Published online 14 December 2006 *Nucleic Acids Research, 2007, Vol. 35, No. 2* ***559–571***
doi:10.1093/nar/gkl1086

Genome-wide analyses of retrogenes derived from the human box H/ACA snoRNAs

Yuping Luo and Siguang Li*

College of Life Sciences, Nanchang University, Nanchang 330047, People's Republic of China

Received August 21, 2006; Revised November 20, 2006; Accepted November 21, 2006

ABSTRACT

The family of box H/ACA snoRNA is an abundant class of non-protein-coding RNAs, which play important roles in the post-transcriptional modification of rRNAs and snRNAs. Here we report the characterization in the human genome of 202 sequences derived from box H/ACA snoRNAs. Most of them were retrogenes formed using the L1 integration machinery. About 96% of the box H/ACA RNA-related sequences are found in corresponding locations on the chimpanzee and human chromosomes, while the mouse shares ~50% of these human sequences, suggesting that some of the H/ACA RNA-related sequences in primate occurred after the rodent/primate divergence. Of the H/ACA RNA-related sequences, 49% are found in intronic regions of protein-coding genes and 64 H/ACA-related sequences can be folded to the typical secondary structure of the box H/ACA snoRNA family, while 30 of them were recognized as functional homologs of their corresponding box H/ACA snoRNAs previously reported. Of the 64 sequences with the typical secondary structure of the box H/ACA RNA family, 11 were found in EST databases and 5 among which were shown to be expressed in more than one human tissue. Notably, U107f is nested in an intron of a protein gene coding for nudix-type motif 13, but expressed from the opposite strand, and the searching of EST databases revealed it can be expressed in liver and spleen, even in melanotic melanoma.

INTRODUCTION

The family of box H/ACA RNA is an abundant class of non-protein-coding RNAs, which includes small nucleolar RNAs (snoRNAs), small Cajal body-specific RNAs (scaRNAs) (1), as well as, a homologous class of RNAs in archaeal organisms (2). Typical box H/ACA RNA exhibits a common hairpin–hinge–hairpin-tail secondary structure with the H (ANANNA) motif in the single-stranded hinge region and an ACA triplet located 3 nt upstream of the 3′ termini (3). The majority of known box H/ACA RNAs play important roles in the post-transcriptional modification of rRNAs and snRNAs (4,5): the box H/ACA snoRNAs direct the conversion of uridine to pseudouridine at specific residues of eukaryotic ribosomal RNAs as well as Pol III-transcribed snRNA U6, whereas box H/ACA scaRNAs guide the formation of Pol II-transcribed spliceosomal nuclear RNA (snRNAs) Ψs (1). However, a few H/ACA RNAs are involved in rRNA processing, for example, U17, an evolutionarily conserved H/ACA snoRNA present in vertebrate, yeasts and the unicellular protozoan *Tetrahymena thermophila* (6), is involved in rRNA processing at the 5′ end of 18S rRNA (7). Most likely, U17 functions as an RNA chaperone that safeguards the correct folding of 18S rRNA during pre-rRNA processing.

Recently, systematic experimental approaches and computational screening programs for H/ACA RNAs have been developed and numerous H/ACA RNAs have been detected in eukaryotes from yeast to human (8–15). In humans, ~100 H/ACA RNAs have been identified, and most of which are located within the introns of protein-encoding genes (16). Some H/ACA RNAs have several copies in different introns of the same genes (17,18), or within introns of different genes (19), suggesting redundant H/ACA RNAs appear to have arisen via duplication or transposition from existing H/ACA RNAs, but the ultimate origin of these RNAs is an open question.

In humans, retrotransposons of the long interspersed element-1 (L1) family and their remnants account for ~17% of the human genome (20,21). The enzymatic machinery of a retrotransposition-competent L1 predominantly transposes its own copies (22). However, L1s are capable of transposing other sequences, mostly Alu retroposons, but also cDNAs of different types of cellular RNAs (23–25), thus forming retrogenes or retropseudogenes. The existence of an H/ACA retrogene, i.e. a non-autonomously transcribed H/ACA RNA-related sequence, was reported previously in the mouse genome (15), but no H/ACA retrogene was characterized in humans. Here we have identified 202 novel box H/ACA RNA-related sequences in the human genome, most of which are retrogenes. Sequence analyses suggest the involvement of the L1 retroposition machinery in the formation of

*To whom correspondence should be addressed. Tel: +86 791 8304099; Fax: +86 791 8302703; Email: siguangli@163.com

human H/ACA RNA retrogenes. In addition, we found that the previously reported genes encoding ACA14a, ACA37, ACA41, ACA58, ACA59a, ACA59b, ACA63, ACA66, ACA67, ACA71a, ACA98b and U109 all appear to have resulted from retrotransposition events of H/ACA RNAs, suggesting retrotransposition mechanisms have played a pivotal role in the mobility and diversification of H/ACA RNA genes.

MATERIALS AND METHODS

Computational search for H/ACA RNA-related genes in *Homo sapiens*

The sequences of human H/ACA sno/scaRNAs were taken from the snoRNA database (http://www-snorna.biotoul/fr). We used the megaBLAST tool on the NCBI website (http://www.ncbi.nlm.nih.gov/BLAST) to find box H/ACA RNA-related genes or pseudogenes on the human genome (NCBI build 36.1). The BLAST hits kept for further analysis contained at least 60% of the corresponding mature H/ACA RNA. H/ACA RNA-related sequences found in *H.sapiens* were retrieved with a 600 nt extension at each extremity and then searched for orthologs in chimpanzee genome (Pan troglodytes; NCBI build 1.1), mouse genome (mouse NCBI build 36.1) and other animal databases.

All H/ACA RNA-related genes or pseudogenes were mapped on human genome using BLAT search (http://genome.ucsc.edu/cgi-bin/hgBLAT).

Sequence identity analysis

All H/ACA RNA-related genes or pseudogenes were sequentially aligned with their corresponding H/ACA RNA gene sequence using Matcher (http://bioportal.cgb.indiana.edu/cgi-bin/emboss/matcher). The percentage of identities for each H/ACA RNA-related sequence compared with its corresponding H/ACA RNA gene was calculated.

Detection of chimeric retrogenes

To look for the eventuality of chimeric retrogenes, flanking regions of the H/ACA RNA-related sequences were sequentially aligned with the sequences of a number of other small non-protein-coding RNA species (e.g. tRNAs, snRNAs, miRNAs, rRNAs, etc.) and then investigated for repetitive elements with the RepeatMasker program (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker).

Prediction of secondary structures of H/ACA RNA-related sequences

The secondary structures of all computationally identified H/ACA-related RNAs were derived using the mfold program (26); http://www.bioinfo.rpi.edu/applications/mfold/old/rna.

RESULTS

Identification of 202 box H/ACA RNA-related genes

Using a computational, genome-wide search strategy for extracting of human sequences with sequence similarities to various box H/ACA RNAs, we found 202 box H/ACA RNA-related sequences (Table 1) when requirements for >80% identity of sequence relative to at least 60% of the length of the corresponding RNA were set. The list of these sequences is appended as Supplementary data. We also searched chimpanzee and mouse genomes and found that ~96% of these human box H/ACA RNA-related genes exist in corresponding locations on the chimpanzee chromosomes, while mouse share ~50% of these human box H/ACA RNA-related sequences (data not shown). The distribution of numbers of different human box H/ACA RNA-related genes is strikingly skewed. U70 has the most copies at 21, ACA40 has the second-most at 13, while 13 H/ACA RNAs have only one copy of ACA-related gene each, and no H/ACA-related gene was found for 28 H/ACA genes.

These box H/ACA RNA-related sequences are not uniformly distributed on human chromosomes. There are 22 and 24 copies on chromosomes 1 and 2, respectively, however, no copy was found on chromosome Y and only two copies were found on chromosome 22, while chromosomes 5, 6, 7, 12, 17, 8 and X had some relative excess density of box H/ACA RNA-related genes. Of the 202 box H/ACA RNA-related genes found in the human genome, 99 (49%) located in intronic regions of protein-coding genes. Interestingly, eight of them were distributed on the antisense orientation of their host genes (Table 1). There were no significant differences between box H/ACA RNA-related genes located in introns and these located in intergenic regions in regard to sequence identity and sequence length (data not shown).

Most of the box H/ACA RNA-related genes are retrogenes

Careful analysis of the upstream and downstream region of these H/ACA snoRNA-related sequences, we found that of the 202 box H/ACA RNA-related genes found in this work, 182 (90%) probably correspond to H/ACA retrogenes (Table 1). All these retrogenes were flanked by direct repeats (target site duplications TSDs) of 7–17 nt, and most of them contained poly (A) tails at their 3′ ends (Figure 1). Figure 1A shows a characteristic retrogene consisting of a 3′ end poly(A) tail and of TSDs. In some cases, the H/ACA RNAs, each along with their original 5′- or 3′- flanking sequences, retrotransposed into a new location on the same or a different chromosome (Figure 1B and C), suggesting these H/ACA retrogenes resulted from somewhat stable H/ACA RNA processing intermediates in H/ACA biogenesis. However, some H/ACA RNA retrogenes originated when partially processed, exon-containing hnRNAs were reverse transcribed and inserted at new locations into the genome (Figure 1D and E), for example, the ACA40 gene hosted in the sixth intron of hypothetical protein gene MGC5306, a fragment of the MGC5306 gene including the host intron of ACA40 together with all 3′-exons, retrotransposed independently into chromosome 2 (ACA40b), chromosome 17 (ACA40c), chromosome 10 (ACA40d), chromosome 6 (ACA40e), chromosome 5 (ACA40i), chromosome 8 (ACA40j) and chromosome 5 (ACA40k).

Most of the retrogenes harbored at their 5′ ends either a T_2A_4 hexanucleotide preferably recognized by L1 nicking endonuclease, or its derivatives with one or two single nucleotide substitutions (Figure 1A–E). These features suggest the

Table 1. Box H/ACA RNA-related genes in human

N	Name	Genomic placement	Chromosome	Chromosome start position	Identity (%)[c]	Type	GenBank accession no.
1	ACA1b[a]	Intronic	8	56977836	94.6	Retrogene	AC046176
2	ACA1c	Intronic	2	203625253	87.7	Retrogene[f]	AC023271
3	ACA1d	Intronic	16	24252145	83.9		AC004125
4	ACA2c	Intronic	2	10212650	88.2	Retrogene	AC007240
5	ACA2d	Intronic	1	84515509	91.4	Retrogene	AL359273
6	ACA3b	Intergenic	21	42175400	91.0[d]	Retrogene	AP001745
7	ACA3-2b[a]	Intergenic	16	2786410	86.3	Retrogene	AC005570
8	ACA3-2c	Intergenic	12	83101233	80.0	Retrogene	AC090679
9	ACA4b[a]	Intronic[b]	2	197977688	83.9	Retrogene[f]	AC010746
10	ACA7c	Intronic	11	3900374	90.0	Retrogene[f]	AC087441
11	ACA7d	Intronic[b]	11	73641107	88.5	Retrogene	AP000577
12	ACA7e	Intergenic	X	15644252	87.1	Retrogene	AC112497
13	ACA7f	Intergenic	8	52090134	81.0	Retrogene[f]	AC090919
14	ACA8b	Intergenic	X	132014448	95.0	Retrogene[f]	Z77249
15	ACA8c[a]	Intergenic	17	62698046	95.0	Retrogene	AC007448
16	ACA8d	Intronic	6	41908578	90.3[d]	Retrogene[f]	AL365205
17	ACA8e	Intronic	6	38898061	81.9[d]	Retrogene	AC022402
18	ACA9b[a]	Intronic[b]	X	99964265	94.0	Retrogene[f]	Z95327
19	ACA9c[a]	Intronic	12	122667209	93.3	Retrogene[f]	AC117503
20	ACA9d	Intronic	13	72058819	85.0	Retrogene[f]	AL356754
21	ACA10b[a]	Intronic	2	30263804	98.5	Retrogene[f]	AC016907
22	ACA10c[a]	Intronic	12	46026215	85.8	Retrogene	AC008083
23	ACA12b[a]	Intronic	1	28888764	88.6	Retrogene[f]	AL645729
24	ACA12c[a]	Intronic	17	68326570	80.0	Retrogene[f]	AC011120
25	ACA15b[a]	Intronic	7	64168351	97.8		AC073210
26	ACA15c[a]	Intronic	7	64862474	97.8		AC073107
27	ACA15d	Intronic	22	17617397	93.7[d]		AC000094
28	ACA16b	Intronic	X	16972424	86.7	Retrogene[f]	AL732371
29	ACA17b	Intergenic	10	115570208	89.5	Retrogene	AL592546
30	ACA17c	Intronic	12	70319416	87.5[d]		AC078860
31	ACA18b	Intergenic	15	30007798	93.3	Retrogene	AC079969
32	ACA18c[a]	Intronic[b]	3	178824764	89.6	Retrogene	AC026355
33	ACA18d[a]	Intergenic	5	78552521	92.3	Retrogene	AC016559
34	ACA18e	Intergenic	X	138819755	89.0[d]	Retrogene[f]	AL590077
35	ACA20b[a]	Intronic[b]	7	39335126	92.5	Retrogene	AC092174
36	ACA20c[a]	Intergenic	8	81391729	90.2	Retrogene[f]	AC104212
37	ACA20d	Intergenic	X	5338163	81.9	Retrogene	AC095353
38	ACA22b[a]	Intronic	7	64163812	98.5		AC073210
39	ACA22c	Intronic	12	38499818	88.9	Retrogene	AC121336
40	ACA22d	Intronic	7	56090552	88.5		AC092579
41	ACA25b[a]	Intergenic	3	32054023	87.4	Retrogene[f]	AC094019
42	ACA25c[a]	Intergenic	7	115008605	84.4	Retrogene[f]	AC092590
43	ACA25d[a]	Intergenic	3	180369567	81.4	Retrogene	AC076966
44	ACA26b[a]	Intergenic	2	131403803	86.9	Retrogene	AY270787
45	ACA27b	Intronic	5	139939437	89.8	Retrogene[f]	AC005214
46	ACA27c	Intergenic	16	12843845	90.6	Retrogene[f]	AC092324
47	ACA27d	Intronic	6	137218530	87.3	Retrogene[f]	AL121933
48	ACA30c	Intronic	17	73907613	88.2	Retrogene[f]	AC061992
49	ACA31b	Intronic	17	19505912	89.3[d]	Retrogene	AC005722
50	ACA31c	Intronic	2	4852459	87.7	Retrogene[f]	AC092169
51	ACA32b	Intergenic	21	41833003	82.7	Retrogene	AP001741
52	ACA32c	Intergenic	14	76737121	80.0		AC007375
53	ACA33b[a]	Intergenic	6	104137703	84.8	Retrogene[f]	AL591387
54	ACA33c	Intronic	21	32832200	89.9	Retrogene	AP000272
55	ACA36c[a]	Intronic	2	69600679	97.0	Retrogene[f]	AC079121
56	ACA36d	Intronic	10	73610624	90.9[e]	Retrogene[f]	AL607035
57	ACA37b	Intronic	9	20776934	94.5	Retrogene[f]	AL445624
58	ACA38b[a]	Intronic	17	63167248	84.3	Retrogene	AC006534
59	ACA38c[a]	Intergenic	12	117810622	80.2	Retrogene[f]	AC087863
60	ACA40b[a]	Intergenic	2	135610668	96.9	Retrogene[f]	AC020602
61	ACA40c	Intronic	17	38346113	93.1	Retrogene[f]	AC055866
62	ACA40d	Intergenic	10	36767595	93.1	Retrogene[f]	AL590730
63	ACA40e[a]	Intergenic	6	111276442	93.1	Retrogene[f]	Z84480
64	ACA40f[a]	Intergenic	20	29427783	91.5	Retrogene[f]	AL031650
65	ACA40g	Intergenic	2	18085349	90.0	Retrogene[f]	AC079802
66	ACA40h	Intergenic	1	118032764	90.6[d]	Retrogene	AL390877
67	ACA40i	Intergenic	5	74214202	91.8[d]	Retrogene[f]	AC010501
68	ACA40j	Intergenic	8	134726022	90.7	Retrogene[f]	AC090821
69	ACA40k	Intronic	5	24008052	92.3[d]	Retrogene	AC026784
70	ACA40l	Intergenic	6	35727585	91.5[d]	Retrogene	AL033519

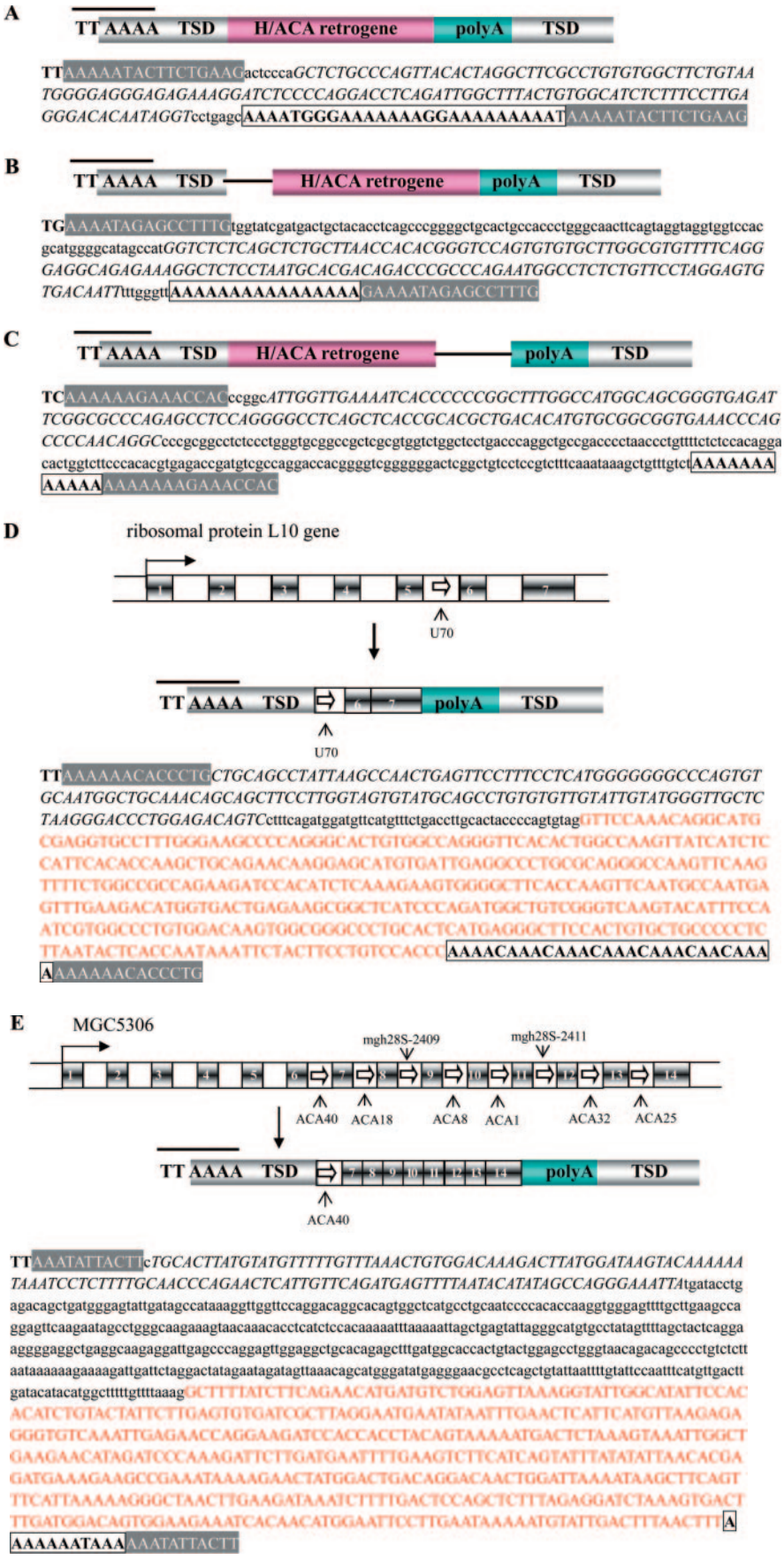
Table 1. *Continued*

N	Name	Genomic placement	Chromo some	Chromosome start position	Identity (%)[c]	Type	GenBank accession no.
71	ACA40m	Intergenic	X	123159276	89.8[d]	Retrogene	AL391241
72	ACA40n	Intergenic	7	99387638	92.6	Retrogene	AC004522
73	ACA41b	Intronic	15	43616740	85.7	Retrogene[f]	AC090527
74	ACA42b	Intronic	14	37259407	85.6[d]	Retrogene[f]	AL136296
75	ACA42c	Intergenic	1	115966016	82.2	Retrogene[f]	AL592436
76	ACA43b	Intergenic	11	65956813	84.2[d]	Retrogene[f]	AP002748
77	ACA43c	Intergenic	2	180507372	80.5	Retrogene[f]	AC096587
78	ACA45b	Intronic	6	43619859	92.9	Retrogene[f]	AL355802
79	ACA45c[a]	Intergenic	20	41366609	90.6	Retrogene[f]	AL021395
80	ACA46b	Intronic	14	77002430	82.6[d]	Retrogene	AF111168
81	ACA47b	Intergenic	1	101371297	92.5	Retrogene[f]	AC093157
82	ACA47c	Intronic	1	9065344	92	Retrogene[f]	AL158048
83	ACA47d	Intergenic	2	53551089	89.9	Retrogene[f]	AC069157
84	ACA47e	Intergenic	2	197259117	85.1	Retrogene[f]	AC068544
85	ACA47f	Intronic	11	12882940	91.6[d]	Retrogene	AC013549
86	ACA48b[a]	Intergenic	X	3460160	91.9	Retrogene[f]	AC141001
87	ACA48c	Intergenic	18	7291021	85.9	Retrogene	AP005270
88	ACA48d	Intergenic	12	55541426	87.5	Retrogene[f]	AC121758
89	ACA48e	Intergenic	4	1102672	86.7	Retrogene[f]	AC092535
90	ACA48f	Intergenic	16	66780691	85.9	Retrogene	AC130462
91	ACA48g	Intronic	2	227541951	83.0	Retrogene	AC073149
92	ACA48h	Intergenic	15	23805182	82.3	Retrogene	AC044913
93	ACA48i	Intergenic	7	101624078	84.8[d]	Retrogene[f]	AF024533
94	ACA48j	Intergenic	6	1347487	87.8[d]	Retrogene[f]	BX322644
95	ACA51b	Intergenic	2	10353624	91.8	Retrogene[f]	AC007240
96	ACA51c	Intronic	4	169629135	90.3	Retrogene[f]	AC079926
97	ACA53b[a]	Intergenic	15	63364858	93.8	Retrogene	AC068213
98	ACA53c	Intergenic	3	129916111	88.3	Retrogene[f]	AC079945
99	ACA57b	Intergenic	12	8641372	84.7	Retrogene	AC092184
100	ACA57c	Intronic	11	116639771	87.5[d]	Retrogene[f]	AP000892
101	ACA58b	Intronic	1	54009278	90.6	Retrogene[f]	AL049745
102	ACA58c[a]	Intronic	1	152498827	86.9	Retrogene	AL590431
103	ACA62b[a]	Intergenic	5	68492391	80	Retrogene	AC022107
104	ACA63b[a]	Intronic	22	18493925	92.8		AC006549
105	ACA63c	Intronic	1	8494442	80	Retrogene[f]	AL096855
106	ACA64b[a]	Intergenic	1	159377621	92.3	Retrogene	AL591806
107	ACA64c[a]	Intergenic	4	140576970	91.6	Retrogene[f]	AC097376
108	ACA64d	Intergenic	X	79999231	83.9	Retrogene	AL590031
109	ACA64e[a]	Intronic[b]	16	12298556	81.6	Retrogene	AC092365
110	ACA66b	Intergenic	1	170003837	87.3	Retrogene[f]	AL135931
111	ACA66c	Intergenic	17	23369957	86.6	Retrogene[f]	AC090287
112	ACA66d	Intergenic	15	73200933	83.0	Retrogene[f]	AC113208
113	ACA67d	Intronic	7	6023034	88.3	Retrogene	AC116348
114	ACA68b[a]	Intronic	1	15741248	90.0	Retrogene[f]	AL121992
115	ACA99b	Intronic	5	40825963	80.0[e]	Retrogene	AC008810
116	E2b	Intronic	8	18881417	87.0[d]	Retrogene[f]	AC009884
117	E2c	Intronic	4	68295578	87.7[d]	Retrogene[f]	AC079880
118	E2d	Intronic	1	35548514	88.6[d]	Retrogene[f]	AL160000
119	E3b[a]	Intronic	1	36656634	91.1	Retrogene[f]	AC119675
120	E3c	Intergenic	7	64429067	89.7	Retrogene[f]	AC092685
121	U17c[a]	Intronic	20	8759832	85.8	Retrogene[f]	AL031683
122	U17d	Intronic	18	17651619	86.4[d]	Retrogene[f]	AC103987
123	U17e	Intergenic	18	51897642	88.6	Retrogene	AC016165
124	U17f	Intronic	6	89480690	89.9[d]	Retrogene	AL160403
125	U19b	Intergenic	2	65239299	84.6	Retrogene	AC007318
126	U19-2b	Intronic	10	51281029	88.0	Retrogene	AL672187
127	U19-2c	Intergenic	17	15418086	86.9	Retrogene[f]	DQ480389
128	U19-2d	Intronic	17	14021201	86.9	Retrogene[f]	DQ075320
129	U23b[a]	Intergenic	4	16931467	91.3	Retrogene[f]	AC006231
130	U23c	Intronic	2	49969510	80.0	Retrogene[f]	AC078994
131	U23d	Intronic	12	9330536	91.5[d]	Retrogene[f]	AC009533
132	U23e	Intergenic	12	31121068	91.5[d]	Retrogene[f]	AC008013
133	U23f	Intronic	12	9488968	88.7[d]	Retrogene[f]	AC006432
134	U23g	Intronic	16	23360667	86.3[d]	Retrogene[f]	AC008915
135	U64b	Intronic	3	40255112	82.1	Retrogene[f]	AC099331
136	U64c	Intergenic	7	12706907	80.0	Retrogene[f]	AC011891
137	U67b[a]	Intergenic	7	87916343	92.3	Retrogene	AC002069
138	U67c[a]	Intronic	1	177437245	90.2	Retrogene[f]	AL512326
139	U67d	Intergenic	8	23719687	80.0	Retrogene[f]	AC012119
140	U67e	Intergenic	6	11818038	82.0	Retrogene	AL022724

Table 1. *Continued*

N	Name	Genomic placement	Chromo some	Chromosome start position	Identity (%)[c]	Type	GenBank accession no.
141	U68b[a]	Intronic	19	37791083	91.1	Retrogene[f]	AC008474
142	U68c	Intergenic	5	158589783	82.2	Retrogene[f]	AC134043
143	U68d[a]	Intergenic	X	24061225	84.5	Retrogene	AC079169
144	U69b	Intergenic	17	8173626	84.2	Retrogene	AC008053
145	U70b[a]	Intronic	1	200214211	93.5	Retrogene[f]	AC099676
146	U70c[a]	Intronic	2	215419918	95.7	Retrogene[f]	AC016708
147	U70d[a]	Intergenic	2	61497882	95.7	Retrogene[f]	AC016894
148	U70e	Intronic	2	165252407	87.7	Retrogene[f]	AL832824
149	U70f	Intergenic	5	170725729	92.8	Retrogene[f]	AC093246
150	U70g	Intergenic	5	87714345	86.3	Retrogene[f]	AC091826
151	U70h	Intergenic	8	8856495	89.9	Retrogene[f]	AC087763
152	U70i[a]	Intronic	8	4973209	87.0	Retrogene[f]	AC019176
153	U70j	Intergenic	8	33517105	93.7[d]	Retrogene[f]	AC013603
154	U70k[a]	Intronic	9	118983203	91.4	Retrogene[f]	AL355608
155	U70l	Intergenic	11	82430163	87.7	Retrogene[f]	AP000893
156	U70m[a]	Intronic	12	67307282	88.4	Retrogene[f]	AC015550
157	U70n	Intronic	12	74369190	92.7[d]	Retrogene[f]	AC078820
158	U70o	Intergenic	12	120029229	85.5	Retrogene	AC079602
159	U70p	Intronic	16	70289971	89.1	Retrogene[f]	AC010653
160	U70q	Intronic	16	48743054	92.0[d]	Retrogene[f]	AC007610
161	U70r	Intronic	17	23373483	86.3	Retrogene[f]	AC090287
162	U70s	Intronic	17	25128801	91.7[d]	Retrogene[f]	AC023389
163	U70t	Intronic	18	3015432	94.8[d]	Retrogene[f]	AP005431
164	U70u	Intronic	19	9791682	99.0[d]	Retrogene[f]	AC008752
165	U70v	Intergenic	21	33136041	90.6	Retrogene[f]	AP000039
166	U71e[a]	Intergenic	10	79797254	82.5	Retrogene	AC012560
167	U72b[a]	Intergenic	3	161897415	91.1	Retrogene[f]	AC069224
168	U72c	Intronic	1	203966973	89.6	Retrogene[f]	AC119673
169	U72d	Intronic	1	222433982	93.3[d]	Retrogene[f]	AC092809
170	U72e	Intergenic	2	139985468	85.8	Retrogene[f]	AC016710
171	U72f	Intronic	3	173971758	83.6	Retrogene[f]	AC108667
172	U72g	Intergenic	2	104716137	83.1	Retrogene[f]	AC068057
173	U72h	Intergenic	8	132514215	87.3	Retrogene[f]	AC104040
174	U87b	Intergenic	16	21506474	90.1[d]	Retrogene[f]	AC005632
175	U107b[a]	Intronic	X	54970463	98.5[e]	Retrogene	AL049732
176	U107c[a]	Intronic	X	51823183	85.5		BX537154
177	U107d[a]	Intronic	X	51950457	85.5		AL928717
178	U107e	Intergenic	15	43294396	92.9[e]		AC051619
179	U107f[a]	Intronic[b]	10	74555844	95.3[e]	Retrogene	AC016394
180	U107g[a]	Intergenic	14	90662522	94.2[e]		AC007374
181	U107h[a]	Intronic	X	47132992	90.7[e]	Retrogene[f]	AL591503
182	U107i	Intergenic	14	69340688	83.8[e]	Retrogene[f]	AL157789
183	U107j	Intergenic	21	34750278	81.4[e]	Retrogene[f]	AP000053
184	U107k	Intronic	4	120710302	86.1[e]	Retrogene[f]	AC080089
185	U108b[a]	Intronic	8	131244403	82.2	Retrogene	AC103725
186	U108c[a]	Intronic[b]	2	55646343	85.8	Retrogene	AC015982
187	U109b[a]	Intronic	1	191293034	89.0	Retrogene[f]	AL136370
188	U109c[a]	Intronic[b]	18	2545357	84.6	Retrogene[f]	AP005061
189	U109d	Intergenic	16	67222295	84.6		AC126773
190	U109e	Intergenic	2	75570241	80.0		AC007099
191	HBI-6b	Intergenic	4	53309060	87.7		AC104066
192	HBI-6c	Intergenic	1	51963035	91.9		AL050343
193	HBI-6d	Intronic	1	210265527	88.7	Retrogene[f]	AC092814
194	HBI-6e	Intergenic	20	39774533	88.7		AL133229
195	HBI-6f	Intergenic	7	151937354	91.0[d]		AC104843
196	HBI-6g[a]	Intergenic	2	10147783	85.4		AC104794
197	HBI-6h	Intergenic	20	5050063	87.9	Retrogene[f]	AL121924
198	HBI-6i	Intergenic	1	154428459	89.6[d]	Retrogene	AL135927
199	HBI-6j	Intergenic	3	53396780	87.7[d]	Retrogene	AC112218
200	HBI-6k	Intergenic	9	89065185	83.8[e]	Retrogene[f]	AL136367
201	HBI-61b	Intronic	21	31958501	86.4[d]	Retrogene	AC026776
202	HBI-61c	Intergenic	18	17545954	87.2[e]	Chimera	AC091038

[a]Retrogenes with common hairpin–hinge–hairpin–tail secondary structure.
[b]Retrogenes distributed on the antisense orientation of protein-coding genes.
[c]Identity to the corresponding consensus sequence.
[d]5′-truncated box H/ACA RNA-related sequences.
[e]3′-truncated or 3′sequences are different from the corresponding consensus sequences.
[f]Retrogenes with poly (A) tails at their 3′ ends.


A
TT AAAA TSD H/ACA retrogene polyA TSD
TTAAAAATACTTCTGAAGactcccaGCTCTGCCCAGTTACACTAGGCTTCGCCTGTGTGGCTTCTGTAA
TGGGGAGGGAGAGAAAGGATCTCCCCAGGACCTCAGATTGGCTTTACTGTGGCATCTCTTTCCTTGA
GGGACACAATAGGTcctgagcAAAATGGGAAAAAAAGGAAAAAAAAATAAAAATACTTCTGAAG
B
TT AAAA TSD H/ACA retrogene polyA TSD
TGAAAATAGAGCCTTTGtggtatcgatgactgctacacctcagcccggggctgcactgccaccctgggcaacttcagtaggtaggtggtccac
gcatggggcatagccatGGTCTCTCAGCTCTGCTTAACCACACGGGTCCAGTGTGTGCTTGGCGTGTTTTCAGG
GAGGCAGAGAAAGGCTCTCCTAATGCACGACAGACCCGCCCAGAATGGCCTCTCTGTTCCTAGGAGTG
TGACAATTtttgggttAAAAAAAAAAAAAAAGAAAATAGAGCCTTTG
C
TT AAAA TSD H/ACA retrogene polyA TSD
TCAAAAAAGAAACCAGccggcATTGGTTGAAAATCACCCCCCGGCTTTGGCCATGGCAGCGGGTGAGAT
TCGGCGCCCAGAGCCTCCAGGGGCCTCAGCTCACCGCACGCTGACACATGTGCGGCGGTGAAACCCAG
CCCCAACAGGCcccgcggcctctcccgggtgcggccgctcgcgtggtctggctcctgacccaggctgccgaccctaaccctgttttctctccacagga
cactggtcttcccacacgtgagaccgatgtcgccaggaccacggggtcgggggggactcggctgtcctccgtctttcaaataaagctgtttgtctAAAAAAA
AAAAAAAAAAAAAAGAAACCAG
D
ribosomal protein L10 gene
U70
TT AAAA TSD polyA TSD
U70
TTAAAAAACACCCTGCTGCAGCCTATTAAGCCAACTGAGTTCCTTTCCTCATGGGGGGGCCCAGTGT
GCAATGGCTGCAAACAGCAGCTTCCTTGGTAGTGTATGCAGCCTGTGTGTTGTATTGTATGGGTTGCTC
TAAGGGACCCTGGAGACAGTCctttcagatggatgttcatgtttctgaccttgcactaccccagtgtagGTTCCAAACAGGCATG
CGAGGTGCCTTTGGGAAGCCCCAGGGCACTGTGGCCAGGGTTCACACTGGCCAAGTTATCATCTC
CATTCACACCAAGCTGCAGAACAAGGAGCATGTGATTGAGGCCCTGCGCAGGGCCAAGTTCAAG
TTTTCTGGCCGCCAGAAGATCCACATCTCAAAGAAGTGGGGCTTCACCAAGTTCAATGCCAATGA
GTTTGAAGACATGGTGACTGAGAAGCGGCTCATCCCAGATGGCTGTCGGGTCAAGTACATTTCCA
ATCGTGGCCCTGTGGACAAGTGGCGGGCCCTGCACTCATGAGGGCTTCCACTGTGCTGCCCCCTC
TTAATACTCACCAATAAATTCTACTTCCTGTCCACCCAAAACAAACAAACAAACAAACAAACAAA
AAAAAAACACCCTG
E
MGC5306
mgh28S-2409
mgh28S-2411
ACA40 ACA18 ACA8 ACA1 ACA32 ACA25
TT AAAA TSD polyA TSD
ACA40
TTAAATATTACTTcTGCACTTATGTATGTTTTTGTTTAAACTGTGGACAAAGACTTATGGATAAGTACAAAAAA
TAAATCCTCTTTTGCAACCCAGAACTCATTGTTCAGATGAGTTTTAATACATATAGCCAGGGAAATTAtgatacctg
agacagctgatgggagtattgatagccataaaggttggttccaggacaggcacagtggctcatgcctgcaatccccacaccaaggtgggagttttgcttgaagcca
ggagttcaagaatagcctgggcaagaaagtaacaaacacctcatctccacaaaaatttaaaaattagctgagtattagggcatgtgcctatagttttagctactcagga
aggggaggctgaggcaagaggattgagcccaggagttggaggctgcacagagctttgatggcaccactgtactggagcctgggtaacagacagcccctgtctctt
aataaaaaagaaaagattgattctaggactatagaatagatagttaaacagcatgggatatgagggaacgcctcagctgtattaattttgtattccaatttcatgttgactt
gatacatacatggctttttgttttaaagGCTTTTATCTTCAGAACATGATGTCTGGAGTTAAAGGTATTGGCATATTCCAC
ACATCTGTACTATTCTTGAGTGTGATCGCTTAGGAATGAATATAATTTGAACTCATTCATGTTAAGAGA
GGGTGTCAAATTGAGAACCAGGAAGATCCACCACCTACAGTAAAAATGACTCTAAAGTAAATTGGCT
GAAGAACATAGATCCCAAAGATTCTTGATGAATTTTGAAGTCTTCATCAGTATTTATATATTAACACGA
GATGAAAGAAGCCGAAATAAAAGAACTATGGACTGACAGGACAACTGGATTAAAATAAGCTTCAGT
TTCATTAAAAAGGGCTAACTTGAAGATAAATCTTTTGACTCCAGCTCTTTAGAGGATCTAAAGTGACT
TTGATGGACAGTGGAAGAAATCACAACATGGAATTCCTTGAATAAAAATGTATTGACTTTAACTTTA
AAAAAATAAAAATATTACTT

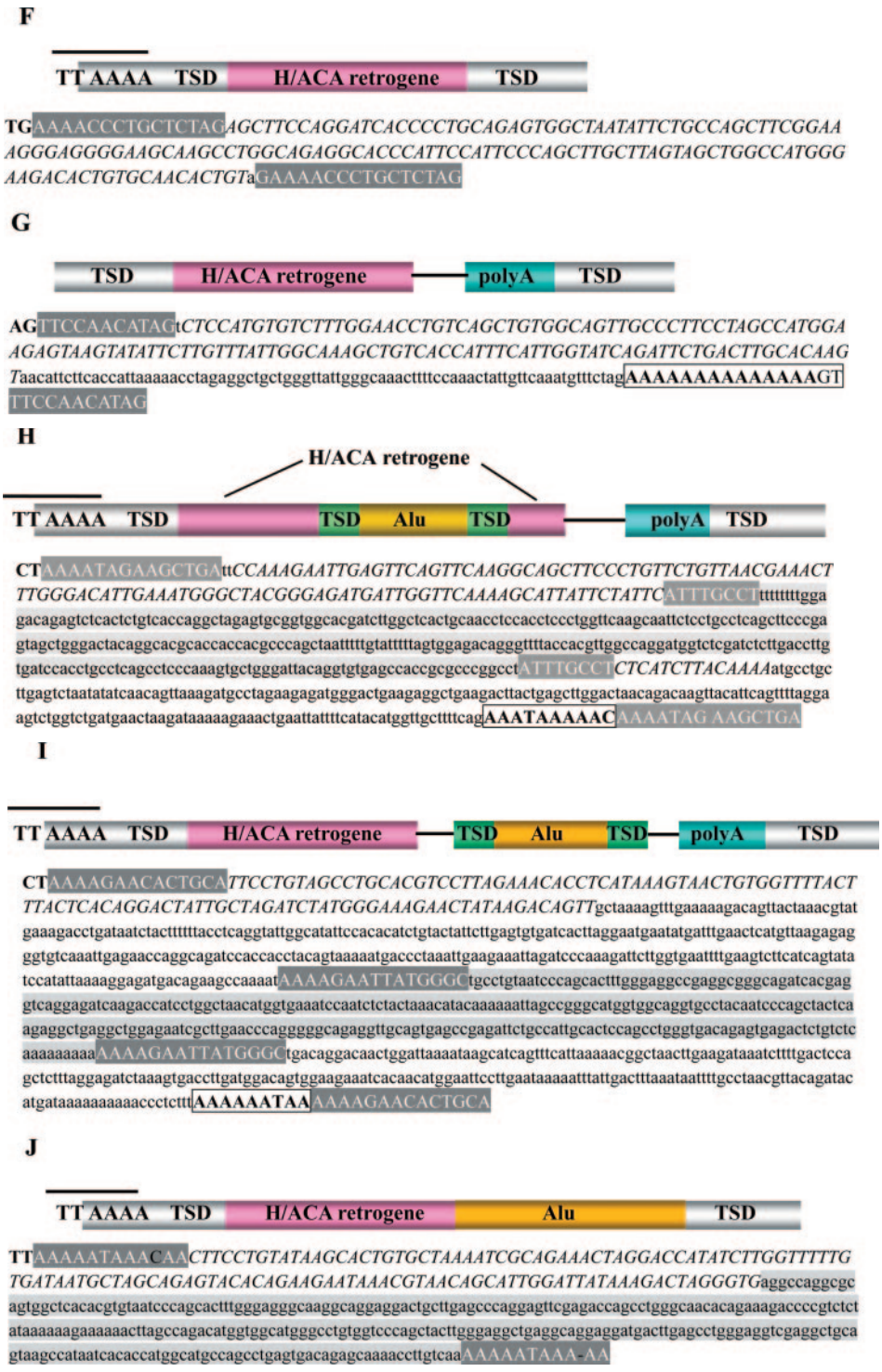


Figure 1. Schematic representation of box H/ACA RNA retrogene examples. (**A**) The sequence below the scheme is retrogene U64b and 55 retrogenes belong to this type. (**B**) The sequence below the scheme is retrogene ACA10b and a number of retroposed nucleotides on the 5′-flanks and 5 retrogenes belong to this type. (**C**) The sequence below the scheme is retrogene ACA64c and a number of retroposed nucleotides on the 3′-flanks and 24 retrogenes belong to this type. (**D**) The sequence below the scheme is retrogene U70m and a number of retroposed nucleotides on the 3′-flanks and 25 retrogenes are similar to this case. (**E**) The sequence below the scheme is retrogene ACA40j and a number of retroposed nucleotides on the 3′-flanks and 12 retrogenes are similar to this case. The exon-derived sequences in (D) and (E) are shown in capital letters. (**F**) The sequence below the scheme is retrogene ACA7d and 6 retrogenes belong to this type. (**G**) The sequence below the scheme is retrogene ACA53c and a number of retroposed nucleotides on the 3′-flanks and 7 retrogenes belong to this type. (**H**) The sequence below the scheme is retrogene ACA36d and a number of retroposed nucleotides on the 3′-flanks and 4 retrogenes belong to this type. (**I**) 6 retrogenes belong to this type. The sequence below the scheme is retrogene ACA18e and a number of retroposed nucleotides on the 3′ flanks. (J) The sequence below the scheme is retrogene HBI-61c and 1 retrogene belongs to this type. In all the cases, the H/ACA RNA sequences are in italics, retroposed nucleotides on the 3′- or 5′-flanks are in lower cases, Alu sequences are shaded, poly(A) and TSD are in opened and closed boxes, respectively. The L1 consensus recognition site (TTAAAA) is indicated at the 5′ end and overlaid by a black bar in the examples.

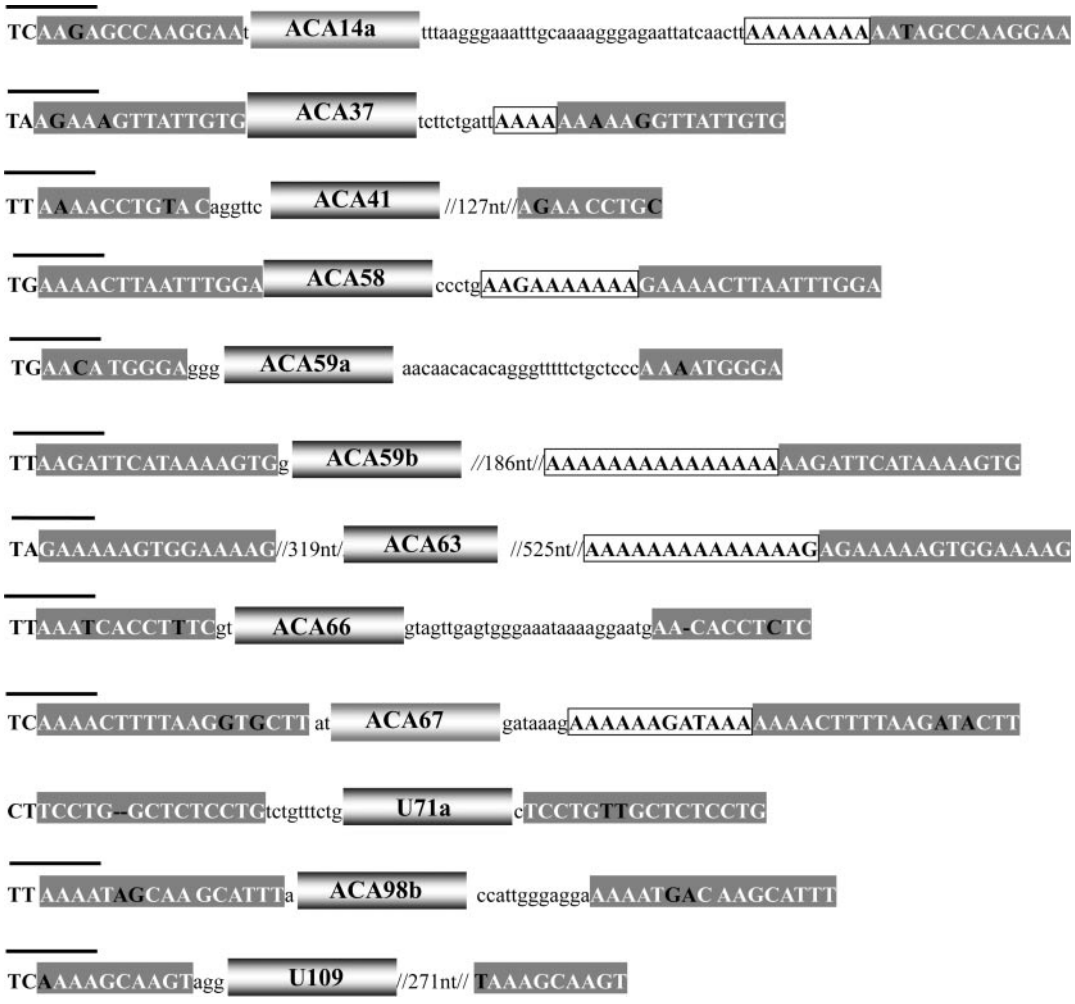


Figure 2. Some previously reported H/ACA snoRNA genes with retrogene hallmarks. Schematic representation of the H/ACA RNA sequences. poly(A) and TSD are in open and closed boxes, respectively. The L1 consensus recognition site (TTAAAA) is indicated at the 5′ end.

involvement of the L1 retroposition machinery in the formation of the H/ACA retrogene. Notably, 39 (19%) of H/ACA RNA-related retrogenes were shortened at their 5′ end (Table 1), presumably because of premature termination of the reverse transcription step. However, there are a few H/ACA RNA-related retrogenes without satisfactory L1 signature, which lack either a poly (A) tail (Figure 1F) or T_2A_4 target site overlapping a TSD (Figure 1G). The existence of tailless retrogenes were reported recently (27), suggesting a variant mechanism for the biogenesis of retrosequences.

Closer inspection of the H/ACA snoRNA-related retrogenes and their flanking sequences revealed that, in same cases, the H/ACA snoRNA-related retrogene had been disrupted by independent integration of an Alu element (Figure 1H). In these cases, allowing for virtual removal of the Alu insertion revealed a 'repaired' retrogene. In other cases, Alu sequence was inserted in the place between H/ACA RNA retrogene and the 3′-TSD (Figure 1I). This suggests that at these sites the H/ACA RNAs were inserted before the integration of the Alu elements. Interestingly, one chimeric retrogene composed of H/ACA sequence fused at its 3′ termini with Alu element, was found (Figure 1J), which was probably formed due to template switching (28) from Alu RNA to H/ACA RNA during reverse transcription and then the fused transcript was integrated into the human genome. A number of retrogenes were reported to result from template switching, including those containing U6, 5S rRNA or 7SL rRNA fused at their 3′ termini with Alu elements (24).

Some previously identified snoRNAs resulted from retrotransposition

Closer analysis of the upstream and downstream region of previously identified snoRNAs showed that ACA14a, ACA37, ACA41, ACA58, ACA59, ACA59b, ACA63, ACA66, ACA67, U71a, ACA98b and U109, are encoded by retrogenes (Figure 2). These box H/ACA RNAs were cloned from a HeLa cell extract immunoprecipitated with an anti-GAR1 antibody (18) or their expression were verified by Northern blot and primer extension (8,13,15). Clearly, these snoRNAs were formed by retrotransposition in the course of primate evolution, for example, the data obtained in this study suggest that the *ACA63* gene originated as the result of retroposition of the ACA63b copy. First, ACA63b is found in corresponding locations on the human, chimpanzee and mouse genomes. Then, human and chimpanzee

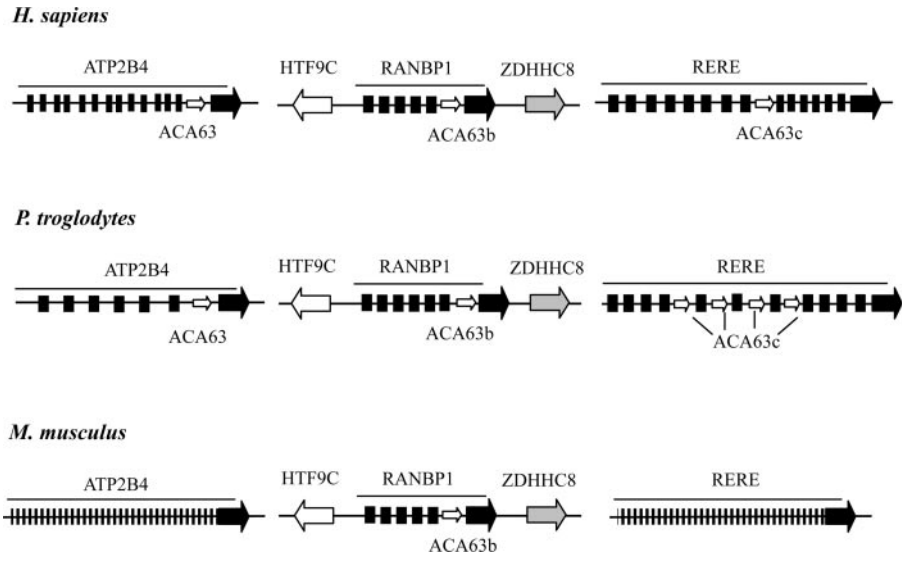


Figure 3. Amplification of ACA63b snoRNA in primate. ACA63 sequence (small arrow) is located within an intron of the orthologous host genes. Additional copies (ACA63 and ACA63c) were generated in the primate lineage. Exons are represented by boxes. The cartoon is not drawn to scale. ATP2B4: ATPase, Ca++ transporting, plasma membrane 4. HTF9C: HpaII tiny fragments locus 9C. RANBP1: RAN binding protein 1. ZDHHC8: zinc finger, DHHC-type containing 8. RERE: arginine-glutamic acid dipeptide (RE) repeats.

ATP2B4 and *RERE* genes encode ACA63 and another retrogene ACA63c in their introns, respectively, while the homologous genes of mouse are devoid of any ACA63-like sequence (Figure 3). Furthermore, comparison and alignment of the two loci ACA63/ACA63b from all available primate sequences revealed that the *Otolemur garnettii* ACA63 locus shows clean absence of the ACA63 along with its retroposed 3′-and 5′-flanking nucleotides (Supplementary Figure 1a). This convincing evidence indicates that human ACA63b that we found in this work is an evolutionary conserved snoRNA widely presented in vertebrates and retrotransposition of ACA63b occurred in primate after the rodent/primate divergence during the course of evolution. Interestingly, there are 4 ACA63c copies with obvious target site duplications (TSDs) in the chimp *RERE* gene, which probably resulted from a single retroposition event into this gene, followed by local segmental duplications.

In vertebrates, sequences encoding H/ACA are generally located in introns of their host gene, in the same orientation. So far, in vertebrates, an intron can carry only one snoRNA gene, but a host gene can carry several different snoRNA genes in different introns (16). The evolutionary analysis of H/ACA RNA genes within the introns of orthologous genes in six vertebrate species showed that a number of snoRNA genes in different introns of a host gene probably resulted from retrotransposition, for example, the *H.sapiens*, *Pan troglodytes, Mus musculus, Rattus norvegicus* and *Canis familiaris* EIF4A2 gene orthologs host three snoRNA genes, HBI-61, E3 and ACA4 in different introns, respectively, while *Gallus gallus* only contains HBI-61 in its orthologous gene (Figure 4A). Similarly, the RPSA genes in all aforementioned mammals host two H/ACA genes, E2 and ACA6 in different introns; however, *G.gallus* is devoid of snoRNAs in the orthologous gene (Figure 4A). Notably, human and chimpanzee ACA4, E2 and E3 are flanked by TSD of >10 nt (data not shown). Although those TSD with a few nucleotide changes, one of these TSDs' ancestral states was present in the tenrec, *Echinops telfairi* ACA4 (Figure 4B), suggesting ACA4 and E3 in EIF4A2 and E2 in RPSA in mammal were resulted from retroposition after the mammal/aves divergence. In addition, there are some host genes which carry several paralogous snoRNA genes in different introns, such as in the *TBRG4* gene (Figure 4A). The amplification of ACA5 in the host gene most likely did not occur via retroposition because insertions of retroposed sequences are virtually random and should not lead to accumulations in neighboring introns (11).

Structures and expression of box H/ACA-related RNAs

Up to date, more than 100 H/ACA RNAs have been found in *H.sapiens* (16). In this study, we found at least two-thirds of these human H/ACA RNA genes have one or more related copies (Table 1). Remarkably, U70 has 21 related copies including six truncated sequences, and another snoRNA gene, U40, exhibits 13 related copies with six truncated sequences. Alignments of these novel H/ACA RNA-related sequences with their orthologs previously reported revealed numerous sequence changes, including small insertions or deletions, which occurred frequently in less important regions, and occasionally in the conserved elements such as box H and ACA. Despite showing sequence variation to some extent, out of 202 box H/ACA RNA-related sequences, 64 can be folded to the typical secondary structure of the box H/ACA RNA family, i.e. the hairpin–hinge–hairpin–tail structure (Supplementary Figure 2), among which 30 were recognized as functional homologs of their corresponding box H/ACA RNAs previously reported according to the relationship between the structure and function of snoRNA, while the remainder did not show any complementarity to either rRNAs or snRNAs due to the sequence diversification and therefore were recognized as orphan H/ACA RNAs.

Retroposition generated for most box H/ACA RNA genes additional copies, quite a number might be functional. Due to

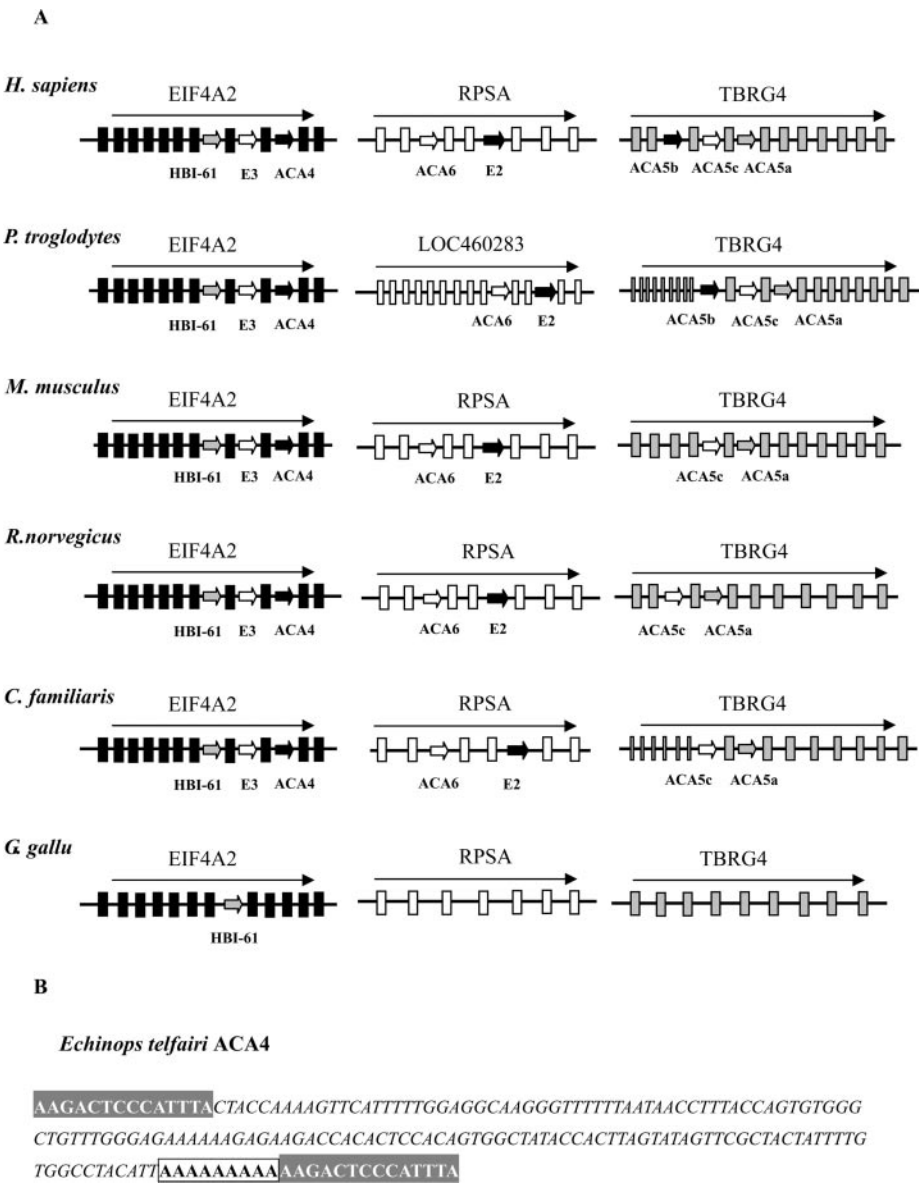


Figure 4. Phylogenetic analysis of some H/ACA RNA genes. (**A**) Presence/absence of H/ACA RNA genes within the introns of orthologous host genes in six vertebrates. Each snoRNA sequence (small arrow) is located within an intron of the indicated genes. Exons are represented by boxes. The cartoon is not drawn to scale. EIF4A2: eukaryotic translation initiation factor 4A, isoform 2. RPSA: ribosomal protein SA. TBRG4: transforming growth factor beta regulator 4. (**B**) Retrogene ACA4 in *Echinops telfairi*. H/ACA RNA sequences are in italics, poly (A) and TSD are in opened and closed boxes, respectively.

cross-hybridization in Northern blot analysis, it could not be assessed if all the 64 box H/ACA RNA-related sequences with typical features of the box H/ACA RNA family are indeed expressed in human tissues. Therefore, we performed BLAST searches of all the 64 box H/ACA RNA-related sequences against EST databases and found that of 11, the corresponding ESTs were detected in EST databases and 5 were shown to be expressed in more than one human tissue (Table 2). Of course, identification of ESTs is not necessarily an indication for the presence of processed and functional snoRNAs. Notably, U107f is located in an intron of a protein gene coding for nudix (nucleoside diphosphate linked moiety X)-type motif 13, but expressed from the opposite strand (Figure 5) and EST database searches revealed that it can be expressed in liver and spleen, even in melanotic melanoma (Table 2). It is not clear whether U107f has a functional role as an antisense regulator for the expression of the protein-coding gene.

DISCUSSION

We have identified in the human genome databases 202 novel box H/ACA RNA-related sequences 0–20% diverged from their corresponding genes reported previously and belonging to 61 box H/ACA RNA types (Table 1), which shows that most human box H/ACA RNA have multiple copies. In contrast to Arabidopsis and rice, where many snoRNAs are found in multiple copies mainly resulting from two different mechanisms: large chromosomal duplications and small tandem duplications producing polycistronic genes (29),

Table 2. Box H/ACA RNA-related genes expressed in human tissues detected in EST databases

Name	GenBank accession no.	EST	Tissue
ACA12b	AL645729	BQ708140	Spleen
ACA15b	AC073107	DB218848	Trachea
ACA15c	AC073107	DB218848	Trachea
ACA58c	AL590431	DW429803	Liver
ACA63b	AC006549	CN275435	Embryonic stem cell, retinoic acid and mitogen-treated hes cell line H7
		AK097659	Testis
		AK094410	Cerebellum
ACA64c	AC097376	DA572426	Whole embryo, mainly body
U68b	AC008474	BQ423961	Retinoblastoma
		BE672593	Lung carcinoid
U107b	AL049732	DB287809	Uterus
		CN267974	Embryonic stem cells, cell lines H1, H7 and H9
U107c	AL928717	H08107	Infant brain
U107d	AL928717	H08107	Infant brain
		AK094541	Amygdala
		CN389247	Embryonic stem cells
U107f	AC016394	CB162932	Liver
		BQ224195	Melanotic melanoma
		BX096147	Liver and spleen

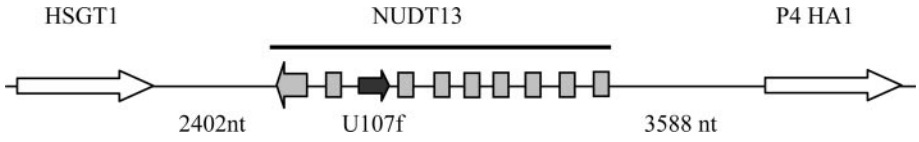


Figure 5. Genomic location of U107f in *H.sapiens*. SnoRNA genes are shown by black arrows, protein-coding genes by non-filled and gray arrows (not drawn to scale). The length of intergenic spacers is also indicated.

human multiple box H/ACA copies mainly result from retroposition. Out of 202 box H/ACA RNA-related sequences identified in this work, 182 have the typical structures of retrogene, and the figure of H/ACA retrogene seems to be underestimated, inasmuch as retrogenes >20% diverged from their corresponding genes are not included in our analysis.

The genomes of the chimpanzee and man share ~96% of box H/ACA RNA-related sequences at identical locations, and only ~4% are thus hominin-specific, having arisen in our genome since the divergence from chimpanzee. On the contrary, the genomes of the mouse contains only ~50% box H/ACA RNA-related sequences relative to man and some sequences were found in different genomic regions, suggesting that most of the H/ACA RNA-related sequences in primate occurred after the rodent/primate divergence. To elucidate the mechanism of H/ACA snoRNA propagation in primates, we analyzed all ape-specific events (those duplicated in human and chimp but not in rhesus monkey) using presence/absence patterns, and found that among nine ape-specific events (ACA1b, ACA10b, ACA40g, ACA40n, ACA43b, ACA51b, ACA57b, ACA64c and U67c), all but one originated from retroposition (Supplementary Figure 1C), suggesting that duplications of most H/ACA snoRNAs in primates are indeed bona fide events mediated by retroposition. In addition, retroposition of different H/ACA RNAs occurred at different stage of primate evolution (Supplementary Figure 1). Notably, the sequence of human-specific retrogene ACA59b is completely identical to ACA59, pointing to a very recent origin of the snoRNA retrogene ACA59b and suggesting, that retrotransposition of snoRNAs still continues to the present day in the human lineage.

Multiple studies have suggested a high rate of retroposition on the primate and rodent lineages (30–32), probably driven by the activity of L1 retrotransposable elements (33). Our results also show the involvement of the L1 retroposition machinery in the formation of human H/ACA retrogenes. Retroposition was commonly thought to generate nonfunctional gene copies (retropseudogenes) that accumulate disablements such as premature stop codons and frameshift mutations for protein-coding genes (34), because the copied mRNA is generally lacking regulatory elements. However, Brosius (35,36) predicted that retrogenes can insert next to resident promoter/enhancer elements and thus escape transcriptional silencing. Indeed, researchers have recently shown that retroposition has generated a significant number of new functional genes (retrogenes) in mammalian genomes (37,38). Similarly, some of the retrogenes derived from H/ACA RNAs appear to be functional genes. First, nearly 50% H/ACA retrogenes found in this work are intronic, encoded within protein-coding genes. Like previously identified intronic snoRNAs (39–41), intronic retrogenes can be co-transcribted with their host genes and then released from excised, debranched introns by exonucleolytic trimming. Furthermore, unlike protein-coding genes, snoRNA retrogenes do not accumulate disablements such as premature stop codons and frameshift mutations. Importantly, some snoRNA retrogenes, even when located in the antisense orientation to their host gene (ACA107f) or in intergenic region (ACA64c), have typical H/ACA RNA structure and can be expressed in human tissues. In addition, for some H/ACA genes retroposition generated more copies and the process may also have provided abundant raw material for the formation of new genes. Therefore it appears that retroposition is one of the

ways of novel snoRNA gene formation. In line with the notion, some previously reported box H/ACA RNA genes apparently resulted from retrotransposition of different box H/ACA RNAs (Figures 2–4).

SUPPLEMENTARY DATA

Supplementary data are available at NAR online.

ACKNOWLEDGEMENTS

The authors thank Donggen Zhou for help with the analysis of secondary structures of RNA. This work was supported by China National Science Foundation 30660042. Funding to pay the Open Access publication charges for this article was provided by the Key Laboratory of Biochemistry and Molecular Biology of Jiangxi Province, China.

Conflict of interest statement. None declared.

REFERENCES

1. Darzacq,X., Jady,B.E., Verheggen,C., Kiss,A.M., Bertrand,E. and Kiss,T. (2002) Cajal body-specific small nuclear RNAs: a novel class of 2′-O-methylation and pseudouridylation guide RNAs. *EMBO J.*, **21**, 2746–2756.
2. Tang,T.H., Bachellerie,J.P., Rozhdestvensky,T., Bortolin,M.L., Huber,H., Drungowski,M., Elge,T., Brosius,J. and Huttenhofer,A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc. Natl Acad. Sci. USA*, **99**, 7536–7541.
3. Ganot,P., Bortolin,M.L. and Kiss,T. (1997) Site-specific pseudouridine formation in eukaryotic pre-rRNAs is guided by small nucleolar RNAs. *Cell*, **89**, 799–809.
4. Bachellerie,J.P., Cavaille,J. and Huttenhofer,A. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
5. Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.
6. Atzorn,V., Fragapane,P. and Kiss,T. (2004) U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production. *Mol. Cell. Biol.*, **24**, 1769–1778.
7. Mishra,R.K. and Eliceiri,G.L. (1997) Three small nucleolar RNAs that are involved in ribosomal RNA precursor processing. *Proc. Natl Acad. Sci. USA*, **94**, 4972–4977.
8. Schattner,P., Barberan-Soler,S. and Lowe,T.M. (2006) A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA*, **12**, 15–25.
9. Torchet,C., Badis,G., Devaux,F., Costanzo,G., Werner,M. and Jacquier,A. (2005) The complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in *Saccharomyces cerevisiae*. *RNA*, **11**, 928–938.
10. Schattner,P., Decatur,W.A., Davis,C.A., Ares,M.,Jr, Fournier,M.J. and Lowe,T.M. (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.*, **32**, 4281–4296.
11. Zemann,A., op de Bekke,A., Kiefmann,M., Brosius,J. and Schmitz,J. (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.*, **34**, 2676–2685.
12. Huttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.
13. Gu,A.D., Zhou,H., Yu,C.H. and Qu,L.H. (2005) A novel experimental approach for systematic identification of box H/ACA snoRNAs from eukaryotes. *Nucleic Acids Res.*, **33**, e194.
14. Li,S.G., Zhou,H., Luo,Y.P., Zhang,P. and Qu,L.H. (2005) Identification and functional analysis of 20 Box H/ACA small nucleolar RNAs (snoRNAs) from *Schizosaccharomyces pombe*. *J. Biol. Chem.*, **280**, 16446–16455.
15. Vitali,P., Royo,H., Seitz,H., Bachellerie,J.P., Huttenhofer,A. and Cavaille,J. (2003) Identification of 13 novel human modification guide RNAs. *Nucleic Acids. Res.*, **31**, 6543–6551.
16. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids. Res.*, **34**, D158–D162.
17. Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.*, **11**, 941–956.
18. Kiss,A.M., Jady,B.E., Bertrand,E. and Kiss,T. (2004) Human box H/ACA pseudouridylation guide RNA machinery. *Mol. Cell. Biol.*, **24**, 5797–5807.
19. Pelczar,P. and Filipowicz,W. (1998) The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5′-terminal oligopyrimidine gene family. *Mol. Cell. Biol.*, **18**, 4509–4518.
20. Ostertag,E.M. and Kazazian,H.H.,Jr (2001) Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.*, **35**, 501–538.
21. Kazazian,H.H.,Jr (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.
22. Wei,W., Gilbert,N., Ooi,S.L., Lawler,J.F., Ostertag,E.M., Kazazian,H.H., Boeke,J.D. and Moran,J.V. (2001) Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.*, **21**, 1429–1439.
23. Esnault,C., Maestre,J. and Heidmann,T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.*, **24**, 363–367.
24. Buzdin,A., Gogvadze,E., Kovalskaya,E., Volchkov,P., Ustyugova,S., Illarionova,A., Fushan,A., Vinogradova,T. and Sverdlov,E. (2003) The human genome contains many types of chimeric retrogenes generated through *in vivo* RNA recombination. *Nucleic Acids Res.*, **31**, 4385–4390.
25. Perreault,J., Noel,J.F., Briere,F., Cousineau,B., Lucier,J.F., Perreaultm,J.P. and Boire,G. (2005) Retropseudogenes derived from the human Ro/SS-A autoantigen-associated hY RNAs. *Nucleic Acids Res.*, **33**, 2032–2041.
26. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
27. Schmitz,J., Churakov,G., Zischler,H. and Brosius,J. (2004) A novel class of mammalian-specific tailless retropseudogenes. *Genome Res.*, **14**, 1911–1915.
28. Brosius,J. (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica*, **107**, 209–238.
29. Barneche,F., Gaspin,C., Guyot,R. and Echeverria,M. (2001) Identification of 66 box C/D snoRNAs in *Arabidopsis thalana*: Extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2′-O-methyltion sites. *J. Mol. Biol.*, **311**, 57–73.
30. Zhang,Z., Harrison,P.M., Liu,Y. and Gerstein,M. (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.*, **13**, 2541–2558.
31. Ohshima,K., Hattori,M., Yada,T., Gojobori,T., Sakaki,Y. and Okada,N. (2003) Wholegenome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol.*, **4**, R74.
32. Zhang,Z., Carriero,N. and Gerstein,M. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.*, **20**, 62–67.
33. Esnault,C., Maestre,J. and Heidmann,T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.*, **24**, 363–367.
34. Mighell,A.J., Smith,N.R., Robinson,P.A. and Markham,A.F. (2000) Vertebrate pseudogenes. *FEBS Lett.*, **468**, 109–114.
35. Brosius,J. (1991) Retroposons—seeds of evolution. *Science*, **251**, 753.
36. Brosius,J. (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, **238**, 115–134.
37. Emerson,J.J., Kaessmann,H., Betran,E. and Long,M. (2004) Extensive gene traffic on the mammalian X chromosome. *Science*, **303**, 537–540.

38. Vinckenbosch,N., Dupanloup,I. and Kaessmann,H. (2006) Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl Acad. Sci. USA*, **103**, 3220–3225.
39. Tycowski,K.T., Shu,M.D. and Steitz,J.A. (1993) A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3. *Genes Dev.*, **7**, 1176–1190.
40. Kiss,T. and Filipowicz,W. (1993) Small nucleolar RNAs encoded by introns of the human cell cycle regulatory gene *RCC1*. *EMBO J.*, **12**, 2913–2920.
41. Kiss,T. and Filipowicz,W. (1995) Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns. *Genes Dev.*, **9**, 1411–1424.