nature portfolio

Corresponding author(s): Jaeyun Sung

Last updated by author(s): Aug 4, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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	Stool metagenome samples used in this study were downloaded from NCBI-SRA (https://www.ncbi.nlm.nih.gov/sra) using SRAToolkit (version 3.0.1).
Data analysis	Code can be accessed at: https://github.com/danielchang2002/GMWI2 Quality control of sequencing data: Potential human contamination was filtered out by removing reads that aligned to the human genome (reference genome GRCh38/hg38) using Bowtie2 v2.4.4 with default parameters. Along with Illumina universal adapter sequences, probable adapter sequences were identified by extracting overrepresented sequences from each metagenome sample using FastQC v0.11.8. Adapter sequence clipping and quality filtration were performed using Trimmomatic v0.39. Specifically, Trimmomatic's "ILLUMINACLIP" step was used, using a maximum seed mismatch count of 2, palindrome clip threshold of 30, simple clip threshold of 10, and minimum adapter length of 2 bp. Additionally, leading and trailing low-quality bases (Phred quality score < 3) of each read were removed, and trimmed reads shorter than 60bp in nucleotide length were discarded. Taxonomic profiling from shotgun metagenomic sequencing data: Taxonomic profiling was carried out using the MetaPhlAn3 v3.0.13 phylogenetic clade identification pipeline using default parameters, which aligns the metagenome reads to clade-specific marker genes in the mpa_v30_CHOCOPhlAn_201901 database. Statistical analyses: R v4.0.3 was used for all statistical analyses. The R packages 'ade4' version 1.7-22 and 'vegan' version 2.6.4 were used to perform principal

coordinate analysis (PCoA) using Bray-Curtis dissimilarity. The “prcomp” function in R was used to perform principal component analysis. The R package ‘vegan’ version 2.6.4 was used to calculate Shannon’s index and species richness. Permutational multivariate analysis of variance (PERMANOVA) on the distance matrix (Bray-Curtis) was done using the “adonis2” function in R. To predict disease presence using gut microbiome profile, a Lasso-penalized logistic regression model (Python library “scikit-learn” v1.0.2) was trained on the binary presence/absence taxonomic profiles of the entire pooled dataset of 8,069 stool metagenomes.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All publicly available stool metagenome samples (and their corresponding studies) used in the analyses of this study are available in Supplementary Data 2. All raw metagenomic reads are available through SRA using the sequencing data accession IDs. Additionally, preprocessed data for this study is downloadable at: https://github.com/danielchang2002/GMWI2/blob/main/manuscript/data.zip

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Our study uses publicly available stool shotgun metagenome data, and was derived on a balanced representation of male (n = 2,436) and female (2,583). Sex was not reported for the remaining 3,050 samples in our study.
Reporting on race, ethnicity, or other socially relevant groupings	Our study analyzes stool shotgun metagenomes from diverse geographies, ethnicities/races, cultures, and balanced sex representation. Please see Fig. 1b and the main text for further information.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

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

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	Analyses to determine sample size (e.g., power analyses) were not performed for our study. All publicly available stool shotgun metagenomes were used, surpassing 8,000 samples. No further sample size estimation were needed.
Data exclusions	Not applicable to our meta-analysis study.
Replication	The source code for the GMWI2 tool, processed datasets, and code notebooks that can be used to reproduce figures and our analyses are freely available online at https://github.com/danielchang2002/GMWI2. We were able to successfully replicate our findings presented in the paper.
Randomization	Not applicable to our meta-analysis study.
Blinding	Not applicable to our meta-analysis study.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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