Widespread RNA Editing of Embedded Alu Elements in the Human Transcriptome

Dennis D.Y. Kim,1 Thomas T.Y. Kim,2 Thomas Walsh,3 Yoshifumi Kobayashi,1 Tara C. Matise,4 Steven Buyske,4,5 and Abram Gabriel1,6

1Department of Molecular Biology and Biochemistry, 2Department of Computer Science, 3Graduate Program in Computer Science, 4Department of Genetics, and 5Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

More than one million copies of the ∼300-bp Alu element are interspersed throughout the human genome, with up to 75% of all known genes having Alu insertions within their introns and/or UTRs. Transcribed Alu sequences can alter splicing patterns by generating new exons, but other impacts of intragenic Alu elements on their host RNA are largely unexplored. Recently, repeat elements present in the introns or 3′-UTRs of 15 human brain RNAs have been shown to be targets for multiple adenosine to inosine (A-to-I) editing. Using a statistical approach, we find that editing of transcripts with embedded Alu sequences is a global phenomenon in the human transcriptome, observed in 2674 (∼2%) of all publicly available full-length human cDNAs (n = 126,406), from >250 libraries and >30 tissue sources. In the vast majority of edited RNAs, A-to-I substitutions are clustered within transcribed sense or antisense Alu sequences. Edited bases are primarily associated with retained introns, extended UTRs, or with transcripts that have no corresponding known gene. Therefore, Alu-associated RNA editing may be a mechanism for marking nonstandard transcripts, not destined for translation.

[Supplemental material is available online at www.genome.org.]
prevailing RNA editing of embedded Alu sequences in the human transcriptome and suggest novel impacts of intragenic Alus on their host genes at the RNA level.

RESULTS AND DISCUSSION

An Excess of A-to-G Substitutions in Human Full-Length cDNAs

To look specifically for multiple-edited targets, we compared the sequences of full-length cDNAs to the reference genome sequence, and noted clusters of A-to-G substitutions. Expressed sequence tags (ESTs) were not used for this study because of their variable sequence quality. In contrast, full-length cDNAs, generated by a method that captures both the 5’-UTR and 3’-poly(A) stretch, represent complete copies of mRNAs and are >99.9% accurate (Ota et al. 2004). Therefore, the major expected source of sequence variation would be single nucleotide polymorphisms (SNPs). To maximize the distinction between SNPs and edited bases, we used a scan statistic method optimized to detect clusters of A-to-G substitutions within one or more region of a given transcript. When each type of substitution (12 overall types) was tabulated and summed for all 128,406 full-length cDNAs, we found an overwhelming excess of A-to-G base changes (n = 62,118; Fig. 1A). For known SNPs, the substitution rate observed for these two transitions is similar per kilobase (n = 254,330,565 bases). The number of cDNAs with a significantly increased likelihood of specific substitutions, as determined by our scan statistic method (see Methods), is shown in Supplemental Table 1. Almost 88% of all A-to-G substitutions occurred within Alu sequences. About 94% of all A-to-G substitutions (n = 62,118; Fig. 1A) were accumulated in a subset of cDNAs (n = 41,848; Table 1). In comparison, within the 125,732 unselected cDNAs, A-to-G substitutions (n = 74,647) accounted for only 20% of all substitutions (n = 380,985). In the selected transcripts, the number of A-to-G substitutions per transcript ranged from 5 to 107, with a mean of 13.12 and a median of 10, whereas the number of the other 11 substitutions combined per transcript ranged from 0 to 37, with a mean of 2.53 and a median of 2. Thus, the selected cDNAs have a disproportionately high number of A-to-G changes (average of 13.12 per transcript for the selected cDNAs vs. 0.59 for the unselected) but show a similar frequency of base changes compared with unselected transcripts for the 11 other substitutions combined (average of 2.53 per transcript for the selected cDNAs vs. 2.44 for the unselected).

Correlation Between A-to-G Substituted Bases and Alu Sequences

The A-to-G substitutions in the selected transcripts were not randomly distributed but showed a striking correlation with embedded Alu sequences. Almost 88% of all A-to-G substitutions occurred within Alu sequences present in these RNAs, although Alu sequences comprised only 20% of the total length of these transcripts (p < 10^{-15}; Table 1). On the contrary, the other types of substitutions showed a random distribution. Of all substitutions within Alus, A-to-G substitutions accounted for 94% (n = 30,744), whereas all other 11 substitutions contributed only 6% (n = 11,077, p < 10^{-15}). Edited sequences were also observed, at much lower frequencies, in LINE-1 and other interspersed repeats, as well as in nonrepetitive sequences (Table 1). Given the statistical improbability for numerous transcripts to randomly cluster A-to-G substitutions almost exclusively within embedded

![Figure 1](https://www.genome.org/content/1720/6/4559/fig1a.png)

**Figure 1** Histograms of transition substitutions from the reference genomic DNA to full-length cDNA. Only transitions, whose frequencies are generally threefold to fourfold higher than transversions, are shown. (A) Total number of transition substitutions in 128,406 cDNAs (total of 254,330,565 bases). (B) Number of cDNAs with a significantly increased likelihood of specific substitutions, as determined by our scan statistic method (see Methods). There is a population of transcripts with excessive A-to-G substitutions that are not apparent in other transitions. A-to-G substitution: 2674 significant transcripts; G-to-A substitution: 35 significant transcripts; T-to-C substitution: 36 significant transcripts; C-to-T substitution: 52 significant transcripts.

<table>
<thead>
<tr>
<th>Repeat element-type</th>
<th>No. of repeat elements</th>
<th>Total bases (% of total)</th>
<th>A-to-G substitutions (% of total)</th>
<th>Distribution of A-to-G substitutions in repeat elements (minimum, 25th percentile, median, 75th percentile, maximum)</th>
<th>Other 11 substitutions combined (% of total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alu</td>
<td>6763</td>
<td>1,570,185 (20.7%)</td>
<td>30,744 (87.6%)</td>
<td>0.0,0,4,7,34</td>
<td>2017 (26.4%)</td>
</tr>
<tr>
<td>Other repeats*</td>
<td>6840</td>
<td>1,472,934 (19.4%)</td>
<td>1826 (5.2%)</td>
<td>0.0,0,0,63</td>
<td>1396 (18.3%)</td>
</tr>
<tr>
<td>Non-repeats*</td>
<td>—</td>
<td>4,555,872 (59.9%)</td>
<td>2515 (7.2%)</td>
<td>—</td>
<td>4235 (55.4%)</td>
</tr>
<tr>
<td>Total</td>
<td>13603</td>
<td>7,598,991</td>
<td>35,085</td>
<td>—</td>
<td>6763</td>
</tr>
</tbody>
</table>

*Other repeats include, but are not limited to LINE-1, LTR, and DNA transposon families.

*Owing to the often interrupted structure of non-Alu repeats, this number is likely to be an overestimate.

*Non-repeats refer to sequences other than interspersed complex repeats and include unique sequences as well as low complexity repeats and satellites.
repeat sequences, as well as the experimental observation of inosine within Alus in some brain transcripts (Morse et al. 2002), we believe that the selected transcripts represent genuinely edited RNAs. Other explanations for the overabundance of targeted A-to-G substitutions are implausible. Almost all Alus, except for the youngest AluH subfamily, are fixed in the human population (Batzer and Deininger 2002). Therefore, the observed overabundance of A-to-G substitutions within Alus cannot be explained by polymorphic Alu insertions. In addition, the ratio of different subfamilies showing Alu editing completely matched their overall genomic distribution. To determine whether Alu DNA sequences are hypermutable for A-to-G transitions, we analyzed nearly one million repeat-element-derived SNPs deposited in the dbSNP database and found no excess of A-to-G substitutions (Supplemental Table 1). Furthermore, the observed A-to-G changes are not likely due to recombination between Alus at different genomic loci. We performed BLAST searches of the human genome sampling 100 Alu sequences with five or more A-to-G substitutions. In each case, the original locus was clearly the best match.

Editing Patterns Within Alu Sequences

Why might Alu sequences be major targets for RNA editing? Although it is conceivable that targeting is based on particular nucleotide motifs intrinsic to Alu sequences, considering the nature of ADARs to recognize and bind any dsRNA structure, an a priori explanation is the capacity for Alus within a primary transcript to form extended RNA duplexes (Kikuno et al. 2002; Morse et al. 2002). The enormous number of Alus present in either orientation within transcribed regions makes the intramolecular formation of dsRNA by inverted Alu repeats the most plausible scenario, although dsRNA formation in trans is also possible. ADARs act on dsRNA without apparent sequence-specific binding, but they do show certain neighbor base preferences. To characterize editing preferences within Alu sequences, we carried out a nearest-neighbor analysis one base upstream or downstream of all substituted sites within Alus. There was substantial underrepresentation of Gs and overrepresentation of Ts in one base 5’ to the edited base, as well as a slight excess of Gs and paucity of As one base 3’ to the edited base. Specific trinucleotides were also relatively favored (TAG, AAG) or disfavored (GAN, AAA; Fig. 2A). These findings are in accord with previous experiments in which purified ADARs were mixed with model dsRNA substrates in vitro, and edited sites were determined (Polson and Bass 1994; Lehmann and Bass 2000). However, one notable difference is an exceptional preference for T as the 5’-neighbor. Given our large sample size, the observed pattern of preferences should help predict potential editing sites in other contexts.

To identify editing hotspots within Alu sense or antisense sequences, we compiled the position of edited bases for Alus, the largest subfamily of Alu elements (Fig. 2B). The overall pattern of editing is bell-shaped, hinting that the central region of the Alu RNA duplex may exhibit higher stability and thus serve as a better substrate for ADAR activity. In support of this idea, the central A-rich linker in Alus is a target for editing, but the 3’-poly(A) tail is not. There are several positions particularly prone to editing. For instance, in the region in the right monomer, with the sense-antisense sequence of 5’-TGT(A/G)(A/G)T-3’ and 5’-ATT(A/G)CA-3’ contains the highest absolute numbers of edited bases. The palindromic dinucleotide TA has the highest likelihood of editing on both strands, consistent with ADARs being dsRNA-specific.

Figure 2 Analysis of preferred editing patterns within Alu sequences. (A) Nearest-neighbor preferences derived from 2,668 Alu sequences with at least five A-to-G substitutions in either orientation. We determined the observed and expected frequencies of di- and trinucleotide patterns using the edited version of Alu sequences and their corresponding genomic sequences, respectively. We calculated percentage excess using the formula, [(observed frequency) − (expected frequency)]/(expected frequency) × 100. Previous studies (Polson and Bass 1994; Lehmann and Bass 2000) showed 5’ preferences of T > A > G and T > A > C = G for ADAR1 and ADAR2, respectively; and 3’ preferences of nothing apparent and T = G > C = A for ADAR1 and ADAR2, respectively. (B) Identification of potential hotspots for editing within the Alu family consensus sequence, generated by separate alignments of 412 sense and 260 antisense sequences with at least 10 A-to-G substitutions using CLUSTALW (http://www.ebi.ac.uk/clustalw). We used Alus with 10 or more A-to-G substitutions to get a reasonable input size for alignment and thus minimize the effects of alignment artifacts. The number of As aligned at each position is not uniform because of sequence divergence in individual Alus. The upper panel represents the sense strand and the bottom panel the antisense strand. The basic Alu structure, as shown in the middle, consists of two nonidentical direct repeats (left and right monomer), linked by a central A-rich region. The right monomer is followed by a 3’-poly(A) that is required for Alu replication. The most frequently edited trinucleotide motif, TAG, is labeled with an *.

Cold Spring Harbor Laboratory Press on January 18, 2018 - Published by genome.cshlp.org

Widespread Alu Editing

Genome Research 1721

www.genome.org
Reduced Frequency of Edited Transcripts in Mouse
The mouse genome is devoid of Alu, which arose during primate radiation. It does contain other repeats, particularly four families of unrelated short interspersed nuclear elements (SINEs), whose total copy number is similar to that of Alu in the human genome (Lander et al. 2001; Waterston et al. 2002). We performed a parallel scan for A-to-G clusters among 112,435 full-length mouse cDNAs and found only 91 transcripts with significantly increased A-to-G substitutions (false discovery rate <0.01). Also, 61 transcripts had evidence of editing within interspersed repeat sequences, the majority being SINEs (Supplemental Table 3). Assuming comparable levels of ADAR activity in human and mouse, the pronounced disparity in frequency may be explained by the differences in repeat length (—300 bp vs. —150 bp for human Alu and mouse SINE, respectively), the copy number of each distinct mouse SINE lineage (—one-fourth the frequency of Alus), and the degree of sequence homogeneity (average divergence <10% vs. >25% for human Alu and mouse SINE, respectively). Thus, from an evolutionary perspective, dimerization of the Alu structure (doubling of Alu length by fusing left and right monomers) and the subsequent burst of Alu retrotransposition during early primate evolution (Batzer and Deininger 2002) have likely contributed to the elevated level of repeat-associated RNA editing in humans.

Tissue Distribution of Edited Transcripts
The ubiquitous expression of ADAR1 and ADAR2 has raised a question as to the existence of possible targets in tissues other than brain, where most edited transcripts have been discovered (Maas et al. 2003). Our analysis indicates that A-to-I editing occurs in at least 30 different organs. The greatest number of edited cDNAs (n = 854) are derived from different brain libraries, where they comprise —5% of these cDNAs. However, several other tissues show higher frequencies of edited transcripts than the brain, including thymus, prostate, spleen, and kidney (Table 2). Even within the brain, different regions showed widely different frequencies of editing, ranging from the cerebellum (12.1%; 181/1497 from three libraries) to the hippocampus (4.5%; 116/2007 from four libraries).

Owing to the wide variability in cDNA sequence acquisition methods, these numbers should be considered preliminary. Using a sample of 38,044 transcripts (including 659 edited transcripts) in which information about tissue or cell line source and normalization procedures were reported, we found about twofold increased frequencies of edited transcripts in tissue samples over cell line samples, normalized libraries over nonnormalized, and normal cells over malignant cells (each with p < 0.001; data not shown). Although these differences may reflect varying degrees of ascertainment bias, they may be indicators of more biologically meaningful differences (Maas et al. 2001).

To characterize editing patterns within a given genomic locus from different tissues, we examined several loci where multiple independent overlapping transcripts showed significant A-to-G substitutions. Approximately 10% of our 2674 edited transcripts were of this type. We found that transcripts from independent tissue sources showed different editing patterns, and the positions of edited bases were only partially superimposable (Fig. 3). This suggests that editing of embedded Alu RNA sequences is stochastic, depending on particular local conditions (i.e., availability of ADAR and RNA structures) at the time of processing.

Note: While our paper was in review, we learned that another group had reached a similar result using a different statistical approach (Levanon et al. 2004). They sequenced the cDNAs derived from multiple tissues of several genes, and confirmed that editing of embedded Alu RNA sequences showed variable patterns from tissue to tissue.

Table 2. Tissue Distribution of the Selected (n = 2674) Transcripts

<table>
<thead>
<tr>
<th>Tissue</th>
<th>No. of libraries</th>
<th>No. of cDNAs</th>
<th>Edited cDNAs</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thymus</td>
<td>5</td>
<td>1234</td>
<td>158</td>
<td>12.80</td>
</tr>
<tr>
<td>Prostate</td>
<td>7</td>
<td>1232</td>
<td>101</td>
<td>8.20</td>
</tr>
<tr>
<td>Spleen</td>
<td>5</td>
<td>1440</td>
<td>113</td>
<td>7.85</td>
</tr>
<tr>
<td>Tongue</td>
<td>3</td>
<td>669</td>
<td>43</td>
<td>6.43</td>
</tr>
<tr>
<td>Mammary gland</td>
<td>4</td>
<td>789</td>
<td>45</td>
<td>5.70</td>
</tr>
<tr>
<td>Peripheral nervous</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>system</td>
<td>3</td>
<td>322</td>
<td>18</td>
<td>5.59</td>
</tr>
<tr>
<td>Larynx</td>
<td>6</td>
<td>1287</td>
<td>69</td>
<td>5.36</td>
</tr>
<tr>
<td>Kidney</td>
<td>14</td>
<td>2866</td>
<td>146</td>
<td>5.09</td>
</tr>
<tr>
<td>Brain</td>
<td>52</td>
<td>18,649</td>
<td>854</td>
<td>4.58</td>
</tr>
<tr>
<td>Uterus</td>
<td>10</td>
<td>3316</td>
<td>142</td>
<td>4.28</td>
</tr>
<tr>
<td>Stomach</td>
<td>3</td>
<td>395</td>
<td>16</td>
<td>4.05</td>
</tr>
<tr>
<td>Other</td>
<td>13</td>
<td>1538</td>
<td>48</td>
<td>3.12</td>
</tr>
<tr>
<td>Eye</td>
<td>5</td>
<td>1963</td>
<td>53</td>
<td>2.70</td>
</tr>
<tr>
<td>Embryo</td>
<td>2</td>
<td>1421</td>
<td>37</td>
<td>2.60</td>
</tr>
<tr>
<td>Lymph node</td>
<td>6</td>
<td>1491</td>
<td>38</td>
<td>2.55</td>
</tr>
<tr>
<td>Small intestine</td>
<td>8</td>
<td>1801</td>
<td>45</td>
<td>2.50</td>
</tr>
<tr>
<td>Cervix</td>
<td>2</td>
<td>891</td>
<td>22</td>
<td>2.47</td>
</tr>
<tr>
<td>Lung</td>
<td>11</td>
<td>3301</td>
<td>53</td>
<td>1.61</td>
</tr>
<tr>
<td>Bone marrow</td>
<td>4</td>
<td>872</td>
<td>14</td>
<td>1.61</td>
</tr>
<tr>
<td>Colon</td>
<td>8</td>
<td>2217</td>
<td>35</td>
<td>1.58</td>
</tr>
<tr>
<td>Heart</td>
<td>8</td>
<td>866</td>
<td>13</td>
<td>1.50</td>
</tr>
<tr>
<td>Ovary</td>
<td>7</td>
<td>1189</td>
<td>16</td>
<td>1.35</td>
</tr>
<tr>
<td>Testis</td>
<td>10</td>
<td>9064</td>
<td>104</td>
<td>1.15</td>
</tr>
<tr>
<td>Bladder</td>
<td>3</td>
<td>479</td>
<td>5</td>
<td>1.04</td>
</tr>
<tr>
<td>Liver</td>
<td>6</td>
<td>2049</td>
<td>21</td>
<td>1.02</td>
</tr>
<tr>
<td>Placenta</td>
<td>12</td>
<td>4317</td>
<td>44</td>
<td>1.02</td>
</tr>
<tr>
<td>Mixed</td>
<td>5</td>
<td>1506</td>
<td>13</td>
<td>0.86</td>
</tr>
<tr>
<td>No information</td>
<td>52</td>
<td>37,377</td>
<td>275</td>
<td>0.74</td>
</tr>
<tr>
<td>Skin</td>
<td>10</td>
<td>2761</td>
<td>12</td>
<td>0.43</td>
</tr>
<tr>
<td>Blood</td>
<td>7</td>
<td>3134</td>
<td>11</td>
<td>0.35</td>
</tr>
<tr>
<td>Muscle</td>
<td>6</td>
<td>1979</td>
<td>6</td>
<td>0.30</td>
</tr>
<tr>
<td>Pancreas</td>
<td>4</td>
<td>752</td>
<td>2</td>
<td>0.27</td>
</tr>
<tr>
<td>Bone</td>
<td>1</td>
<td>50</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>
lian cells, with pre-mRNAs exposed to ADARs at the time of their processing (Raitskin et al. 2001; Bratt and Ohman 2003). Although we do not find evidence that editing has directly altered splice sites, base modifications of \textit{Alu} duplexes could change local RNA structures and hinder the normal splicing process. Conversely, the sheer number of \textit{Alus} present in intragenic regions and their near universal tendency to be removed from pre-mRNAs suggest that editing could mark certain aberrant RNAs for nuclear retention and/or degradation. Cellular machineries exist that process multiple-inosine-containing RNAs. For example, extensively edited transcripts are retained in the nucleus by a protein complex consisting of p54nrb, PSF, and matrin 3 (Zhang and Carmichael 2001). In addition, a cytoplasmic ribonuclease activity that specifically cleaves multiple-edited RNA has been reported (Scadden and Smith 2001b). In both cases, the key requisite for protein–RNA interaction is hyperedited bases, whereas a single A-to-I modified RNA is not a target. Despite the exquisite specificity of these systems, endogenous targets have yet to be identified. It would be of interest to see whether multiple-edited \textit{Alu}-containing RNAs serve as targets for these regulatory proteins.

The true proportion of edited transcripts in the human transcriptome cannot be estimated based on this cDNA survey for several reasons, including the arbitrary minimum threshold inherent in a statistical approach, the likely differential stability of multiple-inosine-containing transcripts and their partitioning between the nucleus and cytoplasm, as well as the variability in methods used to generate cDNA libraries. Nonetheless, given the ubiquity of \textit{Alu} elements and ADARs, edited transcripts are likely to constitute a sizable and widespread population of RNAs. A recent study indicates that the human transcriptome is much more complex than commonly considered, having ∼50% novel transcripts that do not correspond to known genes or ESTs (Kampa et al. 2004). The presence of edited transcripts adds to our appreciation of this complexity. The editing targets identified in the current study may represent a small portion of this “hidden” transcriptome. Consistent with this idea, only two of the 15 previously found \textit{Alu}-associated editing targets (Morse et al. 2002) overlapped with our selected transcripts.

dsRNAs longer than 30 bp induce dsRNA-dependent pathways in mammalian cells such as the interferon response, with activation of PKR and 2′,5′-oligoadenylate (Samuel 2001). Similar to their role in antagonizing RNAi by destabilizing dsRNA (Scadden and Smith 2001a; Tonkin and Bass 2003), ADARs might alleviate the dsRNA pressure generated by inverted \textit{Alus}, probably the biggest reservoir of dsRNAs in the cell. Interferon-induced up-regulation of ADAR1 in lymphocytes leads to an increase of inosine in mRNAs up to 5% of all As (Yang et al. 2003). Thus, the existence of a large pool of targets edited within \textit{Alu} sequences could account for the abundance of inosine found in poly(A) RNAs from many different tissues (Paul and Bass 1998; Maas et al. 2003). The presence of \textit{Alu} sequences within noncoding portions of genes may have an impact on expression of host genes. Recent reports concerning full-length LINE elements within introns in the human genome suggest that they too subtly influence the level of host RNAs by affecting the rate of synthesis through a gene by RNA polymerase II (Perepelitsa-Belancio and Deininger 2003; Han et al. 2004). Identification of potential functions and regulation of \textit{Alu}-associated RNA editing are future challenges in understanding this novel intersection of RNA-modification and retrotransposon biology.

\section*{METHODS}

\subsection*{Statistical Identification of Potentially Edited Transcripts}

We obtained all human and mouse full-length cDNA sequences from the UCSC Genome Browser database (http://genome.ucsc.org).
Characterization of edited transcripts. Below the genomic location is the RefSeq mRNA, in black. Exons are wide blocks, UTRs are narrower blocks, and introns are dashed lines. 

Alu sequences present within edited transcripts are shown above each block. Other (A) APOBEC3G locus exemplifies intron-retaining transcripts. Accession numbers for the transcripts shown are NM021822 for RefSeq mRNA, and (1) AF182420, (2) AK092614, (3) AX748234, (4) AK022802, (5) BC061914 for full-length cDNAs. (B) The F1IR locus exemplifies Alu-associated editing within a UTR. Of note, the positions of edited bases for the four edited transcripts were only partially overlapping. The accession numbers are (1) NM016946, (2) NM144501 for RefSeq mRNAs, and (1) AJ723398, (2) AL136649, (3) AK026665, (4) BC001533, (5) AJ154005, (6) AV58896 for full-length cDNAs. (C) AK126984 exemplifies transcripts that have no corresponding known gene. This particular transcript is extensively edited in four of five Alus. (D) Deviation from known gene structure measured as percentage exon overlap with RefSeq mRNA. (E) Edited transcripts are enriched in Alu elements. The distribution of edited versus nonedited transcripts in each of the histograms was determined to be significantly different (p < 10^-5).

Instead, based on the p-values, we selected the transcripts to limit the false discovery rate (q) to 0.01 (Benjamini and Hochberg 1995; Storey and Tibshirani 2003). Transcripts with fewer than five A-to-G substitutions or <50% A-to-G substitutions among all substitutions were removed to reduce false positives generated by the scan statistic method. Of all the transcripts in which we believe editing has taken place, we expect that <1% are false positives. Computations were done using the R statistical environment with the q-value package (Storey 2002; Team 2003). We identified the location of interspersed repeat elements (excluding simple low complexity repeats and satellites) in each transcript by RepeatMasker (http://www.repeatmasker.org).

Repeat-Element-Derived SNP Analysis

We obtained 704,418 repeat-element-SNPs from the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP, Build #119). Because the direction of the nucleotide change is unknown in this data set, we categorized them into six different polymorphic allele types and calculated the frequency of each allele. Similarly, we obtained an additional 175,097 Alu-derived SNPs with known direction of nucleotide change, separated them into 12 different substitution patterns, and tallied the frequency of each substitution (see Supplemental Table 1).

Characterization of Edited Transcripts

We downloaded tissue, library, histology, normalization protocol, and preparation method information from the UCSC Table Browser as well as the Cancer Genome Anatomy Project (CGAP; http://cgap.nci.nih.gov). We simplified tissue categories according to the UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) tissue classification system. For example, libraries derived from different parts of the brain were combined into a single brain category. To minimize the ascertainment bias, we only used libraries with at least 50 cDNAs for Table 2. We used the UniGene database (Build #164 for Homo sapiens) to cluster edited transcripts that mapped to the same genomic location. We
manually inspected each edited transcript using the UCSC Genome Browser and categorized them into three major classes as described in the text. We examined the extent of deviation of edited transcripts from the known gene structure by calculating the percentage exon overlap with RefSeq mRNA (http://www.ncbi.nlm.nih.gov/RefSeq).

ACKNOWLEDGMENTS

This work was supported, in part, by R01GM60534 (A.G.) and U24MH068457 (S.B.) from the Public Health Service. We thank Xuemei Chen, Marc Gartenberg, Sam Gunderson, Mike Kiledjian, Joseph Naus, Harold Sackrowitz, Steven Sherry, Bento Soares, Ruth Stewart, and Susan Wessler for helpful discussions and/or evaluation of the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


Lehmann, K.A. and Bass, B.L. 2000. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities.


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.


WEB SITE REFERENCES


Received June 4, 2004; accepted in revised form July 9, 2004.
Widespread RNA Editing of Embedded Alu Elements in the Human Transcriptome


*Genome Res.* 2004 14: 1719-1725
Access the most recent version at doi:10.1101/gr.2855504

Supplemental Material

http://genome.cshlp.org/content/suppl/2004/08/12/14.9.1719.DC1

References

This article cites 42 articles, 14 of which can be accessed free at:
http://genome.cshlp.org/content/14/9/1719.full.html#ref-list-1

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions