Evolutionary history of alpha satellite DNA repeats dispersed within human genome euchromatin

Isidoro Feliciello[1,2,] *, Željka Pezer[1], Dušan Kordiš[3], Branka Bruvo Mađarić[1] and Đurđica Ugarković[1,*]

[1] Department of Molecular Biology, Ruđer Bošković Institute, Bijenička 54, HR-10000 Zagreb, Croatia;

[2] Dipartimento di Medicina Clinica e Chirurgia, Universita' degli Studi di Napoli Federico II, via Pansini 5, I-80131 Napoli, Italy;

[3] Department of Molecular and Biomedical Sciences, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia;

***** Correspondence: ugarkov@irb.hr; Tel.: +385 14561197 (Đ.U); ifelicie@unina.it; Tel.: +39 3495250441 (I.F.)

Abstract

Major human alpha satellite DNA repeats are preferentially assembled within (peri)centromeric regions but are also dispersed within euchromatin in the form of clustered or short single repeat arrays. To study the evolutionary history of single euchromatic human alpha satellite repeats (AR) we analysed their orthologous loci across the primate genomes. The continuous insertion of euchromatic ARs throughout the evolutionary history of primates starting with the ancestors of Simiformes (45-60 Myr ago) and continuing up to the ancestors of Homo is revealed. Once inserted, the euchromatic ARs were stably transmitted to the descendant species, some exhibiting copy number variation, while their sequence divergence followed the species phylogeny. Many euchromatic ARs have sequence characteristics of (peri)centromeric alpha repeats suggesting heterochromatin as a source of dispersed euchromatic ARs. The majority of euchromatic ARs are inserted in the vicinity of other repetitive elements such as L1, *Alu* and ERV, or are embedded within them. Irrespective of the insertion context, each AR insertion seems to be unique and once inserted, ARs do not seem to be subsequently spread to new genomic locations. In spite of association with (retro)transposable elements there is no indication that such elements play a role in ARs proliferation. The presence of short duplications at most of ARs insertion sites suggests site directed recombination between homologous motifs in ARs and in the target genomic sequence, probably mediated by extrachromosomal circular DNA, as a mechanism of spreading within euchromatin.

Key words: satellite DNA, euchromatin, heterochromatin, evolution, proliferation, primates

Significance statement:

Satellite DNAs are major constituents of pericentromeric and centromeric regions in many eukaryotes and their role in centromere and kinetochore assembly and heterochromatin formation has been extensively investigated. However, the role of satellite repeats found dispersed in euchromatin, outside of centromere/pericentromere regions, remains largely unexplored. Here we analyse the dynamics of dispersion of human alpha repeats throughout euchromatin during the evolutionary history of primates and the mechanism of their proliferation. The results contribute to our understanding of the possible evolutionary and functional significance of satellite DNA repeats spread throughout euchromatin.

Introduction

Satellite DNAs are tandemly repeated sequences assembled in large arrays within constitutive heterochromatin in (peri)centromeric and/or telomeric regions of eukaryotic chromosomes. Within euchromatin, longer arrays of tandem satellite repeats are generally rare, probably due to the instability caused by intrastrand homologous recombination, although blocks of euchromatic tandem repeats have been found in several species (Kuhn et al. 2012; Pavlek et al. 2015; Pita et al. 2017; Vlahović et al. 2017). Bioinformatic analyses of sequenced genomes however reveal many single repeats or short arrays of satellite DNAs dispersed in the vicinity of genes within euchromatin in diverse species such as mouse (Bulut-Karslioglu et al. 2012) or insects (Brajković et al. 2012; 2018; Kuhn et al. 2012; Ruiz-Ruano et al. 2016). The pattern of dispersion of satellite DNA repeats within euchromatin is very dynamic and differs among related species or even among strains of the same species, as shown for *Drosophila* and *Tribolium castaneum* satellite DNAs, respectively, suggesting that similar to transposable elements, euchromatic satellite repeats are subjected to cycles of proliferation (Feliciello et al. 2015a and b; Sproul et al. 2020). Some euchromatic satellite repeats such as those of a major satellite DNA in the beetle *T. castaneum* modulate the local chromatin environment upon heat stress, affecting the expression of neighbouring genes (Feliciello et al. 2015b). Moreover, euchromatic repeats seem to be in spatial contact with heterochromatin, suggesting that the interplay between euchromatic and heterochromatic repeats could play a role in gene expression modulation (Feliciello et al. 2015b; Lee et al. 2020). In the mosquito *Aedes aegypti* satellite repeats located at a single euchromatic locus promote sequence-specific gene silencing via the expression of abundant PIWI-interacting RNAs (piRNAs; Halbach et al. 2020). Euchromatic satellite repeats also facilitate X chromosome recognition/dosage compensation in *Drosophila* (Menon et al. 2014; Joshi and Meller 2017).

Alpha satellite DNA makes up to 10% of the human genome, it is located in the centromeric and pericentromeric regions of all chromosomes, contributing to essential chromosomal functions such as centromere and kinetochore assembly and heterochromatin formation (McNulty and Sullivan 2018). Based on its wide presence among primates (Willard 1991; Alexandrov et al. 2001; Cacheux et al. 2016) and according to primate phylogeny (Finstermeier et al. 2013; Pozzi et al. 2014), the age of alpha satellite DNA could be estimated to approximately 65-70 Myr. The fundamental unit of human alpha satellite DNA is based on diverged 171 bp monomers which are often organized in complex higher order repeats (Lee et al. 1997), forming chromosome-specific alpha satellite subfamilies (Willard 1985). In addition to their (peri)centromeric location, a bioinformatic search of the human genome revealed the presence of 133 blocks of alpha satellite located >5 Mb from the centromere (Rudd and Willard, 2004). Heterochromatic and euchromatic alpha satellite DNA repeats are

characterized by increased levels of H3K9me3 upon heat stress which can possibly affect neighbouring gene expression (Feliciello et al. 2020).

In this study, using a bioinformatics approach, we characterize annotated alpha satellite repeats dispersed within euchromatin of the human genome, in particular those not organized in clusters but present as single repeat arrays. We trace their evolutionary history using other assembled primate genomes, in order to date when during the evolutionary history each of the extant dispersed alpha repeat was inserted within a particular primate genome and in which context the insertion occurred. We also follow the sequence evolution of the dispersed repeats and compare it to the evolution of species, and analyse the relation of dispersed repeats with centromeric and pericentromeric alpha satellite monomers. Finally, the mechanism of insertion and spreading of alpha repeats along the euchromatic portion of genome is studied. Our study of dynamics of dispersion of human alpha repeats throughout euchromatin during the evolutionary history contributes to the understanding of the possible evolutionary and functional significance of the satellite repeats spreading process.

Materials and methods

Detection and analysis of human ARs and their orthologous sequences

Alpha satellite repeats annotated in the human genome assembly GRCh37/hg19 were extracted from the rmsk table in the UCSC Table Browser (https://genome.ucsc.edu/). UCSC Genome Browser (Kent et al. 2002) was used to retrieve sequences orthologous to dispersed human ARs, in all available primate genome assemblies: human (GRCh38/hg38), chimp (Clint_PTRv2/panTro6), bonobo (MPI-EVA panpan1.1/panPan2), gorilla (GSMRT3/gorGor5), orangutan (Susie_PABv2/ponAbe3), gibbon (Nleu3.0/nomLeu3), green monkey (Chlorocebus_sabeus 1.1/chlSab2),crab-eating macaque (Macaca_fascicularis_5.0/macFas5), rhesus (Mmul_10/rheMac10), baboon (Panu_3.0/papAnu4), proboscis monkey (Charlie1.0/nasLar1), golden snub-nosed monkey (Rrox_v1/rhiRox1), marmoset (WUGSC 3.2/calJac3), squirrel monkey (Broad/saiBol1), tarsier (Tarsius_syrichta-2.0.1/tarSyr2), mouse lemur (Mouse lemur/micMur2) and bushbaby (Broad/otoGar3). Human ARs annotated by RepeatMasker, with 100-200 bp flanking sequences at 5' and 3' site were used for detection of orthologous loci by UCSC Genome Browser and synteny of flanking sequences for each AR was examined to confirm orthologous loci across primate species.

Annotation of human ARs into suprachromosomal families and monomer classes was retrieved using:
https://genome.ucsc.edu/cgibin/hgTracks?db=hg38&hgt.customText=https://raw.github.com/enigene/AS-tracks/master/GRCh38-GCA_000001405.15/human-GRC-hg38-M1SFsv2.2.bed.gz

Analysis of junction regions

Analysis of the junction regions on both sides of ARs annotated by RepeatMasker programme for the presence of short segments of sequence duplication was done through systematic visual examination. We also used program MEME (Bailey, et al. 2015) to computationally detect motifs that are present at the 5' and 3' junction regions of ARs.

Statistical analyses

We used bedtools (Quinlan and Hall 2010) and custom scripts to analyse content of repetitive element classes. Tables ncbiRefSeqCurated and rmsk corresponding to genes and repetitive elements were downloaded from UCSC Genome Browser. Statistical analyses were performed in R. Random introns were chosen such that their length was between 5 kb and 300 kb, matching the range of ARs-containing introns. For analysis of repetitive classes surrounding the intergenic ARs, we extended coordinates of ARs upstream and downstream by 20 kb and intersected the intervals with rmsk data.

Phylogenetic analysis

Alignments of ARs were performed online with MAFFT version 7, using "Auto" or "E-INS-i" strategy (Katoh et al 2019; https://mafft.cbrc.jp/alignment/server/index.html). Junction regions were removed prior to alignment. Uncorrected p-distances were calculated using MEGA 7.0.25 (Kumar et al., 2016). Neighbour-joining (NJ) trees based on the p-distance model were calculated in MEGA 7.0.25 (Kumar et al., 2016), and the robustness of the clades was assessed through 1000 bootstrap replicates. Maximum likelihood (ML) trees were constructed on PhyML 3.0 web-server (Guindon et al., 2010), with automatic model selection by SMS (determined through AIC selection criterion) (Lefort et al., 2017) and aLRT SH-like support (Anisimova and Gascuel 2006). Resulting trees were edited in FigTree v.1.4.3. (http://tree.bio.ed.ac.uk/software/figtree/). To check for higher order repeat organization, we produced self-dot plots of sequences by using Flexidot (Seibt et al. 2018) and setting the word size to 25 bp.

Synteny analysis

To analyse the order of repetitive elements in the vicinity of ARs and compare it between primate species, we first extended the genomic coordinates of each AR in the human assembly: for intergenic AR elements coordinates were extended 2 kb up- and downstream and for intronic elements the coordinates of the whole intron were considered. We used these extended coordinates to find orthologous regions with the LiftOver program (Hinrichs et al. 2006). LiftOver was run on 16 existing primate chain files for assemblies other than human, that were downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/: PanTro6, PanPan2, GorGor5, ChlSab2, PonAbe3, NomLeu3, MacFas5, RheMac10, PapAnu4, NasLar1, RhiRox1, CalJac3, SaiBol1, TarSyr2,

MicMur2, OtoGar3. The resulting coordinates were intersected with coordinates of annotated repetitive elements (rmsk tables) of the respective species downloaded from the UCSC ftp site (ftp://hgdownload.soe.ucsc.edu/goldenPath/). The resulting annotated repetitive elements that overlapped given coordinates were plotted with genoPlotR package (Guy et al. 2010).

Gene ontology analysis

For gene ontology analysis the following tools were used: GOrilla (http://cbl-gorilla.cs.technion.ac.il/; Eden et al. 2009) and Panther (http://www.pantherdb.org/; Mi et al. 2019)

Results

Dispersed ARs in human genome euchromatin

Alpha satellite repeats (AR) annotated in the human genome assembly hg19 were extracted from the rmsk table from the UCSC table browser. In this way 1287 ARs were identified in the assembled chromosomes, ranging in size from 14 bp to 160 602 bp. Since it is possible that some short ARs represent false positives, only hits with at least 50% of 171 bp monomer sequence length were considered for further analyses. Most of the ARs are located within centromeric and pericentromeric regions and 1071 are organized within clusters which are composed of 2 to 52 ARs positioned within a distance less than 1 kb. In order to study the evolutionary history and mode of dispersion of ARs within euchromatin, we focused on ARs located outside of (peri)centromeric regions, preferentially on those not organized within clusters but present as single repeats (suppl. Table 1). For such ARs it is possible to detect orthologous regions within other primate genomes in order to date their insertion and to follow the dynamics throughout the evolutionary history of primates. The presence and organization of euchromatic ARs was additionally checked in the human genome assembly GRCh38.hg38, revealing them on almost all human chromosomes.

There are 32 euchromatic ARs which overlap with genes, all located in introns of 18 protein coding genes, 2 ncRNAs and one pseudogene. Most intronic ARs are short, between 0.5 mer to 4 mer size, except three ARs of several kb size within the *ANKRD30BL* gene, and are preferentially organized as single repeats except in three cases where clusters of two, three and seven repeats are found within genes *LINC01580*, *ANKRD30BL* and *SACS*, respectively (Table 1). In addition to intronic ARs, we analysed the evolutionary history of 36 intergenic ARs, 29 of them organized as single repeats while the others form clusters composed of a few adjacent repeats (Table 1; suppl. Table 1). By examination of clustered ARs we expected to reveal if the ARs within the same cluster share a

common evolutionary history, meaning that they were dispersed simultaneously, possibly together within a single insertion event, or separately during different evolutionary periods. The genes containing intronic ARs as well as genes most proximal to intergenic alpha repeats and their distances relative to ARs which range from 2.3-949 kb are listed in Table 1. Within this set of genes, gene ontology analysis revealed no significantly enriched pathways or molecular functions

Orthologous dispersed ARs in primate species - dating of insertions

We searched for sequence orthologs to all 32 intronic and 36 intergenic dispersed human ARs across the primate genomes in order to trace when their dispersion and insertion in the euchromatin occurred during the evolutionary history. The oldest insertions were traced back 45-60 Myr ago in the ancestors of Simiformes which include Platyrrhini (New World monkeys) and Catarrhini (Old World monkeys and hominoids) (Figure 1A), and are characteristic for 11 single intergenic ARs as well as for 6 single and two clustered intronic ARs (Table 1). The ARs inserted in the ancestors of Simiformes can be detected in almost all available genomes of the descendant species belonging to the branches of Platyrrhini and Catarrhini, revealing that after insertion, ARs are transmitted and preserved in the descendant species. Some AR exhibit copy number variation such as sat_83 which is in the form of 2.3-2.4 mer in Homininae, gibbon, rhesus and new world monkeys, while in baboon, green monkey, proboscis and golden snub-nosed monkey it is found as 3.4 mer and in orangutan as 1.4 mer (Figure 1a). The copy number variation (CNV) of sat_83 can be explained by a single intrastrand recombination event occurring between 2 homologous regions within AR monomers (suppl. Figure 1a; Figure 1b). There is no correlation between the size of sat_83 and species phylogeny suggesting that CNV results from the process of recombination which occurred several times independently during the evolutionary history, but at the same specific site of the sequence. According to species phylogeny, a recombination event which resulted in CNV probably occurred in the ancestor of Platyrrhini (3.4 to 2.4 mer), in the ancestor of Hominoidea (3.4 to 2.4 mer), after separation of orangutan from Homininae (2.4 to 1.4 mer) and in the ancestor of the rhesus-macaque group of old world monkeys (3.4 to 2.4 mer) (Figure 1a).

The next round of AR insertions can be traced 30-45 Myr ago in the ancestors of Catarrhini and includes 9 intronic and 2 intergenic ARs (Table 1). It is interesting that the intergenic sat_706 is detected as a 1.2 mer in old world monkeys (Cercopithecidae), while in Hominoidea it is 2.2 -mer in gibbon, 4.4-mer in orangutan, 3.4 -mer in gorilla, bonobo, chimp and human (Figure 1b). Alignment of sat_706 sequences in different primates indicates that copy number variation could result from intrastrand recombination facilitated by homologous sequence motifs in multimers (suppl. Figure 1b; Figure 1b). The origin of sat_706 in Hominoidea can be explained by a single intrastrand recombination occurring between 2 homologous regions within monomers, while the genesis of a 1.2

mer in old world monkeys is proposed to be due to an additional recombination event (Figure 1b). According to the phylogeny, recombination occurred several times through evolutionary history: in the common ancestor of old world monkeys (4.4 to 1.2 mer), in the ancestor of Hylobatidae (4.4 to 2.2 mer) and in the ancestor of Homininae (4.4 to 3.4 mer) (Figure 1a).

Another round of insertion is dated 20-30 Myr ago in the ancestors of Hominoidea (Hominidae, Hylobatidae) and is characteristic for 7 clustered ARs within an intron of the *SACS* gene as well as for a single intronic and four intergenic ARs (Table 1; Figure 1a). The AR sat_378 exhibits copy number variation, in gibbon (Hylobatidae) it is a 0.9 mer while in Hominidae it is a 1.8 mer (orangutan, human, chimp) except in gorilla and bonobo where a larger region encompassing sat_378 is deleted. The excision of the monomer in gibbon can be explained by a single internal recombination occurring between 2 homologous regions.

Insertions in the ancestors of Hominidae (orangutan, gorilla, pan and human) 18-20 Myr ago are characteristic for 4 intergenic ARs (Table 1). The intergenic sat_372 shows copy number variation: it is 5.9 mer in human, 3.9 mer in bonobo and chimp, 4.9 mer in gorilla and 10.9 mer in orangutan, and can be explained by a single intrastrand recombination event which occurred independently several times, in the lineages of gorilla (10.9 to 4.9 mer), chimp/bonobo (10.9 to 3.9 mer) and human (10.9 to 5.9 mer) (Figure 1a). Three intergenic clusters examined, composed of 2 and 3 ARs respectively, were also inserted in the ancestors of Hominidae (Table 1) and remained preserved in all species except chimpanzee where clustered sat_726 - sat_727 was removed as a part of larger deletion.

Insertions occurring 10-18 Myr ago in the ancestors of Homininae (gorilla, pan, human) include 4 intronic as well as 6 intergenic ARs (Table 1). ARs sat_1145 and sat_1123 show copy number variation: 4.1 mer in human and 2.9 mer in gorilla, while in chimp the large region encompassing these ARs is deleted (Table 1). The insertions occurring in the ancestors of Hominini (pan, human) approx. 7-10 Myr ago are two intergenic ARs, while the most recent insertion is of three clustered ARs within intron of the *ANKRD30BL* gene occurring in the ancestor of Homo (Table 1; Figure 1a).

The analysis of sequences orthologous to dispersed euchromatic human ARs across the primate genomes revealed their continuous insertion throughout evolutionary history of primates starting in the ancestors of Simiformes (45-60 Mys ago) and continuing up to the ancestors of Homo. Once inserted, the ARs were preserved and transferred to the descendant species, while some of them exhibited copy number variation due to recombination occurring independently in different evolutionary periods.

Phylogenetic analysis and sequence evolution of dispersed alpha satellite repeats

Phylogenetic analysis divides human monomeric alpha satellite repeats into different age groups, reflecting the evolution of centromeric alpha satellite DNA sequences in primates (Schueler and Sullivan 2006) which proceeds through proximal expansion of central active centromeric regions and moving of the previous centromeric DNA distally onto each arm (Schueler et al. 2005). According to that model, human alpha repeats proximal to the euchromatin chromosome arms are remnants of the ancestral primate centromere. Based on sequence features, human alpha satellite repeats are classified into 12 distinct monomer types which belong to 5 suprachromosomal families, SF1-5, and their annotation to classes and suprachromosomal families is available (Shepelev et al. 2015).

The analysis of human ARs dispersed in euchromatin shows that half of them are annotated as specific monomer types characteristic for particular suprachromosomal families (Table 1). The annotated euchromatic ARs preferentially belong to SF4+ (monomer type M1+) and SF5 (monomer types R1 and R2) suprachromosomal families which are evolutionary old and constitute pericentromeric regions. Only a few ARs contain monomers of evolutionary new suprachromosomal families which are characteristic of active centromeres, such as sat_84 which belongs to the SF1 (monomer type J1A), while sat_1145 and sat_1123 as well as few kb size AR in intron of *ANKRD30BL* (sat_616-618) contain monomers D1 and D2 of the SF2. The ARs belonging to the new suprachromosomal families were inserted within the last 18 Myr (Table 1).

In order to see if there is any sequence clustering of single dispersed human ARs that could indicate homogenization at the level of chromosome, neighbouring or tandemly arranged repeats, sequence alignment and phylogenetic analyses were performed. Since almost half of the ARs are partial monomers, the alignment was performed on those repeats that mutually overlap in the sequences, while the others were excluded from the analysis. Finally, 77 partial or full sized monomers which derive from 57 dispersed human ARs were aligned (suppl. File1a). Phylogenetic analysis performed by the neighbour joining (NJ) and maximum likelihood (ML) methods resulted in a phylogenetic tree with generally very weak resolution (suppl. Figure 2). Only a few clusters composed of two ARs such as neighbouring sat_380 and sat_381 as well as sat_1145 and sat_1123 are well supported (bootstrap values >0.7). Tandemly arranged monomers within sat_372 (5.9 mer) and sat_706 (3.4 mer) as well as clustered sat_360-366 are partially grouped in the phylogenetic tree, although the groups are not significantly supported. The average sequence divergence (p-distance) of human dispersed alpha satellite monomers is 0.3. We additionally checked the phylogeny of alpha monomers within 3 long clustered ARs (sat_614-618) in the intron of the *ANKRD30BL* gene, which were inserted relatively recently, in the ancestor of Homo. All three phylogenetic trees are characterized by weak resolution with very few supported groups of 2-3 monomers (suppl. Figure 3a; suppl. File1b-d). Although multiple monomers within sat_616-618 belong to the SF2 subfamily which is characteristic of active centromeres, the phylogenetic tree gave no indication for centromere specific higher order organization of monomers and the self-dot plot confirmed this (suppl. Figure 3b). The

average sequence divergence of monomers within long clustered ARs in the *ANKRD30BL* gene is 0.20.

The sequence evolution of single ARs dispersed within euchromatin was followed in different primate species. Six ARs in introns (sat_828, sat_1147, sat_705, sat_685 and two adjacent ARs sat_380-381) and five in intergenic regions (sat_621, sat_704, sat_1, sat_703, sat_722) were examined, all inserted in the ancestors of Simiformes and are therefore widely spread among extant species. In addition, the examined ARs are longer than a 1 mer and do not exhibit copy number variation among primate species. Average sequence divergences (uncorrected p-distances) between respective ARs of different primate species are generally low and range from 0.0468 in sat_1147 to 0.0835 in sat_685. NJ and ML phylogenetic trees reveal evolution of ARs' sequences which generally supports the primate species phylogeny, in particular, major clades of new and old world monkeys as well of hominoids are supported by high confidence, in addition to most of the nodes within clades (Figure 2, Suppl. Figure 4; suppl. File 1e-n). The results reveal that dispersed ARs' sequences are not subjected to abrupt changes and rearrangements except copy number variation as previously described, but evolve gradually, reflecting the evolution of primate species.

Association of ARs with other repetitive elements and synteny analysis

Over 80% of single alpha satellite repeats within human euchromatin have in the vicinity, within 100 bp distance, other repetitive DNA elements (Table 1). To see if there is any preference for specific repetitive families to be nearby ARs we analysed association of all 1287 annotated ARs with annotated TEs. Out of 618 TEs found immediately next to ARs, over half (329) belong to the L1 repeat family and 32% (197) are *Alu* elements, while the majority of the rest are ERV elements (~15%). Only 25 ARs actually overlap other TEs, the majority intersecting ERV elements. To test if ARs are inserted in regions enriched with specific repetitive elements what might also suggest that such elements could mediate ARs dispersion, we analysed repeat composition of 22 introns containing ARs and compared it with the composition of 100 randomly chosen introns without ARs (Figure 3a, b). There is 3.6x higher variance of simple repeats proportions in random introns (p-value = 0.005; F test) compared to ARs-containing introns and the only repetitive class “DNA” (DNA transposons) is significantly underrepresented in introns with ARs (p-value = 0.0038; Welch Two Sample t-test), comprising on average about 8% of repetitive elements compared to about 13% in randomly chosen introns. Overall, the proportion of repetitive element classes seems to be less variable within the 22 introns compared to arbitrarily chosen introns. We also analysed the proportion of repetitive elements 20 kb around intergenic ARs listed in Table 1 (Figure 3c) and found a substantially smaller proportion of SINE elements (26%) compared to ARs-containing introns (40%; p-value = 0.0011; Welch Two

Sample t-test). The average proportion of repeats in introns with ARs, randomly chosen introns or in intergenic regions around ARs is not significantly different from the average proportion of repetitive elements in the human genome overall (Figure 3d). The results indicate that preferential location of ARs near SINE, LINE and ERV elements is more likely due to the general abundance of such repeat elements in introns and intergenic regions rather than to the specific association of ARs with these elements.

To study how the organization of repetitive elements within regions proximal to dispersed ARs evolves among primates, we analysed the distribution of repetitive families within approximately 2 kb at the 5' and 3' of intergenic ARs, as well as within ARs-containing introns. The analysis showed a significant conservation of synteny (order of repeats) within intergenic and intronic regions around many dispersed ARs between human and other primates despite periods of divergent evolution of up to 45-60 Myr and the high abundance of different repetitive families (Figure 4, suppl. Figure 5). The regions proximal to most of dispersed ARs are not prone to the rearrangements and the observed stable transmission of ARs among species is in accordance with the preserved synteny of their neighbouring repeats.

Mechanism of ARs dispersion within euchromatin

Analysis of junctions on both sides of single dispersed euchromatic alpha repeats reveal that short segments of sequence duplication of 2-10 bp size occur at most of the insertion sites (Table 1). The sequence duplication is detected irrespective of the alpha repeat insertion within different context: within other repeats, adjacent to repeats or in the unique regions without repeats within at least 100 bp (Table 1). We propose that regions of short homology of 2-10 bp between an alpha repeat and it's target site facilitated insertion of ARs by homologous recombination (Figure 5). The junction sequences which are duplicated are mutually different and also reflect a difference in the segments or parts of alpha satellite monomers which are inserted at the particular site (Table 1). ARs embedded within other repetitive elements are not inserted within the same position in the repeats of the same type: e.g. sat_605 and sat_497 are inserted at positions 169 and 73 of *Alu* repeat, respectively, while sat_84, sat_612, sat_1147 and sat_703 are inserted at positions 3051, 2147, 5618 and 5900 of L1 repeat, respectively (Table 1). Therefore, each AR insertion seems to be unique, characterized by a specific part of the AR monomer(s) inserted at a specific genomic region.

Since some of ARs are imbedded in other repetitive elements it is possible that these elements are responsible for the dispersion of ARs. To check if ARs and the associated repeats are transferred together to the particular genomic location we analysed the orthologous regions in those primates separated before the insertion of the particular AR occurred (Table 2). For the sat_605 and sat_497

embedded within the *Alu* sequence as well as for sat_60 inserted within ERV and sat_612 in L1, respectively, in the ancestor of Catarrhini, we analysed the orthologous regions in the species of Platyrrhini: marmoset and squirrel monkey. The analyses revealed the presence of *Alu* sequences at sat_605 and sat_497 orthologous loci in new world monkeys but without inserted ARs (Table 2). The sat_60 and sat_612 orthologous loci also contained ERV and L1, respectively in new world monkeys but without ARs insertion. For sat_827 inserted into the Tigger element as well as sat_703 and sat_1147 in L1, respectively, in the ancestors of Simiformes we examined orthologous loci in tarsier (Tarsiidae) and in mouse lemur (Lemuridae). At the sat_827 orthologous locus only the Tigger element was present without sat_827 insertion in both species. At the locus orthologous to sat_703 in the mouse lemur, L1 element without sat_703 is present while in tarsier the large region encompassing orthologous sat_703 locus is absent. The regions encompassing sat_1147 orthologous loci are also absent in both species (Table 2). For sat_58 and sat_358 embedded in ERV in the ancestor of Hominoidea, examination in the old world monkeys (rhesus, baboon) revealed an intact orthologous ERV without an AR insert. Finally, for sat_84 and sat_613 embedded into L1 and ERV in the ancestors of Hominini, examination of orthologous locus in gorilla and orangutan revealed L1 and ERV element without inserted AR. All these observations show that *Alu*, L1, Tigger and ERV elements, respectively were already present at the sites of new AR insertion (Table 2). These suggest that ARs were subsequently inserted within other repetitive elements, probably by the same mechanism used for ARs insertion in unique regions without repeats, or between different repeats.

Once ARs were inserted within other repeats it is possible that such hybrid elements are further spread along the genome by (retro)transposition or by recombination facilitated by larger segments of homology between abundant dispersed repeats such as *Alu*, L1 or ERVs. Most transposons in the human genome have been inactive for the last 500 million years (Lander et al. 2001), but some retrotransposable elements including the prevalent LINE and SINE repeats which dominate the human genome are still active (Mills et al. 2007). However, our search of the human genome assembly for elements homologous to hybrid AR-(retro)transposons gave negative results, suggesting the absence of their spreading to new loci. The same is true for ARs inserted near other repeat elements. Each of the AR insertion is flanked by a specific combination of other repeats or repeat subfamilies and is present at a single genomic locus, as revealed by BLAT search. The only exceptions are sat_1145 and sat_1123 which are highly similar 4.1 mers flanked by MamGypLTR3a and L1PA3, and are present at 2 different genomic loci as a result of duplication of a long chromosomal region of approx. 190 kb within chromosome 9 and not of the specific spreading of the particular AR. Using UCSC browser we searched if segmental duplications annotated in assembled human genome intersect other dispersed ARs. The results reveal that apart from sat_1145 and sat_1123, intergenic sat_372 is associated with the duplication of approx. 280 kb long segment on chromosome 14. Therefore, the results indicate that most AR insertions are unique and present at a

single genomic locus while no evidence of subsequent spreading of inserted ARs across the genome exists.

Considering the dispersion of clustered ARs within euchromatin, sat_614-618 clustered within intron of *ANKRD30BL* gene corresponds to the site of a relic, previously active centromere (Miga 2017). Examination of other clustered ARs within introns of the genes *SACS* and *LINC01580* and of 3 intergenic clusters revealed that all ARs within a cluster seem to be inserted within the same period, suggesting that the cluster could result from a single insertion event. Within clusters, ARs are interspersed with other repetitive elements such as *Alu* in the case of *SACS* intronic cluster or with L1 and ERV-LTRs in intergenic sat_723-727 clusters, respectively (suppl. Figure 6). Regions composed of different types of mutually interspersed repeats are characteristic for pericentromeric heterochromatin and it is possible that they could be transferred from heterochromatin to euchromatin as larger blocks in the same way as shorter single AR repeats, probably by a recombination event relying on short sequence homology. The existence of extrachromosomal circular DNAs (eccDNA) in human cells composed of ARs ranging in size from less than 2 kb to over 20 kb was previously shown (Cohen et al. 2010), revealing that tandemly arranged alpha repeats are prone to generate eccDNA. We suggest the possible involvement of eccDNA in the dispersion of ARs within euchromatin and propose that their insertions within the genome were facilitated by short sequence homology between ARs and their target sequences (Figure 5).

Discussion

Our analysis of alpha satellite repeats annotated in the human genome reveals that majority of them, over 80%, are organized within clusters composed of AR located within a distance less than 1 kb, and such clusters are preferentially located in the pericentromeric and centromeric regions. We focused on ARs not organized within clusters but present as single repeats within euchromatin and followed their evolutionary history among primates. Continuous insertion of ARs from the ancestors of Simiformes (45-60 Myr ago) up to the lineage of Homo is detected. In the available genomes of prosimians, tarsier (*Tarsius syrichta*), mouse lemur (*Microcebus murinus*) and bushbaby (*Otolemur garnettii*) we did not find alpha repeats within regions orthologous to those containing dispersed human AR, suggesting that spreading preferentially started in the ancestors of simians. Alternatively, chromosomal rearrangements may have resulted in deletion of such regions. It is important to mention that the quality of assembly of all primate genomes is not the same and that some of them could contain the unassembled stretches of alpha satellite repeats which can affect our analysis. Long-read sequencing technology recently used for great ape genome assembly (Kronenberg et al. 2018) is expected to improve future studies of satellite DNA evolution.

Based on sequence features, alpha satellite monomers are classified into distinct groups and families (Shelepev et al. 2015). According to such classification, many dispersed alpha repeats belong to evolutionary old alpha satellite families which constitute pericentromeric regions, indicating dispersion of repeats from heterochromatin to euchromatin. The high sequence divergence of dispersed ARs also suggest their origin from the (peri)centromeric alpha repeats which form distinct subfamilies characterized by specific, mutually divergent monomers (Willard 1985; Vissel and Choo 1992). Many extant human single dispersed ARs are up to the monomer size while longer, multimeric ARs often exhibit copy number variation between primate species which can be explained by intrastrand recombination between homologous sequence motifs in the monomers. The recombination process seems to occur randomly and independently in different lineages during different evolutionary periods. Some single ARs composed of multimers such as sat_358 (2.7 mer), sat_685 (1.9 mer), sat_621 (1.8 mer) have remained stable since their insertion in the ancestor of Hominoide, Simiformes and Catarrhini, respectively. However, it is possible that these ARs show copy number variation in some other primate species which were not examined in our study due to the lack of genome assembly data. Excluding copy number variation, single ARs inserted within euchromatin are stably transmitted to descendant species showing gradual sequence evolution which generally follows species evolution. In addition, the organization of repeats in the regions proximal to dispersed alpha repeats seems to be conserved among simian primates. When compared to X chromosome euchromatic satellites in *Drosophila* species which exhibit a high rate of rearrangements and reorganization resulting in change of abundance, location, and composition (Sproul et al. 2020), euchromatic ARs in primate species show greater stability and lower evolutionary dynamics.

Single repeats of a major alpha satellite DNA dispersed within euchromatin are often found adjacent to other abundant repetitive elements such as *Alu*, L1 or ERV, or are embedded within them. Some DNA transposons such as those belonging to the *Helitron* superfamily have a propensity to capture and mobilize flanking DNA sequences (Thomas et al. 2014) and based on such characteristics it was proposed that satellite arrays flanked by *Helitron* transposons can be spread throughout the genome by the process of transposition (Brajković et al. 2012; Šatović et al. 2016). In this study we did not find evidence for (retro)transposable elements to be specifically associated with ARs. In addition, there is no indication that (retro)transposons play a role in spreading of adjacent or embedded alpha repeats throughout the euchromatin, either by recombination with other members of the same family, or by retrotransposition. We also show that segmental duplications within human genome can be associated with dispersion of only a few ARs while *ANKRD30BL* intronic alpha repeat array corresponds to the previously mapped relic chimpanzee centromere (Miga 2017). The most probable mechanism of alpha repeats spreading that we propose is based on extrachromosomal circles of alpha satellite DNA which can be created by intrachromatid recombination of alpha repeats within heterochromatin (Feliciello et al. 2006; 2015a). The presence of short duplications at most AR insertion sites support this model of satellite repeats proliferation. Extrachromosomal circular satellite

DNAs are common across diverse eukaryotic organisms including insects, plants and mammals (Cohen et al. 2006; Navratilova et al. 2008; Cohen and Segal 2009; Paulsen et al. 2018, Sproul et al. 2020), including human cell lines (Cohen et al. 2010). Extrachromosomal satellite DNA circles are proposed to be amplified by rolling circle replication and reintegrated within the genome by a random process of site specific recombination which occurs between short sequence motifs within circularized satellite repeats and homologous motifs at different chromosomal sites, either within euchromatin or heterochromatin.

In conclusion, our research reveals spreading of alpha repeats in the human genome euchromatin which occurred gradually throughout evolutionary history of primates, and discloses the most probable mechanism of alpha repeats proliferation. Considering the influence of dispersed satellite repeats on local chromatin structure and gene expression in different organisms (Feliciello et al. 2015b; Joshi and Meller 2017; Halbach et al. 2020; Feliciello et al. 2020), studies of their evolutionary dynamics and mechanisms of proliferation within euchromatin could contribute to the explanation of their potential role in the evolution of gene regulatory networks.

Acknowledgements: This work was supported by Croatian Science Foundation grant IP-2019-04-6915, by Slovenian-Croatian Bilateral Cooperation and by the Italian Ministry of Education, University and Research (MIUR), fund for Investments on Basic Research (FIRB) and the International Staff Mobility Program of University of Naples Federico II.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability Statements: The data underlying this article are available in the article and in its online supplementary material.

Literature:

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. 2001. Alpha-satellite DNA of primates: old and new families. Chromosoma 110:253-66.

Anisimova M, Gascuel O. 2006. Approximate likelihood ratio test for branchs: A fast, accurate and powerful alternative. System Biol. 55: 539-552

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. Nucleic Acids Res. 43:W39-49.

Brajković J, Feliciello I, Bruvo-Mađarić B, Ugarković Đ. 2012. Satellite DNA-like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. G3: Genes Genomes Genet. 2:931-941.

Brajković J et al. 2018. Dispersion profiles and gene associations of repetitive DNAs in the euchromatin of the beetle *Tribolium castaneum*. G3: Genes, Genomes, Genet. 8: 875-886.

Bulut-Karslioglu A et al. 2012. A transcription factor-based mechanism for mouse heterochromatin formation. Nat Struct Mol Biol. 2012 19:1023-30.

Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, Escudé C. 2016. Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey*: Cercopithecus solatus.* BMC Genomics 17:916.

Cohen S, Segal D. 2009. Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. Cytogenet. Genome Res. 124:327-338.

Cohen S, Agmon N, Sobol O, Segal D. 2010. Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. Mob. DNA 8:11.

Cohen Z, Bacharach E, Lavi S. 2006. Mouse major satellite DNA is prone to eccDNA formation via DNA Ligase IV-dependent pathway. Oncogene 25:4515-4524.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: A Tool For Discovery And Visualization of Enriched GO Terms in Ranked Gene Lists. BMC Bioinformatics 10:48.

Feliciello I, Picariell, O, Chinali G. 2006. Intra-specific variability and unusual organization of the repetitive units in a satellite DNA from *Rana dalmatina*: molecular evidence of a new mechanism of DNA repair acting on satellite DNA. Gene 383: 81-92.

Feliciello I, Akrap I, Brajković J, Zlatar I, Ugarković Đ. 2015a. Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. Genome Biol. Evol. 7: 228-239.

Feliciello I, Akrap I, Ugarković, Đ. 2015b. Satellite DNA Modulates Gene Expression in the Beetle *Tribolium castaneum* after Heat Stress. PLoS Genet. 11: e1005466.

Feliciello I, Sermek A, Pezer Ž, Matulić M, Ugarković Đ. 2020. Heat stress affects H3K9me3 level at human alpha satellite DNA repeats. Genes 11: 663.

Finstermeier K. 2013. A mitogenomic phylogeny of living primates. PLoS One 8:e69504.

Guindon S et al. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. System Biol. 59:307–21.

Guy L, Kultima JR, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26:2334-2335

Halbach R et al. 2020. A satellite repeat-derived piRNA controls embryonic development of *Aedes*. Nature 580: 274–277.

Hinrichs AS et al. 2006. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34:D590-8.

Joshi SS, Meller VH. 2017. Satellite Repeats Identify X Chromatin for Dosage Compensation in *Drosophila melanogaster* Males. Curr Biol. 27:1393-1402.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 20:1160-1166

Kent WJ et al. 2002. The human genome browser at UCSC. Genome Res. 12:996-1006.

Kronenberg ZN et al. 2018. High-resolution comparative analysis of great ape genomes. Science 360:eaar6343.

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

Kumar S, Stecher G, Tamura K. 2016: MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 33:1870–1874.

Lander ES et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC. 1997. Human centromeric DNAs. Hum. Genet. 100: 291–304.

Lee YCG et al. 2020. Pericentromeric heterochromatin is hierarchically organized and spatially contacts H3K9me2 islands in euchromatin. PLoS Genet. 16:e1008673.

Lefort V, Longueville JE, Gascuel O. 2017: SMS: Smart Model Selection in PhyML. Mol Biol Evol. 34: 2422–2424.

McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. Chromosome Res. 26:115-138.

Menon DU, Coarfa C, Xiao W, Gunaratne PH, Meller VH. 2014. siRNAs from an Xlinked satellite repeat promote X-chromosome recognition in Drosophila melanogaster. Proc Natl Acad Sci U S A 111: 16460–16465.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 47:D419-D426.

Miga KH. 2017. Chromosome-Specific Centromere Sequences Provide an Estimate of the Ancestral Chromosome 2 Fusion Event in Hominin Genomes. J Hered. 108 :45-52.

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? Trends Genet. 23: 183–191.

Navratilova A, Koblizkova A, Macas J. 2008. Survey of extrachromosomal circular DNA derived from plant satellite repeats. BMC Plant Biol. 8:90.

Paulsen T, Kumar P, Koseoglu MM, Dutta A. 2018. Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. Trends Genet. 34:270-278.

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

Pita S et al. 2017. Comparative repeatome analysis on *Triatoma infestans* Andean and Non-Andean lineages, main vector of Chagas disease. PLoS One 12:e0181635.

Pozzi L et al. 2014. Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. Mol Phylogenet Evol. 75:165-83.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842

Rudd MK, Willard HF. 2004. Analysis of the centromeric regions of the human genome assembly. Trends Genet. 20:529-33.

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JP. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep. 6: 28333.

Schueler MG et al. 2005. Progressive proximal expansion of the primate X chromosome centromere. Proc Natl Acad Sci U S A 102:10563-8.

Schueler MG, Sullivan BA. 2006. Structural and functional dynamics of human centromeric chromatin. Annu. Rev. Genomics Hum. Genet. 7:301-13

Seibt KM, Schmidt T, Heitkam T. 2018. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. Bioinformatics 34:3575-7.

Shepelev VA et al. 2015. Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. Genom Data 5:139-146.

Sproul JS et al. 2020. Dynamic evolution of euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the simulans clade. Mol Biol Evol. doi: 10.1093/molbev/msaa078

Šatović E, Vojvoda Zeljko T, Luchetti A, Mantovani B, Plohl M. 2016. Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. BMC Genomics 17:997.

Thomas J, Phillips CD, Baker RJ, Pritham EJ. 2014. Rolling-circle transposons catalyze genomic innovation in a Mammalian lineage. Genome Biol. Evol. 6:2595–2610.

Vissel B, Choo KHA. 1992. Evolutionary relationships of multiple alpha satellite subfamilies in the centromeres of human chromosomes 13, 14, and 21. J. Mol. Evol. 35:137–146.

Vlahovic I, Gluncic M, Rosandic M, Ugarkovic Đ, Paar V. 2017. Regular Higher Order Repeat Structures in Beetle *Tribolium castaneum* Genome. Genome Biol Evol. 9:2668-2680.

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

Willard HF. 1991. Evolution of alpha satellite. Curr Opin Genet Dev. 1:509-14

Legends to Figures:

Figure 1. a) Number of single human ARs inserted within euchromatin during evolutionary history is indicated in red on the phylogenetic tree of simians (Simiformes). Species for which genome sequence is available are indicated: human (*Homo sapiens*), chimp (*Pan troglodytes*), bonobo (*Pan paniscus),* gorilla (*Gorilla gorilla gorilla),* orangutan (*Pongo pygmaeus abelii),* gibbon (*Nomascus leucogenys),* rhesus (*Macaca mulatta*), crab-eating macaque (*Macaca fascicularis),* baboon (*Papio anubis),* green monkey (*Chlorocebus sabaeus),* golden snub-nosed monkey (*Rhinopithecus roxellana),* proboscis monkey (*Nasalis larvatus)*, marmoset (*Callithrix jacchus)* and squirrel monkey (*Saimiri boliviensis*). For alpha repeats sat_828, sat_706 and sat_372 which show copy number variation, size of repeats in each species is indicated. **b**) Schematic representation of intrastrand recombination responsible for copy number variation of alpha repeats sat_828 and sat_706. Homologous regions in AR sequences at which recombination occurred are indicated by different signs. In sat_828 there is a single recombination site while in sat_706 there are two sites.

Figure 2. ML trees based on human dispersed ARs, intronic sat_828 and sat_380-381 as well as intergenic sat_704 and sat_703 and their orthologous sequences in different primate assemblies: hg38-human, panTro6-chimp, panPan2-bonobo, gorGor5-gorilla, panAbe3-orangutan, nomLeu3-gibbon, chlSab2-green monkey, macFas5-crab-eating macaque, rheMac-rhesus, papAnu4-baboon, rhiRox1- golden snub-nosed monkey, nasLar1- proboscis monkey, calJac3-marmoset, saiBol1-squirrel monkey. Numbers on the nodes depict ML aLRT / NJ bootstrap support values.

Figure 3. Repeat composition in 22 introns containing ARs (**a**), 100 randomly chosen introns without ARs (**b**) in 20 kb region around intergenic ARs (**c**) and average proportion of repeats in: genome, introns with ARS, 20 kb region around intergenic ARs and randomly chosen introns (**d**). The value of *n* in parentheses denotes the total number of repetitive elements within analysed region.

Figure 4. Organization of repetitive families in the vicinity of intergenic ARs sat_827 and sat_1122 as well as intronic sat_605, in different primate species indicated in the phylogenetic tree. ARs are shown in red and other repetitive families are marked with different colours (see legend). Sat_827 and sat_605 are embedded in Tigger DNA transposon and *Alu* (SINE), respectively, while sat_1122 is inserted within unique region.

Figure 5. Model of the generation of dispersed alpha satellite repeats in euchromatin. The model postulates that alpha satellite repeats are, due to intrastrand homologous recombination, excised from the tandemly arranged heterochromatic repeats in the form of extrachromosomal circular satellite DNA. Short segments of homology, indicated in yellow, between circularized alpha repeats and target regions in euchromatin are necessary for their insertion by site specific homologous recombination. Once inserted, alpha repeats are not further spread throughout euchromatin.

Table 1. List of euchromatic human alpha satellite repeats (ARs) used in this study. Size is expressed as number of monomers, monomer types (if available, according to Shepelev et al. 2015), position relative to the genes and distances to the nearest genes are shown: negative distances mean 5' position of a gene to the AR. The association with other repetitive elements at the insertion site and dating of the insertion is listed as well as target site (TS) duplicated sequences at the insertion sites. For ARs inserted in other repetitive elements, position of insertion within the particular element is shown in brackets. ARs organized in clusters are shaded and ARs embedded in other repeats are indicated in bold.

AR No	size	monomer type	position (associated gene), distance (bp)	Insertion site characteristics	dating of insertion	TS duplicated sequence
sat_1	1,3	--	intergenic (*ADGRL2*), -240820	adjacent to L1PA10	Simiformes	CTT
sat_83	2,4	M1+, X	intron (*PLA2G12B*)	btw L1MB8 and L2c	Simiformes	AGGAGT
sat_368	0,8	--	intergenic (*MYCBP2*), -73091	btw AluSz6 and L1ME1	Simiformes	--
sat_380	1	R1	intron (*LINC0158*0), ncRNA	adjacent to MIR (SINE)	Simiformes	GAA
sat_381	0,9	R1	intron (*LINC01580*), ncRNA	adjacent to MIR (SINE)	Simiformes	GAA
sat_496	1,4	--	intron (*VAV1*)	btw AluSx1 and MLT1C (ERVL)	Simiformes	AAG
sat_623	0,5	--	intron (*PLCB4*)	no adjacent repeats	Simiformes	--
sat_703	1	Xm, X	intergenic (*ACTL6A*), 29696	**L1MC1 (5900)**	Simiformes	TGA
sat_704	1,4	--	intergenic, (*RPL392*), 15390	no adjacent repeats	Simiformes	AGTG
sat_685	1,9	--	intron (*FILIP1L*)	adjacent to L1PA11	Simiformes	AAA
sat_730	0,5	Um	intergenic (*MIR297*), -630154	btw L1PA8A and simple repeat	Simiformes	ATGAAAAAAA
sat_721	0,9	--	intergenic (*RPL21P44*), 37365	btw TAn and ATGGn	Simiformes	ATAT
sat_722	1	--	intergenic (LOC105377247), 129191	btw (ATAAT)n and FordPrefect hAT	Simiformes	GCTA
sat_825	0,7	--	intergenic (*GABRB2*), 125243	btw L1MB2 and AluSx4	Simiformes	GAAA
sat_826	0,8	--	intergenic (*C6orf106*), 7067	no adjacent repeats	Simiformes	--
sat_827	0,7	M1+	intergenic (*FAM83B*), -46144	**Tigger3 (DNA; 205)**	Simiformes	GCT
sat_828	1,7	--	intron (*PRIM2*)	adjacent to AluJr	Simiformes	GAAAAAG
sat_1122	0,5	--	intergenic (*TG*), 2283	no adjacent repeats	Simiformes	GTGA
sat_1147	1,5	M1+	intron (*MAP7*)	**L1ME3 (5618)**	Simiformes	ATC
sat_60	0,5	--	intron (*STAM*)	**MER77B (ERVL; 311)**	Catarrhini	AT
sat_85	0,6	--	intron (*LRRC4C*)	adjacent to MIR (SINE)	Catarrhini	--
sat_87	0,5	--	intron (*LRRC4C*)	adjacent to MER5A hAT-Charlie DNA transp.	Catarrhini	GAG
sat_497	0,6	M1+	intron (*ZNF675*)	**AluSc (73)**	Catarrhini	--
sat_621	1,8	M1+	intergenic, (*SLC40A1*), -33821	no adjacent repeats	Catarrhini	--
sat_605	0,7	--	intron (*SLC30A6*)	**AluSx3 (169)**	Catarrhini	AGA
sat_606	0,5	--	intron (*LINC00486*), ncRNA	no adjacent repeats	Catarrhini	CTT
sat_612	0,7	--	intron (*TMEM131*)	**L1M2c (2147)**	Catarrhini	GCT
sat_705	1,4	--	intron (*AFAP1*)	adjacent to L1PA10	Catarrhini	--
sat_373	0,9	--	intron (*ERO1A*)	no adjacent repeats	Catarrhini	GTTTT
sat_706	3,4	M1+	intergenic (*ACOX*), 23643	btw LTR18B and L2a	Catarrhini	TTA
sat_58	0,5	--	intergenic (*SUSD4*), 12805	**HERVP71A-int (ERV1; 5439)**	Hominoidea	AAC

sat_86	0,3	--	intron (*LRRC4C*)	adjacent to MER5A	Hominoidea	--
sat_358	2,7	M1+	intergenic (*CPNE8*), 59985	**MLT1D (ERVL-MaLR LTR; 216)**	Hominoidea	TCAC
sat_360	1,5	--	intron (*SACS*)	adjacent to AluYjk	Hominoidea	ACA
sat_361	0,9	--	intron (*SACS*)	adjacent to AluYjk	Hominoidea	ACA
sat_362	3,7	--	intron (*SACS*)	adjacent to AluYjk	Hominoidea	TCA
sat_363	1	--	intron (*SACS*)	adjacent to AluYjk	Hominoidea	TTGT
sat_364	1,2	--	intron (*SACS*)	adjacent to AluSp	Hominoidea	ACT
sat_365	0,8	--	intron (*SACS*)	adjacent to AluSp	Hominoidea	TAT
sat_366	0,8	--	intron (*SACS*)	adjacent to AluSp	Hominoidea	AGCT
sat_367	1,5	--	intergenic (*PARP4*), -9495	no adjacent repeats	Hominoidea	TGT
sat_729	0,7	M1+	intergenic, (*CXCL13*), 22197	no adjacent repeats	Hominoidea	TGT
sat_378	1,8	--	intergenic (*SNRPN*), 69307	no adjacent repeat	Hominidae	AAG
sat_410	0,5	--	intergenic (*LINC01566*), 36721	btw L1M3 and AluY	Hominidae	GAT
sat_652	5	M1+	intergenic (*PRAMENP*), -90051	btw AluSq2 and AluSc	Hominidae	AC
sat_653	27	M1+, Um	intergenic, (*PRAMENP*), -91217	btw AluSc and (TG)n	Hominidae	AC
sat_723-25	32	M1+	intergenic (*TECRL*), 949009	btw MSTA (ERVL) and L1PA4	Hominidae	TTG
sat_726	31	M1+	intergenic, (*TECRL*), 933498	btw L1PA4 and MER11C (ERVL)	Hominidae	--
sat_727	15	M1+	intergenic, (*TECRL*), 929843	adjacent to MER11C (ERVL)	Hominidae	TTG
sat_823	0,7	--	intergenic (*LINC02159*), 101431	btw 2 L1PA7	Hominidae	TAA
sat_732	11	M1+	intergenic (*C5orf17*), 64041	adjacent to AluY	Hominidae	AAACCTG
sat_733	26	M1+	intergenic, (*C50rf17*), 59303	btw AluY and L1PA7	Hominidae	TG
sat_864	2	M1+	intergenic (*MAFK*), 11153	btw MER21C and MLT2B3 (ERVL)	Hominidae	TTGG
sat_59	0,5	M1+	intergenic (*DIP2C*), -18185	btw 2 MSTD-int ERVL	Homininae	CTA
sat_257	0,7	--	intron (*DLG2*)	adjacent to L1PA8	Homininae	CAT
sat_372	5,9	M1+	intergenic (*LINC01296*), -19382 bp	no adjacent repeats	Homininae	ACAT
sat_379	1,2	Um	intron (*MYO1E*)	no adjacent repeats	Homininae	ACT
sat_607	0,9	R2, Xm	intergenic (*TGFA*), 15683	btw L1PA2 and Charlie8	Homininae	AAC
sat_731	0,5	M1+	intergenic (*LINC00613*), 290201	adjacent to AluSx1	Homininae	CTCCAA
sat_822	1,1	M1+	intron (*NR3C1*)	btw AluSc and Alu	Homininae	CTT
sat_1123	4,1	D2,3M1	intron, LOC101928195-pseudogene	btw MamGypLTR3a and L1PA3	Homininae	AACAG
sat_1145	4,1	D2,3M1	intergenic, (LOC101928381), ncRNA, 36431	btw MamGypLTR3a and L1PA3	Homininae	AACAG
sat_1146	0,9	--	intergenic (*PPP2R3B*), 15781	adjacent to AluJo	Homininae	TGA
sat_84	1	J1A	intergenic (*OR6A2*), 14531	**L1PA16 (3051)**	Hominini	GTA
sat_613	0,9	R1	intergenic (*DPP10*), 58435	**LTR43 (ERV; 506)**	Hominini	CAT
sat_614	67	M1, R2, X	intron (*ANKRD30BL*)	adjacent to SVA_D –E	Homo	ACACTG
sat_615	49	R1-2, M1, Um	intron (*ANKRD30BL*)	Btw SVA_D-E and L1PA3	Homo	GAA
sat_616-618	77	D1-2; R, M	intron (*ANKRD30BL*)	adjacent to L1PA3	Homo	ATAA

Table 2. List of single alpha satellite repeats located within human euchromatin which are embedded within other repetitive elements in the ancestors of the Catarrhini, Simiformes, Hominoidae and Hominini, as well as a list of repeats at orthologous regions in those primates separated before the insertion of the particular alpha repeat occurred.

Catarrhini	marmoset	squirrel monkey
sat_605-*Alu*	*Alu*	*Alu*
sat_497-*Alu*	*Alu*	*Alu*
sat_60-ERV	ERV	ERV
sat_612-L1	L1	L1

Simiformes	tarsier	mouse lemur
sat_827-Tigger	Tigger	Tigger
sat_703-L1	_	L1
sat_1147-L1	_	_

Hominoidae	rhesus	baboon
sat_58-ERV	ERV	ERV
sat_358-ERV	ERV	ERV

Hominini	gorilla	orangutan
sat_84-L1	L1	L1
sat_613-ERV	ERV	ERV

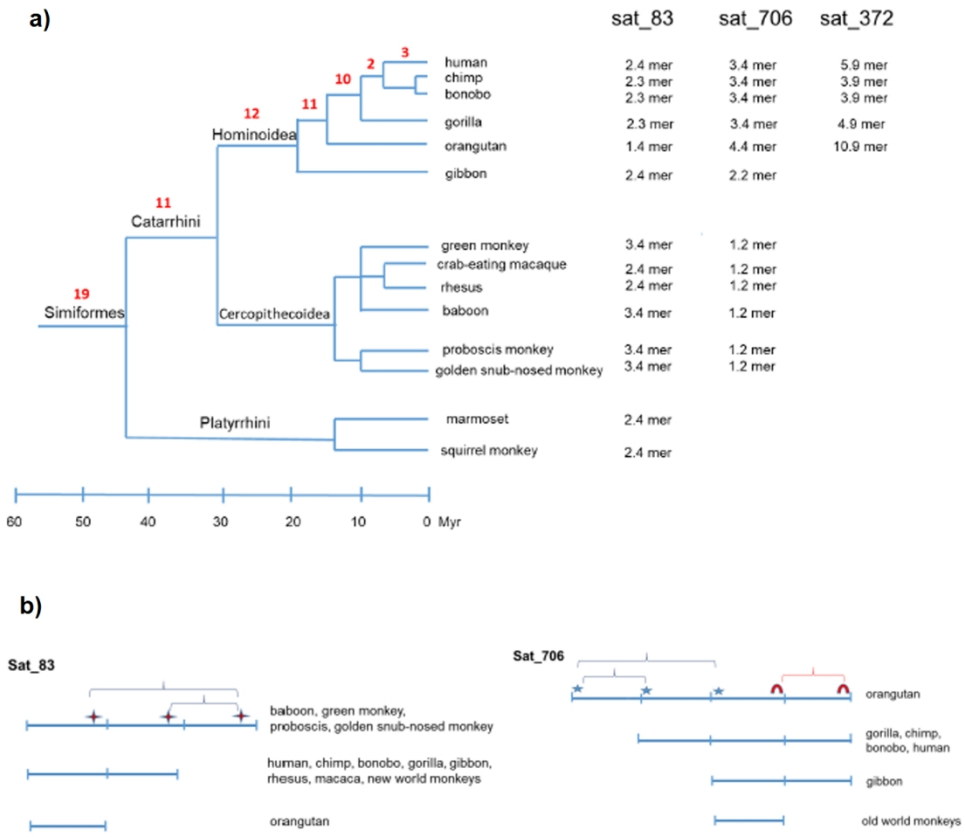

Figure 1

130x116mm (300 x 300 DPI)

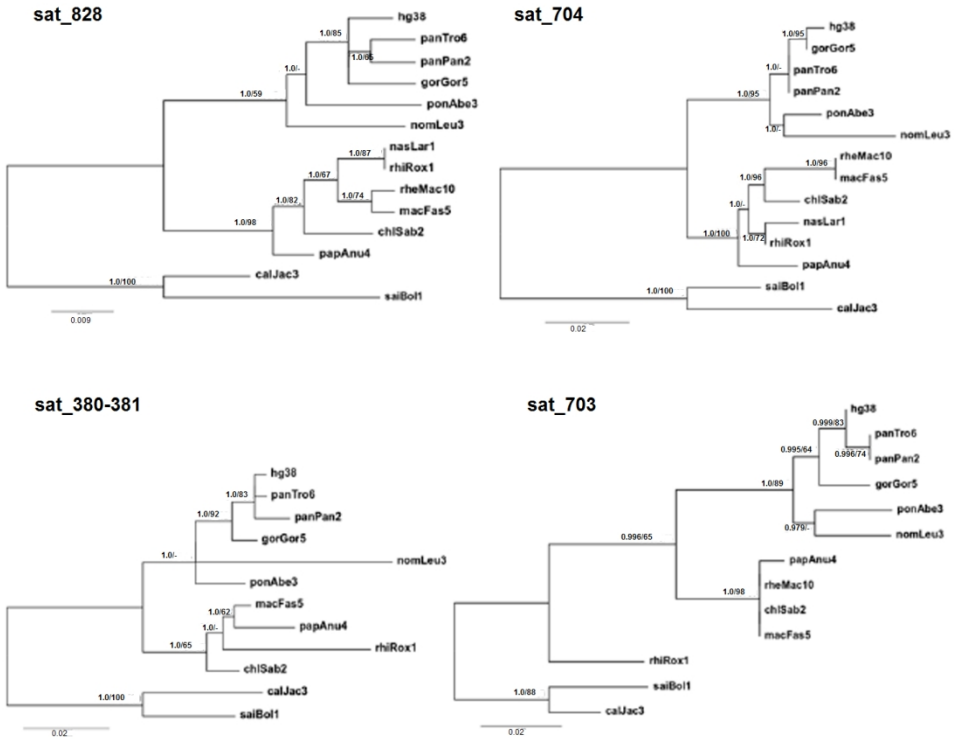

Figure 3

129x102mm (300 x 300 DPI)

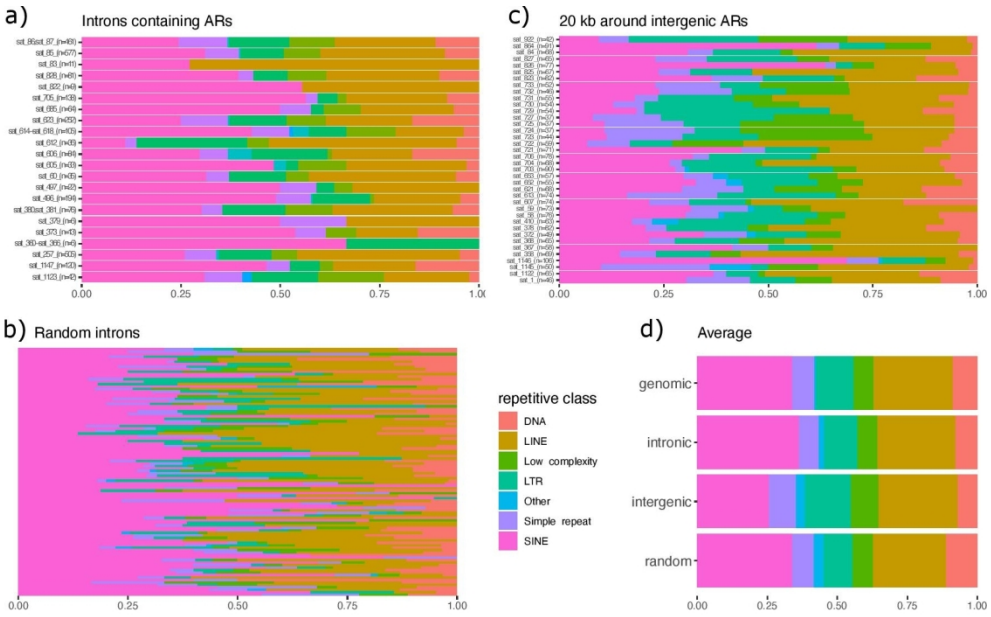

Figure 3

152x95mm (300 x 300 DPI)

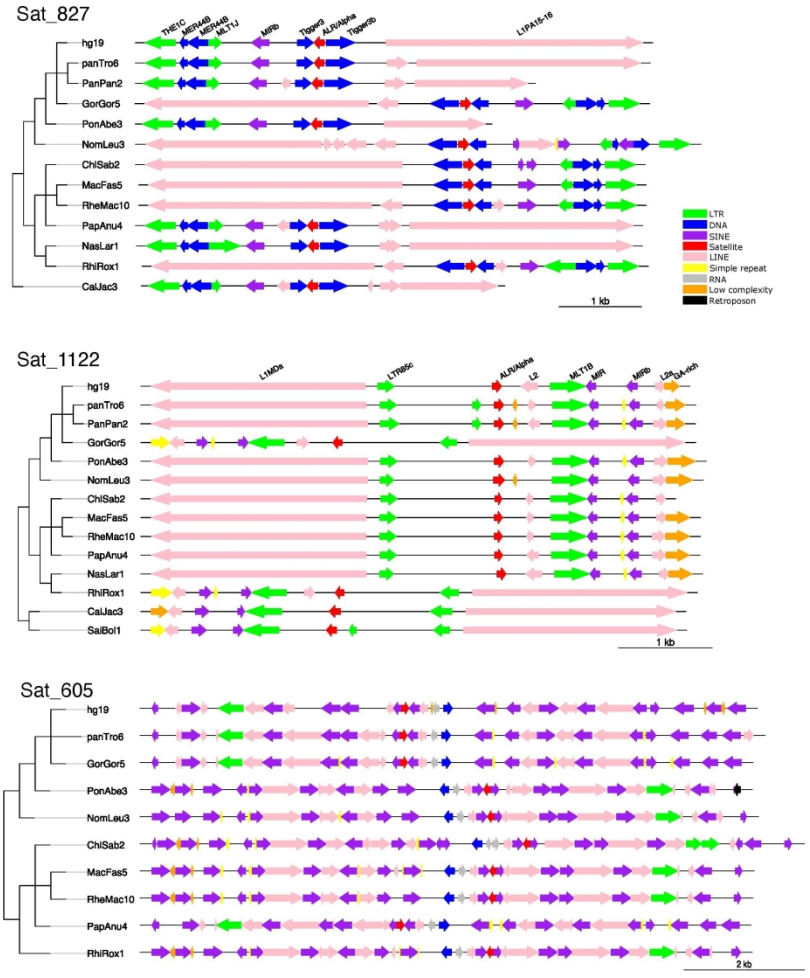

Figure 4

215x209mm (300 x 300 DPI)

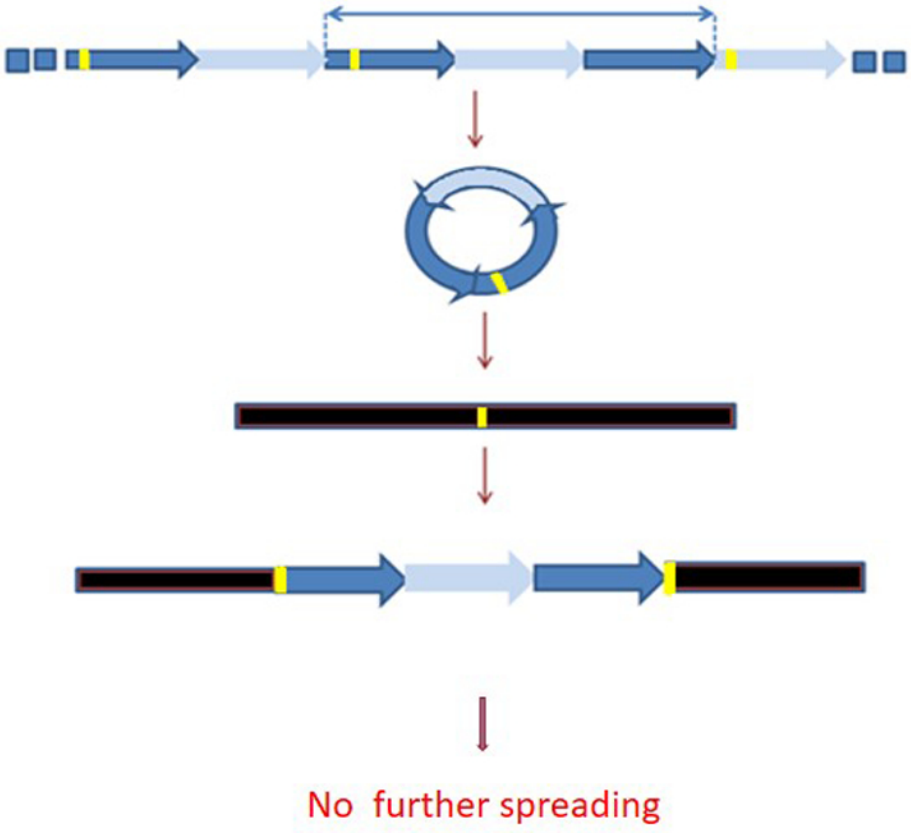

Figure 5

66x58mm (300 x 300 DPI)