Supplemental Material for:

Accurate Identification of A-to-I RNA editing in human by transcriptome sequencing

Jae Hoon Bahn, Jae-Hyung Lee, Gang Li, Christopher Greer, Guangdun Peng and Xinshu Xiao

Supplemental Methods and Results
Supplemental Figures 1-15
Supplemental Tables 1-13

Supplemental Methods and Results

Cell culture, siRNA transfection and RNA purification

U87MG cells used for RNA-Seq were from the same aliquots of cells for genome sequencing obtained from Dr. S. Nelson's lab (UCLA). Cells used for all other experiments were purchased from American Type Culture Collection (ATCC). Cells were maintained in DMEM high glucose medium supplemented with pyruvate, L-glutamine, and 10% fetal bovine serum (FBS) (Hyclone). Control siRNA (ID: D-001210-02-05) and *ADAR* (also known as *ADAR1*) siRNA were obtained from Dharmacon RNAi Tech. The sense strand of the *ADAR* siRNA is 5'-CGCAGAGUUCCUCACCUGUATT-3' (Jayan and Casey 2002), which targets all *ADAR* mRNAs. 50nmol of *ADAR* siRNA and control siRNA were transfected respectively with lipofectamine RNAiMAX (Invitrogen) into $5x10^4$ U87MG cells per single well of 6-well plate via reverse transfection, according to the manufacture's instruction. Cells were harvested 48h following transfection. Total RNA was isolated using the RNeasy micro kit (Qiagen), according to the manufacturer's instructions.

Sanger sequencing of gDNA and cDNA and clonal sequencing

For validation of candidate editing sites, the same aliquots of U87MG cells were used to obtain DNA and RNA samples. Genomic DNA (gDNA) was isolated using the QIAamp DNA mini kit according to the manufacture's instruction. PCR (25 cycles) was performed using 100 ng of gDNA, Taq 2X master mix (NEB) and PCR primers listed in Supplemental Table 4 for genes *CTSB*, *CTSS*, *GNL3L*, *NUP43*, *JUB*, *MDM2*. PCR products were subject to Sanger sequencing (GeneWiz) to confirm the DNA sequences. Reverse transcription (RT) was carried out using 500 ng total RNA and the qScript cDNA kit (Quanta Bio) according to the manufacturer's instructions. PCR (25-32 cycles, primers in Supplemental Table 4) was performed using Taq 2X master mix (NEB) and one-tenth of the RT product. RT-PCR products of the JUB and MDM2 genes were sent for Sanger sequencing to confirm the predicted editing sites (Supplemental Fig. 6). In parallel, *CTSB*, *CTSS*, *GNL3L*, and *NUP43* cDNAs were subcloned into the pcDNA 2.1 TOPO cloning vector. Plasmid DNA was purified using the GeneJET plasmid miniprep kit (Fermentas). Twenty (50 for *CTSB*) clones from each gene were randomly picked and sent for Sanger sequencing. Editing level of each putative editing site was calculated as the fractions of clones with the edited nucleotide among all clones.

Western blot analysis

U87MG cells were harvested with RIPA buffer and protein concentration was measured by BCA protein assay kit (Thermo Fisher Scientific). 50 μg of protein samples were loaded onto 8% SDS-PAGE gel and run at 100V for 2h. The gel was transferred to nitrocellulose membrane at 25V for 1h using a semi-dry transfer kit (Bio-rad). 5% non-fat dry milk in TBS-T buffer was used for blocking. The *ADAR* and *Actin* antibodies (Santa Cruz Biotechnology) were diluted at 1:200 with the blocking solution and incubated with the membrane on a rocking shaker at 4 °C overnight. The membrane was washed with TBS-T for 5 min three times. Secondary antibody conjugated HRP was incubated for 1h. After final washing, the membrane was developed with ECL solution and exposed on the film.

Library preparation and Illumina sequencing

We used the standard Illumina protocol to prepare libraries for RNA-Seq (http://www.illumina.com/support/documentation.ilmn). Briefly, 10μg total RNA was first processed via poly-A selection and fragmentation. We generated first-strand cDNA using random hexamer-primed reverse transcription and subsequently used it to generate second-strand cDNA using RNase H and DNA polymerase. Sequencing adapters were ligated using the Illumina Paired-End sample prep kit. Fragments of ~200 bp were isolated by gel electrophoresis, amplified by 15 cycles of PCR and sequenced on the Illumina Genome Analyzer IIx (Cofactor Genomics) in the paired-end sequencing mode (2x60nt reads).

Enrichment of double-stranded RNA near predicted editing sites

To test whether the predicted A-to-I editing sites are located in or near dsRNA regions, we extracted 4,001bp genomic sequences with the editing sites located in the middle. This sequence was then reverse-complemented and aligned against the immediate neighborhood (200bp flanking each side) of the editing sites using blastn (Altschul et al. 1990). Excluding the case of self-alignment, if the second-best alignment has an alignment length $\geq$ 50 and total identity $\geq$ 80% (parameters chosen to include most known A-I editing with strong dsRNA structures), we concluded that the editing site resides in dsRNA structures. We then calculated the fraction of editing sites that are associated with dsRNA features.

To evaluate the statistical significance of dsRNA enrichment, we first grouped the editing sites according to the type of genomic regions they are located in. We generated four groups with editing sites (1) in known genes and in Alus (2) in known genes but not in Alus (3) in intergenic regions and in Alus (4) in intergenic regions but not in Alus. Subsequently, for each editing site, we selected a random genomic site with the 'A' base from the same type of region as the editing site and similar G+C content in the corresponding 4,001bp region near the editing site. The set of random controls (with the same number of sites as the test set) was then analyzed in the same way as described above. This random selection process was repeated 100 times, and the fractions of sites with dsRNAs were recorded each time. A p-value was calculated for the enrichment of dsRNA features near true editing sites by fitting a normal distribution to the values of the random controls (Fig. 4A).

Sequence motif near predicted editing sites

To find enriched sequence motifs near putative A-to-I editing sites, we analyzed 201bp sequences (100bp flanking each side of the editing sites) using the Multiple Em for Motif Elicitation (MEME) method (Bailey and Elkan 1994). For background control, we used a second-order Markov model generated from random Alu repeat regions. The best motif (21nt-long, Fig. 5A) had highly significant E-value ($<10^{-100}$), whereas one of the best motifs from the shuffled sequence controls had a very large E-value ($>10^{100}$). The same motif was identified if random 201bp Alu sequences were used as a background data set, instead of a Markov model. The score of each motif occurring near editing sites was calculated using the positional weight

matrix from MEME. The consensus structure (Fig. 5A) of the motif was derived using ClustalW (Thompson et al. 1994) and RNAalifold (Bernhart et al. 2008).

Analyses of sequence and structural conservation

To evaluate the conservation level of each editing site and their flanking regions, we downloaded the 46-way multiz alignments from the UCSC browser (Fujita et al. 2010). We focused on the 10 primates among these 46 species, including Human, Chimp, Gorilla, Orangutan, Rhesus, Baboon, Marmoset, Tarsier, Mouse lemur, and Bushbaby. Based on the multiple sequence alignments, the percent identity at each nucleotide position of interest was calculated and shown in Figs. 4C, 4D. As controls, we picked random 'A' sites in the same type of genomic regions as the editing sites. The types of regions considered include coding, introns, UTRs and whether they have Alu elements or not.

Conservation of the motif sequences shown in Fig. 5A was evaluated in the same way as described above. As controls, we picked random 21bp sequences near the identified motif sites and used their conservation levels to normalize those of the actual motif sequences (Supplemental Fig. 10A). Next, we examined the conservation of the base-pairing patterns for corresponding pairs of positions in the motif structure (1 vs 18, 2 vs 17, etc as shown in Fig. 5A). Specifically, the base-pairing conservation level is defined as the percentage of primate genomes with valid base-pairing (AU, GC or GU). Random controls were chosen similarly as described above and their base-pairing conservation levels were used to normalize those of the actual motifs (Fig. 5B and Supplemental Fig. 10B). The above conservation analyses were carried out for 5 groups of editing sites classified according to the score of the strongest motif occurring in the 201bp neighborhood of each editing site. The score cutoffs were: 6.6, 16.8, 21.3 and 24.4 bits, which were chosen so that each group has approximately the same number of editing events.

Comparisons of base-pairing conservation between long-range interactions (inter-motifs) and local hairpin structure (intra- motif)

To evaluate whether multiple motifs in the same gene tend to form inter-motif dsRNA structures, we collected genes with at least two occurrences of the structural motif (motif score > 6.6), at least one of which is within 100bp of an A-to-I editing site. For each motif (*M*) near an editing site, we chose its potential partner motif as the one that can form the maximum number of base pairs with *M* in the same gene. Next, we calculated the base-pairing conservation level between the two motifs as described above but for reversed positions (e.g., position 1 of motif 1 vs. position 18 of motif 2, etc) due to the palindromic nature of the motif. Intra-motif base-pairing conservation was calculated as described above. Supplemental Fig. 11 shows the difference in the conservation levels of intra-motif and inter-motif base-pairing.

Motif enrichment near editing sites in non-Alu regions

For the 461 non-Alu putative A-to-I editing sites, we evaluated whether the structural motif is enriched in their flanking regions. In each range of motif scores (with cutoffs 6.6, 16.8, 21.4 and 24.4), we identified the number of editing sites with a motif (within 100bp) whose score is

greater than the motif score cutoff. We also randomly chose 461 'A' sites outside of Alu regions and carried out the same analysis. The random selection process was repeated 100 times and the number of occurrences of the motif was compared to that near actual editing sites (Supplemental Table 8).

Read coverage of A-to-I editing sites validated via clonal sequencing

It is expected that the higher the read coverage, the more accurate the estimated editing ratios may be. We examined whether the relatively high validation rate shown in Fig. 3A is due to high read coverage of the tested sites that is not representative of the entire set of predicted A-to-I events. About 69% (64 out of 93) of the editing candidates in our validation were associated with relatively low number of reads (<=31 reads per candidate editing site). The rest (29) of candidate sites were located in one gene (*CTSB*) with higher read coverage (35-69 reads per site). With this gene excluded, the read coverage distribution in the validation set is roughly similar, although not identical, to that of the entire set of predicted A-to-I editing (filtered for <=31 reads per site as well, Supplemental Fig. 7). In this case, editing ratios resulted from RNA-Seq and clonal sequencing are still highly correlated (results not shown) and the false discovery rate of our prediction is about 6% (4 false positives out of 62 predictions). If only the candidate sites in the *CTSB* gene were considered, there were no false positive predictions compared to clonal sequencing, indicating that the validation rate would be higher for sites with >31 read coverage. Considering the results on the above two groups of editing sites, we expect our overall validation success rate was not very much augmented due to high read coverage.

Overlap of genes with putative A-to-I editing sites with cancer genes

We compared the set of 1,167 genes with putative A-to-I editing sites in our study with the list of genes related to cancer as annotated by the NCI Cancer Gene Index project. If all expressed genes are included, there were 341 genes in common, suggesting significant enrichment of cancer-related genes among those with A-to-I editing sites ($p < 2.2x10^{-16}$, hypergeometric test). Since the edited genes and the caner genes tend to be highly expressed in the U87MG cells, we carried out another analysis to focus on highly expressed genes. We defined a new background gene set to include only genes whose expression level is in the upper 30% of all expressed genes based on our RNA-Seq data in the control siRNA samples. A total of 885 genes with A-to-I editing and 3,058 genes from the cancer database are present in this background set, among which 292 genes are in common ($p = 0.009$, hypergeometric test). The expression levels of the 3 sets of genes (edited, cancer-related and background genes) are highly similar (data not shown).

References:

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.

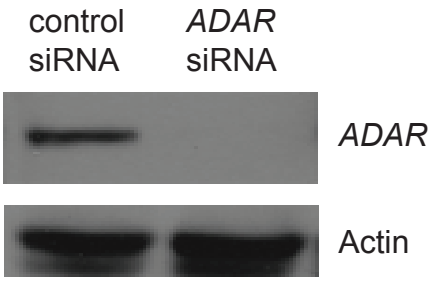
Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.

Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., and Stadler, P.F. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9**: 474.

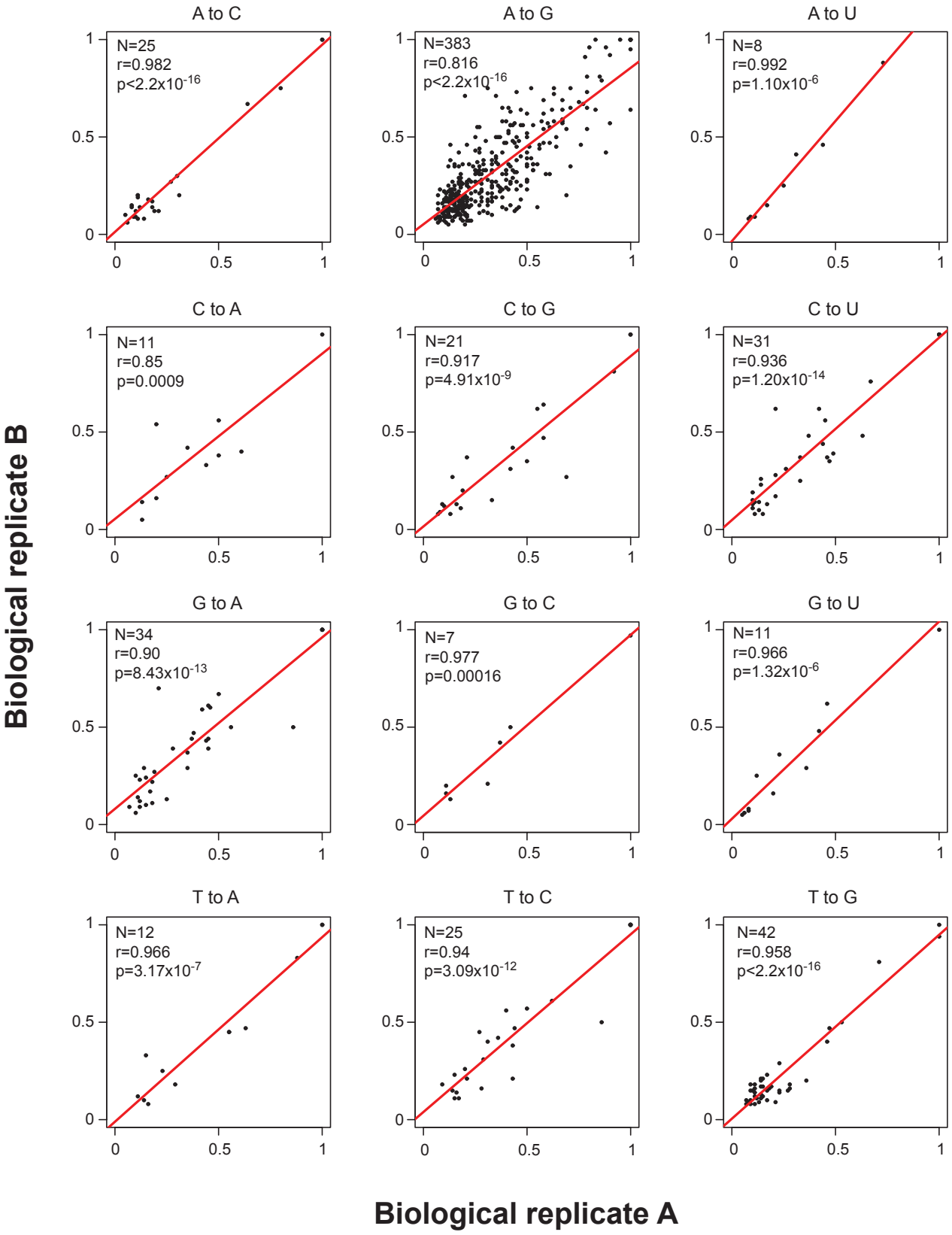
Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A. et al. 2010. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39**(Database issue): D876-882.

Jayan, G.C. and Casey, J.L. 2002. Inhibition of hepatitis delta virus RNA editing by short inhibitory RNA-mediated knockdown of ADAR1 but not ADAR2 expression. *J Virol* **76**(23): 12399-12404.

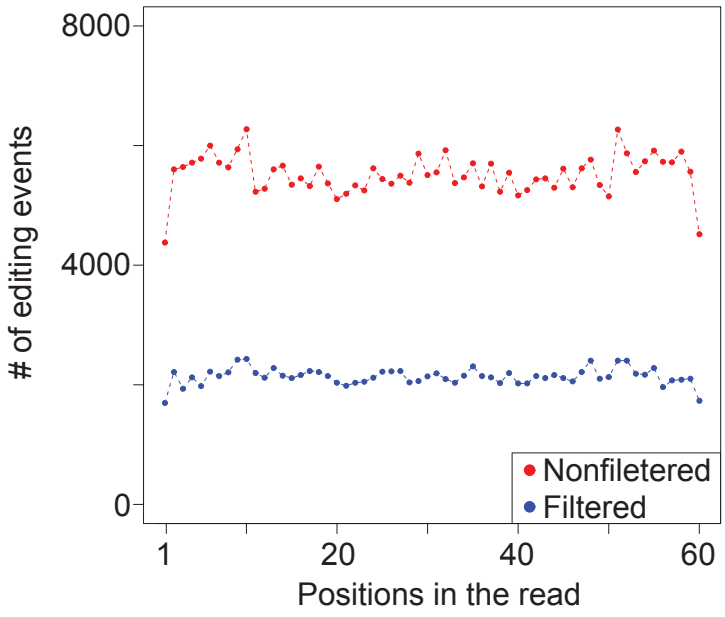
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**(22): 4673-4680.



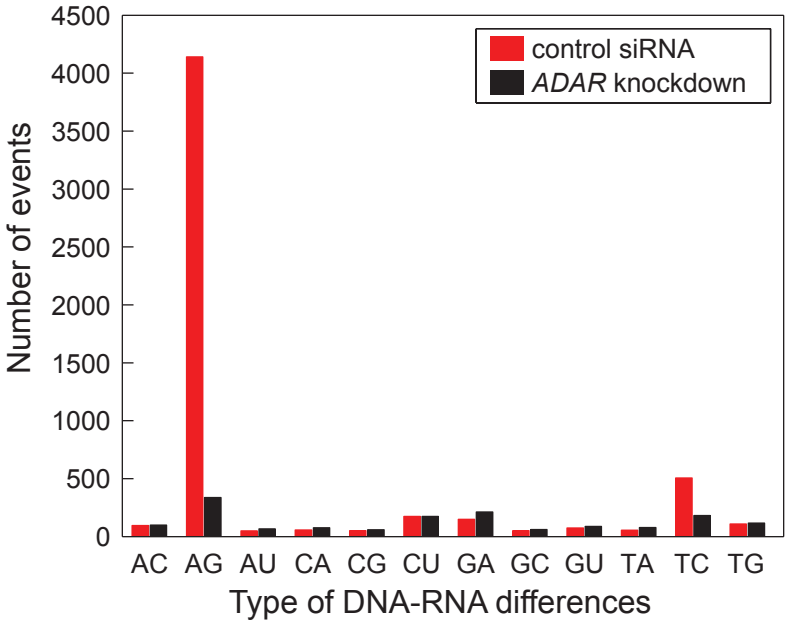
Supplemental Fig. 1. *ADAR* **knockdown.** Western blot of *ADAR* in U87MG cells transfected with the control siRNA or the siRNA targeting *ADAR*. Results of Actin are shown as loading controls.



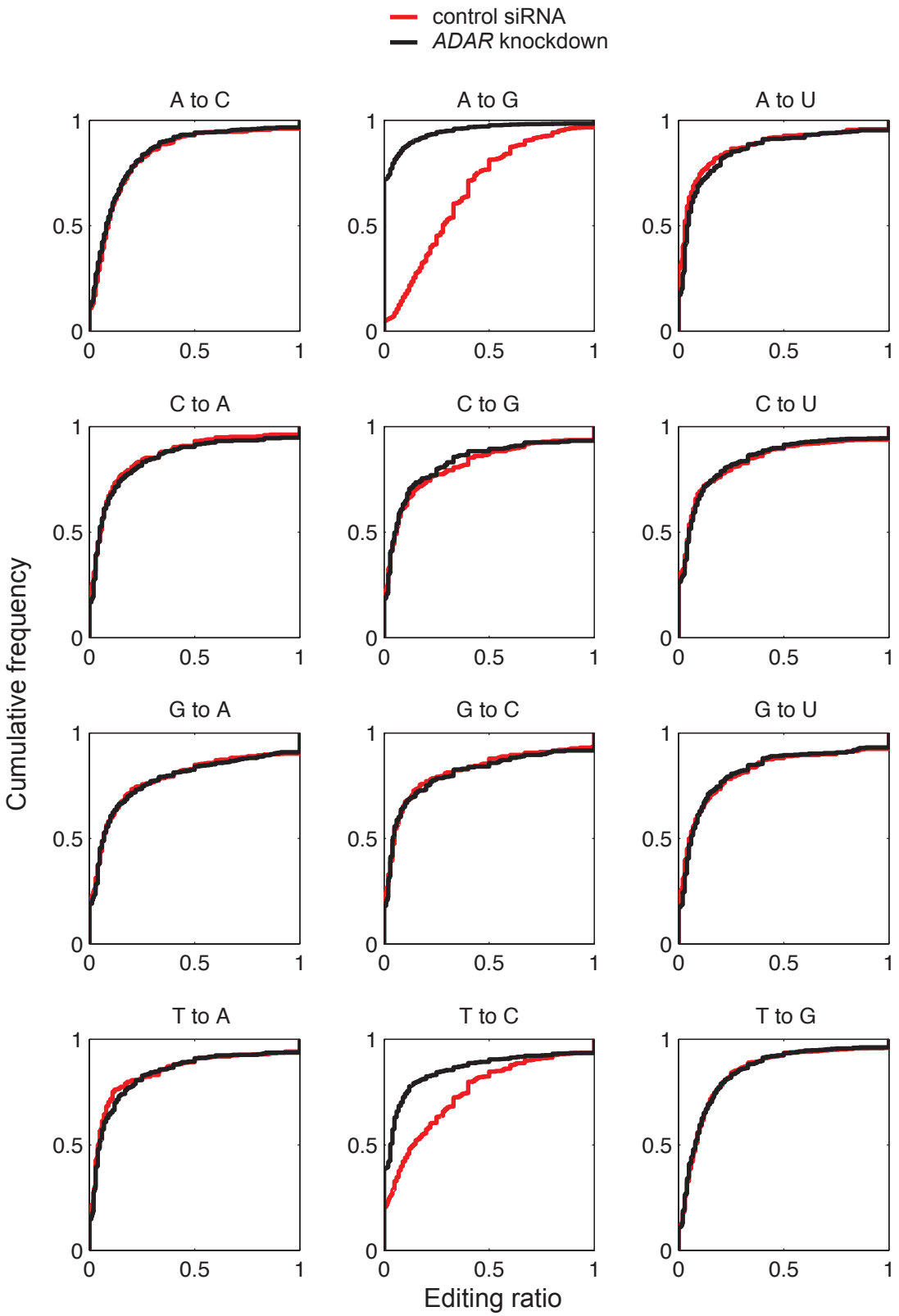
Supplemental Fig. 2. Correlation of editing ratios between biological replicates of control siRNA transfection in U87MG cells. N is the number of the events common to both samples, Pearson correlation coefficient and the corresponding p-values are shown.



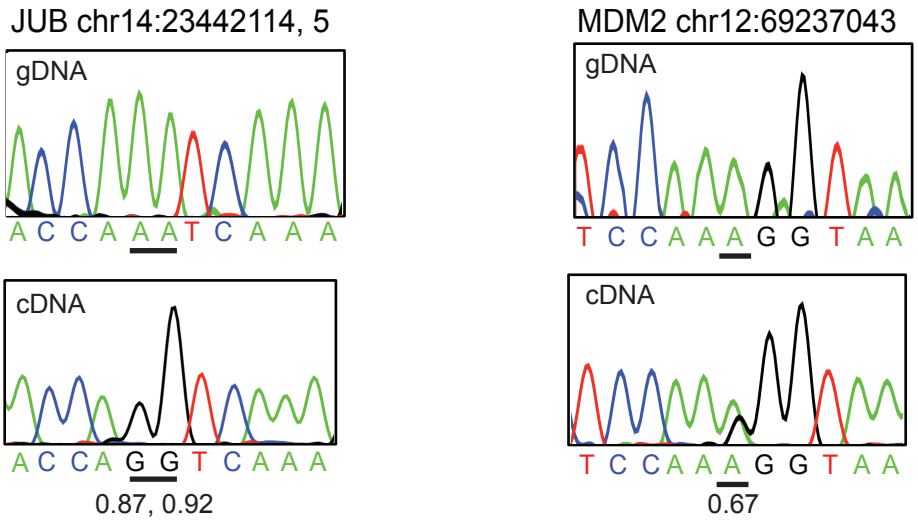
Supplemental Fig. 3. Distribution of the positions of putative A-to-I editing sites in the corresponding RNA-Seq reads. X-axis represents the position in the reads; y-axis shows the total number of editing events in a corresponding read position. Red dots represent all A-to-G events detected in the U87MG cells (control siRNA transfection). Blue dots correspond to those events that had an estimated editing ratio ≥ 0.2.



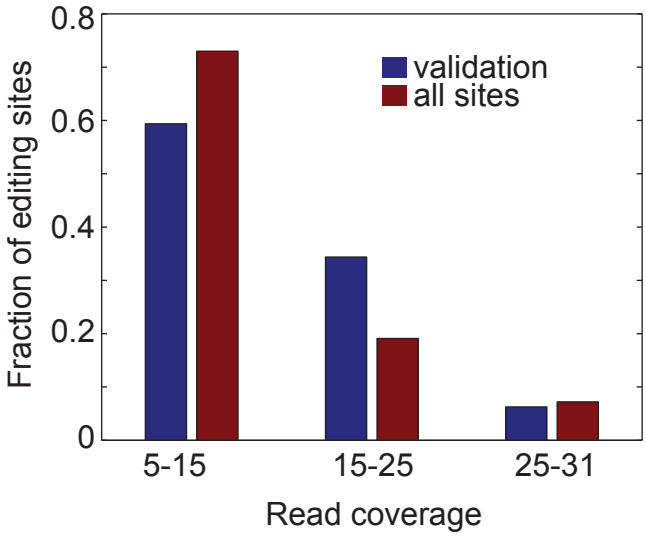
Supplemental Fig. 4. Number of DNA-RNA differences of each type.
A minimum "editing ratio" of 0.2 is required. Results from samples transfected with control siRNA and *ADAR* siRNA are shown.



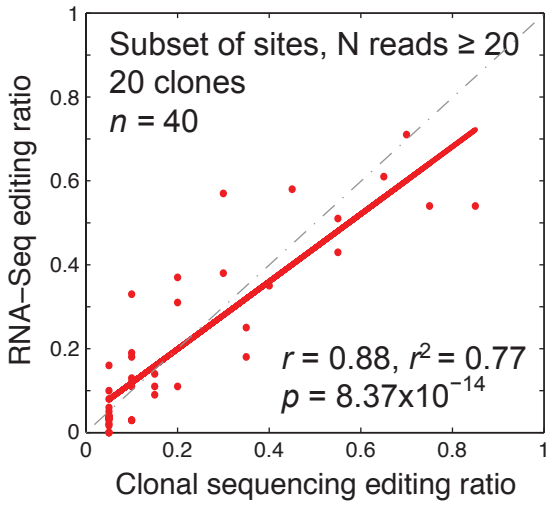
Supplemental Fig. 5. Empirical cumulative distribution function of "editing ratios" of each type of DNA-RNA differences estimated from RNA-Seq. A union of events identified in the two samples (control and *ADAR* knockdown) is included in each curve. For events absent in one sample (those that failed the statistical identification procedure), editing ratio in this sample was calculated as the number of reads with the edited base divided by the total number of reads at that position.



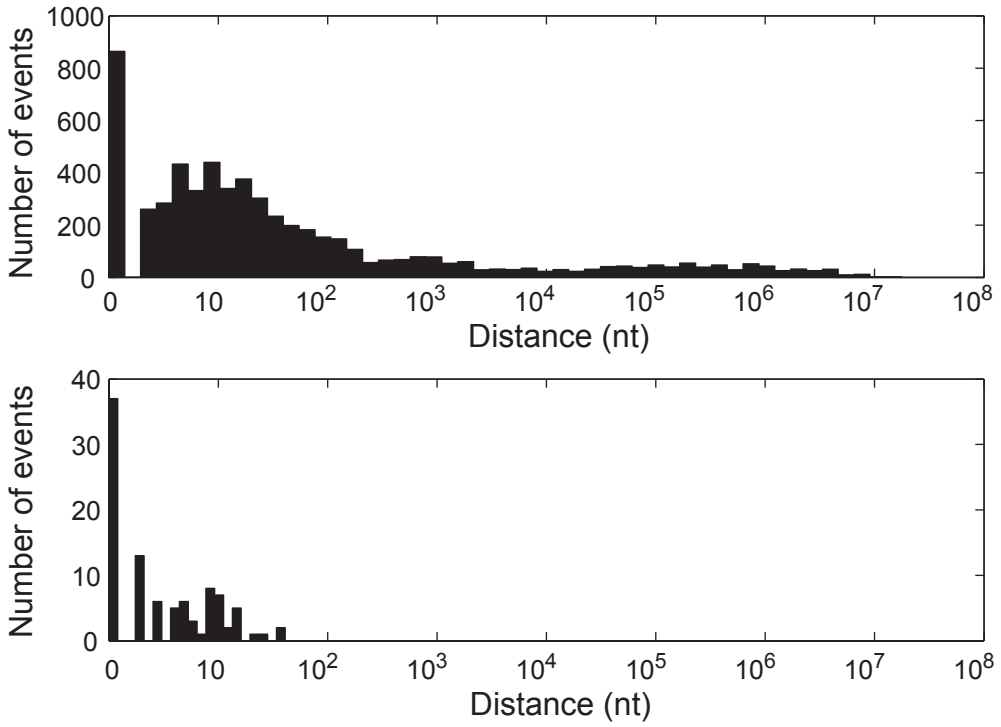
Supplemental Fig. 6. A-to-I editing in two genes validated by Sanger sequencing. Sequencing traces from both genomic DNA (gDNA) and cDNA amplifications are shown. Gene names and genomic coordinates (hg19) of editing sites are listed. The numbers below each editing site (underlined with black bar) in the cDNA traces are estimated editing ratios from RNA-Seq.



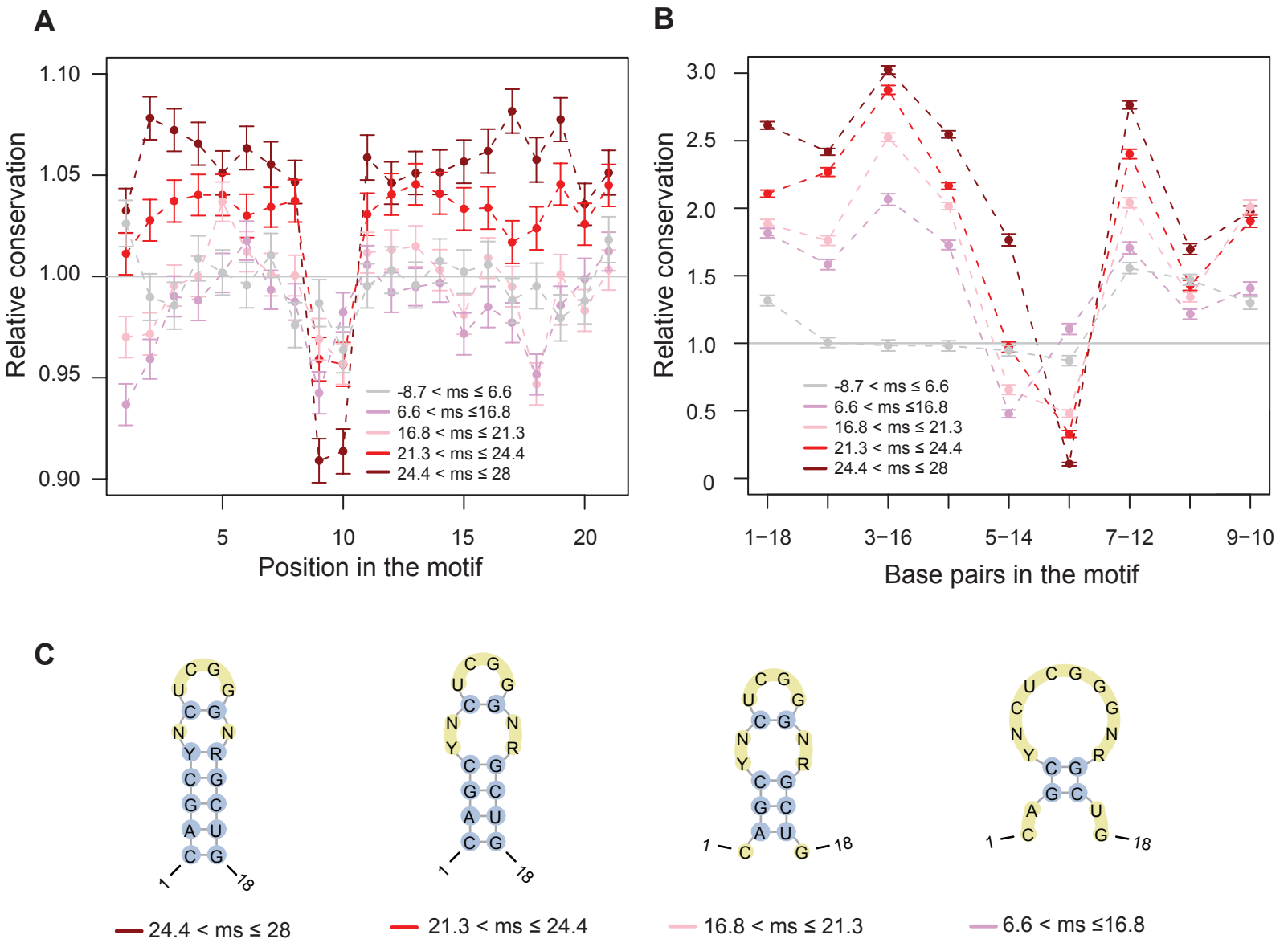
Supplemental Fig. 7. Distribution of read coverage of putative editing sites in the validation data set and all A-to-I sites (both filtered for ≤31 reads per site). X-axis shows number of reads per site.



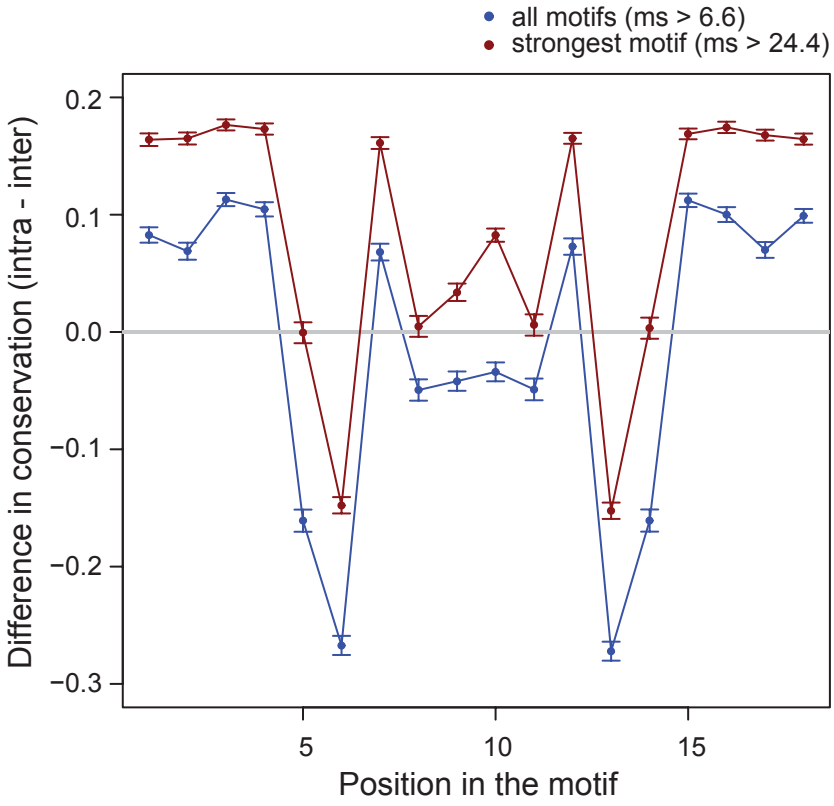
Supplemental Fig. 8. Validation of RNA-Seq-predicted editing sites via clonal sequencing. 40 sites with read coverage of at least 20; 20 clones were randomly picked for Sanger sequencing.



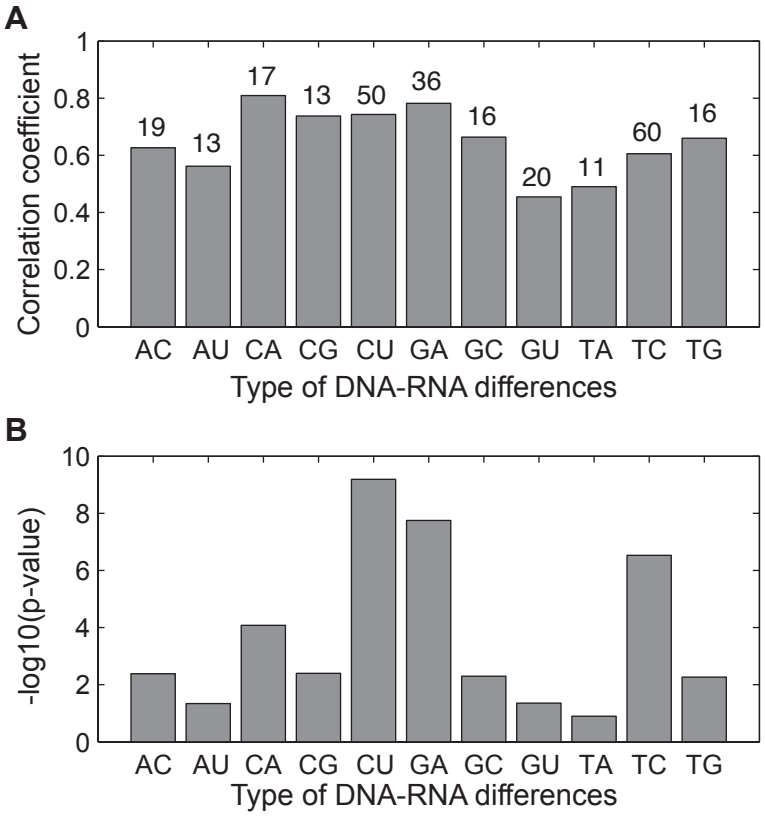
Supplemental Fig. 9. Distance of predicted A-to-I editing sites to their respective closest neighboring sites. (a) All predicted A-to-I editing sites in the U87MG cells with control siRNA transfection. **(b)** The 93 putative A-to-I editing sites subject to validation via clonal sequencing.



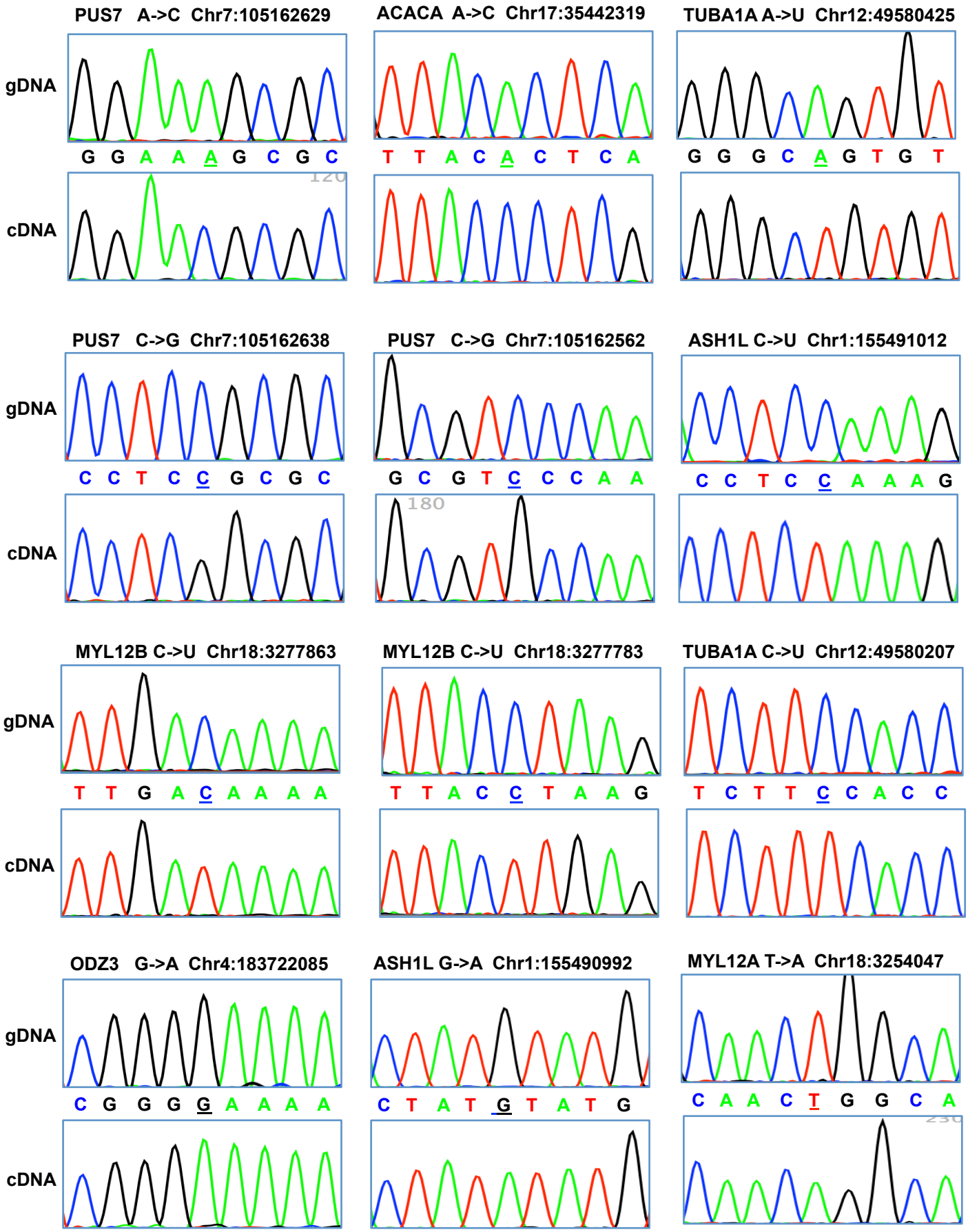
Supplemental Fig. 10. The new structural motif potentially related to A-to-I editing. **(A)** Sequence conservation of motifs grouped according to the motif score (ms). The motif is shown in Fig. 5A. X-axis represents nucleotide positions in the motif. Y-axis is the conservation level relative to random controls. Conservation was evaluated among primate genomes. Error bars represent 95% confidence intervals. **(B)** Structural conservation of the motifs in different groups. X-axis shows the base-pairing positions of the motif. Y-axis represents the co-conservation of each base pair relative to random controls among primates. **(C)** Consensus RNA secondary structures of each group of motifs. The group with the smallest motif score (-8.7 to 6.6) is not shown because they do not form secondary structures.

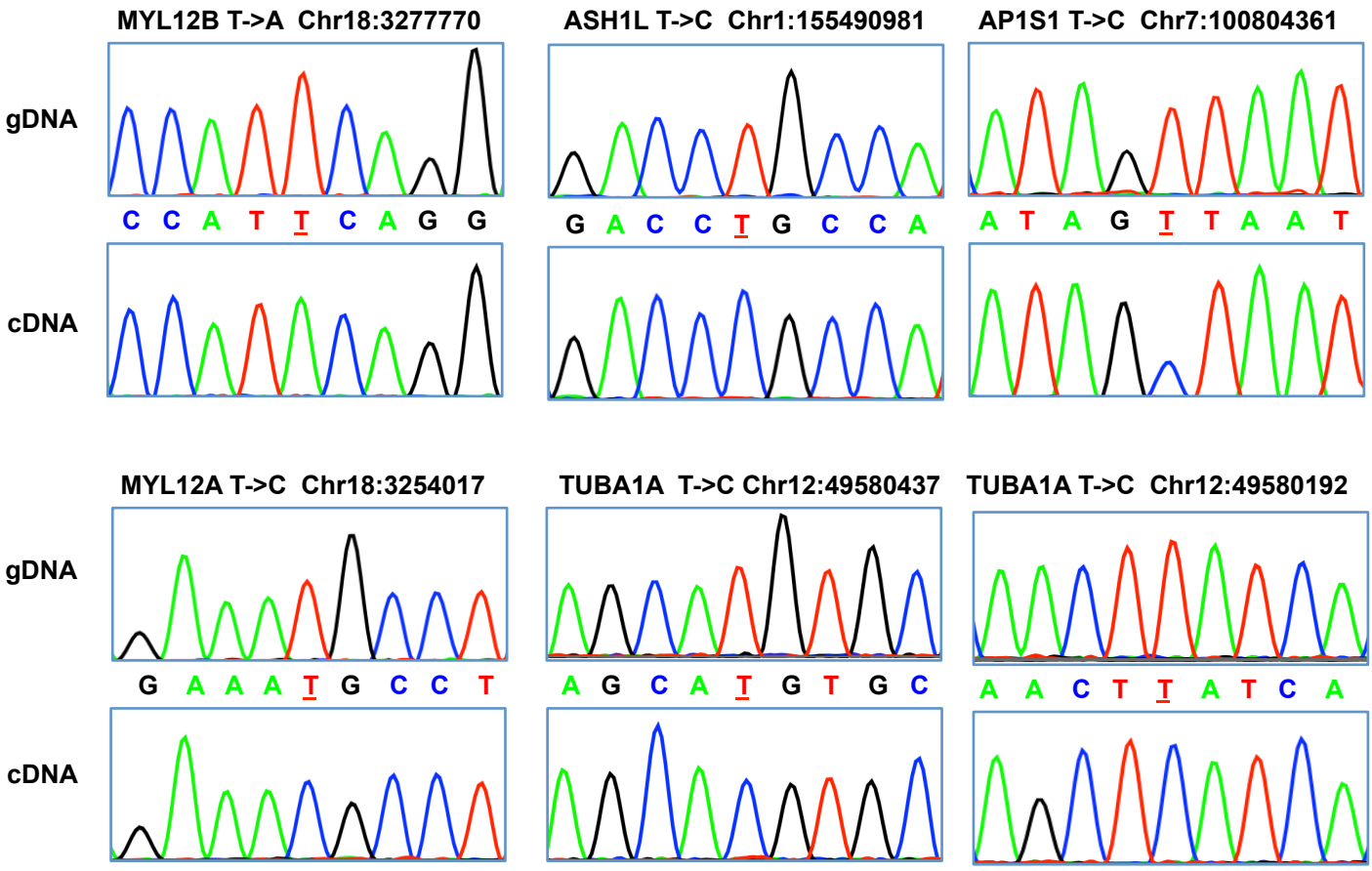


Supplemental Fig. 11. Comparison of base-pairing conservation patterns between positions within a motif (intra) and between two motifs in the same gene (inter). The motif is shown in Fig. 5A. X-axis shows the position of the base in the motif (For intra-motif conservation, the pairing positions were assigned the same conservation value). Y-axis represents the difference of conservation in primates (intra - inter). The analysis was performed for two different groups, 1) all motifs (motif score, ms > 6.6) and 2) strongest motifs (ms > 24.4).

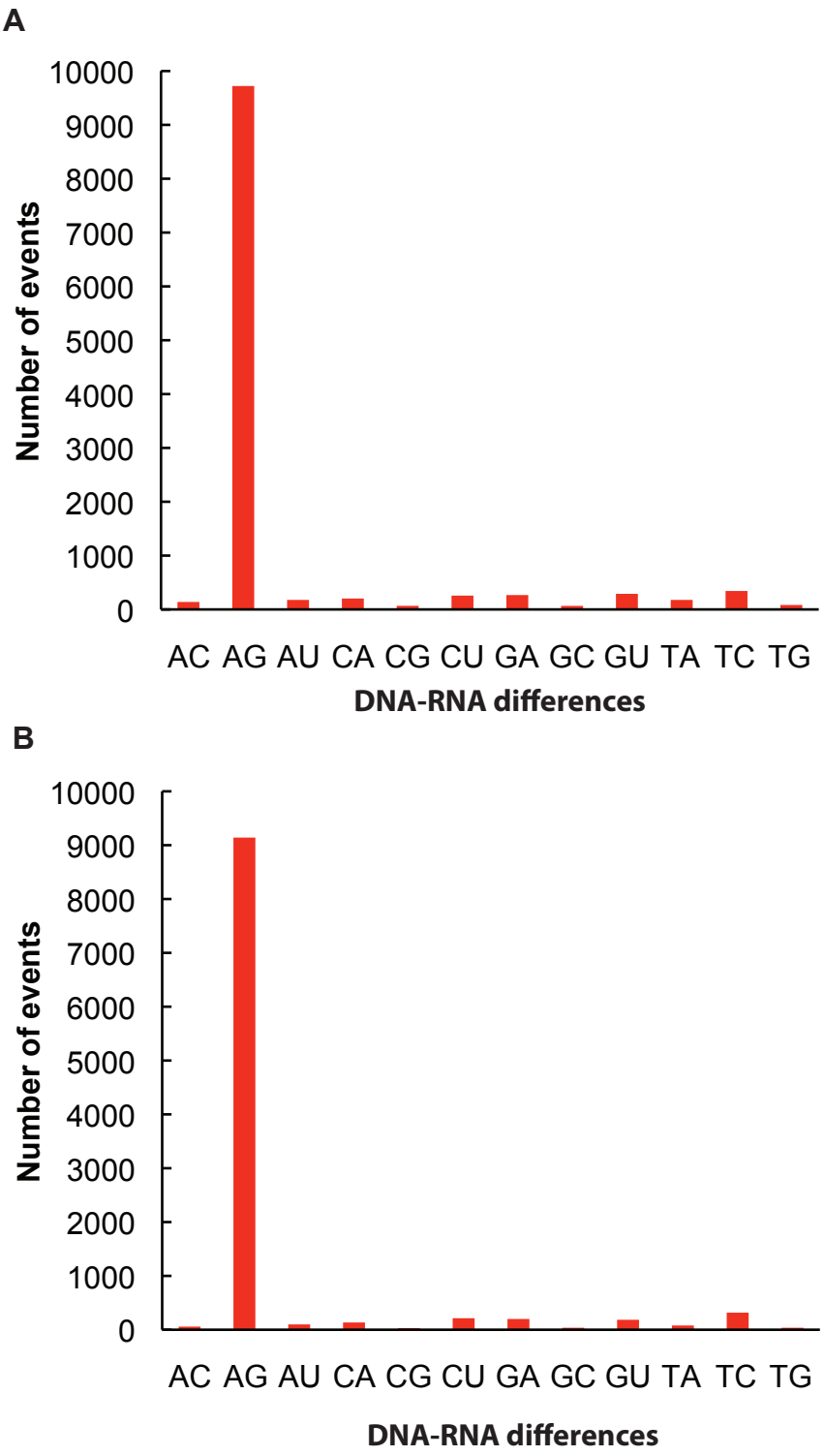


Supplemental Fig. 12. Correlation of "editing ratios" between A-to-I and other types of DNA-RNA differences in the control siRNA transfection sample. (A) Correlation coefficients of "editing ratios" for sites located in the same gene. If multiple sites of one type exist in the same gene, the average "editing ratio" was used. The number of genes in each group is shown above the bar. **(B)** P-values of the correlations in (A).

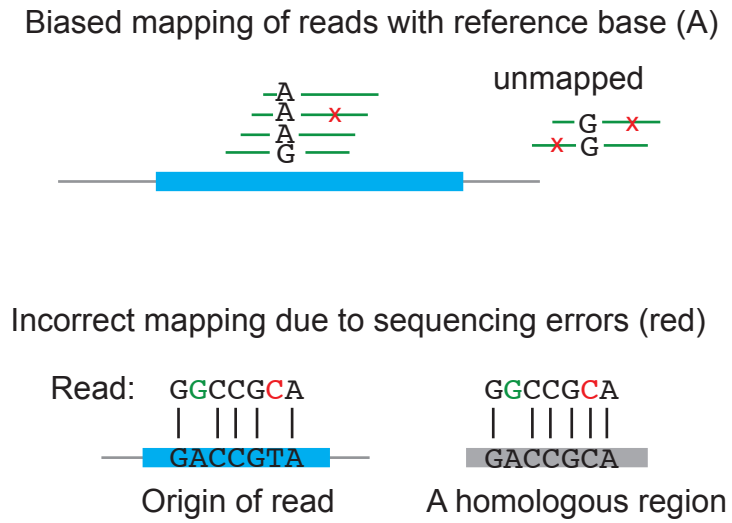

PUS7 A->C Chr7:105162629
ACACA A->C Chr17:35442319
TUBA1A A->U Chr12:49580425
gDNA
G G A A A G C G C
T T A C A C T C A
G G G C A G T G T
cDNA
120
PUS7 C->G Chr7:105162638
PUS7 C->G Chr7:105162562
ASH1L C->U Chr1:155491012
gDNA
C C T C C G C G C
G C G T C C C A A
C C T C C A A A G
cDNA
180
MYL12B C->U Chr18:3277863
MYL12B C->U Chr18:3277783
TUBA1A C->U Chr12:49580207
gDNA
T T G A C A A A A
T T A C C T A A G
T C T T C C A C C
cDNA
ODZ3 G->A Chr4:183722085
ASH1L G->A Chr1:155490992
MYL12A T->A Chr18:3254047
gDNA
C G G G G A A A A
C T A T G T A T G
C A A C T G G C A
cDNA



Supplemental Fig. 13. Sequencing traces (gDNA and cDNA) of the 18 non-A-to-G events confirmed by Sanger sequencing. cDNA traces were generated by cloning of RT-PCR products and Sanger sequencing. The trace of one clone that demonstrated the DNA-RNA difference is shown (and thus there are no multiple peaks at the event site). Gene name, event type and genomic coordinate (hg19) are shown. The gDNA and cDNA sequences are aligned and the nucleotide with DNA-RNA difference is underlined.



Supplemental Fig. 14. Number of DNA-RNA differences of each type in the breat caner samples. (A) All events **(B)** Events with with a minimum “editing ratio” of 0.2.



Supplemental Fig. 15. Two main problems in mapping of RNA-Seq reads for single-nucleotide analysis. Sequencing errors are represented by crosses or nucleotides in red. Upper: inaccurate estimation of editing levels due to suboptimal mapping. Reads carrying the A or the G nucleotide at a hypothetical site with A-to-G difference are shown. When mapped to the reference genome or transcriptome (with an A at this site), the reads with the A nucelotide are favored, whereas those with the G nucleotide are less likely to be mappable in the presence of sequencing errors. Lower: false positive predictions of editing events due to incorrect mapping. The read contains an edited nucleotide (in green) and another nucleotide (in red) affected by sequencing errors. It will map with two mismatches to the presumably correct genomic location, but may map to another homologous locus with only one mismatch.

Supplemental Table 1. Number of pairs of RNA-Seq reads and mapping results in the control U87MG cells and cells with *ADAR* knockdown.

Sample	Total	Uniquely mapped	Mapped to intergenic regions	Mapped to known genes	Mapped to known exons
Control	55,580,381	27,100,589	1,511,053	25,589,536	21,626,325
ADAR Knockdown	59,551,967	32,169,474	1,512,771	30,656,703	26,201,444

Supplemental Table 2. Number of events with DNA-RNA differences (read coverage > 10) identified using the two biological replicates from the control siRNA transfection experiment of U87MG cells.

Type	Biological replicate A	Biological replicate B	Common editing events
A → C	88	113	25
A → G	862	841	383
A → U	31	41	8
C → A	41	43	11
C → G	40	45	21
C → U	131	142	31
G → A	111	129	34
G → C	28	34	7
G → U	40	49	11
T → A	46	43	12
T → C	136	158	25
T → G	113	130	42

Supplemental Table 3. Number of DNA-RNA differences common to our study (U87MG cells) and other studies.

		Number of events overlap with U87MG results												
Reference	**# Sites in reference**	A →C	A →G	A →U	C →A	C →G	C →U	G →A	G →C	G →U	T →A	T →C	T →G	total
DARNED (Kiran and Baranov 2010)	42,045	0	841	0	0	0	0	0	0	0	0	13	0	854
(Blow et al. 2004)	1,727	0	8	0	0	0	0	0	0	0	0	0	0	8
(Li et al. 2009)	710	0	22	0	0	0	0	0	0	0	0	0	0	22
(Sakurai et al. 2010)	5,672	0	173	0	0	0	0	0	0	0	0	6	0	179
(Zaranek et al. 2010)	108,442	0	67	0	0	0	1	1	0	0	0	5	0	74
(Li et al. 2011)	10,210	6	25	1	3	1	3	10	0	3	1	5	15	73
(Ju et al. 2011)	1,809	7	148	0	1	0	2	3	0	0	2	5	4	172

Blow M, Futreal PA, Wooster R, Stratton MR. 2004. A survey of RNA editing in human brain. *Genome Res* **14**(12): 2379-2387.

Ju YS, Kim JI, Kim S, Hong D, Park H, Shin JY, Lee S, Lee WC, Yu SB, Park SS et al. 2011. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat Genet* **43**(8): 745-752.

Kiran A Baranov PV. 2010. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* **26**(14): 1772-1776.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**(5931): 1210-1213.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. 2011. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**(6038): 53-58.

Sakurai M, Yano T, Kawabata H, Ueda H, Suzuki T. 2010. Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. *Nat Chem Biol* **6**(10): 733-740.

Zaranek AW, Levanon EY, Zecharia T, Clegg T, Church GM. 2010. A survey of genomic traces reveals a common sequencing error, RNA editing, and DNA editing. *PLoS Genet* **6**(5): e1000954.

Supplemental Table 4. Primers used in the validation of editing sites (Fig. 3 and Supplemental Fig. 6; The names of those for results in Fig. 3 are in bold).

Primer name	**Sequence (5'-3')**
CTSB-Fw	TCAGCTTGTTGCTGGATTTGTGG
CTSB-Rv	GTGCCTGGTAAGCTTGCCCTTGC
CTSB-Seq	CTCAAAACATAGCTGTGCTC
CTSS-Fw	CACGTGTAGTGGCTCATGCCTG
CTSS-Rv	CCTGGTCAAATTGTTGATGACTACG
GNL3L-Fw	CTGAAGCATCTGCATATTGAAAGAACGC
GNL3L-Rv	CTGCCTCCCAGGTTCAGGCGATTCTCC
JUB-EcoRI-Fw	CCGAATTCCTACCTCTCAGGGACATCCCAC
JUB-EcoRI-Rv	CCGAATTCGTTCTGGTCCTGGATTCCAGG
JUB-Rv2	CCAGGCTAGTCTTGAACTC
JUB-Seq	GCCACTGTGTCTGGCCTAG
MDM2-EcoRI-Fw	CACGAATTCGGTACAGAGTTAAATTTGAAGG
MDM2-EcoRI-Rv	CCCGAATTCCATTAACTATAGTCACCCTAC
MDM2-Fw2	GCAATGATCTAGAAGCAGATG
MDM2-Seq	CAATGTATGGGTAGAACATG
NUP43-Fw	GTTTTGGATCCCACTCACACAC
NUP43-Rv	GTACCTTGGTGATCATCTCCAG
NUP43-Seq	CATCATCCCTTCTATGGTATTC

Supplemental Table 5. Validation of putative A-to-I editing sites that are not clustered with other predicted A-to-I sites.

Coordinate (hg19)	Shortest distance to other A-to-I sites	Editing ratio	Validation method	Validated?
chr3:124731926	7,801	0.02	Clonal Seq	Yes
chr14:53239043	4,351	0.09	Clonal Seq	Yes
chr12:124497392	350,975	0.02	Clonal Seq	Yes
chr3:58260496	43,782	0.02	Clonal Seq	Yes
chr14:103800877	149,686	0.04	Clonal Seq	Yes
chr12:54796788	1,105,643	0.06	Clonal Seq	Yes
chr3:188596827	1,859	0.25	Sanger Seq	Yes
chr2:176789478	1,703	0.53	Sanger Seq	Yes
chr19:55900533	104,211	0.78	Sanger Seq	Yes
chr17:40008468	25,868	0.67	Sanger Seq	Yes

Supplemental Table 6. List of A-to-G events identified in the U87MG cells with a minimum editing ratio of 0.2

Supplied as a separate file due to large table size

Supplemental Table 7. Distribution of DNA-RNA differences (U87MG data) in different types of genomic regions. Percentages are shown in bold if they are the maximum value of a row or a column. "Intergenic" regions were present because we extended the gene boundaries by 1kb each upsteam and downstream of the annotated gene ends, to accommodate cases when the 5' or 3' ends of the gene were not annotated accurately. "Noncoding" category includes the non-coding genes and non-coding transcripts of coding genes.

Type	Total	Coding transcripts				Noncoding	Intergenic
		Coding	Introns	5' UTR	3' UTR		
A→G	4,141	45 1.1%	2,015 **48.7%**	45 1.1%	1,293 31.2%	485 11.7%	258 6.2%
A→C	94	31 **33.0%**	9 9.6%	4 4.3%	38 **40.4%**	5 5.3%	7 7.4%
A→U	48	4 8.3%	16 33.3%	0 0.0%	22 **45.8%**	4 8.3%	2 4.2%
C→A	57	6 10.5%	16 28.1%	2 3.5%	24 **42.1%**	1 1.8%	8 **14.0%**
C→G	50	9 18.0%	11 22.0%	12 **24.0%**	13 **26.0%**	2 4.0%	3 6.0%
C→U	173	26 15.0%	45 26.0%	5 2.9%	64 **37.0%**	18 10.4%	15 8.7%
G→A	149	18 12.1%	46 30.9%	8 5.4%	46 **30.9%**	20 13.4%	11 7.4%
G→C	51	9 17.6%	14 **27.5%**	7 13.7%	11 21.6%	8 15.7%	2 3.9%
G→U	73	9 12.3%	24 32.9%	3 4.1%	31 **42.5%**	2 2.7%	4 5.5%
T→A	54	6 11.1%	16 29.6%	2 3.7%	23 **42.6%**	4 7.4%	3 5.6%
T→C	506	42 8.3%	239 **47.2%**	9 1.8%	48 9.5%	123 **24.3%**	45 8.9%
T→G	109	28 25.7%	19 17.4%	10 9.2%	39 **35.8%**	10 9.2%	3 2.8%

Supplemental Table 8. Motif enrichment near predicted A-to-I editing sites in non-Alu regions.

Motif score (ms) cutoff	**Number of editing sites in non-Alu regions with motif**	**Mean of number of motifs in the random sets**	**P-value**
ms > 6.6	51	56.25	0.755
ms > 16.8	21	7.71	2.047×10^{-7}
ms > 21.4	15	5.09	3.082×10^{-6}
ma > 24.4	6	2.71	0.02

Supplemental Table 9. Examples of genes with A-to-I editing in U87MG involved in cancer-related processes and pathways

Biological function	**Gene names**
Tumor suppressor and cancer marker protein	*BLCAP, CTAGE5, DLC1, DPH1, NF2, PFDN5, PLCD4, PTEN, TPD52L2, UXT*
Apoptosis regulation and induction	*BAX, CASP8, CD70, CFLAR, DAP3, DAPK3, DFFA, DLC1, PDCD5, STAT3, TEP1, TNFRSF9, XIAP*
Metastasis	*ADAM19, CD151, FN1, HPSE, LAMB1, MMP15, MMP17, MTA1, RHOC*
DNA repair	*CUL4C, DCLRE1C, DDB2, DDX52, ERCC4, ERCC6, NEIL2, XPA, XPC, XRCC2*
p53 pathway	*FHL2, GNL3, MDM2, MDM4, TP53BP1*
NF-κB pathway	*CARD8, COPS5, IRAK4, MALT1, RELA*
MAPK signaling pathway	*AKT2, MAPK2K2, MAPK3, MAPK8, MAPK8IP3, MAPK9, MAPK14, NF2*

Supplemental Table 10. Co-occurrence of other types of DNA-RNA differences with the predicted A-to-G events in the same gene (1,167 genes with predicted A-to-G events)

Type	# genes	# genes also with A-to-G events	P- value
A→C	91	19	1.79×10^{-5}
A→U	45	13	4.69×10^{-6}
C→A	56	17	1.19×10^{-7}
C→G	47	13	8.28×10^{-6}
C→U	155	53	$< 10^{-17}$
G→A	123	39	1.55×10^{-15}
G→C	49	17	1.10×10^{-8}
G→U	66	20	1.31×10^{-8}
T→A	50	11	3.54×10^{-4}
T→C	105	20	4.99×10^{-5}
T→G	258	62	1.11×10^{-16}

Supplemental Table 11. Validation of all types of DNA-RNA differences other than the A-to-G type. "Editing ratio" is used for convenience although these events may not be resulted from RNA editing mechanisms.

Type of DNA-RNA difference	Genomic coordinate (hg19)	Editing ratio in RNA-Seq	Gene	Validated?
A→C	chr7:105162629	0.37	PUS7	YES
	chr17:35442319	0.1	ACACA	YES
	chr1:155490984	0.33	ASH1L	NO
A→U	chr12:49580425	0.3	TUBA1A	YES
	chr12:54796109	0.14	ITGA5	NO
C→A	chr3:58260537	0.03	ABHD6	NO
	chr3:58260464	0.18	ABHD6	NO
C→G	chr7:105162638	0.37	PUS7	YES
	chr7:105162562	0.4	PUS7	YES
	chr14:103800841	0.36	EIF5	NO
C→U	chr1:155491012	0.1	ASH1L	YES
	chr14:103800775	0.05	EIF5	NO
	chr18:3277863	0.27	MYL12B	YES
	chr18:3277783	0.33	MYL12B	YES
	chr12:49580207	0.05	TUBA1A	YES
G→A	chr4:183722085	0.2	ODZ3	YES
	chr1:155490992	0.03	ASH1L	YES
	chr17:35442228	0.02	ACACA	NO
	chr12:124497299	0.02	ZNF664	NO
G→C	chr3:58260533	0.05	ABHD6	NO
G→U	chr14:53239034	0.7	STYX	NO
	chr13:98652834	0.18	IPO5	NO
T→A	chr18:3254047	0.06	MYL12A	YES
	chr18:3277770	0.4	MYL12B	YES
	chr12:9102082	0.86	M6PR	NO
T→C	chr1:155490981	0.04	ASH1L	YES
	chr14:53239029	0.3	STYX	NO
	chr12:124497250	0.03	ZNF664	NO
	chr7:100804361	0.01	AP1S1	YES
	chr18:3254017	0.33	MYL12A	YES
	chr12:49580437	0.17	TUBA1A	YES
	chr12:49580192	0.67	TUBA1A	YES
	chr3:124731939	0.21	HEG1	NO
T→G	chr17:35442327	0.39	ACACA	NO
	chr17:35442354	0.03	ACACA	NO
	chr17:35442372	0.03	ACACA	NO
	chr12:124497385	0.02	ZNF664	NO

Supplemental Table 12. List of A-to-G events identified in the human primary breast cancer data with a minimum editing ratio of 0.2

Supplied as a separate file due to large table size

Supplemental Table 13. Comparison of the numbers of DNA-RNA differences identified in the U87MG and breast cancer samples that are included in the background set (defined as all genomic homozygous sites in known genes that are common to the two data sets and with at least 5 RNA-Seq reads). P-values calculated by the hypergeometric test.

Type	# in background set	# in U87MG	# in breast cancer	# Overlap	P-value
A → C	7,701,752	101	52	2	2.26×10^{-7}
A → G	7,701,752	1,883	1,004	351	$< 2.16 \times 10^{-16}$
A → U	7,701,752	95	39	4	1.79×10^{-15}
C → A	6,322,426	94	59	1	0.0009
C → G	6,322,426	55	28	1	0.0002
C → U	6,322,426	247	33	3	3.21×10^{-10}
G → A	6,565,155	197	47	1	0.001
G → C	6,565,155	69	21	0	1
G → U	6,565,155	95	75	2	5.75×10^{-7}
T → A	7,664,306	97	42	2	1.36×10^{-7}
T → C	7,664,306	189	42	5	$< 2.16 \times 10^{-16}$
T → G	7,664,306	133	31	1	0.0005