nature research

Corresponding author(s): Jeroen Raes

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	16S data pre-processing was performed using LotuS (version 1.565, used for demultiplexing sequencing reads) and the DADA2 pipeline (version 1.6.0), with the RDP classifier (version 2.1254) for taxonomy assignment. For analysis of shotgun sequencing data, paired-end reads were quality trimmed with Trimmomatic (version 0.32), decontaminated from phiX and human sequences using DeconSeq (version 0.4.3), and broken pairs were fixed using a custom Biopython script (available at https://github.com/raeslab/raeslab-utils/). Reads were mapped on the integrated gene catalog using BWA (version 0.7.8), and the mapping was summarized to functional profiles with featureCounts (version 1.5.3). The code used to compute GBM abundances from a KO abundance table is shared on GitHub (https://github.com/raeslab/omixer-rpm). The GBM input file and references used to assemble each GBM can be downloaded from http://raeslab.org/software/gbms.html. Data analysis and graphical representations were performed using R, a free software environment for statistical computing.

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 upon request. 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

FGFP 16S sequencing data and metadata on the microbiota covariates used in this study are available at the European genome-phenome archive (EGA, https://www.ebi.ac.uk/ega/) - accession nr EGAS00001003296. The LLD sequence data and age and sex information per sample are also available at EGA with accession number EGAS00001001704, while the rest of microbiota covariates can be requested from the LifeLines cohort study (https://lifelines.nl/lifelines-research/access-to-lifelines) following the standard protocol for data access. FGFP and TR-MDD shotgun sequencing data and metadata are available at EGA - accession nr EGAS00001003298.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	No sample size calculation was performed. All samples from the already published FGFP and LLD datasets after excluding any subject without GP report of depression but taking medication labeled as "antidepressant" by the Anatomical Therapeutic Chemical classification system (ATC codes N06A) were included in the analysis (N=1054 and N=1063 respectively). For the FGFP shotgun dataset, 150 samples (N=80 cases and 70 controls) were selected for shotgun sequencing based on recent literature indicating that groups with N>20 are sufficient for depression vs controls group comparison analysis (PMID: 24888394, 25882912, 27067014, 27491067).
Data exclusions	No data were excluded from the analyses.
Replication	Extensive validation was performed: Microbial taxa associations with QoL scores were validated in the LLD dataset (N=6/28 associations replicated). Taxa associations with depression were validated in the FGFP non-medicated subset (N=2/4), in the LLD dataset (N=2/4), in the LLD non-medicated subset (N=2/4), and compared to published literature (N=2/4). GBM associations with QoL were replicated in the LLD (N=1/9 associations replicated). GBM associations with depression were tested in the LLD dataset and TR-MDD dataset.
Randomization	Samples were only classified according to the original datasets. To partial out the effects of covariates of microbiota composition (age, sex, BMI, bowel transit time, and gastrointestinal disease) in associations between bacterial genera or gut-brain modules (GBMs) and metadata variables, we fitted generalized linear models (GLMs).
Blinding	Investigators were not blinded during data collection and analyses.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

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	FGFP cohort: 576 female/478 male. Mean age: 50.9; BMI: 24.9; BSS: 4. LLD cohort: 616 female/447 male. Mean age: 57.9; BMI: 25.3; BSS: 3.8.
Recruitment	Samples from the already published FGFP (PMID:27126039) and LLD (PMID:27126040) datasets were included in the analysis (N=1054 and N=1063 respectively). Participants in the newly collected TR-MDD dataset (N=7) were balanced on age, sex, and BSS to the FGFP healthy subset selected for shotgun sequencing (N=70), and to the FGFP-depression shotgun subset (N=80).