nature research

Corresponding author(s): Michael Fischbach, Dylan Dodd, Marnix Medema

Last updated by author(s): 2023/01/02

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 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	No software was used for data collection.
Data analysis	Sequence similarity networks were constructed using BiG-SCAPE v1.0 and annotated in Cytoscape v3.0. Read mapping to metabolic gene clusters was performed using BiG-MAP v1.0, which makes use of Bowtie v2.1.0 for mapping. Redundancy of gene clusters was controlled using clustering with MMseqs v2. Phylogenetic trees were constructed using FastTree v2.1 and visualized using iTOL v5, based on alignments generated using Clustal Omega v1.2.2 and curated in JalView v2. Profile Hidden Markov models were built and used to search sequence data using HMMER v3.1b2. Homologous gene clusters were identified using clusterTools v1 and MultiGeneBlast v1.1.14. Raw metatranscriptome/ metagenome reads were filtered using kneadData (v0.4.6.1). The gutSMASH software written for this study is available from https://github.com/victoriapascal/gutsmash.

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 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 LifeLines DEEP cohort raw metagenomic sequencing data, metabolome data and human phenotypes (i.e. age and sex) used for the analysis presented in this study are available at the European Genome-phenome Archive under accession EGAS00001001704. Taxonomic assignments of bacteria were performed according

to the Genome Taxonomy Database release 95 (https://gtdb.ecogenomic.org/). Lists of accessions of genome assemblies used are available in Tables S3 and S4. iHMP multi-omics data were downloaded from https://ibdmdb.org. Raw sequence data of the iHMP are also available from the NCBI sequence read archive (SRA) via BioProject PRJNA398089, metatranscriptome data through GEO Series accession number GSE111889, and metabolomics data at the Metabolomics Workbench (http://www.metabolomicsworkbench.org; Project ID PR000639).

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	This study included 1,135 individuals from the population-based LifeLines-DEEP cohort with fecal metagenomic sequencing data available. 1054 of them also have plasma metabolomics available. In order to ensure the analysis power, the study includes as much as subjects as possible. Thus no sample size calculation was performed.
Data exclusions	Samples with low quality of metagenomics sequencing were excluded.
Replication	There was no direct replication in independent cohorts. However, indirect replication and comparison were performed. For example, we compared associations between plasma and fecal levels of the same metabolites. We also compared the associations between metagenomics from the LLD cohort and metatranscriptomics data of the iHMP cohort.
Randomization	This is human cohort-based analysis. The sample collection, sequencing and metabolomics profiling were performed in a random order. No extra randomization was done for this study. We included age, sex and read depth of sequencing data as covariates in our correlation analyses.
Blinding	This study is a human cohort based, observational study. Thus no blinding was performed.

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
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	The study has included the population-based LifeLines-Deep DEEP (n=1,135), with 58.20% female. The mean age (SD) of participants was 45.04 (13.60) years and their mean (SD) BMI was 25.26 (4.18);
Recruitment	The LifeLines-DEEP cohort was a random subset of the population-based Lifelines cohort. The recruitment of participants was through the LifeLines organization. Eligible participants were invited to participate in the LifeLines Cohort Study through their GP. A large number of GPs within the northern three provinces of The Netherlands (Friesland, Groningen and Drenthe) were involved and invited all their patients between the ages of 25 and 50 years, unless the participating GP considered the patient not eligible based on the following criteria: severe psychiatric or physical illness; limited life expectancy (<5 years);insufficient knowledge of the Dutch language to complete a Dutch questionnaire. Participants were asked to indicate whether their family members, such as partners, parents, parents-in-law and children would also be willing to participate in the study. If so, permission was asked to send them an invitation to participate. Children could only participate if one of their parents was a participant. In addition, inhabitants of the northern provinces could also register themselves via the LifeLines website.

Ethics oversight	All participants signed an informed consent form prior to sample collection. Institutional ethics review board (IRB) approval was available for the Lifelines DEEP (ref. M12.113965).

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