prepRNA: an integrative tool for Illumina RNAseq data filtering

Dragana Dudić[1], Bojana Banović Đeri[2], Željko Stanković[3], Zoran Ž. Avramović[3]

[1]*Faculty of Informatics and Computer Science, University Union Nikola Tesla, Belgrade, Serbia,*
ddudic@unionnikolatesla.edu.rs
[2]*Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia,*
bojanabanovic@imgge.bg.ac.rs
[3]*Pan-European University APEIRON, Banja Luka, Bosnia and Herzegovina,*
{zeljko.z.stankovic, zoran.z.avramovic}@apeiron-edu.eu

Contribution to the State of the Art

https://doi.org/10.7251/JIT2201026D UDC: 547.963.3:577.113

Abstract: The vast amount of currently available transcriptome sequences is comprised of Illumina RNAseq data. Usually, publicly available datasets are provided as raw data and preparing them for the downstream NGS analysis is the first step required. Such preprocessing step, besides the evaluation of the quality of the raw data, includes data filtering, in order to provide high quality results of the downstream analysis. Existing tools for NGS data filtering are either too general or incomplete for the Illumina RNAseq filtering task, which is why a new tool for this endeavor was needed. We present prepRNA, a novel tool intended for Illumina RNAseq data filtering, which was designed as a comprehensive and user-friendly wrapper tool with possibility of further upgrading with a quality control option.

Keywords: RNAseq, data filtering, data preprocessing, NGS data, Illumina.

Introduction

In all eukaryotes genome content is transcribed from DNA molecule to RNA molecule by complex biological process. Entirety of RNA molecules, which may arise in the transcription process, make a complete transcriptome of that organism. Total eukaryotic transcriptome is comprised of two different classes of RNA molecules: 1) coding RNAs - messenger RNAs (mRNAs) and 2) non-coding RNAs – ribosomal RNAs (rRNAs), transport RNAs (tRNAs), short RNAs (miRNAs, piRNAs, siRNAs, snRNAs, snoRNAs) and long non-coding RNAs (pRNAs, eRNAs, gsRNAs, lincRNA, NAT), with different functions. Moreover, total transcriptome has a variable content and structure, and its composition directly depends on internal and external conditions, enabling real-time adaptation of an organism to developmental stage and surrounding environment. The main causes of variability in transcriptome content in eukaryotes include alternative splicing, inclusion, exclusion and up- or down-regulation of expression of individual genes via their regulation pathways.

RNA sequencing (RNAseq) represents determination of the sequence of nucleotides in total transcriptome of one organism. Since tRNAs and rRNAs are considerably more abundant than the rest of RNAs, they are usually physically removed during NGS library preparation in order not to mask the rest of the transcriptome. Further, total transcriptome contains numerous different mRNA transcripts at different levels of intron processing. Determining complete RNA content and presence of individual transcripts in RNA, their number and possible nucleotide changes in different organisms, different tissues of one organism, different stages of development and different environmental conditions are important for gene expression determination, leading to insights in potential genes' roles and pathways of regulation as well as in phylogenetic analysis and evolution of certain molecular pathways, gene families, etc.

For the last 40 years, determining total RNA content and content of individual RNA classes and transcripts has been an active issue, during which several sequencing technologies were developed. These new sequencing technologies included: Roche 454, SOLiD, Ion Torrent, Illumina, PacBio and Oxford Nanopore platforms, based on different approaches to determine precise sequence with deep coverage. Increased availability of such Next Generation Sequencing (NGS) technology provided the vast amount of RNA-seq data, since it became favored for transcriptome analysis due to higher coverage and resolution when compared to Sanger and microarray sequencing [1].

Overall insight in publicly available NCBI GEO datasets revealed that among 161,145 published RNAseq datasets (date of search March 17th 2022), 255 originated from ION Torrent technology, while 160,890 originated from Illumina, implying that Illumina platform is the most popular choice for RNA-seq analysis among researchers.

Filtering sequenced data

First step in any NGS analysis is to preprocess the NGS data: to check the quality of the raw data and, according to results of the quality control, to remove the adapters' sequences, contaminating sequences and to omit low quality portions of the NGS dataset (NGS data filtering). Thus, in the Illumina RNAseq domain, necessary data filtering tasks include [2]: contaminant removal in a narrow sense, adapter removal and trimming of low-quality nucleotides.

One of the steps in the process of preparing a sample for NGS sequencing includes adding short sequences of known and constant nucleotide content, called adapters, at the end of every fragment. Although adapters should not be present in the resulting raw set of NGS reads, sometimes, when the length of sequenced molecules vary, as commonly happens when working with total transcriptome (due to the nature of transcriptome composition), adapters are retained at 3' end of a sequence. Their presence can be noticed during the NGS quality control, where several checkpoints indicate abundant presence of the adapters. Even if the quality control checkpoints do not show the presence of the adapters, it still does not exclude their presence, because the quality control tool may not report adapters if they are not abundant enough to reach a sufficient extent in a randomized sample. Nevertheless, adapters should be removed from the NGS dataset, because their presence degrades the quality of the obtained results. Usually, general tools for trimming/preprocessing of NGS data, like Trimmomatic [7], Cutadapt [8] and AfterQC [9], are used for the adapter removal task, but there are also other, more specialized tools, like Scythe.

Contaminants in a narrow sense include all sequences originating from different organisms than the one that was sequenced and as such should be removed from the sequenced data. The first insight in the level of contamination is given by the quality control checkpoints, but in order to get all information further exploration is needed. Contaminant organisms can be detected by sampling 500-1,000 sequences from the RNAseq dataset and searching against the BLAST database. If contaminant organisms are detected, level of contamination can be determined by specific tools for this task, like VecScreen [3] and FastQScreen [4] or with broader purpose tools that have this functionality, like DeconSeq [5]. Contamination removal may be done by using specialized tools like BioBloom or by tools with broader scope like FastqPuri [10] and DeconSeq or even by mapping the NGS data to the reference genomes of the sources of the contamination.

All currently available NGS platforms have certain technological limitations and the presence of some low-quality nucleotide sequences is expected in the sequence dataset. Illumina NGS sequencing uses approach called sequencing-by-synthesis, which is error-prone in its nature and may be the reason of occasional low-quality nucleotides to occur. Usually, in order to deal with these parts of the NGS sequence, the low-quality base trimming is used. Low quality base trimming is a necessary step in a filtering process and it is included in all general tools for filtering/preprocessing of NGS reads. Initial checkpoints of NGS quality control indicate position and level of low-quality bases, which should be used as guidelines for the lower quality boundary in the NGS reads and decision on their removal.

Related work

There are many available tools for filtering NGS data, but among them there are very few designed for filtering RNAseq data, and none of them is specialized for Illumina data. That is why, until now, the only option was to use more general tools for RNAseq data filtering tasks, like Trimmomatic, Cutadapt and AfterQC. Trimmomatic is a tool for filtering Illumina NGS data, Cutadapt is a fully general NGS filtering data tool, while AfterQC is a tool intended for NGS data preprocessing with partial quality control support.

From the perspective of Illumina RNAseq filtering approach, all three tools have similar functionalities: they all have features for quality trimming, adapter removal and removing of the reads below specified length. While AfterQC tool removes adapters automatically, Trimmomatic and Cutadapt, as general tools, have to be provided with the adapters' sequences specific for the used platform. Among two later general tools the difference is in a way of forwarding of adapters' sequences: Trimmomatic expects an input file with adapters' sequences and Cutadapt accepts adapters' sequences as command line argument. Users may have troubles with providing adapter sequences if they are not familiar with sequencing technology and how to find the list of used adapters. Trimmomatic, as a tool designed for Illumina NGS data, offers several files with usual Illumina adapters, but still, the user may be confused which one to choose if they do not have information about the sequencing process. Also, all three tools have a similar disadvantage by missing the feature for contaminant removal in a narrow sense, because this feature is needed to facilitate the further downstream analysis.

Besides general NGS tools, there are also tools specialized for RNAseq data, like FastqPuri and RNA-QC-Chain [11]. RNA-QC-Chain is a tool for RNAseq data preprocessing with partial quality control support and no contaminant removal in a narrow sense. According to [10], FastqPuri is a tool for Illumina RNA-seq data preprocessing with partial quality control support, but it lacks user instructions, documentation and support. Also, tool installation failed on several Unix machines because of unsupported R packages.

Due to all above mentioned issues, we created prepRNA, as a user-friendly tool meant for Illumina RNA-seq complete filtering, which promotes reusing and repurposing instead of reinventing, through combining the best solutions from the existing sources.

Implementation

prepRNA is a command-line Unix wrapper tool built around Trimmomatic and Bowtie tools, which enables comprehensive and user-friendly complete filtering of Illumina RNAseq data. It was intended to be used after finished quality control of RNAseq data and relies on the information on determined level of contamination. The general architecture of a prepRNA tool is presented on Figure 1.

prepRNA consists of three modules: contaminant-index-download module, contaminant-indexing module and filtering module. All modules are easy-to-use with a clear and concise user manual given under --help option. Contaminant-index-download module represents an optional module that enables downloading reference genomes for commonly rep-

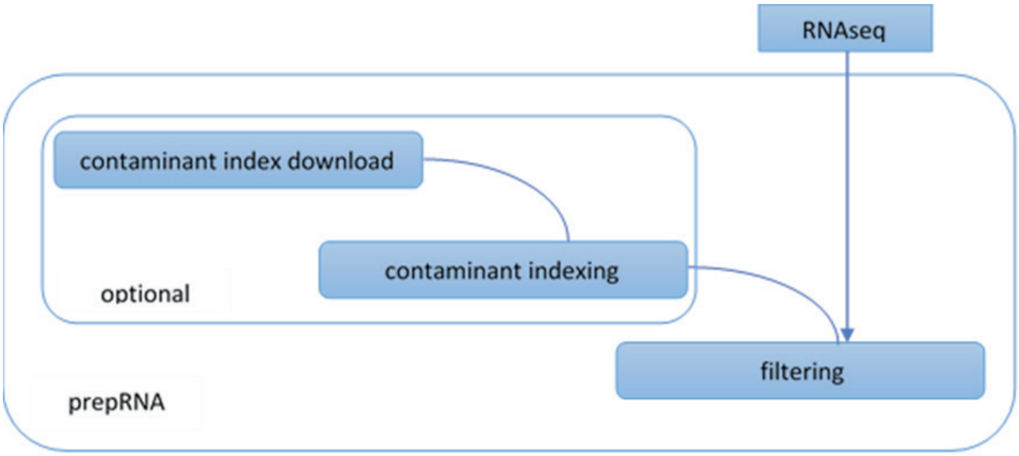


Figure 1. *prepRNA tool architecture*

resented Illumina RNAseq contaminants in a narrow sense, which include: PhiX bacteriophage as control viral DNA, vectors like plasmid, phage, cosmid, BAC, PAC, YAC and transposable elements from the cloning host, which is usually *Escherichia coli* or yeast and human as the NGS library and NGS platform handler. Options offered in this module are straight-forward and every listed contaminant has a separate option which invokes its download. Tool creates a directory inside the current directory where downloaded DNA sequences are stored. After downloading required contaminant reference genomes, they are automatically indexed in order to enable decontamination of Illumina RNAseq data. This is performed within the contaminant-indexing module, which was built as the wrapper around bowtie-build functionality for indexing sequencing data of the Bowtie tool. Bowtie tool is a well-known and broadly used tool for mapping of DNAseq data, but in this case, it was used for indexing and mapping of RNAseq data of selected contaminant organisms, because no splice-awareness is needed for decontamination of RNAseq data. Once created, indexes are stored together with contaminant sequences and could easily be reused in other sequencing projects. Since contaminant-indexing module depends on the contaminant-index-download module, they share the same options.

The filtering module was built as a wrapper around Trimmomatic and Bowtie tools, offering necessary and sufficient options for preprocessing of Illumina RNAseq data: contaminant removal in narrow sense, adapter removal and quality trimming. For all of the aforementioned tasks, a separate option is offered. Contaminant removal option expects value as a string that contains first letters of contaminants in any order. Quality trimming option expects integer value between 1 and 40, covering the full range of quality values available in Illumina RNAseq data. Adapter removal is fully automated step and it has no value, because the domain of interest is specialized to Illumina RNAseq. Additionally, filtering of extremely short nucleotide sequences was set to be done automatically, while duplicate removal and removal of the content bias at the 5’ end were omitted, as not beneficial for the further analysis of the Illumina RNAseq data [2].

Regarding the steps of adapter removal, quality trimming or adapter removal together with quality trimming, prepRNA was created as a wrapper tool around Trimmomatic. Trimmomatic is a widely used tool for filtering Illumina sequencing data characterized by good performance, but not as user-friendly for people not having good computer skills. Moreover, Trimmomatic lacks some important functionalities necessary for preprocessing of RNAseq data, like removal of contaminants in a narrow sense, which is why we added this functionality by using the mapping method with the Bowtie tool, through which RNAseq data are mapped to the selected downloaded and indexed contaminant reference genome. Only unmapped data are then processed and forwarded for further RNAseq filtering or offered for output in adequate format, depending on the task. Also, in prepRNA tool the parameters used in adapter removal, quality trimming or adapter removal together with quality trimming are automatically set to the values which showed the best performance for RNAseq data globally. In that way, prepRNA is easy to use by non-computer specialized users, like biologists, physicists and chemists.

Evaluation

There is a lack of tools for raw Illumina RNAseq data filtering, which is why more general filtering tools are being used for this task, leading to different difficulties for non-computer specialized users. Although most of general tools give good results for selected tasks, none of them offer a complete list of filtering tasks needed for the Illumina RNAseq data. Unlike similar existing tools, prepRNA provides complete and comprehensive filtering of Illumina RNAseq data. Comparison of five existing filtering tools with prepRNA tool is given in the Table 1.

In the matter of domain, among six compared tools, three tools were designed for RNAseq data and two tools were adapted for Illumina data, but only prepRNA was designed for both Illumina and RNAseq data. Regarding user friendliness and easiness for installation three out of six tools met both criteria: Cutadapt, AfterQC and prepRNA, but among them only prepRNA was designed for RNAseq and Illumina data. Features comparison showed that among six tools only two, prepRNA and FastqPuri, provided all the features needed for preprocessing of Illumina RNAseq data, but FastqPuri did not meet two other important criteria of being easy to

***Table 1.** Comparison of tools for filtering RNAseq data. QC – quality control, CR -contaminant removal, AR- adapter removal, QF – quality filtering*

Tool	prepRNA	Trimmomatic	Cutadapt	FastqPuri	AfterQC	RNA-QC-Chain
Language	C, java	java	C,Python	C, R	C,Python	C++
Mode	SE, PE	SE, PE	SE, PE	SE, PE	PE	SE
Format	fq.gz	fq*	fq, fa, gz	fq*	fq	fq,fa
RNAseq	+	-	-	+	-	+
Illumina	+	+	-	-	-	-
easy to install	+	+	+	-	+	+
easy to use	+	-	+	-	+	-
QC	-	-	-	+/-	+/-	+/-
CR	+	-	-	+	-	-
AR	+	+	+	+	+	+
QF	+	+	+	+	+	+

install and easy to use, while prepRNA did. Regarding this last consideration one should have in mind that aforementioned tools are designed for Unix-like systems and mostly used on server machines, where regular users do not have permissions for installing software anywhere but the home directory. This makes installing dependencies difficult, which may disable functionalities of a tool. That is why it is more convenient to use tools developed on programming languages that are built in the system, because compatibility and functionality are guaranteed in that case.

Considering tools' availability to be used with two reading modes, single-end (SE) and paired-end (PE), most of aforementioned tools, including prepRNA, supported both ways. This feature is important because Illumina PE sequencing is widely used RNA sequencing method and supporting PE reading in RNAseq filtering is necessary. In PE reading, obtained sequences are more reliable because sequencer reads every fragment twice, rather than in SE reading, where fragments are read once. In respect of tools support of sequencing formats, it was essential to support the widely used compressed format for Illumina RNAseq data, fastq.gz (fq.gz) format, which stand for all aforementioned tools, including prepRNA. Finally, RNAseq data filtering resides on the results of quality control, which makes convenient for RNAseq quality control and RNAseq data filtering to be combined in one tool. Even though three out of six tested tools - FastqPuri, AfterQC and RNA-QC-Chain tried to implement this task, none of them provided comprehensive quality control needed for RNAseq data. Thus, prepRNA did not tried to implement it, but rather to lean on output of existing comprehensive quality control tools, like FastQC.

CONCLUSION

Existing tools that can be used for Illumina RNAseq filtering tasks lack some functionalities necessary for the completeness of the filtering process. Thus, we developed prepRNA, a user-friendly and comprehensive tool for NGS data filtering designed especially for Illumina RNAseq data. prepRNA is maintained, easy-to-install and easy-to-use tool with clear user manual and installation instructions. This tool was designed as a wrapper tool, combining two broadly used tools in the bioinformatics community, with an aim to fulfill all the requirements of the Illumina RNAseq data filtering process. It provides contaminant removal, adapter removal and quality trimming steps, for which it uses parameters, determined to be the best from global RNAseq practice, as default ones. Moreover, prepRNA can be extended with quality control functionality in order to address the whole process of Illumina RNAseq data preprocessing. Finally, prepRNA takes user's input in short and unambiguous form expecting only RNA sequences files and desired filtering options provided as an input, making it highly desirable among non-computer specialized users.

REFERENCES

[1] Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, *2015*(11), 951–969. https://doi.org/10.1101/pdb.top084970

[2] Dudić, D., Đeri, B. B., Pajić, V., & Pavlović-Lažetić, G. (2021). Demystification of RNAseq Quality Control. JITA, 22(2), 73-86.

[3] Schäffer, A.A., Nawrocki, E.P., Choi, Y., Kitts, P.A., Karsch-Mizrachi, I., McVeigh, R. (2018) Vec-Screen plus taxonomy: imposing a tax(onomy) increase on vector contamination screening. Bioinformatics 34(5), 755–759

[4] Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research, 7, 1338. https://doi.org/10.12688/f1000research.15931.2

[5] R. Schmieder, R. Edwards (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One, 6: e17288, doi: https://doi.org/10.1371/journal.pone.0017288.

[6] Pérez-Rubio, P., Lottaz, C. & Engelmann, J.C. (2019) FastqPuri: high-performance preprocessing of RNA-seq data. BMC Bioinformatics 20, 226 . https://doi.org/10.1186/s12859-019-2799-0

[7] Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

[8] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), pp. 10-12. doi:https://doi.org/10.14806/ej.17.1.200

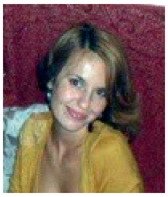
[9] Chen, S., Huang, T., Zhou, Y. et al. (2017) AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics 18, 80 . https://doi.org/10.1186/s12859-017-1469-3

[10] Pérez-Rubio, P., Lottaz, C. & Engelmann, J.C. (2019) FastqPuri: high-performance preprocessing of RNA-seq data. BMC Bioinformatics 20, 226 . https://doi.org/10.1186/s12859-019-2799-0

[11] Zhou, Q., Su, X., Jing, G. et al. (2018) RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC Genomics 19, 144 . https://doi.org/10.1186/s12864-018-4503-6

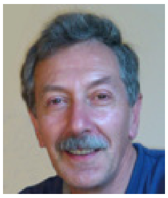
Received: December 12, 2021
Accepted: April 1, 2022

About the authors



Dragana Dudić is an Assistant at the Faculty of Informatics and Computer Science at the University Union Nikola Tesla for several Computer Science subjects. She obtained BSc and MSc in Computer Science at the Faculty of Mathematics, University of Belgrade where she is finishing her PhD in the field of Bioinformatics. Dragana is interested in research of repetitive sequences, especially human and plant transcriptomics.

Bojana Banović Đeri is a Associate Research Professor in the Laboratory for Plant Molecular Biology in the Institute of Molecular Genetics and Genetic Engineering, University of Belgrade. In 2013. obtained PhD in Molecular biology at the Faculty of Biology, University of Belgrade in the field of plants' molecular biology, genetics and proteomics (characterization of buckwheat heteromorphic self-incompatibility), with later specialization in bioinformatics (until 2021: nine international trainings and programme "Bioinformatics for biologists" at the School of Computing, Union University, Belgrade). In 2017 participated in the US expert exchange programme „Open World" for food safety and security. Her scientific interests primarily includes plants' response to stress, single stranded RNA viruses and food safety. Author of 19 research papers (https://orcid.org/0000-0001-7663-1878) cited more than 135 times (Google Scholar). Coordinator of one international and one national project and participant in 14 projects (national and international). Lecturer at Molecular Biology module for PhD students, Faculty of Biology, University of Belgrade. Mentor of one PhD thesis and three master thesis. Active as a member of several scientific societies, reviewer of international scientific papers and projects, organizer of several conferences and workshops. Engaged in science promotion (Researchers' Night, Plant Fascination Day, etc.).



Željko Stanković was born in Belgrade. He finished primary and secondary school in his hometown. He received his higher education in Cleveland, Ohio, USA, where he graduated in 1981. In 2006, he defended his master's thesis at the University of Novi Sad. He defended his doctoral dissertation at Singidunum University in 2010. He has been programming since 1984. making programs for his first computer Commodore 64. Robotics and bioengineering are a long-standing field of work and interest.

Zoran Ž. Avramović was born in Serbia, on September 10, 1953. He finished elementary school and high school in Loznica with great success. He was awarded several diplomas by Nikola Tesla and Mihailo Petrović Alas. He graduated on time at the University of Belgrade - Faculty of Electrical Engineering, with an average grade of 9.72 in five-year studies. He received his master's degree at that faculty (all excellent grades, exams and master's degrees), and then obtained a doctorate in technical sciences (in 1988). As an excellent student of the University, he had the right and at the same time studied mathematics at the Faculty of Mathematics in Belgrade. He was the champion of Serbia in mathematics ("first prize") and Yugoslavia in electrical engineering ("gold medal")

For citation

Dragana Dudić, Bojana Banović Đeri, Željko Stanković, Zoran Ž. Avramović, prepRNA: an integrative tool for Illumina RNAseq data filtering, *JITA – Journal of Information Technology and Applications, Banja Luka*, Pan-Europien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 12(2022) 1:26-31, (UDC: 547.963.3:577.113), (DOI: 10.7251/JIT2201026D), Volume 12, Number 1, Banja Luka, June (1-64), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004