6 Gerlai, R. (2001) Gene targeting: technical confounds and potential solutions in behavioral brain research. *Behav. Brain Res.* 125, 13–21
7 Ledermann, B. and Burki, K. (1991) Establishment of a germ-line competent C57BL/6 embryonic stem cell line. *Exp. Cell Res.* 197, 254–258
8 Kontgen, F. *et al.* (1993) Targeted disruption of the MHC class II Aa gene in C57BL/6 mice. *Int. Immunol.* 5, 957–964
9 Zhou, L. *et al.* (2001) Murine inter-strain polymorphisms alter gene targeting frequencies at the mu opioid receptor locus in embryonic stem cells. *Mamm. Genome* 12, 772–778
10 te Riele, H. *et al.* (1992) Highly efficient gene targeting in embryonic stem cells through homologous recombination with isogenic DNA constructs. *Proc. Natl. Acad. Sci. U. S. A.* 89, 5128–5132
11 Simpson, E.M. *et al.* (1997) Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. *Nat. Genet.* 16, 19–27
12 Banbury Conference on Genetic background in Mice, (1997) Mutant mice and neuroscience: recommendations concerning genetic background. *Neuron* 19, 755–759
13 Mounkes, L. *et al.* (2003) The laminopathies: nuclear structure meets disease. *Curr. Opin. Genet. Dev.* 13, 223–230
14 Schuster-Gossler, K. *et al.* (2001) Use of coisogenic host blastocysts for efficient establishment of germline chimeras with C57BL/6J ES cell lines. *Biotechniques* 31, 1022–1024
15 Auerbach, W. *et al.* (2000) Establishment and chimera analysis of 129/SvEv- and C57BL/6-derived mouse embryonic stem cell lines. *Biotechniques* 29, 1024–1028
16 Le Fur, N. *et al.* (1996) Base substitution at different alternative splice donor sites of the tyrosinase gene in murine albinism. *Genomics* 37, 245–248
17 Liu, P. *et al.* (1998) Embryonic lethality and tumorigenesis caused by segmental aneuploidy on mouse chromosome 11. *Genetics* 150, 1155–1168
18 Fedorov, L.M. *et al.* (1997) A comparison of the germline potential of differently aged ES cell lines and their transfected descendants. *Transgenic Res.* 6, 223–231

0168-9525/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2003.12.006

Genome Analysis

Comparative analysis of processed pseudogenes in the mouse and human genomes

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

[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

Pseudogenes are important resources in evolutionary and comparative genomics because they provide molecular records of the ancient genes that existed in the genome millions of years ago. We have systematically identified ~5000 processed pseudogenes in the mouse genome, and estimated that ~60% are lineage specific, created after the mouse and human diverged. In both mouse and human genomes, similar types of genes give rise to many processed pseudogenes. These tend to be housekeeping genes, which are highly expressed in the germ line. Ribosomal-protein genes, in particular, form the largest sub-group. The processed pseudogenes in the mouse occur with a distinctly different chromosomal distribution than LINEs or SINEs – preferentially in GC-poor regions. Finally, the age distribution of mouse-processed pseudogenes closely resembles that of LINEs, in contrast to human, where the age distribution 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 because of a lack of functional promoters or other regulatory elements [1]. The majority of 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 long interspersed nuclear element 1 (LINE1) [2–4]. These processed pseudogenes typically do not contain introns and sometimes have a recognizable 3′ poly-adenine tail (if it has not decayed). Processed pseudogenes 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’ because 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 these genes [9,10]. Pseudogenes are often erroneously annotated as functional genes in 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 ~8000 processed pseudogenes in the human genome [19]. In this article, we describe the pseudogene population in mouse and present some comparative analysis between the human and the mouse genomes. Details of the pseudogene annotation procedure have been described previously [18,19]; the data described here can be accessed at: http://www.pseudogene.org/. The processed pseudogenes have been indexed at: http://www.pseudogene.org/mouse/.

Pseudogenes in mouse

Pseudogenes were surveyed using the mouse assembly 14.30.1, which was downloaded from the Ensembl website in March 2003 (http://www.ensembl.org) [20,21]. The

Corresponding author: Mark Gerstein (Mark.Gerstein@yale.edu).

numbers of annotated pseudogenes and functional genes are shown in Table 1. We would like to emphasize that all of the pseudogenes included in our analysis 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) the alignment covers $> 70\%$ of the coding sequence (CDS); and (iv) the sequence contains frame disruptions, such as frameshifts or stop codons. Type 1 processed pseudogenes satisfy all four criteria. Type 2 processed pseudogenes satisfy all criteria except (iv); it is likely that Type 2 processed pseudogenes were created recently, so that they have not yet accumulated any frame disruptions. 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) has only 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 in Type 1 regardless of whether they contained frame disruptions. Only Type 1 processed pseudogenes were included in subsequent analysis; the inclusion of Type 2 and Type 3 did not affect our conclusions. We also tested different combinations of sequence identity and E-value cut-off points in the annotation procedure, which also did not affect our conclusions. In addition to the processed pseudogenes, we also identified other types of pseudogenic sequences in the mouse genome, 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 observed in human, even though the mouse genome is only slightly smaller than the human genome (Table 1). However, such observations should not be interpreted to mean that 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 not recognized easily by sequence-similarity based methods. Similar observations have been made for the interspersed repeats: a larger proportion 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 mouse and human processed pseudogenes (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, a processed pseudogene contains, on average, more than five frame disruptions. The mouse pseudogenes appear to be more decayed than human pseudogenes. This is a direct result of the higher neutral mutation rates in the mouse genome [27,28].

Similar to humans [19], the number of processed pseudogenes on each mouse chromosome is proportional to the chromosome length ($R = 0.73$, $P < 0.0003$). However, although the macroscopic distribution appears

Table 1. The number of pseudogenes in the mouse genome

Chromosome	Chromosome 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	24 948 (16 515)	4476 (1134)	236	135	735	3591
Total (human)	3040	22 920 (17 948)	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 main text. The ribosomal protein (RP) processed pseudogenes are considered to be Type 1, and their numbers are listed in parentheses.

[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[e]		
	Average[a]	> 90%[b]	Average[c]	> 90%[d]	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%	3066 (40%)	5.0	6.0	5.3

[a] The average completeness of the pseudogene sequences in mouse and human.

[b,d] The number and percentage of the processed pseudogenes that have sequence completeness or DNA sequence identity >90%.

[c] The average sequence identity for all mouse and human pseudogenes.

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

random and dispersed, microscopically, the density of the processed pseudogenes is uneven among regions of different G + C compositions (Figure 1). The LINE, *Alu*elements, other short interspersed nuclear elements (SINEs) and processed pseudogenes were all processed by the same retrotransposition machinery, which has a preference for GC-poor sites [3,4,29]. However, LINEs have higher density in the GC-poor regions of the genome, whereas, the SINE and *Alu* elements are enriched in the GC-rich regions (note that *Alus* are primate-specific SINEs). This is because SINEs, including the *Alu* elements, have higher G + C content (~57%) than both

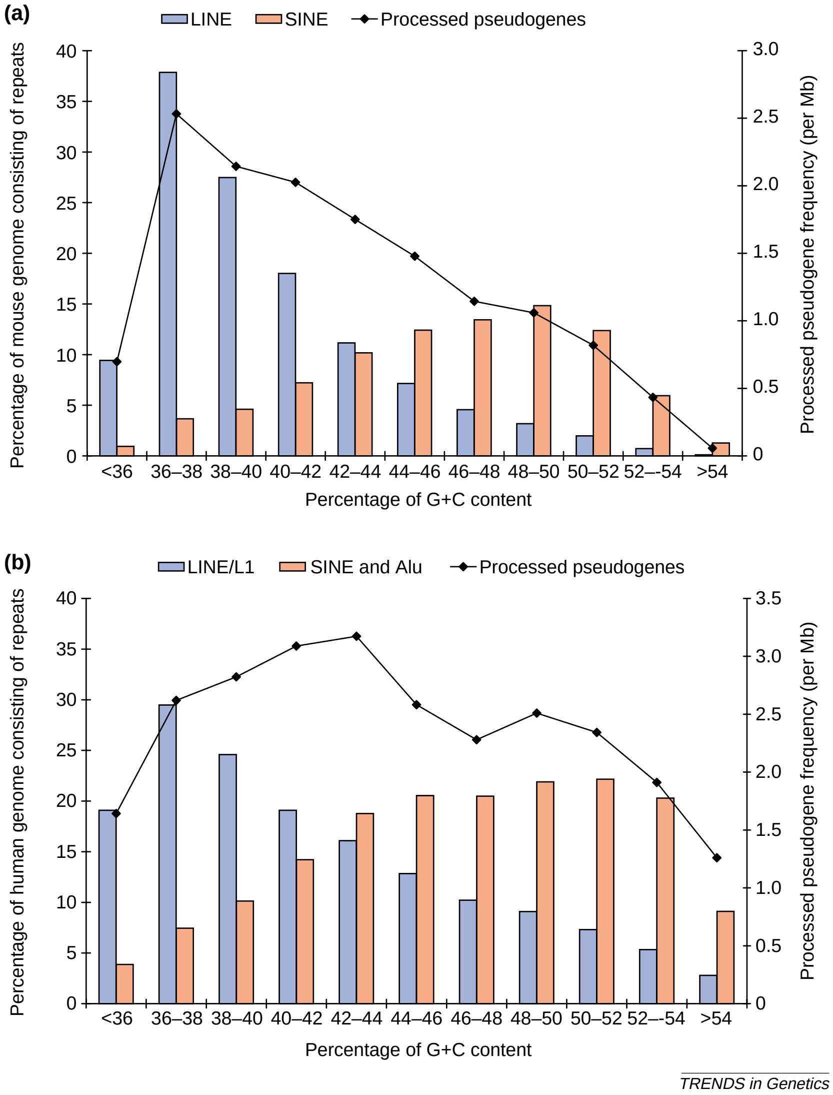

Figure 1. The density of interspersed repeats and processed pseudogenes in **(a)** the mouse and **(b)** the human genomes. Pseudogene and the repeats are grouped according to the G + C content of the surrounding 100-kb DNA.

LINEs (~40%) and the genome background. Therefore, SINE and *Alu* elements have a faster decay rate in the compositionally destabilizing environment and quickly blend into the background [18,30,31]. In the mouse genome, the processed pseudogenes, similar to LINEs, have the highest density in the GC-poor regions and are depleted in the GC-rich regions. The processed pseudogenes in human have a slightly different distribution, and are mostly enriched in the regions of intermediate G + C content. Overall, however, the situation in both genomes is broadly similar with SINE and *Alu* elements enriched in GC-rich regions, LINEs enriched in GC-poor regions, and the pseudogenes in the intermediate regions but shifted more towards GC-poor in the mouse. It has been noted that the GC distribution of the mouse genome is much 'tighter' than in 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 for the different distributions shown in Figure 1.

Pseudogene divergence

Figure 2 shows the distribution of the sequence divergence, or the age distribution, of the processed pseudogenes in comparison with LINEs and SINEs in the human and mouse genomes. The rates of retrotransposition in the

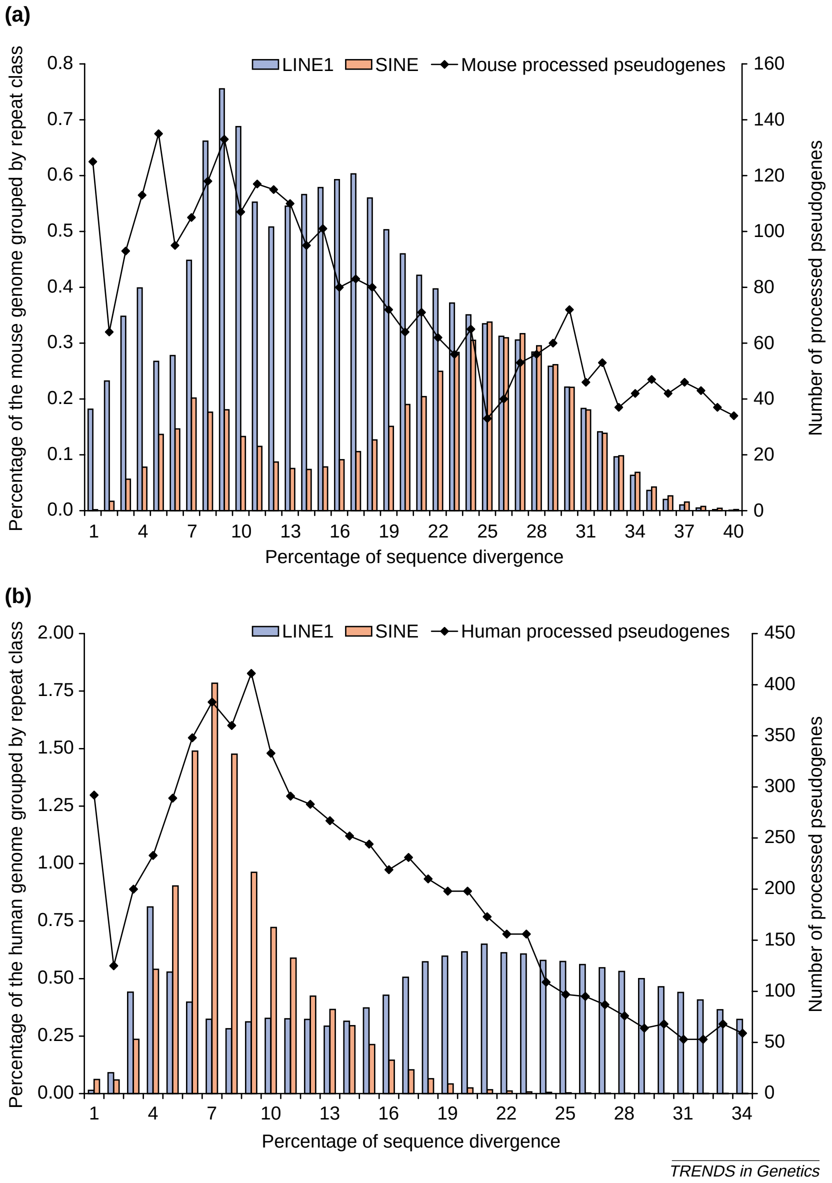

Figure 2. Age distribution of interspersed repeats and processed pseudogenes **(a)** in mouse **(b)** and human. Pseudogenes and repeats are grouped according to their sequence divergence from the present-day functional genes or inferred consensus sequence of the ancient repeats. We used the package PHYLIP [J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.5c, distributed by the author] to calculate the divergence data for the processed pseudogenes, following the Kimura 2-parameter model [35]. The divergence data of the repeats were derived from the program RepeatMasker (A.F. Smit and P. Green, unpublished). Note that, the average rate of nucleotide substitutions (per year) in the mouse lineage is about twofold of that in the human lineage [28,36]. A 1% sequence divergence represents ~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.

mouse and human lineages have evolved differently since they diverged ~75 million years (Myr) ago. In human, the retrotransposition rate peaked ~40 Myr ago and declined rapidly; however, it has been more uniform in mouse [28,30]. The age profiles of the processed pseudogenes also reflect the evolution of the retrotransposition activity in each species. The profile of the human processed pseudogenes is 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 coincides with that of LINE1.

Human and mouse lineages diverged ~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 ~60% of the processed pseudogenes in both the human and the mouse genomes are lineage specific (i.e. they were created after the last human and mouse common ancestor). This concurs with the results we obtained from comparing our pseudogene annotations with the human–mouse complete genome alignment [32]. Approximately 40% of the processed pseudogenes were found to be in a syntenic region that is preserved in both human and mouse. These are likely to be 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]. In Figure 3a, the mouse processed pseudogenes have been divided into subgroups according to the Gene Ontology (GO) functional categories of the functional genes [34]. Similar to the situation in the human genome, the largest subgroup is the RP pseudogenes [18]; other notable groups include DNA and 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;

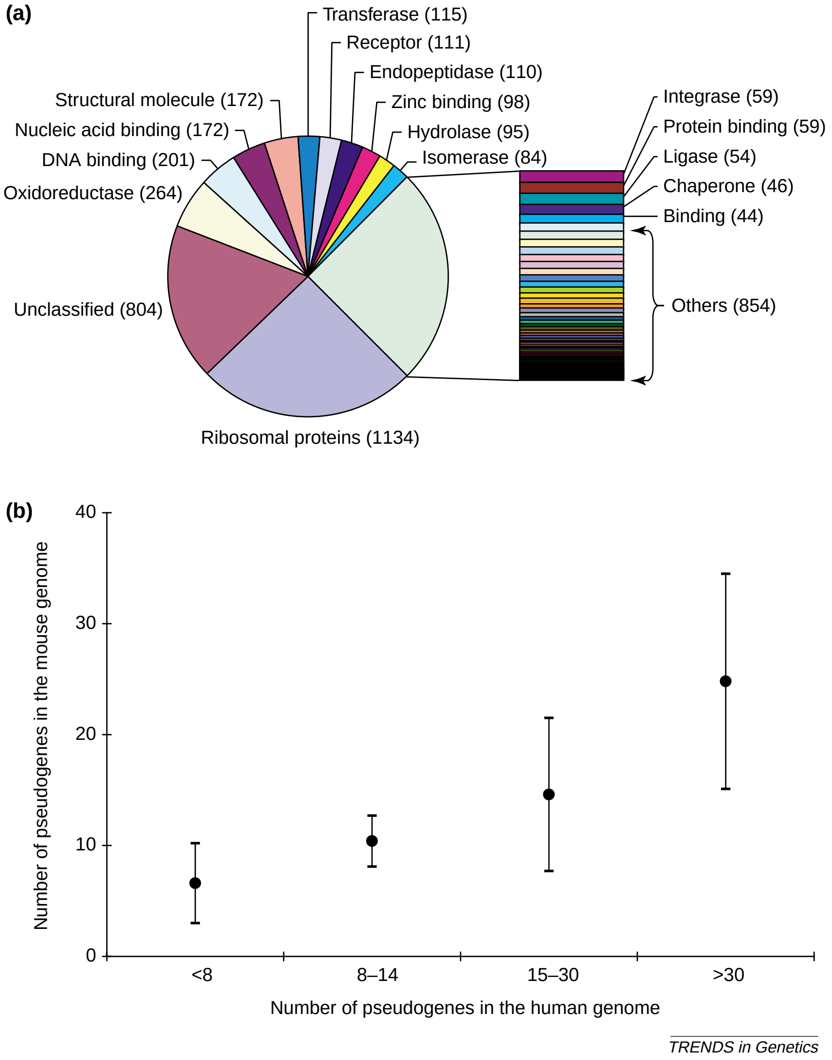

Figure 3. The properties of the parent genes. **(a)** Classification of the processed pseudogenes according to major Gene Ontology (GO) functional categories [34]. 'Unclassified' are those pseudogenes that arose from the functional genes that have not been assigned to a GO category. Categories with fewer members are referred to as 'Others'. **(b)** The numbers of processed pseudogenes of the 79 ribosomal protein (RP) genes in the human and mouse genomes are compared. The RP genes were divided into four groups according to the number of processed pseudogenes in the mouse genome.

these include *gapdh*, (186 copies), cyclophilin A (49 copies) and cytochrome *c* (13 copies). For each RP gene, there is a significant correlation between the numbers of processed pseudogenes in mouse and in human ($R = 0.52$, $P < 1e^{-7}$) (Figure 3b). The correlation is still significant if we only consider the lineage-specific pseudogenes ($R = 0.50$, $P < 1e^{-5}$). This indicates that, in addition to gene expression, other gene-specific factors affect the abundance of the processed pseudogenes. These factors include gene length and gene G + C content [19,33].

Concluding remarks

The results we have described in this article provide, by far, the most comprehensive catalogue of processed pseudogenes in the mouse genome, which will be updated regularly when a new release of the genome draft becomes available. It will be interesting, and informative, to compare the pseudogenes in human, mouse and rat. The rat genome sequence is likely to be available in early 2004.

Acknowledgements

M.G. acknowledges financial support from NIH (NP50 HG02357–01). Z.Z. 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, 109–114
2 Maestre, J. *et al.* (1995) mRNA retroposition in human cells: processed pseudogene formation. *EMBO J.* 14, 6333–6338
3 Feng, Q. *et al.* (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87, 905–916
4 Esnault, C. *et al.* (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* 24, 363–367
5 Petrov, D.A. *et al.* (1996) High intrinsic rate of DNA loss in *Drosophila*. *Nature* 384, 346–349
6 Petrov, D.A. *et al.* (2000) Evidence for DNA loss as a determinant of genome size. *Science* 287, 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, 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, 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, 119–125
11 Mounsey, A. *et al.* (2002) Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. *Genome Res.* 12, 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, 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, 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, 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, 25–33
17 Cole, S.T. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011
18 Zhang, Z. *et al.* (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* 12, 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.* 13, 2541–2558
20 Birney, E. *et al.* (2001) Mining the draft human genome. *Nature* 409, 827–828
21 Hubbard, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41
22 Boeckmann, B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370
23 Wool, I.G. *et al.* (1995) Structure and evolution of mammalian ribosomal proteins. *Biochem. Cell Biol.* 73, 933–947
24 Glusman, G. *et al.* (2001) The complete human olfactory subgenome. *Genome Res.* 11, 685–702
25 Woischnik, M. and Moraes, C.T. (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Res.* 12, 885–893
26 Tourmen, Y. *et al.* (2002) Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* 80, 71–77
27 Graur, D. *et al.* (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J. Mol. Evol.* 28, 279–285
28 Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
29 Jurka, J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1872–1877
30 International Human Genome Sequencing Consortium, (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
31 Pavlicek, A. *et al.* (2001) Similar integration but different stability of *Alus* and LINEs in the human genome. *Gene* 276, 39–45
32 Kent, W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.* 12, 996–1006
33 Goncalves, I. *et al.* (2000) Nature and structure of human genes that generate retropseudogenes. *Genome Res.* 10, 672–678
34 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29
35 Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120
36 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, 1741–1745

doi:10.1016/j.tig.2003.12.005