Coordinated evolution of the SNORD115 and SNORD116 tandem repeats at the imprinted Prader–Willi/Angelman locus

Mathilde Guibert, Hélène Marty-Capelle, Anne Robert, Bruno Charpentier, Stéphane Labialle

TITLE

Coordinated evolution of the SNORD115 and SNORD116 tandem repeats at the imprinted Prader Willi/Angelman locus.

AUTHORS

Hélène Marty-Capelle[1†], Mathilde Guibert[2†], Anne Robert[1], Bruno Charpentier[1], Stéphane Labialle[1]*

[†] Hélène Marty-Capelle and Mathilde Guibert contributed equally to this work.

[1] Université de Lorraine, CNRS, IMoPA, F-54000 Nancy, France.

[2] Université de Lorraine, F-54000 Nancy, France.

* Correspondence: stephane.labialle@univ-lorraine.fr

ABSTRACT

The Prader Willi/Angelman syndrome (PWS/AS) locus is regulated by the epigenetic mechanism of parental genomic imprinting (PGI). This region holds two eutherian-specific, large tandem repeats of box C/D small nucleolar RNA (Snord) genes called SNORD115 and SNORD116, whose loss of paternal expression are key in the development of the Prader Willi Syndrome. Snords represent an ancient class of noncoding RNAs that classically direct the 2'-O-methylation of specific nucleotides of ribosomal RNAs. However, Snord115 and Snord116 belong to the large class of orphan Snords whose functions remain unclear. The constraints that generated and maintained this unusual genetic organization for mammalian genomes has been poorly addressed to date. Here, a comparative analysis of the evolutionary history of both tandem repeats reveals that several genetic events affected them concomitantly, including copy gains and losses between species, emergence of gene subfamilies in catarrhines or partial tandem duplication in rats. Several indications suggest that PGI orchestrated this coordination of events, adding to its roles on gene dosage, meiotic recombination and replication timing a new effect on mammalian genome structure and evolution. Finally, our work provides a functional rationale for the existence of closely located tandem repeats of small RNA genes in mammalian genomes.

KEYWORDS

Eutherian mammals, phylogenetics, tandem repeats, Prader Willi/Angelman locus, SNORD115, SNORD116, C/D box snoRNA.

INTRODUCTION

Box C/D small nucleolar RNAs (Snords) are part of a large group of middle-sized, metabolically stable RNAs that regulate post-transcriptional modification of ribosomal RNA precursors, small nuclear RNAs and transfer RNAs by 2'-O-methylation (1, 2) or, in rare cases, by base acetylation (3). Snords associate with a set of four core proteins, namely SNU13, NOP56, NOP58 and the methyltransferase FBL1, also called Fibrillarin, to form small nucleolar ribonucleoprotein complexes (snoRNPs). A basal RNA stem formed by base pairing of the ends of the molecule and a Kink-turn structure formed by the C and D boxes (consensus RUGAUGA and CUGA, respectively) are required for snoRNP biogenesis and stability (4). Often, a second couple of internal boxes noted C' and D' could also form a Kink-turn. Hybridization with targets classically involves ~9 to 20 nt-long sequences upstream the D and/or D' boxes called antisense elements ASE1 and ASE2, respectively. On the target RNA, the nucleotide that is methylated usually base-pair with the 5th nucleotide upstream of the D and/or D' boxes. Additionally, a few Snords such as U3 and U8 function as chaperones to promote the maturation of ribosomal RNA precursors (reviewed in (5)). In addition, up to one third of human Snords exhibits a lack of complementarity with canonical RNA targets and have been designated orphans. The function, if any, of this class of RNAs is largely mysterious. Nevertheless, studies have proposed that some Snords (orphan or not) target noncanonical RNAs such as pre-messenger RNAs (pre-mRNAs; reviewed in (6)). However, only a few have been experimentally validated and when they were it was usually limited to assays on classical cell lines.

The eutherian-specific SNORD115 family is one the best-known representative of the orphan SNORD gene category. At its discovery, it has been proposed to target the Htr2c pre-mRNA encoding a serotonin receptor due to a perfect 18 nt-long complementarity conserved in several species including Human and mouse (7). Subsequently, two functional effects were reported on the basis of cell line studies: editing of pre-mRNA targets (8) and alternative splicing (9), the latter having been extended to a handful of targets with more limited hybridization complementarities (10). SNORD115 genes are organized in a large tandem repeat siting at human 15q13 region also known as the Prader Willi/Angelman Syndrome (PWS/AS) locus. The region hosts a second tandem repeat constituted by the orphan SNORD116 genes that likely co-emerged with SNORD115 in an ancestor of modern eutherians. The PWS/AS locus is controlled by the mechanism of parental genomic imprinting that authorizes expression of the Snords from the paternal chromosome only. Most patients suffering from Prader Willi syndrome (PWS) lack expression of dozens of genes from the 15q13 locus including SNORD115 and SNORD116. While absence of expression of the former is not sufficient to elicit the disease (11,12), the latter candidates as a major contributor of PWS (13-20). The ability of SNORD116 to hybridize with cellular RNAs was less evident than that of SNORD115, yet recent works have

proposed candidate pre-mRNA targets whose deregulation of steady-state expression and splicing level may explain several PWS features (21, 22).

The reason, if any, of the genetic colocalization of the SNORD115 and SNORD116 repeats have received limited attention to date. Intriguingly, other tandem repeats of small RNA genes have been found in eutherian genomes. Concerning box C/D snoRNA genes, the Dlk1-Dio3 imprinted domain harbors two tandem repeats in eutherians called SNORD113 and SNORD114 (23), and a third tandem repeat called Bsr specific of rats (24). A second class of tandem repeats formed of microRNA genes is also present, e.g. the miR379-410 hosted by the Dlk1-Dio3 domain. As a recurrent feature, these tandem repeats of small non-coding RNA genes are innovations introduced at the basis of the eutherian clade or later. It is also striking that they all locate at loci controlled by parental genomic imprinting. Therefore, the question of the evolutionary constraints that gave rise to and maintained these unusual genetic features emerged several years ago, together with the question of their function in the physiology and evolution of eutherian mammals. Hypotheses on emergence (25, 26) but not on function, if any, in association with parental genomic imprinting have been proposed. Following a recent phylogenetic analysis of the SNORD116 genes (22), here we have analyzed the SNORD115 genes and performed a comparative analysis of the evolutionary history of the two tandem repeats, which opens surprising hypotheses about the genetic dynamics of this unusual locus.

METHODS

Identification of SNORD115 gene sequences

We selected the following eutherians species with genome assemblies exhibiting at the PWS/AS locus very limited or no sequence gaps as well as no truncation in several scaffolds: Human (*Homo sapiens*, Hsa), Chimpanzee (*Pan troglodytes*, Ptr), Orangutan (*Pongo pygmaeus*, Ppy), Rhesus macaque (*Macaca mulatta*, Mml), Mouse Lemur (*Microcebus murinus*, Mim), Mouse (*Mus musculus*, Mmu), Rat (*Rattus norvegicus*, Rno), Rabbit (*Oryctolagus cuniculus*, Ocu), Pig (*Sus scrofa*, Ssc) and Brown bat (*Myotis lucifugus*, Mlu). To obtain SNORD115 gene sequences, we combined data from whole genome annotations using the UCSC (genome.ucsc.edu) and Ensembl (www.ensembl.org/index.html) genome browsers and from the snoRNA databases snoRNA-LBME-db (27) and snOPY (28). We only collected sequences from PWS/AS loci. The genome assemblies used were GRCh38.p13 for Human, Pan_tro_3.0 for Chimpanzee, ponAbe3 for Orangutan, Mmul_10 for Rhesus macaque, Mmur_2.0 for Mouse Lemur, GRCm38 for Mouse, Rnor_6.0 for Rat, OryCun2.0 for Rabbit, Sscrofa11.1 for Pig and Myoluc2.0 for Brown bat. For each species, we checked the accuracy of the annotations and identified gene copies that could have been omitted in datasets using the BLAT/BLAST option of the UCSC and Ensembl browsers. We numbered SNORD115 gene copies from proximal to distal position on the tandem repeat. The SNORD115 gene sequences are listed in Supplementary Information Table 1.

Phylogenetic analyses

The sequences were aligned using the MUSCLE application in EMBL-EBI (29) with default parameters. The number of base substitutions per site was estimated using the MEGA X software by averaging between sequence groups or overall sequence pairs in each group (30) using the Kimura 2-parameter model. The p-distance corresponds to the proportion of nucleotide sites at which two sequences being compared differ and is obtained by dividing the number of nucleotide differences by the total number of nucleotides being compared. All ambiguous positions were removed for each sequence pair (pairwise deletion option). Variance was estimated using the bootstrap method and 1 000 replicates. The phylogenetic network was constructed with the SIMPLE application in EMBL-EBI using a neighbour-joining clustering method and the phylogenetic trees were generated using the iTOL tool (31). Consensus sequences were generated from sequence alignments using the EMBOSS Cons application in EMBL-EBI and the percentage of nucleotide variation was calculated as 100-Nt/(Nt-Nm), where Nt is the total nucleotide count and Nm is the major nucleotide count at a given position in the alignment. For the sake of clarity, nucleotide positions where gaps were equal to or exceeded 75% of NT were removed. The age of common ancestor between two or more species was estimated using TimeTree 5 (32).

Theoretical energy of hybridization

The energy of hybridization between RNA sequences was calculated using the IntaRNA application (33). As the natural context of the sequences was unknown, the hybridization energy was considered but not the unfolding energies.

Prediction of SNORD115-RNA interactions

To identify RNA targets conserved in Human and mouse species, we used the Ensembl interface to perform a BLASTN (ensembl.org/Multi/Tools/Blast) search with distant homologies (maximum hits: 5000; maximal E-value: 10,000; word size for seeding alignment: 5; match/mismatch: 1, -1; gap penalties: opening: 0, extension: 2) testing human and mouse ASE sequences against human and mouse cDNA (transcripts/splice variants) collections, respectively. To identify human candidate RNA targets, we performed a nucleotide BLAST search (blast.ncbi.nlm.nih.gov) using the Human genomic plus transcript (Human G+T) database and blastn algorithm with default parameters. Only hybrid size of 14 nucleotides or more were considered for analysis.

RESULTS

Covariation of SNORD115 and SNORD116 copy numbers

The human PWS/AS locus holds several protein-coding genes, long and short non-coding RNAs and two tandem repeats of orphan SNORD genes (Fig. 1A) whose synteny is pervasively conserved in eutherians. It is possible that this long-standing proximity has been associated with, or facilitated, similar evolutive histories that may extend to shared regulatory and functional mechanisms. Having analyzed the SNORD116 tandem repeat in a previous work (22), we performed a phylogenetic analysis of the SNORD115 tandem repeat to compare them. We collected 496 gene sequences present at the PWS locus of 10 eutherian species. The sequences are listed in Supplementary Information Table 1. The species were chosen for their distribution over the eutherian tree as well as for the reliability of their genomic data. Strikingly, copy count varied largely between species ranging from two in mouse lemur to 140 in mouse (Fig. 1B), which supports the existence of strong gene birth-and-death processes, as previously proposed (34). If variation in paralog number was limited in catarrhines (48 to 56 copies; last common ancestor $\sim$ 28.9 MYA), it could be much more elevated such as observed between mouse and rat (140 vs. 38 copies; last common ancestor $\sim$ 20.9 MYA), which suggests that the process has been highly active in rodents. It should be emphasized that the fact that copy number was less variable in catarrhines compared to rodents is also a feature observed for the SNORD116 genes (22). Plotting together SNORD115 and SNORD116 copy counts confirmed a strong correlation extending to the entire set of species (Fig. 1C; Pearson correlation $r^2 = 0.875$, $P = 0.00096$). As this observation was largely unexpected, we scanned additional eutherian genomes. Using the UCSC Genome Browser, we selected genome assemblies with absence or limited number of sequence gaps as well as absence of splits of the PWS/AS locus in two or more scaffolds. The analysis of 24 eutherian species confirmed the pattern of covariation (Supplementary Figure S1; Pearson correlation $r^2 = 0.889$, $P = 6E\text{-}9$). In addition, the sequence diversity between species was broadly correlated to their evolutionary distance to Human as reference (Fig. 1D; Pearson correlation $r^2 = 0.825$, $P = 0.00334$ for SNORD115; $r^2 = 0.85$, $P = 0.00184$ for SNORD116). Here, concerning SNORD116 from primates, we considered only the ancestral SNORD116 subfamily and not the primate-specific subfamilies that specifically emerged in this clade and distort the global level of sequence diversity (22). Furthermore, the rate of accumulation of variation was identical for both gene families (d=0.017/MYA), suggesting a similar trend towards slight genetic drift.

Non-allelic gene conversion shaped the SNORD115 tandem repeat

Repeated sequences with high homology are prone to recurrent genetic transfer from copy to copy also known as non-allelic gene conversion. This phenomenon pervasively shaped the evolutionary

history of the SNORD116 repeat (22). To test whether a similar phenomenon affected the SNORD115 genes, we first analyzed the occurrence of single nucleotide polymorphisms (SNPs) in human populations. We collected data from the genome aggregation database GnomAD v.3.1.2 that included 1002 SNPs spread on the 48 paralog copies (listed in Supplementary Information Table 2). Of these, only eight SNPs were shared by the nine human populations analyzed. Conversely, there was a high prevalence of rare variants as 94% of the SNPs had a minor allele frequency (MAF) <0.001. Accordingly, 49% of the polymorphisms were found to be singletons and, for the remaining ones, 43.5% were found in only one population. Then, we found three additional evidences in support of the existence of recurrent non-allelic genetic conversion events. First, paralogous sequence variants (PSVs) were overrepresented as they constituted 55.5% of the SNPs. Considering positions occupied by only two nucleotides in a human paralog alignment, PSVs constituted 58.5% of the SNPs, which corresponds to a strong deviation from random distribution (Khi2 test, P = 8.1E-36). This observation supports the hypothesis that SNPs resulted not only from point mutations but also from transfer between donor to acceptor copies. Next, comparison of the frequency of SNPs along the human gene consensus with the level of nucleotide variation between paralogs (Fig. 2A) revealed a positive correlation (Pearson correlation r^2 = 0.469, P = 8E-6), which fulfills the criterion of non-allelic gene conversion whose frequency is likely homogeneous along the gene sequence but whose detection depends on the presence of nucleotide variation between donor and acceptor copies. Finally, in order to test whether the phenomenon occurred in non-human species, we considered the relationship between the genetic distances between paralog sequences and the size of the tandem repeat in the different genomes. We observed a negative correlation for both SNORD115 and SNORD116 genes (Fig. 2B; Pearson correlation r^2 = -0.777, P = 0.0082 and r^2 = -0.809, P = 0.0046, respectively) that is consistent with gene conversion rate being driven by the number of repetitions, as already described (35). As a consequence of this and of the covariation in copy counts, there was a significant correlation between the paralog sequence diversities of the two tandem repeats (Fig. 2C). To reveal it, it was necessary to consider primate and non-primate species separately (Pearson correlation r^2 = 0.957, P = 0.0107 and r^2 = 0.902, P = 0.0364, respectively) due to the prevalence of SNORD116 subfamilies that shifts upward the level of sequence diversity in primates (22) whereas SNORD115 gene subfamilies were limited in size (see below), providing limited contribution to sequence heterogeneity. Collectively, the data strongly suggest that non-allelic gene conversion has strongly shaped SNORD115 and SNORD116 gene sequences during eutherian evolution.

Emergence of SNORD115 subfamilies in catarrhines

A phylogenetic tree presenting the relatedness of the 496 homologs revealed that most paralog copies in mouse, rat, rabbit, pig and bat generated monophyletic leaves confirming a surge of specific

sequences that contrasted with the interleaved pattern of primate genes (Supplementary Figure S2). Indeed, several ortholog copies were more similar than paralog ones, especially in Human, chimp, macaque and marmoset. Interestingly enough, a similar observation was made previously concerning the SNORD116 tandem repeat where three conserved subfamilies could be identified in primates. Yet, the level of overall genetic distance between SNORD115 genes was lower than the one observed for SNORD116, likely due to an overall higher number of gene copies that fuels a higher rate of non-allelic gene conversion and, therefore, genetic homogeneity. Thus, to identify possible SNORD115 subfamilies grouping gene orthologs, we arbitrarily set a low p-distance cutoff of 0.03. We applied it to a pairwise sequence alignment and identity calculation of all SNORD115 gene homologs. We identified only two subgroups present in three species at least and they belonged to primate species (Fig. 3A). Subgroup 1 was the largest group by far as it included 147 genes including half or more of the copies from Human, chimp, macaque, marmoset and mouse lemur (62%, 53%, 83%, 76% and 50% of the genes, respectively). Overall, this subgroup constituted two-third of the catarrhine genes, which explains that global sequence variation was limited in this clade. Subgroup 2 was much smaller, comprising just one gene copy in the catarrhal species, while the remaining gene copies belonged to the outgroup. At sequence level, the consensuses of subgroups 1 and 2 mostly differed at the region flanked by boxes C' and D, (Fig. 3B). To note, they also diverged by a C/T substitution at position 1 of box D', which corresponds to a recurrent substitution that is generally tolerated for Snord's canonical function. Outgroup copies generated a consensus identical to the consensus of subgroup 1 despite the genetic divergence of several copies, suggesting that they represent members of subgroup 1 having accumulated a high number of nucleotide variations. Interestingly, gene subgroups also populate the SNORD116 tandem repeats in catarrhines, and more largely in primates (Fig.3C). Collectively, it suggests that the tandem repeats formed gene subfamilies in a primate ancestor for SNORD116, and/or later, up to the catarrhine ancestor, for SNORD115, as it could be expected that the maintenance of subgroups was more challenging at the larger SNORD115 tandem repeats that likely experienced a higher level of genetic homogenization events compared to SNORD116. Nevertheless, the existence of these subgroups at least since the birth of the catarrhine lineage distinguishes it from non-primate lineages. The formation of gene could be associated to functional changes such as pseudogenization or neofunctionalization (36). As a first line of evidence, we considered the expression level of the different SNORD115 gene copies. We examined a high-quality dataset generated with thermostable group II intron reverse transcriptase (TGIRT) that evaluated the level of expression of SNORD genes in seven human adult tissues (37). For most SNORD115 copies, brain was the tissue with highest level of expression followed by skeletal muscle while breast, ovary and testis were the tissues with lowest level of expression (Supplementary Figure S3A). Nevertheless, the expression level varied greatly among copies. This was also true considering members of subgroup 1 that exhibited poor (e.g.

SNORD115-1, SNORD115-25) to robust (e.g. SNORD115-17, SNORD115-33) expression levels, which could hardly be attributed to their limited differences in nucleotide sequence and suggests the involvement of mechanisms that remain to be identified. It was also the case for the SNORD115-23 copy, i.e. the representative of subgroup 2 in humans, whose sequence variation with subgroup 1 could hardly explain a difference in expression level. Nevertheless, the expression of this copy was remarkable: it was the most highly expressed copy in a majority of tissues including brain or it ranked among the three most expressed copies (Supplementary Figure S3B).

Uneven distribution of variation along SNORD115 gene sequences including at ASE sequences

To further address the distribution of genetic variation, we generated a consensus of the 496 gene homologs and reported the nucleotide variability per position, i.e. the percentage of occurrence of nucleotides that differ from the main one at each position (Fig. 4A). The regions that classically contribute to snoRNA biogenesis and stability, i.e. the basal stem and the C and D boxes, belonged to the most conserved ones. This observation comforted the idea that SNORD115 generate bona fide box C/D snoRNAs. The C' and D' boxes were less conserved as often observed at canonical SNORD genes. Also, sequences framed by boxes C and D' exhibited a greater sequence diversity than sequences framed by boxes C' and D ($d=0.39 \pm 0.09$ versus $d=0.12 \pm 0.03$), even if smaller (15 versus 28 nucleotide-long, respectively), in line with previous findings (34). By homology with canonical Snord elements, we called these sequences antisense element ASE1 and ASE2, respectively. We confirmed that the level of nucleotide variation within species (Fig. 4B) and between species (Fig. 4C) was higher for ASE1 than ASE2 sequences. Using Human as a reference, the inter-species divergence at ASE2 was almost flat with an average of d=0.001/MYA, while it was 0.066/MYA for ASE1 (Fig. 4D). The level of ASE1 sequence variation correlated with the age of the last common ancestor with Human ($p=0.009$), which could be interpreted as a region having experienced a continuous drift. Conversely, the ASE2 sequence was poorly sensitive to the evolutionary distance in agreement with a selective effect. We found 23 ASE1 and 18 ASE2 different sequences common to at least two species in the 496 gene homologs, and only eight ASE1 and eight ASE2 different sequences common to at least three species. The most conserved ASE1 and ASE2 sequences belonged to 14 genes from four simian species and 111 genes from seven eutherian species, respectively (Supplementary Figure S4A). Strikingly, a large core of 25 contiguous nucleotides (TAAAAATCATGCTCAATAGGATTAC at position 48 to 72 of the gene consensus) was present in the ASE region of 202 genes distributed in all species. This highly conserved stretch does not include the nucleotide at position -1 relative to the D box. To note, mismatches and sequence polymorphisms limiting hybridization strength are often found at this position at SNORD genes guiding rRNA modification. This sequence harbored a perfect hybridization capacity with a conserved complementary sequence of Htr2c mRNAs (Supplementary Figure S4B and C). In catarrhines, this ASE2

sequence was harbored by members of subgroup 1, while members of subgroup 2 exhibited a sequence with 11 nucleotide substitutions. In consequence, members of this subgroup such as the Snord115-23 copy in humans exhibited an altered hybridization capacity with Htr2c (Supplementary Figure S5). Therefore, the human SNORD115-23 gene produced a highly expressed RNA with specific ASE2 sequence, opening the possibility that this copy – and more broadly members of subgroup 2 – has experienced neo- or pseudogenization depending on whether this copy targets other RNA(s) or not. More broadly, the SNORD115 and SNORD116 copies possessed several ASE variants that theoretically could target different set of RNAs, as illustrated by searching for human RNAs with best hybridization capacity to the main ASE sequence variants (Supplementary Figure S6), which illustrate the collective potential to functional diversity. Finally, a search for candidate SNORD115 ASE1 and ASE2 hybridization targets by conservation of sequence complementarity in both Human and mouse as done recently for SNORD116 ASEs (22), resulted in no hit for ASE1s and a unique hit for ASE2s corresponding to the Htr2c mRNA (data not shown).

Structure stabilization of SNORD115 and SNORD116 tandem repeats in catarrhines

Variation in copy numbers was limited in simian species for SNORD116 and SNORD115 tandem repeats. To address whether it could be the signature of a process of genetic stabilization, we evaluated synteny conservation by analyzing the relative position of strictly or highly homologous copies in the different tandems. The repetitive structure of the tandems makes it inherently difficult to disentangle vertical inheritance of copies from horizontal transfer by duplication of entire copies or part of them. Nevertheless, it was possible to identify apparent conservations of relative position for several SNORD115 copies in catarrhines. Syntenic conservation was evidenced by copies from subgroup 2 in the middle of the tandems and by the perfect homology or the p-distance < 0.1 that concerned the four copies forming the distal extremity of the tandems (see Fig. 3A). In contrast, the copies forming the proximal part of the tandems mainly belonged to subgroup 1 and exhibited a high sequence homogeneity. Maintenance of SNORD116 tandem synteny in simians have been already identified (22, 34) and we confirmed it by analyzing the position of perfect or near-perfect orthologs (see Fig. 3C). To note, a common pattern of high genetic divergence of the copies forming the distal part of the tandems could be observed (Supplementary Fig. S7). Therefore, in addition to a clade-specific stabilization of genetic variation, the data suggests that SNORD115 and SNORD116 shared similarities in the distribution of nucleotide variation along the tandem repeat.

High genetic dynamics of SNORD115 and SNORD116 in rodents

The genetic stability of the tandem repeats in catarrhines contrasted strongly with the situation in rodents, e.g. the mean p-distance of SNORD115 genes between Human and macaque (d=0.109±0.015; last common ancestor ~ 28.8 MYA) was lower than the one observed between mouse and rat

(d=0.212±0.04; last common ancestor ~ 13.1 MYA). Also, the mean p-distance between Human and mouse genes was lower than the one between Human and rat genes (d=0.193±0.041 versus d=0.283±0.048), suggesting that the mouse gene sequences reminded closer to the eutherian ancestral sequence and highlighting a higher divergence in rat. To note, the situation was similar concerning SNORD116: if the mean p-distance between Human and macaque (d=0.232±0.025) was higher than the one observed between mouse and rat (d=0.131±0.032), it was due to the presence of large gene subgroups in primates. Indeed, when considered independently the variability per subgroup between Human, chimp and macaque orthologs was much lower than the one observed between mouse and rat (d=0.029±0.01, d=0.045±0.01 and d=0.06±0.018 for SNORD116 subgroup 1, subgroup 2 and subgroup 3, respectively). To further explore the evolutive histories of SNORD115 and SNORD116, we thought to compare the patterns of genetic relatedness between paralog copies. In mice, the sharp increase in the size of the two tandems was associated with a strong level of genetic homogeneity. Therefore, the complex pattern of mouse copies with perfect sequence homology was likely the product of gene duplication events and nucleotide modification events – either de novo of through non-allelic gene conversion – that are inherently difficult to disentangle (Supplementary Figure S8). In consequence, if some patterns suggestive of duplication events could be found at the SNORD116 tandem (e.g. involving at least two gene copies with different sequences, such as the 116-25/26 pair identical to the 116-78/79 pair, or the pattern formed by the series 116-44/45/47/48/49 that was almost perfectly replicated in the series 116-52 to 116-57), they were hard to identify at the highly populated SNORD115 tandem that is mostly composed of large series of perfect or almost perfect repeats. As an illustration of this difficulty, the presence of recurrent patterns was discernable, e.g. the 115-92/93 pair that is replicated in the form of the 115-97/98 pair and of the 115-100/101 pair. Yet, locally the number of 115-92 variants outnumber the 115-93 variants, which could also be due to a series of duplications and/or gene conversion events of the former that were followed by emergence of the latter by de novo mutation then spreading by duplication and/or gene conversion. Conversely, in rat the level of genetic divergence was higher in agreement with a smaller number of gene copies (18 for SNORD116, 38 for SNORD115), opening up the possibility of more meaningful analysis. Indeed, strong dynamics can favor the observation of recent events that have not yet been obscured by mechanisms such as non-allelic gene conversion. As a general picture, the pattern of genetic relatedness between rat paralogs showed that identical or very similar copies were mostly placed close together, which suggests their generation by local duplication and/or local non-allelic gene conversion events (Fig. 5). Most interestingly, the analysis also evidenced a large duplication event that affected the proximal extremity of both tandem repeats. To note, the number of duplicated copies relative to the total number of copies was homothetic for both tandem repeats (four duplicated genes over a total of 18 genes for SNORD116, nine duplicated genes over a total of 38 genes for

SNORD115). If not by chance, it would suggest an extremely efficient level of coordination of the two genetic remodeling events.

DISCUSSION

Evolutionary pas de deux of the SNORD115 and SNORD116 tandem repeats

Here, we show that the genetic histories of the SNORD115 and SNORD116 tandem repeats was recurrently marked by common genetic events (summarized in Fig. 6 for euarchontoglire species that include primates and glires) that suggests the existence of one or several mechanisms of coordination. Moreover, not to mention the SNORD116 subfamilies formed from simians, which would distort genetic variation, the overall rate of divergence during eutherian evolution was strongly similar for SNORD115 and SNORD116. The correlation in the copy count of both tandem repeats suggests that simultaneous events of gene gains and losses have dominated their evolutionary history. Rapid changes in tandem size could be expected considering their inherent repetitive nature, yet these are supposed to occur independently at each repeat. The apparent coordination of these events is unexpected and new, as far as we know. Interestingly, a recent study reported a similar conclusion after analyzing SNORD115 and SNORD116 copy numbers in different mouse strains (38). Therefore, our study invites to extend the phenomenon to the other eutherian species. While size variation of tandem repeat by non-allelic recombination events is well understood, its coordination concerning two neighboring repeats is largely unexpected, and puzzling. To the best of our knowledge, nothing is known about the mechanisms supporting it. Nevertheless, if not by chance, two hypotheses can be formulated as to its cause, the first based on the local epigenetic context and the second based on natural selection. The first hypothesis is based on the introduction of a topological constraint at the SNORD115-SNORD116 locus. The PWS/AS domain is controlled by parental genomic imprinting that, in somatic cells, promotes chromatin compaction and relaxation of the maternal and paternal chromosomes, respectively (39). In germinal cells, parent-of-origin effects on DNA demethylation and de novo methylation at imprinted control regions have been observed during gametogenesis, which may reflect the conservation of dissimilar chromatin structures (40-42). Moreover, the transcriptional status of genomic regions has been shown to affect the size of chromatin loops during early meiotic prophase (43), which could influence the alignment of homologous chromosomes at monoallelically-expressed regions. In agreement, an effect of parental genomic imprinting on recombination during mammalian gametogenesis has been experimentally supported in both sexes (44, 45). In consequence, a model can be proposed in which the maintenance of different states of chromatin compaction in the two parental chromosomes would result in a restricted regime of non-allelic crossover favoring coordinated copy gains and losses when the locus is reshuffled: if one parental chromosome is more compact than the other, and the inter-tandem region is correctly aligned, then for the upstream tandem formed by the SNORD116 genes, the gene copies on the compacted maternal chromosome will only face the distal part of the gene cluster of the relaxed paternal chromosome; conversely, for

the downstream tandem formed by the SNORD115 genes, the gene copies on the compacted maternal chromosome will only face the proximal part of the gene cluster on the relaxed paternal chromosome. Then, if a crossover occurs on each tandem repeat, on the paternal chromosome it will tend to occur on the distal part of the SNORD116 tandem and on the proximal part of the SNORD115 tandem, favoring coordinated gene gains and losses. It could also explain the possibility that the two tandem repeats have recently undergone a common duplication event involving nearly a quarter of the copies located at the proximal end of the SNORD115 and SNORD116 tandem repeats in rats. This model infers that coordination could occur when two uncommon (epi)genomic features meets, i.e. the presence of (at least) two neighboring repeats at a locus controlled by parental genomic imprinting or another mechanism generating differential chromatin compaction. Note that in the species we selected, this situation seemed to affect only genomes where each tandem repeat contained at least 18 gene copies (Supplementary Figure S9). It would be interesting in the future to analyze more species as well as other tandem repeats to test whether this number could be confirmed as a lower limit for the process of coordination to be activated. If it is, it will fit well with the chromosomal structure hypothesis, as it could be expected that a minimal size of tandem should be reached in order to favor non-allelic recombination events in a favorable chromosomal pairing configuration. Alternatively, if they are not the by-product of a higher-order mechanism such as genomic imprinting, could these observations be the consequence of a selective effect? It has already been proposed that SNORD115 and SNORD116 modify each other's activity (46). However, the physiological relevance of the study was limited by the experimental setup that utilized transient overexpression of artificially-expressed SNORD genes in the moderately relevant Hek293 cells. Nevertheless, if confirmed, a reciprocal influence would open the possibility that the two gene families function cooperatively, which could imply a selective advantage provided by the conservation of a certain ratio of gene dosage. Yet, if both RNA species cooperate, it should not represent their unique or main function. Indeed, in such a situation one would expect that the absence of expression of Snord115 would roughly phenocopy the absence of Snord116, which is not supported by experimental data. Nevertheless, the existence of a cooperative function with a selective value close to neutrality could not be ruled out. In any cases, further experimental analyses will be needed to distinguish between the different scenarios, which could also help formalize new hypotheses.

Parental genomic imprinting is a fascinating mechanism that emerged at least twice during evolution, in mammals and angiosperms. The benefits it could bring to the organismal fitness have been widely debated, and certainly still holds mysteries. Several competing theories proposed that it could result from unequal interest of parental genomes in allocating resources to offspring (47) or that differences in chromatin structure of parental chromosomes facilitate distinction between homologous copies

during DNA repair and recombination in meiotic and mitotic cells (48), among other assumptions. One consequence of genomic imprinting is the failure to develop gynogenetic and androgenetic embryos, as evidenced in mouse models (49). In addition, imprinted genes experience asynchronous replication timing (50) including at the PWS/AS locus (51), like do other genes with monoallelic expression such as immune or olfactory receptor genes (52, 53). Thus, the data presented here invite us to consider the synchronization of the size of neighboring tandem repeats as a new candidate effect of parental genomic imprinting.

SNORD115 and SNORD116 gene subgroups in catarrhines

The stabilization of size and synteny of both tandem repeats in simians was associated with the emergence of gene subfamilies. Members of these subfamilies diverge in nucleotide sequence, gene expression and, as an overt possibility, in function. Whether this association underly a functional link is an open question and two opposite hypotheses could support it: either the genetic stabilization of the tandems offered the time necessary for genes to segregate in subfamilies or, inversely, the formation of functional subfamilies functioning independently or in a competitive/cooperative way restrained the possibility to change gene dosage without affecting global fitness. This second hypothesis opens the possibility that part or totality of the subfamilies performs a selective function – i.e. evolved in a context of neofunctionalization. Interestingly, the small SNORD115 subfamily was formed by one gene copy per species, which excludes its conservation by gene conversion between highly similar sequences that would dominate events between more divergent sequences, i.e. within vs. between subfamilies as proposed previously at other repeats (54) including SNORD116 (22). To note, the frequency of gene conversion is thought to be dependent on the extent of similarity between donor and acceptor copies, reaching an optimum above 90% (55). Then, it is possible that the copies forming the distal extremity of the SNORD115 tandem escaped genetic homogenization because they rapidly accumulated mutations decreasing substantially their homology with surrounding copies. As an example, in Human, homology with global consensus was 72%, 71%, 66% and 78% for the SNORD115-45, -46, -47 and -48 copies, respectively. However, this explanation does not hold for members of subgroup 2, e.g. the human SNORD115-23 copy that shared 90% of homology with the global consensus, which suggests a remarkable resistance to non-allelic gene conversion that might be of selective origin. Additionally, this copy was one of the most expressed in human tissues whereas its ASE2 sequence was uncommon as it did provide a poor capacity of interaction with the Htr2c mRNA. Therefore, it is possible that this subfamily experienced neofunctionalization to mitigate the activity of other SNORD115 copies and/or to perform other functions. Therefore, the data suggest that, after the emergence of gene subfamilies that could have occurred in the primate ancestor or later, the SNORD115 and SNORD116 tandem repeats have experienced a stabilization of genetic variation in

catarrhines, in contrast to what was proposed before (34). Also, the possibility that SNORD115 and SNORD116 gene families host copies of different functions invites to a paradigmatic shift concerning the nature of the tandem repeats: while it has been mainly supposed that these structures provide high gene dosage, it invites to revoke analyses limited to consensus sequences and to dissect their function and evolutive history at a sub-tandem scale.

Strong conservation of one SNORD115 antisense element and its hybridization capacity with the Htr2c mRNA

The hypothesis that eukaryotic Snords could target mRNAs emerged two decades ago from the description of the hybridization potential of human and mouse Snord115 with the Htr2c mRNA (7). Yet, the demonstration of the existence of this interaction is still pending and its functionality of in vivo remain a matter of debate, despite the attention it received (56). Several approaches have been utilized to attribute functions to Snord115, including gene deletion or overexpression to observe changes in molecular and cellular phenotypes, bioinformatic analysis to identify potential RNA targets, and transcriptome profiling to compare expression levels in various biological contexts. Our data agree with strong conservation of a subset of ASE2 sequences that suggests that this region performs a selective function. Going further, the simplest hypothesis is a function in RNA target binding, as in the canonical functioning of SNORDs. Noteworthy, the 26-nt long ASE2 sequence of highest conservation ended one nucleotide before box D, which is reminiscent of the pattern of canonical Snord-target interaction that often excludes the -1 nucleotide (57) and reinforces the possibility of Snord115 acting in a similar way. Replicating the strategy used to identify RNA targets of Snord116 (22), we found Htr2c as the only RNA target to be theoretically capable to hybridize with an ASE element of Snord115 in both Human and mouse (data not shown). Therefore, these data give full support to the hypothesis of Htr2c targeting. We proposed recently a similar functional mechanism involving the ASE1 element of Snord116 theoretically capable of interacting with a small set of mRNAs. In conclusion, the data show that, since their emergence in the common ancestor of eutherians, the evolutionary histories of the SNORD115 and SNORD116 tandem repeats have shared numerous features and events that suggests the existence of one or several mechanisms of coordination, but also that they likely use a common function of targeting specific mRNAs by hybridization with one of their antisense elements.

REFERENCES

1. Kiss T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.* 2001; 20(14):3617-22.

2. Vitali P, Kiss T. Cooperative 2'-O-methylation of the wobble cytidine of human elongator tRNAMet(CAT) by a nucleolar and a Cajal body-specific box C/D RNP. *Genes Dev.* 2019; 33(13-14):741-6.

3. Sharma S, Yang J, van Nues R, Watzinger P, Kötter P, Lafontaine DLJ, Granneman S, Entian KD. Specialized box C/D snoRNPs act as antisense guides to target RNA base acetylation. *PLoS Genet.* 2017; 13(5):e1006804.

4. Watkins NJ, Ségault V, Charpentier B, Nottrott S, Fabrizio P, Bachi A, Wilm M, Rosbash M, Branlant C, Lührmann R. 2000. A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. *Cell.* 2000; 103(3):457-66.

5. Kiss T. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell.* 2002; 109(2):145-8.

6. Bratkovič T, Božič J, Rogelj B. Functional diversity of small nucleolar RNAs. *Nucleic Acids Res.* 2020; 48(4):1627-51.

7. Cavaillé J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie JP, Brosius J, Hüttenhofer A. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc Natl Acad Sci U S A. 2000; 97(26):14311-6.

8. Vitali P, Basyuk E, Le Meur E, Bertrand E, Muscatelli F, Cavaillé J, Huttenhofer A. ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. *J Cell Biol.* 2005; 169(5):745-53.

9. Kishore S, Stamm S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science.* 2006; 311(5758):230-2.

10. Kishore S, Khanna A, Zhang Z, Hui J, Balwierz PJ, Stefan M, Beach C, Nicholls RD, Zavolan M, Stamm S. The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Hum Mol Genet.* 2010; 19(7): 1153-64.

11. Bürger J, Horn D, Tönnies H, Neitzel H, Reis A. Familial interstitial 570 kbp deletion of the UBE3A gene region causing Angelman syndrome but not Prader-Willi syndrome. *Am J Med Genet.* 2002; 111(3):233-7.

12. Runte M, Varon R, Horn D, Horsthemke B, Buiting K. Exclusion of the C/D box snoRNA gene cluster HBII-52 from a major role in Prader-Willi syndrome. *Hum Genet.* 2005; 116(3):228-30.

13. Skryabin BV, Gubar LV, Seeger B, Pfeiffer J, Handel S, Robeck T, Karpova E, Rozhdestvensky TS, Brosius J. Deletion of the MBII-85 snoRNA gene cluster in mice results in postnatal growth retardation. *PLoS Genet.* 2007; 3(12):e235.

14. Ding F, Li HH, Zhang S, Solomon NM, Camper SA, Cohen P, Francke U. SnoRNA Snord116 (Pwcr1/MBII-85) deletion causes growth deficiency and hyperphagia in mice. PLoS One. 2008; 3(3):e1709.

15. Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE, Garnica A, Cheung SW, Beaudet AL. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat Genet.* 2008; 40(6):719-21.

16. de Smith AJ, Purmann C, Walters RG, Ellis RJ, Holder SE, Van Haelst MM, Brady AF, Fairbrother UL, Dattani M, Keogh JM et al. A deletion of the HBII-85 class of small nucleolar RNAs (snoRNAs) is associated with hyperphagia, obesity and hypogonadism. *Hum Mol Genet.* 2009; 18(17): 3257-65.

17. Duker AL, Ballif BC, Bawle EV, Person RE, Mahadevan S, Alliman S, Thompson R, Traylor R, Bejjani BA, Shaffer LG, Rosenfeld JA, Lamb AN, Sahoo T. Paternally inherited microdeletion at 15q11.2 confirms a significant role for the SNORD116 C/D box snoRNA cluster in Prader-Willi syndrome. *Eur J Hum Genet.* 2010; 18(11):1196-201.

18. Polex-Wolf J, Lam BY, Larder R, Tadross J, Rimmington D, Bosch F, Cenzano VJ, Ayuso E, Ma MK, Rainbow K, Coll AP, O'Rahilly S, Yeo GS. Hypothalamic loss of Snord116 recapitulates the hyperphagia of Prader-Willi syndrome. *J Clin Invest.* 2018; 128(3):960-9.

19. Adhikari A, Copping NA, Onaga B, Pride MC, Coulson RL, Yang M, Yasui DH, LaSalle JM, Silverman JL. Cognitive deficits in the Snord116 deletion mouse model for Prader-Willi syndrome. *Neurobiol Learn Mem.* 2019; 165:106874.

20. Tan Q, Potter KJ, Burnett LC, Orsso CE, Inman M, Ryman DC, Haqq AM. Prader-Willi-Like Phenotype Caused by an Atypical 15q11.2 Microdeletion. *Genes (Basel).* 2020; 11(2):128.

21. Kocher MA, Huang FW, Le E, Good DJ. Snord116 post-transcriptionally increases Nhlh2 mRNA stability: implications for human Prader-Willi Syndrome. *Hum Mol Genet.* 2021; 30(12):1101-10.

22. Baldini L, Robert A, Charpentier B, Labialle S. Phylogenetic and Molecular Analyses Identify SNORD116 Targets Involved in the Prader-Willi Syndrome. *Mol Biol Evol.* 2022;39(1):msab348.

23. Cavaillé J, Seitz H, Paulsen M, Ferguson-Smith AC, Bachellerie JP. Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. *Hum Mol Genet.* 2002; 11(13):1527-38.

24. Cavaillé J, Vitali P, Basyuk E, Hüttenhofer A, Bachellerie JP. A novel brain-specific box C/D small nucleolar RNA processed from tandemly repeated introns of a noncoding RNA gene in rats. *J Biol Chem.* 2001; 276(28):26374-83.

25. Labialle S, Cavaillé J. Do repeated arrays of regulatory small-RNA genes elicit genomic imprinting?: Concurrent emergence of large clusters of small non-coding RNAs and genomic imprinting at four evolutionarily distinct eutherian chromosomal loci. *Bioessays.* 2011; 33(8):565-73.

26. Wang Q, Chow J, Hong J, Smith AF, Moreno C, Seaby P, Vrana P, Miri K, Tak J, Chung ED, Mastromonaco G, Caniggia I, Varmuza S. Recent acquisition of imprinting at the rodent Sfmbt2 locus correlates with insertion of a large block of miRNAs. *BMC Genomics.* 2011; 12:204.

27. Lestrade L and Weber MJ. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* 2006; 34(Database issue):D158-62.

28. Yoshihama M, Nakao A, Kenmochi N. snOPY: a small nucleolar RNA orthological gene database. *BMC Res Notes.* 2013; 6:426.

29. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 2019; 7(W1):W636-41.

30. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol.* 2018; 35(6):1547-9.

31. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2019; 47(W1):W256-9.

32. Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol.* 2022;39(8):msac174.

33. Mann M, Wright PR, Backofen R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res.* 2017; 45(W1):W435-9.

34. Zhang YJ, Yang JH, Shi QS, Zheng LL, Liu J, Zhou H, Zhang H, Qu LH. Rapid birth-and-death evolution of imprinted snoRNAs in the Prader-Willi syndrome locus: implications for neural development in Euarchontoglires. *PLoS One.* 2014; 9(6):e100329.

35. Melamed C, Kupiec M. Effect of donor copy number on the rate of gene conversion in the yeast Saccharomyces cerevisiae. *Mol Gen Genet.* 1992; 235(1):97-103.

36. Ohno S. Evolution by Gene Duplication. *SpringerVerlag,* New York. 1970.

37. Fafard-Couture É, Bergeron D, Couture S, Abou-Elela S, Scott MS. Annotation of snoRNA abundance across human tissues reveals complex snoRNA-host gene relationships. *Genome Biol.* 2021; 22(1):172.

38. Keshavarz M, Savriama Y, Refki P, Reeves RG, Tautz D. Natural copy number variation of tandemly repeated regulatory SNORD RNAs leads to individual phenotypic differences in mice. *Mol Ecol.* 2021; 30(19):4708-22.

39. Leung KN, Vallero RO, DuBose AJ, Resnick JL, LaSalle JM. Imprinting regulates mammalian snoRNA-encoding chromatin decondensation and neuronal nucleolar size. Hum Mol Genet. 2009; 18(22):4227-38.

40. Davis TL*,* Yang GJ*,* McCarrey JR*,* Bartolomei MS. The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development*. Hum. Mol. Genet.* 2000; 9:2885-94*.*

41. Lucifero D*,* Mann MR*,* Bartolomei MS*,* Trasler JM. Gene-specific timing and epigenetic memory in oocyte imprinting*. Hum. Mol. Genet.* 2004; 13:839-49*.*

42. Lee DH, Singh P, Tsai SY, Oates N, Spalla A, Spalla C, Brown L, Rivas G, Larson G, Rauch TA, Pfeifer GP, Szabó PE. CTCF-dependent chromatin bias constitutes transient epigenetic memory of the mother at the H19-Igf2 imprinting control region in prospermatogonia. *PLoS Genet.* 2010; 6(11):e1001224.

43. Zuo W, Chen G, Gao Z, Li S, Chen Y, Huang C, Chen J, Chen Z, Lei M, Bian Q. 2021. Stage-resolved Hi-C analyses reveal meiotic chromosome organizational features influencing homolog alignment. *Nat Commun.* 2021; 12(1):5827.

44. Paigen K, Szatkiewicz JP, Sawyer K, Leahy N, Parvanov ED, Ng SH, Graber JH, Broman KW, Petkov PM. The recombinational anatomy of a mouse chromosome. *PLoS Genet.* 2008; 4(7):e1000119.

45. Ng SH, Madeira R, Parvanov ED, Petros LM, Petkov PM, Paigen K. Parental origin of chromosomes influences crossover activity within the Kcnq1 transcriptionally imprinted domain of Mus musculus. *BMC Mol Biol.* 2009; 10:43.

46. Falaleeva M, Surface J, Shen M, de la Grange P, Stamm S. SNORD116 and SNORD115 change expression of multiple genes and modify each other's activity. *Gene.* 2015;572(2):266-73.

47. Moore T, Haig D. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet.* 1991; 7(2):45-9.

48. Pardo-Manuel de Villena F, de la Casa-Esperón E, Sapienza C. Natural selection and the function of genome imprinting: beyond the silenced minority. *Trends Genet.* 2000; 16(12):573-9.

49. McGrath J, Solter D. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell.* 1984; 37(1):179-83.

50. Simon I, Tenzen T, Reubinoff BE, Hillman D, McCarrey JR, Cedar H. Asynchronous replication of imprinted genes is established in the gametes and maintained during development. *Nature.* 1999; 401(6756):929-32.

51. Knoll JH, Cheng SD, Lalande M. Allele specificity of DNA replication timing in the Angelman/Prader-Willi syndrome imprinted chromosomal region. *Nat Genet.* 1994; 6(1):41-6.

52. Mostoslavsky R, Singh N, Tenzen T, Goldmit M, Gabay C, Elizur S, Qi P, Reubinoff BE, Chess A, Cedar H, Bergman Y. Asynchronous replication and allelic exclusion in the immune system. *Nature.* 2001; 414(6860):221-5.

53. Chess A, Simon I, Cedar H, Axel R. Allelic inactivation regulates olfactory receptor gene expression. *Cell.* 1994; 78(5):823-34.

54. Harpak A, Lan X, Gao Z, Pritchard JK. Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proc Natl Acad Sci U S A.* 2017; 114(48):12779-84.

55. Doronina L, Reising O, Schmitz J. Gene Conversion amongst *Alu* SINE Elements. *Genes (Basel).* 2021; 12(6):905.

56. Hebras J, Marty V, Personnaz J, Mercier P, Krogh N, Nielsen H, Aguirrebengoa M, Seitz H, Pradere JP, Guiard BP, Cavaille J. Reassessment of the involvement of Snord115 in the serotonin 2c receptor pathway in a genetically relevant mouse model. *Elife.* 2020; 9:e60862.

57. Chen CL, Perasso R, Qu LH, Amar L. Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA-rRNA duplexes. *J Mol Biol.* 2007; 369(3):771-83.

FIGURE LEGENDS

Fig. 1.

(A) The human PWS/AS locus contains maternally expressed (orange) and paternally expressed (blue) genes. Protein-coding genes are represented as boxes and arrows indicate the sense of transcription. The C/D snoRNA genes are represented as thin lines. The drawing is not to scale. (B) Intra- and interspecies comparison of the SNORD115 genes in 10 eutherian species. The number of paralogs, within p-distances and mean p-distance to Human are given for each species (SD, standard deviation). Catarrhines and glires are highlighted in blue and green, respectively. (C) Correlation between SNORD115 and SNORD116 copy numbers. (D) Correlation between the p-distance to human homologs and the evolutionary distance to Human in each species. In C and D, a linear regression curve is shown for each set of data.

Fig.2.

SNORD115 tandem repeats have been shaped by non-allelic gene conversion. (A) The proportion of sequence variation between human paralogs and SNP density is reported on the consensus sequence of human genes. (B) Correlation between the p-distance between paralogs and the number of copies of the tandem repeat in each species. (C) Correlation between SNORD115 and SNORD116 p-distances between copies in each species. In C and D, a linear regression curve is shown for each set of data.

Fig.3.

Two conserved gene subfamilies are hosted by the SNORD115 tandem repeats of catarrhines. (A) Distribution of SNORD115 gene subfamilies in primates. Gene copies with p-distance < 0.03 are grouped if involving three species at least (identified by colors). Inter-species copies with perfect homologies are connected by continuous lines, excepted for SNORD115 subgroup 1 forming numerous connections that are omitted for clarity. Singletons with p-distance < 0.1 are connected by dashed lines when involving three species at least. (B) Alignment of the sequence consensus of subgroup 1 (147 genes), subgroup 2 (3 genes) and 63 primate outgroup genes. Nucleotides that differ from human consensus are shown in blue. (C) (A) Distribution of SNORD116 gene subfamilies in primates. Gene copies with p-distance < 0.05 are grouped if involving three species at least (identified by colors). Inter-species copies with perfect homologies are connected by continuous lines and singletons with p-distance < 0.1 are connected by dashed lines when involving three species at least. Hsa: Human, Ptr: chimp, Mml: macaque, Cja: marmoset, Mim: mouse lemur.

Fig.4.

(A) Consensus sequence of the 496 homologs. The most frequent nucleotide is given at each position, and the graph shows the percentage of occurrence of the other nucleotides. (B) Mean p-distance between paralog ASE1 and ASE2 sequences. (C) Distance between species for ASE1 sequences and for ASE2 sequences. (D) Correlation between the p-distance to human homologs and the evolutionary distance to Human for the ASE1 and ASE2 sequences from each species. A linear regression curve is shown for each set of data.

Fg.5.

Distribution of sequence homology along the SNORD115 (A) and SNORD116 (B) tandem repeats in rats. Copies with perfect sequence homology are connected by a dendrogram and multi-copy duplicates are boxed. The distributions of the p-distance to human gene consensus along the repeats are shown below each scheme.

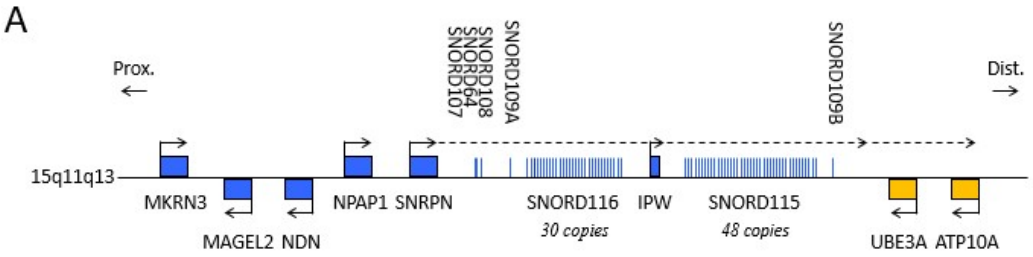
Fig.6.

Common events that shaped the evolutionary history of the SNORD115 and SNORD116 tandem repeats in euarchontoglires. Overall, the data suggest that non-allelic gene conversion events occurred independently at each tandem repeat, while gene gains and losses and gene duplication events exhibited a coordinated behavior, likely due to the presence of parental genomic imprinting at the locus.

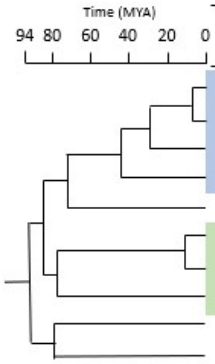
CONFLICTS OF INTEREST

The authors declare no conflict of interest.

Figure 1

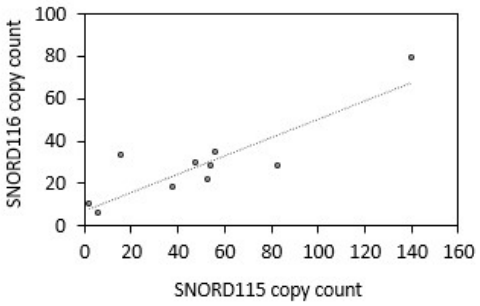


B



Species	Species ID	Gene count	Within-species mean p-distance	Mean p-distance to Human
Homo sapiens	Hsa	48	0.124±0.015	-
Pan troglodytes	Ptr	53	0.118±0.014	0.119±0.014
Macaca mulatta	Mml	54	0.088±0.014	0.106±0.013
Callithrix jacchus	Cja	56	0.079±0.012	0.101±0.012
Microcebus murinus	Mim	2	0.257±0.069	0.180±0.034
Mus musculus	Mmu	140	0.045±0.008	0.178±0.037
Rattus norvegicus	Rno	38	0.175±0.022	0.268±0.046
Oryctolagus cuniculus	Ocu	83	0.137±0.019	0.200±0.036
Sus scrofa	Ssc	6	0.358±0.054	0.291±0.041
Myotis lucifugus	Mlu	16	0.176±0.031	0.220±0.033

C



D

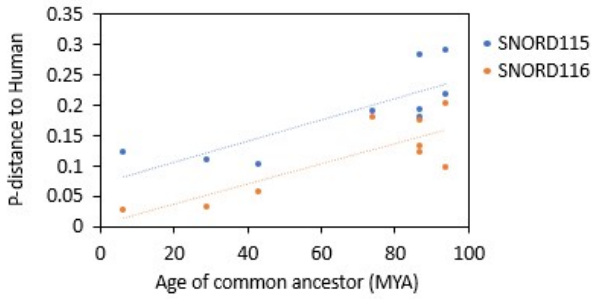


Figure 2

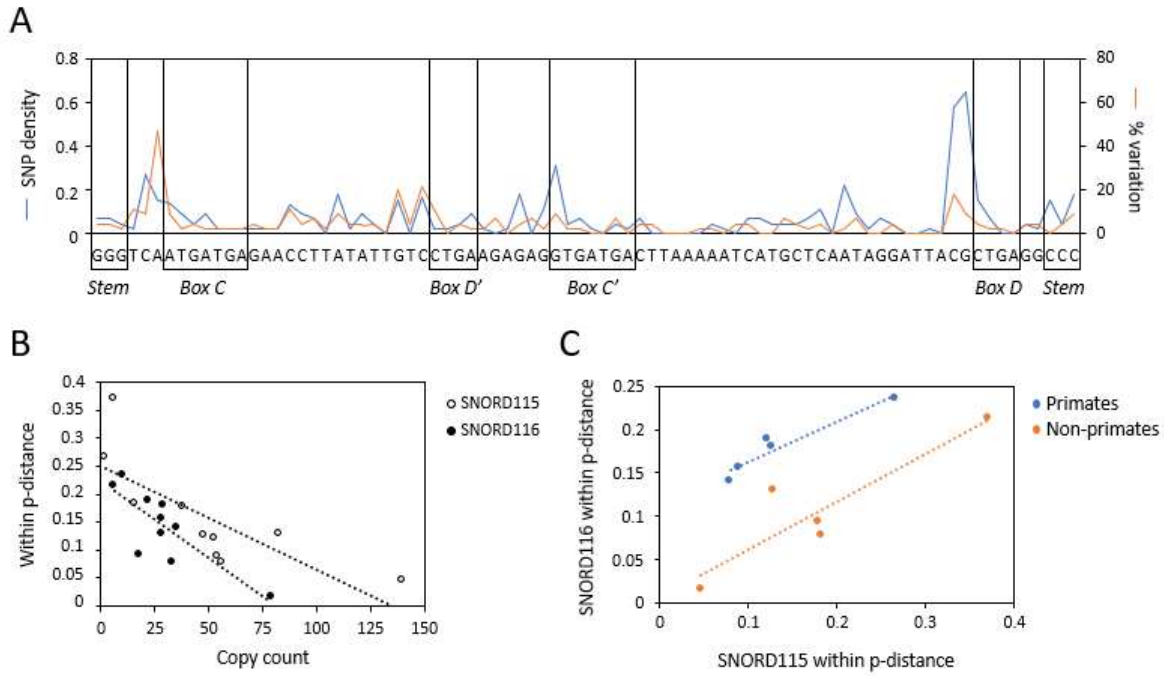

A
SNP density
% variation
GGGTCAATGATGAGAACCTTATATTGTCCTGAAGAGAGGTGATGACTTAAAAATCATGCTCAATAGGATTACGCTGAGGCCC
Stem
Box C
Box D'
Box C'
Box D
Stem
B
Within p-distance
Copy count
SNORD115
SNORD116
C
SNORD116 within p-distance
SNORD115 within p-distance
Primates
Non-primates

Figure 3

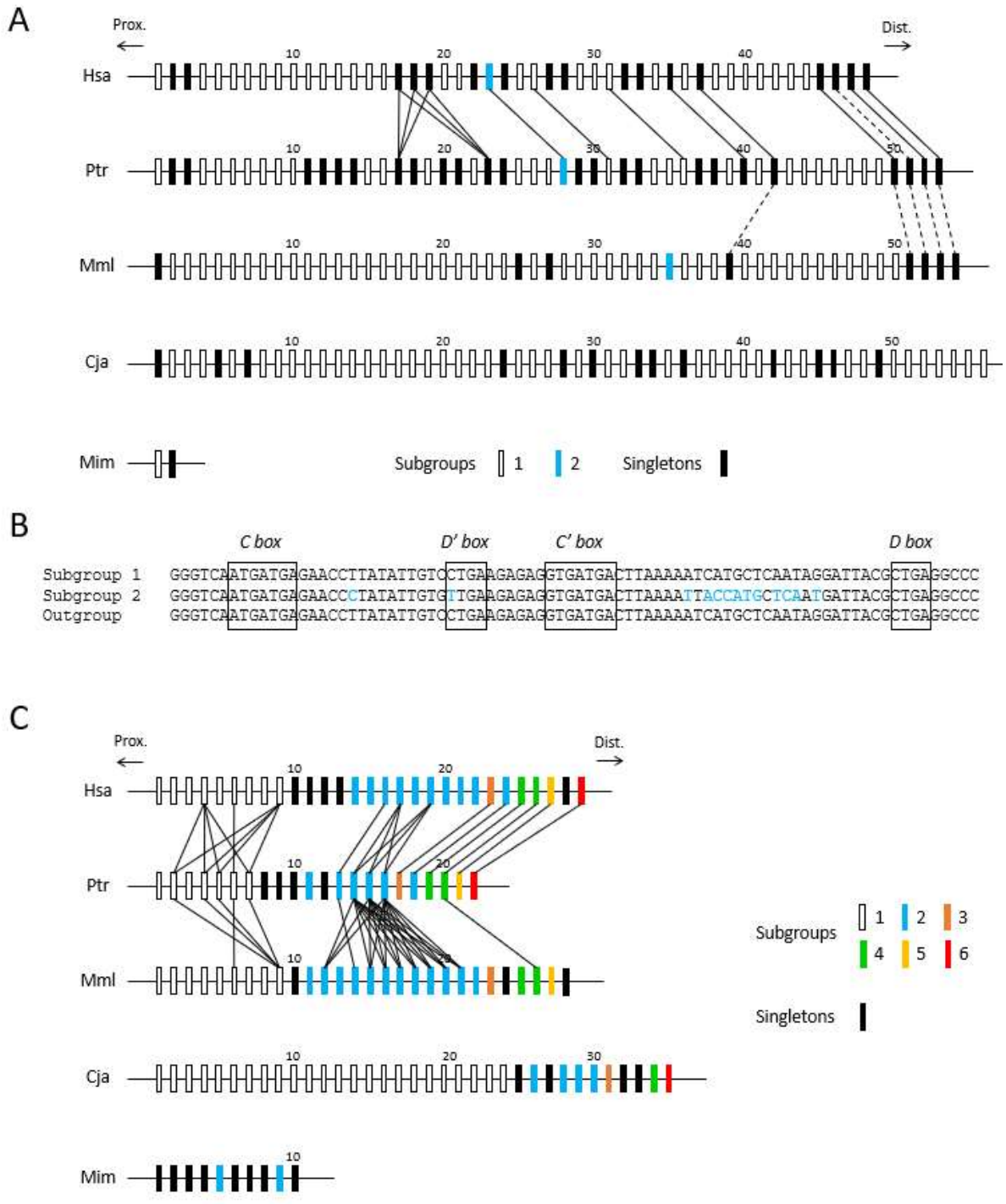

A
Prox.
Dist.
Hsa
Ptr
Mml
Cja
Mim
Subgroups 1 2
Singletons
B
C box
D' box
C' box
D box
Subgroup 1 GGGTCAATGATGAGAACCTTATATTGTCCTGAAGAGAGGTGATGACTTAAAAATCATGCTCAATAGGATTACGCTGAGGCCC
Subgroup 2 GGGTCAATGATGAGAACCCTATATTGTGTTGAAGAGAGGTGATGACTTAAAATTACCATGCTCAATGATTACGCTGAGGCCC
Outgroup GGGTCAATGATGAGAACCTTATATTGTCCTGAAGAGAGGTGATGACTTAAAAATCATGCTCAATAGGATTACGCTGAGGCCC
C
Prox.
Dist.
Hsa
Ptr
Mml
Cja
Mim
Subgroups 1 2 3 4 5 6
Singletons

Figure 4

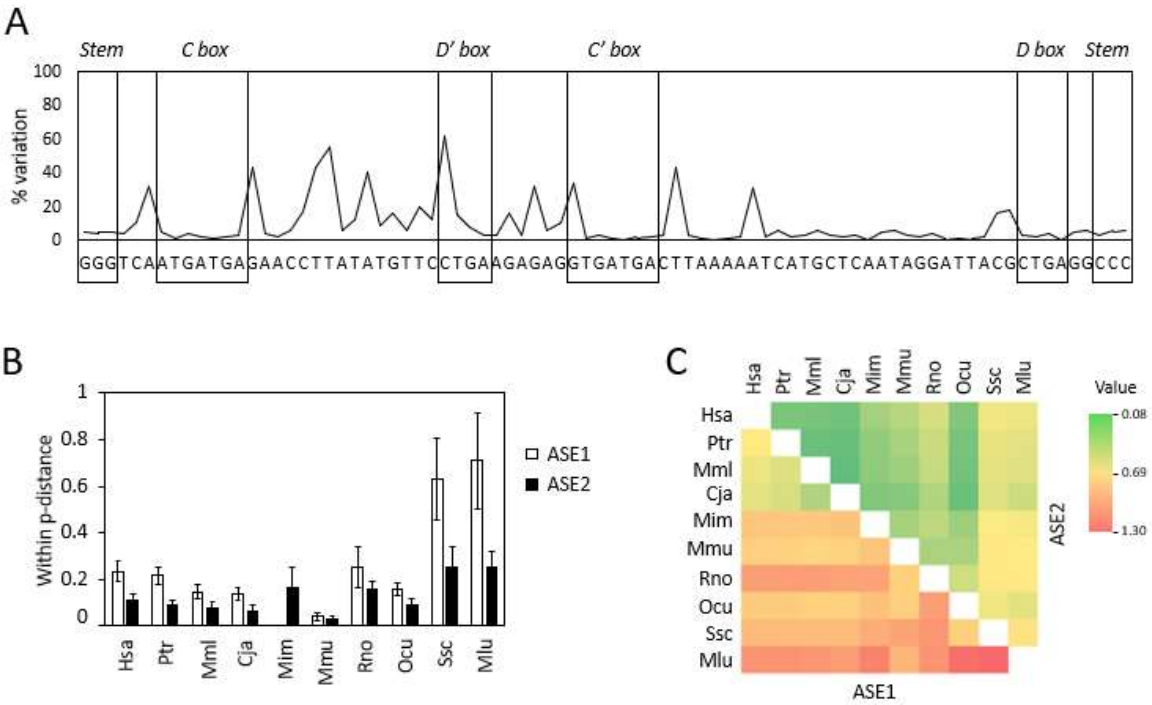
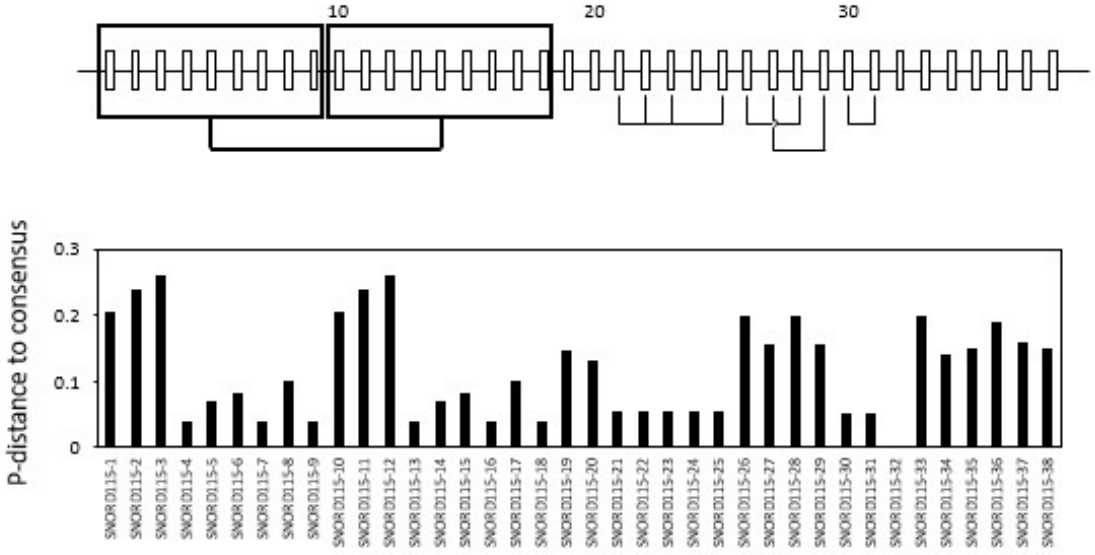

A
Stem
C box
D' box
C' box
D box
Stem
% variation
100
80
60
40
20
0
GGGTCAATGATGAGAACCTTATATGTTCCTGAAGAGAGGTGATGACTTAAAAATCATGCTCAATAGGATTACGCTGAGGCCC
B
Within p-distance
1
0.8
0.6
0.4
0.2
0
ASE1
ASE2
Hsa
Ptr
Mml
Cja
Mim
Mmu
Rno
Ocu
Ssc
Mlu
C
Hsa
Ptr
Mml
Cja
Mim
Mmu
Rno
Ocu
Ssc
Mlu
ASE1
ASE2
Value
0.08
0.69
1.30
D
P-distance to Human
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
Age of common ancestor (MYA)
ASE1
ASE2

Figure 5

A


10
20
30
P-distance to consensus
0.3
0.2
0.1
0
SNOR D115-1
SNOR D115-2
SNOR D115-3
SNOR D115-4
SNOR D115-5
SNOR D115-6
SNOR D115-7
SNOR D115-8
SNOR D115-9
SNOR D115-10
SNOR D115-11
SNOR D115-12
SNOR D115-13
SNOR D115-14
SNOR D115-15
SNOR D115-16
SNOR D115-17
SNOR D115-18
SNOR D115-19
SNOR D115-20
SNOR D115-21
SNOR D115-22
SNOR D115-23
SNOR D115-24
SNOR D115-25
SNOR D115-26
SNOR D115-27
SNOR D115-28
SNOR D115-29
SNOR D115-30
SNOR D115-31
SNOR D115-32
SNOR D115-33
SNOR D115-34
SNOR D115-35
SNOR D115-36
SNOR D115-37
SNOR D115-38

B

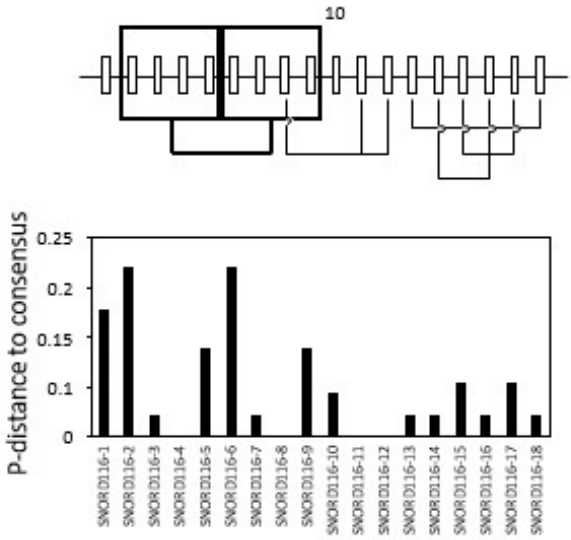
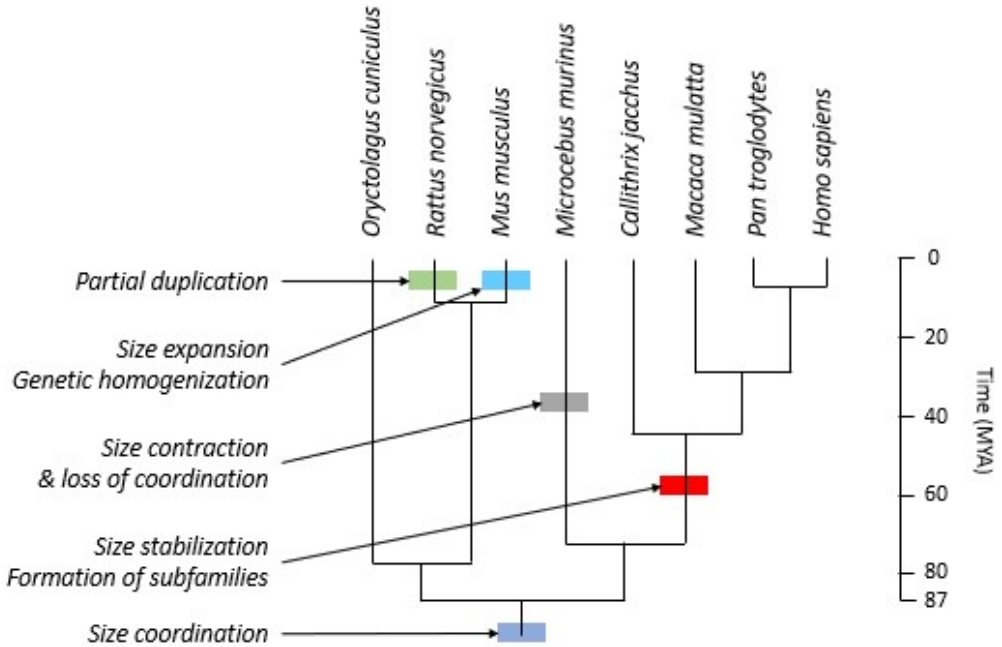

10
P-distance to consensus
0.25
0.2
0.15
0.1
0
SNOR D116-1
SNOR D116-2
SNOR D116-3
SNOR D116-4
SNOR D116-5
SNOR D116-6
SNOR D116-7
SNOR D116-8
SNOR D116-9
SNOR D116-10
SNOR D116-11
SNOR D116-12
SNOR D116-13
SNOR D116-14
SNOR D116-15
SNOR D116-16
SNOR D116-17
SNOR D116-18

Figure 6


Oryctolagus cuniculus
Rattus norvegicus
Mus musculus
Microcebus murinus
Callithrix jacchus
Macaca mulatta
Pan troglodytes
Homo sapiens
Partial duplication
Size expansion
Genetic homogenization
Size contraction
& loss of coordination
Size stabilization
Formation of subfamilies
Size coordination
0
20
40
60
80
87
Time (MYA)