Article type: Protocols for solving a specific problem using different sets of programs

Large-scale Prediction of ADAR-mediated Human A-to-I RNA Editing Effects

Li Yao[3†], Heming Wang[4†], Yuanyuan Song[5†], Zhen Dai[9], Hao Yu[7], Ming Yin[6], Dongxu Wang[7], Xin Yang[8], Jinlin Wang[6], Tiedong Wang[7], Nan Cao[10], Guangqi Song[1,9]*, Yicheng Zhao[2,7]*

[1] Shanghai Institute of Liver Disease, Zhongshan Hospital, Fudan University, Shanghai, China.
[2] College of life science, Northeast Forestry University, Harbin, China.
[3] Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing, China
[4] School of Life Science and Technology, ShanghaiTech University, Shanghai, China
[5] Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China.
[6] Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China.
[7] College of Animal Sciences, Jilin University, Changchun, China.
[8] Department of Life Sciences, Southeast University, Nanjing, China.
[9] Cluster of Excellence Rebirth, Hannover Medical School, Hannover, Germany.
[10] Changchun Sub-branch, the People's Bank of China, Changchun, China.
[†] Co-first authors
* Correspondence may also be addressed to zyc@rnaeditplus.org and guangqisong@rnaeditplus.org.

ABSTRACT

Adenosine-to-inosine (A-to-I) editing by ADAR proteins is one of the most frequent modification during the post- and co-transcription. To facilitate the assignment of biological functions to specific editing sites, we designed an automatic online platform to annotate A-to-I RNA editing sites in pri-mRNA splicing signals, microRNAs, microRNA target regions (3'UTR) from human (*homo sapiens*) high-throughput sequencing data and predict their effects based on large-scale bioinformatic analysis. After analyzing plenty of previously reported RNA editing events and human normal tissues RNA high-seq data, more than 60,000 potentially effective RNA editing events on functional genes were found. The platform named RNA Editing Plus is available for free at https://www.rnaeditplus.org/ and we believe our platform governing multiple optimized methods will improve further studies of A-to-I induced editing post-transcriptional regulation.

KEYWORDs

RNA editing, microRNA targeting, Alternative mRNA splicing, Gene mutation.

INTRODUCTION

The most frequent type of RNA editing is the A-to-I catalyzed by the adenosine deaminase acting on RNA (ADAR) family of enzymes, which occurs mainly within double-stranded RNA regions (dsRNA). Specifically, since inosine (I) residues preferentially base pair with cytidine (C), inosine residues in the coding and noncoding RNA sequences are thereby recognized as guanosine (G), genomically manifested as A-to-G mismatches. High-throughput sequencing technology has greatly accelerated the A-to-I editing research [1], and hundreds of thousands of RNA editing sites are identified yearly (RADAR[2], DARNED[3], HREA[4], and DREAM[5]). As reported, A-to-I RNA editing in human occurs frequently in intron and untranslated regions (UTRs) containing primate-specific inverted Alu repeats [6].

RNA editing in introns may contribute to pre-mRNA alternative splicing, and miRNAs or 3'UTRs editing may change or redirect interactive relationship between certain mRNAs and miRNA [1, 7-9](Fig.1a). Numerous modified nucleotides in functional genes are subjected to A-to-I editing, connecting to various diseases [10, 11]. However, a compact link or rational standard between editing calling and downstream effects are still absent. Here, we developed a one-step analysis system gathering RNA editing calling, miRNA-3'UTR binding evaluation, mRNA alternative splicing prediction and gene mutation scan modules (Fig.1b).

MATERIAL AND METHODs

RNA-seq data mapping. HISAT2[12], STAR[13], BWA[14] were all employed to pre-test the editing calling reproducibility. When conducting the HISAT2 index, we adopted GENCODE V24[15] to annotate exon and pre-mRNA splicing region, dbSNP build 146 from UCSC to annotate SNP. When using STAR, we adopt a two-round mapping, the parameter is –sjdbOverhang 75 when indexing for the second round. To BWA, we use the commands 'bwa aln fastqfile' and 'bwa samse -n4'. According to Ramaswami. et al, we also performed editing calling after incorporating different RNA-seq alignments, 'merged' in Fig.1c means to merge all reads before editing calling, including BWA-REDItools-merged, HISAT2-REDItools(tran)-merged. While, without 'merged' in Fig.1c means to merged all the results after performing editing calling, including BWA-REDItools, HISAT2-REDItools, HISAT2-REDItools (tran), HISAT2-REDItools (tran, SNP), STAR-REDItools, STAR-GATK[16]. To REDItools, we used the commands REDItoolDenovo.py -d -1 -c 2 -C 0 -v 3 -f 0.1 -e. To gatk, we used gatk HaplotypeCaller, the parameter is -dontUseSoftClippedBases -stand_call_conf 20.0. A VariantFilter (hard filter) were also performed with parameter: -window 35 -cluster 3 -filterName FS -filter "FS > 30.0" -filterName QD -filter "QD < 2.0". As a result, HISAT2 was chosen as our default mapping tool because of higher sensitivity to mismatch (Fig.1c and Fig.S2).

Identification and annotation of human A-to-I RNA editing events. REDItool[17] made it possible to perform editing calling without the need for matched genomic DNA sequence, and we prepared common reference genome files using GRCh38 (hg38) in advance and used the default parameter REDItoolsDenovo.py. Via total bases substitution scanned from the mapped reads (BAM file) to the reference genome, an empirical distribution was calculated and further employed to identify genome-wide variations. For each possible RNA editing type, Fisher exact test was used to judge its authenticity by false discovery rate. When using GATK as variant calling, we employed dontUseSoftClippedBases -stand_call_conf 20.0 for HaplotypeCaller and window 35 -cluster 3 -filterName FS -filter "FS > 30.0" -filterName QD -filter "QD < 2.0" for Variant Filtration. We initially purged SNP effects on empirical distribution. The liftOver tool from UCSC (http://genome.ucsc.edu/cgi-bin/hgLiftOver) was utilized to update and filter previously reported RNA editing events according to new hg38 reference file (Fig.S7 and Table.S6). Additional annotations by Repeat Masker database[18] were introduced, subsequently.

Prediction module for editing on human miRNA-targeting. miRBase21[19] was used to annotate miRNAs for the discovered RNA editing sites. Using experimentally validated miRNA-mRNA targeting relationship from both miRmap and miRTarBase (True positive), miRNA-mRNA non-targeting relationship from TargetScan (True negative), we preliminarily evaluated human miRNA-target binding model *in silico* from four aspects: 1.Thermodynamic including *ΔG* duplex, *ΔG* open, *ΔG* binding, *ΔG* seed duplex, *ΔG* seed binding[20], 2.Evolutionary and 3.Probabilistic including binomial distribution method binomial distribution (binominal distribution)[21], exact probability distribution (exact probability distribution)[22], 4.Sequence-based features including TargetScan context score (a/u ratio over g&c, weighted around the seed match (AU content)), and the 3'-compensatory pairing feature (3'-compensatory pairing)[23]. Since *ΔG* duplex, *ΔG* binding, *ΔG* open in non-targeting group and AU content feature in targeting group were more close to normal distribution. For editing in seed region, we employed TargetScan to predict possible miRNA-mRNA interaction, and employed miRanda[24] for predicting editing effects in miRNAs non-seed region (Fig.S5). To further enhance the accuracy of effects prediction, we performed a SVM classification for a second evaluation after TargetScan assessment.

SVM in miRNA and 3'UTR prediction module. We initially summarized 291 experimentally validated data (RNA editing/mutation/SNP) with mature miRNAs and 3'UTRs (Table.S2). Specially, these experimentally validated 3'UTRs data includes multiple editing/mutation/SNP, while only single nucleotide changed miRNAs data were chosen. Then, we calculated nine parameters values before and after RNA editing/mutation/SNP, and four of those were chosen since their high significances including *ΔG* open, *ΔG* binding, *ΔG* seed duplex, *ΔG* seed binding (Fig.S6). After that, 100 random tests (70% for SVM training, 30% for accuracy detecting) were performed to detect the prediction module accuracy (Supplementary data S1). Besides, we selected 19 typical A-to-I and A-to-G experimentally validated data for measuring our miRNA-targeting module against Targetscan and miRanda, detailed analyzing results (9/19 for TargetScan, 4/19 for miRanda, and 12/19 for RNA editing plus) were listed in Table.S3.

Predictive module for RNA editing on RNA Splicing. Since 'GT' and 'AG' are highly conservative (Fig.1d), we only considered the nucleic acid alteration in 5'ss (6nt: +3 to +8) and 3'ss (18nt: -20 to -3) intro regions, while recognizing 'GT' consensus at positions (+1, +2) and 'AG' consensus at positions (-2, -1). According to annotation from GENCODE v24, if editing occurs in 5'/3'ss intro region, MaxEntScan will be directly called for calculating scores for each region (detailed formula is listed below). To predicting editing effects on branch site, we introduced AGEZ[25] to find the first 'AG' in the upstream of 3'ss region, and took the position weight matrix[26] to entirely scan the whole 'AG-BS-AG' region (-21 to -150), determining the region with highest score as branch site (formula is listed below). To facilitate the accuracy when predicting effective editing sites on pre-mRNA splicing, we set up related thresholds, which limited the minimum disparity values between unedited and edited scores and classify the A-to-I editing effects in six aspects including inactivated (or weakened) 5' or 3' splice site, enhanced 5' or 3' splice site, weakened, inactivated, enhanced, new branch site.

$$\text{Splicing site score} = \log2(\mathit{consensus\ score} \times \mathit{Maximum\ Entropy\ Distribution\ Score})$$

$$\text{Branch point score} = \sum_{j=1}^{7} PWM_{i,j} \quad (\mathit{where}\ i = 1, 2, 3\ \mathit{or}\ 4\ \mathit{corresponding\ to}\ A, C, G\ \mathit{and}\ U)$$

In order to optimize related thresholds (5'ss, 3'ss), we introduced a receiver operating characteristic curve (ROC) by highlighting the true positive rate (TPR) against the false positive rate (FPR) (formulas see below) based on 1,713 experimentally validated testing samples (Table.S4).

$$TPR=\frac{TP}{TP+FN}$$

$$FPR=FP/(FP+TN)$$

Sequence Preferences for base positions flanking analysis. Sequence preference detection is performed via a two-sample logo program[27].

Gene Ontology (GO) analysis. DAVID web tool[28] was employed to perform GO analysis, we submitted all potentially effective editing events from 28 human normal tissues (Table.S7) as standard protocol to calculate gene enrichments, the top 10 gene ontology terms significantly associated with each tissue were listed in Fig.2f, Fig.S10 and Table.S8.

Data collection. We collected previously reported human A-to-I editing events from DARNED, RADAR, and HERA. We collected previously reported miRNA-targeting data from TargetScan, miRmap and miRTarBase. We selected 156 normal tissue pair-end Illumina RNA-seq data regarding kidney, heart, liver, lung, brain, etc and YH RNA-seq data, details are described in Table.S10. We have manually scanned plenty of experimental data (more than 2,000 cases in total) regarding miRNA targeting, gene SNP, and RNA splicing, detailed information are available in Table.S2 and S4.

Statistics and Code Availability. All data were analyzed by R (the R Project for Statistical Computing) and GraphPad Prism software. RNA Edit Plus was implemented using a combination of PHP, Python and C codes. The code package is available request.

RESULTS

Accurate identification and annotation of human A-to-I RNA editing sites. Mapping RNA reads to the reference genome and editing calling is the key step for A-to-I RNA editing sites identification and annotation, however, there are different popular mapping tools (HISAT2, STAR, BWA)[17, 29-31]. Using a previously published deeply sequenced Han Chinese RNA-seq data (YH)[32], we employed HISAT2, STAR, BWA combining with REDItools respectively, to test the editing calling reproducibility. As shown in Fig.1c, RNA editing events from the YH data were illustrated by a similarity matrix, indicating that combining HISAT2 with REDItools is able to provide more previously reported editing sites. As a result, a large number of previously identified editing sites were found residing in Alu regions, while the non-Alu RNA editing sites number was relatively low (Fig.S3).

Analysis of A-to-I RNA editing modification on miRNAs and 3'UTR. MicroRNAs (miRNAs) maturation of miRNAs can be divided into two sections, the nucleus primary miRNA (pri-miRNAs) with stem-loop structures are processed at hairpins by Drosha-DGCR8 complex to form precursor miRNAs (pre-miRNAs). In the cytoplasm, pre-miRNAs are further recognized and cleaved by Dicer-TRBP complex to yield about 22 nt-long miRNA duplexes. The strand with more stable 5' of the duplexes are then loaded onto the Argonaute (AGO) proteins within the RNA-induced silencing complex (RISC), and unwound into single-stranded mature miRNAs [33]. As we know, both pre- and pri-miRNAs have dsRNA substrates, allowing ADAR to influence the miRNA function. Editing of pri-miRNAs may affect their processing into mature miRNAs or lead to production of mutated miRNAs, which silence a changed set of target genes. RNA editing occurred in mature hsa-let-7d weakened its inhibitory ability on LIN28B [7].

Here, we focused on A-to-I editing effects on mature miRNA. After comparing the distributions according to various features from existing prediction methods regarding miRNA-targeting, we employed TargetScan for analyzing editing effects in miRNA seed region and miRanda for non-seed region, since related feature distributions seems more closely to Gaussian distribution (Fig.S5). Using experimentally validated data (RNA editing/mutation/SNP)(Table.S2), we calculated nine parameters involved in miRNA-3'UTR binding before and after nucleotides changing in miRNAs (Fig.1f and Fig.S6). As shown in Fig.1g and Fig.S6, random combination of three out of four parameters (*ΔG* open, *ΔG* binding, *ΔG* seed duplex, *ΔG* seed binding) is able to efficiently distinguish the experimentally validated data into two different groups (True Positive and True Negative). To further enhance the prediction accuracy, we selected these four parameters together for a Support Vector Machine (SVM) classification. As a result, we achieved to predict 12/19 altered miRNA-mRNA binding testing examples (Table.S3), which significantly improved the prediction accuracy (Supplementary file 1). To gain further investigation, we calculated all the RNA editing sites in mature miRNA sequences from DARNED, RADAR, HERA databases, and found 74 potentially effective editing events in mature miRNAs (Fig.S7 and Table.S1).

A-to-I editing also occurs in 3'UTR regions of human transcriptome, which affects the existing miRNA binding sites as well as generate novel binding sites (Fig.1a). For instance, A-to-I editing in AHR 3'UTR created a new miR-378 binding site [8]. As mentioned above, we collected single or multiple nucleotide editing/mutation/SNP in human 3'UTR experimental data (Table.S2) and adopted a similar prediction strategy. In DARNED, RADAR and HERA databases, we found 65,841 sites in 3'UTR affecting miRNA-targeting (Fig.S7. and Table.S1).

Analysis of A-to-I RNA editing modification on mRNA alternative splicing. A-to-I editing inside LUSTR/GPR107 intron caused the exclusion of the Alu exon, indicating alternative splicing might be co-regulated by RNA editing [34]. RNA editing sites were found in all three main regions involved with pre-mRNA alternative splicing (donor: 5'splicing site, acceptor: 3'splicing site, and Branch site)[35]. We retrieved short sequence motif distribution via hg38 and GENCODE v24 data around 5 and 3 splicing site (5'ss and 3'ss) and found 'GT' and 'AG' are highly conservative (Fig.1d). To predict potential effects in 5'ss and 3'ss, we employed MaxEntScan based on max entropy theory, which recognizes splicing signal and decoy signal only by defined signal, providing us unbiased prediction [36]. In advance, we updated all data collected in MaxEntScan according to hg38 and GENCODE. We next measured MaxEntScan scores before and after A-to-I editing in 5'ss/3'ss regions, and filtered the altered values by thresholds to judge whether these 5'ss/3'ss regions enhanced or weakened. To facilitate the accuracy, we optimized related thresholds by ROC (Fig.1e) using experimentally validated data (Table.S4), and 1,413 out of 1,713 5'ss/3'ss experimentally validated events were correctly predicted (Table.S9). For editing in branch site, we adopted AG-Exclusion Zone algorithm (AGEZ), combining position weight matrix to entirely evaluate the original and edited 'AG-BS-AG' region. As a result, there are 805 DARNED, and RADAR potential intronic A-to-I editing sites affecting pre-mRNA alternative splicing (Fig.S7 and Table.S1).

Scan of A-to-I RNA editing induced mRNA missense mutation. A-to-I RNA editing was also found in coding sequence (CDS) region, sometimes producing gene missense mutation. The CAG (Glutamine) to CGG (Arginine) mutation committed by ADAR on AMPA receptor subunit GluR-B unspliced transcript has been reported previously [9]. We recognized A-to-I as A-to-G, and compared translation before and after RNA editing to look for effective events on protein coding. We scanned all

RNA editing sites in gene exon or CDS from DARNED, RADAR, HERA and found 1,786 exon A-to-I editing sites affecting protein coding (Fig.S7 and Table.S1).

Analysis of A-to-I RNA editing disruption and effects of human normal tissues. After constructing RNA Editing Plus, we analysed the RNA editing level and the expression of ADAR1&2 in several normal human tissues from 156 RNA-seq data (Fig.2a and Fig.S4), and more potential editing events were found in testis tissue (Table.S5). Comparing the ADAR expression values to overall A-to-G Editing levels from all 156 human tissue samples, we confirmed positive correlation relationship, however, the relationship is nonlinear (Fig.2b and Fig.S9). We also detected the A-to-G editing levels across 28 normal tissue types (Fig.2c). Then, we investigated the sequence context flanking using all potential A-to-I (G) RNA editing events from 156 samples (Table.S5), neighbour sequence preferences of the whole genome, Alu, non-Alu repetitive and Non-repetitive regions are shown in Fig.2d.

Importantly, 60,936 potentially effective A-to-I editing events were predicted by our platform from human normal tissues (Fig.2e, Fig.S8, Table.1, and Table.S7). Different to previous reports, we used these gene functionally effective editing data which eliminating interference from the no-effective editing evets in Gene Ontology (GO) enrichment analysis. As a result, multiple aspects were influenced including Chromatin binding, Ligase activity, etc (Fig.2f & Fig.S10 and Table.S8).

DISCUSSION

To date, several RNA editing-site databases or bioinformatic tools have been developed that include information gathered from the literature or from manually accrued datasets. However, there are few integrated tools providing one-pass identification and annotation, or multifunctional analysis for RNA editing research. We therefore developed this robust platform for integrated acquisition, storage, display and analysis of high-throughput RNA data. Furthermore, our platform allows for seamless integration of multiple, published or locally produced datasets via loading BAM (binary format for storing sequence data) alone.

To our knowledge, several standalone programs and web services are available for the annotation and analysis of RNA editing data. However, the majority of currently available tools have a command-line interface and typically require file conversions between them. Although Galaxy provides the opportunity to run tools without using a command-line interface, users still have to manage file type conversions and select detailed parameters each time, which requires a deep understanding of each tool and file format. In summary, few of the available tools provide a biologist-friendly interface, and none integrate such an interface with data storage, display and analysis.

Occasionally, A-to-I editing events in a certain region are capable to influence multiple aspects, we distinguished the transcripts when processing the annotation to increase additional information, which might prevent the loss prediction of various effects. A series of studies which confirmed the redirection of interactive miRNA when RNA editing or mutation occurred at UTR regions or miRNA mature sequences usually performed a dual-luciferase reporter assay to validate the downstream effects on miRNA-mRNA interaction. However, the truth is that different proportion of UTR length in reporter vectors can lead to completely different results. If the UTR used in dual-luciferase assay is truncated which contains only one target site, the luciferase activity may obviously change between wild type and mutated miRNA-mRNA interactions, but RNA Editing Plus considers the full length of UTRs that possibly containing multiple target sites, recognizing it as 'common targets', which explained the incorrect prediction of IGF1R and AhR [8, 37] of our platform.

Our prediction modules provided more than 60,000 potential editing events affecting mRNA alternative splicing, miRNA target silencing and protein coding, which will contribute significantly to related fields. For convenience, all potentially effective A-to-I editing sites mentioned above have been initially indexed in RNA Editing Plus, users can search their interested events by inputting gene information as well as submitting their open-access RNA-seq data according to our platform tutorial (Supplementary tutorial).

SUPPLEMENTARY DATA

Supplementary Figures, Captions, Data and any associated References can be found online.

FUNDING

This work was financed by the National Natural Science Foundation of China (31472159 to Y.Z., 81601190 to X.Y.), Heilongjiang Postdoctoral Grant (LBH-Z14018 to Y.Z.), Fundamental Research Funds for the Central Universities (2572015BA06 to Y.Z.), and Natural Science Foundation from Jiangsu Province (BK20160665 to X.Y.).

CONFLICT OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

L.Y., H.W., and Y.S. optimized the algorithm, interpreted the results, and scanned the experimental validation data. M.Y., Z.D., X.Y., J.W., N.C., H.Y., and T.W. interpreted the results. Y.Z. and G.S. conceived the study, interpreted the results, and wrote the main manuscript.

Yicheng Zhao is a lecturer of Biology in the Northeast University. His main interest is in understanding regulatory non-coding RNAs mechanisms in pancreatic and liver cancer.

Guangqi Song is an associate professor of Zhongshan Hospital, Fudan University. He is focusing on liver function/development, and related stem cell-based therapies in hepatic diseases.

KEY POINTS

1) Human RNA editing effects can be predicted reliably from human RNA high-seq data alone, and carries relevant biological information with integration of suitable platforms and pipelines.

2) More than 60,000 RNA editing sites potentially effecting microRNA targeting, mRNA alternative splicing and gene CDS missense mutation in human normal tissue were illustrated.

3) HISAT2 is suitable for RNA editing calling since its higher sensitivity to mismatch when mapping the RNA high-seq data.

REFERENCES

1. Nishikura K. A-to-I editing of coding and non-coding RNAs by ADARs, Nat Rev Mol Cell Biol 2016;17:83-96.
2. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing, Nucleic Acids Res 2014;42:D109-113.
3. Kiran A, Baranov PV. DARNED: a DAtabase of RNa EDiting in humans, Bioinformatics 2010;26:1772-1776.
4. Picardi E, Manzari C, Mastropasqua F et al. Profiling RNA editing in human tissues: towards the inosinome Atlas, Sci Rep 2015;5:14941.
5. Alon S, Erew M, Eisenberg E. DREAM: a webserver for the identification of editing sites in

mature miRNAs using deep sequencing data, Bioinformatics 2015;31:2568-2570.

6. Bazak L, Haviv A, Barak M et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes, Genome Res 2014;24:365-376.

7. Zipeto MA, Court AC, Sadarangani A et al. ADAR1 Activation Drives Leukemia Stem Cell Self-Renewal by Impairing Let-7 Biogenesis, Cell Stem Cell 2016;19:177-191.

8. Nakano M, Fukami T, Gotoh S et al. RNA Editing Modulates Human Hepatic Aryl Hydrocarbon Receptor Expression by Creating MicroRNA Recognition Sequence, J Biol Chem 2016;291:894-903.

9. Higuchi M, Single FN, Kohler M et al. RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency, Cell 1993;75:1361-1370.

10. Slotkin W, Nishikura K. Adenosine-to-inosine RNA editing and human disease, Genome Med 2013;5:105.

11. Tariq A, Jantsch MF. Transcript diversification in the nervous system: a to I RNA editing in CNS function and disease development, Front Neurosci 2012;6:99.

12. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements, Nat Methods 2015;12:357-360.

13. Dobin A, Davis CA, Schlesinger F et al. STAR: ultrafast universal RNA-seq aligner, Bioinformatics 2013;29:15-21.

14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 2009;25:1754-1760.

15. Harrow J, Frankish A, Gonzalez JM et al. GENCODE: the reference human genome annotation for The ENCODE Project, Genome Res 2012;22:1760-1774.

16. McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res 2010;20:1297-1303.

17. Picardi E, Gallo A, Galeano F et al. A novel computational strategy to identify A-to-I RNA editing sites by RNA-Seq data: de novo detection in human spinal cord tissue, PLoS One 2012;7:e44184.

18. Saha S, Bridges S, Magbanua ZV et al. Empirical comparison of ab initio repeat finding programs, Nucleic Acids Res 2008;36:2284-2294.

19. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data, Nucleic Acids Res 2014;42:D68-73.

20. Vejnar CE, Zdobnov EM. MiRmap: comprehensive prediction of microRNA target repression strength, Nucleic Acids Res 2012;40:11673-11683.

21. Marin RM, Vanicek J. Efficient use of accessibility in microRNA target prediction, Nucleic Acids Res 2011;39:19-29.

22. Nuel G, Regad L, Martin J et al. Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data, Algorithms Mol Biol 2010;5:15.

23. Grimson A, Farh KK, Johnston WK et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing, Mol Cell 2007;27:91-105.

24. John B, Enright AJ, Aravin A et al. Human MicroRNA targets, PLoS Biol 2004;2:e363.

25. Gooding C, Clark F, Wollerton MC et al. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones, Genome Biol 2006;7:R1.

26. Desmet FO, Hamroun D, Lalande M et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals, Nucleic Acids Res 2009;37:e67.

27. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, Bioinformatics 2006;22:1536-1537.

28. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat Protoc 2009;4:44-57.

29. Ramaswami G, Zhang R, Piskol R et al. Identifying RNA editing sites using RNA sequencing data alone, Nat Methods 2013;10:128-132.

30. Ramaswami G, Lin W, Piskol R et al. Accurate identification of human Alu and non-Alu RNA editing sites, Nat Methods 2012;9:579-581.

31. Zhang Q, Xiao X. Genome sequence-independent identification of RNA editing sites, Nat Methods 2015;12:347-350.

32. Peng Z, Cheng Y, Tan BC et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome, Nat Biotechnol 2012;30:253-260.

33. Zhao Y, Song Y, Yao L et al. Circulating microRNAs: Promising Biomarkers Involved in Several Cancers and Other Diseases, DNA Cell Biol 2017;36:77-94.

34. Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome, PLoS Biol 2004;2:e391.

35. Shao C, Yang B, Wu T et al. Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome, Nat Struct Mol Biol 2014;21:997-1005.

36. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, J Comput Biol 2004;11:377-394.

37. Gilam A, Edry L, Mamluk-Morag E et al. Involvement of IGF-1R regulation by miR-515-5p modifies breast cancer risk among BRCA1 carriers, Breast Cancer Res Treat 2013;138:753-760.

TABLE AND FIGURE LEGENDS

Table 1. Potentially effective RNA editing events from previously reported RNA editing sites.

[a]Potentially effective events on miRNA-target silencing including editing on miRNA and 3'UTR.

[b]Potentially effective events on mRNA alternative splicing. [c]Potentially effective RNA editing events on protein coding. Detailed information please see Fig.S7 & S8 and Table.S1 & S7.

Figure 1. A computational framework to identify A-to-I RNA editing effects. (a) The diagram shows RNA editing affects miRNA targeting and mRNA splicing resulted from nucleaic acid alterations. (b) Basic work principles of RNA Editing Plus, after annotating each A-to-I event from RNA-seq data or RNA editing list, the editing effects will be predicted by three bioinformatic modules (detailed workflow information please see Fig.S1). (c) YH data editing calling comparisons. A-to-I editing sites from each method were compared and the similarity was calculated as S(Row(x),Col(y)) = (Row(x)⋂Col(y)) / Row(x), 'tran' means indexing with transcription annotation, 'SNP' means indexing with dbSNP 146 annotation, 'merged' means to merge all reads before editing calling. 918 (containing 892 SNP sites) non-RNA editing sites were removed from Ramaswami.et al [29] result via hg38 updating and SNP annotating (detailed information see Table.S6). (d) Sequence preferences for base positions flanking 5'ss and 3'ss were calculated using hg38 and GENCODE v24. (e) Pre-mRNA splicing prediction module thresholds optimization. The ROC curve shows the process of determining the optimal thresholds by changing new different parameters. (f-g) Combination of multiple thermodynamic features is more efficiently in evaluating miRNA-mRNA targeting than using single feature from experimentally validated data. The *p*-value for (f) was calculated by one-way ANOVA tests, the *p*-value for (g) was calculated by Welch Two Sample t-test, n=291. (Detailed information see Fig.S6 and Table.S2).

Figure 2. Analyzing ADAR-medtiaed RNA editing of human normal tissue. (a) ADAR1 (p110 & p150) and ADAR2 expression level were calculated as FPKM value (details see Fig.S4). (b) We correlate enzymatic ADAR expression (ADAR p110, p150, ADAR2) values and all A-to-I (G) editing event numbers in all sample groups (n = 156). Spearman rank correlation coefficients (r), Linear regression goodness of fit (r2) and their *p*-values are shown (Related information see Fig.S9). (c) All A-to-I (G) RNA editing levels from 28 human normal tissues were calculated by Hierarchical clustering of Spearman correlation coefficients, the similarity was calculated as S((Row(x),Col(y)) = (Row(x)∩Col(y)) / (Row(x)∪Col(y)). (d) The motif flanking A-to-I (G) RNA editing sites and motif based on Alu, Non-Alu, Non-repetitive. Editing sites are identified from 28 normal tissue and Sequence preference is represented using a two-sample logo program. (e) Analysis of potential effective RNA editing events disruption from 28 types of normal tissues. (Related information see Fig.S8 and Table.S7). (f) Gene Ontology (GO) enrichment analysis of all potential effective RNA editing sites

from normal tissue, the top 10 gene ontology terms were selected. (Detailed GO analysis data for each tissue type please see Fig.S10 and Table.S8).