Bioinformatics, 38(10), 2022, 2934–2936
https://doi.org/10.1093/bioinformatics/btac198
Advance Access Publication Date: 7 April 2022
Applications Note

Sequence analysis

tRNAstudio: facilitating the study of human mature tRNAs from deep sequencing datasets

Marina Murillo-Recio[1], Ignacio Miguel Martínez de Lejarza Samper[1], Cristina Tuñí i Domínguez[1], Lluís Ribas de Pouplana [1,2,*] and Adrian Gabriel Torres [1,*]

[1]Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Catalonia 08028, Spain and [2]Catalan Institution for Research and Advanced Studies, Barcelona, Catalonia 08010, Spain

*To whom correspondence should be addressed.

Associate Editor: Christina Kendziorski

Received on September 10, 2021; revised on March 17, 2022; editorial decision on March 29, 2022

Abstract

Summary: High-throughput sequencing of transfer RNAs (tRNA-Seq) is a powerful approach to characterize the cellular tRNA pool. Currently, however, analyzing tRNA-Seq datasets requires strong bioinformatics and programming skills. tRNAstudio facilitates the analysis of tRNA-Seq datasets and extracts information on tRNA gene expression, post-transcriptional tRNA modification levels, and tRNA processing steps. Users need only running a few simple bash commands to activate a graphical user interface that allows the easy processing of tRNA-Seq datasets in local mode. Output files include extensive graphical representations and associated numerical tables, and an interactive html summary report to help interpret the data. We have validated tRNAstudio using datasets generated by different experimental methods and derived from human cell lines and tissues that present distinct patterns of tRNA expression, modification and processing.

Availability and implementation: Freely available at https://github.com/GeneTranslationLab-IRB/tRNAstudio under an open-source GNU GPL v3.0 license.

Contact: adriangabriel.torres@irbbarcelona.org or lluis.ribas@irbbarcelona.org.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transfer RNAs (tRNAs) are small non-coding RNAs that bring amino acids to the ribosome for protein synthesis. They are transcribed as longer precursor tRNAs (pre-tRNAs) that need to be processed and chemically modified in order to become fully active. Mature tRNAs can also be further processed into tRNA-derived fragments (tRFs) that perform a wide range of non-canonical tRNA functions (Su *et al.*, 2020).

High-throughput sequencing of tRNAs is a powerful approach to study tRNA biology (Torres *et al.*, 2015, 2019). Several methods have been developed to sequence tRNAs, ranging from standard small RNA-Seq to specialized methods such as DM-tRNA-Seq (Zheng *et al.*, 2015), Arm-Seq (Cozen *et al.*, 2015), YAMAT-Seq (Shigematsu *et al.*, 2017), mim-tRNA-Seq (Behrens *et al.*, 2021), AQRNA-Seq (Hu *et al.*, 2021), among others (we herein refer to any deep sequencing method that can detect tRNA reads as 'tRNA-Seq'). However, analyzing tRNA-Seq datasets is computationally challenging and requires specialized bioinformatics and programming skills (Hoffmann *et al.*, 2018).

Here, we present tRNAstudio, an integrative pipeline to analyze human tRNA-Seq datasets that is packaged into a user-friendly graphical user interface (GUI) implemented in local mode. Using publicly available datasets, we show that tRNAstudio can extract information on tRNA expression, processing and post-transcriptional modification status. The pipeline output includes an interactive html summary report, extensive graphical data representations, and spreadsheets useful for custom analyses.

2 Description and implementation

tRNAstudio is implemented as a GUI (Fig. 1), built with the Python library Tkinter, prepared to run in Mac (OS X El Capitan or higher) and Linux-based platforms, and designed for the analysis of human tRNAs using, as input, tRNA-Seq datasets generated by single- or paired-end sequencing. The code was primarily written in Python3 and R. tRNAstudio uses a Conda environment that includes the installation of R package, python modules and all the requirements and dependencies needed to perform tRNA analyses (Bowtie2, Samtools, Bedtools, Pysam, Pysamstats and Picard). To run tRNAstudio, the user will need to be familiar with a command-line interface and simple bash commands. The installation of Conda and the creation of the environment is executed by running the requirements script ('bash requirements.sh'). To activate the Conda environment and to launch the GUI the user needs to run only two

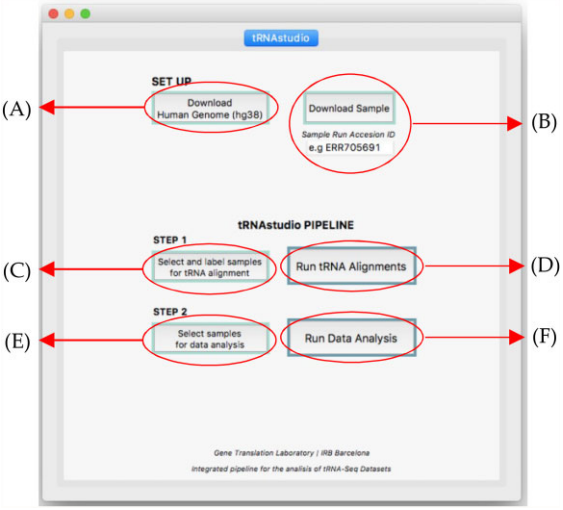

Fig. 1. tRNAstudio GUI visualized in macOS systems

commands: 'conda activate tRNAstudioEnv' and 'python3 tRNAstudioGUI', respectively. Detailed methodological descriptions of tRNAstudio are available in Supplementary Methods. tRNAstudio can be run in standard computers but we recommend at least 8 cores, 16 Gb of RAM and 100 Gb of available ROM. Under these conditions, four samples (around 10 Gb of information per sample) can be analyzed in 2–3 h.

Processing of tRNA-Seq datasets is achieved in six simple steps. The first time tRNAstudio is implemented, the user will download the reference Human Genome hg38 (Fig. 1A). Next, samples (datasets: Fastq files) of interest are directly downloaded from the Gene Expression Omnibus public repository (https://www.ncbi.nlm.nih.gov/gds/) by providing the run accession ID (prefixes SRR . . ., ERR . . ., DRR . . .) (Fig. 1B). Users can also analyze manually downloaded datasets from other repositories, or their own datasets, upon incorporating the desired Fastq files into the 'Fastq_downloaded' folder of tRNAstudio. Samples that are to be aligned against the reference genomes are selected and labeled in a metadata file that is required for comparative analyses among samples (Fig. 1C). Labeling information includes the group to where the sample belongs to (e.g. 'Control' samples and 'Treated' samples) and whether the selected datasets have been generated by single- or paired-end sequencing. The alignment pipeline is then executed (Fig. 1D) and a pop-up dialogue will inform when the alignments are done. The user can choose which of the aligned samples will be used for data analysis (Fig. 1E). Previously aligned samples can also be selected. Last, the data analysis is performed (Fig. 1F), results are saved, and an html summary report is generated to help the user interpret the data. tRNAstudio and detailed instructions of use can be found at https://github.com/GeneTranslationLab-IRB/tRNAstudio.

3 Results

tRNAstudio performs serial alignments against the whole human genome and against custom genomes as depicted in Supplementary Figure S1. Custom genomes are built with all unique human tRNA sequences, either in the form of pre-tRNAs (genomic sequences with 5′- and 3′-flanking regions; only applicable for nuclear-encoded tRNAs) or mature tRNAs (tRNA sequences without flanking and intronic regions, with 3′-CCA trinucleotide addition, and, in the case of $tRNA^{His}$, with a 5′-G addition; applicable to nuclear- and mitochondrial-encoded tRNAs). To aid in the identification of tRNA genes, we provide a file that links each tRNA gene analyzed by tRNAstudio using hg38 to its corresponding gene in hg19, its tRNA 'license plate' (Pliatsika *et al.*, 2016), and a hyperlink to additional gene expression information from MINTbase (Pliatsika *et al.*, 2016; Supplementary Table S1).

Datasets are first aligned against the whole human genome. Reads mapping to non-tRNA genes are discarded and reads mapping to tRNA genes are classified as 'mitochondrial' (if mapped to mitochondrial-encoded tRNA genes) or 'cytosolic' (if mapped to nuclear-encoded tRNA genes). Given the polycistronic nature of mitochondrial transcripts (Ojala *et al.*, 1981), reads classified as derived from mitochondrial tRNAs (mt-tRNA) are then aligned against a custom mature mt-tRNA genome to remove unprocessed mitochondrial transcripts that may partially overlap with mt-tRNA genes. Remaining mapped reads are kept for further analyses. Reads classified as derived from cytosolic tRNAs, and unmapped reads resulting from the initial mapping against the whole genome undergo serial alignments against custom tRNA genomes as described in Supplementary Methods (Supplementary Fig. S1). As a virtue of these serial alignment strategy, tRNAstudio improves both the recovery of reads mapped to nuclear-encoded tRNA genes and the alignment quality of the reads when compared against alignment strategies that use single reference genomes (Supplementary Fig. S2A). Users of tRNAstudio obtain absolute read counts for every nuclear- and mitochondrial-encoded tRNA genes, and their corresponding mapping quality score (MAPQ; Supplementary Methods and Fig. S2B). Of note, tRNAstudio considers all mapped tRNA reads for analysis, regardless of their MAPQ or whether they are derived from tRNAs with or without natural post-transcriptional nucleotide additions (i.e. $tRNA^{His}$ 5′-G or partial or full 3′-CCA) (further details in Supplementary Methods).

Mapped reads are then used for differential tRNA gene expression analyses using two complementary methods: DESeq2 (Love *et al.*, 2014) and iso-tRNA-CP (Torres *et al.*, 2019). Iso-tRNA-CP evaluates the proportional contribution of each tRNA gene to its corresponding isodecoder tRNA set (i.e. individual analyses among all genes having the same tRNA anticodon sequence). Given that mt-tRNA genes are represented by a single isodecoder gene (Juhling *et al.*, 2009), iso-tRNA-CP is only applicable to nuclear-encoded (i.e. cytosolic) tRNAs. Results are accompanied by a principal component analysis and can be visualized through heatmaps and interactive graphs and tables (Supplementary Fig. S3).

tRNAstudio also classifies reads derived from cytosolic pre-tRNAs or processed tRNAs. We validated this function by analyzing datasets enriched in reads derived from pre-tRNAs (Torres *et al.*, 2015) or mature tRNAs (Zheng *et al.*, 2015; Supplementary Fig. S4). The custom tool for the trimming of soft-clipped bases implemented by tRNAstudio aids in the correct assignment of reads to each category, as it specifically detects reads bearing post-transcriptional 3′-CCA additions, or $tRNA^{His}$ 5′-G addition. These modifications are present on processed tRNAs but may otherwise be confused as nucleotides derived from pre-tRNA 3′-trailer or 5′-leader sequences, respectively (further details in Supplementary Methods). We benchmarked this function against a standard tool for trimming soft-clipped bases (Biostar84452 from JVarkit). We find that both 3′-CCA and 5′-G additions are removed from the reads when using JVarkit but are retained when using tRNAstudio's customized tool (Supplementary Fig. S5A and B). Furthermore, tRNAstudio classifies reads as 'likely derived from pre-tRNAs' or 'likely derived from mature tRNAs' (i.e. processed tRNAs), based on the genomic coordinates of the mapped tRNA reads and on the presence or absence of post-transcriptional nucleotide additions (see Supplementary Methods). Using datasets enriched in reads derived from mature tRNAs (Zheng *et al.*, 2015), we find that tRNAstudio assigns 99.3% of tRNA reads to the processed tRNA set when applying its custom classification strategy, while only 68.5% of tRNA reads are classified into this group when using a standard genomic coordinates-based classification method (Supplementary Fig. S5C).

tRNAstudio uses a base-calling function to evaluate tRNA modification levels. Reverse transcriptases generate mutations in the obtained cDNA (and hence, in their derived sequencing reads) when encountering modified tRNA bases. Analyses of datasets with tRNAstudio revealed sequence variations that coincide with tRNA positions known to undergo post-transcriptional modifications such as positions 9 (m^1G), 26 (m^2_2G), 32 (m^3C), 34 (I: inosine), 37 (m^1I) and 58 (m^1A: 1-methyladenosine; de Crecy-Lagard *et al.*, 2019;

Supplementary Fig. S6). Furthermore, analyses of datasets derived from human cell lines depleted of ADAT2, the catalytic subunit of the enzyme that catalyzes A-to-I conversion at positions 34 of tRNAs (Torres *et al.*, 2015), revealed a quantitative decrease in the modification ratio at these positions without changes in the modification ratio of unrelated positions such as 58 (m^1A) (Supplementary Fig. S7A). Likewise, we detected a specific decrease in the modification ratio at positions 58 (m^1A) when analyzing datasets derived from artificially demethylated RNAs (Zheng *et al.*, 2015), without alterations in the modification ratio at positions 34 (I) (Supplementary Fig. S7B). Interactive heatmaps aid in visualizing global base-calling results for each position in every tRNA gene, and evaluating changes in modification ratios at specific tRNA positions when samples are compared (see Supplementary Discussion and Fig. S8).

tRNAstudio also reports on tRNA gene sequence coverages, which can aid in the identification of *bona fide* tRFs. tRFs derived from the 3′-arm of $tRNA^{Arg}_{CCU}$ and $tRNA^{Arg}_{UCG}$, and from the 5′-arm of $tRNA^{Cys}_{GCA}$ were shown to be abundant in human brain (Torres *et al.*, 2019). Using tRNAstudio, we analyzed datasets from human brain and found that tRNA sequence coverages mapped to the abovementioned tRFs (Supplementary Fig. S9).

We show that tRNAstudio can extract biologically relevant information from tRNA-Seq datasets, while allowing the analyses to be performed in local mode and with a user-friendly GUI. This work brings bioinformatics closer to experimental laboratories and will be useful to accelerate the pace at which knowledge on canonical and non-canonical tRNA biology expands (see Supplementary Discussion).

Acknowledgements

We thank Oscar Reina from the Biostatistics and Bioinformatics Core Facility at IRB Barcelona for technical assistance and helpful discussions.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness [PID2019-108037RB-100 to L.R.d.P]. M.M.-R. is funded by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) [2021 FI_B 01053].

Conflict of Interest: none declared.

References

Behrens,A. *et al.* (2021) High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Mol. Cell*, **81**, 1802–1815.e1807.

Cozen,A.E. *et al.* (2015) ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat. Methods*, **12**, 879–884.

de Crecy-Lagard,V. *et al.* (2019) Matching tRNA modifications in humans to their known and predicted enzymes. *Nucleic Acids Res.*, **47**, 2143–2159.

Hoffmann,A. *et al.* (2018) Accurate mapping of tRNA reads. *Bioinformatics*, **34**, 1116–1124.

Hu,J.F. *et al.* (2021) Quantitative mapping of the cellular small RNA landscape with AQRNA-seq. *Nat. Biotechnol.*, **39**, 978–988.

Juhling,F. *et al.* (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Ojala,D. *et al.* (1981) tRNA punctuation model of RNA processing in human mitochondria. *Nature*, **290**, 470–474.

Pliatsika,V. *et al.* (2016) MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments. *Bioinformatics*, **32**, 2481–2489.

Shigematsu,M. *et al.* (2017) YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs. *Nucleic Acids Res.*, **45**, e70.

Su,Z. *et al.* (2020) Noncanonical roles of tRNAs: tRNA fragments and beyond. *Annu. Rev. Genet.*, **54**, 47–69.

Torres,A.G. *et al.* (2015) Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Res.*, **43**, 5145–5157.

Torres,A.G. *et al.* (2019) Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. *Proc. Natl. Acad. Sci. USA*, **116**, 8451–8456.

Zheng,G. *et al.* (2015) Efficient and quantitative high-throughput tRNA sequencing. *Nat. Methods*, **12**, 835–837.