Neutral evolution of “non-coding” cDNAs from the mouse transcriptome

It has been argued that as many as 15,815 of 33,409 non-redundant mouse cDNAs may represent functional RNA genes[1]. The argument was supported by the fact that some of these cDNAs were confirmed by ESTs and found near CpG islands or polyadenylation signals[2], although many were expressed at such low levels that they could not be detected by microarrays[3]. We show that these “non-coding” cDNAs are no better conserved, in rat or human, than an evolutionarily neutral control. Hence, they are either non-functional or, if they are functional, extremely specific to a given species.

We downloaded FANTOM release 2.0 cDNAs from the authors’ website. Table 1 shows the data in the four categories defined by the authors, which we refer to as coding1 (most likely protein), coding2 (marginal protein), non-coding1 (marginal RNA), and non-coding2 (most likely RNA). Overall transcript sizes average about 2-kb in each category. Most known RNA genes are much smaller than this. Traditional RNA genes, like the 587 mouse entries in *Rfam*[4], average 96-bp in size. Larger RNA genes certainly do exist (*e.g.*, H19 and Xist) and many are stored at the *Erdmann* database[5]. Another striking difference between the given categories is the progression from 13.4% single-exon genes in coding1 to 68.7% and 73.1% single-exon genes in non-coding1 and non-coding2.

As an evolutionarily neutral control, we use “intergenic” sequences of length 2-kb that are at least 5-kb distant from genes annotated by *Ensembl*, predicted by *FgeneSH*, or aligned to cDNAs. Transposons identified by *RepeatMasker* are excluded, as is the 5% of highly conserved mouse sequence that is under purifying selection[6]. Conversely, we have two positive controls. One is the coding1 category of protein coding genes. The other is a set of all known mouse RNA genes. To avoid an overt bias towards small RNA genes, we removed genes smaller than 80-bp in *Rfam*, leaving behind many splicing factors like U1 and U6. We then added all the mouse genes in the *Erdmann* database, which total 40. The resultant set of 321 RNA genes is referred to as “ncRNAs”.

Genome sequences were taken from the UCSC Genome Browser with time stamp 28 June 2003 (rat) and 10 April 2003 (human). *BlastZ*[7] was used for the alignments, with default setting K=3000 and H=2200, and the C=2 option enabled to chain exons together. Although the complexities of the chaining procedure may prevent a few multi-exon genes from aligning, this should not be a problem for non-coding cDNAs, since most are single-exon. We further require that the fraction of the transcript length that is aligned by *BlastZ* must exceed a predetermined alignment threshold of 25%. The low threshold ensures that our positive controls almost always pass. Results are shown in Figure 1.

The crucial observation is that the distributions of sequence identity and insertion-deletion (indel) rate are remarkably similar for non-coding1, non-coding2, and intergenic. Even the widths of the distributions, a reflection of the stochastic nature of the underlying evolutionary process, are highly similar. Best conserved are coding1 and ncRNAs. Worst conserved are non-coding1, non-coding2, and intergenic. The bigger effect is observed in mouse-to-human, because it represents 75 million years of divergence, versus only 14~24 million years in mouse-to-rat. For the latter comparison, the shift (δ) is small compared to the width (σ); but it is significant, as it is a shift in an entire distribution, and the oft-cited rule $\delta >> \sigma$ applies to a point sampled from a distribution.

The simplest explanation is that non-functional transcripts can be produced at low copy numbers, escape the cell’s mRNA surveillance system, and yet inflict no damage to the cell. Table 1 highlights two theories. If these are processed pseudogenes, there should be residual similarity to known proteins, especially mouse proteins. Setting to E-values of 10^{-2}, we find that 36.5% and 19.0% of non-coding1 and non-coding2 are similar to mouse coding1. Just 15.7% and 2.4% are similar to SwissProt, because SwissProt does not store translated cDNAs. If random genomic sequence is transcribed, we should find transposon remnants (ignoring SINEs because they are derived from tRNAs). This is indeed the case for 48.4% and 46.4% of non-coding1 and non-coding2. Note too that the ncRNAs control set is mostly negative for pseudogenes and random genomic sequence.

Given that all of the best techniques for detecting RNA genes depend on sequence conservation[8,9], the absence of such cannot be summarily dismissed, even if one can find isolated examples of RNA genes being weakly conserved[10]. Extraordinary claims require extraordinary proof. This is even more true when much of the data supports an alternative interpretation that they are simply non-functional cDNAs.

Authors

Jun Wang[1,2], Jianguo Zhang[2], Hongkun Zheng[2], Jun Li[2], Dongyuan Liu[2], Heng Li[2], Ram Samudrala[3], Jun Yu[1,2], and Gane Ka-Shu Wong[1,2,4].

[1]*James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou Genomics Institute, Hangzhou 310007, China.*

[2]*Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing 101300, China.*

[3]*University of Washington, Computation Genomics Group, Dept. of Microbiology, Seattle, WA 98195, USA.*

[4]*University of Washington Genome Center, Dept. of Medicine, Seattle, WA 98195, USA.*

E-MAIL gksw@genomics.org.cn

References

1. Okazaki, Y., et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563-573 (2002). http://www.gsc.riken.go.jp/e/FANTOM.

2. Numata, K., et al. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res.* **13**, 1301-1306 (2003).

3. Bono, H., et al. Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarray. *Genome Res.* **13**, 1318-1323 (2003).

4. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S.R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439-441 (2003). http://rfam.wustl.edu.

5. Szymanski, M., Erdmann, V.A. & Barciszewski, J. Noncoding regulatory RNAs database. *Nucleic Acids Res.* **31**, 429-431 (2003). http://biobases.ibch.poznan.pl/ncRNA.

6. Waterston, R.H., et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002). http://hgarchive.cse.ucsc.edu/goldenPath/28jun2002/vsMm2/axtTight.

7. Schwartz, S., et al. Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103-107 (2003).

8. Eddy, S.R. Computational genomics of noncoding RNA genes. *Cell* **109**, 137-140 (2002).

9. Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B. & Bartel, D,P. Vertebrate microRNA genes. *Science* **299**, 1540-1540 (2003).

10. Nesterova, T.B., et al. Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res.* **11**, 833-849 (2001).

Legends

Figure 1: Comparisons to rat (a,c,e) and human (b,d,f). Two panels show the number of good alignments (a,b). Four others show the distribution of sequence identities (c,d) and insertion-deletion rates (e,f), restricted to the good alignments. Each solid dot shows the center of the bin over which we signal average. Coloring is red (coding1), blue (coding2), black (non-coding1), green (non-coding2), brown (ncRNAs), and yellow (intergenic). For panels c to f, we add a purple line for the CDS region of coding1.

Table 1: Other attributes of mouse cDNAs. After computing best ORFs, leftover flanking sequences are taken to be untranslated regions. Sizes are described by mean (standard deviation). In the *RepeatMasker* tallies, we do not count SINEs.

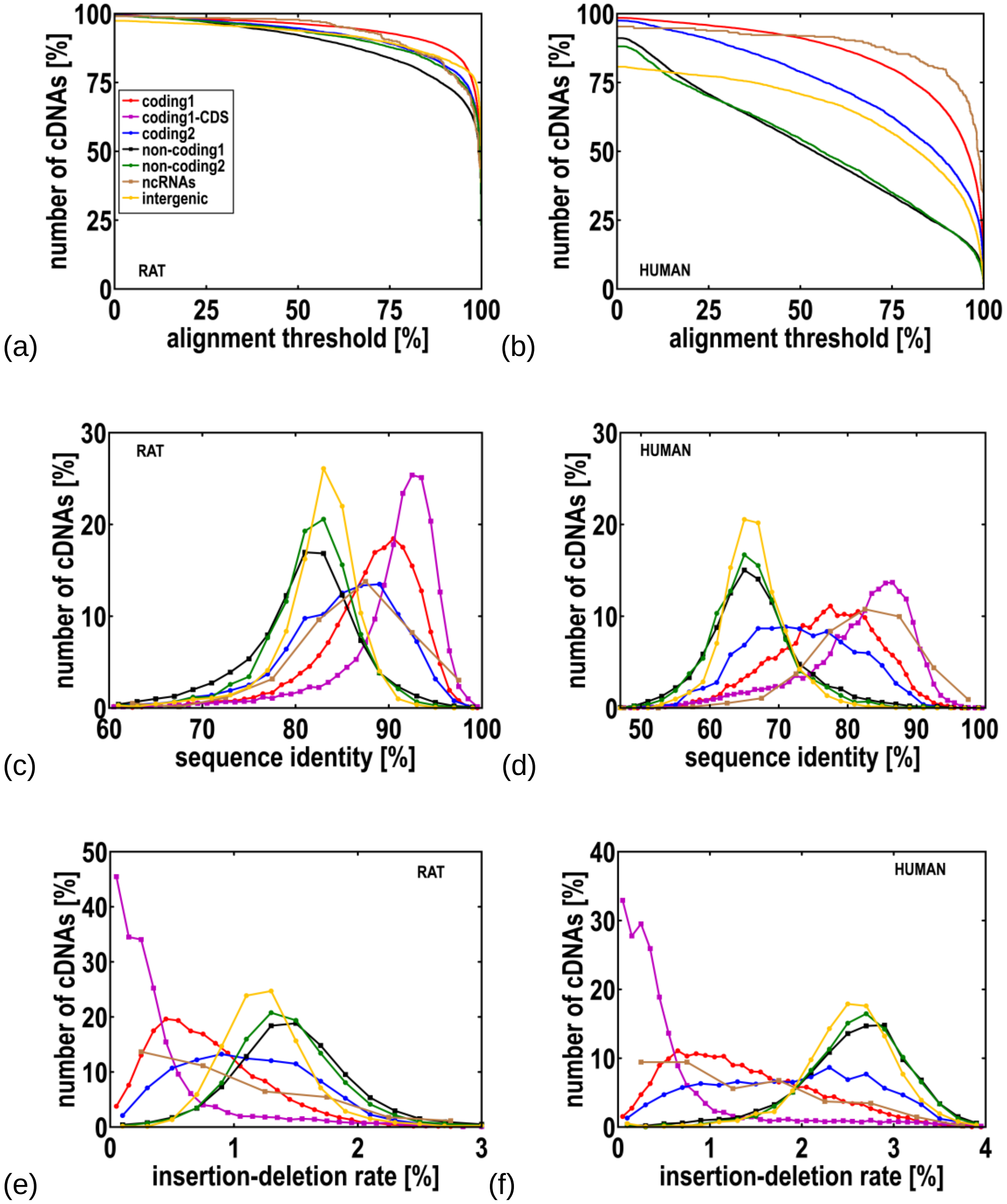
(a)
RAT
number of cDNAs [%]
alignment threshold [%]
0
25
50
75
100
coding1
coding1-CDS
coding2
non-coding1
non-coding2
ncRNAs
intergenic
(b)
HUMAN
number of cDNAs [%]
alignment threshold [%]
(c)
RAT
number of cDNAs [%]
sequence identity [%]
60
70
80
90
100
0
10
20
30
(d)
HUMAN
number of cDNAs [%]
sequence identity [%]
50
60
70
80
90
100
(e)
RAT
number of cDNAs [%]
insertion-deletion rate [%]
0
1
2
3
10
20
30
40
50
(f)
HUMAN
number of cDNAs [%]
insertion-deletion rate [%]
0
1
2
3
4
10
20
30
40

Figure 1.

Table 1.

	FANTOM categories				control data sets	
	coding1	**coding2**	**non-coding1**	**non-coding2**	**ncRNAs**	**intergenic**
number of cDNAs	14,317	3,277	11,526	4,280	321	3,450
# in a single exon	13.4%	35.4%	68.7%	73.1%	90.7%	100%
size of FL cDNA	2146 (1061)	2174 (1091)	1939 (1019)	1790 (996)	325 (1055)	2000 (0)
size of 5' UTR	242 (335)	640 (686)	842 (754)	791 (727)	N/A	889 (523)
size of best ORF	1107 (742)	550 (578)	206 (91)	194 (80)	N/A	213 (88)
size of 3' UTR	836 (746)	983 (807)	891 (770)	805 (718)	N/A	898 (524)
***BlastX* proteins**						
E-value = 1E-2						
SwissProt	72.4%	55.5%	15.7%	2.4%	0.9%	2.9%
mouse coding1	100.0%	59.3%	36.5%	19.0%	4.4%	3.7%
combined	100.0%	68.0%	37.6%	19.5%	4.4%	4.4%
E-value = 1E-4						
SwissProt	68.8%	50.4%	11.1%	0.8%	0.0%	2.0%
mouse coding1	100.0%	53.0%	31.0%	12.6%	3.7%	2.5%
combined	100.0%	62.9%	31.9%	12.8%	3.7%	3.0%
E-value = 1E-6						
SwissProt	65.3%	45.5%	6.1%	0.0%	0.0%	1.6%
mouse coding1	100.0%	47.5%	25.4%	7.7%	2.5%	1.8%
combined	100.0%	58.2%	26.2%	7.7%	2.5%	2.2%
RepeatMasker	13.7%	27.7%	48.4%	46.4%	3.4%	0.0%