nature portfolio

Peer Review File

Gut Microbiome Maturation in Early Childhood Interacts with Host Genetics to Predict Type 1 Diabetes Risk

Corresponding Author: Dr Dong Wang

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Metabolism.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Dong et al analyzed a large number of microbiome samples sequenced metagenomically with 100x2 HiSeq data in relation to host genetic detail and clinical variables. Unfortunately, it not possible to meaningfully evaluate the described microbiome results due to lack of clarity with the methods including databases and parameters used. There further seems to be possible sample independence violations with the staistics. While the authors indicate their methods are on github, the github repository does not address how the metagenomic data were processed or treated. The results and statements derived from them only make sense in light of the methods, so I did not focus much on the results or main text.

Of critical note, the authors do not appear to appropriately filter the metagenomic data for human reads. Most importantly, the absence of appropriate filtering allows for subject identification in human associated samples (see https://www.nature.com/articles/s41564-023-01381-3). Additionally, inclusion of genomes with a complete Y-chromosome during filtering is essential to mitigate technical bias (see https://www.nature.com/articles/s41467-025-56077-5).

Specific comments are below:

Fig1, notes 12151 samples from 887 subjects collected "approximatey monthly" during the first five years of life. 5 * 60 * 887 = 53220. Less than 1/4 of the samples are represented? The main text notes 13 +/- 9 which is not monthly?

Fig1b, if for five years, why does the plot extend to what looks like at least six years?

Fig1c, humans have 23 chromsomes but the plot shows 22?

Fig1e, this plot is misleading. The genetic data on the x-axis are (presumably) singular points per subject whereas the microbiome data are between one and many points per subject on the y-axis. Ordinations are sensitive to the data included, and the dimensionality reduction can be biased if samples are not independent as is the case for subjects with more than one sample. Given the result in fig1b and the implication that < 1/4 of microbiome samples are represented, I would additionally assume that not all subjects are evenly sampled by time point or the number of samples. This plot would be more meaningful if the authors constrained it to a common time point and use a single sample per subject. Additionally, what is typical for multiomic microbiome comparisons between distances is Mantel correlation, and when comparing ordinations, Procrustes. What are the corresponding Mantel and Procrustes statistics at say time point zero, 1 year, 2 years, etc?

Fig1f, PERMANOVA assumes sample independence. Was that controlled for here or are multiple samples from the same subject included?

line 118, is there a statistic to support the statement of "strong correlation"?

line 130, please see the comments on that figure subpanel specifically for suggestions on computing a correlation. This and on line 118, the authors are making statements using statistical language ("correlate" here) but without indicating what the statistics used are, their effect sizes or p-values.

line 142, ah, okay, Mantel was applied. Was this Pearson or Spearman? Were subjects not included at a given time point given a zero distance value, or were the genetic data subset for each test? I can't find clarity in the methods. Procrustes is still recommended in addition to Mantel. Methods aside, this statement is surprising as country is a well known association with the microbiome particularly with methods that account for evolution. The authors show that both the genetic data exhibit a clear separation in fig1d, and a measurable microbiome effect (violating sample independence?) in fig1f. Given that both data layers used with Mantel demonstrate a country effect, it seems surprising the effect is not supported by Mantel, which I believe would only occur of the country effects were inconsistent which would itself be surprising.

line 182 / line 191, using the provided definition of microbiome maturation, what specifically do the authors mean when they stay "...did not achieve full maturation..."? It looks like the read line has samples which are more similar to their baseline than the other trajectories. Or is it the observed variability in the curves? It looks like this is in part based on supplementary figure 1a, where its stated "...a distinct inflection point is observed at k=3" though it seems like an overstatement given the figure. Why do the authors assume disrete rather than continous spaces here?

line 192, how does fig2b demonstrate reproducibility? The plots look quite different and the relationships between the curves are different.

fig2c, aggregating relative abundance data violates the compositional nature of microbiome data. These are at best misleading figures. I urge the authors to adopt compositionally appropriate measures typically used with microbiome data.

extended fig4, were the pathogens listed, like C. bolteae, actually observed in these data? Is it possible that C. bolteae is arising from short reads multimapping with the other clostridia? Indication of genome coverage could be helpful here. I see 52 species here, but the results section mentions 92, what about the other 40?

figure 5cd, are the authors surprised to not find a Bifidiobacterium association with a mutation in LCT for lactose intolerance? That is one of the dominant signals in other microbiome / human genome assessments.

line 501, HUMAnN2 is described in the methods section as capable of detecting virus. Was this virus detected? If not, is that surprising?

line 596, why were 10 individuals omitted, and how were the 594 individuals selected?

line 661, please define insufficient DNA quantity and what high quality means. I would encourage the authors to cite software they use. I additionally recommend noting any deviations from default parameters. From what's shown, the version of bowtie2 is quite old. Were the data processed single or paired end? What specific human genome or genomes was used for host filtering?

line 676, what proportion of the reads were unmapped, and was there a country bias?

line 671, HUMAnN2 0.9.4 is nearly 10 years old!? At least one of the authors on this paper is involved in that software. I'm genuinely confused here, should the research communtiy _not_ use the newer versions? Certainly there have been some bug fixes in the many versions released since 0.9.4 (https://github.com/biobakery/humann/tags)?

line 702, GRCh37 was released in 2009 and is incomplete. Why that genome, and not e.g. T2T-CHM13v2.0 which actually has the Y-chromosome?

line 720, the data were not rarefied? How were the 92 species selected? No coverage filtering was performed? Reference-based approaches are not without value, but what was the reason the authors did not pursue metagenomic assembly? It is extremely unlikely the genomes in their reference database are _the exact_ genomes in the subject.

(Remarks on code availability)
The code does not reflect the actual methods. It is quite incomplete and therefore not useful for reproducibility. It is not a resource for the community as it's not generalized. As a precise example, the methods text states "humann" was used, but a search shows that term only appears in reference to a file (https://github.com/biobakery/teddy_genetics/blob/5e5d0fa33c77c9d80a414515307ab2d3f239f6a7/Figure_scripts/Fig3.R#L156), not the actual use of the software, and unexpectedly the file referenced isn't even in the repository.

Reviewer #3

(Remarks to the Author)
The authors have undertaken a significant revision of the MS in response to three in-depth reviews including from this Reviewer. My assessment is that they have addressed each of the concerns regarding data clarity, analytical approach and interpretation. I have no further suggestions for revision and defer to the Editors regarding the added novelty of this work beyond prior metagenomic analyses in the TEDDY study

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)

I appreciate the strong effort made by the authors to address all review comments.

My main concern lies in the choice of 3 sub groups, as this is a central finding of the paper. The authors provide some explanation for their metrics, but I am left feeling that these choices are arbitrary. My understanding of the field is that while enterotypes are frequently invoked, equally invoked is the notion that gut microbiome profiles belong to a continuum, and that discrete categorization is oversimplifying the microbiome (e.g. see Jeffrey et al. 2012 Nature Reviews Microbiology).

The reviewers kindly provided justification for their three statistics, but, when I look at the plots, I feel I could justify other ks using the same logic. E.g.

“The Elbow Method (Panel A) demonstrates a distinct inflection point at k=3, where the reduction in within-cluster sum of squares begins to plateau, indicating diminishing returns for additional clusters.”

One could argue the inflection point is at 2.

"The Calinski-Harabasz Index (Panel C) is maximized at k=2 but remains high at k=3, before declining sharply at higher k values. This suggests that while a 2-cluster solution provides strong separation, a 3-cluster solution maintains high structural integrity while capturing necessary biological nuance (specifically, the distinction between "Late Matured" and "Early Plateaued" trajectories). "

It's actually higher at 4 and 5.

"The Gap Statistic (Panel B) trends upward toward higher k values. However, choosing a higher k (e.g., k=6) results in highly granular clusters with reduced sample sizes."

OK – what about 4, 5?

I’m not advocating for any one k – in fact, I think moving away from any k is a more robust option.

At minimum, I think the authors could temper their claims about there being specifically three groups – maybe there are more? Or less? Or a continuum?

(Remarks on code availability)
I cannot figure out how supplemental figures are generated.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors are requesting a re-review of results in which there are compelling reasons to not have confidence in the underlying data processing from which the results are derived from. While I appreciate the detailed responses from the authors, and the selection of additional analyses the performed, I am disappointed to see the authors not adopt modern evidenced host filtration procedures. They acknowledge the deficiencies of their current analysis, yet the authors choose to write around the concerns rather than redoing the work. Writing around the problem does not directly address the evidence that the methods used have serious ethical issues and create technical bias. It's unfortunate the TEDDY consortium is using antiquated methods rather than reprocessing the entirety of their data, but that appears to be a policy decision without clear scientific evidence. The authors further seem to suggest that updating software doesn't matter which is a remarkable position to take. Notably, multiple pieces of the software they use are 10 years old with many important bug fixes having been addressed in the intervening years. Confusingly, the authors even acknowledge this in their response: "HUMAnN has undergone substantial development since v0.9.4, including bug fixes..."

(Remarks on code availability)
I see they updated the GitHub repository. The files inside refer to data files that I did not obviously find in the repository. I did not look closely at the actual code given the ethical and technical concerns I have with the data processing.

Response to Editor's Comments

You will see that they all agree your manuscript has improved during revisions, but they have raised a few more points that need to be addressed before we can make a decision on publication.

> Thank you so much for your thoughtful evaluation and for recognizing the improvements we made during the revision process. We also sincerely appreciate the additional comments and suggestions from you and the reviewers, which have helped us further strengthen the manuscript. We have carefully addressed all remaining points in the revised version and provide detailed responses below.

In this case, the referees' reports seem to be quite clear and we will need you to address all of the remaining points raised. In particular, please confirm that human reads were filtered out during the analysis, as highlighted by Reviewer #1.

> In direct response to your specific request: we confirm that human reads were filtered out prior to all downstream analyses. Specifically, quality-controlled reads were aligned with Bowtie 2 (v2.2.3) against a reference database comprising bacterial, viral, human (GRCh38/hg38), and vector sequences, and any read whose highest-scoring alignment was to a non-bacterial sequence — including the human genome — was removed before taxonomic or functional profiling. This procedure follows the centralized TEDDY metagenomic processing pipeline used in the foundational TEDDY publications[1,2], and the 1,238 metagenomes newly added for this study were processed through the identical workflow. We recognize that this filtering step was insufficiently described in the original Methods, and we have substantially expanded the Methods section to document the host-decontamination procedure, the reference assemblies used (GRCh38/hg38 for metagenomic host-read removal; GRCh37/hg19 for ImmunoChip-derived host genotype data), the software versions and parameters used at each step, and our public code repository now includes the full processing pipeline. Full details, together with our responses to Reviewer 1's related concerns regarding participant privacy and sex-chromosome reference completeness, are provided in our response to Reviewer 1, Comment 1 (and the related Comments 16, 17, and 19).
>
> Beyond the human-read question, the principal additional revisions in this resubmission include: (i) clarification and quantification of the longitudinal sampling structure, with a new Supplementary Figure showing per-participant sampling density; (ii) reanalysis of host-genetics–microbiome associations using independence-respecting methods (single-sample-per-subject Mantel and Procrustes tests; subject-level restricted-permutation PERMANOVA); (iii) a sensitivity analysis at k = 2 and k = 4, addressing both reviewers' concerns about the three-trajectory discretization; (iv) a comprehensive update of the Methods with software versions, parameters, reference assemblies, and quality-control thresholds. The manuscript has also been revised to temper language regarding discrete maturation classes, consistent with both reviewers' guidance and the broader gradient view of microbiome variation [3]

Please highlight all changes in the manuscript text file, and upload your revised manuscript as a word file (.doc).

We have highlighted all changes in the manuscript and uploaded the revised version as a Word file.

Sex and Gender Reporting:

Nature Journals have recently announced an update to our guidance on reporting on sex and gender in research studies (see here). We strongly encourage researchers to follow the 'Sex and Gender Equity in Research – SAGER – guidelines' and to include sex and gender considerations for studies involving humans, vertebrate animals and cell lines where relevant to the topic of study (an overview can be found here). Authors should use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms.

When preparing your revised manuscript, please be aware of our guidance on Sex and Gender reporting). Please note that:

1) If the research findings apply to only one sex or gender, that must be indicated in the title and/or abstract.

2a) For studies involving vertebrates animal and cell lines- The Reporting Summary should include whether sex was were considered in the study design.

2b) For studies involving human research participants- The Reporting Summary should include whether sex and/or gender was considered in the study design and whether sex and/or gender of participants was determined based on self-report or assigned (and methodology used).

3) Data should be reported disaggregated for sex and gender where this information has been collected and consent has been obtained for reporting and sharing individual-level data; disaggregated numbers for individual experiments must be provided in the source data as appropriate whereas overall numbers may be provided in the Nature Portfolio Reporting Summary.

Information on the 3 points above should be included in the revised manuscript and detailed in the cover letter.

In addition, please note that if sex- and gender-based analyses have been performed a priori, results should be reported regardless of positive or negative outcome. We discourage conducting post hoc sex- and gender-based analysis if the study design is insufficient (for example, low sample size) to enable meaningful conclusions.

If no sex- and gender-based analyses have been performed, please indicate the reasons for the lack of these analyses in the Reporting Summary.

We encourage you to archive the data reported in your manuscript in an accessible, persistent repository. If your data are archived prior to the acceptance of your manuscript, please provide us with the full citation as soon as you receive it so that a link to the data can be included in the publication. See http://www.nature.com/authors/policies/availability.html for more information.

We have updated the manuscript and the Reporting Summary to comply with Nature Portfolio's Sex and Gender Reporting guidelines. Sex was considered in the study design and was determined for TEDDY participants at enrollment based on clinical records. Because this study evaluated biological sex rather than gender, we use the term sex consistently throughout the manuscript. The distribution of male and female participants is reported in Table 1. Sex was included as a covariate in all multivariable association and survival models. In addition, we performed pre-specified sex-stratified analyses of our primary findings, yielding results consistent with the main analyses (see new Supplementary Table). The study included both male and female participants, and the findings are applicable to both sexes. These details have been incorporated into the revised manuscript and are reflected in the Reporting Summary submitted with this revision.

If your paper is accepted for publication, we will edit your display items electronically so they conform to our house style and will reproduce clearly in print. If necessary, we will re-size figures to fit single or double column width. If your figures contain several parts, the parts should form a neat rectangle when assembled. Choosing the right electronic format at this stage will speed up the processing of your paper and give the best possible results in print. If you are in doubt about the correct format for your figures after reading our guidelines, please ask the art editors for advice metabolism@nature.com.

When submitting the revised version of your manuscript, please pay close attention to our href="https://www.nature.com/nature-portfolio/editorial-policies/image-integrity">Digital Image Integrity Guidelines. and to the following points below:

-- that unprocessed scans are clearly labelled and match the gels and western blots presented in figures.

-- that control panels for gels and western blots are appropriately described as loading on sample processing controls

-- all images in the paper are checked for duplication of panels and for splicing of gel lanes.

EXTENDED DATA FIGURES

When re-submitting your manuscript, please ensure that any supplementary figures and tables that are crucial to the manuscript's conclusions are converted into Extended Data figures and tables to increase visibility of these data. Extended Data figures and tables are online-only (present in the online PDF and full-text HTML versions of the paper), peer-reviewed display items that provide essential background to the article but are not included in the main article due to space constraints. A maximum of ten Extended Data display items (figures and tables) is permitted.

Finally, please ensure that you retain unprocessed data and metadata files after publication, ideally archiving data in perpetuity, as these may be requested during the peer review and production process or after publication if any issues arise.

> We have promoted the supplementary figures and tables most central to the manuscript's conclusions to Extended Data figures and tables, while remaining within the maximum limit of ten Extended Data display items. The remaining supporting materials are provided as Supplementary Information.

We are constantly trying to improve the quality of methods and statistics reporting in the papers we publish. To that end, we ask you to pay particular attention to the detail provided for error bars and uncertainties you report in your work. Definitions of statistical methods and measures must be provided. Error bars should always be defined (stating accompanying sample size n) as standard deviation, standard error on the mean, confidence interval, or other measure of variability as appropriate. Curves drawn on top of data should be described as fits or guides to the eye, as applicable, and details about the fit should be provided. All this information should be made available in the relevant figure caption (or the Methods section if too long).

> We have revised the figure legends and Methods, where appropriate, to clearly define all statistical methods, error bars, sample sizes, and curve descriptions in accordance with the journal's reporting guidelines.

Please include a data availability statement as a separate section after Methods but before references, under the heading "Data Availability". This section should inform readers about the availability of the data used to support the conclusions of your study. This information includes accession codes to public repositories (data banks for protein, DNA or RNA sequences, microarray, proteomics data etc…), references to source data published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", mentioning any restrictions on availability. If DOIs are provided, we also strongly encourage including these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see:

http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

> We have included a Data Availability section as a separate section between the Methods and the References. The section provides access information for all datasets supporting this study, including the controlled-access TEDDY metagenomic data (dbGaP accession phs001443.v1.p1), host genotype data (dbGaP accession phs001037.v2.p1), the associated clinical metadata through the NIDDK Central Repository, and the processed DIABIMMUNE dataset.

Nature's preferred approach to sharing research data is via public repositories (http://www.nature.com/authors/policies/availability.html#data) but when this is not possible a DAS should include, at a minimum, a statement confirming that all relevant data are available from the authors, and/or are included with the manuscript (e.g. as source data or Supplementary Information), listing which data are included (e.g. by figure panels and data types) and mentioning any restrictions on availability. If a dataset generated or analysed during the study is publicly available and has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Methods.

Please note that the minimum DAS is:

"The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request."

Please feel free to amend this statement as appropriate, and please let me know if you have any questions about our data availability policy.

Please ensure your manuscript includes a statement declaring any financial and non-financial competing interests you or your co-authors may have. For more details, please see our Competing Interests policy page.

For figures, please note that mean and standard deviation are not appropriate with small sample sizes, and bar graphs are often misleading. Plotting independent data points is usually more informative and we recommend the use of scatter plots for graphs with less than 5-10 data points.

Figure legends must provide a brief description of the figure and the symbols used, including definitions of any error bars employed in the figures. When technical replicates are reported, error and significance measures reflect the experimental variability, not the variability of the biological process; please ensure the number and nature of replicates (i.e., biological vs. technical replicates) is reported correctly in the figure legend. All abbreviations used should be defined. Figure legends should be no longer than 350 words.

As a guideline, Articles allow up to 50 references (excluding those cited exclusively in Methods).

Please include a statement before the Acknowledgements naming the author to whom correspondence and requests for materials should be addressed.

Please also complete and upload a Reporting Summary form, (found here https://www.nature.com/authors/policies/ReportingSummary.pdf) which collects information on experimental design and reagents. This document is available to referees to aid the evaluation of the manuscript, and will be published with the manuscript.

Please note that this form is a dynamic 'smart pdf' and must therefore be downloaded and completed in Adobe Reader.

Finally, we require authors to include a statement of their individual contributions to the paper -- such as experimental work, project planning, data analysis, etc. -- immediately after the acknowledgements. The statement should be short, and refer to authors by their initials. For details please see the Authorship section of our joint Editorial policies at http://www.nature.com/authors/policies/authorship.html.

> We have revised the manuscript accordingly and ensured that all journal formatting and reporting requirements have been addressed, including the Data Availability Statement, Competing Interests declaration, figure legends, Reporting Summary, Correspondence statement, and Author Contributions.

Responses to Reviewers' Comments

Referee #1: microbiome, bioinformatics

Dong et al analyzed a large number of microbiome samples sequenced metagenomically with 100x2 HiSeq data in relation to host genetic detail and clinical variables. Unfortunately, it not possible to meaningfully evaluate the described microbiome results due to lack of clarity with the methods including databases and parameters used. There further seems to be possible sample independence violations with the staistics. While the authors indicate their methods are on github, the github repository does not address how the metagenomic data were processed or treated. The results and statements derived from them only make sense in light of the methods, so I did not focus much on the results or main text.

Of critical note, the authors do not appear to appropriately filter the metagenomic data for human reads. Most importantly, the absence of appropriate filtering allows for subject identification in human associated samples (see https://www.nature.com/articles/s41564-023-01381-3). Additionally, inclusion of genomes with a complete Y-chromosome during filtering is essential to mitigate technical bias (see https://www.nature.com/articles/s41467-025-56077-5).

We thank the reviewer for their careful evaluation of the manuscript and for raising this important point regarding host read removal and data privacy. We fully agree that rigorous filtering of human reads and clear documentation of processing steps are essential both for data privacy and for the interpretability and reproducibility of microbiome analyses.

We first clarify a point that motivates several of our choices. The TEDDY study is a large multi-site, multi-country consortium in which metagenomic data generation and bioinformatic processing are centralized and standardized across clinical centers in Finland, Germany, Sweden, and the United States. This centralization is a deliberate design feature that ensures uniform quality control and cross-site comparability, but it also means that the core processing pipeline is maintained at the consortium level rather than controlled by individual analysis groups. Of the 12,151 metagenomes analyzed here, 10,913 were generated and processed in prior TEDDY publications[1,2], and the 1,238 newly added samples were further processed through the identical bioBakery 2.0 workflow specifically to maintain comparability with this existing, peer-reviewed corpus. Re-processing against a different reference would not only introduce batch effects between the newly added and previously published samples, but would also depart from the standardized processing applied across the broader TEDDY resource, a substantial undertaking in a consortium of this scale. We have made this rationale explicit in the revised Methods (Line 706).

We confirm that host-associated reads, including human-derived reads, were removed prior to downstream analysis. Specifically, raw metagenomic reads were aligned using Bowtie2 (v2.2.3) to a comprehensive reference database including bacterial, viral, human, and vector genomes derived from the NCBI whole-genome sequencing (WGS) archive (March 2015 release). Reads for which the highest-identity alignment was not bacterial (including those mapping to the human genome) were excluded from subsequent analyses.

We appreciate the reviewer highlighting the re-identification risk posed by residual host reads[4]. Two factors mitigate this risk in our study. First, all analyses in this manuscript are performed on taxonomic and functional feature tables (MetaPhlAn 2 and HUMAnN 2 outputs), which contain no human sequence. Second, the underlying sequencing reads are deposited in dbGaP under controlled access, the access-control mechanism recommended for protecting against re-identification from human-associated metagenomes.

We acknowledge that this procedure was not described with sufficient clarity in the original Methods. We have now revised the Methods section to explicitly state that reads mapping to the human genome were removed, and to clarify the alignment and filtering strategy.

Line 698: “Host and non-bacterial read decontamination was performed using Bowtie2 v2.2.3 with the --end-to-end --sensitive alignment mode. Reads were aligned against a comprehensive reference database derived from the NCBI whole-genome sequencing (WGS) archive (March 2015 release), including bacterial, viral, human, and vector genomes. Human reads were filtered using the GRCh38 (hg38) reference genome, and reads whose best alignment was non-bacterial were excluded from downstream analyses.”

In response to the reviewer’s concern regarding potential technical bias related to sex chromosomes, In response to the reviewer's concern regarding potential technical bias related to sex chromosomes, the human reference used for host-read filtering was GRCh38 (hg38) — the current major human reference — applied via Bowtie 2 alignment as described in our response to Comment 16. GRCh38 includes the Y chromosome, but its Y-chromosome assembly remains incomplete relative to the recently completed telomere-to-telomere assembly (T2T-CHM13v2.0), with substantial gapped regions on the Y long arm and in centromeric/ampliconic sequence. We therefore cannot claim that GRCh38-based host filtering masks the Y chromosome as completely as the most recent reference would.

To assess whether any residual bias could affect our conclusions, we note that sex was included as a covariate in all association and survival models, and we have additionally performed sex-stratified analyses of our primary findings. The association of the Incomplete Maturation pattern with T1D risk was consistent with our main results (**Supplementary Table 7**). This indicates our central conclusions are not driven by sex-correlated technical artifacts.

Supplementary 7: Sex-stratified associations between the microbiome maturational patterns and islet autoantibody seroconversion or T1D during follow-up in the TEDDY Study

	Sex	Early Matured	Late Matured	Early Plateaued	P[4]
Seroconversion or T1D					
Model 1[2]	Male	Ref.	0.77 (0.53,1.12)	3.35 (2.01,5.59)	3.21e-07
Model 2[3]	Male	Ref.	0.77 (0.52,1.14)	4.12 (2.42,7.01)	2.28e-08
Seroconversion					
Model 1	Male	Ref.	0.76 (0.52,1.11)	3.44 (2.06,5.76)	1.70e-07
Model 2	Male	Ref.	0.75 (0.51,1.12)	4.22 (2.47,7.22)	1.10e-08
T1D					
Model 1	Male	Ref.	1.07 (0.52,2.2)	4.52 (1.11,18.45)	1.60e-01
Model 2	Male	Ref.	1.3 (0.59,2.84)	6.29 (1.42,27.93)	9.54e-02
Seroconversion or T1D					
Model 1	Female	Ref.	1.35 (0.86,2.11)	2.38 (1.3,4.34)	2.35e-02
Model 2	Female	Ref.	1.17 (0.74,1.85)	2.04 (1.06,3.9)	1.06e-01
Seroconversion					

Model 1	Female	Ref.	1.42 (0.89,2.26)	2.1 (1.11,3.98)	7.18e-02
Model 2	Female	Ref.	1.27 (0.79,2.05)	1.83 (0.91,3.65)	2.38e-01
T1D					
Model 1	Female	Ref.	2.23 (0.82,6.08)	7.4 (2,27.35)	1.19e-02
Model 2	Female	Ref.	1.54 (0.54,4.35)	7 (1.81,26.99)	1.65e-02

[1] Values are risk ratios (95% confidence intervals) estimated from conditional logistic models.
[2] Model 1 was adjusted for risk set, clinical center, sex, and family history of T1D.
[3] Model 2 was further adjusted for age, delivery mode (cesarean and vaginal deliveries), antibiotic use (Yes or No), probiotic use (Yes or No), time of solid food introduction (Days since birth), breastfeeding cessation (Days since birth) and the top 5 genetic principal components.
[4] P value was calculated using the likelihood ratio test by comparing models with and without maturational patterns.

We further note that the host genetic analyses use ImmunoChip variants on the GRCh37 (hg19) coordinate system, restricted to autosomal variants only (Response to Comment 4); the host-genetics side of our analysis is therefore by construction not subject to sex-chromosome coverage artifacts, regardless of the reference used.

The reviewer correctly notes that our repository did not adequately document metagenomic processing. We have updated the repository to include the complete host-filtering and profiling workflow, including reference databases and versions, Bowtie2 parameters, and the MetaPhlAn 2 / HUMAnN 2 commands, so that the full pipeline from raw reads to feature tables is reproducible.

Specific comments are below:

*1. Fig1, notes 12151 samples from 887 subjects collected "approximatey monthly" during the first five years of life. 5 * 60 * 887 = 53220. Less than 1/4 of the samples are represented? The main text notes 13 +/- 9 which is not monthly?*

We thank the reviewer for this careful observation and for highlighting the potential inconsistency in the sampling description. We agree that the original wording in our manuscript may have been misleading, and we have revised the manuscript accordingly. We also provide a more complete numerical accounting below.

Per the TEDDY protocol, stool samples are collected at approximately monthly intervals from 3 to 48 months of age, then every three months thereafter[5]. The maximum theoretical collection schedule over 5 years is therefore approximately 46 monthly + 4 quarterly samples per child (~50 per fully-followed participant).

Two distinct factors explain why the number of metagenomes per participant is substantially below the ~50-sample theoretical maximum implied by a strict monthly schedule over 5 years.

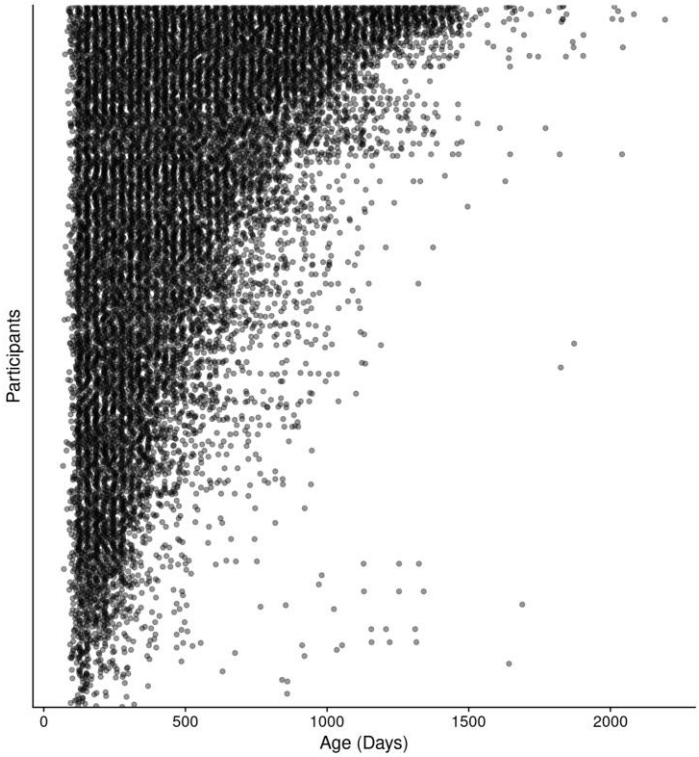
First, TEDDY is a dynamic prospective cohort study: participants were enrolled across a staggered calendar period from birth, while disease-endpoint ascertainment and the follow-up window analyzed here were cut at a fixed calendar time. Per-participant follow-up duration is therefore variable and right-censored — that is, children enrolled later in

the calendar period contributed correspondingly shorter follow-up at the analytic cutoff. As a consequence, most children did not contribute samples across the full age range, and the per-participant sample count reflects each child's available follow-up duration rather than a fixed 5-year monthly series.
Second, of the samples collected, a subset was selected for shotgun metagenomic sequencing based on sample availability, DNA quality and yield, and the sequencing design used across the multiple TEDDY metagenomic studies[1,2] and the present study. Across 887 participants this yielded the 12,151 metagenomes analyzed here. We recognize the original wording ("approximately monthly") conflated the collection protocol with the sequencing dataset, and we have revised the text to be explicit:

"Stool samples were collected approximately monthly through 48 months of age and quarterly thereafter, per the TEDDY protocol. Because TEDDY is a dynamic prospective cohort study with right-censored follow-up at the analytic cutoff, per-participant follow-up duration is variable. A subset of these stool samples was selected for shotgun metagenomic sequencing, yielding 12,151 metagenomes from 887 participants (mean 13.7 ± 8.7 sequenced samples per participant over follow-up)."

As a result, the number of metagenomic samples per participant is substantially lower than the theoretical maximum based on the collection schedule. The observed average of 13 ± 9 sequenced samples per participant over ~5 years is therefore consistent with prior TEDDY microbiome studies, and reflects the sparser post-4-year tail that reflects both real-world sampling and sequencing selection. Because our primary analyses focus on microbiome maturation during the first ~800 days, the more relevant statistic is the sampling density within that window: Participants contributed a mean of 10.7 ± 5.6 sequenced samples during the first 2 years of life (median 11; interquartile range 6–15), corresponding to an average inter-sample interval of approximately 40.7 days during early life. This dense early-life sampling is the basis for the trajectory clustering analyses presented in Figure 2.

In addition, to provide full transparency regarding the per-participant sampling density, we have added a new supplementary figure (**Supplementary Fig. 1**) showing the distribution of metagenomic samples per participant across the cohort, along with the temporal distribution of samples across age. The figure clearly shows the dense, near-monthly sampling pattern during the first ~1,400 days followed by a sparser quarterly cadence thereafter, consistent with the TEDDY protocol.
We hope this clarification resolves the reviewer’s concern.



Supplementary Figure 1. Per-participant distribution of metagenomic stool samples across age. Each dot represents one shotgun-sequenced stool sample from the analyzed dataset (12,151 samples from 887 participants). The y-axis represents individual participants, sorted by sample density; the x-axis represents participant age in days at the time of sample collection. The plot illustrates the longitudinal sampling structure of the cohort, including inter-individual variability in follow-up duration and sampling density. Consistent with the TEDDY protocol and with prior TEDDY microbiome studies[1,2], sampling is densest during early life (approximately 0-1,400 days, corresponding to the first ~4 years) and becomes sparser thereafter.

2. Fig1b, if for five years, why does the plot extend to what looks like at least six years?

> We thank the reviewer for this careful observation. The reviewer is correct that Fig. 1b extends beyond 5 years, with a maximum follow-up of 2,193 days (~6 years) and mean and median follow-up durations of 706.8 and 649 days, respectively. The text's description of the sampling as spanning "5 years" was therefore an imprecise descriptor, and we have corrected it accordingly.
>
> In our analysis, we included all available longitudinal samples from early childhood rather than applying a strict 1,825-day cutoff, so the plotted range reflects the true span of the data; in practice, sampling density is highest during the first ~4 years of life and becomes progressively sparser thereafter, with only a small subset of participants contributing samples beyond 5 years of age.

We have revised the manuscript to describe the sampling window precisely and, importantly, to use identical wording across the main text, Methods, and all figure legends. The revised description reads: The revised text now reads (Line 150): *"…longitudinally collected stool samples spanning early childhood, with dense sampling through the first ~4 years of life and sparser sampling extending up to ~6 years in a subset of participants…"*. We have also updated the Fig. 1b legend to reflect the same window, and verified that all other references to the sampling duration in the main text, Methods, and figure legends are now consistent.

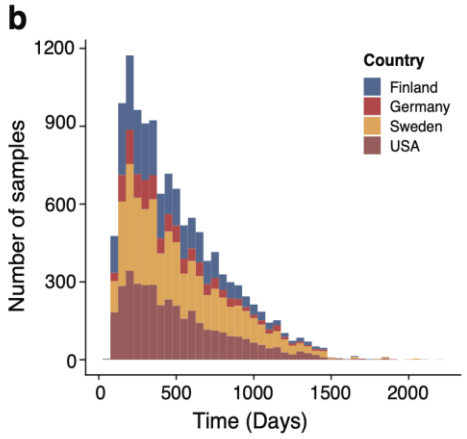


Figure 1: (b) The distribution of metagenomes remains consistent across different countries throughout the follow-up period, with dense sampling during the first ~4 years of life and sparser sampling extending up to ~6 years in a subset of participants. Each bar in the stacked bar plot represents the country-specific distribution of metagenomes within a 50-day interval.

3. Fig1c, humans have 23 chromsomes but the plot shows 22?

We thank the reviewer for pointing this out. The original plot included only autosomes (chromosomes 1-22) and omitted the sex chromosomes, and our downstream analyses were restricted to autosomal variants to ensure consistency and robustness. This was intentional, but as the reviewer notes, it was not stated in the original legend.
We restricted the visualization, and more importantly, all downstream genetic analyses, to autosomal variants for two reasons. First, the Illumina Immunochip[6] provides limited and uneven coverage of the sex chromosomes, making them unsuitable for the variant-density summary shown here. Second, excluding sex chromosomes from genetic analyses is standard practice in population-genetic and host-genome–microbiome studies, as their distinct inheritance and coverage properties can introduce artifacts into principal-component and association analyses. Accordingly, our genetic principal components and all SNP–microbiome association analyses were computed from autosomal variants only. We have now revised the figure legend to explicitly state that Fig. 1c displays autosomes only.

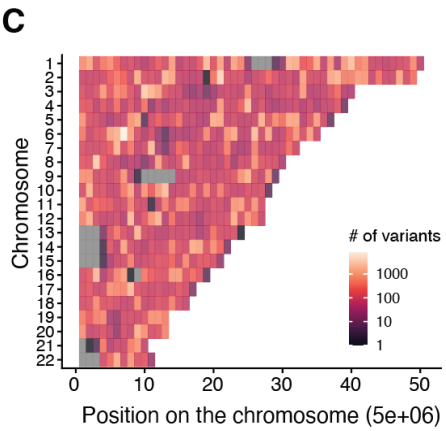


Figure 1: (c) Participants underwent genotyping using the Infinium ImmunoArray-24 v2 BeadChip, which was designed to detect genetic variations in the human immune system. The plot displays autosomal chromosomes only (chromosomes 1–22). Grey indicates bins containing no genotyped variants.

4. Fig1e, this plot is misleading. The genetic data on the x-axis are (presumably) singular points per subject whereas the microbiome data are between one and many points per subject on the y-axis. Ordinations are sensitive to the data included, and the dimensionality reduction can be biased if samples are not independent as is the case for subjects with more than one sample. Given the result in fig1b and the implication that < 1/4 of microbiome samples are represented, I would additionally assume that not all subjects are evenly sampled by time point or the number of samples. This plot would be more meaningful if the authors constrained it to a common time point and use a single sample per subject. Additionally, what is typical for multiomic microbiome comparisons between distances is Mantel correlation, and when comparing ordinations, Procrustes. What are the corresponding Mantel and Procrustes statistics at say time point zero, 1 year, 2 years, etc?

Thank you for this insightful comment. We agree that ordination methods can be sensitive to non-independence when multiple samples per subject are included, and that uneven sampling across subjects has the ability to change the major axes of visualization. We have followed the reviewer's recommended approach and performed additional analyses to formally evaluate the relationship between host genetic and microbiome structure.

Our intention in Fig. 1e was to illustrate the longitudinal trajectory of the microbiome in relation to age, not as a formal test of association between genetic and microbiome distances. We agree this was neither clearly stated nor the appropriate analysis for that question, and we have revised the legend accordingly.

Revised Figure legend: "(e) Host genetic architecture, as summarized by the first principal component (PC1), and gut microbial community structure, as summarized by the first principal coordinate (PCo1) based on species-level Bray-Curtis dissimilarity.

Points are colored by age, illustrating the age-related changes in microbiome community structure during early life."

Following the reviewer's suggestion, we restricted the analysis to a single sample per subject — selecting, for each subject, the sample closest to each target time point. We include the same ordination using only this subset of time points as Supplementary Figure 2. It does not substantively change the interpretation that, as calculated by our original Mantel statistic in the first submission, host genetics and early life microbiome structure are not meaningfully correlated, although a weak but statistically significant correlation was observed at baseline (Spearman ρ = 0.12, *P* = 0.0008).

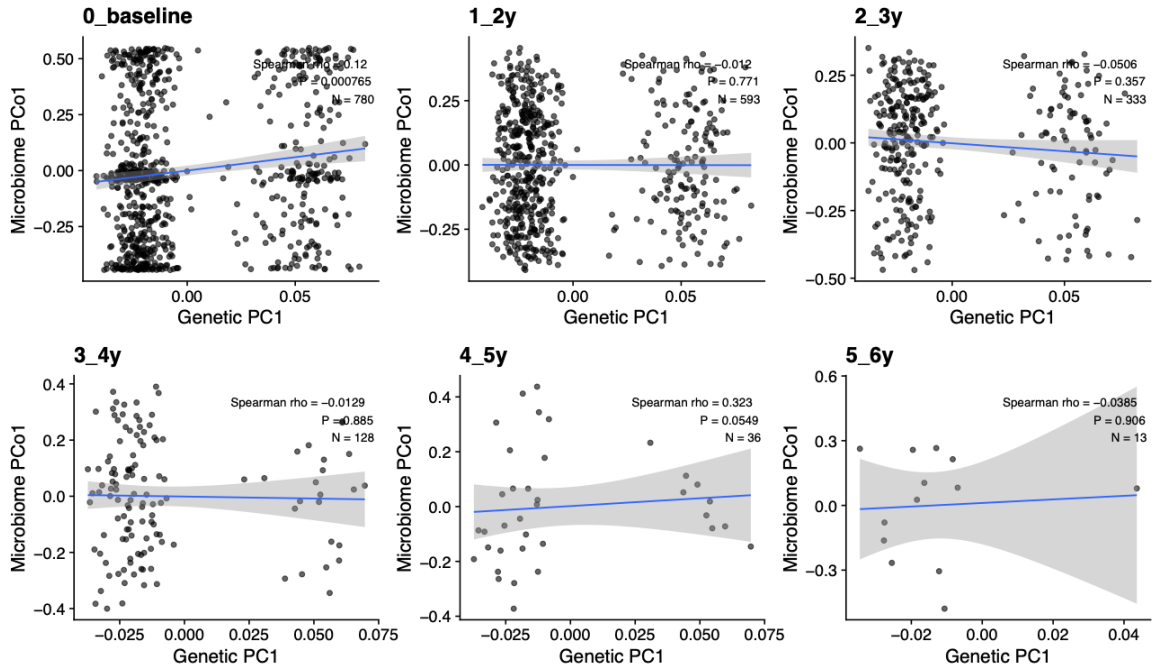
To quantify this, at each time point (baseline, 1 year, 2 years, etc), we then computed (i) Mantel correlations between the host genetic distance matrix and the microbiome distance matrix, and (ii) Procrustes statistics comparing the genetic and microbiome ordinations. The results are summarized below and have been added as **Supplementary Table 2**:

Supplementary Table 2. Mantel and Procrustes analyses comparing host genetic distance and gut microbiome distance across age groups.

Time Bin	N	Mantel_r	Mantel_p	Procrustes_r	Procrustes_p
< 1 Year	780	0.0019	0.56	0.0756	0.12
1-2 Year	593	0.0097	0.27	0.0864	0.16
2-3 Year	333	0.0134	0.25	0.1065	0.35
3-4 Year	128	0.0444	0.09	0.1607	0.53
4-5 Year	36	0.0511	0.23	0.3825	0.07
5-6 Year	13	-0.0576	0.66	0.4573	0.83

Across all age groups, host genetic distance showed weak and non-significant concordance with microbiome distance by both Mantel and Procrustes analyses. These independent, sample-independent tests are fully consistent with our primary finding that host genetics exerts only a modest influence on overall early-life microbiome structure (e.g., the small PERMANOVA effect sizes reported in the main text).

We are grateful to the reviewer, as this analysis strengthens that conclusion using the methods most appropriate for the comparison. As the reviewer anticipated, subjects are not evenly sampled across ages, so the number of subjects with a sample near each target time point differs; we report the per-time-point N in Supplementary Table 2 so that the independence constraint is explicit and results were robust to the width of the age window used to select the nearest sample.



Supplementary Figure 2. Association between host genetics and gut microbiome structure across developmental time points. Each panel shows the association between genetic PC1 and microbiome PCo1 at a given developmental time point. To ensure sample independence, a single sample per participant was included at each time point, defined as the earliest available sample within the corresponding time window. Points represent individual subjects. Blue lines indicate fitted linear regression lines and gray shading indicates 95% confidence intervals. Corresponding correlation coefficients, P values, and sample sizes are shown in each panel.

5. Fig1f, PERMANOVA assumes sample independence. Was that controlled for here or are multiple samples from the same subject included?

Thank you for raising this important point. We agree that PERMANOVA assumes independence among samples, and that repeated measures within subjects need to be properly accounted for.

We did account for this in the original analysis, but we described the permutation scheme imprecisely in the Methods, and we are grateful for the opportunity to correct it. Rather than permuting samples freely (which would treat the longitudinal samples as independent and anti-conservatively inflate significance), we used a restricted permutation design in which whole subjects were the exchangeable unit: subjects were permuted freely while samples were never permuted within a subject, so that each subject's full set of repeated measures was kept intact under permutation. This is the appropriate scheme for testing a between-subject variable and preserves the within-subject correlation structure of the longitudinal data.

Under this independence-respecting design, both between-subject variables explained only small fractions of overall microbiome variation, and neither reached significance: genetic PC1, $R^2 = 0.28\%$, $p = 0.99$; country of origin, $R^2 = 0.87\%$, $p = 0.99$. We note this explicitly because it bears on the reviewer's later comment regarding the Mantel analysis: because the country effect on the early-life microbiome is itself small and non-significant, the weak genetic–microbiome concordance by Mantel/Procrustes is the expected result rather than a contradictory one.

We have now clarified this in the Methods section (Line 771).

“PERMANOVA was performed on Bray–Curtis distances using adonis2 in the vegan R package. To respect the non-independence of repeated longitudinal samples, we used a restricted permutation design in which whole subjects were treated as the exchangeable unit: subjects were freely permuted while no permutation was performed within subjects. This is the appropriate scheme for testing between-subject covariates such as genetic principal components and country of origin, as it keeps each subject's repeated measures intact and prevents anti-conservative inflation of significance from pseudo-replication. Significance was assessed using 999 permutations.”

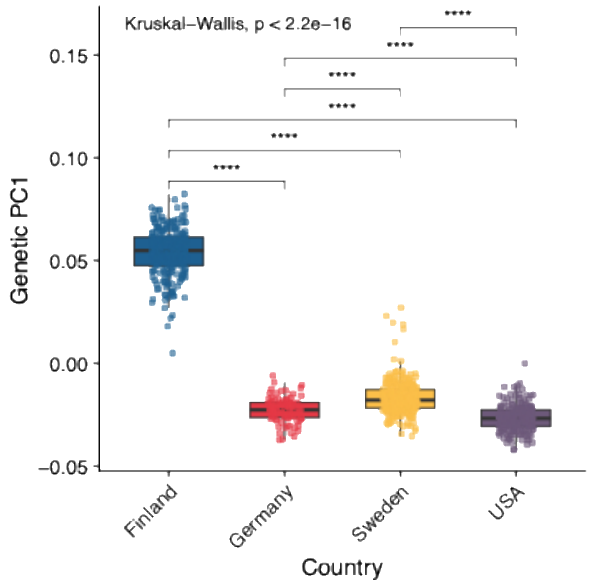
6. line 118, is there a statistic to support the statement of "strong correlation"?

Thank you for this helpful suggestion. We agree that the original phrasing “strong correlation” was not sufficiently supported. We have revised the text to more accurately describe the relationship and added a formal statistical test and an effect size estimate. Specifically, we assessed the association between genetic PC1 and geographical location using a Kruskal–Wallis test across countries, which showed highly significant differences ($\chi^2 = 589.16$, $df = 3$, $p < 2.2 \times 10^{-16}$), with a large effect size (72.8% of the variance in PC1 explained by country), confirming that PC1 strongly reflects population genetic structure aligned with geographic origin. This is also visually evident in Response Figure 1, in which Finnish participants in particular are clearly separated along PC1. We have revised the text accordingly (Line 120):

“As expected, we observed a strong correlation between the first genetic principal component (PC1) and participants' geographical location($\chi^2 = 589.16$, $df = 3$, $p < 2.2 \times 10^{-16}$; 72.8% of the variance in PC1 explained by country).”

We note that this strong geographic structuring of host genetic ancestry is fully consistent with our broader finding that host genetics exerts only a modest direct effect on overall microbiome composition: PC1 captures population structure robustly, yet that structure explains only a small fraction of microbiome variation

(Figure 1f), underscoring that the genetic–microbiome relationship in early life is subtle despite well-defined underlying ancestry.



Response Figure 1: Distribution of genetic principal component 1 (PC1) across countries. Boxplots represent the median and interquartile range, with individual samples shown as points. Differences across countries were assessed using a Kruskal–Wallis test ($\chi^2 = 589.16$, df = 3, $p < 2.2 \times 10^{-16}$; 72.8% of the variance in PC1 explained by country). Pairwise comparisons were performed using Wilcoxon rank-sum tests with Benjamini–Hochberg correction. Significance levels are indicated as *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

7. line 130, please see the comments on that figure subpanel specifically for suggestions on computing a correlation. This and on line 118, the authors are making statements using statistical language ("correlate" here) but without indicating what the statistics used are, their effect sizes or p-values.

Thank you for this helpful suggestion. We agree that our original wording did not clearly specify the statistical tests and effect sizes. We have now revised the text to explicitly report the statistical analyses and results.

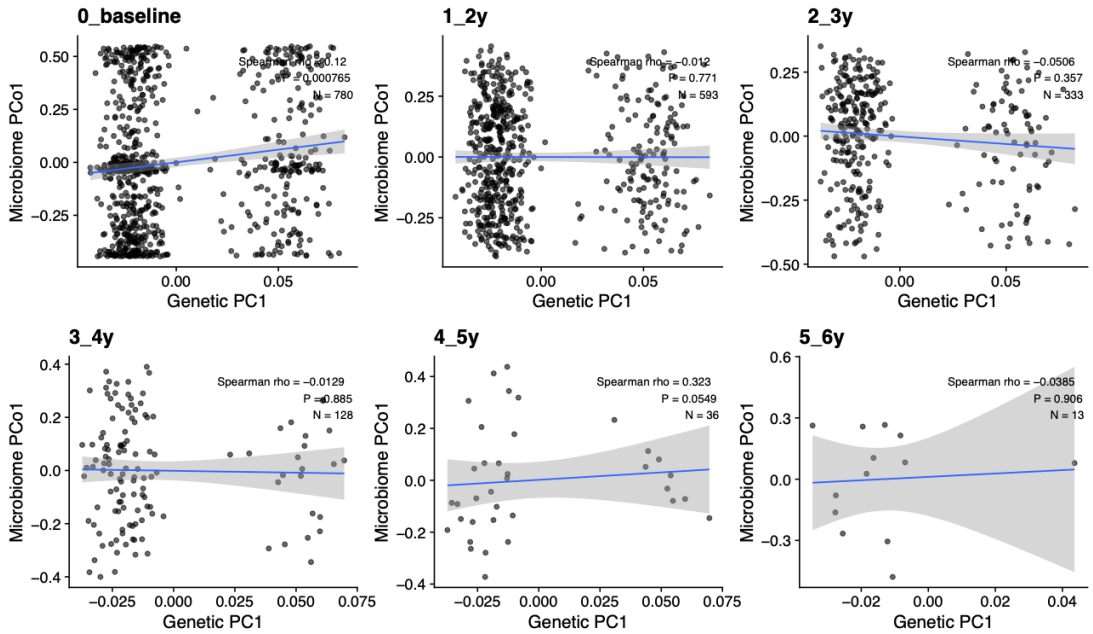
We additionally assessed the relationship between genetic PC1 and microbiome PCo1 using Spearman correlation based on a single-sample-per-subject analysis. A weak correlation was observed at baseline ($\rho = 0.12$, $P = 7.6 \times 10^{-4}$), whereas no significant associations were detected at later time points. Consistent with the Mantel and Procrustes analyses, these results suggest that the overall relationship between host genetic variation and microbiome structure is weak (Supplementary Table 2).

In addition, PERMANOVA analysis, using the subject-level restricted-permutation design described in our response to the Figure 1f comment (Comment 5), showed that genetic

PC1 explained a negligible proportion of the variation in overall microbiome composition (R^2 = 0.28%, P = 0.99). In the same framework, country of origin also explained only a small and non-significant fraction of microbiome variation (R^2 = 0.87%, P = 0.99).

We have updated the manuscript accordingly and clarified the statistical methods and results.

Revised text (Line 131): Based on previous studies on host genetics and the gut microbiome[7], we hypothesized that host genetics would have a smaller impact on microbial community structure than host phenotypic factors. Consistent with this hypothesis, genetic PC1 showed only a weak association with microbiome PCo1 at baseline (Spearman's ρ = 0.12, P = 7.6 × 10^{-4}, **Supplementary Figure 2**), and no significant associations were observed at later time points. Similarly, Mantel and Procrustes analyses comparing host genetic and microbiome distance matrices demonstrated weak and non-significant concordance across all age groups (**Supplementary Table 2**). This was further supported by the non-significant association between genetic PC1 and overall microbial community variation (PERMANOVA: R^2 = 0.28%, P = 0.99; 999 permutations).



Supplementary Figure 2. Association between host genetics and gut microbiome structure across developmental time points. Each panel shows the association between genetic PC1 and microbiome PCo1 at a given developmental time point. To ensure sample independence, a single sample per participant was included at each time point, defined as the earliest available sample within the corresponding time window. Points represent individual subjects. Blue lines indicate fitted linear regression lines and gray shading indicates 95% confidence intervals. Corresponding correlation coefficients, P values, and sample sizes are shown in each panel.

8. line 142, ah, okay, Mantel was applied. Was this Pearson or Spearman? Were subjects not included at a given time point given a zero distance value, or were the genetic data subset for each test? I can't find clarity in the methods. Procrustes is still recommended in addition to Mantel. Methods aside, this statement is surprising as country is a well known association with the microbiome particularly with methods that account for evolution. The authors show that both the genetic data exhibit a clear separation in fig1d, and a measurable microbiome effect (violating sample independence?) in fig1f. Given that both data layers used with Mantel demonstrate a country effect, it seems surprising the effect is not supported by Mantel, which I believe would only occur of the country effects were inconsistent which would itself be surprising.

We thank the reviewer for these detailed and insightful comments, and for the careful reasoning about why a null Mantel result might seem surprising. We address the methodological questions first, then the conceptual point, which we think is important. In brief, the reviewer's statement that geography *can* be associated with microbiome structure is accurate, but it often is not, particularly among relatively culturally homogeneous locations such as developed, urban populations. This is broadly supported by many previous publications[7,8], and we extend its quantification in our data below.

Specifically, we used Spearman rank correlation for all Mantel tests. To respect sample independence, each test was performed separately at a given developmental time point using a single sample per subject. Subjects without a sample at a given time point were simply excluded from that test: both the genetic and microbiome distance matrices were subset to the intersection of subjects present at that time point, so that the two matrices always contained the identical subject set. No imputation or zero-distance assignment was used for absent subjects. We have revised the Methods to state this explicitly (Line 783), and we report the per-time-point subject counts in **Supplementary Table 2**.

As recommended, we have added Procrustes analyses alongside the Mantel tests at each time point; these are reported together in Supplementary Table 2 (see also our response to Comment 4). Again, this did not change our conclusions, inasmuch as both analyses consistently indicated little correspondence between host genetic and microbiome distances across time points.

On why the country signal does not produce a positive Mantel correlation, the reviewer's reasoning is logical, and we want to address it directly rather than generically. The key point is an asymmetry between the two data layers. In this population, host genetics segregate very strongly by country: in our data, this explains a large share of genetic PC1 variance (ε^2 = 72.8%; see response to line 118, Comment 6), consistent with the clear separation in Figure 1d. However, country explains only a small and statistically non-significant share of overall microbiome composition in this early-life cohort (PERMANOVA R^2 = 0.87 %, p = 0.99, subject-level permutation). Again, this is not

surprising when comparing high-income, industrialized, Westernized populations with similar early life cultural and dietary exposures [7,8].

Thus because the microbiome distance matrix is dominated by non-country, individual-level variation, there is no substantial shared country structure for Mantel to detect across the two matrices. Therefore, a weak, non-significant Mantel correlation is the expected result, not a contradictory one. This is therefore not a case of "inconsistent" country effects, but of a strong country effect in one layer and a negligible one in the other.

We agree with the reviewer that country and biogeography are well-established correlates of the adult gut microbiome under other circumstances - such as hunter-gatherer populations or areas with differences in early life hygiene - and this can also be assessed using phylogeny-aware methods. The difference here reflects the early-life setting of our cohort: during the first years of life, microbiome variation is dominated by age/maturation, feeding (breastfeeding and weaning), and delivery mode[9,10], and the geographic/country signal, while present in adults, is comparatively small at these ages. We have added this point to the Discussion as follows:

Line 490: “We also note that, although geographic differences in the gut microbiome are well documented between populations with markedly different environmental, dietary, or lifestyle exposures, most notably in cross-continental comparisons between industrialized and traditional populations, such effects tend to be more modest within high-income, industrialized cohorts such as TEDDY, and are further attenuated in early life, when microbiome variation is dominated by age, feeding practices, and delivery mode[7-9,11].”

9. line 182 / line 191, using the provided definition of microbiome maturation, what specifically do the authors mean when they stay "...did not achieve full maturation..."? It looks like the read line has samples which are more similar to their baseline than the other trajectories. Or is it the observed variability in the curves? It looks like this is in part based on supplementary figure 1a, where its stated "...a distinct inflection point is observed at k=3" though it seems like an overstatement given the figure. Why do the authors assume disrete rather than continous spaces here?

We thank the reviewer for this important comment, which raises three related points: the definition of "full maturation," the basis of the k = 3 solution, and the discrete versus continuous representation. We address each below.

1. **Definition of microbiome maturation.** In our study, microbiome maturation was operationally defined as the divergence from the individual's early-life baseline microbiome composition, quantified using beta diversity. This definition is based on the well-established concept that early-life microbiomes represent an immature ecological state, and that maturation corresponds to progressive compositional shifts over time. One of its earliest high-profile uses was for diverse international populations in [11].

Thus by “did not achieve full maturation” (the Early Plateaued trajectory), we quantify this as divergence from baseline that remains limited and plateaus over time (at a Bray-Curtis dissimilarity of ~0.7), in contrast to the progressive divergence seen in the other trajectories (which reaches ~0.8). We agree that this is not the only possible interpretation of a smaller distance to baseline, but in this context the baseline represents an early developmental stage, and the observed patterns reflect a systematic delay in microbiome succession departing further from that baseline.
We have revised the manuscript to define the term precisely. Line 203: “The first pattern, named "Early Matured," was characterized by rapid early divergence from baseline within the first 400 days and remained relatively stable thereafter (Extended Data Figure 3). The second pattern, termed "Late Matured," exhibited slower initial divergence from baseline and lower microbial diversity during the first 400 days, but subsequently showed accelerated divergence and converged toward the Early Matured trajectory. The third pattern, called "Early Plateaued," maintained low diversity throughout the first 800 days and exhibited more limited divergence from baseline, reaching an earlier plateau than the other trajectories.”

We note that the biological significance of this trajectory is established by its association with downstream T1D risk (below) rather than resting on an a priori interpretation of the pattern as developmental “delay” per se.

2. **Discrete versus continuous representation.** We fully agree that microbiome profiles likely lie along a continuum, and that any discrete clustering represents an approximation of an underlying continuous structure. We adopted clustering as a pragmatic discretization that yields interpretable, clinically tractable groups for risk stratification. Importantly, the central biological conclusion does not depend on discreteness. We have revised the manuscript to frame the clusters explicitly as a discretization of a continuum.
   Revised text: Line 190: “Our algorithm identified three major patterns (**Figure 2a, Supplementary Figures 3 and 4**) that summarize variation across a continuous spectrum of microbiome development”.

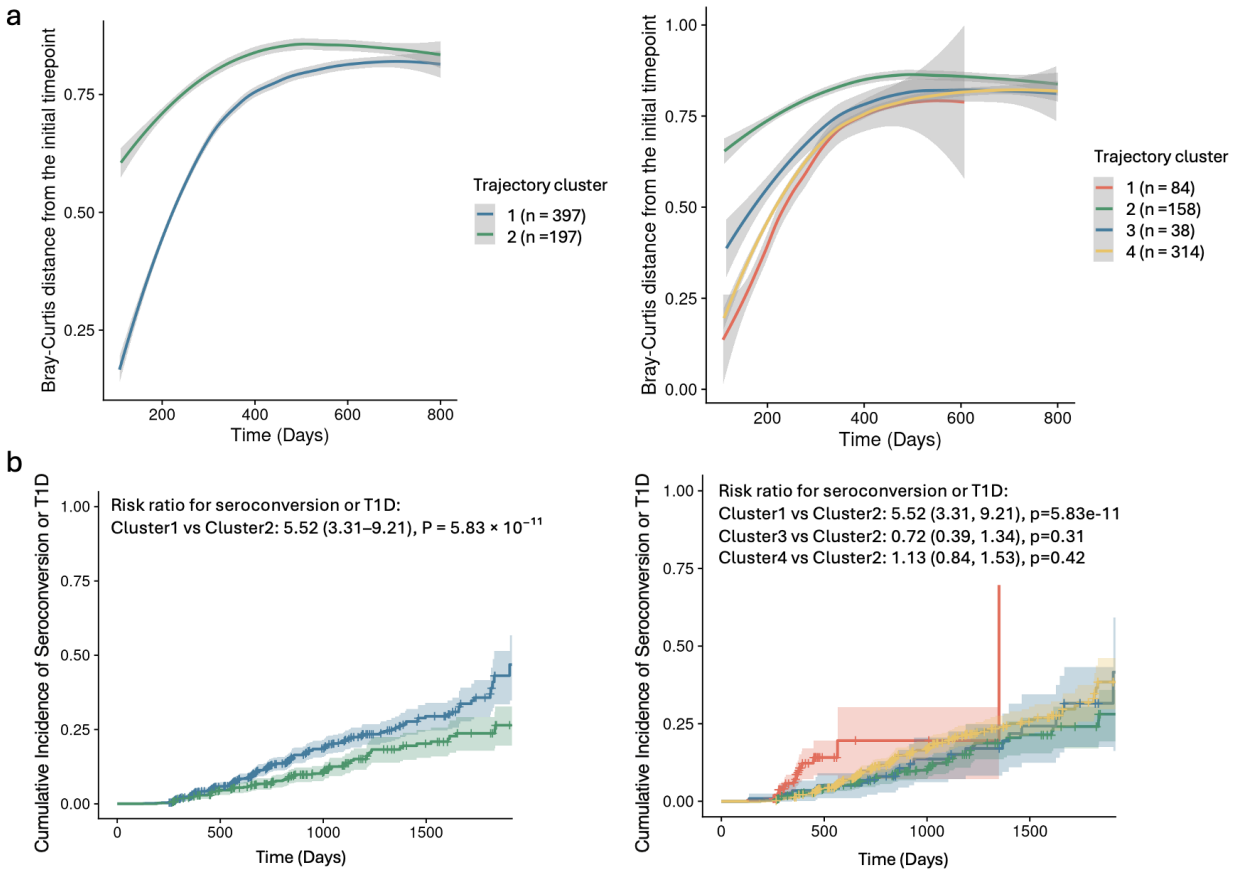
3. **Choice of k = 3.** We selected this value based on a combination of quantitative criteria, as detailed in Supplementary Figure 1:

   - Elbow Method (Panel A): The within-cluster sum of squares shows a marked reduction in slope around k = 3, beyond which additional clusters yield diminishing returns. We note that this transition is gradual rather than sharp, and we have revised the corresponding text in the Supplementary Material to describe it accordingly rather than as a "distinct inflection point."
   Revised text:

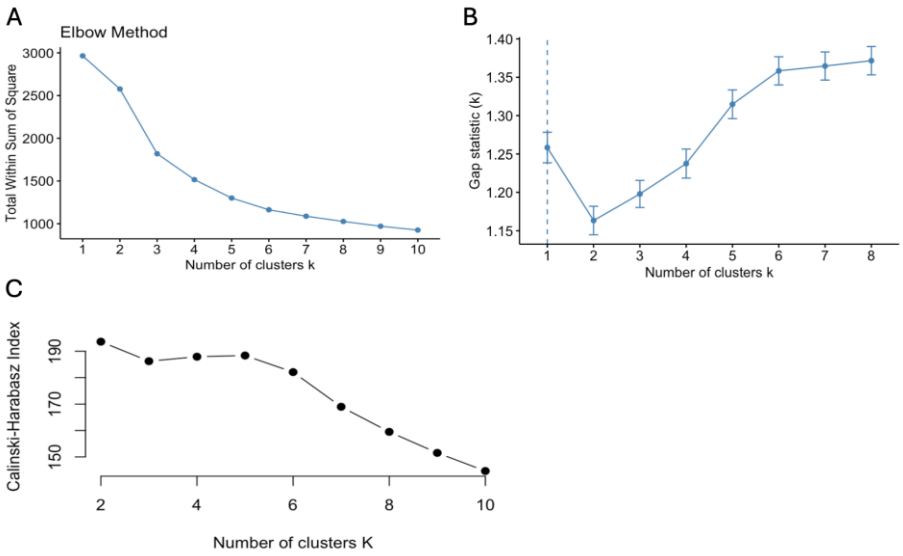
Figure file, line 106: (A) The Elbow Method showing the Total Within Sum of Squares; a marked reduction in slope is observed around k = 3, beyond which additional clusters yield diminishing returns.

- Calinski–Harabasz Index (Panel C): Maximized at k = 2 but remains high at k = 3 before declining sharply at higher k. While k = 2 offers strong separation, it collapses biologically distinctions between "Late Matured" and "Early Plateaued" trajectories that are differentiated by other criteria.
- Gap Statistic (Panel B): Trends upward at higher k, but solutions with k ≥ 4 produced cluster sizes too small to support robust statistical inference for T1D risk associations.

Given this combination of statistics, we therefore do not claim that the data-driven indices uniquely identify three clusters. Rather, k = 3 was chosen as a balance between resolving distinct maturation trajectories and retaining adequate sample sizes for downstream analyses.
Importantly, the association between microbiome maturation and T1D risk was qualitatively similar for adjacent solutions (k = 2 and k = 4), indicating that our conclusions are not dependent on the choice of exactly three clusters. We report this sensitivity analysis in Supplementary Figure 5.
Taken together, we view k = 3 as a parsimonious and interpretable summary of variation in microbiome maturation rather than evidence that microbiome development is inherently discrete.



Supplementary Figure 5. Sensitivity analyses of microbiome maturation trajectories using alternative cluster numbers (a) Sensitivity analyses using alternative numbers of trajectory clusters. Left, trajectory clustering with k = 2 identified two broad microbiome maturation patterns. Right, trajectory clustering with k = 4 identified four maturation patterns. Clusters were derived using a trajectory clustering algorithm based on Bray–Curtis distances from each participant's baseline gut microbiome composition across visits during the first 800 days of life. Trajectories were smoothed and visualized using LOESS curves; shaded bands indicate 95% confidence intervals. (b) Association between alternative maturation patterns and risk of islet autoantibody seroconversion or type 1 diabetes (T1D). Left, cumulative incidence curves and hazard ratio estimates for the two-cluster solution. Right, cumulative incidence curves and hazard ratio estimates for the four-cluster solution. Hazard ratios were estimated using multivariable Cox proportional hazards models adjusting for clinical center, sex, antibiotic use, probiotic use, breastfeeding cessation, age at solid food introduction, and the top five host genetic principal components. Shaded bands indicate 95% confidence intervals. Consistent results across clustering resolutions indicate that the association between impaired/non-progressive microbiome maturation and T1D risk is robust to the choice of k.



Supplementary Figure 3: Determination of the optimal number of microbiome maturational patterns. Diagnostic metrics for identifying the optimal number of clusters (k) using the traj package. (A) The Elbow Method showing the Total Within Sum of Squares; a marked reduction in slope is observed around k = 3, beyond which additional clusters yield diminishing returns. (B) The Gap Statistic (mean±SE). (C) The Calinski-Harabasz Index; the score remains high at k=3 before declining, indicating good cluster separation. The final selection of k=3 balances the inflection in the Elbow method and the high CH index with the need for an adequate sample size per cluster for downstream statistical inference.

10. line 192, how does fig2b demonstrate reproducibility? The plots look quite different and the relationships between the curves are different.

We thank the reviewer for this thoughtful observation and for prompting us to clarify what we mean by "reproducibility" in the context of Fig. 2b. We agree that the two panels in Fig. 2b are not numerically identical and that the exact shapes and relative positions of the three trajectories differ between the two cohorts.

First, we'd like to clarify what Figure 2b shows, and what it does not. We applied the same trajectory-clustering framework, independently and without supervision, to the DIABIMMUNE cohort. In both cohorts, the procedure recovered a set of qualitatively similar maturation patterns: a trajectory that begins at high distance from baseline and remains high; a trajectory that begins low and increases over time toward higher distances; and a trajectory with limited progressive divergence from baseline. We now describe this as a qualitative recapitulation of three maturation-trajectory shapes in an independent cohort, rather than as full reproducibility. We agree with the reviewer that the curves are not numerically equivalent, for instance, DIABIMMUNE Cluster 2 continues to rise across the observation window rather than asymptoting toward Cluster 1 as the corresponding TEDDY group does, and the exact crossover behavior differs

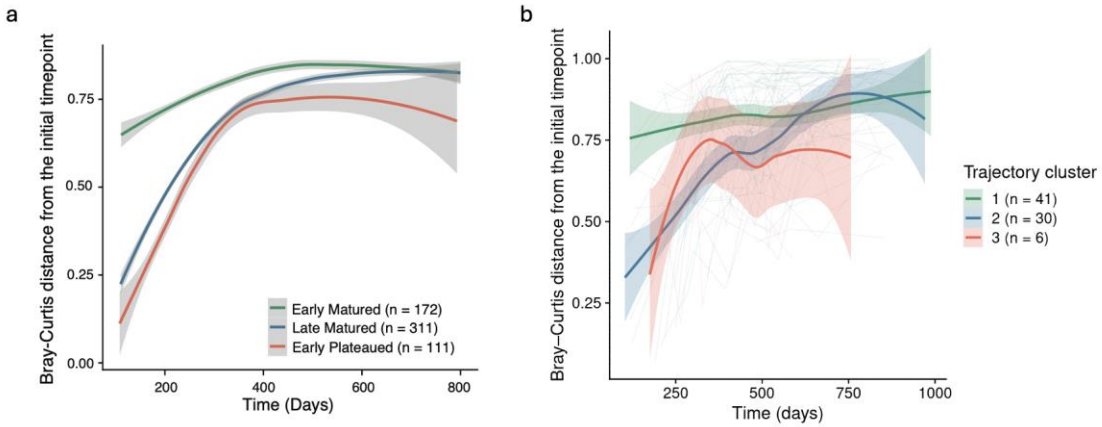
between cohorts. These differences are real and we acknowledge them explicitly in the revised text.

Revised text: Line 200: We applied the same trajectory-clustering framework to the independent DIABIMMUNE three-country cohort, and observed three qualitatively similar microbiome maturation trajectories.

This correspondence is a non-trivial finding: the same three qualitative patterns emerge from two independent cohorts that differ substantially in sample size, geography, and follow-up structure, supporting the generalizability of a three-trajectory framework for characterizing early-life microbiome maturation. Differences in the detailed shapes and crossover points between cohorts are biologically expected — cohort composition, feeding patterns, age distribution, and sampling density all influence the fitted curves — and do not undermine the qualitative consistency of the trajectory structure itself.

We would also note, in fairness to the limitations of any external replication available to us, that publicly available datasets matching the specific combination of features that defines TEDDY — a large, multi-country, genetically high-T1D-risk pediatric cohort with prospective dense longitudinal shotgun metagenomic sampling from infancy and prospectively ascertained T1D outcomes — are effectively absent at present. DIABIMMUNE is the closest available comparator (shared age range, longitudinal early-life sampling, partial geographic overlap with TEDDY), and we chose it for that reason despite the constraints the reviewer has fairly identified. We have noted this in the revised Discussion, framing Figure 2b as a check against the best-available external dataset rather than as a definitive cross-cohort validation.

To make this correspondence more immediately apparent to the reader, we have harmonized the color scheme across the two panels of Fig. 2b so that corresponding trajectory classes now appear in the same color in both cohorts.



Response Figure 2: (a) A data-driven approach revealed three early-life gut microbiome maturational patterns in the TEDDY dataset.: Early Matured, Late Matured, and Early Plateaued. The patterns were identified using a trajectory clustering algorithm that employs the Bray-Curtis distance to quantify changes in the gut microbial composition from baseline across subsequent visits during the initial 800-day period. Trajectories were smoothed and visualized using the LOESS algorithm. (b) Gut microbiome maturation patterns in the DIABIMMUNE cohort. Using the same trajectory clustering framework applied in TEDDY, we identified three data-driven gut microbiome maturation patterns in early life. The patterns were identified using a trajectory clustering algorithm that employs the Bray-Curtis distance to quantify changes in the gut microbial composition from baseline across subsequent visits during the initial 1000-day period.

11. fig2c, aggregating relative abundance data violates the compositional nature of microbiome data. These are at best misleading figures. I urge the authors to adopt compositionally appropriate measures typically used with microbiome data.

We thank the reviewer for raising this important point regarding the compositional nature of microbiome data. We fully agree that microbiome sequencing data are inherently compositional and that compositionally aware methods are essential for statistical inference. However, we respectfully disagree that Fig. 2c is misleading, and we would like to clarify the distinction between (i) descriptive visualization of per-taxon temporal trajectories and (ii) statistical inference on compositional data.

1. All inferential analyses identifying the species shown in Fig. 2c were performed using MaAsLin2, a compositionally aware framework. The taxa visualized in Fig. 2c were not selected on the basis of aggregated relative abundances. They were identified based on differential-abundance testing performed with MaAsLin2[12], which applies total-sum-scaling (TSS) normalization followed by log or arcsine-square-root transformation and fits per-feature mixed-effects models. This is a well-benchmarked and widely adopted framework for compositional microbiome data, and is the de facto standard in the field. No inferential conclusions — including which species are differentially trajected across the Early Matured, Late Matured, and Early Plateaued groups — rely on the aggregated relative abundances displayed in Fig. 2c.

2.Figure 2c is then a descriptive visualization of how each pre-identified taxon changes over age within each trajectory. Per-taxon longitudinal abundance plots of this kind are the established convention in early-life longitudinal microbiome studies, including the two foundational TEDDY metagenomic publications[2]. We adopted this convention to maintain comparability with prior reports and because per-taxon mean trajectories more directly communicate which species rise or fall across maturation groups than CLR-transformed trajectories would. We have revised the legend to state explicitly that Figure

2c is a descriptive visualization of features identified through compositionally appropriate inference (MaAsLin 2), not an inferential analysis itself (Line 275).

12. extended fig4, were the pathogens listed, like C. bolteae, actually observed in these data? Is it possible that C. bolteae is arising from short reads multimapping with the other clostridia? Indication of genome coverage could be helpful here. I see 52 species here, but the results section mentions 92, what about the other 40?

We thank the reviewer for these thoughtful technical questions and welcome the opportunity to clarify several points.

1. On the classification of *C. bolteae* as a "pathogen." We note that the term "pathogen" was used by the reviewer in reference to C. bolteae; our manuscript does not characterize these species as pathogens. It is a gram-positive, spore-forming anaerobic endogenous resident of the human gut, and although it has been associated with several disease states (including inflammatory bowel disease, autism-spectrum disorders, and antibiotic-associated diarrhea), its behavior is better described as a “pathobiont” because of this rather than as a frank pathogen. The manuscript thus lists it descriptively throughout without applying a pathogen/commensal label.

2. On the possibility of multimapping artifacts. This is an important technical concern, and we appreciate the reviewer raising it. MetaPhlAn 2 (and previous / subsequent versions) were specifically designed to mitigate such multimapping concerns as the reviewer raises. Its core algorithm relies on ~1 million clade-specific marker genes selected to be present in target species and absent from their close relatives, and calling a species as present requires concordant signal across multiple distinct markers (low-support single-marker hits are filtered by default[13]). This means that sequences used to identify *C. bolteae* are pre-selected to specifically not occur in other clostridia. Combined with the per-species marker-breadth data above, this provides both methodological and empirical assurance against the multimapping artifact.

3. On the apparent discrepancy between 52 species shown and 92 species mentioned in the Results. The 92 species referenced in the Results correspond to the full set of species tested for differential abundance across maturation groups, while the 52 species shown in Extended Fig. 4 are the subset that reached statistical significance ($q < 0.05$) at least once during the follow-up window. This selection criterion is stated in the original Extended Fig. 4 legend: "The heatmap displays microbial features that were differentially abundant (q-value <0.05) between maturational patterns at least once during the follow-up." The remaining 40 species were tested but did not reach significance under this criterion and were therefore not displayed in the heatmap. Nonetheless, we agree with the reviewer that providing the full results would improve transparency. The full differential abundance results for all 92 species tested were already included in Supplementary Tables 5 and 6.

We hope these clarifications and additions fully address the reviewer's concerns.

13. figure 5cd, are the authors surprised to not find a Bifidiobacterium association with a mutation in LCT for lactose intolerance? That is one of the dominant signals in other microbiome / human genome assessments.

We thank the reviewer for this excellent point, which allows us to highlight an important developmental distinction that is often overlooked in microbiome-genome studies.
The reviewer is correct that the LCT locus (rs4988235 and neighboring variants) represents one of the most robust host-genome–microbiome associations reported in adult populations, with lactase-nonpersistent (GG) genotype adults consistently showing higher Bifidobacterium abundance than lactase-persistent (AA/AG) adults [14-16]. However, this association is not expected to manifest in an early-life cohort such as ours, for reasons grounded in the basic developmental biology of the LCT locus.
The LCT genotype does not produce a lactase-activity phenotype during infancy and early childhood. Intestinal LCT expression and lactase activity are near-universally high at birth and throughout infancy, regardless of rs4988235 genotype, because newborns must digest lactose — the primary carbohydrate in breast milk — for survival [17,18]. The decline in LCT expression that underlies adult lactose non-persistence is developmentally programmed and begins only after weaning, typically from approximately 2 years of age onward, and does not fully distinguish lactase-persistent and lactase-nonpersistent groups until ~2–12 years of age [17-20]. The majority of sampling in TEDDY occurs during the developmental window in which LCT expression remains high across essentially all individuals, irrespective of rs4988235 genotype.
The mechanistic basis of the adult LCT-Bifidobacterium association therefore does not apply during infancy.

To make this distinction explicit for the reader, we have added the following text to the Discussion (Line 529):
"Notably, we did not recover the LCT (rs4988235)–Bifidobacterium association that represents one of the strongest adult host-genome–microbiome signals reported to date[14,15]. This is biologically consistent with the developmental context of our cohort: intestinal LCT expression is near-universally high during infancy regardless of rs4988235 genotype, and the adult LCT-mediated mechanism — in which undigested colonic lactose selectively promotes Bifidobacterium growth in lactase-nonpersistent adults — does not yet operate in this age range. Early-life Bifidobacterium abundance is instead predominantly driven by feeding mode, HMO exposure, and perinatal factors, which we account for in our models."
We thank the reviewer for raising this point, which has allowed us to clarify an important developmental distinction that strengthens the biological interpretation of our results.

14. line 501, HUMAnN2 is described in the methods section as capable of detecting virus. Was this virus detected? If not, is that surprising?

We thank the reviewer for this careful observation, which prompts a useful clarification of both what HUMAnN 2 can detect and what the absence of viral signal in our outputs means biologically.
HUMAnN 2 is designed for bacterial/archaeal functional profiling, not virome analysis. Although HUMAnN 2 can in principle assign reads to viral references when present in its underlying databases, in practice its reference set (ChocoPhlAn at the nucleotide level; UniRef at the translated level) is overwhelmingly bacterial and archaeal, with very limited representation of the uncultured bacteriophages that comprise the large majority of the human gut virome. HUMAnN 2's intended output is community-level functional profiling (pathways, gene families, EC numbers) for cellular microbes, and we use it for exactly that purpose in this study. Accordingly, no viral functional features were retained in our reported HUMAnN 2 outputs.
This is consistent with expectation, not a surprising negative result. The early-life gut virome is real and substantial, but it is dominated by phages targeting the resident bacteria, and these are systematically under-represented in cellular-genome reference databases of the type HUMAnN 2 relies on. Robust virome characterization in metagenomic data typically requires dedicated approaches — e.g., viral-like-particle (VLP) enrichment, assembly-based viral contig identification, and curated viral reference databases (e.g., IMG/VR, RefSeq Viral, the Gut Phage Database) — none of which were applied here. The absence of substantive viral output from HUMAnN 2 therefore reflects the appropriate scope of the tool rather than a biological observation about the cohort.
Manuscript revisions. We have clarified the Methods to state HUMAnN 2's scope precisely, and acknowledged the absence of dedicated virome analysis as a limitation in the Discussion:

Methods: "Functional profiling was performed using HUMAnN 2 (v0.9.4) to characterize community pathway and gene-family abundance. HUMAnN 2 is designed for functional profiling of bacterial and archaeal communities; viral functional content was not analyzed in this study, as the underlying ChocoPhlAn and UniRef reference databases used by HUMAnN 2 contain limited viral content and are not appropriate for virome characterization."
Discussion: "Our analysis focused on the bacterial and archaeal components of the metagenome and does not include dedicated characterization of the gut virome. Robust virome analysis in this cohort would require approaches such as VLP enrichment, viral contig assembly, and curated viral reference databases, and represents a valuable direction for future work."

15. line 596, why were 10 individuals omitted, and how were the 594 individuals selected?

We thank the reviewer for the opportunity to clarify this point. The reported numbers reflect two sequential filtering steps.

(i) Why were 10 individuals omitted? The apparent discrepancy reflects data availability rather than participant exclusion. The final metagenomic dataset comprised

12,151 samples from 887 participants, but host genotype data were available for only 877 participants. Consequently, analyses incorporating host genetic information were restricted to these 877 individuals.

(ii) How were the 594 individuals selected for trajectory analysis? From the genetically QC-passed pool, we restricted the trajectory analysis to participants with at least 4 longitudinal metagenomic samples within the first 800 days of life. This minimum was pre-specified based on the requirements of the traj clustering framework, which fits per-subject summary measures that become unreliable with fewer than 4 observations. The 594 individuals are all participants in the genetically QC-passed pool meeting this criterion; no additional selection was applied. We also confirmed that the inclusion criterion did not differentially exclude T1D cases: the proportion of T1D cases among included (594) and excluded participants was 48.8% vs. 43.8%, p = 0.17.

16. line 661, please define insufficient DNA quantity and what high quality means. I would encourage the authors to cite software they use. I additionally recommend noting any deviations from default parameters. From what's shown, the version of bowtie2 is quite old. Were the data processed single or paired end? What specific human genome or genomes was used for host filtering?

Thank you for this helpful and detailed suggestion. We agree that additional clarification of sequencing processing and quality control steps improves reproducibility and transparency. Our study builds upon the original TEDDY study metagenomic framework developed by Vatanen et al.[1], and therefore we intentionally followed the same data processing strategy to ensure consistency and comparability across analyses. We have now clarified these points in the revised Methods section as detailed below.

First, regarding DNA quantity and quality, we have now provided explicit thresholds for sample inclusion. Specifically, a minimum DNA input of 20 ng was required for whole-genome shotgun (WGS) library preparation. Sequencing runs were considered passing if >60% of reads passed Illumina's standard quality filter, and the per-sample target yield was 1 Gb (with a minimum acceptable threshold of ~500 Mb when necessary). Samples that failed to meet these criteria—i.e., those that did not yield sufficient reads for downstream analysis or failed sequencing QC—were excluded. We have updated the Methods to clarify these thresholds.

Second, we have now explicitly cited all software tools used in the processing pipeline, including versions and corresponding references: Casava v1.8.2 (Illumina) for FASTQ generation, Trim Galore v0.2.8, cutadapt v1.9dev2 for adapter and quality trimming, PRINSEQ v0.20.5 for dereplication and low-complexity filtering, and Bowtie2 v2.2.3 for read alignment. Unless otherwise specified, default parameters were used as implemented in the original pipeline, and we have now explicitly stated this in the revised manuscript. Deviations from default parameters were as follows:

Trim Galore / cutadapt:

-q 20 (default quality trimming threshold)

--stringency 6 (adapter overlap required before trimming; more conservative than the default of 1)

--length 50 (minimum read length after trimming; default is 20)

--retain_unpaired with --length_1 51 and --length_2 51 (retains unpaired reads ≥51 bp after trimming)

PRINSEQ:

-derep 12 (dereplication enabled; no dereplication by default)

-lc_method dust -lc_threshold 5 (low-complexity filtering via the DUST method; no default complexity filtering)

-trim_ns_left 1 -trim_ns_right 1 (trims single Ns from both ends; no default N trimming)

-no_qual_header

Bowtie2:

--end-to-end --sensitive (default alignment mode)

--no-unal --no-sq --no-head (suppresses unaligned reads and SAM header lines; non-default output behavior to streamline downstream parsing)

Unless otherwise specified, all other parameters were set to their defaults as implemented in the original TEDDY pipeline. We have explicitly noted these settings in the revised Methods.

Third, we clarify that sequencing was performed using a 2 × 100 bp paired-end protocol on the Illumina HiSeq 2000 platform, and all downstream processing was conducted on paired-end reads.

Fourth, regarding host and non-bacterial read filtering, reads were aligned using Bowtie2 to a comprehensive reference database derived from the NCBI whole-genome sequencing (WGS) archive (as of March 2015), which included bacterial, viral, human, and vector genomes. Human reads were filtered using the GRCh38 (hg38) reference genome. Reads whose best alignment was not bacterial were excluded from downstream analyses. For clarity, we note that two distinct human references are used in this study for two different purposes: GRCh38 (hg38) for metagenomic host-read removal (as described here), and GRCh37 (hg19) as the coordinate system for the ImmunoChip-derived host genetic variants (the standard reference for ImmunoChip data and the coordinate system on which the relevant T1D genetic literature is reported). This separation reflects the conventional choice for each data modality and the independence of the metagenomic and genotyping pipelines. We have updated the Methods to state both references explicitly. We have clarified this in the Methods.

Finally, we acknowledge openly that v2.2.3 is not the most recent Bowtie 2 release. As detailed in our response to Comment 1.1, the TEDDY consortium employs a centralized, standardized metagenomic processing pipeline across clinical centers, and 10,913 of the 12,151 metagenomes analyzed here were processed in prior TEDDY publications[1,2]

using this pipeline. We processed the 1,238 newly added samples through the same workflow to preserve cross-study comparability and avoid introducing processing-related batch effects across the larger TEDDY metagenomic resource. Updating to a newer Bowtie 2 version on the new samples alone would have compromised this consistency.

Revised Methods (Line 683): Samples were metagenomically sequenced as one library each multiplexed through Illumina HiSeq 2000 machines using a 2 × 100-bp paired-end read protocol, and all downstream processing was conducted on paired-end reads. A minimum DNA input of 20 ng was required for whole-genome shotgun (WGS) library preparation. Sequencing runs were considered passing if >60% of reads passed Illumina quality filtering, with a target sequencing depth of 1 Gb per sample and a minimum acceptable threshold of ~500 Mb when necessary. Samples failing to meet these criteria or yielding insufficient high-quality reads for downstream analysis were excluded.

Raw FASTQ files generated by Casava v1.8.2 (Illumina) were pre-processed using cutadapt v1.9dev2 for adapter trimming and Trim Galore v0.2.8 (Babraham Bioinformatics) for quality trimming. PRINSEQ v0.20.5 was used for dereplication and low-complexity filtering. Unless otherwise specified, default parameters were used throughout the pipeline. Non-default parameters included the following: for Trim Galore/cutadapt, -q 20, --stringency 6, --length 50, and --retain_unpaired with --length_1 51 and --length_2 51; for PRINSEQ, -derep 12, -lc_method dust, -lc_threshold 5, -trim_ns_left 1, -trim_ns_right 1, and -no_qual_header; and for Bowtie2, --no-unal, --no-sq, and --no-head.

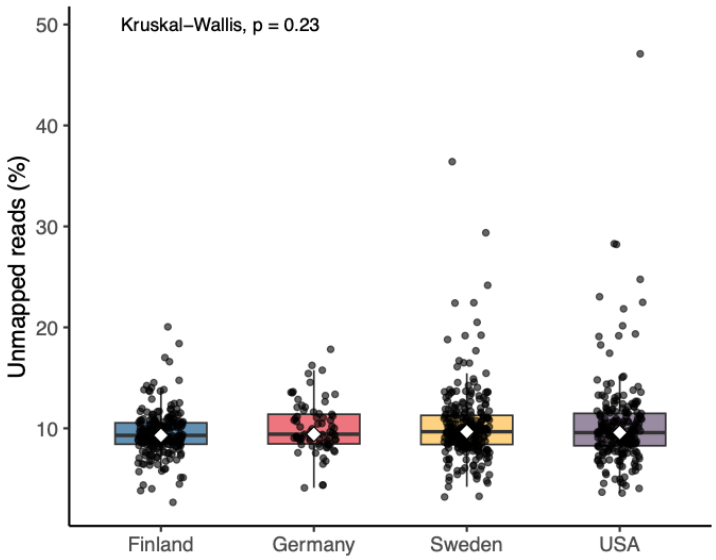
Host and non-bacterial read decontamination was performed using Bowtie2 v2.2.3 with the --end-to-end --sensitive alignment mode. Reads were aligned against a comprehensive reference database derived from the NCBI whole-genome sequencing (WGS) archive (March 2015 release), including bacterial, viral, human, and vector genomes. Human reads were filtered using the GRCh38 (hg38) reference genome, and reads whose best alignment was non-bacterial were excluded from downstream analyses.

17. line 676, what proportion of the reads were unmapped, and was there a country bias?

We thank the reviewer for this helpful question. Across all samples, the proportion of unmapped reads (defined as reads not assigned to any gene family in the HUMAnN2 output) was modest, with a median value of 9.46% at baseline (IQR: 8.39%–11.0%). This is concordant with more recent evaluations of similar values using e.g. MetaPhlAn-based estimates [21].

To assess potential country-level bias, we compared the proportion of unmapped reads across countries using one sample per participant (earliest available time point). As shown in the **Response Figure 3**, we did not observe significant differences across

countries (Kruskal–Wallis test, P = 0.23). Consistently, pairwise comparisons between countries were not significant after multiple testing correction (all FDR-adjusted P > 0.3). These results suggest that unmapped read proportions are comparable across countries and are unlikely to introduce systematic country-level bias in downstream analyses.



Response Figure 3: Distribution of unmapped reads across countries. Boxplots show the percentage of reads not assigned to any gene family (UNMAPPED in HUMAnN2) across baseline samples from different countries. Each point represents one sample. Statistical significance was assessed using the Kruskal–Wallis test.

18. line 671, HUMAnN2 0.9.4 is nearly 10 years old!? At least one of the authors on this paper is involved in that software. I'm genuinely confused here, should the research communtiy _not_ use the newer versions? Certainly there have been some bug fixes in the many versions released since 0.9.4 (https://github.com/biobakery/humann/tags)?

We thank the reviewer for this direct question, which deserves a direct answer. The reviewer is correct on both factual points: HUMAnN has undergone substantial development since v0.9.4, including bug fixes, updated reference databases, and improved translated search; and members of our team are indeed involved in HUMAnN development. We agree that for new analyses beginning from raw reads, current HUMAnN versions are the appropriate choice, and we use them in our other ongoing work.

For this specific study, however, the choice of HUMAnN 2 v0.9.4 was deliberate and grounded in cross-study comparability rather than oversight. The 12,151 metagenomes analyzed here are drawn from a single, integrated TEDDY metagenomic resource: 10,913 samples were profiled in the two foundational TEDDY publications (Vatanen et al., Nature 2018; Stewart et al., Nature 2018)[1,2] using HUMAnN 2 v0.9.4 against the matched ChocoPhlAn and UniRef databases of that release, and the 1,238 samples newly added for this study were processed through the identical workflow to preserve

direct comparability with the existing corpus. Re-profiling with HUMAnN 3 would have introduced version-confounded differences — different UniRef release, different ChocoPhlAn content, different pathway-inference defaults — between the new and previously published profiles, would have created a within-study batch effect between the old and new samples, and would have departed from the canonical functional profiling of this widely used cohort resource on which subsequent TEDDY analyses (including ours) build.

We also note, as a practical matter, that metagenomic processing in TEDDY is centralized and standardized at the consortium level rather than controlled by individual analysis groups (see also our response to Comment 1); the workflow used here is the consortium's established pipeline. We mention this not as a substitute for the scientific rationale above, but to be transparent that re-processing this resource against newer pipelines is a coordinated decision at the consortium scale rather than a choice individual analysis groups can make unilaterally for a single study.

In summary: we agree with the reviewer that current HUMAnN versions are preferable for new data generation, and we make different version choices in our other work that begins from raw reads. For the present analysis, v0.9.4 was the appropriate choice to maintain consistency with the established TEDDY functional-profile corpus and to avoid introducing avoidable batch effects. We have revised the Methods to state this rationale explicitly (Line 706):

"Taxonomic and functional profiling were performed using the standardized bioBakery 2.0 workflow, consistent with previous TEDDY microbiome studies, to ensure comparability across the TEDDY metagenomic resource."

19. line 702, GRCh37 was released in 2009 and is incomplete. Why that genome, and not e.g. T2T-CHM13v2.0 which actually has the Y-chromosome?

We thank the reviewer for raising this point, which we address on two levels: the rationale for using GRCh37 in this specific study, and the relevance of reference completeness to the analyses we actually perform.

Why GRCh37 (hg19) is the appropriate reference for ImmunoChip data. The host genetic data in this study were generated using the Illumina ImmunoChip SNP array[22], which was developed for fine-mapping loci associated with autoimmune diseases. ImmunoChip-derived data are conventionally released, processed, and reported on GRCh37/hg19 coordinates: this is the coordinate system of the array manifest and cluster files, the coordinate system used by the foundational TEDDY genotyping pipeline, and — most importantly — the coordinate system on which the entire body of T1D GWAS literature, including the variants and loci with which our PC3-immune-pathway finding interfaces, is reported. Switching unilaterally to a different reference would, in this study, sever the interpretive connection to that literature while providing

little practical benefit for the analyses we conduct (below). We have made this rationale explicit in the revised Methods.

Reference completeness and the analyses presented here. We agree with the reviewer that T2T-CHM13v2.0 is the most complete current human reference and is the only assembly that fully resolves the Y chromosome. We also note, however, that "incompleteness relative to T2T" is a property GRCh37 shares with GRCh38 — both contain gaps and incomplete sex-chromosome assembly relative to the telomere-to-telomere reference. T2T-CHM13 adoption for ImmunoChip-scale genotype-pipeline workflows and T1D-genetics literature is still nascent at the time of this study, and we are not aware of an established T2T-aligned ImmunoChip processing standard.More importantly for our specific analyses: all host-genetic analyses in this study are restricted to autosomal variants (Comment 4 / Figure 1c). Our genetic principal components, our SNP–microbiome association tests, and our PC × maturation × T1D interaction analyses are all computed exclusively from autosomes. Reference completeness on the Y chromosome — the specific motivation for switching to T2T-CHM13 — therefore does not enter our analyses by construction. The reference choice could in principle matter for autosomal variants if specific autosomal regions were grossly different between GRCh37 and T2T, but the autosomes are largely stable across these references, and the ImmunoChip's tag-SNP design is not concentrated in the regions where T2T provides substantial additional content (centromeres, segmental duplications, acrocentric arms).

Constraints and consistency. As also noted in our responses to Comments 1, 16, and 18, TEDDY genotype data are processed under a centralized, standardized consortium pipeline; the reference assembly is a coordinated decision at consortium scale rather than one individual analysis groups can change for a single study. We mention this as a transparency note rather than as the primary justification — the scientific rationale (ImmunoChip convention; GWAS-literature comparability; autosomal-only analysis scope) is what supports the choice on its merits.

We have revised the Methods to state explicitly that ImmunoChip variants were processed on GRCh37 (hg19), that this reflects the standard convention for ImmunoChip data and the coordinate system of the relevant T1D-genetics literature, and that all downstream analyses are restricted to autosomal variants.

20. line 720, the data were not rarefied? How were the 92 species selected? No coverage filtering was performed? Reference-based approaches are not without value, but what was the reason the authors did not pursue metagenomic assembly? It is extremely unlikely the genomes in their reference database are _the exact_ genomes in the subject.

Thank you for these important comments. Rarefaction was not performed, consistent with current best practice for shotgun metagenomic data. Rarefaction was originally

developed for 16S amplicon data with highly variable per-sample read counts; for shotgun-derived MetaPhlAn 2 species profiles, normalization is intrinsic to the relative-abundance framework (proportions are independent of total sequencing depth), and rarefying discards real signal without statistical benefit. McMurdie & Holmes[23] provide the formal argument against rarefaction for differential-abundance analysis, and our downstream inference relies on compositionally appropriate methods (MaAsLin 2[12]; see Comment 12) rather than depth-based subsampling.

Regarding our selection of the 92 species and coverage filtering, our filtering protocol applies coverage-aware control at two distinct stages, rather than relying on a single filter:

Per-species detection (MetaPhlAn 2). Species presence is determined by MetaPhlAn 2's clade-specific marker-gene framework, in which a species call requires concordant signal across multiple distinct markers from a panel of ~1 million clade-specific genes selected to be conserved within a species clade and absent from sister clades. Low-support hits resting on one or two markers are filtered by default. This is a coverage- and breadth-aware mechanism operating per-species, more discriminating than a depth-based global filter would be, and it is the standard mechanism for guarding against spurious low-coverage species calls in MetaPhlAn 2[13]. Per-species marker-coverage statistics for the species in Extended Figure 4 are reported in Supplementary Table Z (added in response to Comment 13).

Cohort-level prevalence/abundance. Species passing per-sample detection were retained for downstream differential-abundance testing if they had a relative abundance ≥1% in at least one sample from ≥10% of subjects — yielding the 92 species tested. This is the field-standard prevalence/abundance threshold used in the centralized TEDDY metagenomic pipeline and in the foundational TEDDY publications[1,2], and is broadly consistent with thresholds used in comparable large-scale shotgun microbiome differential-abundance analyses. We have made both filtering stages explicit in the revised Methods.

The combination of per-species multi-marker detection and cohort-level prevalence/abundance filtering is, in our view, a more rigorous control against low-coverage and spurious-detection artifacts than a single global depth-based filter would provide.

Regarding reference-based profiling versus assembly, the reviewer raises a fair and important point: reference-based methods such as MetaPhlAn 2 and HUMAnN 2 do not capture strain-level variation or novel genomic content present in subjects but absent from the reference database. We agree that assembly-based approaches — including metagenome-assembled genomes (MAGs), strain-aware profilers (e.g., StrainPhlAn, inStrain), and gene-content profilers — provide complementary information that reference-based methods cannot. We chose reference-based profiling for this study for three specific reasons:

First, our analytic questions are at the species and pathway level — early-life maturation trajectories, cross-trajectory differential abundance, gene-by-microbiome interactions on T1D risk — rather than at the strain or accessory-genome level. Reference-based species/pathway profiling is the appropriate granularity for these questions.

Second, MetaPhlAn 2's species identification is not vulnerable to the exact-strain-match concern the reviewer raises: it identifies species via clade-specific marker genes selected to be conserved across members of a species clade and absent from sister clades, with a species call supported by concordant signal across multiple markers (Comment 1.13 details the per-species marker-coverage evidence). The "are the database genomes the exact ones in the subject?" concern applies primarily to strain-level inference, which we do not perform here.

Third, the 12,151 metagenomes in this study constitute an integrated TEDDY resource, with 10,913 samples profiled in two prior Nature publications[1,2] using MetaPhlAn 2 / HUMAnN 2. We processed the 1,238 newly added samples through the identical workflow to maintain comparability — re-processing the entire corpus via assembly would have departed from the canonical TEDDY processing and introduced batch effects between the new and previously published profiles.

We note in the revised Discussion that strain-resolved and assembly-based analyses of this cohort represent valuable directions for future work, expected to reveal aspects of microbial functional variation — particularly accessory-genome and strain-level dynamics — that the reference-based framework used here does not address.

21.Remarks on code availability:
The code does not reflect the actual methods. It is quite incomplete and therefore not useful for reproducibility. It is not a resource for the community as it's not generalized. As a precise example, the methods text states "humann" was used, but a search shows that term only appears in reference to a file (https://github.com/biobakery/teddy_genetics/blob/5e5d0fa33c77c9d80a414515307ab2d3f239f6a7/Figure_scripts/Fig3.R#L156), not the actual use of the software, and unexpectedly the file referenced isn't even in the repository.

We thank the reviewer for this important comment. The preprocessing pipeline used in this study was assembled using the published workflow described in Stewart et al.[2], rather than custom software developed specifically for this study. We agree, however, that the previous version of the repository did not adequately document the specific implementation used in our analyses. We have therefore updated the repository to include the reference databases and software versions, Bowtie2 parameters, and the MetaPhlAn 2 and HUMAnN 2 commands used in our analyses, thereby making the workflow substantially more transparent and reproducible.

Reviewer #3 (Remarks to the Author):

The authors have undertaken a significant revision of the MS in response to three in-depth reviews including from this Reviewer. My assessment is that they have addressed each of the concerns regarding data clarity, analytical approach and interpretation. I have no further suggestions for revision and defer to the Editors regarding the added novelty of this work beyond prior metagenomic analyses in the TEDDY study

> We thank the reviewer for the careful evaluation of our revised manuscript and for recognizing our efforts to address the reviewers' comments. We appreciate the reviewer's positive assessment and thoughtful feedback throughout the review process.

Reviewer #5 (Remarks to the Author):

1. I appreciate the strong effort made by the authors to address all review comments.

My main concern lies in the choice of 3 sub groups, as this is a central finding of the paper. The authors provide some explanation for their metrics, but I am left feeling that these choices are arbitrary. My understanding of the field is that while enterotypes are frequently invoked, equally invoked is the notion that gut microbiome profiles belong to a continuum, and that discrete categorization is oversimplifying the microbiome (e.g. see Jeffrey et al. 2012 Nature Reviews Microbiology).

The reviewers kindly provided justification for their three statistics, but, when I look at the plots, I feel I could justify other ks using the same logic. E.g.

“The Elbow Method (Panel A) demonstrates a distinct inflection point at k=3, where the reduction in within-cluster sum of squares begins to plateau, indicating diminishing returns for additional clusters.”

One could argue the inflection point is at 2.

"The Calinski-Harabasz Index (Panel C) is maximized at k=2 but remains high at k=3, before declining sharply at higher k values. This suggests that while a 2-cluster solution provides strong separation, a 3-cluster solution maintains high structural integrity while capturing necessary biological nuance (specifically, the distinction between "Late Matured" and "Early Plateaued" trajectories). "

It's actually higher at 4 and 5.

"The Gap Statistic (Panel B) trends upward toward higher k values. However, choosing a higher k (e.g., k=6) results in highly granular clusters with reduced sample sizes."

OK – what about 4, 5?

I’m not advocating for any one k – in fact, I think moving away from any k is a more robust option.

At minimum, I think the authors could temper their claims about there being specifically three groups – maybe there are more? Or less? Or a continuum?

> We thank the reviewer for this thoughtful comment and for the Jeffery et al. citation[3], which we have engaged with directly in the revised manuscript. The reviewer's core point — that microbiome composition is most accurately viewed as a continuum, and that any small-k discretization risks oversimplification — is one we fully agree with, and we have tempered the manuscript accordingly. Note that we explicitly do not describe the space of occupied microbiome configurations discretely, only the collection of trajectories taken through it by individuals - and, as detailed below, even this is an analytic convenience (albeit an accurate one). This concern overlaps substantially with a comment raised by Reviewer 1 (Comment 10), and we have addressed both reviewers' concerns through a common set of revisions. We summarize the changes here.

Tempering the discrete-class framing. We agree that the original text implied stronger discreteness than is supported by either the data or the broader literature. Following the reviewers' comments, we have:

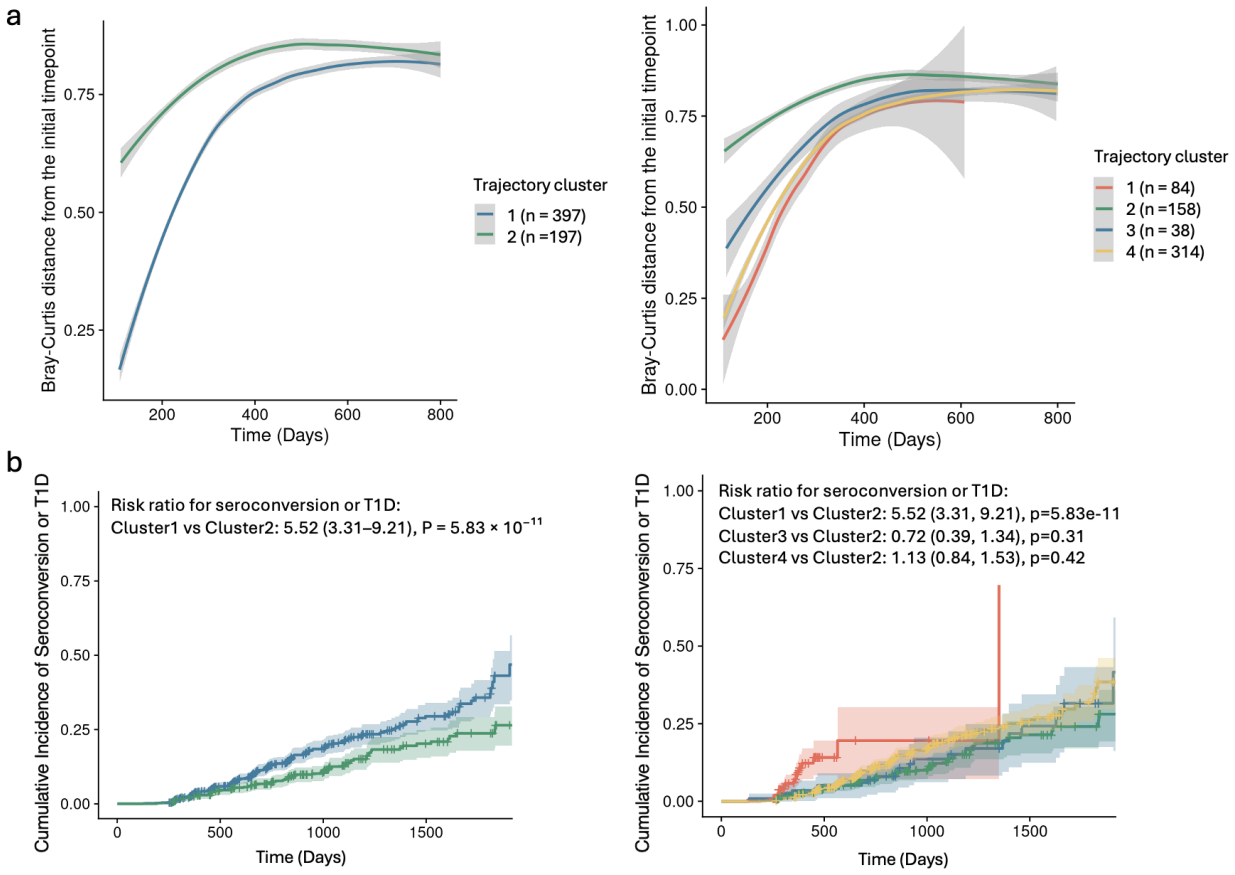
1. Removed the phrase "distinct inflection point" from the Supplementary Figure 1 legend and revised the description of the elbow as gradual rather than sharp, in response to the reviewer's observation that the elbow could reasonably be argued to occur at k = 2.
2. Revised the Methods to acknowledge explicitly that the clustering metrics do not uniquely support k = 3. As the reviewer notes, the Calinski–Harabasz index is higher at k = 4 and k = 5 than at k = 3, and the Gap statistic does not identify a clear optimum in this range. We therefore no longer claim that the statistical metrics uniquely identify three clusters.
3. Reframed the three classes throughout the manuscript as a parsimonious discretization of an underlying continuum rather than evidence for intrinsically discrete microbiome community types, explicitly invoking Jeffery et al.[3] and the broader gradient literature to make this position clear.

Why three classes are retained as the primary presentation. Despite the above, we elected to keep the three-class framing as the primary structure of the paper, for two reasons we believe are defensible:

First, the three classes correspond to long-recognized phenotypes in pediatric medicine — normal, catch-up, and delayed/faltering developmental trajectories — quantified for decades using standardized z-scores relative to WHO growth references, and explicitly transferred to the gut microbiome by Subramanian et al. [24]. via the microbiota-for-age z-score (MAZ) framework. The three-class framing thus has external biological grounding independent of the statistical clustering metrics.

Second, we tested whether the association between microbiome maturation and T1D risk was robust to the specific choice of k. Adjacent clustering solutions (k = 2 and k = 4) yielded qualitatively similar conclusions regarding the relationship between maturation trajectory and T1D risk (**Supplementary Figure 5**). We therefore do not view the choice of three as a load-bearing claim of the paper; the load-bearing claim is that impaired/non-progressive maturation, however it is operationalized, associates with T1D risk.

In summary, we have tempered the language throughout the manuscript so that we do not present the trajectories as discrete biological states, but rather as an interpretable discretization of an underlying continuum. Most importantly, the association between microbiome maturation and T1D risk was observed both when maturation was modeled continuously and when alternative clustering solutions were considered. We therefore conclude that the principal biological finding does not depend on the assumption of exactly three groups, but instead reflects a broader relationship between delayed microbiome maturation and T1D risk.



Supplementary Figure 5. Sensitivity analyses of microbiome maturation trajectories using alternative cluster numbers (a) Sensitivity analyses using alternative numbers of trajectory clusters. Left, trajectory clustering with k = 2 identified two broad microbiome maturation patterns. Right, trajectory clustering with k = 4 identified four maturation patterns. Clusters were derived using a trajectory clustering algorithm based on Bray–Curtis distances from each participant's baseline gut microbiome composition across visits during the first 800 days of life. Trajectories were smoothed and visualized using LOESS curves; shaded bands indicate 95% confidence intervals. (b) Association between alternative maturation patterns and risk of islet autoantibody seroconversion or type 1 diabetes (T1D). Left, cumulative incidence curves and hazard ratio estimates for the two-cluster solution. Right, cumulative incidence curves and hazard ratio estimates for the four-cluster solution. Hazard ratios were estimated using multivariable Cox proportional hazards models adjusting for clinical center, sex, antibiotic use, probiotic use, breastfeeding cessation, age at solid food introduction, and the top five host genetic principal components. Shaded bands indicate 95% confidence intervals. Consistent results across clustering resolutions indicate that the association between impaired/non-progressive microbiome maturation and T1D risk is robust to the choice of k.

2. Remarks on code availability:

I cannot figure out how supplemental figures are generated.

Thank you for pointing this out. We have now uploaded the complete code used to generate the supplementary figures to our GitHub repository, and the repository has been updated accordingly.

Reference

1. Vatanen, T.*, et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589-594 (2018).
2. Stewart, C.J.*, et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583-588 (2018).
3. Jeffery, I.B., Claesson, M.J., O'Toole, P.W. & Shanahan, F. Categorization of the gut microbiota: enterotypes or gradients? *Nat Rev Microbiol* **10**, 591-592 (2012).
4. Tomofuji, Y.*, et al.* Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. *Nat Microbiol* **8**, 1079-1094 (2023).
5. Group, T.S. The Environmental Determinants of Diabetes in the Young (TEDDY) Study. *Ann N Y Acad Sci* **1150**, 1-13 (2008).
6. Parkes, M., Cortes, A., van Heel, D.A. & Brown, M.A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* **14**, 661-673 (2013).
7. Rothschild, D.*, et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210-215 (2018).
8. Falony, G.*, et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560-564 (2016).
9. Bäckhed, F.*, et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690-703 (2015).
10. Bokulich, N.A.*, et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci Transl Med* **8**, 343ra382 (2016).
11. Yatsunenko, T.*, et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222-227 (2012).
12. Mallick, H.*, et al.* Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput Biol* **17**, e1009442 (2021).
13. Truong, D.T.*, et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**, 902-903 (2015).
14. Goodrich, J.K.*, et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19**, 731-743 (2016).
15. Blekhman, R.*, et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* **16**, 191 (2015).
16. Qin, Y.*, et al.* Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *Nat Genet* **54**, 134-142 (2022).
17. Wang, Y.*, et al.* The genetically programmed down-regulation of lactase in children. *Gastroenterology* **114**, 1230-1236 (1998).
18. Heine, R.G.*, et al.* Lactose intolerance and gastrointestinal cow's milk allergy in infants and children - common misconceptions revisited. *World Allergy Organ J* **10**, 41 (2017).
19. Swallow, D.M. Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet* **37**, 197-219 (2003).
20. Misselwitz, B., Butter, M., Verbeke, K. & Fox, M.R. Update on lactose malabsorption and intolerance: pathogenesis, diagnosis and clinical management. *Gut* **68**, 2080-2091 (2019).
21. Blanco-Míguez, A.*, et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, 1633-1644 (2023).
22. Hadley, D.*, et al.* HLA-DPB1*04:01 Protects Genetically Susceptible Children from Celiac Disease Autoimmunity in the TEDDY Study. *Am J Gastroenterol* **110**, 915-920 (2015).

23. McMurdie, P.J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* **10**, e1003531 (2014).
24. Subramanian, S.*, et al.* Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417-421 (2014).

Responses to Reviewers' Comments

Reviewer #1:

Remarks to the Author:

The authors are requesting a re-review of results in which there are compelling reasons to not have confidence in the underlying data processing from which the results are derived from. While I appreciate the detailed responses from the authors, and the selection of additional analyses the performed, I am disappointed to see the authors not adopt modern evidenced host filtration procedures. They acknowledge the deficiencies of their current analysis, yet the authors choose to write around the concerns rather than redoing the work. Writing around the problem does not directly address the evidence that the methods used have serious ethical issues and create technical bias. It's unfortunate the TEDDY consortium is using antiquated methods rather than reprocessing the entirety of their data, but that appears to be a policy decision without clear scientific evidence. The authors further seem to suggest that updating software doesn't matter which is a remarkable position to take. Notably, multiple pieces of the software they use are 10 years old with many important bug fixes having been addressed in the intervening years. Confusingly, the authors even acknowledge this in their response: "HUMAnN has undergone substantial development since v0.9.4, including bug fixes..."

> We are grateful for the reviewer's continued attention to our data processing. We would like to begin by clarifying one point about the procedure we used, as it may be helpful context for the concerns that follow. Human reads were filtered from all metagenomes prior to any downstream analysis. Quality-controlled reads were aligned with Bowtie 2 (v2.2.3) against a reference database containing bacterial, viral, human (GRCh38/hg38) and vector sequences, and any read whose highest-scoring alignment was non-bacterial, including reads aligning to the human genome, was removed before taxonomic or functional profiling. All 12,151 metagenomes, including the 1,238 newly generated for this study, were processed through this identical workflow. Following the previous round of review, we expanded the Methods to document the reference assemblies, software versions and parameters at each step, and updated the public repository to include the pipeline from raw reads to feature tables.
>
> We understand the reviewer's remaining concerns to relate to whether the reference and tooling used for this filtering are current, and we address these in turn.
>
> On re-identification risk, we share the reviewer's view that residual host reads in released data are a genuine concern. Two features of this study help to bound that risk here. All analyses reported in the manuscript are performed on MetaPhlAn 2 and HUMAnN 2 feature tables, which contain no human sequence. The underlying sequencing reads are deposited in dbGaP under controlled access, which is the mitigation recommended in the study the reviewer cites; access requires an approved data use request under the dbGaP authorization process rather than being open to the public.

On technical bias from incomplete sex-chromosome representation, GRCh38 includes chromosome Y, though it is not the telomere-to-telomere assembly now available, and we would not claim that residual Y-derived reads are removed as completely as a current reference would achieve. Three features of the study design may help in judging how far this could affect our conclusions. All host genetic analyses use autosomal ImmunoChip variants only, so the host-genetics side of the study is not subject to sex-chromosome coverage. Every sample was processed through a single identical pipeline, so any residual host content is non-differential with respect to type 1 diabetes case status, which was ascertained independently of sequencing; misclassification of this kind would be expected to attenuate rather than generate the associations we report. Sex was included as a covariate in all adjusted models, with sex-stratified analyses of the primary association yielding results consistent with the main analyses (Supplementary Table 7), which we would not expect if a sex-linked technical artifact were driving the findings.

We agree with the reviewer that MetaPhlAn and HUMAnN have undergone substantial development since v2.6.0 and v0.9.4, and we did not intend to suggest that software currency is unimportant. The TEDDY metagenomic resource is processed centrally and shared across several concurrent analyses, and re-profiling it is a consortium-level decision that sits outside the scope of this manuscript. We have therefore added an explicit statement of this limitation and its likely direction of effect to the Discussion, so that readers encounter the constraint in the paper itself:

"Additionally, taxonomic and functional profiling was performed with bioBakery 2.0 (MetaPhlAn 2 and HUMAnN 2), which has since been updated. The reference databases available at that time were smaller than the most updated ones, so poorly characterized taxa may be underrepresented and our functional annotations less complete. We retained these versions for comparability with prior TEDDY analyses, and because all samples were processed through a single identical pipeline, any resulting misclassification is non-differential with respect to T1D status and would tend to attenuate rather than create associations. All software versions and analytical parameters used in this study have been documented to facilitate reproducibility. "

We recognize that this falls short of the reprocessing the reviewer would prefer, and we have tried to be as transparent as possible in the manuscript about what that means for the interpretation of our findings.

Remarks on code availability:

I see they updated the GitHub repository. The files inside refer to data files that I did not obviously find in the repository. I did not look closely at the actual code given the ethical and technical concerns I have with the data processing.

Thank you for your comments. The source data files underlying the figures have been provided with the manuscript submission. The TEDDY microbiome whole-genome sequencing and SNP data supporting the findings of this study are

available in the NCBI database of Genotypes and Phenotypes (dbGaP) under the primary accession code phs001442.v4.p3, in accordance with the dbGaP controlled-access authorization process. The analytical datasets generated in this study are available upon request through the NIDDK Central Repository (NIDDK-CR) website, Resources for Research (R4R): https://repository.niddk.nih.gov/.