REDITOOLS: efficient RNA editing detection by RNA-SEQ data

Ernesto Picardi[1,2], **Graziano Pesole**[1,2]✉
[1]University of Bari, Bari, Italy
[2]Institute of Biomembrane and Bioenergetics of National Research Council, Bari, Italy

Motivation and Objectives

RNA editing and alternative splicing are post-transcriptional modifications that increase the complexity of eukaryotic transcriptomes and proteomes. Both phenomena have been efficiently investigated by massive sequencing according to recent high-throughput technologies. In particular, the RNA-seq methodology is the de facto technology to investigate entire eukaryotic RNA populations at single nucleotide level providing fruitful snapshots of cell/tissue activities in a variety of normal and non-homeostatic conditions (Wang, *et al.*, 2009).

RNA editing can modify specific RNAs at selected locations (Maas, 2011) and, in human, frequently involves the deamination of adenosines to inosines by the family of ADAR enzymes acting on double RNA strands (Hogg, *et al.*, 2011). Inosine is commonly interpreted as guanosine by splicing and translation machineries other than sequencing enzymes. A-to-I RNA editing has a plethora of biological effects, strictly related to the RNA region involved in the modification. Changes in 5' and 3'UTRs, for example, can lead to altered expression, preventing the efficient ribosome binding at 5'UTR or the recognition by small regulatory RNAs at 3'UTR. In contrast, alterations in coding protein regions can induce amino acid replacements with more or less severe functional consequences (Hood and Emeson, 2011).

RNA editing events can be detected at large scale by adopting the RNA-seq technology and, thus, employing multiple read alignments onto the corresponding reference genome to look at A-to-G changes (Eisenberg, *et al.*, 2010). Recently in human, thousands of candidates have been identified and validated by direct comparison with whole genome sequencing data in order to purge single nucleotide variations (SNPs) (Ramaswami, *et al.*, 2012).

Although a variety of methodologies have been developed to explore the RNA editing impact on eukaryotic transcriptomes, no comprehensive software for this aim has been released to date. The main challenge is to implement effective filters to mitigate the detection of false positives due to sequencing errors, mapping errors and SNPs. Very recently, we released the web service ExpEdit to explore known RNA editing events in RNA-seq experiments and the first computational strategy to predict de novo RNA editing events without any a priori knowledge of the genomic information or the nature of RNA editing process (Picardi, *et al.*, 2012). Here we present REDItools, a suite of python scripts aimed to the study of RNA editing at genomic scale.

The package is freely available at Google Code repository (http://code.google.com/p/reditools/)and released under the MIT license.

Methods

REDItools consist of three main python scripts and several accessory scripts. Main scripts are based on Pysam module (http://code.google.com/p/pysam/) a wrapper of SAMtools (Li, *et al.*, 2009) for easy manipulation of big alignment files. Pysam includes methods and functions to handle read alignments in SAM/BAM format facilitating the browsing of multiple read alignments position by position along a reference genome. REDItools enable the analysis of RNA editing at three levels: 1) REDItoolDnaRna.py identifies RNA editing changes by comparing RNA-seq and DNA-seq reads from the same individual; 2) REDItoolKnown.py explores the RNA editing potential of entire RNA-seq experiments using known sites stored in public databases as DARNED or provided by users; 3) REDItoolDenovo.py implements our methodology to detect RNA editing events using RNA-seq data alone.

Several accessory scripts are also provided in order to facilitate the browsing of results and assisting users through the annotation of predicted positions by using widespread databases from UCSC genome browser. Additional annotation and filtering steps are performed using the tabix program, for which the wrapper is included in the Pysam module.

All results are provided in tabulated tables for easy parsing and filtering.

Results and Discussion

REDItools work on machines running unix/linux operating systems and accept BAM files from whatever sequencing technology or organism.

REDItools have been extensively tested on public human RNA-seq experiments. In particular, we used REDItools to explore the impact of RNA editing on RNA-seq reads from the Illumina Human Body Map 2.0 Project using known events annotated in DARNED database (Kiran, *et al.*, 2013). Fastq files for RNA-seq experiments were downloaded from ArrayExpress database and mapped onto the hg18 human genome by GSNAP (Wu and Nacu, 2011) including known splice sites from UCSC, RefSeq, Ensembl and AspicDB (Martelli, *et al.*, 2011). SAM files were converted to BAM by SAMtools, duplicated reads were marked by Picard MarkDuplicates.jar (http://sourceforge.net/projects/picard/) and quality scores were recalibrated by GATK (McKenna, *et al.*, 2010). We added also further positions not yet present in DARNED and available from Ramaswami *et al.* (2012). The complete set of known RNA editing events comprises 564,135 positions.

Filtering output tables in order to focus only on positions supported by at least 10 reads and RNA editing frequency ^{3} 0.1, yielded the following table ordered by the descending number of RNA editing sites per tissue (Table 1). It is quite interesting to note that the number of RNA editing sites is tissue dependent as expected and the brain is not the human tissue with the predominant number of events. Moreover, this naïve experiment directly demonstrates the power and effectiveness of our REDItools in investigating at genomic level the intriguing phenomenon of RNA editing.

Table 1

Accession	OrganismPart	Editing Sites
ERR030874	ovary	6599
ERR030881	adrenal	5603
ERR030872	thyroid	4573
ERR030873	testes	4122
ERR030882	brain	3992
ERR030880	adipose	3877
ERR030883	breast	3739
ERR030877	prostate	3238
ERR030885	kidney	3146
ERR030886	heart	2393
ERR030887	liver	2376
ERR030884	colon	2320
ERR030875	white blood cells	2288
ERR030878	lymph node	1978
ERR030879	lung	1591
ERR030876	skeletal muscle	536

Acknowledgements

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, Aging Program 2012-2014 and by the Italian Ministry for Foreign Affairs (Italy-Israel actions).

References

Eisenberg, E., Li, J.B. and Levanon, E.Y. (2010) Sequence based identification of RNA editing sites, *RNA biology*, **7**, 248-252.

Hogg, M., *et al.* (2011) RNA editing by mammalian ADARs, *Advances in genetics*, **73**, 87-120.

Hood, J.L. and Emeson, R.B. (2011) Editing of Neurotransmitter Receptor and Ion Channel RNAs in the Nervous System, *Current topics in microbiology and immunology*. **353**, 61-90

Kiran, A.M., *et al.* (2013) Darned in 2013: inclusion of model organisms and linking with Wikipedia, *Nucleic acids research*, **41**, D258-261.

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

Maas, S. (2011) Gene regulation through RNA editing, *Discovery medicine*, **10**, 379-386.

Martelli, P.L., *et al.* (2011) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing, *Nucleic acids research*, **39**, D80-85.

McKenna, A., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome research*, **20**, 1297-1303.

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, *Nature methods*, **9**, 579-581.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature reviews*, **10**, 57-63.

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