Identification of widespread ultra-edited human RNAs

Shai Carmi, Itamar Borukhov, and Erez Y. Levanon

PLoS Genetics **7** (2011) 1–11.

Speaker: Joseph Chuang-Chieh Lin

The Comparative & Evolutionary Genomics/Transcriptomics Lab.
Genomics Research Center, Academia Sinica
Taiwan

28 March 2012

Outline

Outline

Introduction

- RNA editing: a post-transcriptional modification of RNA molecules.
- One of the most frequent RNA-editing in primates: Adenosine-to-Inosine (A-to-I).
  - Inosine is later translated as guanosine (G).
  - A-to-I editing is mediated by adenosine deamisnase proteins acting on dsRNA (ADAR).
  - It increases the transcriptome complexity and constitutes an additional mechanism for controlling gene activity.
  - Crucial for normal life and development & related to diseases (e.g., brain tumors)

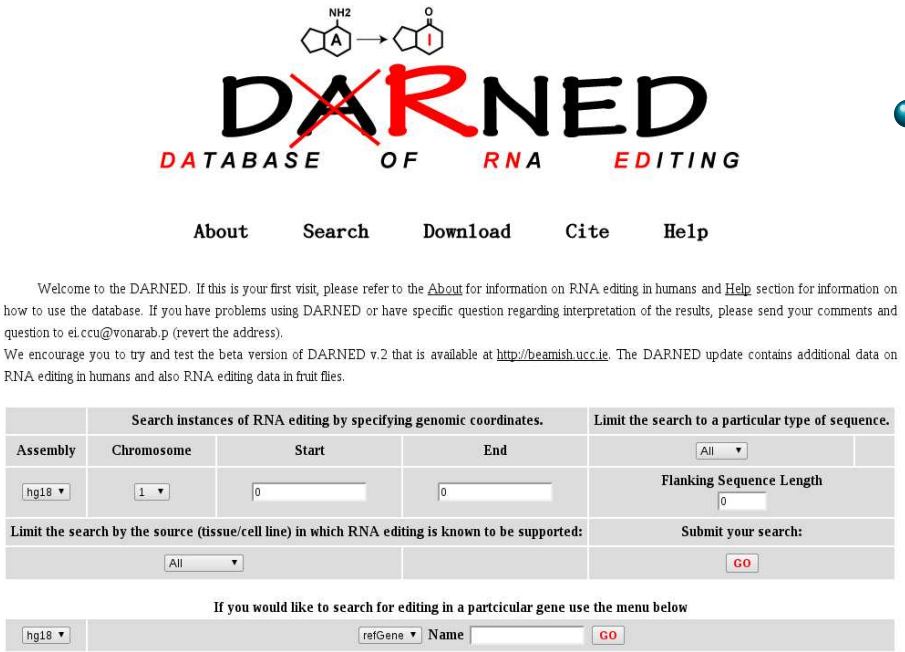
Introduction

- RNA editing: a post-transcriptional modification of RNA molecules.
- One of the most frequent RNA-editing in primates: Adenosine-to-Inosine (A-to-I).
  - Inosine is later translated as guanosine (G).
  - A-to-I editing is mediated by adenosine deamisnase proteins acting on dsRNA (ADAR).
  - It increases the transcriptome complexity and constitutes an additional mechanism for controlling gene activity.
  - Crucial for normal life and development & related to diseases (e.g., brain tumors)

Introduction

- RNA editing: a post-transcriptional modification of RNA molecules.
- One of the most frequent RNA-editing in primates: Adenosine-to-Inosine (A-to-I).
  - Inosine is later translated as guanosine (G).
  - A-to-I editing is mediated by adenosine deamisnase proteins acting on dsRNA (ADAR).
  - It increases the transcriptome complexity and constitutes an additional mechanism for controlling gene activity.
  - Crucial for normal life and development & related to diseases (e.g., brain tumors)

Introduction (contd.)



- DARNED: a database of RNA editing in human [Kiran & Baranov 2010].
  - $>$ 40,000 editing sites.
  - Detected through systematic surveys of cDNA and EST libraries & genome-wide screens.

Introduction (contd.)

- Known A-to-I editing sites can be roughly classified into two categories:
  - At coding sequences.
  - At repetitive elements (mostly in *Alu* elements [Batzer & Deininger 2002]) in UTRs or introns [e.g., Levanon et al. 2004].

Introduction (contd.)

- Ultra-edited RNAs:
  - An extremely large fraction of their adenosines are edited.

Ultra-edited RNAs

We designate an RNA as ultra-edited if it satisfies the following conditions:

1. The alignment has ≥ 12 A-to-G mismatches.
2. The number of A-to-G mismatches is $\geq 90\%$ of all mismatches.
3. The number of A-to-G mismatches was $\geq 20\%$ of the number of As in the (genomic) subsequence extending from the first to the last A-to-G mismatch.

Introduction (contd.)

- Ultra-editing was usually overlooked in systematic RNA editing detection screens.
- However, based on previous evidences and the large amount of cellular inosine, many more ultra-edited RNAs are suspected to exist.

- A computational pipeline to identify ultra-edited RNAs.
  a. Start with RNA sequences that previously could not be aligned to the genome.
  b. Realign the RNA sequences in (a.) after reducing the genomic DNA and RNA sequences by A→G transformation.
  c. Whenever a transformed RNA has successfully aligned to the transformed genome, the original sequences were recovered and the mismatches were examined.
  d. Search for the alignments for large and dense clusters of A-to-G mismatches and filter the results.
- With high confidence, 760 A-to-I ultra-edited RNAs are detected in over 14,000 editing sites, most of which were previously unknown.
  - Total: 15,646 (14,538 are unique).

- A computational pipeline to identify ultra-edited RNAs.
  a. Start with RNA sequences that previously could not be aligned to the genome.
  b. Realign the RNA sequences in (a.) after reducing the genomic DNA and RNA sequences by A→G transformation.
  c. Whenever a transformed RNA has successfully aligned to the transformed genome, the original sequences were recovered and the mismatches were examined.
  d. Search for the alignments for large and dense clusters of A-to-G mismatches and filter the results.
- With high confidence, 760 A-to-I ultra-edited RNAs are detected in over 14,000 editing sites, most of which were previously unknown.
  - Total: 15,646 (14,538 are unique).

- A computational pipeline to identify ultra-edited RNAs.
  a. Start with RNA sequences that previously could not be aligned to the genome.
  b. Realign the RNA sequences in (a.) after reducing the genomic DNA and RNA sequences by A→G transformation.
  c. Whenever a transformed RNA has successfully aligned to the transformed genome, the original sequences were recovered and the mismatches were examined.
  d. Search for the alignments for large and dense clusters of A-to-G mismatches and filter the results.
- With high confidence, 760 A-to-I ultra-edited RNAs are detected in over 14,000 editing sites, most of which were previously unknown.
  - Total: 15,646 (14,538 are unique).

- A computational pipeline to identify ultra-edited RNAs.
  a. Start with RNA sequences that previously could not be aligned to the genome.
  b. Realign the RNA sequences in (a.) after reducing the genomic DNA and RNA sequences by A→G transformation.
  c. Whenever a transformed RNA has successfully aligned to the transformed genome, the original sequences were recovered and the mismatches were examined.
  d. Search for the alignments for large and dense clusters of A-to-G mismatches and filter the results.
- With high confidence, 760 A-to-I ultra-edited RNAs are detected in over 14,000 editing sites, most of which were previously unknown.
  - Total: 15,646 (14,538 are unique).

- A computational pipeline to identify ultra-edited RNAs.
  a. Start with RNA sequences that previously could not be aligned to the genome.
  b. Realign the RNA sequences in (a.) after reducing the genomic DNA and RNA sequences by A→G transformation.
  c. Whenever a transformed RNA has successfully aligned to the transformed genome, the original sequences were recovered and the mismatches were examined.
  d. Search for the alignments for large and dense clusters of A-to-G mismatches and filter the results.
- With high confidence, 760 A-to-I ultra-edited RNAs are detected in over 14,000 editing sites, most of which were previously unknown.
  - Total: 15,646 (14,538 are unique).

Outline

Extraction of candidate sequences

- Query the UCSC Genome Browser for long (> 250 bp) human ESTs or mRNAs from GenBank that did not align to the genome, and downloaded their sequences from NCBI Batch Entrez.
  - ▷ 458,124 sequences.
- Discard the ESTs or mRNAs with:
  - $> 60\%$ or $< 10\%$ percentage of *a single nucleotide*,
  - $\geq 10\%$ of *ambivalent nucleotides* (non-[ACGT]), or
  - $\geq 50\%$ *simple repeats content*.
  - ▷ 438,807 sequences remain.
- MEGABLAST the remaining sequences to the genome (GRCh37/hg19) and eliminate each sequence that aligned with $\geq 98\%$ identity (along $\geq 90\%$ of its length).
  - ▷ 334,344 sequences remain.

Extraction of candidate sequences

- Query the UCSC Genome Browser for long (> 250 bp) human ESTs or mRNAs from GenBank that did not align to the genome, and downloaded their sequences from NCBI Batch Entrez.
  - ▷ 458,124 sequences.
- Discard the ESTs or mRNAs with:
  - $> 60\%$ or $< 10\%$ percentage of *a single nucleotide*,
  - $\geq 10\%$ of *ambivalent nucleotides* (non-[ACGT]), or
  - $\geq 50\%$ *simple repeats content*.
  - ▷ 438,807 sequences remain.
- MEGABLAST the remaining sequences to the genome (GRCh37/hg19) and eliminate each sequence that aligned with $\geq 98\%$ identity (along $\geq 90\%$ of its length).
  - ▷ 334,344 sequences remain.

Extraction of candidate sequences

- Query the UCSC Genome Browser for long (> 250 bp) human ESTs or mRNAs from GenBank that did not align to the genome, and downloaded their sequences from NCBI Batch Entrez.
  - ▷ 458,124 sequences.
- Discard the ESTs or mRNAs with:
  - $> 60\%$ or $< 10\%$ percentage of *a single nucleotide*,
  - $\geq 10\%$ of *ambivalent nucleotides* (non-[ACGT]), or
  - $\geq 50\%$ *simple repeats content*.
  - ▷ 438,807 sequences remain.
- MEGABLAST the remaining sequences to the genome (GRCh37/hg19) and eliminate each sequence that aligned with $\geq 98\%$ identity (along $\geq 90\%$ of its length).
  - ▷ 334,344 sequences remain.

DNA and RNA transformation

- ⋆ A-to-I ultra-edited RNAs harbor a large number of A-to-G mismatches, but *no* or *very few* mismatches of any other type.
- ⋆ Therefore, an ultra-edited RNA would generate a good alignment to the genome if A-to-G mismatches are specifically **masked**.

- Transform every A to G both in the genomic DNA sequence and in the candidate RNA sequences.
- All strand combinations are separately aligned.

DNA and RNA transformation (contd.)

DNA:	T	A	C	C	C	A	C	T	A	C	A	G	G	C	A
	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
RNA:	T	A	C	C	C	G	C	T	G	C	G	G	A	C	G

DNA:	T	G	C	C	C	G	C	T	G	C	G	G	G	C	G
	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
RNA:	T	G	C	C	C	G	C	T	G	C	G	G	G	C	G

DNA and RNA transformation (contd.)

DNA strand	RNA strand	Transformation	Detected editing events
+	+	A→G	A-to-G, G-to-A, T-to-C, C-to-T
+	−	A→G	A-to-G, G-to-A, T-to-C, C-to-T
−	+	A→G	A-to-G, G-to-A, T-to-C, C-to-T
−	−	A→G	A-to-G, G-to-A, T-to-C, C-to-T
+	+	A→C	A-to-C, C-to-A, T-to-G, G-to-T
+	−	A→C	A-to-C, C-to-A, T-to-G, G-to-T
−	+	A→C	A-to-C, C-to-A, T-to-G, G-to-T
−	−	A→C	A-to-C, C-to-A, T-to-G, G-to-T
+	+	G→C	G-to-C, C-to-G
+	−	G→C	G-to-C, C-to-G
+	+	A→T	A-to-T, T-to-A
+	−	A→T	A-to-T, T-to-A

- With A→G transformation, clusters of G-to-A are also detected, though they are served as a negative control.

Alignment of the transformed sequences

- The authors uploaded the candidate RNA sequences and the human genome to a commercial cloud computer (http://aws.amazon.com/ec2) to speed up the computation of the alignments (all strand combinations).
- Only the best alignment which is particularly convincing (E-value $\leq 10^{-50}$, percent identity $\geq 95\%$, length ≥ 100 bp) is retained.
- 690,495 successful alignments are obtained ($\approx 17\%$ of the number of possible alignments).
  - $690{,}495/(334{,}344 \times 12)$.

Identification of ultra-editing

- Recall the definition of A-to-G ultra-edited RNAs:

Ultra-edited RNAs

We designate an RNA as ultra-edited if it satisfies the following conditions:

1. The alignment has ≥ 12 A-to-G mismatches.
2. The number of A-to-G mismatches is $\geq 90\%$ of all mismatches.
3. The number of A-to-G mismatches was $\geq 20\%$ of the number of As in the (genomic) subsequence extending from the first to the last A-to-G mismatch.

- A similar procedure is used to search for RNAs with other possible types of editing (G-to-A, A-to-C, etc.).
- The values of the cutoffs: ≈ 12 mismatches.
  - Expected number of mismatches in an EST aligning to an *Alu* element discarded by UCSC.

Filtering of the results

Remove the following cases:

i. RNAs that appear ultra-edited in more than one transformation/strand combination.

ii. RNAs where the aligning part of the RNA or DNA is too homogeneous.

- For example, a single nucleotide repeat was longer than 36 bp, or one nucleotide frequency was outside the range 10%–60%.

iii. RNAs where the alignment has > 5 gaps overall or > 3 gaps in the RNA or DNA.

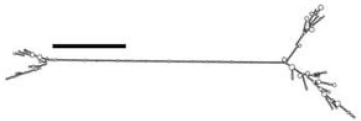
Filtering of the results (contd.)

iv. RNAs where another MEGABLAST search against the (non-transformed) genome yield a better alignment in another locus (over length $\geq 90\%$ of that of the original alignment).

v. RNAs where A-to-G ultra-editing is found on a particular strand of the DNA, but other mRNA sequences (from the UCSC Genome Browser) support, by ≥ 2 sequences, transcription only from the opposite strand.

- This step particularly serves to eliminate T-to-C editing.

Outline



DB347561 vs. chr22:31650658-31650949 (48 A-to-G mismatches)

```
CTGCAGGCCAGGCATAGTGGCTCACACCTGTAATCGCAGCACTTCGGGAGGCCAAGGCGGGCAGATCACTTGAGGTCGGGAGTTCGAGACCAGCCTCACCAACATGGGGAAACCCCGTCT
.........G...G.G..............GG.......G............GG.......G....G.....................G..G.....G..G..G......GG........
CTACTAAAAATAGAAAAATTAGCCGGGTGTGGTGGCATGCGCCAGTAATCCCAGCTACTCAGGAGGCTGAGGTGGGAAAATCACTTGAACTCGGGAAGCAGAGGTTGCAGTGAGCCGAGA
.......GG..G....G...G..............G......G..GG....G...G...G..............GGGG..G....GG.......G..G........G...G.......
TTGTGCCACTGCACTCCAGCCTGGGCGATAAGAGCAAAATTCCATCTCAAA
.......G....G....G.................GGGG....G.......
```

BU180574 vs. chr7:73641384-73641862 (57 A-to-G mismatches)

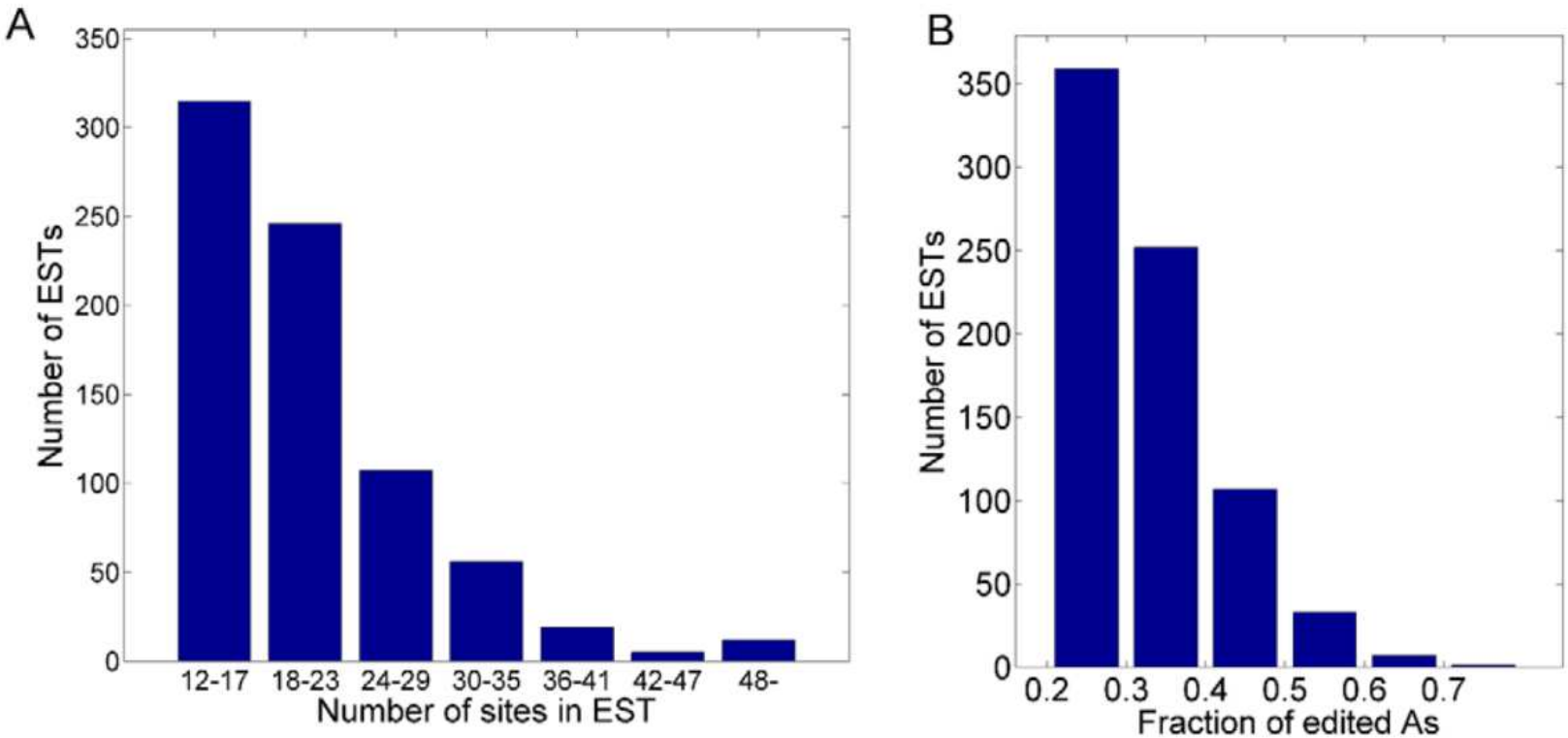
```
TTGTACAAAACCTTCCCATTAGCCAGGCATGGTAGCTCATGCCTGTAATCCCAGCACTTGGGGAGGCTGAGGTGGGCCGATCACTTTAGGTCAGGAGTTTGAGACCAGCCTGGCCAACAT
.........................G.........G............GG....G.................................G.........G..........G...........GG...
GGTGAAACCCCATCTCTACTAAAAATACAAAAATTAGCCAGGTGTTGTGGCGTGTGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAAGAGAAACCTGAACCCAGGAGGCAGAGGTTG
....G......G.........G..G.G.GGG.G..G.........................GG....G...G............G...GG..........GG...G......G.G.....
CAGTGAGCCGAGATCACGCCACTGCACTCCAGCCTGGGTGACAGAGCGAGACTCCCTCTCAAAAAACAAAATAAAATAAAATAAATACAAATTGGCCGGGTGTGGTGGCAGACACCTGTA
.G.........G....G....G....G................G.G.....G.........GGG......G.....................G.....................G...G.....G
ATCGCAGCTACTCGGGAGGCTGAGGCAGGAGCATCGCTTGAACCCAGGAGGCGGAGTTTGCAGTGAGCCAAGATTGCA-CCATTGCACTTCAGCCTGGGCGACAGGAACGAA-AA-AAAA
G.........G................G.................G...............G...G...GG.......C..G.............................G..N....
```

AI937118 vs. chr2:77654132-77654395 (47 A-to-G mismatches)

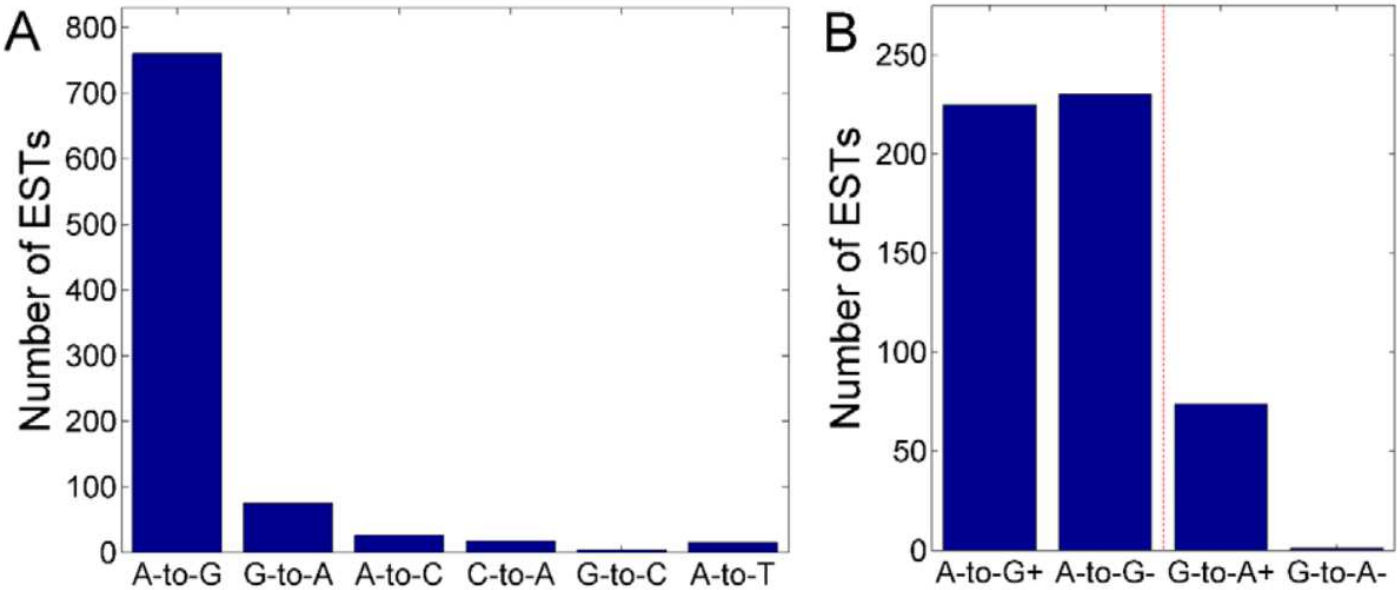
```
TGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAAGCTGAGGTAGGCAGATCATGAGGTC-AAGAGATCAAGACCATCCTGATCAACATGATGAAACCTCATCTCCACTAAAAATAC
...................GG....G...........G...T....G...G....G..G....CG.G.....GG..........G..GG.G.....GGG....G........GGG...G.
AAAAATTAGCTGGGTGTGGTGGTGCGCACCTGTAGTCCCAGCTACTCTGGAGGCTGGGGCAGGAGAATCACTTGAACCCAGGAGGCGGAGGTTGCAGTGAGCAGAAATCACAGCACTGTG
..G....G..................G.....G.....G...G......G.........G..G..G..G....GG...G........G......G...G..G.GGG..G.G........
```

DA079658 vs. chr2:27536658-27537228 (39 A-to-G mismatches)

```
ATTTGAGCACTCTTGTTTTATTACTTTTTTTTTCTTTTTGAGATAGGGTCTTGTTCTGTCGCCCAGGCAGCAGTGTAATGGCGTGATCACTGCAGCCTGGACCTCCTGGGCTCAAACAGT
.............................................G.................................G.....................................GGG.G..
CCTCCTACTTCAGCCTCCTGAGTAGCTGGGACTACAGGCACGTGCCACCGTGCCCAGCTAATTGTTTAAAAAAATTTTTTGTAGAAGTGGGGTCTTGCTATATTGCCCAGGCTGGACTCA
......G....G...........G.........G.....G..............G.............................G..............G......G..........G
AACTCTTGGGCTCAAGCAATCCTCCTGCCTCAGCCTCCCAGAGTGCCGGGATTACAGGTGTGAGCCACCATTCCAGGCCGGCACTCTTGTTTTAGAAAAAGATTTATTCTGGCCAGGCAC
GG............G..GG............G.....................G.....................................................................
AGTGGCTTATGCCTGTAATCTCAGCACTTTGAGAAGCCGAGGTGAGCAGATCACCTGAGGTCAGGGGTTCAAGATCAGCCTGGCTAACATGGCAAAACCCCATCTCTACTAAAAATACAA
G..............G.....G.........G..G.................G.................GG..............G.............................T..
AAAATAGCCAGGCATGGTGGTGGGTGCCTGTAGTCCCAGCTACTTGAGAGGCTGAGGCAGGGGAATCGCTTGAATCGGGAGGCAGAGGTTG
.........G...G.................G.....G...G..............G.............G.................
```


A
Number of ESTs
350
300
250
200
150
100
50
0
12-17
18-23
24-29
30-35
36-41
42-47
48-
Number of sites in EST
B
Number of ESTs
350
300
250
200
150
100
50
0
0.2
0.3
0.4
0.5
0.6
0.7
Fraction of edited As

The number of ultra-editing events by mismatch type and strand

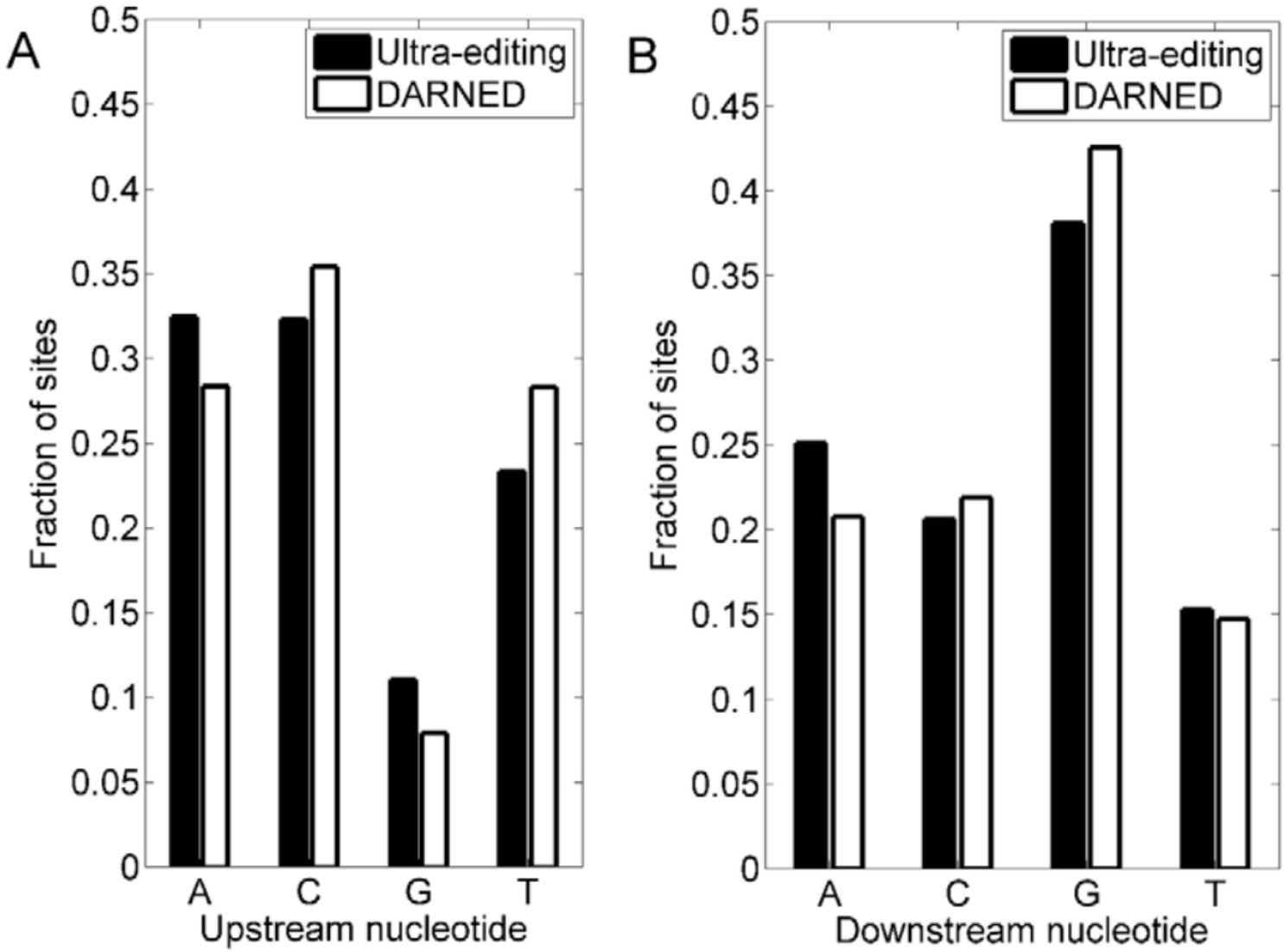


- All (75) but one of the G-to-A ultra-edited ESTs are from the + strand (sequencing error?).
- ▷ 63/75 G-to-A ultra-edited ESTs came from NCI-CGAP libraries, compared to only 99/760 for A-to-G.
- ▷ 65/75 of the G-to-A ultra-edited ESTs were sequenced in 1997, compared to only 114/760 for A-to-G.
- ⋆ Most of the G-to-A clusters are due to technical faults.

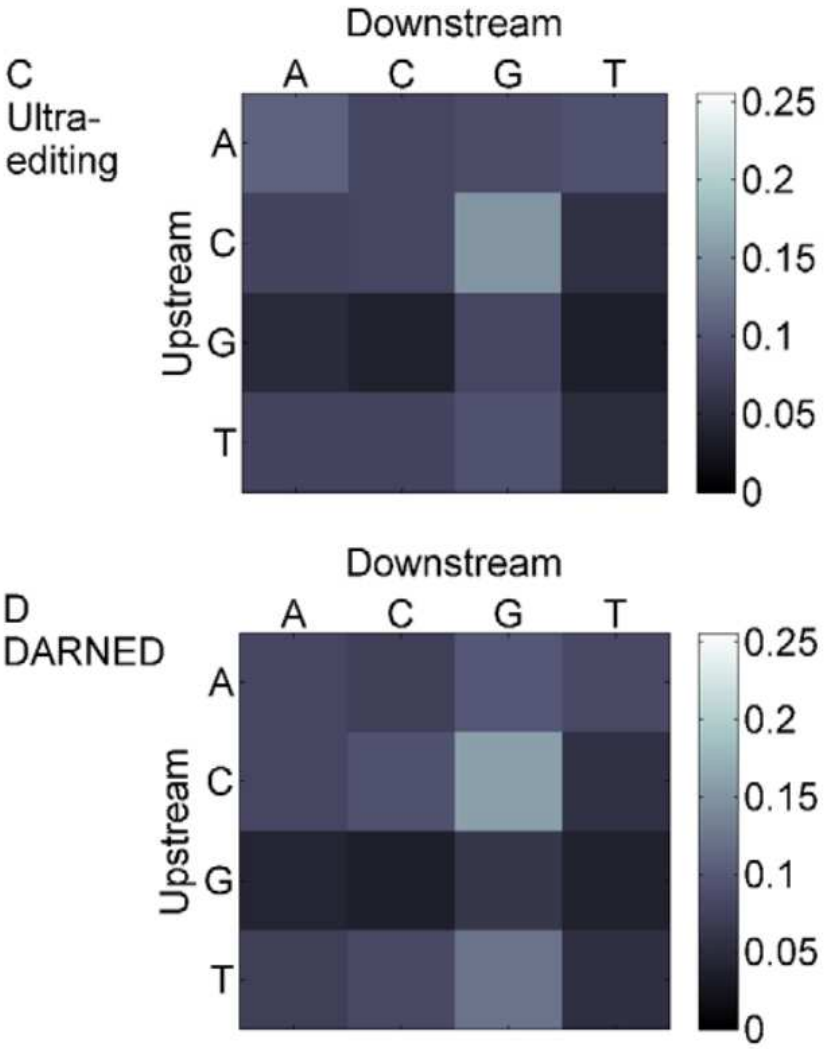
Most A-to-G sites are novel

- 13,668 novel sites (94%), which did not appear in DARNED.
- The 760 ultra-edited ESTs map to 695 distinct genomic regions.
  - 647 are covered by one ultra-edited EST.
  - 41 are covered by two ultra-edited ESTs
  - 1 is covered by 11 ESTs.
- Only 42 sites (0.29%) overlap with genomic SNPs.

The ultra-editing sequence motif is similar to the known ADAR1 motif



The ultra-editing sequence motif is similar to the known ADAR1 motif (contd.)



- ultra-editing is **more** common than DARNED: AAA, GAA, GAG.
- ultra-editing is **less** common than DARNED: CAC, AAG, TAG.

Tissues enriched in ultra-editing

Table 1. Top tissues and health states containing ultra-edited ESTs.

Tissue	Number of ESTs	Enrichment[a]	Health state	Number of ESTs	Enrichment[a]
Liver	312	12.98	Normal	563	1.57
Brain	118	0.97	Lung tumor	13	0.73
Lung	33	0.89	Glioma	9	0.78
Thymus	31	3.74	Soft tissue/muscle tissue tumor	9	0.69
Eye	21	0.93	Non-neoplasia	8	0.72
Muscle	20	1.63	Head and neck tumor	8	0.37
Prostate	20	0.65	Colorectal tumor	8	0.4
Uterus	19	0.74	Kidney tumor	6	0.61
Uncharacterized tissue	15	0.4	Gastrointestinal tumor	6	0.42
Spleen	12	2.17	Uterine tumor	6	0.57

The enrichment is the number of ultra-edited ESTs from the tissue divided by the expected number, which was computed as follows.

(a). For each tissue, we calculate the ratio of the total number of ESTs in the tissue to the total number of ESTs in *all* tissues.

(b). The expected number of ultra-edited ESTs in a tissue is (a) multiplied by the total number of ultra-edited ESTs.

Most ultra-edited RNAs overlap with relatively new *Alu* elements

- The DARNED short clusters:
  - Group adjacent editing sites (not ultra-edited, separated by < 300 bp—the *Alu* length) and eliminate clusters with a single site or with ≥ 12 sites.
  - 4,456 *short clusters* are obtained.

Table 2. The fraction of edited elements from each major Alu sub-family.

Alu sub-family	Number of ultra-edited ESTs[a]	Number of DARNED short clusters	Total number in the entire genome
AluY	91 (11.2%)	415 (9.4%)	143,178 (12.6%)
AluS	601 (73.9%)	2811 (63.6%)	686,962 (60.1%)
AluJ	121 (14.9%)	1194 (27%)	312,138 (27.3%)

[a]Note that the sum of the second column exceeds the number of ultra-edited ESTs because some ESTs overlap with more than one Alu.
doi:10.1371/journal.pgen.1002317.t002

Ultra-editing substrates form relatively long dsRNA structure

Table 3. Secondary structure and repetitive elements in the edited regions.

	Property	Ultra-editing[c,d]	DARNED short clusters[c,d]	P-value[e]
1	Maximum length of dsRNA using BLAST[a].	322±11	212±4	9.6×10^{-23}
2	Maximum length of dsRNA using RNA Fold[b].	400±5	363±2	2×10^{-12}
3	Total repeat content in the region.	63.2%±0.7%	52.6%±0.3%	6×10^{-37}
4	Minimum of (number of +Alus, number of −Alus) in the region.	3.8±0.09	3.59±0.04	2.6×10^{-2}
5	Distance between the edited Alu and the nearest inverted Alu.	855±52	956±21	3.4×10^{-7}

[a]The edited region and its reverse complement were aligned using BLAST. We used the total number of aligning base pairs as an estimate of the length of the dsRNA.
[b]The secondary structure of the RNA was calculated using RNAFold [50]. We used the maximal number of open brackets in the structure as an estimate of the length of the dsRNA.
[c]Regions used: 1.5 kbp flanking upstream and downstream of the edited regions for properties 1,2, and 3; 5 kbp for property 4.
[d]Means are reported along with the standard error of the mean [sqrt(sample variance/n)].
[e]P-values were calculated using Mann-Whitney U test.
doi:10.1371/journal.pgen.1002317.t003

- The reason for the increased dsRNA length is likely the dramatic overabundance of repeats in the ultra-edited flanking region.

Ultra-edited sites are relatively rare in exons

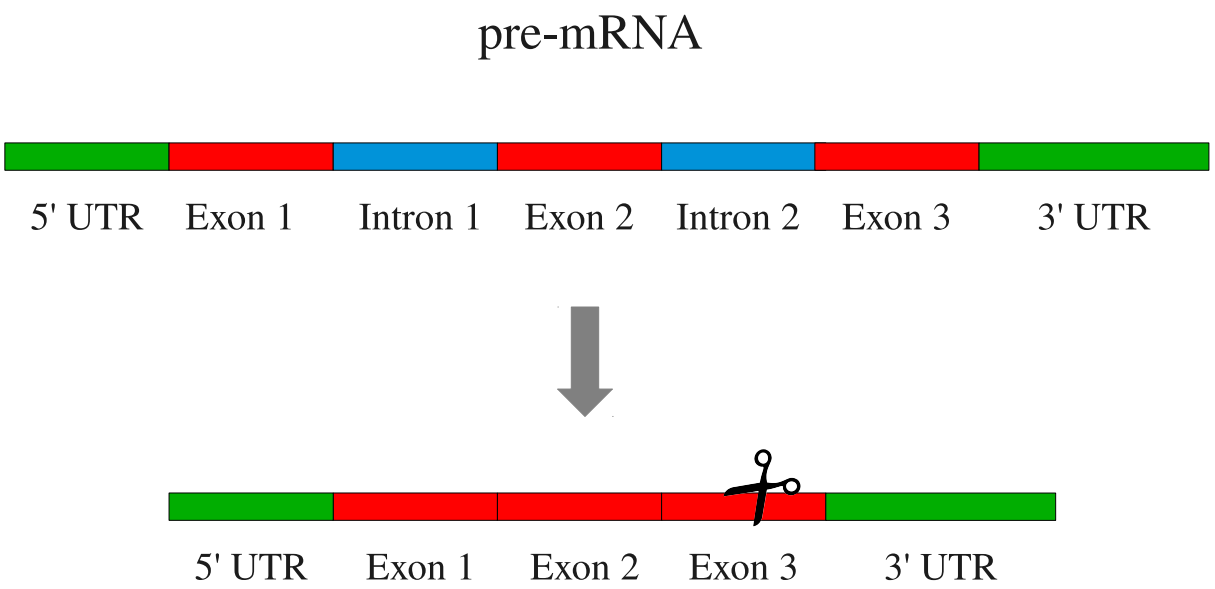
- Most ultra-edited RNAs overlap with genes (547/760 ESTs (72%); the overlap is with 460 genes).
  - 61 (8%) overlap with exons;
    - 38 with 3'UTRs, 4 with 5'UTR, 17 with non-coding RNA, and two with coding sequences.
  - The other 486 ESTs overlap with introns.
- The higher level of editing in 3'UTRs compared to 5'UTRs is most probably due to their larger sizes.
  - avg. length of 3'UTRs: $\approx$ 525 bp; avg. length of 5'UTRs: $\approx$ 145 bp [Milo et al. 2010].

Possible cleavage or NC-splicing of ultra-edited RNAs

- Hyper-edited RNAs can be specifically cleaved [Scadden & Smith 1997, 2001; Scadden 2005].
- Hundreds of putative hyper-editing sites are shown to be non-canonically (NC) spliced out of UTRs [Osenberg et al. 2009].
- To find out if ultra-edited regions are also cleaved or NC-spliced, the authors search for the ultra-edited regions that overlap with both a UTR and an intron.

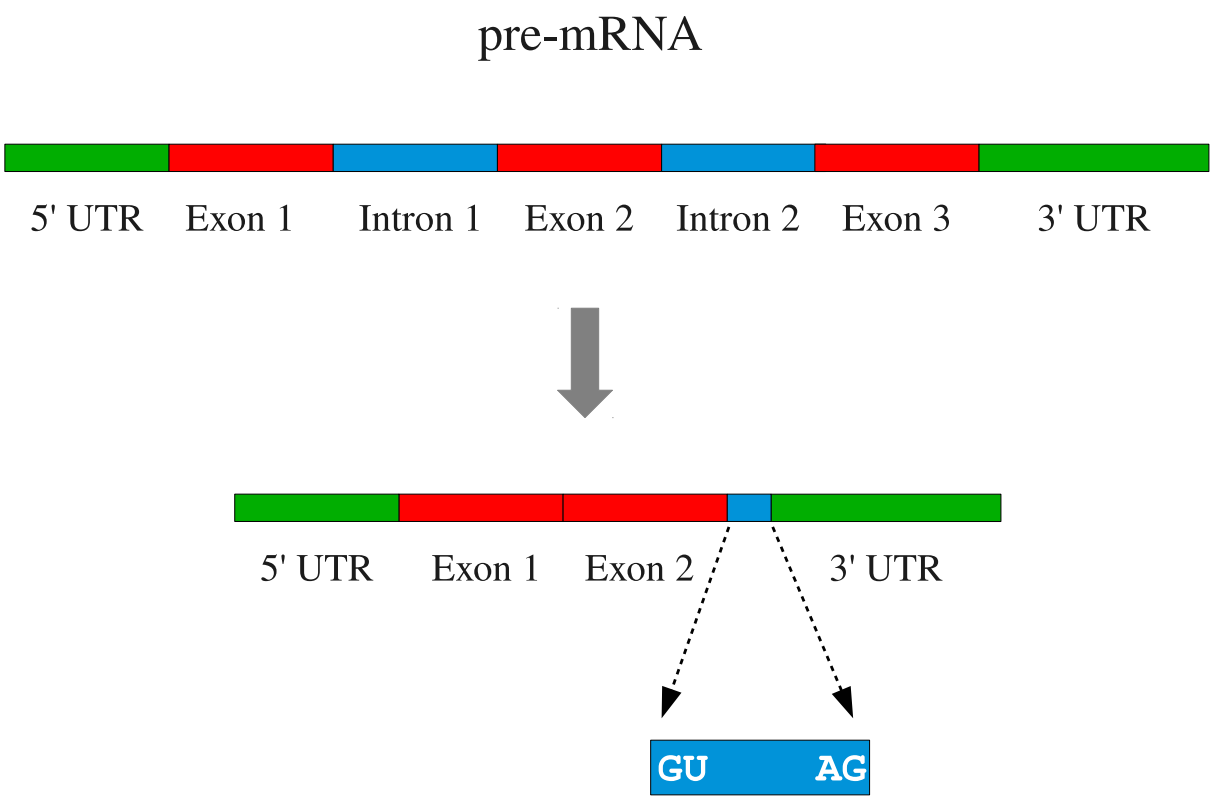
Possible cleavage or NC-splicing of ultra-edited RNAs (contd.)

- Cleaved RNAs appear as properly spliced sequence, up to a certain point where an exon extends abnormally until it is cleaved at the ultra-edited region.



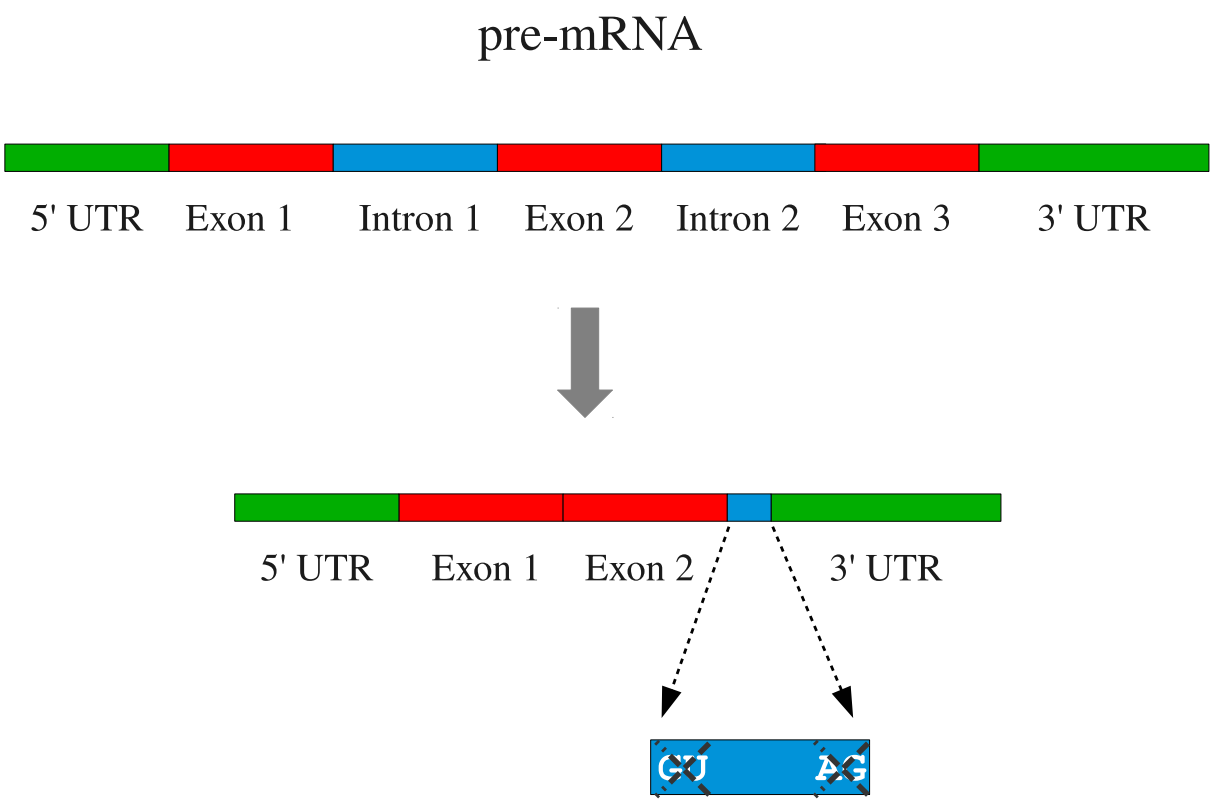
Possible cleavage or NC-splicing of ultra-edited RNAs (contd.)

- NC-spliced RNAs also appear to be normally spliced, except for an additional short intron in their 3'UTR, whose boundaries overlap with the ultra-edited *Alus* but lack the GT-AG canonical splicing signals.



Possible cleavage or NC-splicing of ultra-edited RNAs (contd.)

- NC-spliced RNAs also appear to be normally spliced, except for an additional short intron in their 3'UTR, whose boundaries overlap with the ultra-edited *Alus* but lack the GT-AG canonical splicing signals.



Possible cleavage or NC-splicing of ultra-edited RNAs (contd.)

- 10 cleaved & 5 NC-spliced mRNAs in ultra-edited regions were identified.

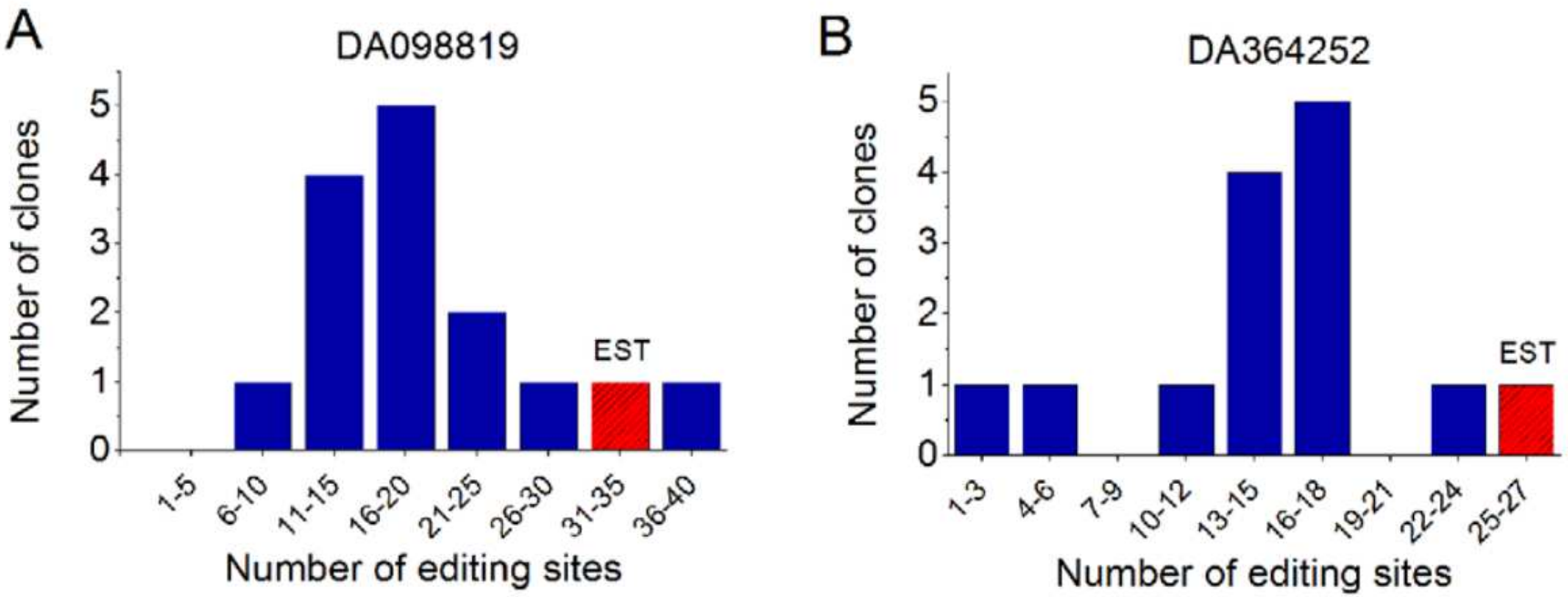
Ultra-edited genomic regions are slightly less conserved

- Ultra-editing substrates are more abundant in introns and in new *Alu* sub-families than the short clusters.
- Hypothesis:
  - The ultra-edited genomic regions are less conserved.

Ultra-edited genomic regions are slightly less conserved (contd.)

- Extract for each edited region with flanking 500 bp upstream and downstream.
- Using PhyloP: less conserved compared to the short clusters.
  - avg. score: 0.008, ± 0.002 standard error of the mean; P-value: 0.004 by t-test.
  - the short clusters: avg. score $\pm(15+1) \times 0.001$.
- Using PhastCons: no difference is observed.

Experimental validation of ultra-edited RNAs



Outline

Discussion

- The coordinates of the ultra-editing sites have been deposited in DARNED.
- The high confidence the authors have in their ultra-edited RNAs stems from:
  - mismatch clusters of types other than A-to-G are extremely less;
  - the sequence motif is typical to the editing by ADAR;
  - the localization of the editing sites in *Alu* elements.

Discussion

- The coordinates of the ultra-editing sites have been deposited in DARNED.
- The high confidence the authors have in their ultra-edited RNAs stems from:
  - mismatch clusters of types other than A-to-G are extremely less;
  - the sequence motif is typical to the editing by ADAR;
  - the localization of the editing sites in *Alu* elements.

Discussion

- The coordinates of the ultra-editing sites have been deposited in DARNED.
- The high confidence the authors have in their ultra-edited RNAs stems from:
  - mismatch clusters of types other than A-to-G are extremely less;
  - the sequence motif is typical to the editing by ADAR;
  - the localization of the editing sites in *Alu* elements.

Discussion (contd.)

- The most edited tissue except livers: the brain.
- However, this is to some extent because of the high coverage of the brain transcriptome.

Discussion (contd.)

- The most edited tissue except livers: the brain.
- However, this is to some extent because of the high coverage of the brain transcriptome.

Discussion (contd.)

- The direct sequence changes induced by editing do not seem to have an important function
- Regardless of the function of ultra-editing, the edited regions are characterized by potential to create particularly long, stable dsRNAs.

Discussion (contd.)

- The direct sequence changes induced by editing do not seem to have an important function
- Regardless of the function of ultra-editing, the edited regions are characterized by potential to create particularly long, stable dsRNAs.

Discussion (contd.)

- Ultra-editing could be **rare** due to the following reasons.
  - Only 0.4% (3/695) of the ultra-edited regions are covered by ≥ 4 ESTs, compared to 10.6% in a previous genome-wide screen.
  - Only 2/27 clones in the study are far more edited than other clones.
  - The probability to encounter an ultra-edited RNA is exponentially small.
    - Assume that each site is edited independently.

Discussion (contd.)

- Ultra-editing could be **rare** due to the following reasons.
  - Only 0.4% (3/695) of the ultra-edited regions are covered by ≥ 4 ESTs, compared to 10.6% in a previous genome-wide screen.
  - Only 2/27 clones in the study are far more edited than other clones.
  - The probability to encounter an ultra-edited RNA is exponentially small.
    - Assume that each site is edited independently.

Thank you.