we performed BLAST searches on unmapped *C. jacchus* contigs and retrieved the five contigs with the highest score values (accession numbers: ACFV01174585.1, ACFV01176989.1, ACFV01177303.1, ACFV01181345.1and ACFV01184555.1). As a result, we obtained 492 copies of MarmoSAT with an average array size of 99.8 copies per contig, suggesting the presence of long arrays of MarmoSAT in the *C. jacchus* unmapped data.

Because MarmoSAT and AS repeat units have similar sizes (~171 bp), we performed extensive sequence comparisons with described AS sequences from several primates, including *C. jacchus* and humans, but dot plot analyses could not find any significant similarity hits between the reads in the MarmoSAT cluster and the ones in the ASor CarB clusters, indicating lack of homology (Supplementary Fig. S2). Moreover, BLAST searches revealed no significant similarity between this sequence and any other deposited in the GenBank or RepBase databases.

In order to investigate the homogenization dynamics of MarmoSAT related to their genomic location (from different arrays or chromosomes), we constructed NJ phylogenetic trees using 458 copies extracted from all chromosome files, three BACs, one from the X (accession number: AC146662.3) and two from the Y chromosome (accession numbers: AC243896.4 and AC243459.3), and 492 copies retrieved from unassembled contigs (Supplementary Fig. S3). The resulting trees showed that MarmoSAT sequences were not clustered in chromosome-specific branches. Similarly, the repeats found in the three BACs did not cluster together with the assembled sequences of the corresponding X and Y chromosomes. We also produced chromosome-specific trees, but still could not find any array specificity for MarmoSAT repeats (data not shown).

3.2. MarmoSAT flanking regions are enriched with Alu-like and L1-Cja-like retrotransposons

Aiming to better understand the genomic distribution and possible association with different genetic elements, we analysed 42 flanking

regions (from 500 bp to 1 kb, when available) of the MarmoSAT arrays found on the assembled chromosome files (Supplementary Table S3). Our analysis revealed that TEs Alu-like and L1-Cja-like were often associated with MarmoSAT arrays. We found 20 insertions of Alu-like elements adjacent to MarmoSAT arrays, whereas 17 L1-Cja-like elements were found neighbouring MarmoSAT sequences. In contrast, we did not find AS sequences associated with MarmoSAT. We found no preferential positions for TE insertion inside MarmoSAT monomers, and no micro-sequence similarities that could indicate any insertion bias (data not shown).

We also found the subtelomeric satDNA family CarB, previously described in *M. argentatus*,[19] flanking the 5' regions of MarmoSAT arrays on chromosomes 14 and 16 (Supplementary Table S3), and one copy flanking MarmoSAT in one BAC mapped on chromosome Y (accession number AC243896.4). Interestingly, we identified CarB as representing only 0.108% of the *C. jacchus* genome. We found low copy numbers of CarB on chromosomes 1, 3–8, 13–20, and it was absent on the other chromosomes.

3.3. Phylogenetic distribution of MarmoSAT

In order to ascertain the distribution of MarmoSAT in *Cebuella pygmaea* and *Mico argentatus* we amplified this sequence by PCR from genomic DNAs of both species. The sequencing of cloned PCR products from both species confirmed the presence of MarmoSAT in their genomes.

The distribution of MarmoSAT in other NWM species was verified by an *in silico* search using a MarmoSAT query against the GenBank non-redundant nucleotide collection (NCBI). BLAST searches retrieved one significant hit with e-value $2e^{-13}$ corresponding to satDNA CgoA (accession number: X52012.1), previously described in *Callimico goeldii*.[18] This 338 bp satDNA family was described as restricted to the *C. goeldii* genome after Southern blot hybridizations with 70% stringency. Dot-plot analysis of the CgoA repeat unit sequences showed that this satDNA is composed of two monomers (Fig. 1), one with 170 bp (CgoA1), and another with 168 bp (CgoA2). Pairwise sequence comparisons between CgoA monomers revealed 32.9% nucleotide divergence, indicating a HOR organization in *C. goeldii*. The comparison of CgoA1 and CgoA2 against MarmoSAT revealed 30.3% and 33.3% of divergence, respectively.

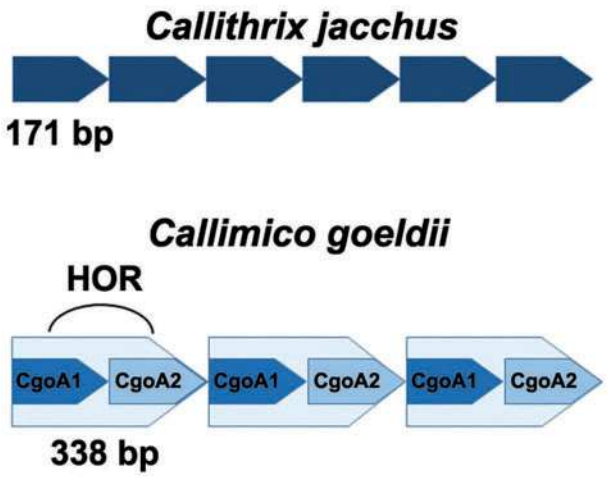

Figure 1. Schematic illustration of MarmoSAT repeat units of 171 bp in *Callithrix jacchus* and 338 bp HOR of CgoA1 and CgoA2 units in *Callimico goeldii*.

We searched for MamoSAT in the outgroup species *Aotus nancymaae* and *Saimiri boliviensis* (accession numbers SRR1692997 and SRR315548), but did not find homologous sequences. The absence of MarmoSAT in the *Aotus* and *Saimiri* lineages suggests that this satDNA family probably amplified after the split of Callitrichinae (Fig. 2).

3.4. Chromosomal location of MarmoSAT on Callitrichini and *Callimico*

The chromosome location of MarmoSAT was investigated after FISH on chromosomes from *Callithrix penicillata, C. geoffroyi, Mico argentatus* and *Callimico goeldii*. CBG-banding was also performed in order to compare the distribution of MarmoSAT in relation to the constitutive heterochromatin in the four species (Figs 3A and C and 4).

The MarmoSAT probe hybridized to both subtelomeric regions of all biarmed chromosomes in *C. penicillata*, with the exception of pair 3 and the sex chromosomes, which had only their short arms labelled (Fig. 3B). Pairs 1 and 18 showed interstitial labelling on their short and long arms, respectively, whereas pairs 16 and 17 had no hybridization signals. *C. geoffroyi* chromosomes presented the same hybridization pattern, with the exception of pair 15, which did not display any signal (Fig. 3D). In *M. argentatus*, MarmoSAT sequences were visualized on both ends of the biarmed pairs 3, 20 and in the sex chromosomes; in the subtelomeric regions of the short arm of pairs 10 and 21 and in the long arms of pairs 2, 5–7, 13 and 14. In addition, interstitial signals were detected in the short arms of pairs 2 and 3 and in the long arms of pairs 5 and 14. Among the acrocentric chromosomes, MarmoSAT sequences hybridized to the long arms of pairs 15, 17 and 19 and to interstitial regions of pairs 15, 16 and 19 (Fig. 4A). In *C. goeldii*, MarmoSAT sequences were located at subtelomeric regions of the short and long arms of biarmed chromosomes, with the exception of pair 11, which showed signals only on its long arms. The acrocentrics had only their long arms labelled (Fig. 4B).

Since MarmoSAT had a telomeric localization, we also performed double FISH with a telomeric probe $(TTAGGG)_4$. Telomeric sequences were detected in all telomeres of the three species (Figs 3B, 3D and 4A) and co-localized with MarmoSAT in most chromosome pairs. Additionally, the Y chromosomes of both *Callithrix* species had their long arms completely labelled with the telomeric probe.

A search for sequencing reads that spanned the interface between MarmoSAT and telomeric sequences on the Trace Archive File of *C. jacchus* (http://www.ncbi.nlm.nih.gov/Traces/home/) revealed two reads composed by MarmoSAT arrays adjacent to telomeric repeats. In both cases, the telomeric repeats were in the terminal 3' position, indicating a subtelomeric location of MarmoSAT (Fig. 5).

Besides the MarmoSAT hybridizations, we also report for the first time the chromosome location of AS in *C. penicillata, C. geoffroyi*, *M. argentatus* and *C. goeldii* (Fig. 6). This satDNA is present in large amounts in the (peri)centromeric regions of these marmosets' chromosomes, corroborating the bioinformatics analysis and previous cytogenetic studies of *C. jacchus*.[11]

3.5. MarmoSAT is transcribed on several tissues from *C. jacchus*

We used the NHPRTR[33] data to investigate the transcription of MarmoSAT in the tissues of *C. jacchus*. This analysis revealed MarmoSAT transcripts in all tissues surveyed albeit at very different abundances (Fig. 7). The 13 tissues analysed displayed over 11-fold variation in the transcription level of MarmoSAT, with a higher-

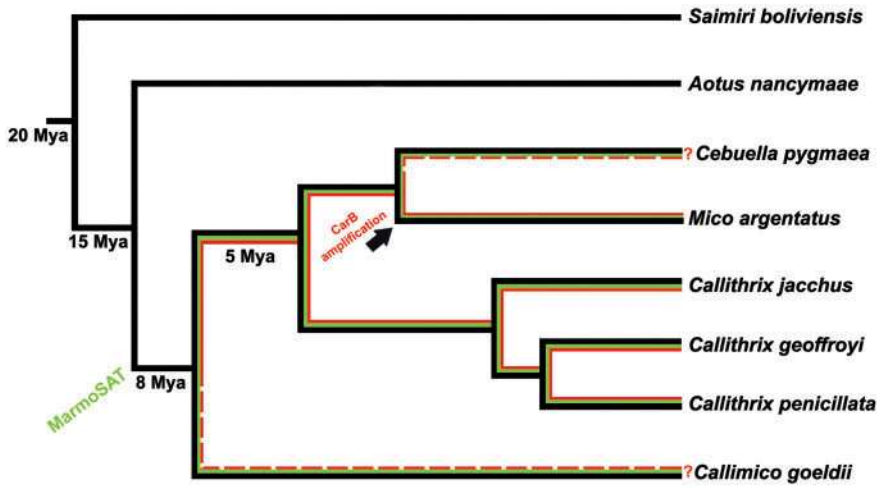

Figure 2. Possible evolutionary pathway of MarmoSAT and CarB amplification in Callitrichini and *Callimico goeldii*. Phylogenetic relationships are based on Perelman et al.[2]. Colored branches indicate the presence of MarmoSAT (green, lighter color) and CarB (red, darker color) satellite DNA families. The traced lines in the *Callimico* and *Cebuella* lineages indicate insufficient data to verify the hypothesis.

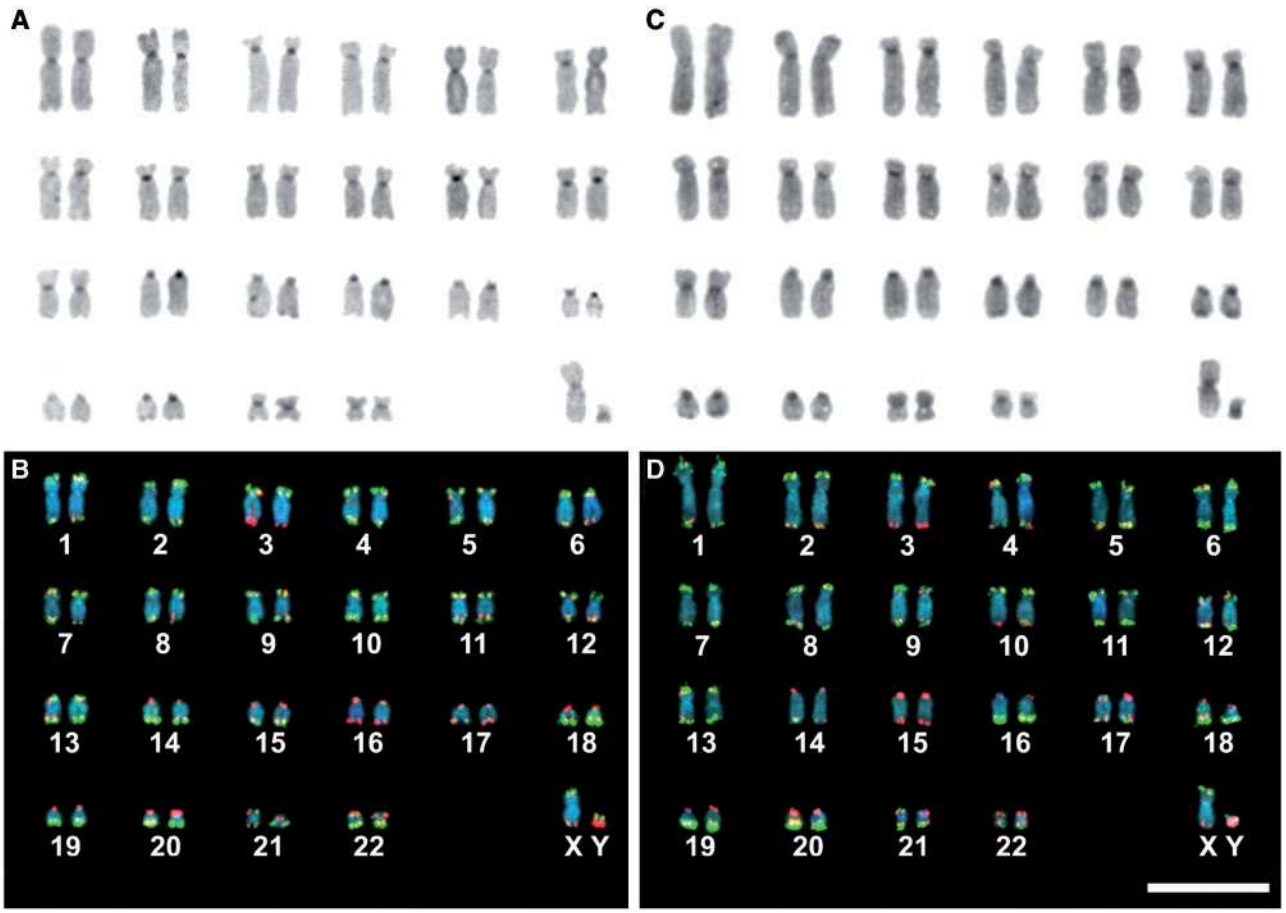

Figure 3. *Callithrix penicillata* metaphases after (A) CBG-banding and (B) FISH with the MarmoSAT (green) and telomeric (red) probes. C-D show the results of the same experiments in *C. geoffroyi*. Bar = 10 μm. Colour visible in online version.

than-average transcript level in brain (left and right hemispheres), heart plus thymus, heart only and spleen (Fig. 7).

Because we found no putative promoter regions in MarmoSAT sequences, since L1 and Alu were the sequences most commonly associated with MarmoSAT arrays, we also analysed their transcripts (Supplementary Table S3). L1 and Alu also displayed ubiquitous transcription, but showed no significant correlation with MarmoSAT transcription levels (Spearman's $R = 0.24692$

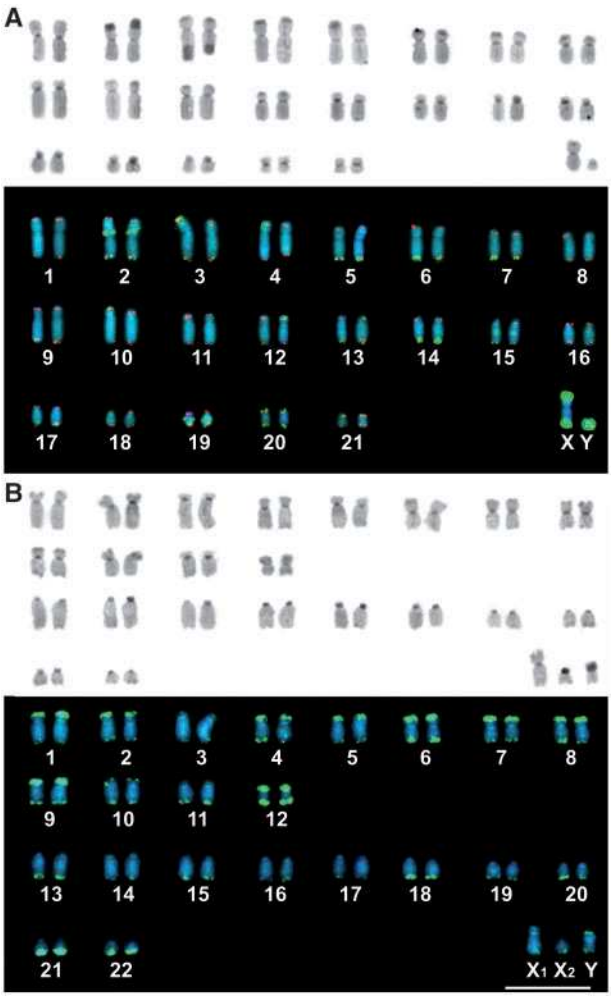

Figure 4. Karyotypes of *Mico argentatus* after (A) CBG-banding and FISH with MarmoSAT (green) and telomeric sequences (red) probes. In (B), CBG-banding and FISH with MarmoSAT sequences probe in *Callimico goeldii*. Bar = 10 μm. Colour visible in online version.

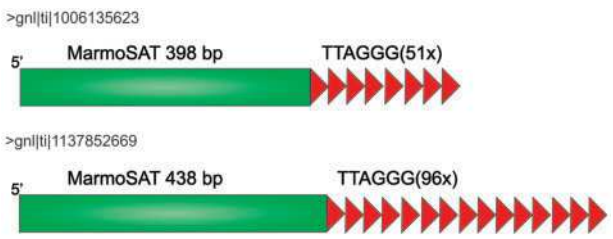

Figure 5. Schematic representation of MarmoSAT repeats adjacent to telomeric repeats on reads *gnl:1006135623* and *gnl:1137852669* found in the NCBI Trace Archive Files database.

and -0.1318, $P = 0.3733$ and 0.6693, respectively), and only moderate correlation with one another ($R = 0.4835$ and $P = 0.097$; Supplementary Fig. S4).

4. Discussion

In this study, we identified a new satDNA with 171 bp monomers in the common marmoset genome, which was named MarmoSAT. Regardless of the same motif size, MarmoSAT and AS do not share any sequence similarity or conserved structure that could suggest a common origin (Supplementary Fig. S2). The presence of two non-homologous complex satDNA families that share the same motif size was not previously reported in primates. It is possible that the convergence to the same motif size reflects the optimal distance for nucleosome positioning, as suggested by Henikoff *et al.*,[39] indicating that despite sequence heterogeneity, different satDNA families may retain structural features important to heterochromatic domains in *C. jacchus*.

4.1. Evolutionary turnover of subtelomeric satDNAs MarmoSAT and CarB in Callitrichini

SatDNAs have been identified in the subtelomeric heterochromatin of most marmoset chromosomes[19, 20, 40]. In the *Callithrix* species analysed herein we found MarmoSAT to be a major component of these regions (Fig. 3). The hybridization of MarmoSAT and telomeric sequences showed very similar patterns in both *C. penicillata* and *C. geoffroyi*, which may be explained by the recent diversification of the clade (2.6 Mya).[3]

Together with the cytogenetic analysis, our searches on assembled genomes and WGS data of NWMs showed that MarmoSAT is apparently restricted to *Callithrix*, *Mico*, *Callimico* and *Cebuella*, while CarB is present in *Callithrix* and *Mico* (Fig. 2). Therefore, MarmoSAT and CarB repeats, a satDNA previously described in *Mico* species,[19,20] were already present in the common ancestor of Callitrichini and *C. goeldii*. The FISH experiments revealed that MarmoSAT is more abundant in *Callithrix* and *C. goeldii* than in *Mico* (Figs 3 and 4). Accordingly, Fanning *et al.*[18] CgoA description supports MarmoSAT high abundancy in *C. goeldii*. Altogether, these results support the hypothesis that MarmoSAT sequences were probably abundant in the ancestor of Callitrichini and *C. goeldii* and remained copious in *Callithrix* and *Callimico*. On the other hand, CarB underwent amplification in *Mico*, as shown by Alves *et al.*[19] and Canavez *et al.*,[20] replacing MarmoSAT. This assumption is supported for instance by comparing the homologous *M. argentatus* chromosome 4 and *Callithrix* chromosome 13, *M. argentatus* chromosome 5 short arm and *Callithrix* pair 20, and *M. argentatus* chromosome 18 and *Callithrix* 19, in which the *Callithrix* counterparts showed MarmoSAT hybridization, whereas in *M. argentatus* chromosomes there were large heterochromatic blocks mainly composed of CarB.

The amplification of specific satDNAs was already reported in *Callimico* and *Cebuella*.[18,40] In marmosets, the differential amplification of unrelated satDNAs does not seem to affect karyotype stability, as this group of NWMs has conserved karyotypes.

4.2. MarmoSAT subtelomeric arrays lack chromosome-specificity

SatDNA repeats on the same array or chromosome tend to present a higher level of sequence similarity to each other when compared with those on non-homologous chromosomes.[41,42] An NJ phylogenetic tree with 980 copies of MarmoSAT revealed that these sequences lack chromosome specificity (Supplementary Fig. S3). The absence of differential homogenization for specific variants between MarmoSAT subtelomeric arrays in *C. jacchus* could be the result of exchange events between arrays and chromosomes. This hypothesis is supported by data in chimpanze, in which it has been shown that the subterminal regions of different chromosomes interact with each other to form stable physical contacts in meiosis,[43] which may result

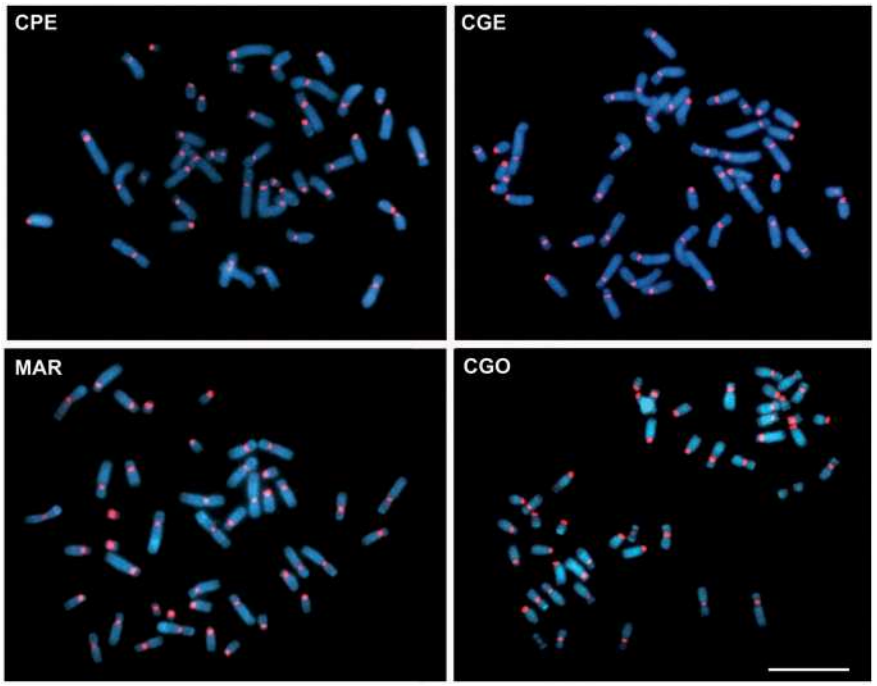

Figure 6. FISH with a digoxigenin-labelled α-satellite DNA in Callitrichini and *Callimico goeldii* cells. CPE = *Callithrix penicillata*, CGE = *C. geoffroyi*, MAR = *Mico argentatus*, CGO = *C. goeldii*. Bar = 10 μm.

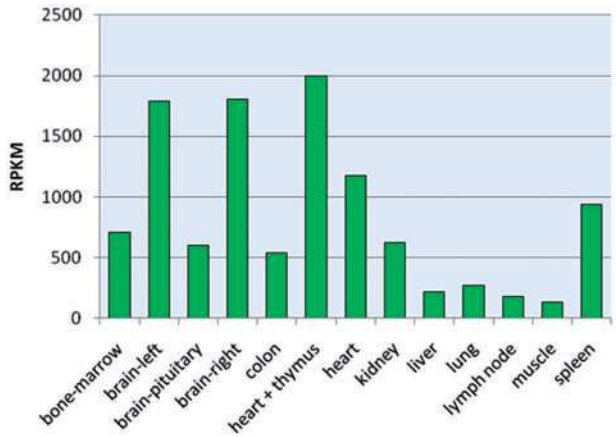

Figure 7. Transcription level of MarmoSAT in several tissues of a female *Callithrix jacchus*. RPKM: *Reads Per Kilobase Million*.

in frequent DNA exchanges between chromatids. Moreover, Rudd *et al.*[44] showed that in humans, 17% of all sister chromatid exchanges occur in the terminal ~100 kb of chromosomes, translating into a recombination rate on subtelomeric regions 160-fold larger than in euchromatic regions.

4.3. MarmoSAT is present as HORs in *Callimico goeldii*

Another significant finding was the identification of MarmoSAT organized in HOR structures in *C. goeldii*. Although HORs are apparently common and widespread in primates,[10,12] they are predominantly located at (peri)centromeric, rather than in subtelomeric regions.[45] Herein we showed that the previously described copies of CgoA satDNA[18] are in fact two highly different monomers of MarmoSAT organized in HORs (Fig. 1). Probably, the establishment of a HOR organization occurred after the divergence of *C. jacchus* and *C. goeldii* ~8 million years ago (Mya) (Fig. 2). Moreover, we observed that the subtelomeric location of MarmoSAT is shared among Callitrichini species and *C. goeldii* (Figs 3 and 4) This is the first report of subtelomeric HORs in primates, indicating that HORs sequences may be found in heterochromatic regions outside of centromeres.

The example of HORs in closely related species such as *C. jacchus* and *C. goeldii* is illuminating since it shows that HORs may develop over a relatively short period of evolutionary time. In contrast with what was found in *C. jacchus* AS,[12] the sequence similarity between *C. goeldii* CgoA1 and CgoA2 repeat units is considerably lower than 70%. This pattern is similar to the observed in previously described HORs from primates and bovids.[21,46]

4.4. Possible role for L1 in MarmoSAT dispersion

The presence of several MarmoSAT arrays adjacent to L1 opens the possibility that their dispersion could have been mediated by these retroelements (Supplementary Table S3). Interestingly, we detected many copies of MarmoSAT located interstitially on *C. penicillata* and *C. geoffroyi* pairs 1 and 18 and *M. argentatus* pairs 2, 3, 5, 14–16 and 19 (Figs 3 and 4). In humans, it has been shown *in vivo* and *in vitro* that the L1 can co-mobilize 3' downstream sequences to other genomic locations as the result of imperfect transcription events[47]. During the transposition event the transcription of an L1-element may bypass its own polyadenylation signal utilizing a second downstream polyadenylation site for 3' end processing, leading to the transcription and later transposition of adjacent flanking sequences. Even though most of L1-elements have truncated sequences, these elements are still capable of retrotransposition,[48] suggesting that even incomplete L1s may play a role in such events.

The finding of Alu-like and L1-Cja-like flanking MarmoSAT may have resulted from a bias in assembling repetitive DNAs with WGS

data, due to the high sequence similarity among repeats.[49] In fact, unmapped sequences have a low frequency of TEs insertions when compared with the sequences present in assembled chromosomes. We also observed that these TEs lack preferential insertion sites in MarmoSAT arrays, either target sites or array positioning, indicating a probable random nature for these events. For example, it has been suggested that L1 elements have a cleavage site preference for sequences rich in AA|TTT,[50] but we did not observe any insertion at these specific sites. Although MarmoSAT sequences have three different AA|TTT regions, the absence of TEs insertions may be affected by the local chromatin structure.[51] Alternatively, TEs loci inside MarmoSAT sequences may have suffered several mutations resulting in different sequences compared with the initial insertion regions.

4.5. Potential functional roles of MarmoSAT

Albeit being devoid of protein-coding capacity, satDNAs may possess structural and/or functional roles, usually via expression of non-coding RNAs (ncRNAs; reviewed by Biscotti *et al.*[52]). Interestingly, we found MarmoSAT-derived transcripts in all tissues analysed (Fig. 7) and because this satDNA does not possess promoter regions we conclude that its transcription is probably initiated in flanking sequences. The most frequent sequences flanking MarmoSAT arrays are L1 and Alu retroelements, but MarmoSAT transcription levels do not correlate with those from these elements (Supplementary Fig. S4). Because of its abundance and presence in almost all chromosomes, MarmoSAT could be transcribed from a number of different loci and have different promoters, inside or outside L1/Alu.

Besides the subtelomeric location determined by FISH, we found Sanger sequencing reads that span both MarmoSAT and the telomeric repeats in *C. jacchus* (Fig. 5). ncRNAs containing telomeric repeats in mammals are known as telomere repeat containing RNAs (TERRAs) and are thought to regulate telomerase and to modulate telomeric chromatin throughout the cell cycle (reviewed in Luke and Lingner[53]). TERRAs transcription has been shown to start at subtelomeric repeats towards the chromosome ends in human, mouse and yeast.[54–56] MarmoSAT abundance and location make it possible to suggest that it could be part of TERRAs. Moreover, MarmoSAT expression profile is similar to the TERRA expression found in mouse, with higher levels of transcripts in spleen, kidney and thymus.[55] Although the results presented herein are suggestive of interesting functional roles for MarmoSAT transcripts, the validity of these hypotheses must be assessed experimentally.

5. Conclusions

In this study, we identified with a combination of bioinformatics and cytogenetics, a subtelomeric satDNA, termed MarmoSAT, in the common marmoset and provided insights into its organization and evolution. Our data suggest that MarmoSAT originated in the common ancestor of Callitrichini and *Callimico goeldii*. Interestingly, the MarmoSAT arrays are organized in HORs in *C. goeldii*. We also propose that MarmoSAT transcripts may play a role in telomeric chromatin.

Supplementary data

Supplementary data are available at *DNARES* Online.

Conflict of interest

None declared.

Accession numbers

Sequencing data generated for this study have been submitted to GenBank under accession numbers KX686899 (MarmoSAT – *Cebuella pygmaea*), KX686900 and KX686901 (MarmoSAT – *Mico argentatus*).

Funding

This work was supported by a grant from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, process 407262/2013-0 to MS and RS); Progetti di Interesse Nazionale 2012 (PRIN) to RS; Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG). NPA, LGL and GBD have doctoral fellowships from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

1. Schneider, H., Sampaio, I. 2015, The systematics and evolution of New World primates – A review. *Mol. Phylogenet. Evol.*, **82**, 348–57.
2. Perelman, P., Johnson, W.E., Roos, C., et al. 2011, A molecular phylogeny of living Primates. *PLoS Genet.*, **7**, e1001342.
3. Schneider, H., Bernardi, J.A.R., da Cunha, D.B., et al. 2012, A molecular analysis of the evolutionary relationships in the Callitrichinae, with emphasis on the position of the dwarf marmoset. *Zool. Scr.*, **41**, 1–10.
4. Nagamachi, C.Y., Pieczarka, J.C., Barros, R.M.S. 1992, Karyotypic comparison among *Cebuella pygmaea, Callithrix jacchus* and *C. emiliae* (Callitrichidae, Primates) and its taxonomic implications. *Genetica*, **85**, 249–57.
5. Sachidanandam, R., Weissman, D., Schmidt, S.C., et al. 2001, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–33.
6. Cordaux, R., Batzer, M.A. 2009, The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.*, **10**, 691–703.
7. Melters, D.P., Bradnam, K.R., Young, H.A., et al. 2013, Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.*, **14**, R10.
8. Plohl, M., Meštrović, N., Mravinac, B. 2014, Centromere identity from the DNA point of view. *Chromosoma*, **123**, 313–25.
9. Alves, G., Seuánez, H.N., Fanning, T. 1994, Alpha satellite DNA in neotropical primates (Platyrrhini). *Chromosoma*, **103**, 262–7.
10. Alkan, C., Ventura, M., Archidiacono, N., Rocchi, M., Sahinalp, S.C., Eichler, E.E. 2007, Organization and evolution of primate centromeric DNA from Whole-Genome Shotgun sequence data. *PLoS Comput. Biol.*, **3**, e181.
11. Cellamare, A., Catacchio, C.R., Alkan, C., et al. 2009, New insights into centromere organization and evolution from the white-cheeked gibbon and marmoset. *Mol. Biol. Evol.*, **26**, 1889–900.
12. Sujiwattanarat, P., Thapana, W., Srikulnath, K., Hirai, Y., Hirai, H., Koga, A. 2015, Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. *Sci. Rep.*, **5**, 10315.
13. Pezer, Ž., Brajković, J., Feliciello, I., Ugarković, Đ. 2012, Satellite DNA-mediated effects on genome regulation. *Genome Dyn.*, **7**, 153–69.
14. Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I., Martienssen, R.A. 2002, Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, **297**, 1833–7.
15. Rošić, S., Köhler, F., Erhardt, S. 2014, Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. *J. Cell. Biol.*, **207**, 335–349.
16. Feliciello, I., Akrap, I., Ugarkovic, D. 2015, Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. *PLoS Genet.*, **11**, e1005466.
17. Chan, F.L., Marshall, O.J., Saffery, R., et al. 2012, Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proc. Natl. Acad. Sci. USA*, **109**, 1979–84.

18. Fanning, T.G., Seuánez, H.N., Forman, L. 1989, Satellite DNA sequences in the neotropical marmoset *Callimico goeldii* (Primates, Platyrrhini). *Chromosoma*, **98**, 396–401.
19. Alves, G., Canavez, F., Seuánez, H., Fanning, T. 1995, Recently amplified satellite DNA in *Callithrix argentata* (Primates, Platyrrhini). *Chromosome Res*., **3**, 207–13.
20. Canavez, F., Alves, G., Fanning, T.G., Seuánez, H.N. 1996, Comparative karyology and evolution of the Amazonian *Callithrix* (Platyrrhini, Primates). *Chromosoma*, **104**, 348–57.
21. Prakhongcheep, O., Chaiprasertsri, N., Terada, S., et al. 2013, Heterochromatin blocks constituting the entire short arms of acrocentric chromosomes of Azara's owl monkey: formation processes inferred from chromosomal locations. *DNA Res*., **20**, 461–70.
22. Novák, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J. 2013, RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–3.
23. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. 2005, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, **110**, 462–7.
24. Bao, W., Kojima, K.K., Kohany, O. 2015, Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, **6**, 11.
25. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*., **27**, 573–80.
26. Junier, T., Pagni, M. 2000, Dotlet: diagonal plots in a web browser. *Bioinformatics*, **16**, 178–9.
27. Edgar, R.C. 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–7.
28. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S. 2011, MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol*., **28**, 2731–9.
29. Sumner, A.T. 1972, A simple technique for demonstrating centromeric heterocromatin. *Exp. Cell Res*., **75**, 305–6.
30. Sherlock, J.K., Griffin, D.K., Delhanty, J.D.A., Parrington, J.M. 1996, Homologies between human and marmoset (*Callithrix jacchus*) chromosomes revealed by comparative chromosome painting. *Genomics*, **33**, 214–9.
31. Neusser, M., Stanyon, R., Bigoni, F., Wienberg, J., Müller, S. 2001, Molecular cytotaxonomy of New World monkeys (Platyrrhini) – comparative analysis of five species by multi-color chromosome painting gives evidence for a classification of *Callimico goeldii* within the family of Callitrichidae. *Cytogenet. Cell Genet*., **94**, 206–15.
32. Dumas, F., Stanyon, R., Sineo, L., Stone, G., Bigoni, F. 2007, Phylogenomics of species from four genera of New World monkeys by flow sorting and reciprocal chromosome painting. *BMC Evol. Biol*., **7(Suppl 2)**, S11.
33. Peng, X., Thierry-Mieg, J., Thierry-Mieg, D., et al. 2014, Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). *Nucleic Acids Res*., **43** (Database issue), D737–42.
34. Langmead, B., Salzberg, S.L. 2012, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–9.
35. Giardine, B., Riemer, C., Hardison, R.C., et al. 2005, Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*., **15**, 1451–5.
36. Goecks, J., Nekrutenko, A., Taylor, J. 2010, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*., **11**, R86.
37. Reese, M.G. 2001, Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem*., **26**, 51–6.
38. Pellicciari, C., Formenti, D., Redi, C.A., Manfredi, M.G. 1982, DNA content variability in primates. *J. Hum. Evol*., **11**, 131–41.
39. Henikoff, S., Ahmad, K., Malik, H.S. 2001, The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293**, 1098–102.
40. Neusser, M., Münch, M., Anzenberger, G., Müller, S. 2005, Investigation of marmoset hybrids (*Cebuella pygmaea* x *Callithrix jacchus*) and related Callitrichinae (Platyrrhini) by cross-species chromosome painting and comparative genomic hybridization. *Cytogenet. Genome Res*., **108**, 191–6.
41. Dover, G. 1982, Molecular drive: a cohesive mode of species evolution. *Nature*, **299**, 111–7.
42. Kuhn, G.C.S., Küttler, H., Moreira-Filho, O., Heslop-Harrison, J.S. 2012, The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. *Mol. Biol. Evol*., **29**, 7–11.
43. Hirai, H., Matsubayashi, K., Kumazaki, K., Kato, A., Maeda, N., Kim, H.S. 2004, Chimpanzee chromosomes: retrotransposable compound repeat DNA organization (RCRO) and its influence on meiotic prophase and crossing-over. *Cytogenet. Genome Res*., **108**, 248–54.
44. Rudd, M.K., Friedman, C., Parghi, S.S., Linardopoulou, E.V., Hsu, L., Trask, B.J. 2007, Elevated rates of sister chromatid exchange at chromosome ends. *PLoS Genet*., **3**, e32.
45. Koga, A., Hirai, Y., Terada, S., et al. 2014, Evolutionary origin of higher-order repeat structure in alpha-satellite DNA of primate centromeres. *DNA Res*., **21**, 407–15.
46. Modi, W.S., Ivanov, S., Gallagher, D.S. 2004, Concerted evolution and higher-order repeat structure of the 1.709 (satellite IV) family in bovids. *J. Mol. Evol*., **58**, 460–5.
47. Goodier, J.L., Ostertag, E.M., Kazazian, H.H. Jr 2000, Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet*., **9**, 653–7.
48. Belancio, V.P., Roy-Engel, A.M., Pochampally, R.R., Deininger, P. 2010, Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Res*., **38**, 3909–22.
49. Treangen, T.J., Salzberg, S.L. 2011, Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet*., **13**, 36–46.
50. Cost, G.J., Boeke, J.D. 1998, Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry*, 18081–93.
51. Cost, G.J., Golding, A., Schlissel, M.S., Boeke, J.D. 2001, Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Res*., **29**, 573–7.
52. Biscotti, M.A., Canapa, A., Forconi, M., Olmo, E., Barucca, M. 2015, Transcription of tandemly repetitive DNA: functional roles. *Chromosome Res*., **23**, 463–77.
53. Luke, B., Lingner, J. 2009, TERRA: telomeric repeat-containing RNA. *EMBO J*., **28**, 2503–10.
54. Azzalin, C.M., Reichenbach, P., Khoriauli, L., Giulotto, E., Lingner, J. 2007, Telomeric repeat-containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science*, **318**, 798–801.
55. Schoeftner, S., Blasco, M.A. 2008, Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nat. Cell Biol*., **10**, 228–36.
56. Porro, A., Feuerhahn, S., Delafontaine, J., Riethman, H., Rougemont, J., Lingner, J. 2014, Functional characterization of the TERRA transcriptome at damaged telomeres. *Nat. Commun.* **5**, 5379.