nature biotechnology

The difficult calls in RNA editing

Brenda Bass, Heather Hundley, Jin Billy Li, Zhiyu Peng, Joe Pickrell, Xinshu Grace Xiao & Li Yang

Accounting for errors arising from different high-throughput sequencing platforms and those arising from the approaches used to call variants are at the center of a controversy in RNA editing.

Detecting rare sequence variation using high-throughput sequencing is fraught with pitfalls relating to sifting true variants from artifacts, as exemplified by a recent controversial study[1] of single-base changes in RNAs after transcription (RNA editing). *Nature Biotechnology* contacted several experts to discuss the current state of the detection of RNA editing events from RNA-seq data and the lessons that may apply to those facing similar analytical challenges, such as in the study of rare cells in a population or low-abundance splice variants.

***Nature Biotechnology:* How has RNA-seq been used to study RNA editing?**

Heather Hundley: RNA-seq has primarily been used to catalog RNA editing sites in both the human and mouse transcriptomes, as well as small RNAs in *Caenorhabditis elegans*, mice and humans.

Brenda Bass: Most studies of RNA editing using RNA-seq have focused on the A-to-I type of editing, the most prevalent type of editing in the nuclear-encoded RNAs of animals. However, RNA-seq approaches have also been used to analyze RNA editing in transcripts encoded in organelle genomes—for example, those of plant chloroplast and *Physarum polycephalum* mitochondria.

Xinshu Grace Xiao: The basic concept is quite simple. A mismatch between a genomic DNA sequence and an RNA transcribed from it is called a candidate RNA editing event if it is not a DNA variant or the result of a sequencing error. Therefore, most approaches involve mapping the RNA-seq reads to a genome, comparing DNA and RNA sequences and calling an RNA editing event. However, these seemingly straightforward steps are not as simple as they appear to be. Close attention must be paid to achieve accurate mapping of the reads and effective removal of potential false positives that result from various artifacts or errors.

What are the advantages of RNA-seq?

Joe Pickrell: The main advantages of RNA-seq are throughput and quantification. In terms of throughput, the entire transcriptome of a cell type can be assayed in a single sequencing experiment. In terms of quantification, the level of editing at each site can be quantified by using coverage levels of each base.

Joe Pickrell, Postdoctoral Fellow, Department of Genetics, Harvard Medical School

Jin Billy Li, Assistant Professor of Genetics, Stanford University

Jin Billy Li: In the past, RNA editing has been studied by large-scale efforts to sequence expressed sequence tags (ESTs) from cDNA using capillary Sanger sequencing. Now, a single lane from an Illumina HiSeq 2000 generates a few times more data than all of the human EST data generated by numerous efforts over many years.

X.G. Xiao: *A priori* knowledge about properties of RNA editing events, such as where they most often occur, is not needed. In this sense, the methods are unbiased. For instance, if a strand-specific RNA-seq protocol is used, editing predictions can be made without relying on known gene annotations.

H. Hundley: The ability to detect editing events in a large number of cellular targets makes identification of trends in editing patterns more apparent. Also, the methods can identify editing events that are rare, such as the small percentage of editing events present in miRNAs.

Heather Hundley, Assistant Professor of Biochemistry and Molecular Biology, Indiana University

What barriers are there to wide adoption of high-throughput sequencing?

B. Bass: High-throughput sequencing–based approaches are being widely adopted to answer questions that require a genome-wide approach, such as "How many miRNAs are edited?" or "How many editing sites are there in human brain?" or "How many editing sites are there in plant mitochondria RNA?" However, most long-time researchers in the RNA editing field are focused on questions that are less amenable to these approaches—specifically, the biological implications of a single editing event in a certain mRNA, or the mechanism by which a certain type of editing is catalyzed by a specific enzyme.

J.B. Li: The amount of data that has become available in the past few years is overwhelming for most researchers. And there is a lack of a user-friendly streamlined computational pipeline that combines the many steps required to achieve high specificity and sensitivity of calling RNA editing. Both of these barriers need to be addressed.

X.G. Xiao: One main concern is still cost. High-throughput sequencing is still expensive, and its application to RNA editing necessitates greater

sequencing depth than is normally required for gene expression studies. Second, it is still an open question whether current bioinformatic approaches can guarantee a low false-positive rate and a high true-positive rate to make the investment worthwhile. Lastly, accurately predicting RNA editing requires knowledge of the DNA sequence. Getting whole-genome sequencing or genotyping data can be expensive if large genomes are involved.

What notable contributions have these approaches made so far?

B. Bass: In all honesty, hard work by RNA editing researchers, using what would now be called archaic techniques, foreshadowed much of what we have learned from high-throughput sequencing so far. As a specific example from the A-to-I editing field, Daniel Morse, as a postdoctoral fellow in my lab, used a differential display technique to identify editing sites in *C. elegans*[2] and human brain[3]. The lesson from his studies was that A-to-I editing sites were most abundant in noncoding regions of mRNAs, such as introns and untranslated regions, and that the double-stranded structures required for editing were formed by pairing of repetitive elements. In fact, these are the overriding lessons of most high-throughput sequencing approaches in the A-to-I editing field to date. Of course, I can't emphasize enough how wonderful it is to know all of the editing events on a genome-wide basis!

Brenda Bass, Distinguished Professor & H.A. and Edna Benning Presidential Endowed Chair, Department of Biochemistry, University of Utah

In my opinion, the most noteworthy insights into RNA editing as revealed by high-throughput sequencing are just starting to appear in the literature, and the best is yet to come. These future studies will not just define editing events in the transcriptome but evaluate how they change during development or, importantly, in disease. In this regard, changes in the levels of A-to-I editing in the brains of patients with certain neuronal diseases have been observed, but are hard to interpret owing to limiting material and small datasets. No doubt future high-throughput sequencing studies will address this.

Zhiyu Peng: We have learned that RNA editing may play an important role in human disease. For example, researchers found that two new RNA editing events alter amino acid sequence of COG3 and SRP9, and showed the possible important role of RNA editing plays in breast cancer[4].

Zhiyu Peng, Vice Director of Research and Cooperation, BGI

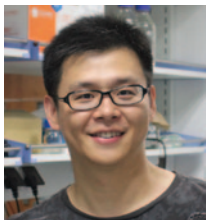
H. Hundley: One noteworthy trend that came out of the ENCODE data was the finding that although the lists of genes with edited transcripts are similar between cell lines, the individual editing sites vary[5]. This suggests that it may be more important for a transcript to be 'marked' by editing than whether or not a specific adenosine undergoes deamination. Because A-to-I editing occurs within double-stranded (ds) regions of RNA, and dsRNAs are potent triggers of the antiviral response, editing may provide a critical means for preventing dsRNA structures within cellular transcripts from improperly activating the immune system.

What sources of error confound variant calling in RNA-seq data?

J.B. Li: Based on our experience, mapping error is the main source, although sequencing errors inevitably affect accuracy. Once the reads are actually mapped, the challenge is to distinguish RNA editing events from genomic SNPs (see **Box 1** for more details).

Are these sources of error unique to RNA editing analysis?

Li Yang: Compared with SNP [single-nucleotide polymorphism] calling, most of these sources of error are similar, except for some features derived from transcriptional processes. For example, some alignment errors, such as the incorrect alignment at RNA splicing junctions, are unique to RNA editing analysis. In addition, genes can be disproportionately transcribed from sister chromosomes and from heterozygous variant sites. In these cases, a genomic variant can be easily misinterpreted as a potential RNA editing site.



Li Yang, Principal Investigator, Group Leader, CAS-MPG Partner Institute for Computational Biology, Shanghai, China

X.G. Xiao: Errors in read mapping are by no means unique to RNA editing analysis. All applications of high-throughput sequencing are faced with mapping inaccuracies. However, the impact of such errors is particularly severe when the data are analyzed at single-nucleotide resolution or when low-abundance reads are sought after. Statistically, low-abundance reads more often reflect mapping errors than do high-abundance ones. Similar problems exist in other applications where low-abundance reads are relied upon to make interpretations. In contrast, there are applications that are more robust to mapping errors, such as the estimation of overall gene expression levels, where mapping errors normally affect a minor fraction of all reads of a gene.

How are these problems being addressed?

J.B. Li: I believe that, in most cases, distinct approaches need to be developed for different problems. However, some unified efforts, such as accurate mapping of RNA-seq reads, can be generalized to different challenges where RNA-seq read mapping is a critical component.

For RNA editing, we and others have pinpointed some major sources of error[6–8] and subsequently developed new approaches to alleviate the problems[5,9–12].

J. Pickrell: There is unlikely to be a 'magic bullet' that solves these problems in a completely unified way. However, the similarities between the issues in, say, identifying RNA editing and identifying SNPs means that many lessons learned in one can be applied to the other. For example, tests for strand and position biases when calling variants from short reads have become standard in both situations.

What kinds of new technologies are needed?

J. Pickrell: Sequencing technologies that both provide long reads (thus avoiding many of the read mapping issues) and perform direct sequencing of RNA (rather than converting to cDNA) would dramatically decrease the false-positive rates in these sorts of studies.

X.G. Xiao: Deeper sequencing depth, once it's more affordable, will enable more rigorous statistical tests to make editing calls. In terms of algorithms, more stringent and accurate mapping approaches need to be developed. Easy-to-use analytic tools with proven accuracy and reliability could make RNA editing analysis a standard procedure in RNA-seq data analysis.

Box 1 Dealing with false positives

A *Science* paper[1] published last year reported large numbers of differences between DNA and messenger RNA in human cells, indicating unprecedented levels of sequence changes through RNA editing, many of which could not have arisen through known RNA editing mechanisms. We asked several investigators to discuss the difficulties of calling true variants from massive quantities of noisy data.

What can cause false positives in RNA editing analysis?

J. Pickrell: There are multiple important sources of false-positive RNA editing sites when using RNA-seq data. First, for most RNA-seq protocols, one must first convert RNA to cDNA; mismatches between the primers used in the reaction and the RNA can introduce mutations (analogous to site-directed mutagenesis) that could be falsely interpreted as RNA editing sites. Additionally, one must often compare RNA-seq reads to a reference genome to identify mismatches between the RNA and DNA. Any errors in the reference genome (such as sections of DNA that are absent from the reference, or unnoticed copy-number variants) can lead to errors in mapping RNA-seq reads and could lead to false positives. Paralogous genes (those with very similar sequences) are another major source of error, as RNA-seq reads mismapped between paralogs can be difficult to distinguish from true editing.

L. Yang: Sequencing errors are systematic and dependent on the technology. These errors are frequently found at the ends of short reads. Alignment errors usually happen at splicing junction sites and introns. Certain repetitive regions, such as *Alu* elements, are hard to map precisely but are highly edited. Thus, many highly edited but possibly important regions are likely to be removed by most current methods of alignment.

How have these problems been dealt with?

B. Bass: The gold standard for dealing with false positives due to sequencing errors is to compare high-throughput sequencing data from wild-type animals with data from animals deficient for the RNA editing enzyme of concern. A true editing site will be absent from the mutant animal. Of course, this option is not available when evaluating organisms in which the editing enzyme is essential for viability. Scientists studying A-to-I editing in mammals are faced with this problem and have used a variety of methods to minimize false positives. These include enriching for sequences that include other characteristics of edited regions, such as double-stranded structures. When evaluating false positives among potential editing sites that may represent a previously uncharacterized type of editing, it is imperative that a subset of the candidate editing events be validated with more conventional 'low-throughput' molecular or biochemical techniques.

J.B. Li: We have learned a great deal to accurately map the short RNA-seq reads. For example, we trim the first six bases of the reads because mismatch errors are introduced during first-strand cDNA synthesis using random hexamers. Furthermore, we map reads to the reference genome and all splice junction sequences and use BLAT to ensure unique mapping of reads. To distinguish SNPs from RNA editing, we remove all known SNPs from the called RNA variants even when the matched genome sequence of the same sample is available.

H. Hundley: Experimentally, as the price of next-generation sequencing has dramatically declined, it will become more feasible to sequence both the genomic DNA and RNA from the same individual, thus eliminating the identification of SNPs as editing sites.

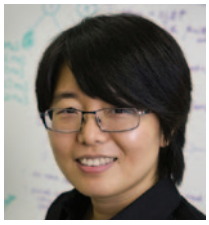
X.G. Xiao: Most methods focused on removing likely false positives after the initial editing calls using heuristic filters. In an ideal world, we want to do accurate read mapping in the first place, which can make a method less dependent on the post-filtering steps. We published one of the few studies that described improved strategies to reduce mapping errors[9].

Z. Peng: Current pipelines cannot reliably call RNA editing events in genes that are expressed at low levels.

L. Yang: As many transcriptome datasets are publicly available without a matched genome, new approaches are needed to predict RNA editing events in the absence of matching genomic DNA sequence.

What issues are most important for the field going forward?

X.G. Xiao: We lack a gold standard data set to validate bioinformatically predicted RNA editing calls. The validation method used most often is Sanger sequencing of RT-PCR products, which is known to have limited sensitivity and low quantification accuracy. We have used clonal sequencing (where we clone RT-PCR products into vectors and Sanger sequence a large number of randomly picked clones), which can have great sensitivity and accuracy if enough clones are sequenced.



Xinshu Grace Xiao, Assistant Professor, University of California, Los Angeles

But the cost can be prohibitive.

There is a need for a unified effort to set up a gold standard database of RNA editing events. Perhaps we can choose a common cell line for which abundant RNA-seq data are available and collectively conduct deep clonal sequencing (or other experimental validation) on a subset of RNA editing sites (for example, 20–30 sites). All bioinformatic methods can then be applied to the RNA-seq data and benchmarked against the experimental gold standard.

J.B. Li: The devil is in the details. This is very true in the recent debate on the RNA editing work. The multitude of existing RNA-seq read mapping tools, while seemingly sufficient for most applications of RNA-seq (for example, to gene expression levels, splicing junctions and fusion genes), perform poorly when challenged to look for single-base mismatches, such as RNA editing sites. A deeper understanding of computation, technology and biology can only be achieved by two-way communication between computational biologists and experimentalists.

1. Li, M. *et al. Science* **333**, 53–58 (2011).
2. Morse, D.P., *et al. Proc. Natl. Acad. Sci. USA* **96**, 6048–6053 (1999).
3. Morse, D.P., *et al. Proc. Natl. Acad. Sci. USA* **99**, 7906–7911 (2002).
4. Shah, S.P. *et al. Nature* **461**, 809–813 (2009).
5. Park, E. *et al. Genome Res.* **22**, 1626–1633 (2012).
6. Pickrell, J.K. *et al. Science* **335**, 1302 (2012).
7. Lin, W. *et al. Science* **335**, 1302 (2012).
8. Kleinman, C.L. & Majewski, J. *Science* **335**, 1302 (2012).
9. Bahn, J.H. *et al. Genome Res.* **22**, 142–150 (2012).
10. Peng, Z. *et al. Nat. Biotechnol.* **30**, 253–260 (2012).
11. Ramaswami, G. *et al. Nat. Methods* **9**, 579–581 (2012).
12. Kleinman, C.L. *et al. RNA* **18**, 1586–1596 (2012).

Interviewed by H. Craig Mak, Associate Editor, Nature Biotechnology