BITS :: Call for Abstracts 2023 - Oral communication

Type	Oral communication
Session	Genomics, transcriptomics, epigenomics and epitranscriptomics
Title	Innovative pipeline for lncRNA detection from small RNA-seq data
All Authors	Consiglio A(1), Alfieri R(2), Licciulli F(1), Mezzelani A(2), Cupaioli FA(2), Chiappori F(2)

Affiliation

(1) National Research Council – Institute for Biomedical Technologies, CNR – ITB, Bari
(2) National Research Council – Institute for Biomedical Technologies, CNR – ITB, Segrate (Mi)

Motivation

The interest of the scientific community on non-coding RNAs (ncRNAs) is increasing due to their role in gene regulation both in physiological and pathological conditions. Tissue and condition specific expression of ncRNAs highpoints their potential role as biomarkers.
RNA sequencing (RNA-seq) is used to investigate transcriptome, including long non-coding RNAs (lncRNAs). The standard workflow of RNAseq analysis requires starting high-quality RNA to be unbiased. However, in several types of samples, such as formalin fixed paraffine embedded tissues and stools, RNA is degraded and then suitable only for small RNAseq (small RNA-seq). Thus, small RNA-seq alone loses the interesting information layer about lncRNAs. Moreover, several sequences obtained from small RNA-seq were discarded because they cannot be aligned with ncRNA reference databases. For all these reasons, some studies focus only on small RNAs, since standard RNA-seq from the same sample is not obtainable.
Reads from small RNA-seq have two main drawbacks to be overcome: reads are too short (less than 50bp) to be uniquely mapped, and this results also in too many multimapping regions along human transcriptome.
Here we propose an innovative pipeline, aiming at identifying lncRNAs starting from small RNAseq performed on degraded RNA of stool samples, considering possible multimapping reads and gene coverage.

Methods

SmallRNA reads were checked for quality control using FastQC package, filtered and then mapped on Ensembl genome, using bowtie and RSEM. This returns two bam files, one for transcripts, used for read-count quantification, and one for genome, for genome-browser visualization of mapped reads. RSEM outputs were processed with MultiDEA [1] to evaluate both uniquely and ambiguously mapped reads associating a counts and error estimation for each gene and sample. Moreover, gene coverage breadth was evaluated and considered to estimate the reliability of lncRNA expression evaluated from short RNA reads. EdgeR was used for differential expression analysis, with abs(log2fc)>1 and FDR<0.05.

Results

We tested the pipeline on a small RNA-seq dataset generated from stool samples from 6 children affected by autism spectrum disorders and 6 neurotypical controls [2].
The identification of a specific lncRNA gene was addressed by combining uniquely- and multi-mapped reads on the same lncRNA coordinates, at genomic and transcriptomic levels.
The proposed pipeline enables the identification of differentially expressed lncRNAs with a uniform coverage on transcript region and allowing also lncRNA isoforms detection.

Furthermore, we are validating the proposed pipeline on available RNA-seq datasets whose lncRNA has already been validated by RT-PCR. Reads from the selected datasets will be randomly cut from long to short reads to simulate a dataset which is comparable to a degraded RNA-seq one, as the test set.

Info

[1] Consiglio A. et al. BMC Bioinform. 2016, 17, 345. https://doi.org/10.1186/s12859-016-1195-2
[2] Chiappori F. et al. Nutrients 2022, 14,1340. https://doi.org/10.3390/nu14071340 IF: 6.706

filename	-

Figure

-

Availability	-

Corresponding Author

Name, Surname	Federica, Chiappori
Email	federica.chiappori@itb.cnr.it

Submitted on	20.04.2023