A bioinformatics workflow for the analysis of transcriptome data generated by deep-sequencing

Licciulli F[1], Caratozzolo FM[2], Cornacchia S[2], D’Elia D[1], D’Erchia AM[1], Fosso B[2], Grillo G[1], Liuni S[1], Mangiulli M[2], Manzari C[2], Mignone F[3], Paluscio AM[1], Picardi E[2], Sbisà[1], Tullo A[1], Pesole G[2,4]

Motivation

The huge amount of transcript data produced by high-throughput sequencing requires the development and implementation of suitable bioinformatic workflows for their analysis and interpretation. These analysis workflows, including different modules, should be specifically designed also based on the sequencing platform (Roche 454, Illumina, SOLiD) and the nature of the data (polyA or total RNA fraction, strand specificity). In the case of cDNA obtained from a total RNA preparation, in addition to polyadenylated protein coding mRNAs, a great variety of transcript sequences can be obtained, including ribosomal RNAs, mitochondrial transcripts and a large variety of functional non coding RNAs (ncRNAs). To deal with these data the analysis workflow should include specific modules to distinguish ncRNAs fractions from the large number of other functional proteincoding transcripts. To this aim we developed an analysis pipeline that, given as input a large collection of reads (particularly from Roche 454), provides the expression profile at qualitative and quantitative level of human mtDNA, ribosomal RNAs, ncRNAs and protein coding mRNAs.

Methods

The identification of reads representing ncRNAs has been carried out through a megaBLAST search against fRNAdb rel. 3.4 [1] a comprehensive database of ncRNAs including data from different sources (e.g. RNAdb v2.0 [2] mirBASE 10.0 [3] NONCODE v1.0 [4] and others), ribosomal RNAs and human mtDNA. All sequences are organized in different categories based on their description using Sequence Ontology (SO) terms [5] . We only considered entries corresponding to experimentally known ncRNAs, excluding predicted ones. The megaBLAST output is then suitably parsed to detect fully and partially mapped reads, as well as possible chimeric reads. A read is considered fully mapped if the unmapped residual sequence is shorted than 50 nt. The bioinformatics workflow dynamically generates a table of expressed ncRNAs, indicating the number of represented

[1] Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, sede di Bari, Bari, Italy [2] University of Bari, Dipartimento di Biochimica e Biologia Molecolare “E. Quagliariello”, Bari, Italy [3] University of Milan, Chimica strutturale e Stereochimica Inorganica, Milan, Italy [4] Istituto Biomembrane e Bioenergetica del Consiglio Nazionale delle Ricerche, Bari 70126, Italy

reads, and give the possibility to extract read collections belonging to a given class or category, based on the SO classification. The collection of unmapped residual reads is then addressed to the bioinformatic modules devoted to the analysis of the expression profile of protein coding mRNAs, at both qualitative and quantitative level. To this aim the 454 Transcriptome Profile Explorer (454Trex, Mignone et al. in preparation) platform has been used that, after performing a GMAP mapping of reads on the human genome, processes mapping results, crossing mapping data with genome annotations, to provide the expression profile of human genes in the samples under investigation. Specific statistical modules have been also implemented to detect differentially expressed at gene and transcript (i.e. splicing isoform) level.

Results

We recently discovered that TRIM8 is a new modulator of the p53 oncosuppressor gene stability and activity. Our preliminary data suggest that the over-expression of TRIM8 promotes the activation of p53 target genes involved in growth arrest and DNA repair. In order to identify more broadly these group of genes, we analysed, by high-troughput sequencing, the transcriptional profile of HCT116 (p53w.t.) colon carcinoma cell line transiently transfected with TRIM8 (TRIM) or the control (FLAG). 48h after transfection, total RNA were extracted, depleted of about 85-90% of ribosomal RNA, retro-transcribed, and amplified. The resulting double strand cDNA libraries were sequenced in the standard 454 System workflow. The pyrosequencing generated 400,879 and 468,381 reads for the TRIM and FLAG samples, respectively with an average length of about 300 bp. These two datasets were then addressed to the bioinformatic pipeline described in the Methods section to detect and annotate reads corresponding to mtDNA, rRNAs and other ncRNAs. Over 60% of the reads were found to map to mtDNA, rRNAs, and other ncRNAs in both samples, whereas about 25% of the reads corresponded to protein coding mRNAs. The observation of a large number of chimeric reads was not surprising in consideration of the specific amplification procedure adopted for the cDNA library preparation. Specific results including differentially expressed ncRNAs and protein coding mRNAs in the TRIM and FLAG sample will be discussed in detail. The bioinformatics workflow developed for this specific case is of general applicability for RNA-Seq experiments of both total and poly-A+ transcriptome analyses.

Contact e-mail

graziano.pesole@biologia.uniba.it

Supplementary information

References

[1] T. Mituyama et al. (2009) “The functional RNA Databases 3.0: databases to support mining and annotation of functional RNAs”. Nucleic Acids Researches, Vol.37, 89-92 [2] Pang et al. (2007) RNAdb 2.0 – an expanded databases of

mammalian non-coding RNAs. Nucleic Acids Res., 35, D178-D182. [3] Griffiths-Jones et al. (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res., 36, D154-D158. [4] He et al., (2008) NONCODE V2.0: deconding the non-coding, Nucleic Acids Res., 36, D170-D172 [5] Eilbeck et al. The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology (2005) 6:R44
Acknowledgements This work was supported by the Ministero dell'Istruzione, dell'Università e della Ricerca: Fondo Italiano Ricerca di Base: ''Laboratorio Internazionale di Bioinformatica" (LIBI); Laboratorio di Bioinformatica per la Biodiversità Molecolare (MBLAB), Progetto Strategico Regionale.