

Dec 29, 2025

FAS2rDNA-Colab: A cloud-based workflow for pan-cancer, isoform-wide miRNome reconstitution across TCGA cohorts

DOI

<https://dx.doi.org/10.17504/protocols.io.14egn1xr6v5d/v1>

Marvin De los Santos¹

¹ChordexBio



Marvin De los Santos

ChordexBio

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.14egn1xr6v5d/v1>

External link: <https://fas2rdna.chordexbio.com/>

Protocol Citation: Marvin De los Santos 2025. FAS2rDNA-Colab: A cloud-based workflow for pan-cancer, isoform-wide miRNome reconstitution across TCGA cohorts. [protocols.io https://dx.doi.org/10.17504/protocols.io.14egn1xr6v5d/v1](https://dx.doi.org/10.17504/protocols.io.14egn1xr6v5d/v1)

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: December 29, 2025

Last Modified: December 29, 2025

Protocol Integer ID: 235975

Keywords: mirna sequences across multiple cancer cohort, wide mirnome reconstitution across tcga cohorts microrna, derived mirna expression data, mirna expression data, formatted mirna sequence, tcga cohorts microrna, mirna dataset, cancer genome atlas, scale resources such as the cancer genome atlas, tcga mirna dataset, wide mirnome sequences from tcga, wide mirnome sequence, standardized foundation for exploratory mirnome research, cdna sequence, biological variability across cancer type, raw sequencing reprocessing, exploratory mirnome research, resulting reconstructed mirnome, reconstructed mirnome, multiple cancer cohort, explicit nucleotide representation, genomic coordinate, using genomic coordinate, wide mirnome reconstitution, cancer type, resolved sequence output, sequence output

Disclaimer

Your usage of different FAS2rDNA versions and other components of the FAS2rDNA suite may be limited. Please refer to the license notice at <https://github.com/mahvin92/FAS2rDNA-Colab> for your review. Note that this protocol was tested using the GDAC-derived TCGA miRNA isoform expression data only. Additional validation may be required when applying this workflow to other studies or datasets. FAS2rDNA-Colab is provided for research use only and has not been validated for clinical or diagnostic use. The software is provided "as is", without warranty of any kind, express or implied. For more information, please visit <https://fas2rdna.chordexbio.com/>.

Abstract

MicroRNA (miRNA) sequence composition and isoform diversity play important roles in post-transcriptional regulation and contribute to biological variability across cancer types. Large-scale resources such as The Cancer Genome Atlas (TCGA) provide a standardized foundation for exploratory miRNome research; however, TCGA miRNA datasets are typically distributed as expression matrices without direct access to reconstructed, isoform-resolved sequence outputs. This limits the application of sequence-based analyses, including pan-cancer comparisons and machine learning workflows that require explicit nucleotide representations.

FAS2rDNA-Colab is a cloud-based workflow that reconstructs FASTA-formatted DNA/cDNA sequences using genomic coordinates/annotations. This protocol extends FAS2rDNA-Colab for the reconstitution of isoform-wide miRNome sequences from TCGA-derived miRNA expression data. By reconstituting FASTA-formatted miRNA sequences across multiple cancer cohorts, the protocol enables pan-cancer and isoform-level comparisons without reliance on predefined probe sets or raw sequencing reprocessing. The resulting reconstructed miRNomes can be used for sequence validation, exploratory comparative analyses, and downstream computational modeling.



Before start

For the purpose of this protocol, the GDAC Broad Institute database will be used to obtain molecular data of the different TCGA cohorts. However, users can use different databases to obtain the miR-Seq annotations of the cohorts, such as in UCSC Xena Browser or GDC Data Portal.

FAS2rDNA exists in two versions: 1) CLI for local implementation, and 2) Colab for cloud implementation using Google Colab. This protocol uses the FAS2rDNA-Colab. Alternatively, users may use the CLI version and modify the protocol here: <https://dx.doi.org/10.17504/protocols.io.rm7vzengxvx1/v1>



Protocol Procedure

- 1 Acquire or download the TCGA miR-Seq annotation data in the GDAC Broad Institute portal.
- 1.1 Navigate to <https://gdac.broadinstitute.org/>
- 1.2 Select the following cancer cohorts from the list:



	Disease Name	Cohort	Cases
	Breast invasive carcinoma	BRCA	1098
	Colon adenocarcinoma	COAD	460
	Colorectal adenocarcinoma	COADREAD	631
	Glioblastoma multiforme	GBM	613
	Glioma	GBMLGG	1129
	Head and Neck squamous cell carcinoma	HNSC	528
	Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973
	Kidney renal clear cell carcinoma	KIRC	537
	Brain Lower Grade Glioma	LGG	516
	Liver hepatocellular carcinoma	LIHC	377
	Lung adenocarcinoma	LUAD	585
	Lung squamous cell carcinoma	LUSC	504
	Ovarian serous cystadenocarcinoma	OV	602
	Prostate adenocarcinoma	PRAD	499



	Disease Name	Cohort	Cases
	ma		
	Skin Cutaneous Melanoma	SKCM	470
	Stomach adenocarcinoma	STAD	443
	Stomach and Esophageal carcinoma	STES	628

List of cohorts included in the miRNome sequence reconstitution using FAS2rDNA-Colab.

- 1.3 Agree to the TCGA guidelines on the responsible use of data to proceed.
- 1.4 From each individual Archive pop-up, download the *illuminahisec_mirnaseq-miR_isoform_expression* data from the 'miR-Seq' section.
- 1.5 After each download, inspect that all the data entries contain valid non-empty data, particularly in the following columns:

	A	B	C	D
	SampleId	miRNA_ID	isoform_c oords	read_coun t
	TCGA-05-4390-01A-02T-1754-13	hsa-let-7a-1	hg19:9:96938243-96938263: +	1

Sample miR-Seq annotation data, showing the critical data types.



1.6 Sort the *read_count* column and remove all rows with invalid or empty data (e.g., N, X, None, N/A). Note that if you are working with normalized expression count, zero (0) does not mean lack of detection but the expression is negligible from a reference count (e.g., normal tissue/cell).



2 Pre-process your data for FAS2rDNA reconstitution and miRNA sequence reconstruction.

2.1 Ensure that the *isoform_coords* of your data follow the FAS2rDNA-required format:

Standard format:
assembly:chromosome:start-end:strand

Example data:
hg19:9:106938220-106938244:+

2.2 Rename the mandatory columns below:

From -> To
SampleId -> sample_id
miRNA_ID -> gene_id
isoform_coords -> seq_loc
read_count -> descriptions

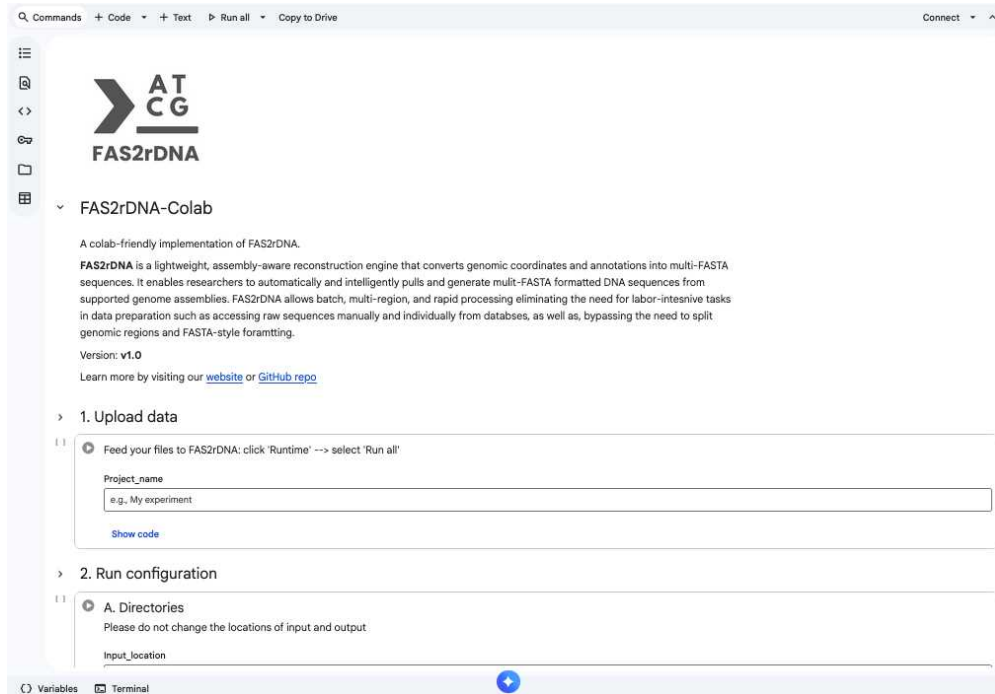


Note that you can delete the remaining columns or keep them as they are.

2.3 Save your data in tab-delimited text file (.txt or .tsv file types are recommended).

3 Reconstitute the miRNome sequences of all the cohorts using FAS2rDNA-Colab.

3.1 Launch FAS2rDNA-Colab here: <https://fas2rdna.chordexbio.com/colab-access> or visit the FAS2rDNA-Colab GitHub repo here: <https://github.com/mahvin92/FAS2rDNA-Colab>.



FAS2rDNA-Colab interface on Google Colab.

3.2 Type the name of your experiment in the *Project_name* field.

3.3 Run FAS2rDNA-Colab:

Click 'Runtime' -> select 'Run all' -> Upload all your miR-Seq data files

3.4 Once the run is completed, the results will be automatically downloaded. If download does not automatically starts, manually get them from the following directory:

/content/fas2rdna/outputs

4 Validate the resulting multi-FASTA miRNome sequences of the cohorts.

4.1 Check that the individual .fasta file from all text inputs, including the combined .fasta file are all generated in the output folder.

- 4.2 Inspect all the files, ensuring that they are not empty, not corrupted, and complete. You can run a quick entry count by searching the occurrence of the symbol '>' and match that count with the number of items in your text files.
- 4.3 Assess that the sequences are in multi-FASTA format.

Sample multi-FASTA result:

```
>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1_1
AAATCGGCGGACTCGGCAC ...
>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1_4
TTTAAACGCCCCACGCCT ...
>>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1_3
GGGCGCGTTACGTGCACGT ...
>>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1_13
TGCATTGACACCACTTCGG ...
```

Application Note

- 5 The reconstituted miRNome profiles are ready for sequence-level validation against accessioned miRNA references (e.g., miRBase). Beyond validation, the reconstructed FASTA-formatted sequences can be directly leveraged for:
 - Exploratory pan-cancer, sequence-based miRNA expression analyses, enabling comparative studies across tumor types without reliance on predefined probe sets;
 - Machine learning-driven miRNome modeling, where full-length and isoform-resolved sequences can be encoded for classification, clustering, or predictive tasks;
 - Isoform-specific correlation and variability studies, facilitating the investigation of isomiR diversity, sequence heterogeneity, and their associations with biological or clinical variables.

Collectively, these applications position FAS2rDNA-reconstituted miRNomes as versatile inputs for downstream computational genomics and integrative cancer research workflows. However, please note that FAS2rDNA-derived miRNome sequence reconstitution does not contextually capture miRSNPs yet. Current works and future iterations of FAS2rDNA will include custom accession referencing to capture base variations and miRNA sequences polymorphisms.