



RESEARCH ARTICLE

**REVISED** **Nasopharyngeal Dysbiosis Precedes the Development of Lower Respiratory Tract Infections in Young Infants, a Longitudinal Infant Cohort Study [version 2; peer review: 1 approved, 2 approved with reservations]**

Rotem Lapidot <sup>1,2</sup>, Tyler Faits<sup>3</sup>, Arshad Ismail <sup>4</sup>, Mushal Allam<sup>4</sup>, Zamantungwak Khumalo<sup>4,5</sup>, William MacLeod <sup>6</sup>, Geoffrey Kwenda <sup>7</sup>, Zachariah Mupila<sup>8</sup>, Ruth Nakazwe<sup>9</sup>, Daniel Segrè<sup>10-13</sup>, William Evan Johnson<sup>3</sup>, Donald M Thea<sup>6</sup>, Lawrence Mwananyanda <sup>8</sup>, Christopher J Gill <sup>6</sup>

- <sup>1</sup>Pediatric Infectious Diseases, Boston Medical Center, Boston, MA, 02118, USA
- <sup>2</sup>Pediatrics, Boston University School of Medicine, Boston, MA, 02118, USA
- <sup>3</sup>Computational Biomedicine, Boston University School of Medicine, Boston, MA, 02118, USA
- <sup>4</sup>Sequencing Core Facility, National Institute for Communicable Diseases, Johannesburg, 2131, South Africa
- <sup>5</sup>Department of Veterinary Tropical Diseases, University of Pretoria, Pretoria, 0002, South Africa
- <sup>6</sup>Department of Global Health, Boston University School of Public Health, Boston, MA, 02118, USA
- <sup>7</sup>Department of Biomedical Sciences, School of Health Sciences, University of Zambia, Lusaka, Zambia
- <sup>8</sup>Right To Care, Lusaka, Zambia
- <sup>9</sup>Department of Pathology and Microbiology, University Teaching Hospital, Lusaka, Zambia
- <sup>10</sup>Bioinformatics Program and Biological Design Center, Boston University, Boston, MA, 02118, USA
- <sup>11</sup>Department of Physics, Boston University, Boston, MA, 02118, USA
- <sup>12</sup>Department of Biology, Boston University, Boston, MA, 02118, USA
- <sup>13</sup>Department of Biomedical Engineering, Boston University, Boston, MA, 02118, USA

**v2** **First published:** 12 Apr 2022, **6:48**  
<https://doi.org/10.12688/gatesopenres.13561.1>  
**Latest published:** 20 Mar 2024, **6:48**  
<https://doi.org/10.12688/gatesopenres.13561.2>

**Abstract**

**Background**

Infants suffering from lower respiratory tract infections (LRTIs) have distinct nasopharyngeal (NP) microbiome profiles that correlate with severity of disease. Whether these profiles precede the infection or are a consequence of it, is unknown. In order to answer this question, longitudinal studies are needed.

**Methods**

We conducted a retrospective analysis of NP samples collected in a

**Open Peer Review**

**Approval Status** ? ✓ ?

	1	2	3
<b>version 2</b> (revision) 20 Mar 2024	? view	✓ view	? view
	↑	↑	↑
<b>version 1</b> 12 Apr 2022	✗ view	✗ view	? view

1. **Wouter A.A. De Steenhuijsen Piters** ,  
 University Medical Center Utrecht, Utrecht,  
 The Netherlands

longitudinal birth cohort study of Zambian mother-infant pairs. Samples were collected every two weeks from 1-week through 14-weeks of age. Ten of the infants in the cohort who developed LRTI were matched 1:3 with healthy comparators. We completed 16S rRNA gene sequencing on the samples each of these infants contributed and compared the NP microbiome of the healthy infants to infants who developed LRTI.

## Results

The infant NP microbiome maturation was characterized by transitioning from *Staphylococcus* dominant to respiratory-genera dominant profiles during the first three months of life, similar to what is described in the literature. Interestingly, infants who developed LRTI had distinct NP microbiome characteristics before infection, in most cases as early as the first week of life. Their NP microbiome was characterized by the presence of *Novosphingobium*, *Delftia*, high relative abundance of *Anaerobacillus*, *Bacillus*, and low relative abundance of *Dolosigranulum*, compared to the healthy controls. Mothers of infants with LRTI also had low relative abundance of *Dolosigranulum* in their baseline samples compared to mothers of infants that did not develop an LRTI.

## Conclusions


Our results suggest that specific characteristics of the NP microbiome precede LRTI in young infants and may be present in their mothers as well. Early dysbiosis may play a role in the causal pathway leading to LRTI or could be a marker of underlying immunological, environmental, or genetic characteristics that predispose to LRTI.


## Keywords

Lower Respiratory Tract Infection, Nasopharyngeal Microbiome, Dysbiosis, Longitudinal Cohort study

National Institute for Public Health and the Environment, Bilthoven, The Netherlands

**Mari-Lee Odendaal**, National Institute for Public Health and the Environment, Bilthoven, The Netherlands  
Utrecht University, Utrecht, The Netherlands

2. **Frank A Scannapieco** , University at Buffalo, NY, USA

3. **Carter Merenstein** , University of Pennsylvania, Pennsylvania, USA

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Rotem Lapidot ([rotemlapidot@gmail.com](mailto:rotemlapidot@gmail.com))

**Author roles:** **Lapidot R:** Conceptualization, Data Curation, Methodology, Project Administration, Writing – Original Draft Preparation; **Faits T:** Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Ismail A:** Data Curation, Formal Analysis, Investigation, Writing – Review & Editing; **Allam M:** Data Curation, Formal Analysis, Investigation, Writing – Review & Editing; **Khumalo Z:** Data Curation, Formal Analysis, Investigation, Writing – Review & Editing; **MacLeod W:** Data Curation, Formal Analysis, Investigation, Writing – Review & Editing; **Kwenda G:** Investigation, Writing – Review & Editing; **Mupila Z:** Investigation, Writing – Review & Editing; **Nakazwe R:** Investigation, Writing – Review & Editing; **Segrè D:** Validation, Writing – Review & Editing; **Johnson WE:** Formal Analysis, Funding Acquisition, Methodology, Supervision, Visualization, Writing – Review & Editing; **Thea DM:** Supervision, Writing – Review & Editing; **Mwananyanda L:** Supervision, Writing – Review & Editing; **Gill CJ:** Conceptualization, Funding Acquisition, Methodology, Supervision, Visualization, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by The Southern Africa Mother Infant Pertussis Study, PI Gill funded by the Gates Foundation, OPP1105094 and the SAMIPS – Nasopharyngeal Carriage (SAMIPS-NPC). PI Gill. Funder NIH/NIAID (1R01AI133080). WEJ and TF were supported by funds from the NIH, U01CA220413 and R01GM127430.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2024 Lapidot R *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Lapidot R, Faits T, Ismail A *et al.* **Nasopharyngeal Dysbiosis Precedes the Development of Lower Respiratory Tract Infections in Young Infants, a Longitudinal Infant Cohort Study [version 2; peer review: 1 approved, 2 approved with reservations]** Gates Open Research 2024, 6:48 <https://doi.org/10.12688/gatesopenres.13561.2>

**First published:** 12 Apr 2022, 6:48 <https://doi.org/10.12688/gatesopenres.13561.1>

**REVISED Amendments from Version 1**

This updated version of the manuscript has been revised as follows:

The abstract was changed to address reviewers' comments.

We have re-organized sections in the manuscript, shortened the background and removed redundant text.

In the method section we have added information on the study population and study design and clarified how infants with lower respiratory tract infections were defined, based on the WHO criteria (adjusted to the purpose of this analysis).

Importantly, we have addressed the concern for sample contamination in our study, given the nature of respiratory samples in a very young cohort, both in the methods and the limitation section.

We have removed all Extended data and figures, corrected Table 2 and Table 4 and added a figure of the mothers NP microbiome (Figure 5).

In our discussion and conclusions, we have revised the wording to reflect a more cautious interpretation of the results given the limitations of this study as detailed in the manuscript.

**Any further responses from the reviewers can be found at the end of the article**

## Background

Lower respiratory tract infections (LRTI), including pneumonia and bronchiolitis, are the leading cause of death in children under five years of age, accounting for 1.3 million deaths each year, with 81% concentrated in children 2 years or younger (Cao *et al.*, 2019; Fischer Walker *et al.*, 2013).

Increasingly, LRTI is seen as a consequence of the interaction between the pathogen and other contextual factors. Such factors include the immune state of the host, intercurrent viral infections and may also be the microbial ecosystem in which the pathogen exists, *i.e.*, the nasopharyngeal microbiome (Fujiogi *et al.*, 2022; Man *et al.*, 2017). Interactions between the microbiome and a specific potential pathogen (*i.e.*, a pathobiont), could influence the behavior of that pathogen to either impede or promote LRTI (Brugger *et al.*, 2016; Stewart *et al.*, 2017).

Several cross-sectional studies have found that children with LRTIs often have distinct nasopharyngeal (NP) microbiome profiles at time of infection compared with healthy children. The NP microbiome profiles appear to be dominated by bacterial genera that differ between respiratory infections and health. For example, NP microbiomes dominated by *Streptococcus* and *Haemophilus* are associated with LRTI, whereas microbiome profiles dominated by *Moraxella*, *Corynebacterium* and/or *Dolosigranulum* characterize healthy children. Furthermore, NP microbiome characteristics correlate with the severity of respiratory disease and with clinical outcomes (de Steenhuisen Pipers *et al.*, 2015; Hasegawa *et al.*, 2017; Man *et al.*, 2017). While provocative, such observations largely rest on cross-sectional studies, and so cannot resolve the direction

of cause and effect: we do not know whether these microbial profiles are a result of the infection or whether they preceded it. If the latter is true, then differences in the NP microbiome could potentially represent a state of vulnerability, participating in a causal pathway leading to LRTI.

To draw such inferences, it is necessary to have longitudinal data, with sampling of infants before the development of the LRTI. Since LRTI is a rare event, collecting longitudinal data is complicated by the large number of infants needed to be followed. Between 2015 and 2016, our team conducted a prospective cohort study in Zambia, The Southern Africa Mother Infant Pertussis Study – SAMIPS (Gill *et al.*, 2016) and was able to create a biological sample library that allowed a longitudinal analysis of this kind.

Within this cohort of 1,981 healthy infants a sub-set of 10 infants developed LRTI based on standard WHO clinical criteria (*Revised WHO classification and treatment of childhood pneumonia at health facilities • EVIDENCE SUMMARIES •*, 2014), and adjusted for the purpose of this analysis. We focused on the following fundamental analyses: 1) what is the 'normal' pattern of NP microbiome maturation over the first several months of life? 2) how does this contrast with the maturation of NP microbiome of infants who developed LRTI? 3) is there evidence that NP dysbiosis precedes the onset of LRTI? 4) are there distinct microbiome profiles that characterize sickness and health and other infant characteristics? 5) Is there also evidence of specific NP microbiome characteristics among the mothers of infants who later developed LRTI?

## Methods

### Study population

This is a nested time-series case comparator study within the prospective longitudinal Southern Africa Mother-Infant Pertussis study (SAMIPS). SAMIPS was a study of the burden of pertussis in Zambian infants in which infants and their mothers were followed over the first three months of life. Full methods description is previously detailed by Gill *et al.* (Gill *et al.*, 2016), in short: All infants enrolled to SAMIPS were from the large periurban slum called Chawama compound, were less than ten days of age, born term, via normal vaginal delivery, were not underweight and were deemed healthy after birth. All infants received scheduled vaccines. Immunization schedule in Zambia includes the Bacillus Calmette-Guèrin (BCG) vaccine at birth. The Diphtheria, Tetanus and whole cell pertussis vaccine, H. influenza B, and Hepatitis B vaccines (DTwP-Hib-HepB), the 10-valent pneumococcal conjugate vaccine (PCV10) and the Oral Poliomyelitis Vaccine (OPV) at one month, two months and three months of age. The inactivated polio vaccines is scheduled at 3 months of age, and the 2 doses of Rotavirus vaccine at one and two months of age. Median age of the mothers was 25 (IQR 21-29), more than 90% were married and those who were HIV positive were enrolled only if they were receiving antiretroviral treatment. Written informed consent was obtained as appropriate from mothers of infants enrolled in the study.

The study was approved by the ethical review committees at the ERES Converge IRB in Lusaka, Zambia, and at Boston University Medical Center. All mothers provided written informed consent, with consent provided in English, Bemba or Nyanja as preferred by the participant.

### Study design

Mother-infant pairs were enrolled when mothers returned for their first postpartum well-child visit at one week of age. At enrollment, and 2–3-week intervals thereafter, through 14 weeks, we obtained a posterior nasopharyngeal (NP) swab from both mother and baby, with additional unscheduled visits and swabs obtained adventitiously if either returned seeking care for an acute respiratory infection. In each visit, clinical symptoms were documented as well as prescribed antibiotics.

Within the SAMIPS cohort, we identified ten infants who during the study period suffered from symptoms of lower respiratory tract infection (LRTI) as adopted from the WHO (*Revised WHO classification and treatment of childhood pneumonia at health facilities • EVIDENCE SUMMARIES •*, 2014). Since the aim of the WHO classification is to identify all possible cases of LRTI/pneumonia and guide management and treatment of these infants, the sensitivity of the constellation of symptoms is high. The WHO classifies pneumonia as cough/cold symptoms AND fast breathing and/or chest indrawing, and severe pneumonia as pneumonia with general danger signs including inability to drink, persistent vomiting, convulsions, lethargy etc. Within the SAMIPS cohort we identified 247 infants who developed LRTI during the study period based on these definitions. To increase the specificity of LRTI cases for our analysis we identified 10 infants who had both chest indrawing *and* fast breathing, most of which also had other danger signs (lethargy, poor feeding, vomiting or convulsions), and thus presented with severe LRTI. We then matched these case infants 1:3 by season of birth, and number of siblings with healthy comparators. All longitudinal samples of infants in our cohort, as well as the sample taken from the mothers at the first study visit were included in the microbiome analysis.

### Sample processing and storage

NP swabs were obtained from the posterior nasopharynx using a sterile flocked tipped nylon swab (Copan Diagnostics, Merrieta, California). The swabs were then placed in universal transport media, put on ice and transferred to our onsite lab on the same campus, where they were aliquoted and stored at -80°C until DNA extraction. DNA was extracted using the NucliSENS EasyMagG System (bioMérieux, Marcy l’Etoile, France). Extracted DNA was stored at our lab located at the University Teaching Hospital in Lusaka at -80°C. Sample collection, processing and storage were previously described (Gill *et al.*, 2016).

### 16S ribosomal DNA amplification and MiSeq sequencing

For 16S library preparations, two PCR reactions were completed on the template DNA. Initially the DNA was amplified with primers specific to the V3–V4 region of the 16S rRNA gene

(Klindworth *et al.*, 2013). The 16S primer pairs incorporated the Illumina overhang adaptor (16S Forward primer

5’-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACAGCCTACGGGNGGCWGCAG-3’;

16S reverse primer

5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3’)

Each PCR reaction contained DNA template (~12 ng), 5µl forward primer (1µM), 5 µl reverse primer (1µM), 12.5 µl 2 X Kapa HiFi Hotstart ready mix (KAPA Biosystems Woburn, MA), and PCR grade water to a final volume of 25µl. PCR amplification was carried out as follows: heated lid 110°C, 95°C for 3 min, 25 cycles of 95°C for 30s, 55°C for 30s, 72°C for 30s, then 72°C for 5 min and held at 4°C. Negative control reactions without any template DNA were carried out simultaneously.

PCR products were visualized using Agilent TapeStation (Agilent Technologies, Germany). Successful PCR products were cleaned using AMPure XP magnetic bead-based purification (Beckman Coulter, IN). The IDT for Illumina Nextera DNA UD Indexes kit (Illumina, San Diego, CA) with unique dual index adapters were used to allow for multiplexing. Each PCR reaction contained purified DNA (5 µl), 10 µl index primer mix, 25 µl 2X Kapa HiFi Hot Start Ready mix and 10 µl PCR grade water. PCR reactions were performed on a Bio-Rad C1000 Thermal Cycler (Bio-Rad, Hercules, CA) Cycling conditions consisted of one cycle of 95°C for 3 min, followed by eight cycles of 95°C for 30 s, 55°C for 30 s and 72°C for 30 s, followed by a final extension cycle of 72°C for 5 min. PCR products of negative controls were confirmed negative on Agilent TapeStation (no band observed).

Prior to library pooling, the indexed libraries were purified with Ampure XP beads and quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA). Purified amplicons were run on the Agilent TapeStation (Agilent Technologies, Germany) for quality analysis before sequencing. The sample pool (2 nM) was denatured with 0.2N NaOH, then diluted to 4 pM and combined with 10% (v/v) denatured 20 pM PhiX, prepared following Illumina guidelines. Samples were sequenced on the MiSeq sequencing platform at the NICD Sequencing Core Facility, using a 2 x 300 cycle V3 kit, following standard Illumina sequencing protocols. Negative controls were sequenced as well, resulting in extremely low reads that were not analyzed further.

In addition to the negative controls, we processed all samples in random, blinding for timing of collection as well as clinical data.

### Data processing

We assessed the quality of the sequencing data using FastQC (Andrews, 2010), which indicated that the overall sequencing quality was excellent, with mean Phred quality

scores remaining greater than 30 (>99.9% accuracy) for over 200bp for both forward and reverse reads. We used *Trimmomatic* (Bolger *et al.*, 2014) to trim Illumina adapters and remove low-quality sequences, setting the tool's parameters to LEADING:6, TRAILING:6, SLIDINGWINDOW:6:15, and MINLEN:36. This quality filtering removed less than 0.5% of reads from each sample.

Sequencing data were processed using Pathoscope2 (Hong *et al.*, 2014; Odom *et al.*, 2023). Samples with less than 10,000 reads were excluded from further analysis.

We used PathoScope 2 to assign sequencing reads to bacterial genomes. We used all of RefSeq's representative bacterial genomes (downloaded November 2, 2018) as a PathoScope reference library. From PathoScope's subspecies-level final best hit read numbers, we compiled counts tables and relative abundance tables for each sample at the phylum, genus, and where possible, to the species level. Although we have established that species-level classification is made more accurate by metagenomic methods such as PathoScope (Odom *et al.*, 2023), genus level classification is much more reliable, so we decided to focus only on the genus level.

## Data and statistical analysis

### *NP microbiome characteristics and evolution over time.*

We describe the normal evolution of the NP microbiome in healthy infants over the first three months of life. We calculated microbial richness using Chao1 index, and diversity of microbial taxa using the Shannon diversity index. We report the individual evolution of NP microbiome of each of the 10 infants who develop LRTI. In order to establish statistical significance, we used the *lmer* function from the *lme4* package for R (Bates *et al.*, 2015) to apply a mixed-effects linear model to the log counts per million (logCPM) value of each genus, including age and HIV exposure as fixed effects and the study subject as a random effect. All p-values generated by these linear models are reported after False Discovery Rate (FDR) adjustment for multiple comparisons using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). We only generated mixed-effects models for genera which had an average relative abundance of at least 0.5% across all healthy infant samples.

For visualization of the development of healthy NP microbiota, we grouped all infant samples by age (in days) into 7 bins, each comprising a 16-day age window (0–15 days, 16–31 days, etc). We only visualized genera which had an average relative abundance of at least 1% across all samples. The relative abundances of all genera which did not meet this threshold were summed into a group labelled "Other/Low abundance" for plotting purposes only.

We calculated estimates of the alpha diversity within each sample based on the species-level counts tables generated by PathoScope 2. We calculated alpha diversity using two methods: the Chao1 index, which estimates the total

number of species present within a sample, and the Shannon index, an entropy-based metric which incorporates both the number of species present and the evenness of abundance among those species. The Chao1 index was calculated using the R package *fossil* (Vavrek, 2011) and the Shannon index was calculated using the R package *vegan* (available via CRAN) (Oksanen *et al.*, 2019) each with a rarefaction depth of 10,000. We constructed a mixed-effects linear model as described above, except using each alpha diversity metric as a response variable, in order to test whether alpha diversity changed as infants aged.

Nasopharyngeal samples, particularly from young infants, have low DNA density, making them susceptible to contamination, which in our analysis could not be entirely eliminated (Salter *et al.*, 2014). Samples of all infants (both cases and comparators) were processed at the same time and under similar conditions, lowering the likelihood of contamination impacting the results.

### *Analysis of the association between the NP microbiome and the development of LRTI.*

We used the *lmer* function from the *lme4* package (described above) to build mixed-effects linear models to compare the development of the NP microbiomes of infants who developed LRTIs to those of healthy infants. This time, we included infection status and the interaction of infection status with age as fixed-effect covariates in addition to age and HIV exposure, as well as study subject as a random effect. Once again, p-values were generated using the *Anova* function of the *car* package (Fox & Weisberg, 2019) and then FDR corrected.

We similarly modified the models we had used to test alpha diversity in order to see if either Shannon or the Chao1 index values were different in LRTI infants, once again adding infection status and the interaction between infection status and age as fixed effects.

### *Differential abundance analysis at first timepoints.*

We performed differential abundance between the first samples from healthy and LRTI infants using the R package *DESeq2* (Love *et al.*, 2014) available via Bioconductor (Huber *et al.*, 2015). We imported our unnormalized genus counts table compiled from PathoScope2 as a *DESeqDataSet* and ran the function *DESeq*, using a design model that included infants' HIV exposure (from an HIV infected mother) as a covariate. For microbiome data, *DESeq2* has been shown to return lower false discovery rates than other differential tests (McMurdie & Holmes, 2014), and performs particularly well for smaller experiments (Weiss *et al.*, 2017).

To test whether the presence or absence of certain genera at the first sampled timepoint were associated with LRTI, we performed Fisher's exact test to determine if healthy and LRTI infants are equally likely to have each genus in their NP microbiome. Because very low-abundance genera could be the result of spurious alignments or contamination, we explored

both a high threshold (>1% relative abundance) and a low threshold (>0.1% relative abundance) for defining presence of a genus.

**Beta diversity and clustering.** We computed a Bray-Curtis dissimilarity matrix between samples using *vegan*'s `vegdist` function. When applied to relative abundance values, Bray-Curtis dissimilarity between two samples *i* and *j* is defined as  $BC_{ij} = 1 - \sum_{n=0}^N \min(g_{in}, g_{jn})$  where  $g_{in}$  is the relative abundance of genus *n* in sample *i*. In order to identify specific microbial profiles we performed hierarchical clustering of samples based on this dissimilarity matrix using R's `hclust` function with the method set to "ward.D". We defined clusters using R's `cutree` function, with the value for *k* selected by maximizing the Silhouette and Frey indexes as calculated by the package *NbClust* (Charrad *et al.*, 2014). For each cluster, we performed Fisher's exact tests to determine whether that cluster was enriched for LRTI samples generally, pre- LRTI samples, active LRTI samples, or HIV-exposed samples.

We used the `metaMDS` function from the R package *vegan* to perform non-metric multidimensional scaling (NMDS) ordination on our Bray-Cutris dissimilarity matrix, using as parameters `k=3`, `try=50`, and `trymax=1000`. Scaling our data onto just two dimensions using NMDS yielded a stress value greater than 0.2, indicating a poor fit; we instead scaled the data onto three dimensions (stress=0.13), and used the *vegan*'s `envfit` function to project the age and LRTI status of each sample into the NMDS ordination.

**Differential analysis of maternal NP microbiomes.** We used Spearman correlation coefficients between the mother and child at the genus level, on the first time point of sampling. We chose Spearman correlation, which utilizes rank order rather than continuous values, due to the compositional nature of bacterial abundance data. We calculated Spearman's  $\rho$  for the relative abundance of each genus between mothers and their infants. We tested the significant of these correlations by comparing the

distribution of  $\rho$  values to 1000 null distributions of the same metric, generated by randomly permuting the mother/infant labels.

We used DESeq2 to test for differential abundance of genera in the NP microbiomes of mothers of LRTI infants and mothers of control infants. For this analysis, we only included samples taken from mothers at the earliest pediatric visits, before their infants began exhibiting LRTI symptoms. We included the HIV status of the mothers as a covariate in DESeq2's regression model. We report p-values after FDR correction via Benjamini-Hochberg procedure, and consider adjusted p-values below 0.1 to be significant.

**Results**

With ten infants with LRTI and 3:1 matching, our analysis set consisted of 40 infants at ~seven time points each. All infants were born healthy via vaginal delivery. Male sex was more common in infants who developed LRTI ( $p= 0.067$ ). A third of infants with LRTI were born to mothers with HIV (receiving anti-retroviral treatment), compared to 40% of infants in the healthy group. Basic characteristics of the 40 infants are shown in Table 1. The symptoms and timing of sampling of the ten infants who developed LRTI are shown in Table 2.

**16S ribosomal DNA amplicon sequencing data and processing**

We successfully sequenced 265 NP swabs from 40 infants, capturing a median of seven samples from each infant. The median age at first sampling was seven days, and the median age at final sampling was 104 days. We also sequenced two NP swabs from each infant's mother at first and last time points, for a total of 345 samples from mothers and infants combined. In six of these samples, fewer than 10,000 reads aligned to RefSeq reference genomes and were excluded from further analysis. The remaining 339 samples had a median of 101,979 reads per sample assigned to reference genomes

**Table 1. Characteristics of healthy infants and infants with LRTI.**

Characteristics	Healthy Infants (N=30)	Infants with LRTI (N=10)	All Subjects (N=40)
Sex, n (%)			
Females	16 (53.3%)	2 (20.0%)	18 (45.0%)
Males	14 (46.7%)	8 (80%)	22 (55%)
Season of enrollment			
Dry Season (May–Oct), n (%)	28 (93.3%)	8 (80.0%)	36 (90.0%)
Rainy Season (Nov–Apr), n (%)	2 (6.7%)	2 (20.0%)	4 (10.0%)
Median age at enrollment in days (IQR)	7.0 (6 - 9)	7.0 (6 - 10)	7.0 (6 - 9)
HIV exposed, n (%)	13 (43.3%)	3 (30.0%)	16 (40.0%)
Mean number of samples collected (SE)	6.6 (0.2)	6.6 (0.6)	6.7 (0.2)

**Table 2. Clinical symptoms and age of 10 infants with LRTI at each study visit/NP sampling.**

Infants with LRTI	Sample Number and Infant age at sampling (days)								
	1	2	3	4	5	6	7	8	9
Infant 1	7 days	27 days	42 days	62 days	79 days				
Infant 2	7 days	27 days	35 days	42 days	59 days	73 days	88 days	104 days	
Infant 3	7 days	11 days	62 days						
Infant 4	7 days	19 days	45 days	60 days	68 days	107 days			
Infant 5	7 days	28 days	42 days	56 days	69 days	84 days	100 days		
Infant 6	7 days	21 days	42 days	56 days	59 days	73 days	87 days	96 days	103 days
Infant 7	7 days	50 days	59 days	73 days	87 days	106 days			
Infant 8	7 days	24 days	27 days	42 days	61 days	73 days	90 days	104 days	
Infant 9	7 days	24 days	39 days	44 days	65 days				
Infant 10	7 days	23 days	40 days	61 days	83 days	99 days	113 days		

- No symptoms
- Mild upper respiratory symptoms (cough/runny or blocked nose)
- Diagnosis of LRTI (cough/runny or blocked nose with or without fever AND fast breathing with indrawing of the chest)

and were included in the analysis. From these, we detected 421 unique genera, spanning 14 unique phyla, which were assigned at least 100 sequence reads across all samples. Based on these results, we were confident in our ability to proceed with the ensuing analyses.

**Analysis One: What is the NP microbiome maturation in healthy infants in the first three months of life?**

Given our ultimate goal of identifying characteristics of the NP microbiome in infants who develop LRTI, as a first step we describe the characteristics and development of NP microbiome of the healthy infants. We analyzed the NP samples from all the infants who remained free of any respiratory symptoms through the end of observation, using linear regression to track changes in relative abundance of genera over time spanning the period between enrollment after birth and 14 weeks of age. Since we used linear mixed models, with log counts/million to transform the data, the curvilinear relationship of the data has been accounted for statistically.

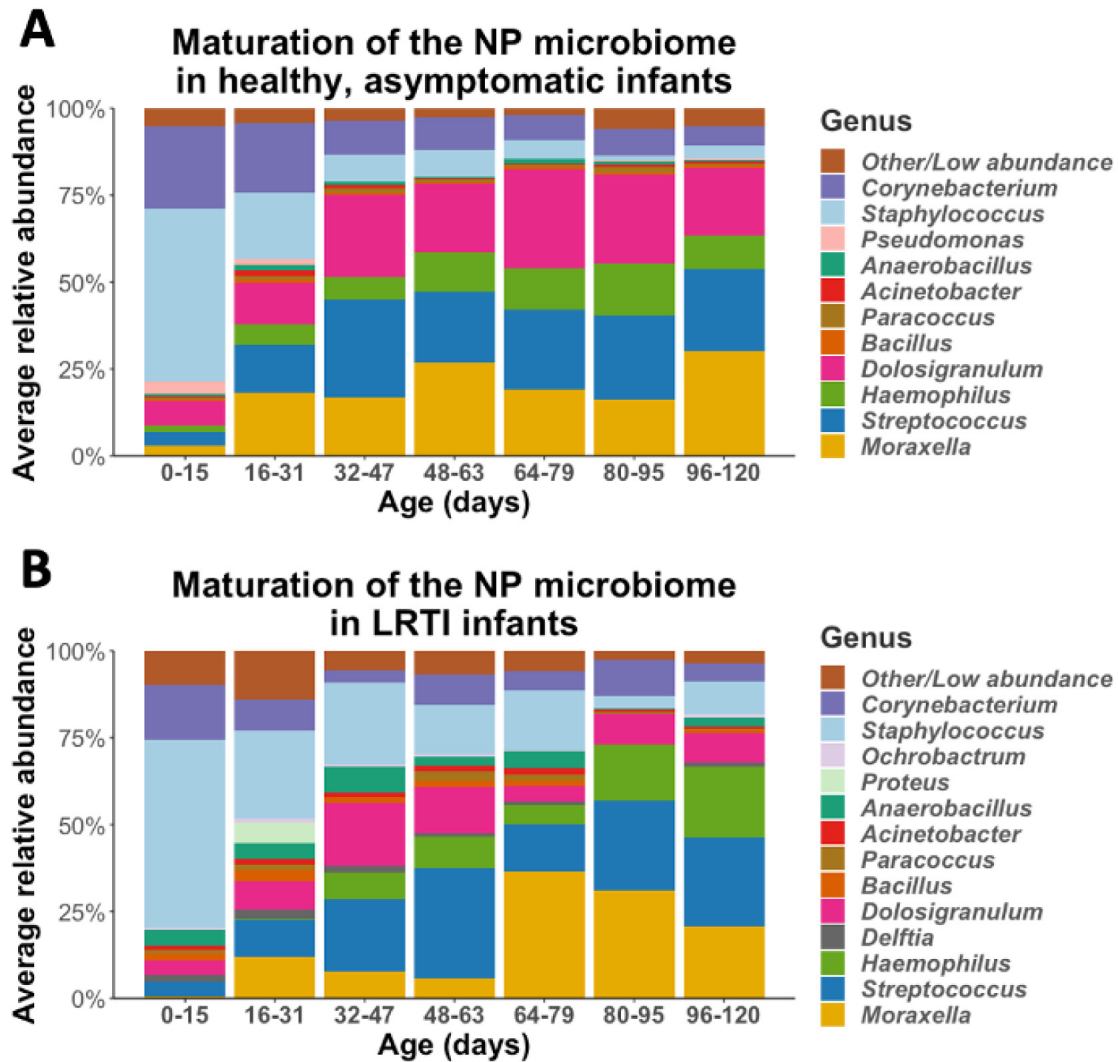
We observed a stepwise pattern of maturation as the infants aged, summarized in Figure 1a, showing the relative abundance of different genera across each age group. As can be seen, there is a clear shift in the abundance of dominant genera with time, with some dominating early in life, and others becoming more prominent as the children aged. Early in life, the dominant genera were *Staphylococcus* and *Corynebacterium*. According to a mixed-effects model, these genera declined in relative abundance as infants aged (*Staphylococcus*:  $p < 10E-7$ , *Corynebacterium*:  $p < 0.001$ ) and were replaced primarily by *Streptococcus* ( $p < 0.001$ ) *Dolosigranulum* ( $p < 0.001$ ), *Moraxella* ( $p < 0.001$ ), and *Haemophilus* ( $p = 0.02$ ).

We did not measure any significant change in the alpha diversity (richness within a given sample) of NP microbiomes as healthy infants aged, measured either by Shannon index ( $p = 0.32$ ) or Chao1 index ( $p = 0.15$ ). When we clustered samples based on beta diversity (between sample diversity), measured as the Bray-Curtis dissimilarity between pairs of samples, we identified a distinct profile associated with samples from very young infants that contrasted against several profiles for more mature infant NPs. While each cluster is dominated by one or several of the most common genera, very few samples from healthy infants had high abundance of genera outside of the six most prominent genera. The primary axis of a Principal Coordinate Analysis (PCoA) correlated with the age of the infants at the time of sampling, and stratified samples mainly by relative abundance of *Staphylococcus* and *Corynebacterium* in younger infants vs. the genera which were more common at older ages. The second PCoA axis distinguished between samples that were rich in *Moraxella* or *Dolosigranulum* from those rich in *Streptococcus* or *Haemophilus*. In summary, this analysis showed that the microbiomes of early infancy were highly dynamic over time, but that these shifts occurred in a structured and stereotypical pattern.

**Analysis Two: Does the maturation of the NP microbiome differ among infants who developed LRTI compared with healthy infants?**

Given evidence from prior literature that during LRTIs the NP microbiome of infants is different than that compared to healthy infants, we set out to test whether the general maturation pattern of the NP microbiome in the first months of life is altered in infants who went on to develop an LRTI. We repeated our analysis as described for healthy infants, stratifying into



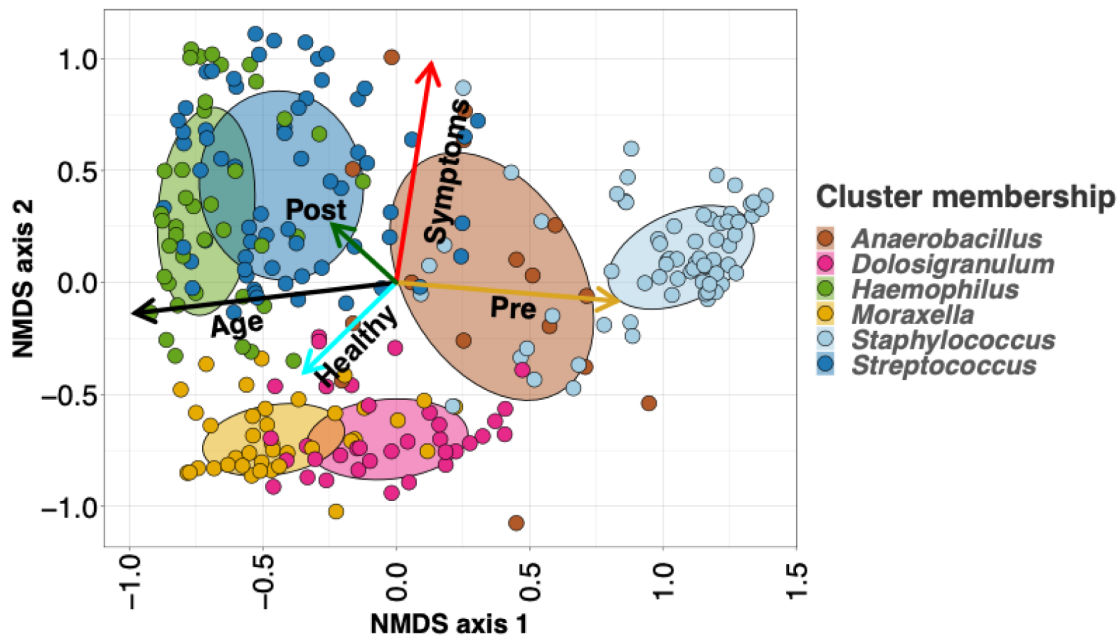


**Figure 1.** The maturation of the NP microbiomes of **A**) healthy, asymptomatic infants (n=30), and **B**) LRTI infants (n=10) over three months of observation. These stacked bar plots show the average relative abundance of the most common genera found in infant NPs, with samples binned by age.

age groups and mapping the evolution of the NP microbiome over the first three months of life (Figure 1b). Infants who developed LRTI had similar general succession patterns as described for healthy infants, with high relative abundance of *Staphylococcus* early in life replaced by relative abundance of *Streptococcus*, *Haemophilus*, *Corynebacterium*, *Dolosigranulum* and *Moraxella*. Even though the general succession pattern of NP microbiome in infants with LRTI was similar to succession patterns of healthy infants, they exhibited distinct characteristics. Notably, the NP microbiome of infants who developed LRTI had, on average, higher relative abundance of specific genera including *Bacillus* (p=0.05) and *Delftia* (p<0.001) and lower relative abundance of *Dolosigranulum* (p<0.001).

As with the healthy control infants in our analysis 1, we did not observe any change in alpha diversity in LRTI infants as they aged (Shannon: p=0.08, Chao1: p=0.74). Analysis of the beta diversity between LRTI infant samples once again revealed a cluster of samples taken at very early time points, dominated by *Staphylococcus*, with samples taken from older timepoints exhibiting profiles rich in *Streptococcus*, *Dolosigranulum*, *Moraxella*, and *Haemophilus*. However, in LRTI infants we also observed a large sixth cluster, characterized by a high abundance of *Anaerobacillus* as well as various other rare genera

Since each infant developed an LRTI at a different age, stratifying the infants into age groups resulted in grouping



**Figure 2. Nonmetric multidimensional scaling (NMDS) ordination plots of all infants' (n=40) nasopharyngeal (NP) samples.** We applied 3-dimensional NMDS ordination to the Bray-Curtis dissimilarity matrix between all infants' NP swabs, and projected vectors into that ordination space representing the best fit correlations for the age at sampling (the black arrows) and LRTI status (the cyan arrows represent control infants). Age is highly correlated with the first NMDS axis, and samples on the young end of the age vector mostly belong to the *Staphylococcus*-dominated profile, whereas samples on the older end tend to belong more to the *Haemophilus* and *Moraxella*-dominated profiles. The *Dolosigranulum*-dominated profile is associated with the healthy end of the vector for LRTI status, while the *Anaerobacillus*-dominated profile is associated with disease.

together infants at different time points in relation to their disease – before the LRTI, at time of the LRTI, and following the LRTI.

### Analysis three: *Can we identify specific characteristics of the NP microbiome that precede the development of LRTI?*

A key limitation of the previous analyses is that they present the average across each time point, and as the children age, more and more of the data in the LRTI group will represent an LRTI event or post LRTI timepoints. This is particularly important since all these children received antibiotics with their diagnosis, which will have obvious impacts on the microbiome. To address this source of confounding, we conducted several analyses. First, we compared the microbiomes at the baseline visit (at enrollment), which preceded all LRTI events when the infants were healthy, and before any antibiotic was given.

We analyzed the earliest NP samples taken from each infant at 7 days of age, comparing the microbiomes of those infants who eventually developed LRTIs to those who did not. At enrollment all infants were healthy by definition (based on enrollment inclusion/exclusion criteria), and therefore, infants who developed LRTI could collectively be grouped as “prior to infection” at that time point.

We identified three options by which a genus could be different between the 2 groups: First, a genus that was identified exclusively in infants who developed LRTI, such as *Novosphingobium* (4/10). Second, genera that were more common in infants with LRTI (but were present in both groups), such as *Delftia* (8/10 in LRTI infants vs 13/30 in healthy infants). And third, there were genera that were detected in both groups, but were present with higher relative abundance in infants with LRTI compared to the healthy infants, such as *Anaerobacillus*, *Bacillus*, *Blastococcus*, *Brachybacterium*, *Ochrobactrum*, *Ornithinimicrobium*, and *Sphingomonas*. Overall, ten genera were significantly different in infants who later developed LRTI at the first time point (Table 3). Notably, *Dolosigranulum*, which has been identified in prior studies as being associated with a healthy microbiome, as was the case among the healthy infants here, had significantly lower relative abundance in infants who developed LRTIs than in healthy counterparts prior to the LRTI and even at the first sample time point.

Nonmetric multidimensional scaling (NMDS) of the beta diversity dissimilarity matrix between all samples allows us to visualize more holistic structural differences in the NP microbial communities of healthy vs LRTI infants (Figure 2). When we project the age and eventual LRTI status of the infants

**Table 3. Differential abundance between control and LRTI infants at earliest observed timepoint.**

Genus	Log Foldchange	Frequency in control infants	Frequency in LRTI infants	Adjusted p-value
<i>Anaerobacillus</i>	2.66	70%	70%	0.013
<i>Bacillus</i>	2.54	60%	70%	<0.01
<i>Blastococcus</i>	5.36	0%	10%	<0.01
<i>Brachybacterium</i>	5.22	3%	30%	<0.01
<i>Delftia</i>	2.81	43%	80%	<0.01
<i>Dolosigranulum</i>	-4.14	57%	50%	<0.01
<i>Novosphingobium</i>	6.80	0%	40%	<0.01
<i>Ochrobactrum</i>	2.62	27%	60%	<0.01
<i>Ornithinimicrobium</i>	4.77	3%	20%	<0.01
<i>Sphingomonas</i>	2.72	17%	40%	<0.01

into the NMDS ordination space, we can see that age correlates closely the primary NMDS axis, whereas LRTI status is mostly correlated with the secondary, indicating differences in NP microbiomes between healthy and LRTI infants independent of the aging process.

When comparing specific genera abundance relative to the time of infection, i.e. comparing the time points preceding the LRTI (for the case infants) and the same time points for the control infants we confirmed lower relative abundance of *Dolosigranulum*, and higher relative abundance of *Anaerobacillus* in the LRTI infants before their infection. (Figure 3).

#### Analysis four: Are there distinct microbiome profiles that characterize sickness and health and other infant characteristics?

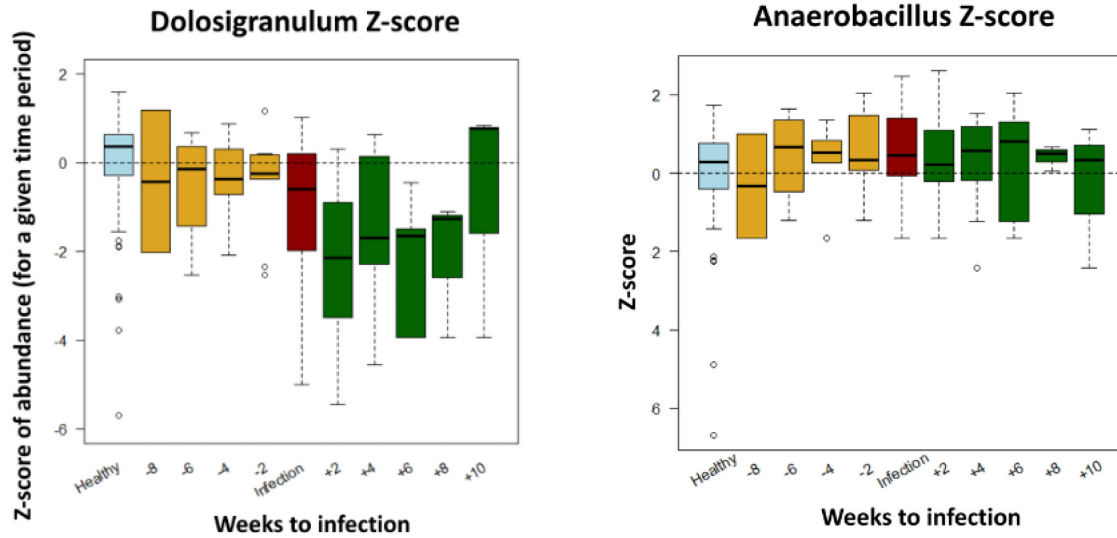
Using the Silhouette and Frey clustering indexes (NbClust) our samples were split into six primary clusters (Silhouette index) and 13 sub-clusters (Frey index). These six primary profiles were then named after the dominant genus within each cluster (the highest relative abundance genus). The Bray-Curtis dissimilarity matrix analysis yielded the following clusters (Figure 4): *Staphylococcus* dominant, *Streptococcus* dominant, *Moraxella* dominant, *Dolosigranulum* dominant, *Haemophilus* dominant, and *Anaerobacillus* dominant profiles, corresponding to six of the seven most abundant genera across all our samples, as shown in Table 4. *Corynebacterium* is the only highly-abundant genus that does not compose the majority (or plurality) of relative abundance within any cluster; instead of being dominant in a subset of samples, *Corynebacterium* often co-occurred alongside the more dominant *Staphylococcus*, or to a lesser extent *Dolosigranulum*. For ease of reporting, we shall henceforth refer to each cluster by its most abundant genus.

Figure 4 shows the microbial composition of each of the 262 infant samples in our study which passed sample quality filters, grouped by the six primary profiles (Figure 4A) and the 13 sub-profiles (Figure 4B).

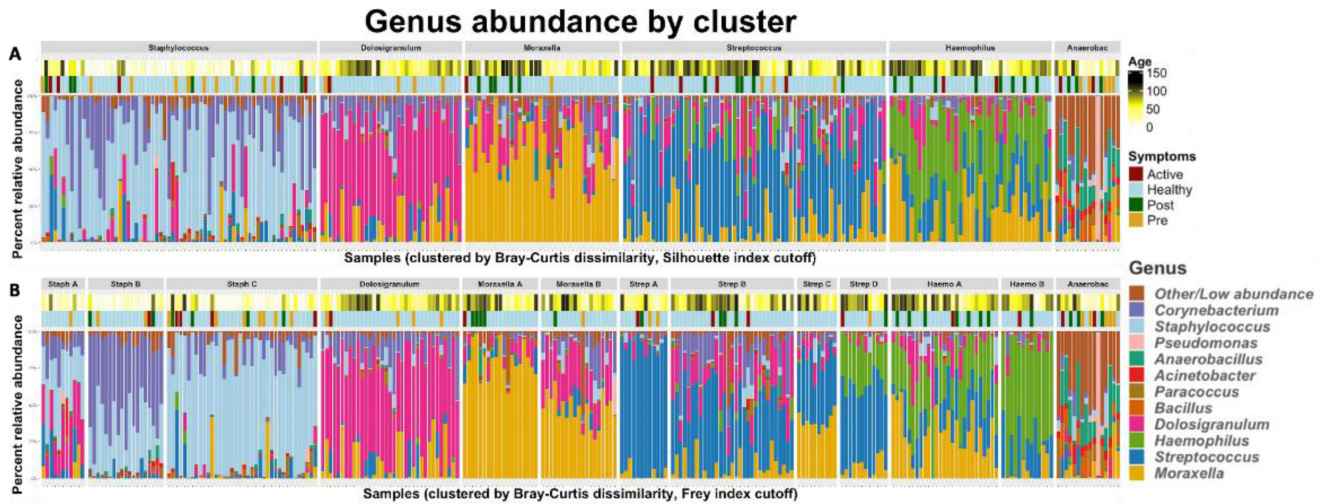
Fisher's exact tests revealed that the *Anaerobacillus* dominant profile was highly associated with infants who developed LRTIs, ( $p < 0.01$ , estimated odds-ratio=5.74). The *Staphylococcus* sub-profile Staph-C was associated with LRTI infants ( $p = 0.04$ , estimated odds-ratio=2.26), and the *Streptococcus* sub-cluster Strep-C (which is also rich in *Moraxella*) was associated with healthy infants ( $p = 0.07$ ). Using ANOVA to assess the association of each profile with age, the *Staphylococcus* dominant profile was clearly associated with samples from younger infants compared to all other profiles, and the *Anaerobacillus* dominant profile was associated with younger samples when compared to the *Haemophilus* and *Streptococcus* profiles (Table 5).

We visualized the association between LRTI status and NP taxonomic profiles using NMDS ordination (Figure 2). By projecting infants' LRTI status and age into the ordination space, we can see that the vector corresponding to healthy samples points towards the *Dolosigranulum* profile (and to a lesser extent towards the *Moraxella* profile), while the LRTI vector points towards the *Anaerobacillus* profile.

Together, these results reinforce a number of our previous observations; in particular, we can see that there is a general trend for infant NP microbiome profiles to shift from being *Staphylococcus* dominant shortly after birth towards several other profiles. We also see a clear pattern in the LRTI infants, comprising higher than normal relative abundance of *Anaerobacillus* as well as higher prevalence of rare genera which



**Figure 3.** Relative abundance Z-scores of specific genera of LRTI infants compared to healthy controls (light blue), by weeks from infection.



**Figure 4.** The taxonomic profiles of all infant NP samples (n=40), clustered by pairwise Bray-Curtis dissimilarity. Clusters were defined by performing hierarchical clustering on the beta diversity matrix and then cutting the resulting dendrogram into an optimal number of clusters according to the **A)** Silhouette index (6 clusters) and **B)** Frey index (13 clusters). The color bars above the stacked bar plots indicate the infants’ ages at the time of each sample and their LRTI status – “healthy” indicates an infant which did not develop LRTI symptoms during our observation.

typically make up an extremely low portion of (or are completely absent from) healthy NP microbiomes.

**Analysis five:** Is the NP microbiome of mothers of infants who develop LRTI different than mothers of healthy infants?

Observing distinct characteristics of the NP microbiome of infants as early as age 7 days, suggested that these profiles might be related to in-utero exposures, transmittable immunologic factors, and/or host genetics. That led us to question whether mothers of infants who develop LRTI have themselves

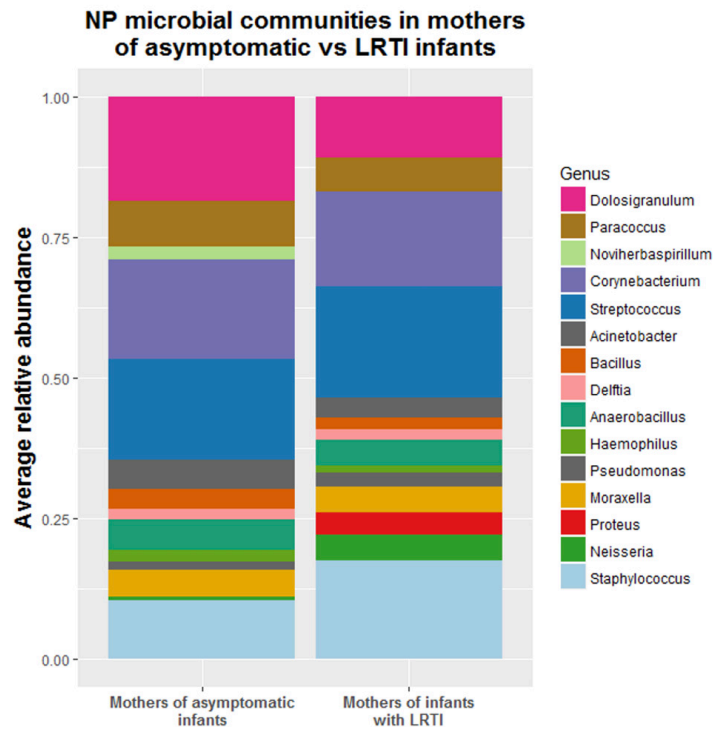
distinct characteristics of the NP microbiome. We analyzed the first NP swabs from each of the mothers enrolled in our study taken at the infants’ day seven enrollment visits, correlated their microbiomes to those of their infants, and used DESeq2 to establish which genera were differentially abundant between mothers of LRTI infants and mothers of healthy infants. Similar to the pattern seen in the infants themselves, the mothers of infants who developed LRTIs had significantly decreased relative abundance of *Dolosigranulum* (p=0.05) as compared to mothers of healthy infants at 7 days of infant’s life (Figure 5).

**Table 4. Relative abundance and frequency of the most common genera observed in the NP microbiome of healthy control and LRTI infants.**

Healthy infants				LRTI infants			
Genus	Mean Relative Abundance	Frequency (Samples)	Frequency (Subjects)	Genus	Mean Relative Abundance	Frequency (Samples)	Frequency (Subjects)
Streptococcus	19.4%	93.4%	100%	Staphylococcus	22.1%	90.8%	100%
Dolosigranulum	19.1%	84.3%	100%	Streptococcus	18.8%	93.8%	100%
Moraxella	18.3%	77.8%	100%	Moraxella	14.8%	80%	100%
Staphylococcus	14.0%	88.9%	100%	Dolosigranulum	9.6%	58.5%	100%
Corynebacterium	12.0%	94.4%	100%	Corynebacterium	8.3%	86.1%	100%
Paracoccus	0.95%	66.7%	100%	Anaerobacillus	3.9%	58.5%	100%
Acinetobacter	0.84%	66.2%	100%	Bacillus	1.7%	64.6%	100%
Bacillus	0.55%	52.5%	100%	Delftia	1.5%	63.1%	100%
Anaerobacillus	1.1%	53.0%	96.7%	Acinetobacter	1.4%	73.8%	100%
Pseudomonas	0.89%	38.9%	93.3%	Pseudomonas	0.43%	46.2%	100%
Delftia	0.32%	37.9%	90%	Paracoccus	1.1%	52.3%	90%
Aeromonas	0.13%	26.3%	80%	Ochrobactrum	0.64%	43.1%	90%
Haemophilus	8.7%	38.4%	73.3%	Haemophilus	8.0%	38.5%	80%
Kocuria	0.11%	18.7%	63.3%	Novosphingobium	0.37%	29.2%	80%
Ochrobactrum	0.16%	16.2%	53.3%	Kocuria	0.30%	26.2%	80%
Escherichia	0.25%	8.1%	40%	Sphingomonas	0.38%	30.8%	70%
Enterobacter	0.18%	9.6%	40%	Aeromonas	0.17%	21.5%	60%
Klebsiella	0.12%	6.6%	30%	Janibacter	0.13%	12.3%	60%
				Brachybacterium	0.11%	23.1%	60%
				Agrobacterium	0.20%	15.4%	50%
				Veillonella	0.19%	15.4%	50%
				Cutibacterium	0.15%	12.3%	50%
				Stenotrophomonas	0.14%	13.8%	50%
				Halolactibacillus	0.11%	15.4%	50%
				Proteus	0.92%	9.2%	40%
				Nocardioides	0.12%	13.8%	40%
				Variovorax	0.11%	12.3%	40%
				Klebsiella	0.18%	7.7%	30%
				Marmoricola	0.18%	9.2%	20%
				Blastococcus	0.14%	6.2%	20%
				Knoellia	0.11%	7.7%	20%
				Anaerococcus	0.10%	6.2%	10%

**Table 5. Associations between NP microbiome profiles with LRTI status and age.**

Associations with LRTI status			
Cluster/Subcluster	Odds ratio estimate	Odds ratio range	P-value
Anaerobacillus	5.74	1.80-20.11	<0.01
Staphylococcus C	2.26	1.02-4.92	0.04
Streptococcus C	0.00	0.00-1.34	0.07
Associations with age (in days)			
Cluster	Cluster	Difference in age	Adjusted P-value
Staphylococcus	Moraxella	39	<0.01
Staphylococcus	Dolosigranulum	52	<0.01
Staphylococcus	Streptococcus	41	<0.01
Staphylococcus	Haemophilus	44	<0.01
Staphylococcus	Anaerobacillus	21	0.05
Anaerobacillus	Streptococcus	20	0.1
Anaerobacillus	Haemophilus	24	0.05



**Figure 5. Stacked bar plots showing the average relative abundance of the most common genera found in mothers NPs at first time point.**

## Discussion

In this analysis, we show that the NP microbiome of infants with LRTI differs from that of healthy infants and that there is evidence suggesting dysbiosis precedes the onset of LRTI. Intriguingly, we observed different microbiome patterns in the mothers of infants who later developed LRTI and those whose children remained healthy. That, and the fact that the microbiome of mother-infant pairs is more closely correlated within pairs than across pairs, suggests that some of the infant dysbiosis has transgenerational origins. As an overall synthesis, our data suggest that there are quantitative and qualitative differences between infants (and their mothers) who do and do not develop LRTI. This supports the hypothesis that LRTI is not a random event, but rather may reflect predispositions that are generally unobserved but may nonetheless play an essential or contributory role in the pathogenesis of childhood LRTI.

The nasopharynx is the ecologic niche of respiratory pathobionts, and in this ecosystem they will either become invasive or remain merely colonizers. The characteristics of the NP microbiome at time of infection is associated with LRTI and its severity. But there is also good reason to believe that the maturation of the NP microbiome in the first months of life, and not only its characteristics at time of infection, is associated with respiratory health and development of disease later in life. For example, maturation of the gut microbiome is known to regulate the immune system evolution and is associated with the development of diseases later in life such as obesity and type 1 diabetes (Bokulich *et al.*, 2016; Stewart *et al.*, 2018). Gut microbial dysbiosis in children often predisposes to recurrent *C. difficile* infections (Ihekweazu & Versalovic, 2018). Thus, a similar association between the NP microbiome and risk of respiratory infections is a plausible theory for which there is precedent. This has been shown to be true for mild respiratory infections, and we assume this is also true for lower respiratory tract infections. (Bosch *et al.*, 2017; de Steenhuijsen Piters *et al.*, 2022; Teo *et al.*, 2015, Teo *et al.*, 2018)

We have characterized the normal, healthy maturation of the NP microbiome over the first months of life, and showed how this maturation is different in infants who develop early LRTI. While the evolution of the normal microbiome is highly dynamic, it proceeds in a stereotypical fashion, with stepwise shifts from a flora dominated by skin organisms (*Staphylococcus*), to one that is dominated by genera more typically associated with the respiratory tract (*Dolosigranulum*, *Streptococcus*, *Haemophilus* and *Moraxella*). Similar microbial succession patterns were previously described in other birth cohorts (Biesbroek *et al.*, 2014). Importantly, our data describing the maturation of the NP microbiome of infants in a low-middle-income country adds to what's currently known from developed countries around the world.

By contrast, infants who develop LRTI have similar general succession patterns as healthy infants; transitioning from high relative abundance of *Staphylococcus* to high relative

abundance of genera associated with the respiratory tract, but unlike healthy infants, the evolution of their NP microbiome is characterized by low relative abundance of specific genera associated with 'health', such as *Dolosigranulum*, and high relative abundance of other genera that appear unique, such as *Anaerobacillus*, *Bacillus*, and a mixture of 'other' uncommon genera. Additionally, the LRTI infants' microbiomes include a larger number of uncommon and transient genera, presenting a picture that is more chaotic than what is seen in the healthy infants. Since the maturation analysis included post-LRTI samples, the differences observed in the NP microbiome maturation of LRTI infants could also be attributed to antibiotics and not only to the LRTI itself.

Case-control studies have consistently demonstrated an association between NP microbiome characteristics and LRTIs at time of disease, though interpretation in terms of causality could not be shown. The relatively high abundance of *Dolosigranulum/Corynebacterium* and *Moraxella* are correlated with healthy states (Mansbach *et al.*, 2016), whereas NP microbiomes enriched with *Streptococcus* and *Haemophilus* are associated with LRTI and also correlate with severity of disease (de Steenhuijsen Piters *et al.*, 2016; Kelly *et al.*, 2017). But are these microbial profiles a result of the infection? Or were they present before the infection?

We were able to identify several microbiome profiles which appear to cluster by chronological age, LRTI and health. Our results indicate that young infants who developed LRTI, had NP microbiome dysbiosis prior to acquiring the infection, and as early as 7 days of life. These infants have NP microbiome enriched with *Aneorobacillus/Bacillus*, *Acinetobacter*, and other uncommon/unspecified genera, and also have relatively lower abundance of *Dolosigranulum*. Our intriguing results suggest that their mothers NP microbiome at the same early time point also differed from that of mothers of healthy infants.

The interaction between host, microbiome and pathobionts is complex and most probably multidirectional. The NP microbiome, known to be associated with environmental factors (breastfeeding, mode of delivery) (Bosch *et al.*, 2017; Brugger *et al.*, 2016) could also very well be a reflection or marker of host genetics and immune system function, which would explain why so early in life "high risk" profiles are observed. New acquisition of a pathobiont in the nasopharynx initiates interactions between the pathobiont and other organisms residing in the nasopharynx. These interactions modify metabolic activity and gene expressions of the pathobiont that influence whether the pathobiont becomes invasive. The interactions themselves between organisms in the nasopharynx also modify host immune response which underscores the complex relationship between host, microbiome and pathogens (de Steenhuijsen Piters *et al.*, 2019).

The key unresolved question is what role dysbiosis plays in the causal pathway leading to LRTI: is dysbiosis a marker of other unobserved forces that lead to LRTI, such as underlying host genetic or immunologic factors? Or does dysbiosis play

a role in the causal pathway leading to LRTI? While our data cannot resolve this question, the implication of our findings are substantial. Our findings suggest that distinct NP microbiome characteristics identified in the first days of life are associated with higher risk of developing LRTI in early infancy. This suggests that there is an important window of opportunity for identifying these infants and intervene. According to our findings, it may even be that we can identify these infants, by examining the mothers.

Our study has several limitations. Infants were followed until the age of three months, and thus our findings could not be generalized to older age groups. On the other hand, it is possible that infants included in our healthy control group developed LRTI after the study period, in that case our results are biased towards the null, possibly underestimating differences between the two groups. Since microbiome analysis was done retrospectively on an existing library of samples, our ability to have appropriate control analysis was limited, and thus we cannot completely exclude the possibility of contamination in our samples.

A further limitation is that we do not know the causative pathogen of the LRTIs, and whether these were viral, bacterial, or mixed pathogen LRTIs. LRTI is a heterogeneous set of conditions, and it is plausible that dysbiosis can interact in pathogen-specific ways. The diagnosis of LRTI was based only on clinical data. Even though different pathogens interact in different ways with the NP microbiome and the host immune system, our data suggests that there is a common NP microbiome risk profile, regardless of the causative pathogen. Lastly, while our analysis included a very large number of longitudinal samples, our sample size only included 10 infants who developed LRTI (by our conservative definition). However, LRTI is a comparatively rare event and requires longitudinal surveillance of thousands of subjects over an extended period to identify even a few cases, which accounts for the paucity of research on this topic. Logistically, it is immensely challenging and resource intense to create and sample a cohort in the way we have done. Nonetheless, further research will be needed to confirm or refine these initial observations. If confirmed, these findings are not only critical to our understanding of factors that lead to the development of LRTI, and why one infant develops an LRTI while others do not, it also suggests that we have a window of opportunity to identify these “at-risk” infants before their infection, and to potentially intervene. These prevention measures could have a high impact on decreasing burden of LRTI in infancy.

## Conclusions

Specific characteristics of the NP microbiome in infants may precede LRTIs, suggesting at minimum a signal of infants at higher risk for LRTIs, and possibly a causative role in the development of these infections. Specific NP microbiome profiles which could be identified perinatally and appear to be associated with a higher risk of developing LRTIs in early infancy, present a potential window of opportunity for interventions. Our findings should be confirmed by large scale longitudinal studies.

## Declarations

### Ethics approval and consent to participate

The study was approved by the ethical review committees at the ERES Converge IRB in Lusaka, Zambia, and at Boston University Medical Center. All mothers provided written informed consent, with consent provided in English, Bemba or Nyanja as preferred by the participant.

## Data availability

### Underlying data

GitHub. Infant\_Nasopharyngeal\_Dysbiosis. DOI: [https://github.com/tfaits/Infant\\_Nasopharyngeal\\_Dysbiosis](https://github.com/tfaits/Infant_Nasopharyngeal_Dysbiosis)

This project contains the following underlying data:

- All code, processed data, and the sample information metadata
- Taxon counts tables are called "species.RDS", "genus.RDS", and "phylum.RDS". For strain/subspecies-level counts, "PathoScopeTable.txt" has the unfiltered/unprocessed outputs from PathoScope.

The raw and processed sequencing data from this study are available in the SRA repository, under NIH Sequence Read Archive, BioProject: PRJNA817266.

### Extended data

Harvard Dataverse. Nasopharyngeal Dysbiosis Precedes the Development of Lower Respiratory Tract Infections in Young Infants, a Longitudinal Infant Cohort Study. DOI: <https://doi.org/10.7910/DVN/BWGTEQ> (Lapidot, 2022)

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

## References

Andrews S: **FastQC**. Babraham Bioinforma, 2010.  
 Bates D, Mächler M, Bolker BM, *et al.*: **Fitting linear mixed-effects models using lme4**. *J Stat Softw.* 2015; **67**(1): 1–48.  
[Publisher Full Text](#)

Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *J R Stat Soc Ser B.* 1995; **57**(1): 289–300.  
[Reference Source](#)



- Biesbroek G, Bosch AATM, Wang X, *et al.*: **The impact of breastfeeding on nasopharyngeal microbial communities in infants.** *Am J Respir Crit Care Med.* 2014; **190**(3): 298–308.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bokulich NA, Chung J, Battaglia T, *et al.*: **Antibiotics, birth mode, and diet shape microbiome maturation during early life.** *Sci Transl Med.* 2016; **8**(343): 343ra82.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**(15): 2114–2120.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bosch AATM, De Steenhuijsen Piters WAA, Van Houten MA, *et al.*: **Maturation of the infant respiratory microbiota, environmental drivers, and health consequences. A Prospective Cohort Study.** *Am J Respir Crit Care Med.* 2017; **196**(12): 1582–1590.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Brugger SD, Bomar L, Lemon KP: **Commensal-Pathogen Interactions along the Human Nasal Passages.** *PLoS Pathog.* 2016; **12**(7): e1005633.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cao B, Ho J, Retno Mahanani W, *et al.*: **London School of Hygiene & Tropical Medicine Trevor Croft.** The Demographic and Health Surveys (DHS) Program, ICF, 2019.
- Charrad M, Ghazzali N, Boiteau V, *et al.*: **Nbclust: An R package for determining the relevant number of clusters in a data set.** *J Stat Softw.* 2014; **61**(6): 1–36.  
[Publisher Full Text](#)
- de Steenhuijsen Piters WAA, Heinonen S, Hasrat R, *et al.*: **Nasopharyngeal Microbiota, Host Transcriptome, and Disease Severity in Children with Respiratory Syncytial Virus Infection.** *Am J Respir Crit Care Med.* 2016; **194**(9): 1104–1115.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- de Steenhuijsen Piters WAA, Jochems SP, Mitsi E, *et al.*: **Interaction between the nasal microbiota and *S. pneumoniae* in the context of live-attenuated influenza vaccine.** *Nat Commun.* 2019; **10**(1): 2981.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- de Steenhuijsen Piters WAA, Sanders EAM, Bogaert D: **The role of the local microbial ecosystem in respiratory health and disease.** *Philos Trans R Soc B Biol Sci.* 2015; **370**(1675): 20140294.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- de Steenhuijsen Piters WAA, Watson RL, de Koff EM, *et al.*: **Early-life viral infections are associated with disadvantageous immune and microbiota profiles and recurrent respiratory infections.** *Nat Microbiol.* 2022; **7**(2): 224–237.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Fischer Walker CL, Rudan I, Liu L, *et al.*: **Global burden of childhood pneumonia and diarrhoea.** *Lancet.* 2013; **381**(9875): 1405–1416.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Fox J, Weisberg S: **An R Companion to Applied Regression.** Third. ed. Thousand Oaks (CA): SAGE Publications, 2019.  
[Reference Source](#)
- Fujiogi M, Raita Y, Pérez-Losada M, *et al.*: **Integrated relationship of nasopharyngeal airway host response and microbiome associates with bronchiolitis severity.** *Nat Commun.* 2022; **13**(1): 4970.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gill CJ, Mwananyanda L, MacLeod W, *et al.*: **Incidence of severe and nonsevere pertussis among HIV-exposed and -unexposed zambian infants through 14 weeks of age: Results from the southern Africa mother infant pertussis study (samips), a longitudinal birth cohort study.** *Clin Infect Dis.* 2016; **63**(Suppl 4): S154–S164.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hasegawa K, Linnemann RW, Mansbach JM, *et al.*: **Nasal Airway Microbiota Profile and Severe Bronchiolitis in Infants: A Case-control Study.** *Pediatr Infect Dis J.* 2017; **36**(11): 1044–1051.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hong C, Manimaran S, Shen Y, *et al.*: **PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples.** *Microbiome.* 2014; **2**: 33.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Huber W, Carey VJ, Gentleman R, *et al.*: **Orchestrating high-throughput genomic analysis with Bioconductor.** *HHS Public Access. Nat Methods.* 2015; **12**(2): 115–121.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ihekweazu FD, Versalovic J: **Development of the Pediatric Gut Microbiome: Impact on Health and Disease.** *Am J Med Sci.* 2018; **356**(5): 413–423.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kelly MS, Surette MG, Smieja M, *et al.*: **The Nasopharyngeal Microbiota of Children with Respiratory Infections in Botswana.** *Pediatr Infect Dis J.* 2017; **36**(9): e211–e218.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Klindworth A, Pruesse E, Schweer T, *et al.*: **Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.** *Nucleic Acids Res.* 2013; **41**(1): e1.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lapidot R: **Nasopharyngeal Dysbiosis Precedes the Development of Lower Respiratory Tract Infections in Young Infants, a Longitudinal Infant Cohort Study.** Harvard Dataverse, V1, 2022.  
<http://www.doi.org/10.7910/DVNBWGTEQ>
- Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 550.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Man WH, de Steenhuijsen Piters WAA, Bogaert D: **The microbiota of the respiratory tract: Gatekeeper to respiratory health.** *Nat Rev Microbiol.* 2017; **15**(5): 259–270.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mansbach JM, Hasegawa K, Henke DM, *et al.*: **Respiratory syncytial virus and rhinovirus severe bronchiolitis are associated with distinct nasopharyngeal microbiota.** *J Allergy Clin Immunol.* 2016; **137**(6): 1909–1913. e4.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McMurdie PJ, Holmes S: **Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.** *PLoS Comput Biol.* 2014; **10**(4): e1003531.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Odom AR, Faits T, Castro-Nallar E, *et al.*: **Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data.** *Sci Rep.* 2023; **13**(1): 13957.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Oksanen J, Blanchet FG, Friendly M, *et al.*: **Package “vegan”.** *Community Ecol Packag.* 2019; **2**: 1–297.
- Revised WHO Classification and Treatment of Pneumonia in Children at Health Facilities: Evidence Summaries.** Geneva: World Health Organization, 2014.  
[Reference Source](#)
- Salter SJ, Cox MJ, Turek EM, *et al.*: **Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.** *BMC Biol.* 2014; **12**: 87.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Stewart CJ, Ajami NJ, O'Brien JL, *et al.*: **Temporal development of the gut microbiome in early childhood from the TEDDY study.** *Nature.* 2018; **562**(7728): 583–588.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Stewart CJ, Mansbach JM, Wong MC, *et al.*: **Associations of nasopharyngeal metabolome and microbiome with severity among infants with bronchiolitis. A multiomic analysis.** *Am J Respir Crit Care Med.* 2017; **196**(7): 882–891.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Teo SM, Mok D, Pham K, *et al.*: **The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development.** *Cell Host Microbe.* 2015; **17**(5): 704–715.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Teo SM, Tang HHF, Mok D, *et al.*: **Airway Microbiota Dynamics Uncover a Critical Window for Interplay of Pathogenic Bacteria and Allergy in Childhood Respiratory Disease.** *Cell Host Microbe.* 2018; **24**(3): 341–352. e5.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Vavrek MJ: **fossil: Palaeoecological and palaeogeographical analysis tools.** *Palaeontol Electron.* 2011; **14**(1): 16.  
[Reference Source](#)
- Weiss S, Xu ZZ, Peddada S, *et al.*: **Normalization and microbial differential abundance strategies depend upon data characteristics.** *Microbiome.* 2017; **5**(1): 27.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

## Open Peer Review

Current Peer Review Status: ? ✓ ?

---

### Version 2

Reviewer Report 23 July 2024

<https://doi.org/10.21956/gatesopenres.16786.r36205>

© 2024 De Steenhuijsen Piters W. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Wouter A.A. De Steenhuijsen Piters** 

<sup>1</sup> Department of Paediatric Immunology and Infectious Diseases, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, The Netherlands

<sup>2</sup> Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

We recognize the significance and novelty of the study, where the authors study the nasopharyngeal microbiota of Zambian infants, particularly those suffering from lower respiratory tract infections (LRTIs), in a longitudinal context. However, as was also discussed in our initial review, we still have two major concerns, which we believe are unsatisfactorily addressed.

First and foremost, we fully agree with the concerns raised by the Reviewer 3 (Carter Merenstein) regarding potential contamination issues. The absence of laboratory controls (especially DNA isolation negative controls) raises doubts about the validity of the lab procedures and possibility of contamination. Additionally, the authors do not report on a measure for bacterial biomass, which could have provided crucial insights in the extent of contamination. This issue is of relevance since, as also Reviewer 3 notes, some of the genera described to be associated with LRTIs (including those making up the disease-associated *Anaerobacillus*-cluster) have not been described as part of the nasopharyngeal microbiome previously, instead being mentioned as contaminants in literature. This may also be the case for *Delftia*, *Bacillus*, *Paracoccus*, *Acinetobacter*, *Proteus* and *Ochrobactrum*.

As said, we believe the study is of interest for the reasons stated in the introduction of our review. At the same time, we recognize this is a retrospective microbiome-study, where samples that were previously collected for a different purpose were sequenced at a later point in time, which brings many technical challenges, some of which cannot be amended. While the authors briefly acknowledge contamination as a limitation in their discussion (in a rather imprecise statement), we believe its importance has not been adequately emphasized throughout the article and - importantly - may not be clear to novice readers. At the very least, the likelihood of contamination should be underscored more prominently, particularly in the methods, results and discussion sections (see also major remark 1 in our initial review).

Our second concern revolves around the title and analysis three. The title appears to derive from the results of analysis three, depicting the fluctuation in the relative abundance of *Dolosigranulum* and *Anaerobacillus* before (see notes on *Anaerobacillus* above), during and after LRTI infections. We would like to reiterate our previous remarks (major remark 2 and minor remark 42) regarding the lack of statistical support for these findings and the failure to account for the influence of age (especially when considering the *Dolosigranulum* dynamics). Given the absence of statistical validation and the inability to adjust for age in analysis three, we argue that the title does not appropriately capture the contents of the manuscript. Here, we feel that the authors did not sufficiently address our previous remarks and would urge them to reconsider addressing these concerns.

Minor comments:

- Still unclear if PathoScope2 has been validated in the context of 16S-microbiota analyses
- Still unclear why the authors did not look at ASV/OTU level results
- Still use “evolution” instead of dynamics or development
- “Samples of all infants (both cases and comparators) were processed at the same time and under similar conditions, lowering the likelihood of contamination impacting the results” à processing the samples together does not necessary combat contamination; this merely standardizes the environmental factors during sample handling and processing.
- The genera in table 3 are all common contaminants
- It is unclear how the genera in Table 4 were selected? Top? I would suggest to include this in the table caption

Apart from these minor comments, there is also a significant number of (mostly minor) comments in our previous review the authors were not yet able to address. We would again urge them to critically go through our previous comments and see what can be done to take away our concerns/answer our questions.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Our group's expertise is on early-life development of respiratory microbiota in the context of health and disease. Both reviewers specialise in microbiota data analysis.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Reviewer Report 13 April 2024

<https://doi.org/10.21956/gatesopenres.16786.r36206>

© 2024 Scannapieco F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Frank A Scannapieco** 

University at Buffalo, NY, USA

No further comments.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Oral microbiology; Relationships between oral disease and systemic disease.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 12 April 2024

<https://doi.org/10.21956/gatesopenres.16786.r36204>

© 2024 Merenstein C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Carter Merenstein** 

University of Pennsylvania, Pennsylvania, USA

This study leverages a prospective longitudinal cohort to identify changes in the nasopharyngeal microbiome that may precede the development of a lower respiratory tract infection (LRTI) in infants under 3 months. The prospective nature of this cohort is valuable, and provided the unique opportunity to separate causes and effects of LRTI development. However, as discussed in my previous review, the analysis of these low-biomass samples leads to some ambiguous results.

The authors mention that negative controls had extremely low read counts, but do not quantify this or mention what taxa these few reads came from. Even if the counts are low, if they are from the same taxa as our found in other samples, it might be informative. Additionally, as mentioned previously, these negative controls should be included in the SRA upload for the sake of transparency. Finally, the nature of the negative controls, namely that they are PCR controls rather than blank swabs from the same site as the collection, run through the same DNA extraction kit, limits their utility. It is still entirely possible that contamination is present upstream of their negative controls.

This does not eliminate the value of this study, but should widen the interpretation of some of their findings. Namely, they find an increase in rare taxa in the samples preceding LRTI, and ascribe this to dysbiosis in the nasopharynx. An equally likely interpretation is that LRTI is preceded by reduced bacterial biomass in the nasopharynx, resulting in a higher proportion of sequencing reads coming from contamination. If the authors had information on the biomass of these samples (e.g. qPCR data, or even the DNA concentration post-amplification) it might address whether this phenotype of rare taxa is actually a result of lowered biomass. Otherwise, they should make this interpretation clear.

Specific revisions:

1. Upload all sequencing data to SRA. Presently only 172 samples are public, which excludes the negative controls and the swabs from the mother.
2. Revise language around the findings from analysis three and four to include the possibility that LRTI samples are lower biomass and have a higher proportion of contamination
3. Explicitly mention the lack of extraction and collection blanks as a limitation. The discussion mentions "our ability to have appropriate control analysis was limited" but this could be clarified for readers not familiar with low-biomass analysis.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Microbiome, respiratory tract microbiome

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

---

### Version 1

Reviewer Report 26 September 2023

<https://doi.org/10.21956/gatesopenres.14828.r34801>

© 2023 Merenstein C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Carter Merenstein** 

University of Pennsylvania, Pennsylvania, USA

Major comments:

1. The authors mention sequencing negative controls, but do not include these in any analysis. It would be useful to have some mention of how they were used, and whether there was any overlap between negative controls and experimental samples.
2. Relatedly, I can't find the negative controls in the SRA upload. These need to be included in the bioproject so that future researchers can also account for potential background.

Minor comments:

1. Figure 3 could use significance bars for clarity. From the surrounding context, the p values in Table 3 must apply to Figure 3, but by eye they don't seem to match (i.e. in the table *Anaerobacillus* has a positive LogFC, but the earliest time point in the boxplot seems to be equal or below the healthy control). Significance bars would highlight which groups and which timepoints are actually being compared here.
2. As mentioned in major point 1, Figure 4 would be strengthened by inclusion of the negative

control samples to ensure that none of these profiles actually represent the background of the extraction process.

3. The two sections of table 4 (Healthy Infants and LRTI infants) should be presented side by side, rather than one on top of the other, to allow for easy comparison across by taxa. This table also could likely be in the extended data, but that may be more up to the editor.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Partly

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Microbiome, respiratory tract microbiome

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 02 Mar 2024

**Rotem Lapidot**

We greatly appreciate the comments and advice received from the peer reviewer. We have attempted to address each of the issues raised and provide below a point-by-point summary of our responses. In each case we provide the comment verbatim, followed by our responses, heralded by '\*\*\*' and in italics.

Major comments:

The authors mention sequencing negative controls, but do not include these in any analysis. It would be useful to have some mention of how they were used, and whether there was any overlap between negative controls and experimental samples.

Relatedly, I can't find the negative controls in the SRA upload. These need to be included in the bioproject so that future researchers can also account for potential background.

*\*\*\*We thank the reviewer for this important comment. Regrettably, due to very low sequencing reads the control samples were not analyzed further and are no longer available to us. This compromises our ability to control for contamination, and we have addressed this both in the limitations section in the manuscript and in our response to reviewers #1.*

Minor comments:

Figure 3 could use significance bars for clarity. From the surrounding context, the p values in Table 3 must apply to Figure 3, but by eye they don't seem to match (i.e. in the table *Anaerobacillus* has a positive LogFC, but the earliest time point in the boxplot seems to be equal or below the healthy control). Significance bars would highlight which groups and which timepoints are actually being compared here.

*Figure 3 shows the relative abundance compared to time of infection. For some infants 8 weeks prior to infection would be the first sample taken, and for other infants, the first sample taken would be 2 weeks prior to the LRTI (and their first sample would be represented in the fourth yellow boxplot). The first boxplot of LRTI infants is the average of infants who had an LRTI 8 weeks after their first sampling and therefor does not match the results in table 3 which provides results of first sample of all 10 LRTI infants.*

As mentioned in major point 1, Figure 4 would be strengthened by inclusion of the negative control samples to ensure that none of these profiles actually represent the background of the extraction process.

*\*\*\*We agree, but unfortunately do not have the data to be included in the figure.*

The two sections of table 4 (Healthy Infants and LRTI infants) should be presented side by side, rather than one on top of the other, to allow for easy comparison across by taxa. This table also could likely be in the extended data, but that may be more up to the editor.

*\*\*\*We thank the reviewer for the suggestion and agree that these tables should be presented side by side.*

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 26 September 2023

<https://doi.org/10.21956/gatesopenres.14828.r34798>

© 2023 Scannapieco F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Frank A Scannapieco** 

University at Buffalo, NY, USA

The authors analyzed nasopharyngeal (NP) samples from a longitudinal, prospective cohort study of 1,981 Zambian mother-infant pairs who underwent NP sampling from 1-week through 14-weeks of age at 2-3-week intervals. The idea was to determine if substantial differences were apparent between the NP microbiome of children destined for a lower respiratory tract infection (LRTI) (N=10) when compared to normal infants (N=30), from essentially birth to 14 weeks. The microbiome was compared using 16S rRNA gene sequencing on the samples, as well as from baseline samples of the infants' mothers.

The authors interpret the results to suggest that microbial dysbiosis of the NP microbiome in infants precedes LRTIs. The use of the term dysbiosis implies that the microbiome of infants destined to suffer LRTI showed inherent differences after a few weeks following birth, suggesting that the bacteria present early on somehow direct later risk for infection.

This study is certainly of interest as there are virtually no longitudinal studies of the NP microbiome of infants who suffer LRTI. The existence of the sample bank built for the Southern Africa Mother-Infant Pertussis study (SAMIPS) conducted in Zambia of infants and their mothers were followed over the first three months of life. The authors are applauded for attempting to glean information about risk factors for LRTI using this biobank.

Assumptions that the differences in the microbiome seen to precede the LRTI may not entirely explain the child's risk for LRTI. First, with only 10 samples in the LRTI group, it is quite possible that the microbes identified may not be consistently found using a larger sample size. Also, the rather brief period of observation (14 weeks) also limits the generalizability of the findings.

LRTI infection is a catch-all diagnosis that includes many etiologic agents, including both viral and bacterial agents. The inability to pinpoint the etiologic agents in this study makes it more difficult to assign risk based on the observed results. The definitions of LRTI used in this study are also fairly non-specific and need to be kept in mind when considering results.

There is no mention, either in the Background or in the Discussion, that the oral cavity, in addition to the nasopharynx, serves as an important source of microbes into the lower airway. There have been quite a few studies in adults, and far fewer in children, that show that the oral cavity is an important source of lung microbes. At the very least, the oral cavity should be discussed as a possible source of microbes in this context.

The authors assume that the shifts noted in the proportion of various genera over time in some way influences vulnerability to LRTI. How might this happen? For example, *Anaerobacillus* was noted as elevated in proportions in subjects with LRTI risk. This is a rather recently described genus of spore forming strict anaerobes. How might this group impact other microbial groups?

The term dysbiosis may not be the most accurate description for the observations made in this study. The shifting of the flora to be dominated by one or a few bacterial groups at one time and by others later may be "normal"; inclusion of a larger sample size may have attenuated the observed effects. Also, most descriptions of the dysbiotic state note that microbial diversity diminishes with dysbiosis. It seems the opposite is true here, since many more genera were



associated with LRTI compared to healthy subjects. Also, limiting analysis to the genus level may have reduced the chance of identifying sub-groups of organisms associated with each state of health. A more granular sequence analysis might be worthwhile.

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Oral microbiology; Relationships between oral disease and systemic disease.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.**

Author Response 02 Mar 2024

**Rotem Lapidot**

We greatly appreciate the comments and advice received from the peer reviewer. We have attempted to address each of the issues raised and provide below a point-by-point summary of our responses. In each case we provide the comment verbatim, followed by our responses, heralded by '\*\*\*' and in italics.

The authors analyzed nasopharyngeal (NP) samples from a longitudinal, prospective cohort study of 1,981 Zambian mother-infant pairs who underwent NP sampling from 1-week through 14-weeks of age at 2-3-week intervals. The idea was to determine if substantial differences were apparent between the NP microbiome of children destined for a lower respiratory tract infection (LRTI) (N=10) when compared to normal infants (N=30), from essentially birth to 14 weeks. The microbiome was compared using 16S rRNA gene sequencing on the samples, as well as from baseline samples of the infants' mothers.

The authors interpret the results to suggest that microbial dysbiosis of the NP microbiome in infants precedes LRTIs. The use of the term dysbiosis implies that the microbiome of infants destined to suffer LRTI showed inherent differences after a few weeks following birth, suggesting that the bacteria present early on somehow direct later risk for infection.

*\*\*\* Our findings of distinct NP microbiome within days of birth in infants who later developed LRTI in the first 3 months of life, if conformed on larger scale studies, can only suggest an association. We have softened our conclusions accordingly. In our manuscript we very broadly and shortly discuss hypotheses that could explain this association*

*Here, detailed are several hypotheses:*

*1. The microbiome plays a causative role by one or more of these mechanisms (i.e., the microbiome directs later risk of infection):*

*i. The microbiome influences the environment in the nasopharynx (pH, metabolites, antibiotics exerted by the residing bacteria etc.). These environmental changes can indirectly influence a newly acquired pathobiont to be more virulent.*

*ii. The microbiome in the nasopharynx influences host responses to the pathobiont indirectly through immunological pathways.*

*iii. Direct interaction between pathobionts and the microbiome in the NP influencing virulent of the pathobiont (such as in the case with *S. pneumoniae* and RSV).*

*2. The microbiome is a result of underlying immunological host characteristics which are the actual reason for the higher risk of LRTI development.*

*We assume that there is probably more than one explanation for the observed association between the NP microbiome and later risk of LRTI, and there could be several roles the NP microbiome plays in the development of LRTI.*

This study is certainly of interest as there are virtually no longitudinal studies of the NP microbiome of infants who suffer LRTI. The existence of the sample bank built for the Southern Africa Mother-Infant Pertussis study (SAMIPS) conducted in Zambia of infants and their mothers were followed over the first three months of life. The authors are applauded for attempting to glean information about risk factors for LRTI using this biobank.

*\*\*\* We thank the reviewer for his kind words. We believe that this unique sample library could provide valuable insights into respiratory health in infants, and we have already completed several other analyses on these samples which are currently submitted for publication.*

Assumptions that the differences in the microbiome seen to precede the LRTI may not entirely explain the child's risk for LRTI. First, with only 10 samples in the LRTI group, it is quite possible that the microbes identified may not be consistently found using a larger sample size. Also, the rather brief period of observation (14 weeks) also limits the generalizability of the findings.

*\*\*\* We completely agree with the reviewer's comment. This is a pilot study, that was aimed to see if there was a signal supporting our hypothesis, with the plan to continue exploring this question on a larger scale if such signal was found.*

LRTI infection is a catch-all diagnosis that includes many etiologic agents, including both viral and bacterial agents. The inability to pinpoint the etiologic agents in this study makes it more difficult to assign risk based on the observed results. The definitions of LRTI used in this study are also fairly non-specific and need to be kept in mind when considering results.

*\*\*\* Once again, we completely agree with the reviewer's comment. Given the rarity of LRTI (especially with the modified restricted definition we have used to identify the most severely ill infants) we have a relatively small number of infants. We thus did not further divide into the different assumed etiologies. This is of course one of the limitations of this study (as we discuss in our limitations section), but non the less, biases our results to the null. If certain NP microbiome profiles are associated with specific etiologies of LRTI, then by clumping all etiologies together we might have lost some data suggesting this association. We are very eager to continue to explore this question, and we currently are in the process of analyzing these data sets for possible etiologies.*

There is no mention, either in the Background or in the Discussion, that the oral cavity, in addition to the nasopharynx, serves as an important source of microbes into the lower airway. There have been quite a few studies in adults, and far fewer in children, that show that the oral cavity is an important source of lung microbes. At the very least, the oral cavity should be discussed as a possible source of microbes in this context.

*\*\*\* We agree with the reviewer that the oral cavity and the oral microbiome is of interest and importance. In the pediatric population (unlike adults) the anatomical niche for the common respiratory pathogens such as *S. pneumoniae* and *H. influenzae* is the nasopharynx and not the oral cavity. In this analysis we did not aim to explore the pathogens directly causing the LRTI, but rather explore the respiratory anatomical sight where first interactions with the pathobionts occur. In future studies it would be of interest to collect both NP swabs and oral swabs and explore these associations. A similar analysis of the OP microbiome over time could yield different results, but since we did not sample it, we cannot comment further.*

The authors assume that the shifts noted in the proportion of various genera over time in some way influences vulnerability to LRTI. How might this happen? For example, Anaerobacillus was noted as elevated in proportions in subjects with LRTI risk. This is a rather recently described genus of spore forming strict anaerobes. How might this group impact other microbial groups?

*\*\*\* It is important to emphasize that we do not assume influences of the microbiome. We are strictly describing our observations. We can try to hypothesize how these observations could be associated with the development of LRTI (as detailed above in the first response), but these would only be hypothesis. Our concept regarding microbiome is that there is significance to the context in which a certain microbe is found. As a simplified example, detecting *S. pneumoniae* together with *Moraxella* could have a completely different meaning then detecting *S. pneumoniae* together with *H. influenzae*, and therefore hypothesizing regarding a specific microbe would be of little value. We should also emphasize that given the comments we received from other reviewers, these rare genera (such as *Anaerobacillus*) could also be a result of contamination that we were not able to control for.*

The term dysbiosis may not be the most accurate description for the observations made in this study. The shifting of the flora to be dominated by one or a few bacterial groups at one time and by others later may be “normal”; inclusion of a larger sample size may have attenuated the observed effects. Also, most descriptions of the dysbiotic state note that microbial diversity diminishes with dysbiosis. It seems the opposite is true here, since many more genera were associated with LRTI compared to healthy subjects. Also, limiting analysis to the genus level may have reduced the chance of identifying sub-groups of organisms associated with each state of health. A more granular sequence analysis might be worthwhile.

*\*\*\* We thank the reviewer for these comments. As suggested, we have taken out the term dysbiosis from the manuscript in appropriate places.*

*We and others have noted that the NP microbiome in health is correlated with lower diversity, as opposed to the fecal microbiome where health states are associated with higher diversity. Our finding of what we termed ‘dysbiosis’ referred to the distinct microbiome characteristics we observed in the first sample taken from infants (those who developed LRTI compared to healthy controls) and was not related to the shifting of dominant bacteria over time. When comparing the changes of the NP microbiome during the first 3 months of life between LRTI infants and controls we observed similar succession patterns.*

*We appreciate the comment that a more granular analysis might be useful. Although we have established that species-level classification is made more accurate by metagenomic methods such as PathoScope (Odom et al, 2023), genus level classification is much more reliable. In addition, our results would not be either changed or strengthened by the delineation of specific species (e.g. *Dolosigranulum pigrum*). So we decided to focus only on the genus level. We have included this justification in the manuscript.*

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 16 May 2022

<https://doi.org/10.21956/gatesopenres.14828.r31973>

© 2022 De Steenhuijsen Piters W et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Wouter A.A. De Steenhuijsen Piters** 

<sup>1</sup> Department of Paediatric Immunology and Infectious Diseases, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, The Netherlands

<sup>2</sup> Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

**Mari-Lee Odendaal**

<sup>1</sup> Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

<sup>2</sup> Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, The Netherlands

### Summary:

In the original article entitled “*Nasopharyngeal Dysbiosis Precedes the Development of Lower Respiratory Tract Infections in Young Infants, a Longitudinal Infant Cohort Study*”, Lapidot and co-authors investigated the maturation of the nasopharynx microbiome of 40 Zambian infants, part of whom developed (severe) LRTIs ( $n=10$ ). This paper is of importance, as it contrasts earlier work, which is largely based on European/Australian cohorts. Regardless, similar signals were identified, with an enrichment of *Dolosigranulum* in healthy infants compared to those developing LRTIs. In addition, the authors suggest *Dolosigranulum* abundance already diminishes prior to LRTI and is found in lower abundance in mothers of children with LRTIs, which is very provocative.

However, especially these latter analyses need to be further substantiated by statistics, among others appropriately controlling for age in order to draw these conclusions. Apart from that, we are concerned some of the genera related to ‘dysbiosis’ (before/during LRTI; e.g. *Novosphingobium*, *Delftia*, *Anaerobacillus* and *Bacillus*) may not represent true biological signals, instead reflecting background contamination. These species are typically not observed in the nasopharynx (neither in health, nor in disease; see among others Man *et al.*, 2019<sup>1</sup>) and instead are reported as part of the ‘reagent kitome’ (Salter *et al.*, 2014<sup>2</sup>). More information/in depth analyses are required to make a distinction between true signal/contamination, among others inspecting/reporting blank profiles, running decontam (in case of sufficient controls), and by assessing genera by DNA-isolation run/date. Apart from these main points, we encountered a high number of ‘minor’ points, which largely revolve around unclarities/discrepancies in methods/results, unclear structuring of methods/results, definition of LRTI vs LRTI symptoms and notes on superfluous/repetitive statements/analyses. The paper would benefit from extensive restructuring based on these points in our view.

### Major comments:

1. As the authors are aware, nasopharyngeal samples are particularly sensitive to contamination due to their low bacterial density, especially in early life (see among others our previous work; Bosch *et al.*, 2017<sup>3</sup> and De Steenhuijsen Piters *et al.*, 2022<sup>4</sup>). This underlines the importance of carefully examining the profiles of NP samples and their blanks. In this current study, the authors included negative controls throughout their laboratory analysis, but discharged the samples after the laboratory steps due to a low number of reads. Although samples were processed at random (allowing authors to correct for batch effects), more global contamination (which may be quite stable in a given lab over time) cannot be corrected for in this manner. In addition, the authors did not report information on bacterial density, which would have been helpful to assess the risk of contamination in these samples (and for example link low biomass with profiles enriched for contaminants).

A sufficient number of reagent controls and density measurements allows for the use of R-packages like decontam (Davis *et al.*, 2018<sup>5</sup>), allowing for filtering of contaminating taxa. We admit this may be challenging in the context of this study, given that low biomass is likely associated with young age and with higher numbers of contaminants. Inspection of decontam-results would therefore be warranted in addition to running decontam in the first place.

Together, the points raised above open up the discussion on whether contamination has impacted

the microbiota profiles reported by the authors. Generally, I believe the average profiles, as well as individual profiles look roughly adequate and in line with previously published literature. However, especially the enrichment of genera including *Novosphingobium*, *Delftia*, *Anaerobacillus* and *Bacillus* in pre-LRTI/LRTI is troublesome, as these are all well-known contaminants (Salter *et al.*, 2014<sup>2</sup>) and were not reported in previously published studies (e.g. Teo *et al.*, 2015<sup>6</sup>, Kelly *et al.*, 2017<sup>7</sup>). Furthermore, among others Extended Figure 1 (healthy infants) shows an early-life (likely low biomass) and highly diverse cluster with *Anaerobacillus* and *Pseudomonas*, which could also reflect possible signs of contamination, rather than true biological signals. This is less problematic, as these genera are not prominently reported as differentially abundant.

Taken together, we would urge the authors to further assess this issue, for example by further inspecting blank profiles, running decontam (if the number of blanks allows this), assessing DNA density, and assessing samples by DNA-isolation run/DNA-isolation date. If the possibility of contamination cannot be excluded, this should be at least clearly discussed in the paper and conclusions based on possible contaminants should be down-toned.

2. Some of the conclusions the authors draw are insufficiently supported by their analysis. Among others, although highly provocative, the notion that *Dolosigranulum* abundance already diminishes prior to LRTI is not substantiated by statistical analyses. See for details the minor comments below.

3. Important clinical information is currently not provided. This includes information on LRTI phenotype, etiology (causative pathogen? Viral data?), clinical information on LRTI symptoms/severity, treatment (antibiotics/immunosuppressants) and co-morbidities.

#### Minor comments:

1. The term 'dysbiosis' is generally considered vague. If possible, be specific on what is perturbed compared to healthy controls (e.g. alpha-/beta-diversity, composition of specific taxa, etc).
2. Abstract: 'Whether these profiles precede the infection or a consequence of it, ...', should be '...or are a consequence of it...'
3. Abstract: it is somewhat misleading to report the total number of participants to this study in the abstract, while microbiota analysis is performed on a (much) smaller subset of these infants. Please report adequate sample size numbers here, if any.
4. Abstract: the methods section could be written a bit more 'to the point'/shorter.
5. Abstract: '... or could be a marker of other pathogenic forces that directly lead to LRTI.'. What is meant by this?
6. Background: second section of the first paragraph on *S. pneumoniae* seems too detailed. I think the second paragraph also already conveys this message. Consider omitting/shortening this.
7. Background: 'LRTI is seen ... impede or promote LRTI.'; consider simplifying this section a bit. I think the point should be that microbial development is impacted by various

host/environmental factors (birth mode/viral infection/nutritional status), which can have direct/indirect impact on pathogen colonization/infection susceptibility.

8. Background: consider shortening the last 2 paragraphs, since especially sample numbers, timing of sampling etc. should be reported in methods.
9. Methods: importantly, was this nested microbiome study part of the initial study plan? What was the primary goal of the SAMIPS study? Could the authors expand on what type of mothers/families were samples; i.e. urbanization level, nutritional status, education level etc.
10. Methods: could the authors be more specific/summarize the (adjusted) WHO-criteria used to determine LRTI?
11. Methods: is the DNA extraction method used based on mechanical lysis/chemical lysis of cells? Especially, chemical lysis could result in an underrepresentation of gram-positives. Is this method benchmarked for use with respiratory samples?
12. Methods: methods for DNA amplification and MiSeq PCR are very detailed, if possible, could the authors refer to other papers using the same protocols? Or state something along the lines of 'performed in accordance to manufacturer's instructions' (if that is the case)?
13. Methods: this paragraph: 'Sequencing data were processed using QIIME2 (Bolyen et al., 2019) and Pathoscope2 (Hong et al., 2014)'. Samples with less than 10,000 reads were excluded from further analysis.' should be moved to the 'Data processing' section.
14. Methods: this statement 'To account for reagent ... as well as clinical data.', is a bit odd. Processing in random fashion/blinding is not done to account for contamination in my view.
15. Methods: information on processing using QIIME is missing. QIIME incorporates many different programs/tools, so please make sure to report the vital tools/parameters used or refer to previous work. Apart from information on filtering and trimming, which was already provided, information on denoising/error correction, merging of paired reads, ASV/OTU-calling, removal of chimeras and taxonomic annotations should be included (or referenced).
16. Methods: why did the authors choose to use PathoScope2 to annotate their 16S-reads? I do not encounter this often; has this been validated in the context of 16S-microbiota analyses? The paper on PathoScope seems to mostly focus on annotation of reads generated through metagenomic sequencing. Why did the authors not use a more standard approach (implemented in QIIME), like a naive bayesian classifier/DECIPHER to annotate reads?
17. Methods: the authors conducted all the analyses at genus level, it is however unclear why the decision was made to focus on genus and not on individual taxa (ASVs/OTUs). Looking at a lower taxonomic level can be of relevance since specific strains or species within a genus can have a very different function and thus associations with outcomes. We therefore encourage the authors to clearly explain their decision to look at genus level.

18. Methods: the authors used linear regression in their analyses. Among others, linear regression models assume a linear relationship between the predictor (e.g. age; I assume this was modeled as a continuous variable, please specify) and outcome variable (diversity/genus abundance). Did the authors check these assumptions? In figure 1 for instance, age does not appear to be linearly correlated with the relative abundance of the top taxa. We encourage the authors to elaborate on this further, also considering that some genera show a non-linear abundance over time. They could therefore consider fitTimeSeries-analyses (metagenomeSeq-package) or use GAM/spline-based models.
19. Methods: for visualization and modeling, several thresholds for inclusion of genera were used. What was the rationale behind these thresholds? Were these defined post-hoc (after running the analyses) or up front?
20. Methods: p-values of mixed linear models were calculated using ANOVA-tests. Against what model did the authors test their 'full model' (including infection status, age, the interaction age:infection status and HIV status)? An empty model, only including an intercept? Also, this model does not account for non-linear microbiota development over time (see previous comment). In addition, given that authors matched for HIV exposure status (according to the results section), why did they add that to their model (while not adding season/maternal age, other factors they matched for)? Could the authors clarify and align information on matching in Methods and Results?
21. Methods: could the authors provide any information on antibiotic usage in these infants (especially around birth/LRTI). Were all infants breastfed?
22. Methods: generally, the description of models is detailed, but also seems repetitive. Please check whether condensing this information a bit more is possible. Also, it is not clear why different modeling frame works were used (e.g. DESEQ2 vs mixed linear models). Also, in DESEQ2 it seems the authors did not account for repeated measures (i.e. subject as random effect).
23. Methods: 'For each cluster, we performed Fisher's exact tests to determine whether that cluster was enriched for LRTI samples generally, pre-symptomatic samples, active symptom samples, or HIV-exposed samples.'. It is unclear how these sample-types are defined (particularly 'active symptom samples'). Are these the same as LRTI-samples? Please clarify. In addition, the authors should consider adjusting for age when assessing enrichment of LRTI-samples in specific clusters, given that these samples are typically collected at older age. This would imply running (mixed) logistic regression models including age.
24. Methods: was the Spearman correlation-analysis between mothers/infants performed on the vector of genus abundances for each mother-infant pair (separately per time point)?
25. Results: first section on definition of LRTIs should be moved to the Methods section.
26. Results/table 1: could any more information on LRTI phenotype be included, what symptoms did infants have, how severe were these LRTIs (severity score), what treatment did infants receive, could the authors share any information on LRTI etiology?



27. Results: 'A third of infants with LRTI were born to mothers with HIV (receiving anti-retroviral treatment), compared to 40% of infants in the healthy group.' I do not follow; few lines back the authors describe matching for HIV exposure. Similarly, it seems the authors matched for season, yet there are 2 LRTI infants enrolled in rainy season vs 2 healthy controls enrolled in the same season. According to the 3:1-scheme this should be 6 healthy controls. Was the matching imperfect?
28. Results/table 2; the distinction between symptomatic (define) and non-symptomatic routine visits should be made more clearly in the methods-section.
29. Results/table 2: 'Diagnosis of LRTI (cough/runny or blocked nose with or without fever AN 0 fast breathing with indrawing of the chest)'. Is 'AN 0 fast breathing' a typo?
30. Results/analysis one: I suggest using the term 'development' or 'dynamics' instead of 'evolution'.
31. Results/analysis one: '... different genera across each age averaged stratum'. Please rephrase.
32. Results/analysis one: please use '*Corynebacterium*' instead of '*Corynebacteria*'.
33. Results/general: the results on age dynamics in (healthy) NP microbiota composition/diversity lack a measure of effect size. If possible, add this, at least for important findings. Also, as noted in the methods, please check if the assumption that relative abundance is linearly related to age is valid. For example, inspecting scatter plots of relative abundance across age would be helpful.
34. Results/analysis one: 'However, alpha diversity only reflects the number of dominant genera, and not whether the dominant genera are themselves diverse.'. I politely disagree, dominant genera should be indicated by low richness and low evenness. Consider rephrasing this sentence. I think assessing beta-diversity is valid regardless of alpha-diversity results.
35. Extended Figure 1A: consider removing white lines around bars and annotating the clusters (e.g. MIX-cluster or *Haemophilus* (HAE)-cluster). Addendum: this was done for the full dendrogram-analysis I see; as suggested consider only presenting that analysis instead (i.e. running clusterin once).
36. Extended Figure 1B: explain RCE on x-/y-axis; consider converting to a % of explained variance.
37. Extended Figure 2: also see methods, please define pre-/post- symptom groups more explicitly.
38. Extended Figure 3: in line with previous comments; be clear on LRTI vs LRTI-symptoms and whether there is a difference. Is there anything known on RTI symptoms in healthy controls?

39. Extended Figure 4: I applaud the authors for adding both 2D and 3D NMDS-plots, yet a visualization of the 3D analysis in 2D (i.e. plotting NMDS1 vs NMDS2, NMDS3 vs NMDS2 and NMDS1 vs NMDS3), would be easier to interpret.
40. Results/analysis three: this section (especially the first sections) is generally difficult to follow, it seems partly mixed with methods and seems to discuss several analyses at the same time. Please try to restructure this section, discussing any analyses one-by-one.
41. Figure 2: methods are lacking information on how clinical variables were projected into ordination space (envFit?).
42. Figure 3: why did the authors choose to report these two genera? Please provide a rationale, for example based on other analysis supporting this choice. Also, it would be helpful to know/visualize the number of samples at each time point. Did the authors consider running any statistics on these results?
43. Results/analysis three: the statement 'Concentrating on the samples taken prior to infection (and prior to antibiotic administration), this analysis confirmed lower relative abundance of *Dolosigranulum*, and higher relative abundance of *Anaerobacillus* in the LRTI infants before their infection.', although very interesting, is not supported sufficiently by the authors' analyses I believe. Time points prior to infection are likely related to earlier age and therefore higher *Dolosigranulum*, while time points after infections are related to older age and lower *Dolosigranulum*. Therefore, the pattern observed may merely reflect age dynamics. To support the statement the authors are trying to make, it may be interesting to look at time points before/after infection vs and age-matched control samples. Or include age when modeling these effects. Apart from that, from eyeballing this figure, I would also conclude the *Dolosigranulum* dramatically drops upon infection. In fact, it could be both, i.e. lower *Dolosigranulum* prior to infection and a more significant drop after infection. In addition, also given the question of the authors in the discussion ('But are these microbial profiles a result of the infection? Or were they present before the infection?'), it would make sense to expand/deepen this analysis further, including pathobionts like *Haemophilus*/*Streptococcus*/*Moraxella*.
44. Results/analysis four: again, this section includes too many details on methods in my view.
45. Results/analysis four: the authors refer to table 4, yet this seems to not include data on clustering. Should this be Figure 4?
46. Results/analysis four: the authors present many dendrograms/clustering in their paper (LRTI only, healthy only and all, both clustered based on two indices). If possible, I would advise to combine these analyses and properly name clusters according to the most dominant taxon, so that throughout the paper, the same clusters are discussed. In addition, if clustering based on Frey's index does not add anything to the message the authors want to convey, I would omit it. Similarly, consider presenting one figure showing per-individual microbiota profiles over time (grouped by LRTI/non-LRTI; currently Extended Figure 3 and 5).
47. Table 5: I suggest to rerun this analysis using cluster/no clusters as response variable and

LRTI status + age as predictors.

48. Results/analysis five: very interesting analysis, yet results are presented without any information on effect size. Also, it is not clear what number of genera are tested and whether other genera were also significantly different. Please expand on this analysis a bit further if possible.
49. Results/general: did the authors consider running further analyses on symptomatic/non-LRTI samples?
50. Discussion: make sure all relevant papers are cited, among others Man *et al.* (2019<sup>1</sup>) is currently missing, while this to date is one of the largest studies on (severe) LRTI and NP microbiota.
51. Discussion: I think some statements are insufficiently supported by the data; for example '... clear evidence of dysbiosis preceding the onset of LRTI.'. This seems based on figure 3, where no statistical support was provided. Also, as said 'dysbiosis' is a bit vague, notably, it seems the authors observed specific differentially abundant genera between LRTI vs no LRTI. Same goes for '... we observed different microbiome patterns', this seems based on analysis five, which requires more detail.
52. Discussion: 'The NP microbiome at time of infection is associated with the risk of development of LRTI and its severity.'. This does not make sense to me, do the authors refer to NP microbiome prior to infection?
53. Discussion: 'Thus, a similar association between the NP microbiome and risk of respiratory infections is a plausible theory for which there is precedent.' In fact, this has been shown for NP microbiota in the context of mild infections by our group (Bosch *et al.*, 2017<sup>3</sup> and De Steenhuijsen Piters *et al.*, 2022<sup>4</sup>, *Nat Microbiol*, 2022) and others (Teo *et al.*, 2015<sup>6</sup> and Teo *et al.*, 2018<sup>8</sup>).
54. Discussion: discussion on the comparison/possible differences between Zambian children and existing European/Australian cohorts is lacking. This is one of the big strengths of this study; we need data from a wider range of cohorts and this study may contribute to that goal in my view. Another big strength of this study is that data on healthy microbiota are available from children who will develop a severe LRTI (this contrasts among others work from our group, where we assessed longitudinal microbiota data in context of mild respiratory infections).

## References

1. Man W, van Houten M, Mérelle M, Vlieger A, et al.: Bacterial and viral respiratory tract microbiota and host characteristics in children with lower respiratory tract infections: a matched case-control study. *The Lancet Respiratory Medicine*. 2019; **7** (5): 417-426 [Publisher Full Text](#)
2. Salter SJ, Cox MJ, Turek EM, Calus ST, et al.: Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014; **12**: 87 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Bosch A, Piters W, van Houten M, Chu M, et al.: Maturation of the Infant Respiratory Microbiota,

- Environmental Drivers, and Health Consequences. A Prospective Cohort Study. *American Journal of Respiratory and Critical Care Medicine*. 2017; **196** (12): 1582-1590 [Publisher Full Text](#)
4. de Steenhuijsen Piters WAA, Watson RL, de Koff EM, Hasrat R, et al.: Early-life viral infections are associated with disadvantageous immune and microbiota profiles and recurrent respiratory infections. *Nat Microbiol*. **7** (2): 224-237 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Davis N, Proctor D, Holmes S, Relman D, et al.: Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. 2018; **6** (1). [Publisher Full Text](#)
6. Teo SM, Mok D, Pham K, Kusel M, et al.: The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe*. 2015; **17** (5): 704-15 [PubMed Abstract](#) | [Publisher Full Text](#)
7. Kelly MS, Surette MG, Smieja M, Pernica JM, et al.: The Nasopharyngeal Microbiota of Children With Respiratory Infections in Botswana. *Pediatr Infect Dis J*. 2017; **36** (9): e211-e218 [PubMed Abstract](#) | [Publisher Full Text](#)
8. Teo S, Tang H, Mok D, Judd L, et al.: Airway Microbiota Dynamics Uncover a Critical Window for Interplay of Pathogenic Bacteria and Allergy in Childhood Respiratory Disease. *Cell Host & Microbe*. 2018; **24** (3): 341-352.e5 [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

No

**If applicable, is the statistical analysis and its interpretation appropriate?**

No

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

No

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Our group's expertise is on early-life development of respiratory microbiota in the context of health and disease. Both reviewers specialise in microbiota data analysis.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.**

Author Response 02 Mar 2024

**Rotem Lapidot**

We greatly appreciate the comments and advice received from the peer reviewers. We have attempted to address each of the issues raised and provide below a point-by-point summary of our responses. In each case we provide the comment verbatim, followed by our responses, heralded by '\*\*\*' and in italics.

In the original article entitled "*Nasopharyngeal Dysbiosis Precedes the Development of Lower Respiratory Tract Infections in Young Infants, a Longitudinal Infant Cohort Study*", Lapidot and co-authors investigated the maturation of the nasopharynx microbiome of 40 Zambian infants, part of whom developed (severe) LRTIs ( $n=10$ ). This paper is of importance, as it contrasts earlier work, which is largely based on European/Australian cohorts. Regardless, similar signals were identified, with an enrichment of *Dolosigranulum* in healthy infants compared to those developing LRTIs. In addition, the authors suggest *Dolosigranulum* abundance already diminishes prior to LRTI and is found in lower abundance in mothers of children with LRTIs, which is very provocative.

*\*\*\* We thank the reviewer for this clear assessment of the importance of our work*

However, especially these latter analyses need to be further substantiated by statistics, among others appropriately controlling for age in order to draw these conclusions. Apart from that, we are concerned some of the genera related to 'dysbiosis' (before/during LRTI; e.g. *Novosphingobium*, *Delftia*, *Anaerobacillus* and *Bacillus*) may not represent true biological signals, instead reflecting background contamination. These species are typically not observed in the nasopharynx (neither in health, nor in disease; see among others Man *et al.*, 2019<sup>1</sup>) and instead are reported as part of the 'reagent kitome' (Salter *et al.*, 2014<sup>2</sup>). More information/in depth analyses are required to make a distinction between true signal/contamination, among others inspecting/reporting blank profiles, running decontam (in case of sufficient controls), and by assessing genera by DNA-isolation run/date. Apart from these main points, we encountered a high number of 'minor' points, which largely revolve around unclarities/discrepancies in methods/results, unclear structuring of methods/results, definition of LRTI vs LRTI symptoms and notes on superfluous/repetitive statements/analyses. The paper would benefit from extensive restructuring based on these points in our view.

*\*\*\* We thank the reviewer for these assessments and comments. We have updated our manuscript to address the individual points (as detailed below), including issues of statistical significance, contamination, and clarification of methods/results.*

**Major comments:**

1. As the authors are aware, nasopharyngeal samples are particularly sensitive to contamination due to their low bacterial density, especially in early life (see among others our previous work; Bosch *et al.*, 2017<sup>3</sup> and De Steenhuijsen Pijters *et al.*, 2022<sup>4</sup>). This underlines the importance of carefully examining the profiles of NP samples and their blanks. In this current study, the authors included negative controls throughout their laboratory analysis, but discharged the samples after the laboratory steps due to a low

number of reads. Although samples were processed at random (allowing authors to correct for batch effects), more global contamination (which may be quite stable in a given lab over time) cannot be corrected for in this manner. In addition, the authors did not report information on bacterial density, which would have been helpful to assess the risk of contamination in these samples (and for example link low biomass with profiles enriched for contaminants).

A sufficient number of reagent controls and density measurements allows for the use of R-packages like decontam (Davis *et al.*, 2018<sup>5</sup>), allowing for filtering of contaminating taxa. We admit this may be challenging in the context of this study, given that low biomass is likely associated with young age and with higher numbers of contaminants. Inspection of decontam-results would therefore be warranted in addition to running decontam in the first place.

Together, the points raised above open up the discussion on whether contamination has impacted the microbiota profiles reported by the authors. Generally, I believe the average profiles, as well as individual profiles look roughly adequate and in line with previously published literature. However, especially the enrichment of genera including *Novosphingobium*, *Delftia*, *Anaerobacillus* and *Bacillus* in pre-LRTI/LRTI is troublesome, as these are all well-known contaminants (Salter *et al.*, 2014<sup>2</sup>) and were not reported in previously published studies (e.g. Teo *et al.*, 2015<sup>6</sup>, Kelly *et al.*, 2017<sup>7</sup>). Furthermore, among others Extended Figure 1 (healthy infants) shows an early-life (likely low biomass) and highly diverse cluster with *Anaerobacillus* and *Pseudomonas*, which could also reflect possible signs of contamination, rather than true biological signals. This is less problematic, as these genera are not prominently reported as differentially abundant.

Taken together, we would urge the authors to further assess this issue, for example by further inspecting blank profiles, running decontam (if the number of blanks allows this), assessing DNA density, and assessing samples by DNA-isolation run/DNA-isolation date. If the possibility of contamination cannot be excluded, this should be at least clearly discussed in the paper and conclusions based on possible contaminants should be down-toned.

*\*\*\* We thank the reviewers for this important comment. The reviewer's point about contamination is well taken, and of course such events are known to occur, and we cannot completely eliminate the possibility that contamination events occurred. However, several features of our data suggest that these results are valid. First, the case and comparator samples were processed (DNA extraction, amplicon sequencing) at the same time, in the same lab, using the same batch of reagents. All samples were multiplexed and sequenced on the same flow cell. Note that we did not see these supposed "contamination" organisms in the controls. Contamination should be a consistent event, and so we would have expected to see these microbes in both groups. Second, the longitudinal sampling structure of these data sets also show that these rare organisms were present repeatedly within subjects over repeated independent samplings, and again more abundantly in the case samples, again in a way that is not consistent with contamination events, and which would be nearly impossible to detect in a cross-sectional analysis. Finally, we note that Salter *et al.*, did not mention *Anaerobacillus* and *Delftia* and as common contaminants, the Teo and Kelly studies were conducted in different countries (Teo in particular was in Australia), and they did not report rare microbes (e.g. Kelly only reported OTUs with 1,000+ reads). In addition, we used our PathoScope metagenomic*

*processing pipeline to identify microbes, which we have recently demonstrated to be more accurate than OTU-based methods for taxonomic classifications, especially for less-abundant microbes (Odom et al., Scientific Reports, 2023). Thus, the fact that we are finding additional rare microbes compared to these other studies is expected. However, to further address the reviewer's concern, we have added further discussion regarding possible contamination, as well as toned down our conclusions with respect to these microbes.*

2. Some of the conclusions the authors draw are insufficiently supported by their analysis. Among others, although highly provocative, the notion that *Dolosigranulum* abundance already diminishes prior to LRTI is not substantiated by statistical analyses. See for details the minor comments below.

*\*\*\* We have provided responses to the minor concerns below and in the manuscript. For the specific concern with Dolosigranulum, a linear mixed model for pre-LRTI timepoints that accounts for time and individual provides a statistically significant FDR and p-value (<0.001) that the abundance of Dolosigranulum is lower in the infants that later develop LRTI. We interpret this model/p-value as compelling statistical evidence that supports our conclusions.*

3. Important clinical information is currently not provided. This includes information on LRTI phenotype, etiology (causative pathogen? Viral data?), clinical information on LRTI symptoms/severity, treatment (antibiotics/immunosuppressants) and co-morbidities.

*\*\*\* Table 2 summarizes the clinical data of all infants with LRTI. For this analysis we did not include causative pathogen, but this data is further analyzed in other studies we have done on this population. The logic behind this was to see if regardless of a specific pathogen we can identify infants at risk for severe respiratory infections, and we believe our data suggest so. By definition, all infants enrolled were healthy, with no known immunocompromised states or comorbidities. HIV exposed infants were enrolled only if their mother was treated for HIV during her pregnancy (added clarifying sentence). As well, antibiotic was not prescribed prior to the event of LRTI (some infants were treated at time of LRTI and that is why we did not include analysis of the samples after the LRTI event).*

#### **Minor comments:**

1. The term 'dysbiosis' is generally considered vague. If possible, be specific on what is perturbed compared to healthy controls (e.g. alpha-/beta-diversity, composition of specific taxa, etc).

*\*\*\*We changed the term "dysbiosis" to a more specific description of the observed changes of the microbiome throughout the manuscript, where appropriate.*

2. Abstract: 'Whether these profiles precede the infection or a consequence of it, ...', should be '...or are a consequence of it...'

*\*\*\*We have added "are" as the reviewers suggested.*

3. Abstract: it is somewhat misleading to report the total number of participants to this

study in the abstract, while microbiota analysis is performed on a (much) smaller subset of these infants. Please report adequate sample size numbers here, if any.

*\*\*\*We have taken out the general number of participants in the original cohort and included only the number of infants included in this analysis.*

4. Abstract: the methods section could be written a bit more 'to the point'/shorter.

*\*\*\*We have shortened the methods section in the abstract as the reviewer suggested*

5. Abstract: '... or could be a marker of other pathogenic forces that directly lead to LRTI.' What is meant by this?

*\*\*\*Changed to "marker of underlying immunological, environmental or genetic characteristics that predispose to LRTI."*

6. Background: second section of the first paragraph on *S. pneumoniae* seems too detailed. I think the second paragraph also already conveys this message. Consider omitting/shortening this.

*\*\*\*We omitted this section as the reviewers suggest.*

7. Background: 'LRTI is seen ... impede or promote LRTI.'; consider simplifying this section a bit. I think the point should be that microbial development is impacted by various host/environmental factors (birth mode/viral infection/nutritional status), which can have direct/indirect impact on pathogen colonization/infection susceptibility.

*\*\*\*We have simplified this section to be more concise.*

8. Background: consider shortening the last 2 paragraphs, since especially sample numbers, timing of sampling etc. should be reported in methods.

*\*\*\*We have shortened the last two paragraphs as the reviewers recommended*

9. Methods: importantly, was this nested microbiome study part of the initial study plan? What was the primary goal of the SAMIPS study? Could the authors expand on what type of mothers/families were sampled; i.e. urbanization level, nutritional status, education level etc.

*\*\*\* We have added the requested changes.*

10. Methods: could the authors be more specific/summarize the (adjusted) WHO-criteria used to determine LRTI?

*\*\*\*We have added the definition of the WHO for pneumonia and the changes we have made to adjust for a more specific diagnosis.*

11. Methods: is the DNA extraction method used based on mechanical lysis/chemical lysis of cells? Especially, chemical lysis could result in an underrepresentation of gram-positives.



Is this method benchmarked for use with respiratory samples?

*\*\*\* As noted in the methods, we used the EasyMAG system for DNA extraction, which is benchmarked for respiratory samples, and this uses a chemical lysis buffer.*

12. Methods: methods for DNA amplification and MiSeq PCR are very detailed, if possible, could the authors refer to other papers using the same protocols? Or state something along the lines of 'performed in accordance to manufacturer's instructions' (if that is the case)?

*\*\*\* Given that there is no word limit on Gates Open Access, we see no reason to reduce the details about how the PCR and sequencing was performed.*

13. Methods: this paragraph: 'Sequencing data were processed using QIIME2 (Bolyen et al., 2019) and Pathoscope2 (Hong et al., 2014)'. Samples with less than 10,000 reads were excluded from further analysis.' should be moved to the 'Data processing' section.

*\*\*\* We have moved these sentences to the "Data processing" section.*

14. Methods: this statement 'To account for reagent ... as well as clinical data.', is a bit odd. Processing in random fashion/blinding is not done to account for contamination in my view.

*\*\*\* We agree that this sentence is confusing and have rewritten it accordingly.*

15. Methods: information on processing using QIIME is missing. QIIME incorporates many different programs/tools, so please make sure to report the vital tools/parameters used or refer to previous work. Apart from information on filtering and trimming, which was already provided, information on denoising/error correction, merging of paired reads, ASV/OTU-calling, removal of chimeras and taxonomic annotations should be included (or referenced).

*\*\*\* While QIIME2 was used for some preliminary analyses, all of our final results were generated with PathoScope. Thus, we have rewritten this section to eliminate the reference to the older QIIME2 software.*

16. Methods: why did the authors choose to use PathoScope2 to annotate their 16S-reads? I do not encounter this often; has this been validated in the context of 16S-microbiota analyses? The paper on PathoScope seems to mostly focus on annotation of reads generated through metagenomic sequencing. Why did the authors not use a more standard approach (implemented in QIIME), like a naïve bayesian classifier/DECIPHER to annotate reads?

*\*\*\* We and others have recently established that metagenomic processing methods such as PathoScope and Kraken provide more accurate taxonomic characterization of microbes in 16S data than OTU/ASV based methods such as QIIME2, DADA2, and Mothur (Odom et. al., Scientific Reports, 2023). The increased accuracy, especially for low-level taxonomies and low-abundance microbes was particularly important for our research goals. Of note, our team developed PathoScope, which is why we selected that metagenomic approach.*

17. Methods: the authors conducted all the analyses at genus level, it is however unclear

why the decision was made to focus on genus and not on individual taxa (ASVs/OTUs). Looking at a lower taxonomic level can be of relevance since specific strains or species within a genus can have a very different function and thus associations with outcomes. We therefore encourage the authors to clearly explain their decision to look at genus level.

*\*\*\* Although we have established that species-level classification is made more accurate by metagenomic methods such as PathoScope (Odom et al, 2023), genus level classification is much more reliable. In addition, our results would not be either changed or strengthened by the delineation of specific species (e.g. Dolosigranulum pigrum). So, we decided to focus only on the genus level. We have included this justification in the manuscript.*

18. Methods: the authors used linear regression in their analyses. Among others, linear regression models assume a linear relationship between the predictor (e.g. age; I assume this was modeled as a continuous variable, please specify) and outcome variable (diversity/genus abundance). Did the authors check these assumptions? In figure 1 for instance, age does not appear to be linearly correlated with the relative abundance of the top taxa. We encourage the authors to elaborate on this further, also considering that some genera show a non-linear abundance over time. They could therefore consider fitTimeSeries-analyses (metagenomeSeq-package) or use GAM/spline-based models.

*\*\*\* We appreciate this concern by the reviewer. We note that our linear mixed models and most other analyses were applied to the log-counts per million (logCPM). The log transformation manages the non-linearity in the abundances quite well and was selected based on appropriate diagnostics as suggested by the reviewer. This approach (linear modeling on logged data) is quite common and standard in these data types, so we feel it does not need special justification in our methods. In addition, because the mixed model on the logged data fits so well, and because our number of longitudinal measurements is relatively small/moderate, we feel that it would be overkill to apply time-series or GAM-based models for these data. For figure 1 in particular, we note that the "stacked" nature of the bar chart accentuates the look of non-linearity in the microbes, and again, what non-linearity is there is easily handled by a log transformation.*

19. Methods: for visualization and modeling, several thresholds for inclusion of genera were used. What was the rationale behind these thresholds? Were these defined post-hoc (after running the analyses) or up front?

*\*\*\* Yes, thresholds were selected to optimize visual appeal and quality in our results/figures, and were empirically derived/decided. However, in each case we labored over multiple thresholds and were careful that the selection of thresholds did not impact the conclusions to be drawn – but rather were to clarify the results and conclusions.*

20. Methods: p-values of mixed linear models were calculated using ANOVA-tests. Against what model did the authors test their 'full model' (including infection status, age, the interaction age:infection status and HIV status)? An empty model, only including an intercept? *\*\*\* We did not use 'ANOVA-tests' to calculate p-values for our mixed models. Our Methods section states that we used the "Anova function of the car package" to calculate our p-values; car::Anova is a versatile generic analysis function that can input multiple (>13) distinct R model objects (e.g., lm, glm, multinom, coxph, lme mixed models) and calculates appropriate p-*

*values for the coefficients for these models based on their method/type. We note that `car::Anova` is a very commonly used function for the calculation of  $p$ -values for coefficients for lme mixed model objects in R—and does not use a full/reduced model approach as suggested by the reviewer. We refer the reviewer to the reference in the paper (Weisenberg, et al. 2019) or to the `car` package user manual (<https://cran.r-project.org/web/packages/car/car.pdf>) for more details on how `car::Anova` calculates  $p$ -values for mixed models.*

Also, this model does not account for non-linear microbiota development over time (see previous comment).

*\*\*\*The analyses account for non-linearity through the use of a log transformation (see previous comment)*

In addition, given that authors matched for HIV exposure status (according to the results section), why did they add that to their model (while not adding season/maternal age, other factors they matched for)? Could the authors clarify and align information on matching in Methods and Results?

*\*\*\* There is a strong body of literature, including our own work (Brennan et al. J Acquir Immune Defic Syndr, 2019, Odom et al., Gates Open Research, 2022), that establishes that HIV-exposure can lead to significant differences in the risk to develop respiratory infections and in the airway microbiome of infants and children. HIV exposure was a significant factor in our data for this paper as well. In contrast, we didn't observe strong differences in our data based on our other matching factors (season/maternal age, etc). We point out that unless a paired analysis is conducted (e.g. paired t-test), statistically significant matching/confounding variables should always be included as a covariate in a statistical model—hence the inclusion of HIV status in our model. We refer the reviewer to the following reference (or a linear modeling textbook) for more information on why/when matching variables should be included in models: <https://www.bmj.com/content/352/bmj.i969>.*

21. Methods: could the authors provide any information on antibiotic usage in these infants (especially around birth/LRTI). Were all infants breastfed?

*\*\*\*All infants were deemed healthy when enrolled in the study (and therefore were not given antibiotic by definition). We added a sentence in the study design regarding antibiotic documentation. Since these infants were healthy at enrollment and were followed in the clinic where the study was conducted, we assumed that infants with LRTI did not receive antibiotics prior to their symptoms and that the healthy cohort did not receive antibiotics at all. We also refer to antibiotics' prescription in our analysis three.*

22. Methods: generally, the description of models is detailed, but also seems repetitive. Please check whether condensing this information a bit more is possible. Also, it is not clear why different modeling frame works were used (e.g. DESEQ2 vs mixed linear models). Also, in DESEQ2 it seems the authors did not account for repeated measures (i.e. subject as random effect).

*\*\*\* As mentioned above, because Gates does not have a word/page limit, we feel a detailed*

*(albeit repetitive) methods section is preferred to enhance the replicability and reproducibility of our work. We note that DESeq2 was used for two analyses: The analysis of the infant microbiomes at the first timepoint, and 2) the analysis of the maternal microbiomes. DESeq2 was not used for any analyses with repeated measures. Rather, mixed models were used to manage repeated measures in analyses that used longitudinal measures across individuals. Although the DESeq2 and mixed modeling approaches utilize different techniques and error models, they are both valid and standard methods for their respective applications, which is why they were applied in their contexts in the paper.*

23. Methods: 'For each cluster, we performed Fisher's exact tests to determine whether that cluster was enriched for LRTI samples generally, pre-symptomatic samples, active symptom samples, or HIV-exposed samples.'. It is unclear how these sample-types are defined (particularly 'active symptom samples'). Are these the same as LRTI-samples? Please clarify. In addition, the authors should consider adjusting for age when assessing enrichment of LRTI-samples in specific clusters, given that these samples are typically collected at older age. This would imply running (mixed) logistic regression models including age.

*\*\*\* Each measurement of the microbiome came from a sample that was either HIV exposed or not exposed and came from a child that either experienced an LRTI or not. For the LTRI infants, their points were further classified to their pre-symptomatic (pre-LRTI symptoms), (actively) symptomatic (LRTI symptoms), and post-LRTI. We thank the reviewers for pointing this out and clarified this in the methods. In this unsupervised dimension reduction analysis, we were trying to identify whether a cluster was enriched HIV status, LTRI samples, or pre-, active, post status, or age. We acknowledge that age can be a confounding factor here, and the reviewers make an excellent suggestion on how to adjust for this possible confounding using multivariate analyses. However, here we were just generally exploring the data from a univariate perspective, and did not plan to make any strong statistical conclusions on group membership (or its meaning), therefore for our needs we feel a univariate exploration was sufficient.*

24. Methods: was the Spearman correlation-analysis between mothers/infants performed on the vector of genus abundances for each mother-infant pair (separately per time point)?

*\*\*\* Yes, the spearman correlation analysis was conducted between the mother and child at the genus level, at only the first time point for the infant. This is already stated in the results section—we have now added this to the methods section.*

25. Results: first section on definition of LRTIs should be moved to the Methods section.

*\*\*\* We have moved the first section of results to the method section.*

26. Results/table 1: could any more information on LRTI phenotype be included, what symptoms did infants have, how severe were these LRTIs (severity score), what treatment did infants receive, could the authors share any information on LRTI etiology?

*\*\*\* The LRTI status was inferred based on whether infants presented with symptoms consistent with WHO's clinical definition of severe pneumonia, which is based on respiratory rate, fever, and chest wall indrawing, and adapted to the purpose of this study as described. Symptoms of the*

*infants is detailed in table 2, and in general, we identified the 10 infants with the most severe presentation (as discussed in the methods, study design). As we mention, infants were often prescribed antibiotics at the time of diagnosis. However, our focus is on the events prior to the LRTI. In our view, the antibiotic exposure impact on the microbiome is a distinct question from whether dysbiosis preceded LRTI, which is the key finding of the paper.*

27. Results: 'A third of infants with LRTI were born to mothers with HIV (receiving anti-retroviral treatment), compared to 40% of infants in the healthy group.' I do not follow; few lines back the authors describe matching for HIV exposure. Similarly, it seems the authors matched for season, yet there are 2 LRTI infants enrolled in rainy season vs 2 healthy controls enrolled in the same season. According to the 3:1-scheme this should be 6 healthy controls. Was the matching imperfect?

*\*\*\* We thank the reviewers for pointing out this important point, and we have deleted matching by HIV status. Matching of some comparisons was performed to the best of our ability, considering that the study was not designed to answer that specific question directly.*

28. Results/table 2; the distinction between symptomatic (define) and non-symptomatic routine visits should be made more clearly in the methods-section.

*\*\*\*The distinction is between scheduled and unscheduled visits, since infants could be symptomatic on a scheduled visit, but could also be seen for an unscheduled visit if they developed symptoms in between study visits. This is detailed in the method section Table 2 summarizes whether an infant had symptoms in each visit, and if they were upper respiratory symptoms or Lower respiratory symptoms, with the definition given in the legend of the figure.*

29. Results/table 2: 'Diagnosis of LRTI (cough/runny or blocked nose with or without fever AN 0 fast breathing with indrawing of the chest)'. Is 'AN 0 fast breathing' a typo?

*\*\*\*We thank the reviewers for pointing out, yes this is a typo (should be AND fast breathing) and was corrected.*

30. Results/analysis one: I suggest using the term 'development' or 'dynamics' instead of 'evolution'.

*\*\*\*The word evolution was replaced with development.*

31. Results/analysis one: '... different genera across each age averaged stratum'. Please rephrase.

*\*\*\*Changed "age averaged stratum" to "age group".*

32. Results/analysis one: please use 'Corynebacterium' instead of 'Corynebacteria'.

*\*\*\*We have changed corynebacteria to Corynebacterium.*

33. Results/general: the results on age dynamics in (healthy) NP microbiota

composition/diversity lack a measure of effect size. If possible, add this, at least for important findings. Also, as noted in the methods, please check if the assumption that relative abundance is linearly related to age is valid. For example, inspecting scatter plots of relative abundance across age would be helpful.

*\*\*\* We added p-values for the microbes that change significantly over time, but did not add an effect size because we feel that effect sizes in this case did not add any value to our narrative. Since we used linear mixed models, with log counts/million to transform the data, the curvilinear relationship of the data has been accounted for statistically. We apologize for the confusion on this point and have clarified this in the results.*

34. Results/analysis one: 'However, alpha diversity only reflects the number of dominant genera, and not whether the dominant genera are themselves diverse.' I politely disagree, dominant genera should be indicated by low richness and low evenness. Consider rephrasing this sentence. I think assessing beta-diversity is valid regardless of alpha-diversity results.

*\*\*\*We have deleted this sentence and agree with the reviewer's comment.*

35. Extended Figure 1A: consider removing white lines around bars and annotating the clusters (e.g. MIX-cluster or *Haemophilus* (HAE)-cluster). Addendum: this was done for the full dendrogram-analysis I see; as suggested consider only presenting that analysis instead (i.e. running clusterin once).

*\*\*\* We have removed extended figure 1*

36. Extended Figure 1B: explain RCE on x-/y-axis; consider converting to a % of explained variance.

*\*\*\* We have removed extended figure 1*

37. Extended Figure 2: also see methods, please define pre-/post- symptom groups more explicitly.

*\*\*\* We have removed extended figure 2*

38. Extended Figure 3: in line with previous comments; be clear on LRTI vs LRTI-symptoms and whether there is a difference. Is there anything known on RTI symptoms in healthy controls?

*\*\*\* We have removed extended figure 3*

39. Extended Figure 4: I applaud the authors for adding both 2D and 3D NMDS-plots, yet a visualization of the 3D analysis in 2D (i.e. plotting NMDS1 vs NMDS2, NMDS3 vs NMDS2 and NMDS1 vs NMDS3), would be easier to interpret.

\*\*\* *We have removed extended figure 4*

40. Results/analysis three: this section (especially the first sections) is generally difficult to follow, it seems partly mixed with methods and seems to discuss several analyses at the same time. Please try to restructure this section, discussing any analyses one-by-one.

\*\*\**We have shortened and re-organized this section*

41. Figure 2: methods are lacking information on how clinical variables were projected into ordination space (envFit?).

\*\*\* *The Methods section already includes this detail: "... used the vegan's envfit function to project the age and LRTI status of each sample into the NMDS ordination"*

42. Figure 3: why did the authors choose to report these two genera? Please provide a rationale, for example based on other analysis supporting this choice. Also, it would be helpful to know/visualize the number of samples at each time point. Did the authors consider running any statistics on these results?

\*\*\* *We focused on Dolosigranulum because it has previously been identified as a marker of a healthy NP ecosystem, and because our analysis showed a strong signal about changes in the abundance of this pathogen. For Anaerobacillus, this was a frequent and novel finding within the case infants.*

43. Results/analysis three: the statement 'Concentrating on the samples taken prior to infection (and prior to antibiotic administration), this analysis confirmed lower relative abundance of *Dolosigranulum*, and higher relative abundance of *Anaerobacillus* in the LRTI infants before their infection.', although very interesting, is not supported sufficiently by the authors' analyses I believe. Time points prior to infection are likely related to earlier age and therefore higher *Dolosigranulum*, while time points after infections are related to older age and lower *Dolosigranulum*. Therefore, the pattern observed may merely reflect age dynamics. To support the statement the authors are trying to make, it may be interesting to look at time points before/after infection vs and age-matched control samples. Or include age when modeling these effects. Apart from that, from eyeballing this figure, I would also conclude the *Dolosigranulum* dramatically drops upon infection. In fact, it could be both, i.e. lower *Dolosigranulum* prior to infection and a more significant drop after infection. In addition, also given the question of the authors in the discussion ('But are these microbial profiles a result of the infection? Or were they present before the infection?'), it would make sense to expand/deepen this analysis further, including pathobionts like *Haemophilus*/*Streptococcus*/*Moraxella*.

\*\*\* *The reviewer is correct that our results are not definitive. Confirmatory studies will be required to see if this finding can be replicated in other studies/contexts. We make no claims about the interpretation of Dolosigranulum post LRTI diagnosis as this will certainly be influenced by antibiotic exposure. That effect is likely to be profound and of course unsurprising, and is not the key finding of our analysis, which focused on events preceding the LRTI event. We*

*clarify that these models do control for age (see NMDS plots). The effect of age is controlled for by the inclusion of age-matched comparator samples within the time series analysis. This seeks to control for potential confounder introduced by the fact that LRTI occurred at different ages, as the reviewer points out.*

44. Results/analysis four: again, this section includes too many details on methods in my view.

*\*\*\* We have rearranged this section to include results and moved the methods to the method section*

45. Results/analysis four: the authors refer to table 4, yet this seems to not include data on clustering. Should this be Figure 4?

*\*\*\* We apologize for not being clearer. The references are correct as stated, but we have changed the sentence to improve clarity*

46. Results/analysis four: the authors present many dendrograms/clustering in their paper (LRTI only, healthy only and all, both clustered based on two indices). If possible, I would advise to combine these analyses and properly name clusters according to the most dominant taxon, so that throughout the paper, the same clusters are discussed. In addition, if clustering based on Frey's index does not add anything to the message the authors want to convey, I would omit it. Similarly, consider presenting one figure showing per-individual microbiota profiles over time (grouped by LRTI/non-LRTI; currently Extended Figure 3 and 5).

*\*\*\* We appreciate the reviewer's comments and have adjusted the results accordingly, and omitted the extended figures for simplification.*

47. Table 5: I suggest to rerun this analysis using cluster/no clusters as response variable and LRTI status + age as predictors.

*\*\*\* We thank the reviewer for this excellent suggestion. We indeed plan to perform such analysis once we increase our sample size sufficiently to provide meaningful statistical power for such an approach.*

48. Results/analysis five: very interesting analysis, yet results are presented without any information on effect size. Also, it is not clear what number of genera are tested and whether other genera were also significantly different. Please expand on this analysis a bit further if possible.

*\*\*\* Regarding maternal data – this is a critical issue and our findings are provoking and intriguing. We chose not to extend on this data in the manuscript and to investigate this topic further separately. We have added a figure of the mothers microbiome (Figure 5).*



49. Results/general: did the authors consider running further analyses on symptomatic/non-LRTI samples? \*\*\**Yes, this was a pilot study, and we are currently working on this sample library looking into different respiratory viruses, HIV exposed infants, mildly symptomatic children, and more.*

50. Discussion: make sure all relevant papers are cited, among others Man *et al.* (2019<sup>1</sup>) is currently missing, while this to date is one of the largest studies on (severe) LRTI and NP microbiota.

\*\*\* *We have added relevant studies, including the suggested important study by Man et al*

51. Discussion: I think some statements are insufficiently supported by the data; for example ‘... clear evidence of dysbiosis preceding the onset of LRTI.’ This seems based on figure 3, where no statistical support was provided. Also, as said ‘dysbiosis’ is a bit vague, notably, it seems the authors observed specific differentially abundant genera between LRTI vs no LRTI. Same goes for ‘... we observed different microbiome patterns’, this seems based on analysis five, which requires more detail.

\*\*\* *We respectfully disagree with this comment. The Figure 3 is only one of many evidentiary features used to make our conclusions: so are the other figures, the mixed model and DESeq2 analyses, etc. There is clear evidence that (in our data) there is a difference between LRTI and control children, and that this precedes the LRTI.*

52. Discussion: ‘The NP microbiome at time of infection is associated with the risk of development of LRTI and its severity.’ This does not make sense to me, do the authors refer to NP microbiome prior to infection?

\*\*\**We have rephrased the sentence to correct it. The NP microbiome at time of infection and not prior to it has specific characteristics that are correlated with severity of LRTI.*

53. Discussion: ‘Thus, a similar association between the NP microbiome and risk of respiratory infections is a plausible theory for which there is precedent.’ In fact, this has been shown for NP microbiota in the context of mild infections by our group (Bosch *et al.*, 2017<sup>3</sup> and De Steenhuijsen Pijters *et al.*, 2022<sup>4</sup>, Nat Microbiol, 2022) and others (Teo *et al.*, 2015<sup>6</sup> and Teo *et al.*, 2018<sup>8</sup>).

\*\*\**We thank the reviewers for pointing this out, we added a sentence and the relevant references*

54. Discussion: discussion on the comparison/possible differences between Zambian children and existing European/Australian cohorts is lacking. This is one of the big strengths of this study; we need data from a wider range of cohorts and this study may contribute to that goal in my view. Another big strength of this study is that data on healthy microbiota are available from children who will develop a severe LRTI (this contrasts among others work from our group, where we assessed longitudinal microbiota data in context of mild respiratory infections).

*\*\*\* We appreciate and thank the reviewers for their comments. We have edited the discussion to reflect these comments.*

**Competing Interests:** No competing interests were disclosed.

---