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 (Villani et al. 2017; Bao et al. 2021). 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. 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` (**???**). `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

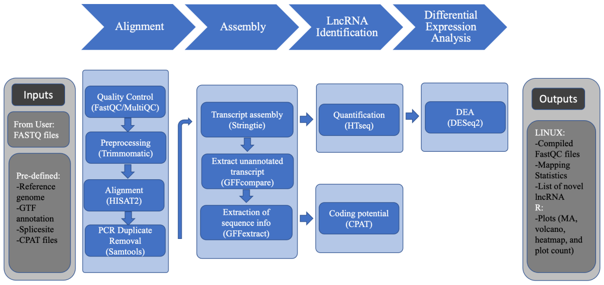
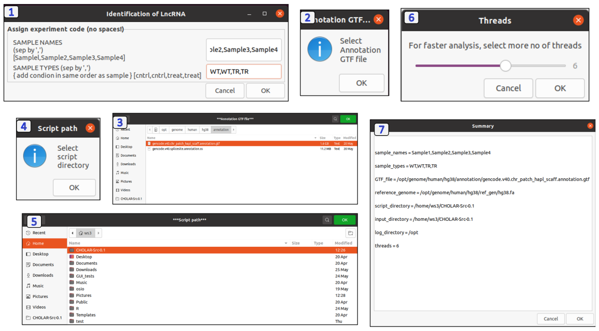

Alignment
Assembly
LncRNA Identification
Differential Expression Analysis
Inputs
From User: FASTQ files
Pre-defined: -Reference genome -GTF annotation -Splicesite -CPAT files
Quality Control (FastQC/MultiQC)
Preprocessing (Trimmomatic)
Alignment (HISAT2)
PCR Duplicate Removal (Samtools)
Transcript assembly (Stringtie)
Extract unannotated transcript (GFFcompare)
Extraction of sequence info (GFFextract)
Quantification (HTseq)
Coding potential (CPAT)
DEA (DESeq2)
Outputs
LINUX: -Compiled FastQC files -Mapping Statistics -List of novel lncRNA R: -Plots (MA, volcano, heatmap, and plot count)

Figure 1: Schematic of the tool CHOLAR. It starts from input files, performs alignment; assembly; LncRNA identification; differential expression analysis and provides output files and plots



1 Identification of LncRNA

Assign experiment code (no spaces!)

SAMPLE NAMES (sep by ',') [Samplel,Sample2,Sample3,Sample4] — ɔle2,Sample3,Sample4

SAMPLE TYPES (sep by ',') { add condion in same order as sample } [cntrl,cntrl,treat,treat] — WT,WT,TR,TR

2 ıotation GTF...

Select Annotation GTF file

6 Threads

For faster analysis, select more no of threads — 6

4 Script path

Select script directory

3 ***Annotation GTF file***

opt / genome / human / hg38 / annotation

Name	Size	Type	Modified
gencode.v40.chr_patch_hapl_scaff.annotation.gtf	1.6 GB	Text	20 May
gencode.v40.splicesite.annotation.ss	11.2 MB	Text	20 May

5 ***Script path***

ws3

Name	Size	Type	Modified
CHOLAR-Src-0.1			12:26
Desktop			20 Apr
Documents			20 Apr
Downloads			25 May
GUI_tests			24 May
Music			20 Apr
osio			19 May
Pictures			12:28
Public			20 Apr
R			24 May
Templates			20 Apr
test			Thu

7 Summary

```
sample_names = Sample1,Sample2,Sample3,Sample4

sample_types = WT,WT,TR,TR

GTF_file = /opt/genome/human/hg38/annotation/gencode.v40.chr_patch_hapl_scaff.annotation.gtf

reference_genome = /opt/genome/human/hg38/ref_gen/hg38.fa

script_directory = /home/ws3/CHOLAR-Src-0.1

input_directory = /home/ws3/CHOLAR-Src-0.1

log_directory = /opt

threads = 6
```

Figure 2: 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

Acknowledgements

References

Bao, X. W., R. Shi, T. Y. Zhao, Y. F. Wang, N. Anastasov, M. Rosemann, and W. J. Fang. 2021. "Integrated Analysis of Single-Cell Rna-Seq and Bulk Rna-Seq Unravels Tumour Heterogeneity Plus M2-Like Tumour-Associated Macrophage Infiltration and Aggressiveness in Tnbc." Journal Article. *Cancer Immunology Immunotherapy* 70 (1): 189–202. https://doi.org/10.1007/s00262-020-02669-7.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al. 2021. "Twelve Years of Samtools and Bcftools." Journal Article. *Gigascience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

Engreitz, J. M., J. E. Haines, E. M. Perez, G. Munson, J. Chen, M. Kane, P. E. McDonel, M. Guttman, and E. S. Lander. 2016. "Local Regulation of Gene Expression by lncRNA Promoters, Transcription and Splicing." Journal Article. *Nature* 539 (7629): 452–55. https://doi.org/10.1038/nature20149.

Fiancette, R., C. M. Finlay, C. Willis, S. L. Bevington, J. Soley, S. T. H. Ng, S. M. Baker, S. Andrews, M. R. Hepworth, and D. R. Withers. 2021.

"Reciprocal Transcription Factor Networks Govern Tissue-Resident Ilc3 Subset Function and Identity." Journal Article. *Nature Immunology* 22 (10): 1245–+. https://doi.org/10.1038/s41590-021-01024-x.

Frankish, A., M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, et al. 2021. "GENCODE 2021." Journal Article. *Nucleic Acids Research* 49 (D1): D916–D923. https://doi.org/10.1093/nar/gkaa1087.

Gupta, R., J. Kleinjans, and F. Caiment. 2021. "Identifying Novel Transcript Biomarkers for Hepatocellular Carcinoma (Hcc) Using Rna-Seq Datasets and Machine Learning." Journal Article. *Bmc Cancer* 21 (1). https://doi.org/10.1186/s12885-021-08704-9.

Love, M. I., W. Huber, and S. Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with Deseq2." Journal Article. *Genome Biology* 15 (12). https://doi.org/10.1186/s13059-014-0550-8.

Pertea, M., G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, and S. L. Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from Rna-Seq Reads." Journal Article. *Nature Biotechnology* 33 (3): 290–+. https://doi.org/10.1038/nbt.3122.

Putri, G. H., S. Anders, P. T. Pyl, J. E. Pimanda, and F. Zanini. 2022. "Analysing High-Throughput Sequencing Data in Python with Htseq 2.0." Journal Article. *Bioinformatics* 38 (10): 2943–5. https://doi.org/10.1093/bioinformatics/btac166.

Villani, A. C., R. Satija, G. Reynolds, S. Sarkizova, K. Shekhar, J. Fletcher, M. Griesbeck, et al. 2017. "Single-Cell Rna-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors." Journal Article. *Science* 356 (6335). https://doi.org/10.1126/science.aah4573.

Wang, L., H. J. Park, S. Dasari, S. Q. Wang, J. P. Kocher, and W. Li. 2013. "CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model." Journal Article. *Nucleic Acids Research* 41 (6). https://doi.org/10.1093/nar/gkt006.

Zhang, Y., C. Park, C. Bennett, M. Thornton, and D. Kim. 2021. "Rapid and Accurate Alignment of Nucleotide Conversion Sequencing Reads with Hisat-3N." Journal Article. *Genome Research* 31 (7). https://doi.org/10.1101/gr.275193.120.

Zhu, J., Y. T. Wang, W. Yu, K. S. Xia, Y. L. Huang, J. J. Wang, B. Liu, H. M. Tao, C. Z. Liang, and F. C. Li. 2019. "Long Noncoding Rna: Function and Mechanism on Differentiation of Mesenchymal Stem Cells and Embryonic Stem Cells." Journal Article. *Current Stem Cell Research & Therapy* 14 (3): 259–67. https://doi.org/10.2174/1574888x14666181127145809.