CHOLAR: Characterization of LncRNA from raw reads

13 June 2022

Summary

RNA-sequencing has found numerous implementations in research, from distinguishing immune cell subtypes to differential gene expression between cancer versus normal tissue types [@Villani2017][@Bao2021]. Another application of RNA-seq is to identify novel transcripts involved in various biological processes (Gupta, Kleinjans and Caiment 2021). The most relevant is context and cell-type-specific non-coding RNAs, such as long non-coding RNAs (lncRNAs), which have become a case-point for most transcriptomic studies proving their role in regulating gene expression, post-transcriptional regulation, and epigenetic regulation (Engreitz et al. 2016, Zhu et al. 2019). It is becoming crucial to check the relative expression of lncRNAs in transcriptome-wide studies. Our group has developed an automated lncRNA expression pipeline. The only requirement from the user-side is raw data in FASTQ format (Paired-end or Single-end). The user will get a list of known and novel lncRNAs, and differential gene expression between condition(s). The pipelines come with a user-friendly GUI, thereby eliminating the need for the user to be versed in complex transcriptome analysis and UNIX environment. The source code is available under an open-source licence at https://github.com/schosio/CHOLAR.

Statement of need

The number of inferences generated from RNA-seq datasets is countless. The software used in the RNA-seq analysis pipeline requires a UNIX-based command-line interactive (CLI) environment, with each software executed in succession. Installing multiple CLI tools, handling various file formats, and plotting graphs require understanding of Linux and R programming language. Moreover, no tool identifies novel lncRNAs from raw transcriptomic data to the best of our knowledge. LncRNA identification tools such as `CPAT` (Wang et al. 2013) (and other tools) take either transcript sequence (FASTA) or transcript coordinate (BED, GTF) as input and provide a list of predicted lncRNAs. To address these

issues, we developed CHOLAR which is a tool for characterization of LncRNA from raw reads . `CHOLAR` i) identifies novel lncRNAs from raw reads ii) provides a user-friendly GUI interface to make changes at every step iii) allows to identify differentially expressed genes and lists known and novel lncRNAs iv) generates publication-quality plots such as MA plot, Volcano plot and heatmap.

Implementation

The `CHOLAR` pipeline is implemented in bash and R, where it first reads the input FASTQ files(s) to check the quality of reads using `FastQC` (Fiancette et al. 2021). Bad quality (< 28) reads and adaptors are removed using `Trimmomatic` (Bolger, Lohse and Usadel 2014). `HISAT2` performs the mapping of reads on the human reference genome (hg38) (Zhang et al. 2021). The SAM files generated from `HISAT2` are converted to BAM, and PCR duplicates are removed utilising the `samtools` toolkit (Danecek et al. 2021). The transcript assembly is done using Stringtie, and the resulting GTF files are merged using the merge utility of `Stringtie` (Pertea et al. 2015). The merged GTF file is compared against the reference annotation file from GENCODE (Frankish et al. 2021) to filter novel transcripts using `GFFCOMPARE`. The coding potential of novel transcripts is predicted using CPAT (Wang et al. 2013). The gene counts are calculated using `HTSeq` (Putri et al. 2022), and subsequent differential gene expression analysis (DGEA) is done using packages from R statistical language. The `DESeq2` package is used for performing DGEA (Love, Huber and Anders 2014), and `ggplot` and `dplyr` libraries for plotting graphs. The Graphical user interface is built using zenity in bash. The schematic of the tool is given in figure 1.

Example

We chose the GSE147761 dataset from the GEO database (NCBI) to showcase the CHOLAR tool. A sample of results and plots generated by the tool are given in figure 2. The GUI of the tool is shown in figure 3.

Citations

Figures

Fig1: Schematic of the tool CHOLAR. It starts from input files, performs alignment; assembly; LncRNA identification; differential expression analysis and provides output files and plots

Fig3: From top left clockwise: sample name input, gtf file dialog, threads slider, summary of all inputs, directory selection for script, script dialog, file selection for gtf

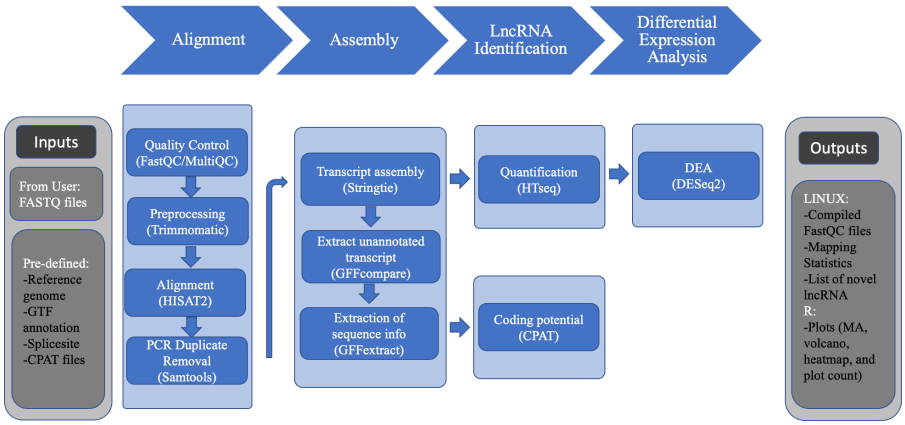

Figure 1: fig1.png

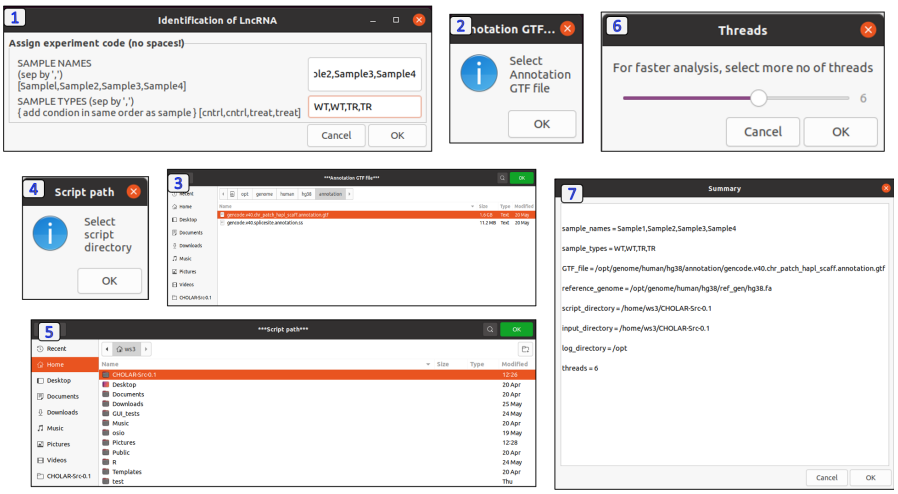

Figure 2: fig3.png

Acknowledgements

References