Dec 29, 2025

High-throughput isoform-wide miRNome sequence reconstruction in the TCGA-LUAD cohort using FAS2rDNA

DOI
https://dx.doi.org/10.17504/protocols.io.rm7vzenqxvx1/v1

Marvin De los Santos[1]

[1]ChordexBio

Marvin De los Santos
ChordexBio

DOI: https://dx.doi.org/10.17504/protocols.io.rm7vzenqxvx1/v1

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

Protocol Citation: Marvin De los Santos 2025. High-throughput isoform-wide miRNome sequence reconstruction in the TCGA-LUAD cohort using FAS2rDNA. **protocols.io** https://dx.doi.org/10.17504/protocols.io.rm7vzenqxvx1/v1

Protocol status: Working
We use this protocol and it's working

Created: December 28, 2025

Last Modified: December 29, 2025

Protocol Integer ID: 235950

Keywords: isoform-wide sequence reconstruction, miRNome profiling, miRNA isoforms, TCGA-LUAD, lung adenocarcinoma, genomic coordinate-based reconstruction, FASTA generation, batch genomic workflows, multi-assembly genome support, multi-loci sequence extraction, automated sequence formatting, multi-species genome analysis, machine learning–ready sequences, isoform-level molecular discovery, cancer genomics, scalable bioinformatics pipelines, FAS2rDNA, wide reconstruction of mirna sequence, resolution exploration of mirna isoform diversity, wide mirnome sequence reconstruction, generating mirna isoform, mirna isoform diversity, wide mirnome sequence reconstruction in the tcga, mirna sequence, reconstructed mirnome sequence, mirna isoform, integrative mirna landscape analysis, molecular understanding of lung adenocarcinoma, reconstructing sequence, sequencing dataset, aware nucleotide sequence, sequencing data, scale mirnome study, genomic coordinate, improving molecular understanding, level molecular discovery

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 for your review. Note that this protocol was tested using the GDAC-derived TCGA-LUAD miRNA isoform expression data only. Additional validation may be required when applying this workflow to other cohorts or datasets. FAS2rDNA 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

Large-scale miRNome studies frequently rely on coordinate-based annotations or raw sequencing datasets that are computationally expensive to reprocess and difficult to integrate into sequence-centric analytical workflows. This protocol presents an isoform-wide reconstruction of miRNA sequences from the TCGA-LUAD cohort using FAS2rDNA, enabling direct derivation of strand-aware nucleotide sequences without reanalyzing bulk sequencing data. By reconstructing sequences directly from genomic coordinates, the workflow provides a faster, more scalable, and reproducible alternative for generating miRNA isoform–resolved FASTA datasets.

The reconstructed miRNome sequences generated through this protocol are directly applicable to machine learning–based modeling, isoform-level molecular discovery, and integrative miRNA landscape analysis. Applied to the TCGA-LUAD cohort, this workflow facilitates high-resolution exploration of miRNA isoform diversity with the broader objective of improving molecular understanding of lung adenocarcinoma and supporting data-driven strategies aimed at reducing cancer-related mortality.

Protocol Procedure

1 Acquire or download the TCGA-LUAD miRNome annotation data.

For the purpose of this exercise, the GDAC Broad Institute database will be used to obtain molecular data.

1.1 Visit the GDAC portal of Broad Institute at **https://gdac.broadinstitute.org/**.

1.2 Select 'Lung adenocarcinoma' or LUAD cohort from the list.

Note: You must agree to the TCGA guidelines on the responsible use of data to proceed.

1.3 From the LUAD Archive pop-up, navigate to the 'miRSeq' section and select 'illuminahiseq_mirnaseq-miR_isoform_expression'

1.4 Wait until the download is finished (the file is normally in the form of tab-delimited .txt file type).

2 Preprocess and normalize the genomic coordinate annotations.

2.1 Open the file using a spreadsheet program.

2.2 Inspect the data and filter out any missing values or incorrectly formatted valued.

2.3 Ensure that the 'isoform_coords' header of your data follow the FAS2rDNA-required format:

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

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

2.4 Once quality check is done, rename the 'SampleId', 'miRNA_ID', 'isoform_coords', and 'read_count' headers to 'sample_id', 'gene_id', 'seq_loc', and 'description', respectively. This will ensure the data is properly formatted for FAS2rDNA.

3 Perform isoform-wide sequence reconstruction using FAS2rDNA.

For the purpose of this exercise, the FAS2rDNA-CLI version will be used (available at **https://fas2rdna.chordexbio.com/**)

3.1 Install the required python dependencies.

```
pip install pandas pyfaidx tqdm
apt-get update -qq
apt-get install -y samtools
```

3.2 Install or download the FAS2rDNA-CLI script.

Note that the technical documentation of FAS2rDNA can be accessed publicly on GitHub (available at https://github.com/mahvin92/FAS2rDNA)

3.3 Run FAS2rDNA using the the format below, specifying the location of the TCGA-LUAD miRNA annotation text file:

```
python3 fas2rdna.py \
  --input-dir /Users/Desktop/FAS2rDNA
```

or build a custom FASTA header by running the following line:

```
python3 fas2rdna.py \
  --input-dir /Users/Desktop/FAS2rDNA \
  --header "{sample_id}|{gene_id}|{seq_loc}|{description}"
```

3.4 After FAS2rDNA finishes running, inspect the resulting .fasta ouput files. FAS2rDNA will generate the individual .fasta file from multiple text files and the combined .fasta file, compiling all multi-FASTA sequences in one file. The result map is in the following structure:

```
/Users/Desktop/FAS2rDNA/
├── LUAD.txt
├── test.txt
└── fas2rdna_output/
    ├── genomes/
    ├── fasta/
    │   ├── LUAD.fasta
    │   └── test.fasta
    └── All_sequences.fasta
```

4 Cross-reference the generated sequences with the data input, ensuring all samples were represented, the sequences are in FASTA-format and the .fasta files are not empty.

Sample LUAD data:

```
sample_id       gene_id seq_loc description
TCGA-05-4390-01A-02T-1754-13    hsa-let-7a-1    hg19:9:96938243-
96938263:+      1
TCGA-05-4390-01A-02T-1754-13    hsa-let-7a-1    hg19:9:96938243-
96938264:+      4
TCGA-05-4390-01A-02T-1754-13    hsa-let-7a-1    hg19:9:96938243-
96938265:+      1
TCGA-05-4390-01A-02T-1754-13    hsa-let-7a-1    hg19:9:96938243-
96938266:+      4
TCGA-05-4390-01A-02T-1754-13    hsa-let-7a-1    hg19:9:96938244-
96938263:+      13

...
```

Sample reconstructed LUAD miRNome FASTA-formatted sequences (FAS2rDNA result):

```
>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1
ATGAGGTAGTAGGTTGTATAG
>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1
ATGAGGTAGTAGGTTGTATAGT
>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1
ATGAGGTAGTAGGTTGTATAGTT
>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1
ATGAGGTAGTAGGTTGTATAGTTT
>TCGA-05-4390-01A-02T-1754-13_hsa-let-7a-1
TGAGGTAGTAGGTTGTATAG

...
```

Additional note

5 FAS2rDNA operates in batch (multiple experiments), multi-assembly (detects different assembly versions like hg17, hg18, hg19, hg38 in humans), multi-loci (can reconstruct sequence from different DNA locations), auotomated formatting (compile multi-FASTA-formatted results, ready for downstream analyses), and multi-species (supports assemblies from humans to yeast genomes) workflows, giving users the speed, scalability, confidence and convenience needed to analyze and perform experiments using large genomic data.