

# Nasopharyngeal Dysbiosis Precedes the Development of Lower Respiratory Tract Infections in Young Infants, a Longitudinal Infant Cohort Study

Rotem Lapidot<sup>1,2\*</sup>, Tyler Faits<sup>3</sup>, Arshad Ismail<sup>4</sup>, Mushal Allam<sup>4</sup>, Zamantungwak T.H Khumalo<sup>4,5</sup>, William MacLeod<sup>6</sup>, Geoffrey Kwenda<sup>7</sup>, Zacharia Mupila<sup>8</sup>, Ruth Nakazwe<sup>9</sup>, Daniel Segrè<sup>10,11,12,13</sup>, W. 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>Division of Pediatric Infectious Diseases, Boston Medical Center, Boston

<sup>2</sup>Boston University School of Medicine, Boston

<sup>3</sup>Division of Computational Biomedicine, Boston University School of Medicine, Boston

<sup>4</sup>Sequencing Core Facility, National Institute for Communicable Diseases of the National Health Laboratory Service, South Africa

<sup>5</sup>Department of Veterinary Tropical Diseases, Faculty of Veterinary Science, University of Pretoria, South Africa

<sup>6</sup>Boston University School of Public Health, Department of global health, Boston

<sup>7</sup>Department of Biomedical Sciences, School of Health Sciences, University of Zambia, Lusaka Zambia

<sup>8</sup>Right to Care, 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

<sup>11</sup>Department of Biology, Boston University, Boston

<sup>12</sup>Department of Biomedical Engineering, Boston University, Boston

<sup>13</sup>Department of Physics, Boston University, Boston

**Corresponding author:** Rotem Lapidot, MD, MSCI Boston University Medical Campus  
[rotem.lapidot@bmc.org](mailto:rotem.lapidot@bmc.org)

**Keywords:** Lower Respiratory Tract Infection, Nasopharyngeal Microbiome, Dysbiosis, Longitudinal Cohort study

## This file includes:

Main Text

Tables 1 to 5

Figures 1 to 3

Supplemental Figures 1 to 3

39 **Abstract:**

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

44 **Methods:** We conducted an analysis of a longitudinal prospective cohort study of 1,981  
45 Zambian mother-infant pairs who underwent NP sampling from 1-week through 14-weeks of age  
46 at 2-3-week intervals. Ten of the infants in the cohort developed LRTI and were matched 3:1  
47 with healthy comparators. We completed 16S rRNA gene sequencing on the samples each of  
48 these infants contributed, as well as from baseline samples of the infants' mothers, and  
49 characterized the normal maturation of the healthy infant NP microbiome, compared to infants  
50 who developed LRTI.

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

60 **Conclusions:** Our results suggest that NP microbiome dysbiosis precedes LRTI in young infants  
61 and may be present in their mothers as well. Early dysbiosis may play a role in the causal  
62 pathway leading to LRTI or could be a marker of other pathogenic forces that directly lead to  
63 LRTI.

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

67

## 68 **Background**

69 Lower respiratory tract infections (LRTI), including pneumonia and bronchiolitis, are the leading  
70 cause of death in children under five years of age, accounting for 1.3 million deaths each year,  
71 with 81% concentrated in children 2 years or younger (Cao et al., 2019; Fischer Walker et al.,  
72 2013). A necessary step leading to LRTI is the acquisition of a respiratory pathogen, such as  
73 *Streptococcus pneumoniae*. However, pneumococcal carriage is nearly universal among infants,  
74 only a few of whom develop severe invasive disease (Balsells et al., 2018; Yildirim et al., 2017,  
75 2010). This indicates that the presence of the pathogen, while necessary, does not adequately  
76 address the more fundamental question of why some infants develop LRTI while most do not.

77 Increasingly, LRTI is seen as a consequence of the interaction between the pathogen and other  
78 contextual factors. Such factors include the net immune state of the host, intercurrent viral  
79 infections that may act transiently, or, in the case of HIV, for extended periods. Another factor  
80 may also be the microbial ecosystem in which the pathogen exists, *i.e.*, the nasopharyngeal  
81 microbiome. Looking at the microbiome as an ecosystem model considers individual members  
82 of that ecosystem to exist in some dynamic equilibrium characterized by reciprocal loops of  
83 interaction. As such, the interaction between the microbiome and a specific potential pathogen  
84 (*i.e.*, a pathobiont), could influence the behavior of that pathogen to either impede or promote  
85 LRTI (Brugger et al., 2016; Stewart et al., 2017).

86 In support of this ecosystem model, several cross-sectional studies have found that children with  
87 LRTIs often have distinct nasopharyngeal (NP) microbiome profiles at time of infection  
88 compared with healthy children. The NP microbiome profiles appear to be dominated by  
89 bacterial genera that differ between respiratory infections and health. For example, NP  
90 microbiomes dominated by *Streptococcus* and *Haemophilus* are associated with LRTI, whereas  
91 microbiomes profiles dominated by *Moraxella*, *Corynebacterium* and/or *Dolosigranulum*  
92 characterize healthy children. Further, NP microbiome characteristics correlate with the severity  
93 of respiratory disease and with clinical outcomes (de Steenhuijsen Piters et al., 2015; Hasegawa  
94 et al., 2017). While provocative, such observations largely rest on cross-sectional studies, and so  
95 cannot resolve the direction of cause and effect: we do not know whether these microbial profiles  
96 are a result of the infection or whether they preceded it. If the latter is true, then differences in

97 the NP microbiome could potentially represent a state of vulnerability, participating in a causal  
98 pathway leading to LRTI.

99 To draw such inferences, it is necessary to have longitudinal data, with sampling of infants  
100 before the development of the LRTI. Since LRTI is a rare event, collecting longitudinal data is  
101 complicated by the large number of infants needed to be followed. Between 2015 and 2016, our  
102 team conducted a prospective cohort study and was able to create a biological sample library that  
103 allowed a longitudinal analysis of this kind. The study took place in Lusaka, Zambia, among  
104 1,981 mother-infant pairs: The Southern Africa Mother Infant Pertussis Study – SAMIPS (Gill et  
105 al., 2016). The pairs were enrolled one-week post-partum. All enrolled infants were healthy and  
106 born term. At baseline, and every two-three weeks thereafter through 14 weeks of age, we  
107 obtained NP samples from mother and baby.

108 Within this cohort of 1,981 healthy infants a sub-set of 10 infants developed severe LRTI based  
109 on standard WHO clinical criteria (*Revised WHO classification and treatment of childhood  
110 pneumonia at health facilities • EVIDENCE SUMMARIES •*, n.d.). By comparing the infants who  
111 developed LRTI to matched healthy infants, we were able to conduct a time series analyses of  
112 NP microbiome of both infant populations, using 16S ribosomal DNA sequencing. We focused  
113 on the following fundamental analyses: 1) what is the ‘normal’ pattern of NP microbiome  
114 maturation over the first several months of life? 2) how does this contrast with the maturation of  
115 NP microbiome of infants who developed LRTI? 3) is there evidence that NP dysbiosis precedes  
116 the onset of LRTI? 4) are there distinct microbiome profiles that characterize sickness and health  
117 and other infant characteristics? 5) Is there also evidence of NP dysbiosis among the mothers of  
118 infants who later developed LRTI?

119

## 120 **Results**

121 Within the SAMIPS cohort we identified ten infants who developed LRTI during the study  
122 period as defined by the WHO clinical criteria: cough, cold and fast breathing, chest indrawing  
123 or other general danger signs (lethargy, difficulty feeding, persistent vomiting, and convulsions)  
124 (*Revised WHO classification and treatment of childhood pneumonia at health facilities •  
125 EVIDENCE SUMMARIES •*, n.d.). We then matched these case infants by season of birth,  
126 number of siblings and HIV exposure status, with healthy comparators. With ten infants with

127 LRTI and 3:1 matching, our analysis set consisted of 40 infants at ~7 time points each. All  
128 infants were born healthy via vaginal delivery. Male sex was more common in infants who  
129 developed LRTI ( $p= 0.067$ ). A third of infants with LRTI were born to mothers with HIV  
130 (receiving anti-retroviral treatment), compared to 40% of infants in the healthy group. Basic  
131 characteristics of the 40 infants are shown in **Table 1**. The symptoms and timing of sampling of  
132 the ten infants who developed LRTI are shown in **Table 2**.

133

Table 1. Characteristics of healthy infants and infants with LRTI

Characteristics	Healthy Infants (N=30)	Infants with LRTI (N=10)	All Subjects (N=40)	<i>p</i> *
Sex, n (%)				$p=0.067$
Females	16 (53.3%)	2 (20.0%)	18 (45.0%)	
Males	14 (46.7%)	8 (80%)	22 (55%)	
Season of enrollment				$p=0.224$
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)	$p=0.634$
HIV exposed, n (%)	13 (43.3%)	3 (30.0%)	16 (40.0%)	$p=0.456$
Mean number of samples collected (SE)	6.6 (0.2)	6.6 (0.6)	6.7 (0.2)	$p=0.958$

\*P values calculated using student t-test

134

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

Infants with LRTI	Sample Number (and age at sampling)								
	1	2	3	4	5	6	7	8	9
1.	7 days	27 days	42 days	62 days	79 days				
2.	7 days	27 days	35 days	42 days	59 days	73 days	88 days	104 days	
3.	7 days	11 days	62 days						
4.	7 days	19 days	45 days	60 days	68 days	107 days			
5.	7 days	28 days	42 days	56 days	69 days	84 days	100 days		
6.	7 days	21 days	42 days	56 days	59 days	73 days	87 days	96 days	103 days
7.	7 days	50 days	59 days	73 days	87 days	106 days			
8.	7 days	24 days	27 days	42 days	61 days	73 days	90 days	104 days	
9.	7 days	24 days	39 days	44 days	65 days				
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)  
 LRTI symptoms (cough/runny or blocked nose with or without fever AND labored breathing/poor feeding/ inwarding of the chest/lethargy)

135

136 *16S ribosomal DNA amplicon sequencing data and processing*

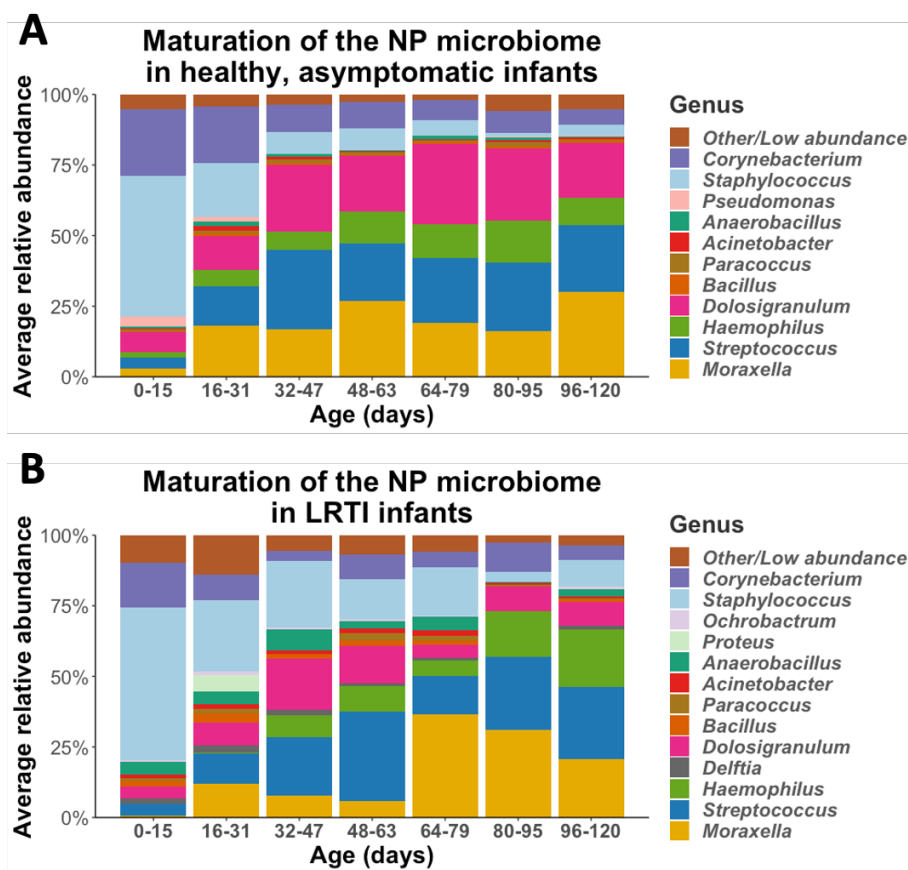
137 We successfully sequenced 265 NP swabs from 40 infants, capturing a median of seven samples  
 138 from each infant. The median age at first sampling was seven days, and the median age at final  
 139 sampling was 104 days. We also sequenced two NP swabs from each infant's mother at first and  
 140 last time points, for a total of 345 samples from mothers and infants combined. In six of these  
 141 samples, fewer than 10,000 reads aligned to RefSeq reference genomes and were excluded from  
 142 further analysis. The remaining 339 samples had a median of 101,979 reads per sample assigned  
 143 to reference genomes and were included in the analysis. From these, we detected 421 unique  
 144 genera, spanning 14 unique phyla, which were assigned at least 100 sequence reads across all  
 145 samples. Based on these results, we were confident in our ability to proceed with the ensuing  
 146 analyses.

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

149 Given our ultimate goal of identifying characteristics of the NP microbiome in infants who  
 150 develop LRTI, as a first step we describe the characteristics and evolution of NP microbiome of

151 the healthy infants. We analyzed the NP samples from all of the infants who remained free of  
152 LRTI through the end of observation, using linear regression to track changes in relative  
153 abundance of genera over time spanning the period between enrollment after birth and 14 weeks  
154 of age.

155 We observed a stepwise pattern of maturation as the infants aged, summarized in **Figure 1a**,  
156 showing the relative abundance of different genera across each age averaged stratum. As can be  
157 seen, there is a clear shift in the abundance of dominant genera with time, with some dominating  
158 early in life, and others becoming more prominent as the children aged. Early in life, the  
159 dominant genera were *Staphylococcus* and *Corynebacterium*. According to a mixed-effects model,  
160 these genera declined in relative abundance as infants aged (*Staphylococcus*:  $p < 10E-7$ ,  
161 *Corynebacterium*:  $p < 0.001$ ) and were replaced primarily by *Streptococcus* ( $p < 0.01$ )  
162 *Dolosigranulum* ( $p < 0.001$ ), *Moraxella* ( $p < 0.001$ ), and *Haemophilus* ( $p = 0.02$ ).



163

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

167

168 We did not measure any significant change in the alpha diversity (richness within a given  
169 sample) of NP microbiomes as healthy infants aged, measured either by Shannon index ( $p=0.32$ )  
170 or Chao1 index ( $p=0.15$ ). However, alpha diversity only reflects the number of dominant genera,  
171 and not whether the dominant genera are themselves diverse. Thus, when we clustered samples  
172 based on beta diversity (between sample diversity), measured as the Bray-Curtis dissimilarity  
173 between pairs of samples, we identified a distinct profile associated with samples from very  
174 young infants that contrasted against several profiles for more mature infant NPs (**Supplemental**  
175 **Figure 1a**). While each cluster is dominated by one or several of the most common genera, very  
176 few samples from healthy infants had high abundance of genera outside of the six most  
177 prominent genera. The primary axis of a Principal Coordinate Analysis (PCoA) (**Supplemental**  
178 **Figure 1b**) correlated with the age of the infants at the time of sampling, and stratified samples  
179 mainly by relative abundance of *Staphylococcus* and *Corynebacterium* in younger infants vs. the  
180 genera which were more common at older ages. The second PCoA axis distinguished between  
181 samples that were rich in *Moraxella* or *Dolosigranulum* from those rich in *Streptococcus* or  
182 *Haemophilus*. In summary, this analysis showed that the microbiomes of early infancy were  
183 highly dynamic over time, but that these shifts occurred in a structured and stereotypical pattern.

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

186 Given evidence from prior literature that during LRTIs the NP microbiome of infants is different  
187 than that compared to healthy infants, we set out to test whether the maturation of the NP  
188 microbiome in the first months of life is altered in infants who develop an LRTI. We repeated  
189 our analysis as described for healthy infants, stratifying into age groups and mapping the  
190 evolution of the NP microbiome over the first three months of life (**Figure 1b**). Infants who  
191 developed LRTI had similar general succession patterns as described for healthy infants, with  
192 high relative abundance of *Staphylococcus* early in life replaced by relative abundance of  
193 *Streptococcus*, *Haemophilus*, *Corynebacterium*, *Dolosigranulum* and *Moraxella*. Even though  
194 the general succession pattern of NP microbiome in infants with LRTI were similar to succession  
195 patterns of healthy infants, they exhibited distinct characteristics. Notably, the NP microbiome of  
196 infants who developed LRTI had, on average, higher relative abundance of specific genera

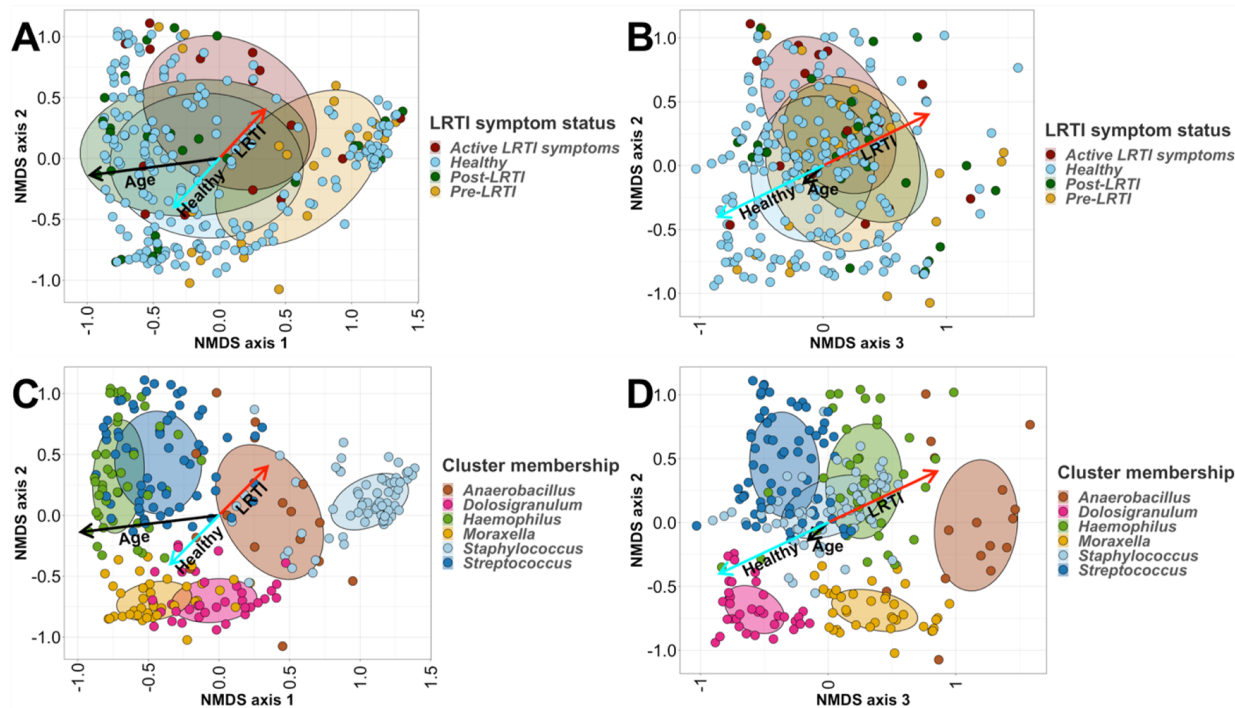


197 including *Bacillus* ( $p=0.05$ ) and *Delfia* ( $p<0.001$ ) and lower relative abundance of  
198 *Dolosigranulum* ( $p<0.001$ ).

199 As with the healthy control infants in our analysis 1, we did not observe any change in alpha  
200 diversity in LRTI infants as they aged (Shannon:  $p=0.08$ , Chao1:  $p=0.74$ ). Analysis of the beta  
201 diversity between LRTI infant samples once again revealed a cluster of samples taken at very  
202 early time points, dominated by *Staphylococcus*, with samples taken from older timepoints  
203 exhibiting profiles rich in *Streptococcus*, *Dolosigranulum*, *Moraxella*, and *Haemophilus*.  
204 However, in LRTI infants we also observed a large sixth cluster, characterized by a high  
205 abundance of *Anaerobacillus* as well as various other rare genera (**Supplemental Figure 2**).

206 Nonmetric multidimensional scaling (NMDS) scaling of the beta diversity dissimilarity matrix  
207 between all samples allows us to visualize more holistic structural differences in the NP  
208 communities of healthy vs LRTI infants (**Figure 2**). When we project the age and eventual LRTI  
209 status of the infants into the NMDS ordination space, we can see that age correlates closely the  
210 primary NMDS axis, whereas LRTI status is mostly correlated with the secondary and tertiary  
211 axes, indicating differences in NP microbiomes between healthy and LRTI infants independent  
212 of the aging process.

213



214

215 **Figure 2: Nonmetric multidimensional scaling (NMDS) ordination plots of all infants' (n=40)**  
 216 **nasopharyngeal (NP) samples.** We applied 3-dimensional NMDS ordination to the Bray-Curtis dissimilarity  
 217 matrix between all infants' NP swabs, and projected vectors into that ordination space representing the best fit  
 218 correlations for the age at sampling (the black arrows) and LRTI status (the cyan arrows represent control  
 219 infants, the red arrows represent LRTI infants). Plots A and C show the first two NMDS axes, while plots B  
 220 and D show the second and third axes. Samples in plots A and B are colored based on their LRTI symptom  
 221 progression, whereas samples in plots C and D are colored by their primary taxonomic profile cluster  
 222 membership (see Figure 3 for details). Age is highly correlated with the first NMDS axis, and samples on the  
 223 young end of the age vector mostly belong to the *Staphylococcus*-dominated profile, whereas samples on the  
 224 older end tend to belong more to the *Haemophilus* and *Moraxella*-dominated profiles (A). The  
 225 *Dolosigranulum*-dominated profile is associated with the healthy end of the vector for LRTI status, while the  
 226 *Anaerobacillus*-dominated profile is associated with disease (A,C).

227

228 Since each infant developed an LRTI at a different age, stratifying the infants into age groups  
 229 resulted in grouping together infants at different time points in relation to their disease – before  
 230 the LRTI, at time of the LRTI, and following the LRTI. In order to describe maturation of the  
 231 microbiome up to the time of LRTI, we created individual plots of the NP microbiome for each  
 232 one of the ten sick infants (**Supplemental Figure 3**). These emphasize the high degree of  
 233 heterogeneity of patterns over time across individuals.

234

235

236 *Analysis three: Is dysbiosis detectable at birth among infants who later develop LRTI?*

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

242 We used the R package DESeq2 to perform differential abundance tests. To be considered  
243 significant, a given genus would need to be differentially abundant with an FDR-adjusted p-  
244 value of less than 0.1 and also a mean relative abundance of at least 0.1% among either healthy  
245 or sick infants. We identified three options by which a genus could be different between the 2  
246 groups: First, a genus that was identified exclusively in infants who developed LRTI, such as  
247 *Novosphingobium* (4/10). Second, genera that were more common in infants with LRTI (but  
248 were present in both groups), such as *Delftia* (8/10 in LRTI infants vs 13/30 in healthy infants).  
249 And third, there were genera that were detected in both groups, but were present with higher  
250 relative abundance in infants with LRTI compared to the healthy infants, such as *Anaerobacillus*,  
251 *Bacillus*, *Blastococcus*, *Brachybacterium*, *Ochrobactrum*, *Ornithinimicrobium*, and  
252 *Sphingomonas*. Overall, ten genera were significantly different in infants who later developed  
253 LRTI at the first time point (**Table 3**). Notably, *Dolosigranulum*, which has been identified in  
254 prior studies as being associated with a healthy microbiome, as was the case among the healthy  
255 infants here, had significantly lower relative abundance in infants who developed LRTIs than in  
256 healthy counterparts prior to the LRTI and even at the first sample time point.

257

258

259

260

261

262

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

263

264

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

267 In order to identify specific microbial profiles, we applied hierarchical clustering to the Bray-  
 268 Curtis dissimilarity matrix between each pair of samples from all infants. The Bray-Curtis  
 269 dissimilarity matrix is a common tool in ecology for measuring the distance between different  
 270 populations in terms of beta-diversity, and is bounded between 0 and 1, spanning ‘no  
 271 dissimilarity’ to ‘complete dissimilarity’. We calculated at the genus level, using the *hclust*  
 272 package in R. We used the Silhouette and Frey clustering indexes (NbClust) to determine the  
 273 optimal heights at which to trim the dendrogram produced by the *hclust* function in R, splitting  
 274 our samples into six primary clusters (Silhouette index) and 13 sub-clusters (Frey index). These  
 275 six primary profiles were then named after the dominant genus within each cluster (the highest  
 276 relative abundance genus). This yielded the following clusters: *Staphylococcus* dominant  
 277 *Streptococcus* dominant, *Moraxella* dominant, *Dolosigranulum* dominant, *Haemophilus*  
 278 dominant, and *Anaerobacillus* dominant profiles, corresponding to six of the seven most  
 279 abundant genera across all our samples, as shown in **Table 4**. *Corynebacterium* is the only  
 280 highly-abundant genus that does not compose the majority (or plurality) of relative abundance  
 281 within any cluster; instead of being dominant in a subset of samples, *Corynebacterium* often co-

282 occurred alongside the more dominant *Staphylococcus*, or to a lesser extent *Dolosigranulum*. For  
 283 ease of reporting, we shall henceforth refer to each cluster by its most abundant genus.

284

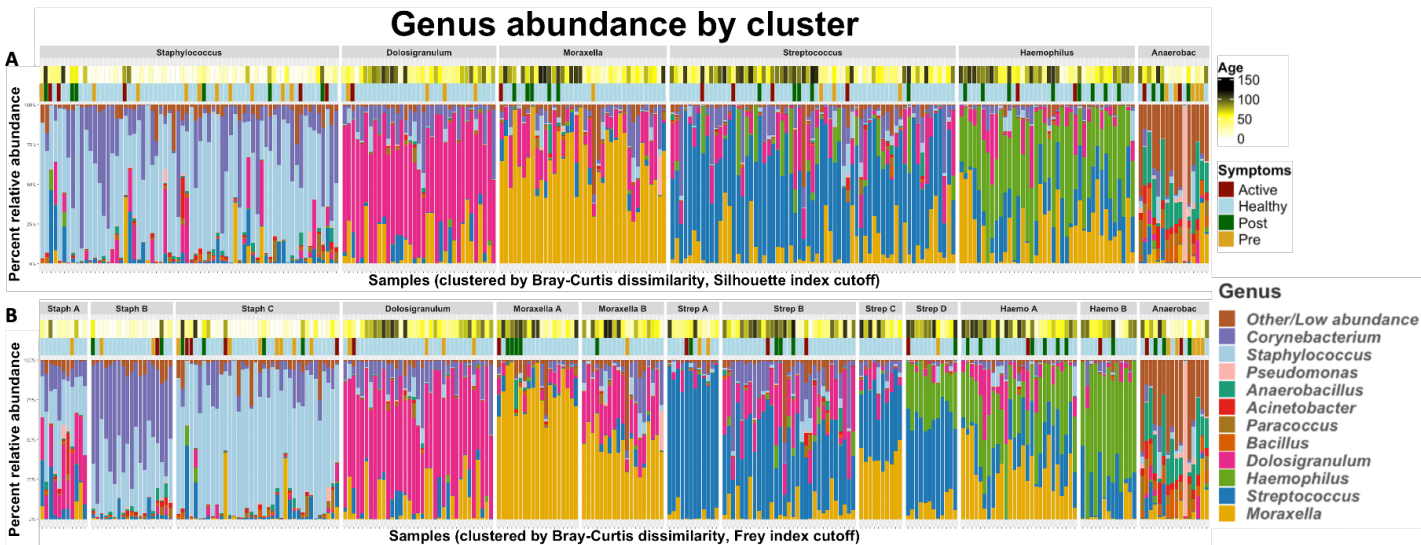
Table 4. Relative abundance and frequency of the most common genera observed in the NP microbiome of healthy control and LRTI infants

285

<b>Healthy infants</b>			
<b>Genus</b>	<b>Mean Relative Abundance</b>	<b>Frequency (Samples)</b>	<b>Frequency (Subjects)</b>
<i>Streptococcus</i>	19.4%	93.4%	100%
<i>Dolosigranulum</i>	19.1%	84.3%	100%
<i>Moraxella</i>	18.3%	77.8%	100%
<i>Staphylococcus</i>	14.0%	88.9%	100%
<i>Corynebacterium</i>	12.0%	94.4%	100%
<i>Paracoccus</i>	0.95%	66.7%	100%
<i>Acinetobacter</i>	0.84%	66.2%	100%
<i>Bacillus</i>	0.55%	52.5%	100%
<i>Anaerobacillus</i>	1.1%	53.0%	96.7%
<i>Pseudomonas</i>	0.89%	38.9%	93.3%
<i>Delftia</i>	0.32%	37.9%	90%
<i>Aeromonas</i>	0.13%	26.3%	80%
<i>Haemophilus</i>	8.7%	38.4%	73.3%
<i>Kocuria</i>	0.11%	18.7%	63.3%
<i>Ochrobactrum</i>	0.16%	16.2%	53.3%
<i>Escherichia</i>	0.25%	8.1%	40%
<i>Enterobacter</i>	0.18%	9.6%	40%
<i>Klebsiella</i>	0.12%	6.6%	30%
<b>LRTI infants</b>			
<b>Genus</b>	<b>Mean Relative Abundance</b>	<b>Frequency (Samples)</b>	<b>Frequency (Subjects)</b>
<i>Staphylococcus</i>	22.1%	90.8%	100%
<i>Streptococcus</i>	18.8%	93.8%	100%
<i>Moraxella</i>	14.8%	80%	100%
<i>Dolosigranulum</i>	9.6%	58.5%	100%
<i>Corynebacterium</i>	8.3%	86.1%	100%

286

287 **Figure 3** shows the microbial composition of each of the 262 infant samples in our study which  
288 passed sample quality filters, grouped by the six primary profiles (**Figure 3A**) and the 13 sub-  
289 profiles (**Figure 3B**).



290

291 **Figure 3:** The taxonomic profiles of all infant NP samples (n=40), clustered by pairwise Bray-Curtis  
292 dissimilarity. Clusters were defined by performing hierarchical clustering on the beta diversity matrix and then  
293 cutting the resulting dendrogram into an optimal number of clusters according to the A) Silhouette index (6  
294 clusters) and B) Frey index (13 clusters). The color bars above the stacked bar plots indicate the infants' ages  
295 at the time of each sample and their LRTI status – “healthy” indicates an infant which did not develop LRTI  
296 symptoms during our observation.

297

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

306

307

308

309

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

310

<b>Associations with LRTI status</b>			
<b>Cluster/Subcluster</b>	<b>Odds ratio estimate</b>	<b>Odds ratio range</b>	<b>P-value</b>
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
<b>Associations with age (in days)</b>			
<b>Cluster</b>	<b>Cluster</b>	<b>Difference in age</b>	<b>Adjusted P-value</b>
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

311

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

317 Together, these results reinforce a number of our previous observations; in particular, we can see  
 318 that there is a general trend for infant NP microbiome profiles to shift from being *Staphylococcus*  
 319 dominant shortly after birth towards several other profiles. We also see a clear dysbiotic pattern,  
 320 comprising higher than normal relative abundance of *Anaerobacillus* as well as higher  
 321 prevalence of rare genera which typically make up an extremely low portion of (or are  
 322 completely absent from) healthy NP microbiomes.

323

324

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

327 Observing distinct characteristics of the NP microbiome of infants as early as age 7 days,  
328 suggested that these profiles might be related to in-utero exposures, transmittable immunologic  
329 factors, and/or host genetics. That led us to question whether mothers of infants who develop  
330 LRTI have themselves distinct characteristics of the NP microbiome. We analyzed the first NP  
331 swabs from each of the mothers enrolled in our study taken at the infants' day seven enrollment  
332 visits, correlated their microbiomes to those of their infants, and used DESeq2 to establish which  
333 genera were differentially abundant between mothers of LRTI infants and mothers of healthy  
334 infants. Similar to the pattern seen in the infants themselves, the mothers of infants who  
335 developed LRTIs had significantly decreased relative abundance of *Dolosigranulum* ( $p=0.05$ ) as  
336 compared to mothers of healthy infants at 7 days of infant's life.

## 337 **Discussion**

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

348 The nasopharynx is the ecologic niche of respiratory pathobionts, and in this ecosystem they will  
349 either become invasive or remain merely colonizers. The NP microbiome at time of infection is  
350 associated with the risk of development of LRTI and its severity. But there is also good reason  
351 to believe that the maturation of the NP microbiome in the first months of life, and not only its  
352 characteristics at time of infection, is associated with respiratory health and development of  
353 disease later in life. For example, maturation of the gut microbiome is known to regulate the  
354 immune system evolution and is associated with the development of diseases later in life such as



355 obesity and type 1 diabetes (Bokulich et al., 2016; Stewart et al., 2018). Gut microbial dysbiosis  
356 in children often predisposes to recurrent *C. difficile* infections (Ihekweazu and Versalovic,  
357 2018). Thus, a similar association between the NP microbiome and risk of respiratory infections  
358 is a plausible theory for which there is precedent.

359 We have shown that we can characterize the normal, healthy maturation of the NP microbiome  
360 over the first months of life, and how this maturation is different in infants who develop early  
361 LRTI. While the evolution of the normal microbiome is highly dynamic, it proceeds in a  
362 stereotypical fashion, with stepwise shifts from a flora dominated by skin organisms  
363 (*Staphylococcus*), to one that is dominated by genera more typically associated with the  
364 respiratory tract (*Dolosigranulum*, *Streptococcus*, *Haemophilus* and *Moraxella*). Similar  
365 microbial succession patterns were previously described in other birth cohorts (Biesbroek et al.,  
366 2014).

367 By contrast, infants who develop LRTI have similar general succession patterns as healthy  
368 infants; transitioning from high relative abundance of *Staphylococcus* to high relative abundance  
369 of genera associated with the respiratory tract, but unlike healthy infants, the evolution of their  
370 NP microbiome is characterized by low relative abundance of specific genera associated with  
371 ‘health’, such as *Dolosigranulum*, and high relative abundance of other genera that appear  
372 unique, such as *Anaerobacillus*, *Bacillus*, and a mixture of ‘other’ uncommon genera.  
373 Additionally, the LRTI infants’ microbiomes include a larger number of uncommon and  
374 transient genera, presenting a picture that is more chaotic than what is seen in the healthy infants.

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

382 We were able to identify several microbiome profiles which appear to cluster by chronological  
383 age, LRTI and health. Our results indicate that young infants who developed LRTI, had NP  
384 microbiome dysbiosis prior to acquiring the infection, and as early as 7 days of life. These

385 infants have NP microbiome enriched with *Aneorobaccillus/Bacillus*, *Acinetobacter*, and other  
386 uncommon/unspecified genera, and also have relatively lower abundance of *Dolosigranulum*.  
387 Our intriguing results suggest that their mothers NP microbiome at the same early time point also  
388 differed from that of mothers of healthy infants.

389 The interaction between host, microbiome and pathobionts is complex and most probably  
390 multidirectional. The NP microbiome, known to be associated with environmental factors  
391 (breastfeeding, mode of delivery) (Bosch et al., 2017; Brugger et al., 2016). could also very well  
392 be a reflection or marker, of host genetics and immune system function, which would explain  
393 why so early in life “high risk” profiles are observed. New acquisition of a pathobiont in the  
394 nasopharynx initiates interactions between the pathobiont and other organisms residing in the  
395 nasopharynx. These interactions modify metabolic activity and gene expressions of the  
396 pathobiont that influence whether the pathobiont becomes invasive. The interactions themselves  
397 between organisms in the nasopharynx also modify host immune response which underscores the  
398 complex relationship between host, microbiome and pathogens (de Steenhuijsen Piters et al.,  
399 2019).

400 The key unresolved question is what role dysbiosis plays in the causal pathway leading to  
401 pneumonia: is dysbiosis a marker of other unobserved forces that lead to pneumonia, such as  
402 underlying host genetic or immunologic factors? Or does dysbiosis play a role in the causal  
403 pathway leading to LRTI? While our data cannot resolve this question, the implication of our  
404 findings are substantial. Our findings suggest that NP dysbiosis identified in the first days of life  
405 is associated with higher risk of developing LRTI in early infancy. This suggests that there is an  
406 important window of opportunity for identifying these infants and intervene. According to our  
407 findings, it may even be that we can identify these infants, by examining the mothers.

408 Our study has several limitations. Infants were followed until the age of three months, and thus  
409 our findings could not be generalized to older age groups. On the other hand, it is possible that  
410 infants included in our healthy control group developed LRTI after the study period, in that case  
411 our results are biased towards the null, possibly underestimating differences between the two  
412 groups.

413 A further limitation is that we do not know the causative pathogen of the LRTIs, and whether  
414 these were viral, bacterial, or mixed pathogen LRTIs. LRTI is a heterogeneous set of conditions,

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

#### 430 **Conclusions:**

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

#### 437 **Materials and Methods**

##### 438 **Study population**

439 This is a nested time-series case comparator study within the prospective longitudinal Southern  
440 Africa Mother-Infant Pertussis study (SAMIPS). SAMIPS was a study conducted in Zambia in  
441 which infants and their mothers were followed over the first 3 months of life. Full methods  
442 description is previously detailed by Gill et al (Gill et al., 2016), in short: All infants enrolled to  
443 SAMIPS were less than ten days of age, born term, via normal vaginal delivery, and deemed

444 healthy after birth. All infants received scheduled vaccines. Written informed consent was  
445 obtained as appropriate from mothers of infants enrolled in the study.

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

#### 449 **Study design**

450 Mother-infant pairs were enrolled when mothers returned for their first postpartum well-child  
451 visit at one week of age. At enrollment, and 2-3-week intervals thereafter, through 14 weeks, we  
452 obtained a posterior nasopharyngeal (NP) swab from both mother and baby, with additional  
453 swabs obtained adventitiously if either returned seeking care for an acute respiratory infection.

454 Within the SAMIPS cohort, we identified ten infants who during the study period suffered from  
455 symptoms of lower respiratory tract infection (LRTI) as adopted from the WHO (*Revised WHO*  
456 *classification and treatment of childhood pneumonia at health facilities • EVIDENCE*  
457 *SUMMARIES* •, n.d.). Sick infants were matched 1:3 with healthy comparators by season of  
458 enrollment, maternal age and household composition.

#### 459 **Sample processing and storage**

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

#### 467 **16S ribosomal DNA amplification and MiSeq sequencing.**

468 For 16S library preparations, two PCR reactions were completed on the template DNA. Initially  
469 the DNA was amplified with primers specific to the V3–V4 region of the 16S rRNA gene  
470 (Klindworth et al., 2013). The 16S primer pairs incorporated the Illumina overhang adaptor (16S  
471 Forward primer

472 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3';

473 16S reverse primer

474 5'-

475 GTCTCGTGGGCTCGGAGATGTGTATAAAGAGACAGGACTACHVGGGTATCTAATCC-

476 3')

477 Each PCR reaction contained DNA template (~12 ng), 5 $\mu$ l forward primer (1 $\mu$ M), 5  $\mu$ l reverse  
478 primer (1 $\mu$ M), 12.5  $\mu$ l 2 X Kapa HiFi Hotstart ready mix (KAPA Biosystems Woburn, MA),  
479 and PCR grade water to a final volume of 25 $\mu$ l. PCR amplification was carried out as follows:  
480 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  
481 72°C for 5 min and held at 4°C. Negative control reactions without any template DNA were  
482 carried out simultaneously.

483 PCR products were visualized using Agilent TapeStation (Agilent Technologies, Germany).  
484 Successful PCR products were cleaned using AMPure XP magnetic bead-based purification  
485 (Beckman Coulter, IN). The IDT for Illumina Nextera DNA UD Indexes kit (Illumina, San  
486 Diego, CA) with unique dual index adapters were used to allow for multiplexing. Each PCR  
487 reaction contained purified DNA (5  $\mu$ l), 10  $\mu$ l index primer mix, 25  $\mu$ l 2X Kapa HiFi Hot Start  
488 Ready mix and 10  $\mu$ l PCR grade water. PCR reactions were performed on a Bio-Rad C1000  
489 Thermal Cycler (Bio-Rad, Hercules, CA) Cycling conditions consisted of one cycle of 95°C for  
490 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  
491 final extension cycle of 72°C for 5 min.

492 Prior to library pooling, the indexed libraries were purified with Ampure XP beads and  
493 quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA).  
494 Purified amplicons were run on the Agilent TapeStation (Agilent Technologies, Germany) for  
495 quality analysis before sequencing. The sample pool (2 nM) was denatured with 0.2N NaOH,  
496 then diluted to 4 pM and combined with 10% (v/v) denatured 20 pM PhiX, prepared following  
497 Illumina guidelines. Samples were sequenced on the MiSeq sequencing platform at the NICD  
498 Sequencing Core Facility, using a 2 x 300 cycle V3 kit, following standard Illumina sequencing  
499 protocols.

500 Sequencing data were processed using QIIME2 (Bolyen et al., 2019) and Pathoscope2 (Hong et  
501 al., 2014). Samples with less than 10,000 reads were excluded from further analysis.

502

### 503 **Data processing**

504 We assessed the quality of the sequencing data using FastQC (Andrews, 2010), which indicated  
505 that the overall sequencing quality was excellent, with mean Phred quality scores remaining  
506 greater than 30 (>99.9% accuracy) for over 200bp for both forward and reverse reads. We used  
507 *Trimmomatic* (Bolger et al., 2014) to trim Illumina adapters and remove low-quality sequences,  
508 setting the tool's parameters to LEADING:6, TRAILING:6, SLIDINGWINDOW:6:15, and  
509 MINLEN:36. This quality filtering removed less than 0.5% of reads from each sample.

510 We used PathoScope 2 to assign sequencing reads to bacterial genomes. We used all of RefSeq's  
511 representative bacterial genomes (downloaded November 2, 2018) as a PathoScope reference  
512 library. From PathoScope's subspecies-level final best hit read numbers, we compiled counts  
513 tables and relative abundance tables for each sample at the phylum, genus, and where possible, to  
514 the species level.

### 515 **Data and statistical analysis**

#### 516 NP microbiome characteristics and evolution over time

517 We describe the normal evolution of the NP microbiome in healthy infants over the first three  
518 months of life. We calculated microbial richness using Chao1 index, and diversity of microbial  
519 taxa using the Shannon diversity index. We report the individual evolution of NP microbiome of  
520 each of the 10 infants who develop LRTI. In order to establish statistical significance, we used  
521 the *lmer* function from the *lme4* package for R (Bates et al., 2015) to apply a mixed-effects linear  
522 model to the log counts per million (logCPM) value of each genus, including age and HIV  
523 exposure as fixed effects and the study subject as a random effect. All p-values generated by  
524 these linear models are reported after False Discovery Rate (FDR) adjustment for multiple  
525 comparisons using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). We only  
526 generated mixed-effects models for genera which had an average relative abundance of at least  
527 0.5% across all healthy infant samples.

528 For visualization of the development of healthy NP microbiota, we grouped all infant samples by  
529 age (in days) into 7 bins, each comprising a 16-day age window (0-15 days, 16-31 days, etc). We  
530 only visualized genera which had an average relative abundance of at least 1% across all

531 samples. The relative abundances of all genera which did not meet this threshold were summed  
532 into a group labelled “Other/Low abundance” for plotting purposes only.

533 We calculated estimates of the alpha diversity within each sample based on the species-level  
534 counts tables generated by PathoScope 2. We calculated alpha diversity using two methods: the  
535 Chao1 index, which estimates the total number of species present within a sample, and the  
536 Shannon index, an entropy-based metric which incorporates both the number of species present  
537 and the evenness of abundance among those species. The Chao1 index was calculated using the  
538 R package *fossil* (Vavrek, 2011) and the Shannon index was calculated using the R package  
539 *vegan* (available via CRAN) (Oksanen et al., 2019) each with a rarefaction depth of 10,000. We  
540 constructed a mixed-effects linear model as described above, except using each alpha diversity  
541 metric as a response variable, in order to test whether alpha diversity changed as infants aged.

#### 542 Analysis of the association between the NP microbiome and the development of LRTI

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

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

#### 552 Differential abundance analysis at first timepoints

553 We performed differential abundance between the first samples from healthy and LRTI infants  
554 using the R package *DESeq2* (Love et al., 2014) available via Bioconductor (Huber et al., 2015).  
555 We imported our unnormalized genus counts table compiled from PathoScope2 as a  
556 *DESeqDataSet* and ran the function *DESeq*, using a design model that included infants’ HIV  
557 exposure (from an HIV infected mother) as a covariate. For microbiome data, *DESeq2* has been  
558 shown to return lower false discovery rates than other differential tests (McMurdie and Holmes,  
559 2014), and performs particularly well for smaller experiments (Weiss et al., 2017).

560 To test whether the presence or absence of certain genera at the first sampled timepoint were  
561 associated with LRTI, we performed Fisher's exact test to determine if healthy and LRTI infants  
562 are equally likely to have each genus in their NP microbiome. Because very low-abundance  
563 genera could be the result of spurious alignments or contamination, we explored both a high  
564 threshold (>1% relative abundance) and a low threshold (>0.1% relative abundance) for defining  
565 presence of a genus.

#### 566 Beta diversity and clustering

567 We computed a Bray-Curtis dissimilarity matrix between samples using *vegan*'s *vegdist*  
568 function. When applied to relative abundance values, Bray-Curtis dissimilarity between two  
569 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  
570 of genus  $n$  in sample  $i$ . We performed hierarchical clustering of samples based on this  
571 dissimilarity matrix using R's *hclust* function with the method set to "ward.D". We defined  
572 clusters using R's *cutree* function, with the value for  $k$  selected by maximizing the Silhouette  
573 and Frey indexes as calculated by the package *NbClust* (Charrad et al., 2014). For each cluster,  
574 we performed Fisher's exact tests to determine whether that cluster was enriched for LRTI  
575 samples generally, pre-symptomatic samples, active symptom samples, or HIV-exposed samples.

576 We used the *metaMDS* function from the R package *vegan* to perform non-metric  
577 multidimensional scaling (NMDS) ordination on our Bray-Cutris dissimilarity matrix, using as  
578 parameters  $k=3$ ,  $try=50$ , and  $trymax=1000$ . Scaling our data onto just two dimensions using  
579 NMDS yielded a stress value greater than 0.2, indicating a poor fit; we instead scaled the data  
580 onto three dimensions (stress=0.13), and used the *vegan*'s *envfit* function to project the age and  
581 LRTI status of each sample into the NMDS ordination. 2-dimensional plots of our NMDS  
582 ordinated data can be found in **Figure 2**, and a 3-dimensional plot can be found in **Supplemental**  
583 **Figure 4**.

#### 584 Differential analysis of maternal NP microbiomes

585 We used Spearman correlation coefficients to verify that the composition of infant NP  
586 microbiomes is related to their mother's NP microbiome. We chose Spearman correlation, which  
587 utilizes rank order rather than continuous values, due to the compositional nature of bacterial  
588 abundance data. We calculated Spearman's  $\rho$  for the relative abundance of each genus between  
589 mothers and their infants. We tested the significant of these correlations by comparing the



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

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

598 **Declarations:**

599 **Ethics approval and consent to participate:** The study was approved by the ethical review  
600 committees at the ERES Converge IRB in Lusaka, Zambia, and at Boston University Medical  
601 Center. All mothers provided written informed consent, with consent provided in English,  
602 Bemba or Nyanja as preferred by the participant.

603 **Consent for publication:** N/A

604 **Availability of data and materials:** The raw and processed sequencing data from this study are  
605 available in the SRA repository, under accession number pending. Furthermore, all code,  
606 processed data, and the sample information metadata are available in the following GitHub  
607 repository: [https://github.com/tfaits/Infant\\_Nasopharyngeal\\_Dysbiosis](https://github.com/tfaits/Infant_Nasopharyngeal_Dysbiosis). Taxon counts tables are  
608 called "species.RDS", "genus.RDS", and "phylum.RDS". For strain/subspecies-level counts,  
609 "PathoScopeTable.txt" has the unfiltered/unprocessed outputs from PathoScope.

610 **Competing interest:** All authors declare no competing interests.

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

614 **Acknowledgments:** Not applicable.

615

616

617

## 618 **References**

- 619
- 620 Andrews S. 2010. FastQC. *Babraham Bioinforma*. <https://doi.org/citeulike-article-id:11583827>
- 621 Balsells E, Dagan R, Yildirim I, Gounder PP, Steens A, Muñoz-Almagro C, Mameli C, Kandasamy R,  
622 Givon Lavi N, Daprai L, van der Ende A, Trzciński K, Nzenze SA, Meiring S, Foster D, Bulkow  
623 LR, Rudolph K, Valero-Rello A, Ducker S, Vestheim DF, von Gottberg A, Pelton SI, Zuccotti GV,  
624 Pollard AJ, Sanders EAM, Campbell H, Madhi SA, Nair H, Kyaw MH. 2018. The relative invasive  
625 disease potential of *Streptococcus pneumoniae* among children after PCV introduction: A systematic  
626 review and meta-analysis. *J Infect* **77**:368–378. doi:10.1016/j.jinf.2018.06.004
- 627 Bates D, Mächler M, Bolker BM, Walker SC. 2015. Fitting linear mixed-effects models using lme4. *J*  
628 *Stat Softw* **67**. doi:10.18637/jss.v067.i01
- 629 Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful  
630 Approach to Multiple Testing. *J R Stat Soc Ser B* **57**:289–300. doi:10.1111/j.2517-  
631 6161.1995.tb02031.x
- 632 Biesbroek G, Bosch AATM, Wang X, Keijser BJB, Veenhoven RH, Sanders EAM, Bogaert D. 2014. The  
633 impact of breastfeeding on nasopharyngeal microbial communities in infants. *Am J Respir Crit Care*  
634 *Med* **190**:298–308. doi:10.1164/rccm.201401-0073OC
- 635 Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, Lieber AD, Wu F, Perez-Perez GI, Chen  
636 Y, Schweizer W, Zheng X, Contreras M, Dominguez-Bello MG, Blaser MJ. 2016. Antibiotics, birth  
637 mode, and diet shape microbiome maturation during early life. *Sci Transl Med* **8**.  
638 doi:10.1126/scitranslmed.aad7121
- 639 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data.  
640 *Bioinformatics* **30**:2114–2120. doi:10.1093/bioinformatics/btu170
- 641 Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ,  
642 Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT,  
643 Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC,  
644 Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM,  
645 Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H,  
646 Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang K Bin, Keefe  
647 CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolk T, Kreps J, Langille MGI, Lee J, Ley R,  
648 Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ,

- 649 Melnik A V., Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF,  
650 Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A,  
651 Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford  
652 AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ,  
653 Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J,  
654 Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R,  
655 Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science  
656 using QIIME 2. *Nat Biotechnol*. doi:10.1038/s41587-019-0209-9
- 657 Bosch AATM, De Steenhuijsen Piters WAA, Van Houten MA, Chu MLJN, Biesbroek G, Kool J, Pernet  
658 P, De Groot PKCM, Eijkemans MJC, Keijser BJF, Sanders EAM, Bogaert D. 2017. Maturation of  
659 the infant respiratory microbiota, environmental drivers, and health consequences. *Am J Respir Crit  
660 Care Med* **196**:1582–1590. doi:10.1164/rccm.201703-0554OC
- 661 Brugger SD, Bomar L, Lemon KP. 2016. Commensal–Pathogen Interactions along the Human Nasal  
662 Passages. *PLOS Pathog* **12**:e1005633. doi:10.1371/journal.ppat.1005633
- 663 Cao B, Ho J, Retno Mahanani W, Louise Strong World Bank Group Emi Suzuki K, Andreev K,  
664 Bassarsky L, Gaigbe-Togbe V, Gerland P, Gu D, Hertog S, Li N, Spoorenberg T, Ueffing P,  
665 Wheldon M, Bay G, Cruz Castanheira H, Alkema L, Black R, Hopkins J, Guillot M, Hill K,  
666 Pedersen J, Jon Wakefield F, Liu L, Perin J, Villavicencio F, Yeung D, Ganesh Director V, Zhang  
667 Y, Hancioglu A, Avanesyan K, Bania S, Carter K, Carvajal L, Coskun Y, Delamónica E, Hanafy A,  
668 Hassfurter K, Jaques Y, Khan S, Kumapley R, Noeva R, Olivetti D, Quintana E, Ranck A, Requejo  
669 J, Unalan T, Young U. 2019. London School of Hygiene & Tropical Medicine Trevor Croft, The  
670 Demographic and Health Surveys (DHS) Program, ICF.
- 671 Charrad M, Ghazzali N, Boiteau V, Niknafs A. 2014. Nbclust: An R package for determining the relevant  
672 number of clusters in a data set. *J Stat Softw* **61**:1–36. doi:10.18637/jss.v061.i06
- 673 de Steenhuijsen Piters WAA, Heinonen S, Hasrat R, Bunsow E, Smith B, Suarez-Arrabal M-C,  
674 Chaussabel D, Cohen DM, Sanders EAM, Ramilo O, Bogaert D, Mejias A. 2016. Nasopharyngeal  
675 Microbiota, Host Transcriptome, and Disease Severity in Children with Respiratory Syncytial Virus  
676 Infection. *Am J Respir Crit Care Med* **194**:1104–1115. doi:10.1164/rccm.201602-0220OC
- 677 de Steenhuijsen Piters WAA, Jochems SP, Mitsi E, Rylance J, Pojar S, Nikolaou E, German EL,  
678 Holloway M, Carniel BF, Chu MLJN, Arp K, Sanders EAM, Ferreira DM, Bogaert D. 2019.  
679 Interaction between the nasal microbiota and *S. pneumoniae* in the context of live-attenuated  
680 influenza vaccine. *Nat Commun* **10**:1–9. doi:10.1038/s41467-019-10814-9

- 681 de Steenhuijsen Piters WAA, Sanders EAM, Bogaert D. 2015. The role of the local microbial ecosystem  
682 in respiratory health and disease. *Philos Trans R Soc B Biol Sci*. doi:10.1098/rstb.2014.0294
- 683 Fischer Walker CL, Rudan I, Liu L, Nair H, Theodoratou E, Bhutta ZA, O'Brien KL, Campbell H, Black  
684 RE. 2013. Global burden of childhood pneumonia and diarrhoea. *Lancet*. doi:10.1016/S0140-  
685 6736(13)60222-6
- 686 Gill CJ, Mwananyanda L, MacLeod W, Kwenda G, Mwale M, Williams AL, Siazeele K, Yang Z,  
687 Mwansa J, Thea DM. 2016. Incidence of severe and nonsevere pertussis among HIV-exposed and-  
688 unexposed zambian infants through 14weeks of age: Results from the southern Africa mother infant  
689 pertussis study (samips), a longitudinal birth cohort study. *Clin Infect Dis* **63**:S154–S164.  
690 doi:10.1093/cid/ciw526
- 691 Hasegawa K, Linnemann RW, Mansbach JM, Ajami NJ, Espinola JA, Petrosino JF, Piedra PA, Stevenson  
692 MD, Sullivan AF, Thompson AD, Camargo CA. 2017. Nasal Airway Microbiota Profile and Severe  
693 Bronchiolitis in Infants: A Case-control Study. *Pediatr Infect Dis J* **36**:1044–1051.  
694 doi:10.1097/INF.0000000000001500
- 695 Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, Crandall KA, Johnson WE.  
696 2014. PathoScope 2.0: A complete computational framework for strain identification in  
697 environmental or clinical sequencing samples. *Microbiome* **2**. doi:10.1186/2049-2618-2-33
- 698 Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L,  
699 Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, Macdonald J,  
700 Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L,  
701 Morgan M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor HHS Public  
702 Access. *Nat Methods* **12**:115–121. doi:10.1038/nmeth.3252
- 703 Ihekweazu FD, Versalovic J. 2018. Development of the Pediatric Gut Microbiome: Impact on Health and  
704 Disease. *Am J Med Sci* **356**:413–423. doi:10.1016/j.amjms.2018.08.005
- 705 Kelly MS, Surette MG, Smieja M, Pernica JM, Rossi L, Luinstra K, Steenhoff AP, Feemster KA,  
706 Goldfarb DM, Arscott-Mills T, Boiditswe S, Rulaganyang I, Muthoga C, Gaofiwe L, Mazhani T,  
707 Rawls JF, Cunningham CK, Shah SS, Seed PC. 2017. The Nasopharyngeal Microbiota of Children  
708 with Respiratory Infections in Botswana. *Pediatr Infect Dis J* **36**:e211–e218.  
709 doi:10.1097/INF.0000000000001607
- 710 Klindworth A, Pruesse E, Schweer T, Rg Peplies J, Quast C, Horn M, Glö Ckner FO. 2013. Evaluation of  
711 general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based

- 712 diversity studies. *Nucleic Acids Res* **41**. doi:10.1093/nar/gks808
- 713 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq  
714 data with DESeq2. *Genome Biol* **15**:550. doi:10.1186/s13059-014-0550-8
- 715 Mansbach JM, Hasegawa K, Henke DM, Ajami NJ, Petrosino JF, Shaw CA, Piedra PA, Sullivan AF,  
716 Espinola JA, Camargo CA. 2016. Respiratory syncytial virus and rhinovirus severe bronchiolitis are  
717 associated with distinct nasopharyngeal microbiota. *J Allergy Clin Immunol* **137**:1909-1913.e4.  
718 doi:10.1016/j.jaci.2016.01.036
- 719 Mcmurdie PJ, Holmes S. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.  
720 *PLoS Comput Biol* **10**:1003531. doi:10.1371/journal.pcbi.1003531
- 721 Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, Minchin PR, O'hara RB,  
722 Simpson GL, Solymos P, Henry M, Stevens H, Szoecs E, Maintainer HW. 2019. Package “vegan”  
723 Title Community Ecology Package. *Community Ecol Packag* **2**:1–297.
- 724 Revised WHO classification and treatment of childhood pneumonia at health facilities • EVIDENCE  
725 SUMMARIES • n.d.
- 726 Stewart CJ, Ajami NJ, O'Brien JL, Hutchinson DS, Smith DP, Wong MC, Ross MC, Lloyd RE,  
727 Doddapaneni HV, Metcalf GA, Muzny D, Gibbs RA, Vatanen T, Huttenhower C, Xavier RJ,  
728 Rewers M, Hagopian W, Toppari J, Ziegler AG, She JX, Akolkar B, Lernmark A, Hyoty H, Vehik  
729 K, Krischer JP, Petrosino JF. 2018. Temporal development of the gut microbiome in early childhood  
730 from the TEDDY study. *Nature* **562**:583–588. doi:10.1038/s41586-018-0617-x
- 731 Stewart CJ, Mansbach JM, Wong MC, Ajami NJ, Petrosino JF, Camargo CA, Hasegawa K. 2017.  
732 Associations of nasopharyngeal metabolome and microbiome with severity among infants with  
733 bronchiolitis: A multiomic analysis. *Am J Respir Crit Care Med* **196**:882–891.  
734 doi:10.1164/rccm.201701-0071OC
- 735 Vavrek MJ. 2011. fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontol Electron*  
736 **14**:16.
- 737 Weisberg JF and S. 2019. An R Companion to Applied Regression, Third. ed. Thousand Oaks (CA):  
738 SAGE Publications.
- 739 Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-  
740 Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential  
741 abundance strategies depend upon data characteristics. *Microbiome* **5**:1–18. doi:10.1186/s40168-  
742 017-0237-y

743 Yildirim I, Hanage WP, Lipsitch M, Shea KM, Stevenson A, Finkelstein J, Huang SS, Lee GM,  
744 Kleinman K, Pelton SI. 2010. Serotype specific invasive capacity and persistent reduction in  
745 invasive pneumococcal disease. *Vaccine* **29**:283–8. doi:10.1016/j.vaccine.2010.10.032

746 Yildirim I, Little BA, Finkelstein J, Lee G, Hanage WP, Shea K, Pelton SI. 2017. Surveillance of  
747 pneumococcal colonization and invasive pneumococcal disease reveals shift in prevalent carriage  
748 serotypes in Massachusetts’ children to relatively low invasiveness. *Vaccine* **35**:4002–4009.  
749 doi:10.1016/j.vaccine.2017.05.077

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

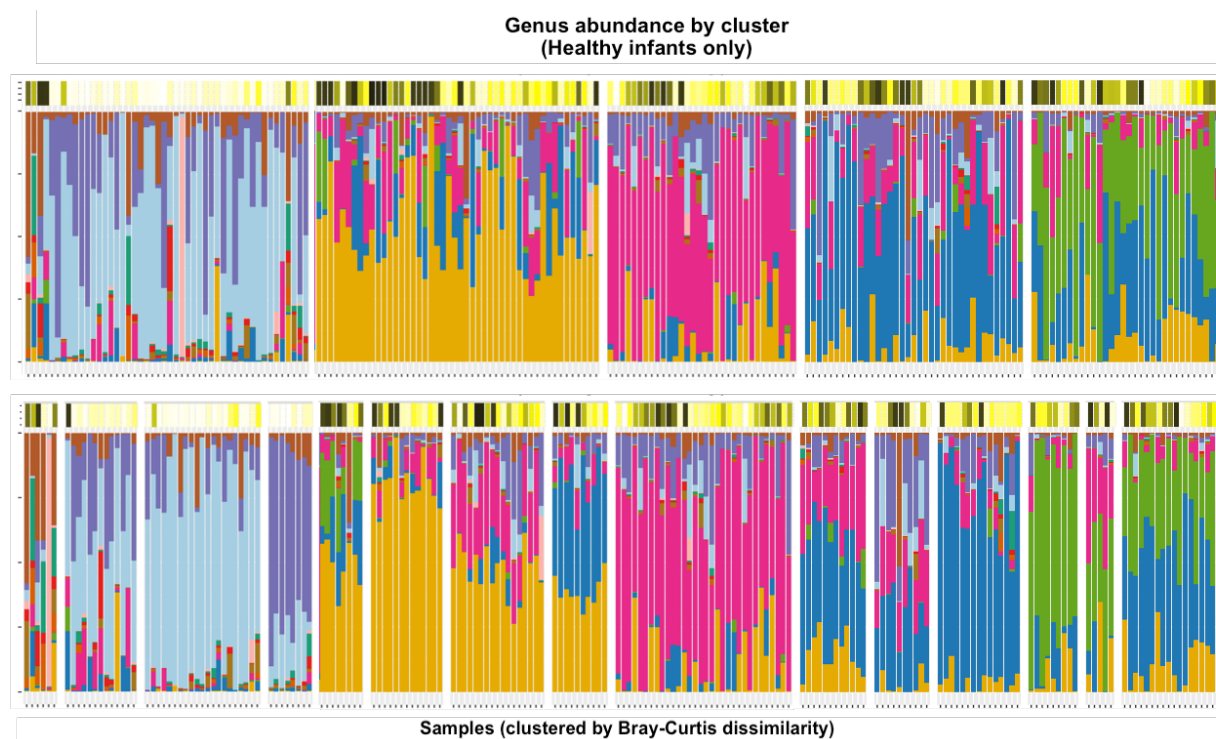
767

768

769

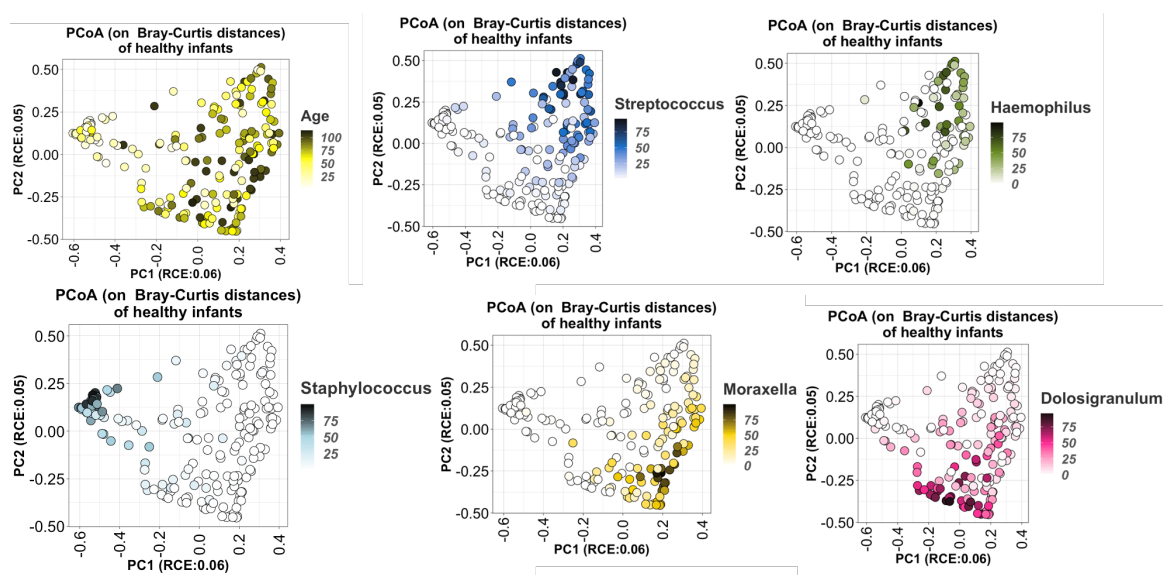
770 **Supplemental figures**

771



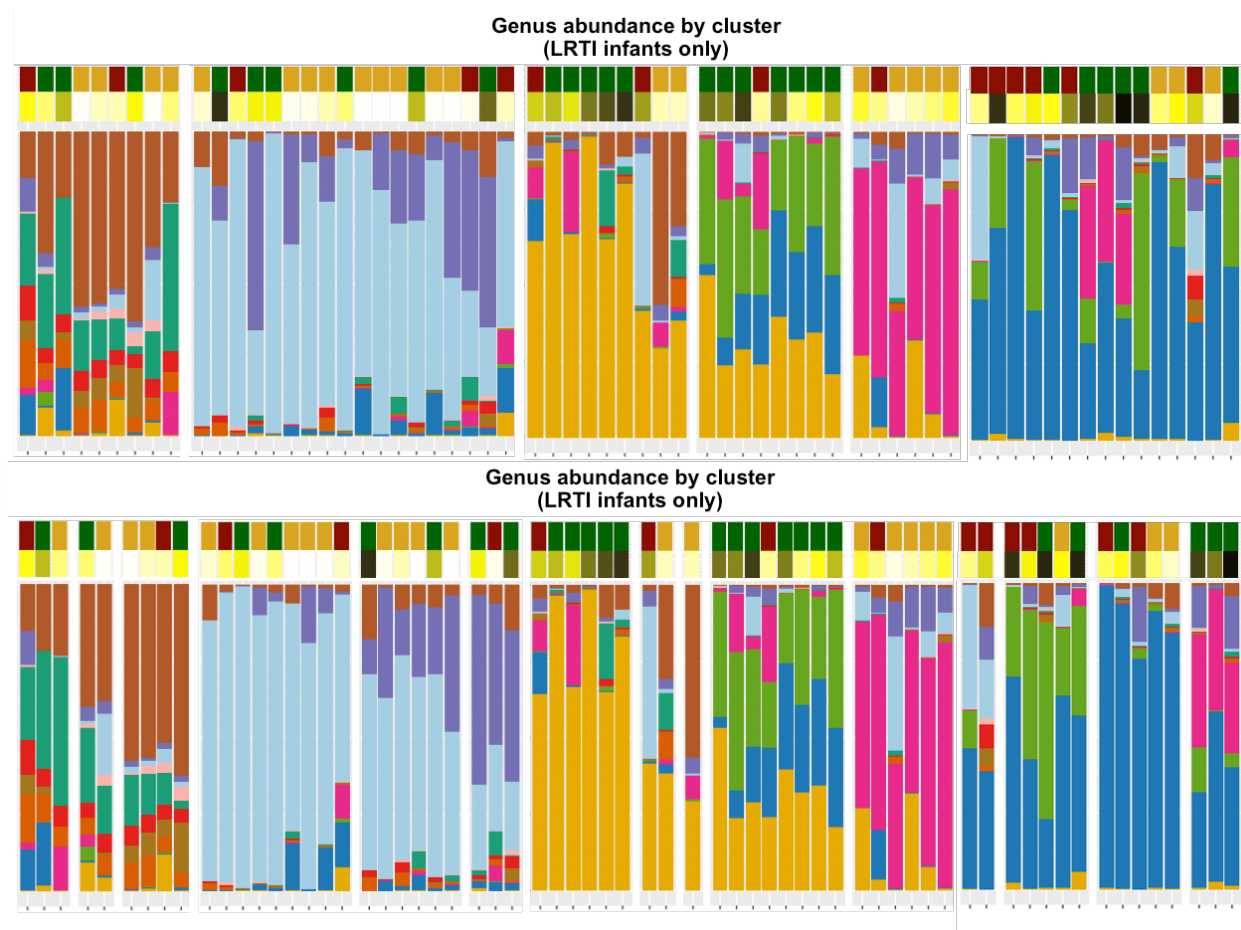
772

773 **Supplemental Figure 1A)** The taxonomic profiles of healthy infants' NP samples (n=30), clustered by  
774 pairwise Bray-Curtis dissimilarity. Clusters were defined by performing hierarchical clustering on the beta  
775 diversity matrix and then cutting the resulting dendrogram into an optimal number of clusters according to the  
776 Silhouette index (5) and Frey index (15). The color bars above the stacked bar plots indicate the infants' ages  
777 at the time of each sample.



778

779 **Supplemental Figure 1B)** Principal Coordinate Analysis (PCoA) of the Bray-Curtis dissimilarity matrix  
780 between healthy infants' samples, colored by age and relative abundance of the dominant genera.



781

782 **Supplemental Figure 2:** The taxonomic profiles of LRTI infants' NP samples (n=10), clustered by pairwise  
783 Bray-Curtis dissimilarity. Clusters were defined by performing hierarchical clustering on the beta diversity  
784 matrix and then cutting the resulting dendrogram into an optimal number of clusters according to the  
785 Silhouette index (6) and Frey index (15). The color bars above the stacked bar plots indicate the infants' ages  
786 and symptom status at the time of each sample.

787

788

789

790

791

792

793

794

795

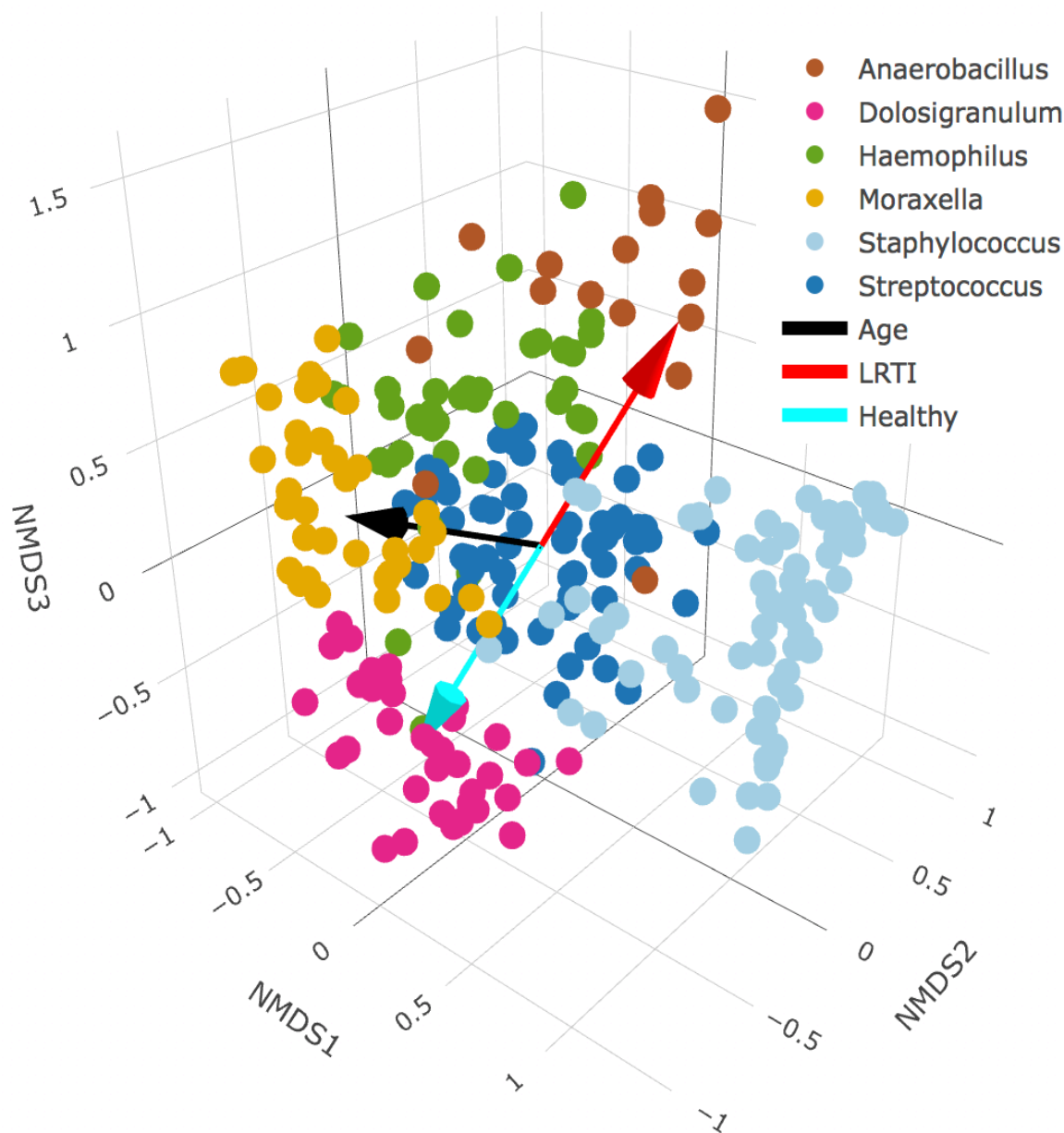
796

797

798

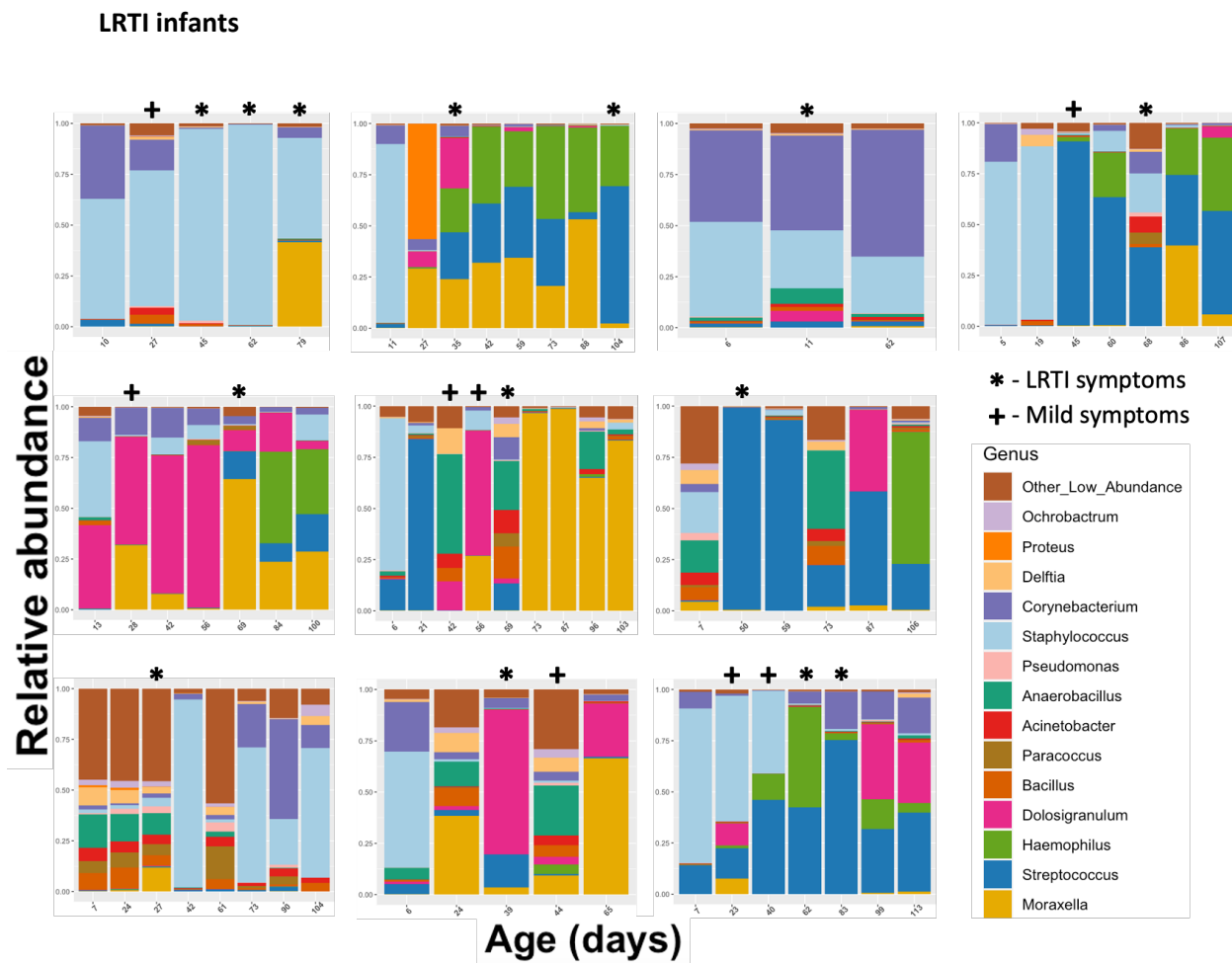


799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827



828 **Supplemental Figure 3:** A 3-D scatterplot of infant nasopharyngeal samples projected into nonmetric  
829 multidimensional scaling (NMDS) ordination space. We applied 3-dimensional NMDS ordination to the Bray-  
830 Curtis dissimilarity matrix between all infants' NP swabs, and projected vectors into that ordination space  
831 representing the best fit correlations for the age at sampling (the black vector) and LRTI status (the cyan vector  
832 vector represent control infants, the red vector represent LRTI infants). Samples are colored by their primary  
833 taxonomic profile cluster membership (see Figure 3 for details). Age is highly correlated with the x-axis, and  
834 samples on the young end of the age vector mostly belong to the *Staphylococcus*-dominated profile, whereas  
835 samples on the older end tend to belong more to the *Haemophilus* and *Moraxella*-dominated profiles. The  
836 *Dolosigranulum*-dominated profile is associated with the healthy end of the vector for LRTI status, while the  
837 *Anaerobacillus*-dominated profile is associated with disease.

838  
839  
840



841  
842

Tracking the NP maturation of infants who develop LRTI

**Supplemental Figure 4:** Each facet represents the relative abundance of bacterial genera in one infant's nasopharynx over time. Samples

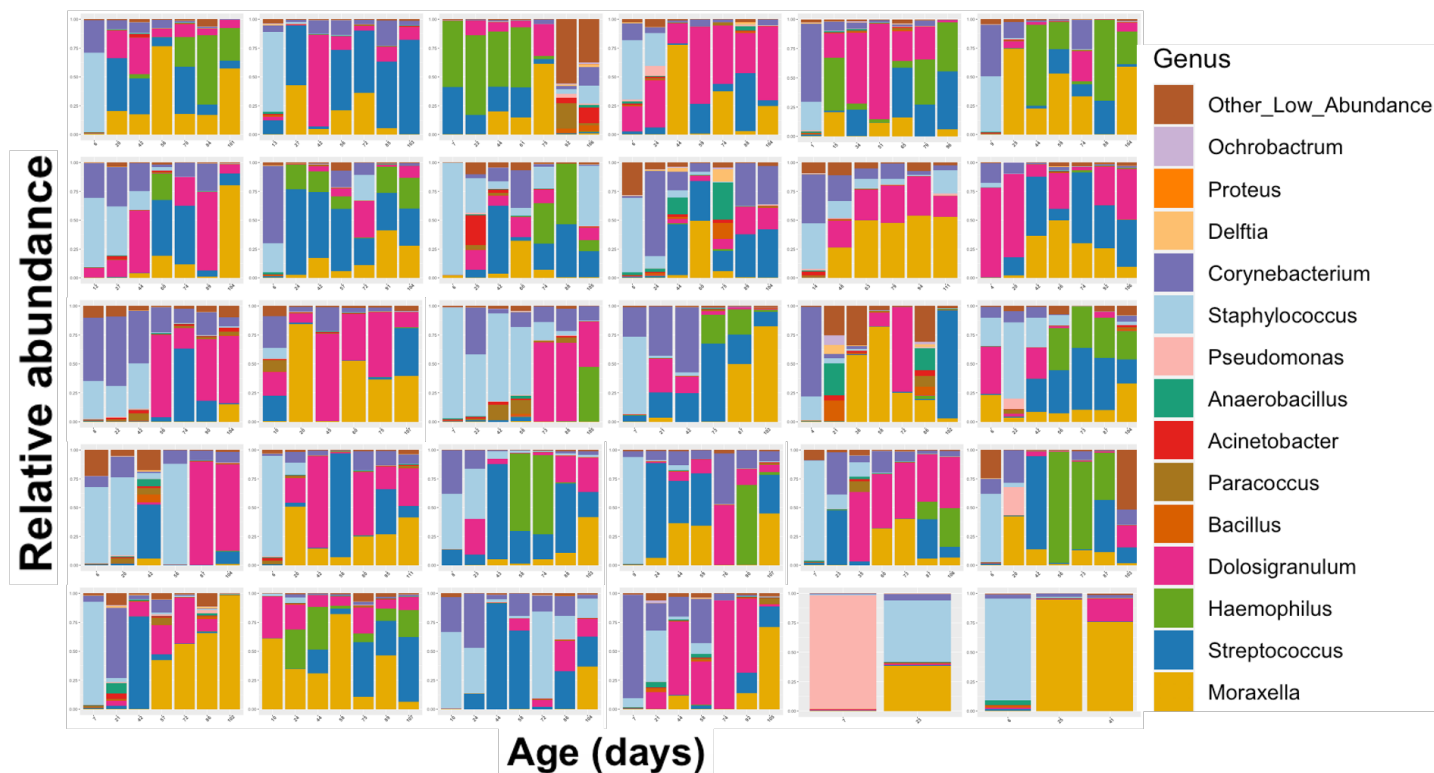
846 are ordered chronologically from left to right, with labels on the x-axis displaying the infant's age at the time  
847 of the NP swab. Time points where infants were experiencing severe respiratory symptoms are marked with a  
848 \*. Time points where infants were experiencing mild symptoms which did not qualify as LRTI are marked  
849 with a +.

850 Note that the order of these infants (reading rows left to right, top row first) is the same as the order in Table 2  
851 (Clinical symptoms)

852  
853  
854  
855  
856  
857

858  
859  
860  
861  
862

### Healthy infants



863  
864  
865  
866  
867

**Supplemental Figure 5:** Tracking the NP maturation of healthy infants. Each facet represents the relative abundance of bacterial genera in one infant's nasopharynx over time. Samples are ordered chronologically from left to right, with labels on the x-axis displaying the infant's age at the time of the NP swab.