POSTER - RNA AND TRANSCRIPTOMICS

COMPUTATIONAL APPROACHES FOR LNCRNA DISCOVERY USING LONG-READ RNA-SEQ DATA

Otávio Pereira Carreiro De Sousa (opcsousa@gmail.com)

Caio César Maia Medeiros (caio.maia@academico.ufpb.br)

Maria Bárbara Borges De Santana (mbarborgess@gmail.com)

Profa. Dra. Thaís Gaudencio Do Rêgo (gaudenciothais@gmail.com)

Vinicius Maracaja Coutinho (vinicius.maracaja@uchile.cl)

The advent of third-generation sequencing technologies has significantly advanced identification and investigation of long non-coding RNAs (lncRNAs), enabling broader and more accurate sequencing. This has deepened the understanding of lncRNA functions and interactions with other molecules. The ability to generate long, high-precision reads has been instrumental in uncovering the structural and functional complexity of lncRNAs in the context of both research and clinical applications. A wide range of computational tools can be employed to process the massive volume of data generated by long-read sequencing technologies, particularly focused on RNA-seq assays. Therefore, we present a literature review highlighting tools and databases related to lncRNA studies using long-read RNA-seq approaches. Key workflow steps include demultiplexing, quality control (QC), filtering, trimming, and mapping of RNA long-reads. Tools identified for demultiplexing include ONTbarcoder and Guppy. ONTbarcoder has some advantages such as the use features of real-time barcoding, giving a rapid overview of barcades obtained within minutes of

sequencing. Guppy is a traditional demultiplexing tool with great performance but should be used specific for Oxford Nanopore technologies. FastQC offers fast, versatile visualization of quality metrics across platforms. For filtering and trimming, Nanoq offers fast processing for nanopore reads and fits automated pipelines, while Filtlong has slower processing and higher memory usage but allows more complex read filtering. On the other hand, Fastplong is designed for a single and fast overall QC, filtering and trimming. Subsequently, MultiQC concatenates analyses of multiple tools into a single report. For mapping, Minimap2, Graphmap2, and deSALT were identified. Minimap2 is a general-purpose aligner, also supporting short-read alignment to long-read assemblies with splice-aware mapping. Graphmap2 targets long error-prone reads, while deSALT provides fast analysis of large data volumes but demands greater computational resources. Subsequent steps in lncRNA transcriptome assembly include identifying and characterizing lncRNA isoforms, normalization, and differential expression analysis. Key tools here are Bambu, IsoQuant, and StringTie2. Bambu focuses on reference-guided transcript reconstruction but has limitations in de novo detection. StringTie2 excels at de novo assembly, while IsoQuant is versatile for both annotation-free discovery and reference-guided analyses. Functional characterization of lncRNAs was also addressed, including enrichment and co-expression analysis, lncRNA-protein interactions, RNA-RNA interactions, and RNA-DNA interactions. ClusterProfiler and CEMITool were highlighted for enrichment and co-expression analysis. ClusterProfiler supports a wide range of gene set enrichment analyses, while CEMITool offers an automated pipeline for co-expression module identification and analysis. For lncRNA-protein interactions, LncADeep and LPIH2V were identified. ASSA, RIBlast, and LASTAL are prominent for RNA-RNA interaction studies, providing efficient search algorithms for interaction prediction. Triplexator and LongTarget were the main tools cited for RNA-DNA interactions, facilitating the discovery of triplex-forming sites between RNA and DNA molecules. Additionally, a survey of 21 frequently cited databases relevant to lncRNA research was conducted, covering multiple organisms and data types, such as LncBook, LNCipedia 5 and lncATLAS. The lack of standardized nomenclature within the academic community poses a challenge to the use of these databases. Finally, this review aims to assist researchers in selecting the most appropriate tools, considering different usage purposes and biological contexts.

Palavras-chave: bioinformatics; transcriptomics; lncrna; long-read; rna-seq.