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

All external datasets come from Cancer DepMap portal, which are open datasets. The Cancer Dependency Map contains an extensive collection of genomics data, including measurements of gene expression, RNAi and CRISPR dependency, and drug sensitivity gathered from over 1000 cancer cell types.

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:
☒ 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):

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:

In the Github repository https://github.com/BobZiyangDing/FAB-Correlation-Structure-Testing , datasets are in the following directories:

- ./data/coexpressed_genes.csv : This is the correlated gene pairs we use as ground truth for real data experiment as presented in section 5 of our manuscript
- ./data/depmap_crispr_data.csv: This is this gene expression for breast cancer gene expression dataset used in section 5, case 1, as the auxiliary dataset
- ./data/depmap_rnai_data.csv: This is this gene expression for breast cancer gene expression dataset used in section 5, case 1, as the testing dataset
- ./data/depmap_breast_gene_expression_data.csv: This is this gene expression for breast cancer gene expression dataset used in section 5, case 2, as the auxiliary dataset
- ./data/depmap_lung_gene_expression_data.csv: This is this gene expression for lung cancer gene expression dataset used in section 5, case 2, as the testing dataset

Additional information (optional)

Part 2: Code

Abstract

We provide R scripts and Rmd scripts. The R scirpts include algorithms, figure generators, and simulated data generators. Rmd scipts doesn’t carry any such functionality but calls functions in R scripts to form reproducibility files. All results in the the paper are generated using Rmd files. The knitted html outputs of Rmd files are also in the github repository.

Description

Code format(s)

☒ Script files

☒ R ☐ Python ☐ Matlab

☐ Other: Rcpp

☒ 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

We use R version 4.0.5

Libraries and dependencies used by the code

packageVersion("FedData") = 2.5.7

```
packageVersion("Rcpp")            = 1.0.7
packageVersion("fabricatr")       = 0.14.0
packageVersion("mltools")         = 0.3.5
packageVersion("onehot")          = 0.1.1
packageVersion("mgcv")            = 1.8.34
packageVersion("tidyverse")       = 1.3.1
packageVersion("foreach")         = 1.5.1
packageVersion("doParallel")      = 1.0.16
packageVersion("kableExtra")      = 1.3.4
packageVersion("boot")            = 1.3.27
packageVersion("tensorr")         = 0.1.1
packageVersion("reshape2")        = 1.4.4
packageVersion("dvmisc")          = 1.1.4
packageVersion("tidyverse")       = 1.3.1
packageVersion("hash")            = 2.2.6.1
packageVersion("adaptMT")         = 1.0.0
packageVersion("sgof")            = 2.3.2
packageVersion("HDtweedie")       = 1.1
packageVersion("ggpubr")          = 0.4.0
packageVersion("latex2exp")       = 0.5.0
packageVersion("gridExtra")       = 2.3
packageVersion("cowplot")         = 1.1.1
packageVersion("grid")            = 4.0.5
packageVersion("MASS")            = 7.3.53.1
packageVersion("matrixcalc")      = 1.0.5
packageVersion(“clusterGeneration") = 1.3.7
packageVersion("pracma")          = 2.3.3
packageVersion("Matrix")          = 1.3.2
packageVersion("tidyverse")       = 1.3.1
packageVersion("foreach")         = 1.5.1
packageVersion("doParallel")      = 1.0.16
packageVersion("hash")            = 2.2.6.1
packageVersion("MLmetrics")       = 1.1.1
packageVersion("abind")           = 1.4.5
packageVersion(“wesanderson”) = 0.3.6
```

Supporting system/hardware requirements (optional)

Parallelization used

☐ No parallel code used
☒ Multi-core parallelization on a single machine/node
Number of cores used: 16
☐ 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)

The repo is at https://github.com/BobZiyangDing/FAB-Correlation-Structure-Testing
Feel free to clone
git clone https://github.com/BobZiyangDing/FAB-Correlation-Structure-Testing.git

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

Location

The workflow is available:
☐ As part of the paper's supplementary material
☒ In this Git repository: https://github.com/BobZiyangDing/FAB-Correlation-Structure-Testing
☐ Other:

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

To reproduce individual simulation section results: open individualExperiments.Rmd, simply run all chunks
To reproduce real data section results: open real_data_experiments.Rmd, simply run all chunks

To reproduce table and bootstrap grid figure in our paper: open MassiveExperiments.Rmd, simply run all chunks

Please note: please restart R session and clear all output everytime run another Rmd file. Otherwise, results could be slightly different.

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)

Packages don't need to be installed one by one. Just install package "FedData" by Install.package("FedData")

Then, all the rest packages will be auto-installed and loaded once running the Dependency.R file. This dependency R file will also be run everytime running any Rmd file.

Notes (optional)

For parallel computing: the scripts will automatically choose the maximum-1 cores available for parallel computing. Thus, no manual adjustment is needed

Timing for reproducible files: MassiveExperiments.rmd will take the most time to run, as the same algorithm is repeatedly run many times to form an averaged result. The rest 2 files can be reproduced relatively quickly.