nature research

Corresponding author(s): Jingyuan Fu

Last updated by author(s): Jul 5, 2021

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 specific software was used for data collection.
Data analysis	All data analyses were conducted using publicly available tools. For this study the following software was used: kneadData (v0.4.6.1), Bowtie2 (v2.1.0), MetaPhlAn2 (v2.7.2), HUMAnN2 (v0.10.0), gutSMASH, SGV-Finder, vegan (v2.5.5), glmnet (v2.0.16), TwoSampleMR (v0.5.5), mediation (v4.5.0). Code used for data analyses is publicly available at: [https://github.com/GRONINGEN-MICROBIOME-CENTRE/Groningen-Microbiome/tree/master/Projects/LLDeep_plasma_GeneralMeta].

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 raw metagenomic sequencing data and plasma metabolome of the Lifelines-DEEP and replication cohorts are available from the European Genome-Phenome Archive (EGA, https://www.ebi.ac.uk/ega/home) via study accession number EGAS00001001704.

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	In total, this study includes 1,368 individuals of the Lifelines-DEEP (LLD) and GoNL cohorts for whom we had collected extensive phenotypic datasets. In detail, dietary habits, genetics, the gut microbiome and plasma metabolome information were available for 1,054 LLD baseline participants, with 331 of them also followed-up 4 years later. In addition, we included two replication cohorts: 237 LLD participants for whom we had dietary habits, genetics and plasma metabolome information and 77 GoNL participants for whom only genetics and plasma metabolome information were 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	Only samples with metagenomics, metabolomics, genetics or dietary habits missing were excluded in analyses.
Replication	We replicated the significant metabolites associations to genetics and microbiome, our metabolite variance estimation and our MR results in corresponding omics datasets of the Genome of the Netherlands (GoNL) cohort, the LLD 4-year follow-up cohort and the LLD2 cohort.
Randomization	This is human cohort-based analysis. The sample collection and sequencing were performed in a random order. No extra randomization was done for this study.
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 cohort Lifelines-DEEP from the Netherlands, n=1,500, 58.20% female, the mean age (SD) of participants is 45.04 (13.60) years and their mean BMI is 25.26 (4.18).
Recruitment	The Lifelines-DEEP cohort is a random subset of the population-based Lifelines cohort. All participants were collected in the Netherlands, thus the reported results could be region-specific. However, independent replication carried in both the GoNL and LLD sub cohorts illustrated the robustness of results.
Ethics oversight	All Lifelines participants signed an informed consent form prior to sample collection. The ethics review board of University Medical Center Groningen (UMCG) has approved the study with reference number M12.113965.

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