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

Abstract

This paper uses the 16S rRNA sequencing data from the T1D cohort of the DIABIMMUNE study (Kostic et al., 2015). The sample-level data, including the OTU table, taxonomy table and the covariate information, is available at https://diabimmune.broadinstitute.org/diabimmune/t1d-cohort/resources/16s-sequence-data. The (subject-level) cohort information is available at https://www.cell.com/cms/10.1016/j.chom.2015.01.001/attachment/1f0883f8-1df7-447d-a47b-c1aa2bb2bbaf/mmc2.xlsx.

Availability

- [x] 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 data* section, below.

Publicly available data

- [x] Data are available online at: https://diabimmune.broadinstitute.org/diabimmune/t1d-cohort/resources/16s-sequence-data and https://www.cell.com/cms/10.1016/j.chom.2015.01.001/attachment/1f0883f8-1df7-447d-a47b-c1aa2bb2bbaf/mmc2.xlsx
- [ ] 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

Description

File format(s)

- [x] CSV or other plain text.
- [x] Software-specific binary format (.Rda, Python pickle, etc.): .RData
- [ ] Standardized binary format (e.g., netCDF, HDF5, etc.):
- [ ] Other (please specify):

Data dictionary

- [x] Provided by authors in the following file(s): data_dictionary.md
- [ ] Data file(s) is(are) self-describing (e.g., netCDF files)
- [ ] Available at the following URL:

Additional Information (optional)

Part 2: Code

Abstract

R package for the logistic-tree normal models is available at https://github.com/MaStatLab/LTN.git . Code for reproducing all results in the paper is available at https://github.com/MaStatLab/LTN_analysis.git , where the folder "src" includes code for data processing, simulation studies, and case study. For more detailed description of the R scripts, see https://github.com/MaStatLab/LTN_analysis/blob/main/README.md .

Description

Code format(s)

- [x] Script files
- [x] R
- [ ] Python
- [ ] Matlab
- [ ] Other:
- [x] Package
- [x] R
- [ ] Python
- [ ] MATLAB toolbox
- [ ] Other:
- [ ] Reproducible report
- [ ] R Markdown
- [ ] Jupyter notebook
- [ ] Other:
- [ ] Shell script
- [ ] Other (please specify):

Supporting software requirements

Version of primary software used

R version 3.6.0

Libraries and dependencies used by the code

VGAM_1.1-5, statmod_1.4.36, phyloseq_1.30.0, philr_1.12.0, mvtnorm_1.1-2, ggplotify_0.0.8, ggplot2_3.3.5, data.tree_1.0.0, BayesLogit_2.1, ape_5.5, reshape2_1.4.4, GetoptLong_1.0.5, MASS_7.3.51.4, ROCR_1.0.11

Supporting system/hardware requirements (optional)

Parallelization used

- [x] No parallel code used
- [ ] Multi-core parallelization on a single machine/node
- Number of cores used:
- [ ] Multi-machine/multi-node parallelization
- Number of nodes and cores used:

License

- [x] MIT License (default)
- [ ] BSD
- [ ] GPL v3.0
- [ ] Creative Commons
- [ ] Other: (please specify below)

Additional information (optional)

Scope

The provided workflow reproduces:

- [ ] Any numbers provided in text in the paper
- [x] All tables and figures in the paper
- [ ] Selected tables and figures in the paper, as explained and justified below:

Workflow

Format(s)

- [ ] Single master code file
- [ ] Wrapper (shell) script(s)
- [ ] Self-contained R Markdown file, Jupyter notebook, or other literate programming approach
- [x] Text file (e.g., a readme-style file) that documents workflow
- [ ] Makefile
- [ ] Other (more detail in *Instructions* below)

Instructions

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

- [x] Not feasible to run on a desktop machine, as described here:

The simulation study and case study is run on the computing cluster. For each simulation setting, an example of running a single simulation is provided in the readme file. It typically takes less than 6 hours for a certain simulation round. (We tested the run-time on a machine with 8 GB memory and 1.8 GHz Intel Core i5 processor.)

Additional information (optional)

Notes (optional)