nature portfolio

Corresponding author(s): Chen-Ching Lin

Last updated by author(s): Apr 28, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

16S amplicon and metagenomic shotgun datasets from Nearing's and Kennedy's studies were directly sourced from their repositories, while others were downloaded from NCBI SRA and processed by us. Additionally, the miRNA RPM table was downloaded from the TCGA GDC portal. The real-sim dataset was directly downloaded from [https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html].

Data analysis

The raw 16S amplicon sequencing dataset was processed using the dada2 pipeline (version 1.16.0) and taxonomy was assigned using SILVA (version 138.1). Shotgun reads underwent processing with KneadData (version 0.10.0) to eliminate low-quality or host-contaminated reads based on the hg37 reference genome. Taxonomic profiling and quantification were carried out using MetaPhlAn (version 4.0.2) with the database updated as of October 2022. For miRNA processing, 5' end isoforms relied on miRBase (version 21).

All benchmarked methods are list in Supplementary Note 3. and the version of R or pathway package are list below.

R packages include :
ALDEx2 Version 1.26.0
compositions Version 2.0-6
DESeq2 Version 1.34.0
dplyr Version 1.1.2
edgeR Version 3.36.0
effsize Version 0.8.1
exactRankTests Version 0.8-35
ggplot2 Version 3.4.2
lme4 Version 1.1-34

MASS Version 7.3-55
metagenomeSeq Version 1.36.0
NBZIMM Version 1.0
nlme Version 3.1-155
patchwork Version 1.1.2
phyloseq Version 1.38.0
RColorBrewer Version 1.1-3
scales Version 1.2.1

python packages include :
numpy Version 1.24.1
mrmr Version 0.2.7
pandas Version 2.0.2
PreLect Version 0.0.1
scipy Version 1.11.3
scikit-learn Version 1.3.2
xgboost Version 1.7.5

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The tables and metadata of all analyzed datasets, including row count and VST tables, are uploaded to the Zenodo repository [https://doi.org/10.5281/zenodo.10062237], with additional depictions listed in Supplementary Data.

The scripts for the main figures, data processing, and feature selection are available on our GitHub repository [https://github.com/YinchengChen23/PreLect_manuscript].

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

Sex and gender were not applicable in this study as published data were used.

Reporting on race, ethnicity, or other socially relevant groupings

Race, ethnicity, or other socially relevant groupings was not applicable in this study, since we using published data.

Population characteristics

Not applicable, since we using published data.

Recruitment

Not applicable, since we using published data.

Ethics oversight

Not applicable, since we using published data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Sample sizes were determined based on previously openly shared datasets from Nearing's and Kennedy's studies, as well as datasets available on NCBI SRA and TCGA GDC.

Data exclusions

We did not exclude any samples or features from datasets sourced from Nearing's study. However, we excluded samples with a total raw read count of under 50,000 from six of our processed datasets to remove low-quality samples.

Replication

Since PreLect uses 0 for weight initialization, as long as it is the same data and the same lambda, the results obtained will be exactly the same. In addition, we checked the prediction performance of each benchmarked method multiple times and still maintained a certain pattern.

Randomization

To assess PreLect's capacity. The synthetic data was generated. Randomization was employed to create true negative features by shuffling the non-zero values of features, excluding the top 100 prevalent features. This randomization aims to mitigate biological signals, ensuring a fair comparison between PreLect and benchmarking methods. Our findings illustrates that ALDEx2, ANCOM2, metagenomeSeq, and MI excel in identifying true positive features, likely due to their sensitivity to case-control variations. Nonetheless, PreLect outperforms ML-based methods.

Blinding

Blinding wasn't applicable in this study. We only excluded samples with low sequencing quality during data collection. All machine learning-based benchmarking methods utilized the same data type (VST with z-standardization) as input, and the input data type for statistical-based models followed official recommendations (ANCOM2 and edgeR included pseudocounts, while rast utilized raw count tables).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern
☒	□	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Plants

Seed stocks

Not applicable.

Novel plant genotypes

Not applicable.

Authentication

Not applicable.