Published online April 7, 2005

2032–2041 *Nucleic Acids Research, 2005, Vol. 33, No. 6*
doi:10.1093/nar/gki504

Retropseudogenes derived from the human Ro/SS-A autoantigen-associated hY RNAs

Jonathan Perreault[1,2], Jean-François Noël[1,4], Francis Brière[1,2], Benoit Cousineau[1,5], Jean-François Lucier[1], Jean-Pierre Perreault[1,2] and Gilles Boire[1,3,*]

[1]RNA group/Groupe ARN, [2]Department of Biochemistry, [3]Department of Medicine, [4]Department of Microbiology and Infectiology, Faculty of Medicine, Université de Sherbrooke, Sherbrooke, Quebec, J1H 5N4, Canada and [5]Department of Microbiology and Immunology, McGill University, 3775 University Street, Montréal, Quebec, H3A 2B4, Canada

Received January 10, 2005; Revised March 7, 2005; Accepted March 22, 2005

ABSTRACT

We report the characterization in the human genome of 966 pseudogenes derived from the four human Y (hY) RNAs, components of the Ro/SS-A autoantigen. About 95% of the Y RNA pseudogenes are found in corresponding locations on the chimpanzee and human chromosomes. On the contrary, Y pseudogenes in mice are both infrequent and found in different genomic regions. In addition to this rodent/primate discrepancy, the conservation of hY pseudogenes relative to hY genes suggests that they occurred after rodent/primate divergence. Flanking regions of hY pseudogenes contain convincing evidence for involvement of the L1 retrotransposition machinery. Although Alu elements are found in close proximity to most hY pseudogenes, these are not chimeric retrogenes. Point mutations in hY RNA transcripts specifically affecting binding of Ro60 protein likely contributed to their selection for direct *trans* retrotransposition. This represents a novel requirement for the selection of specific RNAs for their genomic integration by the L1 retrotransposition machinery. Over 40% of the hY pseudogenes are found in intronic regions of protein-coding genes. Considering the functions of proteins known to bind subsets of hY RNAs, hY pseudogenes constitute a new class of L1-dependent non-autonomous retroelements, potentially involved in post-transcriptional regulation of gene expression.

INTRODUCTION

Ro ribonucleoproteins (RNPs) are low-abundance autoantigens that are frequently targeted by antibodies from patients with connective tissue diseases, but not from animals with spontaneous autoimmune diseases (1,2). Ro RNPs consist of the non-covalent association of short (70–115 nt) non-coding RNAs of the Y family with a 60 kDa protein (Ro60). The Y RNAs vary in numbers among species (e.g. two in mice, mY1 and mY3; four in humans, hY1, hY3, hY4 and hY5; see Figure 1) and cell types (hY1 and hY4 in erythrocytes; hY1 and hY3 in platelets). The hY3 RNA is the most conserved Y RNA among mammals (3). Proposed roles for Ro60 protein include regulation of translation of ribosomal mRNAs (4), as well as quality control of small RNAs and enhancement of cell survival after exposure to ultraviolet irradiation [reviewed in (5)]. Nonetheless, homozygous animals for deletion of the genes coding for Ro60 exhibit mild phenotypic abnormalities (6), and an autoimmune syndrome that shares some features with systemic lupus erythematosus (7). Still unidentified functions related to the Y RNAs and/or the Ro RNPs themselves are suspected. Indeed, the La protein and additional proteins [heterogeneous nuclear ribonucleoproteins (hnRNP) K and I, nucleolin, and Ro binding protein I (RoBPI)] associate with specific subsets of Y RNAs and/or Ro RNPs (8–11). Most hY RNA-associated proteins are involved in alternative splicing and in regulation of translation of specific mRNAs (12–16).

Mobile (or transposable) elements have largely contributed to shape mammalian genomes. In humans, retrotransposons of the long interspersed element-1 (L1) family and their remnants account for ~17% of the human genome [reviewed in (17,18)].

*To whom correspondence should be addressed. Tel: +1 819 564 5261; Fax: +1 819 564 5265; Email: Gilles.Boire@USherbrooke.ca
Correspondence may also be addressed to Jean-Pierre Perreault. Tel: +1 819 564 5310; Fax: +1 819 564 5340; Email: Jean-Pierre.Perreault@usherbrooke.ca

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

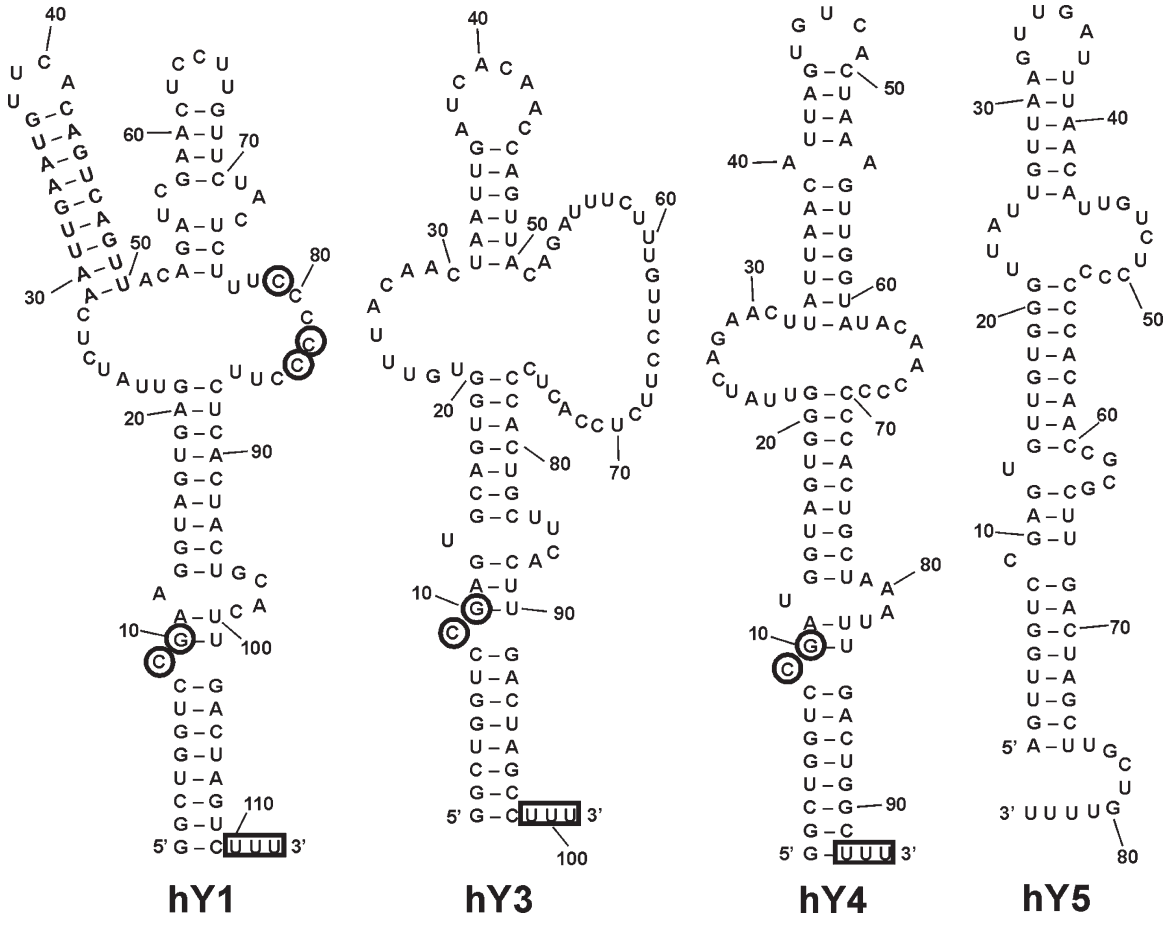

Figure 1. Proposed secondary structures and nucleotide sequences of the four hY RNA, i.e. hY1, hY3, hY4 and hY5 RNAs. The circled nucleotides correspond to the most frequently mutated positions and boxed nucleotides are those most frequently missing in the pseudogenes.

The vast majority (more than 99.8%) of L1s are not mobile, but the average human and mouse genomes contain 60–100 and 3000 retrotransposition-competent L1s, respectively (19,20). Although L1 reverse transcriptase (RT) has a marked *cis* preference (i.e. preferentially retrotransposes L1 elements) (21), it is able to mobilize in *trans* non-autonomous sequences, such as short interspersed nucleotide elements (SINEs) (22). In humans, the most abundant SINE is the ~300 bp Alu element that constitutes 11% of the human genome (i.e. 1.1 million copies). The L1 retrotransposition machinery also participates in genome integration of processed pseudogenes and chimeric retrogenes. Processed pseudogenes arise when cellular mRNAs are reverse transcribed and reinserted at new locations into the genome by the L1 integration machinery (23). Chimeric retrogenes are generated through template switching of the L1 protein (ORF2) during reverse transcription, generating fusions of L1 or Alu elements (3′ end) with small nuclear RNAs (snRNAs), such as U6 (5′ end) (24). The existence of Y pseudogenes, i.e. non-autonomously transcribed Y RNA-related sequences, was previously reported both in mouse and in man genomes (25,26); only a handful of these were characterized (26–28). In addition, the functional gene encoding hY5 RNA itself was proposed to result from a retrotransposition event of the hY3 RNA (29).

We characterized close to 1000 copies of Y RNA pseudogenes in the human and chimpanzee genomes, while mY pseudogenes were seldom found in the mouse. Convincing evidence indicated that the hY retrotransposition events occurred in *trans* using the L1 machinery, likely when point mutations preventing Ro60 protein (and possibly La protein) binding were present in the Y RNA transcripts. Chimeric retrogenes involving hY RNAs were distinctly rare. The age distribution and genomic distribution of hY pseudogenes parallel those of Alu elements, including a preferential localization in gene-rich regions and evidence that these integration events are relatively recent. Similar to Alu elements that recently acquired a novel respectability as mediators of genomic evolution (30), hY RNAs may represent a novel class of L1-dependent non-autonomous retrotransposable elements with potential biological significance.

MATERIALS AND METHODS

Search for homologies

We used the megaBLAST tool on the NCBI website (http://www.ncbi.nlm.nih.gov/BLAST) with a word size of 11 to do our main search for hY pseudogenes on the human genome (NCBI build 34.1). We used a low complexity filter but did not mask for repeats to avoid missing hY pseudogenes. The BLAST hits kept for further analysis were at least half the respective hY length. The same procedure was used to look

for pseudogenes of other non-coding small RNAs. The comparison between different species was performed using BLAT search (http://genome.cse.ucsc.edu/cgi-bin/hgBlat) with the corresponding Y sequences. Y pseudogene sequences found in man were retrieved with a 500 nt offset at each extremity and then searched for in chimpanzee genome (*Pan troglodytes*; NCBI build 1 version 1) using BLAT to compare man and chimpanzee Y pseudogenes.

Sequence variation analysis

All pseudogenes were sequentially aligned with their corresponding hY RNA gene sequence using Matcher (http://www.sanger.ac.uk), and a script allowed the analysis of mutations. Positions varying more than two standard deviations above overall sequence variation were more closely analysed. The percentage of mutations for a given position was calculated. Only the pseudogenes with a nucleotide at that specific position were considered in this statistics (e.g. a pseudogene missing 14 nt at its 5′ end would not be used in statistics for position 9).

GC content

A window of 5 kb was retrieved on each side of the pseudogenes and analysed with a script to calculate the percentage of GC. This script was also used to find the GC content in hY pseudogenes.

Retrieving pseudogene position

Pseudogenes were mapped on human genome using ENSEMBL (http://www.ENSEMBL.org/Homo sapiens; version 18.34.1). Genomic positions were used to look for features in ENSEMBL database. Features retrieved were classified into three groups: exon, intron and intergenic. Pseudogenes found in genes were manually screened for their orientation compared with the gene.

Retrotransposition signature analysis

Based on the signature elements shown in Figure 4A, we designed the 'RTAnalyzer' program to evaluate the probability of retrotransposition by the L1 machinery for a given pseudogene, expressed as the RetroScore (see below). After the 5′ and 3′ ends of the hY pseudogenes were determined, Matcher (http://www.sanger.ac.uk) was used by RTAnalyzer to identify the target site duplications (TSDs) on each side. A poly(A) tail was also looked for between the 3′ end of the hY pseudogene and the 3′ TSD. The consensus endonuclease cleavage site was searched at the 5′ end of the pseudogene with four bases overlapping with the TSD. The TSDs were more heavily weighted in the RetroScore (60% of its maximum) because they represent the most characteristic feature of L1 retrotransposed elements. Because a poly(A) tail is often found in L1 signature, but it tends to shrink with time and is prevalent in the genome, the poly(A) tail accounted for 30% of RetroScore. The endonuclease target site sequence represented only 10% of RetroScore because it is a relatively non-conserved short sequence. In some cases, manual adjustments were required to compensate for false hits owing to AT-rich sequences in putative TSD alignments. The total percentage of missing 5′ ends in pseudogenes was calculated. Only pseudogenes in which at least 10% of the 5′ end of the hY sequence were missing were considered truncated.

The formula to calculate the 'RetroScore' was divided in three 'subscores', each addressing separately the poly(A) tail, the TSDs and the consensus cleavage site, as follows:

$$[A(5\sqrt{L^{A}}) - D^{A}] + [(100L^{TSD})/15 - (13M + 26G) - (10\sqrt{{}_{5'}D^{TSD}}) - (2_{3'}D^{TSD})] + [10S_{freq} - D^{T}].$$

A is the proportion of adenines in the poly(A) tail, L^{A} the length of the poly(A) tail and D^{A} the distance (in nucleotides) between the 3′ end of the sequence homologous to the hY and the poly(A) tail. L^{TSD} is the length of the TSDs, M the number of mismatches and G the number of gaps between the two TSDs, ${}_{5'}D^{TSD}$ is the distance between the 5′ end of the sequence homologous to the hY and the TSD, and ${}_{3'}D^{TSD}$ the distance between the 3′ end of the poly(A) and the TSD. S_{freq} is the relative frequency of a given sequence as a target site of the L1 endonuclease (31) and D^{T} the distance between the 5′ end of the TSD and the target site, considering four overlapping nucleotides. The minimum allowed for each subscore was zero. Poly(A) sequences located more than 15 bases apart from the sequence homologous to hY were not considered. Multipliers were defined arbitrarily according to the relative importance of each variable and adjusted by empirical testing to represent best what might be true retrotransposition events. However, we recognize that true retrotransposition events with very divergent or very small TSDs did not score well. For example, only 70% of a random selection of Alu sequences had a RetroScore of 40 or more (data not shown). Examples of the use of our RetroScore for assessing the degree of conformity of signatures of L1 retrotransposition are attached as Supplementary Material.

Surrounding Alu elements

To look for the eventuality of chimeric retrogenes, 500 bases on each side of the hY pseudogenes were sequentially aligned with the sequences of a number of other small non-coding RNA species (e.g. tRNAs, U RNAs, snoRNAs, miRNAs, rRNAs, etc.). Occasionally, Alu sequences were close enough to hY pseudogenes to move the TSD beyond the window used for our first analysis. RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) was also used to this end.

RESULTS

Identification of 966 hY RNA pseudogenes

A computational, genome-wide search using MegaBlast (http://www.ncbi.nlm.nih.gov/BLAST) (32) was performed for human sequences homologous to various small non-coding RNAs, including the four hY RNAs. When requirements for >75% identity of sequence relative to at least 50% of the length of the corresponding RNA were set, a large number of potential pseudogenes were identified (Table 1). The list included 5 and 35 homologues of the $tRNA^{Arg}$ and $tRNA^{Lys}$, respectively, and 1366 homologues of the multicopy snRNAs (U1, U2, U4, U5 and U6), including 1085 for U6 RNA alone. This was in good agreement with previous studies reporting a high prevalence of U6 RNA pseudogenes, relative to the other U snRNAs (24). The same genome search identified a total of 966 hY RNA-related sequences with a low probability of occurrence (E-value < 0.001 for 99% of the sequences).

The list of these sequences and their genomic positions is appended as Supplementary Material. Careful analysis of the upstream region of the hY-related sequences failed to identify known promoters, indicating that these were likely pseudogenes. The hY homologous sequences were derived from all four hY RNAs, with a marked preponderance (84%) for pseudogenes derived from hY3 and hY1, the most conserved Y RNAs in vertebrates. Pseudogenes related to hY5 RNA, a primate-specific acquisition, represented <1% of the total.

Each of the four hY RNAs is encoded by a single gene, and all four genes are found in close proximity on chromosome 7 (Figure 2, small box and arrow) (29). On the contrary, hY pseudogenes are widely distributed on all human chromosomes (Figure 2). Globally, the number of hY pseudogenes per chromosome was proportional to its DNA length with a maximum of 94 copies on chromosome 1 (Figure 2, inset). However, only one copy was found on chromosome Y, while chromosomes 1, 12 and 17 had some relative excess density of hY pseudogenes. This distribution is very similar to that of Alu repeats and other L1-mediated pseudogenes (33,34), with the exceptions of a greater relative density of hY pseudogenes on chromosome 7 and a lesser relative density on chromosome 19.

Table 1. Number of small non-coding RNAs homologous sequences in the human genome[a]

Pseudogenes	Number
hY1	368
hY3	442
hY4	148
hY5	8
U1	91
U2	46
U3	45
U4	70
U5	29
U6	1085
$tRNA^{Arg}$	5
$tRNA^{Lys}$	35
5S rRNA	659

[a]MegaBlast results using requirements for >75% identity of sequence relative to at least 50% of the length of the corresponding RNA gene, excluding full-length sequences that are 100% identical.

Conservation and age distribution of the hY pseudogenes

None of the hY pseudogene sequences was 100% identical to the corresponding hY functional genes (Figure 3A). This observation may point to the mechanism underlying retrotransposition and help to date the genetic events (see below). Sequence differences between hY pseudogenes and the corresponding hY genes were scattered all over their length and were usually found at each individual position in 10 ± 4% of the homologues (Figure 3A). However, three significant deviations from this apparently random mutation pattern were observed. The first deviation consisted of an almost constant mutation in at least one position of the CG dinucleotide at positions 9 and 10; at least one of these two residues was mutated in over 80% of the pseudogenes corresponding to hY1, hY3 and hY4 RNAs (Figure 1). Positions 9 and 10 correspond to the conserved region known to be essential for Ro60 binding to Y RNAs (35). The second deviation from background was an 80% mutation rate in residues forming the polypyrimidine-rich region in the middle of hY1 RNA (Figure 1), a region involved in the binding of hnRNP K and PTB to this RNA [(10); F. Brière and G. Boire,

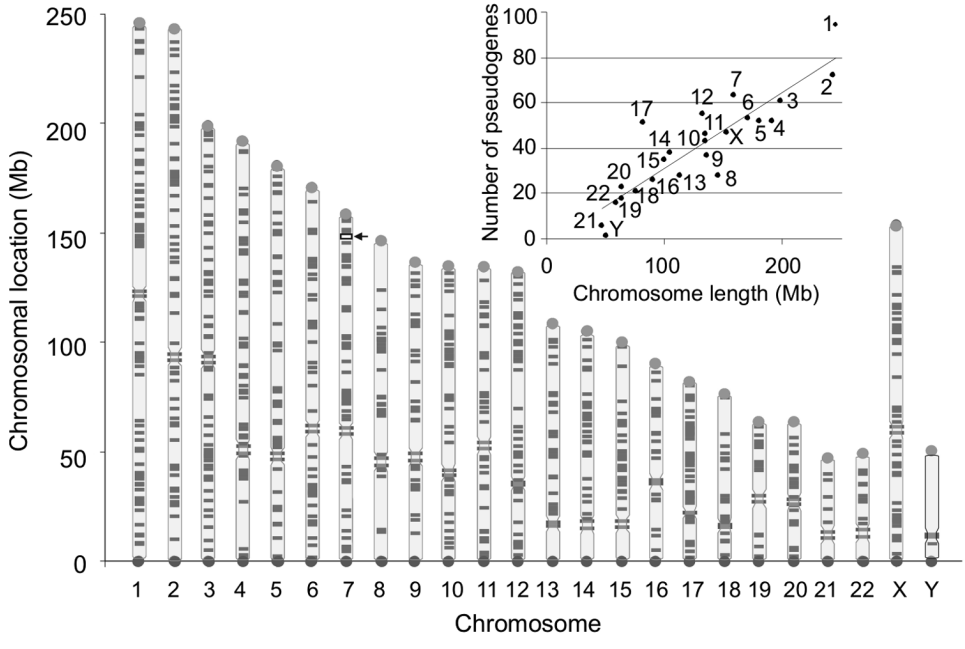

Figure 2. Distribution of the hY pseudogenes in the human genome. The 24 human chromosomes are shown vertically from left to right. Pseudogenes are represented by short black horizontal bars, telomeres by dots and centromeres by long horizontal bars. The position of the cluster formed by the functional hY genes is indicated by the arrow and small box. The inset presents the correlation between the numbers of hY pseudogenes on chromosomes relative to their length.

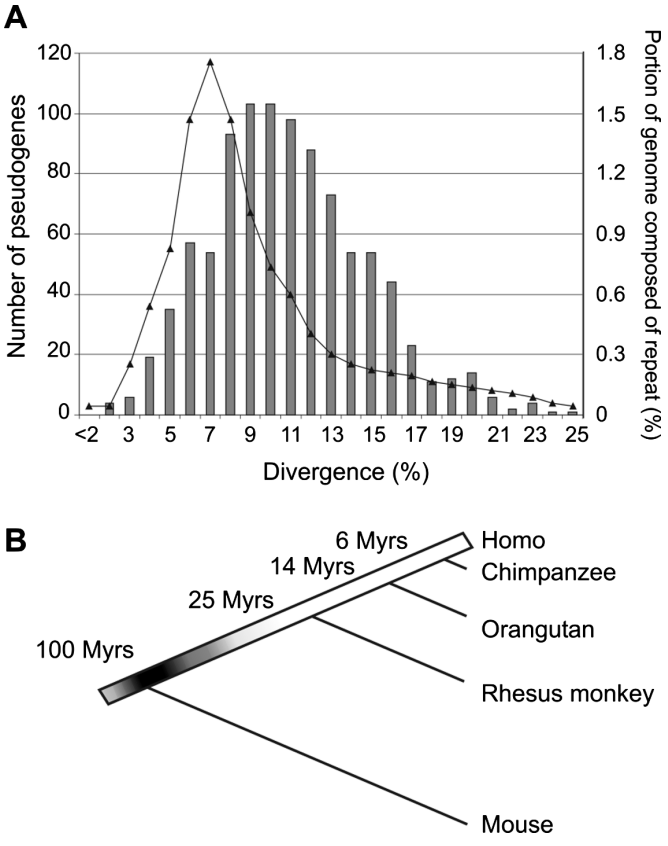

Figure 3. Conservation and age distribution of the hY pseudogenes. (**A**) Distribution of pseudogenes according to the percentage of divergence. The scale for hY pseudogenes (thick lines) and Alu curve (line with triangles) allows the comparison of identity peaks, but not absolute numbers. Alu data are reproduced from (47). (**B**) Phylogenetic tree of a few primates and of mouse adapted from (48). The thickness of the line grossly indicates the frequency of hY integration events.

unpublished data]. Finally, one, two and three of the U residues at the 3′ end were missing in 80, 40 and 25% of the hY pseudogenes, respectively. The La protein binds to RNA polymerase III transcripts through a characteristic poly(U) sequence at their 3′ end; a minimum of 3 U is preferred for binding of the La protein (36). Unlike most RNA polymerase III transcripts, mature hY RNAs retain a short $U_{3–4}$ tail at their 3′ end, and consequently maintain La protein binding. Taken together, these observations suggest that at least one and most probably two point mutations disabling cognate protein binding to the hY RNA contributed to enhance their retrotransposition efficiency. Formal testing of this hypothesis, as well as analysis of the mechanisms involved in this enhanced efficiency of retrotransposition with loss of specific protein binding, is in progress.

The degree of divergence in hY pseudogene sequence followed a normal distribution, with a peak at 9% (Figure 3A). In the absence of selection pressure, a mutation rate of 1% at individual nucleotides is estimated to occur every 6.7 million years (Myrs) (37). The distribution of divergence across Alu sequences exhibits a peak value at 7% (38). At first sight, this difference in the peak distribution of divergence suggested that most hY pseudogenes would have been generated before Alu insertions. However, since some positions in hY pseudogenes are almost uniformly mutated (see above), and these

Table 2. Relative numbers[a] of Y pseudogenes in different species

	Human	Chimp	Mouse
Y1	1.00	0.89	0.06
Y3	1.02	0.96	0.01
Y4	0.60	0.52	NA[b]
Y5	0.03	0.02	NA[b]

[a]Numbers are relative to the number of Y1 pseudogenes in man, according to BLAT results.
[b]Non applicable.

mutations are likely to be present at the RNA level before its integration into DNA, it appears that Alu elements and hY pseudogenes may be essentially contemporaneous. As most of the hY pseudogenes showed a divergence below 15%, an important proportion of the hY retrotranposition events occurred <100 Myrs ago, i.e. after primate and rodent divergence. A search for Y RNA homologues in the mouse genome using the BLAT software (http://genome.ucsc.edu/cgi-bin/hgBlat) (39) confirmed that hypothesis. Contrary to man, the mouse contains only two Y RNA genes: the mY1 and mY3 genes (40). Mouse mY1 and human hY1 RNAs, and mY3 and hY3 RNAs are divergent at 3 and 5 positions, respectively. Using BLAT, only 24 mY1 and 3 mY3 pseudogenes were identified (Table 2). In addition, mY and hY pseudogenes were not present at the corresponding genomic positions (data not shown). Clearly, the integration events of hY RNAs occurred after the divergence of rodents and primates.

On the contrary, man and chimpanzee (Figure 3B, estimated divergence 6 Myrs ago) should have most Y pseudogenes in common. The sequences of Y1, Y3 and Y4 genes are identical between man and chimpanzee, while the sequence of Y5 presents a single nucleotide insertion in the chimpanzee's version. The distribution of Y pseudogenes was almost identical between chimpanzee and man (data not shown), but the number of Y pseudogenes found in the chimpanzee genome was slightly lower, i.e. ~90% of the number found in man (Table 2). This small difference in prevalence suggests that retransposition of hY RNAs may still have occurred over the most recent 6 Myrs of evolution. Alternatively, some of the missing Y pseudogenes in the chimpanzee likely represent holes in the still incompletely assembled chimp genome.

Genomic localization of hY pseudogenes

We then looked at the localization of the hY-derived sequences relative to protein-coding genes (Table 3). Out of the 966 hY pseudogenes found in the human genome, 403 (42%) were located in protein-coding genes. However, only three of these hY pseudogenes were found in exons, and all three in untranslated regions of the ALDH9A1, CCL19 and MPRG genes (aldehyde dehydrogenase, small inducible cytokine A19 precursor and membrane progestin receptor γ, respectively). No common function or structure between these genes is apparent at this time. There were no significant differences between hY pseudogenes located in introns and those located in intergenic regions in regard to sequence identity, sequence length and specific mutations (data not shown). Intronic hY pseudogenes were almost randomly distributed between the sense and antisense orientation of the genes. Again, no differences between sense and antisense

Table 3. Overall statistics of hY pseudogenes

	Pseudogenes position				Ave. sequence identity[e] (%)	Ave. relative sequence length[f] (%)	GC content		
	Intragenic[a]								
	Sense[b]	Antisense[c]	Total	Intergenic[d]			hY genes (%)	hY pseudogenes (%)	Flanking region[g] (%)
hY1	80	82	162	206	87.9	91.5	44.6	42.1	42.5
hY3	103	71	174	268	90.6	90.3	45.5	42.6	40.0
hY4	24	38	62	86	90.7	87.2	42.6	40.5	40.6
hY5	4	1	5	3	89.4	86.3	45.2	42.9	40.6
Total	211	192	403	563	89.6	88.8	44.5	42.0	41.0

[a]Number of hY pseudogenes located in genes.
[b]Number of hY pseudogenes located in genes and in the same orientation.
[c]Number of hY pseudogenes located in genes and in opposite orientation.
[d]Number of hY pseudogenes located in intergenic regions.
[e]Average sequence identity between the pseudogene sequences and the corresponding hY RNA cDNA sequence.
[f]Length of pseudogenes divided by the length of hY genes, averaged over the entire pseudogene population.
[g]GC content of 5 kb flanking regions of hY pseudogenes.

hY pseudogenes were observed (data not shown). As expected, owing to the relatively high rate of spontaneous deamination of C to T in DNA, the GC content of the hY pseudogenes was slightly lower than the average 44% content of the hY genes (Table 3). The GC content of 5 kb regions surrounding hY pseudogenes was 41% (close to the GC content of the human genome), not significantly different for intronic and intergenic pseudogenes. This GC content is intermediate between those surrounding Alu (GC rich) and L1 (AT rich) (Table 3), as is the case for processed pseudogenes (33).

Mutant hY RNAs were directly retrotransposed in *trans* by the L1 machinery

The abundance of hY pseudogenes suggested that they were integrated through an efficient mechanism of retrotransposition, most likely the L1 machinery. Indeed, L1 encodes an endonuclease/reverse transcriptase that can bind and mobilize other RNAs in *trans* (17). The L1 retrotransposition machinery leaves behind a characteristic signature consisting of a 3′ end poly(A) tail and of TSDs (~15 bp long) flanking the pseudogenes (Figure 4A) (17). The TSD on the 5′ side usually starts at a T_2A_4-related sequence, the preferred substrate of L1 endonuclease.

In order to score the L1 signature flanking hY pseudogenes, a computer program was written (RTAnalyzer, see Materials and Methods). Initially, the program identified the 5′ and the 3′ extremities of hY pseudogenes to look for a 3′ end adjacent poly(A) sequence and for flanking TSDs. A scoring system ('RetroScore', see Materials and Methods) was established considering several parameters, including the length of the poly(A) tail and TSD sequences, their position, the homology between the TSDs, etc. A RetroScore cut-off of 40 was used to identify those sequences with a high probability of retrotransposition through L1. To confirm that a RetroScore cut-off of 40 was specific to identify bona fide L1 signatures, negative control sequences, such as a 120 nt intronic fragment of the Ro60 gene, $tRNA^{arg}$ and $tRNA^{lys}$, gave scores well below the cut-off, with the exception of a few tRNA pseudogenes with obvious L1 signatures that gave very good results (data not shown). Similarly, the genuine hY1, hY3, hY4 and hY5 genes gave scores of 0, 10, 15 and 22, respectively. Figure 4B presents a typical example of an hY1 pseudogene with an excellent L1 signature in close proximity of the hY1 pseudogene (i.e. a sequence with >85% identity to hY1, a T_2A_4 target site overlapping a TSD between 8 and 18 nt long located <2 nt away from the 5′ end of the hY1-homologous

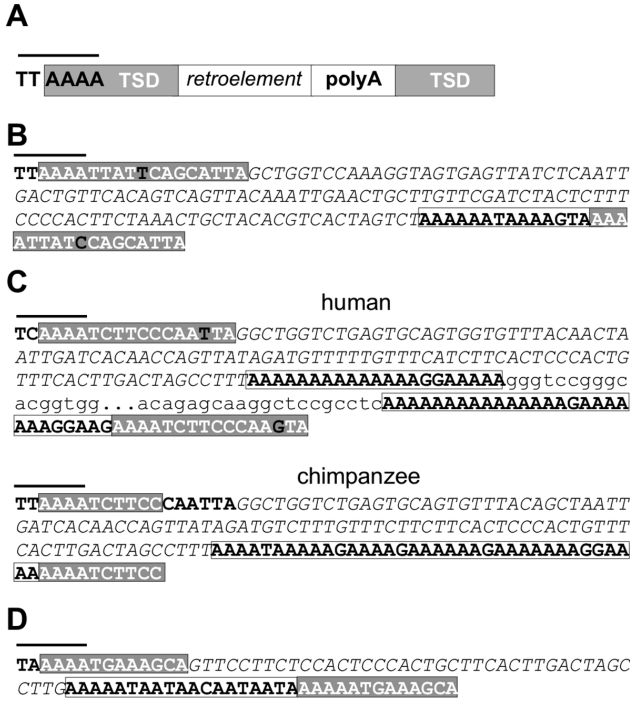

Figure 4. Examples of retrotransposed hY pseudogenes. (**A**) Schematic representation of a L1 retrotransposition signature. (**B**) Typical hY1 pseudogene found on chromosome 12, position 59021988; RetroScore 133. (**C**) Example of a hY3 pseudogene with an Alu element retrotransposed in its 3′ end (upper portion) found on chromosome 22, position 14601960; RetroScore when not considering the Alu insertion: 21; RetroScore when considering the Alu insertion: 115. The lower portion shows the corresponding Y3 pseudogene in the chimpanzee genome, where no Alu element is present (RetroScore 79). (**D**) Illustration of a hY3 sequence lacking a significant portion of the 5′ end of hY3 RNA found on chromosome 6, position 31568914; RetroScore 102. (Note that this insertion, shorter than the criteria established for pseudogenes, is shown here to clearly show 5′ truncation.) In all the cases, the hY sequences are in italics, Alu sequences are in lower cases, 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.

Table 4. Pseudo-hYs with L1 retrotransposition signatures

	poly(A)[a]	TSD[b]	Above-threshold RetroScore[c]	
			pseudo-hY[d]	hY-Alu[e]
hY1	319 (87%)	197 (54%)	212 (58%)	247 (67%)
hY3	377 (85%)	259 (59%)	285 (64%)	307 (69%)
hY4	116 (80%)	85 (59%)	105 (71%)	109 (74%)
hY5	5 (63%)	4 (50%)	5 (63%)	6 (75%)
Total	817 (84%)	545 (56%)	607 (63%)	669 (69%)

[a]Requiring at least four adenines in 3′.
[b]Target site duplication at least 4 nt long with little discordance.
[c]For the calculation of RetroScore and the definition of minimal requirements, see Materials and Methods.
[d]Percentage of pseudogenes directly fulfilling minimal requirements.
[e]Percentage of pseudogenes fulfilling minimal requirements, taking Alu insertions into account.

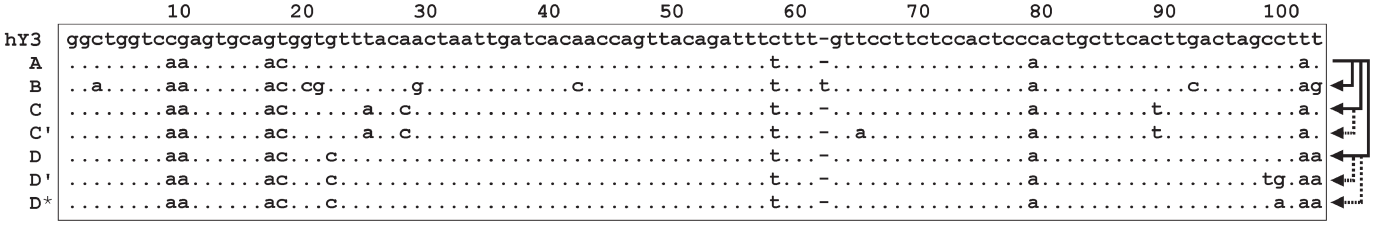

Figure 5. Proposed scenario for the evolution of the PMS2 gene family. Arrows at right indicate the hypothetical evolution of the PMS2 gene family, based on conserved divergences between subsets of hY3-derived pseudogenes. Fifteen PMS2 homologous sequences, all on Chromosome 7, contain a highly related hY3 pseudogene in the same orientation and in the same region of the same intron. Sequence variations in this hY3 pseudogene were used to propose a sequence of duplications events of a putative original PMS2 gene 'A', the hypothetic ancestor of the family, which is no more present as such in the human genome. 'B', 'C' and 'D' would represent the first duplication events of 'A', characterized by the introduction of new mutations. 'Primed' letters, as well as 'asterisk', indicate subsequent duplications and mutations that would have occurred following the initial duplications of PMS2. B is the actual PMS2 gene (NM_000535.2). There are three copies of the C gene (XM_374461, XM_498220 and XM_379906). C′ corresponds to gene PMS2L9 (NM_005395), derived from C. D is no longer present and would represent the ancestor of D′ and D*. There are nine copies of D′, six of which are annotated (two portions present in PMS2L2, NM_002679; LOC441259, XM_496900; LOC392729; LOC441263, XM_496904 and PMS2L5, NM_174930) and three are not (positions 71954462, 76315785 and 74157499). D* is gene PMS2L1 (XM_377962). The annotations and positions above correspond to NCBI build 35.1.

sequence and from a poly(A) tail between 10 and 30 nt long containing >70% adenines). Among all hY pseudogenes identified, 63% were flanked by such conserved L1 signatures (Table 4).

Closer inspection of the hY pseudogenes without satisfactory L1 signatures revealed that, in many cases, one of the TSDs (usually on the 3′ side of the hY pseudogenes) had been disrupted or displaced by the independent integration of an Alu element (Figure 4C). In these cases, allowing for removal of the Alu insertion revealed a 'repaired' two-part 3′ TSD that could then be matched with its corresponding sequence on the 5′ side of the hY pseudogene. This suggests that at these sites the hY RNAs were inserted before the integration of the Alu elements. Confirming this hypothesis, the corresponding pseudogene within the chimpanzee genome does not have an Alu sequence at its 3′ end and harbours typical TSDs (Figure 4C). The presence of a number of diverging Alu insertions close to hY pseudogenes between chimp and man indicated that Alu insertions frequently occurred in the 3′ TSD of pre-existing hY pseudogenes, either in man or in chimpanzee. The high prevalence of Alu insertions close to hY pseudogenes likely results from the fact that TSDs and the adjacent poly(A) stretch of L1-retrotransposed elements are good AT-rich nests for subsequent retrotransposition events using the L1 machinery. The addition of those hY pseudogenes with an Alu insertion to those that initially had convincing L1 signatures totalled up to 69% of all hY pseudogenes, similar to the percentage obtained with a random sample of Alu sequences (Table 3). This indicated that the vast majority of integration events of hY pseudogenes could still be recognized as L1-mediated insertion events. Moreover, up to 12% of hY pseudogenes were shortened by at least 10% at their 5′ end, presumably because of premature termination of the reverse transcription step (Figure 4D). Since the great majority of the L1 elements found in the human genome are 5′ truncated, this is additional evidence that hY pseudogenes were created by the L1 retrotransposition machinery.

Chimeric retrogenes formed of the 5′ end of snRNAs and of the 3′ end of Alu elements, most frequently U6-Alu retrogenes, were recently reported (24). Of the more than 300 small noncoding RNA species that were screened, only Alu sequences were found in close proximity to hY pseudogenes. As discussed above, such Alu elements were found in proximity (<300 nt) in more than half the hY pseudogenes. Nonetheless, <5% of the hY pseudogenes may correspond to true chimeric retrogenes. Similarly, we found no evidence for specific pairing of an internal sequence of hY RNAs with DNA sequences resulting from L1 endonuclease cuts, as has been recently proposed for tRNA-derived retropseudogenes (41). This negative result was expected, as this mechanism would lead to the absence of a poly(A) tail, a feature present in most hY pseudogenes.

hY pseudogene duplication

While the majority of hY pseudogenes resulted from independent retrotransposition events, ~5% represented duplications of DNA segments already containing hY pseudogenes. For example, we observed that 15 members of the Postmeiotic Segregation increased 2 [PMS2, a member of the DNA mismatch repair machinery (42)] gene family contain a highly related hY3 pseudogene in the same orientation, in the same region of the same intron. Single base differences between these duplicated pseudogenes were used to track the sequence of duplication events of the PMS2 gene family (Figure 5). According to this hypothetical reconstruction, the initial retrotransposition of hY3 RNA produced an ancestral (no more recognizable as such in the human genome) PMS2-like gene containing six mutations relative to the hY3 gene sequence. A first round of duplication events of the original PMS2-like gene yielded three different PMS2 genes characterized by the presence of a set of unique mutations in the hY3 pseudogene. Afterwards, two of the three PMS2 genes were again duplicated to yield three clusters of related genes characterized by mutations in the hY3 pseudogene specific to each cluster. Fifteen additional families of similarly duplicated DNA regions containing from 2 to 7 members were identified, but the majority of these duplications occurred in intergenic regions.

DISCUSSION

A new class of retroelements

We report here that almost 1000 pseudogenes derived from hY RNAs are widely scattered within all human chromosomes. As suggested by the example in Figure 4D, there seems to be smaller insertions and the 966 pseudogenes is most likely an underestimation of the total number of hY sequence insertions in the genome. Each of the four hY RNAs gave rise to pseudogenes, although most were derived from hY1 and hY3 RNAs. The genomes of the chimpanzee and man share ~95% of Y pseudogenes at identical locations. On the contrary, the mouse genome contains only 27 mY pseudogenes, found in non-conserved genomic locations relative to man. The degree of conservation of the hY pseudogenes' sequences compared with their corresponding genes also suggests that most of the Y pseudogenes were retrotransposed after the rodent/primate divergence. However, this explanation does not rule out completely the possibility that the lower prevalence of Y pseudogenes in mice results from a faster mutation rate leading to rapid decay of retrotransposed sequences.

At this point in time, we may only speculate on the physiological or pathological importance of the presence of a large number of Y pseudogenes in the human genome. Up to 42% of hY pseudogenes were found in intronic regions of genes. A simple explanation is that insertion of additional genetic material in introns is less detrimental than in coding regions. However, the role of Alu elements located in introns in the generation of alternative splice sites and their contribution to cell and tissue diversity is increasingly recognized (18). Because hY RNAs are bound by a number of proteins with proven or putative roles in alternative splicing, it is tempting to speculate a role for some of these intronic hY-related sequences in alternative splicing. However, these speculations need more detailed analysis.

hY pseudogenes result from L1-mediated retrotransposition of mutated hY RNAs

Careful analysis of the sequences and flanking regions of pseudogenes and of their corresponding functional hY genes revealed that most hY pseudogenes are not the result of duplication events. On the contrary, overwhelming evidence suggests that hY pseudogenes were integrated in *trans* using the L1 retrotransposition machinery. Up to 70% of the hY pseudogenes presented highly conserved signatures of the L1 retrotransposition machinery. The true number of hY-related retrotransposed sequences is probably higher, because our strict scoring criteria did not take random mutations into account and would have rejected poor TSD signatures, such as those of only 2 nt reported previously (43).

Retrotransposition of hY pseudogenes using L1 most likely occurred directly in *trans* from the hY RNA themselves. Contrary to what is observed with some U RNAs, we found little evidence for chimeric hY retrogenes. Most Alu elements close to hY pseudogenes proved to result from subsequent integration in the 3′ poly(A) tail or TSD of a pre-existing hY pseudogene. Only in rare cases, intercalating sequences were so short between the hY and Alu elements that a template switch of the L1 RT could not be formally excluded.

Identification of a large number of pseudogenes resulting from direct retrotransposition of hY RNAs was quite surprising. The number of hY pseudogenes is significantly larger than the corresponding numbers derived from 100- to 1000-fold more abundant non-coding RNAs, such as 5S RNA and U1 RNA. Bona fide hY RNAs also lack a 3′ poly(A) tail known to be important for efficient L1-mediated retrotransposition (43), although L1-mediated retrotransposition may still occur at a reduced level without poly(A) tail (44). However, the high frequency of targeted point mutations in hY pseudogenes suggested that prevention of binding of Ro60 as well as La proteins on hY RNAs may have made these mutated RNAs better substrates for retrotransposition. We have previously shown that human Ro RNPs have distinct biochemical and immunological properties that are determined by the hY RNA they contain (45,46). Only hY1 and hY3 RNAs are bound by proteins such as hnRNP K and PTB (10). The association of hY1 and hY3 RNAs with specific proteins might have contributed to the relative abundance of the hY1- and hY3-related pseudogenes. In addition, the high prevalence of point mutations in the middle region of hY1-related pseudogenes, a feature not present in hY3-related pseudogenes, may also relate to the differential binding properties of hnRNP K and/or PTB proteins to hY1 and hY3 RNAs [(10); F. Brière and G. Boire, unpublished data]. On the contrary, hY5 RNA appears to have been the source of only a few pseudogenes. In addition to its relatively recent introduction in primates, the specific association of hY5 RNA with RoBPI (8) may have played a role in that relative paucity of retrotransposition events. We hypothesize that the presence of specific point mutations, introduced either at the genomic or at the transcription levels, and the resulting lack of Ro60 and La protein binding (and possibly, in the case of hY1 RNA, of hnRNP K and/or PTB proteins), may increase the odds for retrotransposition. This hypothesis, as well as other

mechanistic questions, is amenable to direct testing in cellular models (43).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Sherif Abou Elela and Dr Raymund Wellinger for their many insights in the project. This work was supported by a grant from the Canadian Institute of Health Research (CIHR) to the RNA group and to J.-P.P. The RNA group is also supported by grants from both the Université de Sherbrooke and Génome Québec (bioinformatics core grant). J.P. was the recipient of a predoctoral fellowship from FRSQ (Québec). J.-P.P. is an investigator from the CIHR. B.C. is a CIHR new investigator and a McGill William Dawson Scholar. Funding to pay the Open Access publication charges for this article was provided by CIHR grant number MOP-44002.

Conflict of interest statement. None declared.

REFERENCES

1. Tan,E.M. (1989) Antinuclear antibodies: diagnostic markers for autoimmune diseases and probes for cell biology. *Adv. Immunol*., **44**, 93–151.
2. Bouffard,P., Laniel,M.A. and Boire,G. (1996) Anti-Ro (SSA) antibodies: clinical significance and biological relevance. *J. Rheumatol*., **23**, 1838–1841.
3. Farris,A.D., O'Brien,C.A. and Harley,J.B. (1995) Y3 is the most conserved small RNA component of Ro ribonucleoprotein complexes in vertebrate species. *Gene*, **154**, 193–198.
4. Pellizzoni,L., Lotti,F., Rutjes,S.A. and Pierandrei-Amaldi,P. (1998) Involvement of the *Xenopus laevis* Ro60 autoantigen in the alternative interaction of La and CNBP proteins with the 5′ UTR of L4 ribosomal protein mRNA. *J. Mol. Biol*., **281**, 593–608.
5. Chen,X. and Wolin,S.L. (2004) The Ro 60 kDa autoantigen: insights into cellular function and role in autoimmunity. *J. Mol. Med*., **82**, 232–239.
6. Labbe,J.C., Burgess,J., Rokeach,L.A. and Hekimi,S. (2000) ROP-1, an RNA quality-control pathway component, affects *Caenorhabditis elegans* dauer formation. *Proc. Natl Acad. Sci. USA*, **97**, 13233–13238.
7. Xue,D., Shi,H., Smith,J.D., Chen,X., Noe,D.A., Cedervall,T., Yang,D.D., Eynon,E., Brash,D.E., Kashgarian,M., Flavell,R.A. and Wolin,S.L. (2003) A lupus-like syndrome develops in mice lacking the Ro 60-kDa protein, a major lupus autoantigen. *Proc. Natl Acad. Sci. USA*, **100**, 7503–7508.
8. Bouffard,P., Barbar,E., Brière,F. and Boire,G. (2000) Interaction cloning and characterization of RoBPI, a novel protein binding to human Ro ribonucleoproteins. *RNA*, **6**, 66–78.
9. Fabini,G., Rutjes,S.A., Zimmermann,C., Pruijn,G.J. and Steiner,G. (2000) Analysis of the molecular composition of Ro ribonucleoprotein complexes. Identification of novel Y RNA-binding proteins. *Eur. J. Biochem*., **267**, 2778–2789.
10. Fabini,G., Raijmakers,R., Hayer,S., Fouraux,M.A., Pruijn,G.J. and Steiner,G. (2001) The heterogeneous nuclear ribonucleoproteins I and K interact with a subset of the Ro ribonucleoprotein-associated Y RNAs *in vitro* and *in vivo*. *J. Biol. Chem*., **276**, 20711–20718.
11. Fouraux,M.A., Bouvet,P., Verkaart,S., van Venrooij,W.J. and Pruijn,G.J. (2002) Nucleolin associates with a subset of the human Ro ribonucleoprotein complexes. *J. Mol. Biol*., **320**, 475–488.
12. Van Buskirk,C. and Schupbach,T. (2002) Half pint regulates alternative splice site selection in Drosophila. *Dev. Cell*, **2**, 343–353.
13. Wolin,S.L. and Cedervall,T. (2002) The La protein. *Annu. Rev. Biochem*., **71**, 375–403.
14. MacMorris,M., Brocker,C. and Blumenthal,T. (2003) UAP56 levels affect viability and mRNA export in *Caenorhabditis elegans*. *RNA*, **9**, 847–857.
15. Bomsztyk,K., Denisenko,O. and Ostrowski,J. (2004) hnRNP K: one protein multiple processes. *Bioessays*, **26**, 629–638.
16. McCutcheon,I.E., Hentschel,S.J., Fuller,G.N., Jin,W. and Cote,G.J. (2004) Expression of the splicing regulator polypyrimidine tract-binding protein in normal and neoplastic brain. *Neuro-oncol*., **6**, 9–14.
17. Ostertag,E.M. and Kazazian,H.H.,Jr. (2001) Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet*., **35**, 501–538.
18. Kazazian,H.H.,Jr. (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.
19. Goodier,J.L., Ostertag,E.M., Du,K. and Kazazian,H.H.,Jr. (2001) A novel active L1 retrotransposon subfamily in the mouse. *Genome Res*., **11**, 1677–1685.
20. Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V. and Kazazian,H.H.,Jr. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. USA*, **100**, 5280–5285.
21. 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.
22. Weiner,A.M. (2000) Do all SINEs lead to LINEs? *Nature Genet*., **24**, 332–333.
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. Weiner,A.M., Deininger,P.L. and Efstratiadis,A. (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem*., **55**, 631–661.
26. O'Brien,C.A. and Harley,J.B. (1992) Association of hY4 pseudogenes with Alu repeats and abundance of hY RNA-like sequences in the human genome. *Gene*, **116**, 285–289.
27. Crouch,D. and Liebke,E.H. (1989) The molecular cloning of a mouse Ro RNA, my1-like sequence. *Nucleic Acids Res*., **17**, 4890.
28. Jurka,J., Smith,T.F. and Labuda,D. (1988) Small cytoplasmic Ro RNA pseudogene and an Alu repeat in the human alpha-1 globin gene. *Nucleic Acids Res*., **16**, 766.
29. Maraia,R., Sakulich,A.L., Brinkmann,E. and Green,E.D. (1996) Gene encoding human Ro-associated autoantigen Y5 RNA. *Nucleic Acids Res*., **24**, 3552–3559.
30. Jurka,J. (2004) Evolutionary impact of human Alu repetitive elements. *Curr. Opin. Genet. Dev*., **14**, 603–608.
31. Jurka,J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci. USA*, **94**, 1872–1877.
32. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol*., **7**, 203–214.
33. Zhang,Z., Harrison,P. and Gerstein,M. (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res*., **12**, 1466–1482.
34. Grover,D., Mukerji,M., Bhatnagar,P., Kannan,K. and Brahmachari,S.K. (2004) Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. *Bioinformatics*, **20**, 813–817.
35. Green,C.D., Long,K.S., Shi,H. and Wolin,S.L. (1998) Binding of the 60-kDa Ro autoantigen to Y RNAs: evidence for recognition in the major groove of a conserved helix. *RNA*, **4**, 750–765.
36. Stefano,J.E. (1986) Purified lupus antigen La recognizes an oligouridylate stretch common to the 3′ termini of RNA polymerase III transcripts. *Cell*, **36**, 145–154.
37. Kumar,S., Tamura,K., Jakobsen,I.B. and Nei,M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1245.
38. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. (2001)

Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
39. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
40. Hendrick,J.P., Wolin,S.L., Rinke,J., Lerner,M.R. and Steitz,J.A. (1981) Ro small cytoplasmic ribonucleoproteins are a subclass of La ribonucleoproteins: further characterization of the Ro and La small ribonucleoproteins from uninfected mammalian cells. *Mol. Cell. Biol.*, **1**, 1138–1149.
41. Schmitz,J., Churakov,G., Zischler,H. and Brosius,J. (2004) A novel class of mammalian-specific tailless retropseudogenes. *Genome Res.*, **14**, 1911–1915.
42. Shin-Darlak,C.Y., Skinner,A.M. and Turker,M.S. (2005) A role for *Pms2* in the prevention of tandem CC→TT substitutions induced by ultraviolet radiation and oxidative stress. *DNA Repair*, **4**, 51–57.
43. Dewannieux,M., Esnault,C. and Heidmann,T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet.*, **35**, 41–48.
44. Roy-Engel,A.M., Salem,A.H., Oyeniran,O.O., Deininger,L., Hedges,D.J., Kilroy,G.E., Batzer,M.A. and Deininger,P.L. (2002) Active Alu element 'A-tails': size does matter. *Genome Res.*, **12**, 1333–1344.
45. Boire,G. and Craft,J. (1990) Human Ro ribonucleoprotein particles: characterization of native structure and stable association with the La polypeptide. *J. Clin. Invest.*, **85**, 1182–1190.
46. Gendron,M., Roberge,D. and Boire,G. (2001) Heterogeneity of human Ro ribonucleoproteins (RNPS): nuclear retention of Ro RNPS containing the human hY5 RNA in human and mouse cells. *Clin. Exp. Immunol.*, **125**, 162–168.
47. Deininger,P.L. and Batzer,M.A. (2002) Mammalian retroelements. *Genome Res.*, **12**, 1455–1465.
48. Goodman,M., Porter,C.A., Czelusniak,J., Page,S.L., Schneider,H., Shoshani,J., Gunnell,G. and Groves,C.P. (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.*, **9**, 585–598.