Visual map of A-to-I editing sites on human long non-coding RNAs.

David S. Fischer[1,2]

[1]Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom
[2]UTRIP Program, Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, Japan

RNA editing increases transcriptome complexity and regulates gene expression. Adenosine deaminases acting on RNA (ADARs) are RNA-editing enzymes that convert adenosine (A) to inosine (I) on stretches of double-stranded RNA (dsRNA). In humans, the ADAR family contains three active isoforms with different subcellular localizations: ADAR1-p150 (cytoplasmic), ADAR1-p110 (nuclear), and ADAR2 (nuclear). Previous sequencing data in our lab included size-fractionated RNAs that co-immunoprecipitated with ADAR isoforms overexpressed in human HeLa cells. Interestingly, this revealed that long non-coding RNAs (lncRNAs) are only targeted by the nuclear isoforms ADAR1-p110 and ADAR2, but not by cytoplasmic ADAR1-p150 (Ishii R. *et al.*, unpublished). We estimated that at least one third of human lncRNAs are produced from computationally-predicted bidirectional transcription sites (Galipon J. *et al.*, unpublished). Overall, 25% of ADAR-targeted lncRNAs were found to be transcribed from the antisense strand of an annotated mRNA or lncRNA, meaning that edited lncRNAs are not especially enriched in bidirectional transcription sites. I was able to confirm the expression of four of these edited lncRNAs by strand-specific real-time PCR. Furthermore, I found that lncRNA editing sites seem to be clustered slightly differently depending on which ADAR isoform was overexpressed, and the edits were almost exclusively localized in exons.

Introduction

Adenosine deaminases acting on RNA (ADARs) convert an adenosine (A) to an inosine (I) by hydrolytic deamination on stretches of double-stranded RNA (dsRNA). ADAR editing is known to occur mainly in various repetitive sequences outside of protein-coding sequences and can have regulatory functions, notably in miRNA biogenesis[3]. In human, the ADAR protein family has three active isoforms: ADAR1-p150, ADAR1-p110 and ADAR2. ADAR3 is enzymatically inactive, mainly expressed in the brain, and thought to act as a competitive inhibitor against other ADAR isoforms. While ADAR1-p110 and ADAR2 are mainly localized in the nuclei, ADAR1-p150 is mainly in the cytoplasm[1]. Previous research in our laboratory produced a genome-wide list of RNA editing sites by deep sequencing of RNA fragments that co-immunoprecipitated (RIP-seq) with each ADAR isoform, overexpressed separately in HeLa cells. Computational analysis revealed that long non-coding RNAs (lncRNAs) were enriched in the IP fractions of the nuclear isoforms ADAR1-p110 and ADAR2, but not in the fraction of the cytoplasmic ADAR1-p150 (Ishii *et al.*, unpublished).

DsRNA stretches can originate either from intra- or inter-molecular hybridization. Intermolecular hybridization can recognize complementary stretches either in *cis* or in *trans*. If ADAR proteins preferentially target lncRNAs coming from sites of bidirectional transcription, this may point towards a pattern of hybridization in *trans*. I calculated the percentage of edited lncRNAs which transcribed antisense from either a lncRNA, an mRNA, or both, and compared this percentage to the predicted genome-wide average percentage of lncRNAs coming from computationally predicted bidirectional sites. Finally, I created a visual map of lncRNA editing sites, which can be explored using the Integrated Genome Viewer (IGV). I found that when different ADAR isoforms target the same lncRNA, they seem to form slightly different clusters of editing sites. Furthermore, editing sites in lncRNA were localized mainly in exons.

Results

1. Identification of ADAR-targeted bidirectional loci

Bidirectional transcription sites are characterized by the presence of transcripts derived from both strands. In this study, we picked up edited lncRNAs that have one or more of the following annotations on the antisense strand: another lncRNA transcript (database: NONCODE v3), a messenger RNA (REFseq) or a miRNA (UCSC table browser). All three files were uploaded concurrently into the IGV genome browser and all 80 previously identified ADAR-edited lncRNAs were checked one by one for bidirectional transcription. Among them, 20 lncRNAs (25%) came from bidirectional loci. To compare with the genomic average, a random sample of 100 lncRNAs (NONCODE v3) was also checked by hand for bidirectional transcription. 34% of these randomly picked lncRNAs originated from bidirectional transcription sites.

My collaborator developed an R code to test both the random list and full list of lncRNAs for overlapping antisense mRNA or lncRNA, and found 38% of bidirectional loci, in close agreement with my manual result of 34% on the random sample. In a whole genome prediction, at least one third of the lncRNAs listed in NONCODE came from bidirectional transcription sites (Galipon J., unpublished). These findings indicate that ADAR-targeted lncRNAs are not enriched in bidirectional transcription sites.

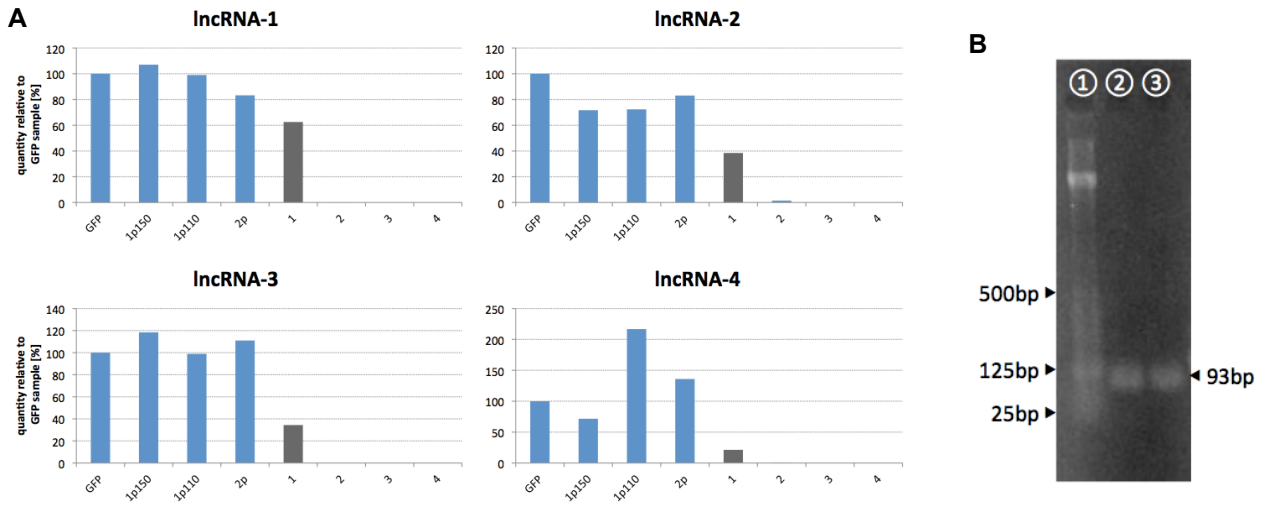

Figure 1. Expression levels of edited lncRNAs were confirmed by strand-specific RT-qPCR. [A] Strand-specific RT-qPCR was performed on total RNA extracted from cells overexpressing each ADAR isoform; signals were normalized to the control overexpressing GFP. Technical controls are as follows: (1) no primer RT; (2) no RTase RT; (3) no primer qPCR; (4) no template qPCR (NTC). [B] 1% agarose gel stained with ethidium bromide showing PCR amplification of total random 6-mer-generated cDNA, using the qPCR primers designed to detect lncRNA-1. Lanes are as follows: (1) 25 bp DNA ladder (Invitrogen 10597-011); (2) 1µg random 6-mer total cDNA as template; (3) 1/20th dilution of the cDNA in lane (2).

2. Confirmation of ADAR-edited lncRNA expression by gene-specific RT-qPCR

To confirm the expression of edited lncRNAs in HeLa cells overexpressing each ADAR isoform, we designed primers for gene-specific reverse transcription (RT) followed by quantitative real-time PCR (RT-qPCR). The primers were designed to amplify lncRNA-specific exon stretches, to make sure we can detect the edited lncRNA identified by RIP-seq, even in the presence of multiple overlapping transcripts in the same direction. Out of the 20 lncRNAs coming from bidirectional sites, 13 had exon sequences suitable for gene-specific RT-qPCR (data not shown).

LncRNA-specific RT was performed with the reverse primer. The resulting cDNA was digested with RNase H and used as a template for real-time PCR (*cf.* Methods). We could confirm the expression of all 4 edited lncRNAs that were checked (Figure 1A).

However, a high signal in the negative (no primer) RT control suggested a high background signal level. The no RT controls worked well (Figure 1A). The RT reaction was performed at 55ºC instead of 48ºC to increase specificity[2]. However, this did not significantly reduce the background levels in the no-primer controls. To rule out the possibility of non-specific amplification during qPCR, an ordinary PCR was performed on total cDNA generated from random hexamers, using the primers that gave the worst background (lncRNA-1). The PCR product was resolved on a 2% agarose gel. We detected a single band at the expected size (93bp), indicating that the primers are specific and thus suitable for qPCR (Figure 1B).

Despite high signals in the no-primer controls, all sample signals were significantly above background, suggesting that the 4 lncRNAs are indeed expressed in HeLa cells. Furthermore, none of these 4 lncRNA expression levels seemed to depend on the overexpression of ADAR1-p150, ADAR1-p110, or ADAR2.

3. Generating BED files for visualization of introns, exons and editing sites in the IGV genome browser

The Integrated Genome Viewer (IGV) accepts files in the standard Browser Extensible Data (BED) format and displays each file as a single track aligned against a reference genome (human genome, hg19). To visualize edited sites in the browser, a BED file was created which contains the positions of all edited sites detected in ADAR1-p150, ADAR1-p110, and ADAR2 samples, respectively. The editing sites appear in IGV as small vertical lines, and the editing sites for the three different ADAR isoforms were displayed in three separate tracks, respectively.

As can happen with the human genome, a lot of the lncRNAs we are looking at have a multitude of other annotated overlapping transcripts in the same direction. To avoid confusion with overlapping transcripts, I created another BED file to display the ADAR-targeted lncRNAs in a separate track. This file was generated using R code to output the lines in the lncRNA database (NONCODE) BED file, in which information of edited lncRNA names also appeared. So, now we have five tracks: all lncRNAs (NONCODE), edited lncRNAs, ADAR1-p150 edits, ADAR1-p110 edits, and ADAR2 edits (Figure 2).

Globally, editing sites seem to form clusters at certain regions along the lncRNAs, suggesting that these regions are prone to form dsRNA structures. However, if we look at lncRNAs that are targeted by multiple ADAR isoforms, we can see that different ADARs do not always edit at the same positions, suggesting that there are preferable editing sites for a certain ADAR isoform.

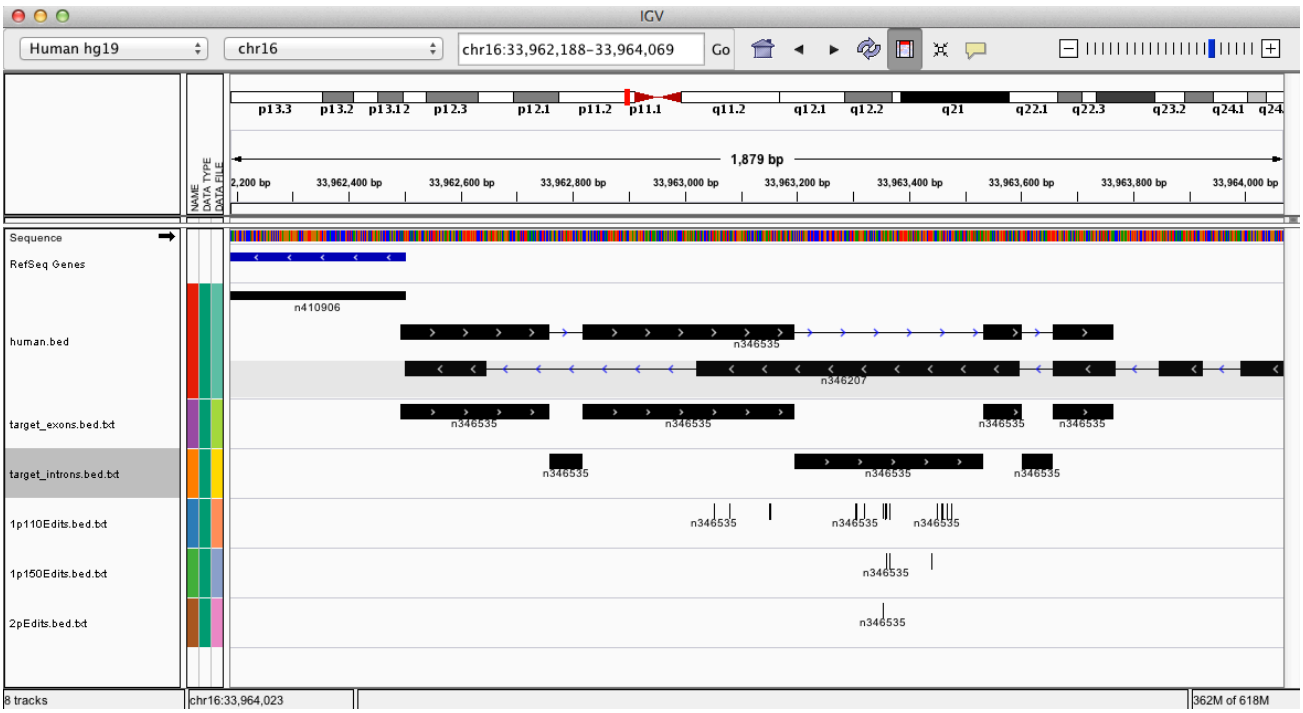

Figure 2. Browsing lncRNA editing sites in the Integrated Genome Browser (IGV). This is a screenshot showing an example bidirectional locus with discrete editing sites for each specific ADAR isoform. From top to bottom, the tracks are as follows: RefSeq Genes (UCSC Genome Browser), human lncRNAs (NONCODE v3), custom-made track displaying only human lncRNA exons, similar custom-made track for introns, editing sites observed in ADAR1-p110, ADAR1-p150, and ADAR2 overexpression samples, respectively. The reference genome is the hg19 build.

4. Development of an algorithm to count editing sites located in introns and exons

Are editing sites located preferentially in exons or introns? According to the annotated database, lncRNAs tend to harbour extremely long introns. If editing occurs at random, we expect to find a majority of editing sites in introns. To further simplify analysis, we obtained two BED files from the NONCODE database, one containing only lncRNA exons and another containing only lncRNA introns. We obtained them by uploading the NONCODE file into the UCSC table browser, and then downloading exon and intron files separately. We further processed these files using R and confirmed that our list only contains exons or introns of edited lncRNAs.

To count the editing sites in exons and introns respectively, the files that contain only ADAR editing target lncRNA exons or introns were compared against the files containing all the edit information. The following algorithm was developed in R: all exons or introns of the input file were tested against all edited positions. Several outcomes were counted iteratively: edited positions (1) within an intron, (2) within an exon or (3) outside of annotated intron/exon borders. The analysis of edited sites showed a strong enrichment in exons (80%). Not all edited sites could be mapped, but the fraction of non-mapped edits (7%) was negligible. This could be due to the presence of editing sites in reads that overlap the annotated transcription start site. In this case, the read would be annotated as belonging to that lncRNA, even if the editing site itself lies outside of transcription start or end site (Ishii, personal communication).

Discussion

ADAR editing targets are not enriched in bidirectional loci. The presence of 4 edited lncRNAs could be confirmed in HeLa cells. The RT-qPCR protocol used to amplify lncRNAs showed a high backround signal, which might be due to unspecific DNA oligonucleotides from the DNA digestion step and to unspecific priming caused by the RNase H activity of the used RTase. Therefore, the following protocol changes are suggested: Discarding DNA oligomers from the solution mixture before the RT reaction with a G25 column, that binds specifically DNA oligomers <25bp and the use of a recombinant RTase that doesn't have RNase H activity (e.g. superscript RTase).

The edit mapping data have to be corrected by the stability of pre- and mature RNA, since the reduced read-coverage of introns might be depended on the lower pre-RNA content compared to mature RNA.

Our data also suggests that ADAR1-p150 RIP-seq (Galipon, Ishii, unpublished data), may not reveal specific editing sites. Indeed, all the lncRNAs edited by ADAR1-p150 are also edited by other active ADAR isoforms, and the editing positions were conserved between RNA samples extracted from cells overexpressing different ADAR isoforms.

Methods

HeLa cell culture

HeLa cells were kept at 37ºC and 5% CO_2 in Gibco

DMEM medium supplemented with 10% FBS and antibiotics and replated once they reached 100% confluency. Replating was performed as follows (for a 9cm diameter dish): The medium was discarded, the cells washed with Wako D-PBS(-) (10ml), incubated with trypsin-EDTA (1ml) for 2 minutes at the above culture conditions and cell detachment was visually checked. DMEM (4ml) was added and aliquots taken to count cells. The cells were diluted as required and replated.

HeLa cell transfection

HeLa cells were transfected with Lipofectamine 2000 following the corresponding Invitrogen protocol using Gibco Opti-MEM as a transfection medium and GFP-tagged ADAR1-p150, ADAR1-p110, ADAR2 and a control GFP overexpression vectors. WT untransfected cells were kept as a negative control. Transfection was assessed by fluorescence microscopy via the overexpressed GFP/GFP-tag with the Zeiss Axiovert 200.

RNA extraction

Total RNA was extracted from the HeLa cells with the Quiagen RNeasy kit, following the corresponding protocol of Quiagen. A DNase digestion was performed on the purifired RNA, by adding Promega 10xDNase buffer (5µl) and RQ1 Promega DNase (3µl) to the purified RNA solution (42µl). The mix was incubated at 37ºC for 30 minutes. Subsequently, RNA extraction with the Quiagen RNeasy kit was performed again. The RNA concentration, 260/280 and 260/230 values were measured with Thermo Scientific NANODROP 2000 chosing RNA-40 as the sample type.

Primer design

Primers were designed for specific exon stretches. The sequence of the lncRNA exons was taken from the IGV Genome Browser into which the NONCODE library (downloaded 10th July 2013) was loaded. In the case of overlapping sense transcripts, ApE (Universal) software was used to create alignments and find non-overlapping regions suitable for primer design. 4 specific primer pairs were designed using Primer BLAST.

Reverse Transcription (RT)

RT was performed with the Roche Transcription high fidelity cDNA synthesis kit in the Gene Amp PCR system 9700. The total reaction volume was 20µl. A 2.5µM reverse primer mix (2µl) and 0.25µg purified total RNA in PCR-grade water (9.4µl) were mixed and incubated at 65ºC for 10 min. Then 5x Buffer (4µl), 2mM each dNTP mix (2µl), DTT (1µl), RTase (1.1µl) and PCR grade water (0.5µl) were added in a master-mix. No-primer and no-RTase controls were performed by substituting these components with PCR-grade water. The mix was incubated 10 min at 29ºC, 60 min at 48ºC/55ºC and 5 min at 85ºC. RNase H digestion was performed on the RT products by adding RNase H (1µl) and incubating for 30 min at 37ºC.

Quantitative (q) PCR

qPCR was performed with the Applied Biosystems ABI PRISM 7000 Sequence Detection System and the FastStart Universal SYBR Green Master (ROX) from Roche. The total reaction volume was 25µl: 12.5µl FastStart, 0.5µl of a primer mix containing the forward and reverse primer at 30µM, 11µl RNase free water and 1µl template at a 1:10 dilution of the RT product. No-primer and no-template controls were performed by substituting the respective component by RNase free water. The primers mentioned under primers were used. The cycle parameters were: 10 min at 95ºC, 40x(15 sec at 95ºC, 1 min at 60ºC) and the samples were kept at 4ºC until they were frozen.

PCR

The Toyobo KOD –Plus- polymerase was used and the PCR mix prepared according to the manufacturer's instructions but with a total volume of 20µl. 1µl template was added undiluted and at 1:20 dilution of the cDNA sample. The cycle was: 2 min at 95ºC, 4x(15 sec at 95ºC, 15 sec at 68ºC, 30 sec at 72ºC).

Agarose Gels

Agarose gels were casted from 0.5x TAE buffer with 1~2% agarose. The gels were run at 50V. Nucleic acid was stained by incubating the gel for 15 min in 0.5µg/ml ethidium bromide solution and visualized with UV-B (302nm) radiation.

10% Polyacrylamide Gel (PAG)

The final composition of PAGs was as follows: 0.5x TBE, 10% acrylamide solution and fill up with milliQ water and then adding 10% APS (1:100 v/v) and TEMED (1:1000 v/v). PAG were run for 40 min at 15mA and 200V. Nucleic acid bands were stained by incubating the gel for 4 min in ethidium bromide solution and visualized with UV radiation.

R programming language

The RStudio IDE was used to code in R.

Acknowledgements

I would like to thank Prof. Ui-Tei and Dr. Josephine Galipon for providing me with guidance and the opportunity to conduct this project and everybody else in the Ui-Tei group for helping me on a daily basis. I would also like to thank Ms. Soeda and ILO team for their help and the organization of UTRIP.

References

[1]Nishikura K., Annu. Rev. Biochem. 2010; 79:321–49

[2]Große C. *et al.*, J. Mol. Microbiol. Biotechnol. 2007; 12:227–240