Supplementary material

Is abundant A-to-I RNA editing primate-specific?

Eli Eisenberg[1,2], Sergey Nemzer[1], Yaron Kinar[1], Rotem Sorek[1], Gideon Rechavi[3] and Erez Y. Levanon[1,3]

[1]Compugen Ltd, 72 Pinchas Rosen St, Tel-Aviv 69512, Israel
[2]School of Physics and Astronomy, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[3]Department of Pediatric Hematology-Oncology, Chaim Sheba Medical Center and Sackler School of Medicine, Tel Aviv University, Tel Aviv 52621, Israel

Calculating the estimated accuracy

The overrepresentation of A-to-G mismatches in the various analyses conducted was considered a signature of editing. To obtain an estimate for the number of mismatches resulting from sources other than editing, we used the count of G-to-A mismatches, which are at least as common as A-to-G mismatches in sequencing errors, mutations and single nucleotide polymorphisms (SNPs), the most significant sources of mismatches. Throughout the article, the accuracy of a set of editing sites is defined as the percentage of the sites or sequences that were found in excess of the background level. The background level is estimated by the number of sites or sequences found when replacing A-to-G by G-to-A. This estimation of the accuracy level was experimentally confirmed [1].

Comparison of editing levels in specific tissues

Editing levels vary between different tissues [1,2]. Thus, the differences between the editing level observed in human and mouse could have resulted from differences in the tissue distributions of the available human and mouse RNA sequences. To rule out this possibility, we repeated the human–mouse comparison for RNA sequences of homogeneous tissue origin. We used three different tissues that have a significant and similar number of sequences for both organisms: brain, thymus and testis. We found that for all three tissues the level of editing in human is significant (at least 3% of sequences are edited), whereas in mouse editing is undetectable for such small data sets (Tables 1–3). Notably, the editing level in RNAs that originate from the human thymus is exceptionally high: ~17% of sequences (1bp per 1000bp) are edited.

Table 1. The number of mismatches in brain RNAs[a]

	Human (5398 sequences)			Mouse (1242 sequences)		
Mismatch	**Number of consecutive mismatch**			**Number of consecutive mismatch**		
	1	3	5	1	3	5
AG	6469 (2298)	974 (391)	431 (219)	756 (353)	19 (19)	2 (2)
GA	2485 (1698)	38 (37)	1 (1)	707 (321)	19 (16)	1 (1)
CT	2652 (1806)	43 (37)	5 (5)	792 (366)	21 (21)	3 (3)
TC	3274 (2108)	74 (66)	5 (4)	768 (334)	17 (16)	1 (1)
Percentage of A-to-G mismatches	33.8%	85.1%	97.1%	16.2%	20.4%	25.0%

[a]Number of single (or stretches of consecutive) mismatches in RNAs of a specific tissue origin, for the most common mismatches. The numbers in parentheses are the number of distinct RNA sequences in which the associated instances occur. The last row presents the percentage of A-to-G instances among the total number of all 12 mismatches.

Table 2. The number of mismatches in thymus RNAs[a]

	Human (1090 sequences)			Mouse (3746 sequences)		
Mismatch	**Number of consecutive mismatch**			**Number of consecutive mismatch**		
	1	3	5	1	3	5
AG	3036(670)	633(193)	304(134)	779(554)	28(18)	8(5)
GA	573(374)	9(9)	0(0)	783(575)	10(10)	1(1)
CT	601(394)	7(7)	0(0)	561(371)	7(7)	0(0)
TC	810(511)	18(11)	5(1)	455(310)	8(8)	0(0)
Percentage of A-to-G mismatches	50.5%	94.5%	98.1%	14.4%	26.9%	72.7%

[a]Number of single (or stretches of consecutive) mismatches in RNAs of a specific tissue origin, for the most common mismatches. The numbers in parentheses are the number of distinct RNA sequences in which the associated instances occur. The last row presents the percentage of A-to-G instances among the total number of all 12 mismatches.

Table 3. The number of mismatches in testis RNAs[a]

	Human (6217 sequences)			Mouse (6881 sequences)		
Mismatch	**Number of consecutive mismatch**			**Number of consecutive mismatch**		
	1	3	5	1	3	5
AG	5142 (2674)	394 (209)	133 (64)	1717 (1180)	32 (28)	2 (2)
GA	2769 (1918)	39 (37)	4 (3)	1707 (1079)	49 (42)	8 (7)
CT	2674 (1894)	27 (26)	0 (0)	1699 (1040)	52 (46)	6 (5)
TC	3037 (2024)	44 (43)	0 (0)	1148 (729)	30 (27)	3 (3)
Percentage of A-to-G mismatches	27.6%	75.9%	95.7%	12.6%	11.7%	5.9%

[a]Number of single (or stretches of consecutive) mismatches in RNAs of a specific tissue origin, for the most common mismatches. The numbers in parentheses are the number of distinct RNA sequences in which the associated instances occur. The last row presents the percentage of A-to-G instances among the total number of all 12 mismatches.

Alu and the number of potential dsRNAs

The vast majority of A-to-I editing detectable using our algorithm in human occurs within *Alu* elements, which are the most abundant short interspersed elements (SINEs) in primates. *Alu* elements tend to accumulate within genes [3], and are present in ~75% of all human genes [4]. Thus, *Alu* repeats occurring within the RNA transcript facilitates the formation of the dsRNA substrates required for ADARs action by pairing with other *Alu* repeats within the pre-mRNA.

It is surprising to find such a substantial difference in global A-to-I editing patterns between human and rodents, because the number of SINEs in the human genome is similar to the total number of rodent SINEs [5]. One major reason for the difference is the fact that only one SINE is dominant in human, making a dsRNA formation out of two consecutive and oppositely oriented SINEs more probable. Furthermore, the dsRNAs formed in human are longer (thus contain more adenosines to be edited) because *Alu* is longer than the equivalent rodent B1. We tested whether this effect alone can explain the ~20 fold difference in the number of editing sites. For this purpose, we extracted all human ESTs and cDNAs and aligned them to the genome (details of this procedure are given in Sorek *et al.* [6]). Following the protocol of Levanon *et al.* [1], we aligned the expressed part of the gene with the corresponding genomic region, looking for reverse complement alignments longer than 32nt with identity levels >85%. We found 429 000 such potential dsRNAs structures, covering 4.69Mb. In comparison, the number of such potential dsRNAs structures in the mouse genome was 81 000, covering 969kb. Thus, the number of potential dsRNA stems in human is only fivefold greater than that of mouse, and similarly, the total size of expressed DNA regions potentially creating such stems is only five times larger, compared with the 20-to-40-fold increase in editing observed in

humans. It is thus possible that the *Alu* repeat, in addition to being more abundant than each of the mouse repeats, is also preferentially targeted by ADARs. In fact, we have previously reported that *Alu* elements contain 'hotspots' for editing [1].

A comparison of editing levels in sequences of a single source

Another potential bias that could have led to the differences between the editing levels observed in human and mouse might result from the different sources used for RNA purification and sequencing in different research centres. To rule out this possibility, we repeated the above analysis for RNA sequences of homogeneous origin. We scanned >22 000 full-length human and mouse cDNA sequences coming from the Mammalian Gene Collection (MGC) Program [7], and repeated the analysis for these sequences only. Comparing the number of sequences with three consecutive A-to-G mismatches to the corresponding number with G-to-A mismatches, we found that ~180 human sequences were edited. By contrast, editing is undetectable in the mouse data.

Library distribution of edited RNAs

To make sure the overrepresentation of edited RNAs in human is not a result of a small number of faulty libraries, we have studied the library assignment of human RNAs that exhibit editing. In particular, we studied the 2513 RNA sequences that show five consecutive A-to-G mismatches. Library assignment is available for 1832 of these sequences. We find that they come from 207 different libraries. In addition, we found that at least one sequence with five or more consecutive A-to-G mismatches is present in 170 out of the 230 libraries (74%) containing at least 50 RNA sequences.

Materials and methods

Employing the algorithm of Ref. [8], human and mouse ESTs and cDNAs were obtained from NCBI GenBank version 136 (June 2003; www.ncbi.nlm.nih.gov/dbEST). The genomic sequences were taken from the human genome build 33, June 2003, and mouse genome build 32, November 2003 (data can be downloaded from www.ncbi.nlm.nih.gov/genome/guide/human).

Sequences were cleaned from terminal vector sequences, and low-complexity stretches and repeats (including *Alu* repeats) in the expressed sequences were masked. Then, expressed sequences were compared with the genome to find likely high-quality hits. They were then aligned to the genome by use of a spliced alignment model that allows long gaps. Only sequences that had >94% similarity to a stretch in the genome were used in further stages. For further details, see Sorek *at al.* [6].

In addition, RNA sequences and their pair-wise alignments to the genome were downloaded from UCSC genome browser site http://genome.ucsc.edu (human assembly July 2003, mouse assembly October 2003, rat assembly June 2003, chicken assembly February 2004 and fly assembly January 2003). We retained only sequences with a unique alignment to the genome, and then recorded all mismatches.

References

1 Levanon, E.Y. *et al.* (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol* 22, 1001–1005

2 Paul, M.S. and Bass, B.L. (1998) Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *Embo J* 17, 1120–1127

3 Medstrand, P. *et al.* (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* 12), 1483–1495

4 Grover, D. *et al.* (2004) Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. *Bioinformatics* 20, 813–817

5 Gibbs, R.A. *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493–521

6 Sorek, R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res* 12, 1060–1067

7 Strausberg, R.L. *et al.* (2002) Generation and initial analysis of more than 15 000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci U S A* 99, 16899–16903

8 Morse, D.P. *et al.* (2002) RNA hairpins in noncoding regions of human brain and *Caenorhabditis elegans* mRNA are edited by adenosine deaminases that act on RNA. *Proc Natl Acad Sci U S A* 99, 7906–7911