BITS :: Call for Abstracts 2023 - Oral communication

Type	Oral communication
Session	Genomics, transcriptomics, epigenomics and epitranscriptomics
Title	Resolving complex transcription models with a new complementary sequencing and bioinformatics pipeline
All Authors	S. Cancellieri (1), D. Pollini (1), Roberto Bertorelli (1), Veronica De Sanctis (1), Chiara Valentini (1), Paolo Cavallerio (1), G. Piccoli (1), E. Domenici (1)

Affiliation

(1) CIBIO, University of Trento, Trento

Motivation

Identification and classification of alternative transcripts is a very demanding and complex task even with modern RNA sequencing techniques and computational approaches.
Many software developed to discover novel transcripts uses a probabilistic approach to identify transcripts evaluating the presence of splicing sites between different RNA sequences aligned on the same genomic region.
New wet lab techniques and sequencing methodologies have been developed to identify transcripts and their sequence using long read sequencing proposed by Pacbio (IsoSeq) and Oxford Nanopore Technologies (Long Nanopore), but these methods are costly, complex and difficult to operate in many facilities when interested in only one gene.
To overcome these problems we have developed a bioinformatic analysis pipeline to have an insight on the information that can be obtained through tailor-made short RNA sequencing approaches on alternative splicing and transcript analysis.

Methods

Our approach takes into consideration three steps: i) collect annotated transcripts for a specific gene, produce an annotated gene model and perform a barcode linked-reads RNA sequencing (TELL-Seq) to identify annotated transcripts and inconsistencies in the model; ii) perform alignments and identification of inconsistencies such novel introns or exon skipping events; iii) a final short RNA sequencing to identify and annotate the inconsistencies to generate a final corrected model.
Firstly, we generate a merged transcript model for a gene by collecting data from the most popular databases for transcriptomics, RefSeq and Gencode. After data collection and merge, we produce a graphical representation of the model with previously annotated transcripts.
After this pre-analysis step, we extract all the known sequences from the reference transcriptome and start designing the possible primers for correctly identifying each specific transcript.
Sequencing is performed using a technique called TELL-Seq™, designed to generate de-novo genome assemblies that in this pipeline is adapted to perform reconstruction of transcripts and identification of splicing events using transposons tagged cDNA sequences.
The second step of the pipeline consists in alignment of the obtained sequences to validate annotated transcripts and possible new transcripts or alternative splicing events. This step is performed with two different software, the first one is an aligner software designed specifically to work with TELL-Seq reads which generate a graph of overlapping sequences and a long read using the ends of identical tagged sequences. The second software used is RSEM, used to correctly align the short sequences on the genome and compare them to the long read sequences generated by tagged reads. This process helped to identify unknown events and undiscovered elements such as exon skipping or new introns.
The third and last step of the pipeline consists in a second, more focussed, RNA sequencing based on the novel information extracted from the second step.
This adapted sequencing technique (Isoform Specific Amplicon Short Sequencing, ISASS for short) targets the exons we identified as transcript specific and tested all the novel elements we discovered in the second step, giving a sound and clear picture of the total gene transcripts model.

Results

We tested our designed pipeline on a gene under investigation as a putative pharmacological target. The gene taken into account has an unusual transcription panel, with different shorts non-coding transcripts and two protein coding transcripts with a very uneven exons length and a nearly identical sequence between transcripts.
We applied the above described pipeline by adapting it to the data.
After the pre-analysis and first alignment of TELL-Seq reads, we discovered that the canonical transcript contains an intron in the longer exon, justifying the abnormal behavior observed in the publicly available RNA sequences.
In addition, our analysis did not reveal any evidence supporting the existence of the second protein coding transcript annotated in the gene model produced by merging the transcripts in the public databases.
After these discoveries, we moved to the third and final step of the pipeline, performing a targeted sequencing, trying to find characterizing exons (exons specific to a single transcript) followed by a focused alignment.
The result of the third step was a new gene model, which takes into account the newly identified intron in the canonical transcript and the absence of the protein coding transcript described in the annotation databases.
To summarize, our method is designed to work with complex transcription panels, for genes

with many splicing events, transcripts with unusual conformation and many shared exons between transcripts. The method generates a corrected gene model accounting for transcription events and unseen elements, focusing on target sequencing and specific alignment, necessary steps to obtain reliable gene models.

Info	
-	
filename	-
Figure	
-	
Availability	-

Corresponding Author	
Name, Surname	Samuele, Cancellieri
Email	samuele.cancellieri@unitn.it
Submitted on	02.05.2023