Genome-wide analyses of retrogene s 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 archael 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 euarkyotic 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 euarkyotes 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

© 2006 The Author(s).

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
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%) are retrogenes, 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 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<T>A hexanucleotide preferably recognized by L1 nicking endonuclease, or its derivatives with one or two nucleotide substitutions (Figure 1A–E). These features suggest the
| N | Name     | Genomic placement | Chromosome | Chromosome start position | Identity (%) | Type     | GenBank accession no. |
|---|----------|-------------------|------------|---------------------------|--------------|----------|----------------------|
| 1 | ACA1b    | Intronic          | 8          | 56977836                  | 94.6         | Retrogene | AC046176             |
| 2 | ACA1c    | Intronic          | 2          | 203625253                 | 87.7         | Retrogene | AC022371             |
| 3 | ACA1d    | Intronic          | 16         | 24252145                  | 83.9         | Retrogene | AC004125             |
| 4 | ACA2c    | Intronic          | 2          | 10212650                  | 88.2         | Retrogene | AC022371             |
| 5 | ACA2d    | Intronic          | 16         | 56977836                  | 94.6         | Retrogene | AC046176             |
| 6 | ACA3b    | Intergenic        | 21         | 42175400                  | 91.0         | Retrogene | AL359273             |
| 7 | ACA3-2b  | Intergenic        | 12         | 83172025                  | 80.0         | Retrogene | AC004125             |
| 8 | ACA4b    | Intergenic        | 2          | 36977837                  | 91.0         | Retrogene | AC022371             |
| 9 | ACA4c    | Intronic          | 2          | 19797768                  | 83.9         | Retrogene | AC004125             |
| 10| ACA5c    | Intronic          | 17         | 42175400                  | 91.0         | Retrogene | AC022371             |
| 11| ACA5d    | Intergenic        | 21         | 42175400                  | 91.0         | Retrogene | AC022371             |
| 12| ACA6c    | Intronic          | 11         | 3900374                   | 87.1         | Retrogene | AL359273             |
| 13| ACA6d    | Intronic          | 11         | 73641025                  | 87.1         | Retrogene | AL359273             |
| 14| ACA7b    | Intergenic        | X          | 132014448                 | 95.0         | Retrogene | Z77249               |
| 15| ACA7c    | Intronic          | 12         | 83172025                  | 80.0         | Retrogene | AC004125             |
| 16| ACA7d    | Intronic          | 13         | 73641025                  | 87.1         | Retrogene | AL359273             |
| 17| ACA7e    | Intronic          | X          | 132014448                 | 95.0         | Retrogene | Z77249               |
| 18| ACA7f    | Intronic          | 12         | 83172025                  | 80.0         | Retrogene | AC004125             |
| 19| ACA8b    | Intergenic        | X          | 132014448                 | 95.0         | Retrogene | Z77249               |
| 20| ACA8c    | Intronic          | 17         | 62908764                  | 91.0         | Retrogene | AC004125             |
| 21| ACA8d    | Intronic          | 17         | 62908764                  | 91.0         | Retrogene | AC004125             |
| 22| ACA8e    | Intronic          | 17         | 62908764                  | 91.0         | Retrogene | AC004125             |
| 23| ACA9b    | Intronic          | 8          | 52090134                  | 80.0         | Retrogene | AC004125             |
| 24| ACA9c    | Intronic          | 8          | 52090134                  | 80.0         | Retrogene | AC004125             |
| 25| ACA9d    | Intronic          | 9          | 20776934                  | 89.3         | Retrogene | AC004125             |
| 26| ACA9e    | Intronic          | 9          | 20776934                  | 89.3         | Retrogene | AC004125             |
| 27| ACA10d   | Intronic          | 5          | 118032764                 | 80.0         | Retrogene | AC004125             |
| 28| ACA10e   | Intronic          | 5          | 118032764                 | 80.0         | Retrogene | AC004125             |
| 29| ACA11b   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 30| ACA11c   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 31| ACA11d   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 32| ACA11e   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 33| ACA11f   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 34| ACA11g   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 35| ACA11h   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 36| ACA11i   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 37| ACA11j   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 38| ACA11k   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| 39| ACA11l   | Intronic          | 1          | 28052145                  | 80.0         | Retrogene | AC004125             |
| N  | Name       | Genomic placement | Chromosome | Chromosome start position | Identity (%) | Type   | GenBank accession no. |
|----|------------|-------------------|------------|---------------------------|--------------|--------|----------------------|
| 71 | ACA40m     | Intergenic        | X          | 123159276                 | 89.8d        | Retrogene | AL391241             |
| 72 | ACA40n     | Intergenic        | 7          | 99387638                  | 92.6         | Retrogene | AC004522             |
| 73 | ACA41b     | Intronic          | 14         | 43616740                  | 85.7         | Retrogene | AC090527             |
| 74 | ACA42b     | Intronic          | 11         | 37259407                  | 86.5d        | Retrogene | AL136296             |
| 75 | ACA43c     | Intronic          | 2          | 180507372                 | 80.5         | Retrogene | AC096587             |
| 76 | ACA44b     | Intronic          | 6          | 43619859                  | 92.9         | Retrogene | AL355902             |
| 77 | ACA45c     | Intergenic        | 20         | 41366609                  | 90.6         | Retrogene | AL021395             |
| 78 | ACA46b     | Intronic          | 14         | 77002430                  | 82.6d        | Retrogene | AF111168             |
| 79 | ACA47b     | Intronic          | 1          | 101371297                 | 92.5         | Retrogene | AC093157             |
| 80 | ACA47c     | Intron            | 2          | 19759117                  | 85.1         | Retrogene | AC091546             |
| 81 | ACA47d     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 82 | ACA47e     | Intergenic        | 2          | 115966016                 | 82.2         | Retrogene | AL595213             |
| 83 | ACA47f     | Intergenic        | 1          | 115966016                 | 82.2         | Retrogene | AL595213             |
| 84 | ACA48b     | Intergenic        | X          | 3460160                   | 91.9         | Retrogene | AC141001             |
| 85 | ACA48c     | Intergenic        | 18         | 7291021                   | 85.9         | Retrogene | BX326644             |
| 86 | ACA48d     | Intergenic        | 12         | 55541426                  | 87.5         | Retrogene | AC009270             |
| 87 | ACA48e     | Intergenic        | 16         | 66780691                  | 85.9         | Retrogene | AC013549             |
| 88 | ACA48f     | Intergenic        | 2          | 122541951                 | 83.0         | Retrogene | AC097149             |
| 89 | ACA48g     | Intergenic        | 15         | 101624078                 | 82.3         | Retrogene | AC044913             |
| 90 | ACA48h     | Intergenic        | 7          | 101624078                 | 82.3         | Retrogene | AC044913             |
| 91 | ACA48i     | Intergenic        | 1          | 101371297                 | 82.5         | Retrogene | AC044913             |
| 92 | ACA48j     | Intergenic        | 20         | 41366609                  | 85.1         | Retrogene | AC044913             |
| 93 | ACA48k     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 94 | ACA48l     | Intergenic        | 2          | 115966016                 | 82.2         | Retrogene | AL592436             |
| 95 | ACA48m     | Intergenic        | 1          | 115966016                 | 82.2         | Retrogene | AL592436             |
| 96 | ACA48n     | Intergenic        | 2          | 115966016                 | 82.2         | Retrogene | AL592436             |
| 97 | ACA48o     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 98 | ACA48p     | Intergenic        | 2          | 115966016                 | 82.2         | Retrogene | AL592436             |
| 99 | ACA48q     | Intergenic        | 1          | 115966016                 | 82.2         | Retrogene | AL592436             |
| 100| ACA49a     | Intergenic        | 2          | 101371297                 | 82.5         | Retrogene | AC044913             |
| 101| ACA49b     | Intergenic        | 1          | 101371297                 | 82.5         | Retrogene | AC044913             |
| 102| ACA49c     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 103| ACA49d     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 104| ACA49e     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 105| ACA49f     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 106| ACA49g     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 107| ACA49h     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 108| ACA49i     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 109| ACA49j     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 110| ACA49k     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 111| ACA49l     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 112| ACA49m     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 113| ACA49n     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 114| ACA49o     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 115| ACA49p     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 116| ACA49q     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 117| ACA49r     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 118| ACA49s     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 119| ACA49t     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 120| ACA49u     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 121| ACA49v     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 122| ACA49w     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 123| ACA49x     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| 124| ACA49y     | Intergenic        | 11         | 65956813                  | 84.2         | Retrogene | AL592436             |
| 125| ACA49z     | Intergenic        | 2          | 115966016                 | 82.5         | Retrogene | AC044913             |
| N  | Name   | Genomic placement | Chromosome | Chromosome start position | Identity (%) | Type   | GenBank accession no. |
|----|--------|-------------------|------------|---------------------------|--------------|--------|----------------------|
| 141| U68b   | Intronic          | 19         | 37791083                  | 91.1         | Retrogene | AC008474             |
| 142| U68c   | Intronic          | 5          | 158590783                 | 82.2         | Retrogene | AC134043             |
| 143| U68d   | Intronic          | X          | 24061225                  | 84.5         | Retrogene | AC079169             |
| 144| U69b   | Intronic          | 17         | 8173626                   | 84.2         | Retrogene | AC008053             |
| 145| U70b   | Intronic          | 1          | 200214211                 | 93.5         | Retrogene | AC099676             |
| 146| U70c   | Intronic          | 2          | 215419918                 | 95.7         | Retrogene | AC016708             |
| 147| U70d   | Intronic          | 2          | 61497882                  | 95.7         | Retrogene | AC016894             |
| 148| U70e   | Intronic          | 5          | 17025729                  | 92.8         | Retrogene | AL832824             |
| 149| U70f   | Intronic          | 5          | 87714345                  | 86.3         | Retrogene | AC091826             |
| 150| U70g   | Intronic          | 8          | 8856495                   | 89.9         | Retrogene | AC087763             |
| 151| U70h   | Intronic          | 8          | 3517105                   | 93.7         | Retrogene | AC013603             |
| 152| U70i   | Intronic          | 5          | 118883203                 | 91.4         | Retrogene | AL835608             |
| 153| U70j   | Intronic          | 11         | 20430163                  | 87.7         | Retrogene | AP000893             |
| 154| U70k   | Intronic          | 11         | 67307282                  | 84.2         | Retrogene | AC016394             |
| 155| U70l   | Intronic          | 12         | 79797254                  | 82.5         | Retrogene | AC002224             |
| 156| U70m   | Intronic          | 12         | 203966973                 | 89.6         | Retrogene | AC119673             |
| 157| U70n   | Intronic          | 12         | 74369190                  | 92.7         | Retrogene | AC015550             |
| 158| U70o   | Intronic          | 12         | 120029229                 | 85.5         | Retrogene | AC078820             |
| 159| U70p   | Intronic          | 16         | 70289971                  | 89.1         | Retrogene | AC010653             |
| 160| U70q   | Intronic          | 16         | 48743054                  | 92.0         | Retrogene | AC127610             |
| 161| U70r   | Intronic          | 17         | 23373483                  | 86.3         | Retrogene | AC090287             |
| 162| U70s   | Intronic          | 17         | 25128801                  | 91.7         | Retrogene | AC023389             |
| 163| U70t   | Intronic          | 17         | 3015432                   | 94.8         | Retrogene | AP005431             |
| 164| U70u   | Intronic          | 17         | 9791682                   | 99.0         | Retrogene | AC008752             |
| 165| U70v   | Intronic          | 17         | 3136041                   | 90.6         | Retrogene | AP000039             |
| 166| U71e   | Intronic          | 10         | 79797254                  | 82.5         | Retrogene | AC012560             |
| 167| U71b   | Intronic          | 10         | 161897415                 | 91.1         | Retrogene | AC000922             |
| 168| U72c   | Intronic          | 1          | 203966973                 | 89.6         | Retrogene | AC119673             |
| 169| U72d   | Intronic          | 1          | 222433982                 | 93.4         | Retrogene | AC092809             |
| 170| U72e   | Intronic          | 2          | 139985468                 | 85.8         | Retrogene | AC016710             |
| 171| U72f   | Intronic          | 3          | 173971758                 | 83.6         | Retrogene | AC108667             |
| 172| U72g   | Intronic          | 2          | 104716137                 | 83.1         | Retrogene | AC068057             |
| 173| U72h   | Intronic          | 8          | 132514215                 | 87.3         | Retrogene | AC104040             |
| 174| U78b   | Intronic          | 8          | 21506474                  | 90.1         | Retrogene | AC005632             |
| 175| U107b  | Intronic          | X          | 54970463                  | 98.5         | Retrogene | AC049732             |
| 176| U107c  | Intronic          | X          | 51823183                  | 85.5         | Retrogene | BX537154             |
| 177| U107d  | Intronic          | X          | 51930457                  | 85.5         | Retrogene | AL928717             |
| 178| U107e  | Intronic          | 15         | 43294396                  | 92.0         | Retrogene | AC051619             |
| 179| U107f  | Intronic          | 16         | 74555844                  | 95.3         | Retrogene | AC016394             |
| 180| U107g  | Intronic          | 14         | 90662522                  | 94.2         | Retrogene | AC007374             |
| 181| U107h  | Intronic          | X          | 47132992                  | 90.7         | Retrogene | AC091503             |
| 182| U107i  | Intronic          | 14         | 69340688                  | 83.8         | Retrogene | AC157789             |
| 183| U107j  | Intronic          | 21         | 34750278                  | 81.4         | Retrogene | AP000053             |
| 184| U107k  | Intronic          | 4          | 120710302                 | 86.1         | Retrogene | AC080089             |
| 185| U108b  | Intronic          | 8          | 131244032                 | 82.5         | Retrogene | AC103725             |
| 186| U108c  | Intronic          | 2          | 55646343                  | 85.8         | Retrogene | AC015982             |
| 187| U108d  | Intronic          | 1          | 191293034                 | 89.0         | Retrogene | AC136370             |
| 188| U109a  | Intronic          | 16         | 2545357                   | 84.6         | Retrogene | AP005061             |
| 189| U109b  | Intronic          | 16         | 6722295                   | 84.6         | Retrogene | AC126773             |
| 190| U109c  | Intronic          | 16         | 75570241                  | 80.0         | Retrogene | AC007099             |
| 191| U109d  | Intronic          | 18         | 52594922                  | 87.1         | Retrogene | AC104066             |
| 192| HBI-6a | Intronic          | 18         | 52594922                  | 87.1         | Retrogene | AC104066             |
| 193| HBI-6b | Intronic          | 18         | 17545945                  | 87.2         | Chimera   | AC091038             |

* Retrogene with common hairpin–hinge–hairpin–tail secondary structure.
* Retrogene distributed on the antisense orientation of protein-coding genes.
* Identity to the corresponding consensus sequence.
* 5'-truncated box H/ACA RNA-related sequences.
* 3'-truncated or 3' sequences are different from the corresponding consensus sequences.
* Retrogenes with poly (A) tails at their 3' ends.
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 ACA18e and a number of retroposed nucleotides on the 3'-flanks. (H) 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.
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 T2A4 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 some 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 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 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
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 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; 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 revealed 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
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),
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 retrotroposition 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 disablers 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-transcribed 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

| 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                  |
| ACA64c | AC097376              | DA572426 | Cerebellum               |
| U68b   | AC008474              | BQ423961 | Retinoblastoma          |
| U107b  | AL049732              | BE672593 | Lung carcinoid          |
| U107c  | AL928717              | CN267974 | Embryonic stem cells, cell lines H1, H7 and H9 |
| U107d  | AL928717              | H08107  | Infant brain            |
|        |                       | H08107  | Infant brain            |
|        |                       | AK094541 | Amygdala                |
|        |                       | CN389247 | Embryonic stem cells     |
|        |                       | CB162932 | Liver                   |
|        |                       | BQ224195 | Melanotic melanoma      |
|        |                       | BX096147 | Liver and spleen        |

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

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.
ways of novel snRNA 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.

Conflicts of interest statement. None declared.

REFERENCES
1. Darzaqz,X., Jady,B.E., Verheggen,C., Kiss,A.M., Bertrand,E. and Kiss,T. (2002) Caja body-specific small nuclear RNAs: a novel class of 2′-O-methylation and pseudouridylidy linkage RNAs. EMBO J., 21, 2746–2756.
2. Tang,T.H., Bachellerie,J.P., Rozhdestvensky,T., Bortolin,M.L., Huber,H., Drungowski,M., Elg,T., Brosius,J. and Hüttenhofer,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 nuclear RNAs. Cell, 89, 799–809.
4. Bachellerie,J.P., Cavaillé,J. and Hüttenhofer,A. (2002) The expanding snoRNA world. Biochimie, 84, 775–790.
5. Kiss,T. (2002) Small nuclear 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 component of the H/ACA pseudouridylation guide RNA machinery. Mol. Cell. Biol., 24, 1429–1439.
7. Schattner,P., Decatur,W.A., Davis,C.A., Ares,M.,Jr, Fournier,M.J. and Huttenhofer,A. (2006) A computational screen for mammalian pseudouridylation guide H/ACA RNAs in mouse. Genetica, 133, 2032–2041.
8. Zaker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31, 3406–3415.
9. Schmitz,J., Churakov,G., Zischler,H. and Brosius,J. (2004) A novel class of mammalian-specific tailless retropseudogenes. Genome Res., 14, 1911–1915.
10. Brosius,J. (1999) Genomes were forged by massive bombardments with retroelements and retropseudogenes. Genetica, 107, 209–238.
11. Barneche,F., Gaspin,C., Guyot,R. and Echeverria,M. (2001) Identification of 66 box C/D snoRNAs in Arabidopsis thaliana: Extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2′-O-methylation sites. J. Mol. Biol., 311, 57–73.
12. Ohshima,K., Hattori,M., Yada,T., Gojobori,T., Sakaki,Y. and Okada,N. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol., 4, R74.
13. 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.
14. Ozolina,K., Hattori,M., Yada,T., Gojobori,T., Sakaki,Y. and Okada,N. (2005) A novel approach for systematic identification of box H/ACA snoRNAs from eukaryotes. Nucleic Acids Res., 33, e194.
15. 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 nuclear RNAs (snORANs) from Schizosaccharomyces pombe. J. Biol. Chem., 280, 16446–16455.

570 Nucleic Acids Research, 2007, Vol. 35, No. 2
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.