Bioinformatics analysis reveals novel long non-coding RNA candidates in Ewing sarcoma

Caroline Brunetto de Farias, André Tesainer Brunetto, Marialva Sinigaglia, Ney Lemke, Rafael Luiz Buogo Coan

UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO, ICI - Instituto do Câncer Infantil

Abstract

Long non-coding RNAs (lncRNAs) are defined as RNA molecules with more than 200 nucleotides, which don't encode proteins. They are present exclusively on the nucleus, cytoplasm or in both. They can interact with cellular components to form RNA-DNA, RNA-RNA and RNA-protein complexes, modulating gene wide expression. LncRNAs are also players in several diseases, including Ewing sarcoma (ES), which is a childhood malignant neoplasm that affects bones and soft tissues. A vital molecular alteration in ES is the translocation between chromosomes 11 and 22, resulting in the fusion protein EWS-FLI1, which acts as a transcription factor, altering genome-wide gene expression. In this study, our goal was to establish a bioinformatics pipeline to determine new lncRNA in ES patient samples. We used Illumina RNA-Seq data from dbGaP (phs000768v2p1) to identify novel lncRNA candidates in 26 ES patient samples consisting of EWS-FLI1 types I, II and III fusions. Raw reads were trimmed and quality filtered with Trimmomatic 0.36, then aligned to the human genome (hg38) with HISAT2 2.1.0. We then performed a new guided assembly on each sample with Stringtie 1.3.4, which was followed by merging the 26 new transcriptomes into a single file. We used the filter module from FEELnc 0.1.1 to exclude transcripts overlapping sense protein-coding genes from Gencode v.30 (150, 140 transcripts) and lncRNA from RNAcentral v.16 (554, 174 transcripts). After filtering, 527 candidate lncRNA remained, which were subjected to two coding potential estimators to computationally evaluate their protein-coding ability. We used FEELnc coding potential module and PLEK 1.2 for this task, only keeping the consensus transcripts found on both programs. A total of 459 transcripts lasted. We quantified the expression of the candidate lncRNA with Salmon v.1.1.0 and made between sample normalization with DESeq2. There are several novel transcripts with various levels of transcription, which may indicate a level of activity in ES. Our next steps include further genomic characterization of candidate lncRNA, plus in vitro and in vivo validation of potential transcripts involved in ES pathology. Although in early steps, our results show the potential of bioinformatics analysis to identify new candidate lncRNA that may be involved in ES biology.

Funding: CNPq 142129/2018-6

Link to Video: