BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS. VOL 5. NO 1. 43–45 doi:10.1093/bfgp/ell014

Algorithmic approaches for identification of RNA editing sites

Erez Y. Levanon and Eli Eisenberg

Advance Access publication date 22 February 2006

Abstract

Recently a number of groups have introduced computational methods for the detection of A-to-I RNA editing sites. These approaches have resulted in finding thousands of editing sites within the genomic repeats, as well as a few novel genetic recoding sites. We review these recent advancements, emphasizing the principles underlying the various methods used. Possible directions for extending these methods are discussed.

Keywords: *RNA editing; genomic repeats; human-mouse comparison*

Adenosine-to-inosine (A-to-I) RNA editing is a modification in the RNA molecule that alters the original DNA content. It occurs immediately following transcription and before splicing. When the newly formed RNA has a double-stranded RNA structure (dsRNA), a member of the adenosine deaminases that act on RNA (ADARs) protein family can attach and deaminate some of the adenosines (A) within the double-stranded region into inosines (I). The ribosome and the splicing enzymes, as well as sequencing machines, recognize the inosine as guanosine (G) [1].

Till recently, a very small number of A-to-I editing targets were identified in the human genome, mostly due to chance discoveries. Nevertheless, the functional importance of this mechanism was established by showing that mouse lacking ADARs die *in utero* or shortly after birth [2–4]. In addition, a number of neurological diseases were associated with altered editing patterns [5–8]. Not all these phenotypes are explained by the limited number of editing targets identified, suggesting one should continue looking for more editing sites. Experimental approaches to find additional editing events were developed [9]. However, within current technology only a small fraction of editing sites has been detected by these methods.

In principle, computational identification of editing events should be straightforward. The sequencing machinery reads an edited site within an expressed sequence as a 'G', where the corresponding genome position will be an 'A'. Thus, one has only to compare the millions of publicly available expressed sequences with the genome and look for such inconsistencies. However, this naive approach is bound to fail due to the large number of mismatches between the genome and the expressed sequences due to other reasons. Major sources for such a mismatch are genomic polymorphisms. As different expressed data are derived from different individuals, they are known to have millions of sites along the genome where two sequences do not agree. In addition, tens of millions of random sequencing errors in the expressed sequences may look as editing sites when aligning them to the genome. Additional causes of variance between RNA and the genome include mutations and inaccurate alignment of the RNA sequence data to the genome due to duplications.

The known recoding editing sites—where editing affects the resulting protein—have a common characteristic. The vicinity of these sites is highly conserved between species [10]. This is due to the evolutionary constraint to keep the dsRNA structure intact, in addition to maintaining the coding information. This constraint leads to conservation in the DNA level, and has proven to be very useful for bioinformatic searches for more

Corresponding author. Eli Eisenberg, School of Physics and Astronomy, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Tel: +972 3 640 7723; Fax: +972 3 640 2979. E-mail: eli.eisenberg@gmail.com

Erez levanon is a PhD student in the Sackler School of Medicine, Tel Aviv University (Tel Aviv, Israel) under the guidance of Prof. Gideon Rechavi. He also works as a research scientist in the biotechnology company Compugen Ltd.

Eli Eisenberg is a Senior Lecturer in the School of Physics and Astronomy in Tel Aviv University. His interests include mesoscopic physics and statistical mechanics as well as bioinformatics.

candidates [10–12]. The conservation of the editing site is used as a sieve through which one sifts the few editing recoding sites out of the tens of millions of mismatches between expressed sequence tags (ESTs)/RNAs and the genome, using the observation that the sequencing errors and the single nucleotide polymorphisms (SNPs) are not evolutionarily conserved between species while editing recoding sites are. In a recent study, we employed this strategy, looking for such conserved mismatches located in the exactly same position in human and mouse. The search resulted in four additional A-to-I editing substrates [11]. We note that using this approach, one seems to be better off not implementing the requirement for a dsRNA structure, as the typical dsRNA structures of the few known targets are rather weak and hard to predict computationally [13]. Another point to note is that these four editing sites all appear in dbSNP, since the variability of the expressed sequences in these sites was (erroneously) interpreted as a sign for an SNP [14]. Thus, one should use dbSNP carefully when searching for editing sites. On the other hand, dbSNP might be actually used as an alternative starting point in looking for additional editing targets.

In addition to the few isolated sites within the coding sequence (editing of which might result in an amino acid substitution), a large number of clusters of editing events were recently found in non-coding regions. Recently, three computational methods of identification of such clusters of mismatches in the alignments of the clean RNA set were published [15–18]. The methods differ by the clustering criterion used, ranging from a detailed statistical model to a simple count of consecutive mismatches of the same type. Impressively, all three procedures have yielded highly similar results: A-to-G substitutions, standing for A-to-I editing events account for more than 80% of the 12 possible types of mismatches in the selected set of transcripts.

Almost all of these clusters occur within Alu repetitive elements, which are short interspersed elements (SINEs). There are about a million copies of Alu in the human genome, roughly 300 bp long each, together accounting for ~10% of the genome [19]. Since they are so common, especially in gene-rich regions, pairing of two nearby, oppositely oriented, Alus in the same pre-mRNA structure is likely, resulting in a long and stable dsRNA structure. Such structures are ideal targets for the ADARs.

Editing events occur before splicing, thus they may occur in introns as well. However, computational approaches based on expressed sequences are obviously limited in their ability to detect editing within introns. Therefore, it is anticipated that the actual number of editing sites in the human genome is even much higher than the tens of thousands of sites reported in the above works. Indeed, direct sequencing of human brain total RNA has revealed that up to 1 in 1000 bp of the expressed regions are being edited [20].

Analysis of the editing events detected has taught us more about the nature of the process. Weak sequence preferences for the nucleotides preceding and following the editing sites are observed, presumably attesting for ADAR binding preferences. There is also some evidence that the local dsRNA structure may play a role in targeting of the ADARs. Analysis of the distance between edited Alus and their nearest reverse complement Alu have shown that effective editing requires a distance of roughly 2000 bp or less between the two Alus. Further support for the paired Alus model comes from the observation that the more reverse complement Alus within this distance, the higher is the level of editing [16–18, 20]. Finally, it was shown that the edited adenosines within the dsRNA structure are paired with a 'U' or a 'C' in the reverse strand, meaning that editing is either strengthening or weakening the dsRNA structure, but virtually never has a neutral effect on the dsRNA pairing energy [18]. This last result suggests a regulatory role for RNA editing in controlling dsRNA stability, in accordance with recent observations suggesting that editing is involved in molecular mechanisms based on dsRNA structure, like RNAi [21] and miRNA [22]. Such knowledge on the characteristics of ADAR targets might turn out to be instrumental in future searches for editing targets.

Alu repetitive elements are unique to the primates, but the occurrence of repetitive elements in general is common to all metazoa. Interestingly, applying the same methods in looking for clusters of editing sites in other organisms have shown that there are about 40 times fewer editing events in mouse as compared with the human genome [15, 16]. A similar picture was observed in rat, chicken and fly [15]. The reason for this huge difference is likely the fact that in human there is only one dominant SINE, which is relatively less diverged (~12% average divergence). In mouse, for example, there

are four different SINEs, which are shorter and more divergent (~20% average divergence). It is tempting to link the over-representation of editing in brain tissues and the association of aberrant editing with neurological diseases, and speculate that the massive editing of brain tissues is responsible in part for the brain complexity and thus the massive invasion of Alus to the primate genomes, which allowed this abundant editing may have played a role in the evolution of primates.

What are the next challenges for computational identification of editing sites? First, a full account for the role of the characteristics of repetitive elements determining the editing level is still lacking. One would like to be able to predict which elements are likely to be edited, and what is the expected level of editing in a given organism. Second, more work is required in order to supply experimentalists with hints for the mysterious role of the abundant editing phenomena. One can hope that analysis of the tissue-origin of the edited sequences might provide us with directions to attack this question. Finally, there are almost no strategies yet for computational search for other types of RNA editing, in particular C-to-U editing. Very few examples of this process are known in mammals [23], but it is anticipated that the use of expressed data and evolutionary conservation accompanied by additional unique features of these types of editing will be of use to reveal the full spectrum of the transcriptome.

Key Points

- Recent bioinformatic studies have shown that RNA editing is very common in the human transcriptome.
- Clusters of mismatches between RNA sequences and the corresponding genomic DNA sequences may distinguish between RNA editing and 'noise'.
- Most of A-to-I editing events occur in the primate-specific Alu repeat.
- Few editing events are at genetic recoding sites, modifying the resulting protein. Such sites are typically located in highly conserved genomic regions.
- Better and more efficient algorithms are still required to reveal the full spectrum of editing in the genome.

References

1. Bass BL. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* 2002;**71**:817–46.
2. Hartner JC, Schmittwolf C, Kispert A, *et al*. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. *J Biol Chem* 2004; **279**:4894–902.
3. Higuchi M, Maas S, Single FN, *et al*. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* 2000;**406**:78–81.
4. Wang Q, Khillan J, Gadue P, *et al*. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science* 2000;**290**:1765–8.
5. Maas S, Patt S, Schrey M, *et al*. Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc Natl Acad Sci USA* 2001;**98**:14687–92.
6. Kawahara Y, Ito K, Sun H, *et al*. Glutamate receptors: RNA editing and death of motor neurons. *Nature* 2004;**427**: 801.
7. Gurevich I, Tamir H, Arango V, *et al*. Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron* 2002;**34**:349–56.
8. Brusa R, Zimmermann F, Koh DS, *et al*. Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. *Science* 1995;**270**:1677–80.
9. Morse DP, Bass BL. Long RNA hairpins that contain inosine are present in *Caenorhabditis elegans* poly(A)+ RNA. *Proc Natl Acad Sci USA* 1999;**96**:6048–53.
10. Hoopengardner B, Bhalla T, Staber C, *et al*. Nervous system targets of RNA editing identified by comparative genomics. *Science* 2003;**301**:832–6.
11. Levanon EY, Hallegger M, Kinar Y, *et al*. Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res* 2005;**33**:1162–8.
12. Clutterbuck DR, Leroy A, O'Connell MA, *et al*. A bioinformatic screen for novel A-I RNA editing sites reveals recoding editing in BC10. *Bioinformatics* 2005;**21**: 2590–5.
13. Bhalla T, Rosenthal JJ, Holmgren M, *et al*. Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nat Struct Mol Biol* 2004;**11**:950–6.
14. Eisenberg E, Adamsky K, Cohen L, *et al*. Identification of RNA editing sites in the SNP database. *Nucleic Acids Res* 2005;**33**:4612–7.
15. Eisenberg E, Nemzer S, Kinar Y, *et al*. Is abundant A-to-I RNA editing primate-specific? *Trends Genet* 2005;**21**:77–81.
16. Kim DD, Kim TT, Walsh T, *et al*. Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res* 2004;**14**:1719–25.
17. 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.
18. Levanon EY, Eisenberg E, Yelin R, *et al*. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol* 2004;**22**:1001–5.
19. Lander ES, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–21.
20. Blow M, Futreal PA, Wooster R, *et al*. A survey of RNA editing in human brain. *Genome Res* 2004;**14**:2379–87.
21. Tonkin LA, Bass BL. Mutations in RNAi rescue aberrant chemotaxis of ADAR mutants. *Science* 2003;**302**:1725.
22. Luciano DJ, Mirsky H, Vendetti NJ, *et al*. RNA editing of a miRNA precursor. *Rna* 2004;**10**:1174–7.
23. Wedekind JE, Dance GS, Sowden MP, *et al*. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet* 2003;**19**:207–16.