Mobile Genetic Elements 1:4, 291–295; November/December 2011; © 2011 Landes Bioscience

The evolution and consequences of snaR family transposition in primates

Andrew M. Parrott and Michael B. Mathews*
Department of Biochemistry and Molecular Biology; New Jersey Medical School; University of Medicine and Dentistry of New Jersey; Newark, NJ USA

The small NF90 associated RNA (snaR) family of small noncoding RNAs (ncRNA) appears to have evolved from retrotransposon ancestors at or soon after pivotal stages in primate evolution. snaRs are thought to be derived from a FLAM C-like (free left Alu monomer) element through multiple short insertion/deletion (indel) and nucleotide (nt) substitution events. Tracing snaR's complex evolutionary history through primate genomes led to the recent discovery of two novel retrotransposons: the Alu/snaR related (ASR) and catarrhine ancestor of snaR (CAS) elements. ASR elements are present in the genomes of Simiiformes, CAS elements are present in Old World Monkeys and apes, and snaRs are restricted to the African Great Apes (Homininae, including human, gorilla, chimpanzee and bonobo). Unlike their ancestors, snaRs have disseminated by multiple rounds of segmental duplication of a larger encompassing element. This process has produced large tandem gene arrays in humans and possibly precipitated the accelerated evolution of snaR. Furthermore, snaR segmental duplication created a new form of chorionic gonadotropin β subunit (CGβ) gene, recently classified as Type II CGβ, which has altered mRNA tissue expression and can generate a novel short peptide.

Keywords: snaR, ncRNA, segmental duplication, retrotransposition, primate

Abbreviations: ASR, Alu/snaR related; CAS, catarrhine ancestor of snaR; CGβ, chorionic gonadotropin beta subunit; DHX34, DEAH box polypeptide 34; FLAM, free left Alu monomer; HAR1, human accelerated region 1; indel, insertion/deletion; LH, luteinizing hormone; LINE, long interspersed sequence; ncRNA, noncoding RNA; nt, nucleotide; NF90, nuclear factor 90; PCR, polymerase chain reaction; SINE, short interspersed sequence; snaR, small NF90 associated RNA; SNP, single nucleotide polymorphism

Submitted: 09/01/11

Revised: 09/30/11

Accepted: 10/04/11

http://dx.doi.org/10.4161/mge.18301

*Correspondence to: Michael B. Mathews; Email: mathews@umdnj.edu

The Biochemistry and Expression of snaR Non-Coding RNA

snaR non-coding RNAs were discovered in a screen of binding partners for the double-stranded RNA binding protein nuclear factor 90 (NF90).[1-3] Two snaR family subsets (snaR-A and -B) were first identified, and subsequent analysis of human and chimpanzee genomes expanded the snaR family to 30 loci in human and 12 in chimpanzee,[4] with approximately half in human (14) belonging to the snaR-A subset (**Fig. 1**; denoted in red). snaRs are typically ~120 nt long, terminate in a 3' oligo A/oligo U tract, and are predicted to adopt stable secondary structures.[4] In vitro transcription analysis of snaR-A constructs established that these genes contain an intragenic promoter and are transcribed by RNA polymerase III.[1] Following its transcription, snaR-A appears to undergo processing and translocation to the cytoplasm, possibly through NF90-mediated transport.[2,5] In the cytoplasm, the majority of snaR-A RNA co-sediments with ribosomes, raising the possibility that these non-coding RNAs function in translational control.[5]

snaR-A is induced upon cellular transformation and is expressed at high levels in most immortal cell lines, including HEK 293 cells from which they were first isolated.[1,5] In contrast, RNA gel blot and quantitative RT-PCR analyses have found snaRs to have restricted and discrete human tissue expression. snaR-A is highly expressed in testis and notably the pituitary gland, while other snaR species are well-expressed in testis, but are differentially expressed in regions of the brain.[5]

The Multiple Stages of snaR Evolution Parallels Primate Evolution

A preliminary search of annotated genomes found snaR genes to be present in human and chimpanzee, but not in other mammals such as mouse or rhesus macaque.[1] To determine their genetic

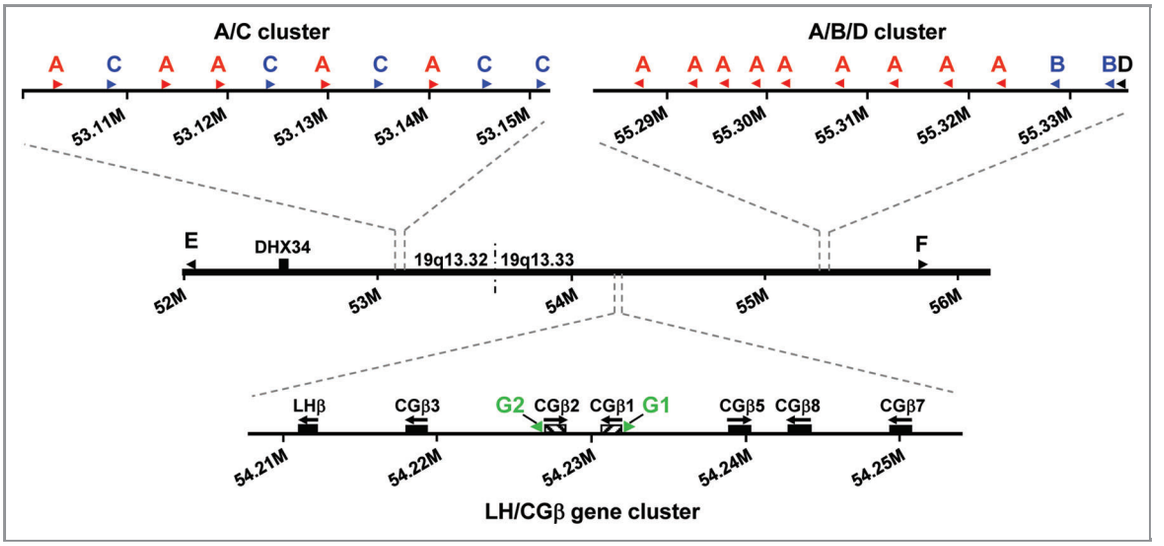

Figure 1. Human snaR gene clusters. Shown is a region of human chromosome 19q13.32–33 (middle line). Expansions detail the *snaR* A/C and A/B/D clusters (top line) and the *LH/CGβ* cluster (bottom line). The locations and direction of transcription of *snaR-A*, *-B*, *-C*, *-D*, *-E*, *-F*, *-G1* and *-G2* are denoted by colored arrow heads. *CGβ1* and *CGβ2* are denoted by hatched boxes, while *DHX34* and *LH/CGβ* cluster genes are denoted by closed boxes and their direction of transcription by arrows. Distances are in megabases (M). Adapted from Parrott et al.[5] with permission.

extent in primates and trace their molecular origin we exploited a genetic feature, namely that nearly all snaR genes are surrounded by highly conserved sequence. PCR was conducted on Great Ape genomic DNA using primers complementary to sequence encompassing *snaR*.[5] All African Great Apes were found to contain snaR genes, but a shorter element (~100 nt) was present in the orangutan PCR duplicon.[5] A search for the orangutan PCR duplicon discovered two orthologous segments separated by ~1.9 Kb on chromosome 19 of the rhesus macaque genome.[5] Searches of the intervening macaque sequence revealed a triplicate tandem repeat of 1.9 Kb demarcated by Alu sequence.[5] The syntenic repeat pattern is present in the human genome, but the 5' repeat contains *snaR-F*, rather than the truncated element present in orangutan and macaque.[5] We concluded that this element is the ancestor of *snaR*. Further searches found the novel element to be restricted to Old World Monkeys and apes, hence its name Catarrhine ancestor of snaR (CAS).[5] A discontiguous search of the 1.9 Kb rhesus macaque segment found orthologous sequence in New World Monkeys and in prosimians.[5] In New World Monkeys, the position of the *CAS* locus was occupied by a slightly longer element with similarity to FLAM C and its descendant the left monomer of Alu: this element was named Alu/snaR-related (ASR).[5]

Sequence alignment with its ancestors revealed that multiple indels (**Fig. 2**, left side) and nucleotide substitutions shaped the molecular evolution of *snaR*.[5] Each stage accompanies or follows a major primate speciation event (**Fig. 2**): *ASR* evolved from FLAM-C via an internal deletion of 19 nt after speciation of Simiiformes from tarsiers; *CAS* evolved from *ASR* after a further 3' internal deletion of 13 nt that occurred in Catarrhines after their geographical separation from Platyrrhines; and *snaR* are recently evolved from *CAS* via two sequential internal insertions of 8 nt in the African Great Apes (Hominidae). Remarkably, these events in *snaR* evolution appear to have occurred in primates at a single locus on chromosome 19, named the parent locus (**Fig. 2**, right side).

The Transposition of *snaR* Differs from that of its Ancestors

Certain genetic properties of *ASR* and *CAS* strongly suggest they are novel retrotransposons: (1) *ASR* and *CAS* are derived from a retrotransposon; (2) the majority of their loci are flanked by short direct repeats, a hallmark of target site duplication and characteristic of retrotransposition;[6,7] and (3) their loci are scattered randomly throughout the genomes of Simiiformes. All human *CAS*, and all but one chimpanzee *CAS*, have an ortholog in at least another species, whereas a substantial proportion of orangutan *CAS* and the majority of macaque *CAS* are species-specific.[5] Therefore, *CAS* retrotransposition appears to have slowed or even stopped in the African Great Apes. By contrast, few *snaR* appear to be derived from retrotransposition; instead, most are flanked by sequence identical to the parent locus. This suggests that *snaR* have disseminated through duplication of a larger encompassing segment or 'duplicon'. Segmental duplication leads to non-random, largely intrachromosomal distributions[8] and is notably more active in the African Great Apes than in orangutan or lesser apes and monkeys.[9] Indeed *snaRs* dislay a non-random distribution, and in human are chiefly located on chromosome 19 in two large inverted tandem arrays (**Fig. 1**).[1]

To trace the dispersal of *snaR* from its parent locus, the ~1.9 Kb segment containing *snaR-F* on chromosome 19 (**Fig. 2**), we searched for this segment on the well annotated human genome.[5] This analysis revealed two partial duplications on chromosomes 2 and 3, multiple short duplications in the two large tandem

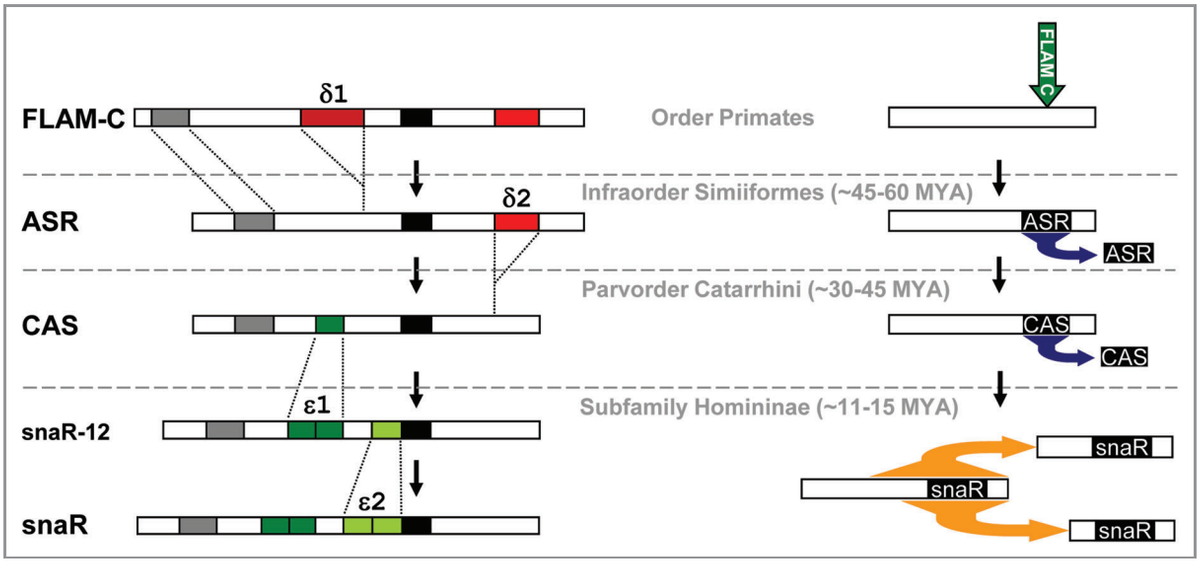

Figure 2. The step-wise evolution of *snaR* in the parent locus. Left: Schematic of the major deletions (δ1 and δ2 in red) and expansions (ε1 and ε2 in green) relating FLAM-C, ASR, CAS and snaR genes. These molecular events apparently occurred contemporaneously with major events in primate evolution (dashed horizontal gray lines), allowing their timing to be estimated. Pol III A and B boxes are represented by gray and black boxes, respectively. snaR-12 is an intermediate species between CAS and snaR. Right: Schematic illustrating the insertion of a progenitor *FLAM C* element (green arrow) into chromosome 19 sequence in primates (open box), and its stepwise evolution to *ASR*, *CAS* and *snaR*. The dissemination of *ASR* and *CAS* by retrotransposition (dark blue arrows) and of *snaR* by segmental duplication (orange arrows) from the parent locus is depicted. Adapted from Parrott and Mathews[4] with permission.

arrays on chromosome 19q13.32–33, and two short duplications in the *LH/CGβ* gene cluster situated between the tandem arrays (**Fig. 3**). Each paralogous duplicon contained a *snaR* gene and a variable amount of flanking sequence. Based on the sequences flanking the duplicons, *snaR* genes appear to have diversified along two independent duplication pathways: a major pathway that gave rise to most *snaRs*, and a second pathway which impacted *CGβ* evolution (**Fig. 3**).[10]

The Major Pathway of *snaR* Duplication Led to its Accelerated Evolution

The major pathway of snaR gene duplication includes genes for *snaR-A*, *-B*, *-C* and *-D* present in the chromosome 19 clusters (**Fig. 1**), and *snaR-H* and *-I* on chromosomes 2 and 3. These appear to have undergone a series of genomic rearrangements involving insertion of a fragment of the *snaR* parental locus into a DEAH box polypeptide 34 (*DHX34*) gene, followed by segmental duplications (**Fig. 3**). All of these *snaR* genes retain flanking sequence derived from *DHX34*. This gene is located at 52.56 Mb on chromosome 19, ~0.6 Mb downstream of the *snaR* cluster near 53 Mb (**Fig. 1**). We infer that a fragment of the parental *snaR* locus was inserted into a copy of a region of the *DHX34* gene, giving rise to a hypothetical intermediate (**Fig. 3**, step i). The insertion point is in long interspersed element (LINE) sequence (hatched in **Fig. 3**), dividing the *DHX34* duplicate into '5DX' (2.1 kbp) and '3DX' (1.2 kbp) fragments. This intermediate presumptively served as the source of H- and I-duplicons, which retain 5DX sequence, and of the A/B/C/D-duplicon which retains 3DX sequence (**Fig. 3**). All these duplicons contain a 'Core' sequence consisting of a *snaR* gene surrounded by ~0.5 Kb of parent sequence (**Fig. 3**). The Core, which has 5' *Alu* sequence and a 3' end 6 bp upstream of that of the I-duplicon (**Fig. 3**), is reminiscent of 'Duplication Cores' identified as the foci of complex interspersed duplication blocks[11] and as a source of rapidly evolving genes undergoing positive selection.[8,12,13] It is not known whether *snaR* acts as the 'driver' or is merely a 'passenger' of this duplication process.

The A/B/C/D-duplicons that encompass the *snaR-A*, *-B*, *-C* and *-D* genes consist of the Core linked to variably sized fragments of 3DX and additional chromosome 19 sequence, termed R19 (**Fig. 3**). Both the 3DX and R19 sequences are flanked by *Alu*. Initial insertion of the Core-3DX composite (**Fig. 3**, step ii) appears to have been followed by its tandem duplication with R19, and both events were possibly facilitated by *Alu-Alu*-mediated recombination (**Fig. 3**, step iii), thought to be a common recombination mechanism in primates.[14] Uniform tandem repeats are evident in the A/C cluster (**Fig. 1**). In the A/B/D cluster the uniformity of the tandem repeats is altered, especially at the 5' end (55.29–55.30 Mb) as a result of variability in the lengths of their constituent 3DX (~0.7–1.2 Kb) and R19 (~0.7–3.6 Kb) sequences.

Expansion through segmental duplication often confers genetic freedom on the resultant homologous genes, allowing them to undergo accelerated evolution.[11] Interestingly, the most abundant human *snaR* genes, *snaR-A* and *snaR-B/C*, form two distinct subsets and appear to have diverged rapidly from each other.[4,5] *snaR-A* is present in human and gorilla, arguing that they originated in the common ancestor of African Great Apes followed

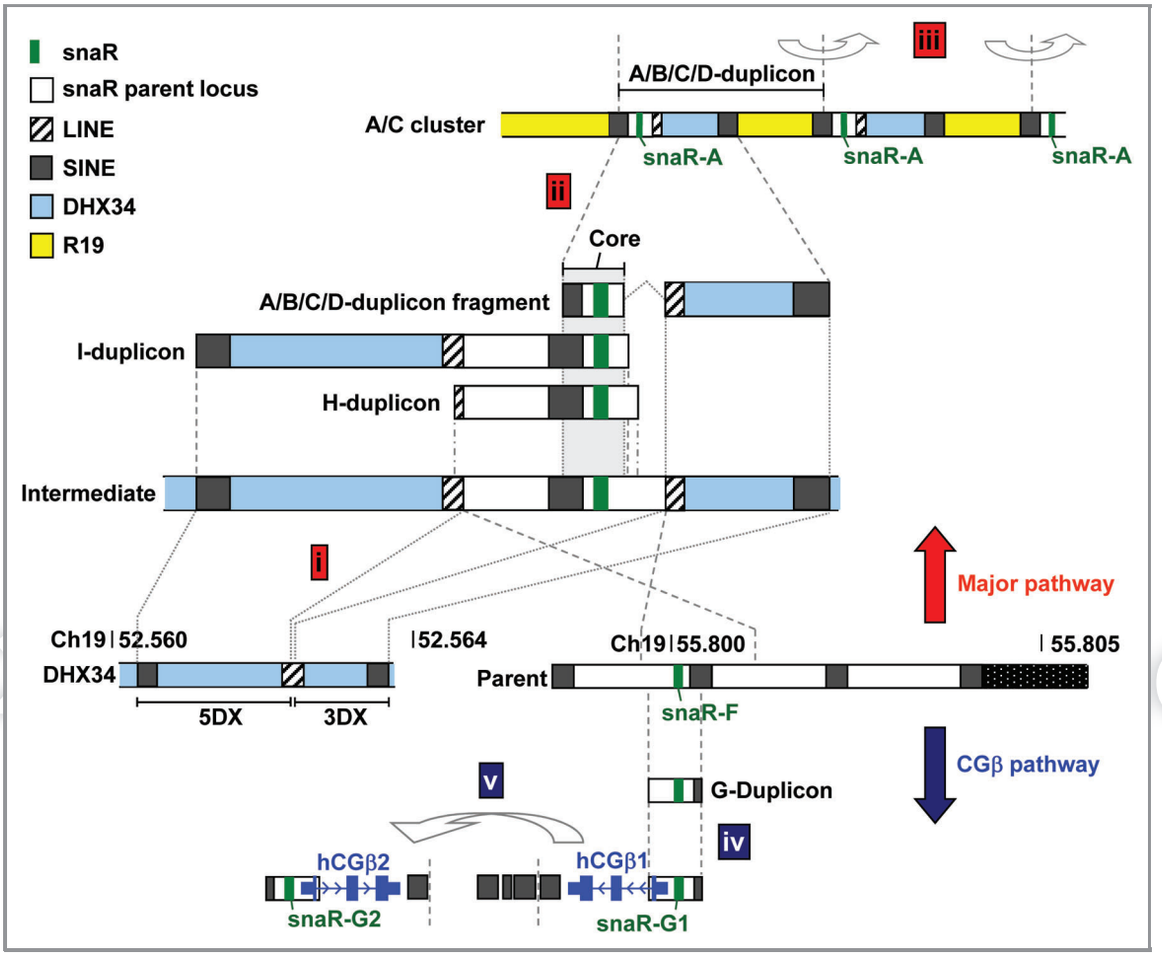

Figure 3. Dissemination of *snaR* by segmental duplication. The major pathway of snaR duplication (red arrow) entails formation of a hypothetical intermediate. A fragment of the *snaR* parent locus (open box, *snaR* in green) inserted into a copy of the *DHX34* gene (blue boxes; step i). The insertion point is in long interspersed element (LINE) sequence (hatched), dividing the *DHX34* duplicate into '5DX' and '3DX' fragments. The intermediate presumptively served as the source of H- and I-duplicons, which retain 5DX sequence, and of the A/B/C/D-duplicon which retains 3DX sequence. Insertion of the Core-3DX composite into chromosome 19, at a site adjacent to R19 (yellow), forms the A/B/C/D-duplicon (step ii), which subsequently underwent tandem duplication via *Alu-Alu*-mediated recombination (*Alu* sequences are denoted as dark gray boxes) to form the A/C cluster (step iii). The A/B/D cluster appears to have arisen as an inverted duplication of the A/C cluster. The CGβ pathway (blue arrow) envisions generation of the G-duplicon from the *snaR* parental locus and its insertion into a *CGβ* gene (step iv). The substitution of common *CGβ* gene sequence gave rise to the *CGβ1* gene (dark blue), and its subsequent inverted segmental duplication (large curved arrow; step v) generated *CGβ2*. Adapted from Parrott et al.[5] with permission.

by loss in the genus *Pan*; on the other hand, *snaR-B* and *-C* are unique to human and presumably evolved after the *Homo-Pan* species divergence. *snaR-A* underwent copy number expansion (**Fig. 1**) giving rise to 14 paralogous alleles in human and to an unknown number in gorilla. The distribution of *snaR-B* and *-C*, interspersed with *snaR-A* in the two *snaR* clusters in a tandem repeat pattern, strongly suggests that *snaR-B/C* evolved recently from redundant *snaR-A* copies (**Fig. 1**). The A/B/C/D-duplicon contains multiple SINE and LINE elements and the average substitution rate outside of the *snaR* locus is ~3.7 nt per 120 nt. In contrast, there are ~17 nt changes between human *snaR-A* (121 nt) and *snaR-C* (119 nt) species (14.0–14.2% difference), and consequently these two subsets are predicted to fold into distinctly different structures.[1] This rate of substitution approaches that observed between the human and chimpanzee orthologs of Human Accelerated Region 1 (*HAR1*) small non-coding RNA (18 substitutions in 118 nt, 15.3% difference), which is considered to be one of the most rapidly evolved RNAs in humans.[15] Single nucleotide polymorphisms (SNPs) also demonstrate recent genetic mutation; interestingly both *snaR-G2* (Human Genome Diversity Project, rs3810177)[5] and *snaR-I* (rs13323015) contain SNPs that display transition mutations between human ethnic groups.

snaR Transposition Impact on Chorionic Gonadotropin Genes

Chorionic gonadotropin (CG) is a glycoprotein hormone essential to primate reproduction. It is a heterodimer composed of an α subunit common to luteinizing hormone (LH) and other gonadotropins and a unique β subunit. In human there are 6 *CGβ* genes and a single *LHβ* gene arranged in a ~40 Kb gene cluster between the two large *snaR*

arrays on chromosome 19q13.33 (**Fig. 1**). The *CGβ* genes of African Great Apes can be classified as Type I and Type II.[10] Type I *CGβ* genes are thought to have arisen through duplication and minor mutation of the *LHβ* gene.[16] The Type II *CGβ* genes, *hCGβ1* and *hCGβ2* in humans, are unique to African Great Apes and evolved through substitution of ancestral Type I sequence with the ~0.7 Kb G-duplicon containing the *snaR-G* gene (**Fig. 3**, step iv).[10] This substitution gave rise to *hCGβ1*, which underwent inverted segmental duplication to yield *hCGβ2* (and *snaR-G2*; **Fig. 3**, step v). Human and gorilla appear to have retained both *CGβ1* and *CGβ2* and their respective *snaR-Gs*,[10] while a further duplication of *CGβ1* and deletion of *CGβ2* occurred in the genus *Pan*.[10,17]

The G-duplicon replaced the Type I proximal promoter and nearly all the 5' untranslated region, resulting in alternative Type II mRNA splicing and altered tissue expression in testis rather than placenta.[10] Furthermore, although hCGβ remains as the major protein product of Type II genes, alternatively spliced mRNA yields a novel 60 amino acid polypeptide of unknown function.[10] Thus, *snaR* transposition has dramatically altered the biology of an essential hormone gene, possibly redirecting its placental function to a role in the male reproductive system.

References

1. Parrott AM, Mathews MB. Novel rapidly evolving hominid RNAs bind nuclear factor 90 and display tissue-restricted distribution. Nucleic Acids Res 2007; 35:6249-58; PMID:17855395; http://dx.doi.org/10.1093/nar/gkm668
2. Parrott AM, Walsh MR, Mathews MB. Analysis of RNA:protein interactions in vivo: identification of RNA-binding partners of nuclear factor 90. Methods Enzymol 2007; 429:243-60; PMID:17913627; http://dx.doi.org/10.1016/S0076-6879(07)29012-3
3. Parrott AM, Mathews MB. Novel human non-coding RNAs bind to nuclear factor 90. Miami Winter Symposium 2008, 19:S15. http://www.med.miami.edu/mnbws/documents/PARROTT.pdf
4. Parrott AM, Mathews MB. snaR genes: recent descendants of Alu involved in the evolution of chorionic gonadotropins. Cold Spring Harb Symp Quant Biol 2009; 74:363-73; PMID:20028844; http://dx.doi.org/10.1101/sqb.2009.74.038
5. Parrott AM, Tsai M, Batchu P, Ryan K, Ozer HL, Tian B, et al. The evolution and expression of the snaR family of small non-coding RNAs. Nucleic Acids Res 2011; 39:1485-500; PMID:20935053; http://dx.doi.org/10.1093/nar/gkq856
6. Maraia RJ, Chang DY, Wolffe AP, Vorce RL, Hsu K. The RNA polymerase III terminator used by a B1-Alu element can modulate 3' processing of the intermediate RNA product. Mol Cell Biol 1992; 12:1500-6; PMID:1549107
7. Grindley ND. IS1 insertion generates duplication of a nine base pair sequence at its target site. Cell 1978; 13:419-26; PMID:350412; http://dx.doi.org/10.1016/0092-8674(78)90316-1
8. Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet 2007; 39:1361-8; PMID:17922013; http://dx.doi.org/10.1038/ng.2007.9
9. Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, et al. A burst of segmental duplications in the genome of the African great ape ancestor. Nature 2009; 457:877-81; PMID:19212409; http://dx.doi.org/10.1038/nature07744
10. Parrott AM, Sriram G, Liu Y, Mathews MB. Expression of type II chorionic gonadotropin genes supports a role in the male reproductive system. Mol Cell Biol 2011; 31:287-99; PMID:21078876; http://dx.doi.org/10.1128/MCB.00603-10
11. Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 2006; 7:552-64; PMID:16770338; http://dx.doi.org/10.1038/nrg1895
12. Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, et al. Positive selection of a gene family during the emergence of humans and African apes. Nature 2001; 413:514-9; PMID:11586358; http://dx.doi.org/10.1038/35097067
13. Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, Bork P. Complex genomic rearrangements lead to novel primate gene function. Genome Res 2005; 15:343-51; PMID:15710750; http://dx.doi.org/10.1101/gr.3266405
14. Bailey JA, Liu G, Eichler EE. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 2003; 73:823-34; PMID:14505274; http://dx.doi.org/10.1086/378594
15. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, et al. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 2006; 443:167-72; PMID:16915236; http://dx.doi.org/10.1038/nature05113
16. Fiddes JC, Goodman HM. The cDNA for the beta-subunit of human chorionic gonadotropin suggests evolution of a gene by readthrough into the 3'-untranslated region. Nature 1980; 286:684-7; PMID:6774259; http://dx.doi.org/10.1038/286684a0
17. Hallast P, Saarela J, Palotie A, Laan M. High divergence in primate-specific duplicated regions: Human and chimpanzee Chorionic Gonadotropin Beta genes. BMC Evol Biol 2008; 8:195; PMID:18606016; http://dx.doi.org/10.1186/1471-2148-8-195