Non-coding RNA Research 8 (2023) 602–604

Contents lists available at ScienceDirect

Non-coding RNA Research

journal homepage: www.keaipublishing.com/en/journals/non-coding-rna-research

Original Research Article

Firalink: A bioinformatics pipeline for long non-coding RNA data analysis

Louis Chauviere [a], Lucien Hoffbeck [a], Muhammad Shoaib [b], Florent Tessier [a], Huseyin Firat [a], Venkata Satagopam [b], Yvan Devaux [c,*], on behalf of COVIRNA consortium

[a] From Firalis SA, Huningue, France
[b] Luxembourg Centre for Systems Biomedicine, Bioinformatics Core, University of Luxembourg, Belvaux, Luxembourg
[c] Cardiovascular Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, Luxembourg

A B S T R A C T

Summary: The Firalink bioinformatics pipeline has been developed to analyse long non-coding RNA (lncRNA) data generated by targeted sequencing. This pipeline has been first implemented for use with the FIMICS panel containing 2906 lncRNAs useful for investigations in cardiovascular disease. It has been subsequently tested and validated using a panel of lncRNAs targeting brain disease. The pipeline can be adapted to other targeted sequencing panels or other transcriptomics data (e.g. whole transcriptome) through a change of the reference genome/panel. Therefore, Firalink can be applied to different lncRNA panels and transcriptomics data targeting multiple diseases.

Availability and implementation: The Firalink pipeline works on Linux and is freely available to non-commercial users at https://gitlab.lcsb.uni.lu/covirna/covirna-ext/covirna-firalink-pipeline. Access will be granted after contacting bioinformatics@firalis.com. The pipeline is implemented with the Nextflow workflow manager using Python and R scripts. It will remain available for at least two years following publication and will be regularly updated and upgraded.

Supplementary information: For an example of the application of the Firalink pipeline using the FIMICS panel, see www.covirna.eu.

1. Introduction

Long non-coding RNAs (lncRNAs) represent a large group of RNA molecules longer than 200 nucleotides and lacking protein-coding potential. Instead, they regulate gene expression mostly at the epigenetics level and are involved in the development and progression of multiple types of diseases such as cardiovascular disease [1]. During the past few years, multiple tools and databases have been developed to identify lncRNAs and study their expression (Sun et al., 2017, Zhao et al., 2018, Iyer et al., 2015, Zhao et al., 2015, Bryzghalov et al., 2021). Hence, lncRNAs are attracting a lot of attention by the biomedical research community. To facilitate the research on the role and therapeutic potential of lncRNAs in cardiovascular disease, we conducted a deep RNA sequencing experiment in human failing and non-failing hearts through which we identified 2906 lncRNAs that were either enriched in the cardiac tissue or differentially expressed between failing and non-failing hearts. With these 2906 lncRNAs, we developed the FIMICS panel which can be used to identify novel disease markers or therapeutic targets [2]. The FIMICS panel is based on targeted sequencing to increase the sensitivity of the detection of lncRNAs, which are often weakly expressed, especially in blood samples. Therefore, the FIMICS panel allows detecting novel lncRNA biomarkers in blood samples that can be translated to the clinic and help in patient management [2]. Translation of research findings to clinical application in the RNA field is a challenge which can be overcome by attention to methodological issues [3] and collaborative work between researchers with complementary expertise [4]. To facilitate this translation and improve the quality of the analysis of sequencing data generated by the FIMICS panel, we developed the Firalink bioinformatics pipeline, which was applied to RNA sequencing (RNA-seq) data obtained from blood samples of COVID-19 patients.

1.1. RNA-seq data generation

We conducted targeted sequencing of 2906 lncRNAs using the FIMICS panel [2] in whole blood samples obtained from COVID-19 patients involved in the COVIRNA project aiming to discover lncRNAs predicting cardiovascular dysfunction in COVID-19 patients (www.covirna.eu). FIMICS libraries were prepared from total RNA extracted from whole blood samples and sequenced at 50bp reads paired ends (2x50bp) using the Illumina NextSeq2000 sequencer in an ISO17025 accredited lab. Flow cells of 400 M reads allowed the simultaneous quantification of 48 patients and resulted in an average of 8 million reads per sample. As described in details in Ref. [2], targeted sequencing uses both amplification and capture to enrich lncRNAs which are mostly weakly expressed. A first amplification occurs as the last step of library preparation (15 PCR cycles). Then the lncRNAs capture panel is used to

* Corresponding author. Cardiovascular Research Unit, Luxembourg Institute of Health, L1445, Luxembourg.
E-mail address: yvan.devaux@lih.lu (Y. Devaux).

https://doi.org/10.1016/j.ncrna.2023.09.002
Received 14 July 2023; Received in revised form 8 September 2023; Accepted 8 September 2023
Available online 12 September 2023

capture the lncRNAs of interest using biotinylated capture probes. Finally, a second amplification step is realized on captured sequences (14 PCR cycles). This allows targeted sequencing to be equivalent to 100X coverage since using a non-targeted approach, more than 550 million reads are needed to achieve equivalent coverage.

1.2. Pipeline description

Similar to other RNA-seq data analysis workflows, the Firalink pipeline is composed of five steps: 1) quality control of FASTQ files; 2) removal of low-quality reads; 3) evaluation of potential contamination with sequences from other species; 4) alignment of reads against the FIMICS panel (sequences of the 2906 lncRNAs) and generation of lncRNA counts using Kallisto method [5]; 5) compilation of count files into a matrix and quality control.

The following tools are used for each step.

1. *Quality control of FASTQ files.* FastQC software generates a report on descriptive metrics of the sequences contained in FASTQ [6]. Information about the number of reads and their duplication level, the average Phred quality score, GC composition, and the presence of adapters in sequences is provided.
2. *Removal of low-quality sequences.* Trimmomatic software removes low-quality sequences (reads with the Phred score below 30), adaptors or low-quality read extremities [7]. Sequences are compared to the classical Illumina adaptors to detect contaminations.
3. *Evaluation of potential contamination.* Kraken is a taxonomic classification tool used for the detection of inter-species contamination (bacterial, fungal or viral) in samples [8]. It proceeds by alignment of each sequence against the provided database and labels each sequence with a taxonomic level.
4. *Alignment of reads and generation of count values.* Kallisto [5] is a pseudo-alignment software used to obtain the count tables for each sample against a reference in FASTA format that contains the sequences of the FIMICS panel. The tool generates K-mer from the reference and compares sequences to determine the most likely position of the sequence on the given reference. The R package Tximport then concatenates the count tables of each sample into a single table.
5. *Compilation of quality control (QC) files.* MultiQC is a tool that compiles the QC information from the softwares used in the pipeline. It creates a full quality assessment (QA)/QC report which is finally used for the creation of the Firalink:FIMICS report.

An overall workflow of the Firalink pipeline is presented in Fig. 1A.

Construction of lncRNA matrix. We have written a python script that combines lncRNA counts into a single matrix file, which can be exported as comma separated file (CSV), tab separated file (TSV) or Excel file. The script reads all count files in a directory and combines them into a single file. Two files are generated as the result of this 1) count matrix and 2) QC matrix, which can be used for further analysis.

Quality assessment of lncRNA counts. In order to assess the quality of the lncRNAs from count matrix, we have written another python script that generates reports providing the list of zero variance lncRNAs, and lncRNAs with low read counts. This step generates a plot showing on the x-axis filter threshold and on the y-axis the number of filtered lncRNAs, which satisfy the filtering criteria in defined number of samples (Fig. 1B). Each coloured line in Fig. 1B shows the number of lncRNAs that satisfy the threshold criteria in different number of samples. For example, the green line shows the number of lncRNAs that satisfy the filtering criteria in minimum 500 samples. The filtering criteria is defined in term of sum of all lncRNAs that are greater than threshold values. For example, x = 10 and green line means that ~1020

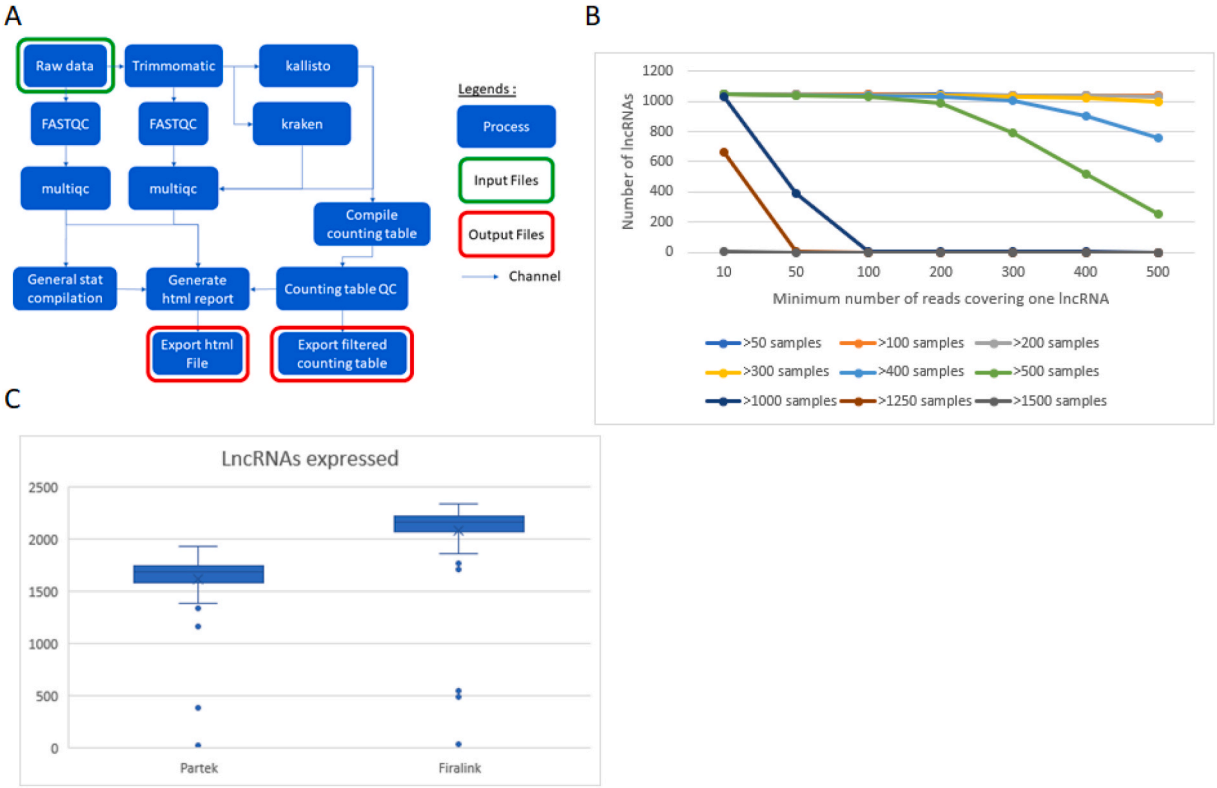

Fig. 1. A. Overall workflow of the Firalink pipeline. **B**. Relationship between count threshold of lncRNAs and number of lncRNAs that satisfy the count threshold in given number of samples indicated by colour code. **C**. Number of lncRNAs detected from the FIMICS panel using the Partek workflow and the Firalink pipeline. Data are from 140 whole blood samples from the PrediCOVID study [9] included in the COVIRNA project. $P < 0.001$ for the comparison between Partek and Firalink (Welch Two Sample *t*-test).

lncRNAs have greater than 10 values in more than 500 samples. Similarly, x = 50 and dark blue line means that ~400 lncRNAs have greater than 50 values in more than 1000 samples.

2. Implementation and results

The Firalink pipeline has been tested using Fastqc v0.11.9, Trimmomatic 0.39, Kraken 2.1.2, Kalisto 0.44.0 and MultiQC 1.12 in whole blood samples collected from 140 COVID-19 patients enrolled in the PrediCOVID study [9] and included in the COVIRNA project (www.covirna.eu). All samples were sequenced with 2x50 paired-end and 1 bp on average was removed from each sequence because of poor quality. GC content remained unchanged during the process of trimming. As shown in Fig. 1C, the use of Firalink pipeline allowed to detect more lncRNAs from the FIMICS panel as compared to the Partek workflow (https://www.partek.com/). The Firalink report used for CE-marking of the FIMICS panel is provided as Supplemental material 1. To test the applicability of the Firalink pipeline to other settings and disease conditions, we conducted an independent targeted sequencing experiment with a panel of lncRNAs dedicated to study brain disease. The analytical report provided in Supplemental material 2 shows the suitability of the Firalink pipeline to be used for other lncRNA panels and targeted sequencing experiments.

3. Novelty and benefits

While the tools used in the Firalink pipeline are mostly standards, they are applied in a coordinated and sequential manner for the first time to lncRNAs quantification from a targeted sequencing panel. Using targeted sequencing and analysis with Firalink pipeline has a major advantage over traditional whole genome sequencing is that it avoids multi-mapping and miss alignments. Due to its enrichment of reads and high sensitivity, combining targeted sequencing with Firalink pipeline allows detecting more lncRNAs than classical sequencing approaches, which represents an important asset when studying lncRNAs in samples with limited input material, such as plasma or serum in human studies. In addition, the use of the Kallisto pseudo-aligner allows a faster analysis as compared to other techniques. Firalink facilitates the use and strengthens the benefit of targeted sequencing approaches which could be helpful not only for discovery but also for clinical application of molecular diagnostic tests, a major interest of the EU-CardioRNA COST Action network [10]. Firalink pipeline has been developed at first instance to work with the FIMICS panel of cardiac-enriched lncRNAs but it is applicable to any targeted sequencing data, as shown using a lncRNA panel for brain disease. The pipeline can be adapted to other targeted sequencing panels or other transcriptomics data (e.g. whole transcriptome) through a change of the reference genome/panel. The use of the Firalink pipeline can therefore be extended to any research project and any disease, thereby covering a wide range of application of targeted sequencing.

Funding

This work was supported by the EU Horizon 2020 project COVIRNA (grant agreement # 101016072). Y.D. has also received funding from the National Research Fund (grants #C14/BM/8225223, C17/BM/11613033 and COVID-19/2020-1/14719577/miRCOVID), the Ministry of Higher Education and Research, and the Heart Foundation-Daniel Wagner of Luxembourg.

CRediT authorship contribution statement

Louis Chauviere: Conceptualization, Methodology, Software, Writing – review & editing. **Lucien Hoffbeck:** Software, Methodology. **Muhammad Shoaib:** Software, Resources. **Florent Tessier:** Software, Methodology. **Huseyin Firat:** Project administration, Supervision, Resources. **Venkata Satagopam:** Project administration, Supervision. **Yvan Devaux:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

Y.D. owns patents related to diagnostic and therapeutic applications of RNAs. Y.D. and H.F. filed a patent on long noncoding RNAs for diagnostic, prognostic and therapeutic uses for pathologies and toxicities inducing heart disorders (WO2018229046). Firalis SA is commercializing the FIMICS and NeuroLINCs panels of long noncoding RNAs for heart and brain diseases.

Acknowledgments

The authors thank all members of COVIRNA project for their contribution. The Predi-COVID study was supported by the Luxembourg National Research Fund (FNR) (Predi-COVID, grant number 14716273), the André Losch Foundation and by European Regional Development Fund (FEDER, convention 2018-04-026-21). We are thankful to all the participants of the Predi-COVID study. We also acknowledge the involvement of the interdisciplinary and inter-institutional study team that contributed to Predi-COVID. The full list of the Predi-COVID team can be found here: https://sites.lih.lu/the-predi-covid-study/about-us/project-team/.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ncrna.2023.09.002.

References

[1] Y. Devaux, et al., Long noncoding RNAs in cardiac development and ageing, Nat. Rev. Cardiol. 12 (7) (2015) 415–425.
[2] H. Firat, et al., FIMICS: a panel of long noncoding RNAs for cardiovascular conditions, Heliyon (2023).
[3] D. de Gonzalo-Calvo, et al., Methodological considerations for circulating long noncoding RNA quantification, Trends Mol. Med. 28 (8) (2022) 616–618.
[4] M. Vanhaverbeke, et al., Peripheral blood RNA biomarkers for cardiovascular disease from bench to bedside: a position paper from the EU-CardioRNA COST action CA17129, Cardiovasc. Res. 118 (16) (2022) 3183–3197.
[5] N.L. Bray, et al., Near-optimal probabilistic RNA-seq quantification, Nat. Biotechnol. 34 (5) (2016) 525–527.
[6] S.W. Wingett, S. FastQ Screen Andrews, A tool for multi-genome mapping and quality control, F1000Res 7 (2018) 1338.
[7] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120.
[8] D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, Genome Biol. 20 (1) (2019) 257.
[9] G. Fagherazzi, et al., Protocol for a prospective, longitudinal cohort of people with COVID-19 and their household members to study factors associated with disease severity: the Predi-COVID study, BMJ Open 10 (11) (2020), e041834.
[10] E.L. Robinson, et al., Leveraging non-coding RNAs to fight cardiovascular disease: the EU-CardioRNA network, Eur. Heart J. 42 (48) (2021) 4881–4883.