MethodsX 10 (2023) 101991

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

tncRNA Toolkit: A pipeline for convenient identification of RNA (tRNA)-derived non-coding RNAs

Shafaque Zahra, Ajeet Singh, Shailesh Kumar*

Bioinformatics Lab, National Institute of Plant Genome Research (NIPGR), Aruna Asaf Ali Marg, New Delhi 110067, India

ARTICLE INFO

Method name:
tncRNA Toolkit: A pipeline for identification of tRNA-derived non-coding RNAs

Keywords:
Non-coding RNAs
sRNA-seq
Pipeline
tRFs
tncRNAs
Transcription regulation

ABSTRACT

Insights into the eukaryotic gene regulation networks have improved due to the advent of diverse classes of non-coding RNAs. The transfer RNA (tRNA)-derived non-coding RNAs or tncRNAs is a novel class of non-coding RNAs, shown to regulate gene expression at transcription and translation levels. Here, we present a pipeline 'tncRNA Toolkit' for accurately identifying tncRNAs using small RNA sequencing (sRNA-seq) data. Previously, we identified tncRNA in six major angiosperms by utilizing our pipeline and highlighted the significant points regarding their generation and functions. The 'tncRNA Toolkit' is available at the URL: http://www.nipgr.ac.in/tncRNA. The scripts are written in bash and Python3 programming languages. The program can be efficiently run as a standalone command-line tool and installed in any Linux-based Operating System (OS). The user can run this program by providing the input of sRNA-seq data and genome file.

The various features of the 'tncRNA Toolkit' are as follows:

- Major tncRNA classes identified by this tool include tRF-5, tRF-3, tRF-1, 5′tRH, 3′tRH, and leader tRF. Also, it categorizes miscellaneous tncRNAs as other tRF.
- It provides the following information for each identified tncRNA viz. tncRNA class, raw and normalized read count (RPM), read length, progenitor tRNA information (amino acid, anticodon, locus, strand), tncRNA sequence, and tRNA modification sites.
- We hope to facilitate quick and reliable tncRNA identification, which will boost the exploration of this novel class of non-coding RNAs and their relevance in the living world, including plants.

Specifications Table

Subject Area:	Biological sciences
More specific subject area:	Bioinformatics
Method name:	tncRNA Toolkit: A pipeline for identification of tRNA-derived non-coding RNAs
Name and reference of original method:	Shafaque Zahra, Ajeet Singh, Nikita Poddar, Shailesh Kumar. Transfer RNA-derived non-coding RNAs (tncRNAs): Hidden regulation of plants' transcriptional regulatory circuits. Computational and Structural Biotechnology Journal, Volume 19, 2021, Pages 5278–5291, ISSN 2001–0370 https://doi.org/10.1016/j.csbj.2021.09.021.
Resource availability:	https://github.com/skbinfo/tncRNA-Toolkit http://www.nipgr.ac.in/tncRNA https://zenodo.org/badge/latestdoi/387762887

* Corresponding author.
E-mail address: shailesh@nipgr.ac.in (S. Kumar).

https://doi.org/10.1016/j.mex.2022.101991
Received 31 October 2022; Accepted 28 December 2022
Available online 29 December 2022

Introduction and background

The abundance of non-coding RNAs (ncRNAs) repertoire with various regulatory functions has been well identified and documented over the last ten years due to a significant development in molecular biology. They are effective post-transcriptional, epigenetic, and transcriptional regulators of gene expression in living systems [1]. The discovery of short untranslated RNAs, other than microRNAs (miRNAs) and small interfering RNAs (siRNAs), has been dramatically improved by developing high-throughput sequencing technologies. In all three domains of life, tRNA-derived non-coding RNAs (tncRNAs) have been identified [2]. This particular class of regulatory RNAs is generated by the endonucleolytic cleavage of mature or precursor transfer RNAs (tRNAs) [3]. In addition to the well-known shorter tRNA-derived RNA fragments (tRFs) [4] or tRNA-derived RNAs (tDRs) in a size range of 14 to 30 nucleotide (nt) [5], longer tRNA halves (tRHs) of length 30 to 40 nt [6], are also considered as tncRNAs [7]. Their production depends on the nature of tRNA, cell type, developmental stage, stress, and tissue [8]. Also, nuclear and organellar tRNAs can be the source of these molecules [9]. In different types of cancers [10, 11], and plant stresses [12], this family of short RNAs has received much attention. By association with their cognate target mRNAs or proteins, some tRFs alter the cell's gene expression and translation process [13].

Identifying tncRNAs in small RNA-seq (sRNA-seq) datasets is extremely difficult and error-prone [14]. Extensively altered tRNA nucleosides impact reverse transcription, leading to nucleotide truncation or misincorporation during library preparation. This, in turn, raises the likelihood of a mismatch. But enabling a single mismatch to triumph over sequencing faults could result in base misidentification, increasing the number of false negatives [15]. That's because the 20 primary tRNA isotypes include both isodecoders (tRNAs carrying the same anticodon with variations in the body sequence) and isoacceptors (tRNAs with various anticodons but charged with the same amino acid), share a high degree of sequence similarity. For the identification of tRNA-derived reads, various mapping techniques have been proposed. Some computational methods, such as tDRmapper [16], MINTmap [17], and tRF2Cancer [18], are available for their detection; however, they were developed and evaluated on human datasets and are only appropriate for use with human data. Compared to humans, tncRNAs have been less explored in the plant domain. While specialized plant databases, such as tRex [19] and PtRFdb [20], containing data related to the tRFs, have also recently been developed, a practical technique for the precise identification and study of tncRNAs in plants is still absent. We have developed a pipeline for the quick and accurate detection of tncRNAs [7] in plants, but it can be used for any species.

Features of 'tncRNA toolkit'

Genome index building, tRNA prediction, and read mapping

At first, the tncRNA Toolkit takes the genome containing DNA sequences in FASTA format as input, runs tRNAscan-SE [21], and performs tRNA prediction using the genome sequence. For the nuclear genome, tRNAscan-SE runs on default mode. For the organellar genome, the "-O" mode is used. Due to the high number of tRNA gene copies predicted by tRNAscanSE, we first filter the predicted tRNA gene pool by deleting pseudogenes and keeping true tRNAs based on the score. Only high-quality tRNAs with a score of 50 or more are chosen for mapping since the score is a crucial indicator of the structural propensity of tRNA [22].

The genome is modified by masking the tRNA genes and their corresponding 50 nt upstream (leader) and downstream (trailer) in the genome, thus called the masked genome. Pre-tRNA (5′ leader and 3′ trailer portion) and mature tRNA (addition of CCA tail at 3′ end) set are inserted as artificial chromosomes. Masked genome and artificial chromosomes combine to form an artificial genome, and a bowtie index is created using bowtie build (bowtie v1.3) [23]. This index is further utilized for reads mapping. Filtered reads (trimmed and high-quality reads) are provided as input to the tncRNA Toolkit. To prevent ambiguous reads from non-tRNA regions, only those reads aligning to tRNA regions (artificial chromosomes) are considered for further tncRNA identification analysis. Also, it makes the normalization of tncRNAs better, as it calculates most of the mapped reads. HAMR [24] is used for tRNA nucleoside modification detection from the mapped reads.

The mapped reads from the previous step are utilized to create a FASTA file containing unique reads with the count, using SAM flags 0 & 16 for single-end reads. Mapped unique reads are aligned to + strand only without mismatch, and multi-mapped reads are discarded if alignment occurs more than 50 times, using bowtie1 with "–norc -v 0 -m 50″ arguments [23]. The bowtie combinatorial arguments "–best" and "–strata" are used, which guarantees that reported singleton alignments are best in terms of the stratum. Then, the output is used to identify locus, location, length of reads mapped to mature tRNA, and their 50 bp upstream and downstream flank; based on that, reads are categorized into various tncRNA subtypes. The tRNA halves are classified by cleavage at the anticodon loop (2 nt + 3 nt of anticodon + 2 nt = 7 nt), and information for the modification site from HAMR is added.

tncRNA identification and classification

The excision of pre-tRNA and mature tRNA leads to the biogenesis of several tncRNA subtypes. The pre-tRNA gives rise to leader tRF and tRF-1/tsRNA [3], whereas mature tRNA generates tRF-5, tRF-3 (with or without CCA), 5′tRH, 3′tRH (with or without CCA), and internal tRF [25]. Thus, tncRNAs are classified based on tRNA origin and cleavage site. For accurate identification of tRNA-derived reads, those reads will be classified as tncRNAs that map exclusively to the artificial chromosomes (i.e., tRNA set consisting of CCA containing mature tRNA, 50 nt leader, and 50 nt trailer sequences). The reads identified as tncRNAs are categorized into tRF-5 (reads from 5′ end of mature tRNA), tRF-3 (reads from 3′ end of mature tRNA, with or without CCA), tRF-1 (from trailer portion of pre-tRNA), leader-tRF (from leader region of pre-tRNA), 5′tRH (from 5′ end of mature tRNA containing anticodon loop portion),

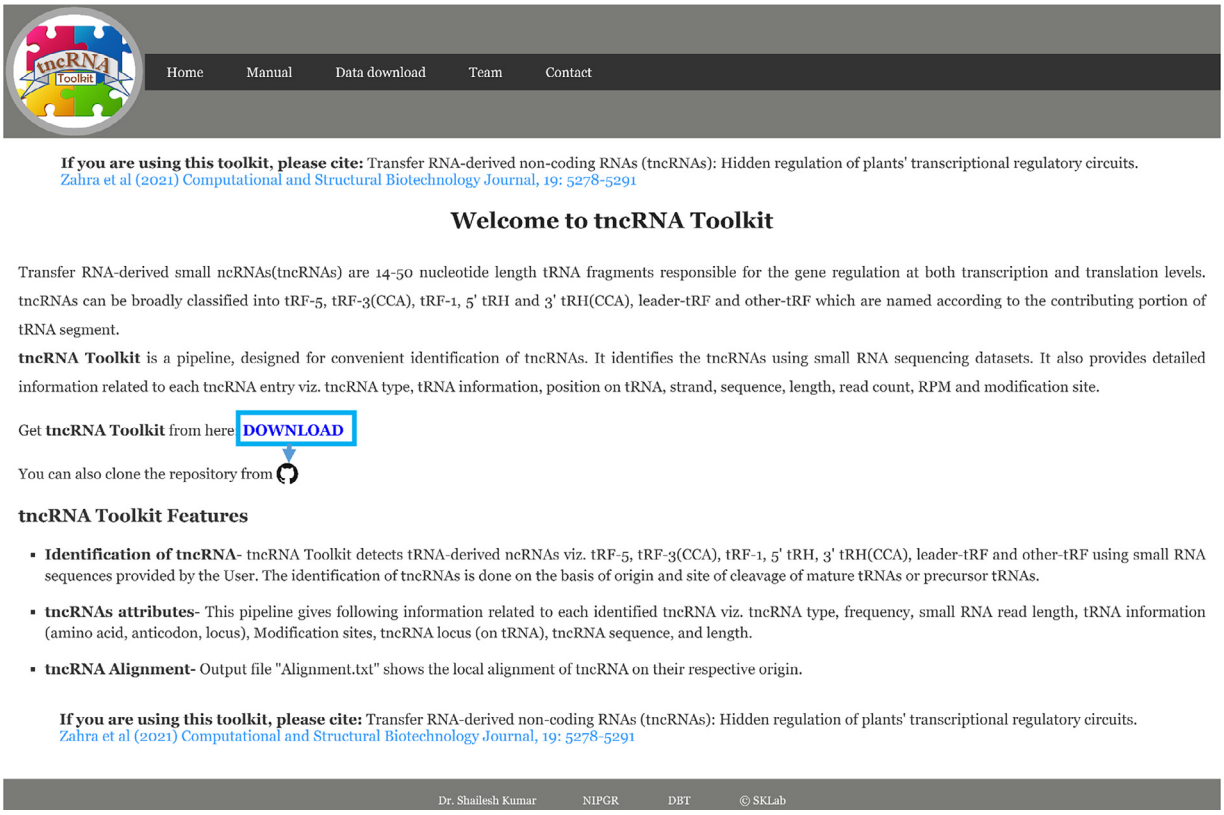

Fig. 1. The user interface of the homepage of the tncRNA Toolkit showing the download link.

3′tRH (from 3′ end of mature tRNA containing anticodon loop portion, with or without CCA), and other tRF (from internal region of mature tRNA excluding extreme ends).

Structure and usage of the tncRNA Toolkit

The tncRNA-Toolkit can be downloaded from the tncRNA website, http://www.nipgr.ac.in/tncRNA/ (shown in Fig. 1). This distribution includes the python3 script "tncRNAs.py" and additional scripts in the 'util' folder (Fig. 2). The prerequisites and installation steps have been thoroughly explained in the manual page of tncRNA website. After completing the installation steps, this pipeline is ready to run.

First step:

Building the bowtie index using the genome of interest (given as a fasta file) using the command below:

python3 tncRNAs.py -g ⟨genome fasta⟩ -s ⟨species name⟩

This command will generate the bowtie index and needed files in "lib/indexes/<provided species name>."
The example is shown below:

python3 tncRNAs.py -g At_genome.fa -s Arabidopsis_thaliana

It should be noted that the genome fasta header should start with ">chr[N*um*]". Mitochondrial and plastid fasta headers also should begin as ">chrMt" & ">chrPt", respectively. This format will be helpful for the automation of scripts and the distinction of nuclear & organellar regions. Once the index is built, the user can further analyze the processed sRNA single-end data for that species, as shown in the second step.

Second step:

tncRNA identification using single end (processed and quality checked) small RNA-seq data in fastq format using the command below:

python3 tncRNAs.py -s ⟨species name⟩ -i ⟨processed small RNA reads⟩ -o ⟨output dir⟩

It should be noted that the species name provided should be identical to the name provided for building the index in the first step.
Example:

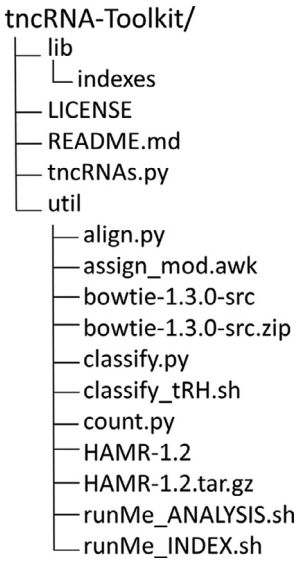

Fig. 2. Recursive list of tncRNA Toolkit directory in Tree format.

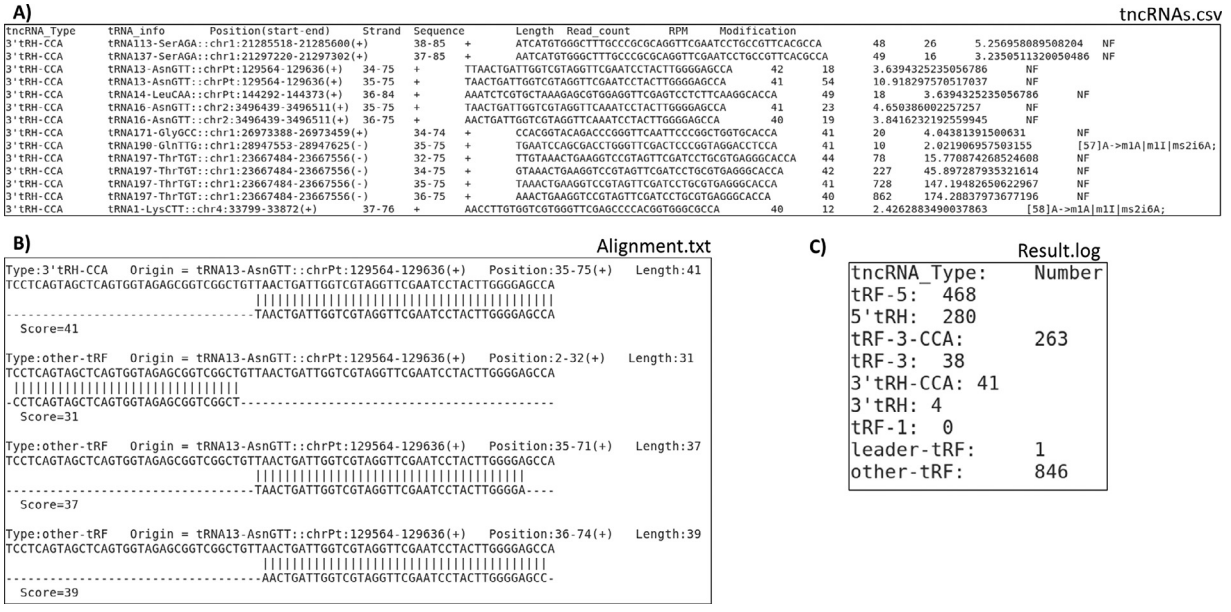

Fig. 3. Screenshots of output files generated by tncRNA Toolkit, (A) tncRNAs.csv file showing the list of tncRNAs identified along with relevant information, (B) Alignment.txt file showing the tncRNA alignment on its progenitor tRNA, (C) Result.log file showing the summary of various class-wise tncRNA counts.

python3 tncRNAs.py -s Arabidopsis_thaliana -i SRR1693713_trimmed.fq -o test_dir

After the successful completion of this command, three output files will be generated in the output directory, viz. tncRNAs.csv, Alignment.txt, and Result.log.

tncRNA Toolkit output files

In the provided directory with the '-o' option as described above, the tncRNA Toolkit provides three result files:

1. tncRNAs.csv

Various helpful information related to each tncRNA identified is provided in the output file (tncRNAs.csv) viz. tncRNA type, parental tRNA locus information, tncRNA position on tRNA (start-end), strand, sequence, length, read count, RPM, nature, and position of modified nucleosides (Fig. 3A).

2. Alignment.txt

The alignment of each tncRNA sequence over its parental tRNA sequence, which includes mature tRNA, 5′ leader, and 3′ trailer precursor tRNA sequences, is shown in this file, including the alignment score, along with some basic meta-data from tncRNAs.csv (Fig. 3B).

3. Result.log

This log file (Fig. 3C) keeps track of statistical data regarding various tncRNA sub-types identified in the sample.

Conclusion

Here, we have introduced the tncRNA Toolkit, an open-source program to make tncRNA identification and analysis easier. It was created in Python3 and Bash. It serves as a starting point for future exploration of tncRNA research. It is simple to run for a user with basic bioinformatics skills. Any organism can be selected for tncRNA analysis using our software.

Additionally, accessory choices are offered for a user-customized examination with various parameters. It can provide an accurate and speedy tncRNA prediction using the two processes outlined above. This method is helpful for mining tncRNAs and carrying out extensive analysis, as was done in our recent study [26]. Our pipeline significantly advances the understanding of tncRNA detection in plants by utilizing methods that can enhance tncRNA identification and provides flexibility using miscellaneous options while understanding a complex sRNA molecule of great significance as potential modulators of gene expression. Our goal is to produce a tool that steadily quickens the large-scale analysis of tncRNAs.

Fundings

This research is supported by the BT/PR40146/BTIS/137/4/2020 project grant from the Department of Biotechnology (DBT), Government of India, and a core research grant of NIPGR, New Delhi.

Authors contributions

SZ and SK wrote the manuscript. AS helped in the data collection and compilation. SK designed all the analyses & experiments, and conceived the idea, and coordinated the project.

Data and resource availability

All pipeline scripts, codes, and results analyzed from our previous study are freely available on our website (URLs: http://nipgr.ac.in/tncRNA). The codes and usage are also available at https://github.com/skbinfo/tncRNA-Toolkit. We have used the *Arabidopsis thaliana* genome (TAIR10.1) and sRNA-Seq data from accession SRR1693713 to show the results. Fasta headers of the genome were transformed for starting with “chr” (e.g.,>chr1, >chr2, and likewise) to make them convenient for secondary analysis.

Declaration of Competing Interest

The authors have no conflicts of interest to declare.

Data Availability

No data was used for the research described in the article.

Acknowledgment

SZ and AS thank the council of scientific and industrial research (CSIR), India, for research fellowships. DBT (Department of Biotechnology)-eLibrary Consortium (DeLCON) is acknowledged for providing e-resources. SK acknowledges the Computational Biology and Bioinformatics Facility (CBBF) at NIPGR, New Delhi.

References

[1] K.V. Morris, J.S. Mattick, The rise of regulatory RNA, Nat. Rev. Genet. 15 (6) (2014) 423–437 Jun., doi:10.1038/nrg3722.
[2] S.P. Keam, G. Hutvagner, tRNA-derived fragments (tRFs): emerging new roles for an ancient RNA in the regulation of gene expression, Life 5 (4) (2015) 1638–1651 *(Basel)*Nov., doi:10.3390/life5041638.
[3] Y.S. Lee, Y. Shibata, A. Malhotra, A. Dutta, A novel class of small RNAs: tRNA-derived RNA fragments (tRFs), Genes Dev. 23 (22) (2009) 2639–2649 Nov., doi:10.1101/gad.1837609.

[4] L. Zhu, J. Ge, T. Li, Y. Shen, J. Guo, tRNA-derived fragments and tRNA halves: the new players in cancers, Cancer Lett. 452 (2019) 31–37 Jun., doi:10.1016/J.CANLET.2019.03.012.
[5] C. Kuscu, P. Kumar, M. Kiran, Z. Su, A. Malik, A. Dutta, tRNA fragments (tRFs) guide ago to regulate gene expression post-transcriptionally in a Dicer-independent manner, RNA 24 (8) (2018) 1093–1105 May, doi:10.1261/rna.066126.118.
[6] J.M. Dhahbi, 5' tRNA halves: the next generation of immune signaling molecules, Front. Immunol. 6 (2015) 74, doi:10.3389/fimmu.2015.00074.
[7] S. Zahra, A. Singh, N. Poddar, S. Kumar, Transfer RNA-derived non-coding RNAs (tncRNAs): hidden regulation of plants' transcriptional regulatory circuits, Comput. Struct. Biotechnol. J. 19 (2021) 5278–5291 Jan., doi:10.1016/J.CSBJ.2021.09.021.
[8] M. Raina, M. Ibba, tRNAs as regulators of biological processes, Front. Genet. 5 (2014) 171Jun., doi:10.3389/fgene.2014.00171.
[9] V. Cognat, et al., The nuclear and organellar tRNA-derived RNA fragment population in Arabidopsis thaliana is highly dynamic, Nucleic Acids Res. 45 (6) (2017) 3460–3472 Apr., doi:10.1093/nar/gkw1122.
[10] P. Kumar, C. Kuscu, A. Dutta, Biogenesis and function of transfer RNA-related fragments (tRFs, Trends Biochem. Sci. 41 (8) (2016) 679–689 Aug., doi:10.1016/J.TIBS.2016.05.004.
[11] A.G. Telonis, et al., Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies, Oncotarget 6 (28) (2015) 24797–24822, doi:10.18632/oncotarget.4695.
[12] C.S. Alves, F.T.S. Nogueira, Plant small RNA world growing bigger: tRNA-derived fragments, longstanding players in regulatory processes, Front. Mol. Biosci. 8 (2021) 1–7 no. June, doi:10.3389/fmolb.2021.638911.
[13] A.R. Soares, M. Santos, Discovery and function of transfer RNA-derived fragments and their role in disease, WIREs RNA 8 (2017) 1423, doi:10.1002/wrna.1423.
[14] A. Hoffmann, J. Fallmann, E. Vilardo, M. Mörl, P.F. Stadler, F. Amman, Accurate mapping of tRNA reads, Bioinformatics 34 (7) (2018) 1116–1124 Apr., doi:10.1093/bioinformatics/btx756.
[15] A.G. Telonis, P. Loher, Y. Kirino, I. Rigoutsos, Consequential considerations when mapping tRNA fragments, BMC Bioinform. 17 (2016) 123 Mar., doi:10.1186/s12859-016-0921-0.
[16] S.R. Selitsky, P. Sethupathy, tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data, BMC Bioinform. 16 (1) (2015) 354 Dec., doi:10.1186/s12859-015-0800-0.
[17] P. Loher, A.G. Telonis, I. Rigoutsos, MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data, Sci. Rep. 7 (1) (2017) 41184 Mar., doi:10.1038/srep41184.
[18] L.-.L. Zheng, et al., tRF2Cancer: a web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers, Nucleic Acids Res. 44 (W1) (2016) W185–W193 Jul., doi:10.1093/nar/gkw414.
[19] A. Thompson, et al., tRex: a web portal for exploration of tRNA-derived fragments in Arabidopsis thaliana, Plant Cell Physiol. 59 (1) (2018) e1–e1Jan., doi:10.1093/pcp/pcx173.
[20] N. Gupta, A. Singh, S. Zahra, S. Kumar, PtRFdb: a database for plant transfer RNA-derived fragments, Database 2018 (2018) *(Oxford)*, doi:10.1093/database/bay063.
[21] T.M. Lowe, S.R. Eddy, TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Res. 25 (5) (1996) 955–964, doi:10.1093/nar/25.5.0955.
[22] T. Pan, Modifications and functional genomics of human transfer RNA, Cell Res. 28 (4) (2018) 395–404 Apr., doi:10.1038/s41422-018-0013-y.
[23] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (3) (2009) R25 Mar., doi:10.1186/gb-2009-10-3-r25.
[24] P. Ryvkin, et al., HAMR: high-throughput annotation of modified ribonucleotides, RNA 19 (12) (2013) 1684–1692 Dec., doi:10.1261/rna.036806.112.
[25] P. Kumar, J. Anaya, S.B. Mudunuri, A. Dutta, Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets, BMC Biol. 12 (1) (2014) 78 Dec., doi:10.1186/s12915-014-0078-0.
[26] S. Zahra, R. Bhardwaj, S. Sharma, A. Singh, S. Kumar, PtncRNAdb: plant transfer RNA-derived non-coding RNAs (tncRNAs) database, 3 Biotech 12 (5) (2022) 1–7, doi:10.1007/s13205-022-03174-7.