JASA ACS Reproducibility Initiative - Author Contributions Checklist Form

The purpose of the Author Contributions Checklist (ACC) Form is to document the code and data supporting a manuscript, and describe how to reproduce its main results.

As of Sept. 1, 2016, the ACC Form must be included with all new submissions to JASA ACS.

This document is the initial version of the template that will be provided to authors. The JASA Associate Editors for Reproducibility will update this document with more detailed instructions and information about best practices for many of the listed requirements over time.

Data

Abstract (Mandatory)

We collected 22 major cancer types with sufficient patients from the The Cancer Genome Atlas (TCGA) project, where three molecular profiles are used: RNA expression (RNA-seq V2), miRNA, and copy number alterations (CNA).

Availability (Mandatory)

These datasets are all Open Access and available in the Genomic Data Commons (GDC). The 22 TCGA cancer data sets saved in the matlab file can be downloaded from the dropbox directory https://www.dropbox.com/s/v22fx0j2gnpeta6/all_data.mat?dl=0.

Description (Mandatory if data available)

The results in this study are based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. All the datasets in this paper are Open Access, the link to these datasets is https://portal.gdc.cancer.gov/. To avoid batch effects due to use of different platforms, we only consider molecular profiles from one platform. These datasets consist of:

- De-identified clinical and demographic data
- RNA expression (RNA-seq V2): measured using the Illumina sequencing technology with the log2(x + 1) transformed RSEM (RNA-Seq by Expectation Maximization) values.
- miRNA: miRNA mature strand expression RNAseq datasets measured by Illumina miRNA-seq.
- Copy Number Alterations (CNA): estimated using the GISTIC2 threshold method.

The original file formats of these datasets are in .txt format. We collected 22 major cancer types with sufficient patients from the TCGA project (downloaded in October 2017). Patients with missing molecular profiles were removed, with a total of 6,976 patients included in the clustering and survival analysis as follows:

• Bladder Cancer (BLCA), n (sample size) =400.
• Breast Invasive Carcinoma (BRCA), n=741.

• Cervical Squamous Cell Carcinoma (CESC), n=289.
• Colon Adenocarcinoma (COAD), n=249.
• Esophageal Cancer (ESCA), n=180.
• Head and Neck Squamous Cell carcinoma (HNSC), n=471.
• Kidney Renal Clear cell carcinoma (KIRC), n=236.
• Kidney Papillary Cell Carcinoma (KIRP), n=283.
• Lower Grade Glioma (LGG), n=505.
• Liver Cancer (LIHC), n=357.
• Lung Adenocarcinoma (LUAD), n=442.
• Lung Squamous Cell carcinoma (LUSC), n=333.
• Mesothelioma (MESO), n=87.
• Ovarian Cancer (OV), n=297.
• Pancreatic Cancer (PAAD), n=176.
• Prostate Adenocarcinoma (PRAD), n=484.
• Rectum Adenocarcinoma (READ), n=87.
• Sarcoma (SARC), n=250.
• Stomach Adenocarcinoma (STAD), n=365.
• Thyroid Cancer (THCA), n=494.
• Uterine Corpus Endometrioid Carcinoma (UCEC), n=170.
• Uveal Melanoma (UVM), n=80.

The matlab files for these datasets including the following information are also available at https://github.com/ishspsy/MKerW-A :

- “all_clin" includes clinical information of the patients from the 22 cancer types.
- “all_exp" is the RNA data for all the patients.
- “all_mirna" is the MicroRNA data for all the patients.
- “all_cna" is the CNA data for all the patients.
- “all_pat” is the index vector of patients indicating corresponding cancer types (See the first column of “all_clin" for the original cancer name).

Optional Information (complete as necessary)

Code

Abstract (Mandatory)
The proposed algorithm *MKerW-A* learns the weight of each data type as well as a similarity measure between patients via a non-convex optimization framework. It solves the proposed non-convex problem iteratively using the ADMM algorithm. The codes of *MKerW-A* were mainly performed on the Matlab.

Description (Mandatory)
The codes of *MKerW-A* were mainly delivered with the Matlab, while R package was only used to generate the survival curves. The project is licensed under the MIT License. The link to the codes is https://github.com/ishspsy/MKerW-A.

Optional Information (complete as necessary)
Most of the simulations and TCGA data applications could be implemented on an Apple Mac-Book Pro (2.7 GHz, 8 GB of memory) using the MATLAB 2016b.

Instructions for Use

Reproducibility (Mandatory)
All figures from this paper can be reproduced based on the codes and data files provided at https://github.com/ishspsy/MKerW-A. The detailed description of implementation can be also found at this link. The following are details for reproducing the results:

- All the functions used in the proposed algorithm MKerW-A are located in the directory “Main_Code”.
- All the other supplementary files are located in the directory "Other_codes".
- All the codes generating figures presented in the manuscript are located in the directory “Generating_Figures".
- All the resulting files (e.g., .MAT and .eps) are located in the directory "Results_files".