Large-scale Analysis of Pseudogenes in the Human Genome

ZhaoLei Zhang and Mark Gerstein*

Address

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

*Corresponding author

Tel: (203) 432 6105; Fax: (360) 838 7861; Email: Mark.Gerstein@yale.edu

Abbreviations

UTR	untranslated region
LINEs	long interspersed nuclear elements
SINEs	short interspersed nuclear elements
CDS	coding sequences
Ka	non-synonymous rate of substitution
Ks	synonymous rate of substitution

Keywords

Pseudogene, genomics, genome, bioinformatics

Summary

Pseudogenes are considered as gene fossils, i.e. they are disabled copies of functional genes that were once active in the ancient genome. Recently, whole-genome computational approaches have revealed thousands of pseudogenes in the human and other eukaryotic genomes. Identification of these pseudogenes can improve the accuracy of gene annotation. It also offers new insight on the evolution history of human genes and the stability of genome as a whole.

Introduction

Mammalian genomes, such as human and mouse, contain large number of gene-like sequences called pseudogenes. These pseudogenes are inheritable, non-functional, gene homologies that are generally disabled at transcriptional level [1,2]. In most cases, pseudogenes cannot produce transcripts due to the lack of functional promoters. Very rarely, some pseudogenes have retained or acquired a functional promoter so they can be transcribed, but these transcripts are not translated due to lack of translational or splicing signal sequences. As the result of their non-functionality, pseudogenes are generally released from selective pressure and often accumulate mutations such as frameshifts, in-frame stop codons, or interspersed repeats in the original protein-coding sequence (CDS) (see Figure 1). Consequently, we can identify pseudogenes operationally through finding regions of homology that have these non gene-like features (Table 2).

Depending on the mechanism by which they were generated, majority of the mammalian pseudogenes can be divided into duplicated pseudogenes and retrotransposed pseudogenes (also called processed pseudogenes). Duplicated pseudogenes arose from tandem duplication or unequal crossing-over, thus they often have retained the original exon-intron structures of the parental genes, though sometimes incompletely. In contrast, retrotransposed pseudogenes were created from retrotransposition, i.e. the reverse transcription of mRNA transcript followed by integration into the genome [3,4]. Therefore, retrotransposed pseudogenes are often considered as a special type of retrotransposons, just like long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) in the mammalian genomes [5]. Retrotransposed pseudogenes also share some of the common characteristics of the LINEs and SINEs, which include completely lack of introns, the presence of small flanking direct repeats, and a polyadenine tail near the 3'-end. Because of their close homology to functional genes, pseudogenes often introduce errors or contaminations in the sequence databases (Figure 1). In addition to retrotransposed and duplicated pseudogenes, other types of pseudogenes also exist in the human genome (see below).

Over the years, pseudogenes have been comprehensively surveyed in several completely sequenced genomes (Table 1). In 2002, a preliminary survey reported about 400 pseudogenes on the two smallest human chromosomes, 21 and 22 [6]. Several other studies have focused on the pseudogene population of selected gene families [7-10]. Year 2003 proved to be an exciting year for pseudogenes, as three research groups independently published comprehensive surveys of pseudogenes in the entire human

genome [11-13]. It was also discovered in the same year that a mouse pseudogene actually has a regulatory role [14].

Whole-Genome identification of pseudogenes

Traditionally, pseudogenes were often discovered as by-products of gene sequencing or PCR experiments. It is only after the whole-genome sequencing projects that large number of pseudogenes were identified and annotated. Using a homology based approach, Zhang and colleagues identified ~8,000 retrotransposed pseudogenes and ~3,000 duplicated pseudogenes in the human genome draft (Build 28, April, 2002 release) [12]. Ohshima and colleagues [11] used basically the same approach in their survey except that they used an older release of the human genome (April, 2001).

In addition to just relying on existence of truncation or frame disruptions to ascertain the non-functionality of the pseudogenes, Torrents and colleagues [13] developed a neutrality test by computing the ratio of synonymous to non-synonymous substitution rates (K_A/K_S) for each pseudogene. The K_A/K_S ratio measures how often nucleotide substitutions in a DNA sequence change the amino acid and this ratio is often used to test whether a sequence is under selective constraints [15]. These researchers reported ~20,000 potential human pseudogenes. By correlating with sequence conservation in the mouse syntenic regions, they estimated 70% of these were retrotransposed pseudogenes. The pseudogene annotations can also be validated by the absence "CpG islands" in their 5' upstream regions; this is because these "CpG islands" are often associated with the 5' end of the functional genes [16]. Operationally, as we have pointed out earlier, pseudogenes can be

defined by a variety of different sequence features. The three research groups (Tokyo, Yale, and EMBL) have taken somewhat different approaches towards the definition, resulting in different numbers. The differences are summarized in Table 2. Some features listed in Table 2 were not used in the identifying pseudogenes, but rather in the later stages of analysis and inferences.

Exact number of the pseudogenes in the human genome

It is a little surprising that the total numbers of human pseudogenes reported by the three research groups are quite different. Much of the discrepancy can be attributed to the different criteria used by individual groups. Ohshima *et al* [11] applied the stringiest criteria in their procedures as they only presented those pseudogenes that are 90% complete in comparison with their parental genes. Zhang and colleagues [12] counted those candidates that are 70% complete in coding region as pseudogenes and designate those shorter than 70% as "pseudogenic fragments". In contrast, Torrents and colleagues [13] did not apply any sequence completeness threshold in their procedures. If the 70% completeness cutoff is applied to the pseudogene set derived by Torrents *et al*., approximately 7,800 of them are indeed longer than this threshold. This is actually remarkably close to the number reported by Zhang *et al* [12]. Thus, even though the reported numbers differ, the results from the three groups are actually consistent with each other.

Pseudogenes in other organisms

In addition to human, large numbers of pseudogenes were also identified in the genomes of other eukaryotes including *C. elegans* [17], budding yeast [18], puffer fish [19], and fruitfly [20]. Some prokaryotic genomes also reportedly have many pseudogenes [21-23]. Generally pseudogenes are less common in prokaryotes since their genomes are more compact and have higher DNA deletion rates [24].

The initial annotation of the mouse genome reported about 14,000 putative pseudogenes [25]. A more recent study revealed about 5,000 retrotransposed pseudogenes in mouse [26], was based on the same criteria that was used for the human pseudogenes [12]. This is significantly less than the number of retrotransposed pseudogenes in human, even though the mouse genome is only slightly smaller than the human genome. However, this does not mean that retrotransposition is less active in mouse. The mouse genome has higher nucleotide substitution, insertion and deletion rates than human [25,27], thus the pseudogenes in mouse decay faster and are not recognized as easily as those in the human genome.

It is interesting to estimate what fraction of the pseudogenes in the human and mouse genomes are lineage-specific, i.e. those pseudogenes that were created after the primate and rodent lineages split at about 75-80 million years (Myr) ago [25]. Torrents and colleagues [13] reported that ~76% of their "pseudocoding regions" in the human genome can also be found within the corresponding mouse synetnic regions. Using an

alternative approach, Zhang et al. [26] estimated sequence divergence between the pseudogenes and the parent gene and converted the divergence data to evolutionary time. These researchers concluded that about 60% of the retrotransposed pseudogenes in the human and mouse genomes are lineage-specific.

Retrotransposed pseudogenes is a special type of retrotransposons

Human genome contains about several millions of copies of LINE and SINE elements which comprise >30% of the entire human genomic DNA [5]. While LINEs are autonomous (i.e. they can retrotranspose their own transcripts), SINEs have to rely on active LINEs to propagate. It is believed that LINE retrotransposons are also responsible of mobilizing mRNA transcripts and generating retrotransposed pseudogenes [4]. Macroscopically, the distribution of the retrotransposed pseudogenes in the human genome is random and dispersed, with the pseudogene abundance on each chromosome proportional to its length. Despite the common mechanism in their biogenesis, LINEs, SINEs and retrotransposed pseudogenes have distinct distributions in the genomic regions of different G+C composition [12]. This discrepancy has been explained by the different stability of the retrotransposons and pseudogenes in different regions [28].

By calculating the sequence divergence between the sequence of pseudogene and the parental functional gene, one can estimate the age of a pseudogene, i.e. the time that has elapsed since it became non-functional [12] [11,29]. The retrotransposed pseudogenes in the human genome have an overall age profile that is similar to that of the *Alu* elements,

the predominant SINE elements in primates. The rate of new retrotransposed pseudogenes generated in human peaked at approximately 40 million years ago, which coincided with the onset of the higher primates radiation [11,12].

Highly expressed genes tend to have multiple retrotransposed pseudogenes

The number of retrotransposed pseudogenes per gene is highly uneven among human genes. In fact, only 10% of the human genes have at least one retrotransposed pseudogene identified [11,12]. Ribosomal proteins, which have 79 genes in the human genome, account for nearly 20% of the entire retrotransposed pseudogenes population [7]. Other genes that have multiple retrotransposed pseudogenes include housekeeping genes, genes that code for structure protein and metabolic enzymes. In general, the genes that have multiple retrotransposed pseudogenes tend to be highly expressed, have short transcripts, and have lower G+C composition [12,30]. Figure 2 shows the functional categories of the human and mouse genes that gave rise to multiple retrotransposed pseudogenes. These also include some genes that are involved in cancer or have other medical implications such as cyclophilin, nucleophosmin and prohibitin [12]. For those human genes that have multiple retrotransposed pseudogenes, their mouse homologues also tend to have many pseudogenes in the mouse genome.

Pseudogenes as tools to study gene and genome evolution

Pseudogenes are often considered as "genomic fossils" as they provide snapshots of the ancient genes that were active millions of years ago. They can be analyzed to infer the evolutionary history of particular genes or gene families. By comparing the sequences of human cytochrome *c* (*cyc*) pseudogenes with the functional *cyc* gene from human and mouse, it became obvious that accelerated evolution in *cyc* gene had occurred in the primate lineage leading to human [31]. In another case, it is found that the orthologs of a human keratin pseudogene in the chimpanzee and gorilla are still functional [32].

Since pseudogenes are free to accumulate mutations, they are also very valuable in studying nucleotide substitution, insertion and deletions [33,34]. On a related note, retrotransposition of mRNA transcripts has been suggested as an important mechanism of generating new genes [35-37]. Brosius and colleagues have argued that mammalian genomes were forged and shaped by "massive bombardments" of retrotransposed sequences [38,39].

Some pseudogenes are transcribed

Because pseudogenes have high sequence similarity with their parental genes, they can potentially introduce contaminations in hybridization or amplification experiments. Special cautions need to be taken to prevent such interferences [40]. It has been reported that a cytokeratin-19 pseudogene may have interfered with diagnostic assays used to detect micrometastatic tumor cells [41]. In another instance, a novel pseudogene of *phox,*

component of phagocyte NADPH oxidase complex, complicates the detection of chronic granulomatous disease [42,43].

The original definition of pseudogenes implies that they are transcriptional silent, however over the years there have been many reported cases where a pseudogene can indeed be transcribed (for a complete list, see [2]). In one instance, it was found that a tumor suppressor gene, PTEN, has a transcribed retrotransposed pseudogene that has more transcripts than the parental functional gene [44]. In another case, a pseudogene even has developed a tissue-specific expression pattern [45].

Potential functional roles of pseudogenes

Because of their close similarities to the functional genes and high level of sequence conservation, pseudogenes, especially those are transcribed, have been hypothesized to have regulatory roles [46]. Korneev and colleagues has reported that, in the neurons of mollusk *Lymnaea stagnalis,* a transcribed pseudogene of neural *nitric oxide synthase* (nNOS) suppresses the synthesis of nNOS protein in an RNAi-like mechanism [47]. The transcript of the pseudogene contains a region with significant antisense homology to the nNOS mRNA transcript and binds to the nNOS transcript to form a stable RNA/RNA duplex. In another example, the pseudogene of the mouse gene *Makorin1* modulates the expression of the homologous functional gene in either an RNA-mediated or a DNA-mediated mechanism [14]. At the RNA level, the pseudogene RNA transcript could compete with the functional mRNA for an RNA-digesting enzyme. At the DNA level, the

pseudogene locus could potentially compete with the functional *Makorin1* gene for transcription repressors [48].

Pseudogenes have also been proposed to serve as a sequence pool for generating genetic diversity [2]. Genes and pseudogenes can recombine and produce new genes; such processes have been reported in the human immune system [49].

Conclusions

Pseudogenes are ubiquitous and abundant in the mammalian genomes. Their importance and implications have captured the interests of researchers from very diverse disciplines. The fact that pseudogenes have regulatory roles further demonstrates that these sequences should not be treated as "junk DNA". With more mammalian genomes such as that of chimpanzee being sequenced, a more complete picture of pseudogenes and their functional roles is starting to emerge.

Acknowledgements

M.G. acknowledges financial support from NIH (NP50 HG02357–01). Z.Z. acknowledges Dr. Paul Harrison and Dr. Duncan Milburn for helpful discussions.

Figure Legends

Figure 1

(*A*) A screen shot from the Ensembl website showing the contamination of pseudogenes in the genomic databases. The human cytochrome *c* functional gene (*cyc*) is located in the chromosome 7. Many retrotransposed *cyc* pseudogenes exist in the human genome. The red arrows point to those pseudogenes that are mistakenly annotated as genes By Ensembl. The functional *cyc* gene contains 1 intron in the coding region while the pseudogenes have no introns. (*B*) The amino acid and nucleotide sequence alignments between the functional cyc gene and a pseudogene. The pseudogene contains frame shifts and stop codons. (*C*) Retrotransposed pseudogene can be used to verify the exon structure predictions. The exon structures predicted by RefSeq and Ensembl are compared with a retrotransposed pseudogene. The inconsistency between the predictions and the pseudogene sequence could represent alternative splicing or erroneous predictions.

Figure 2

Functional classification of the retrotransposed pseudogenes in the human genome (*A*) and mouse genome (*B*), according to Gene Ontology functional categories. "Unclassified" are those pseudogenes that arose from genes that were not yet assigned to a GO category. Less populated categories are lumped together into "Others."

Table 1 Annotated pseudogenes in the completely sequence genomes.

Organism	Genome size [Mb]	No. of genes	No. of pseudogenes	No. of retrotransposed pseudogenes	Reference
R. prowazekii	1.1	834	241	0	[21]
M. leprae	3.3	1,604	1,116	0	[22]
Y. pestis	4.6	4,061	160	0	[50]
E. coli, K-12	4.6	4,400	95	0	[23]
E. coli, O157	5.5	6,000	101	0	[23]
S. cerevisiae	12.1	6,340	241	0	[18]
C. elegans	102.9	20,009	2,168	208	[17]
D. melanogaster	128.3	14,332	110	34	[20]
A. thaliana	115.4	25,464	> 700	?	[51]
H. sapiens	3,040	~35,000			
			~14,000	~7,800	[12]
			~3,600	~3,600	[11]
			~19,000	~13,300	[13]
M. musculus	2,493	~22,000			
			~10,000	~4,500	[26]
			~13,000	N/A	[25]

Table 2

Features of pseudogenes (or potential pseudogenes)	Tokyo [11]	Yale [12,26]	EMBL [13]	Others *
Level of sequence homology to parent gene	■	■	■	
Sequence completeness relative to parent gene	■	■	○	
Absence of introns	■	■	◆	
Ratio of the non-synonymous to synonymous substitution rates (Ka/Ks)	○	▲	■	[25,52]
Chromosomal location (in relation to parent gene)	○	▲	◆	
Existence of frame disruptions (frameshifts and stops)	○	■	○	
G+C content of pseudogenes and background	▲	▲	○	
Expression level of the parent gene	○	▲	○	
Occurrence of regulatory regions such as CpG islands	○	○	○	
Codon composition and nucleotide substitutions in relation to parental gene	○	▲	○	[34,53]
Occurrence of polyadenine tail	○	■	○	
Conservation with mouse genome	○	▲	■	
Association with evidence of transcription such as EST matches or micro-array data	○	○	○	[54,55]
Occurrence of SNPs	○	○	○	[56]
Number of pseudogenes per gene family	▲	▲	▲	[7-10,57]

■ Main feature used for the assignment of pseudogenes.

◆ Minor feature used for the assignment of pseudogenes.

▲ Surveyed after assignment of pseudogenes (in comparison to genes).

○ Not currently used or surveyed but potentially could be.

***** Analysis performed by others.

References

*1. Mighell AJ, Smith NR, Robinson PA, Markham AF: Vertebrate pseudogenes. *FEBS Lett* 2000, 468:109-114.

A concise and comprehensive introductory review on pseudogenes. A good beginner's guide on the subject.

**2. Balakirev ES, Ayala FJ: Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* 2003, 37:123-151.

This is a more detailed and up-to-date review on pseudogenes. The authors also discussed pseudogenes in an evolutionary perspective.

3. Maestre J, Tchenio T, Dhellin O, Heidmann T: mRNA retroposition in human cells: processed pseudogene formation. *EMBO J.* 1995, 14:6333-6338.

**4. Esnault C, Maestre J, Heidmann T: Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* 2000, 24:363-367.

For the first time, these authors demonstrated *in vivo* that human LINE retrotransposons were able to generate retrotransposed pseudogenes from mRNA transcripts.

*5. Deininger PL, Moran JV, Batzer MA, Kazazian HH, Jr.: Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* 2003, 13:651-658.

mammalian genome evolution. *Curr Opin Genet Dev* 2003, **13**:651-658.

This is a comprehensive review on mobile elements in the mammalian genome.

*6. Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, Johnson T, Gerstein M: Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res* 2002, 12:272-280.

Serving as a pilot project for the whole-genome search, the authors surveyed for pseudogenes in the two smallest human chromosomes.

* 7. Zhang Z, Harrison P, Gerstein M: Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* 2002, 12:1466-1482.

The authors searched for retrotransposed pseudogenes of the ribosomal proteins in the entire human genome.

8. Glusman G, Yanai I, Rubin I, Lancet D: The complete human olfactory subgenome. *Genome Res.* 2001, 11:685-702.

9. Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P: Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* 2002, 80:71-77.

10. Woischnik M, Moraes CT: Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Res.* 2002, 12:885-893.

**11. Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N: Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* 2003, 4:R74.

These authors reported the retrotransposed pseudogenes in the human genome that are longer than 90% of the complete cDNA length. They also devised a more rigorous approach in calculating the ages of the pseudogenes

**12. Zhang Z, Harrison PM, Liu Y, Gerstein M: Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* 2003, 13:2541-2558.

This is a continuation of these authors' work on ribosomal protein pseudogenes. They described the overall properties of the human pseudogenes, such as genomic distribution, functional categories, and synteny in the mouse genome.

**13. Torrents D, Suyama M, Zdobnov E, Bork P: A genome-wide survey of human pseudogenes. *Genome Res* 2003, 13:2559-2567.

These authors designed a functionality test, by calculating Ka/Ks ratios, to identify ~19,000 pseudogene candidates in the human genome. Retrotransposed pseudogenes were separated from duplicated pseudogenes based on mouse synteny information.

14. Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A: An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 2003, 423:91-96.

15. Hurst LD: The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* 2002, 18:486.

16. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.

17. Harrison PM, Echols N, Gerstein MB: Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Res.* 2001, 29:818-830.

18. Harrison P, Kumar A, Lan N, Echols N, Snyder M, Gerstein M: A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J Mol Biol* 2002, 316:409-419.

19. Dasilva C, Hadji H, Ozouf-Costaz C, Nicaud S, Jaillon O, Weissenbach J, Crollius HR: Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the Tetraodon nigroviridis genome. *Proc Natl Acad Sci U S A* 2002, 99:13636-13641.

20. Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M: Identification of pseudogenes in the Drosophila melanogaster genome. *Nucleic Acids Res* 2003, 31:1033-1037.

21. Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D, Roux V, Cossart P, Weissenbach J, Claverie JM, et al.: Mechanisms of evolution in Rickettsia conorii and R. prowazekii. *Science* 2001, 293:2093-2098.

22. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, et al.: Massive gene decay in the leprosy bacillus. *Nature* 2001, 409:1007-1011.

23. Homma K, Fukuchi S, Kawabata T, Ota M, Nishikawa K: A systematic investigation identifies a significant number of probable pseudogenes in the Escherichia coli genome. *Gene* 2002, 294:25.

24. Lawrence JG, Hendrix RW, Casjens S: Where are the pseudogenes in bacterial genomes? *Trends Microbiol* 2001, 9:535-540.

25. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al.: Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002, 420:520-562.

**26. Zhang Z, Carriero N, Gerstein M: Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet* 2004, 20:62-67.

Using the same approach as for human pseudogenes, these researchers identified ~ 5,000 retrotransposed pseudogenes in the mouse genome. Based on the age profile and the synteny information, it was estimated that about 60% of the pseudogenes were lineage specific.

27. Graur D, Shuali Y, Li WH: Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J Mol Evol* 1989, 28:279-285.

28. Pavlicek A, Jabbari K, Paces J, Paces V, Hejnar JV, Bernardi G: Similar integration but different stability of Alus and LINEs in the human genome. *Gene* 2001, 276:39-45.

29. Fleishman SJ, Dagan T, Graur D: pANT: a method for the pairwise assessment of nonfunctionalization times of processed pseudogenes. *Mol Biol Evol* 2003, 20:1876-1880.

30. Goncalves I, Duret L, Mouchiroud D: Nature and structure of human genes that generate retropseudogenes. *Genome Res.* 2000, 10:672-678.

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

It was found that cytochrome c gene has 49 pseudogenes in the human genome. An evolutionary history of this gene was revealed from analysis of the pseudogenes.

32. Winter H, Langbein L, Krawczak M, Cooper DN, Jave-Suarez LF, Rogers MA, Praetzel S, Heidt PJ, Schweizer J: Human type I hair keratin pseudogene phihHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence. *Hum Genet* 2001, 108:37-42.

33. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL: Evidence for DNA loss as a determinant of genome size. *Science* 2000, 287:1060-1062.

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

**35. Long M, Langley CH: Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* 1993, 260:91-95.

These authors described the possible ways that new genes are generated in the genome, which include retrotransposition.

36. Brosius J: Many G-protein-coupled receptors are encoded by retrogenes. *Trends Genet* 1999, 15:304-305.

37. Long M, Deutsch M, Wang W, Betran E, Brunet FG, Zhang J: Origin of new genes: evidence from experimental and computational analyses. *Genetica* 2003, 118:171-182.

38. Brosius J: Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* 1999, 107:209-238.

39. Brosius J: The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* 2003, 118:99-116.

40. Harper LV, Hilton AC, Jones AF: RT-PCR for the pseudogene-free amplification of the glyceraldehyde-3-phosphate dehydrogenase gene (gapd). *Mol Cell Probes* 2003, 17:261-265.

41. Ruud P, Fodstad O, Hovig E: 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* 1999, 80:119-125.

42. Heyworth PG, Noack D, Cross AR: Identification of a novel NCF-1 (p47-phox) pseudogene not containing the signature GT deletion: significance for A47 degrees chronic granulomatous disease carrier detection. *Blood* 2002, 100:1845-1851.

43. Harbord M, Hankin A, Bloom S, Mitchison H: Association between p47phox pseudogenes and inflammatory bowel disease. *Blood* 2003, 101:3337.

44. Fujii GH, Morimoto AM, Berson AE, Bolen JB: Transcriptional analysis of the PTEN/MMAC1 pseudogene, psiPTEN. *Oncogene* 1999, 18:1765-1769.

45. Zhang J, Pontoppidan B, Xue J, Rask L, Meijer J: The third myrosinase gene TGG3 in Arabidopsis thaliana is a pseudogene specifically expressed in stamen and petal. *Physiol Plant* 2002, 115:25-34.

46. McCarrey JR, Riggs AD: Determinator-inhibitor pairs as a mechanism for threshold setting in development: a possible function for pseudogenes. *Proc Natl Acad Sci U S A* 1986, 83:679-683.

**47. Korneev SA, Park JH, O'Shea M: Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J Neurosci* 1999, 19:7711-7720.

These authors reported that a pseudogene regulate the functional gene transcripts in a RNAi-like mechanism.

48. Lee JT: Molecular biology: Complicity of gene and pseudogene. *Nature* 2003, 423:26-28.

49. Vargas-Madrazo E, Almagro JC, Lara-Ochoa F: Structural repertoire in VH pseudogenes of immunoglobulins: comparison with human germline genes and human amino acid sequences. *J Mol Biol* 1995, 246:74-81.

50. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, et al.: Genome sequence of Yersinia pestis, the causative agent of plague. *Nature* 2001, 413:523-527.

51. Arabidopsis Genome Initiative: Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 2000, 408:796-815.

**52. Nekrutenko A, Makova KD, Li WH: The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res* 2002, 12:198-202.

These authors explored the possibility of using Ka/Ks ratio to validate gene predictions.

53. Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, Gerstein M: Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res* 2002, 30:2515-2523.

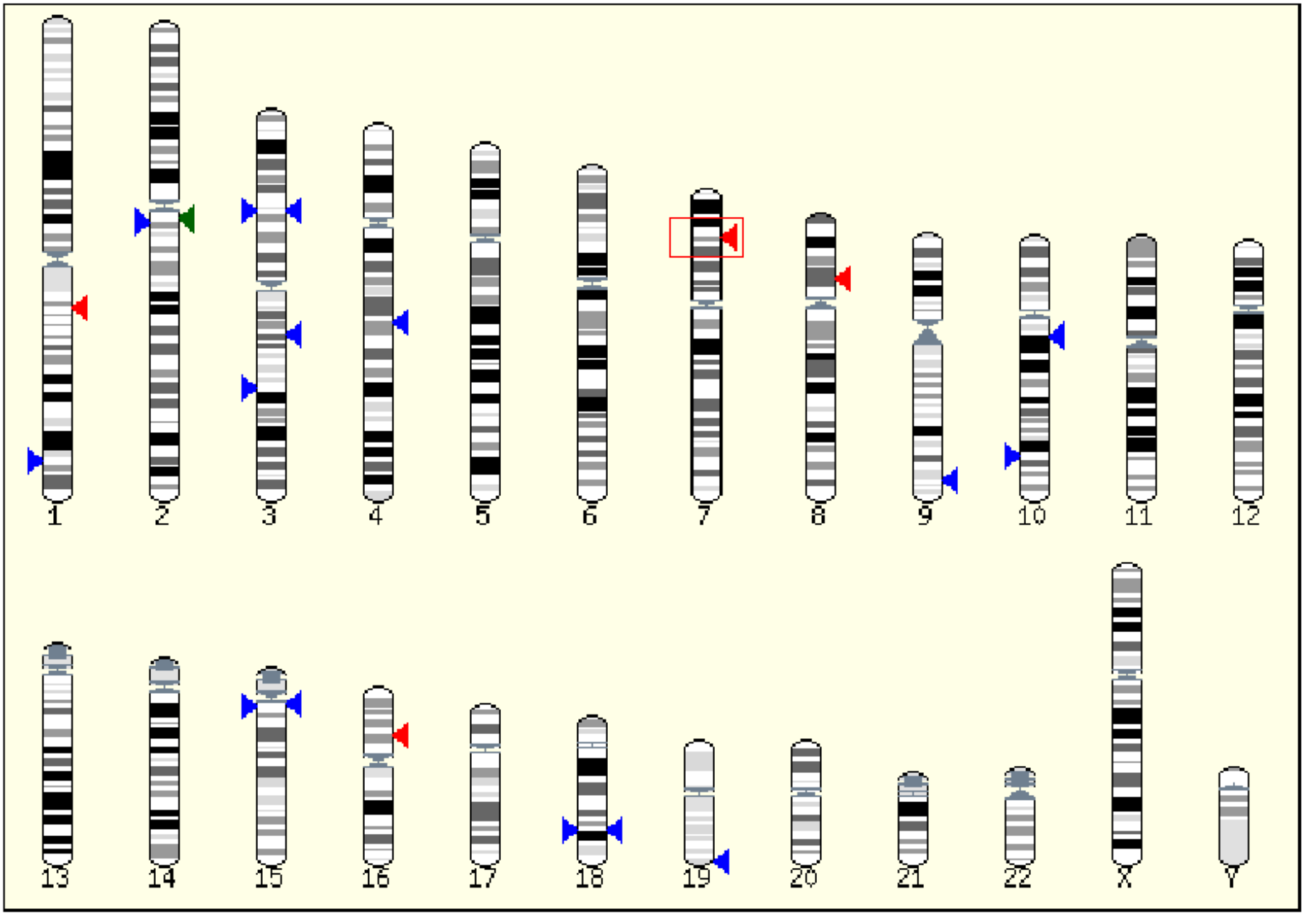
54. Mounsey A, Bauer P, Hope IA: Evidence Suggesting That a Fifth of Annotated Caenorhabditis elegans Genes May Be Pseudogenes. *Genome Res.* 2002, 12:770-775.

55. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, et al.: The transcriptional activity of human Chromosome 22. *Genes Dev* 2003, 17:529-540.

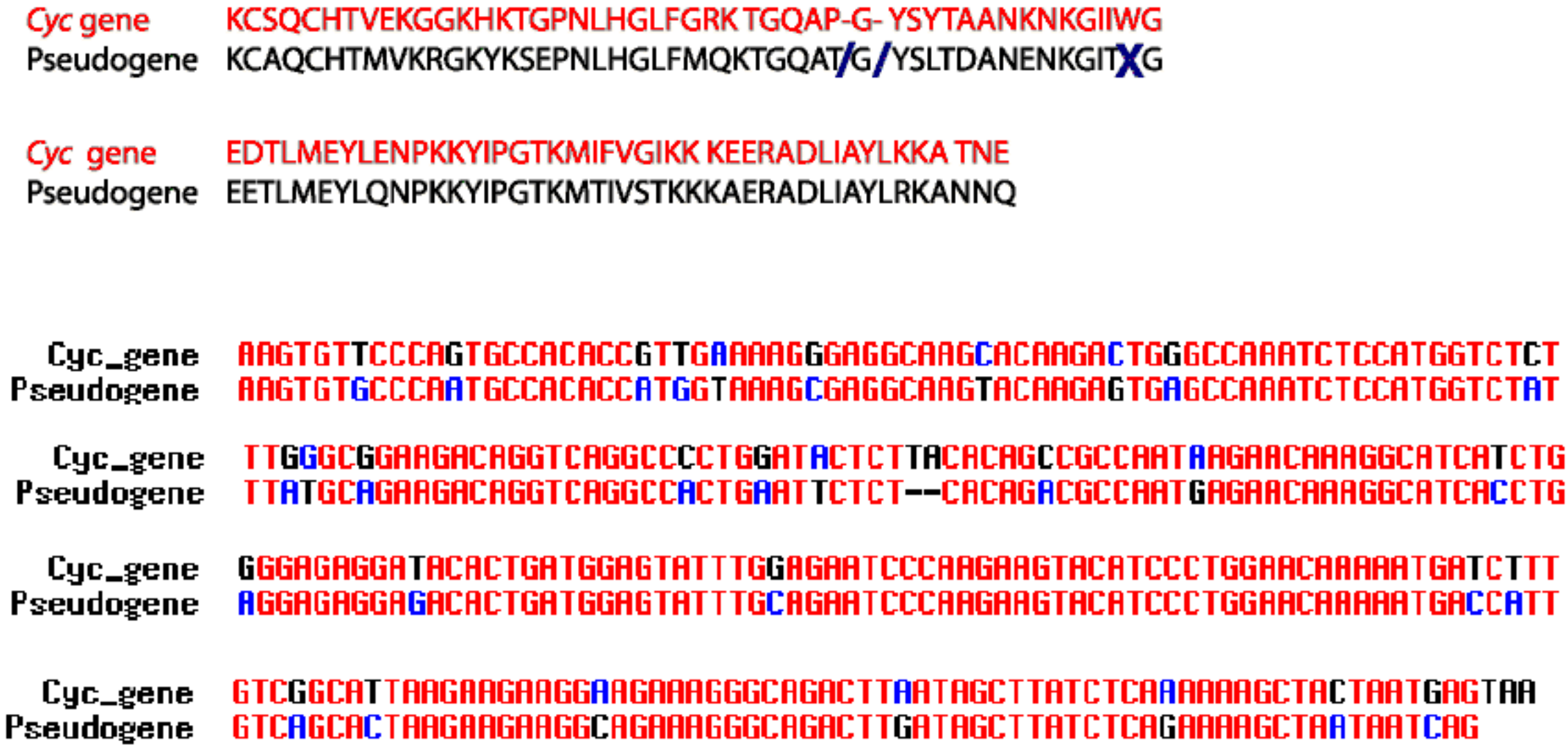
56. Balasubramanian S, Harrison P, Hegyi H, Bertone P, Luscombe N, Echols N, McGarvey P, Zhang Z, Gerstein M: SNPs on human chromosomes 21 and 22 -- analysis in terms of protein features and pseudogenes. *Pharmacogenomics* 2002, 3:393-402.

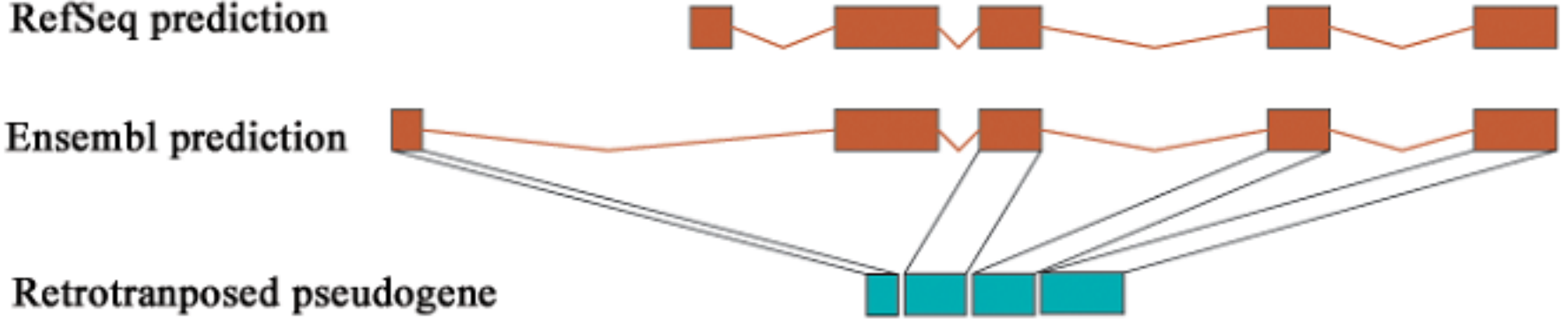
57. Strichman-Almashanu LZ, Bustin M, Landsman D: Retroposed copies of the HMG genes: a window to genome dynamics. *Genome Res* 2003, 13:800-812.

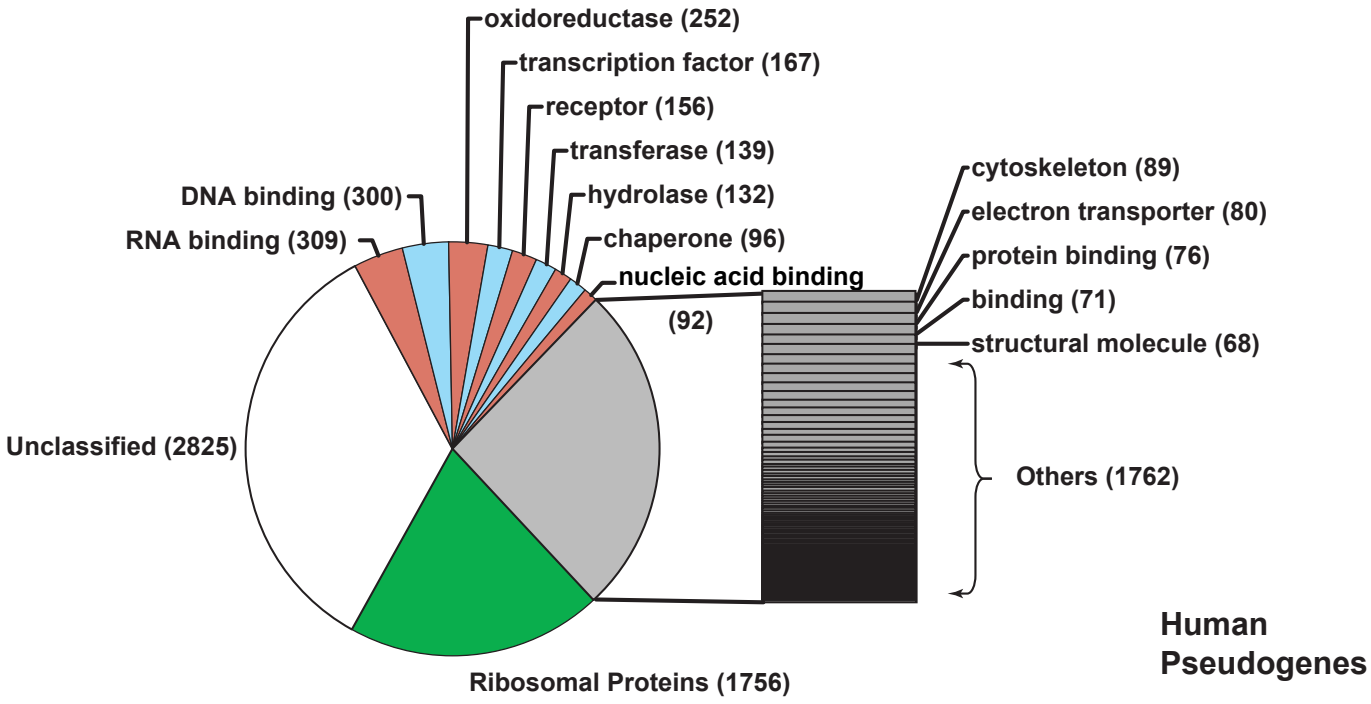
A

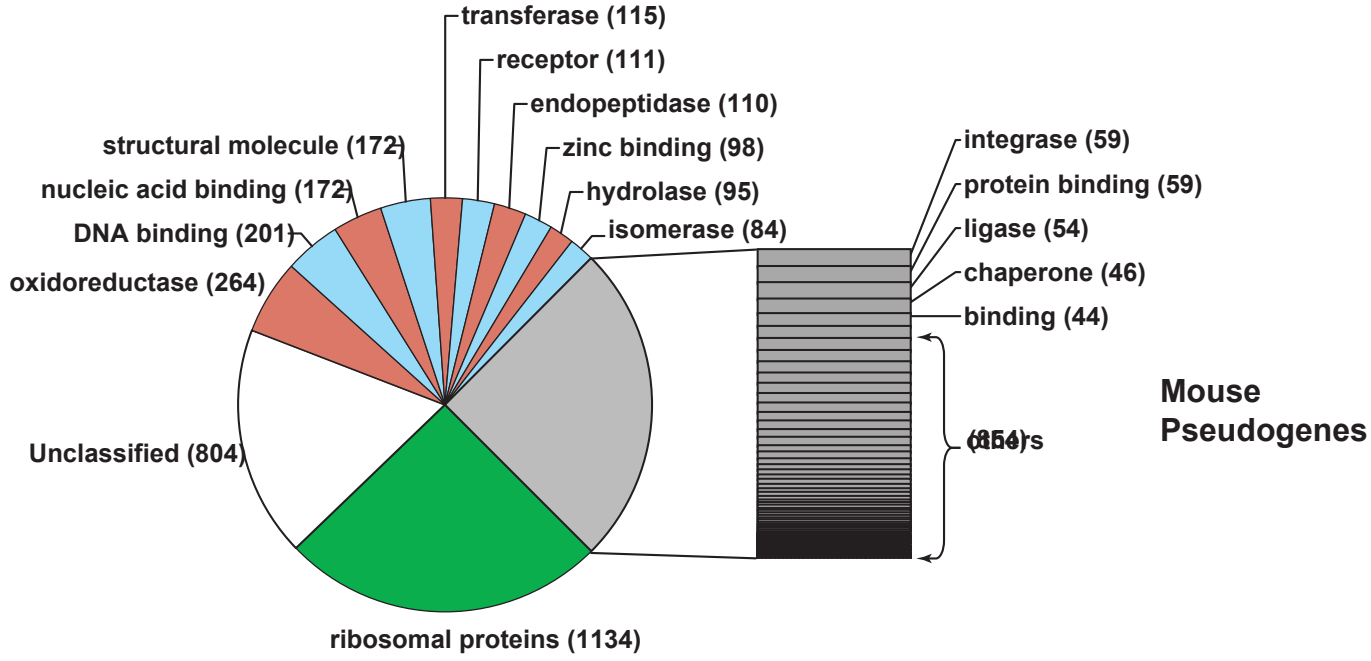


B


Cyc gene KCSQCHTVEKGGKHKTGPNLHGLFGRK TGQAP-G- YSYTAANKNKGIIWG
Pseudogene KCAQCHTMVKRGKYKSEPNLHGLFMQKTGQAT/G/YSLTDANENKGITXG
Cyc gene EDTLMEYLENPKKYIPGTKMIFVGIKK KEERADLIAYLKKA TNE
Pseudogene EETLMEYLQNPKKYIPGTKMTIVSTKKKAERADLIAYLRKANNQ
Cyc_gene AAGTGTTCCCAGTGCCACACCGTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAATCTCCATGGTCTCT
Pseudogene AAGTGTGCCCAATGCCACACCATGGTAAAGCGAGGCAAGTACAAGAGTGAGCCAAATCTCCATGGTCTAT
Cyc_gene TTGGGCGGAAGACAGGTCAGGCCCCTGGATACTCTTACACAGCCGCCAATAAGAACAAAGGCATCATCTG
Pseudogene TTATGCAGAAGACAGGTCAGGCCACTGAATTCTCT--CACAGACGCCAATGAGAACAAAGGCATCACCTG
Cyc_gene GGGAGAGGATACACTGATGGAGTATTTGGAGAATCCCAAGAAGTACATCCCTGGAACAAAAATGATCTTT
Pseudogene AGGAGAGGAGACACTGATGGAGTATTTGCAGAATCCCAAGAAGTACATCCCTGGAACAAAAATGACCATT
Cyc_gene GTCGGCATTAAGAAGAAGGAAGAAAGGGCAGACTTAATAGCTTATCTCAAAAAAGCTACTAATGAGTAA
Pseudogene GTCAGCACTAAGAAGAAGGCAGAAAGGGCAGACTTGATAGCTTATCTCAGAAAAGCTAATAATCAG


RefSeq prediction
Ensembl prediction
Retrotranposed pseudogene


oxidoreductase (252)
transcription factor (167)
receptor (156)
transferase (139)
hydrolase (132)
chaperone (96)
nucleic acid binding (92)
DNA binding (300)
RNA binding (309)
Unclassified (2825)
Ribosomal Proteins (1756)
cytoskeleton (89)
electron transporter (80)
protein binding (76)
binding (71)
structural molecule (68)
Others (1762)
Human Pseudogenes


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)
Mouse Pseudogenes