Review Article

J Genet Med 2013;10(1):27-32
http://dx.doi.org/10.5734/JGM.2013.10.1.27
ISSN 1226-1769(Print) 2233-9108(Online)

JGM Journal of Genetic Medicine

Bioinformatics Approaches for the Identification and Annotation of RNA Editing Sites

Soo Youn Lee and Ju Han Kim*

Seoul National University Biomedical Informatics (SNUBI) and Systems Biomedical Informatics Research Center

Post-transcriptional nucleotide sequence modification of transcripts by RNA editing is an important molecular mechanism in the regulation of protein function and is associated with a variety of human disease phenotypes. Identification of RNA editing sites is the basic step for studying RNA editing. Databases and bioinformatics resources are used to annotate and evaluate as well as identify RNA editing sites. No method is free of limitations. Correctly establishing an analytic pipeline and strategic application of both experimental and bioinformatics methods constitute the first step in investigating RNA editing. This review summarizes modern bioinformatics approaches and related resources for RNA editing research.

Key words: RNA editing, RNA-Seq, Bioinformatics

Introduction

RNA editing is a molecular process in which nucleotide sequence change occurs within a RNA molecule after it has been transcribed by an RNA polymerase.[1] RNA editing is a rather rare molecular event and other common forms of RNA processing including splicing, 5'-capping and 3'-polyadenylation events are not usually considered as RNA editing.

While extensive RNA editing (or pan-editing) can occur in some organisms, it is rather rare and usually consists of a small number of changes to the sequence of affected molecules in vertebrates. While RNA editing has been observed in different forms of RNAs including tRNA, rRNA, mRNA and miRNA in eukaryotes and their viruses, it has not been seen in prokaryotes. RNA editing occurs in some cellular organelles like mitochondria and plastids as well as in nuclei and the cytoplasm.

RNA editing is mainly classified into two categories; substitution editing (chemical alteration of individual nucleotides) and insertion/deletion editing (insertion or deletion of nucleotides). With the development of high-throughput next-generation sequencing technologies, increasing numbers of RNA editing sites are rapidly being discovered. Accordingly, many bioinformatics tools and algorithms to identify novel RNA editing sites and semantic annotation of known information are increasingly being introduced. Each method, however, is not free from limitations. For example, numerous false positives are found in the newly discovered sites.[2-4] It is in part due to the limitation of the currently available high-throughput sequencing technologies and in part due to the incompleteness of bioinformatics analysis methods.

This review summarizes modern bioinformatics approaches that discover and annotate RNA editing sites. Analytic methods do not work alone solely on RNA sequencing data. This review also includes the strategic steps used in the analysis of RNA editing, related bioinformatics resources that are useful in the analysis of RNA editing and methods to compare the analysis to further improve the analysis results.

Received: 4 June 2013, Revised: 13 June 2013, Accepted: 15 June 2013, Published: 30 June 2013
*Corresponding author: Ju Han Kim, M.D., Ph.D.
Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea
Tel: +82-2-740-8320, Fax: +82-2-747-8928, E-mail: juhan@snu.ac.kr

RNA editing

The RNA modification phenomenon of uridine (U) insertion/deletion RNA editing was discovered in the mitochondria of trypanosomatid protists in 1986.[5] RNA editing can be grouped into two basic classes-insertion/deletion editing and substitution editing. While extensive RNA editing (pan-editing) may occur in some organisms, editing in vertebrates is rare and usually consists of a small number of changes to the sequences of affected molecules.

A-to-I editing is the most common form of RNA editing in mammals.[5, 6] ADAR (Adenosine DeAminases acting on RNA) are RNA-editing enzymes involved in the hydrolytic deamination of Adenosine to Inosine (A-to-I editing)[7-9] in higher eukaryotes.[10] C-to-U editing, another editing by deamination, involves cytidine deaminase that deaminates a Cytidine base into a Uridine base. Apolipoprotein B in humans is an example of C-to-U editing with the Apo B100 form having the CAA sequence which is edited to UAA, a stop codon, to create Apo B48. Apo B100, the unedited form, is expressed in the liver and Apo B48, edited by the APOBEC enzyme, is expressed in the intestines.[11] C-to-U editing is much rarer than A-to-I editing.

RNA editing sites are located within introns, 5′UTRs, 3′UTRs, non-coding RNA sequences[6] and coding regions.[12] Especially, thousands of RNA editing sites have been discovered in Alu in human mRNAs.[13, 14] RNA editing may cause non-synonymous protein coding substitutions,[15] alternative splicing, alteration of the miRNA seed regions, and gene expression.[6]

Many RNA editing sites have been reported to be present in a variety human diseases like epilepsy, brain ischemia, depression and brain tumors.[12, 16-18] In particular, diverse cancer-related RNA editing sites have been identified in many oncogenes like glioma associated oncogene 1 (*GLI1*).[19] Some RNA editing sites have an impact on drug discovery.[20] Gene products like isoforms are created by RNA editing affecting drug activities.[19]

Recent development of high-throughput DNA and RNA sequencing technologies have contributed to the identification of new RNA editing sites and editing types such as T to C, T to A, C to A, C to T, G to T, C to G and G to C.[21] The results from these new techniques suggest that RNA editing is strongly connected to human phenotypes including several diseases.[22] Currently, one big challenge in RNA editing research is the correct identification of RNA editing sites from noisy RNA-Seq data by discriminating true RNA editing sites from Single Nucleotide Polymorphisms (SNPs) and technical artifacts caused by sequencing and analysis error.[23] For these reasons, bioinformatics approaches for RNA-Seq data analysis and identifying RNA editing sites are becoming more important. Thus, the following section discusses RNA-Seq data analysis pipelines to identify RNA editing sites and introduce bioinformatics tools and databases.

DNA-Seq and RNA-Seq data preprocessing for RNA editing research

Fig. 1 shows the steps for identifying RNA editing sites from RNA-Seq and DNA sequence (from exome sequencing or whole

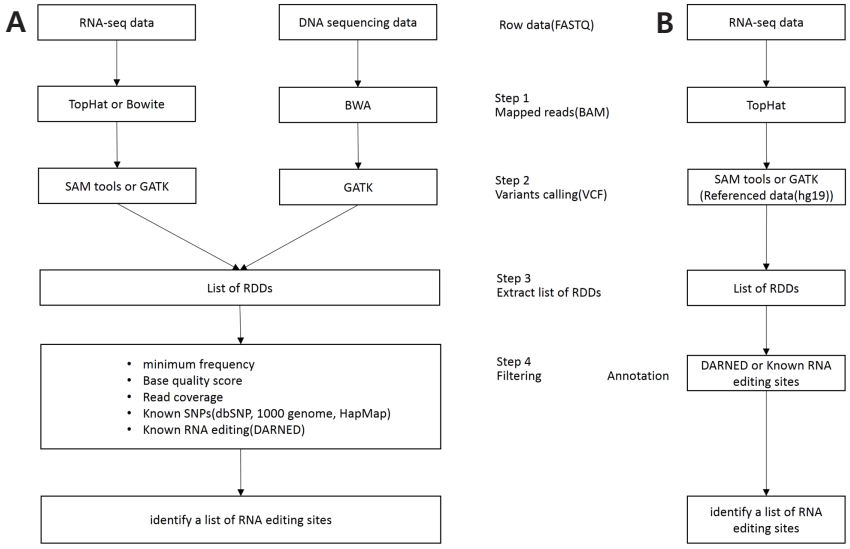

Fig. 1. Steps for identifying RNA editing sites. (A) Flowchart diagram illustrating the identification of novel RNA editing sites using both DNA and RNA sequences. (B) Flowchart diagram illustrating the annotation of known RNA editing sites to RNA-Seq data.

genome sequencing) data. It shows an analysis pipeline for the discovery of novel RNA editing sites and another for systematically annotating RNA-Seq data with known editing sites. A variety of bioinformatics resources, databases, methods and algorithms that are listed in Table 1 are being widely used in analysis pipelines these days.

Identification of novel RNA editing sites consists of two steps, read mapping and variant calling. Both Bowtie[24] and TopHat[25] are popular sequence read mapping tools. Bowtie is a fast, memory-efficient short read aligner, aligning short DNA reads to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index for memory efficiency. TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the fast short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. While TopHat is used more frequently in directly comparing RNA-seq data with reference genome alone,[26] SOAP2[27] and Burrows-Wheeler Aligner (BWA)[23] are used to identify RNA editing sites by comparing DNA and RNA sequence pairs from the same sample. BWA alone does not map reads over introns. Mapping RNA-seq reads without a method for mapping over junctions will result in exon 'islands' and will not map any reads which span an intron. For the second step of variants calling, SAMtools[28] and GATK[29] widely used. SAM files and BAM files contain the same information, but in a different format. The SAM format is a text format for storing sequence data in a series of tab delimited ASCII columns. The BAM format stores the same data in a compressed, indexed, binary form. Currently, most SAM format data are the output of aligners that read FASTQ files and assign sequences to a position with respect to a known reference genome. SAMtools receives SAM or BAM file format and sorts BAM files. Pileup or mpileup command is used to convert sorted BAM files to the VCF (Variant Call Format) format. Then VCFtools[30] performs VCF quality checking and filtering. In identifying RNA editing sites, filtering steps are very important. In most experiments, researchers check

Table 1. Resources for RNA editing research

Tools	URL	Function	Species	Ref.
Data Preprocessing				
TopHat	http://tophat.cbcb.umd.edu/	It aligns RNA-Seq reads to mammalian-sized genomes using the ultrahigh-throughput short read aligner Bowtie.	All species	Picardi et al.[27]
Bowtie	http://bowtie-bio.sourceforge.net	Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.	All species	Langmead et al.[24]
BWA	http://bio-bwa.sourceforge.net/	Fast and accurate short read alignment with Burrows-Wheeler Transform.	All species	Ramaswami et al.[23]
SAMtools	http://samtools.sourceforge.net/	Utilities for manipulating alignments in the SAM format.	All species	Li et al.[28]
GATK	http://www.broadinstitute.org/gatk/	The Genome Analysis Toolkit.	All species	McKenna et al.[29]
Databases for RNA editing				
DARNED	http://darned.ucc.ie	Collect RNA editing sites from the literature	Human, mouse	Kiran et al.[31] Kiran and Baranov[32]
dbRES	http://bioinfo.au.tsinghua.edu.cn/dbRES/		Plants, metazoan, protozoa, fungi, and virus	He et al.[33]
REDIdb	http://biologia.unical.it/py_script/REDIdb	A database for organellar RNA editing sites	198 organisms	Picardi et al.[35]
GOBASE	http://gobase.bcm.umontreal.ca/		3 bacterial genomes (*Escherichia coil* K12)	O'Brien et al.[36]
miR-EdiTar	http://microrna.osumc.edu/mireditar/	A prediction database of A to I edited miRNA target sites	Human	Laganà et al.[37]
Novel RNA editing site identification				
Rddchecker	http://genomics.jhu.edu/software/rddChecker/	Software that detects RNA-DNA differences	All species	
Annotation for RNA editing sites				
ExpEdit	http://www.caspur.it/ExpEdit	A webserver for human RNA editing in RNA-seq experiments.	Human	Picardi et al.[27]
RCARE	http://www.snubi.org/software/rcare	RNA-seq comparison and annotation for RNA editing	Human	

the minimum frequency (about 10%),[1] base quality score (a base quality score approximately greater than 25) and read coverage (a read coverage approximately greater than 10).[1,22] GATK can be used to recalibrate the base quality score.[23]

Database resources

Since the development of high-throughput DNA and RNA sequencing methods, large numbers of novel editing sites have been discovered. Most of the database resources (Table 1) provide known RNA editing site information curated from the literature. dbRES[31] is a web-oriented database for annotated RNA editing sites manually collected from the literature with related experiment results and the GeneBank database.[32] DARNED (DAtabase of RNa EDiting in humans)[33,34] is the largest database of human RNA editing sites with approximately 500,000 human RNA editing sites providing a centralized access to published data. RNA editing locations are mapped on the reference human genome with the annotation information including the region, gene, reference source and reference PubMed id. DARNED also contains 8,500 RNA editing events in Drosophila melanogaster and Mus musculus.[33] It generates a Wikipedia subsection on RNA editing entries for 16 genes and Alu repeats. REDIdb[35] and GOBASE[36] contain RNA editing sites in mitochondrion and chloroplast-encoded sequences. miR-EdiTar[37] contains predicted A-to-I RNA editing sites in miRNA binding regions.

Discovery of novel RNA editing sites

Novel RNA editing sites were discovered from a variety of human samples including B cells from 27 individuals[22] and the Han Chinese population[27] as well as from cell lines.[1] One of the popular methods is to compare RNA and DNA sequences obtained from the same sample, returning RNA-DNA differences (RDDs). After detecting the RDDs, a variety of biological resources like the 1000 genome,[38] HapMap[39] and dbSNP[40] are used to filter known SNPs from the RDDs. DARNED[33,34] is also used to filter known RNA editing sites.

Two methods for identifying RNA editing sites have recently been proposed based on RNA-Seq data from multiple samples of a single species.[23] The first method performs RNA variant calling for each RNA-Seq data after mapping sequence reads to a genomic reference sequence and filtering known SNPs. The second method performs sequence alignments using pooled reads from different RNA-Seq samples to select a higher read coverage. After alignment, it performs RNA variant calling and filters known SNPs. Given the fact that the DNA-RNA paired dataset is very rare, these methods have merits. RddChecker (http://genomics.jhu.edu/software/rddChecker/) is a representative tool for identifying RDDs and novel RNA editing sites using DNA and RNA sequencing data (Table 1). A large number of false positives are the major problem with these methods. Li et al., for example, reported that they found 28,848 RDDs with 12 different RNA-editing types including A to G (I)[22] but many studies have claimed false positives from the same dataset.[2-4]

Annotation of RNA-Seq data with known editing sites

RDDs obtained from the comparison a RNA-Seq with a reference genome sequence are annotated with known RNA editing sites queried from RNA editing databases like DARNED[33,34] to determine true RDDs. In 2011, Picardi et al. presented Expedit,[26] a web application that maps data and, given individual sequence reads as input, executes a comparative analysis against DARNED editing sites. It provides a user-friendly web interface for uploading raw RNA-Seq data like FASTQ, BAM and SAM format files and explores RNA editing sites. It annotates each RNA editing site with 10 types of information including location, gene information, source, etc, but with some limitations. It deals with the hg18 reference only. Uploading raw RNA-Seq data cannot be completed within a practical time. For example, it takes about two hours for a 700 Mb BAM file. RCARE (RNA-Seq comparison and annotation for RNA editing; http://www.snubi.org/software/rcare/) is a useful tool for identifying RNA editing sites from a variety of RNA-seq data, providing 22 types of biological information including synonymous/non-synonymous changes, splicing junctions, non-coding RNAs and gene information. It also provides seven summary plots including the rate of RNA editing type, distribution in each chromosome and origin of the samples.

Conclusion

RNA Editing is an important post-transcriptional mechanism, altering the sequence of primary RNA transcripts by deleting, inserting or modifying residues. Many studies are performing experiments to discover RNA editing sites. Each method for identifying RNA editing sites, however, has limitations. Novel RNA editing site detection methods suffer from false positives.

Known RNA editing site annotation methods are not sufficient in discovering novel RNA editing sites. Researchers should carefully consider experimental conditions and analysis methods. Due to the limitations of the current state of the art tools used in investigating RNA editing sites, one has to benchmark and test drive different methods and knowledge bases to achieve the best results. Better bioinformatics tools will emerge. We hope this review suggests a reasonable guideline for the identification of RNA editing sites using DNA and/or RNA sequencing data.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (2013-005540). S.L. is partly supported by the educational grant of NRF funded by Korea government (No. 2012M3A9D1054622).

References

1. Park E, Williams B, Wold BJ, Mortazavi A. RNA editing in the human ENCODE RNA-seq data. Genome Res 2012;22:1626-33.
2. Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science 2012;335:1302; author reply 1302.
3. Lin W, Piskol R, Tan MH, Li JB. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science 2012;335:1302; author reply 1302.
4. Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science 2012;335:1302; author reply 1302.
5. Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell 1986;46:819-26.
6. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem 2010;79:321-49.
7. Kim U, Wang Y, Sanford T, Zeng Y, Nishikura K. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. Proc Natl Acad Sci U S A 1994;91:11457-61.
8. Kumar M, Carmichael GG. Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. Proc Natl Acad Sci U S A 1997;94:3542-7.
9. Wagner RW, Smith JE, Cooperman BS, Nishikura K. A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and Xenopus eggs. Proc Natl Acad Sci U S A 1989;86:2647-51.
10. Basilio C, Wahba AJ, Lengyel P, Speyer JF, Ochoa S. Synthetic polynucleotides and the amino acid code. V. Proc Natl Acad Sci U S A 1962;48:613-6.
11. Gerber AP, Keller W. RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem Sci 2001;26:376-84.
12. Maas S, Kawahara Y, Tamburro KM, Nishikura K. A-to-I RNA editing and human disease. RNA Biol 2006;3:1-9.
13. 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.
14. Kim DD, Kim TT, Walsh T, Kobayashi Y, Matise TC, Buyske S, et al. Widespread RNA editing of embedded alu elements in the human transcriptome. Genome Res 2004;14:1719-25.
15. Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference. Nat Rev Mol Cell Biol 2006;7:919-31.
16. Cenci C, Barzotti R, Galeano F, Corbelli S, Rota R, Massimi L, et al. Down-regulation of RNA editing in pediatric astrocytomas: ADAR2 editing activity inhibits cell migration and proliferation. J Biol Chem 2008;283:7251-60.
17. Paz N, Levanon EY, Amariglio N, Heimberger AB, Ram Z, Constantini S, et al. Altered adenosine-to-inosine RNA editing in human cancer. Genome Res 2007;17:1586-95.
18. Peng PL, Zhong X, Tu W, Soundarapandian MM, Molner P, Zhu D, et al. ADAR2-dependent RNA editing of AMPA receptor subunit GluR2 determines vulnerability of neurons in forebrain ischemia. Neuron 2006;49:719-33.
19. Huang WH, Tseng CN, Tang JY, Yang CH, Liang SS, Chang HW. RNA editing and drug discovery for cancer therapy. ScientificWorldJournal 2013;2013:804505.
20. Decher N, Netter MF, Streit AK. Putative impact of RNA editing on drug discovery. Chem Biol Drug Des 2013;81:13-21.
21. Chakravarti A. Widespread promiscuous genetic information transfer from DNA to RNA. Circ Res 2011;109:1202-3.
22. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science 2011;333:53-8.
23. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, et al. Identifying RNA editing sites using RNA sequencing data alone. Nat Methods 2013;10:128-32.
24. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25.
25. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009;25:1105-11.
26. Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol 2012;30:253-60.

27. Picardi E, D'Antonio M, Carrabino D, Castrignanò T, Pesole G. ExpEdit: a webserver to explore human RNA editing in RNA-Seq experiments. Bioinformatics 2011;27:1311-2.

28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078-9.

29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20:1297-303.

30. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156-8.

31. Kiran AM, O'Mahony JJ, Sanjeev K, Baranov PV. Darned in 2013: inclusion of model organisms and linking with Wikipedia. Nucleic Acids Res 2013;41(Database issue):D258-61.

32. Kiran A, Baranov PV. DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics 2010;26:1772-6.

33. He T, Du P, Li Y. dbRES: a web-oriented database for annotated RNA editing sites. Nucleic Acids Res 2007;35(Database issue):D141-4.

34. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res 2013;41(Database issue):D36-42.

35. Picardi E, Regina TM, Brennicke A, Quagliariello C. REDIdb: the RNA editing database. Nucleic Acids Res 2007;35(Database issue):D173-7.

36. O'Brien EA, Zhang Y, Wang E, Marie V, Badejoko W, Lang BF, et al. GOBASE: an organelle genome database. Nucleic Acids Res 2009;37 (Database issue):D946-50.

37. Laganà A, Paone A, Veneziano D, Cascione L, Gasparini P, Carasi S, et al. miR-EdiTar: a database of predicted A-to-I edited miRNA target sites. Bioinformatics 2012;28:3166-8.

38. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-73.

39. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010;467:52-8.

40. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 2000;28:352-5.