



Documentation

CmiRClustFinder v2.0 Tool

Department of Bioinformatics
Manipal School of Life Sciences,
Manipal Academy of Higher Education,
Manipal – 576104

Last Revised: July 2022



If you are using this pipeline, please cite:

- Ware, A.P., Kabekkodu, S.P., Chawla, A., Paul, B., Satyamoorthy K. Diagnostic and prognostic potential clustered miRNAs in bladder cancer. 3 Biotech 12, 173 (2022). <https://doi.org/10.1007/s13205-022-03225-z>

Example Dataset

To download example datasets (.zip), please click [here](#)

File Details:

File	Description
miRNA-Clusters.bed	Four-column simple bed file contain clustered miRNA and their genomic location
Segmented-SCNA.seg	Segmented SCNA from the TCGA-CHOL (cholangiocarcinoma) patients

Thoughts behind the development of the CmiRClustFinder Tool

Copy number aberration (CNA) events are prevalent in cancer patients. Recurrent and aberrant copy number regions can occur in genetic regulators such as microRNA (miRNA) clusters. The miRNA clusters are groups of miRNAs that are co-expressed, with similar expression patterns and jointly regulate target genes that are associated with various normal biological phenomena and tumor-associated pathways. Several studies have demonstrated CNA-mediated miRNA dysregulation; however, the interplay between clustered miRNAs and CNAs in cancer progression is largely unknown.

We hypothesize that the abnormal expression of clustered miRNAs due to aberrant CNAs and genomic rearrangements leads to the initiation and progression of several cancer types. Therefore, an understanding of miRNA clusters associated with CNAs could shed light on CNA-regulated tumor initiation and progression. Hence, we have developed a user-friendly computational pipeline, called CmiRClustFinder v1.0 by integrating multiple R-based and Linux command-line packages. CmiRClustFinder integrates CNA, gene and miRNA expression datasets from TCGA to compute CNA co-localized miRNA clusters, miRNAs, genes and user-defined genetic elements from 35 cancer types. The automated pipeline CmiRClustFinder v1.0 can be downloaded from the GitHub repository (https://github.com/msls-bioinfo/CmiRClustFinder_v1.0). The pipeline can be effectively used for integrated high throughput data analytics and identification of signatures for cancer diagnosis.

“CmiRClustFinder is not limited to only miRNAs/miRNA clusters, Users can use it for other genomic elements/regions.”



MANIPAL SCHOOL OF LIFE SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

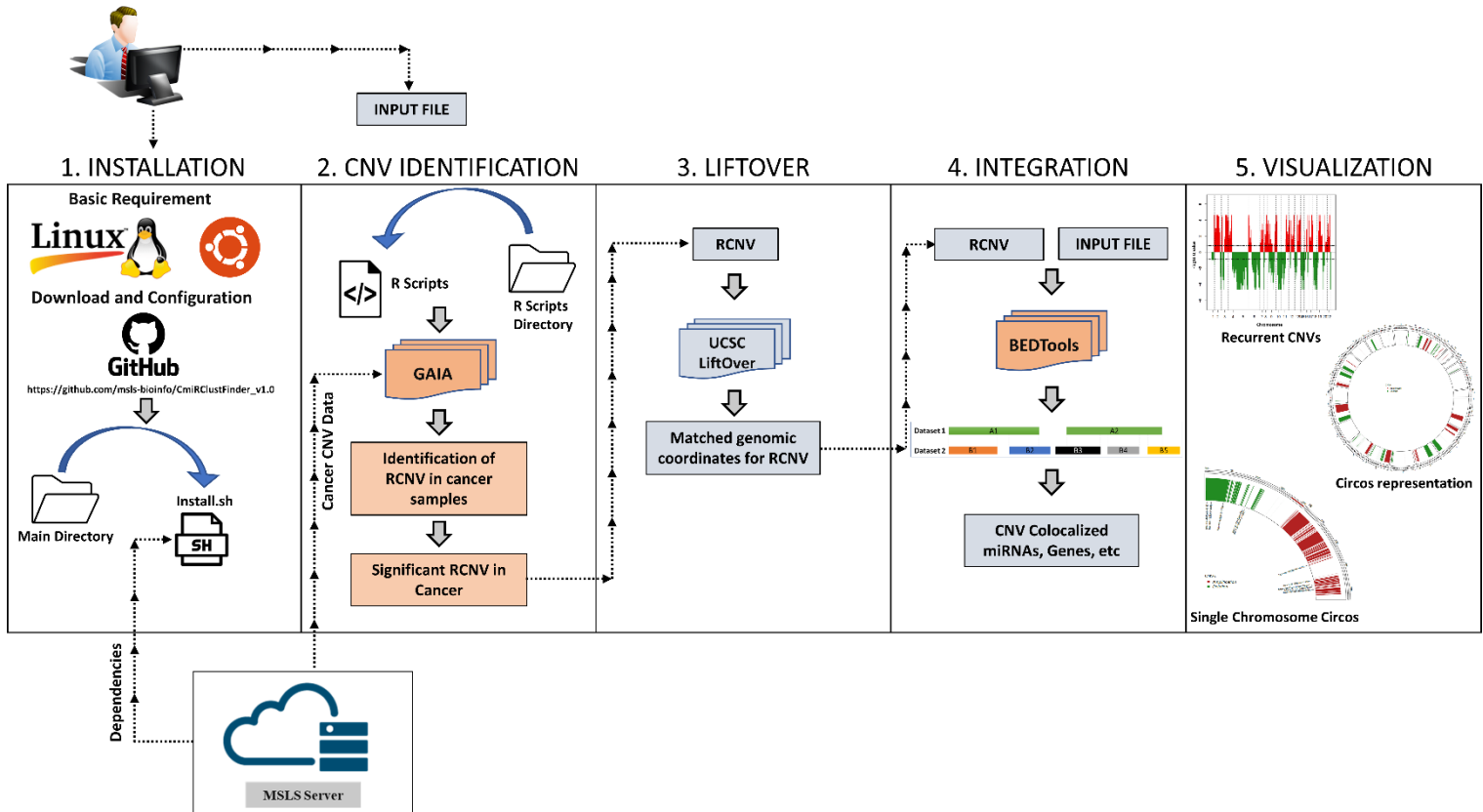


MANIPAL
ACADEMY of HIGHER EDUCATION
(Deemed to be University under Section 3 of the UGC Act, 1956)

Standalone CmiRClustFinder v1.0

(A Linux based pipeline)

How does it work?



CmiRClustFinder v1.0 computes the integrated data within five steps. The installation script will download all the necessary resources and prepare the pipeline for use in the first step. In the second step, the GAIA package finds frequent aberrations in chromosomal regions among cancer patients' datasets. In the third step, the LiftOver tool matches the genomic build for RCNVs and user-defined genetic elements. We have integrated BEDTools to find colocalization of significant RCNV and genomic elements in the fourth step. Lastly, the circos package generates a circos representation of the data.

Installation and Prerequisite

CmiRClustFinder pipeline is designed for Linux operating system. If you wish to use the standalone version of this pipeline, follow the instructions below.

The following Linux utilities are required to run this pipeline. Please make sure the following are installed and available on your system prior to running `install.sh` from the source directory.

1. R = 4.0 (or higher)
2. git
3. unzip

If the above prerequisites are satisfied, you are ready to install dependencies and build the program. Note during the building procedure, `install.sh` will attempt to download and install several packages, so an active internet connection is required.

To obtain *CmiRClustFinder*, Use:

```
git clone https://github.com/mpls-bioinfo/CmiRClustFinder\_v1.0.git
```

```
cd CmiRClustFinder_v1.0/
```

Or

```
wget https://github.com/mpls-bioinfo/CmiRClustFinder\_v1.0/archive/refs/heads/main.zip
```

```
unzip main.zip
```

```
cd CmiRClustFinder_v1.0-main/
```

If you have downloaded the source code and it is in a directory `CmiRClustFinder/`, to install all dependencies follow the procedure

```
cd CmiRClustFinder/
```

```
sh install.sh
```

Running the pipeline

After the successful execution of `install.sh` you are ready to run the main pipeline script `CmiRClustFinder.r` which is in `RScript/` directory

Please navigate in to the `Rscripts/` directory to start the pipeline

General usage:

Rscript CmiRclustFinder.r <TCGA cohort abbreviation> <BED file specifying user interested genomic regions>

Arguments / Parameters	Description
<i><TCGA cohort abbreviation></i>	This argument required TCGA cohort abbreviation, which you can select from the list below
<i><BED file specifying user interested genomic regions></i>	BED file which contains the specific genomic regions, to check their co-localization with RCNV.



TCGA Cancer Abbreviations

Sr. No.	Cohort Abbreviation	Cohort Name
1	TCGA-ACC	Adrenocortical carcinoma
2	TCGA-BLCA	Bladder Urothelial Carcinoma
3	TCGA-BRCA	Breast invasive carcinoma
4	TCGA-CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
5	TCGA-CHOL	Cholangiocarcinoma
6	TCGA-COAD	Colon adenocarcinoma
7	TCGA-COADREAD	Colorectal adenocarcinoma
8	TCGA-DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
9	TCGA-ESCA	Esophageal carcinoma
10	TCGA-GBM	Glioblastoma multiforme
11	TCGA-GBMLGG	Glioma
12	TCGA-HNSC	Head and Neck squamous cell carcinoma
13	TCGA-KICH	Kidney Chromophobe
14	TCGA-KIPAN	Pan-kidney cohort (KICH+KIRC+KIRP)
15	TCGA-KIRC	Kidney renal clear cell carcinoma
16	TCGA-KIRP	Kidney renal papillary cell carcinoma
17	TCGA-LGG	Brain Lower Grade Glioma
18	TCGA-LIHC	Liver hepatocellular carcinoma
19	TCGA-LUAD	Lung adenocarcinoma
20	TCGA-LUSC	Lung squamous cell carcinoma
21	TCGA-MESO	Mesothelioma
22	TCGA-OV	Ovarian serous cystadenocarcinoma
23	TCGA-PAAD	Pancreatic adenocarcinoma
24	TCGA-PCPG	Pheochromocytoma and Paraganglioma
25	TCGA-PRAD	Prostate adenocarcinoma
26	TCGA-READ	Rectum adenocarcinoma
27	TCGA-SARC	Sarcoma
28	TCGA-SKCM	Skin Cutaneous Melanoma
29	TCGA-STAD	Stomach adenocarcinoma
30	TCGA-TGCT	Testicular Germ Cell Tumors
31	TCGA-THCA	Thyroid carcinoma
32	TCGA-THYM	Thymoma
33	TCGA-UCEC	Uterine Corpus Endometrial Carcinoma
34	TCGA-UCS	Uterine Carcinosarcoma
35	TCGA-UVM	Uveal Melanoma

The second argument required for Rscript is the BED file which contains the specific genomic regions, to check their co-localization with RCNV. Below is an example of the BED file.

NOTE: The table header is for descriptive purposes, the BED file should not have a header

CHROM	START	END	IDENTIFIER
chr19	53666679	53706336	hsa-miR-526a-1/miR-512-1
chr14	1.01E+08	1.01E+08	hsa-miR-1185-1/miR-379
chr14	1.01E+08	1.01E+08	hsa-miR-136/miR-493
chrX	50003148	50014683	hsa-miR-502/miR-532
chr9	1.35E+08	1.35E+08	hsa-miR-3689f/miR-3689c
chr13	91350605	91351391	hsa-miR-92a-1/miR-17
chrX	1.34E+08	1.34E+08	hsa-miR-106a/miR-363
chrX	1.35E+08	1.35E+08	hsa-miR-424/miR-450b
chrX	1.46E+08	1.46E+08	hsa-miR-891b/miR-892c
chr20	63919449	63919939	hsa-miR-941-5/miR-941-1

Know more about the bed file format: (<http://genome.ucsc.edu/FAQ/FAQformat#format1>)



MANIPAL SCHOOL OF LIFE SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)



MANIPAL
ACADEMY of HIGHER EDUCATION
(Deemed to be University under Section 3 of the UGC Act, 1956)

CmiRClustFinder v2.0

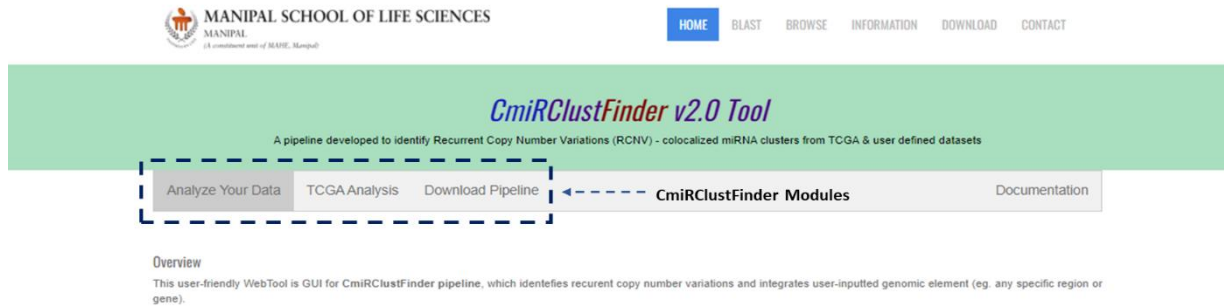
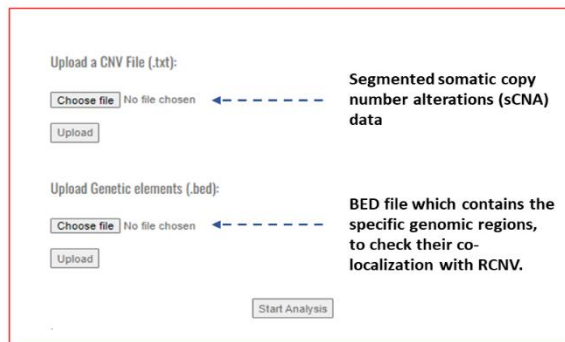
(A Webserver with GUI)

Module for analyzing user input data

CmiRClustFinder 2.0 is the upgraded version with a graphical user interface (GUI) for non-Linux users.

Data Submission

The data submission form on the CmiRClustFinder webserver

Dataset Example

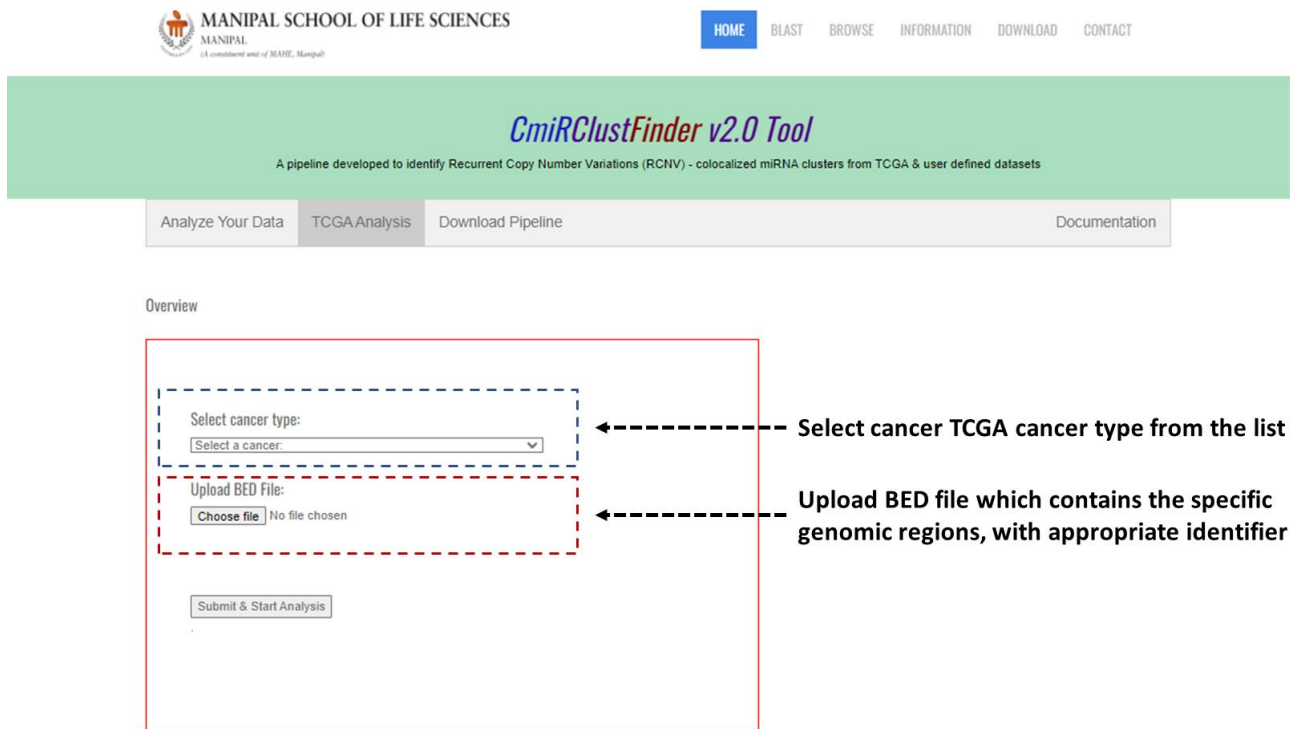
Segmented copy number data

Sample	Chromosome	Start	End	Num_Probes	Segment_Mean
TCGA-3X-AAV9-10A-01D-A419-01	1	3218610	247813706	129096	0
TCGA-3X-AAV9-10A-01D-A419-01	2	484222	9948475	5699	0.0088
TCGA-3X-AAV9-10A-01D-A419-01	2	9949576	9950371	3	-1.5053
TCGA-3X-AAV9-10A-01D-A419-01	2	9953949	70256255	35199	0.0041
TCGA-3X-AAV9-10A-01D-A419-01	2	7025897	70259495	3	-0.9245
TCGA-3X-AAV9-10A-01D-A419-01	2	70263331	76769258	3626	0.0022
TCGA-3X-AAV9-10A-01D-A419-01	2	76789156	76799126	6	-1.014
TCGA-3X-AAV9-10A-01D-A419-01	2	76808048	222751806	74929	0.005
TCGA-3X-AAV9-10A-01D-A419-01	2	222754101	222759558	4	-1.11
TCGA-3X-AAV9-10A-01D-A419-01	2	222760279	242476062	11782	0.0028
TCGA-3X-AAV9-10A-01D-A419-01	3	2212571	10559238	4761	0.0039
TCGA-3X-AAV9-10A-01D-A419-01	3	10560942	10562533	2	-1.5264
TCGA-3X-AAV9-10A-01D-A419-01	3	10565125	65047603	31780	0.002
TCGA-3X-AAV9-10A-01D-A419-01	3	65047633	65048037	2	-1.3193
TCGA-3X-AAV9-10A-01D-A419-01	3	65051762	65432030	333	-0.0053
TCGA-3X-AAV9-10A-01D-A419-01	3	65432464	65432483	2	-1.1371
TCGA-3X-AAV9-10A-01D-A419-01	3	65434647	80933596	9483	-0.0029
TCGA-3X-AAV9-10A-01D-A419-01	3	80935171	80936580	3	-1.267
TCGA-3X-AAV9-10A-01D-A419-01	3	80940139	197538677	59974	0.0029
TCGA-3X-AAV9-10A-01D-A419-01	4	1053934	36536256	21807	0.001
TCGA-3X-AAV9-10A-01D-A419-01	4	36540003	36540033	2	-1.44

Bed file

chr16	16300159	16309966	hsa-miR-3180-2/miR-3179-2
chr16	18402178	18411977	hsa-miR-3179-3/miR-3180-3
chr1	220117853	220118241	hsa-miR-194-1/miR-215
chr12	69584722	69584822	hsa-miR-3913-2/miR-3913-1
chr16	14303967	14309371	hsa-miR-365a/miR-193b
chr16	14925937	14930879	hsa-miR-6770-1/miR-6511a-1
chr16	16324588	16329364	hsa-miR-6770-2/miR-6511a-2
chr16	18488301	18494576	hsa-miR-3179-4/miR-3670-4
chr18	21825698	21829088	hsa-miR-1-2/miR-133a-1
chr20	34990376	34990472	hsa-miR-499b/miR-499a
chr3	160404588	160404825	hsa-miR-16-2/miR-15b
chr9	124692442	124693798	hsa-miR-181b-2/miR-181a-2
chr9	128244721	128245030	hsa-miR-3154/miR-199b
chr9	128392618	128392708	hsa-miR-219b/miR-219a-2
chr22	46112749	46113768	hsa-let-7b/let-7a-3
chr8	12719132	12727299	hsa-miR-3926-2/miR-5692a-2
chr1	51059837	51060103	hsa-miR-6500/miR-4421
chr1	62078786	62078856	hsa-miR-3116-2/miR-3116-1

Module for analyzing user input genetic element against TCGA patient datasets



The screenshot shows the web interface for the CmiRClustFinder v2.0 Tool. At the top, there is a navigation bar with links for HOME, BLAST, BROWSE, INFORMATION, DOWNLOAD, and CONTACT. Below this is a green banner with the tool's name and a description: "A pipeline developed to identify Recurrent Copy Number Variations (RCNV) - colocalized miRNA clusters from TCGA & user defined datasets". A secondary navigation bar contains buttons for "Analyze Your Data", "TCGA Analysis", "Download Pipeline", and "Documentation". The "TCGA Analysis" button is highlighted. Below this is an "Overview" section containing a form with two main input fields: "Select cancer type:" with a dropdown menu and "Upload BED File:" with a "Choose file" button. A "Submit & Start Analysis" button is located at the bottom of the form. Two dashed boxes highlight the input fields, with arrows pointing to explanatory text on the right.

Select cancer TCGA cancer type from the list

Upload BED file which contains the specific genomic regions, with appropriate identifier

Data Processing

After submitting the data in the proper format, it will be redirected to the data processing window. The time required for data processing depends on the size of data files submitted and the processing load on the server.

Data processing page, which will be redirected to results after completion

Please wait while we are processing your data....



Date: Thu Jun 9 19:54:22 IST 2022

You will be auto-redirected to results!

Starting Analysis

#####

```
+--+--+--+--+--+--+--+--+
| CmiRclustFinder v2.0 |
+--+--+--+--+--+--+--+--+
```

Developed by:
Department of Bioinformatics,
Manipal School of Life Sciences,
Manipal Academy of Higher Education

For more information, visit:
https://github.com/msls-bioinfo/CmiRclustFinder_v1.0

For help, inquiries and suggestions, please contact:
bioinfo.sls@manipal.edu


#####

Loading Dependencies....

[1] "Identifying RCNV...."

Interpreting CmiRClustFinder Output

The output from CmiRClustFinder contains Circos plots for the representation of genomic regions/RCNV regions/their colocalization. The .tsv master file will provide specific details of each genetic element that overlap with significant RCNVs from the specified dataset.



MANIPAL SCHOOL OF LIFE SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

[HOME](#) [BLAST](#) [BROWSE](#) [INFORMATION](#) [DOWNLOAD](#) [CONTACT](#)

CmiRClustFinder v2.0 Tool





A pipeline developed to identify Recurrent Copy Number Variations (RCNV) - colocalized miRNA clusters from TCGA & user defined datasets

Results
Documentation

Here are the result of your analysis...

CmiRClustFinder pipeline, which identifies recurrent copy number variations and integrates user-inputted genomic element (eg. any specific region or gene).

Click on icons to visualize/download results

File	Description	Visualize/Download
Circos genetic elements (PNG)	Circos represented of user provided genetic elements	
Circos (PDF)	Circos represented of recurrent copy number variation co-localized genetic elements	
Result (TSV)	Circos represented of recurrent copy number variation co-localized genetic elements	
Result (ZIP)	Compressed file of results	

[Go Back](#)



Credits

1. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages (<https://f1000research.com/articles/5-1542>)
2. TCGAbiolinks R package (<https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html>)
3. BEDTools (<https://bedtools.readthedocs.io/en/latest/>)
4. UCSC liftOver (<https://genome-store.ucsc.edu/>)