BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 9 2011, pages 1311–1312
doi:10.1093/bioinformatics/btr117

Genome analysis

Advance Access publication March 22, 2011

ExpEdit: a webserver to explore human RNA editing in RNA-Seq experiments

Ernesto Picardi[1], Mattia D'Antonio[2], Danilo Carrabino[2], Tiziana Castrignanò[2] and Graziano Pesole[1,3,*]

[1]Dipartimento di Biochimica e Biologia Molecolare 'E. Quagliariello', Università di Bari, 70125 Bari, [2]Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, 00185 Rome and [3]Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, 70125 Bari, Italy

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: ExpEdit is a web application for assessing RNA editing in human at known or user-specified sites supported by transcript data obtained by RNA-Seq experiments. Mapping data (in SAM/BAM format) or directly sequence reads [in FASTQ/short read archive (SRA) format] can be provided as input to carry out a comparative analysis against a large collection of known editing sites collected in DARNED database as well as other user-provided potentially edited positions. Results are shown as dynamic tables containing University of California, Santa Cruz (UCSC) links for a quick examination of the genomic context.

Availability: ExpEdit is freely available on the web at http://www.caspur.it/ExpEdit/.

Contact: graziano.pesole@biologia.uniba.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 27, 2010; revised on February 26, 2011; accepted on February 28, 2011

1 INTRODUCTION

RNA editing is a post-transcriptional mechanism challenging the central dogma of molecular biology. According to such phenomenon, genetic information is enzymatically overwritten at specific sites to generate transcripts different from their corresponding genomic templates [for a comprehensive review see Gott and Emeson (2000)]. Historically, RNA editing has been described for the first time as the insertion and the deletion of uridine residues in mitochondrial RNAs of kinetoplastid protozoa (Benne *et al*., 1986). Nowadays, the term RNA editing is also used to indicate post-transcriptional changes due to specific base substitutions (Gott and Emeson, 2000). Such alterations may affect coding as well as non-coding RNAs located in different cellular compartments and occur in a variety of organisms including prokaryotes, animals, plants and viruses, through unrelated biochemical mechanisms (Gott and Emeson, 2000).

In mammals and, in particular in human nuclei, the A-to-I conversion is the most frequent type of RNA editing catalyzed by the ADAR (adenosine deaminase acting on RNA) family of enzymes that requires double stranded RNA as a substrate to carry out the deamination (Keegan *et al*., 2001). Inosines are commonly interpreted as guanosines by splicing and translational machineries as well as sequencing enzymes. Several A-to-I changes have been described in coding regions causing amino acid substitutions with functional consequences. The vast majority of editing sites, however, have been recently found in repeated regions of the human genome and especially in *Alu* rich segments (Barak *et al*., 2009).

Although the functional and physiological role of many human edits is currently elusive, the molecular mechanism of RNA editing is indispensable to preserve the cellular homeostasis. Deregulation of RNA editing has been associated to a variety of pathological phenotypes including several neurodegenerative diseases (Maas *et al*., 2006). Moreover, mice lacking the ADAR activity die during or soon after the weaning (Keegan *et al*., 2001).

A-to-I conversions in human have been mainly discovered computationally by analyzing alignments of messenger RNA (mRNA)/expressed sequence tag (EST) sequences against their source genome (Levanon *et al*., 2004). Padlock experiments in combination with deep sequencing have also been used to confirm computationally selected editing candidates in different human tissues (Li,J.B. *et al*., 2009). Nowadays, >40 000 editing changes are known in human and they are freely available through the specialized DARNED database (Kiran and Baranov, 2010).

The extraordinary coverage provided by RNA-Seq methodology can now provide a comprehensive transcriptome snapshot dramatically improving our understanding of gene expression dynamics in different tissues, developmental stages or pathological conditions (Wang *et al*., 2009). Massive transcriptome sequencing offers the great opportunity to explore and investigate post-transcriptional changes due to RNA editing (Picardi *et al*., 2010). Indeed, the huge amount of short reads generated by the RNA-Seq provides significant support for individual genomic positions after the appropriate mapping strategy and facilitates the identification of A-to-G conversions. RNA-Seq experiments, therefore, may allow large-scale detection of editing sites thus contributing to elucidate their functional roles in a variety of tissues, developmental stages and physiological conditions. A large amount of RNA-Seq data is now publicly available through the SRA repository at National Center for Biotechnology Information (NCBI). However, no tool has been developed so far to investigate such data, as well as newly produced RNA-Seq data. To fill this gap we developed an *ad hoc* web

*To whom correspondence should be addressed.

service freely available at http://www.caspur.it/ExpEdit, specifically designed for exploring the editing pattern in human.

2 IMPLEMENTATION

The core of ExpEdit has been developed in python programming language by using the SAMtools package (Li,H. *et al*., 2009) and hypertext preprocessor (PHP) scripting to dynamically generate web pages. To run the application, the user needs to upload one or more short read datasets in the standard FASTQ or SRA format (gzip or bz2 compression for FASTQ is also admitted to speed up the upload process). Short reads are then quickly mapped onto the complete human genome by Bowtie (Langmead *et al*., 2009), tolerating by default at most three mismatches in the simpler end-to-end alignment mode and requiring uniquely mapping reads. Nonetheless, Bowtie parameters can be freely adjusted before the submission. As an alternative, pre-aligned reads by other software can be uploaded in SAM or BAM format. ExpEdit can handle paired-end reads and performs spliced alignments by means of Tophat program based on Bowtie (Trapnell *et al*., 2009).

In the next step, unique read alignments in SAM/BAM format are converted in the pileup format using SAMtools in order to explore all mappings position by position. Finally, the resulting pileup file is parsed to extract known RNA editing positions annotated in the specialized DARNED database (Kiran and Baranov, 2010). For each position, the nucleotide distribution of supporting reads is calculated as well as the percentage of editing. A specific filter that excludes bases with a phred-like quality score lower than a prefixed cut-off (by default equal to 20) has been also introduced. In addition, read nucleotides supporting a specific genomic position can be filtered according to the strand of the reference to take into account the effect of antisense transcription (recommended for strand-oriented reads).

The most relevant step of the ExpEdit workflow is the accurate read-to-genome mapping, because erroneous alignments can seriously affect editing detection and quantification. Alignment biases can be mitigated using reads mapping in unique genome locations and increasing the number of mismatches. To improve the mapping reliability in proximity of repeated regions, reads longer than 50 bases or paired-ends should be used. In any case, users are completely free to refine offline their alignments and then upload the relevant SAM or BAM files.

At the end of the analysis, ExpEdit returns a summary table with relevant information for each editing site including genomic location, base coverage, editing extent, gene name and region (CDS, intron or UTR) as well as an external link to the University of California, Santa Cruz (UCSC) genome browser showing the relevant genomic context. Potential single nucleotide polymorphisms (SNPs) are also indicated in order to be filtered out if required.

Output tables can be dynamically sorted, filtered and exported as textual/excel files for offline downstream analyses. ExpEdit can also explore RNA editing candidates not yet annotated in the DARNED database uploading as input a text file with a list of putative editing positions. The analysis of multiple FASTQ (from different lanes of the same run) or SRA or SAM or BAM files or a combination of them is also permitted in a single job.

Although there are no real technical limitations (except for the 2 GB threshold for direct upload of input files), potential errors related to the methodology (the cross mapping is the most relevant) should be taken into account. Moreover, ExpEdit is not a prediction tool and, thus, new editing candidates detected providing user-submitted candidate positions should be experimentally confirmed to exclude mapping errors or nucleotide changes due to unknown SNPs.

The current release of ExpEdit has been extensively tested on Illumina FASTQ files from the SRA archive at the NCBI. SOLiD short reads cannot presently analyzed as such but standard SAM/BAM files generated after the appropriate color space mapping can be uploaded as well. Similarly, SAM/BAM files containing alignments of longer reads from 454 or Sanger sequencing can be also provided.

To facilitate and speed-up the upload process, ExpEdit can automatically retrieve input files from valid ftp or http links by using the *wget* application. In this way there are no limitations on the size of input files. Examples of the ExpEdit application as well as details about input and output files are described in Supplementary Materials.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Angela Gallo for critical reading of the manuscript and fruitful suggestions.

Funding: Ministero dell'Istruzione, Università e Ricerca (MIUR, Italy); 'Laboratorio di Bioinformatica per la Biodiversità Molecolare' (Progetto FAR, DM19410)' Progetto Strategico Regione Puglia PS_012; AriSLA - Fondazione Italiana di Ricerca per la SLA.

Conflict of Interest: none declared.

REFERENCES

Barak,M. *et al.* (2009) Evidence for large diversity in the human transcriptome created by Alu RNA editing. *Nucleic Acids Res.*, **37**, 6905–6915.

Benne,R. *et al.* (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, **46**, 819–826.

Gott,J.M. and Emeson,R.B. (2000) Functions and mechanisms of RNA editing. *Annu. Rev. Genet.*, **34**, 499–531.

Keegan,L.P. *et al*. (2001) The many roles of an RNA editor. *Nat. Rev. Genet.*, **2**, 869–878.

Kiran,A. and Baranov,P.V. (2010) DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics*, **26**, 1772–1776.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Levanon,E.Y. *et al*. (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.*, **22**, 1001–1005.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,J.B. *et al.* (2009) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science*, **324**, 1210–1213.

Maas,S. *et al.* (2006) A-to-I RNA editing and human disease. *RNA Biol.*, **3**, 1–9.

Picardi,E. *et al.* (2010) Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. *Nucleic Acids Res.*, **38**, 4755–4767.

Trapnell,C. *et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.