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

Genome Research **22** (2012) 142–150.

Speaker: Joseph Chuang-Chieh Lin

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

29 February 2012

Outline

Introduction

- Use transcriptome sequencing data (RNA-seq) for global identification of RNA editing.
- The RNA-seq data:
  a human glioblastoma cell line: U87MG.
  - Samples are transfected with either a siRNA that targets the ADAR gene or a control siRNA.

Introduction

- Use transcriptome sequencing data (RNA-seq) for global identification of RNA editing.
- The RNA-seq data:
  a human glioblastoma cell line: U87MG.
  - Samples are transfected with either a siRNA that targets the ADAR gene or a control siRNA.

Introduction (contd.)

- 9,636 DNA-RNA differences (RDDs) were identified, and 62% (5,965) are putative A-to-I editing sites.
- Estimation editing levels from RNA-seq correlated well with those based on traditional clonal sequencing.
- Genes with predicted A-to-I editing were significantly enriched with those known to be involved in cancer.
- Similar results are obtained from primary breast cancer samples despite their difference in cell type, cancer type, and genomic backgrounds.

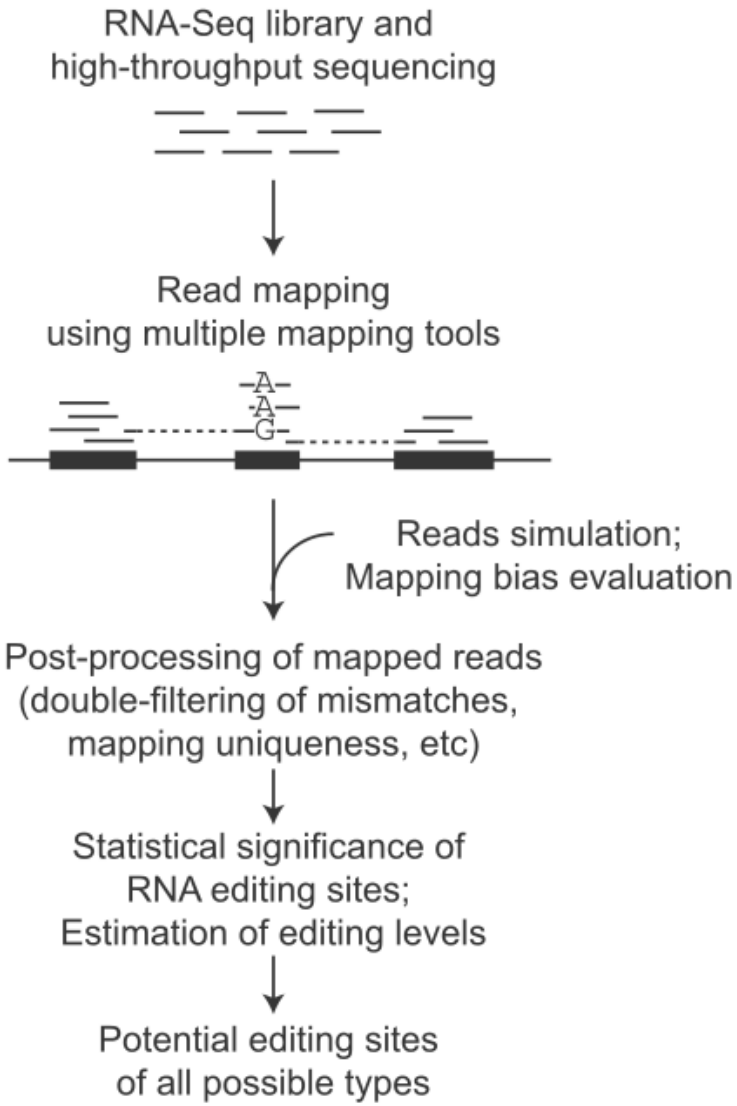
Restrictions of previous bioinformatic methods

Identify disparities between DNA and RNA sequences by analyzing cDNA, EST, and gDNA.

- Require priori knowledge of editing patterns to restrain the search.
  - The feature of clustering of putative editing sites;
  - The presence of dsRNA structure;
  - ...
  - ⋆ However, incorporation of such constraints often *limits* the results to editing sites with the corresponding characteristics.
- The estimation of RNA editing levels is usually not afforded.

Outline

Identification of RNA-editing sites



Outline

Reads mapping

- Map each end of the paired-end reads to hg19 genome using a combination of tools (∵ they could differ significantly for some reads):
  - Nowtie, BLAT, TopHat.
  - Exon-exon junction allowed: BLAT and TopHat.
  - The mapping parameters are given in the paper (p. 149).

Reads mapping (contd.)

- Initial mapping: ≤ 12 mismatches in each 60-nt read.
- All mappings of each pair of reads were examined to determine **if they pair correctly** (with the expected orientation & the distance between the pair being $< 500{,}000$ bp in the genome).
- Require that the pair of reads:
  - map uniquely (as a pair, not necessarily individually) with ≤ 5 mismatches on each reads,
  - do NOT map to anywhere else in the genome as a pair with ≤ 12 mismatches.

Outline

Identification of RNA editing sites (I)

- For homozygous sites derived from the U87MG genome sequencing data,
  - pile up reads overlapping these sites;
  - examine whether mismatches to the genome sequence existd in the RNA reads;
  - Remove all duplicate reads within each RNA-seq library.
    - ∵ amplication bias in the RT-PCR process ⇒ for the accuracy of the estimated editing ratio.

Identification of RNA editing sites (I)

- For homozygous sites derived from the U87MG genome sequencing data,
  - pile up reads overlapping these sites;
  - examine whether mismatches to the genome sequence existd in the RNA reads;
  - Remove all duplicate reads within each RNA-seq library.
    - ∵ amplication bias in the RT-PCR process ⇒ for the accuracy of the estimated editing ratio.

Identification of RNA editing sites (II)

- Infer the strand of the reads based on the strand of genes they were mapped to.
  - Reads mapped to regions with bidirectional transcription (sense & antisense gene pairs) were discarded.
  - for comprehensive gene annotation: Ensembl, RefSeq, UCSC KnownGenes, Gencode genes, and VegaGenes.
  - Extend teh gene boundaries by 1kb each beyond the two ends.

Identification of RNA editing sites (III)

- A statistical approach to see whether RDDs are likely authentic.
- Calculate the prob. of observing the specific nucleotide (n) for A-to-I editing assuming that
  - the site is *edited* with the true editing ratio r;
  - the quality score of the observed n is q;
  - the position of n in the read is i.

$$\mathbf{Pr}[n \mid r, q, i] = \mathbf{Pr}[n \mid \text{freq}(A) = 1 - r, \text{freq}(G) = r, q, i].$$

- Assume that q and i affect the likelihood of a base-call being a sequencing error (similar to the approach used by SNP calling algorithm by Li & Durbin 2009; Li et al. 2009).
- The optimal r: the one maximizing the above function.

Identification of RNA editing sites (III contd.)

- LLR to evaluate the significance of a predicted event:
- $$\mathrm{LLR} = \log_{10}\left(\frac{\max_r\{\mathbf{Pr}[n \mid r, q, i]\}}{\mathbf{Pr}[n \mid r = 0, q, i]}\right).$$
  - ⋆ $r = 0$: not editing.
- Use LLR ≥ 2.
  - Indicating that the site is 100 times more likely being a true locus with RDD than a result of sequencing error.
- Require ≥ 2 edited reads and ≥ 5 reads in total for each considered site.
- Mismatches within the first and last five bases of a read were discarded.

Outline

Evaluation of mapping bias

Relatve ratio:

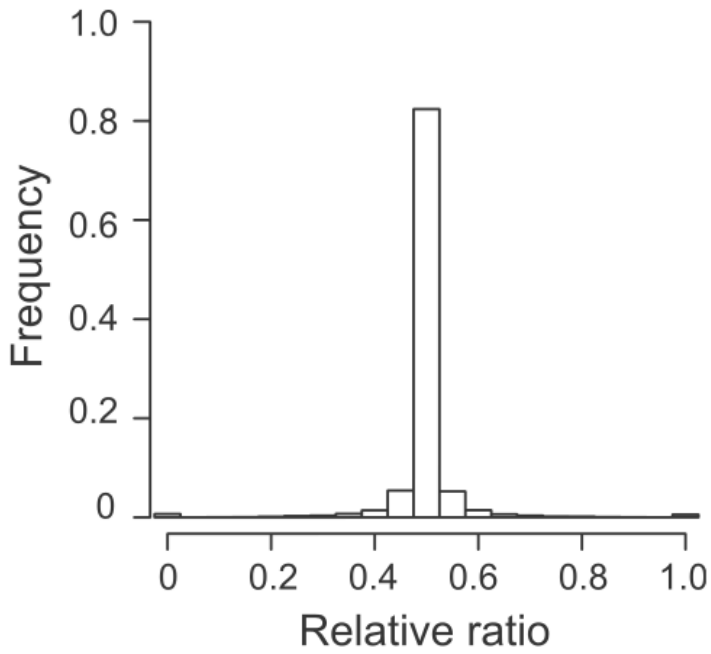
$$\frac{\frac{N_{\text{mapped\_ref}}}{N_{\text{simulated\_ref}}}}{\frac{N_{\text{mapped\_ref}}}{N_{\text{simulated\_ref}}} + \frac{N_{\text{mapped\_edit}}}{N_{\text{simulated\_edit}}}} := \frac{\alpha}{\alpha + \beta}.$$

Hence,

$$\frac{\alpha}{\alpha + \beta} = \frac{1}{2} \Rightarrow \alpha : \beta = 1 : 1$$

That is,

$$\frac{N_{\text{mapped\_ref}}}{N_{\text{mapped\_edit}}} = \frac{N_{\text{simulated\_ref}}}{N_{\text{simulated\_edit}}}.$$

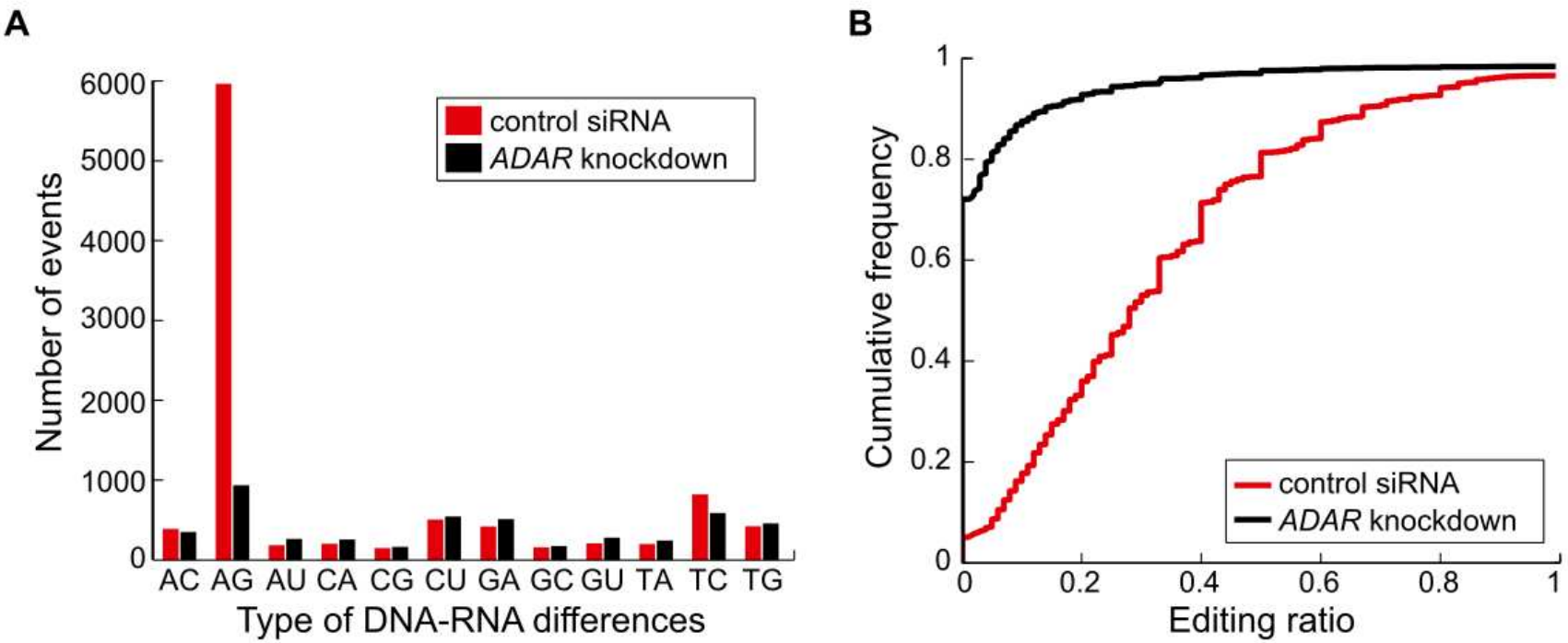


Evaluation of mapping bias (contd.)

- Simulate 870,280 reads (60nt in length) covering 21,757 heterozygous genomic sites assumed to have alternative alleles (1:1 ratio).
- 40 pairs of reads were generated to overlap each genomic site with a random (uniformly) insert size in the range of $[60, 240]$ bp and random start position relative to the site.
- The base at the heterozygous site was chosen as one of the alternative alleles with *equal probability*.

Outline

RDD identified via RNA-seq



Sanger sequencing of gDNA and cDNA & PCR

- ⋆ gDNA sequencing: confirm that it's not a heterozygous SNP.
- ⋆ cDNA sequences: enable detection of edited nucleotides.
- However, cDNA is not sensitive and quantitive enough to detect low-level editing or to provide accurate estimates of editing ratios (?).
- Instead, the traditional clonal sequencing approach is used to analyze the cDNA sequences and PCR sequencing is only used to confirm the gDNA sequences only.
- Four genes were randomly picked where a number of A-to-I editing sites are located within 400 bases.
  - Their cDNA sequences were amplified and cloned into a TOPO vector.
  - 20 clones for each gene were randomly picked and analyzed by Sanger sequencing.

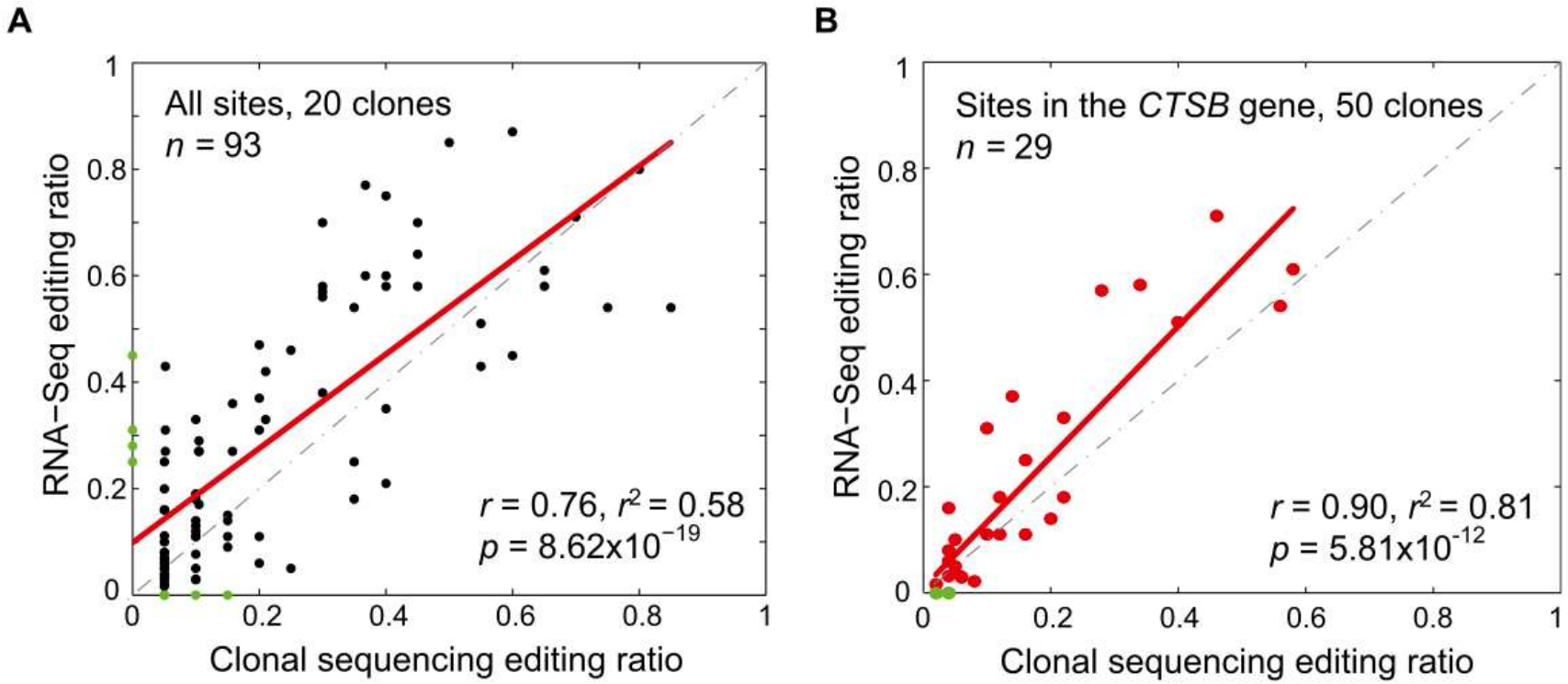
Sanger sequencing of gDNA and cDNA & PCR

- ⋆ gDNA sequencing: confirm that it's not a heterozygous SNP.
- ⋆ cDNA sequences: enable detection of edited nucleotides.
- However, cDNA is not sensitive and quantitive enough to detect low-level editing or to provide accurate estimates of editing ratios (?).
- Instead, the traditional clonal sequencing approach is used to analyze the cDNA sequences and PCR sequencing is only used to confirm the gDNA sequences only.
- Four genes were randomly picked where a number of A-to-I editing sites are located within 400 bases.
  - Their cDNA sequences were amplified and cloned into a TOPO vector.
  - 20 clones for each gene were randomly picked and analyzed by Sanger sequencing.

Sanger sequencing of gDNA and cDNA & PCR

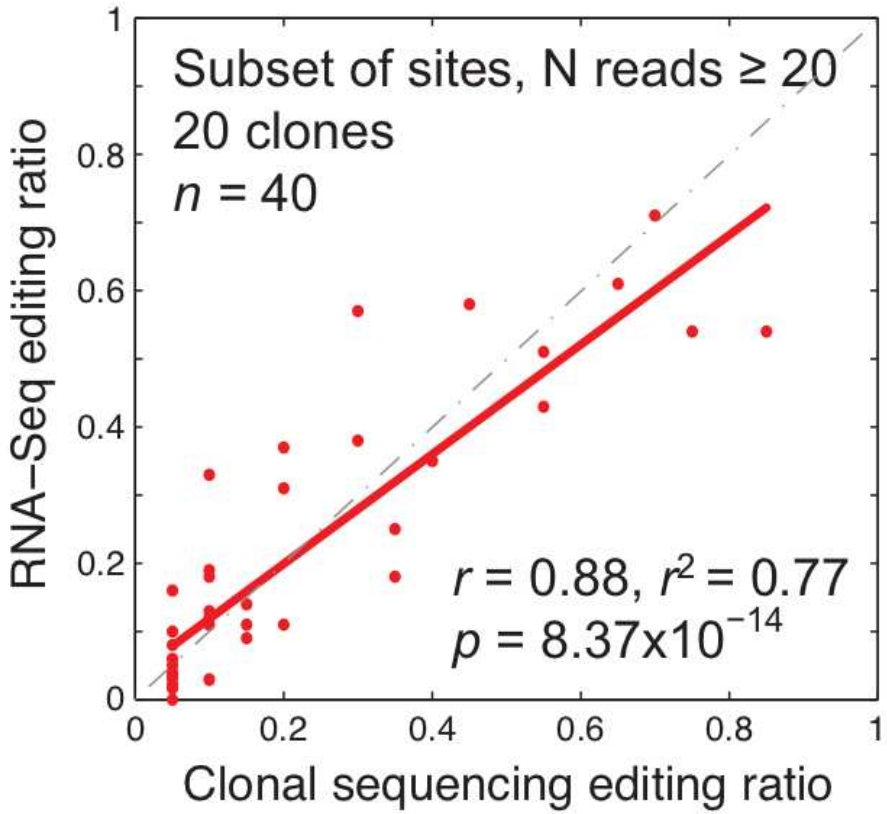
- ⋆ gDNA sequencing: confirm that it's not a heterozygous SNP.
- ⋆ cDNA sequences: enable detection of edited nucleotides.
- However, cDNA is not sensitive and quantitive enough to detect low-level editing or to provide accurate estimates of editing ratios (?).
- Instead, the traditional clonal sequencing approach is used to analyze the cDNA sequences and PCR sequencing is only used to confirm the gDNA sequences only.
- Four genes were randomly picked where a number of A-to-I editing sites are located within 400 bases.
  - Their cDNA sequences were amplified and cloned into a TOPO vector.
  - 20 clones for each gene were randomly picked and analyzed by Sanger sequencing.

Sanger sequencing of gDNA and cDNA & PCR (contd.)



FDR (false-discovery rate): $4/(93-4) \approx 4.5\%$.

Sanger sequencing of gDNA and cDNA & PCR (contd.)

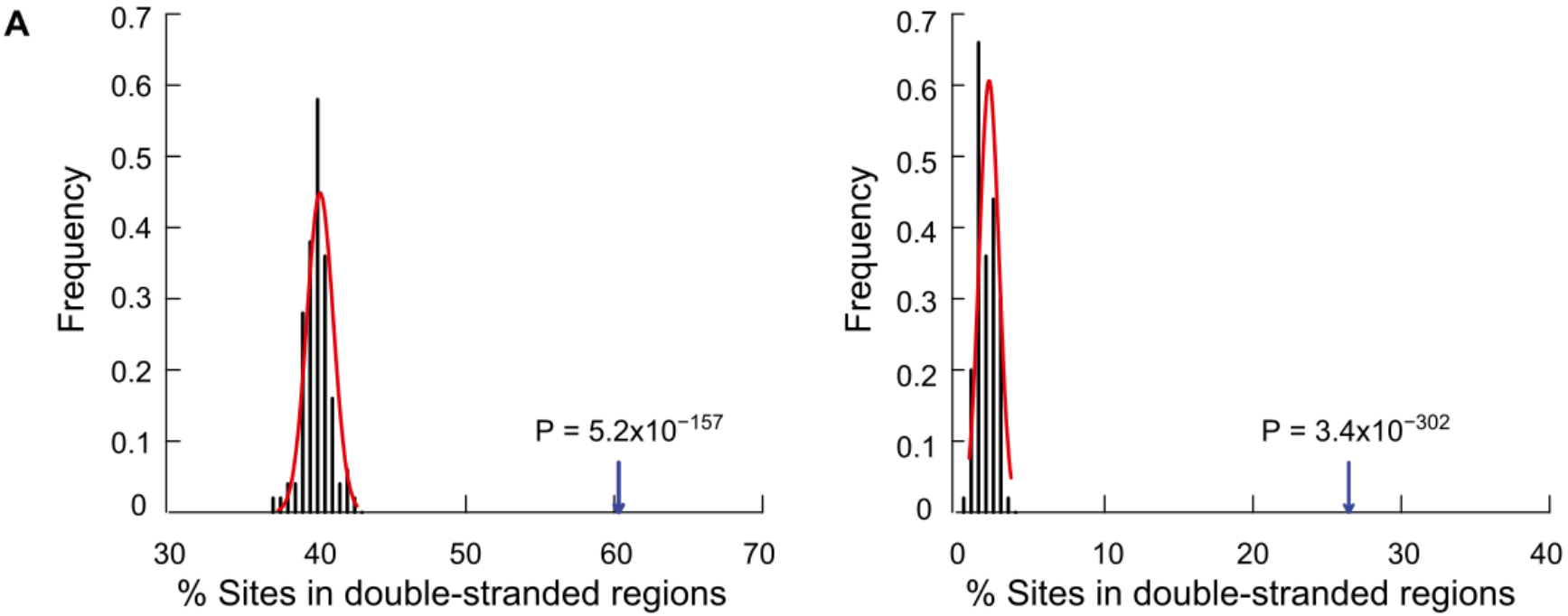


Characterization of predicted A-to-I editing events

- Consider 4,141 A-to-I editing sites with $\geq 20\%$ editing level identified from the control siRNA samles.

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%

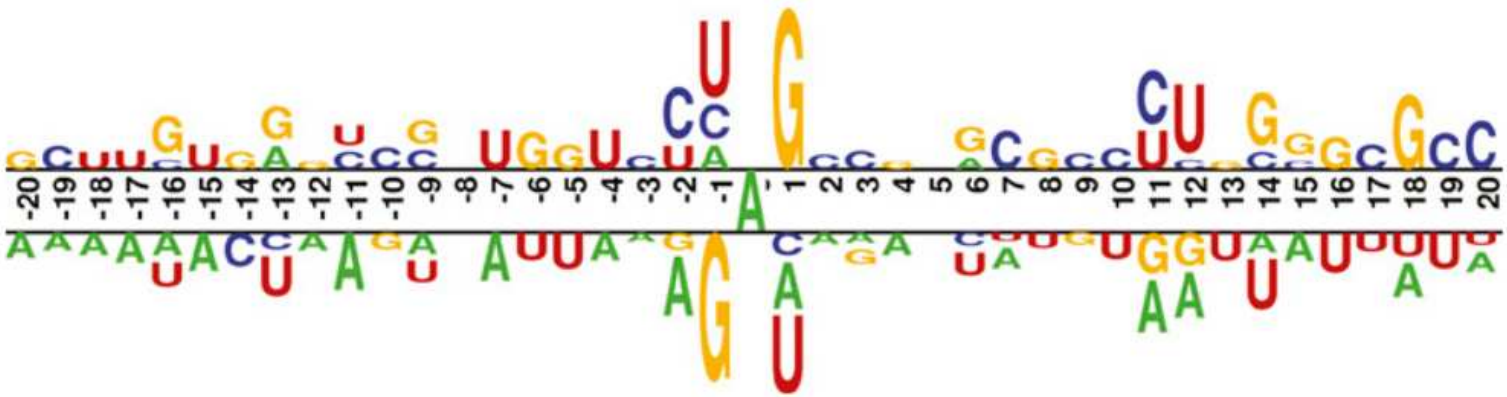
Characterization of predicted A-to-I editing events (contd.)



In *Alu* elements vs. outside of Alu elements.

Characterization of predicted A-to-I editing events (contd.)

B

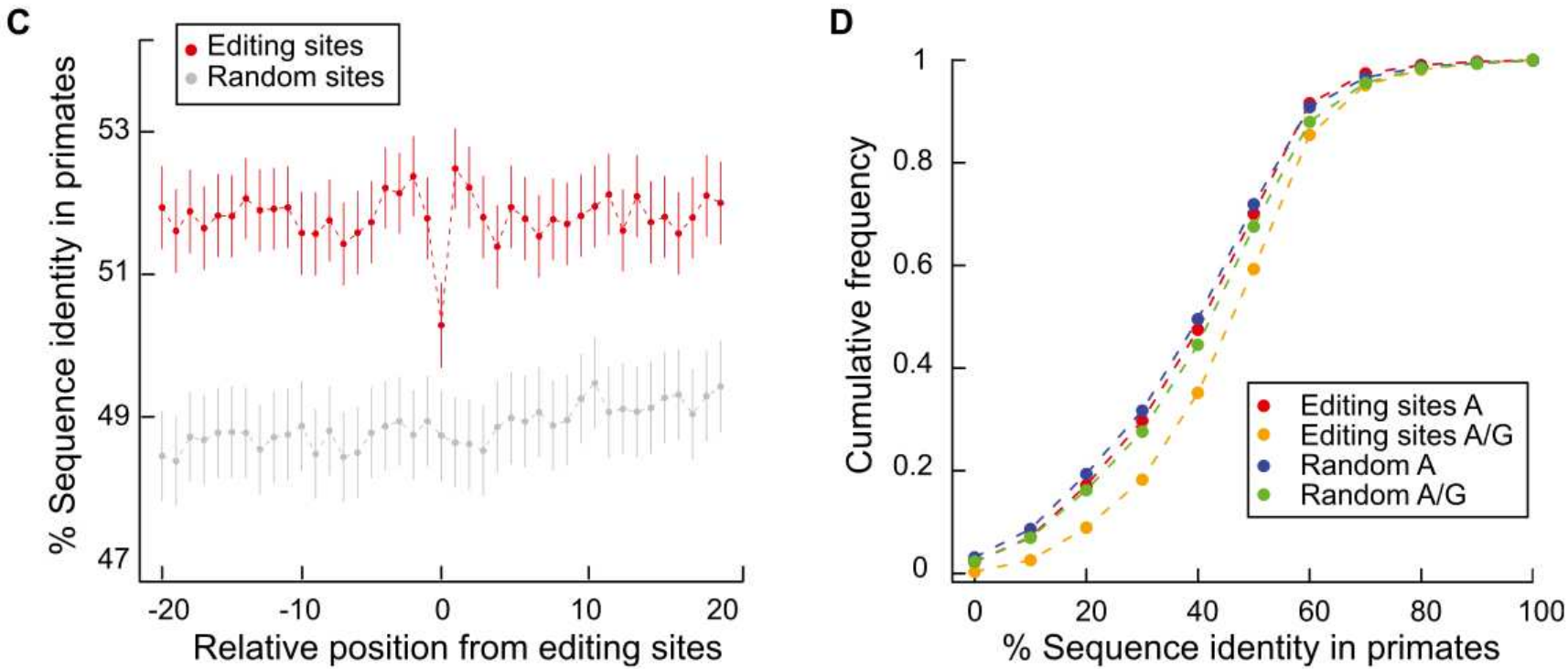


Motifs near editing sites far away from *Alus*

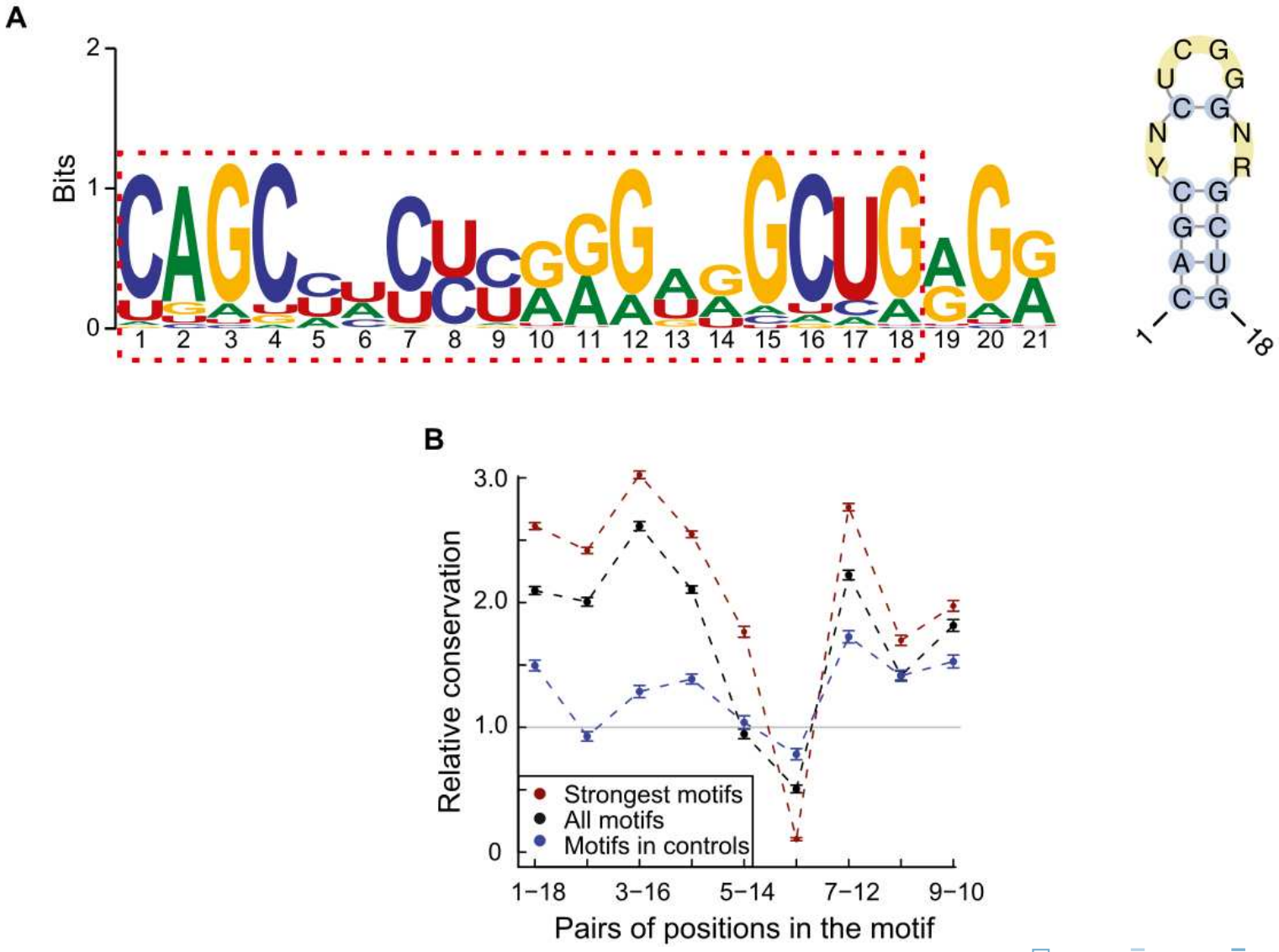
Supplemental Table 8. Motif enrichment near predicted A-to-I editing sites in non-Alu regions.

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

Conservation of neighborhood of predicted A-to-I editing sites



A structural motif in ADAR editing



Other types of DNA-RNA differences

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}

Other types of DNA-RNA differences (contd.)

- Regions with unknown sense-antisense transcription may lead to confusion of an actual A-to-G events as T-to-C events, vice versa.
- Indeed, if most T-to-C events were resulted from A-to-I editing on the opposite strand, then they are expected to be as highly enriched in *Alus* as the A-to-G events.
- Yet, 63% of T-to-C events occur in *Alus*, significantly lower than the 88% among A-to-G events ($p < 1 \times 10^{-10}$).

Outline

Discussion

- It is still possible to have false-positive prediction due to sequencing or mapping errors.
  - Mapping errors arise due to highly homologous regions in mammalian genomes.
- Increased read coverage at putative editing sites enable better accuracy in the esitmation of editing ratios.

Discussion (contd.)

- The predicted A-to-I editing sites are often associated with lower genomic conservation compared with their flanking regions.
- However, changing the A to I (G) via editing *increases sequence conservation* in primates.
- G-to-A genomic mutationss may be corrected by RNA editing.

Discussion (contd.)

- Editing levels of the A-to-I editing sites tend to be relatively low (mean, 0.35; median, 0.33).
- Among all 5,965 A-to-G sites in U87MG cells,
  - 31%: editing level ≤ 0.2;
  - 5%: editing level ≥ 0.8.
  - ▷ Consistent with the continuous probing (COP) hypothesis (Gommans *et al.* 2009).
    - Low-level editing is prevalent due to COP of the transient and dynamic RNA secondary structures by the editing machinary.

Thank you.