nature research

Corresponding author(s): Pieter Dorrestein

Last updated by author(s): Dec 9, 2019

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	'ili 3D-cartography modeling code is available here https://github.com/MolecularCartography/ili Single spectrum search is available here https://gnps.ucsd.edu/ProteoSAFe/index.jsp?params=%7B%22workflow%22%3A%22SEARCH_SINGLE_SPECTRUM%22%7D 'ili 3D-cartography modeling code is available here https://github.com/MolecularCartography/ili The MASST search custom code is available here: https://github.com/CCMS-UCSD/GNPS_Workflows/tree/master/search_single_spectrum Invesalius Version 3 was used to process MRI data and 3D models were generated with the ili software available here: ili.embl.de
Data analysis	Invesalius software (https://invesalius.github.io/) was used for generation of mouse 3D-model from MRI dicoms. 'ili software was used for 3D-mapping https://github.com/MolecularCartography/ili Ivesalius Version 3 was used to process MRI data and 3D models were generated with the ili software available here: ili.embl.de

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

All metabolomics data is available at GNPS (gnps.ucsd.edu) under the MassIVE ID numbers: MSV000079949 (GF and SPF mouse data). Additional MS datasets:

MSV000082480, MSV000082467, MSV000079134, MSV000082406, MSV000083032, MSV000083004,MSV000083446.
The sequencing data for the GF and SPF mouse study is available on the Qiita microbiome data analysis platform at Qiita.ucsd.edu under study ID 10801 and through the European Bioinformatics Institute accession number ERP109688.

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	No intial sample sizes were predetermined for the GF and SPF mice. We were searching for molecules present only in colonized mice our n=4 enabled a higher level of confidence that these molecules were not due to individual variation but in fact presence of a microbiome. Upon FXR agonism tests during gavage experiments initial data using n=4 was used to project required n for subsequent validation experiments. No samples sizes were calculated for the clinical experiments. All of these were searching of publicly available data for the new bile acids MS/MS spectra. We subsequently tested abundance of these compounds in the subjects datasets and if significance was found it was reported, thus, the sample sizes were sufficient.
Data exclusions	There are no data exclusions from the clinical experiments or any other aspect of the manuscript.
Replication	The multiple mice tested in each group enabled a verification that molecules found were consistent and reproducibly from the microbiome. Searching the public datasets also provided validation for the molecules association with guts of mammals. Other aspects of this study were broad scale mining of publicly available mass spectrometry data, thus, we did not attempt to reproduce the findings, however, links to the workflows and data are available such this can be easily replicated
Randomization	The datasets are randomized in the context of our molecular MASST searching. We have no prior knowledge of what any of the samples are and the searches report the results regardless of grouping of samples. But again, we don't believe mining public data has relevance in this context.
Blinding	The investigators were blinded to the study subjects. There was no prior knowledge of what studies the compounds were present in, and no groups of patients/subjects were known prior to the MS/MS searches.

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

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Germ-free (GF) C57Bl/6J mice and Conventionally-colonized specific pathogen free (SPF) mice (C57Bl/6J)
Wild animals	This study did not involve wild animals.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	The California Institute of Technology and University of California at San Diego provided guidance/oversight.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	The human populations used were not relevant as covariates as they were only mined to identify the presence of a molecule in their sample sets. The cystic fibrosis patient population was a pediatric group age 0-18 years.
Recruitment	Patients were recruited during routine visits to UCSD clinics there was no bias towards selection of any particular group.
Ethics oversight	UCSD Institutional Review Board, University of Michigan, and Yale University Institutional Review Boards

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJEguidelines for publication of clinical research and a completedCONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*