BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 14 2013, pages 1813–1814
doi:10.1093/bioinformatics/btt287

Sequence analysis

Advance Access publication June 5, 2013

REDItools: high-throughput RNA editing detection made easy

Ernesto Picardi[1,2,*] and Graziano Pesole[1,2,3,*]

[1]Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari, [2]Institute of Biomembranes and Bioenergetics, National Research Council and [3]Center of Excellence in Comparative Genomics, University of Bari, 70125 Bari, Italy

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: The reliable detection of RNA editing sites from massive sequencing data remains challenging and, although several methodologies have been proposed, no computational tools have been released to date. Here, we introduce REDItools a suite of python scripts to perform high-throughput investigation of RNA editing using next-generation sequencing data.

Availability and implementation: REDItools are in python programming language and freely available at http://code.google.com/p/reditools/.

Contact: ernesto.picardi@uniba.it or graziano.pesole@uniba.it

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

Received on February 19, 2013; revised on May 6, 2013; accepted on May 16, 2013

1 INTRODUCTION

RNA editing is an important post-transcriptional mechanism that alters primary RNAs through the insertion/deletion or modification of specific nucleotides (Gott and Emeson, 2000). It occurs in a wide range of organisms and, concurrently with alternative splicing, increases the diversity of eukaryotic transcriptomes and proteomes (Mallela and Nishikura, 2012).

In humans, RNA editing affects both coding and non-coding transcripts by the deamination of cytosine to uridine (C-to-U) or, more commonly, through the conversion of adenosine to inosine (A-to-I) by members of adenosine deaminase (ADAR) family of enzymes, which act on double-stranded RNA (dsRNA) (Gott and Emeson, 2000). Editing events can lead to a variety of biological effects depending on the RNA type (mRNA or ncRNA) or region (5′-UTR, 3′-UTR, CDS, intron) involved in the RNA editing modification (Keegan *et al.*, 2001). For example, changes in UTRs can lead to altered expression, preventing efficient ribosome binding or recognition by small regulatory RNAs whereas alterations in coding protein regions can induce amino acid replacements with variable functional consequences. In addition, RNA editing can influence the activity of ncRNAs such as siRNAs, miRNAs and potentially of piwiRNAs by affecting base-pairing interactions within RNA secondary structures (Wulff and Nishikura, 2012). The importance of RNA editing is underlined by the fact that its deregulation in humans is associated with a variety of neurological or neurodegenerative disorders and cancer (Gallo and Locatelli, 2011; Silberberg *et al.*, 2012).

Although the identification of sites subject to RNA editing is simple, in principle, requiring the direct comparison between transcripts and their corresponding genomic loci of origin, accurate genome-wide implementation remains a challenging task (Bass *et al.*, 2012). Few years ago, we published the first attempt to use massive transcriptome sequencing (RNA-Seq) for exploring the RNA editing landscape in mitochondria of grapevine (Picardi *et al.*, 2010). Since then, RNA-Seq has become the *de facto* standard approach to study RNA editing potential in whole eukaryotic genomes, with sometimes questionable results (Hayden, 2011; Ramaswami *et al.*, 2012). The major challenge in identifying RNA editing changes by RNA-Seq data is the discrimination of true RNA editing sites from genome-encoded SNPs and technical artefacts caused by sequencing or read-mapping errors (Eisenberg, 2012; Ramaswami *et al.*, 2012). The use of DNA-Seq data from single individuals, annotations in dbSNPs and several stringent filters can minimize the detection of false RNA editing sites (Eisenberg, 2012; Ramaswami, *et al.*, 2012).

Although several methodologies to reveal the RNA editing potential in eukaryotic transcriptomes have been proposed, no comprehensive software devoted to this aim has been released. In the meantime, massive transcriptome sequencing data are routinely produced worldwide and a large number of experiments from different organisms and conditions are freely available through specialized databases such as the Short Read Archive at NCBI.

To promote the investigation of RNA editing at large scale in the next-generation sequencing era, we developed the package REDItools—comprising of a suite of scripts written in the portable python language. REDItools are available under the MIT license at http://code.google.com/p/reditools/, are not organism oriented and work with RNA-Seq and DNA-Seq data from any sequencing platform.

2 FEATURES

REDItools include three main scripts to study RNA editing using both RNA-Seq and DNA-Seq data from the same sample/individual or RNA-Seq data alone. *REDItoolDnaRNA.py* detects RNA editing candidates by comparing pre-aligned RNA-Seq and DNA-Seq reads in the standard BAM format. The script explores genomic positions site by site and returns a table containing the coverage depth, the mean quality score, the observed base distribution, the strand if available and the list of observed

*To whom correspondence should be addressed.

substitutions as well as the frequency of variation. If DNA-Seq data are also available, the same information, except for the strand, is provided to assess genomic sequence and, thus, exclude potential SNPs. In addition, individual positions can be filtered according to read coverage, base quality score, mapping quality, bases supporting the variation, type of substitution and frequency. *REDItoolDnaRna.py* can exclude positions in homopolymeric regions of predefined length, in intronic sequences surrounding known splice sites, invariant RNA-Seq positions, sites not supported by DNA-Seq and positions near read ends. Additionally, users can provide genomic regions to include or exclude in RNA editing searches. Additional read level filters include the possibility of excluding ambiguously mapped reads, possible PCR duplicates and discordant paired-end reads. In the case of directional RNA-Seq data, *REDItoolDnaRna.py* can infer the strand per position, improving the reliability of RNA editing calls by excluding noise from antisense transcription or mapping errors. Alternatively, the strand of individual positions can be inferred from user specified annotations.

REDItoolKnown.py, has been developed to explore the RNA editing potential of RNA-Seq experiments by looking at known events only, following an improved methodology implemented in our ExpEdit web service (Picardi *et al.*, 2011) (see Supplementary Material for details).

Finally, *REDItoolDenovo.py* performs the *de novo* detection of RNA editing candidates using RNA-Seq data alone without resequencing data for the individual donor genome (Picardi *et al.*, 2012) (see Supplementary Material for details).

REDItools includes some accessory scripts to post-process output tables, filtering candidates position-by-position using known annotations and identifying ambiguous alignments by Blat. All candidate positions can also be easily and quickly annotated using relevant databases as those in the UCSC genome browser.

3 IMPLEMENTATION

REDItools are in the portable python programming language and based on the Pysam module, a wrapper for the widely used SAMtools (Li *et al.*, 2009), which includes methods and functions to handle read alignments in SAM/BAM format—facilitating the browsing of multiple read alignments site by site along a reference genome. All scripts have low memory request enabling RNA editing calling on standard desktop computers. Because the search for RNA editing candidates is time-consuming requiring the processing of individual genomic positions, REDItools can speed up the process by distributing calculations over independent cores using the native python multiprocessing module. Filtering and annotation of individual positions is extremely fast and memory efficient, as it is performed by the *tabix* program (Li *et al.*, 2009) wrapped in the *Pysam* module.

REDItools require inputs in BAM format from any platform/organism, and results are provided in tab-formatted tables facilitating downstream analyses.

We tested REDItools on the lymphoblastoid cell line GM12878 used in Ramaswami *et al.* (2012) for which BAM files produced by the BWA mapper (Li and Durbin, 2009) were kindly provided by the authors. Limiting the analysis to chr21 and using the same filtering scheme as the original work, we obtained remarkably similar results (see Supplementary Material for details). Remapping all reads by GSNAP (Wu and Nacu, 2011) to preserve the paired end information and reapplying REDItools, we obtained 221 401 and 2089 candidate editing sites in Alu regions and non-Alu non repetitive regions, respectively, with respect to 147 029 and 1451 sites detected by Ramaswami *et al.* (2012) (see Supplementary Material for details).

In conclusion, REDItools is a unique and effective resource for the investigation of RNA editing from NGS data. It is highly flexible, including a variety of filters and quality checks, and may provide very reliable sets of editing candidate sites according to the user requirements (see Supplementary Material for a further benchmark application on Illumina Body Map RNA-Seq data).

ACKNOWLEDGEMENTS

Authors thank G. Ramaswami and the Li lab for providing BAM files from GM12878 lymphoblastoid cell line and David Horner for critical reading of the manuscript.

Funding: This work was supported by the Italian Ministero dell'Istruzione, Università e Ricerca (MIUR): PRIN 2009 and 2010; Consiglio Nazionale delle Ricerche: Flagship Project Epigen, Medicina Personalizzata and Aging Program 2012-2014 and by the Italian Ministry for Foreign Affairs (Italy-Israel actions).

Conflict of Interest: none declared.

REFERENCES

Bass,B. *et al.* (2012) The difficult calls in RNA editing. *Nat. Biotechnol.*, **30**, 1207–1209.

Eisenberg,E. (2012) Bioinformatic approaches for identification of A-to-I editing sites. *Curr. Top. Microbiol. Immunol.*, **353**, 145–162.

Gallo,A. and Locatelli,F. (2011) ADARs: allies or enemies? The importance of A-to-I RNA editing in human disease: from cancer to HIV-1. *Biol. Rev. Camb. Philos. Soc.*, **87**, 95–110.

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

Hayden,E.C. (2011) Evidence of altered RNA stirs debate. *Nature*, **473**, 432.

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

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Mallela,A. and Nishikura,K. (2012) A-to-I editing of protein coding and noncoding RNAs. *Crit. Rev. Biochem. Mol. Biol.*, **47**, 493–501.

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.

Picardi,E. *et al.* (2011) ExpEdit: a webserver to explore human RNA editing in RNA-Seq experiments. *Bioinformatics*, **27**, 1311–1312.

Picardi,E. *et al.* (2012) A novel computational strategy to identify A-to-I RNA editing sites by RNA-Seq data: de novo detection in human spinal cord tissue. *PLoS One*, **7**, e44184.

Ramaswami,G. *et al.* (2012) Accurate identification of human Alu and non-Alu RNA editing sites. *Nat. Methods*, **9**, 579–581.

Silberberg,G. *et al.* (2012) Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. *Hum. Mol. Genet.*, **21**, 311–321.

Wu,T.D. and Nacu,S. (2011) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

Wulff,B.E. and Nishikura,K. (2012) Modulation of microRNA expression and function by ADARs. *Curr. Top. Microbiol. Immunol.*, **353**, 91–109.