Ro-Associated Y RNAs in Metazoans: Evolution and Diversification

Jonathan Perreault,† Jean-Pierre Perreault,* and Gilles Boire†*

*Département de Biochimie, Université de Sherbrooke, Sherbrooke, Québec, Canada and †Service de Rhumatologie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Sherbrooke, Québec, Canada

The Y genes encode small noncoding RNAs whose functions remain elusive, whose numbers vary between species, and whose major property is to be bound by the Ro60 protein (or its ortholog in other species). To better understand the evolution of the Y gene family, we performed a homology search in 27 different genomes along with a structural search using Y RNA specific motifs. These searches confirmed that Y RNAs are well conserved in the animal kingdom and resulted in the detection of several new Y RNA genes, including the first Y RNAs in insects and a second Y RNA detected in *Caenorhabditis elegans*. Unexpectedly, Y5 genes were retrieved almost as frequently as Y1 and Y3 genes, and, consequently are not the result of a relatively recent apparition as is generally believed. Investigation of the organization of the Y genes demonstrated that the synteny was conserved among species. Interestingly, it revealed the presence of six putative "fossil" Y genes, all of which were Y4 and Y5 related. Sequence analysis led to inference of the ancestral sequences for all Y RNAs. In addition, the evolution of existing Y RNAs was deduced for many families, orders and classes. Moreover, a consensus sequence and secondary structure for each Y species was determined. Further evolutionary insight was obtained from the analysis of several thousand Y retropseudogenes among various species. Taken together, these results confirm the rich and diversified evolution history of Y RNAs.

Introduction

Ro ribonucleoproteins (RNPs) are low abundance autoantigens that are frequently targeted by autoantibodies present in the serum of patients with connective tissue diseases (reviewed in Chen and Wolin 2004). The Ro RNPs result from the noncovalent association of small noncoding RNAs of the Y family with the 60 kDa Ro protein (Ro60), or its orthologs. The size of Y RNAs varies from 70 to 115 nucleotides, and the numbers of genes per species from 0 to 4. For example, four Y RNAs are known to exist in humans (hY1, hY3, hY4, and hY5), two in mice and rats, only one in *Caenorhabditis elegans* and none in *Drosophila melanogaster*. Y RNAs are present in many metazoans (Farris et al. 1999; Teunissen et al. 2000), and were even identified in the bacterium *Deinococcus radiodurans* where an ortholog of Ro60 was also detected (Chen, Quinn, and Wolin 2000). The Y3 RNA was proposed to be the most conserved among vertebrates (Farris, O'Brien, and Harley 1995). The sequence and secondary structure of the human hY3 RNA are illustrated in figure 1A. The stable stem formed by the base-pairing of the 5′ and 3′ extremities of the Y RNA is the binding site of the Ro60 protein, and is the most conserved element of these RNAs (Wolin and Steitz 1984; O'Brien, Margelot, and Wolin 1993; Van Horn et al. 1995; Farris et al. 1999; Teunissen et al. 2000). In addition to other important sequence requirements, the binding of Ro protein to this stem requires a characteristic bulged C in its middle portion (in either position 8 or 9; Green et al. 1998; Stein et al. 2005). Unlike most RNA polymerase III transcripts, mature Y RNAs conserve a 3′ poly(U) tail to which the La protein may still bind. Additional proteins may be present in human Ro RNPs, or bind Y RNAs, including hnRNP K, PTB, RoBP1 and nucleolin (Wolin and Steitz 1984; Bouffard et al. 2000; Fabini et al. 2000, 2001; Fouraux et al. 2002). Contrary to Ro60 and La, the binding of the additional proteins varies depending on the Y RNA and on the experimental conditions.

Despite the remaining uncertainties about the function(s) of Y RNAs and of their corresponding Ro RNPs, several hints at their multiple roles are available. For example, the Ro60 protein has been suggested to play a role in the regulation of the translation of ribosomal mRNA (Pellizzoni et al. 1998), in an ill-defined quality control system for small RNAs (Shi et al. 1996; Labbé et al. 2000; Chen et al. 2003; Stein et al. 2005; Fuchs et al. 2006) and in the enhancement of cell survival after exposure to ultraviolet irradiation (Chen, Quinn, and Wolin, 2000; Lawley et al. 2000; Chen and Wolin 2004). Recently, Y RNAs were also proposed to be essential for efficient DNA replication in a manner independent of Ro60 binding (Christov et al. 2006). Finally, because most hY RNA-associated proteins are implicated in alternative splicing and in the regulation of the translation of specific RNAs (Wolin and Steitz 1984; Bouffard et al. 2000; Fabini et al. 2000, 2001; Fouraux et al. 2002), the involvement of the noncoding Y RNAs in these physiological mechanisms has been proposed on several occasions.

Through a bioinformatic approach, we recently identified a large number of Y pseudogenes within the human genome (Perreault et al. 2005). Approximately a thousand hY RNA pseudogenes were retrieved in humans and chimpanzees, while mY pseudogenes were found to be rare in mice. Moreover, we reported that the hY retrotransposition events had occurred in *trans* using the L1 machinery. By analogy with the 7SL RNA that initially gave rise to the Alu elements, it was suggested that hY RNAs represent a novel class of nonautonomous, L1-dependent, retrotransposable elements (Perreault et al. 2005). Here, in order to better understand the phylogeny of the Y RNA genes and pseudogenes alike, we analyzed Y sequences in 27 metazoan genomes. This resulted in the identification of up to 20 complete new Y genes and thousands of new pseudogenes from which several analyses were performed.

Key words: Y RNA, retroposition, retrotransposition, pseudogene, ancestral genes.

E-mail: Gilles.Boire@USherbrooke.ca; Jean-Pierre.Perreault@USherbrooke.ca.

Mol. Biol. Evol. 24(8):1678–1689. 2007
doi:10.1093/molbev/msm084
Advance Access publication April 29, 2007

Materials and Methods

Y RNA Pseudogene Identification and Analysis

Genomes were downloaded from http://hgdownload.cse.ucsc.edu/downloads.html (University of California at

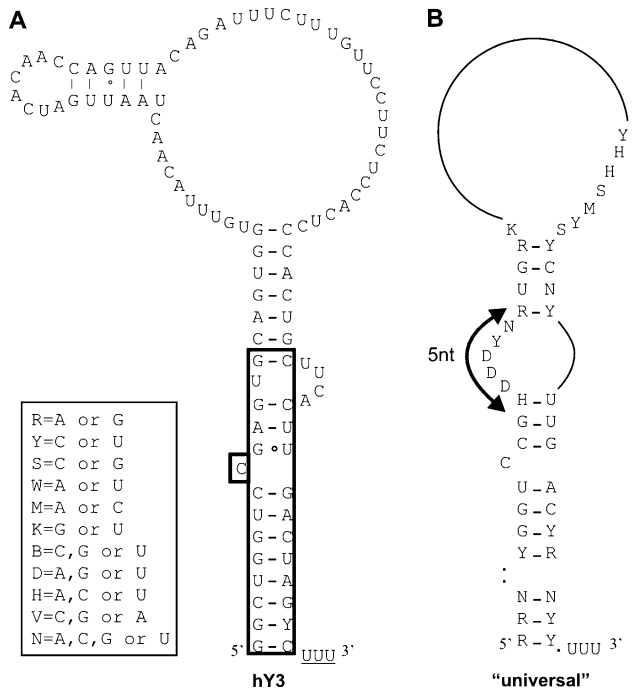


FIG. 1.—Predicted secondary structure of Y RNAs. (*A*) Sequence and secondary structure of the human Y3 RNA (hY3). The protein binding sites are boxed for Ro60 and underlined for La. B) Sequence and secondary structure of a putative ancestral Y RNA species derived from all known Y1, Y3, Y4, and Y5 RNAs (the only exceptions being the *Deinococcus* Y RNAs and a single mispair found in both *Xenopus* Y5 and YαRNAs). The arrow indicates the constant number of nucleotides in the loop located between 2 stables stems; a single dot indicates the possibility of an additional nucleotide at that position, while a solid line indicates the possibility of multiple nucleotides at that position. Note that any nucleotides unpaired in the consensus structure may be paired in some Y RNAs. The inset shows the IUPAC code for nucleotide nomenclature.

Santa Cruz) as well as from ftp://ftp.ncbi.gov/genomes and ftp://ftp.ncbi.gov/pub/TraceDB (National Center for Biotechnology Information). BLATz software was obtained from W. James Kent (unpublished software based on BLAT and BLASTz [Jegga et al. 2002; Kent 2002; Kent et al. 2003]). All known Y sequences (i.e., a total of 38 sequences as of July 1st, 2006) were independently used for the BLATz searches (with a cut off score of 2500) on each genome with a Perl script allowing automated use of the software. Additional 5′ and 3′ offset sequences (500 bp) were retrieved with this script according to the genome coordinates. The resulting sequences were compared to each other using a matcher local alignment tool adapted with a Perl script. This permitted the building of a large collection of DNA sequences from 27 genomes containing putative Y-related sequences.

Finding and Confirmation of Y RNA Genes and Deduction of Ancestral Genes

When the sequence homology of the putative Y-related sequences with known sequences of Y RNAs was relatively high, we searched for conserved structural features so as to be able to determine whether this represented a true gene and not a pseudogene. These required structural characteristics include the presence of an RNA polymerase III terminator (i.e., a stretch of Ts) and of 2 candidate promoter boxes located at the appropriate distances from position +1 (–8 to –12 and –45 to –47, depending on the Y type). When the appropriate structural features were present, the positions of those candidate Y genes relative to the other genes in the same genome were examined in order to look for synteny, a feature further supporting the conclusion that they represented putative new Y genes. When only partial sequences of Y RNAs were available from the NCBI databases (e.g., Y1 in dog and guinea pig), these sequences were used to validate the potential new Y genes. Alternatively, consensus Y sequences and predicted consensus Y RNA structures (see fig. 1B) were entered into rnamotif (Macke et al. 2001) in order to search for Y RNA genes in species where none were found by homology search.

All identified Y sequences were aligned using ClustalW (Chenna et al. 2003), and clustered into one of the four major types of Y RNAs (Y1, Y3, Y4, or Y5) according to these alignments. The alignment was then manually adjusted so as to permit inference analysis of ancestral gene sequences.

Northern Blot of Y5 in Mammals

Total liver RNA was extracted from mouse, rat, rabbit, and guinea pig tissue. Total testis RNA was also extracted from guinea pig tissue. The tissue was initially homogenized in Qiazol (Qiagen, Mississauga, ON, Canada) using a glass homogenizer, and RNA was then extracted according to manufacturer's recommendations. As a control, total RNA was extracted from the human hepatocyte cell line Huh-7. The following 5′ labeled DNA oligonucleotide probes were used for the "fossil" murine mY5 (5′-GTTGTGGGTTATTGTTAAATTGTTTAACACTGT-3′); the human hY5 (5′-GTTGTGGGTTATTGTTAAGTTGATTTAAC-3′); and, the guinea pig Y5 (5′-GTTAACCATGTATAACAATAACCCACAGC-3′), Y3 (5′-TGGTTGTGATCAATTAGTTGTAAACACCACT-3′) and 5S (5′-AAAGCCTACAGCACCCGGTATTCCC-3′).

Results

Extraction of a Collection of Y-related Sequences from Multiple Genomes

As a first step investigating for the presence of Y RNA genes within various genomes, we retrieved all complete and partial Y RNA sequences (i.e., 20 complete and 18 partial, respectively) available from the NCBI nucleotide databases. These sequences were then individually used to perform homology searches using BLATz software on 27 metazoan genomes. Sixteen of these genomes were fully sequenced and annotated, with the others being at various stages of completeness. Thus, the wallaby and the marmoset genomes were still very incomplete, while the genomes of the guinea pig, armadillo, cow, elephant, opossum, rabbit, rhesus monkey, tenrec and the frog *Xenopus tropicalis* were not fully assembled into chromosomes. These initial searches generated more than one hundred thousand hits. Only 6 genomes did not generate a single

Table 1
List of the Y RNA Genes Found in 27 Different Genomes

	Known (putative new) Y Genes								
	Homology Search							RNAmotif	
Species	Y1	Y3	Y4	Y5	Yα	CeY	DrY	Y1	New Y
Homo sapiens (man)	1	1	1	1					
Pan troglodytes (chimpanzee)	1	1	1	1					
Macaca mulatta (rhesus)	(1)	(1)	(1)	(1)					
Callithrix jacchus (marmoset)									
Cavia porcellus (guinea pig)	(1)[a]	(1)	(1)	(1)[b]					
Loxodonta africana (elephant)		(1)	(1)	(1)					
Canis familiaris (dog)	1[a]	1[a]	1[a]	1					
Rattus norvegicus (rat)	1	1							
Mus musculus (mouse)	1	1							
Echinops telfairi (tenrec)	(1)		(1)	(1)					
Dasypus novemcinctus (armadillo)	(1)		(1)	(1)					
Bos taurus (cow)	(1)[a]	(1)[a]	(1)[a]	(1)					
Oryctolagus cuniculus (rabbit)	(1)[a]	(1)	(1)						
Monodelphis domestica (opossum)	(1)	(1)	(1)						
Macropus eugenii (wallaby)	(1)								
Gallus gallus (chicken)	1	1							
Xenopus tropicalis (western clawed frog)		(1)		(1)	(2)				
Danio rerio (zebrafish)	(1)	(1)							
Fugu rubripes (Japanese pufferfish)								(1)	
Tetraodon nigroviridis (spotted green puffer)								(1)	
Caenorhabditis elegans (roundworm)						1			(1)
Caenorhabditis briggsae (roundworm)									(2)
Anopheles gambiae (mosquito)									(1)
Drosophila melanogaster (fruitfly)									
Apis mellifera (bee)									
Ciona intestinalis (sea squirt)									
Deinococcus radiodurans							4		

NOTE:—The existence of the four guinea pig Y RNAs, but not their sequence, was previously shown by immunoprecipitation (Itoh, Kriet, and Reichlin 1990).
[a] Previous work uncovered partial sequences of these Y RNAs.
[b] The previously found partial sequence was an artifact not matching genomic data.

hit: the fish *Tetraodon nigroviridis*, the fruitfly, the mosquito *Anopheles gambiae*, the bee, the roundworm *Caenorhabditis briggsae* and the seasquirt. Redundant hits were removed from this database using a Perl script, reducing the number of unique putative sequences to a little more than 5,000 (tables 1 and 2). More than 70% of these hits corresponded to homologous Y sequences retrieved from four primates: *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, and *Callitrix jacchus*. Of the nonprimate genomes, nearly 70% of the Y sequences retrieved were found in *Cavia porcellus* alone, a striking exception among the rodents (see below).

Table 2
Numbers and Identity Percentages of Y Pseudogenes in Mammals

	Y1		Y3		Y4		Y5	
Species	nb	Id	nb	Id	nb	Id	nb	Id
Homo sapiens (man)	415	89%	484	90%	155	91%	19	88%
Pan troglodytes (chimpanzee)	401	89%	461	90%	155	91%	21	87%
Macaca mulatta (rhesus)	392	88%	437	89%	145	90%	22	87%
Callithrix jacchus (marmoset)[b]	370	88%	400	91%	160	90%	6	89%
Cavia porcellus (guinea pig)[b]	310	94%	160	96%	630	94%	8	82%
Loxodonta africana (elephant)	124[a]	77%	5	81%	4	87%	4	85%
Canis familiaris (dog)	18	89%	30	88%	6	92%	7	88%
Rattus norvegicus (rat)	48	88%	18	86%	2	80%	1	81%
Mus musculus (mouse)	40	84%	13	86%	3	79%	1	85%
Echinops telfairi (tenrec)	31	80%	8	85%	25	91%	2	84%
Dasypus novemcinctus (armadillo)	4	91%	10	82%	3	96%	3	84%
Bos taurus (cow)	13	94%	9	87%	5	95%	2	80%
Oryctolagus cuniculus (rabbit)	4	96%	13	87%	7	96%	1	87%
Monodelphis domestica (opossum)	7	86%	2	82%	4	81%	1	78%

nb and Id indicate number of pseudogenes and identity percentages, respectively.
[a] The elephant Y1 pseudogenes are most probably an artifact due to a similarity with another, larger, repeated element
[b] Marmoset and guinea pig results are very rough estimates. The guinea pig genome is fully sequenced, but as yet it is not annotated. As the coverage of the genome is approximately 2 for the guinea pig, the crude results of homology search were divided by 2. On the other hand, as only a third of the marmoset genome is sequenced, the number of hits was multiplied by 3 in order to estimate the total number of Y pseudogenes in that species.

Identification of Putative Y RNA Genes

Several true Y genes were directly identified from the database of homologous Y sequences because these hits exhibited the highest levels of homology with known RNA sequences from other species. As additional requirements before these sequences were called Y RNA genes, they had to fold into a predicted secondary structure consistent with that of a Y RNA, to possess a 3′ T stretch sufficient to serve as RNA polymerase III terminator and to be preceded by an upstream sequence consistent with a candidate promoter and present at an appropriate distance from position +1. Only the sequences fully satisfying all three of these criteria were considered putative Y genes. This allowed the identification of 57 Y RNA genes, of which 38 were novel (table 1). This list includes 50 genes that were homologous to Y1, Y3, Y4 or Y5 RNA, as well as 7 Y RNA genes that could not be classified in the "classic" Y number fashion. Each of the Y RNA genes described previously (i.e., including both complete and partial sequences) was retrieved from the appropriate genome, serving as proof-of-principle for our strategy. In the eight cases where only partial Y RNA sequences were available from Genbank complete sequences of the Y RNA genes were identified from the corresponding genomes. However, there was one exception: the partial sequence of Y5 RNA from guinea pig did not match with the genomic Y5 gene sequence we identified. Errors inherent to the cloning strategy used to obtain the partial Y5 RNA sequence likely explain this discrepancy.

A second strategy, using the predicted secondary structures of the putative Y RNAs, was developed in order to identify additional Y RNA genes from genomes in which no Y sequence (or a number lower than expected) were identified. A consensus Y RNA sequence and structure was first established from the 38 complete and partial known sequences from the NCBI databases (fig. 1B). This consensus includes a portion of the RNA that was already known to be important for Ro binding (Green et al. 1998; Stein et al. 2005) as well as additional requirements towards the middle of the Y RNA. Briefly, the predicted structure of the "universal" Y RNA corresponds to a stem with a few bulges resulting from the base-pairing of the 5′ and 3′ ends of the sequence, some sequence constraints and a poly(U) 3′ tail that is not always found in the RNA gene but is part of the polymerase III terminator signal. This consensus sequence was then used to write an RNAmotif descriptor (Supplementary Material 1) to search for new Y RNA putative genes. When this strategy was applied to the human genome, the four hY genes were readily retrieved. When applied to the other genomes the same search identified six putative new Y RNA genes (table 1). These six putative Y genes also possess an RNA polymerase III terminator and are preceded by an appropriate candidate promoter, thus fulfilling the requirements to be considered as potential Y RNA genes. These 6 potential Y genes identified by structure rather than by homology included two representatives of Y1 RNA genes in fish and four unclassified Y RNA genes in *Anopheles gambiae*, *Caenorhabditis elegans*, and *Caenorhabditis briggsae* (two genes). The newly discovered *C. elegans* Y RNA gene overlapped with a recently sequenced 92-nt small noncoding RNA lacking a Ro60-binding region (National Center for Biotechnology Information [NCBI] accession number AM286260). The fact that the same search strategy identified two Y RNA genes in the related nematode *C. briggsae* strengthened the assumption that the second gene we identified in *C. elegans* may be a still unreported Y RNA gene, although this remains to be confirmed experimentally. Clearly, the use of a second search strategy using structural requirements allowed us to expand the collection of putative genes coding for Y RNAs.

Altogether, this search provided a compilation of 63 Y genes (see table 1 and online Supplementary Material 2 for a complete list). Of these 63 genes, 52 can be classified as: Y1 (17), Y3 (14), Y4 (11), and Y5 (10) RNA genes. The 11 remaining genes exhibited structural and sequence homology to the "universal" Y sequence, but were found mostly in invertebrates such as roundworms (4 genes), insects (1 gene) and bacteria (4 genes) (see online Supplementary Material 1). The only exceptions were putative Yα RNA genes identified in frog *X. tropicalis* by sequence homology with the Yα RNA previously described in *X. laevis*. One of these Yα homologous genes in *X. tropicalis* was almost identical to the Yα gene from *X. laevis*, and the second contained a small deletion. Interestingly, we identified a large set (10 in all, including 8 new ones) of putative Y5 RNA genes. This observation was unexpected, since Y5 RNA was initially found only in primates and thought to be of recent origin, although it was later found in the frog (O'Brien, Margelot, and Wolin 1993). It is now clear the origin of Y5 RNA goes back to the common ancestor of primates and frogs.

Synteny of Y RNA Genes

Subsequently, the organization of Y RNA genes at the chromosomal level was investigated in various genomes. In humans, the four Y genes are located in a relatively small region on chromosome 7 (fig. 2; Wolin and Steitz 1983; Maraia et al. 1996). More specifically, the four hY RNA genes are found in a region of ~70 kb located between the EZH2 and PDIA4 genes. The distance between each gene varies from ~3 kb between the Y3 and Y1 genes to ~20 kb between the Y5 and Y4 genes. The Y gene synteny observed in the human genome was found in most other species. All tetrapods possessed a cluster of Y RNA genes between the EZH2 and PDIA4 genes (fig. 2), except for zebrafish, for which no Y RNA gene was found between those two genes. Instead of Y RNA genes, the CUL1 gene was found in this location in zebrafish (this gene is found on the other side of EZH2 in the other species). The distances between each of the Y RNA genes varied from ~3 kb to ~102 kb. The order and orientation of the Y genes was also conserved in most species, with the exception of the opossum and the chicken. Usually the Y5, Y4, and Y3 genes are found in consecutive order, and their transcription is in the same orientation. Conversely, the transcription of the Y1 RNA gene appears to be in the opposite orientation. The fact that all of the newly discovered Y RNA genes fit into the conserved gene organization of the Y genes provided additional support for their putative expression as Y RNAs.

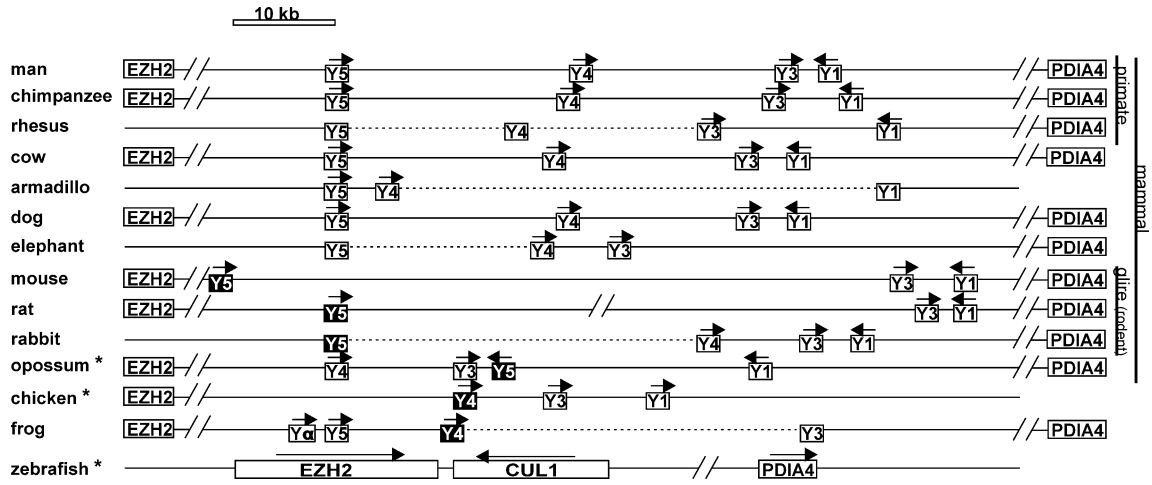


FIG. 2.—Synteny of the known and putative Y genes in various genomes. The relative positions of the Y genes (white boxes) and of the "fossil"Y genes (black boxes) are shown. The polarity of the genes is indicated by arrows (except where the assembly is incomplete and the orientation could not be determined). The lines are broken where the distance exceeds that which can be illustrated. The boxed EZH2 and PDIA4 represent genes that surround the Y RNA cluster in most species. The dotted lines correspond to regions of insufficient data due to incomplete assembly, rendering impossible to assign a position relative to the other genes. The same is also the case for any missing PDIA4 and EZH2 genes. The asterisks indicate species presenting an organization of their Y genes that differs from the majority. The scale bar represents 10 kb. Zebrafish is only included here in order to illustrate the absence of Y genes in this region.

Interestingly, sequences homologous to Y4 were found at the positions expected for their corresponding genes within the genomes of both the frog (*X. tropicalis*) and the chicken. The same was also true for Y5 in mouse, rat, rabbit, and opossum (fig. 2, in the dark boxes). However, these Y-related sequences all lacked important Y RNA features such as the bulged cytosine essential for Ro60 binding. Moreover, as previously reported, we were unable to detect any expression of Y5 RNA in rat and mouse by Northern blot hybridization, even when using probes fully complementary to the predicted Y5 RNA sequence (data not shown). These Y4 and Y5 sequences thus likely represent "fossil" genes that have lost their function.

Evolution of the Current Y RNA Genes from Hypothetical Ancestors

Based on sequence homologies between Y genes of the same type (i.e., Y1, Y3, Y4, and Y5) across various species, a presumptive path of evolution that would require the least number of events to occur from a proposed "ancestral" Y RNA gene was established (fig. 3). Since some Y RNAs are not found in some species, there are slight differences between the putative evolutionary trees of each Y RNA gene. For example, we could not define a tetrapod ancestor for Y4 and Y5, since these genes were not found before the divergence of amniotes and frogs (there is no outgroup available).

Based on figure 3, it is clear that the Y1, Y3, and Y4 RNA genes are well conserved in amniotes, while Y5 diverges slightly more. In addition, some positions in each gene vary much more frequently, such as position 101 in hY1, which is A in rabbit, C in marsupials, and T in the other species. It must also be remembered that different nucleotide variations (e.g., substitution, addition, or deletion) may have occurred independently in taxons or subtaxons originating from a common ancestor. For example, the putative amniote ancestor Y3 most likely harbored a T at position 13 which was passed on to early diapsids and mammals, but later mutated to a G in birds and an A in muridea (see fig. 3). On other occasions, independent "same-position variations" may have given rise to the same nucleotide. This may especially apply to CG dinucleotides that will most often drift to TG or CA (e.g., C57 in hY1 that is a T in rodents or A40 in hY4 that is a G in ancient mammals).

Incorporating the sequences of pseudogenes in the analysis helped, in some cases, to infer theoretical ancestral sequences that may be more reliable, at least in the case of mammals where pseudogenes are more numerous (see below and table 2). In addition, including pseudogene sequences permitted a higher resolution on the evolution timescale. For instance, rat Y3 pseudogenes all possess a T at position three, as do rat and mouse Y3s. As most other Y3 RNA genes (including all genes from mammals) encode a C at position 3, this strongly suggests a origin of these pseudogenes from the ancestral gene of both mouse and rat. Furthermore, the C at position 11 in the rat Y3 gene is not found in the Y3 pseudogenes found in the rat. As rat Y3 pseudogenes contain an A at position 11, as is observed in other species including the mouse, it can be suggested that mutation of the A/C at position 11 of the rat Y3 gene only happened after all Y3 pseudogene retroposition events occurred in rat. In conclusion, combining data on all Y genes, and adding in data on pseudogenes when needed, made it possible to suggest a putative ancestral gene and a hypothetical evolutionary tree for each of the 4 Y RNA genes (fig. 3).

A Putative "Universal" Y RNA Secondary Structure

In the « universal » Y RNA (fig. 1B), the double-stranded stem resulting from base-pairing of the 5′ and

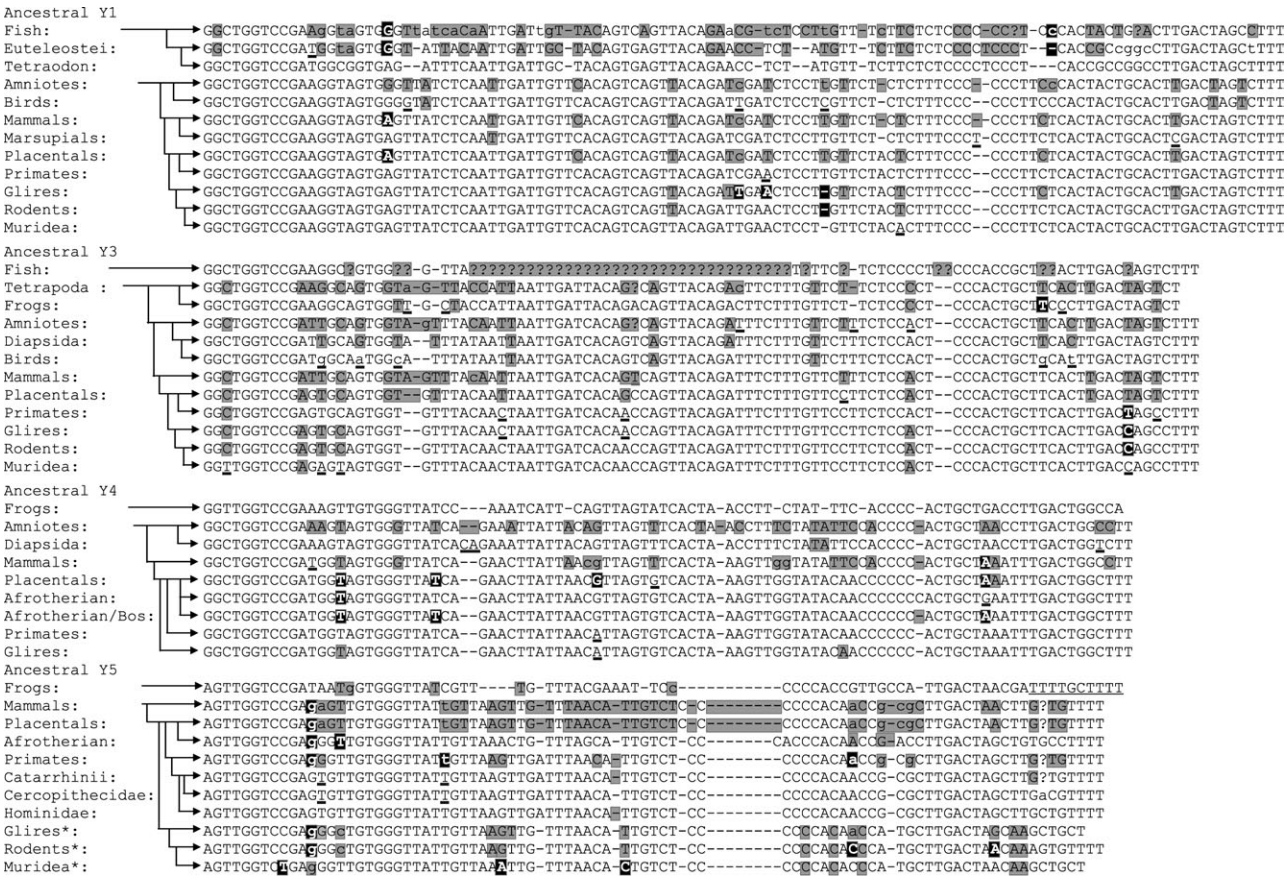


FIG. 3.—Hypothetical reconstitution of the four ancestral Y genes. For each taxon indicated on the left, the sequence of the most ancient common ancestor has been deduced from at least two species within the taxon and one outside of it (with the exception of fish for which no outgroup was available). The taxons denoted with an asterisk have a putative fossil Y. The arrows on the right indicate the order of descent. The gray boxes indicate that one or more species differ from the proposed ancestor, while the black boxes indicate that the variant in question is also found in other taxons. The underlined bases are different from those of the presumed immediate ancestor. Positions that are ambiguous because a single outgroup species was available are indicated in lower case. Positions are replaced with question mark (?) when no single base appears more probable than the others.

3′ ends was highly conserved, as was the 3′ poly(U) tail. These regions are also essential for binding the Ro60 and La proteins, respectively (Wolin and Steitz 1984; Farris et al. 1999; Teunissen et al. 2000). However, the short 3′ poly(U) tail is not always present in the RNA gene itself, but rather originates from the polymerase III terminator signal, and it may be lost or conserved in the transcript, depending on the species. Conversely, the regions forming the central loop and other domains were relatively divergent between Y RNAs of the same type (e.g., Y1), and they exhibited very little homology between different types of Y RNAs. A few conserved features were observed within the central portions, including a small stem in both Y4 and Y3, as well as a polypyrimidine stretch found in all four Y types. To better illustrate this, the "consensus" sequence and structure for each Y1, Y3, Y4, and Y5 RNA was deduced (fig. 4). Moreover, the secondary structures of the Yα and some of the new Y genes found in roundworm and mosquito are also illustrated. The Y3 and Y4 central stems appear to be relatively well conserved. A similar stem is also found in Y5 and, although its sequence and length are less well conserved, it is composed of at least 4 bp in all species except cattle. Similarly, the Y1 left stem appears to be relatively well conserved, even in zebrafish and pufferfish. The presence of this stem received additional support from the observation of base pair covariation where only the trout Y1 RNA was found to differ significantly.

Analysis of the Y Pseudogenes

The BLATz search for Y homologous sequences revealed the existence of Y pseudogenes in almost all mammals (table 2). The only exception was the wallaby that appears to have only one Y gene and no detectable pseudogenes. However, this negative result needs to be re-evaluated once the wallaby genome is completely sequenced. In contrast, human and chimpanzee present a very large number of Y pseudogenes, with a large predominance for those derived from Y1 and Y3. The pseudogenes derived from Y5 genes are present in much smaller amounts (i.e., at least 20-fold less as compared to either Y1 or Y3). We found at least two other genomes that exhibited a high density of Y pseudogenes: the guinea pig and the elephant. In the case of the elephant, careful analysis of the 124 Y1 sequences revealed that most, if not all, were in fact fragments of a larger repeat element specific to this animal. As a consequence, the number of true Y1 pseudogenes in the elephant is much smaller.

In the case of the guinea pig, this animal contrasts sharply with other rodents. For example, all 4 Y RNA genes are still active in the guinea pig, while Y5 is silent in all other rodents tested and Y4 is inactivated in rat and mouse. The guinea pig genome also contains nearly 20 times more Y pseudogenes than do the other rodents. This will also need to be re-evaluated, since at the time of this study the guinea pig genome was sequenced but not annotated. Since the current coverage of the guinea pig genome is approximately 2, and since it is not yet assembled, the approximate number of guinea pig Y pseudogenes was grossly estimated to be about half the number of BLATz hits, or about one thousand.

The type of Y RNA giving rise to the largest number of Y pseudogenes in other mammals varies depending on the species. For example, Y1 pseudogenes are more abundant in mouse and rat, but Y3 pseudogenes predominate in rabbit and dog. Regardless of the species, Y5 pseudogenes were observed to be markedly less abundant than the others. Two possible explanations for this observation were a very low expression level in germ cells and Y5 RNA possessing unique biochemical or physiological properties that render it less prone to L1 retroposition. In order to verify the first hypothesis, RNA samples from several organisms were obtained and analyzed by Northern blot hybridization using a probe specific for hY5. Y5 RNA appeared to be abundantly expressed in guinea pig testis (fig. 5). Therefore, the first hypothesis appeared unlikely. Conversely, the second hypothesis is supported by the fact that the Ro RNPs containing hY5 RNA were shown to be specifically bound by the protein RoBPI (Bouffard et al. 2000), and were found to be localized mostly in the nucleus, rather than in the cytoplasm (Gendron, Roberge, and Boire 2001). Clearly, the forces that govern the formation of Y pseudogenes are variable between organisms and Y RNA type, and they remain to be identified. However, the absence of Y pseudogenes in taxa other than mammals appears unambiguous, consistent with the lower activity of retroposition in most of these species in which LINE1 elements are absent (Ohshima and Okada 2005).

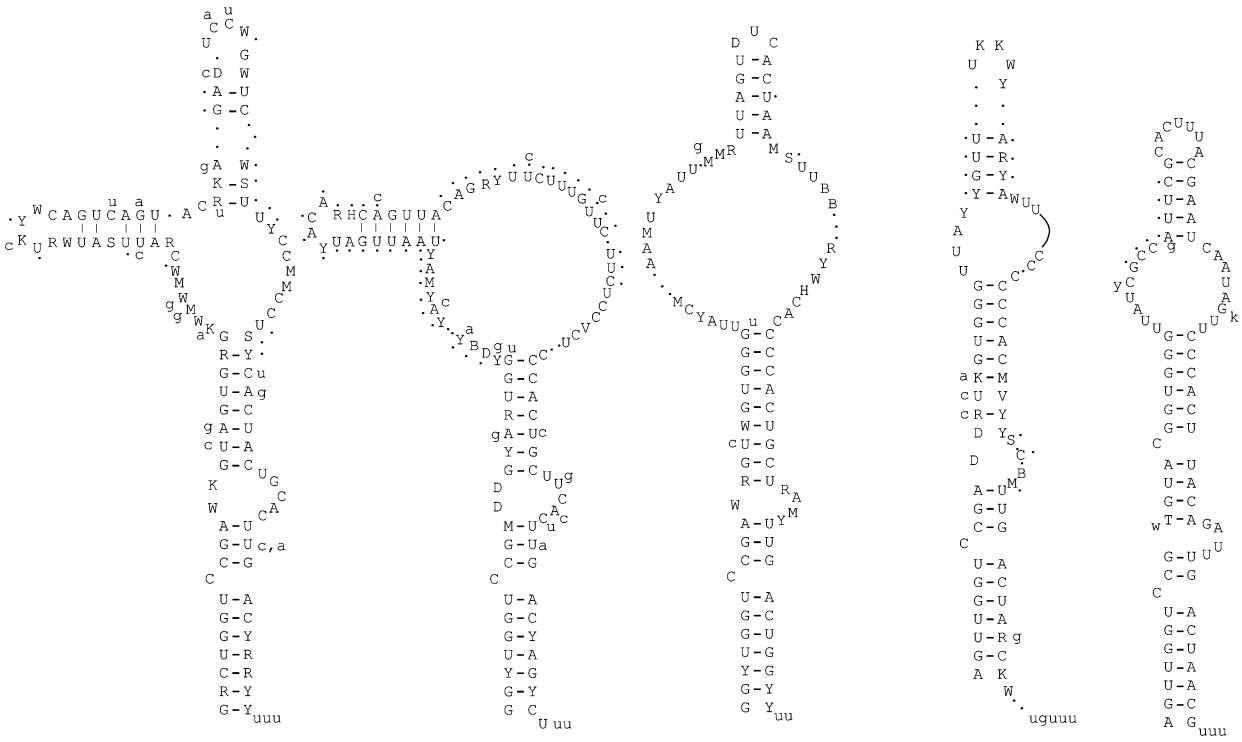
Most Y pseudogenes have similar degrees of homology to their gene of origin, irrespective of the type of Y RNA. However, among the species with a number of pseudogenes sufficiently large for statistical analysis, the primates and the guinea pig appear to be markedly different. Guinea pig pseudogenes are about 95% identical to their Y RNA gene of origin, while primate pseudogenes are less than 90% identical to theirs. This may represent a more recent burst of retroposition of Y RNAs in guinea pig, about 35 MYA, in comparison to that of 70 MYA for the primate pseudogenes. However, these numbers must be viewed with caution, as nonessential DNA from species with shorter life spans (e.g., the muridea lineage) may drift more rapidly than is the case in primates. In conclusion, even if the degree of divergence of pseudogenes may not be directly proportional to the time elapsed since the formation of the pseudogenes, there are obvious differences in Y RNA retroposition between species.

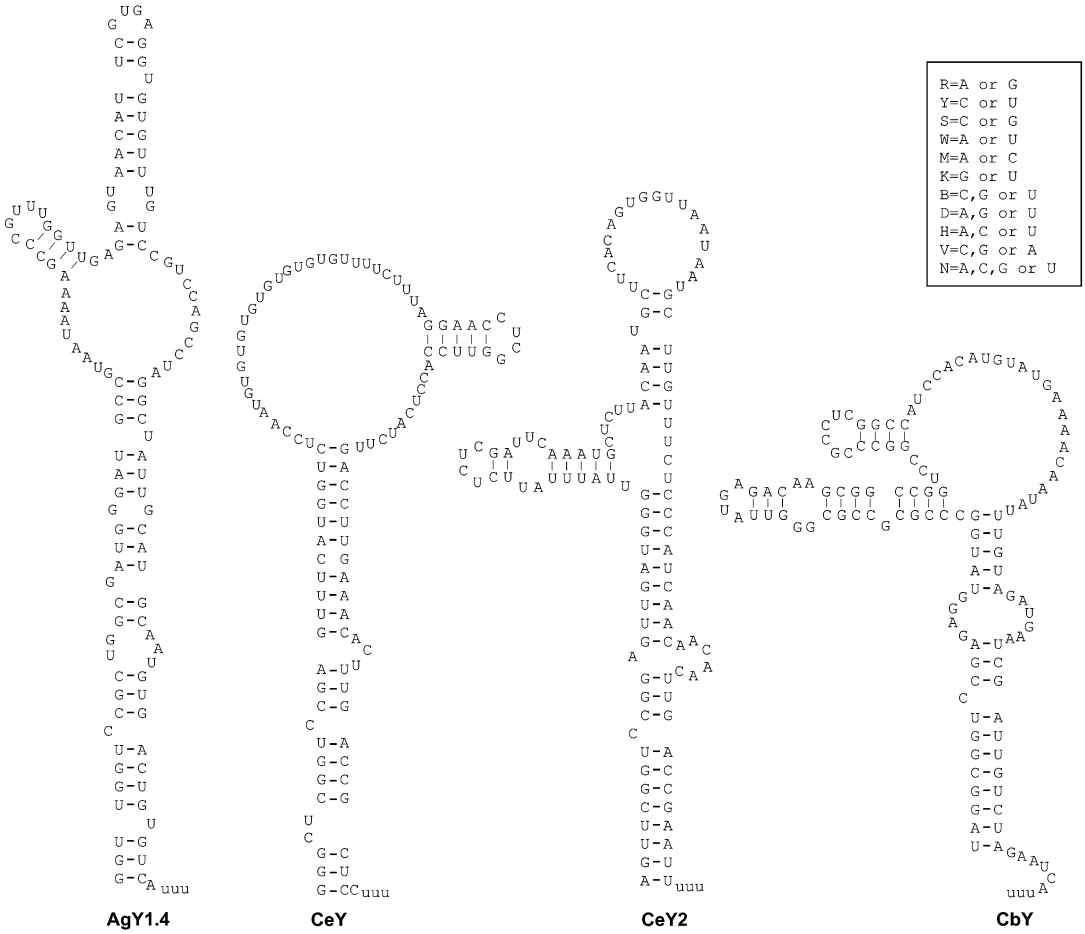
To better assess the timescale of insertion of Y pseudogenes in mammalian genomes, we compared the first 500 bp of their sequences located 5′ and 3′ of the Y homologous sequences. A good alignment of the Y pseudogene and its adjacent regions in a given species with its couterpart Y pseudogene in another species strongly indicated that they originated from the same retroposition event. This analysis was performed in primates (i.e., human, chimpanzee and rhesus monkey) as well as in rodents (i.e., mouse and rat). At first, marmoset was included with the primates, while guinea pig and rabbit were included with the rodents, but these more divergent species provided little useful information and their inclusion tended to obscure any interesting results. Thus, they were excluded from the final analysis. Only three examples of potential "pre-insertion" sites were found in our databank, specifically pseudogenes found in both humans and chimpanzees, but not in rhesus monkeys. In two of these cases, the "pre-insertion" sites in the rhesus genome harbored a sequence corresponding to the terminal 5′ portion of Y3 followed by a poly(A) (see online Supplementary Material 3). Furthermore, in one of these two pseudogenes a 47 nucleotides Y3 pseudogene appears to have been truncated due to similar sequence between RNA in Y3 midportion and the chromosome. This hints towards the mechanism by which L1 retroposed the Y RNAs, even if three cases are not sufficient evidence for any convincing conclusion to be drawn. Although these three pseudogenes likely represent a more recent insertion in humans and chimpanzees, we cannot exclude the possibility of their recent deletion from the rhesus genome.

Finally, in order to refine the resolution of the analysis of Y pseudogene insertion during evolution in primates, 11 pseudogenes were amplified from genomic DNA extracted from mandrill and saki monkeys, two primates relatively distant from humans whose genome is still unsequenced (see online Supplementary material 4). No PCR product suggestive of the presence of a "pre-insertion" site in saki or mandrill was amplified. A pre-insertion site would be suggested by the presence of a PCR product shortened by the number of base pairs corresponding to the pseudogene, its polyA tail and the target site duplication. Thus, even primates from the more distant lineages appear to share most of their Y pseudogenes with humans, further supporting the hypothesis of a burst of Y insertion around 70 MYA.

→

FIG. 4.—Nucleotide sequences and secondary structures of the known and putative metazoan Y RNAs. The secondary structures for the previously known Y RNAs come from previous studies (see text), and the consensus sequence is derived from all previously known metazoan sequences (Y1, Y3, Y4, Y5, Yα, and CeY). Lowercase letters indicate differences between this consensus sequence and the new sequences found in this study. A dot indicates the possibility of a nucleotide at that position (when placed in the structure), or the possibility of the absence of a nucleotide (when placed outside of the structure). Solid lines indicate the presence of varying numbers of nucleotides. The structures of the new types of putative Y RNAs (i.e., AgY1.4, CeY2, and CbY) have been predicted with Mfold (Zuker 2003) using constraints similar to those used to build the consensus sequence. The terminal 3′"uuu"is in lowercase because it is always found in the terminator, but not always in RNA. The inset shows the IUPAC code for the nucleotide nomenclature.


Y1
Y3
Y4
Y5
Yα


R=A or G
Y=C or U
S=C or G
W=A or U
M=A or C
K=G or U
B=C,G or U
D=A,G or U
H=A,C or U
V=C,G or A
N=A,C,G or U
AgY1.4
CeY
CeY2
CbY

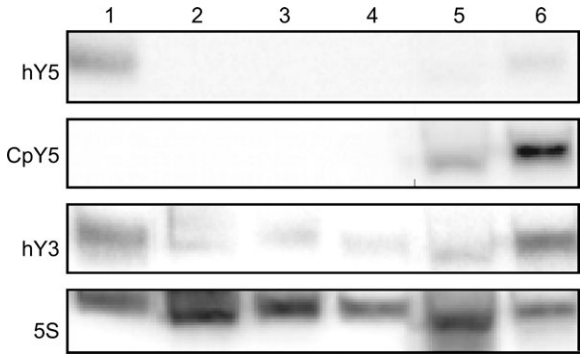


FIG. 5.—Autoradiogram of a Y5 Northern blot hybridization. The probes used are shown on the left: human hY5, guinea pig CpY5, human hY3 and human 5S. Lane 1 is total RNA isolated from the Huh-7 cell line (human liver). Lanes 2 to 5 correspond to total RNA isolated from the liver of mouse, rat, rabbit and guinea pig, respectively. Lane 6 is total RNA isolated from guinea pig testis.

Discussion

Through homology and structural motif bioinformatic searches, we have identified an impressive number of previously unknown Y RNA genes, and have completed the partial sequencing of eight additional genes. The confirmation of synteny across species strongly suggested that these new Y gene sequences were putative genes, most of them being found within 100 kb of each other. In particular, we found a sequence that perfectly fits the consensus Y structure in *Anopheles gambiae*. This might represent the first insect found to encode Y RNAs, and could imply that the model organism *Drosophila melanogaster* might also possess Y genes. Indeed, a less stringent rnamotif search revealed the existence of potential Y RNA sequences in this organism (see online Supplementary Material 5). The possibility that these sequences are putative Y RNA genes is also supported by the presence of an ortholog of Ro60 protein in *Drosophila* and other insects (Holt et al. 2002; Stapleton et al. 2002; Honeybee genome sequencing consortium 2006). The existence of Y RNA genes in insects would open up new possibilities for genetic research on Y RNAs and Ro RNPs. Similarly, the finding of a second putative Y RNA gene within the *C. elegans* genome, as well as of two Y RNA sequences in *C. briggsae*, is of major interest for researchers studying the functions of Ro RNPs in nematodes. For its part, Ro protein is not found in plants, and fungi, and is present in only a few prokaryotes, including *Deinococcus radiodurans*. The Ro ortholog in this bacteria likely comes from a horizontal transfer from a metazoan (Chen and Wolin 2004). On the other hand, the Y RNA genes in *D. radiodurans* appear to originate from the bacteria itself, given their absence of homology with any other Y RNA.

The ancestral sequences deduced from all the compiled Y RNA genes (and some of their pseudogenes) also yielded some results that may contribute to the controversy surrounding the classification of rodents as either monophyletic or polyphyletic. Although the Y genes alone provide too little sequence data to permit robust conclusions, many nucleotides in rodent and glire Y RNAs are dark shaded, indicating a divergence within the family (fig. 3). Some species even have more closely related Y RNA sequences in another taxon. For example, very distinct differences exist between guinea pig and the muridea. The guinea pig has conserved the four Y genes, while the muridea only express the Y1 and Y3 RNAs. Their pseudogene pattern is also astonishingly different: guinea pigs possess about twenty times more Y pseudogenes than do the muridea, more closely resembling the primates in that regard. Other aspects of the biology of guinea pigs more closely resemble that of primates than rodents, such as hematopoiesis and thymus development (Katz and Altman 1979).

To provide a general view on all Y RNA genes and pseudogenes retrieved in the present work, a phylogenic tree was constructed (fig. 6). A combination of NCBI and "Tree of Life" taxonomy was used to decide which species (and their corresponding Y sequences) were to be grouped together. Clearly, all mammals possess Y genes, and most have four Y genes, although only inactivated Y5 and occasionally Y4 genes remain as fossil genes in some species. Conversely, some invertebrates possess putative Y genes of types that cannot be classified as Yα, Y1, Y3, Y4, or Y5. Between these two extremes, the oldest vertebrate class (i.e., fish) possesses only putative Y1 and Y3 genes, with highly divergent intra-class sequences in some cases. It seems reasonable to propose that the first Y RNA gene was of very ancient origin and has diverged to give rise to the two putative Y RNAs present in nematodes and to Y1 and Y3 RNAs in the ancestor of fish. Later on, the Y4 and Y5 genes would have appeared in amphibians. The four types of Y genes would then be conserved to a varying extent in mammals, birds, and lizards. During their evolution, amphibians (or at least frogs) would have lost the Y1 gene and gained a new one, specifically Yα. Concurrent with the Y gene diversification, pseudogenes appeared as illustrated in figure 6. This illustration indicates that an abundant number of pseudogenes appeared in the common primate ancestor, and persisted in all descendant species. Surprisingly, the only other organism whose genome is currently sequenced having a large amount of Y pseudogenes was the guinea pig.

The phylogenic tree also shows that the putative Y4 and Y5 genes are more broadly distributed than previously speculated (see fig. 6). In fact, these genes are present in most mammals, although some, such as the mouse and rat, do not express the corresponding RNAs. Nonetheless, both the mouse and rat genomes contain a degenerated Y5 sequence located about 100 kb from the mY3 gene at the position expected for a bona fide Y5 gene (fig. 2). This "fossil" Y5 RNA gene contains mutations that are likely to affect the function of any expressed RNA, such as disruption of the predicted Y secondary structure (data not shown), thus preventing the binding of the Ro60 protein. A role for Y RNAs independent of the Ro60 protein was recently described (Christov et al. 2006), and a transition state may be proposed in which the mutated mY5 RNA was still useful to the organism despite the "deleterious" mutations preventing Ro60 binding, thus allowing the RNA to be retained for some time in the genome. However, due to an eventual compensation by the Y RNAs still expressed, shown to be able to complement its DNA replication function, additional mutations might have suppressed

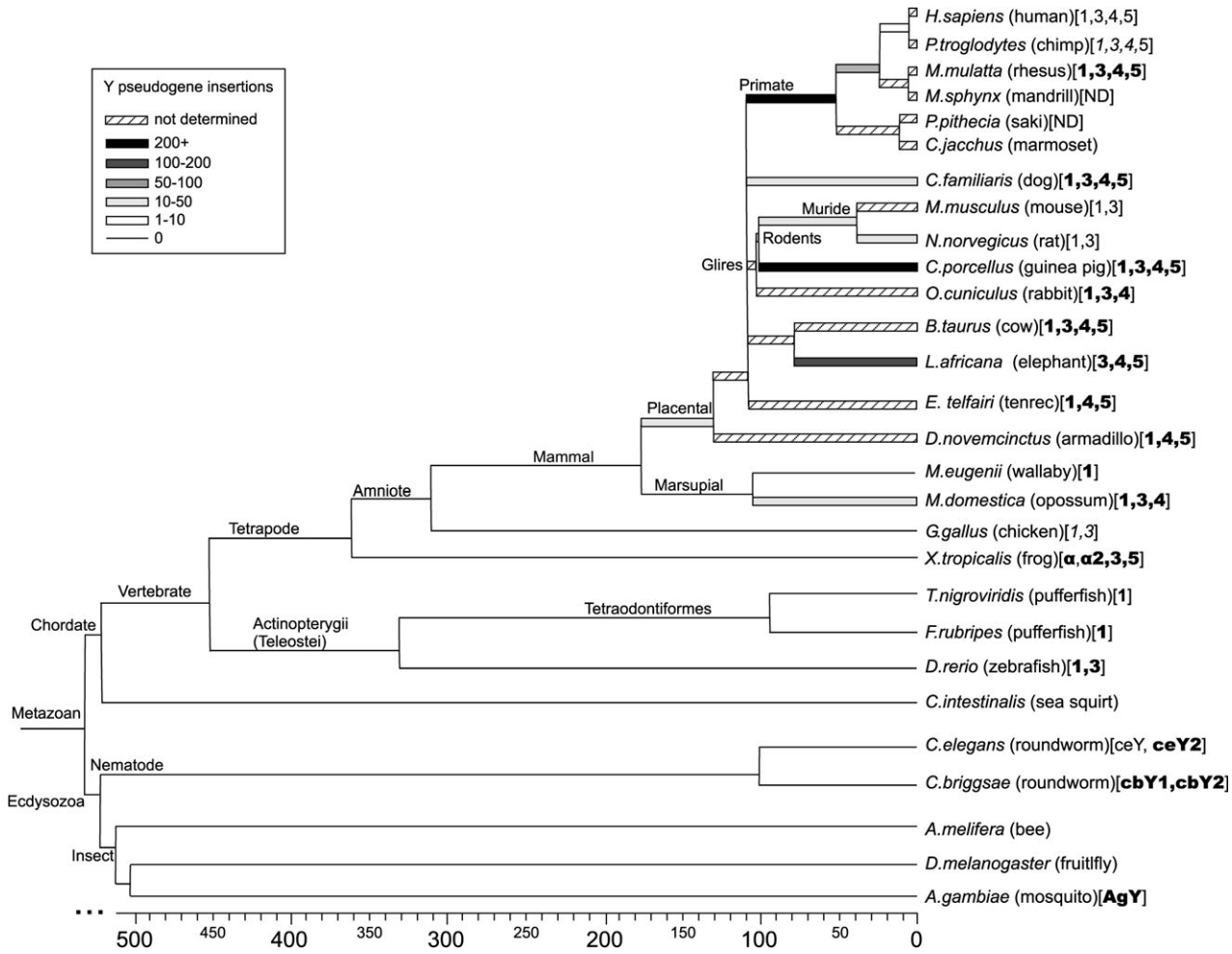


FIG. 6.—Hypothetical phylogenic tree of the species analyzed. The systematic names are followed by the common names in parentheses. The identified Y genes are indicated in brackets. Normal font indicates well known complete genes, *italic*previously predicted genes, **bold** is used for new genes. Most phylogenic relationships were adapted from "The Tree of Life Web Project"(http://tolweb.org), and the timescale was derived from Kumar and Hedges (1998), Hedges (2002), Stein et al. (2003) and Yamanoue et al. (2006). Divergence time before 450 million years ago could not be determined and can go as far as 1 billion years ago. Note that this timescale is controversial and is used only as an approximation. Similarly, the tree topology is also controversial, especially for mammalian orders and families. The shades of the thick lines in mammals indicate the number of Y pseudogene insertions in the corresponding phylum as indicated in the insert.

the expression of the mutated Y5 RNA in mouse. A similar scenario seems to be plausible with Y4 in both chicken and *Xenopus tropicalis*, as well as with Y5 in opossum and rabbit, species where a probable "fossil" gene in correct synteny can still be detected.

Our data indicate the presence of the four Y RNA genes (Y1, Y3, Y4, and Y5) among most tetrapods, from frog to human, although Y1, Y4, or Y5 RNA genes were lost in some species. This suggests that these four Y RNAs already existed in the common ancestor of tetrapods. This is in sharp contrast with the hypothesis that the hY5 RNA gene only recently arose early in primate evolution (Maraia et al. 1996), possibly through retroposition of the hY4 RNA gene (O'Brien and Harley 1992). This broad distribution and long presence during evolution of the Y5 RNA gene contrasts with the fact that Y5 genes are more divergent within primate lineage than are the other Y genes. Among Y RNA genes, the Y5 RNA genes might thus be subject to distinct evolutionary pressure that might be related to specific functions of the RNA or of the proteins with which it associates.

Our proposition for the long existence of the Y4 and Y5 genes during evolution also shatters the previous explanations for the very large differences in the numbers of pseudogenes derived from each Y genes. In humans, Y4 pseudogenes are much less prevalent than Y1 and Y3 pseudogenes, while Y5 pseudogenes represent less than 2% of all Y pseudogenes. We previously proposed an explanation primarily based on temporal reasons (Perreault et al. 2005), which does not stand anymore. The differential expression of Y RNAs in germ line cells is also not likely to be at the origin of the low abundance of Y5 pseudogenes, according to the data presented in figure 5. However, the differential localization that has been reported for Y5 RNA (i.e., nuclear as compared to cytoplasmic for the other hY RNAs) (Gendron, Roberge, and Boire 2001) might contribute to this contrasting retroposition pattern. Alternatively, differences in the proteins binding each of the Y RNAs could also be responsible for their contrasting retroposition patterns. For example, Ro RNPs containing the hY5 RNA were shown to bind RoBP1 (Bouffard et al.

2000), while ribonucleoprotein complexes containing hY1 and hY3 RNAs bound the hnRNP K and PTB proteins (Wolin and Steitz 1984; Fabini et al. 2000, 2001; Fouraux et al. 2002). Such differences suggest distinct biochemical, and possibly physiological, properties. For example, the La, hnRNP K, and PTB proteins that are linked to the Y1 and Y3 RNAs have been shown to be involved in IRES (internal ribosome entry site) dependent translation (Isoyama et al. 1999; Pilipenko et al. 2001; Evans et al. 2003). Hence, it is tempting to speculate that the Y1 and Y3 RNAs may be brought close to the ribosomes translating L1 protein, and may therefore be in a good position to be bound by the retrotransposon machinery, as was previously proposed for Alu retroposition (Boeke 1997; Dewannieux, Esnault, and Heidmann 2003). Clearly, a better understanding of the Y RNA function and of their interaction with the cellular mechanism of retroposition will be instructive in this regard.

In addition to providing important informations on the history of Y RNA genes, our data provide some additional insight into retroposition in the mammalian lineage, the most striking feature of which is the complete lack of a common rule. The true Y RNAs are apparently responsible for the observed insertions of Y pseudogenes. This contrasts with what is observed with either the Alu insertions present primarily in primates or with the B2 elements present in mouse (Serdobova and Kramerov 1998; Nishihara, Terai, and Okada 2002) that both derive from a different gene, the 7SL RNA. In addition, the patterns of proliferation of Y pseudogenes are extremely different from one genome to another. For example, muridea species have only a few pseudogenes (including a few derived from the Y4 gene that is now inactivated), while primates and guinea pig have approximately 1,000 (table 2). To add to this absence of rules, we observed that Y pseudogenes of some mammals will often miss a 3′ portion of the inserted Y sequence (e.g. glires and dog). This is particularly puzzling since retroposition is theoretically more likely to be accompanied by a 5′ truncation due to early termination of transcription by the L1 reverse transcriptase, a phenomenon known to be widely present in primates. Whether the burst of Y pseudogenes in primates and guinea pig was the source of some new function(s) for these sequences in these mammals, or if these sequences are simply the consequence of different attempts to repress retroposition in the different mammal families remains to be found.

In conclusion, we hope that this contribution to Y RNA evolution will suggest some testable hypotheses regarding their functions and their mechanism of retroposition, and that the expression of the novel putative Y RNA sequences we report will be confirmed experimentally.

Note added in proof

After this work was submitted, we became aware of the submission to another journal of an independent study by Mosig et al. 2007, which reaches similar conclusions.

Acknowledgments

The authors thank Dr. Marie-Josée Limoges (Director of Conservation and Animal Health, Granby Zoo, Québec, Canada) for blood samples from mandrill, Japanese macaques, and white face saki monkeys. We also thank Dr. J. W. Kent for the latest version of BLATz software he made available to us prior to its publication, and Emmanuel Tremblay-Rousseau for help with "pre-insertion" site analysis. G.B. and J.P.P are members of the *Fonds de la recherche en santé du Québec* (*FRSQ*)-funded *Centre de recherche clinique Étienne-Le Bel.*

This work was supported by grants from the Canadian Institutes of Health Research (CIHR) to J.P.P. (EOP-38322) and to the RNA group (PRG-80169). The RNA group is also supported by grants from both the Université de Sherbrooke and CIHR (infrastructure grant). J.P. was the recipient of a predoctoral fellowship from FRSQ (Québec). J.P.P. holds the Canada Research Chair in Genomics and Catalytic RNA.

Literature Cited

Boeke JD. 1997. LINEs and Alus-the polyA connection. Nat Genet. 16:6–7.

Bouffard P, Barbar E, Briere F, Boire G. 2000. Interaction cloning and characterization of RoBPI, a novel protein binding to human Ro ribonucleoproteins. RNA 6:66–78.

Chen X, Quinn AM, Wolin SL. 2000. Ro ribonucleoproteins contribute to the resistance of Deinococcus radiodurans to ultraviolet irradiation. Genes Dev. 14:777–782.

Chen X, Smith JD, Shi H, Yang DD, Flavell RA, Wolin SL. 2003. The Ro autoantigen binds misfolded U2 small nuclear RNAs and assists mammalian cell survival after UV irradiation. Curr Biol. 13:2206–2211.

Chen X, Wolin SL. 2004. The Ro 60kDa autoantigen: insights into cellular function and role in autoimmunity. J Mol Med. 82:232–239.

Chenna R, Sugamara H, Koike T, Lopey R, Gibson TJ, Higgins DJ, Thompson JD. 2003. Multiple sequence alignment with the Clustal serie program. Nucleic Acids Res. 31:3497–3500.

Christov CP, Gardiner TJ, Szuts D, Krude T. 2006. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. Mol Cell Biol. 26:6993–7004.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 35:41–48.

Evans JR, Mitchell SA, Spriggs KA, Ostrowski J, Bomszyktyk K, Ostarek D, Willis AE. 2003. Members of the poly (rC) binding protein family stimulate the activity of the c-myc internal ribosome entry segment in vitro an in vivo. Oncogene. 22:8012–8020.

Fabini G, Rutjes SA, Zimmermann C, Pruijn GJ, 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.

Fabini G, Raijmakers R, Hayer S, Fouraux MA, Pruijn GJ, 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.

Farris AD, O'Brien CA, Harley JB. 1995. Y3 is the most conserved small RNA component of Ro ribonucleoprotein complexes in vertebrate species. Gene. 154:193–198.

Farris AD, Koelsch G, Pruijn GJ, van Venrooij WJ, Harley JB. 1999. Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis. Nucleic Acids Res. 27:1070–1078.

Fouraux MA, Bouvet P, Verkaart S, van Venrooij WJ, Pruijn GJ. 2002. Nucleolin associates with a subset of the human Ro ribonucleoprotein complexes. J Mol Biol. 320:475–488.
Fuchs G, Stein AJ, Fu C, Reinisch KM, Wolin SL. 2006. Structural and biochemical basis for misfolded RNA recognition by the Ro autoantigen. Nat Struct Mol Biol. 13:1002–1009.
Gendron M, Roberge D, 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.
Green CD, Long KS, Shi H, Wolin SL. 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.
Hedges SB. 2002. The origin and evolution of model organisms. Nat Rev Genet. 3:838–849.
Honeybee genome sequencing consortium. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 26:931–949.
Holt RA, Subramanian GM, Halpern A, et al. (120 co-authors). 2002. The genome sequence of the malaria mosquito Anopheles gambiae. Science. 298:129–149.
Isoyama T, Kamoshita N, Yasui K, Iwai A, Shiroki K, Toyoda H, Yamada A, Takasaki Y, Nomoto A. 1999. Lower concentration of La protein required for internal ribosome entry on hepatitis C virus RNA than on poliovirus RNA. J Gen Virol. 80:2319–2327.
Itoh Y, Kriet JD, Reichlin M. 1990. Organ distribution of the Ro (SS-A) antigen in the guinea pig. Arthritis Rheum 33: 1815–1821.
Jegga AG, Sherwood SP, Carman JW, Pinski AT, Phillips JL, Pestian JP, Aronow BJ. 2002. Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. Genome Res. 12:1408–1417.
Katz DD, Altman PL. 1979. Guinea pig. In: Katz DD, Altman PL, editors. Inbred and Genetically Defined Strains of Laboratory Animals. Bethesda, MD: Federation of American Societies for Experimental Biology.
Kent JW. 2003. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664.
Kent JW, Baertsch R, Hinrich RA, Miller W, Haussler D. 2002. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA. 100:11484–11489.
Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. Nature. 392:917–920.
Labbé JC, Burgess J, Rokeach LA, Hekimi S. 2000. ROP-1, an RNA quality-control pathway component, affects Caenorhabditis elegans dauer formation. Proc Natl Acad Sci USA. 97:13233–132338.
Lawley W, Doherty A, Denniss S, Chauhan D, Pruijn G, van Venrooij WJ, Lunec J, Herbert K. 2000. Rapid lupus autoantigen relocalization and reactive oxygen species accumulation following ultraviolet irradiation of human keratinocytes. Rheumatology. 39:253–261.
Macke T, Ecker D, Gutell R, Gautheret D, Case DA, Sampath R. 2001. RNAMotif—A new RNA secondary structure definition and discovery algorithm. Nucleic Acids Res. 29:4724–4735.
Maraia RJ, Sasaki-Tozawa N, Driscoll CT, Green ED, Darlington GJ. 1994. The human Y4 cytoplasmic RNA gene is controlled by upstream elements and resides on chromosome 7 with all other hY scRNA genes. Nucleic Acids Res. 22:3045–3052.
Maraia R, Sakulich AL, Brinkmann E, Green ED. 1996. Gene encoding human Ro-associated autoantigen Y5 RNA. Nucleic Acids Res. 24:3552–3559.
Nishihara H, Terai Y, Okada N. 2002. Characterization of novel Alu- and tRNA-related SINEs from the tree shrew and evolutionary implications of their origins. Mol Biol Evol. 19:1964–1972.
O'Brien C, Harley JB. 1992. Association of hY4 pseudogenes with Alu repeats and abundance of hY RNA-like sequences in the human genome. Gene. 116:285–289.
O'Brien CA, Margelot K, Wolin SL. 1993. Xenopus Ro ribonucleoproteins: Members of an evolutionarily conserved class of cytoplasmic ribonucleoproteins. Proc Natl Acad Sci USA 90:7250–7254.
Ohshima K, Okada N. 2005. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res. 110:475–490.
Pellizzoni L, Lotti F, Rutjes SA, 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.
Perreault J, Noel JF, Briere F, Cousineau B, Lucier JF, Perreault JP, Boire G. 2005. Retropseudogenes derived from the human Ro/SS-A autoantigen-associated hY RNAs. Nucleic Acids Res. 33:2032–2041.
Pilipenko EV, Viktorova G, Guest ST, Agol VL, Roos RP. 2001. Cell-specific proteins regulate viral RNA translation and virus-induced disease. EMBO J. 20:6899–6908.
Serdobova IM, Kramerov DA. 1998. Short retroposons of the B2 superfamily: evolution and application for the study of rodent phylogeny. J Mol Evol. 46:202–214.
Shi H, O'Brien CA, Van Horn DJ, Wolin SL. 1996. A misfolded form of 5S rRNA is complexed with the Ro and La autoantigens. RNA. 2:769–784.
Stapleton M, Liao G, Brokstein P, et al. (12 co-authors). 2002. The Drosophila Gene Collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. Genome. 12:1294–1300.
Stein LD, Bao Z, Blasiar D, et al. (36 co-authors). The genome sequence of Caenorhabditis briggsae: a platform for comparative genomicsPLoS Biol20031E45.
Stein AJ, Fuchs G, Fu C, Wolin SL, Reinisch KM. 2005. Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. Cell. 121:529–539.
Teunissen SW, Kruithof MJ, Farris AD, Harley JB, Venrooij WJ, Pruijn GJ. 2000. Conserved features of Y RNAs: a comparison of experimentally derived secondary structures. Nucleic Acids Res. 28:610–619.
Van Horn DJ, Eisenberg D, O'Brien CA, Wolin SL. 1995. *Caenorhabditis elegans* embryos contain only one major species of Ro RNP. RNA. 1:293–303.
Wolin SL, Steitz JA. 1983. Genes for two small cytoplasmic Ro RNAs are adjacent and appear to be single-copy in the human genome. Cell. 32:735–744.
Wolin SL, Steitz JA. 1984. The Ro small cytoplasmic ribonucleoproteins: identification of the antigenic protein and its binding site on the Ro RNAs. Proc Natl Acad Sci USA. 81:1996–2000.
Yamanoue Y, Miya M, Inoue JG, Matsuura K, Nishida M. 2006. The mitochondrial genome of spotted green pufferfish Tetraodon nigroviridis (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. Genes Genet Syst. 81:29–39.
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31: 3406–3415.

Sudhir Kumar, Associate Editor

Accepted April 20, 2007