

# DeFusion: a denoised network regularization for multi-omics integration

## Dependencies

Source codes of **DeFusion** are written using *MATLAB*, but we run preprocessing steps and downstream tasks in *Python*. The Cox Proportional-Hazards model is fitted by the *coxph* function in R.

*Python* 3.7.1

- scikit-learn 0.20.1
- numpy 1.15.4
- pandas 0.23.4
- rpy2 2.9.5

*R* 3.5.3

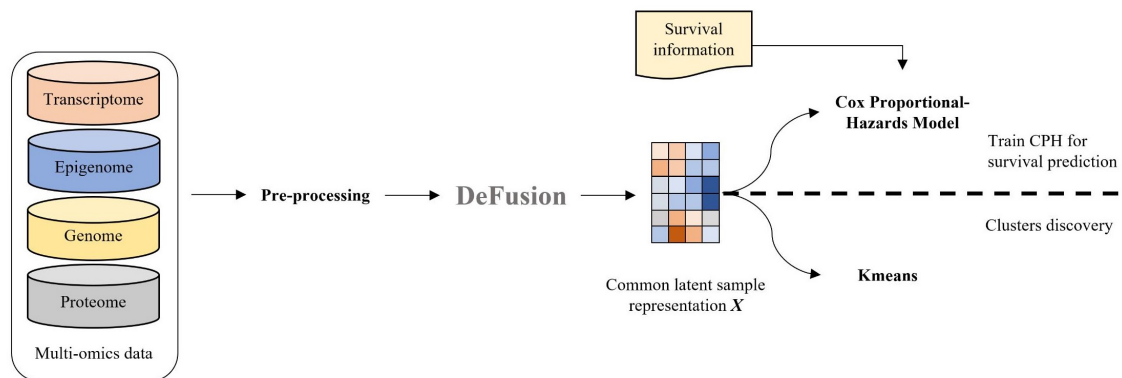
- survival 2.44.1
- survcomp 1.32.0

*MATLAB*

- [Similarity network fusion](#)
- [Network enhancement](#)
- [npy-matlab](#)

Note that the directories *Network\_Enhancement* and *SNFmatlab* store codes downloaded from the official websites

## Work flow



## Data availability

- The data used in simulation study was generated by R scripts provided in [1].
- TCGA-BRCA, TCGA-KIRC, and TCGA-LIHC are publicly available at <https://portal.gdc.cancer.gov/> and we downloaded from Data Release v13.0 (September 27, 2018). The mRNA expression data matrices were constructed by HTseq-FPKM files and miRNA expression data matrices by miRNA expression quantification files. The DNA methylation matrices consisted of both illumina 27K and 450K data at level 3.
- TCGA-KIDNEY was provided by [2].
- TCGA-LAML, TCGA-SARC, and TCGA-SKCM were acquired from [http://acgt.cs.tau.ac.il/multi\\_omic\\_benchmark/download.html](http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html) [3].
- We downloaded the external validation dataset--microarray gene expression profiles of 242 patients from Gene Expression Omnibus (GEO) with access number GSE14520.

- Data used in proteomics-and-phosphoproteomics integration was retrieved from supplementary data in [4]

In seven TCGA (**The Cancer Genome Atlas**) cancer cohorts, we have tried to integrate mRNA and miRNA expression and DNA methylation corresponding to transcriptome and epigenome, respectively.

We also have tried to integrate genome, transcriptome, and epigenome using copy number estimate, mRNA and miRNA expression, and DNA methylation in TCGA-BRCA. In our paper, we have labeled this cohort as TCGA-BRCA<sub>+CNV</sub>.

We have tried proteomics-and-phosphoproteomics integration by using proteomics and phosphoproteomics data of normal and tumor tissues.

[1] Chauvel C, Novoloaca A, Veyre P, et al. Evaluation of integrative clustering methods for the analysis of multi- omics data. Brief Bioinform 2019; Feb:bbz015.

[2] Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11(3):333-337.

[3] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 2018;46(20):10546-10562.

[4] Xu JY, Zhang C, Wang X, et al. Integrative proteomic characterization of human lung adenocarcinoma. Cell 2020;182(1):245-261.

## Downloading data from the TCGA official website (<https://portal.gdc.cancer.gov/> )

We downloaded data from TCGA (<https://portal.gdc.cancer.gov/> ) using the following steps, which we illustrate with downloading HTseq-FPKM files of TCGA-BRCA as a running example. Downloading files of miRNA expression, copy number estimate, and DNA methylation follows the same steps.

- Step 1: Search for TCGA projects in the main page

The screenshot displays the TCGA Genomic Data Commons Data Portal interface. At the top, there's a navigation bar with links for Home, Projects, Exploration, Analysis, and Repository. Below this, a search bar contains the text 'TCGA-BRCA'. To the right of the search bar, there's a 'Cases by Major Primary Site' bar chart showing the distribution of cases across various cancer types. Below the search bar, a table lists search results for TCGA-BRCA, including sample IDs and their corresponding file counts. At the bottom, there's a section for 'GDC Applications' with links to various tools and resources.

Sample ID	Files	Genes	Mutations
CA 607650c8-3495-461c-b748-237353c23416	596,758	23,399	3,287,299
CA 86c8f993-3271-4525-9963-29c55625593a			
CA fc2cd810-aa52-4789-aac2-7683281bb22f			
CA 5ed79093-1571-4d71-8136-0d84ccabdcac			

- Step 2: Click the link in the 'Files' column corresponding to the RNA-seq

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

TCGA-BRCA

Explore Project Data Biospecimen Clinical Manifest

Summary

The project has controlled access data which requires dbGaP Access. See instructions for Obtaining Access to Controlled Data.

Project ID: TCGA-BRCA  
dbGaP Study Accession: phs000178  
Project Name: Breast Invasive Carcinoma  
Disease Type: 9 Disease Types  
Primary Site: Breast  
Program: TCGA

CASES: 1,098  
FILES: 33,766  
ANNOTATIONS: 121

Cases and File Counts by Data Category

Data Category	Cases (n=1,098)	Files (n=33,766)
Sequencing Reads	1,098	4,679
Transcriptome Profiling	1,097	6,090
Simple Nucleotide Variation	1,044	8,648
Copy Number Variation	1,098	9,822
DNA Methylation	1,095	3,234
Clinical	1,098	1,183
Biospecimen	1,098	5,315

Cases and File Counts by Experimental Strategy

Experimental Strategy	Cases (n=1,098)	Files (n=33,766)
Diagnostic Slide	1,062	1,133
Tissue Slide	1,093	1,979
WXS	1,050	10,623
RNA-Seq	1,092	2,359
mRNA-Seq	1,079	3,621
ATAC-Seq	25	25
Genotyping Array	1,098	6,627
Methylation Array	1,095	1,234

- Step 3: Check the "HTSeq-FPKM" checkbox

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart GDC Apps

Files Cases

Add a File Filter

Search Files: e.g. 142952 bam, 4f6e27a-b...

Data Category: transcriptome profiling 1,222

Data Type: Gene Expression Quantification 1,222

Experimental Strategy: RNA-Seq 1,222

Workflow Type: HTSeq - FPKM 1,222

Data Format: ht 1,222

Platform: No data for this field

Access: open 1,222

Files (1,222) Cases (1,092)

Primary Site Project Data Category Data Type Data Format

Showing 1 - 20 of 1,222 files 635.71 MB

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	44ca855-4056-4f6e-c320003c0f4.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	523.49 KB	0
open	de70478-4915-4749-8729-e6ee06a0c359.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.23 KB	0
open	432d963-45d0-46c0-96d5-c7d9f6ec3b6.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507 KB	0
open	996000c-0804-4559-8084-c5059838622a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.53 KB	0
open	02e17d87-c0dc-4051-a2e5-85eab14591b.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	533.83 KB	0
open	8dfe19ae-acc1-4de5-8694-1a1805c11c5.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	536.46 KB	0
open	ab2d5a3d-4253-4c80-b4e5-0d8fe163140.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	514.74 KB	0
open	59325a83-799e-4f58-aaf2-a29b34e7e563.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	495.24 KB	0
open	5498129-1f2c-458b-ba7f-4c76ea2538cc.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	521.16 KB	0
open	74118954-46ea-4099-9709-45175603a1.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.52 KB	0
open	36bcb02-58b0-4731-5a2c-09eaaac15c29.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.12 KB	0
open	b8e7b7b5-54e8-4459-5a21-ea3472b013d7.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	506.3 KB	1
open	57e8b21e-eeb7-4ffe-a8b3-a6018d3c7d96.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	522 KB	0
open	e6ba3664-1968-4f6e-5a11-0157a9d81.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.07 KB	0
open	edeae45b-a454-4a7b-9695-38d29424698.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	542.11 KB	0
open	c4489317-3e8a-4159-b32b-4d9e723359a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.35 KB	0
open	b638ae-65f9cc-42d2-bd7f-5dc14e8b372.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507.88 KB	0
open	0626f07-749e-4788-a815-500b9ca778.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	511.74 KB	0
open	36440974-4445-4728-b08f-10a0329d06.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	519.12 KB	0
open	a18f26b8-9497-4824-a9ba-7b1925ca987.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507.86 KB	0

- Step 4: Add all files to the cart

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart GDC Apps

Files Cases

Add a File Filter

Search Files: e.g. 142952 bam, 4f6e27a-b...

Data Category: transcriptome profiling 1,222

Data Type: Gene Expression Quantification 1,222

Experimental Strategy: RNA-Seq 1,222

Workflow Type: HTSeq - FPKM 1,222

Data Format: ht 1,222

Platform: No data for this field

Access: open 1,222

Files (1,222) Cases (1,092)

Primary Site Project Data Category Data Type Data Format

Showing 1 - 20 of 1,222 files 635.71 MB

Add all files to the Cart

Remove all from the Cart

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	44ca855-4056-4f6e-c320003c0f4.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	523.49 KB	0
open	de70478-4915-4749-8729-e6ee06a0c359.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.23 KB	0
open	432d963-45d0-46c0-96d5-c7d9f6ec3b6.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507 KB	0
open	996000c-0804-4559-8084-c5059838622a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.53 KB	0
open	02e17d87-c0dc-4051-a2e5-85eab14591b.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	533.83 KB	0
open	8dfe19ae-acc1-4de5-8694-1a1805c11c5.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	536.46 KB	0
open	ab2d5a3d-4253-4c80-b4e5-0d8fe163140.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	514.74 KB	0
open	59325a83-799e-4f58-aaf2-a29b34e7e563.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	495.24 KB	0
open	5498129-1f2c-458b-ba7f-4c76ea2538cc.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	521.16 KB	0
open	74118954-46ea-4099-9709-45175603a1.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.52 KB	0
open	36bcb02-58b0-4731-5a2c-09eaaac15c29.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.12 KB	0
open	b8e7b7b5-54e8-4459-5a21-ea3472b013d7.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	506.3 KB	1
open	57e8b21e-eeb7-4ffe-a8b3-a6018d3c7d96.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	522 KB	0
open	e6ba3664-1968-4f6e-5a11-0157a9d81.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.07 KB	0
open	edeae45b-a454-4a7b-9695-38d29424698.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	542.11 KB	0
open	c4489317-3e8a-4159-b32b-4d9e723359a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.35 KB	0
open	b638ae-65f9cc-42d2-bd7f-5dc14e8b372.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507.88 KB	0
open	0626f07-749e-4788-a815-500b9ca778.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	511.74 KB	0
open	36440974-4445-4728-b08f-10a0329d06.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	519.12 KB	0
open	a18f26b8-9497-4824-a9ba-7b1925ca987.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507.86 KB	0

- Step 5: Click the "Cart"

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart 1,222 GDC Apps

Files (1,222) Cases (1,092)

Primary Site Project Data Category Data Type Data Format

Showing 1 - 20 of 1,222 files 635.71 MB

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	44ca0b5-005f-496f-b8ec-c32d0003c594.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	523.49 KB	0
open	de704076-e915-4749-9729-e6ee0a60d359.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.23 KB	0
open	432d8863-b5d0-46c0-96b8-d705f6dc3b6.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507 KB	0
open	996000bc-0804-4559-80bf-c5058838022a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.53 KB	0
open	02e47987-c6c8-4051-a2e6-85ab814591b.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	533.83 KB	0
open	8dfe19ae-aec1-4d65-8d64-1a480c5c11c5.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	536.46 KB	0
open	ab2d5a3d-4253-4c80-b6f6-d38f6163140.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	514.74 KB	0
open	59325a83-799e-4f08-aaf2-a20b34e7e563.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	495.24 KB	0
open	54f68129-1f2c-458b-ba6f-4c78ea2538cc.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	521.16 KB	0
open	7411d895-b4ea-4099-9709-45175603af1.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.52 KB	0
open	36bc602-58b0-4731-8d2c-090eac15d29.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.12 KB	0
open	b9e7b7b5-54e8-4459-8a21-ea3472b013d7.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	506.3 KB	1
open	57e8b21e-eeb7-4f8e-a8b3-b6016b37d3d6.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	522 KB	0
open	e8ba3864-1068-4f8b-9a11-01f37a6d81.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.07 KB	0
open	ed8a4e5b-ad4d-4a0b-9656-3a82a424698.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	542.11 KB	0
open	c4489317-3e8a-4159-b32b-4c9e723359a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.35 KB	0
open	b938ae6d-f8cc-42d2-bc4f-5dc14e8b372.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507.98 KB	0
open	0826cf7-749e-4788-a814-f506b9a77f.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	511.74 KB	0
open	36d40974-444f-4726-b88f-1c0ab329d0df.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	519.12 KB	0
open	a1826c8-9497-4824-a5ba-7b1924ca887.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507.86 KB	0

- Step 6: Download clinical and sample information in “tsv” format

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart 1,222 GDC Apps

Files (1,222) Cases (1,092)

Primary Site Project Data Category Data Type Data Format

Showing 1 - 20 of 1,222 files 635.71 MB

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	44ca0b5-005f-496f-b8ec-c32d0003c594.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	523.49 KB	0
open	de704076-e915-4749-9729-e6ee0a60d359.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.23 KB	0
open	432d8863-b5d0-46c0-96b8-d705f6dc3b6.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507 KB	0
open	996000bc-0804-4559-80bf-c5058838022a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.53 KB	0
open	02e47987-c6c8-4051-a2e6-85ab814591b.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	533.83 KB	0
open	8dfe19ae-aec1-4d65-8d64-1a480c5c11c5.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	536.46 KB	0
open	ab2d5a3d-4253-4c80-b6f6-d38f6163140.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	514.74 KB	0
open	59325a83-799e-4f08-aaf2-a20b34e7e563.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	495.24 KB	0
open	54f68129-1f2c-458b-ba6f-4c78ea2538cc.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	521.16 KB	0
open	7411d895-b4ea-4099-9709-45175603af1.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.52 KB	0
open	36bc602-58b0-4731-8d2c-090eac15d29.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.12 KB	0
open	b9e7b7b5-54e8-4459-8a21-ea3472b013d7.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	506.3 KB	1
open	57e8b21e-eeb7-4f8e-a8b3-b6016b37d3d6.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	522 KB	0
open	e8ba3864-1068-4f8b-9a11-01f37a6d81.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.07 KB	0
open	ed8a4e5b-ad4d-4a0b-9656-3a82a424698.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	542.11 KB	0
open	c4489317-3e8a-4159-b32b-4c9e723359a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.35 KB	0
open	b938ae6d-f8cc-42d2-bc4f-5dc14e8b372.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507.98 KB	0
open	0826cf7-749e-4788-a814-f506b9a77f.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	511.74 KB	0
open	36d40974-444f-4726-b88f-1c0ab329d0df.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	519.12 KB	0

- Step 7: Download the manifest file for all files in the cart.

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart 1,222 GDC Apps

Files (1,222) Cases (1,092)

Primary Site Project Data Category Data Type Data Format

Showing 1 - 20 of 1,222 files 635.71 MB

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	44ca0b5-005f-496f-b8ec-c32d0003c594.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	523.49 KB	0
open	de704076-e915-4749-9729-e6ee0a60d359.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.23 KB	0
open	432d8863-b5d0-46c0-96b8-d705f6dc3b6.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507 KB	0
open	996000bc-0804-4559-80bf-c5058838022a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.53 KB	0
open	02e47987-c6c8-4051-a2e6-85ab814591b.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	533.83 KB	0
open	8dfe19ae-aec1-4d65-8d64-1a480c5c11c5.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	536.46 KB	0
open	ab2d5a3d-4253-4c80-b6f6-d38f6163140.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	514.74 KB	0
open	59325a83-799e-4f08-aaf2-a20b34e7e563.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	495.24 KB	0
open	54f68129-1f2c-458b-ba6f-4c78ea2538cc.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	521.16 KB	0
open	7411d895-b4ea-4099-9709-45175603af1.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.52 KB	0
open	36bc602-58b0-4731-8d2c-090eac15d29.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	510.12 KB	0
open	b9e7b7b5-54e8-4459-8a21-ea3472b013d7.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	506.3 KB	1
open	57e8b21e-eeb7-4f8e-a8b3-b6016b37d3d6.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	522 KB	0
open	e8ba3864-1068-4f8b-9a11-01f37a6d81.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	529.07 KB	0
open	ed8a4e5b-ad4d-4a0b-9656-3a82a424698.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	542.11 KB	0
open	c4489317-3e8a-4159-b32b-4c9e723359a.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	512.35 KB	0
open	b938ae6d-f8cc-42d2-bc4f-5dc14e8b372.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	507.98 KB	0
open	0826cf7-749e-4788-a814-f506b9a77f.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	511.74 KB	0
open	36d40974-444f-4726-b88f-1c0ab329d0df.FPKM.txt.gz	1	TCGA-BRCA	Transcriptome Profiling	TXT	519.12 KB	0



- Step 8: Install the **GDC Data Transfer Tool**. The GDC Data Transfer Tool can be downloaded from <https://gdc.cancer.gov/access-data/gdc-data-transfer-tool>. In Windows, the GDC Data Transfer Tool is used in the terminal. First, go to the directory where “gdc-clinet.exe” locates in command line. Then it should be ready for used in the terminal. An alternative could be adding location of the “gdc-client.exe” to the environment variables.

```
D:\data\TCGA_download_example\gdc-client>ls
gdc-client.exe

D:\data\TCGA_download_example\gdc-client>gdc-client download --help
usage: gdc-client download [-h] [--debug] [--log-file LOG_FILE] [--color_off]
                          [-t TOKEN_FILE] [-d DIR] [-s server]
                          [--no-segment-md5sums] [--no-file-md5sum]
                          [-n N_PROCESSES]
                          [--http-chunk-size HTTP_CHUNK_SIZE]
                          [--save-interval SAVE_INTERVAL] [-k]
                          [--no-related-files] [--no-annotations]
                          [--no-auto-retry] [--retry-amount RETRY_AMOUNT]
                          [--wait-time WAIT_TIME] [--latest] [--config FILE]
                          [-m MANIFEST]
                          [file_id [file_id ...]]
```

- Step 9: Download all files in the cart using the GDC Data Transfer Tool with the manifest file. A user manual can be found in [https://docs.gdc.cancer.gov/Data Transfer Tool/Users Guide/Data Download and Upload/](https://docs.gdc.cancer.gov/Data_Transfer_Tool/Users_Guide/Data_Download_and_Upload/). More details about the GDC Data Transfer Tool's download functionality can be found in the “Download help menu” section.

The command below shows downloading all files listed in the manifest file to a directory specified by the location following the -d option. The location and name of the manifest file follow the -m option.

```
gdc-client download -d D:\data\TCGA_download_example\ -m
D:\data\TCGA_download_example\gdc_manifest_201912012_014201.txt
```

```
D:\data\TCGA_download_example>gdc-client download -d D:\data\TCGA_download_example\ -m D:\data\TCGA_download_example\gdc_manifest_201912012_014201.txt
100% [#####]
N/A%
```

- Step 10: Expression data and clinical information can be matched with samples under the relation between the manifest file, sample sheet and clinical file. In the sample sheet, samples and the names of their expression data files listed in manifest file are given.

Other datasets which are not obtained from the TCGA official website are well-organized tables and can be downloaded directly from the given links.

## Pre-processing steps

- In TCGA cancer cohorts, patients with missing survival information are removed.
- Features with over 20% missing values are removed in mRNA, miRNA, and DNA methylation data downloaded from TCGA.
- Missing values are filled with zeros in mRNA, miRNA, and DNA methylation data.
- Top 2000 most variant features are selected in copy number estimate, mRNA and miRNA expression, DNA methylation, and proteins and phosphoproteins' relative intensities.

- mRNA and miRNA expression are transformed by  $\log_2(x+1)$ .
- Copy number estimate are normalized to the range of [0, 1].
- All features in proteomics and phosphoproteomics are divided by their maximum for normalization.

## Running examples

In the repository, we give three examples to show how to run **DeFusion**.

```
[X, Z, E, convergence] = DeFusion(dataCell, lowDim, alpha, gamma, K, fout)
% @Input:
% dataCell: multiple data matrices stored a cell.
% lowDim: the number of dimensionality of latent sample representation.
% alpha and gamma: Parameters in DeFusion
% K: parameter for NE and SNF, usually set to be 20.
% fout: data path to save the output.

% @Output
% X: latent sample representation, a N x lowDim matrix.
% Z: latent variables of features, a cell.
% convergence: loss of objective function in each iteration, a structure array.
```

- A toy example

- run script 'runToyExample.m' to attain latent sample representation.
- run script 'plotToyExample.m' for visualization.

- Integration of proteomics and phosphoproteomics data

- run script 'runProPhosphoIntegration.m' to attain latent sample representation.
- run script 'visualizeProPhosphoIntegration.m' for visualization.

- Integration of transcriptomics and epigenomics data

In this example, we fuse mRNA and miRNA expression and DNA methylation in TCGA-LIHC.

- run script 'runLIHC.m' to obtain latent sample representation.
- run script 'LIHC3CV.py' to train a Cox Proportional-Hazards model in three-fold cross-validation.