Mouse transcriptome

Neutral evolution of 'non-coding' complementary DNAs

Arising from: Y. Okazaki *et al. Nature* **420,** 563–573 (2002)

Okazaki *et al.* have argued that as many as 15,815 of 33,409 non-redundant mouse complementary DNAs may represent functional RNA genes[1], on the basis of their findings that some of these cDNAs are confirmed by expressed sequence tagging and are found near CpG islands or polyadenylation signals[2] — although many are expressed at such low levels that they could not be detected by microarray analysis[3]. We show here that conservation of these 'non-coding' cDNAs in rats or humans is no better than in an evolutionarily neutral control. Our results indicate that they are either non-functional or, if they are functional, are specific to a given species.

We downloaded FANTOM release 2.0 cDNAs from the authors' website. Table 1 shows the data from the four categories defined by the authors, which we refer to as coding 1 (probably protein), coding 2 (marginal protein), non-coding 1 (marginal RNA), and non-coding 2 (probably RNA). Overall transcript sizes average about 2 kilobases (kb) in each category; most known RNA genes are much smaller than this — for example, the 587 mouse entries in the Rfam database[4] average 96 base pairs (bp) in length. Larger RNA genes do exist (such as *H19* and *Xist*) and many are stored in the Erdmann database[5]. Another striking difference between the given categories is the increase from 13.4% single-exon genes in coding 1 to 68.7% and 73.1% single-exon genes in non-coding 1 and non-coding 2, respectively.

As an evolutionarily neutral control, we use 'intergenic' sequences of 2 kb in length 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 coding 1 category of protein-coding genes and 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 encoding splicing factors such as *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 (ref. 7) was used for the alignments, with default settings $K = 3{,}000$ and $H = 2{,}200$. The $C = 2$ option enabled us to chain exons together. Although the complexities of the chaining procedure may prevent a few multi-exon genes from aligning, this

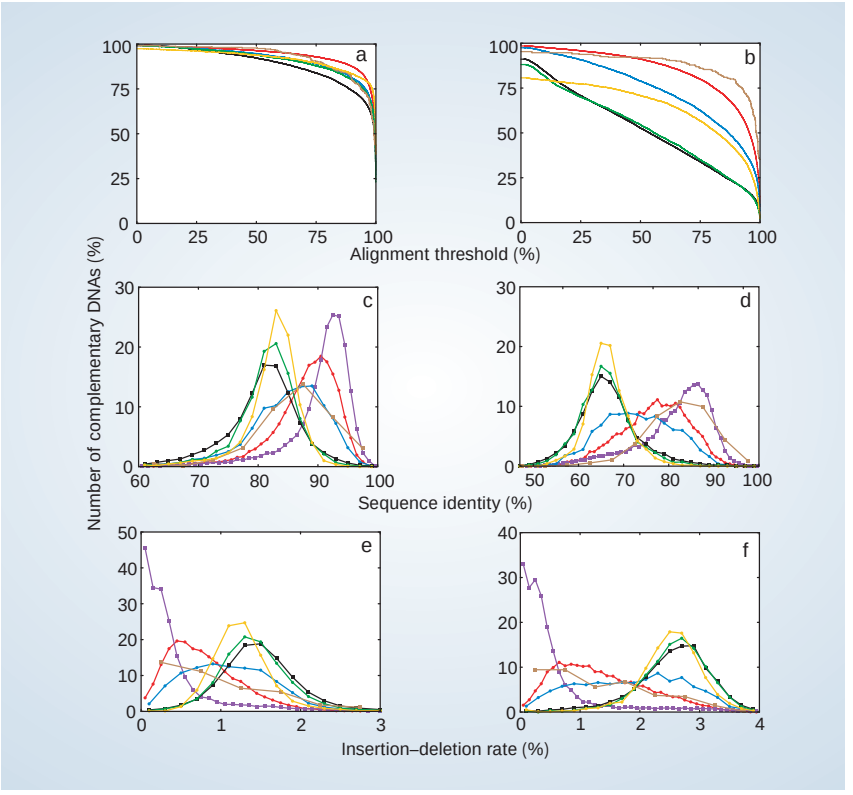

Figure 1 Comparisons between rat (left) and human (right) data. **a, b,** The number of good alignments. **c–f,** Distribution of sequence identities (**c,d**) and insertion–deletion rates (**e,f**) restricted to the good alignments Each solid dot shows the centre of the bin over which signals were averaged. Red, coding 1; blue, coding 2; black, non-coding 1; green, non-coding 2; brown, ncRNAs; and yellow, intergenic. For panels **c** to **f**, a purple line is added for the CDS region of coding 1.

Table 1 **Other attributes of mouse complementary DNAs**

	FANTOM categories				*Control data sets*	
	Coding 1	*Coding 2*	*Non-coding 1*	*Non-coding 2*	*ncRNAs*	*Intergenic*
No. of cDNAs	14,317	3,277	11,526	4,280	321	3,450
No. in a single exon	13.4%	35.4%	68.7%	73.1%	90.7%	100%
Size of FL cDNA	2,146 (1,061)	2,174 (1,091)	1,939 (1,019)	1,790 (996)	325 (1,055)	2,000 (0)
Size of 5′ UTR	242 (335)	640 (686)	842 (754)	791 (727)	NA	889 (523)
Size of best ORF	1,107 (742)	550 (578)	206 (91)	194 (80)	NA	213 (88)
Size of 3′ UTR	836 (746)	983 (807)	891 (770)	805 (718)	NA	898 (524)
BlastX proteins						
E-value = 10^{-2}						
SwissProt	72.4%	55.5%	15.7%	2.4%	0.9%	2.9%
Mouse coding 1	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 = 10^{-4}						
SwissProt	68.8%	50.4%	11.1%	0.8%	0.0%	2.0%
Mouse coding 1	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 = 10^{-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%

After computing the best open-reading frames (ORFs), left-over flanking sequences are taken to be untranslated regions. Sizes (in base pairs) are described as mean (standard deviation). In the RepeatMasker tallies, we do not count short interspersed elements.
UTR, untranslated terminal repeat; NA, not applicable; FL, full-length.

should not be a problem for non-coding cDNAs as most are single-exon. We specified that the fraction of transcript length that is aligned by BlastZ must exceed a predetermined alignment threshold of 25%: this low threshold ensures that our positive controls almost always pass (Fig. 1).

The crucial observation is that the distributions of sequence identity and insertion–deletion ('indel') rate are remarkably similar for non-coding 1, non-coding 2 and intergenic. Even the widths of the distributions, a reflection of the stochastic nature of the underlying evolutionary process, are highly similar. The most well conserved are coding 1 and ncRNAs, and the least well conserved are non-coding 1, non-coding 2 and intergenic. The larger effect is observed in mouse-to-human, because it represents 75 million years of divergence, compared with only 14–24 million years in mouse-to-rat. For the latter comparison, the shift (δ) is small compared with the width (σ); however, it is significant, as it is a shift in an entire distribution, and the oft-cited rule $\delta \gg \sigma$ 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 messenger RNA surveillance system, and yet inflict no damage on 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-coding 1 and non-coding 2 are similar to mouse coding 1. 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 short interspersed elements because they are derived from transfer RNAs). This is indeed the case for 48.4% and 46.4% of non-coding 1 and non-coding 2. 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 this cannot be summarily dismissed, even if isolated examples of RNA genes being weakly conserved can be found[10]. Extraordinary claims require extraordinary proof — this is particularly true when much of the data support an alternative interpretation that they are simply non-functional cDNAs.

Jun Wang*†, Jianguo Zhang†, Hongkun Zheng†, Jun Li†, Dongyuan Liu†, Heng Li†, Ram Samudrala‡, Jun Yu*†, Gane Ka-Shu Wong*†§

**James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou Genomics Institute, Hangzhou 310007, China*
e-mail: gksw@genomics.org.cn
†Beijing Institute of Genomics of the Chinese Academy of Sciences, Beijing 101300, China
‡Computation Genomics Group, Department of Microbiology, University of Washington, Seattle, Washington 98195, USA
§University of Washington Genome Center, Department of Medicine, Seattle, Washington 98195, USA
doi:10.1038/nature03016

1. Okazaki, Y. *et al. Nature* **420,** 563–573 (2002).
2. Numata, K. *et al. Genome Res.* **13,** 1301–1306 (2003).
3. Bono, H. *et al. Genome Res.* **13,** 1318–1323 (2003).
4. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. *Nucleic Acids Res.* **31,** 439–441 (2003).
5. Szymanski, M., Erdmann, V. A. & Barciszewski, J. *Nucleic Acids Res.* **31,** 429–431 (2003).
6. Waterston, R. H. *et al. Nature* **420,** 520–562 (2002).
7. Schwartz, S. *et al. Genome Res.* **13,** 103–107 (2003).
8. Eddy, S. R. *Cell* **109,** 137–140 (2002).
9. Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. *Science* **299,** 1540 (2003).
10. Nesterova, T. B. *et al. Genome Res.* **11,** 833–849 (2001).

Okazaki et al. *reply* — Wang *et al.*[1] challenge our suggestion that almost half of the 33,407 representative putative full-length cDNAs identified by RIKEN are probably non-coding RNAs[2]. The challenge is based on bioinformatic analyses showing that the level of conservation of these sequences is no greater than that observed for intergenic sequences, and less than that observed in a control set of documented ncRNAs.

The analyses of Wang *et al.* have several problems. Their positive control set is heavily biased towards structural and catalytic RNAs. It contains only 19 non-redundant regulatory ncRNAs from the Erdmann database[3], with the majority being drawn from the Rfam RNA database[4], which — despite the 80-nucleotide cutoff used to remove very small RNAs — contains mainly infrastructural RNAs, such as spliceosomal and small nucleolar RNAs, many of which are covariance-model structure-based predictions from large genomic data sets. The large dichotomy in the control set is shown by their average size of 325 nucleotides, with standard deviation of 1,055 nucleotides[1].

There is ample evidence that known regulatory ncRNAs are, in the main, much less conserved than protein-coding sequences. For example, the 110-kb murine ncRNA Air, which is involved in imprinting of the *Igf2r* locus (and is represented in the RIKEN database), has no equivalent transcript in humans[5], and shows no significant homology with either the human or rat genomes. The rat ncRNA Bsr does not occur in human or mouse. The human ncRNA DISC2 is not conserved in mouse or Fugu, despite the fact that its surrounding transcripts (including its putative antisense target DISC1) are conserved across mammals and Fugu[6]. Xist shows low homology (60%) between mammalian species, despite its identical function in X-chromosome inactivation[7]. The antisense transcript Tsix, also important to this process, is poorly conserved between species[7]. Indeed, alignment of known regulatory ncRNAs among mammalian species shows that there is large divergence in the percentage of alignable sequences and that their overall homology is low and indistinguishable from intergenic sequences.

The validity of 'intergenic' sequences as a negative control is also questionable. Although it is thought that only 5% of the mammalian genome is under purifying selection, this figure was obtained from comparison of the mouse and human genomes, compared with sequences assumed to be evolving neutrally[8]. However, multivariate analyses of the *CFTR* (ref. 9) and *SIM2* (ref. 10) loci showed that intergenic and intronic sequences exhibit patterns of conservation that are not evident from pairwise comparisons alone. That is, the sequences exhibiting conservation depend on which species are being compared, implying that more of the genome is under evolutionary selection (both positive and negative) than has been appreciated, with many non-coding sequences, presumably regulatory, being differently conserved in lineage-specific ways.

The possible sources of contamination of cDNA libraries are genomic and pre-mRNA sequences, and 'transcriptional noise'. Our analyses have shown that there is insignificant genomic contamination in the RNA preparations used to construct cDNA libraries, and also that most putative ncRNA sequences are not derived from introns of protein-coding genes. The concept of transcriptional noise derives from studies of stochastic transcription[11], but this does not mean that such transcription occurs from illegitimate promoters, nor is there any evidence for this. Our published[12] and unpublished analyses show that many of the putative ncRNA transcripts exhibit both tissue-specific and dynamic regulation of their expression in relation to external cues. Our findings agree with independent analyses using genomic arrays, which conclude that the human genome contains comparable numbers of protein-coding and non-coding genes that are under the control of common transcription factors and environmental signals[13,14]. They also agree with molecular genetic analyses of well studied loci, which invariably show that at least half of documented transcripts are non-coding[15].

(This response was prepared with input from Shintaro Katayama, Harukazu Suzuki,