nature research

Corresponding author(s): Gary Siuzdak

Last updated by author(s): Mar 31, 2020

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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	Data sets were obtained from publicly available repositories. Software used: XCMS Online, IBM Watson For Drug Discovery, METLIN, Microsoft Academic, Semantic Scholar, SciFinder, HUPO BD-HPP are all available online (not for download). Tools were use regularly between October 2018 and March 2020.
Data analysis	Analysis was performed in the above programs or Microsoft Excel.

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets analysed during the current study that were not generated by the authors but mined from public sources are available in the MetaboLights repository, MTBLS298 and Human Metabolome Database repository, Nonalcoholic fatty liver disease, and in the main text, supplemental information or request of the author of these publications: Metabolomics Identifies Perturbations in Human Disorders of Propionate Metabolism, Metabolism Links Bacterial Biofilms and Colon Carcinogenesis, Systems biology guided by XCMS Online metabolomics. Additional data generated by the authors or analysed during this study are included in this published article (and its supplementary information files).

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	Data sets were retrieved from publicly available repositories or previously published papers.
Data exclusions	No data were excluded.
Replication	In the development of this protocol, we had lab members who were not involved in the development run the protocol one the same data sets. Each time, the results were consistently reproducible. Literature databases update daily with new publications and slight variability can occur day to day if a new paper is published in the search terms.
Randomization	Subsets of the mined datasets were chosen at random by Excel random function or within each software used to serve as the control groups or training sets for validation testing.
Blinding	Since the investigators used publicly available data sets, they were blind to the data collection. Blinding during data analysis was achieved by investigators not interacting with the data before processing, which could have introduced bias from presorting.

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
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

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