Biochemical and Biophysical Research Communications 412 (2011) 407–412

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Genome-wide evaluation and discovery of vertebrate A-to-I RNA editing sites

S. Maas [a,*], C.P. Godfried Sie [a], I. Stoev [b], D.E. Dupuis [a], J. Latona [a], A.M. Porman [a], B. Evans [a], P. Rekawek [a], V. Kluempers [a], M. Mutter [a], W.M. Gommans [a], D. Lopresti [b]

[a] Department of Biological Sciences, Lehigh University, Bethlehem, PA, United States
[b] Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, United States

ARTICLE INFO

Article history:
Received 18 July 2011
Available online 2 August 2011

Keywords:
RNA editing
ADAR
Inosine
Adenosine deamination

ABSTRACT

RNA editing by adenosine deamination, catalyzed by adenosine deaminases acting on RNA (ADAR), is a post-transcriptional modification that contributes to transcriptome and proteome diversity and is widespread in mammals. Here we administer a bioinformatics search strategy to the human and mouse genomes to explore the landscape of A-to-I RNA editing. In both organisms we find evidence for high excess of A/G-type discrepancies (inosine appears as a guanosine in cloned cDNA) at non-polymorphic, non-synonymous codon sites over other types of discrepancies, suggesting the existence of several thousand recoding editing sites in the human and mouse genomes. We experimentally validate recoding-type A-to-I RNA editing in a number of human genes with high scoring positions including the coatomer protein complex subunit alpha (COPA) as well as cyclin dependent kinase CDK13.

Published by Elsevier Inc.

1. Introduction

The post-transcriptional modification of adenosine to inosine in RNA molecules is a widespread mechanism in multicellular animals for generating RNA and protein variation [1,2]. In primates, Alu-repeats have been shown to constitute a major target of adenosine deamination [3,4]. Furthermore, since inosine is interpreted as guanosine during translation, the site-selective alteration of single adenosines within protein-coding sequences can critically modulate protein function as a result of non-synonymous codon changes that cause amino acid substitutions [2].

A-to-I editing is mediated by adenosine deaminases acting on RNA (ADAR), of which ADAR1 and ADAR2 are thought to mediate all known editing events [1]. They recognize partially double-stranded (ds) RNA targets through several dsRNA binding domains. However, the exact mode of interaction and how site-selectivity is achieved are unknown. Repeat element mediated editing such as in Alus exhibits low specificity and is mainly driven by the strong ds-character of the RNA secondary structure. In contrast, in recoding editing single adenosine residues in pre-mRNA molecules are targeted with high specificity. It is currently not possible to predict an RNA editing site based on RNA sequence or predicted RNA secondary structure. In fact, most characterized recoding targets have been identified serendipitously.

Still, the increasing number of validated RNA editing target sequences and secondary structures reveal a bias toward certain molecular characteristics. For example, the local sequence environment influences whether editing occurs by ADAR1 or ADAR2, respectively, due to enzyme-specific preferences for certain bases preceding and following the targeted A. Another critical feature is a partially base-paired RNA secondary structure involving sequences flanking the editing site(s) [1]. Also, such sequences are more highly conserved across species since their involvement in forming a functional RNA secondary structure exerts increased selection pressure [5].

Even though above features contribute to establishing an editing target, none is sufficient to allow for straight-forward prediction of *bona fide* editing sites. Despite these limitations, several new targets of editing have recently been identified using bioinformatics strategies that employ a combination of these molecular features as filter criteria [6–8]. In addition, high-throughput sequencing has also helped to validate potential cases of editing [9,10]. However, the ad hoc nature of previously adopted bioinformatics strategies and the observation that very little if any overlap between predicted candidate lists emerged strongly limits the genome-wide evaluation of the prevalence and importance of A-to-I editing.

We developed the bioinformatics-based search tool RNA Editing Dataflow System (REDS) that allows for a more comprehensive screening, which, combined with standard or high-throughput experimental validation, would facilitate mapping the A-to-I RNA editing landscape and defining the overall impact of editing on gene expression. We recently explored the feasibility of our search

* Corresponding author. Permanent address: National Institute of General Medicine, NIH, 45 Center Drive, Bethesda, MD 20892, United States, Current address: InteRNA Technologies BV, Padualaan 8, 3584 CH Utrecht, The Netherlands. Fax: +1 610 758 4004.
E-mail address: swm3@lehigh.edu (S. Maas).

0006-291X/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.bbrc.2011.07.075

strategy using a subset of human genes and experimentally confirmed several predicted novel editing sites [7]. Here, we applied our search strategy to the genomic scale to analyze the overall landscape of base discrepancies in two species followed by experimental validation of novel recoding editing sites in the human transcriptome.

2. Results and discussion

2.1. Evidence for abundant, site-selective recoding A-to-I editing in human and mouse

In eukaryotes, A-to-I RNA editing is the only known mechanism for generating inosine residues in RNA molecules. First we asked whether an excess of A/G discrepancies versus other types of base differences between cDNA and genomic DNA is detectable, even when excluding repetitive element mediated editing. Such a finding would support the hypothesis that many more editing sites within protein-coding sequences remain to be identified and may provide an estimate on the total number of existing sites. We aligned all sequences available in the human and mouse mRNA databases (from UCSC genome browser) [11] to their genomic counterparts and mapped the positions of A/G or other discrepancies between genomic and expressed sequence (Fig. 1). Since there is no known mechanism for causing G/A or T/C transitions, we expect that such discrepancies will be due to polymorphisms (SNPs) and/or sequencing errors and can therefore be considered background noise.

In the combined coding and non-coding regions of mRNA sequences there are substantially more A/G-type differences in human than either G/A, C/T or T/C discrepancies (68,000 A/G versus 57,000 G/A; 58,000 C/T and 60,000 T/C). In mouse, the total numbers of discrepancies are slightly fewer and the excess of A/G discrepancies is less striking (62,000 A/G versus 56,000 G/A; 58,000 C/T and 55,000 T/C). The difference to human can probably be accounted for by the widespread Alu-repeat mediated editing in primates, as we and others have previously mapped thousands of such editing sites in non-coding human mRNA sequences [3,4]. In contrast, rodents lack Alu-repeats and display only low levels of repeat element mediated editing. This conclusion is further supported when we exclude all discrepancies located within non-coding regions from our analysis (Fig. 1). In both human and mouse, A/G discrepancies are still the most prominent type of discrepancy and the observed excess of A/G in human is now comparable to that in the mouse dataset.

We observe a very strong A/G-discrepancy bias when further restricting our analysis to non-synonymous codon changes (Fig. 1). Editing at such sites will change the meaning of the codon and lead to amino acid substitutions in the resulting protein. There is now a substantial overrepresentation of A/G discrepancies in both human and mouse (1.4–2.1-fold in human and 1.3–2.0-fold in mouse) compared to all other types of transitions. When all known, genomically validated SNPs are subtracted from the list, the excess of A/G discrepancies further increases (1.6–2.2-fold in human and 1.4–2.1-fold in mouse). Moreover, when we eliminate redundant sites identified in other RNA sequences (Fig. 1A and B; # of mRNAs), the excess of A/G discrepancies persists (1.5–2.0-fold in human and 1.4–2.0-fold in mouse). Therefore, A/G discrepancy sites are distributed across many genes and not dominated by a small number of highly expressed genes that are overrepresented in the expression databases.

Our analysis shows that even when just considering protein coding sequences, thereby eliminating the impact of repeat-mediated editing, there is still a substantial excess of A/G versus other transitions in the human and mouse transcriptomes. We see a surplus of ~4700 (human) and 5800 (mouse) unique sites compared to the average of all other types of changes. As many editing sites *in vivo* are modified to only a small extent or in a cell-type specific fashion, the number of discrepancies detected at this time may still represent an underestimate of the total number of editing sites in these species. Similarly, negative results from experimental evaluation of potential editing sites cannot rule out editing in another cell type or at another time point. Alternatively, editing may occur at such a low level that the experimental assay is not sensitive enough for detection. In fact, two recent high-throughput sequencing studies suggest that the bulk of target sites may be edited to a low extent [10,12]. Such a result is also predicted by the continuous probing (COP) hypothesis regarding the possible mechanism of how novel editing sites in the transcriptome emerge [13]. Due to the high accuracy sequence databases we utilize for the analysis (not including EST-type sequences), we expect that among the predicted ~10,000 potential sites almost half may reflect real RNA editing events. Since A-to-I RNA editing is the only eukaryotic mechanism known to generate A/G-discrepancies, the excess over the background of SNPs and sequencing errors points to the existence of thousands of additional editing sites to be characterized.

2.2. Experimental validation of novel editing sites in the human transcriptome

High-scoring sites predicted by REDS that were obtained when applying stringent parameter settings (window1 = 500nt, window2 = 20nt, minimal base-pairs = 11) were experimentally analyzed (Supplementary Tables T1 and T2). Out of 32 known and validated A-to-I editing sites (considered the ‘true positives’ for REDS analysis (Supplementary Table T3)), 30 positions are detectable by REDS due to the presence of at least one mRNA sequence in the database that is of the edited variant (24 in case of mouse).

Editing for several genes after parallel analysis of cDNA and gDNA from human specimen was confirmed *in vivo* (Supplementary Table 1). In addition, 10 sites recently experimentally validated by Li et al. [10] as well as one site validated by Shah et al. [14] are also predicted as high-scoring targets by REDS, as are the other human validated RNA editing sites known at the time. Supplementary Table T2 lists the top ten candidates in human and mouse based on a single, continuously base-paired segment.

Figs. 2 and 3 show four novel human editing targets that harbor a total of 15 validated editing sites. ATP6V0E2 is predicted to form an extended, highly base-paired dsRNA structure with two neighboring segments of 72 and 68 bp, respectively, which are not part of any repetitive-type elements. Five prominent A-to-I RNA editing sites were experimentally validated, four of which lead to non-synonymous codon changes (Fig. 2A). Editing levels are between 30% and 70% and the codons are all located within a single alternatively spliced exon. Based on *in silico* analysis, only about 1.1% of transcripts (EST and mRNA sequences), all of which show evidence of multiple editing sites, contain the alternative exon, which appears to be brain-specific.

Fig. 2B documents editing in transcripts of a gene (BC027448) of unknown function at a total of eight sites with efficiencies between 5% and 95%. It is not known whether it is translated *in vivo*, or what part of the sequence may constitute a functional open reading frame. The predicted RNA secondary structure forms a hairpin with adjacent segments of 23 and 15 continuous base-pairs (Fig. 2B), entirely positioned within the same exon.

The human coatomer protein complex subunit alpha (COPA) was predicted as a candidate editing target from analysis of zebrafish databases and experimentally validated (Maas et al., unpublished). With no edited version of COPA annotated in the human mRNA database, it would have evaded prediction by REDS, had