nature research

Corresponding author(s): Ramnik J. Xavier

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	Bowtie2 v2.2.3, BWA v0.7.12, kneadData v0.4.6.1, Trim Galore v0.4.4 (Babraham Bioinformatics), MetaPhlAn2 v2.6.0, HUMAnN2 v0.10.0, BLAST+ v2.6.0, DIAMOND v0.8.22, Crass v0.3.865, R v3.1.1, Prodigal v2.6.3, CD-HIT v4.6.5, Python v2.7.1. Additional details are given in Methods.

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

DIABIMMUNE microbiome 16S rRNA and metagenomic sequencing data supporting the findings of this study are available in NCBI Sequence Read Archive under BioProject PRJNA497734 and through the DIABIMMUNE microbiome website at https://pubs.broadinstitute.org/diabimmune/.

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	All 16S and metagenomic sequencing samples from DIABIMMUNE study were being analyzed. For metagenomic sequencing data (n=1154 samples) and correlative associations, this provides 90% power given alpha=0.001 and Pearson's r=0.014.
Data exclusions	No data exclusion
Replication	No replication was done.
Randomization	Randomization was not used.
Blinding	No blinding was used, DIABIMMUNE is an observational follow-up study.

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	Participants are newborns from Espoo, Finland; Tartu, Estonia; and Petrozavodsk, Russia with HLA DR-DQ alleles conferring increased risk for autoimmunity (n=156 males, n=133 females). They were being follow until age three.
Recruitment	The DIABIMMUNE cohort recruitment took place between September 2008 and July 2011 in Espoo, Finland; Tartu, Estonia; and Petrozavodsk, Russia. Families with a newborn with HLA DR-DQ alleles conferring increased risk for autoimmunity, determined by a cord blood test, were invited to join the study. The parents gave their written informed consent prior to sample collection. We observed slight self-selection bias in recruitment: there were more mothers, fathers and siblings with an atopic disease and more fathers with type 1 diabetes or some other autoimmune disease in the recruited group compared to the group that chose not to join the study. These differences may increase the number of participants with a clinical outcome (e.g. beta-cell autoimmunity, celiac disease autoimmunity, atopic sensitisation).