Comparative analysis of processed pseudogenes in the mouse and human genomes

Zhaolei Zhang[1], Nick Carriero[2] and Mark Gerstein[1, 2]

(To appear in *Trends in Genetics*, Feb. 2004)

[1]Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA

[2]Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06520, USA

Corresponding author: Mark Gerstein

Tel: (203) 432 6105

Fax: (360) 838 7861

Email: Mark.Gerstein@yale.edu

Summary

Pseudogenes are important resources in evolutionary and comparative genomics. We have systematically identified ~5000 processed pseudogenes in the mouse genome (indexed at pseudogene.org/mouse). We estimate ~60% are lineage-specific, created after mouse and human diverged. In both mouse and human, similar types of genes give rise to many processed pseudogenes. These tend to be housekeeping genes, highly expressed in the germ line. Ribosomal-protein genes, in particular, form the largest sub-group of processed pseudogenes. The processed pseudogenes in the mouse occur with a distinctly different chromosomal distribution than LINEs or SINEs, preferentially in the GC-poor regions. Finally, the age distribution of mouse processed pseudogenes closely resembles that of LINEs, in contrast to human, where it closely follows Alus (SINEs).

Mammalian genomes contain many non-functional, gene-like sequences known as pseudogenes. A pseudogene has close sequence similarity to a functional gene but generally is not transcribed for lack of functional promoters or other regulatory elements [1]. The majority of the mammalian pseudogenes are processed pseudogenes (also known as retropseudogenes). They were inserted into the genome by the retrotransposition of the mRNAs of functional genes by the LINE1 elements [2-4]. These processed pseudogenes typically do not contain introns and sometimes still have a recognizable 3' poly-adenine tail (if has not decayed). They were released from selective pressure and thus many have accumulated diagnostic frame disruptions in their sequences such as frameshifts, stop codons or interspersed repeats.

Pseudogenes are considered as "molecular fossils" as they have proven to be important resources in evolutionary and comparative genomics [5-8]. Because of their high sequence similarity to their "parent" genes, pseudogenes often interfere with PCR or hybridization experiments that are intended for genes [9,10]. Pseudogenes are often erroneously annotated as functional genes in the sequence databanks [11]. Several genome-wide surveys were undertaken to identify pseudogenes in the completely sequenced genomes [12-18]. In a previous survey, we identified ~ 8,000 processed pseudogenes in the human genome [19]. In this paper, we describe the pseudogene population in mouse and present some comparative analysis between human and mouse. Details of the pseudogene annotation procedure have been described previously [18,19]; the data described here can be accessed at http://www.pseudogene.org/.

We did our pseudogene survey on the mouse assembly 14.30.1 downloaded from Ensembl website in March 2003 [20,21]. Table 1 lists, for each chromosome, the numbers of annotated pseudogenes and functional genes. We like to emphasize that all of our pseudogenes have been filtered to remove overlaps with the functional genes. We used the following criteria to decide whether a genomic locus is a processed pseudogene: (i) it shares high sequence similarity with a known mouse protein from Swiss-Prot or TrEMBL [22] (BLAST E-value < 1e-10 and amino acid sequence identity > 40%); (ii) the sequence alignment with the functional gene does not contain gaps longer than 60 bp; (iii) it covers longer than 70% of the coding sequence (CDS); (iv) it contains frame disruptions such as frameshifts or stop codons. Type 1 processed pseudogenes satisfy all four criteria. Type 2 satisfies all criteria except (iv); it is likely that these are the processed pseudogenes that were created recently so that they have not accumulated any frame disruptions yet. There are also some sequences (Type 3) that satisfy all criteria except (ii); these are the processed pseudogenes that are disrupted by repetitive elements. It has been shown that, in human and rodents, each ribosomal protein (RP) only has one functional, multiple-exon gene in the genome [23]. Therefore, we were certain that all the “single-exon” RP similarity loci in the mouse genome are processed pseudogenes and they were included into Type 1 regardless of the existence of frame disruptions. Only the Type 1 processed pseudogenes were included in the subsequent analysis that we described here; inclusion of the Type 2 and Type 3 did not affect the conclusions. We also tested different combinations of sequence identity and E-value cutoffs in the annotation procedure, which did not affect the conclusions either. Other than the processed pseudogenes, we also identified other types of pseudogenic sequences in the mouse including duplicated pseudogenes, pseudogenic fragments, olfactory receptor pseudogenes [24] and nuclear mitochondrial pseudogenes [25,26].

The number of processed pseudogenes in mouse is only half of that in human even though the mouse genome is only slightly smaller than the human genome (Table 1). However, such observation should not be interpreted as that the retrotransposition is less active in mouse. The mouse genome has higher nucleotide substitution, insertion and deletion rates than human [27,28], thus the pseudogenes in mouse decay faster and are more difficult to be recognized by sequence-similarity based methods. Similar observations have been made for the interspersed repeats: a larger fraction of the human genome (46%) than the mouse genome (37.5%) can be recognized as transposon-derived sequences even though transposition has actually been more active in the mouse lineage [28].

Table 2 compares some overall statistics of the mouse and human processed pseudogene (Type 1 only). The majority of the processed pseudogenes have retained a recognizable coding region even though generally they are not under selection pressure. Despite the sequence similarity, on average, a processed pseudogene still contains more than five frame disruptions. The mouse pseudogenes appear to be more decayed than human ones, which is the direct result of the higher neutral mutation rates in the mouse genome [27,28].

Like in human [19], the number of the processed pseudogenes on each mouse chromosome is proportional to the chromosome length ($R = 0.73$, $P < 0.0003$). However, even though the macroscopic distribution appears random and dispersed, microscopically, the density of the processed pseudogenes is highly uneven among regions of different (G+C) compositions (Figure 1). The LINE elements, SINE/Alu elements and processed pseudogenes were all processed by the same retrotransposition machinery, which has a preference for (G+C)-poor sites [3,4,29]. However, while the LINE elements have higher density in the (G+C)-poor regions of the genome, in contrast, the SINE/Alu elements are enriched in the (G+C)-rich regions (note that Alus are primate-specific SINE elements). This is because the SINE/Alu elements have higher (G+C) content (~57%) than both the LINE elements (~40%) and the genome background, therefore they have a faster decay rate in the compositionally destabilizing environment and quickly blend into the background [18,30,31]. In the mouse genome, similar to the LINE elements, the processed pseudogenes have the highest density in the (G+C)-poor regions and are depleted in the (G+C)-rich regions. The processed pseudogenes in human have a slightly different distribution, which are mostly enriched in the regions of intermediate (G+C) content. Overall, however, the picture between the genomes is broadly similar with SINE/Alu elements enriched in (G+C)-rich, LINEs in (G+C)-poor, and the pseudogenes in between, but shifted more towards LINEs in the mouse. It has been noted that the (G+C) distribution of the mouse genome is much tighter than human: the human genome has more regions with extreme (G+C) content whereas the mouse genome is much more uniform [28]. In addition to the higher mutation rate in mouse, this is likely to be the reason that caused the different distributions shown in Figure 1.

Figure 2 shows the distribution of the sequence divergence, or the age distribution, of the processed pseudogenes in comparison with the LINE and SINE elements in the human and mouse genomes. The rates of retrotransposition in the mouse and human lineages have evolved very differently since they diverged about 75 million years (Myr) ago. In human, the rate peaked at about 40 Myr ago and declined rapidly; meanwhile it has been more uniform in mouse [30] [28]. The age profiles of the processed pseudogenes also reflect the evolvement of the retrotransposition activity in each species. The profile of the human processed pseudogenes is very similar to that of the Alu elements, which is the dominant class of repeats in the human genome; the profile of the mouse processed pseudogenes coincide with that of the LINE1 elements.

Human and mouse lineages diverged at approximately 75 Myr ago, this corresponds to 16-17% sequence divergence in human and 33-34% in mouse [28] (Figure 2). From the divergence data, we estimated that about 60% of the processed pseudogenes in both the human and mouse genomes are lineage-specific, i.e. they were created after the last human and mouse common ancestor. This is in good agreement with the similar numbers we obtained from comparing our pseudogene annotations with the human-mouse whole genome alignment [32]. About 40% of the processed pseudogenes were found to be in a syntenic region that is preserved in both human and mouse, i.e. these are likely the ancestral pseudogenes that were created before human and mouse diverged.

The mouse and human genes that have multiple copies of processed pseudogenes are mostly housekeeping genes that are highly expressed in the germ line or embryonic cells

[19,33]. Figure 3(a) divides the mouse processed pseudogenes into subgroups according to the Gene Ontology (GO) functional categories of the functional genes [34]. As in the human genome, the largest subgroup is the ribosomal protein (RP) pseudogenes [18]; other notable groups include DNA/RNA binding proteins, structural molecules and metabolic enzymes. Some genes that are known to have many processed pseudogenes in human also have multiple processed pseudogenes in the mouse genome; these include *gapdh*, (186 copies), cyclophilin A (49) and cytochrome *c* (13). Figure 3b shows a significant correlation between, for each RP gene, the numbers of processed pseudogenes in mouse and in human (R = 0.52, P < 1e-7). The correlation is still strong if we only consider the lineage-specific pseudogenes (R = 0.50, P < 1e-5). This indicates that other than gene expression, other gene-specific factors also affect the abundance of the processed pseudogenes. These factors include gene length and gene (G+C) content [19,33].

The results we described here provide by far the most comprehensive catalogue of processed pseudogenes in the mouse genome, which will be regularly updated when new release of the genome draft becomes available. A three-species comparison will be more informative when the rat genome becomes available in the near future.

Acknowledgement

MG acknowledges financial support from NIH (NP50 HG02357-01). ZZ thanks Duncan Milburn, Hong Jiang, John Karro and Ted Johnson for computational assistances.

References

1 Mighell, A.J. et al. (2000) Vertebrate pseudogenes. *FEBS Lett.* 468 (2-3), 109-114

2 Maestre, J. et al. (1995) mRNA retroposition in human cells: processed pseudogene formation. *EMBO J.* 14 (24), 6333-6338

3 Feng, Q. et al. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87 (5), 905-916

4 Esnault, C. et al. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* 24 (4), 363-367

5 Petrov, D.A. et al. (1996) High intrinsic rate of DNA loss in Drosophila. *Nature* 384 (6607), 346-349.

6 Petrov, D.A. et al. (2000) Evidence for DNA loss as a determinant of genome size. *Science* 287 (5455), 1060-1062.

7 Zhang, Z. and Gerstein, M. (2003) The human genome has 49 cytochrome c pseudogenes, including a relic of a primordial gene that still functions in mouse. *Gene* 312, 61-72

8 Zhang, Z. and Gerstein, M. (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* 31 (18), 5338-5348

9 Guo, N. et al. (1998) The human ortholog of rhesus mannose-binding protein-A gene is an expressed pseudogene that localizes to chromosome 10. *Mamm Genome* 9 (3), 246-249.

10 Ruud, P. et al. (1999) 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 (1), 119-125.

11 Mounsey, A. et al. (2002) Evidence Suggesting That a Fifth of Annotated Caenorhabditis elegans Genes May Be Pseudogenes. *Genome Res.* 12 (5), 770-775.

12 Harrison, P.M. et al. (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Res.* 29 (3), 818-830

13 Harrison, P. et al. (2002) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J Mol Biol* 316 (3), 409-419.

14 Harrison, P.M. et al. (2002) Molecular fossils in the human genome: Identification and analysis of processed and non-processed pseudogenes in chromosomes 21 and 22. *Genome Res.* 12, 272-280

15 Harrison, P.M. et al. (2003) Identification of pseudogenes in the Drosophila melanogaster genome. *Nucleic Acids Res* 31 (3), 1033-1037.

16 Homma, K. et al. (2002) A systematic investigation identifies a significant number of probable pseudogenes in the Escherichia coli genome. *Gene* 294 (1-2), 25.

17 Cole, S.T. et al. (2001) Massive gene decay in the leprosy bacillus. *Nature* 409 (6823), 1007-1011.

18 Zhang, Z. et al. (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* 12 (10), 1466-1482.

19 Zhang, Z. et al. (2003) Millions of years of evolution preserved: a comprehensive catalogue of the processed pseudogenes in the human genome. *Genome Res (2003, in press)*

20 Birney, E. et al. (2001) Mining the draft human genome. *Nature* 409 (6822), 827-828

21 Hubbard, T. et al. (2002) The Ensembl genome database project. *Nucleic Acids Res.* 30 (1), 38-41

22 Boeckmann, B. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* 31 (1), 365-370

23 Wool, I.G. et al. (1995) Structure and evolution of mammalian ribosomal proteins. *Biochem. Cell Biol.* 73 (11-12), 933-947

24 Glusman, G. et al. (2001) The complete human olfactory subgenome. *Genome Res.* 11 (5), 685-702

25 Woischnik, M. and Moraes, C.T. (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Res* 12 (6), 885-893.

26 Tourmen, Y. et al. (2002) Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* 80 (1), 71-77.

27 Graur, D. et al. (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J Mol Evol* 28 (4), 279-285.

28 Waterston, R.H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420 (6915), 520-562.

29 Jurka, J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci.* 94 (5), 1872-1877

30 International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860-921

31 Pavlicek, A. et al. (2001) Similar integration but different stability of Alus and LINEs in the human genome. *Gene* 276 (1-2), 39-45

32 Kent, W.J. et al. (2002) The human genome browser at UCSC. *Genome Research* 12 (6), 996-1006

33 Goncalves, I. et al. (2000) Nature and structure of human genes that generate retropseudogenes. *Genome Res.* 10 (5), 672-678

34 Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25 (1), 25-29.

35 Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c, Distributed by the author. Department of Genetics, University of Washington, Seattle. *PHYLIP (Phylogeny Inference Package) version 3.5c*

36 Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16 (2), 111-120

37 Smit, A.F. and Green, P. unpublished data. *RepeatMasker Program (unpublished)*

38 Wu, C.I. and Li, W.H. (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci U S A* 82 (6), 1741-1745.

Figure Legends

Figure 1. The density of interspersed repeats and processed pseudogenes in the mouse (a) and the human (b) genomes. Pseudogene and the repeats are grouped according to the (G+C) content of the surrounding 100-kb DNA.

Figure 2. Age distribution of interspersed repeats and processed pseudogenes in mouse (a) and human (b). Pseudogenes and repeats are grouped according to their sequence divergence from the present-day functional genes or inferred consensus sequence of the ancient repeats. We used the package PHYLIP [35] to calculate the divergence data for the processed pseudogenes, following the Kimura 2-parameter model [36]. The divergence data of the repeats were derived from the program RepeatMasker [37]. Note that the average rate of nucleotide substitutions (per year) in the mouse lineage is about two-fold of that in the human lineage [28,38]; 1% sequence divergence represents roughly 4.5 Myr in human but only 2.2 Myr in mouse. There are many more ancient pseudogenes that have very high divergence values; these are not shown in the figures. Different scales on the Y-axis were used for the mouse and human data.

Figure 3. (a) Classification of the processed pseudogenes according to major GO functional categories [34]. "Unclassified" are those pseudogenes that arose from the functional genes that have not been assigned to a GO category. Less populated categories are lumped together into "Others". (b) Numbers of processed pseudogenes of the 79 ribosomal protein (RP) genes in the human and mouse genomes respectively. The RP genes are put into four groups according to the number of processed pseudogenes in the mouse genome.

Table 1 Number of pseudogenes in the mouse genome.

Chr.	Chr. length (Mb)	Ensembl genes [a]	Processed ΨG [b]			Dup. ΨG [c]	Frag. ΨG [d]
		Total (Known)	Type 1 (RP)	Type 2	Type 3		
1	197	1504 (931)	325 (70)	14	6	44	271
2	180	2058 (1429)	292 (87)	5	9	30	210
3	161	1237 (808)	293 (86)	15	7	49	235
4	152	1499 (968)	250 (69)	9	6	44	199
5	151	1478 (975)	238 (65)	8	6	60	191
6	150	1282 (885)	290 (61)	41	11	44	208
7	136	2003 (1384)	297 (62)	24	5	74	272
8	129	1189 (818)	197 (46)	11	6	22	132
9	126	1394 (992)	96 (14)	4	5	28	194
10	131	1205 (768)	219 (58)	9	6	44	172
11	123	1877 (1341)	181 (58)	16	4	30	169
12	114	852 (531)	174 (45)	9	9	18	145
13	117	1012 (653)	221 (57)	13	7	30	158
14	116	898 (566)	196 (46)	9	5	19	155
15	105	953 (610)	146 (40)	8	2	39	129
16	99	811 (533)	159 (48)	5	7	23	115
17	94	1157 (790)	192 (42)	10	5	53	171
18	91	628 (390)	175 (47)	11	5	25	109
19	61	795 (586)	89 (33)	3	5	13	64
X	147	1116 (557)	446 (100)	12	19	46	292
Total (mouse)	2581	24948 (16515)	4476 (1134)	236	135	735	3591
Total (human)	3040	22920 (17948)	7819 (1756)	737	1191	3015	6531

[a] Functional genes annotated by Ensembl (Release 15.30.1), which include known genes and novel genes.

[b] Processed pseudogenes. Definitions of the Type 1, 2 and 3 processed pseudogenes are described in the text. The RP processed pseudogenes are considered as Type 1 and their numbers are listed in the parenthesis.

[c] Duplicated pseudogene candidates. These duplicated pseudogenes were created by segment duplication or unequal crossing-over [1] therefore they have retained the original exon structure.

[d] Pseudogenic fragments. These are the protein similarity loci in the genome that are incomplete (< 70%) in the coding region.

Table 2 Overall statistics of the mouse and human processed pseudogenes

	Completeness (CDS only)		**DNA sequence identity**		**Average insertions, deletions or frame disruptions per pseudogene [c]**		
	Average	> 90% [a]	Average	> 90% [b]	Insertions (bp)	Deletions (bp)	Stop codons, frame shifts
Mouse	94%	3227 (72%)	77%	1202 (27%)	5.4	7.9	5.6
Human	95%	6054 (77%)	86%	3,066 (40%)	5.0	6.0	5.3

[a,b] The number and fraction of the processed pseudogenes that have sequence completeness or DNA sequence identity greater than 90%.

[c] Average number of inserted/deleted base pairs and number of frame disruptions in a processed pseudogene, compared with the corresponding functional mouse or human gene.

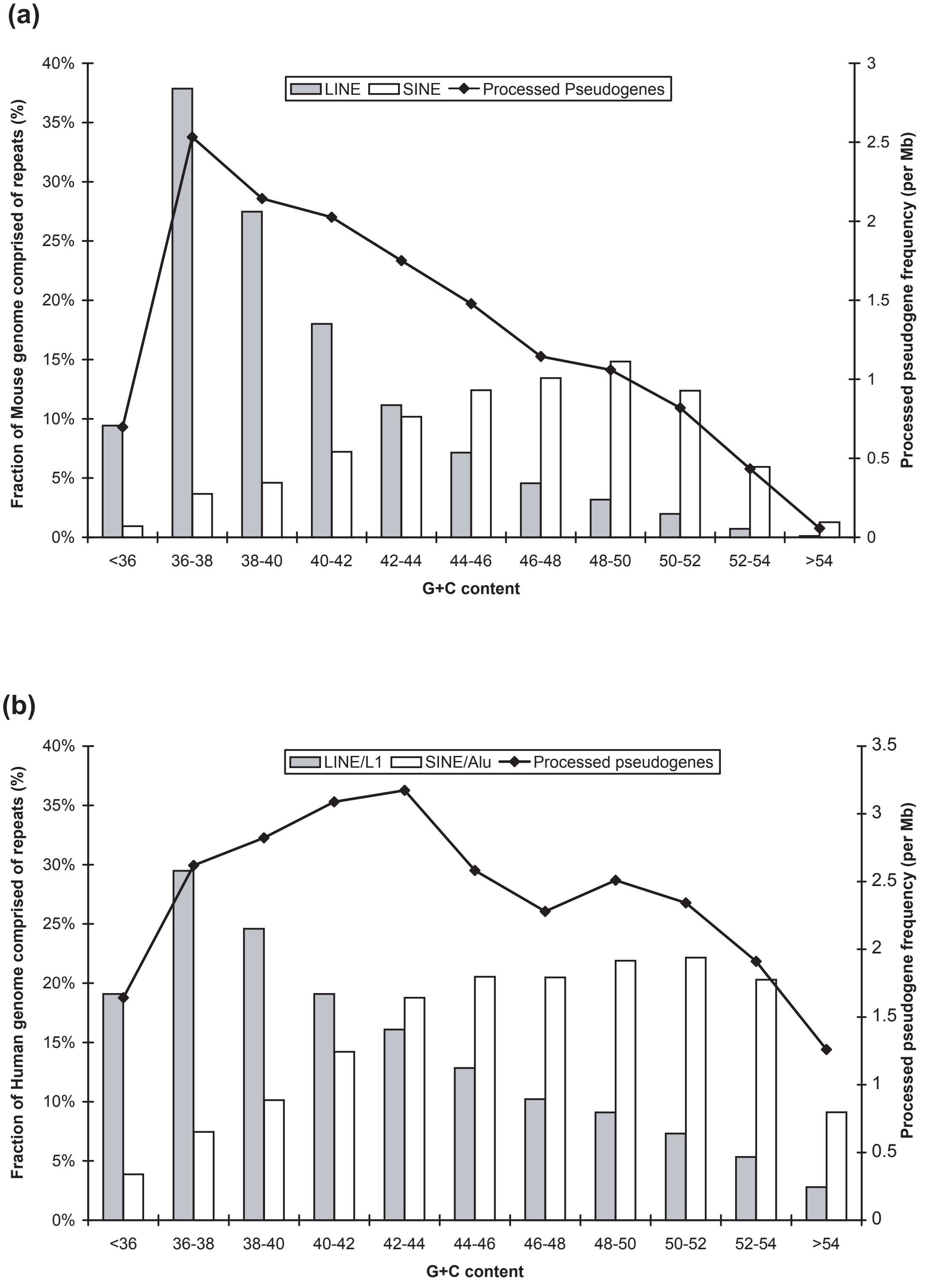
(a)
LINE
SINE
Processed Pseudogenes
Fraction of Mouse genome comprised of repeats (%)
40%
35%
30%
25%
20%
15%
10%
5%
0%
Processed pseudogene frequency (per Mb)
3
2.5
2
1.5
1
0.5
0
<36
36-38
38-40
40-42
42-44
44-46
46-48
48-50
50-52
52-54
>54
G+C content
(b)
LINE/L1
SINE/Alu
Processed pseudogenes
Fraction of Human genome comprised of repeats (%)
40%
35%
30%
25%
20%
15%
10%
5%
0%
Processed pseudogene frequency (per Mb)
3.5
3
2.5
2
1.5
1
0.5
0
<36
36-38
38-40
40-42
42-44
44-46
46-48
48-50
50-52
52-54
>54
G+C content

Fig 1

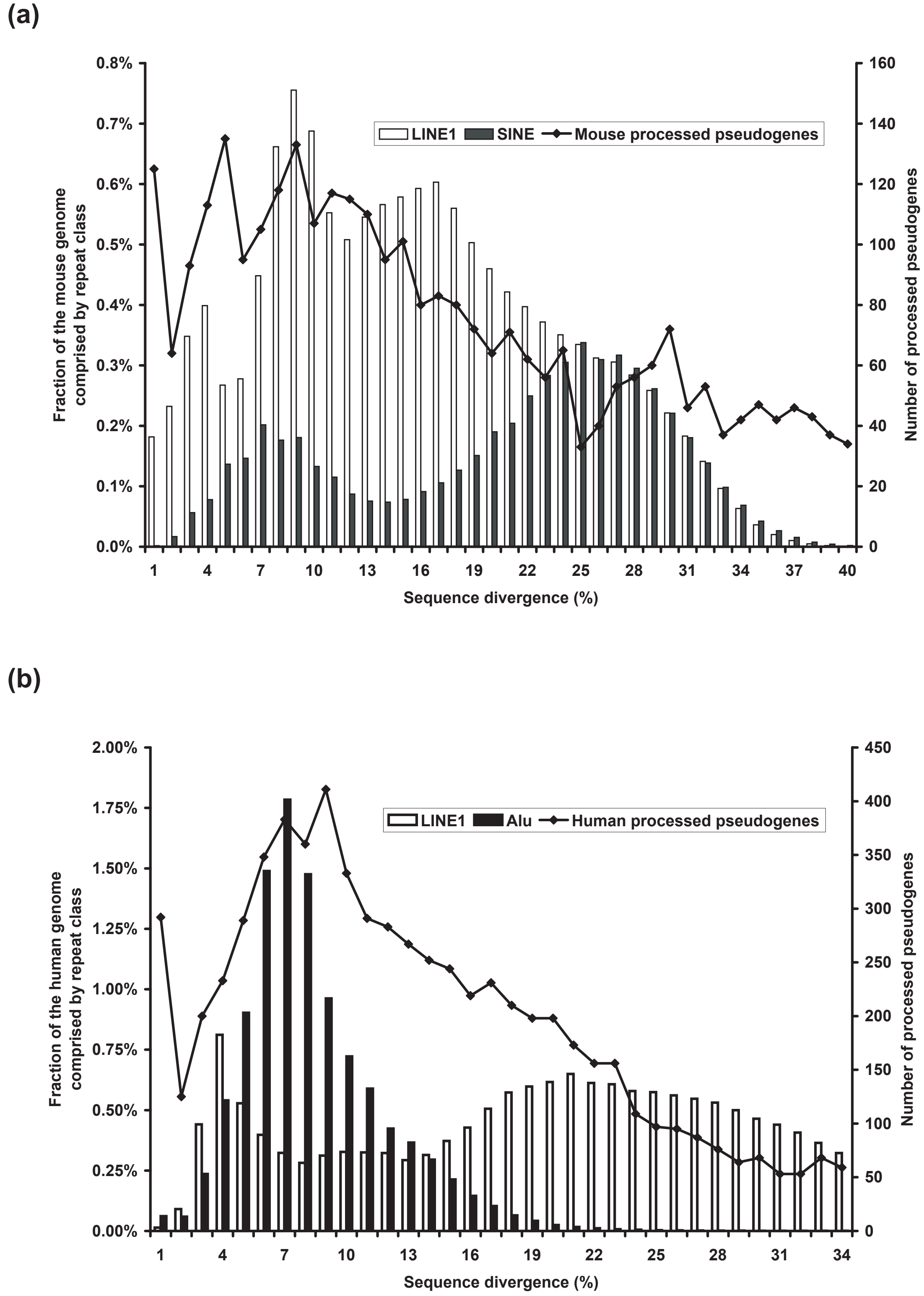
(a)
0.8%
0.7%
0.6%
0.5%
0.4%
0.3%
0.2%
0.1%
0.0%
Fraction of the mouse genome comprised by repeat class
LINE1
SINE
Mouse processed pseudogenes
160
140
120
100
80
60
40
20
0
Number of processed pseudogenes
1
4
7
10
13
16
19
22
25
28
31
34
37
40
Sequence divergence (%)
(b)
2.00%
1.75%
1.50%
1.25%
1.00%
0.75%
0.50%
0.25%
0.00%
Fraction of the human genome comprised by repeat class
LINE1
Alu
Human processed pseudogenes
450
400
350
300
250
200
150
100
50
0
Number of processed pseudogenes
1
4
7
10
13
16
19
22
25
28
31
34
Sequence divergence (%)

Fig 2

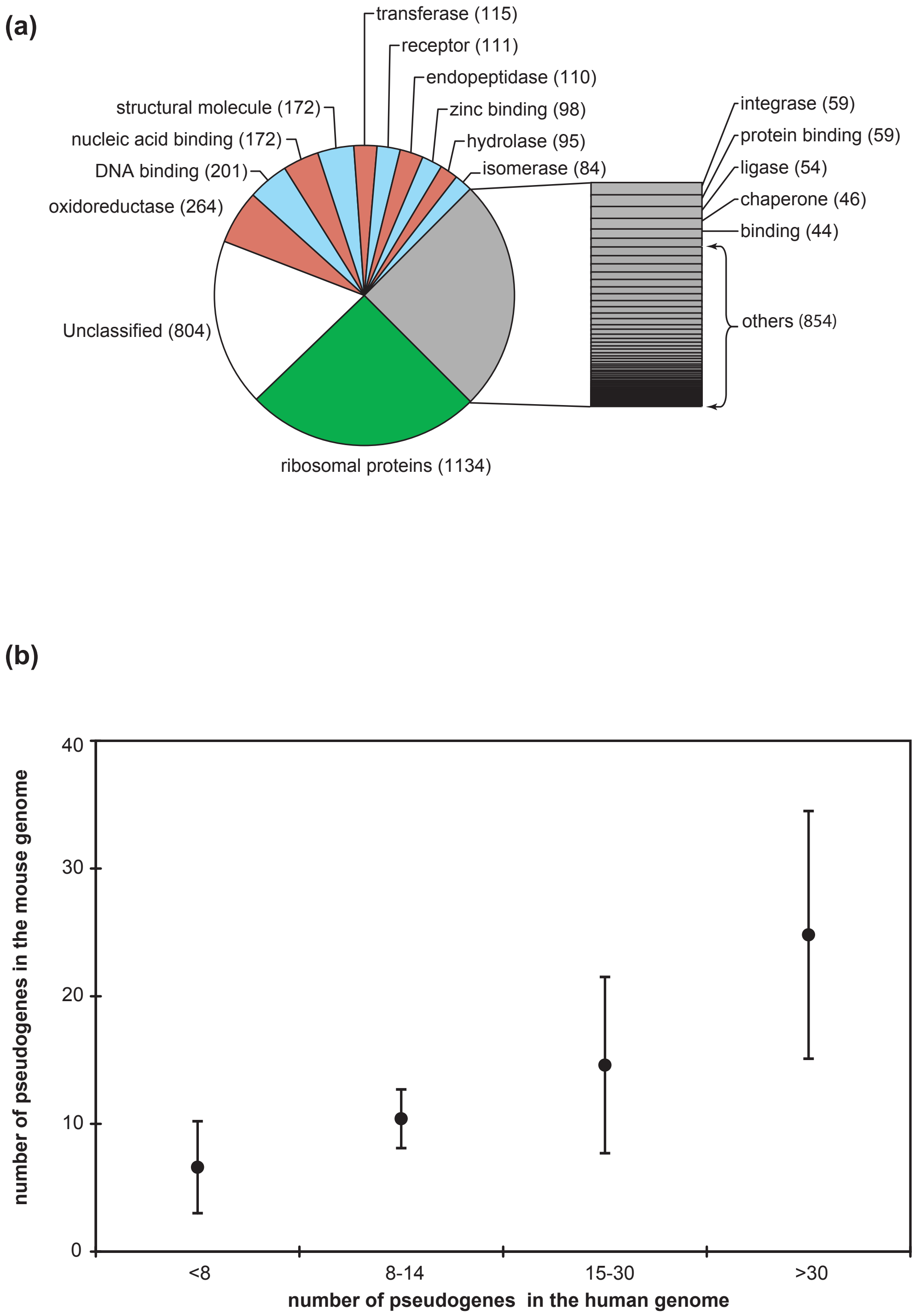
(a)
transferase (115)
receptor (111)
endopeptidase (110)
zinc binding (98)
hydrolase (95)
isomerase (84)
structural molecule (172)
nucleic acid binding (172)
DNA binding (201)
oxidoreductase (264)
Unclassified (804)
ribosomal proteins (1134)
integrase (59)
protein binding (59)
ligase (54)
chaperone (46)
binding (44)
others (854)
(b)
40
30
20
10
0
number of pseudogenes in the mouse genome
<8
8-14
15-30
>30
number of pseudogenes in the human genome

Fig 3