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

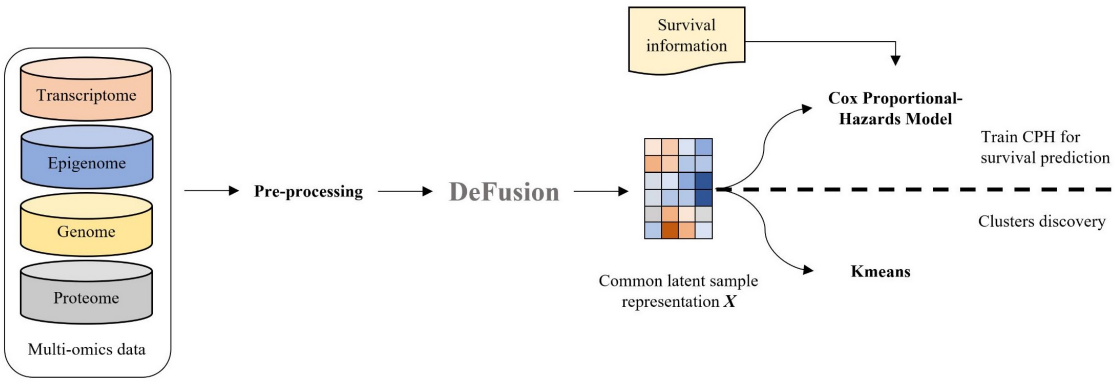
- survival 2.44.1
- survcomp 1.32.0

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 [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$_{+cnv}$.

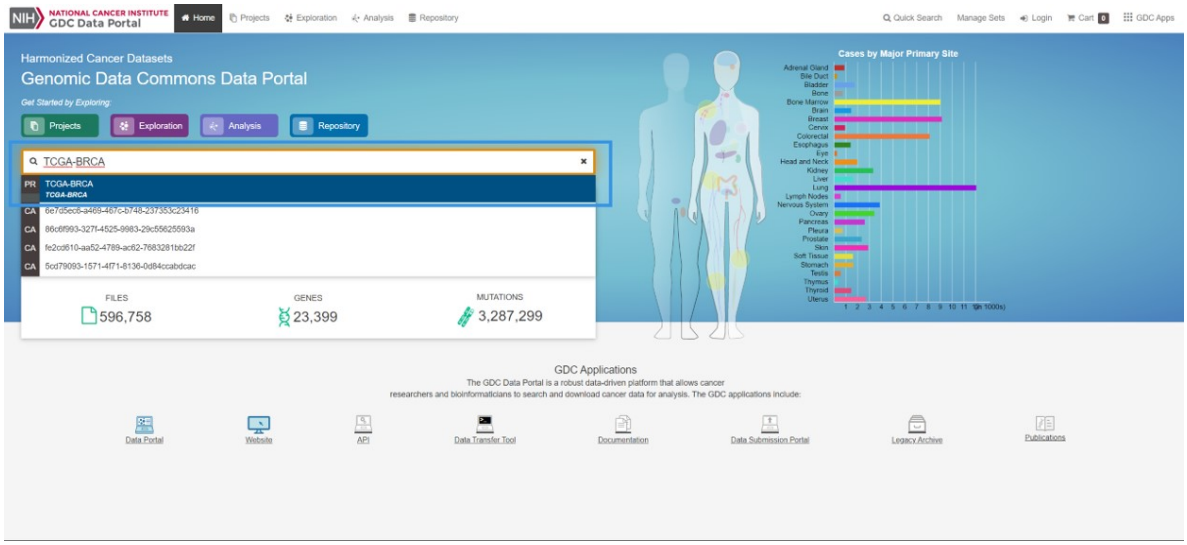
We have tried proteomics-and-phosphoproteomics integration by using proteomics and phosphoproteomics data of normal and tumor tissues.

> [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.

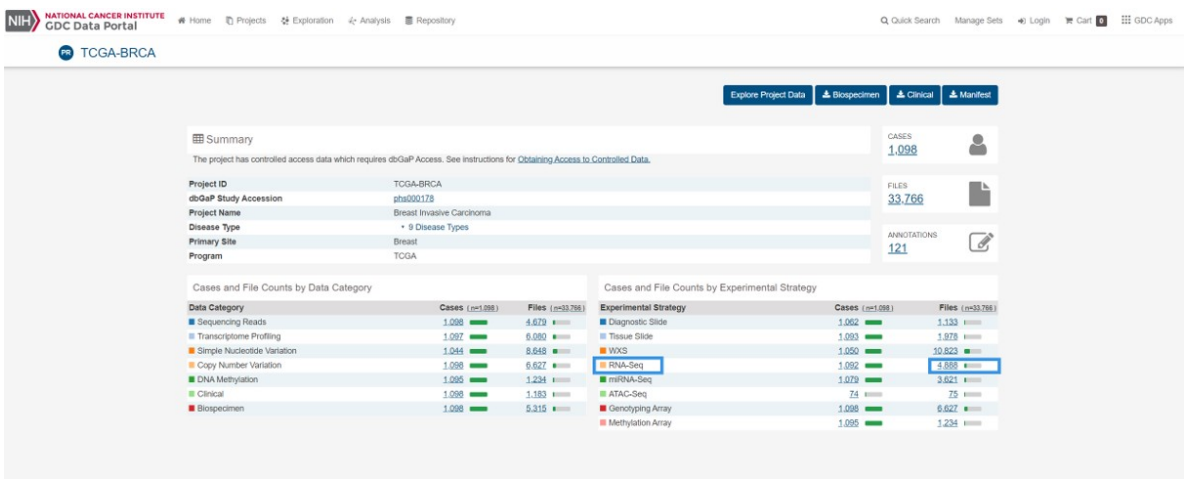
Downloading data from the TCGA official website (https://portal.gdc.cancer.gov/)

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.

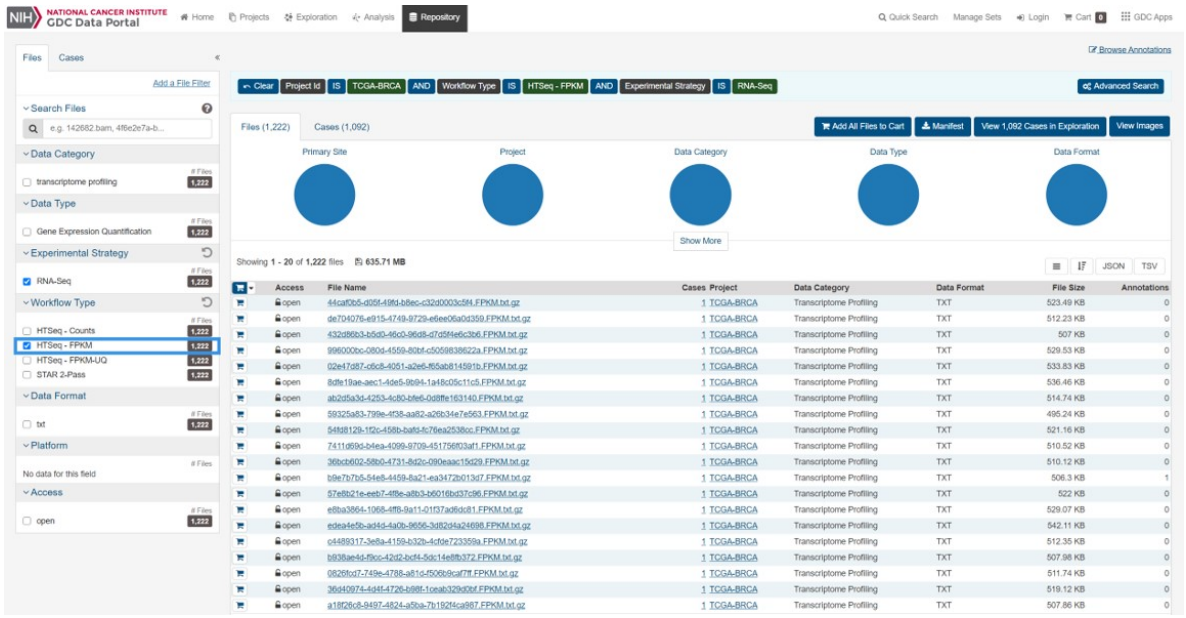
- Step 1: Search for TCGA projects in the main page



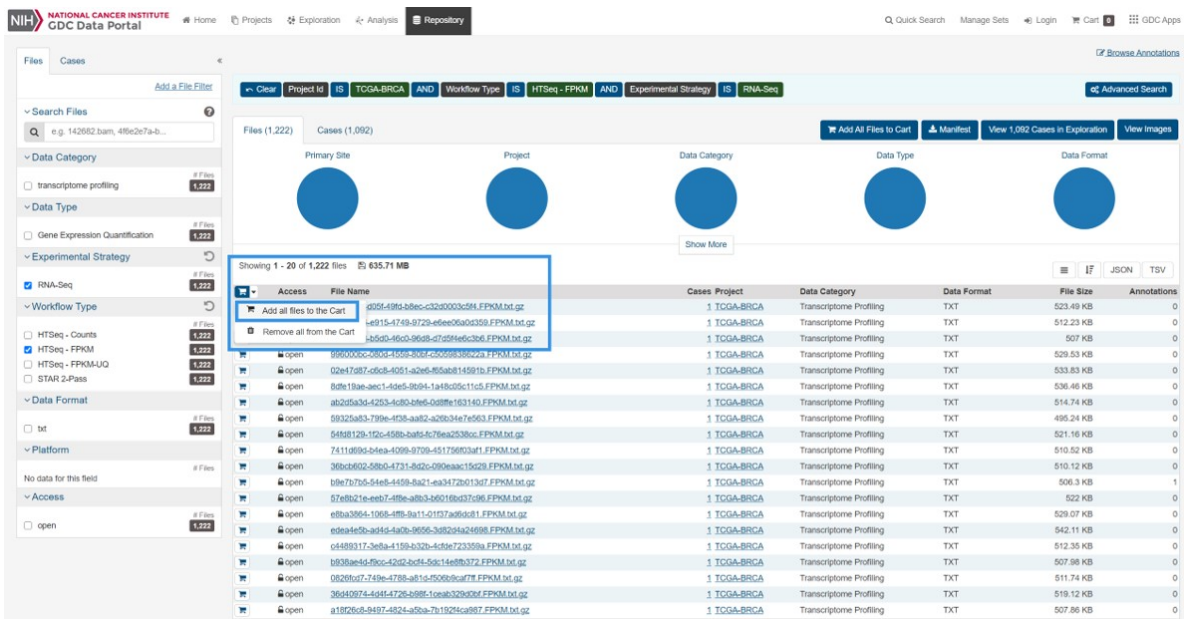
- Step 2: Click the link in the ‘Files’ column corresponding to the RNA-seq



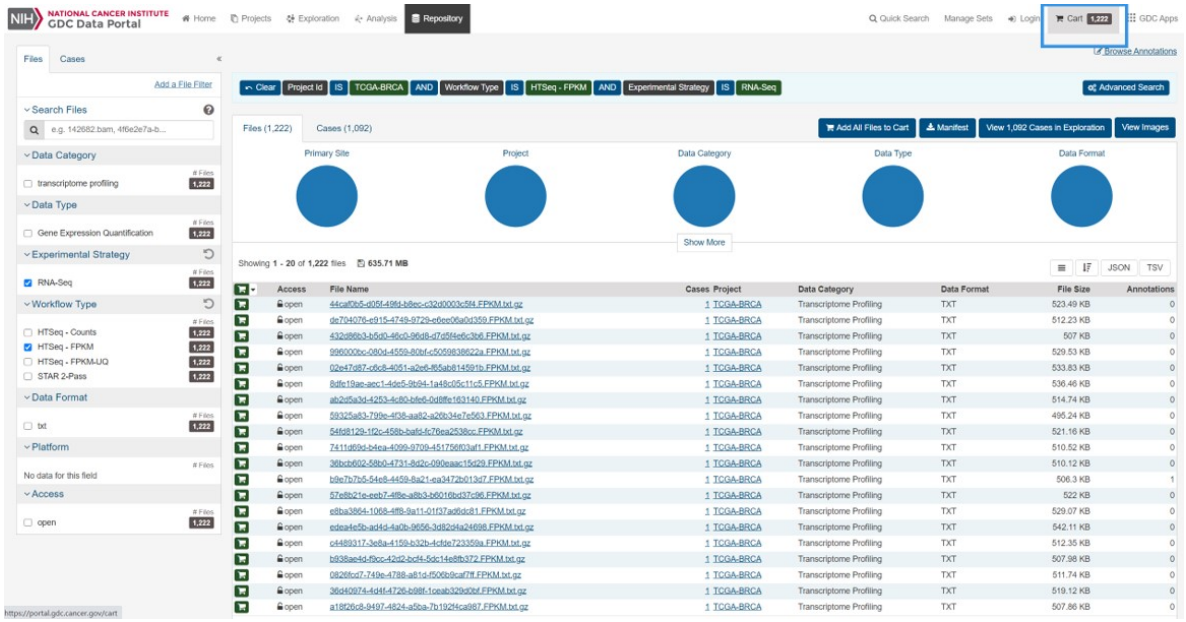
- Step 3: Check the “HTSeq-FPKM” checkbox



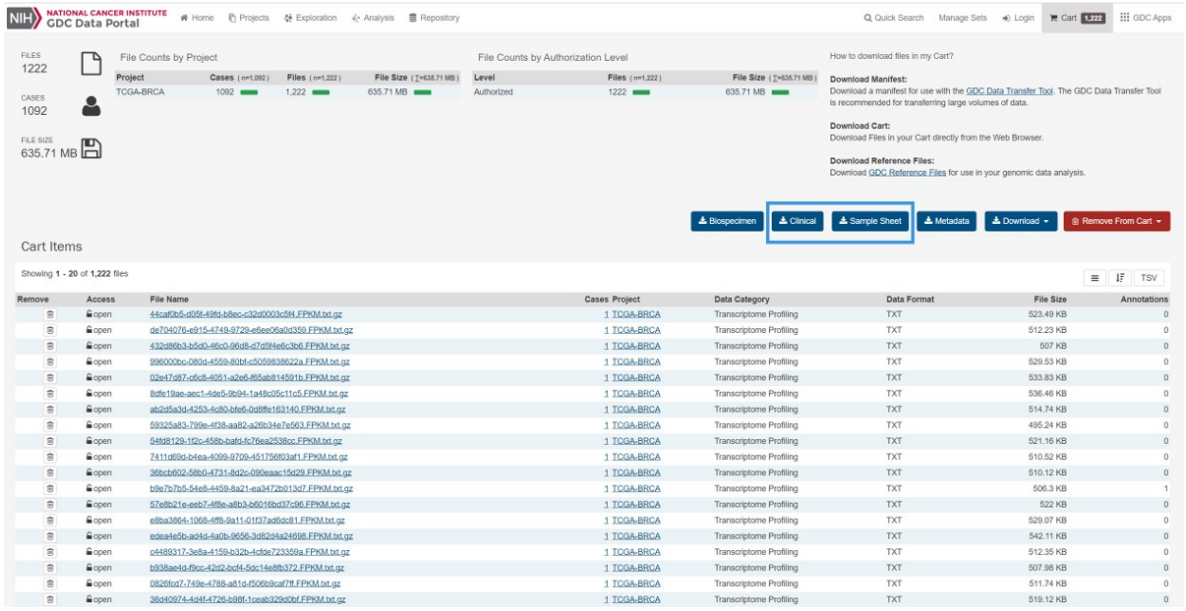
- Step 4: Add all files to the cart



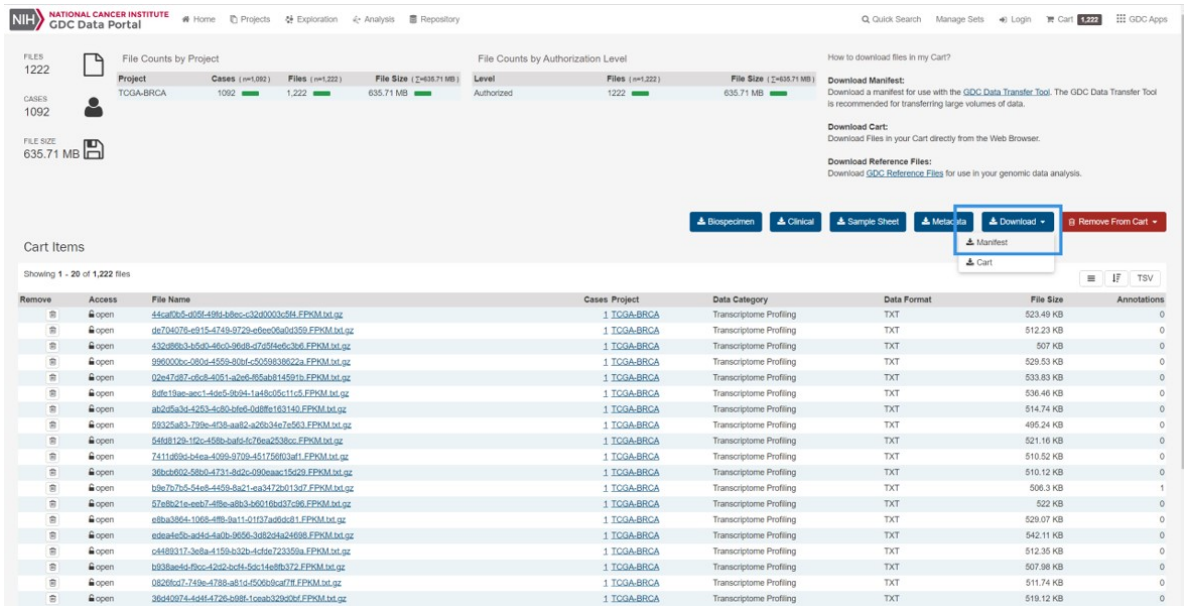
- Step 5: Click the “Cart”



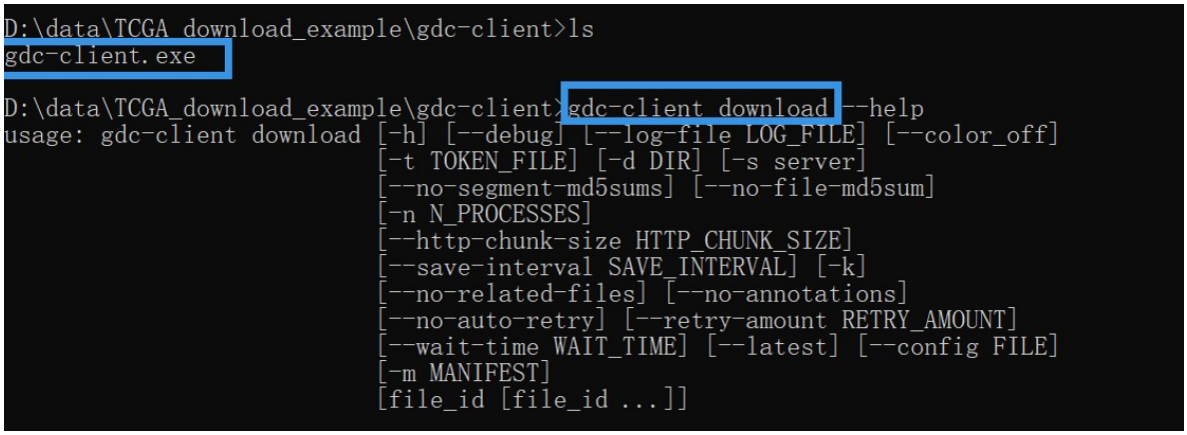
- Step 6: Download clinical and sample information in “tsv” format



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



- 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.



- 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/. 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
```

- 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, paitents 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 log2(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.