nature portfolio

Peer Review File

Gut microbiota wellbeing index predicts overall health in a cohort of 1000 infants

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you for allowing me to review this interesting manuscript.

Here, Hickman and colleagues present the results from their HELMi birth cohort of the relationship between early life (2yrs) gut microbiome development and various environmental and social exposures together with later health outcomes. They find that the microbiome trajectories are to some extend predictable. Overall, I want to congratulate the authors on their fantastic cohort. The dataset is impressive and the analyses interesting, however it also comes with severe limitations in the analytical framework affecting the generalizability of conclusions drawn. Main limitation is that everything is built on a single PCA approach of the entire microbiome at genus level. Every result downstream is built on this representation of the data, with extrapolation on extrapolation from this, from indicator taxa to clusters and trajectories. This is not necessarily problematic, but it will only reveal the part of the microbiome captured in the first two dimensions of this PCA. Correlations based on clusters are always a bit difficult to interpret as various factors are captured and lumped together. The authors end up trying to make a microbiome wellbeing index, which is suggestive of a healthy trajectory. while interesting, it is however being selected from the most healthy microbiome trajectory (based on previous associations) with the healthiest children, therefore it seems a bit circular that indicator genera for this specific pattern is related with overall health, and beneficial factors such as siblings and vaginal birth.

More specific points:

How did the QC look for the samples? were all samples randomized before wet lab procedures? Check technical variables against time. How much did the batch effects affect the compositions and results (e.g. MiSeq vs HiSeq)? Please show these details. Describe sequencing depth, cut offs etc.

The PCA approach of log-transformed relative abundances of genera is generally looking good in e.g. Figure 1, but the authors rely on this also for all downstream analyses. I would

like to see some additional analyses of more microbiome specific tools for beta-diversity (weighted and unweighted) as well (Bray-curtis, Unifrac etc.).

For illustrations, I understand that it is interesting to look at PC1 and PC2 in relation to various exposures, however it would make sense to do all PCA statistics for the manuscript using all dimensions of the data with e.g. adonis2.

I might have missed this, but how were repeated measures (the same child has multiple samples represented) handled in the beta-diversity analyses on exposures?

It is a bit problematic that the authors use this one PCA for so many of the downstream analyses as it will enhance and emphasize certain features of the composition and hide others. Similar, the use of envfit will again only rely on these PC1 and PC2 features across all samples. I would suggest doing individual exposure analyses for beta-diversity and differential analyses at each time point, not restraining analyses (distances) to that of the main PCA.

Most analyses are done at genus-level, but given the changes over time in many species (e.g. bifido going from HMO utilizers to more mature taxa) it could have been interesting to recapitulate at least some of the findings on ASV level. Also, many taxa have some very beneficial species mixed with what could be considered bad guys belonging to the same genus. With V3-V4 sequenced quite a number of taxa can be suggested at species level.

In Figure 2a-c the y-axis scale makes it very difficult to see details of each factor. Mainly because previous timepoint is so dominating.

For all figures the authors show age in weeks, while they in text refer to e.g. 6 months or 24 months sample. I would suggest aligning the labeling in the figures for easier communication.

Extended data figure 1 is difficult to read/comprehend.

Line 130 seems to refer to something else than Extended data figure 2?

Line 137 – I can't find the described key taxa in Extended data figure 2. Instead of looking at indicator organisms for the PCs, I would suggest plotting the relative abundances of genera at each time point. This will show a more clean representation.

I like the details around the splitting of vaginal birth into vaginal with and without ABX and how that affects the different taxa over time – could CS deliveries be split into emergency and elective as well?

Also, the notion that siblings drive early Bifidobacterium and Lactobacillus abundance only in vaginal delivered is interesting, and I like the interpretation that it could be due to different maternal compositions.

For the community types from clustering, it seems a bit odd to make the clustering at genus level and then afterwards (lines 184+) describe taxa on family level.

Please describe deeper the number of features used for training the 5-feature model for describing the community types. How many features were tested, how did you select these 5?

It is unclear how the community types vs disease were tested, how were repeated measures handled and how was multiple testing accounted for?

When evaluating the trajectories over time in relation to health outcomes, it is important only to evaluate outcomes happening after age 2 years. Else, you will have no idea about the direction of associations. One would think that fever episodes and infections could easily affect the gut microbiome development perhaps through antibiotics prescriptions or inflammation (then they become shaping factors). It would likewise be interesting to adjust

your clinical results for the main shaping factors such as siblings and delivery mode. Was multiple testing accounted for?

Were additional exposures included in the trajectory evaluation such as: parental allergies and education level, maternal BMI, parental smoking, gestational diabetes, pregnancy weight gain?

Please indicate how many children were in each category in Fig 5g+h.

Figure 3c+d have switched legends 3d should be richness.

Definition of pathobionts should be referenced.

In general the Supplementary tables appear extremely rough and difficult to read....

Reviewer #2 (Remarks to the Author):

This study examines the role of the early-life human gut microbiota in regulating various aspects of host physiological development with implications for long-term health. The authors address the lack of longitudinally monitored cohorts by describing natural gut microbiota development in infants during their first years of life. They introduce a gut microbiota wellbeing index based on microbiota development and health data from nearly 1000 infants, demonstrating its predictability in determining developmental trajectories and later health outcomes. Key findings include the identification of distinct microbiota development trajectories, the influence of early exposures on these trajectories, and the predictive value of certain genera like Bifidobacterium and Bacteroides in health outcomes. Additionally, the study highlights the potential for using early gut microbiota

information to enhance individual health risk predictions, emphasizing the malleability and predictability of gut microbiota succession during infancy.

-What are the noteworthy results?

The authors identified five distinct microbiota development trajectories and the influence of infant exposures on these trajectories

They pointed out the predictive value of certain genera like Bifidobacterium and Bacteroides in health outcomes

They highlighted the predictability of gut microbiota succession during infancy, as a way to enhance individual health risk predictions

-Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.

Yes, there is originality in this study, and as the authors claim this is the first study proposing a gut microbiota index associated with wellbeing and health in children. With that said, there have been other important infant microbiome cohort studies previously published, although analyzed through different statistical methods. However, it is critical that authors state how their findings may be only applicable to a specific population: Scandinavian children. While general patterns of infant microbiome development are similar in terms of beta-diversity (weighed by similar abundances of predominant species like Bifidobacterium and Bacteroides), other ecological metrics differ between industrialized and non-industrialized pediatric populations, such as alpha-diversity and the abundance of many other taxa.

General Comments:

-The manuscript reads nicely.

-The study is well-designed. There is a sufficient sample size, and both infant exposures and health outcomes have been carefully selected.

-In the Results section, I recommend including the results (like numbers and p-values) at the end of the sentences to avoid breaking the sentences. Some of these results could also

be better organized and more focused if they were presented in tables (not only in the supp. but also in the main manuscript).

It is critical that authors state how their findings may be only applicable to a specific population: Scandinavian children. While general patterns of infant microbiome development are similar in terms of beta-diversity (weighed by similar abundances of predominant species like Bifidobacterium and Bacteroides), other ecological metrics differ between industrialized and non-industrialized pediatric populations, such as alpha-diversity and the abundance of many other taxa.

-Please make sure all bacterial names, regardless of taxonomic level, are italicized, as suggested by the American Society of Microbiology (see here: https://journals.asm.org/writing-your-paper).

-Authors should include biological sex of infants in their models some of the microbial variance.

-Extended Data Figure 2 should include the impact of exposures on age-specific ordination, yet it shows something else. Please include the correct figure referred to in lines 126-140.

-Authors should be specific when describing ecological metrics. For instance, when referring to microbiota composition (line 83) they should use the specific metric they are referring, which is beta-diversity measured by xx distance.

-While framing the analysis on bet-adiversity findings makes sense, alpha-diversity is a very important feature to study ecosystems. It is therefore important that the authors include: (i) the differences in alpha-diversity (Chao1 and Shannon or Simpson) throughout age (ii) how it is impacted by some of the same exposures analyzed for b-diversity.

-In the clustering analysis, it is misleading to calculate the percentage of pathobionts based on genus-level information. Microbial virulence does not occur at the genus level so this needs reassessment. Moreover, this potentially misleading result is not described in the results section. If authors decide to include the level of pathogens in each cluster, they should calculate species of known infectious potential.

-It is also unclear which richness index was used in the cluster analysis (Fig 3d). Regardless of which was used, it is unclear how clusters 3 and 4 have such low a-diversity if these include most of the older infant samples. It will be necessary to explore several a-diversity indices.

-The recursive partitioning analysis is most interesting. It would also be important to understand what % of infants of a specific age fit within each of the clusters, in order to

understand how frequent is for microbiomes of a specific age to fit outside of the typical or ‘normal’ cluster. Also, while the decision tree in Figure 4a is nice, there are other visualization techniques to better display the proportion of the exposures in each of the microbial clusters. Here is an example: https://paulvanderlaken.com/2020/03/31/visualizing-decision-tree-partition-and-decision-boundaries/

-The developmental trajectories are quite revealing and a welcoming addition to this field.

-Is it known what types of probiotics were mostly taken by the infants? If not the brand, the probiotic organism(s)?

Of the 388 infants in T1, almost half had a healthy childhood based on the clinical outcomes used, but the rest did not. Why is this, what factors may be associated with this and how does this impact the accuracy of the wellness index in predicting eubiosis and a healthy infancy?

-It is important to show a plot depicting the relative abundances of the indicator microbes of the microbiota wellness index as this can only be inferred from the extended data figures. This may be one of the most important aspects of this paper so this should be part of the final main figure.

-Please revise the Results section and make sure the figures are clear and accurate Extended Data Figure 3 is completely illegible. Authors may need a figure per timepoint for this.

Extended Data Figures 4 is also impossible to read. Also, the title of this figure should more than “Associations”

- This study comes up with interesting and novel findings. However, it is important to acknowledge that the study does not consider many different factors, such as diet, genetics, lifestyle, ethnicity, and geographical location, that have been shown to notably impact gut microbiome. I strongly recommend pointing this out, particularly in the Discussion and Conclusion sections.

- It is important to mention the importance of other gut microbial communities, including microbial eukaryotes, such as fungi, phages, viruses and protists that have been shown to interact with gut bacteria and significantly influence the host's developmental pathways. It is necessary to emphasize that when you speak of gut microbiota in this study, you are talking about only bacteria and the role of other microbial communities should be studied

in the future to improve the prediction of health-associated infant gut microbiota development patterns.

I strongly recommend noting that, considering the last two matters, these results cannot be generalized to all infants.

Comments on Main (Introduction):

I suggest referring to the cohort in the last paragraph of the introduction.

Comments on Results:

When using GLM to test the effect of multiple predictors or conducting multiple comparisons, it is recommended to adjust the p-values to control the likelihood of false positives. Alternatively, you can run a differential expression analysis using R packages like DESeq2 or EdgeR, which perform multiple testing corrections for p-values by default.

Specify the AIC values in the manuscript (page 5) and on the related plots.

Include the p-values of the X2 test (page 6).

Whenever a p-value is presented, it is a good idea to mention the statistical test as well. This information can be brought to the end of the sentence to prevent breaking the sentences, which reduces the fluency of reading.

Additional comments on Figures & Tables:

Figure 1-a: In the caption, please indicate what each point represents.

Figure 1-a: I recommend adding the results of the statistical test to the plot.

Figures 1-b, 1-c, 2-d: Indicate which comparisons are represented by the stars and their colors.

Figures 2a-c: Side-by-side stacking seems to be a more appropriate way to display multivariate regression results (See figure 2 in the following article as an example: https://www.nature.com/articles/s41586-018-0617-x#Fig2, which is also very similar to the extended data figure 2 in your own manuscript).

Figure 2d: It is not clear enough if the family names shown in the middle plot titles refer to the entire row. I suggest moving these titles to the right or left end of each row.

Figure 3a: Fix the overlapped texts (numbers) on both axes.

Figures 3a legend, 3b-d x-axes: Community types are presented as 1-4 rather than C1-C4, as mentioned in the manuscript (page 4) and caption.

Figure 3b: Why aren't the stacks reaching relative abundance 1.00? If it's because NAs were removed, please specify in the caption.

Figures 3c-d: The y-axis titles seem to be reversed! In accordance with the caption and the manuscript (lines 189-191), the y-axis in these two plots should be labeled as "Microbial Richness" in Figure 3c and "Relative abundance" of pathobionts in Figure 3d.

Figures 5g-h: Enlarge the legend and Include significance levels (stars) in the caption.

Figure 6: Layout the figure.

Extended Data Fig. 5: Consider adding community-type labels.

Extended Data Figs 6 and 7: Consider adding trajectory labels.

Extended Data Fig. 11: Indicate which time point each plot represents.

-Extended Data 10. The colored bars on top of the heatmap are unclear because a legend is missing.

Extended Data Fig 11 is also impossible to read. The vertical bars on the left of each heatmap are also unclear as a legend is missing.

Comments on Discussion:

Line 370: While technical artifacts may be one reason for these results, I suggest you point out some of the other potential factors as well. It is true that in the study you've cited (Xiao et al, 2021), only 35% of the infants were not breastfed (formula-fed babies have significantly higher proportions of Bacteroides, Fallani et al, 2010), and only 8% were delivered through C-section (as you mention later in your discussion, C-section is associated with a high abundance of Bacteroides). However, as I quickly glanced through their data, many other factors that are known to be potential drivers of infants' gut microbial composition such as maternal dietary intake, breast milk composition, weaning time, and type of diet were not measured. You may find other examples in the literature, as well. As we learn more about these factors and many others that could affect the microbial composition of the infants' gut, we may be able to explain these differences in findings.

Line 423: Although these results are intriguing, it is important to acknowledge that most (if not all) of these studies have been conducted in industrialized countries, so the results might differ for less industrialized and rural populations.

Comments on Materials and Methods:

Well explained

Data and code are available

Statistical analysis: Describe how the stool color was detected and analyzed, based on observation or any other method.

Statistical analysis: Specify that the partitioning tree was used to determine the community types.

Statistical analysis: Based on what resource did you define the maturity index? If it was an article, could you cite it? In that case, are there any other factors or ways to measure maturity that you think should be considered?

Other Comments:

Italicize: Bacteroidia (line 172), Bacteroidia and Actinobacteria (line 183), Actinobacteria, Bacteroidia, Enterobacteria, Negativicutes, Bacilli and Clostridia (lines 283-284), staphylococci and streptococci (line 401), and any other bacterial names that are not italicized throughout the manuscript, regardless of the taxonomic level.

Lines 177-178, 196: These lines belong to the Statistical analysis in the Methods section.

Lines 251-256: The Discussion section would be a better place for this level of interpretation.

Lines 261-264: Rephrase these lines to clearly state that T1 was negatively associated with all the following variables after the comma.

Please revise the manuscript for the typos, as I have noticed a few of them; for example, postiive (line 455) associtio (456).

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks on code availability):

I have reviewed the code to make sure it makes sense, which it does. However, I have not run the code to know if it is reproducible or not.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you for allowing me to review this interesting manuscript.
Here, Hickman and colleagues present the results from their HELMi birth cohort of the relationship between early life (2yrs) gut microbiome development and various environmental and social exposures together with later health outcomes. They find that the microbiome trajectories are to some extend predictable. Overall, I want to congratulate the authors on their fantastic cohort. The dataset is impressive and the analyses interesting, however it also comes with severe limitations in the analytical framework affecting the generalizability of conclusions drawn. Main limitation is that everything is built on a single PCA approach of the entire microbiome at genus level. Every result downstream is built on this representation of the data, with extrapolation on extrapolation from this, from indicator taxa to clusters and trajectories. This is not necessarily problematic, but it will only reveal the part of the microbiome captured in the first two dimensions of this PCA. Correlations based on clusters are always a bit difficult to interpret as various factors are captured and lumped together. The authors end up trying to make a microbiome wellbeing index, which is suggestive of a healthy trajectory. while interesting, it is however being selected from the most healthy microbiome trajectory (based on previous associations) with the healthiest children, therefore it seems a bit circular that indicator genera for this specific pattern is related with overall health, and beneficial factors such as siblings and vaginal birth.

We thank the reviewer for the congratulations and this positive appraisal. We indeed propose a healthy microbiome trajectory based on the large amount of metadata that we collected in the HELMi cohort. Some circularity may be unavoidable but we think that there is no way around this and want to stimulate other scientists in the field to address this.

More specific points:

- How did the QC look for the samples? were all samples randomized before wet lab procedures? Check technical variables against time. How much did the batch effects affect the compositions and results (e.g. MiSeq vs HiSeq)? Please show these details. Describe sequencing depth, cut offs etc.
  - We thank the reviewer for raising this very important point. We explored the impact of the technical variations on the same dataset in this paper: (10.1016/j.ebiom.2023.104695), where we discuss extensively the difficulties and biases of the dataset. As often in large longitudinal cohorts, sequencing was performed in several rounds, preventing a complete randomization of the samples through the sequencing runs. Additionally, the DNA extractions of the samples were performed over a period of 43 months, making some technical variables such as the

date of extraction (and storage time), the DNA extraction batch, the sequencing technology and the person performing the extraction highly interlinked variables, whose individual effects are impossible to detangle. In our previous study, we show that the cumulative variance explained by the technical variables on the bacterial composition ranged between 6.8% and 15% in infant samples and 13% and 11% in mothers' and fathers' samples, respectively, which is in line with the technical variance reported by previous studies of smaller cohorts (10.1016/j.eclinm.2022.101403). The sequencing depth cut-off was chosen to be 3000 reads after QC to samples collected at 3 months or before and to 5000 paired reads for the remaining samples based on species richness to sequencing depth evaluations.

The method section has now been updated to reflect this additional information.

"The impact of the technical variation on this dataset is thoroughly explored in 10.1016/j.ebiom.2023.104695. The sequencing depth cut-off was chosen to be 3000 reads after QC to samples collected at 3 months or before and to 5000 paired reads for the remaining samples based on species richness to sequencing depth evaluations."

- The PCA approach of log-transformed relative abundances of genera is generally looking good in e.g. Figure 1, but the authors rely on this also for all downstream analyses. I would like to see some additional analyses of more microbiome specific tools for beta-diversity (weighted and unweighted) as well (Bray-curtis, Unifrac etc.).
    - There was a typo in methods saying we used the PC results for the clustering, we do not. This was corrected (line 591). We utilize the clustering results for the trajectories. The clustering was performed on the log-Pearson distance matrix, and the trajectories utilized the clustering scores (Line 229-230 and line 597). We have discovered that B-C distance does often not well represent true biological gradients in microbiota data, because it may overestimate the influence of a few high-abundant taxa, underestimating the effects of most microbes that are at a low abundance, due to the lack of log-transformation. The B-C distance was originally created for plant communities, which are typically a lot less dimensional (a few to some tens of species) with a lot smaller differences in relative abundances (minimal values being typically 1% or more, as less is not possible to estimate visually from a plant survey square). Microbiota data contain more features (up to thousands), and the range is several orders of magnitude higher. These facts render the plant-survey-based distance metric suboptimal for microbiota data. A log-transformed data matrix overcomes this problem and gives a more accurate representation of all taxa. Because of these reasons, we

have discovered that the log-Pearson distance nearly always outperforms the typically used beta-diversity metrics. We are currently writing a separate manuscript on this.

- For illustrations, I understand that it is interesting to look at PC1 and PC2 in relation to various exposures, however it would make sense to do all PCA statistics for the manuscript using all dimensions of the data with e.g. adonis2.

  We used adonis2 for the analyses where we included all time points at the same time but wanted to fit linear models for each time point to obtain accurate parameter estimates, which was needed to simulate microbiota development in order to estimate the effect of stochasticity vs. determinism. For this reason, we chose to fit models for the first two PCs. The first PCs represent 42.9% (PC2:6.7% and PC1 36.2%) of variation so they are a good proxy for the microbiota composition. We have now added a comment on this to the discussion and included a disclaimer like: "The predictability of microbiota development analysis only included the first two PCs and thus some variation is absent from the model". (line 462-463)

- I might have missed this, but how were repeated measures (the same child has multiple samples represented) handled in the beta-diversity analyses on exposures?
  - We analysed each time point separately. This was corrected (line 613).

- It is a bit problematic that the authors use this one PCA for so many of the downstream analyses as it will enhance and emphasize certain features of the composition and hide others. Similar, the use of envfit will again only rely on these PC1 and PC2 features across all samples. I would suggest doing individual exposure analyses for beta-diversity and differential analyses at each time point, not restraining analyses (distances) to that of the main PCA.
  - We agree that this is a valuable analysis, but that type of analysis was done in our recent paper (Jokela et al. 2023, 10.1016/j.ebiom.2023.104695), so we did not want to repeat the same analysis. This paper aimed not to exhaustively describe all exposure associations but to identify general patterns. PC scores were only used on the envfit analysis, and the predictability of microbiota development. The rest of the analysis didn't rely on the PCA. We also used adonis2, which includes all variation, not only the first 2 PCs.

- Most analyses are done at genus-level, but given the changes over time in many species (e.g. bifido going from HMO utilizers to more mature taxa) it could have been interesting to recapitulate at least some of the findings on ASV level. Also, many taxa

have some very beneficial species mixed with what could be considered bad guys belonging to the same genus. With V3-V4 sequenced quite a number of taxa can be suggested at species level.

- We agree fully with the reviewer that a lot of functionally important taxonomic variation is lost when using genus level data. However, the microbial genera do represent a good compromise between accuracy of taxonomic annotation and ecologically meaningful grouping. We chose the use the genus level in order to be more general rather than overly specific with the results. By not using ASV level data we are not focusing on specific variants in our own data rather than genera that would more likely be present in other datasets. Furthermore, testing the data at the ASV level unnecessarily splits the taxonomic groups into a large number of items, considerably increasing the risk of spurious correlations and making interpretation and generalisation more difficult. We wanted to have reliable and generalizable results and thus felt that going beyond genus level would compromise the accuracy of the taxonomic annotations.

- In Figure 2a-c the y-axis scale makes it very difficult to see details of each factor. Mainly because previous timepoint is so dominating.
  - We have altered the figures to make them more readable:
    - 1)Fig. 2a has had the y-axis changed.
    - 2)In 2b and 2c have changed the figure to show only "previous microbiota" and "Other" factors. As the emphasis of these two figures is to show the impact of previous microbiota compared to other factors. For transparency we have created a new extended data figure showing only the other factors as a heat map.

- For all figures the authors show age in weeks, while they in text refer to e.g. 6 months or 24 months sample. I would suggest aligning the labeling in the figures for easier communication.
  - When referring to figures we have changed the text to weeks to correspond to the figures. However, due to the broader viewpoint taken in the discussion section we have maintained the months and years as this offers a more coherent picture of early life development.

- Extended data figure 1 is difficult to read/comprehend.
  - This figure was removed.

- Line 130 seems to refer to something else than Extended data figure 2?

- Extended data figure 2 and Extended data figure 4 got mixed up and were corrected

- Line 137 – I can't find the described key taxa in Extended data figure 2. Instead of looking at indicator organisms for the PCs, I would suggest plotting the relative abundances of genera at each time point. This will show a more clean representation.
  - This is due to Extended data figure 2 and Extended data figure 4 being mixed up, which has now been corrected.

- I like the details around the splitting of vaginal birth into vaginal with and without ABX and how that affects the different taxa over time – could CS deliveries be split into emergency and elective as well?
  - This is not possible as the reasons (elective or emergency) for Cesarean sections were not collected.

- Also, the notion that siblings drive early Bifidobacterium and Lactobacillus abundance only in vaginal delivered is interesting, and I like the interpretation that it could be due to different maternal compositions.

- For the community types from clustering, it seems a bit odd to make the clustering at genus level and then afterwards (lines 184+) describe taxa on family level.
  - For clustering we tested both family and genus level. We present the family level results for simplicity because the patterns are clearly visible at the family level. However, the supplementary table 5 is at both genus and family.

- Please describe deeper the number of features used for training the 5-feature model for describing the community types. How many features were tested, how did you select these 5?
  - This has been added to the methods section (line 603-605) and reference the appropriate supplementary table (1) with a list of used variables.

- It is unclear how the community types vs disease were tested, how were repeated measures handled and how was multiple testing accounted for?
  - This is explained in the methods line 609-613.
    "Associations between health outcomes and clusters and trajectories was done by performing a logistic regression model adjusting for mother's BMI, mother's and father's allergies, mother's and father's education level, infant's sex, presence of pets in the household, household size (number of people in the home), mother's weight gain during pregnancy, and gestational diabetes." This was done separately for each

time point. No multiple testing adjustment was used because the number of tests was limited (31 tests).

- When evaluating the trajectories over time in relation to health outcomes, it is important only to evaluate outcomes happening after age 2 years. Else, you will have no idea about the direction of associations. One would think that fever episodes and infections could easily affect the gut microbiome development perhaps through antibiotics prescriptions or inflammation (then they become shaping factors). It would likewise be interesting to adjust your clinical results for the main shaping factors such as siblings and delivery mode. Was multiple testing accounted for?
    - To account for this, we only tested associations before the outcome was measured. This was clarified (line 614): "We only tested associations for health outcomes before the outcome took place." We agree that some associations may be driven by reverse causation and have now noted this in the discussion (line 615) We did not want to adjust for the main shaping factors, as that would likely "throw the baby out with the bathwater" - if most of the microbiota variation is eliminated statistically, one would not expect strong associations to remain. No multiple testing adjustment was used because the number of tests was not very high.

- Were additional exposures included in the trajectory evaluation such as: parental allergies and education level, maternal BMI, parental smoking, gestational diabetes, pregnancy weight gain?
    - A large number of variables, including those mentioned, were tested. However we only reported the significant variables. Supplementary table 1.
- Please indicate how many children were in each category in Fig 5g+h.
    - If by each category you mean in each trajectory it is both in Fig 5a-e (written inside each subplot), and in the text
        - T1 (N=388); T2 (N=95); T3 (N=78); T4 (N=151); and T5 (N=116). line 599;

- Figure 3c+d have switched legends 3d should be richness.
    - These have been corrected.

- Definition of pathobionts should be referenced.
    - References added in methods for pathobionts (line 600-610)

- In general the Supplementary tables appear extremely rough and difficult to read….
    - We attempted to make them more readable by reducing the number of decimal points, cleaning up the headers, optimizing bar sizes, and clarifying the column names.

Reviewer #2 (Remarks to the Author):

This study examines the role of the early-life human gut microbiota in regulating various aspects of host physiological development with implications for long-term health. The authors address the lack of longitudinally monitored cohorts by describing natural gut microbiota development in infants during their first years of life. They introduce a gut microbiota wellbeing index based on microbiota development and health data from nearly 1000 infants, demonstrating its predictability in determining developmental trajectories and later health outcomes. Key findings include the identification of distinct microbiota development trajectories, the influence of early exposures on these trajectories, and the predictive value of certain genera like Bifidobacterium and Bacteroides in health outcomes. Additionally, the study highlights the potential for using early gut microbiota information to enhance individual health risk predictions, emphasizing the malleability and predictability of gut microbiota succession during infancy.

- What are the noteworthy results?
    - The authors identified five distinct microbiota development trajectories and the influence of infant exposures on these trajectories
    - They pointed out the predictive value of certain genera like Bifidobacterium and Bacteroides in health outcomes
    - They highlighted the predictability of gut microbiota succession during infancy, as a way to enhance individual health risk predictions
- Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.
  Yes, there is originality in this study, and as the authors claim this is the first study proposing a gut microbiota index associated with wellbeing and health in children. With that said, there have been other important infant microbiome cohort studies previously published, although analyzed through different statistical methods. However, it is critical that authors state how their findings may be only applicable to a specific population: Scandinavian children. While general patterns of infant microbiome development are similar in terms of beta-diversity (weighed by similar abundances of predominant species like Bifidobacterium and Bacteroides), other ecological metrics differ between industrialized and non-industrialized pediatric populations, such as alpha-diversity and the abundance of many other taxa.

We are grateful to the reviewer for these helpful comments that are all addressed below.

General Comments:

- The manuscript reads nicely.

- The study is well-designed. There is a sufficient sample size, and both infant exposures and health outcomes have been carefully selected.

- In the Results section, I recommend including the results (like numbers and p-values) at the end of the sentences to avoid breaking the sentences. Some of these results could also be better organized and more focused if they were presented in tables (not only in the supp. but also in the main manuscript). a table?

    - Thank you for the suggestion where it was possible, we have moved the p values to the end of the sentences. In regard to a table we believe it would be unnecessary as all pvalues can be seen in their respective figures as well.

- It is critical that authors state how their findings may be only applicable to a specific population: Scandinavian children. While general patterns of infant microbiome development are similar in terms of beta-diversity (weighed by similar abundances of predominant species like Bifidobacterium and Bacteroides), other ecological metrics differ between industrialized and non-industrialized pediatric populations, such as alpha-diversity and the abundance of many other taxa.

-We attempted to keep the analyses as general as possible to enable comparison to other cohorts. In Fig. 5f we show the global normal follows T1, and the global compromised follows other trajectories. This indicates that there appears to be a true biological trajectory, which is not governed by geography or cultural practices. Thus, we believe the results in terms of gut microbiota development are somewhat generalizable to other populations. We specifically did not include alpha diversity or very uncommon taxa in the wellbeing index development as these indeed may vary between cohorts. We have added a sentence on the limitations to the discussion.

    - "In order to have a more general MWI that is more accessible we did not use alpha diversity or very uncommon taxa in the wellbeing index development as these indeed may vary between cohorts." (line 436-438)

- Please make sure all bacterial names, regardless of taxonomic level, are italicized, as suggested by the American Society of Microbiology (see here: https://journals.asm.org/writing-your-paper).
    - These were italicized as requested.

- Authors should include biological sex of infants in their models some of the microbial variance.
    - Sex was a variable used in the analysis and is listed in Supplementary table 1. The methods refer the reader to this table (line 544)
    - Specifically, sex was included as a confounder for community type and trajectory associations with health outcomes. (line 611). We tested infant sex for associations with community type, trajectory, and well-being index, however it had no significant effect.

- Extended Data Figure 2 should include the impact of exposures on age-specific ordination, yet it shows something else. Please include the correct figure referred to in lines 126-140.

    - Extended Data Figure 2 and Extended Data Figure 4 were mixed up and are corrected now.

- Authors should be specific when describing ecological metrics. For instance, when referring to microbiota composition (line 83) they should use the specific metric they are referring, which is beta-diversity measured by xx distance.
    - We have added the appropriate distance to this sentence (log-Pearson distance) line 81.

- While framing the analysis on beta-diversity findings makes sense, alpha-diversity is a very important feature to study ecosystems. It is therefore important that the authors include: (i) the differences in alpha-diversity (Chao1 and Shannon or Simpson) throughout age (ii) how it is impacted by some of the same exposures analyzed for b-diversity.
    - We show alpha diversity (richness) by cluster in fig 3c. However, we chose to not focus on alpha diversity as that is generally not a robust indicator of the ecology of the system due to a) it is strongly dependent on sampling depth and other technical biases and this would render our analysis less comparable to other studies and within our study due to batch effects, b) diversity may be driven equally by pathogens or beneficial microbes and thus diversity is very difficult to interpret. The interpretation of alpha diversity especially in infants is very difficult, reflected by the often-conflicting associations between alpha diversity and infant health. We have now added a sentence of alpha diversity and its association with age.

        - Microbial richness varied significantly between community types, being highest in C4 and lowest in C2 (Fig. 3c), and generally showing an increasing association with infant age. (Line 199)

- In the clustering analysis, it is misleading to calculate the percentage of pathobionts based on genus-level information. Microbial virulence does not occur at the genus level so this needs reassessment. Moreover, this potentially misleading result is not described in the results section. If authors decide to include the level of pathogens in each cluster, they should calculate species of known infectious potential.
    - We fully agree. However, since we used 16S rRNA data, we cannot identify taxa at species or strain level, which would be required to identify true pathogens. We therefore do not assume that the microbes are pathogens, but rather make the assumption that they are potential pathobionts, which may not express any virulence. Certain bacterial taxa have generally a potential to cause negative effects on the host, even if they do not cause a direct infection (which would be the case of pathogens). We have added list of the potential pathobionts and a reference to justify each. We have also added the word "potential" to reflect that we do not have the resolution to identify true pathogens.

- It is also unclear which richness index was used in the cluster analysis (Fig 3d). Regardless of which was used, it is unclear how clusters 3 and 4 have such low a-diversity if these include most of the older infant samples. It will be necessary to explore several a-diversity indices.

    - We calculated the richness as the number of species-level taxonomic groups in the sample. This was added to the Methods (line 583). Additionally, Fig 3c is the richness and 3d was the pathobionts. These were mistakes in the figure and have been corrected

- The recursive partitioning analysis is most interesting. It would also be important to understand what % of infants of a specific age fit within each of the clusters, in order to understand how frequent is for microbiomes of a specific age to fit outside of the typical or 'normal' cluster. Also, while the decision tree in Figure 4a is nice, there are other visualization techniques to better display the proportion of the exposures in each of the microbial clusters. Here is an example: https://paulvanderlaken.com/2020/03/31/visualizing-decision-tree-partition-and-decision-boundaries/
    - This information is already given in the figure as the number inside the "()". We have added a clarification of this to the figure legend (line 977). Due to the high

number of variables and their interactions, we felt that the decision tree was the only meaningful way to represent the results and provides relevant information also for non-experts.

- The developmental trajectories are quite revealing and a welcoming addition to this field.

  Thank you !

- Is it known what types of probiotics were mostly taken by the infants? If not the brand, the probiotic organism(s)?
  - The HELMi cohort has recorded whether the child has received probiotics and what kind/organism: Bifidobacterium, lactobacillus, or a mixture of both. These are listed in Supplementary table 1 and were analysed separately.

- Of the 388 infants in T1, almost half had a healthy childhood based on the clinical outcomes used, but the rest did not. Why is this, what factors may be associated with this and how does this impact the accuracy of the wellness index in predicting eubiosis and a healthy infancy?
  - For some of them, we did not have full health information and thus they may have been excluded from the healthy reference not because they were not healthy, but because of lack of data. This also likely reflects that microbiome is not a determining factor in overall health. Other factors play a role, and thus not all illnesses are fully predicted by the gut microbiome. A child can get ill even with the perfect microbiome and a healthy child can have a compromised microbiome.

- It is important to show a plot depicting the relative abundances of the indicator microbes of the microbiota wellness index as this can only be inferred from the extended data figures. This may be one of the most important aspects of this paper so this should be part of the final main figure.
  - Fig. 6 was updated to include this image.

- Please revise the Results section and make sure the figures are clear and accurate Extended Data Figure 3 is completely illegible. Authors may need a figure per timepoint for this.
  - Yes we agree several figures were not easy to read. This is due to putting a pdf image into word, which for some reason on a windows computer makes the text blurry. However, on the original figure file, which is the final version uploaded, they are much better quality than the word version. In regard to the size of the

figures and splitting them into multiple supplementary figures, we believe as this is a Supplementary figure, and would only be viewed digitally the figure is readable as it is possible to zoom in on the subplots and then they are perfectly legible. If we split the figure, it would be less comphrehensible as the figures represent a time point so it is more logical that they are combined into a single file. We have increased the quality of several figures, and they should be more legible now. (Supp. Fig 3, 4, and 11).

- Extended Data Figures 4 is also impossible to read. Also, the title of this figure should more than "Associations"
    - See previous comment above in regard to illegibility above.
    - The title of the figure was updated. "**Age-related associations of principal components**"

- This study comes up with interesting and novel findings. However, it is important to acknowledge that the study does not consider many different factors, such as diet, genetics, lifestyle, ethnicity, and geographical location, that have been shown to notably impact gut microbiome. I strongly recommend pointing this out, particularly in the discussion
    - We agree and have added a segment in the discussion (line 443): "As all children from our cohort are located entirely in Finland, little variation in terms of diet, genetics, lifestyle, ethnicity or geographic location are taken into account in our study. While race and geography are possible sources of variation, its believed that human behavior, such as birth mode and breastfeeding, have larger impacts on microbial colonization[70]"

Discussion and Conclusion sections.

- It is important to mention the importance of other gut microbial communities, including microbial , such as fungi, phages, viruses and protists that have been shown to interact with gut bacteria and significantly influence the host's developmental pathways. It is necessary to emphasize that when you speak of gut microbiota in this study, you are talking about only bacteria and the role of other microbial communities should be studied in the future to improve the prediction of health-associated infant gut microbiota development patterns.
  I strongly recommend noting that, considering the last two matters, these results cannot be generalized to all infants.
    - We have addressed this in the discussion (line 501-502)

- "Our study focuses entirely on the bacteriome of the gut and does not consider the other gut microbiota including fungi, eukaryotes, and viruses."

Comments on Main (Introduction):

- I suggest referring to the cohort in the last paragraph of the introduction.
  - Yes, that is a good idea. We have added the reference.

Comments on Results:

- When using GLM to test the effect of multiple predictors or conducting multiple comparisons, it is recommended to adjust the p-values to control the likelihood of false positives. Alternatively, you can run a differential expression analysis using R packages like DESeq2 or EdgeR, which perform multiple testing corrections for p-values by default.
  Specify the AIC values in the manuscript (page 5) and on the related plots.
  Include the p-values of the X2 test (page 6).
  - We have observed that using DESeq2 and EdgeR does not improve the accuracy of microbiome findings, and therefore we chose not to use these methods. Multiple testing adjustment was done when analysing associations to microbial taxa (Benjamini-Hochberg correction). These are presented in the supplementary files. However, we chose to present the raw p-values in the text for easier interpretation and consistency.

- Whenever a p-value is presented, it is a good idea to mention the statistical test as well. This information can be brought to the end of the sentence to prevent breaking the sentences, which reduces the fluency of reading.
  - Thank you. We have attempted this approach where possible. .

Additional comments on Figures & Tables:

- Figure 1-a: In the caption, please indicate what each point represents.
  - We added "from 3-104 weeks" to the caption for Fig1 (line 951)
- Figure 1-a: I recommend adding the results of the statistical test to the plot.
  - As the figure is already busy, it would become too crowded if further information was added. We have indicated the p value description in the legend.

- Figures 1-b, 1-c, 2-d: Indicate which comparisons are represented by the stars and their colors.

- This was added to the legend.

- Figures 2a-c: Side-by-side stacking seems to be a more appropriate way to display multivariate regression results (See figure 2 in the following article as an example: https://www.nature.com/articles/s41586-018-0617-x#Fig2, which is also very similar to the extended data figure 2 in your own manuscript).
    - We have altered the figure to better show the bars. plot a) has its y axis altered, b) and c) have changed the variables to "previous microbiota" and "other" variables to emphasize the importance of early microbiota against other variables. To provide further clarity a new Supplementary figure 1, was created to elucidate these two plots showing the other variables in a heat map.

- Figure 2d: It is not clear enough if the family names shown in the middle plot titles refer to the entire row. I suggest moving these titles to the right or left end of each row.
    - We have added a line to separate the different Families in the figure.

- Figure 3a: Fix the overlapped texts (numbers) on both axes.
    - This was corrected.

- Figures 3a legend, 3b-d x-axes: Community types are presented as 1-4 rather than C1-C4, as mentioned in the manuscript (page 4) and caption.
    - The abbreviation "C" is meant to represent "community type", so C1 would be "community type" 1. We have added an explanation to the figure legend.

- Figure 3b: Why aren't the stacks reaching relative abundance 1.00? If it's because NAs were removed, please specify in the caption.
    - They don't reach 100% because there are many more genera (above 200) required to reach 100%, and these are not plotted. We have included this explanation in the Figure legend.

- Figures 3c-d: The y-axis titles seem to be reversed! In accordance with the caption and the manuscript (lines 189-191), the y-axis in these two plots should be labeled as "Microbial Richness" in Figure 3c and "Relative abundance" of pathobionts in Figure 3d.
    - This was corrected in the figure

- Figures 5g-h: Enlarge the legend and Include significance levels (stars) in the caption.

- Legend changed to be more readable. Significance stars explanation added to caption (line 990).

- Figure 6: Layout the figure.
  - We have carefully checked the layout of the Figure 6 and think it is OK now.
- Extended Data Fig. 5: Consider adding community-type labels.
  - EDF5 community type labels have been added.

- Extended Data Figs 6 and 7: Consider adding trajectory labels.
  - EDF6 is community type and not trajectory and each column is labelled such at the bottom.
  - EDF7 Trajectories labels have been added.

- Extended Data Fig. 11: Indicate which time point each plot represents.
  - The legend was updated with time point.

- Extended Data 10. The colored bars on top of the heatmap are unclear because a legend is missing.
  - We have added a color bar for the heatmap bars, and extended data 10 has now been combined with Fig6 as per your suggestion.

- Extended Data Fig 11 is also impossible to read. The vertical bars on the left of each heatmap are also unclear as a legend is missing.
  - We have added a legend for the color bars.

Comments on Discussion:

- Line 370: While technical artifacts may be one reason for these results, I suggest you point out some of the other potential factors as well. It is true that in the study you've cited (Xiao et al, 2021), only 35% of the infants were not breastfed (formula-fed babies have significantly higher proportions of Bacteroides, Fallani et al, 2010), and only 8% were delivered through C-section (as you mention later in your discussion, C-section is associated with a high abundance of Bacteroides). However, as I quickly glanced through their data, many other factors that are known to be potential drivers of infants' gut microbial composition such as maternal dietary intake, breast milk composition, weaning time, and type of diet were not measured. You may find other examples in the literature, as well. As we learn more about these factors and many others that could affect the microbial composition of the infants' gut, we may be able to explain these differences in findings.

- We have added a comment on this to the text: "This may reflect the uniqueness of the particular cohort..." (line 387).

- Line 423: Although these results are intriguing, it is important to acknowledge that most (if not all) of these studies have been conducted in industrialized countries, so the results might differ for less industrialized and rural populations.
    - It is true that non-industrialised populations have a different gut microbiota composition, and likely also a different developmental pattern. We have added a comment on this: "It is however, possible that our results do not generalise to populations with very different genetic backgrounds or lifestyle patterns." (line 390-391)

Comments on Materials and Methods:

Well explained

Data and code are available

- Statistical analysis: Describe how the stool color was detected and analyzed, based on observation or any other method.
    - Collection and detection of stool color is described in Suppl. table 1, line 42. "Typical classification of the appearance of stool over the past weeks at home (in the bristol scale) at home: yellow, green, light brown or dark brown". Statisitcal analysis of differences between community types was performed using the χ^2 test. This was shown on line 202, but was added to lines 599 in the methods.

- Statistical analysis: Specify that the partitioning tree was used to determine the community types.
    - The partioning tree was not used to determine the community types, but to describe the variation of the community types. K-means clustering on pearson distance was used to classify the community types. The partitioning tree describes how four variables: age, delivery mode, siblings, and use of a bifidobacterial probiotic described 56% of the variation in community types.
- Statistical analysis: Based on what resource did you define the maturity index? If it was an article, could you cite it? In that case, are there any other factors or ways to measure maturity that you think should be considered?
    - We created our own maturity index and it is explained in the methods section line 634-638.

Other Comments:

- Italicize: Bacteroidia (line 172), Bacteroidia and Actinobacteria (line 183), Actinobacteria, Bacteroidia, Enterobacteria, Negativicutes, Bacilli and Clostridia (lines 283-284),

staphylococci and streptococci (line 401), and any other bacterial names that are not italicized throughout the manuscript, regardless of the taxonomic level.

- We have italicized all official taxonomic names regardless of level, however we don't agree with italicizing some names (e.g. staphylococci and streptococci), as these are common names as well as genus names in the plural and are lower case non-italicized ("Do not italicize or capitalize genus name when used in the plural." https://wwwnc.cdc.gov/eid/page/scientific-nomenclature)

- Lines 177-178, 196: These lines belong to the Statistical analysis in the Methods section.
  - We believe this line makes for a clear introduction to the topic, and in the methods section the explanation is elaborated on.

- Lines 251-256: The Discussion section would be a better place for this level of interpretation.
  - Thank you for the suggestion, but we believe that this fits better here than in the discussion.
- Lines 261-264: Rephrase these lines to clearly state that T1 was negatively associated with all the following variables after the comma.
  - Added "T1 was negatively associated with the following:" line 273

- Please revise the manuscript for the typos, as I have noticed a few of them; for example, postiive (line 455) assocition (456).
  - We have proofread the article again and hope we have corrected all the typos.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks on code availability):

I have reviewed the code to make sure it makes sense, which it does. However, I have not run the code to know if it is reproducible or notThank you – we have run the code to make sure it is reproducible.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you for allowing me to review this again. I find the manuscripts improved, however there are still some major points remaining to be sufficiently incorporated in the manuscript.

Response 1. Thank you for the thorough explanation of the technical variation in the data. With such high amounts of variation explained by technical covariates, the authors should include a supplementary table showing how their predictors and endpoints are associated with these technical variables, and if batch-endpoint overlaps exist, this needs to be accounted for in all analyses.

Response 2. I have no objections against using the log-pearson distance as main analysis. As the authors say, every distance holds different information. I would therefore still argue that a transparent demonstration showing their results applying other weighted as well as unweighted beta-diversity distances would be a great strength, then a reader would be able to interpret whether it is the larger taxa (as B-C / weighted unifrac would emphasize) or the presence of taxa (as Jaccard / unweighted unifrac would emphasize) that carries most signal.

Response 6. Genus vs ASV/species. I do not agree with the authors on this point. Especially when evaluating early life microbiota in the gut there are dramatic changes going on. I would still recommend recapitulating some of their findings on ASV or species level.

Response 16. Trajectories vs health outcomes. I don't understand the comment "We only tested associations for health outcomes before the outcome took place." Your trajectories are built on clusters throughout the first two years of life, dividing the children into trajectory groups – and these are used for associations with health outcomes both later than 2 years which is ok and before 2 years which is not ok. I would suggest removing these

early endpoints or represent them in a different way not claiming them to be outcomes but could just as well be shaping factors.

Also, I would still like to see the associations adjusted for shaping factors. “Throwing the baby out with the bathwater” is definitely not what you are doing when adding these to the models, rather being transparent about potential mechanisms behind your findings, which makes good points for discussion and hypothesis building.

Again, multiple test corrections should be included and indicate how many outcomes were tested for in the manuscript. This point is also brought up by Reviewer 2 and should not be ignored. Results can be shown with both nominal and adjusted p-values.

Response 17. Exposures vs trajectories. Please add to the manuscript how many exposures were tested and why only the significant were presented. Include multiple test adjusted p-values.

Reviewer #2 (Remarks to the Author):

Your efforts to improve the manuscript are greatly appreciated. Here are a few minor suggestions:

Line 583: Thanks for clarifying the definition of "richness index" as the number of species-level taxonomic groups in the sample. However, the name of the richness index used to calculate is still missing.

Figures 1-b, 1-c, 2-d: Thank you for including the significance levels for the stars. However, I think these figures also need a legend explaining what the colors of the stars indicate. My guess is that the star colors are related to comparing the group with the same color as the star with the group next to it. However, it needs to be clearly stated in the caption or legend.

I still do not see any reference to the cohort name in the introduction, despite it being mentioned in the responses. Knowing about the cohort earlier in the paper will be helpful to the readers.

I understand your point about including raw p-values when reporting the results of GLM test. However, there should be a clarification somewhere in the text or supplementary material that you have adjusted the p-values using the BH correction in order to avoid false positives.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks on code availability):

I had already reviewed the code on the first review.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you for allowing me to review this again. I find the manuscripts improved, however there are still some major points remaining to be sufficiently incorporated in the manuscript.

Response 1. Thank you for the thorough explanation of the technical variation in the data. With such high amounts of variation explained by technical covariates, the authors should include a supplementary table showing how their predictors and endpoints are associated with these technical variables, and if batch-endpoint overlaps exist, this needs to be accounted for in all analyses.

-We have now analyzed the impact that the technical variables on both background variables and later life health outcomes. These are added as a Supplementary table (Supplementary table 10). Associations of the technical variables (Run ID, Sample storage days at home, Sample storage days at the laboratory, and sample concentration) were analysed using a generalized linear model by time point. We see that for the background variables the run id was significant for C section deliveries at weeks 13 and 39 and week 78 significant for Lactobacillus probiotic and No formula before 6 months. While for health outcomes it is significant for several variables from week 3 to week 52. Other technical variables were only significant sporadically.

We therefore reran the analyses using Run id as a confounder and have updated the figures and tables. Some minor changes were seen in associations with health outcomes and no differences were seen with the background variables. The health outcomes figure was additionally updated with birth mode (Response 16). The results (lines 271-293) have been updated for the new figure.

Response 2. I have no objections against using the log-pearson distance as main analysis. As the authors say, every distance holds different information. I would therefore still argue that a transparent demonstration showing their results applying other weighted as well as unweighted beta-diversity distances would be a great strength, then a reader would be able to interpret whether it is the larger taxa (as B-C / weighted unifrac would emphasize) or the presence of taxa (as Jaccard / unweighted unifrac would emphasize) that carries most signal.

-A supplementary figure (Supplementary figure 5), has been added which shows the pca at genus level with a) Bray-curtis distance, b) Jaccard distance and c) Robust Aitchinson distances (lines 185-190). For each image, the existing community types developed with the log-pearson distance is shown. We show the existing (genus level) community types work well even with other distances. Discussed in lines 381-383.

Response 6. Genus vs ASV/species. I do not agree with the authors on this point. Especially when evaluating early life microbiota in the gut there are dramatic changes going on. I would still recommend recapitulating some of their findings on ASV or species level.
-We have performed an additional analysis on the species level and added this as a supplementary figure (Supplementary figure 11). The PCA shows a similar pattern as the genus level data, and the trajectories hold similar associations as the genus level for both the background variables and health outcomes. This was added to line 305-310.

Response 16. Trajectories vs health outcomes. I don't understand the comment "We only tested associations for health outcomes before the outcome took place." Your trajectories are built on clusters throughout the first two years of life, dividing the children into trajectory groups – and these are used for associations with health outcomes both later than 2 years which is ok and before 2 years which is not ok. I would suggest removing these early endpoints or represent them in a different way not claiming them to be outcomes but could just as well be shaping factors.
-We feel that I this is the best location as they are health outcomes, however we have addressed the concerns of reverse causation by discussing this in the results line 273-274 and discussion lines 409-415.

Also, I would still like to see the associations adjusted for shaping factors. "Throwing the baby out with the bathwater" 🤔 is definitely not what you are doing when adding these to the models, rather being transparent about potential mechanisms behind your findings, which makes good points for discussion and hypothesis building.
-We had already adjusted the models for maternal BMI, weight gain, and gestational diabetes, parents' smoking, allergic disease and education, infant sex, siblings, pets, household size, house type. We have now adjusted the model for sequencing run ID (technical variable, reviewer Response1), and birth mode as the reviewer requested. We see some minor changes in the updated figure 5h.

Again, multiple test corrections should be included and indicate how many outcomes were tested for in the manuscript. This point is also brought up by Reviewer 2 and should not be ignored. Results can be shown with both nominal and adjusted p-values.
-We have performed the multiple test corrections as requested and added the FDR adjusted p values to a supplementary table (Supplementary table 8).

-The number of outcomes were added for each test (632 & 641)

Response 17. Exposures vs trajectories. Please add to the manuscript how many exposures were tested and why only the significant were presented. Include multiple test adjusted p-values.

The number of exposures has been added (line 650). Multiple test adjusted p values have been added in a Supplementary table 8.

Reviewer #2 (Remarks to the Author):

Your efforts to improve the manuscript are greatly appreciated. Here are a few minor suggestions:

Line 583: Thanks for clarifying the definition of "richness index" as the number of species-level taxonomic groups in the sample. However, the name of the richness index used to calculate is still missing.

-The definition of richness “as number of species-level taxonomic groups in the sample.” is on line 565.

Figures 1-b, 1-c, 2-d: Thank you for including the significance levels for the stars. However, I think these figures also need a legend explaining what the colors of the stars indicate. My guess is that the star colors are related to comparing the group with the same color as the star with the group next to it. However, it needs to be clearly stated in the caption or legend.

-Thank you, we have updated the legend for both figure 1 and 2 to describe what the stars represent on lines 1016-1018 and 1028-1031

I still do not see any reference to the cohort name in the introduction, despite it being mentioned in the responses. Knowing about the cohort earlier in the paper will be helpful to the readers.

-I apologise for that and thank you. We have now referenced it in the introduction (line 69).

I understand your point about including raw p-values when reporting the results of GLM test. However, there should be a clarification somewhere in the text or supplementary material that you have adjusted the p-values using the BH correction in order to avoid false positives.
-We have added this to the methods lines 645-648

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review

and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks on code availability):

I had already reviewed the code on the first review.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you for addressing my remaining points. I am satisfied with the responses.