nature portfolio

Corresponding author(s): Klepac-Ceraj, V.

Last updated by author(s): Jul 30, 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	We processed the sequence data from all participating studies applying the bioBakery 3 meta'omics worflow and generated taxonomic and functional profiles. The workflow followed an already established protocol 55 (Preprocessing: KneadData v0.10.0; Taxonomic profiles: MetaPhlAn v3.1.0; microbial markers: mpa_v31_CHOCOPhlAn_201901 database; Functional profiles: HUMAnN v3.7). The workflow includes Bowtie2, DIAMOND, and MinPath softwares.
Data analysis	All the data analysis and plotting were performed using julia v.1.10.3 (Main packages: BiobakeryUtils.jl v0.7.0, Microbiome.jl v0.10.1, Distances.jl v0.10.12, MultivariateStats.jl v0.10.3, PERMANOVA.jl v0.1.1, MLJ.jl v0.19.2, DecisionTree.jl v0.12.4, GLM.jl v1.8.3, Makie.jl v0.21.9.

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

The processed datasets generated and/or analyzed during the current study have been deposited in Data Dryad under DOI: https://doi.org/10.5061/

dryad.dbrv15f9z. The raw sequencing data for the Khula study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1128723. All other relevant data supporting the key findings of this study and instruction on how to obtain it are available within the article and its Supplementary Information files, or are available from the corresponding author (Dr. Vanja Klepac-Ceraj - vklepacc@wellesley.edu) upon reasonable request.

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	We have available as sex assigned at birth. However, for this study, we do not use or report any sex information of the participants.
Reporting on race, ethnicity, or other socially relevant groupings	We have reported the summary demographics of Khula study participants (mothers). These characteristics include maternal place of birth, primary spoken language, maternal education attainment, mother monthly income (reported in South African Rand, ZAR), mother depression score, infant biological sex. However, none of these variables were used in the model as we aimed to build a model that can be used across many different cohorts and studies. Our model excluded all additional participant and biospecimen metadata, using only participant age and microbial data. This decision was made due to the lack of uniformity in metadata collection and annotation across studies.
Population characteristics	This study included data from 12 different countries from three different continents The countries included Italy, Sweden, Finland, Estonia, El Salvador, Great Britain, Russia, USA, South Africa, Brazil, Ireland and Bangladesh. Publicly available metagenome metadata was obtained from the CuratedMetagenomicsData database, Wellcome LEAP cohorts: Germina (Brazil), Khula (South Africa), M4eFAD (Bangladesh), Combine (Ireland), and from ECHO-Resonance (USA). The cohorts' sample sizes and mean ages are summarized in Table 1, and include: 3 samples from Italy with a mean age of 2.95 (0.0) mo published in Asnicar et al. (2017), 180 samples from Sweden with a mean age of 8.03(4.04) mo published in Backhed et al (2015), 59 samples from Estonia and Finland with a mean age of 12.28 (3.88) mo published in Kostic et al (2015), 3 samples from El Salvador, 285 samples from United Kingdom with a mean age of 8.65 (1.95) mo published in Shao et al (2019), 479 samples from Finland, Estonia, and Russia with a mean age of 10.90 (4.28) mo published by Vatanen et al. (2016), 104 samples from Finland with a mean age of 7.4 (4.25) mo published by Yassour et al (2018), 224 samples from the United States of America with a mean age of 7.50 (4.37) mo published by Bonham et al. (2023), 427 samples from South Africa with a mean age of 9.36 (4.46) mo from this study, 963 samples from Brazil with a mean age of 5.41 (2.13) mo by Fator et al (2024), 353 samples from Ireland with a mean of 6.75 (3.11) mo by Hemmingway et al (2020), 74 samples from Bangladesh with a mean of 11.88 (0.51) mo by O'Sullivan et al. (2024).
Recruitment	The recruitment of the published datasets were described in the corresponding papers. LEAP 1kD Gemina and Khula recruited pregnant mothers at their prenatal visits during the first or second trimester of pregnancy.
Ethics oversight	This study included de-identified data from participants whose parents have consented to the use of their anonymized data and information for research purposes. Approvals for all study protocols for these studies was granted by the respective Institutional Review Boards. For the Khula cohort, this study was approved by the Health Research Ethics Committees (study number: 666/2021). Informed consent was collected from mothers on behalf of themself and their infants. Demographic information, including maternal place of birth, primary spoken language, maternal age at enrollment, maternal educational attainment, and maternal income, was collected at enrollment (and is reported in Table 2). As all data and samples are de-identified, Wellesley College did not need to set up an IRB Authorization Agreement with the University of Cape Town as indicated by the Wellesley's IRB. There is a Material Transfer Agreement with University of Cape Town, South Africa.

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	We did not perform a sample size calculation because our study included preexisting populations. We included 3,154 metagenomes from 1,827 infants across 12 countries. We included infants with samples collected prior to 18 months of age. Our study did not have a predefined effect estimates because we were not studying any effects of treatments or medication as one would in a clinical trial. By combining existing datasets with the new datasets, we were able to obtain larger sample sizes compared to studies that published on the similar research topics.
Data exclusions	Samples identified as belonging to premature-born children were excluded. We also excluded samples belonging to children suffering from acute infectious conditions - including sepsis - at the time of sample collection. Samples from children older than 18 months were also excluded.
Replication	For each type, we performed several replicates of model training, with different seeds. To evaluate the generalizability of our model across different data sources and test the predictive ability of each data source toward age, we performed a leave-one-datasource-out cross-

validation (LOOCV) experiment.

Randomization	This is an observational study without any randomized intervention. Therefore, randomization is not applicable.
Blinding	This is an observational study. Therefore, blinding is not applicable to this study as our study did not contain any interventions.

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

Plants

Seed stocks	There are no plants involved in this study.
Novel plant genotypes	There are no novel plant genotypes reported in this study.
Authentication	This was an observational study and no seed or plant genotypes were used here.