nature research

Corresponding author(s): Aimee Baumann-Dudenhoeffer, Gautam Dantas

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	Sequence data processing and extraction of taxonomic data and functional pathways was performed using Trimmomatic (trimmomatic/0.33), Deconseq/0.4.3-chr38, MetaPhlAn2 (metaphlan2/2.2.0), and Humann2 (humann2/0.9.4)
Data analysis	Statistical analysis was performed in R using the vegan, ape, ggplot2, lme4, lmerTest, MuMin, and multcomp packages. R code used in the study is described in the Supplement, and sample code is available at https://bitbucket.org/alaricwdsouza/twindiet/src/master/TwinDiet_ModelFittingExample.R

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

Sequence data supporting these findings have been deposited, along with relevant clinical metadata, in the SRA under Bioproject ID PRJNA473126 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA473126), with the primary accession codes SAMN09259835-SAMN09260236 (https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP148966). Source data for Figures 1-4 are provided online with the paper. Any additional data generated and analyzed in this study are available from the corresponding author upon reasonable request.

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	Because the expected effect sizes for clinical variables of interest (e.g. formula ingredients) correlating with microbial taxa and functions is unknown, a true power calculation could not be performed. However, our population cohort (N=60), sampling interval (monthly from 0-8 months), and samples successfully whole-genome sequenced (N=402) , is comparable to other published studies of gut microbiome development that were able to detect statistically significant associations of clinical variables with microbiome taxa and functions. For example, Backhed et al, Cell Host and Microbe, 2015 had a larger population (N=98 mothers and infants), but lower resolution longitudinal sampling, with samples collected at birth, 4 months, and 12 months. Similarly, Chu et al, Nature Medicine, 2017 had a larger population (two matched cohorts of N=81), but a smaller number of fecal samples selected for whole-genome sequencing (N=69 meconium samples, and infant and maternal stools). Due to these authors' successful analysis of comparable datasets, it was reasonable to conclude that the sample size of our study was appropriate for investigating the effects of pre- and post-natal clinical factors on early gut microbiome maturation.
Data exclusions	No data that met our predetermined sequencing threshold (>= 5 million total reads prior to processing) were excluded.
Replication	In this longitudinal cohort study, statistical modeling of multiple distinct taxonomic groups and metabolic pathways showed similar patterns and trends, consistently identifying major determinants of functional microbiome maturation (breastfeeding, soy formula exposure, prebiotics, antibiotics, domestic water source, maternal gestational weight gain). Our taxonomic modeling independently confirmed several known determinants of gut microbiome establishment, replicating others' prior work in the field, and confirming the utility of our approach.
Randomization	This was a longitudinal cohort study with no intervention applied; randomization is not applicable
Blinding	This was a longitudinal cohort study with no intervention applied; blinding is not applicable

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	402 samples were included from 60 twin infants in thirty-one families, for a median of 7 samples per infant (IQR 6-8). Infant age at stool collection ranged from the day of delivery to 253 days. The median gestational age at delivery was 37 weeks (IQR 36-38), 43% of infants were delivered vaginally, and 57% were born via Cesarean section. 48% of infants were male and 52% were female. 17% of the infants were Black, 83% were white. 3% of infants were Hispanic, and 97% were non-Hispanic. and 47% of twins were monozygotic, 50% dizygotic, and 3% of unknown zygosity. Four infants' mothers were diabetic (7%), six infants' mothers developed preeclampsia (10%), and two infants were born to a mother with both conditions. Additional detailed population characteristics are included in Supplemental Table S1 and S6
Recruitment	This study was approved by the Human Research Protection Office of Washington University School of Medicine in St. Louis, and complied with all ethical regulations. Written consent was obtained from each adult and a parent or guardian of each minor subject. We used fecal samples that had been frozen at -80 C since collection at monthly intervals from a birth cohort of healthy twins in St. Louis, in which the mothers had consented to monthly fecal sample collection from birth until two years of age. Selecting a population of twins may introduce bias as the mothers are at higher risk for complications of pregnancy (e.g. preeclampsia, diabetes), more likely to deliver via Cesarean section, and more likely to deliver prematurely. To avoid potential bias from early illness and antibiotic administration, we excluded any infants who received antibiotics in the immediate postnatal period.