nature research

Corresponding author(s): Kendra Vehik

Last updated by author(s): Oct 17, 2019

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.

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 software was used for data collection.
Data analysis	VirMAP v1.0 (paper version) was used to generate the virome data and has been deposited in GitHub, https://github.com/cmmr/virmap. diamond version 0.9.22 BBMap version 37.58 blastn: 2.7.1+ lbzip2 version 2.5 MEGAHIT v1.1.3 khmer 2.0+706.g1745464 (normalize-by-median.py) pigz 2.3.3 vsearch v2.9.1_linux_x86_64 zstd v1.3.1 perl v5.24.0 All of the code and dependencies are listed on the GitHub site. SAS 9.4 (SAS Institute, Cary, NC, USA) was used for the statistical analysis and GraphPad Prism 8.0 (San Diego, CA, USA) was used to for the figures.

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. 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

TEDDY virome sequencing data that support the findings of this study have been deposited in the NCBI database of Genotypes and Phenotypes (dbGaP) with the primary accession code phs001442.v1.p1, in accordance with the dbGaP controlled-access authorization process. Clinical metadata and virome results data analyzed for the current study will be made available in the NIDDK Central Repository at https://www.niddkrepository.org/studies/teddy.

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	Longitudinal stool samples from 3 months of age from 383 matched islet autoimmunity (IA) case-control pairs and 112 matched type 1 diabetes (T1D) case-control pairs were analyzed by metagenomic sequencing (IA, n=8,654; T1D, n=3,380) and processed using VirMap. The design includes two matched nested case-control studies. All persistent confirmed IA positive children as of May 31, 2012 were included as case in the IA nested case control study. In a 1:1 matched-nested case-control study, a control was matched to each case based on the case's event time. The control was required to be event-free within +-45 days of the matching case's event-time. The matching criteria were clinical site, sex and family history of T1D. A similar 1:1 matched-nested case-control was designed for T1D where all children who developed T1D as of May 31, 2012 were included in the nested case-control study with a matching control that was chosen on the case's event-time (within +-45 days of case event-time) matched on clinical site, sex and family history of T1D. Power calculations showed both matched-nested case-control studies had 80% power at a significance level of 5% to detect an OR>=2.28 for the IA and an OR>=4.58 for the T1D matched-nested case-control.
Data exclusions	For the nested case-control analysis, some samples were removed so that exactly the same number of samples was included between case and control pairs. This prevented skewing data due to the generally increased number of samples from cases and missing matched pair control data. This lack of matching stool sampling availability resulted in a 9% reduction in the number of pairs for the 1:1 viral metagenomic study for the islet autoimmunity nested-matched case-control, which left 383 pairs (n = 4,327 age matched stool samples in each group) and a 2% reduction in the number of pairs for the 1:1 T1D nested-matched case-control resulting in 112 pairs (n=1,690 age matched stool samples in each group). All outcome analyses were conducted on matched-pair sample data.
Replication	Observational cohort. No replication.
Randomization	Controls were matched individually to cases as described in the manuscript Methods text. Cases were sampled based on specific case-control design (i.e., until either development of islet autoimmunity or diagnosis of T1D) and matched control samples were included up until the corresponding age of event.
Blinding	TEDDY is an observational follow-up study, thus no overall blinding was used. However, the selection of samples sent to the Alkek Center for Metagenomics and Microbiome Research at the Baylor College of Medicine, Houston, TX, were determined by the Data Coordinating Center (USF Health Informatics Institute, Tampa, FL) without the laboratory knowing the case-control status.

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Hela - ATCC Vero – Mary Estes, Baylor College of Medicine RD-CAR – Nora Chapman, University of Nebraska HEK-293 – ATCC Hela - ATCC Vero – Mary Estes, Baylor College of Medicine RD-CAR – Nora Chapman, University of Nebraska HEK-293 – ATCC
Authentication	VERO cells and RD-CAR were not authenticated. HEK-293 and HeLa are ATCC authenticated.
Mycoplasma contamination	All cell lines are certified mycoplasma-free by PCR assay.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cells were used. Finally, no experiments were performed with cells, they were only used to amplify virus.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Children were a mean(stddev) of 2.1(1.2) years of age in the IA case-control and 2.7(1.4) years of age in the T1D case-control. Children samples were obtained from six geographical locations (Finland, Germany, Sweden in Europe and Washington, Colorado and Georgia in the United States). These children are at high HLA genetic risk for developing IA or T1D, with half of the cases for IA or T1D and the other half controls (Supplementary Table 1).
Recruitment	Children were recruited based on specific type 1 diabetes risk human leukocyte antigen (HLA) genotypes and family history of T1D risk. Recruitment began in September of 2004 and was completed in February 2010. Six clinical centers in the USA (Colorado, Washington, and Florida/Georgia) and Europe (Germany, Sweden, and Finland) randomly HLA-screened 424,788 children at birth in hospitals in the four countries. A total of 418,367 general population infants were screen, of which 20,152 (4.8%) were eligible, and 1,437 of the 6,421 screened infants (22.4%) with a first-degree relative with type 1 diabetes were eligible. There were 8,676 children enrolled as participants in the study.
Ethics oversight	The samples and clinical information were obtained under conditions of informed consent and with the approval of the National Institute of Diabetes and Digestive and Kidney Diseases External Executive Committee and participating clinical center Institutional Review Boards.

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	NCT00279318
Study protocol	Full protocol can be accessed at https://teddy.epi.usf.edu/documents/TEDDY_Protocol.pdf.
Data collection	Six clinical research centers - three in the U.S. (Colorado, Georgia/Florida, Washington), and three in Europe (Finland, Germany, and Sweden) participated in a population-based HLA screening of newborns between 2004 and 2010. Children with high risk HLA genotypes were enrolled (n=8,676) and prospectively followed from three months of age to 15 years with study visits that include a blood draw every three months until four years and every three or six months thereafter depending on islet autoimmunity positivity. Stool samples were collected monthly from ages 3 to 48 months and then quarterly until the age of 10 years. A nested-matched case-control was conducted through risk-set sampling using metadata and islet autoimmunity sample results as of 31 May 2012. In a separate nested-matched case-control, each child diagnosed with T1D had a control selected based on their event time from birth. Metadata were collected using validated questionnaires that have been either published or

extensively scrutinized by experts. TEDDY provides many tools, such as 'The TEDDY book', to the parents to assist in real-time collection of all events in their child's life to ensure bias and error are minimized.

Outcomes

Persistent confirmed autoimmunity was defined by the presence of a confirmed islet autoimmunity (glutamic acid decarboxylase (GADA), insulinoma-associated 2 (IA-2A) or insulin (IAA)) at each of the two TEDDY laboratories on two or more consecutive visits. T1D diagnosis was defined according to American Diabetes Association criteria. Discordance between cases and matched controls, both in the number of stool samples positive for each common virus (positive >2% of samples) and in the presence of consecutive samples positive for the virus (allowing for one missed monthly visit), were evaluated using conditional logistic regression models adjusting for HLA genotypes. The magnitude of the associations were assessed by odds ratios with 95% confidence intervals.