Author Contributions Checklist Form

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Part 1: Data

☐ This paper **does not** involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

In the data set, the gene expression analysis was performed for 156 samples from the enrolled patients in the NOAH trail, which consisted of 114 patients with HER2+ locally advanced or inflammatory breast cancer and 42 patients with HER2- disease.

Availability

☑ Data **are** publicly available

☐ Data **cannot be made** publicly available

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available dat*a section, below.

Publicly available data

☑ Data are available online at:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50948

☐ Data are available as part of the paper's supplementary material.

☐ Data are publicly available by request, following the process described here:

☐ Data are or will be made available through some other mechanism, described here:

Non-publicly available data

Discussion of lack of publicly available data:

Description

File format(s)

☐ CSV or other plain text:
☐ Software-specific binary format (.Rda, Python pickle, etc.):
☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
☑ Other (described here):

The data can be downloaded directly in R, as shown in application.R.

Data dictionary

☐ Provided by the authors in the following file(s):
☐ Data file(s) is (are) self-describiing (e.g., netCDF files)
☑ Available at the following URL:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50948

Additional information (optional)

Part 2: Code

Abstract

We include the necessary codes for all the simulations and application. Specifically, for simulation study in Section 5.1, and Appendix D.1-D.2 of the Supplementary Material, the data is generated based on Main_code.R and Ser.R, and the figures are derived based on plot_main.R, plot_kernel.R, and the files in the folder D2_Plot_power. For the application in Section 5.2, the result is obtained by application.R. For simulation study in Appendix D.3- D.5 of the Supplementary Material, the results are derived based on the files in the folder Supplementary. We encourage readers to read the README file accompanying the code for more details.

Description

Code format(s)

☑ Script files

- ☑ R ☐ Python ☐ Matlab
- ☐ Other:

☐ Package

- ☐ R ☐ Python ☐ MATLAB toolbox
- ☐ Other:

☐ Reproducible report

- ☐ R Markdown ☐ Jupyter notebook
- ☐ Other:

☐ Shell script

☐ Other (described here):

Supporting software requirements

Version of primary software used

R version 4.1.2

Libraries and dependencies used by the code

glmnet version 4.1.4
parallel version 4.1.2
Biobase version 2.54.0
GEOquery version 2.62.2
stingr version 1.4.0
ggplot2 version 3.3.6
extrafont version 0.18
gridExtra version 2.3

Supporting system/hardware requirements (optional)

Parallelization used

☐ No parallel code used
☑ Multi-core parallelization on a single machine/node
Number of cores used: 50
☐ Multi-machine/multi-node parallelization
Number of nodes and cores used:

License

☐ MIT License (default)
☐ BSD
☐ GPL v3.0
☐ Creative Commons
☐ Other (described here):

Additional information (optional)

Part 3: Reproducibility workflow

Scope

The provided workflow reproduces:

☐ Any numbers provided in text in the paper

☑ The computational method(s) presented in the paper (i.e., code is provided that implements the method(s))

☑ All tables and figures in the paper

☐ Selected tables and figures in the paper, as explained and justified here:

Workflow details

Format(s)

☐ Single master code file

☐ Wrapper (shell) script(s)

☐ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach

☐ Text file (e.g., a readme-style file) that documents workflow

☐ Makefile

☑ Other (more detail in 'Instructions' below)

Instructions

The application result in Section 5.2 is obtained by directly running application.R. The workflows of simulation studies follow a similar procedure, hence the workflow of the main simulation study in Section 5.1 and Appendix D.1-D.2 of the Supplemantray Material is introduced as an exmaple. The simulation study is based on Main_code.R and Ser.R, where parallel computing is used. According to the setting of simulation, corresponding parameters need to be specified in Main_code.R before running Ser.R. We encourage readers to read the README file accompanying the code for more details.

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

☐ <1 minute

☐ 1-10 minutes
☐ 10-60 minutes
☐ 1-8 hours
☑ >8 hours
☐ Not feasible to run on a desktop machine, as described here:

Additional documentation (optional)

Notes (optional)