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

BrainSpan Dataset (Microarray samples)

- **Abstract**
    - This dataset contains microarray gene expression samples from brain tissue of varying spatiotemporal properties. This dataset was first used in “Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. Nature, 478(7370):483–489.”
- **Availability**
    - The specific dataset we use is derived from a publicly available dataset (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25219). Our lab restructured that data so it would be amendable for a convenient analysis in R.
- **Description**
    - Permissions: The original data is publicly available. Future authors who wish to use this dataset should cite Kang et al. (2011).
    - Licensing information: None
    - Link to data: https://github.com/linnylin92/covarianceSelection/raw/master/newGenexp.RData
    - Data provenance: As stated above, the original data is posted at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25219.
    - File format: R Data format
    - Metadata: Provided in more detail in https://github.com/linnylin92/covarianceSelection/blob/master/covarianceSelection/R/data.R, specifically for “brainspan_id” (stating information on the brains used to gather this data).
    - Version information: None

TADA Dataset (Risk scores) – De Rubeis et al. (2014)

- **Abstract**
    - This dataset contains the risk scores (p-values) for each gene. This dataset was derived by our lab using the methodology in “He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M.

J., Buxbaum, J. D., et al. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genetics, 9(8):e1003671" on the dataset published in "De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature, 515(7526):209–215."

- **Availability**
    - The dataset is public.
- **Description**
    - Permissions: Future authors who wish to use this dataset should cite De Rubeis et al. (2014).
    - Licensing information: None
    - Link to data: https://github.com/linnylin92/covarianceSelection/raw/master/covarianceSelection/data/tada.rda
    - Data provenance: The dataset that we derive our p-value calculations from are located in https://www.nature.com/articles/nature13772 (primarily Supplementary Table 2).
    - File format: R Data format
    - Metadata: Provided in more detail in https://github.com/linnylin92/covarianceSelection/blob/master/covarianceSelection/R/data.R, specifically for "tada".
    - Version information: None

TADA Dataset (Risk scores) – Satterstrom et al. (2019)

- **Abstract**
    - This dataset contains the risk scores (p-values) for each gene. This dataset was derived by our lab using the methodology in "He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., Buxbaum, J. D., et al. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genetics, 9(8):e1003671" on the dataset published in "Satterstrom, F. Kyle, et al. "Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism." (2019)."
- **Availability**
    - The dataset (a list of 102 genes) is shown in Figure 4E in https://www.biorxiv.org/content/biorxiv/early/2019/04/24/484113.full.pdf. A more user-friendly version will be available when the paper is published.
- **Description**
    - Permissions: Future authors who wish to use this dataset should cite Satterstrom et al. (2019).
    - Licensing information: None

  - Link to data: https://github.com/linnylin92/covarianceSelection/raw/master/covarianceSelection/data/validated_genes.rda
  - Data provenance: The dataset is located in Figure 4E in https://www.biorxiv.org/content/biorxiv/early/2019/04/24/484113.full.pdf
  - File format: R Data format
  - Metadata: Provided in more detail in https://github.com/linnylin92/covariance_selection/blob/master/covarianceSelection/R/data.R, specifically for “validated_genes”.
  - Version information: None

Additional information can be found in the R documentation we created, https://github.com/linnylin92/covarianceSelection/blob/master/covarianceSelection/R/data.R.

Code

- **Abstract**
  - We provide the code to reproduce the simulations, analysis, results and figures, all located and documented in our GitHub repository, https://github.com/linnylin92/covarianceSelection.

- **Description**
  - Delivery: Our functions critical for the analysis are bundled into an R package called covarianceSelection, and we provide and document the R code to reproduce the simulations, analysis, results and figures.
  - Licensing information: MIT +file LICENSE
  - Link to code/repository: https://github.com/linnylin92/covarianceSelection
  - Version information: Currently at commit #275, possibly higher at the time of review for minor fixes/cleanup and updates in the README.

- **Optional Information**
  - Hardware requirements: Computer that can run Git, R and install the necessary R packages. Our analysis was run on a server with 10 cores, but our GitHub README provides instructions to tweak the number of cores used if needed.
  - The procedure we develop in our paper (Section 6) has two parts: the Stepdown method and the largest quasi-clique method. The Stepdown procedure shown in https://github.com/linnylin92/covarianceSelection/blob/master/main/step4_subjectselection.R, which takes about 8 hours to complete on the server for our analysis, while the quasi-clique method takes a few minutes.
  - Supporting software requirements: Our R package (and subsequent analyses, figures, etc.) depend on the following R packages:
    - DBI (1.0.0)
    - dequer (2.0.1)

    - devtools (2.2.1)
    - doMC (1.3.6)
    - foreach (1.4.7)
    - glmnet (3.0)
    - hash (2.2.6.1)
    - huge (1.3.4)
    - igraph (1.2.4.1)
    - MASS (7.3.51.4)
    - Matrix (1.2.17)
    - org.Hs.eg.db (3.8.2)

Instructions for Use

Reproducibility
All the simulations, analyses, results and figures can be reproduced. The specific instructions are provided in the README file in our GitHub repository (https://github.com/linnylin92/covarianceSelection/blob/master/README.md).

Replication
Since we provide numerous unit tests for our code, it should be suitable for other analyses. The functions in the NAMESPACE of our R package (https://github.com/linnylin92/covarianceSelection/blob/master/covarianceSelection/NAMESPACE) are meant to be able to used by other researchers, and we have provided documentation for each of these functions.