MIRNETCLASSIFIER REQUIREMENTS

Due to Github file size limit some files required by the packages have not been uploaded. These files should be downloaded and placed in their location manually by the user before running the library.

The **GTF file** and the **Structural Engine files** are compulsory to run the library. It will not work without them. The **TCGA Raw databases** are required only to run the code of the PLOS article.

GTF FILE (COMPULSORY FOR MIRNETCLASSIFIER)

https://www.gencodegenes.org/human/release_45.html

The document is periodically updated. miRNetClassifier works with the 45th version. Higher or lower versions could have compatibility problems.

GTF / GFF3 files

Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only	GTF GFF3
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes)	GTF GFF3
Comprehensive gene annotation	PRI	• It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions	GTF GFF3
Basic gene annotation	CHR	• It contains the basic gene annotation on the reference chromosomes only • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene • This is the **main annotation file** for most users	GTF GFF3
Basic gene annotation	ALL	• It contains the basic gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene • This is a **superset** of the main annotation file	GTF GFF3
Basic gene annotation	PRI	• It contains the basic gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3
Long non-coding RNA gene annotation	CHR	• It contains the comprehensive gene annotation of lncRNA genes on the reference chromosomes	GTF GFF3

Download from the specified link. Extract the compressed document. Place the .gtf file in the next location - **Library/miRNetClassifier/gtf/** with the **original file name**.

STRUCTURAL ENGINE (COMPULSORY FOR MIRNETCLASSIFIER)

TARGETSCAN

https://www.targetscan.org/cgi-bin/targetscan/data_download.vert80.cgi

All predictions for representative transcripts*

File	Description	Fields	# of Rows
Conserved Family Info, all predictions - (75.99 MB)	Positions in UTRs (without gaps) and UTR multiple sequence alignments (MSA; with gaps) of conserved and nonconserved sites corresponding to **conserved miRNA families**	miR Family, Gene ID, Gene Symbol, Transcript ID, Species ID, UTR start, UTR end, MSA start, MSA end, Seed match, and P_{CT}	5,913,100
Nonconserved Family Info, all predictions - (346.57 MB)	Positions in UTRs (without gaps) and UTR multiple sequence alignments (MSA; with gaps) of conserved and nonconserved sites corresponding to **nonconserved miRNA families**	miR Family, Gene ID, Gene Symbol, Transcript ID, Species ID, UTR start, UTR end, MSA start, MSA end, Seed match, and P_{CT}	25,306,796
Conserved site context++ scores - (18.6 MB)	Context++ scores, KDs for all **conserved miRNA sites**	Gene ID, Gene Symbol, Transcript ID, Species ID, miRNA, Site type, UTR start, UTR end, context++ score, context++ score percentile, weighted context++ score, weighted context++ score percentile, predicted relative KD -- **updated 23 May 2022**	1,468,778
Nonconserved site context++ scores - (542.49 MB)	Context++ scores, KDs for all **nonconserved miRNA sites**	Gene ID, Gene Symbol, Transcript ID, Species ID, miRNA, Site type, UTR start, UTR end, context++ score, context++ score percentile, weighted context++ score, weighted context++ score percentile, predicted relative KD -- **updated 23 May 2022**	38,497,660
Summary Counts, all predictions - (296.29 MB)	Counts of every Gene:miRNAfamily pair, including total context++ scores, weighted context++ scores, aggregate P_{CT}s, and predicted occupancies	Transcript ID, Gene Symbol, miRNA family, Species ID, Total num conserved sites, Number of conserved 8mer sites, Number of conserved 7mer-m8 sites, Number of conserved 7mer-1a sites, Total num nonconserved sites, Number of nonconserved 8mer sites, Number of nonconserved 7mer-m8 sites, Number of nonconserved 7mer-1a sites, Representative miRNA, Total context++ score, Cumulative weighted context++ score, Aggregate PCT, Predicted occupancy (low miRNA), Predicted occupancy (high miRNA), Predicted occupancy (transfected miRNA)	23,014,944
Genome coordinates of all predicted sites (168.23 MB)	Genome (hg19) locations of all targets, partitioned into files by conservation of miRNA family and site	BED format fields, where score = context++ score percentile	[8 files]

* The representative transcript of a gene is the transcript variant with the most 3P-seq tags.

Download from the specified link. Extract the compressed document. Place the .txt file in the next location - **Library/miRNetClassifier/StructuralEngine/** with the **Targetscan_targets.txt name**.

DIANA

https://dianalab.e-ce.uth.gr/microt_webserver/#/download

File	Size (MB)	MD5sum
interactions_human.microT.mirbase.txt.gz	308	b851dd0b21b32a7c0252ad04c857a4e4
interactions_human.microT.mirgenedb.txt.gz	122	a4122a96e1895e651d0f325b310d57e2
interactions_mouse.microT.mirbase.txt.gz	215	f2d3c7a2046bf515d1e21ce76569db94
interactions_mouse.microT.mirgenedb.txt.gz	93	3b0249a85c329484ab042c3aca1fcccc
interactions_rat.microT.mirbase.txt.gz	51	b6bb484e7e63233e1ce88a933bc7341f
interactions_rat.microT.mirgenedb.txt.gz	53	a3623858c02926bd0b40c75442ed6e98
interactions_chicken.microT.mirbase.txt.gz	75	075b3e03260b7ac3bb5bbb5818e1994f
interactions_chicken.microT.mirgenedb.txt.gz	32	a9119797deede3b3225c07c118b9fe9c
interactions_drosophila.microT.mirbase.txt.gz	19	5e9be511247d68a50c35c0b06b566729
interactions_drosophila.microT.mirgenedb.txt.gz	13	bccd839a5f67757138f8276bf338c9ce
interactions_celegans.microT.mirbase.txt.gz	14	2eadd3fa68ad23233269e4ad8672d40f
interactions_celegans.microT.mirgenedb.txt.gz	9	bbc724d1919f7d6103c5e8ebc7072b49

Download from the specified link. Extract the compressed document. Place the .txt file in the next location - Library/miRNetClassifier/StructuralEngine/ with the **DIANA_targets.txt name**.

MIRTARBASE

https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2019/php/download.php

Download from the specified link. Extract the compressed document. Place the .csv file in the next location **- Library/miRNetClassifier/StructuralEngine/** with the **MTB_targets.csv name**.

TCGA RAW DATA (Only for PLOS code)

Full data is downloaded from UCSC Xena Functional Genomics Browser. Specifically, data from TCGA is downloaded through the next link:

https://xenabrowser.net/datapages/?hub=https://gdc.xenahubs.net:443

Next ones are the links to download data for each specific cancer.

BRCA -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Breast%20Cancer%20(BRCA)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

COAD -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Colon%20Cancer%20(COAD)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

HNSC -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Head%20and%20Neck%20Cancer%20(HNSC)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

KIRC -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Kidney%20Clear%20Cell%20Carcinoma%20(KIRC)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

LAML -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Acute%20Myeloid%20Leukemia%20(LAML)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

LGG -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Lower%20Grade%20Glioma%20(LGG)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

LIHC -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Liver%20Cancer%20(LIHC)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

LUAD -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Lung%20Adenocarcinoma%20(LUAD)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

LUSC -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Lung%20Squamous%20Cell%20Carcinoma%20(LUSC)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

OV -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Ovarian%20Cancer%20(OV)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

PRAD -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Prostate%20Cancer%20(PRAD)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

SARC -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Sarcoma%20(SARC)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

SKCM -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Melanoma%20(SKCM)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

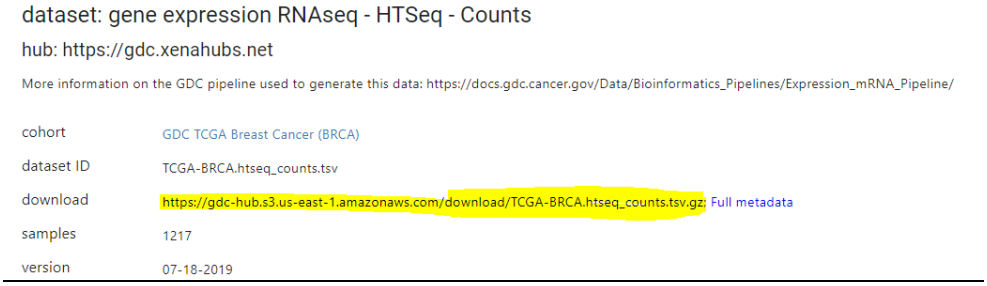
STAD -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Stomach%20Cancer%20(STAD)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

THCA -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Thyroid%20Cancer%20(THCA)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

UCEC -
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Endometrioid%20Cancer%20(UCEC)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

Survival, miRNA expression and mRNA expression is required. Survival and miRNA data are already in the Github repository. mRNA expression data should be downloaded. The file is called HTSeq – Counts (mRNA)

The download link should be selected inside each file.



All three files should be located inside the folder of the specific cancer in his directory: **Library/PLOS Code/TCGA Raw Data/BRCA** (for example). The file should be saved **with its original name**. For example: