ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

GENOMICS

Genomics 81 (2003) 468–480

www.elsevier.com/locate/ygeno

Identification and characterization of over 100 mitochondrial ribosomal protein pseudogenes in the human genome☆

Zhaolei Zhang and Mark Gerstein*

Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520-8114, USA

Received 29 August 2002; accepted 13 December 2002

Abstract

The human (nuclear) genome encodes at least 79 mitochondrial ribosomal proteins (MRPs), which are imported into the mitochondria. Using a comprehensive approach, we find 41 of these give rise to 120 pseudogenes in the genome. The majority of the MRP pseudogenes are of processed origin and can be aligned to match the entire coding region of the functional MRP mRNAs. One processed pseudogene was found to have originated from an alternatively spliced mRNA transcript. We also found two duplicated pseudogenes that are transcribed in the cell as confirmed by screening the human EST database. We observed a significant correlation between the number of processed pseudogenes and the gene CDS length ($R = -0.40$; $p < 0.001$), i.e., the relatively shorter genes tend to have more processed pseudogenes. There is also a weaker correlation between the number of processed pseudogenes and the gene CDS GC content. Our study provides a catalogue of human MRP pseudogenes, which will be useful in the study of functional MRP genes. It also provides a molecular record of the evolution of these genes. More details are available at http://pseudogene.org/.

Keywords: Mitochondrial ribosomal protein; MRP; Pseudogene; Evolution; Bioinformatics

Introduction

Mitochondria are descendents of ancient eubacteria, as they were incorporated into eukaryotes from endosymbiosis during early evolution. These organelles have their own genome and translation machinery: mitochondrial ribosomes. Like their cytoplasmic counterpart, mitochondrial ribosomes consist of two subunits: a small 28S subunit and a large 39S subunit. In total, a mitochondrial ribosome is composed of two RNA molecules and at least 70 different mitochondrial ribosomal proteins (MRPs) [1–4]. Progress has been made recently in determining the exact protein composition of mammalian ribosomes using 2-D electrophoresis and mass spectroscopy [5–8]. At present, 47 proteins are confirmed to be part of the large subunit and 31 proteins of the small subunit. All the mammalian MRPs are encoded by nuclear genes, which were part of the original mitochondrial genome but later migrated into the nuclear genome. These proteins are synthesized by the cytoplasmic ribosomes and imported into mitochondria to be assembled together with the mitochondrial rRNA molecules, which are still encoded by the mitochondrial genome. Despite their common evolutionary origin, many of the MRPs have no apparent homologues in the bacterial or eukaryotic cytoplasmic ribosomes, probably because of their rapid evolution rate [4,9]. The chromosomal locations of most of the human MRP genes have been determined by radiation hybrid mapping [10] and are placed on a megabase map now [11]. Other than protein biosynthesis, some of the MRPs are also implicated in other cellular processes such as GTP binding [12] and apoptosis [13,14].

Unlike the cytoplasmic ribosomal protein genes, there are very few reports of pseudogene sequences for MRP genes. In fact, as of June 2002, there are only 24 MRP pseudogene sequences in GenBank, of which 3 are from human (NG_000968, NG_000977, and AL161788). This is in sharp contrast to the cytoplasmic ribosomal proteins

☆ Sequence data from this article have been deposited with the GenBank Data Library under Accession Nos. AY135236–AY135355.

* Corresponding author. Fax: +1-360-838-7861.

E-mail address: Mark.Gerstein@yale.edu (M. Gerstein).

0888-7543/03/$ – see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0888-7543(03)00004-1

(RP), of which over 2000 pseudogenes have been discovered in the human genome [15,16]. Pseudogenes are disabled copies of functional genes that do not produce a functional, full-length protein [17–19]. The disablements can take the form of premature stop codons or frame shifts in the protein-coding sequence (CDS) or, less obviously, deleterious mutations in the regulatory regions that control gene transcription or splicing. There are two types of pseudogenes: the duplicated (nonprocessed) and the processed. Duplicated pseudogenes arose from tandem DNA duplication or unequal crossing over [17,18], so they have retained the identical, though often incomplete, exon structure of the original functional gene. Processed pseudogenes resulted from retrotransposition, i.e., reverse-transcription of mRNA transcript followed by integration into genomic DNA in germ-line cells [20,21]. Processed pseudogenes are typically characterized by the complete lack of introns, the presence of small flanking direct repeats, and a polyadenine tract near the 3′ end (provided that they have not decayed). Duplicated pseudogenes can be transcribed if the duplicated sequence includes the intact promoter sequence and other essential regulatory elements. Processed pseudogenes in general are not transcribed; however, in very rare cases, transcripts of some pseudogene have been reported though the functional relevance of these pseudogene transcripts remains unclear [22–24].

It is important to identify and characterize human pseudogenes since they can often interfere with studies on the functional genes [25]. They can also introduce errors in gene prediction and single-nucleotide polymorphism mapping efforts [16,26]. Knowing the exact nucleotide sequence and chromosomal localization of these pseudogenes is very important in correctly interpreting experimental results. Pseudogenes also provide a fossil record of gene and genome evolution.

In our previous surveys on the pseudogenes of cytoplasmic ribosomal proteins [16] and cytochrome *c* (Zhang and Gerstein, Gene, in press), most of the pseudogenes we discovered were new sequences that were overlooked by traditional experimental approaches. In the case of the cytoplasmic ribosomal proteins, over 2000 pseudogenes (mostly processed) were discovered, compared with 200 previously reported in GenBank. We conducted a similar survey for mitochondrial ribosomal proteins in the human genome; the detailed characterizations of these pseudogenes are described in this report.

Results

We followed the nomenclature approved by the HUGO Gene Nomenclature Committee (HGNC) (http://www.gene.ucl.ac.uk/nomenclature/genefamily/MRPs.html) [27]. As of June 2002, there are 79 MRP genes listed in this database, as well as in the Mammalian Mitochondrial Ribosomal Consortium (http://www.med.ufl.edu/biochem/tobrien/mmrc/index.htm). Although the goal of our study is to discover MRP pseudogenes, we have also recovered most of the functional human MRP genes in the genome. This indicates that our procedures are very comprehensive, e.g., there is little chance that some MRP pseudogenes were missed in our study.

The human MRP pseudogene population

Of the 79 human MRPs that we searched for pseudogenes, 41 have at least one pseudogenic sequence identified in the human genome. In total, 120 MRP pseudogene sequences were found. Although this is much fewer than the over 2000 pseudogenes for cytoplasmic RPs [16], it is still far more than the 3 human MRP pseudogenes previously deposited in GenBank. Table 1 lists the human MRP pseudogenes as named according to the functional MRP genes they originated from and also sequentially according to their locations on the chromosomes. The NCBI RefSeq ID [28] and cytogenic location for each functional MRP gene that has pseudogenes are also listed for comparison and reference. We divided MRP pseudogenes into three groups according to their origin and sequence completeness. (i) The first group, processed pseudogenes, consists of those sequences of obvious processed origin that also can be translated to a polypeptide longer than 70% of the functional MRP sequence. (ii) The second group, duplicated pseudogenes or nonprocessed pseudogenes, consists of sequences that have clear evidence of gene duplication such as existence of intron structure. (iii) The remaining sequences are short pseudogenic fragments (less than 70% of the corresponding MRP sequence) that are of uncertain origin. It is likely that these fragments are either ancient processed pseudogenes that have been truncated or duplicated individual exons. Of the total 120 MRP pseudogenes sequences, 84 were classified as processed pseudogenes (Table 2), 8 as duplicated pseudogenes, and 27 as pseudogenic fragments. Details on the classification are described under Materials and methods. In Table 1, these three groups are designated “Proc,” “Dupl,” and “Frag.”

Duplicated MRP pseudogenes

Except for *MRPS18CP3* and *MRPS31P1,* six of the eight duplicated MRP pseudogenes are on the same chromosome as the original functional gene, likely the result of unequal crossing over or tandem duplication. Five of the eight duplicated pseudogenes arose from the duplications of *MRPS31,* which has no processed pseudogenes; we will focus on this group of pseudogenes in detail. Four of the five *MRPS31* pseudogenes are on chromosome 13 (Fig. 1A). These *MRPS31* pseudogenes have retained the original exon structure of the functional gene, though they do not have all the exons (Fig. 1B). The two shortest pseudogenes, *MRPS31P1* and *MRPS31P3,* are merely single exons, whereas *MRPS31P4* and *MRPS31P5* contain six of the total seven exons, including exon 1. The protein sequences pre-

Table 1
The human MRP pseudogenes

Name	GenBank Accession No.	Cytogenic band	Chromosomal location[a]	Sequence divergence[b]	Deduced polypeptide[c]			Class[d]
					Interval	Fraction	ID	
MRPS5	NM_031902	2p11.2–q11.2	117.57M (−)					
MRPS5P1	AY135346	2q14.3	117.75M (−)	0.021 ± 0.012	311–356	11%	100%	Frag
MRPS5P2	AY135347	5Proc14.1	32.82M (+)	0.20 ± 0.028	316–430	27%	68%	Frag
MRPS5P3	AY135348	5q23.2	127.55M (−)	0.096 ± 0.009	1–430	100%	83%	Proc
MRPS5P4	AY135349	18q21.33	61.77M (+)	0.20 ± 0.028	316–430	27%	68%	Frag
MRPS6	NM_032476	21q21.3–q22.1	32.07M (+)					
MRPS6P1	AY135350	1p35.2	31.36M (−)	0.18 ± 0.026	1–125	100%	70%	Proc
MRPS6P2	AY135351	1Proc34.1	45.50M (−)	0.11 ± 0.018	1–125	100%	80%	Proc
MRPS6P3	AY135352	3q13.33	122.17M (+)	0.18 ± 0.026	1–125	100%	70%	Proc
MRPS6P4	AY135353	12q21.33	93.37M (−)	0.30 ± 0.039	1–108	86%	61%	Proc
MRPS7	NM_015971	17q23–q25	77.21M (+)					
MRPS7P1	AY135354	8p11.22	38.32M (−)	0.36 ± 0.03	67–242	73%	51%	Proc
MRPS7P2	AY135355	12p13.1	16.52M (+)	0.20 ± 0.019	20–242	92%	70%	Proc
MRPS10	NM_018141	6p21.1–p12.1	28.88M (−)					
MRPS10P1	AY135284	1q23.2	172.50M (+)	0.11 ± 0.03	109–145	18%	97%	Frag
MRPS10P2	AY135285	3p26.3	0.50M (+)	0.31 ± 0.034	61–201	70%	60%	Proc
MRPS10P3	AY135286	3p26.3	0.72M (+)	0.31 ± 0.034	61–201	70%	60%	Proc
MRPS10P4	AY135287	3p26.1	6.70M (−)	0.31 ± 0.034	61–201	70%	60%	Proc
MRPS10P5	AY135288	9p12	39.95M (−)	0.037 ± 0.01	1–201	100%	96%	Proc
MRPS11	NM_022839	15q25	85.05M (−)					
MRPS11P1	AY135289	20p11.23	20.79M (−)	0.17 ± 0.021	41–194	79%	73%	Proc
MRPS15	NM_031280	1p35–p34.1	32.42M (−)					
MRPS15P1	AY135290	15q33.33	69.91M (−)	0.23 ± 0.024	1–257	100%	57%	Proc
MRPS15P2	AY135291	19q13.32	60.97M (+)	0.33 ± 0.026	1–257	100%	60%	Proc
MRPS16	NM_016065	10q22.1	76.03M (−)					
MRPS16P1	AY135292	8q21.3	91.39M (+)	0.15 ± 0.021	1–137	100%	80%	Proc
MRPS16P2	AY135293	20q13.32	57.49M (+)	0.1 ± 0.017	1–137	100%	84%	Proc
MRPS16P3	AY135294	22q13.1	32.84M (+)	0.22 ± 0.027	1–137	100%	67%	Proc
MRPS17	NM_015969	7p11–q11.21	58.96M (−)					
MRPS17P1	AY135295	1p34.3	40.47M (−)	0.062 ± 0.013	1–130	100%	92%	Proc
MRPS17P2	AY135296	1p34.2	42.81M (−)	0.059 ± 0.013	1–130	100%	92%	Proc
MRPS17P3	AY135297	3p12.1	87.93M (−)	0.26 ± 0.032	1–130	100%	63%	Proc
MRPS17P4	AY135298	4p16.3	2.77M (−)	0.087 ± 0.016	1–130	100%	85%	Proc
MRPS17P5	AY135299	6q22.33	134.39M (+)	0.42 ± 0.063	48–130	64%	56%	Frag
MRPS17P6	AY135300	14q11.2	18.45M (+)	0.25 ± 0.041	62–130	53%	60%	Frag
MRPS17P7	AY135301	18q21.31	57.91M (−)	0.087 ± 0.016	1–130	100%	85%	Proc
MRPS17P8	AY135302	18q21.31	58.01M (−)	0.087 ± 0.016	1–130	100%	85%	Proc
MRPS17P9	AY135303	Xq24	111.64M (+)	0.076 ± 0.015	1–130	100%	87%	Proc
MRPS18A	NM_018135	6p21.3	30.34M (−)					
MRPS18AP1	AY135304	3p21.31	49.80M (+)	0.11 ± 0.015	1–196	100%	82%	Proc
MRPS18B	NM_014046	6p21.3	36.04M (+)					
MRPS18BP1	AY135305	1q41	224.06M (−)	0.14 ± 0.03	197–258	24%	81%	Frag
MRPS18BP2	AY135306	2q22.1	132.86M (+)	0.16 ± 0.016	1–258	100%	74%	Proc
MRPS18C	NM_016067	4q21.23	83.51M (+)					
MRPS18CP1	AY135307	3q26.1	170.51M (+)	0.067 ± 0.026	1–33	23%	100%	Frag
MRPS18CP2	AY135308	8p23.1	6.84M (−)	0.029 ± 0.01	1–142	100%	63%	Proc
MRPS18CP3	AY135309	8p21.3	18.74M (+)	1.5 ± 0.36	95–142	34%	82%	Dupl
MRPS18CP4	AY135310	12p13.31	8.49M (−)	N/A	123–142	14%	85%	Frag
MRPS18CP5	AY135311	15q11.2	22.14M (−)	0.08 ± 0.014	1–142	100%	83%	Proc
MRPS18CP6	AY135312	22q13.31	41.61M (−)	1.36 ± 0.25	62–139	57%	87%	Frag
MRPS21	NM_018997	1q21.2	151.90M (+)					
MRPS21P1	AY135313	1p34.3	36.59M (+)	0.095 ± 0.02	1–87	100%	80%	Proc
MRPS21P2	AY135314	1q22	159.92M (+)	0.11 ± 0.022	1–87	100%	80%	Proc
MRPS21P3	AY135315	1q31.2	203.45M (+)	0.06 ± 0.016	1–87	100%	85%	Proc
MRPS21P4	AY135316	9p13.1	38.53M (+)	0.20 ± 0.042	44–87	51%	68%	Frag
MRPS21P5	AY135317	10p12.1	30.34M (+)	0.17 ± 0.029	1–87	100%	66%	Proc
MRPS21P6	AY135318	10q23.1	88.05M (+)	0.098 ± 0.021	1–87	100%	82%	Proc
MRPS21P7	AY135319	16q12.1	48.60M (+)	0.25 ± 0.037	1–87	100%	62%	Proc
MRPS21P8	AY135320	16q12.1	48.66M (+)	0.25 ± 0.038	1–87	100%	59%	Proc
MRPS21P9	AY135321	17q22	50.82M (−)	0.37 ± 0.048	1–87	100%	63%	Proc

Table 1 (*continued*)

Name	GenBank Accession No.	Cytogenic band	Chromosomal location[a]	Sequence divergence[b]	Deduced polypeptide[c]			Class[d]
					Interval	Fraction	ID	
MRPS22	NM_020191	3q23						
MRPS22P1	AY135322	Xq21.31	80.43M (−)	0.42 ± 0.040	75–246	48%	39%	Frag
MRPS23	NM_016070	17q22–q23	58.33M (−)					
MRPS23P1	AY135323	7p13	45.90M (+)	0.22 ± 0.024	1–190	100%	69%	Proc
MRPS24	NM_032014	7p14	44.89M (−)					
MRPS24P1	AY135324	11p15.4	7.63M (−)	0.15 ± 0.019	1–167	100%	70%	Proc
MRPS25	NM_022497	3p25						
MRPS25P1	AY135325	4q21.23	80.52M (+)	0.16 ± 0.02	1–173	100%	71%	Proc
MRPS29	NM_004632	1q21.3	157.65(+)					
MRPS29P1	AY135326	1q21.3	157.83 (+)	1.032 ± 0.150	124–199	19%	88%	Dupl
MRPS29P2	AY135327	2q31.2	172.44M (+)	0.059 ± 0.011	227–398	43%	88%	Frag
MRPS31	NM_005830	13q13.3	39.91M (−)					
MRPS31P1	AY135328	3p21.33	42.48M (−)	1.10 ± 0.15	330–395	17%	79%	Dupl
MRPS31P2	AY135329	13q12.11	17.31M (−)	0.075 ± 0.015	200–320	31%	85%	Dupl
MRPS31P3	AY135330	13q12.11	19.12M (+)	0.13 ± 0.031	272–320	12%	94%	Dupl
MRPS31P4	AY135331	13q14.11	44.96M (+)	0.13 ± 0.013	1–320	81%	74%	Dupl
MRPS31P5	AY135332	13q14.2	52.97M (+)	0.12 ± 0.012	1–320	80%	79%	Dupl
MRPS33	NM_016071, NM_053035	7q32–34	144.27M (−)					
MRPS33P1	AY135333	1q21.3	156.34M (+)	0.25 ± 0.045	32–81	47%	76%	Frag
MRPS33P2	AY135334	4p14	40.68M (−)	0.13 ± 0.022	1–106	100%	78%	Proc
MRPS33P3	AY135335	4q26	117.37M (+)	0.12 ± 0.021	1–106	100%	79%	Proc
MRPS33P4	AY135336	20q13.13	50.93M (+)	0.28 ± 0.0038	19–106	83%	64%	Proc
MRPS35	NM_014018	12q21.1–q21.2	29.07M (+)					
MRPS35P1	AY135337	3p25.3	7.74M (+)	0.25 ± 0.053	145–180	16%	61%	Frag
MRPS35P2	AY135338	5q21.3	111.09M (−)	1.4 ± 0.32	171–227	25%	74%	Frag
MRPS35P3	AY135339	10q23.1	87.00M (+)	2.2 ± 0.52	30–227	87%	53%	Proc
MRPS36	NM_033281	5q12.1	103.32M (+)					
MRPS36P1	AY135340	3p25.3	9.06M (−)	0.03 ± 0.01	1–103	100%	95%	Proc
MRPS36P2	AY135341	4q35.1	186.51M (+)	0.23 ± 0.052	72–103	31%	78%	Frag
MRPS36P3	AY135342	8q24.13	121.05M (+)	0.13 ± 0.034	59–97	38%	85%	Frag
MRPS36P4	AY135343	11q23.2	116.62M (+)	0.13 ± 0.022	1–103	100%	79%	Proc
MRPS36P5	AY135344	12q12	44.25M (−)	0.20 ± 0.03	1–103	100%	77%	Proc
MRPS36P6	AY135345	20p12.1	13.34M (−)	0.31 ± 0.045	25–93	67%	70%	Frag
MRPL2	NM_015950	6p21.3	29.72 (−)					
MRPL2P1	AY135252	12q21.33	93.67M (+)	0.11 ± 0.012	1–305	100%	79%	Proc
MRPL3	NM_007208	3q21–23	135.77M (+)					
MRPL3P1	AY135256	13q12.11	17.06M (−)	0.045 ± 0.007	1–348	100%	91%	Proc
MRPL9	NM_031420	1q21	153.41M (−)					
MRPL9P1	AY135283	8q21.11	76.16M (−)	0.14 ± 0.016	60–267	78%	79%	Proc
MRPL11	NM_016050	11q13.3	71.61M (−)					
MRPL11P1	AY135246	2p16.3	50.04M (+)	0 ± 0	159–192	18%	100%	Frag
MRPL11P2	AY135247	5q31.3	145.64M (−)	0.063 ± 0.012	1–192	100%	64%	Proc
MRPL11P3	AY135248	12q21.2	82.54M (−)	0.20 ± 0.048	161–192	17%	84%	Frag
MRPL14	NM_032111	6p21.1	47.58M (+)					
MRPL14P1	AY135249	17p13.3	1.11M (+)	0.27 ± 0.046	68–130	43%	56%	Frag
MRPL15	NM_014175	8q11.2-q13	54.48M (+)					
MRPL15P1	AY135250	15q26.1	87.58M (−)	0.14 ± 0.014	19–296	94%	79%	Proc
MRPL15P2	AY135251	15q26.1	87.76M (−)	0.14 ± 0.014	19–296	94%	79%	Proc
MRPL20	NM_017971	1p36.3–36.2	8.36M (−)					
MRPL20P1	AY135253	21q22.2	34.94M (+)	0.16 ± 0.021	1–149	100%	72%	Proc
MRPL22	NM_014180	5q33.1–33.3	155.76M (−)					
MRPL22P1	AY135254	4q12	57.53M (−)	0.25 ± 0.023	1–228	100%	70%	Proc
MRPL22P2	AY135255	5q33.1	155.88M (−)	0.029 ± 0.02	66–88	10%	96%	Frag
MRPL30	NM_016503	2q11.2	93.34M (+)					
MRPL30P1	AY135257	6p12.1	60.82M (+)	0.048 ± 0.012	12–141	86%	89%	Proc
MRPL30P2	AY135258	12p11.22	33.23M (+)	0.19 ± 0.026	13–135	81%	76%	Proc
MRPL30P3	AY135259	12p11.22	33.44M (−)	0.20 ± 0.026	13–135	81%	76%	Proc
MRPL32	NM_031903	7p14	43.87M (+)					
MRPL32P1	AY135260	Xp11.23	43.67M (−)	0.16 ± 0.027	104–188	45%	74%	Frag

(*continued on next page*)

Table 1 (*continued*)

Name	GenBank Accession No.	Cytogenic band	Chromosomal location[a]	Sequence divergence[b]	Deduced polypeptide[c]			Class[d]
					Interval	Fraction	ID	
MRPL35	NM_016622	2p11.2	85.25M (+)					
MRPL35P1	AY135261	6p23	14.55M (+)	0.096 ± 0.017	1–122	100%	83%	Proc
MRPL35P2	AY135262	10q21.3	64.69M (−)	0.072 ± 0.015	1–122	100%	88%	Proc
MRPL35P3	AY135263	10q22.2	77.61M (+)	0.069 ± 0.014	1–122	100%	88%	Proc
MRPL35P4	AY135264	Xp22.31	10.25M (−)	0.12 ± 0.019	1–122	100%	79%	Proc
MRPL36	NM_032479	5p15.3	25.10M (+)					
MRPL36P1	AY135265	2p13.2	68.66M (−)	0.029 ± 0.01	1–103	100%	95%	Proc
MRPL42	NM_014050	12q22	104.81M (+)					
MRPL42P1	AY135266	4q27	119.61M (−)	0.16 ± 0.022	1–142	100%	73%	Proc
MRPL42P2	AY135267	6p22.3	16.83M (+)	0.18 ± 0.023	1–142	100%	71%	Proc
MRPL42P3	AY135268	6q24.2	151.16M (+)	0.17 ± 0.025	33–142	77%	75%	Proc
MRPL42P4	AY135269	7p12.1	52.97M (−)	0.28 ± 0.034	24–142	84%	64%	Proc
MRPL42P5	AY135270	15q13.3	35.66M (−)	0.15 ± 0.021	1–142	100%	78%	Proc
MRPL45	NM_032351	17q21.31	39.99M (+)					
MRPL45P1	AY135271	2p11.2	86.39M (−)	0.10 ± 0.013	1–306	100%	82%	Proc
MRPL45P2	AY135272	17q21.33	46.50M (−)	0.002 ± 0.002	1–153	50%	95%	Dupl.
MRPL48	NM_016055	11q13.2	74.70M (+)					
MRPL48P1	AY135273	6p24.1	10.94M (−)	0.066 ± 0.011	1–212	100%	89%	Proc
MRPL49	NM_004927	11q13	65.64M (−)					
MRPL49P1	AY135274	5q12.1	64.51M (+)	0.27 ± 0.028	1–166	100%	69%	Proc
MRPL49P2	AY135275	8p22	16.77M (+)	0.30 ± 0.032	1–166	100%	61%	Proc
MRPL50	NM_019051	9q31.1	93.10M (−)					
MRPL50P1	AY135276	2p22.3	33.92M (−)	0.25 ± 0.03	25–158	85%	60%	Proc
MRPL50P2	AY135277	2q34	205.04M (−)	0.069 ± 0.013	1–158	100%	86%	Proc
MRPL50P3	AY135278	5p12	52.43M (+)	0.10 ± 0.016	1–158	100%	82%	Proc
MRPL50P4	AY135279	10q23.1	86.05M (−)	0.088 ± 0.015	1–158	100%	82%	Proc
MRPL51	NM_016497	12p13.3–p13.1	6.55M (−)					
MRPL51P1	AY135280	4p15.2	30.59M (+)	0.38 ± 0.05	42–128	68%	54%	Frag
MRPL51P2	AY135281	21q22.3	41.08M (−)	0.11 ± 0.019	1–128	100%	79%	Proc
MRPL53	NM_053050	2p12	73.30M (−)					
MRPL53P1	AY135282	1p13.2	119.64M (−)	0.26 ± 0.035	1–112	100%	67%	Proc
MRP63	NM_024026	13q12.11	19.03M (+)					
MRP63P1	AY135236	1p13.1	122.10M (−)	0.40 ± 0.053	1–102	100%	46%	Proc
MRP63P2	AY135237	1q42.13	235.10M (−)	0.37 ± 0.055	43–102	59%	55%	Frag
MRP63P3	AY135238	3p21.31	48.56M (−)	0.23 ± 0.041	44–102	58%	71%	Frag
MRP63P4	AY135239	3p21.31	50.32M (+)	0.36 ± 0.049	11–102	90%	46%	Proc
MRP63P5	AY135240	4p16.3	2.37M (−)	0.36 ± 0.049	11–102	90%	41%	Proc
MRP63P6	AY135241	5q34	168.31M (+)	0.33 ± 0.041	1–102	100%	61%	Proc
MRP63P7	AY135242	8q22.2	97.82M (−)	0.30 ± 0.042	1–102	100%	55%	Proc
MRP63P8	AY135243	14q13.2	32.63M (−)	0.34 ± 0.048	1–102	100%	50%	Proc
MRP63P9	AY135244	14q22.1	49.16M (+)	0.28 ± 0.047	44–102	58%	58%	Frag
MRP63P10	AY135245	Yp11.2	9.91M (+)	0.40 ± 0.055	11–102	90%	41%	Proc

[a] Chromosomal coordinate of the gene/pseudogene and chromosomal strand.
[b] Nucleotide sequence divergence from the functional MRP gene.
[c] Interval, the residue range in the MRP protein sequence to which the pseudogene matches. Fraction, percentage of the MRP sequence to which the pseudogene matches. ID, amino acid sequence identity between the functional MRP and the predicted pseudogene sequence.
[d] Pseudogene class. Proc, processed; Frag, fragment; Dupl, duplicated.

dicted from the five pseudogenes range in length from 17 to 80% of the functional protein, and all have high sequence identity (>74%) to the functional *MRPS31* protein (Table 1).

We are interested in tracing the evolutionary path of these duplicated pseudogenes. Fig. 1C shows a phylogenetic tree constructed by applying the neighbor-joining (NJ) method [29,30] to the CDS of the pseudogenes and human functional *MRPS31* gene. Also shown in the graph are the estimated ages of the pseudogenes in millions of years, calculated following Kimura's two-parameter model [31] using a mutation rate of 1.5×10^{-9} per site per year. *MRPS31P1* is very ancient and incomplete (its sequence corresponding to residues 330–395 of the intact gene); it is likely the model we used overestimated the sequence divergence and the actual age of this pseudogene.

Potential transcription of two duplicated pseudogenes

Among the *MRPS31* pseudogenes, *MRPS31P4* and *MRPS31P5* are the most intriguing. Alignment of their sequences with the mRNA transcript of the functional gene indicates they are disabled at the translation level since their

Table 2
Number of processed pseudogenes among MRP genes

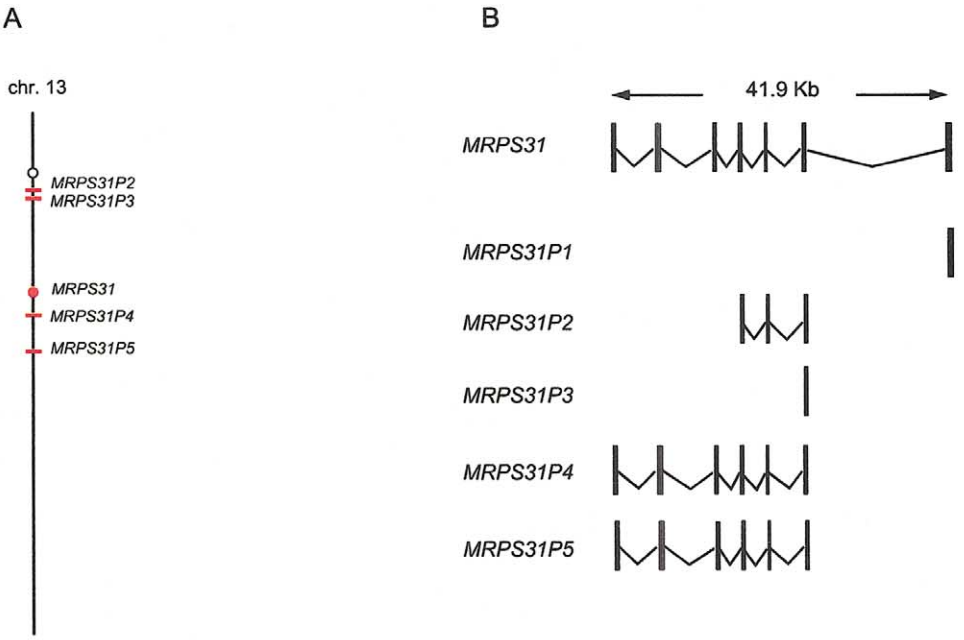
MRP gene	Number of processed pseudogenes
MRPS21	8
MRP63, MRPS17	7
MRPL42	5
MRPL35, MRPL50, MRPS6, MRPS10	4
MRPL30, MRPS16, MRPS33, MRPS36	3
MRPL15, MRPL49, MRPS15, MRPS18C, MRPS7	2
MRPL2, MRPL3, MRPL9, MRPL11, MRPL20, MRPL22, MRPL36, MRPL45, MRPL48, MRPL51, MRPL53, MRPS11, MRPS18A, MRPS18B, MRPS23, MRPS24, MRPS25, MRPS35, MRPS5	1
MRPL1, MRPL4, MRPL12, MRPL13, MRPL14, MRPL16, MRPL17, MRPL18, MRPL19, MRPL23, MRPL24, MRPL27, MRPL28, MRPL32, MRPL33, MRPL34, MRPL37, MRPL39, MRPL43, MRPL44, MRPL46, MRPL47, MRPL55, MRPS2, MRPS9, MRPS12, MRPS14, MRPS22, MRPS26, MRPS27, MRPS28, MRPS29, MRPS30, MRPS31, MRPS34, LACTB	0
Total	84

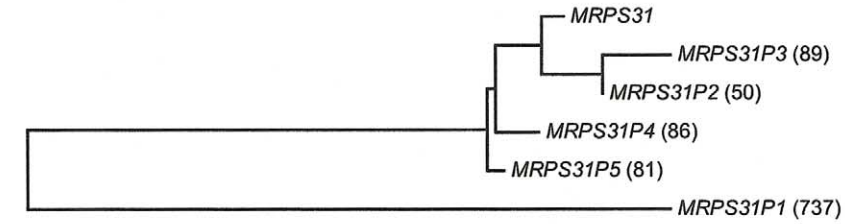
translation initiation codons have mutated from ATG to GTG (Fig. 1D). It was interesting to see whether these two pseudogenes are transcribed in human cells since it is possible they could have retained the intact promoter sequence upstream of the first exon. After conducting a BLAST search [32] on the GenBank human EST database, we found unambiguous matches for each of the pseudogenes. Fig. 1D shows the alignment of the mRNA sequence of the functional *MRPS31* gene, the two pseudogenes, and the top matches from the EST search; only the sequence of the first exon and some of the 5′ upstream sequence are shown. It is clear from the alignment that the EST BM465470 and the pseudogene *MRPS31P4* share the same nucleotide substitutions in comparison with the functional *MRPS31;* so do the EST BG504721 and the pseudogene *MRPS31P5*. Furthermore, both the two ESTs and the two pseudogenes have a mutated translation initiation codon GTG instead of ATG, marked by a downward arrow in Fig. 1D. All this evidence suggests that the two ESTs are the transcripts of the pseudogenes *MRPS31P4* and *MRPS31P5,* respectively. To exclude the possibility that these EST hits are transcripts from some other human genes that happen to share an identical sequence motif with *MRPS31* and the pseudogenes, we did a nucleotide–nucleotide BLAST search on human genomic DNA using the sequence of the first exon of *MRPS31*. The BLAST search did not find any significant matches in the genome except for three loci on chromosome 13 that correspond to the *MRPS31* and the pseudogenes; thus it is unambiguous that the two pseudogenes gave rise to these two ESTs. The sequence alignment shown in Fig. 1D also raises the possibility that exon 1 of the functional *MRPS31* gene is actually longer than is predicted, as indicated by the extra nucleotides at the 5′ end of EST BG504721 and the two pseudogenes. No EST matches were found for other duplicated pseudogenes.

Processed MRP pseudogenes

Similar to cytoplasmic RPs, the majority of the human MRP pseudogenes are of processed origin. Table 2 lists the number of processed pseudogenes for each MRP functional gene, sorted in ascending order. About half of the mitochondrial ribosomal proteins (36 of 79) have at least one processed pseudogene identified in the genome. The CDS regions of these pseudogenes are in general well preserved, as 63 pseudogenes or 75% of the total processed pseudogenes can be translated conceptually to a polypeptide longer than 95% of the full-length functional protein (Fig. 2A). Most of these processed pseudogenes can be extended in both directions beyond the CDS region to match some of the mRNA untranslated (UTR) sequences. A polyadenine tail can be found unambiguously for a third (28 of 84) of the total processed pseudogenes. Such high level of sequence preservation was also previously observed for cytoplasmic RP pseudogenes [16]. Fig. 2B shows the distribution of the amino acid sequence identity between the MRP processed pseudogene and the corresponding functional gene; on average the sequence identity is 73%. We also plotted the age distribution of the pseudogenes by calculating the nucleotide sequence divergence between the pseudogenes and the functional MRP genes (Table 1); the distribution pattern closely resembles that of cytoplasmic RP pseudogenes and *Alu* repeats [16,33].

Some of the processed pseudogenes also contain inserted retrotransposons. An *Alu* repeat is found in the middle of *MRPS15P2*, *MRPS23P1*, and *MRPL15P2*, and a MER1B repeat is found in the middle of *MRPS7P2*. Pseudogene *MRPS10P5* on chromosome 9 consists of two immediately adjacent fragments that correspond to residues 3–77 and 85–201, respectively, but on the opposite strands. It is likely this pseudogene has undergone "5′ inversion," which is common for LINE1-mediated retrotransposition [34–36]. The sequences of these two fragments were merged together in the phylogenetic analysis. The pseudogene for *MRPL11* on chromosome 5, *MRPL11P2,* is a full-length processed pseudogene with a 19-bp-long polyadenine tail. The CDS region of the functional *MRPL11* gene consists of five exons; interestingly, exon 4 is completely missing in the pseudogene sequence. Fig. 3 shows the alignment of the CDS of *MRPL11P2* and the functional *MRPL11* genes of mouse and human. The intron positions are derived from the tBLASTn results and confirmed by the Ensembl Mouse Genome annotation (Ensembl gene ID ENSMUSG-00000024902), as it is apparent that the exon structure or intron positions are conserved between mouse and human for this particular gene. It is likely that *MRPL11P2* originated from an alternatively spliced *MRPL11* mRNA transcript that lacked exon 4. Similar alternative splicing vari-





D

```
                                                                                                  G GTTCATTTCG TGTCTCGGCG ATGTTTCCTA
MRPS31P4  CGCGGGGGAT GTGTTTGGCG TGGGG-CGCA CTGTT-CAGT TGCTCTCCTG GATGGCTTTA TGTCTCCGCG GTGTTTCCTA
BM465470                                                                                TCTCCGCG GTGTTTCCTA
MRPS31P5  CGCGGGGGAT GTGTTTGGCA TGGGGACGCA CTGTTACAGT TGCGCTCCTG GTTGGCTTTG TGTTTCCGCG GTGTTGG-TA
BG504721                GGGCA TGGGGACGCA CTGTTACAGT TGCGCTCCTG GTTGGCTTTG TGTTTCCGCG GTGTTGG-TA

MRPS31    GAGTCTCGAC GTTCCTACCT CTTCGCCCCC TTTCCCGCCA CCCTTTGTCC TCTGGAAGCC CGGAGACATC AGCGGCTGCG
MRPS31P4  GAGTCTCGGC GTTTCTACCT CTTCGCACCC TTTCCTGCCA CCCTTTGTCC CCTGGAAGCC CGGAGACATC AGCGGCTGCA
BM465470  GAGTCTCGGC GTTTCTACCT CTTCGCACCC TTTCCTGCCA CCCTTTGTCC CCTGGAAGCC CGGAGACATC AGCGGCTGCA
MRPS31P5  GAGTCTCGGT GTTTCTACCT CTTAGCACCC TTTCCTGCCA CCCTTTGTCC TGTGGAAGCC CGGAGACATC AGCGGCTGCA
BG504721  GAGTCTCGGT GTTTCTACCT CTTAGCACCC TTTCCTGCCA CCCTTTGTCC TGTGGAAGCC CGGAGACATC AGCGGCTGCA

MRPS31    ATTATGCTAC TCACTGTTCG GCACGGAACA GTCAGG
MRPS31P4  ATTTTGCTAC TCGCTGCTCG GCGTGGAACT GTCAGG
BM465470  ATTTTGCTAC TCGCTGCTCG GCGTGGAACT GTCAGG
MRPS31P5  ATTTTGCTAC TCGCTGCTCG GCATGGAACG GTCAGG
BG504721  ATTTTGCTAC TCGCTGCTCG GCATGGAACG GTCAGG
```

Fig. 1. Duplicated pseudogenes of *MRPS31*. (A) Schematic localization of the *MRPS31* functional gene and pseudogenes on chromosome 13. Open circle, filled red sphere, and red horizontal bars indicate centromere, functional *MRPS31* gene, and the four duplicated pseudogenes. (B) Exon structure of the functional *MRSP31* gene and pseudogenes. The exon structure of the functional gene was obtained from the Ensembl Web site [11] (Ensembl ID: ENSG00000102738). (C) Phylogenetic tree of duplicated pseudogenes of *MRPS31*. Numbers in the brackets are the estimated ages of the pseudogenes in millions of years. (D) Multiple sequence alignment of the *MRPS31* mRNA transcript (*MRPS31*), the two pseudogenes (*MRPS31P4*, *MRPS31P5*), and the closest matches for the pseudogene sequences from NCBI human EST database (BG504721 and BM465470). The downward arrow marks the translation initiation codon ATG (GTG in the pseudogenes and EST sequences).

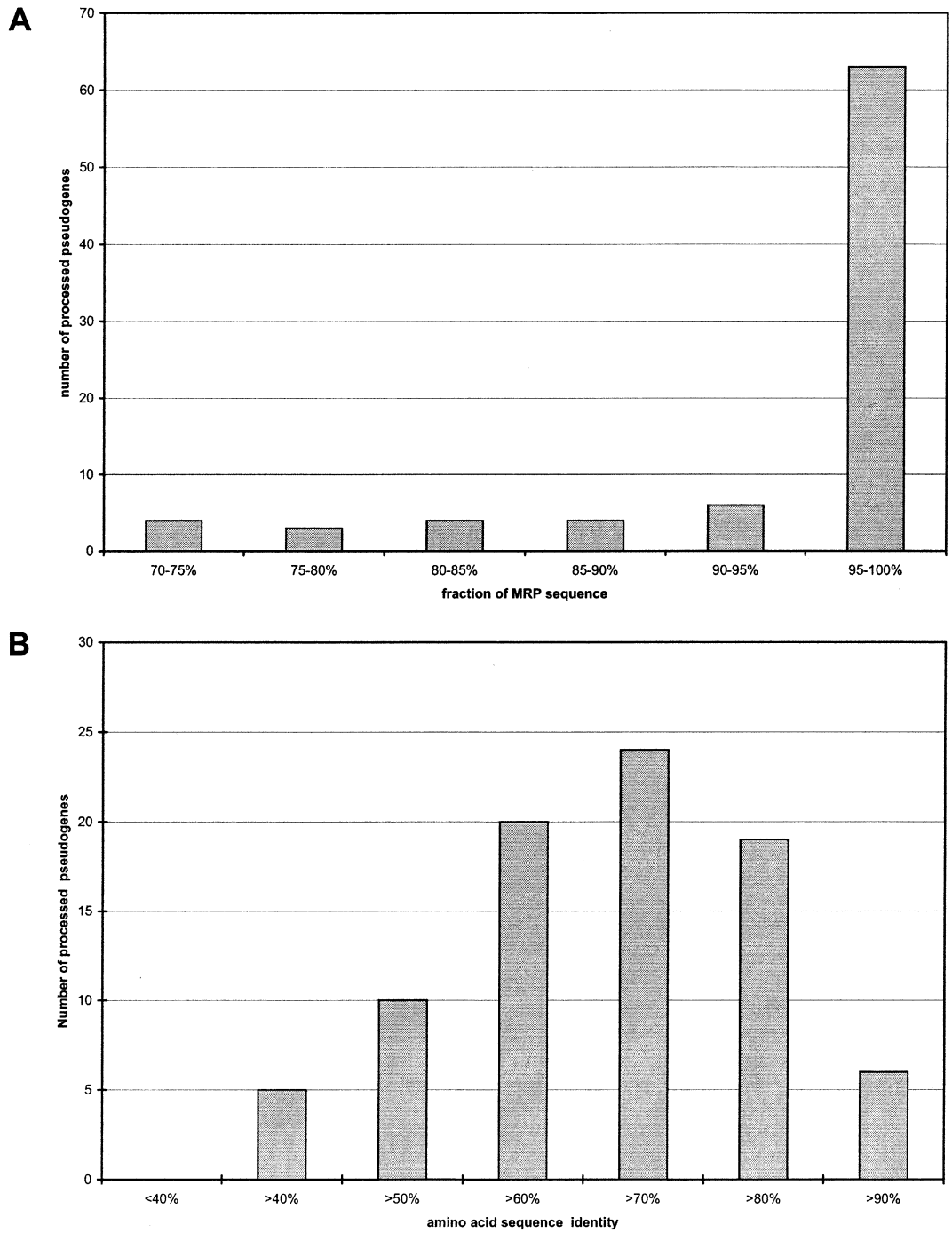


Fig. 2. Statistics of human MRP processed pseudogenes. (A) Distribution of the completeness in the CDS region for processed pseudogenes, i.e., the ratio between the lengths of translated pseudogene and the corresponding functional MRP gene. (B) Distribution of the amino acid sequence identity between processed pseudogenes and the functional MRP sequences.

ation has been recently reported among the human endogenous retrovirus sequences [37].

Number of processed pseudogenes is correlated with gene length

It is obvious that the numbers of processed pseudogenes among MRP genes are highly uneven (Table 2); some MRP genes such as *MRPS11* and *MRPS21* have seven or eight copies while 36 other MRP genes have none in the genome. We are interested in studying the mechanism behind such skewed distribution. Goncalves et al. have proposed that those genes that have produced processed pseudogenes tend to be widely expressed, short in sequence, and low in GC content [38]. For the MRP pseudogenes, we have indeed observed a correlation be-

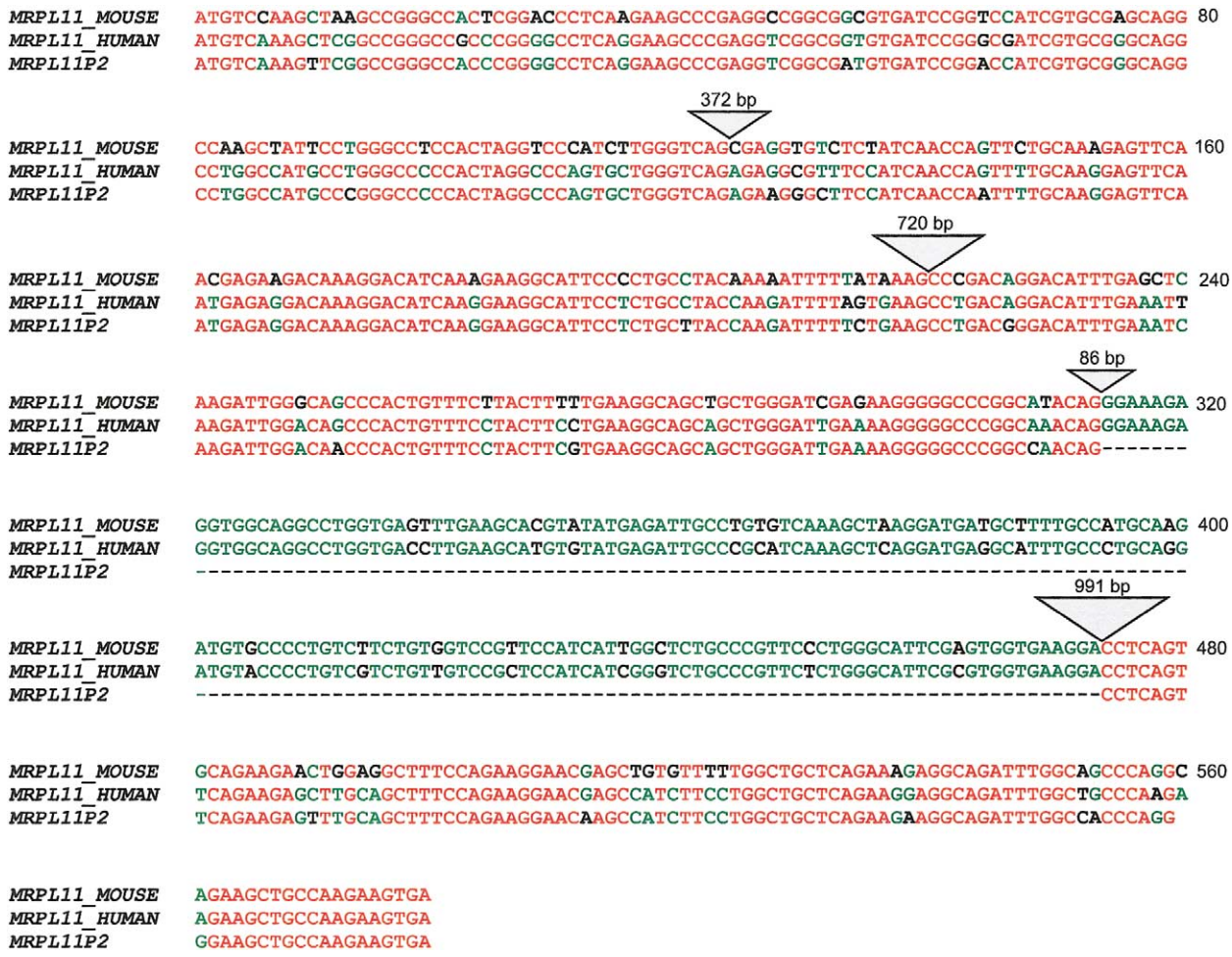


Fig. 3. Alignment of the protein coding sequence of the pseudogene *MRPL11P2* and functional genes from mouse and human. The size and location of the introns are indicated above the genes by triangles, the surface of which is proportional to its size. As can be seen, exon 4 is completely missing in *MRPL11P2*.

tween the number of processed pseudogenes and the gene CDS length ($R = -0.40$; $p < 0.001$) (Fig. 4A). The correlation dropped to -0.25 ($p < 0.037$) if we used the length of the mRNA transcript instead of the CDS sequence. This is probably due to the fact that some of the MRP mRNA sequences are truncated beyond the CDS region. Three functional genes, *MRPS17*, *MRPS21*, and *MRP63*, have the highest numbers of processed pseudogenes, at 7 or 8; they are also among the shortest MRP genes with CDS length of 264, 393, and 309 bp, respectively. Note that such negative correlation simply reflects the fact that retrotransposition for short genes is more efficient and more likely to succeed than for longer genes; it does not necessarily indicate that the retrotransposition machinery has higher binding affinity for the short mRNA transcripts. Long genes tend to have long UTRs in their mRNAs and they are more likely to have processed pseudogenes terminated within the 3′ UTR and thus undetectable by our approaches. We also observed a weaker correlation between the number of processed pseudogenes and the gene CDS GC content ($R = -0.19$, $p < 0.1$) (Fig. 4B), i.e., the MRP genes that gave rise to processed pseudogenes tended to have lower GC content than the genes that had no processed pseudogenes.

As an example, Fig. 5 shows the phylogenetic trees of the pseudogenes of *MRPS17* and *MRPS21*, two MRP genes that have the largest numbers of processed and total pseudogenes. The topologies of the trees are in quite good agreement with the estimated ages, as on average the older pseudogenes were placed near the bottom of the tree. As mentioned in the previous section, the ages of the oldest pseudogenes, *MRPS17P5* and *MRPS21P9*, are probably overestimated due to simplification of the model.

Online database

The human MRP pseudogene sequences have been deposited with GenBank under Accession Nos. AY135236–AY135355. The more detailed data and results, such as sequence alignments and phylogenetic analysis, can also be accessed online at http://www.pseudogene.org/or http://bioinfo.mbb.yale.edu/genome/pseudogene/.

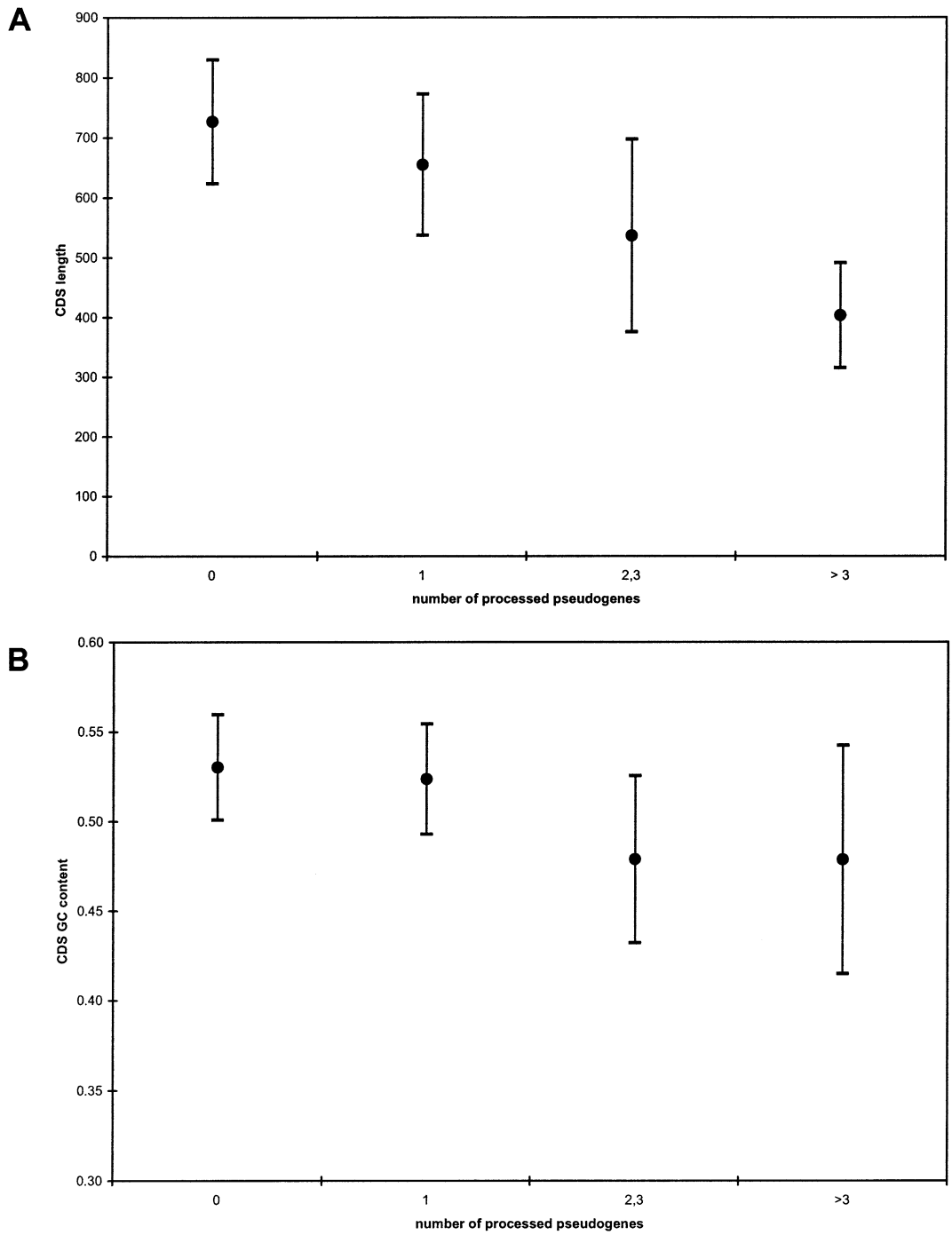


Fig. 4. The correlation between the number of processed pseudogenes and the average CDS length (A) and CDS GC content (B) of corresponding functional MRP genes. The MRPs are binned together according to their number of processed pseudogenes. The error bars represent 95% confidence interval.

Discussion

Potential interference by pseudogenes of functional genes

Except for the three pseudogenes on chromosome 20, most of the MRP pseudogenes we described here were reported for the first time. It is likely that the decay in their sequences has caused them to be overlooked in the systematic MRP gene-mapping project [10]. It is also likely that the particular experimental condition was so optimal and discriminatory that none of the pseudogenes were amplified in the PCR. Nevertheless, the existence of such large numbers of processed MRP pseudogenes in the human genome and the fact that they were discovered only by computational approaches are alarming, as it suggests that other human genes could also have many pseudogene sequences

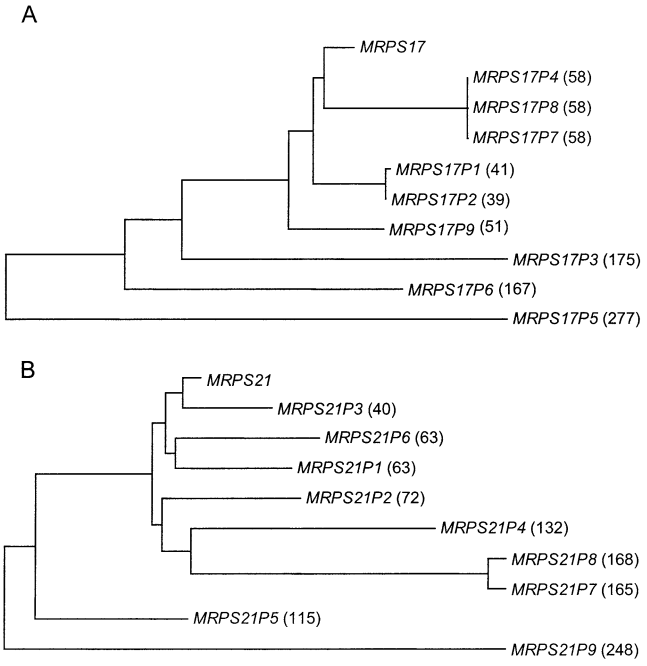


Fig. 5. Phylogenetic trees of pseudogenes of *MRPS17* (A) and *MRPS21* (B). Numbers in parentheses are the estimated ages of the pseudogenes in millions of years.

in the genome that have yet to be discovered. These pseudogenes could potentially interfere with the genomic mapping of the functional genes and hybridization-based experiments in general. In recent years, reverse transcriptase–polymerase chain reaction assays targeting genes with tissue-specific expression patterns have been extensively used in cancer diagnostics. The concept behind such approaches is that if expression of a target gene is detected in a cell that under normal conditions does not express the gene, then it is likely that the cell has become cancerous. However, it has been reported that, at least in one assay that targets gene cytokeratin 19, the existence of a pseudogene may have interfered with almost all published assay results [25]. The discovery of two transcribed MRP pseudogenes, *MRPS31P4* and *MRPS31P5*, further raises concerns about contamination of duplicated pseudogenes in the human EST database. Without definite identification of these sequences as transcribed pseudogenes, they could have been mistaken as functional gene transcripts with sequencing errors. The possibility that *MRPL11P12* is derived from an alternatively spliced mRNA transcript could offer insight into the evolution of the functional MRP gene and the biogenesis of mitochondrial ribosomes.

Properties of genes with processed pseudogenes

It has been demonstrated that the protein machinery encoded by LINE1 has a *cis* preference, i.e., it has higher affinity to wild-type LINE1 mRNA transcripts than to mutant LINE1 transcripts [20,39]. Among non-LINE1 human genes, it is obvious the distribution of number of processed pseudogenes was highly uneven, as has been observed for MRP genes and cytoplasmic RP genes [16]. In the case of RP genes, the number of pseudogenes was reverse-correlated with the gene CDS GC content, i.e., the relatively GC-poor RP genes have more processed pseudogenes than the GC-rich RP genes; however, no correlation between pseudogene numbers and gene length was observed. It may seem contradictory that, here for MRP genes, we found a reverse correlation between the gene CDS length and pseudogene numbers (Fig. 4A). However, cytoplasmic RP genes are a very unique gene family in many ways, as they are the most highly expressed genes in the cell. It is likely that the expression level of RP genes has reached such a saturated level in the cell that the reverse transcription process, which generates RP pseudogenes, has lost the sensitivity or preference toward the gene length.

The much higher expression level of the cytoplasmic RP genes than of the MRP genes can also explain the vast differences in the number of pseudogenes for the otherwise homologous gene families, as on average cytoplasmic RP genes have over 20 times more processed pseudogenes in the human genome than MRP genes. It is expected that the genes with higher expression level, e.g., with more mRNA transcripts in the cell, would have more chance to be taken by the LINE1 machinery and reverse-transcribed to give rise to pseudogenes [18,19]. In a previous study, Goncalves and colleagues [32] have proposed that those genes that have produced processed pseudogenes tend to be widely expressed, short in sequence, and low in GC content. However, the group of genes (total 249) that these authors studied did not necessarily have the same expression level, thus it is difficult to separate the effect of expression level from the effects of other gene properties such as sequence length and GC content. Mitochondrial ribosomal proteins provide a nice opportunity to solve this problem as described below. It is generally assumed that, at least in the yeast cell, the expression of mitochondrial ribosomal proteins is strictly regulated to avoid a huge waste of metabolic energy [2], thus the expression level of individual MRP genes must be in stoichiometry. Therefore, with each individual gene having a similar expression level, the MRP gene family provides a perfect model system to study what factors or what properties of a gene, barring the influences of expression level, may affect the number of processed pseudogenes it has in the genome. The results we obtained from mitochondrial ribosomal proteins confirmed what has been previously proposed [31].

Materials and methods

Details on the pseudogene discovery procedures have been described elsewhere [16]. A brief overview is given below. We used the human genome draft freeze of August 6, 2001, downloaded from the Ensembl Web site (http://www.ensembl.org). Subsequently, all the chromosomal coordinates were based on these sequences. The amino acid

and nucleotide sequences of mitochondrial ribosomal proteins are from the HGNC Web site (http://www.gene.ucl.ac.uk/nomenclature/genefamily/MRPs.html) [27]. Each human chromosome was split into smaller overlapping chunks of 5.1 million bp, and the tBLASTn program of the BLAST package 2.0 [32] was run on these sequences. The default SEG [40] low-complexity filter parameters (12 2.2 2.5) were used in the homology search. The significant homology matches (e value $< 10^{-4}$) were picked out and reduced for mutual overlap.

After the BLAST matches were sorted according to their starting positions on the chromosomes, they were examined and neighboring matches were merged together if they were part of the same pseudogene sequence. The merged matches were then extended on both sides to equal the length of MRP genes they match to, plus 30-bp buffers. For each extended match, the MRP amino acid sequence was realigned to the genomic DNA sequence following the Smith–Waterman algorithm [41] by using the program FASTA [42].

We then examined each pseudogene candidate for existence of exon structures. The majority of the functional MRP genes contain introns, so the pseudogenes for these genes were easily confirmed by their lack of introns. The genes *MRPS33*, *MRPL36*, and *MRP63* have no introns within the coding sequence; their pseudogenes were distinguished from the functional genes by comparing their chromosomal locations with that of the functional MRP genes obtained from HGNC [27] and the NCBI UniGene Web site (http://www.ncbi.nlm.nih.gov/UniGene/). The processed pseudogene *MRPL36P1* has no obvious disablement in the CDS region; we used two lines of evidence to establish that this is a processed pseudogene rather than a functional gene or a duplicated pseudogene. First, a polyadenine tail of 13 bp was found at the end of the 3′ UTR; second, a BLAST query against the NCBI dbEST database [43] found no matches for this pseudogene.

Multiple sequence alignments were performed using the programs ClustalW [44] and MultAlin [45]. The software MEGA2 [46] was used for all the phylogenetic analysis. Phylogenetic trees were constructed by applying the NJ method [29,30] to the CDS of the genes and pseudogenes. For each pseudogene, we also calculated the nucleotide sequence divergence from the functional MRP genes using Kimura's two-parameter model [31], which corrects for multiple hits and also takes into account different substitution rates for transition and transversion. The pseudogene ages were calculated using the formula $T = D/(k)$, where D is the divergence and k is the mutation rate per year per site. A mutation rate of 1.5×10^{-9} per site per year for pseudogenes was used [47].

Acknowledgments

M.G. acknowledges an NIH CEGS grant (P50HG02357-01) and the Keck Foundation for financial support. Z.Z. thanks Paul Harrison for critical comments and Duncan Milburn and Ted Johnson for computational help.

References

[1] D.E. Matthews, R.A. Hessler, N.D. Denslow, J.S. Edwards, T.W. O'Brien, Protein composition of the bovine mitochondrial ribosome, J. Biol. Chem. 257 (1982) 8788–8794.
[2] H.R. Graack, B. Wittmann-Liebold, Mitochondrial ribosomal proteins (MRPs) of yeast, Biochem. J. 329 (1998) 433–448.
[3] S. Goldschmidt-Reisin, M. Kitakawa, E. Herfurth, B. Wittmann-Liebold, L. Grohmann, H.R. Graack, Mammalian mitochondrial ribosomal proteins: N-terminal amino acid sequencing, characterization, and identification of corresponding gene sequences, J. Biol. Chem. 273 (1998) 34828–34836.
[4] T.W. O'Brien, Evolution of a protein-rich mitochondrial ribosome: implications for human genetic disease, Gene 286 (2002) 73–79.
[5] E. Cavdar Koc, W. Burkhart, K. Blackburn, A. Moseley, L.L. Spremulli, The small subunit of the mammalian mitochondrial ribosome: identification of the full complement of ribosomal proteins present, J. Biol. Chem. 276 (2001) 19363–19374.
[6] E.C. Koc, W. Burkhart, K. Blackburn, M.B. Moyer, D.M. Schlatzer, A. Moseley, et al., The large subunit of the mammalian mitochondrial ribosome: analysis of the complement of ribosomal proteins present, J. Biol. Chem. 276 (2001) 43958–43969.
[7] E.C. Koc, W. Burkhart, K. Blackburn, H. Koc, A. Moseley, L.L. Spremulli, Identification of four proteins from the small subunit of the mammalian mitochondrial ribosome using a proteomics approach, Protein Sci. 10 (2001) 471–481.
[8] T. Suzuki, M. Terasaki, C. Takemoto-Hori, T. Hanada, T. Ueda, A. Wada, et al., Proteomic analysis of the mammalian mitochondrial ribosome: identification of protein components in the 28 S small subunit, J. Biol. Chem. 276 (2001) 33181–33195.
[9] S.F. Pietromonaco, R.A. Hessler, T.W. O'Brien, Evolution of proteins in mammalian cytoplasmic and mitochondrial ribosomes, J. Mol. Evol. 24 (1986) 110–117.
[10] N. Kenmochi, T. Suzuki, T. Uechi, M. Magoori, M. Kuniba, S. Higa, et al., The human mitochondrial ribosomal protein genes: mapping of 54 genes to the chromosomes and implications for human disorders, Genomics 77 (2001) 65–70.
[11] T. Hubbard, D. Barker, E. Birney, G. Camero, Y. Chen, L. Clark, et al., The Ensembl genome database project, Nucleic Acids Res. 30 (2002) 38–41.
[12] N.D. Denslow, J.C. Anders, T.W. O'Brien, Bovine mitochondrial ribosomes possess a high affinity binding site for guanine nucleotides, J. Biol. Chem. 266 (1991) 9586–9590.
[13] J.L. Kissil, O. Cohen, T. Raveh, A. Kimchi, Structure–function analysis of an evolutionarily conserved protein, DAP3, which mediates TNF-alpha- and Fas-induced cell death, EMBO J. 18 (1999) 353–362.
[14] E. Cavdar Koc, A. Ranasinghe, W. Burkhart, K. Blackburn, H. Koc, A. Moseley, et al., A new face on apoptosis: death-associated protein 3 and PDCD9 are mitochondrial ribosomal proteins, FEBS Lett. 492 (2001) 166–170.
[15] I.G. Wool, Y.L. Chan, A. Gluck, Structure and evolution of mammalian ribosomal proteins, Biochem. Cell Biol. 73 (1995) 933–947.
[16] Z. Zhang, P. Harrison, M. Gerstein, Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome, Genome Res. 12 (2002) 1466–1482.
[17] A.J. Mighell, N.R. Smith, P.A. Robinson, A.F. Markham, Vertebrate pseudogenes, FEBS Lett. 468 (2000) 109–114.
[18] E.F. Vanin, Processed pseudogenes: characteristics and evolution, Annu. Rev. Genet. 19 (1985) 253–272.
[19] P.M. Harrison, H. Hegyi, S. Balasubramanian, N.M. Luscombe, P. Bertone, N. Echols, et al., Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22, Genome Res. 12 (2002) 272–280.

[20] C. Esnault, J. Maestre, T. Heidmann, Human LINE retrotransposons generate processed pseudogenes, Nat. Genet. 24 (2000) 363–367.
[21] H.H. Kazazian Jr., J.V. Moran, The impact of L1 retrotransposons on the human genome, Nat. Genet. 19 (1998) 19–24.
[22] G.H. Fujii, A.M. Morimoto, A.E. Berson, J.B. Bolen, Transcriptional analysis of the PTEN/MMAC1 pseudogene, psiPTEN, Oncogene 18 (1999) 1765–1769.
[23] J.R. McCarrey, M. Kumari, M.J. Aivaliotis, Z. Wang, P. Zhang, F. Marshall, et al., Analysis of the cDNA and encoded protein of the human testis-specific PGK-2 gene, Dev. Genet. 19 (1996) 321–332.
[24] M.A. Olsen, L.E. Schechter, Cloning, mRNA localization and evolutionary conservation of a human 5-HT7 receptor pseudogene, Gene 227 (1999) 63–69.
[25] P. Ruud, O. Fodstad, E. Hovig, Identification of a novel cytokeratin 19 pseudogene that may interfere with reverse transcriptase–polymerase chain reaction assays used to detect micrometastatic tumor cells, Int. J. Cancer 80 (1999) 119–125.
[26] P.C. Ng, S. Henikoff, Accounting for human polymorphisms predicted to affect protein function, Genome Res. 12 (2002) 436–446.
[27] H.M. Wain, M. Lush, F. Ducluzeau, S. Povey, Genew: the human gene nomenclature database, Nucleic Acids Res. 30 (2002) 169–171.
[28] K.D. Pruitt, D.R. Maglott, RefSeq and LocusLink: NCBI gene-centered resources, Nucleic Acids Res. 29 (2001) 137–140.
[29] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[30] M. Nei, S. Kumar, Molecular Evolution and Phylogenetics (2000, Oxford Univ. Press, New York.
[31] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, J. Mol. Evol. 16 (1980) 111–120.
[32] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[33] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[34] E.M. Ostertag, H.H. Kazazian Jr., Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition, Genome Res. 11 (2001) 2059–2065.
[35] H.H. Kazazian Jr., J.L. Goodier, LINE drive: retrotransposition and genome instability, Cell 110 (2002) 277–280.
[36] D.E. Symer, C. Connelly, S.T. Szak, E.M. Caputo, G.J. Cost, G. Parmigiani, et al., Human L1 retrotransposition is associated with genetic instability in vivo, Cell 110 (2002) 327–338.
[37] A. Pavlicek, J. Paces, D. Elleder, J. Hejnar, Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution, Genome Res. 12 (2002) 391–399.
[38] I. Goncalves, L. Duret, D. Mouchiroud, Nature and structure of human genes that generate retropseudogenes, Genome Res. 10 (2000) 672–678.
[39] W. Wei, N. Gilbert, S.L. Ooi, J.F. Lawler, E.M. Ostertag, H.H. Kazazian Jr., et al., Human L1 retrotransposition: Cis preference versus trans complementation, Mol. Cell. Biol. 21 (2001) 1429–1439.
[40] J.C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, Comput. Chem. 17 (1993) 149–163.
[41] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[42] W.R. Pearson, Comparison of DNA sequences with protein sequences, Genomics 46 (1997) 24–36.
[43] M.S. Boguski, T.M. Lowe, C.M. Tolstoshev, dbEST—Database for "expressed sequence tags", Nat. Genet. 4 (1993) 332–333.
[44] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[45] F. Corpet, Multiple sequence alignment with hierarchical clustering, Nucleic Acids Res. 16 (1988) 10881–10890.
[46] S. Kumar, K. Tamura, I.B. Jakobsen, M. Nei, MEGA2: molecular evolutionary genetics analysis software, Bioinformatics 17 (2001) 1244–1245.
[47] Li, W.-H. (1997). "Molecular Evolution".