Supplementary methods S1

Viral analysis

Nasal swabs were analyzed by real-time polymerase chain reaction (PCR) assays targeting five different respiratory viruses (human rhinovirus/enterovirus (HRV/EV), respiratory syncytial virus (RSV), human metapneumovirus (HMPV), human coronavirus (HCoV), human parainfluenzavirus (HPIV)), which have been shown previously to be the most common viruses in our population (1). One part of the samples has been analysed previously as described by Regamey et al. (1). The other part was studied using commercially available real-time PCR kits (HCoV/HPIV r-gene®, Rhino/EV/Cc r-gene®, RSV/hMPV r-gene®, Argene, bioMérieux, Marcy-l'Etoile, France); the analysis was performed according to the manufacturer's recommendation.

Definition of acute respiratory tract infection and symptom score

An acute respiratory tract infection (ARI) was defined as more than two consecutive days of cough and/or wheeze accompanied with fever (>38°C), rhinitis, otitis media or pharyngitis (1). Wheeze was defined as a squeaky noise or whistling from the chest audible to parents. Respiratory symptoms were assessed using a standardized scoring system (2). Throughout the first year of life i.e. also during the time of the ARI any use of antibiotics was recorded (3).

Nasal swab collection

Nasal swabs were collected at the onset of the first ARI (swab A) and three weeks later (swab B) at the homes of the infants by a research nurse using FLOQSwabs with universal transport medium (UTM, Copan, Italy) and subsequently stored at -80 °C.

Definition of host factors

A detailed pre- and postnatal questionnaire allowed collecting extensive data on several host factors which were of interest in our study (3). Atopy was defined as positive if the infants' mother, father or siblings were suffering from eczema, allergic rhinitis or asthma. Siblings were counted when living in the same household as the study participant. Environmental tobacco smoke (ETS) exposure was defined as positive if there was at least one parent smoking tobacco in the same household. Data on nutrition including breastfeeding were collected throughout the first year of life in weekly telephone interviews.

Microbiota analysis

Bacterial 16S ribosomal ribonucleic acid PCR amplification and 454 Titanium FLX sequencing

Amplification of bacterial 16S ribosomal ribonucleic acid (rRNA) by PCR and 454 Titanium FLX sequencing has been described previously (4, 5). In summary, after DNA extraction the variable regions V3 to V5 of bacterial 16S rRNA gene were amplified by using the multiplex identifier tagged primer pair 341F/907R. After elution of the PCR reactions by 40 μL of double-distilled water, the concentration was measured with the Agilent 2100 Bioanalyzer (Agilent Technologies, Basel, Switzerland). Amplification products with a concentration of less than 1.0 ng/μL which corresponds to less than 1 pg/μL bacterial DNA were excluded from this study according to recent recommendations (6). Samples were pooled using 40 ng/ μL of purified PCR product, resulting in 8 amplicon pools labelled with unique multiplex identifiers. The reads from 454 Titanium amplicon sequencing were submitted to the European Nucleotide Archive (ENA) (accession number PRJEB20454). The average number of reads is 1246.6 with a SD 684.0. There were only 7/334 samples with less than 250 reads. Bacterial density was described as the concentration of PCR product within a sample.

Estimation of relative bacterial abundances and microbiota clustering

Sequencing products were analyzed using Pyrotagger, with the definition of operational taxonomic units (OTUs) based on 97% similarity, estimation of chimeras, and taxonomic assignments (7). Bacterial abundances were calculated using Pyrotagger, as described elsewhere (8). Resulting OTUs were assigned to the five most abundant families with the remaining families grouped together as “Others”.

Our group has previously used the same protocol and have calculated the estimated versus observed richness in a similar set of samples (5). We revealed that the observed species richness covered in average (mean (± SD)) 63·3% (±17·6%) of the estimated species richness (see Figure s3 in (5)).

Calculation of α and β diversity

Alpha diversity analysis was performed using the "diversity" function of the "vegan" software package of R, version 3.02 (http://www.R-project.org) calculating the Shannon Diversity Index (SDI) to describe within-community diversity. For community comparisons between swab A and swab B the β diversity was calculated using the "vegdist" function of the "vegan" software package of R. We chose the Manhattan-type Jaccard dissimilarity index as weighted (abundance-based) β diversity.

Clustering of microbiota

The hierarchical clustering was performed using Matlab 2015b (Mathworks, Natick Massachusetts, United States). The relative abundances of the five most frequent bacterial families with the remaining families grouped together as "Others" were used as input variables for clustering of both swabs. The optimal number of clusters was determined by Silhouette index (9) and KL index (10) for both swabs. Both indices suggested 6 clusters for swab A. Since the sixth cluster was too small for phenotyping purposes (n = 2 infants), it was thus excluded from the study, resulting in five clusters, including 165 infants. For swab B, KL index suggested five clusters and Silhouette index suggested seven clusters. Also taking into account the clusters' sizes, the optimal cluster number for swab B was six. However, as the last three clusters were too small for phenotyping purposes (n ≤ 6 infants), they were not considered for the current study, resulting in three clusters, including 155 infants. The

clustering was also validated with the K-means clustering method, which resulted in similar cluster numbers and classes for both swabs.

References

1. Regamey N, Kaiser L, Roiha HL, Deffernez C, Kuehni CE, Latzin P, et al. Viral etiology of acute respiratory infections with cough in infancy: a community-based birth cohort study. *Pediatr Infect Dis J* 2008; 27(2): 100-105.

2. Silverman M, Wang M, Hunter G, Taub N. Episodic viral wheeze in preschool children: effect of topical nasal corticosteroid prophylaxis. *Thorax* 2003; 58(5): 431-434.

3. Fuchs O, Latzin P, Kuehni CE, Frey U. Cohort profile: the Bern infant lung development cohort. *Int J Epidemiol* 2012; 41(2): 366-376.

4. Hilty M, Qi W, Brugger SD, Frei L, Agyeman P, Frey PM, et al. Nasopharyngeal microbiota in infants with acute otitis media. *J Infect Dis* 2012; 205(7): 1048-1055.

5. Mika M, Korten I, Qi W, Regamey N, Frey U, Casaulta C, et al. The nasal microbiota in infants with cystic fibrosis in the first year of life: a prospective cohort study. *Lancet Respir Med* 2016; 4(8): 627-635.

6. Biesbroek G, Sanders EA, Roeselers G, Wang X, Caspers MP, Trzcinski K, et al. Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *PLoS One* 2012; 7(3): e32942.

7. Kunin V, Hugenholtz P. PyroTagger: a fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. *Open J* 2010; 1–8.

8. Hilty M, Qi W, Brugger SD, Frei L, Agyeman P, Frey PM, et al. Nasopharyngeal microbiota in infants with acute otitis media. *J Infect Dis* 2012; 205(7): 1048-1055.

9. Krzanowski WJ, Lai YT. A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics* 1988; 44(1): 23-34.

10. Rousseeuw P (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J Comput Appl Math* 1987; 20: 53-65.