nature research

Corresponding author(s): Dr Guillaume Meric
A/Prof Michael Inouye

Last updated by author(s): Nov 4, 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 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 particular software was used for data collection.
Data analysis	Data analysis was performed using the following freely-available software: Eagle2 v2.3 (https://alkesgroup.broadinstitute.org/Eagle/), IMPUTE2 v2.3.0 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html), Plink v2.0 (https://www.cog-genomics.org/plink/2.0/), Centrifuge v1.0.4 (https://ccb.jhu.edu/software/centrifuge/), BOLT-LMM v2.3.2 (https://alkesgroup.broadinstitute.org/BOLT-LMM/BOLT-LMM_manual.html), FlashPCA v2.0 (https://github.com/gabraham/flashpca), GCTA-COJO v1.91.3 (https://cnsgenomics.com/software/gcta/#Overview), matSpDlite v1.0 (https://gump.qimr.edu.au/general/daleN/matSpDlite/), ANNOVAR v2018Apr16 (https://annovar.openbioinformatics.org/en/latest/), Run_dbCAN2 v2.0.11 (https://github.com/linnabrown/run_dbcan), GreenAlgorithms v1.0 (https://www.green-algorithms.org/). Notably and aside from regular data manipulation packages, the packages "compositions" v1.40-2 (https://cran.r-project.org/web/packages/compositions/index.html), "vegan" v2.5-6 (https://cran.r-project.org/web/packages/vegan/index.html), "survival" v3.1-8 (https://cran.r-project.org/web/packages/survival/index.html) and "TwoSampleMR" v0.5.4 (https://mrcieu.github.io/TwoSampleMR/articles/introduction.html) were used. Scripts used to analyze non-identifiable data in this study have been made available on Zenodo (doi: 10.5281/zenodo.5641303).

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

Complete summary statistics of microbial taxa with genome-wide significant hits are available in the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/), GCP ID: GCP000228; study accession numbers GCST90032172-GCST90032644. The metagenomic data from FINRISK 2002 samples are available from the European Genome-Phenome Archive (study ID: EGAS00001005020). The phenotype data contain sensitive information from healthcare registers and are not publicly available because to avoid compromising research participant privacy/consent. They are available through the THL biobank upon submission of a research plan and signing a data transfer agreement (https://thl.fi/en/web/thl-biobank/for-researchers/application-process). Additional databases used in this work include GTDB release 89 (https://gtdb.ecogenomic.org/) and CAZy (last accessed 31/07/2019) (http://www.cazy.org/).

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	The FINRISK 2002 study population included a total of n=8,798 individuals. Among these, a total of 5,959 individuals had matching array genotyping and stool metagenomic sequencing information and were included in our study, the largest single-cohort dataset of matched genotype-stool metagenomes to date.
Data exclusions	For the genotyping dataset, SNPs with call rate <95%, sex discrepancies, excess heterozygosity and non-European ancestry were excluded. After SNP imputation, samples with >10% missing rate were removed. Individuals with extreme height or BMI values were further excluded (31 individuals with height<1.47m, which is the standard definition for dwarfism; 5 with BMI >50, which is generally defined as "extremely obese" were removed), as both conditions are known to affect the gut microbiome. See Methods for further details.
Replication	A method on "Replication of previously reported associations" is included in the Methods section. Briefly, to evaluate the reproducibility of our results with previously reported associations, we collected GWAS summary results from 8 studies published in peer-reviewed journals at the time of this work and examined which associations could be replicated. More than a quarter (149/547) previously reported SNPs were associated with gut microbiome in our study surpassing Bonferroni-corrected significant level (p<9.12×10^-5).
Randomization	No randomization was performed. In the GWAS step, the top 10 genetic principal components were included in as covariates. Age, gender, and genotyping batch were adjusted for.
Blinding	The study describes associations between human genotypes and microbes and does not include any particular treatment requiring group allocation, and therefore no blinding.

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	As described in the text, the 5,959 individuals included in the GWAS varied according to the following: age varied from 24 to 74 years old (mean=49.61), sex distribution was 55.1% women and 44.9% men; prescription record linkage indicated that 250 individuals were prescribed antibiotic 1 month before baseline sampling. Microbiome sequencing was performed in 3 batches: RKL004 (n=2,348), RKL001 (n=1,192) and RKL003 (n=2,418). Genotyping was performed in 4 batches: Illumina HumanCoreExome BeadChip (n=3,757), Illumina Human610-Quad BeadChip (n=506), Illumina HumanOmniExpress (n=480) and Illumina GSAMD (n=1,216).
Recruitment	FINRISK population surveys have been performed every 5 years since 1972 to monitor trends of cardiovascular and other non-communicable disease risk factors in the Finnish population. The study population of this study consists in the participants of FINRISK 2002 (FR02) study, including men and women aged between 25 and 74 years from six geographical areas of Finland. The sampling was stratified by sex, region and 10-year age group so that each stratum had 250 participants. The overall participation rate was 65.5% (n = 8,798). Of those, a subset consisting of a total of 5,959 individuals had matching genotype and stool metagenomics of suitable quality and subsequently used in this study (see Methods). Self-selection bias was expected to be very limited, as the FINRISK Study had been established and running for more than 40 years by 2002, time of baseline sampling.
Ethics oversight	The study protocol of FR02 was approved by the Coordinating Ethical Committee of the Helsinki and Uusimaa Hospital District (Ref. 558/E3/2001).

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