1-5-2024

MINRbase: A Comprehensive Database of Nuclear- and Mitochondrial-Ribosomal-RNA-Derived Fragments (rRFs)

Venetia Pliatsika

Tess Cherlin

Phillipe Loher

Panagiotis Vlantis

Parth Nagarkar

See next page for additional authors

Authors

Venetia Pliatsika, Tess Cherlin, Phillipe Loher, Panagiotis Vlantis, Parth Nagarkar, Stepan Nersisyan, and Isidore Rigoutsos

Nucleic Acids Research, 2024, **52**, D229–D238
https://doi.org/10.1093/nar/gkad833
Advance access publication date: 16 October 2023
Database issue

MINRbase: a comprehensive database of nuclear- and mitochondrial-ribosomal-RNA-derived fragments (rRFs)

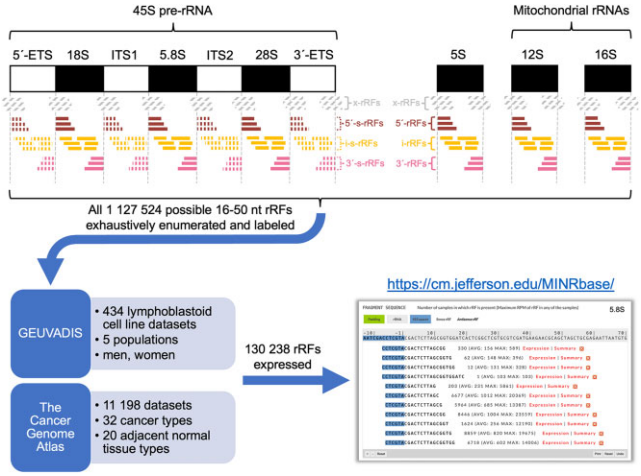
Venetia Pliatsika, Tess Cherlin, Phillipe Loher, Panagiotis Vlantis, Parth Nagarkar, Stepan Nersisyan and Isidore Rigoutsos*

Computational Medicine Center, Thomas Jefferson University, Philadelphia, PA 19107, USA
*To whom correspondence should be addressed. Tel: +1 215 503 6152; Fax: +1 215 503 0466; Email: Isidore.Rigoutsos@jefferson.edu

Abstract

We describe the Mitochondrial and Nuclear rRNA fragment database (MINRbase), a knowledge repository aimed at facilitating the study of ribosomal RNA-derived fragments (rRFs). MINRbase provides interactive access to the profiles of 130 238 expressed rRFs arising from the four human nuclear rRNAs (18S, 5.8S, 28S, 5S), two mitochondrial rRNAs (12S, 16S) or four spacers of 45S pre-rRNA. We compiled these profiles by analyzing 11 632 datasets, including the GEUVADIS and The Cancer Genome Atlas (TCGA) repositories. MINRbase offers a user-friendly interface that lets researchers issue complex queries based on one or more criteria, such as parental rRNA identity, nucleotide sequence, rRF minimum abundance and metadata keywords (e.g. tissue type, disease). A 'summary' page for each rRF provides a granular breakdown of its expression by tissue type, disease, sex, ancestry and other variables; it also allows users to create publication-ready plots at the click of a button. MINRbase has already allowed us to generate support for three novel observations: the internal spacers of 45S are prolific producers of abundant rRFs; many abundant rRFs straddle the known boundaries of rRNAs; rRF production is regimented and depends on 'personal attributes' (sex, ancestry) and 'context' (tissue type, tissue state, disease). MINRbase is available at https://cm.jefferson.edu/MINRbase/.

Graphical abstract



Introduction

Ribosomal RNAs (rRNAs) are non-coding RNAs with highly conserved secondary structures. They are a cell's most abundant transcripts (1,2) and an indispensable component of the ribosome (3). There are six human rRNAs: four (18S, 5.8S, 28S and 5S) are encoded in the nuclear genome (2,4), and two (12S and 16S) are encoded in the mitochondrial (MT) genome (5). Unlike 12S and 16S, the four nuclear-encoded rRNAs have hundreds of genomic copies. Three (18S, 5.8S and 28S) of the four nuclear-encoded rRNAs are transcribed by RNA polymerase I as part of a longer cassette, the 45S pre-RNA. The cassette also includes four 'spacers': 5′-ETS (also known as ETS1), ITS1, ITS2 and 3′-ETS (also known as ETS2). The fourth nuclear-encoded RNA, 5S, has most of its genomic copies on chromosome 1 (1,6) and is transcribed by RNA polymerase III (7).

During the last decade, analyses of small RNA-seq datasets from humans and other organisms revealed diverse populations of abundant molecules that map to rRNA sequences (8–10). These molecules are referred to as 'rRNA-derived fragments' or 'rRFs' for short. Initially, rRFs were dismissed as likely degradation products. However, research into them

Received: August 13, 2023. **Revised:** September 17, 2023. **Editorial Decision:** September 20, 2023. **Accepted:** September 20, 2023

accelerated in recent years, and now there is strong support for their functional importance; see, e.g. the excellent review in (10). This includes evidence that at least some of them act like miRNAs (10,11), in analogy with tRNA-derived fragments (tRFs) that are derived from the 5′ and 3′ regions of tRNAs (12,13).

Our own studies showed that 'similar' samples, both cell lines and tissues, produce very similar rRF profiles and at similar levels of abundance, strongly suggesting that their production is regimented (14). Moreover, we showed that the identity and levels of human rRFs also depend on the donor's sex and ancestry ('personal attributes'), as well as tissue type and disease type ('context') (14). Notably, these rRF dependencies mirror precisely the dependencies we previously reported for two other types of small RNAs, the miRNA isoforms (isomiRs) and the tRFs (15–18).

The Mitochondrial and Nuclear rRNA fragment database (MINRbase) is the first step towards developing a framework to facilitate the study of rRFs and the design of experiments aimed at understanding their biogenesis and function. MINRbase borrows several conceptual, interface and visualization strategies from MINTbase, our previously published state-of-the-art database that catalogs tRFs (19,20). MINRbase offers easy access to results from our systematic analyses of numerous samples from healthy donors and patients, multiple tissue types and many diseases. Central to our design decisions was our intent to provide a comprehensive and intuitive resource that can be used by researchers with little or no experience in computational biology or bioinformatics.

Materials and methods

Deterministic and exhaustive profiling of rRFs in small RNA-seq datasets

We used as reference rRNAs the sequences contained in the following GenBank records: NR_145819.1 (45S), NR_023374.1 (5S), NR_137294.1 (12S) and NR_137295.1 (16S). To identify rRFs that arise from these reference RNAs, we mined small RNA-seq datasets using the deterministic and exhaustive procedure we described previously (14). Briefly, we first flanked each reference rRNA sequence (including the spacer regions) with 49 nucleotides (nts) on each side to allow the possibility of rRFs straddling the known boundaries of rRNAs. We then enumerated all possible 16–50 nt (inclusive) segments and their reverse-complement sequences and exhaustively searched each segment in the human genome while recording all the locations where each segment matched exactly. We did not include fragments below 16 nts since they can occur with high probability in genomic regions that are unrelated to rRNAs (14,21). We also did not include fragments longer than 50 nts since we are unaware of any public reports of their existence. No insertions, deletions or replacements were permitted during this search – thus, all the rRFs could be *exactly* mapped to the parental rRNA sequences.

Labeling of rRFs

We originally introduced 'license plates' as a naming nomenclature for tRFs (19,20). Because license plates are a flexible labeling system with many desirable features, we have since extended it to isomiRs (22) and rRFs (14). License plates do not require a dedicated broker to assign unique labels to each small RNA, as is the case with miRNAs. Researchers who discover new and important small RNAs can easily assign a unique label to the respective sequences themselves. Importantly, the label is derived from the sequence of the small RNA and is thus unchangeable over time, genome-assembly agnostic and reversible, i.e. each small RNA can be assigned a single label and vice versa. Since there is a unique correspondence between a nucleotide sequence and a license plate, different researchers will always assign the same labels to the same RNA, avoiding redundant labels and confusion. The system is automatable and does *not* require more resources as the number of small RNAs increases. For example, the unique license plate for GGGCTACGCCTGTCTGAGCGTCGC, a 24 nt i-rRF from 5.8S, is rRF-24-RIE1975HJM ('rRF' is a prefix indicating the type of the molecule, '24' is the length of the molecule, and 'RIE1975HJM' is a base-32 encoding of the rRF's sequence). Of note, license plates have already been adopted by the tRFTar (23) and tRFtarget (24) databases and the mirGFF3 proposed standard (25). To convert nucleotide sequences to license plates and vice versa, researchers can use our web application (https://cm.jefferson.edu/LicensePlates/) or download and use locally standalone codes (https://cm.jefferson.edu/license-plates-download/).

Normalized abundances

MINRbase uses normalized abundances and reports rRF levels in Reads Per Million (RPM), i.e. an rRF's RPM abundance in a given dataset equals the number of reads that support it in this sample divided by the total number of reads that 'survive' after quality trimming and adapter removal, times one million. Using the number of surviving reads (as opposed to the total number of reads mapped to the rRFs) in the denominator implicitly accounts for the fact that rRFs are one of several small RNA classes that co-exist in a given sample; thus, it prevents overestimating rRFs' abundances.

Processing of the GEUVADIS and TCGA datasets

GEUVADIS datasets: The GEUVADIS Consortium generated RNA-seq datasets for a subset of the samples that had been collected by the 1000 Genomes Project (1KGP) (26). The source of the GEUVADIS datasets were the lymphoblastoid cell lines (LCLs) of participants to the 1KGP belonging to five (*genetically-determined*) populations: CEU—Utah Residents with Northern and Western European Ancestry, FIN—Finnish in Finland, GBR—British from England and Scotland, TSI—Toscani in Italia, YRI—members of the Yoruba tribe from the city of Ibadan, Nigeria. Both sexes are equally represented in each population. We used cutadapt (27) (options: -q 15 -e 0.12 -m 15 -a TGGAATTCTCGGGTGCCAAGG - -match-read-wildcards - -discard-untrimmed) to quality trim reads and remove adapters from the raw FASTQ data. Since quality trimming is applied before adapter removal, and reads without adapters are discarded, the resulting rRFs' 3′ ends are annotated with high confidence. The GEUVADIS collection contained a group of samples that we excluded from the downstream analysis. It comprised 48 datasets generated in laboratory 'number six' that used more than the 36 cycles of sequencing used by the other facilities. This left 434 samples that were sequenced at 36 cycles; after adapter removal, the longest sequence had a length of 33 nucleotides.

TCGA datasets: A total of 11 198 pre-trimmed TCGA datasets were downloaded on 16 October 2015 from TCGA's Cancer Genomic Hub. Each dataset is labeled with tissue

name and tissue type (tumor-adjacent normal, primary tumor, recurrent tumor, new primary tumor, metastatic). A total of 32 cancer types and 20 tumor-adjacent normal tissue types are included. The additional metadata include the patient's sex, age and self-reported ancestry. In contrast to the GEUVADIS project, in TCGA the sequencing stage was capped at 30 cycles; as a result, in the TCGA datasets of MINRbase all rRFs are at most 30 nts long.

Data visualization

We implemented the interactive plots of the 'rRNA alignment' and 'summary' vistas of MINRbase using the Highcharts library for JavaScript (https://www.highcharts.com/).

Results

Overview: rRF types, labels and summary statistics

MINRbase contains information about rRFs produced from the four nuclear rRNAs (18S, 5.8S, 28S, 5S), the two MT rRNAs (12S, 16S) and the four spacers of the 45S pre-RNA (Figure 1). In addition, we also analyzed and entered in MINRbase rRFs that either straddle the known boundaries of these six rRNAs or are fully contained in the spacers of 45S. MINRbase recognizes seven types of rRFs analogously to the notation MINTbase uses with tRFs (14,19,28); the seven types include: (i) 5′-rRFs: these are rRFs whose 5′ end matches the 5′ end of the parental rRNA; (ii) 3′-rRFs: these are rRFs whose 3′ end matches the 3′ end of the parental rRNA; (iii) i-rRFs ('internal' rRFs): these rRFs lie wholly inside the parental rRNA and are neither 5′-rRFs nor 3′-rRFs; (iv): 5′-s-rRFs: these are rRFs whose 5′ end matches the 5′ end of a 45S spacer; (v) 3′-s-rRFs: these are rRFs whose 3′ end matches the 3′ end of a 45S spacer; (vi) i-s-rRFs: these rRFs lie wholly inside a spacer and are neither 5′-s-rRFs nor 3′-s-rRFs and (vii) x-rRFs ('crossing' rRFs): these are rRFs that straddle an rRNA and either an adjacent spacer or a genomic region other than a spacer.

To label the rRFs of MINRbase, we used 'license plates' (20)—see Methods. As mentioned above, license plates are a portable and flexible labeling scheme that is not affected by genomic assembly updates, or the release of multiple genomes for the same organism (e.g. human pangenome (29)). The scheme uses only the small RNA's nucleotide sequence to derive a *unique* label. Because the label is unique, license plates are 'reversible', i.e. each sequence maps to a single label and *vice versa*. We originally introduced this scheme several years ago to label tRFs (19) and have since extended it to isomiRs (22) and rRFs (14).

Across the 434 GEUVADIS and 11 198 TCGA datasets, we find 130 238 rRFs with an abundance ≥1 RPM in at least one dataset. They comprise: 105 5′-rRFs; 100 3′-rRFs; 117 007 i-rRFs; 31 5′-s-rRFs; 12 3′-s-rRFs; 11 712 i-s-rRFs; 1340 x-rRFs (the numbers include 68 rRFs that belong to more than one category). There are 61 379 rRFs that satisfy a 5 RPM threshold and 40 386 that satisfy a 10 RPM threshold. In anticipation of expanding MINRbase by including datasets from additional projects, we enumerated and stored all 1 127 524 possible rRFs (sense and antisense) that can be formed from the six reference rRNAs and four spacers and have lengths 16–50 nts. MINRbase can report only those rRFs that satisfy a user-specified minimum level of abundance; if not provided, this value defaults to 1 RPM.

Overview: the graphical user interface

Similar to MINTbase (19), MINRbase also provides access to the rRF information through five 'vistas' that include: 'genomic loci', 'RNA molecule', 'rRNA alignment', 'expression' and 'summary'. The vistas present different types of information for all the rRFs that satisfy the user's query parameters. All the vistas share the same input query form (center of Figure 2). The user can specify any valid combination of one or more of the following features: rRF type, parental rRNA, nucleotide sequence, rRF license plate, tissue type, minimum abundance or metadata keyword (see below). For example, the user can restrict focus to 5′-rRFs and i-rRFs. The user can further restrict the focus by sub-selecting a combination of one or more among the six reference rRNAs and four spacers. Moreover, the user could enforce a minimum abundance constraint for the previous parameter combination. Alternatively, the user can leave the rRF type and parental rRNA identity unspecified altogether and restrict the minimum abundance instead.

The 'Search by metadata' window allows the optional use of keywords to limit the rRF search to only a subset of the MINRbase datasets. For example, the metadata keyword 'breast' will only report rRFs from datasets obtained from breast tissues. The keyword can also be a compound word such as 'TCGA-UVM' in which case the search will be limited to only the TCGA datasets from uveal melanoma patients. If a keyword is misspelled, the input form will provide alternative suggestions from which the user can choose.

In the above examples, we described searches that sought *all* rRFs satisfying the user-specified criteria. However, there are instances where one wishes to restrict the search to a single rRF. This can be done by providing the rRF's nucleotide sequence or its license plate. As in the case of mistyped keywords, when an rRF is mistyped the input form will suggest one or more candidates with the proposed insertions, deletions and replacements shown.

Using the 'genomic loci' and 'RNA molecule' vistas

For rRFs that satisfy the user query, these two vistas allow retrieving the start/end coordinates of the rRFs within the parental rRNA and the list of their parental rRNAs, respectively. Both outputs are presented in the form of a table whose contents can be sorted either lexicographically or numerically by clicking on a column's header. The user can select the number of rows included in each page of output through the 'Results/Page' drop-down menu. All the output tables could be downloaded from MINRbase. The top-left panel of Figure 2 shows the first two rows of the 'genomic loci' vista output for the default user query (all rRFs with minimum abundance ≥ 1 RPM); rows are sorted by the 'Maximum RPM' column (highlighted in blue). 'Genomic loci' and 'RNA molecule' vistas represent similar information for most rRFs, with rRFs of ambiguous origin comprising the exceptions. For example, rRF-17-QHLVQHP can originate from different parental molecules: 5′-ETS, 28S and 3′-ETS. For this particular rRF, the 'genomic loci' vista will report three rows, one for each parental RNA, whereas the 'RNA molecule' vista will only report one row with the rRF sequence and the number of potential parental locations.

Using the 'expression' vista

This vista lists all rRFs that satisfy the user query together with their expression levels and the available metadata for the

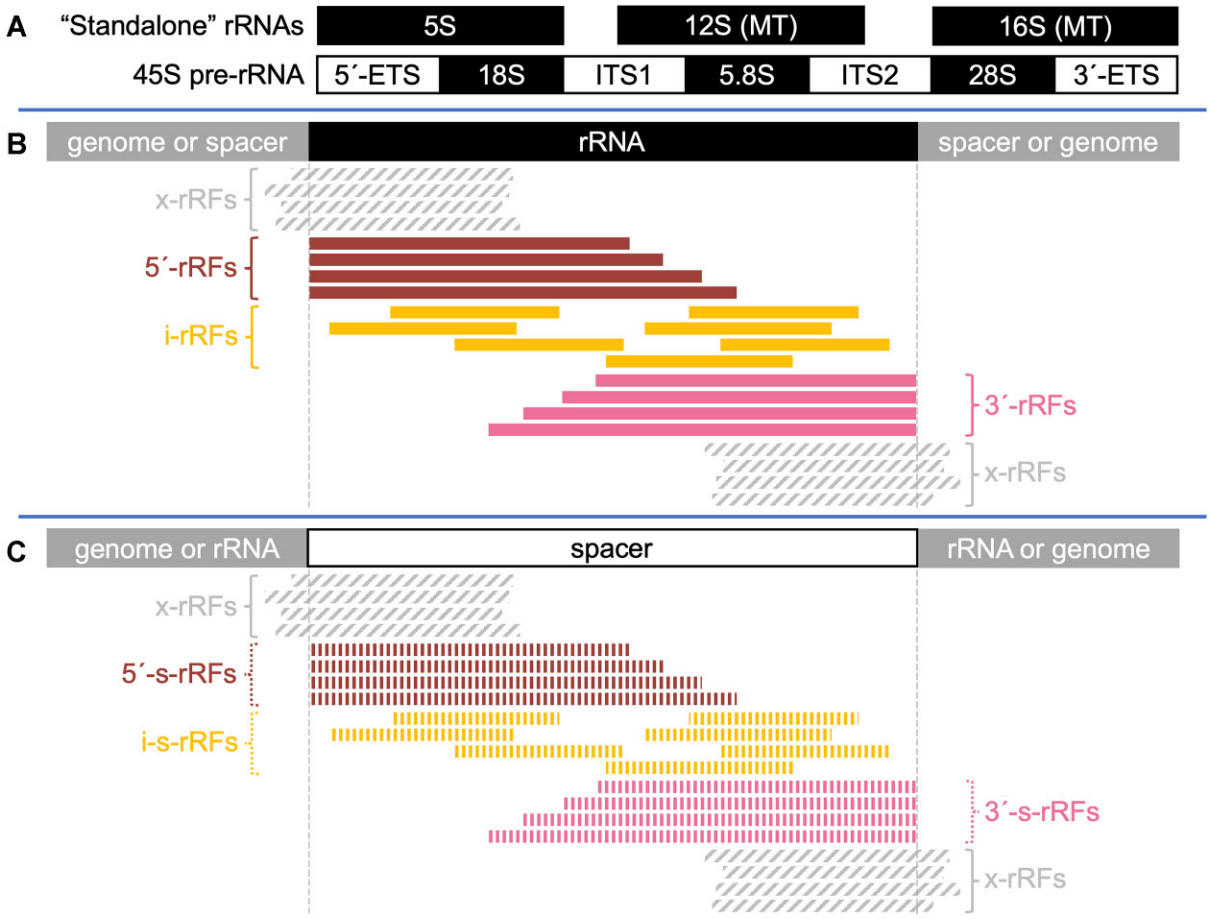


Figure 1. Human rRNAs and rRF types. (**A**) The six human rRNAs: three 'standalone' rRNAs and 45S pre-rRNA composed of three rRNAs and four spacers. (**B**, **C**) Types of rRFs produced by mature rRNAs (**B**) and 45S spacers (**C**).

datasets that contain them. The metadata currently include tissue, tissue type (e.g. normal, primary tumor, metastatic), cancer type, patient's sex, race and age for TCGA datasets and donor's sex and population for GEUVADIS datasets. The upper right panel of Figure 2 shows the 'expression' vista for the highly abundant rRF-18-087ISKDQ. The 'Run' column contains the official external identifiers of each dataset in TCGA and GEUVADIS repositories. The output panel includes several more columns that are not shown in this Figure: the dates each dataset was downloaded and processed, the name of the submitter and hyperlinks to the resources that deposited the raw sequencing data and metadata.

Using the 'summary' vista

This vista is rRF-centric and summarizes in one place the information about a queried rRF that we generated by analyzing the datasets in MINRbase. The reported information is very rich and includes: the identity of the parental rRNA(s); the number of datasets in the database where the rRF is present; and multiple summary plots showing the rRF's distribution and abundance levels by tissue name, tissue type (e.g. 'tumor-adjacent normal', 'cancer' or 'metastatic'), disease type, sex and ancestry (self-reported in the case of TCGA, genetically-determined in the case of GEUVADIS). The bottom panel of Figure 2 shows representative box plots for the rRF-18-087ISKDQ rRF in TCGA tumor-adjacent sample datasets. This page offers three important capabilities. First, if the user changes the minimum abundance requirement for the shown rRF through the provided drop-down menu, all results and plots shown on this page will be updated accordingly. This makes it easy to determine which samples, tissues, tissue types, sexes, ancestries, etc. express the rRF most or least abundantly. Second, all of the plots are interactive: by clicking on a bar in bar plots or a box in box plots, the user can see summary statistics and metadata about the underlying datasets in a pop-up window. In the case of box plots, the pop-up additionally includes minimum, median and maximum RPM values. Third, any of the shown plots can be individually exported as camera-ready images in several formats including JPG, PNG, PDF and SVG (see the dropdown menu in Figure 2).

Using the 'rRNA alignment' vista

This vista shows how the rRFs that satisfy the user query are spread along their parental rRNA(s). In cases where there are multiple parental RNAs, the vista prompts the user to first choose the parental rRNA before showing the alignment. Figure 3 shows two key components of the 'rRNA alignment' vista for 5.8S rRNA. Figure 3A shows the distribution of two of six possible measures across the 5.8S rRNA and its flanking spacers: the *number of rRFs containing the i-th nucleotide* (blue curve) and the *maximum RPM value among rRFs that contain the i-th nucleotide* (magenta curve). The vertical line pointed to by the cursor marks the first nucleotide of 5.8S

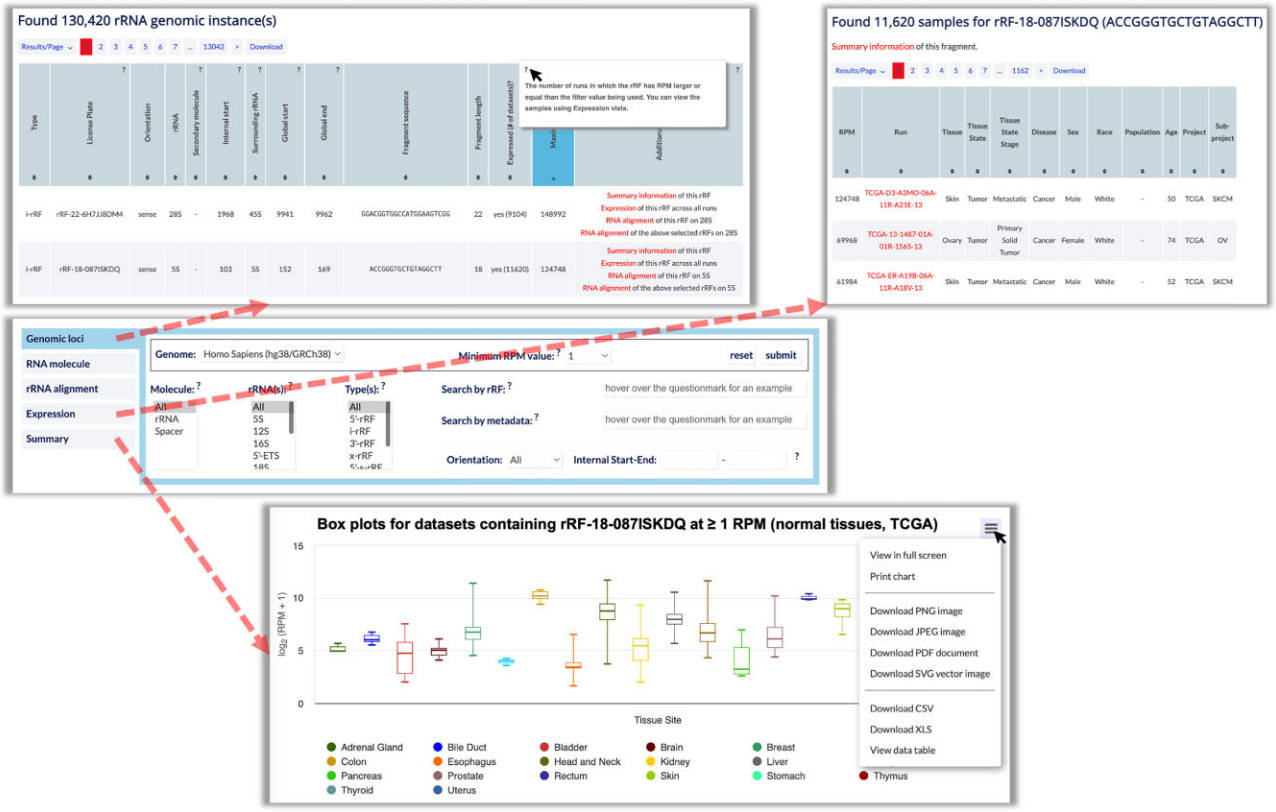


Figure 2. Input form, the 'genomic loci', 'expression' and 'summary' vistas of MINRbase. The output of the 'genomic loci' vista is sorted by maximum RPM (corresponding column name is highlighted in light blue). The tooltip with help in the 'genomic loci' panel is shown as a result of the cursor hovering over the question mark. The 'expression' vista is shown for the abundant rRF of 5S rRNA—the i-rRF rRF-18-087ISKDQ. The 'summary' vista shows the abundance of the same rRF in different human tissues. The pop-up at the top-right corner of the box plot allows the user to export the plot in multiple formats.

rRNA; the spacer ITS1 is immediately to its left. The other four measures, which are not shown here for clarity and whose curves the users can superimpose with the click of a button, include the *number of rRFs starting at the i-th nucleotide*, the *number of rRFs ending at the i-th nucleotide*, the *maximum number of datasets that express rRFs containing the i-th nucleotide* and the *average RPM*. The plot is interactive, and users can zoom in on any subregion of interest. Figure 3B includes a listing of all the rRFs that satisfy the user criteria aligned against the parental RNA. For rRFs that straddle the known rRNA boundaries (see Figure 1), their non-rRNA portion is shown in a different color to facilitate its identification: we use blue for adjacent spacers and green for adjacent genomic regions. Also, rRFs that are antisense to the parental sequence are shown in lowercase boldface italic letters. For each aligned rRF we also report the number of processed datasets that express it and the rRF's average and maximum abundances (in RPM) across the user-specified datasets, or all datasets if the user did not restrict the search. The user can also remove one or more rRFs from the alignment by clicking on the red '**x**' button located at the end of the corresponding line (removed rRFs can be restored by clicking 'Undo', see bottom-right corner of the panel).

Tabular outputs: 'genomic loci', 'RNA molecule', 'expression' and 'rRNA alignment' vistas

In these four vistas, the output pages are tabular. The results can be sorted in either lexicographic or numerical order by clicking on the respective column's header. This is particularly useful as it allows the user to perform multiple actions such as quickly identifying the sample with the highest or lowest expression of an rRF, the age of the youngest or oldest donor, and others.

Vista cross-linking and context-sensitive help

To help users explore the available information from various vantage points seamlessly without having to issue queries anew, we provide extensive cross-linking in the output pages of all the vistas. For example, the user can initiate the exploration through the 'alignment' vista of an rRNA and transition with one click to a specific rRF's 'summary' page (see Figure 3B). The input and output pages of all the vistas have built-in help that is accessible through the provided 'question mark' symbols. For example, hovering the cursor over question marks on top of the column names of the output pages opens a pop-up window with additional information about the column (see the 'genomic loci' vista in Figure 2).

An example use case: discovering that the spacers of 45S produce many abundant rRFs

The number of rRFs matching specific criteria is listed at the top of the output page of the 'RNA molecule' vista. For example, one can calculate the number of i-rRFs originating from 28S simply by selecting these options in the 'rRNA(s)' and 'Type(s)' input fields and clicking 'submit.' Users can also

Figure 3. The view on 5.8S rRNA in the 'rRNA alignment' vista. (**A**) Dependency of different rRF expression measures (Y axes) on the nucleotide position (X axis) in 5.8S rRNA. Only two variables are shown—the number of rRFs that contain a given nucleotide (blue line) and the maximum RPM value among all rRFs that contain a given nucleotide (magenta line). Hovering the cursor over the plot (in this case—over the first nucleotide of 5.8S) shows the exact values on the tooltip. (**B**) Alignment of x-RFs overlapping 5′ end of 5.8S and 3′ end of ITS1. The rectangle with a red dashed border highlights one of the most abundant rRFs.

Table 1. Summary statistics of rRFs production

	Parental molecule	5′-(s)-rRFs	i-(s)-rRFs	3′-(s)-rRFs	x-rRFs	All rRFs	Length (nt)	All rRFs/length
Nuclear rRNAs	5′-ETS	6	600	0	0	606	3654	0.17
	18S	16	8458	15	42	8531	1869	4.56
	ITS1	1	123	0	0	124	1077	0.12
	5.8S	18	937	17	142	1114	157	7.10
	ITS2	0	255	0	0	255	1167	0.22
	28S	18	23 185	4	88	23 295	5066	4.60
	3′-ETS	0	11	0	0	11	361	0.03
	5S	17	892	18	52	979	121	8.09
MT rRNAs	12S	1	1973	6	0	1980	954	2.08
	16S	18	3452	1	47	3518	1559	2.26

Only rRFs exceeding the threshold of 10 RPM in at least one dataset are included. Columns 2–4 represent numbers of distinct 5′-rRFs, i-rRFs and 3′-rRFs for mature rRNAs (18S, 5.8S, 28S, 5S, 12S, 16S) and numbers of distinct 5′-s-rRFs, i-s-rRFs and 3′-s-rRFs for spacers (5′-ETS, ITS1, ITS2, 3′-ETS).

Table 2. Per-unit-length rates of rRFs production at different abundance thresholds

	Parental molecule	Length (nt)	All rRFs/length ≥ 1 RPM	All rRFs/length ≥ 10 RPM	All rRFs/length ≥ 100 RPM
Nuclear rRNAs	5′-ETS	3654	1.88	0.17	0.02
	18S	1869	13.23	4.56	0.83
	ITS1	1077	1.79	0.12	0.01
	5.8S	**157**	**14.55**	**7.1**	**2.18**
	ITS2	1167	2.41	0.22	0.01
	28S	5066	12.87	4.6	0.89
	3′-ETS	361	0.42	0.03	0.01
	5S	**121**	**14.35**	**8.09**	**2.39**
MT rRNAs	12S	954	9.26	2.08	0.26
	16S	1559	10.1	2.26	0.30

For columns 3–5, only rRFs exceeding the threshold of 1/10/100 RPM, respectively, in at least one dataset are counted. Two rRNAs—5.8S and 5S—with outstandingly high rRF production rates at 10 and 100 RPM are marked with bold font.

download a table with the rRFs of a vista's output page and process it later in their favorite table editor. Using the latter approach, we generated Table 1 which lists the number of rRFs by type that are produced by each rRNA and spacer, and whose abundance is ≥10 RPM. Importantly, Table 1 shows that there are multiple abundant rRFs (5′-s-rRFs, 3′-s-rRFs and i-s-rRFs) that lie fully within the 45S spacers. 5′-ETS generates the majority of these spacer-derived fragments. We are unaware of previous systematic analyses that reported these spacer-derived rRFs.

On a related note, Table 2 shows the number of rRFs produced *per unit length* for three abundance thresholds: 1, 10 or 100 RPM. At a threshold of 1 RPM, the per-unit-length rRF production is comparable among the six mature rRNAs and approximately ten-fold higher than that of the four spacers. However, this balance is disrupted when the abundance threshold is raised to 10 RPM or 100 RPM: 5.8S and 5S (marked with bold font) stand out with drastically higher rRF production. These observations mirror our previously reported findings (14).

An example use case: discovering that many abundant rRFs straddle the known boundaries of rRNAs

Table 1 also shows that among the abundantly produced molecules, there are hundreds of x-rRFs whose endpoints are beyond the known boundaries of mature rRNAs. At a threshold of 10 RPM, there are 142 x-rRFs that straddle the 5′ and 3′ boundaries of 5.8S rRNA. Figure 3A, generated using the 'rRNA alignment' vista, shows how the 'maximum RPM' curve peaks near the junction of 5.8S rRNA (to the right of the vertical line pointed to by the cursor) and ITS1 (to the left of the vertical line); the peak values are as high as 50 000 RPM. Figure 3B shows in detail the alignment of the corresponding x-rRFs. Both panels of Figure 3 indicate the existence of many highly abundant x-rRFs that overlap with ITS1 by 1–6 nucleotides (these overlaps are colored in blue in Figure 3B). One of the most abundant of these rRFs, rRF-30-4WEDWXQ36MES (indicated by the red rectangle in Figure 3B), is expressed in as many as 11 170 samples with maximum and average RPM values of 50 582 and 321, respectively.

The cross-link to this rRF's page in the 'summary' vista allows the users to further investigate its distribution across healthy tissues (Figure 4A, B) and diseases (Figure 4C, D). The four panels of Figure 4 were compiled by setting the abundance threshold to 100 RPM. Figures 4A and 4B show the abundance of the x-rRF in tumor-adjacent normal tissues. Panel A shows the percentage of samples in each tissue type that express the rRF at ≥100 RPM. Panel B shows the respective distribution of RPM values using box plots. Similarly, Figures 4C and 4D show this x-rRF's expression patterns across different cancer types. As mentioned above, the plots of the 'summary' vista are interactive: note the pop-up windows at the tip of the cursor in each panel of Figure 4. Figure 4 makes evident that this x-rRF is expressed in a tissue- and disease-specific manner.

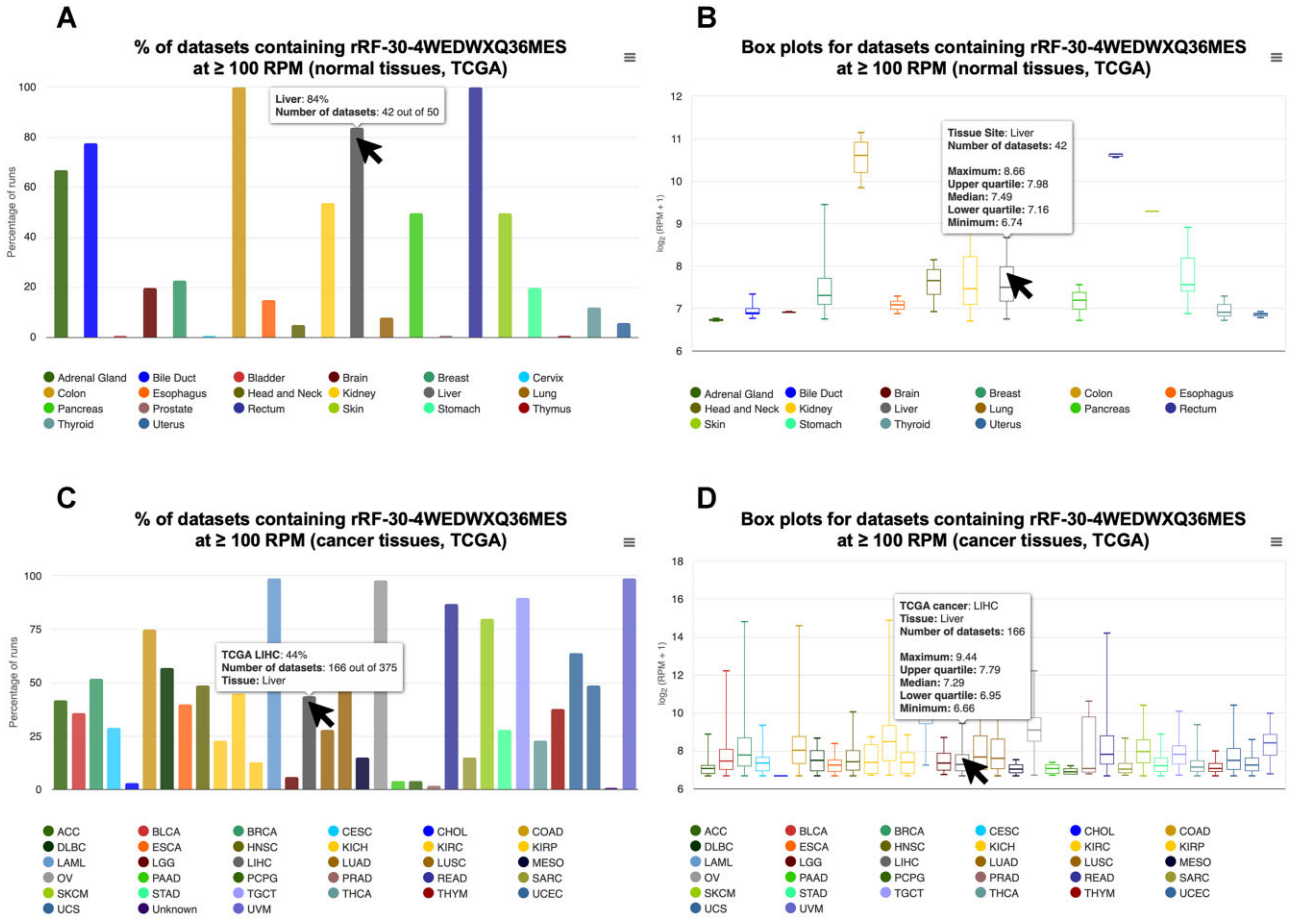


Figure 4. The abundance of rRF-30-4WEDWXQ36MES in human normal and cancer tissues. (**A**, **B**) normal tissues; (**C**, **D**) cancer tissues. Panels (**A**) and (**C**) show the percentage of samples in which the rRF exceeds the 100 RPM threshold; panels (**B**) and (**D**) show box plots for the samples with rRF's abundance over 100 RPM. Hovering the cursor over different parts of the bar/box plots shows a tooltip with exact values.

An example use case: discovering population-specific rRFs

In previous work, we discovered and validated multiple rRFs whose expression levels differ between men and women, and among human populations (14). In particular, we showed that 16S is a rich source of ancestry-specific rRFs; e.g. see Figures 2 and 4 of (14). MINRbase makes it easy to identify such rRFs and visualize population differences in their abundance. A simple approach would proceed as follows. First, we subselect '16S' in the 'rRNA?' drop-down menu. Then, we focus on the GEUVADIS subset of the 1000 Genomes Project (1KGP) by typing '1KGP' in the 'Search by metadata' window. This will restrict the search to samples with genetically-determined ancestries. Next, we enforce a minimum threshold of 500 RPM, select the 'RNA molecule' vista and click on the 'submit' button. As the output table is sorted by abundance by default, we select the first and most abundant rRF—the i-rRF rRF-27-BFVONB4725J (expressed in 265 datasets at ≥500 RPM, maximum abundance 9007 RPM). Clicking on the 'Summary information' link at the end of this rRF's row brings us to its 'summary' vista, a portion of which we reproduced in Figure 5. Figure 5A shows the percentage of the GEUVADIS datasets in which this rRF is present at an abundance of ≥500 RPM. Figure 5B shows the distribution of the rRF's abundance in each population and sex combination; note that the Y-axis here is logarithmic. As can be seen, this rRF is highly abundant in the four European populations (CEU, FIN, GBR, TSI) but not in the African population (YRI), and this holds true for both sexes.

Discussion

We developed MINRbase, a knowledge repository, to facilitate the study of rRFs that we mined from 11 632 public datasets. To the best of our knowledge, MINRbase represents the first attempt to systematically catalog human rRFs across diverse datasets. MINRbase is the companion to MINTbase (19), our previously released database of tRF expression profiles.

MINRbase provides interactive access to the profiles of 130 238 expressed rRFs that arise from the four nuclear rRNAs, two MT rRNAs or the four spacers of 45S. These rRFs are expressed at a level of ≥1 RPM in at least one of the analyzed datasets and are a subset of a theoretically possible collection of 1 127 524 rRFs that can arise from the six rRNAs and four spacers and have lengths between 16 and 50 nts. Almost one-half (61 379) of the expressed rRFs have abundance ≥5 RPM. Clearly, how many and which rRFs are expressed above a given abundance threshold depends on the nature and diversity of the datasets that populate MINRbase.

One of our design goals for MINRbase was for it to be user-friendly and intuitive to use. To this end, we relied on feedback

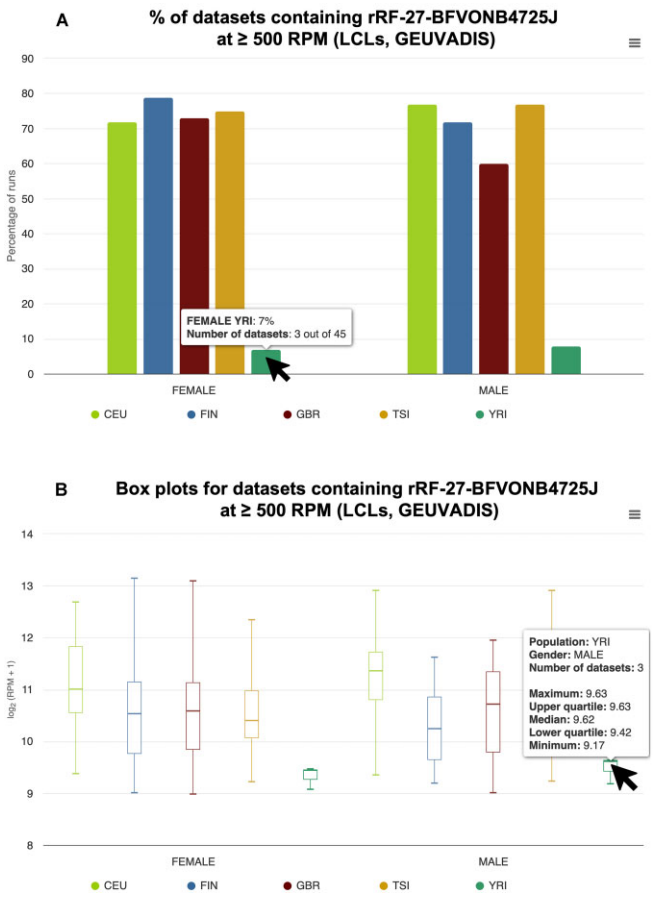


Figure 5. The population-specific abundance of rRF-27-BFVONB4725J. Panel (**A**) shows the percentage of LCL samples in which the rRF exceeds the 500 RPM threshold; panel (**B**) shows the box plot for the LCL samples with rRF's abundance over 500 RPM. Hovering the cursor over different parts of the bar/box plots shows a tooltip with exact values.

that we received from the users of MINTbase, our popular repository for studying tRFs. From a user-experience perspective, we note that even though MINRbase has to sift through much more information than MINTbase to respond to the user query, even complex searches are typically completed in 1–2 s. We were able to provide such very fast response times through an improved database design.

MINRbase provides large-scale support to earlier findings by us (14) and others (10,11) that the production of rRFs is regimented. Further supporting this view is our previous finding that the expression of rRFs depends on 'personal attributes' (sex, ancestry) and 'context' (tissue type, tissue state, disease type) (14), which mirrors our earlier similar findings on isomiRs (15,16,22,30,31) and tRFs (17,18,31). We expect that the number of functionally important rRFs will increase further as we augment MINRbase by including more datasets.

MINRbase helped us generate several new findings that we also reported in this presentation. First, we extended our previous work by reporting a new class of rRFs that arise from the four spacers of 45S. If their 5′ terminus coincided with the 5′ terminus of the spacer, we labeled them 5′-s-rRFs. If their 3′ terminus coincided with the 3′ terminus of the spacer, we labeled them 3′-s-rRFs. Lastly, if they were wholly internal to the spacer, we labeled them i-s-rRFs. These three new rRF types add to the four types we reported previously: 5′-rRFs, 3′-rRFs, i-rRFs and x-rRFs (14). Figure 1 shows the seven rRF types.

Our analyses also show that the four spacers are an unexpectedly rich source of rRFs, just like the 18S, 5.8S and 28S rRNAs. Moreover, these spacer-derived rRFs are very abundant. Previous work (11) found little evidence that these molecules enter the RNA interference pathway, suggesting a different mode of action that remains to be determined.

Another finding pertains to the discovery of x-rRFs that straddle the known boundaries of rRNAs. The x-rRFs may overlap a spacer or an adjacent genomic region. Again, we found an unexpectedly high number of abundant x-rRFs, most of which straddle the boundaries of 5.8S: 142 x-rRFs do so. The recurrent presence of specific x-rRFs across many diverse datasets suggests important roles that are currently unknown.

In summary, we expect that our systematic analyses of numerous (11 632) public datasets from 32 cancer types, 20 tumor-adjacent normal tissue types and lymphoblastoid cell lines obtained from healthy individuals will serve as a rich reference of which rRFs are expressed in which settings at which level of abundance. This information should help prioritize among the many rRFs that are produced in different settings: e.g. finding the most abundant rRFs in a tissue of interest or finding sex-/population-specific rRFs. Thus, it will help facilitate the design of targeted experiments aimed at determining the biogenesis and functions of these novel and very intriguing molecules.

We expect that future releases of MINRbase will provide the ability to generate and report rRFs that are differentially abundant between select combinations of groups of samples (e.g. tumor versus tumor-adjacent normal tissues for a specific cancer type). This task is not easily automatable because one needs to account for multiple documented dependencies and domain-specific considerations. These include batch effect corrections, partitioning the samples by cancer subtype or by tumor purity, excluding samples with evidence of degradation or infections, accounting for a patient's sex and ancestry, and interactions among these variables.

Data availability

MINRbase can be accessed freely at https://cm.jefferson.edu/MINRbase/.

Acknowledgements

Author contributions: V.P., P.L., I.R., T.C., P.N., P.V. and S.N. designed and implemented MINRbase. V.P. and P.L. mined all datasets for rRFs. S.N., I.R., P.L., V.P., T.C., P.N. and P.V. extensively tested MINRbase prior to its release. I.R. and S.N. wrote the manuscript with contributions from V.P. and P.L. All authors have read and approved the final manuscript. I.R. oversaw the study and development of MINRbase.

Funding

Thomas Jefferson University Funds; NIH/NHGRI [R01HG012784 to I.R.]. Funding for open access charge: Thomas Jefferson University Funds.

Conflict of interest statement

None declared.

References

1. Yu,S. and Lemos,B. (2016) A portrait of ribosomal DNA contacts with Hi-C reveals 5S and 45S rDNA anchoring points in the folded human genome. *Genome Biol. Evol.*, **8**, 3545–3558.
2. Moss,T., Langlois,F., Gagnon-Kugler,T. and Stefanovsky,V. (2007) A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis. *Cell. Mol. Life Sci.*, **64**, 29–49.
3. Wilson,D.N. and Doudna Cate,J.H. (2012) The structure and function of the eukaryotic ribosome. *Cold Spring Harb. Perspect. Biol.*, **4**, a011536.
4. Goffova,I. and Fajkus,J. (2021) The rDNA loci-intersections of replication, transcription, and repair pathways. *Int. J. Mol. Sci.*, **22**, 1302.
5. Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457–465.
6. Steffensen,D.M., Duffey,P. and Prensky,W. (1974) Localisation of 5S ribosomal RNA genes on human chromosome 1. *Nature*, **252**, 741–743.
7. Ciganda,M. and Williams,N. (2011) Eukaryotic 5S rRNA biogenesis. *Wiley Interdiscip. Rev. RNA*, **2**, 523–533.
8. Wei,H., Zhou,B., Zhang,F., Tu,Y., Hu,Y., Zhang,B. and Zhai,Q. (2013) Profiling and identification of small rDNA-derived RNAs and their potential biological functions. *PLoS One*, **8**, e56842.
9. Chen,Z., Sun,Y., Yang,X., Wu,Z., Guo,K., Niu,X., Wang,Q., Ruan,J., Bu,W. and Gao,S. (2017) Two featured series of rRNA-derived RNA fragments (rRFs) constitute a novel class of small RNAs. *PLoS One*, **12**, e0176458.
10. Lambert,M., Benmoussa,A. and Provost,P. (2019) Small non-coding RNAs derived from eukaryotic ribosomal RNA. *Noncoding RNA*, **5**, 16.
11. Guan,L. and Grigoriev,A. (2021) Computational meta-analysis of ribosomal RNA fragments: potential targets and interaction mechanisms. *Nucleic Acids Res.*, **49**, 4085–4103.
12. Kumar,P., Anaya,J., Mudunuri,S.B. and Dutta,A. (2014) 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**, 78.
13. Kuscu,C., Kumar,P., Kiran,M., Su,Z., Malik,A. and Dutta,A. (2018) tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA*, **24**, 1093–1105.
14. Cherlin,T., Magee,R., Jing,Y., Pliatsika,V., Loher,P. and Rigoutsos,I. (2020) Ribosomal RNA fragmentation into short RNAs (rRFs) is modulated in a sex- and population of origin-specific manner. *BMC Biol.*, **18**, 38.
15. Loher,P., Londin,E.R. and Rigoutsos,I. (2014) IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. *Oncotarget*, **5**, 8790–8802.
16. Telonis,A.G., Magee,R., Loher,P., Chervoneva,I., Londin,E. and Rigoutsos,I. (2017) Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res.*, **45**, 2973–2985.
17. Telonis,A.G., Loher,P., Honda,S., Jing,Y., Palazzo,J., Kirino,Y. and Rigoutsos,I. (2015) Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget*, **6**, 24797–24822.
18. Telonis,A.G., Loher,P., Magee,R., Pliatsika,V., Londin,E., Kirino,Y. and Rigoutsos,I. (2019) tRNA Fragments Show Intertwining with mRNAs of Specific Repeat Content and Have Links to Disparities. *Cancer Res.*, **79**, 3034–3049.
19. Pliatsika,V., Loher,P., Magee,R., Telonis,A.G., Londin,E., Shigematsu,M., Kirino,Y. and Rigoutsos,I. (2018) MINTbase v2.0: a comprehensive database for tRNA-derived fragments that includes nuclear and mitochondrial fragments from all The Cancer Genome Atlas projects. *Nucleic Acids Res.*, **46**, D152–D159.
20. Pliatsika,V., Loher,P., Telonis,A.G. and Rigoutsos,I. (2016) MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments. *Bioinformatics*, **32**, 2481–2489.
21. Telonis,A.G., Loher,P., Kirino,Y. and Rigoutsos,I. (2016) Consequential considerations when mapping tRNA fragments. *BMC Bioinf.*, **17**, 123.
22. Loher,P., Karathanasis,N., Londin,E., Bray,P., Pliatsika,V., Telonis,A.G. and Rigoutsos,I. (2021) IsoMiRmap-fast, deterministic, and exhaustive mining of isomiRs from short RNA-seq datasets. *Bioinformatics*, **37**, 1828–1838.
23. Zhou,Y., Peng,H., Cui,Q. and Zhou,Y. (2021) tRFTar: prediction of tRF-target gene interactions via systemic re-analysis of Argonaute CLIP-seq datasets. *Methods*, **187**, 57–67.
24. Li,N., Shan,N., Lu,L. and Wang,Z. (2021) tRFtarget: a database for transfer RNA-derived fragment targets. *Nucleic Acids Res.*, **49**, D254–D260.
25. Desvignes,T., Loher,P., Eilbeck,K., Ma,J., Urgese,G., Fromm,B., Sydes,J., Aparicio-Puerta,E., Barrera,V., Espin,R., *et al.* (2020) Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API. *Bioinformatics*, **36**, 698–703.
26. Lappalainen,T., Sammeth,M., Friedlander,M.R., t Hoen,P.A., Monlong,J., Rivas,M.A., Gonzalez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G., *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
27. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, **17**, 3.
28. Magee,R. and Rigoutsos,I. (2020) On the expanding roles of tRNA fragments in modulating cell behavior. *Nucleic Acids Res.*, **48**, 9433–9448.
29. Liao,W.W., Asri,M., Ebler,J., Doerr,D., Haukness,M., Hickey,G., Lu,S., Lucas,J.K., Monlong,J., Abel,H.J., *et al.* (2023) A draft human pangenome reference. *Nature*, **617**, 312–324.
30. Telonis,A.G., Loher,P., Jing,Y., Londin,E. and Rigoutsos,I. (2015) Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. *Nucleic Acids Res.*, **43**, 9158–9175.
31. Telonis,A.G. and Rigoutsos,I. (2018) Race disparities in the contribution of miRNA isoforms and tRNA-derived fragments to triple-negative breast cancer. *Cancer Res.*, **78**, 1140–1154.

Received: August 13, 2023. **Revised:** September 17, 2023. **Editorial Decision:** September 20, 2023. **Accepted:** September 20, 2023