THE LANCET
Respiratory Medicine

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Man WH, van Houten MA, Mérelle ME, et al. Bacterial and viral respiratory tract microbiota and host characteristics in children with lower respiratory tract infections: a matched case-control study. *Lancet Respir Med* 2019; published online March 15. http://dx.doi.org/10.1016/S2213-2600(18)30449-1.

Supplementary appendix for

Bacterial and viral microbiota, and host characteristics in children with lower respiratory tract infections: results from a matched case-control study

Wing Ho Man, Marlies A. van Houten, Marieke E. Mérelle, Arine M. Vlieger, Mei Ling J.N. Chu, Nicolaas J.G. Jansen, Elisabeth A.M. Sanders, Debby Bogaert*

*Correspondence to: d.bogaert@ed.ac.uk

This PDF file includes:

Extended Methods

Study design PICU cohort

We conducted a prospective study from September 2013 to September 2016 in which patients aged 4 weeks to 5 years who became hospitalized at the pediatric intensive care unit of a Dutch university hospital for a WHO-defined LRTI (acute respiratory symptoms [cough, tachypnea or dyspnea] with clinical signs of LRTI, e.g. abnormal lung auscultation and/or chest radiography)[14] requiring mechanical ventilation were enrolled. Exclusion criteria were a language barrier or severe comorbidity, i.e. congenital heart disease, bronchopulmonary dysplasia, prematurity <32 weeks, cystic fibrosis, sickle-cell disease, immunodeficiency, cardiovascular disease, neuromuscular disease, oncological disease or a severe congenital disorder. Transnasal nasopharyngeal swabs and endotracheal aspirates were obtained within four hours after intubation by trained nurses. All samples were stored immediately in a -20°C freezer followed by transportation on dry ice to a -80°C freezer until further processing.

Data on medical history as well as data on demographic, lifestyle and environmental characteristics were obtained by questionnaires and pharmacy printouts. Clinical data of the cases were obtained from medical charts.

Study design case-control cohort

Next to the PICU cohort, we conducted a prospective, matched case-control study from September 2013 to September 2016. Patients were enrolled under the same inclusion and exclusion criteria as the PICU cohort and were age-, gender-, and time-matched with healthy controls in a 1:2 ratio. Cases were recruited from three Dutch teaching hospitals. Healthy children were recruited through well-baby clinics and the local municipalities. In addition to the exclusion criteria of the cases, healthy controls were also excluded if the child had fever (≥38°C) or a respiratory tract infection (except rhinitis) in the previous four weeks, or antibiotic treatment in the previous three months. Transnasal nasopharyngeal swabs were taken of cases generally within 1 hour after admission, and of controls during a home visit within 2 weeks after admission of the matching case by trained nurses and physicians. All samples were directly stored in a -20°C freezer followed by transport on dry ice to a -80°C freezer until further processing. Samples of cases were always stored in the same box as that of their matching controls. Metadata of this cohort was obtained similar to that of the PICU cohort. Both cohorts were registered as one study (www.trialregister.nl, NTR5132) and were approved by the Dutch National Ethics Committee (NL42019·094·12). Written informed parental consent was obtained from all participants.

Study size

When we first designed the current studies, there were no previous studies investigating the nasopharyngeal microbiome in children with a LRTI. Little was known about the composition, variability and diversity of the nasopharyngeal microbiota, except for one study performed by our group recruiting 96 children of 18 months of age.[46] We used the data from this study for the sample size calculation of the current studies. We selected four bacterial taxa that were different in abundance and variability. In order to obtain full depth information on power calculation, we selected bacteria with approximately the widest variability in distribution i.e. high and low abundance and high and low variability. We selected two bacterial taxa with high abundance: one with high variability (*Streptococcus*, mean relative abundance 13·1%, standard deviation [SD] 1·6 times the mean*)* and one with low variability (*Moraxella*, mean abundance 38·5%, SD 0·8 times the mean). In addition, we selected two bacterial taxa with low abundance and high variability (*Prevotella Shahii*, mean abundance 0·3%, SD 6·4 times the mean*)* and low variability (*Helcococcus*, mean abundance 0,4%, SD 0·9 times the mean*)*. Based on these data and on a p-value of 0·002 (a p-value of 0·05, corrected by Bonferroni for multiple testing because of evaluation of at least top 25 species), we calculated that 150 cases and 300 controls were needed to achieve a sufficient power (80% or more) to detect more than threefold shifts in abundance of bacterial taxa with high abundance and low and high variability. Furthermore, these numbers of cases and controls would obtain a power of 80% to detect more than two-fold shifts in bacterial taxa with low abundance and low variability, and for low abundant bacteria with high variability we would be sufficiently powered to detect eight-fold shifts. A case-control ratio of 1:3 did not significantly improve the power and, therefore, we decided for a 1:2 case-control design to be optimal in our setting.

Matching factors

Regarding matching variables, we decided to match in particular for time because of the findings of the above-mentioned study that respiratory microbiota profiles varied strongly with season.[46] Our choice for age and sex was based on several keystone papers reporting that the developing gut microbiota varies across age and sex.[47–49]

Expert review panel

Two expert pediatricians independently classified all cases of the case-control cohort in three major disease phenotypes, i.e. pneumonia, bronchiolitis, and wheezing illness. Cases with a mixed or unclear phenotype were deemed mixed. All disagreements were resolved by consensus. The classification of the expert panelists was based on the entire medical record of the child, including all clinical notes at and during admission, laboratory assessments and imaging.

Detection of respiratory viruses and confirmation of bacterial species

All samples of cases controls were tested using qualitative multiplex realtime-PCR (RespiFinder® SMARTfast 22) specific for human bocavirus (BoV), RSV, human influenza virus (IV), parainfluenza virus (PIV), human rhinoviruses (HRV), adenoviruses (AdV), bocaviruses (BoV), human coronavirus (CoV), and human metapneumovirus (hMPV; **Supplementary Figure 1**).[50] Identification of *Streptococcus pneumoniae* was done by quantitative qPCR targeting the autolysin (*lytA*) gene.[51] The *lytA* C_T-values had a strong correlation with the OTU annotated as *Streptococcus pneumoniae* (3) confirming its origin (Pearson's r=0·82, p<0·001). In addition, we performed a multiplex qPCR (Fast Track Diagnostics) to confirm or further nuance the OTUs annotated as *Staphylococcus, Moraxella or Haemophilus*. We found that the OTU annotated as *S. aureus/epidermidis* (7) was significantly related with the *S. aureus* C_T-values (Pearson's r=-0·34, p<0·001) and that the OTU annotated as *M. catarrhalis/nonliquefaciens* (1) was significantly related with the *M. catarrhalis* C_T-values (Pearson's r=-0·51, p<0·001). We therefore did not change their original annotation. The OTU annotated as *H. haemolyticus* (2) was significantly related with the *H. influenzae* C_T-values (Pearson's r=-0·71, p<0·001); therefore, we changed its annotation into *H. influenzae/haemolyticus* (2).

16S rRNA gene amplification and sequencing

Bacterial DNA was isolated from samples and quantified as previously described.[33,52] In short, an aliquot of 200µl of each sample was added to 650µl lysis buffer with 0·1 mm zirconium beads and 550µl phenol. All samples were mechanically lysed with a bead beater procedure. Almost all samples fulfilled our quality control standards for reliable analyses, having DNA levels of ≥0·3 pg/µl over negative controls (>98%, **Supplementary Figure 1**). Amplification of the V4 hypervariable region of the 16S rRNA gene was performed using barcoded universal primer pair 533F/806R. Amplicons were quantified by PicoGreen (Thermofisher) and pooled in equimolar amounts. Amplicon pools of samples and controls were sequenced in five Illumina MiSeq runs (San Diego, CA, USA). Samples from cases were always processed simultaneously with that of their matching controls to uphold identical storage and processing conditions within each of the case-control sets.

Bioinformatics analysis

Raw sequences were trimmed using an adaptive, window-based trimming algorithm (Sickle, Q>20, length threshold of 150 nucleotides).[53] We aimed to further reduce the number of sequence errors in the reads by applying an error correction algorithm (BayesHammer, SPAdes genome assembler toolkit).[54] Forward and reverse reads were then assembled into contigs using PANDAseq.[55] Merged reads were demultiplexed using QIIME v1·9.[56] After removal of singleton sequences, we removed chimeras using both *de novo* and reference (against Gold database) chimera identification (UCHIME algorithm in VSEARCH).[57,58] VSEARCH abundance-based greedy clustering was used to pick OTUs at a 97% identity threshold.[59] Taxonomic annotation was executed using the RDP-II naïve Bayesian classifier on SILVA v119 training set.[60] Taxonomic assignment was validated by blasting against the NCBI database, using a 100% identity cut-off. After aligning the node representative sequences to the Silva v119 core alignment database using the PyNAST method,[61] a rooted phylogenetic tree was calculated using FastTree.[62] Two samples were

below our threshold of 15,000 reads/sample and were therefore excluded from further analysis (**Supplementary Figure 1**). Removing these left 33,020,647 reads in total (mean 63,945 ± 26,270 reads/sample). The Good's estimator of all samples was above 99·9% and rarefaction curves on raw count data approached plateau (**Supplementary Figure 12**), indicating that the coverage degree of our sequencing was very high.[63] We generated an abundance-filtered dataset by including only those OTUs that were present at or above a confidence level of detection (0·1% relative abundance) in at least 2 samples, retaining 306 OTUs in total.[64] To avoid OTUs with identical annotations, we refer to OTUs using their taxonomical annotations combined with a rank number based on the abundance of each given OTU. The raw OTU-counts table was used for calculations of local diversity (α-diversity) and analyses using the *metagenomeSeq* package.[65] The OTU-proportions table was used for all other downstream analyses, including hierarchical clustering and random forest modelling. Moreover, the Bray-Curtis (dis)similarity metric was consistently used to express ecological distance (β-diversity) in all analyses because it includes proportional abundance information and excludes joint-absence information, and thereby yields useful insights into the specific structure of our data.[66]

Quality control of 16S rRNA gene amplicon sequencing

In total, we had 34 negative controls during DNA isolation. The DNA load of these DNA isolation blanks were two orders of magnitude lower compared to the samples analyzed in this study (median 0·26 vs 49·7 pg/µl; (**Supplementary Figure 13A**). In addition, 14 DNA isolation and PCR blanks were sequenced along with the study samples. All blanks yielded <15,000 reads and the number of reads was at least one orders of magnitude lower compared to that of the samples (median 4,618 vs 67,276 reads; **Supplementary Figure 13B**). There were no taxa shared across all blanks. Total microbiota community of blanks were highly diverse from samples (*adonis* R^2=9·1%, p<0·001) and hierarchical clustering clearly separated the blanks (**Supplementary Figure 13C**). These results robustly indicate that our strict sequencing protocol resulted in no apparent contamination. However, to ensure our data was of the highest quality, we used the bacterial biomass to identify and remove contaminants. We identified 13 OTUs as contaminating species when relating the frequency of each OTU to the bacterial biomass of the samples (**Supplementary Figure 14** and **Supplementary Table 5**). In addition, we used the Decontam R-package to additionally identify 8 OTUs as contaminants by evaluating the prevalence (presence/absence across samples) of each OTU in true positive samples compared to the prevalence in negative controls.[29,30] These 21 OTUs were removed from our dataset prior to all analyses.

Statistical analysis

Data analysis was performed in R v3·2[67] within R studio v1·0.[68] All analyses assessing matched samples accounted for the matched nature of the samples. Our questionnaires contained minimal missing data (<1%), allowing analyses of individuals with complete data on all variables required for a particular analysis. A P-value of less than 0·05 or a Benjamini-Hochberg (BH) adjusted q less than 0·05 was considered statistically significant.

To compare baseline characteristics and viral detection between cases and controls, we used conditional logistic regression analysis.

To assess the concordance between the bacterial microbiota of the nasopharynx with that of endotracheal aspirates, we compared the intra-individual and inter-individual Bray-Curtis similarity. The Bray-Curtis similarity was calculated using the formula: 1 - Bray-Curtis dissimilarity. Statistical significance was determined using Wilcoxon rank-sum test.

Nonmetric multidimensional scaling (NMDS) plots based on Bray-Curtis dissimilarity were used to visualize differences of total microbiota communities and statistical significance was calculated by *adonis* (*vegan,* 1,999 permutations). Because there was substantial multicollinearity (*vif.cca*-function; two covariates were aliased; range variance inflation factor of other covariates, 1·1 - 8·8) between some of the covariates, we performed a stepwise

selected multivariable distance-based redundancy analysis (*ordiR2step*, 1,999 permutations).[69,70] Directionality of these covariates were projected in NMDS plots using *envfit* (*vegan*).

Unsupervised average linkage hierarchical clustering was performed as described previously (*hclust* and *vegdist*, *vegan* package).[13] The optimal number of clusters contained in the microbiota dataset was determined based on Silhouette and Calinski-Harabasz indices (*cluster.stats*, *fpc* package). Clusters including less than five individuals were excluded for further downstream analysis. Random forest classifier analyses was performed to determine biomarker species and factors that most discriminate between clusters by the interpretation step of the two-stage feature selection procedure of *VSURF* (10,000 trees each iteration, with 100 thresholding iterations and 50 interpretation iterations).[71] A chi-square test was used to test for the association between clusters and presence/absence of acute disease.

We used *metagenomeSeq* to identify specific microbial taxa associated with cases or controls (filtered on taxa present in >5% of the samples, 100 maximum iterations, mixed model design).[65] In addition, using a 10-fold cross-validated *VSURF* procedure we performed random forest classifier analysis to determine which microbial taxa best distinguish disease from health. Taxa that were selected at least 1 time during the interpretation step, were deemed minor classifier taxa, while taxa that were selected 9 or 10 times were defined as major classifier taxa. Variable importance was estimated by calculating the mean decrease in Gini after randomly permuting the values of each given variable (*randomForest*, 10,000 trees, 100 replicates).[72]

A similar random forest classifier analysis was performed including not only the bacterial abundance data but also viral presence and metadata of host and lifestyle/environmental factors. Variable importance was estimated by calculating the mean importance after randomly permuting the values of each selected variable (100 replicates, *caret* package).[73] The direction of the variable associations was crudely estimated *post-hoc* using the point biserial correlation coefficients, where the relative abundances of bacterial taxa were transformed with the arcsine square-root transformation for proportional data. Performance of the sparse random forest models was evaluated by calculating the area under the ROC curve (AUC) using the out-of-bag predictions for classification (*pROC* package)[74] as previously described.[75] Using the above procedures, we also build a sparse random forest prediction model to investigate to what extend hospitalization duration could be predicted with all available data. Predictions were determined by a cross-validated random forest procedure (*train* function, 10 folds, 100 iterations, *caret* package).[73] The individual association of these duration-predictive variables with hospitalization duration is depicted as a heatmap of their mean values against short (1-3 days, $\leq 33^{rd}$ percentile), medium (4-5 days, $\leq 67^{th}$ percentile) and long (>5 days, $>67^{th}$ percentile) hospitalization, except for disease phenotype as this is a categorical variable. Colours of the heatmap correspond with row wise normalized values (i.e. white indicates the overall minimum value of a variable, purple indicates the overall maximum value).

Above analyses were carried out for the entire case-control cohort and were in part repeated for each of the phenotypes independently. Additionally, to study the role of nasopharyngeal microbiota in the disease severity, we stratified the cases according to the physicians' judgment whether antibiotics were needed during admission and performed separate analyses on both groups. Finally, a subset of these analyses was performed in the assessment of the concordance between nasopharyngeal and endotracheal samples of the PICU cohort, as well as in the comparison of the PICU cohort nasopharyngeal microbiota with that of age and season-matched participants of the case-control cohort.

Data availability

Sequence data that support the findings of this study have been deposited in the NCBI Sequence Read Archive (SRA) database with BioProject ID PRJNA428382.

Table S1. Baseline characteristics and clinical data for the PICU cohort.

Data on medication use was acquired by pharmacy printouts, whereas the rest of the data was acquired by parent questionnaires. Breastfeeding was nonexclusive. Educational level was classified into three categories: low level (primary school education or pre-vocational education as highest qualification), intermediate (selective secondary education or vocational education) and high level (university of applied sciences and research university). Smoke exposure included children who were exposed to second-hand tobacco smoke. P values were determined by univariate conditional logistic regression. IQR = interquartile range; RTI = respiratory tract infection; LRTI = any parental-reported lower RTI.

	PICU cohort
n	29
Basics	
Girl (%)	14 (48·3)
Age (months) (median [IQR])	2·2 [1·6, 3·6]
Born at term (%)	20 (74·1)
Mode of delivery (%)	
vaginal	18 (66·7)
elective C-section	3 (11·1)
emergency C-section	6 (22·2)
Season of sampling (%)	
Spring	7 (24·1)
Summer	3 (10·3)
Autumn	1 (3·4)
Winter	18 (62·1)
Medical History	
LRTI (%)	2 (14·3)
Wheezing (%)	3 (10·3)
Otitis (%)	1 (3·4)
Hospitalization for RTI (%)	2 (14·3)
Medication	
Antibiotics past 6 months (%)	3 (12·5)
Feeding	
Breastfeeding current and/or >3 months (%)	10 (37·0)
Family	
Education level parents (%)	
high	17 (63·0)
intermediate	8 (29·6)
low	2 (7·4)
Siblings (median [IQR])	1·0 [1·0, 2·0]
Environment	
Smoke exposure (%)	16 (55·2)
Clinical data	
Chest x-ray abnormality (%)	29 (100)
Positive blood culture (%)	1 (3·4)
Positive endotracheal aspirate culture (%)	15 (51·7)
Antibiotic treatment prior to sampling (%)	5 (17·2)
Antiviral treatment prior to sampling (%)	0 (0)
Vasopressor support (%)	3 (10·3)

Table S2. Baseline characteristics and clinical data for the case cohort stratified by phenotype.

[1] = Respiratory rate >40 breaths/min for infants <1 year; >35 breaths/min for children 1-2 years; >30/min for older children. IQR = interquartile range.

	Pneumonia	Bronchiolitis	Wheezing illness	Mixed
n	37	57	34	26
Basics				
Girl (%)	14 (37·8)	25 (43·9)	9 (26·5)	13 (50·0)
Age (months) (median [IQR])	26·1 [16·4, 37·4]	4·9 [2·5, 10·1]	23·3 [15·7, 34·1]	9·6 [4·5, 18·9]
Born at term (%)	35 (94·6)	53 (93·0)	31 (91·2)	23 (88·5)
Mode of delivery (%)				
vaginal	29 (78·4)	44 (77·2)	26 (76·5)	25 (96·2)
elective C-section	5 (13·5)	6 (10·5)	3 (8·8)	1 (3·8)
emergency C-section	3 (8·1)	7 (12·3)	5 (14·7)	0 (0·0)
Season of sampling (%)				
Spring	13 (36·1)	17 (29·8)	11 (32·4)	8 (30·8)
Summer	4 (11·1)	3 (5·3)	9 (26·5)	6 (23·1)
Autumn	1 (2·8)	1 (1·8)	5 (14·7)	1 (3·8)
Winter	18 (50·0)	36 (63·2)	9 (26·5)	11 (42·3)
CLINICAL DATA				
Symptoms at admission				
Duration of symptoms (days) (median [IQR])	3·0 [2·0, 5·0]	3·0 [1·8, 6·0]	2·0 [1·0, 2·0]	3·0 [2·0, 5·0]
Cough (%)	29 (78·4)	49 (86·0)	28 (82·4)	24 (92·3)
Wheeze (%)	2 (5·4)	9 (15·8)	12 (35·3)	3 (11·5)
Common cold (%)	19 (51·4)	38 (66·7)	24 (70·6)	23 (88·5)
Respiratory distress (%)	20 (54·1)	45 (78·9)	31 (91·2)	21 (80·8)
Feeding difficulty (%)	23 (62·2)	38 (66·7)	9 (26·5)	10 (38·5)
Clinical findings at admission				
Tachypnea[1] (%)	23 (62·2)	41 (71·9)	27 (79·4)	18 (69·2)
Oxygen saturation (median [IQR])	95·0 [93·0, 96·0]	95·0 [94·0, 98·0]	94·5 [91·2, 96·0]	95·0 [92·2, 98·8]
Fever at home or during presentation (%)	36 (97·3)	43 (75·4)	21 (61·8)	16 (61·5)
Temperature (°C) (median [IQR])	39·2 [38·2, 39·8]	38·3 [37·7, 39·1]	37·8 [37·4, 38·5]	38·1 [37·4, 39·0]
Nasal flaring (%)	10 (27·0)	17 (29·8)	7 (20·6)	4 (15·4)
Chest retractions (%)	20 (54·1)	44 (77·2)	31 (91·2)	16 (61·5)
Grunting (%)	12 (32·4)	11 (19·3)	7 (20·6)	3 (11·5)
Auscultatory abnormalities (%)	35 (94·6)	55 (96·5)	34 (100·0)	25 (96·2)
Ronchi (%)	9 (24·3)	38 (66·7)	13 (38·2)	15 (57·7)
Wheezing (%)	5 (13·5)	24 (42·1)	30 (88·2)	17 (65·4)
Crepitations (%)	29 (78·4)	29 (50·9)	8 (23·5)	8 (30·8)
Decreased breath sounds (%)	10 (27·0)	3 (5·3)	13 (38·2)	5 (19·2)
Radiology				
Endpoint consolidation (%)	16 (43·2)	2 (3·5)	2 (5·9)	1 (3·8)
Non-endpoint consolidation (%)	9 (24·3)	4 (7·0)	3 (8·8)	6 (23·1)
No chest x-ray abnormality (%)	3 (8·1)	4 (7·0)	3 (8·8)	2 (7·7)
Chest x-ray not performed (%)	10 (27·0)	47 (82·5)	27 (79·4)	18 (69·2)
Laboratory				
Blood tests (%)	22 (59·5)	18 (31·6)	2 (5·9)	4 (15·4)
White blood cell count (x 10^9/L) (median [IQR])	13·8 [11·2, 17·6]	14·8 [9·8, 17·4]	11·1 [10·5, 11·6]	13·5 [9·7, 18·2]
CRP (mg/L) (median [IQR])	27·0 [17·0, 138·0]	17·0 [7·0, 37·0]	19·0 [19·0, 19·0]	22·0 [12·5, 55·2]
Blood culture (all negative) (%)	10 (27·0)	2 (3·5)	0 (0·0)	0 (0·0)
Treatment				
Hospitalization duration (days) (median [IQR])	4·0 [3·0, 5·0]	5·0 [3·0, 7·0]	3·0 [3·0, 4·0]	3·0 [2·2, 5·0]
Supplemental oxygen (%)	26 (70·3)	42 (73·7)	20 (58·8)	12 (46·2)
Supplemental oxygen duration (days) (median [IQR])	3·0 [2·0, 4·0]	4·0 [3·0, 4·8]	2·0 [1·8, 3·0]	4·0 [2·0, 4·0]
Antibiotic treatment during admission (%)	29 (78·4)	4 (7·0)	4 (11·8)	6 (23·1)
Albuterol (%)	10 (27·0)	26 (45·6)	34 (100·0)	15 (57·7)
Prednisone (%)	2 (5·4)	0 (0·0)	18 (52·9)	5 (19·2)
Nasogastric tube feeding (%)	10 (27·0)	27 (47·4)	0 (0·0)	3 (11·5)
Follow Up				
Follow up sampling (days after discharge) (median [IQR])	40·0 [36·0, 46·0]	41·5 [34·8, 46·2]	39·0 [35·2, 43·5]	42·0 [36·0, 50·8]
New episode of respiratory complaints (%)	18 (52·9)	36 (64·3)	25 (73·5)	18 (69·2)

204 **Table S3. STROBE Statement.**

	Item No	Recommendation	Comment
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	Indicated in both title and abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	Provided in abstract
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	Explained in Introduction, paragraphs 1-4
Objectives	3	State specific objectives, including any prespecified hypotheses	Stated in Introduction, paragraph 4
Methods			
Study design	4	Present key elements of study design early in the paper	Presented in Abstract, sub-section "Methods", and in Methods, first sub-section "Study design and procedures"
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Described in Methods, first sub-section "Study design and procedures", and in Extended Methods, sub-sections "Study design PICU cohort" and "Study design case-control cohort"
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls	Described in Methods, sub-section "Study design and procedures", and in Extended Methods, sub-sections "Study design PICU cohort", "Study design case-control cohort", and "Matching factors"
		(*b*) For matched studies, give matching criteria and the number of controls per case	Described in Methods, sub-section "Study design and procedures", and in Extended Methods, sub-sections "Study design case-control cohort" and "Matching factors"
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	Described in Methods and Extended Methods, sub-section "Statistical analysis"
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Described in Methods, sub-section "Microbiota analysis", and in Extended Methods, sub-sections "Detection of respiratory viruses and confirmation of bacterial species", "16S rRNA gene amplification and sequencing", and "Bioinformatics analysis"
Bias	9	Describe any efforts to address potential sources of bias	Described in Methods, sub-section "Study design and procedures", and in Extended Methods, sub-sections "Study

			design case-control cohort" and "Matching factors"
Study size	10	Explain how the study size was arrived at	Described in Extended Methods, sub-section "Study size"
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	Described in Methods and Extended Methods, sub-section "Statistical analysis"
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	Described in Methods and Extended Methods, sub-section "Statistical analysis"
		(*b*) Describe any methods used to examine subgroups and interactions	Described in Methods and Extended Methods, sub-section "Statistical analysis"
		(*c*) Explain how missing data were addressed	Described in Extended Methods, first paragraph of sub-section "Statistical analysis"
		(*d*) If applicable, explain how matching of cases and controls was addressed	Described in Methods and Extended Methods, sub-section "Statistical analysis"
		(*e*) Describe any sensitivity analyses	na
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	Described in Methods, sub-section "Study design and procedures", paragraphs 1-2; Extended Methods, sub-sections "Detection of respiratory viruses and confirmation of bacterial species" and "16S rRNA gene amplification and sequencing"; and Supplementary Figure 1
		(b) Give reasons for non-participation at each stage	Described in Methods, sub-section "Study design and procedures", paragraphs 1-2; Extended Methods, sub-sections "Detection of respiratory viruses and confirmation of bacterial species" and "16S rRNA gene amplification and sequencing"; and Supplementary Figure 1
		(c) Consider use of a flow diagram	Supplementary Figure 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Described in Methods, sub-section "Study design and procedures", paragraphs 1-2; Results, sub-section "Host, lifestyle and environmental factors are associated with risk of disease"; Table 1; and Supplementary Table 1
		(b) Indicate number of participants with missing data for each variable of interest	na

Outcome data	15*	Report numbers in each exposure category, or summary measures of exposure	Described in Results, sub-sections “Viral and bacterial profile differences between cases and controls” and “Clinical presentation independent viral and bacterial differences between cases and controls”
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	Described in Results, sub-sections “Nasopharyngeal microbiota profiles correlate with lower respiratory tract microbiota during childhood LRTI”, “Viral and bacterial profile differences between cases and controls”, and “Combined importance for disease”
		(*b*) Report category boundaries when continuous variables were categorized	na
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	na
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Described in Results, sub-sections “Clinical presentation independent viral and bacterial differences between cases and controls” and “Microbiota and severity of disease”
Discussion			
Key results	18	Summarise key results with reference to study objectives	Described in Discussion, paragraphs 1-2
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Described in Discussion, paragraph 8
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Described in Discussion, paragraphs 4-6
Generalisability	21	Discuss the generalisability (external validity) of the study results	Described in Discussion, paragraph 5 and 8
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Described in Abstract, sub-section “Funding”, and in Methods, sub-section “Role of funding source”

*Give information separately for cases and controls.

Table S4. Summary of intra-individual concordance between the microbiota of the nasopharynx and the lower respiratory tract for the PICU-subcohorts stratified by suspicion or confirmation of bacterial infection and of the cohort excluding the PICU cases that received antibiotics prior to sampling.

	Suspected	Not suspected	Confirmed	Not confirmed	No prior ABx
Viral presence					
Mean agreement	97%	95%	97%	95%	96%
(95% CI)	93-100%	88-100%	93-100%	90-100%	93-100%
Shannon diversity					
Pearson's r	0.59	0.71	0.54	0.73	0.59
p-value	0.01	0.07	0.04	0.02	0.008
Bray Curtis similarity					
Intra-individual	0.59	0.68	0.68	0.48	0.60
Inter-individual	0.13	0.04	0.17	0.06	0.14
p-value	<0.001	<0.001	<0.001	<0.001	<0.001
Correlated taxa (p<0.05)					
Number	45	51	48	42	65
Combined relative abundance	70%	77%	69%	61%	78.6%
Median Pearson's r	0.95	1.0	0.96	0.98	0.81
IQR	0.74-1.0	0.94-1.0	0.79-0.99	0.79-1.0	0.59-100

Table S5. Potential contaminants.

OTUs were identified as potential contaminants based on their relation with bacterial biomass (Frequency) or their presence/absence in samples vs. controls (Prevalence).

OTU	Method
Tepidimonas (28)	Both
Schlegelella (10)	Both
Acidovorax (66)	Both
Vogesella (69)	Both
Acinetobacter (31)	Both
Acinetobacter seohaensis (64)	Both
Phyllobacteriaceae (52)	Both
Pseudomonas stutzeri (95)	Both
Tardiphaga robiniae (106)	Both
Mesorhizobium (81)	Both
Shewanella (30)	Both
Massilia (88)	Frequency
Pseudomonas aeruginosa (79)	Frequency
Rhizobiales (169)	Prevalence
Xanthomonadales (114)	Prevalence
Cyanobacteria (126)	Prevalence
Hydrotalea (205)	Prevalence
Cyanobacteria (143)	Prevalence
Acinetobacter (139)	Prevalence
Modestobacter (167)	Prevalence
Cupriavidus metallidurans (156)	Prevalence

Figure S1. Flow diagram for subject enrolment.

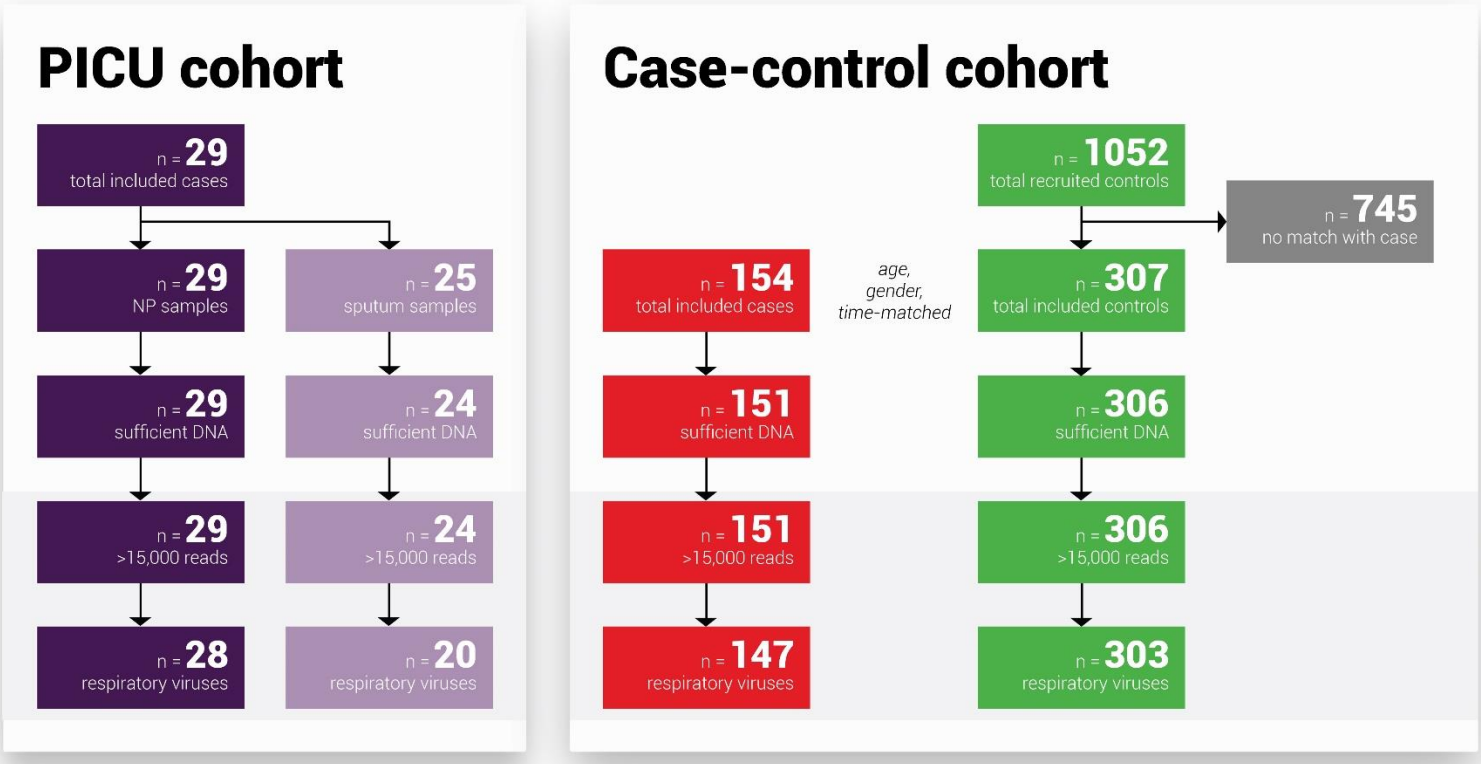
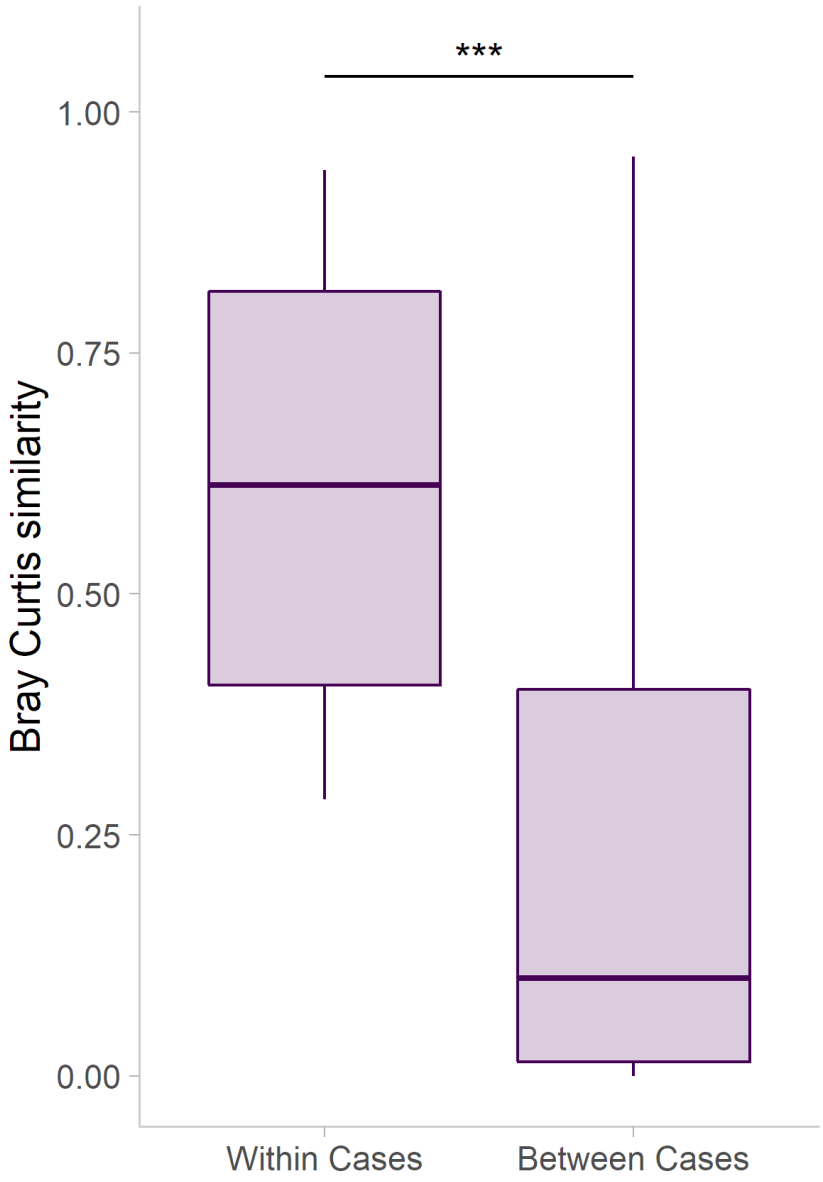


Figure S2. Nasopharyngeal and endotracheal samples harbor similar microbiota community compositions. **(A)** The Bray Curtis similarities of paired nasopharyngeal and endotracheal samples (Within Cases) was significantly higher than the similarities between cases. Significance symbol: *** = p<0·001. **(B)** Relative abundances of the 15 most abundant taxa of the nasopharynx and endotracheal aspirate. Taxa are ordered by phylum, i.e. Proteobacteria (green), Firmicutes (orange), Actinobacteria (purple) and Bacteroides (pink).

A



B

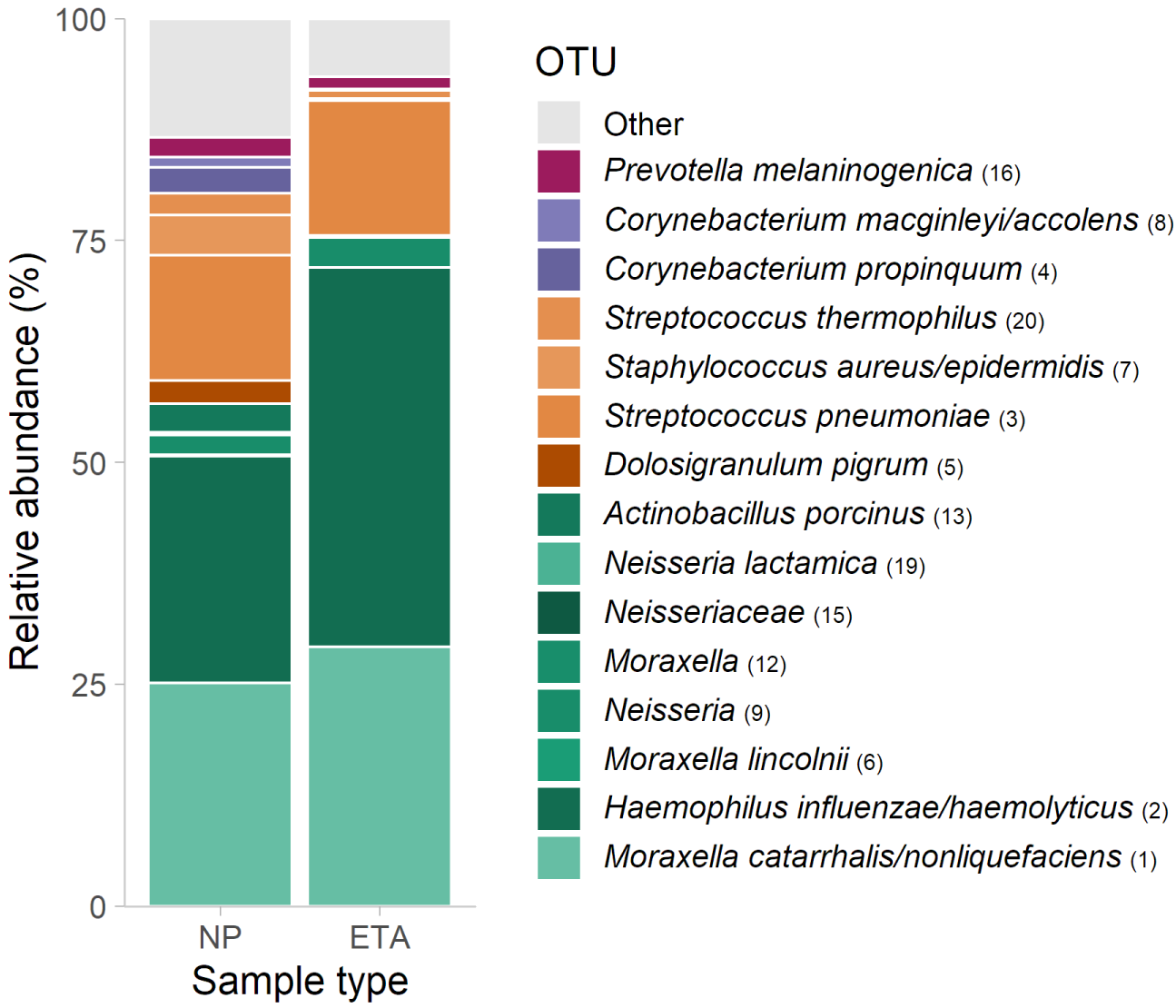

100
75
50
25
0
Relative abundance (%)
NP
ETA
Sample type
OTU
Other
Prevotella melaninogenica (16)
Corynebacterium macginleyi/accolens (8)
Corynebacterium propinquum (4)
Streptococcus thermophilus (20)
Staphylococcus aureus/epidermidis (7)
Streptococcus pneumoniae (3)
Dolosigranulum pigrum (5)
Actinobacillus porcinus (13)
Neisseria lactamica (19)
Neisseriaceae (15)
Moraxella (12)
Neisseria (9)
Moraxella lincolnii (6)
Haemophilus influenzae/haemolyticus (2)
Moraxella catarrhalis/nonliquefaciens (1)

Figure S3. Covariate impact on microbial diversity.

Information of 209 covariates were collected during the study. Data on medication use was acquired by pharmacy printouts, whereas the rest of the data was acquired by parent questionnaires. We depicted all covariates that were significantly univariately associated with the diversity of the nasopharyngeal microbiota (using *adonis*-analysis) based on the healthy control cohort only to avoid confounding effects by disease. The plot shows from left to right for each factor the Spearman correlation coefficients with alpha diversity (Chao1 estimate and Shannon index) and the explained variation in beta diversity (bars representing effect size (R^2) and Benjamini-Hochberg adjusted P-values [Q]). Significant positive and negative correlations with richness and diversity are colored red and blue, respectively, while non-significant correlations are greyed out. Only covariates with a P-value <0·05 in *adonis*-analysis are depicted. Because there was substantial multicollinearity between some of the covariates, a multivariable stepwise distance-based redundancy analysis (db-RDA) model was built in addition: covariates that were selected by this model are represented by *** = p<0·001; ** = p<0·01; * = p<0·05; ns = p≥0·05.

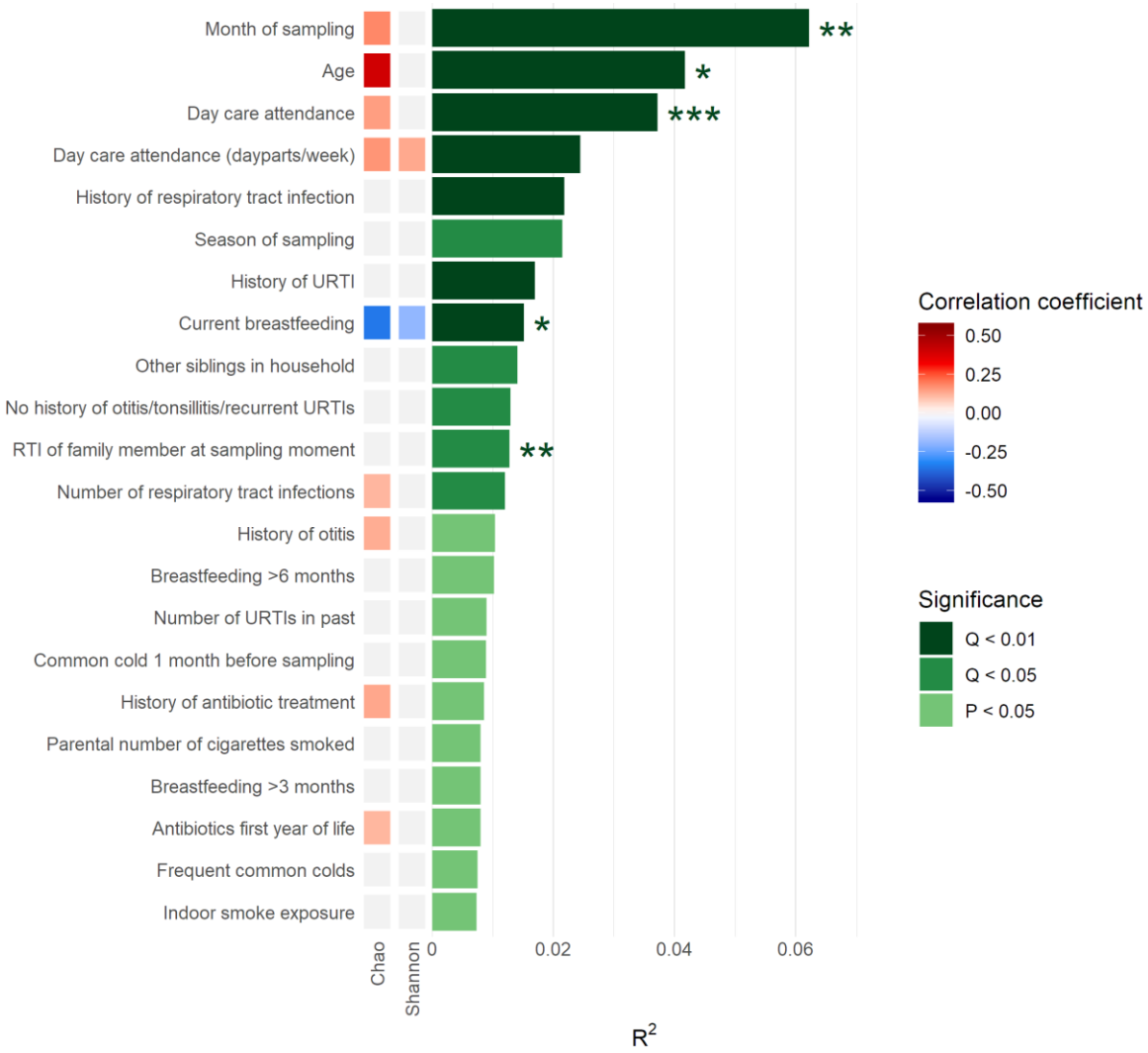
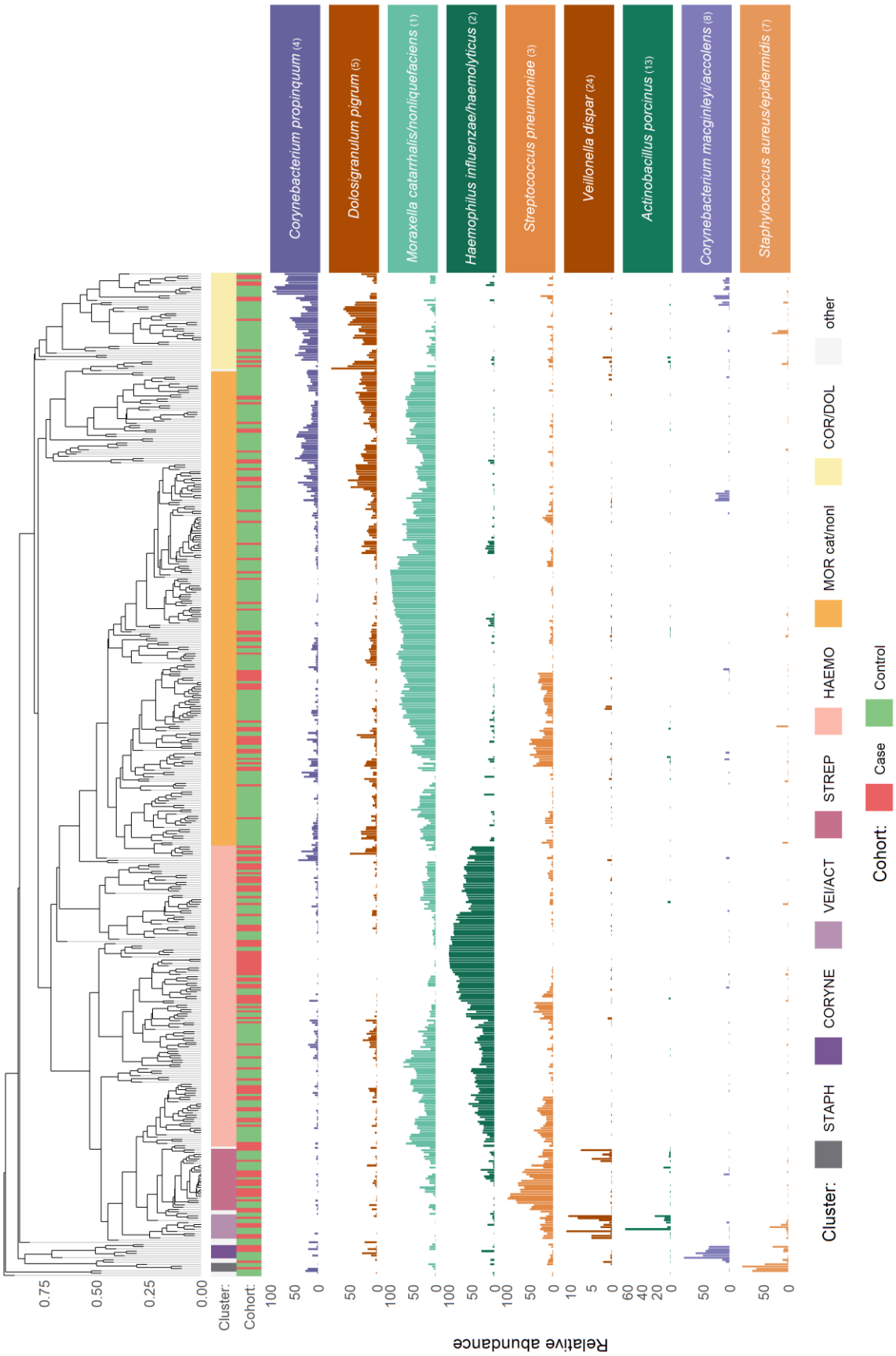


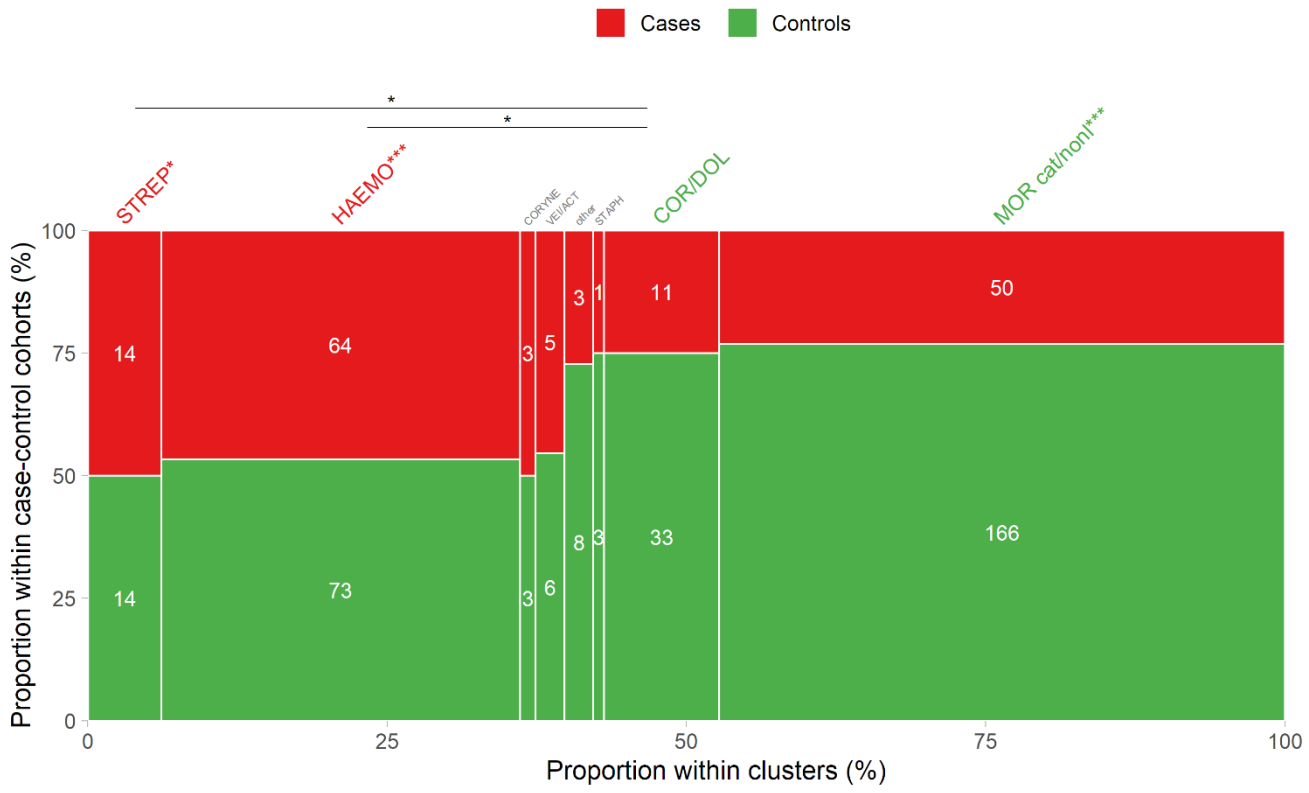
Figure S4. Hierarchical clustering of all study samples identified 15 clusters, of which 7 had 5 or more samples.

Classifier taxa for these 7 clusters were: *Staphylococcus aureus/epidermidis*; *Corynebacterium macginleyi/accolens*; *Haemophilus influenzae/haemolyticus*; *M. catarrhalis/nonliquefaciens*; *Veillonella dispar* & *Actinobacillus porcinus*; *Streptococcus pneumoniae*; *C. propinquum & D. pigrum*. Figure **A** visualizes the clustering dendrogram, including information on the distribution of the subcohorts and a heatmap of the relative abundance of the 9 classifier taxa defined by random forest analysis. **(B)** Mosaic plot showing distribution of case-control cohorts within clusters. Note that each case was matched to two controls. The *M. catarrhalis/nonliquefaciens* cluster was associated with health (green: chi-square test, p<0·001), while the *H. influenzae/haemolyticus* and *S. pneumoniae* clusters were associated with active disease (red: chi-square test, p<0·001 and p=0·049 resp.). Additional comparisons of the remaining clusters against these health and disease associated clusters, demonstrated that the *C. propinquum/D. pigrum* cluster is also related to health (green). **(C)** Boxplots of the bacterial biomass (determined by 16S qPCR) across the distinct microbiota profiles from high to low density. Significantly different: solid line = p<0·001; dashed line = p<0·01; dotted line = p<0·05. The health and disease associated clusters are shown in green and red, respectively.

A


Corynebacterium propinquum (4)
Dolosigranulum pigrum (5)
Moraxella catarrhalis/nonliquefaciens (1)
Haemophilus influenzae/haemolyticus (2)
Streptococcus pneumoniae (3)
Veillonella dispar (24)
Actinobacillus porcinus (13)
Corynebacterium macginleyi/accolens (8)
Staphylococcus aureus/epidermidis (7)
Relative abundance
Cluster:
Cohort:
STAPH
CORYNE
VEI/ACT
STREP
HAEMO
MOR cat/nonl
COR/DOL
other
Case
Control

B


Cases
Controls
*
*
STREP*
HAEMO***
CORYNE
VEI/ACT
other
STAPH
COR/DOL
MOR cat/nonl***
Proportion within case-control cohorts (%)
100
75
50
25
0
14
64
3
5
3
1
11
50
14
73
3
6
8
3
33
166
0
25
50
75
100
Proportion within clusters (%)

C

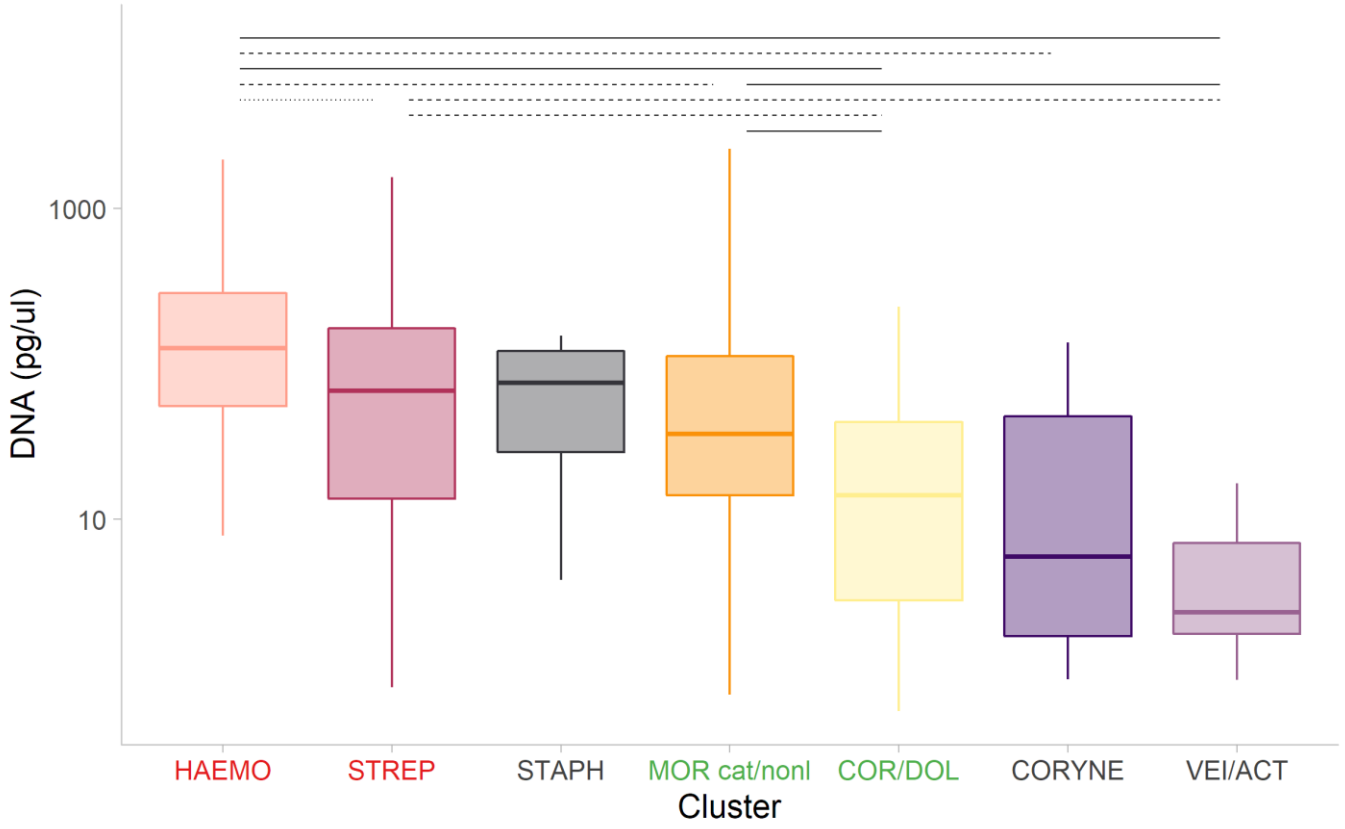
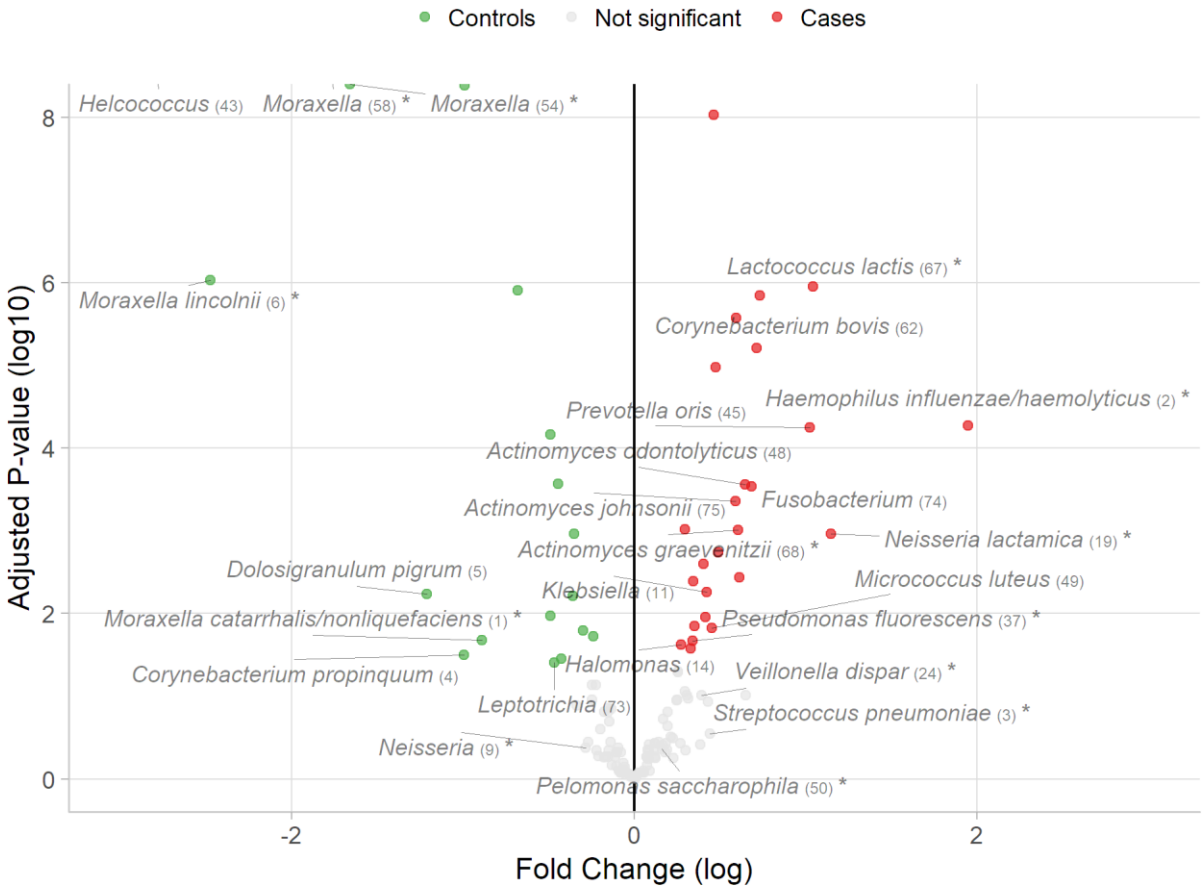
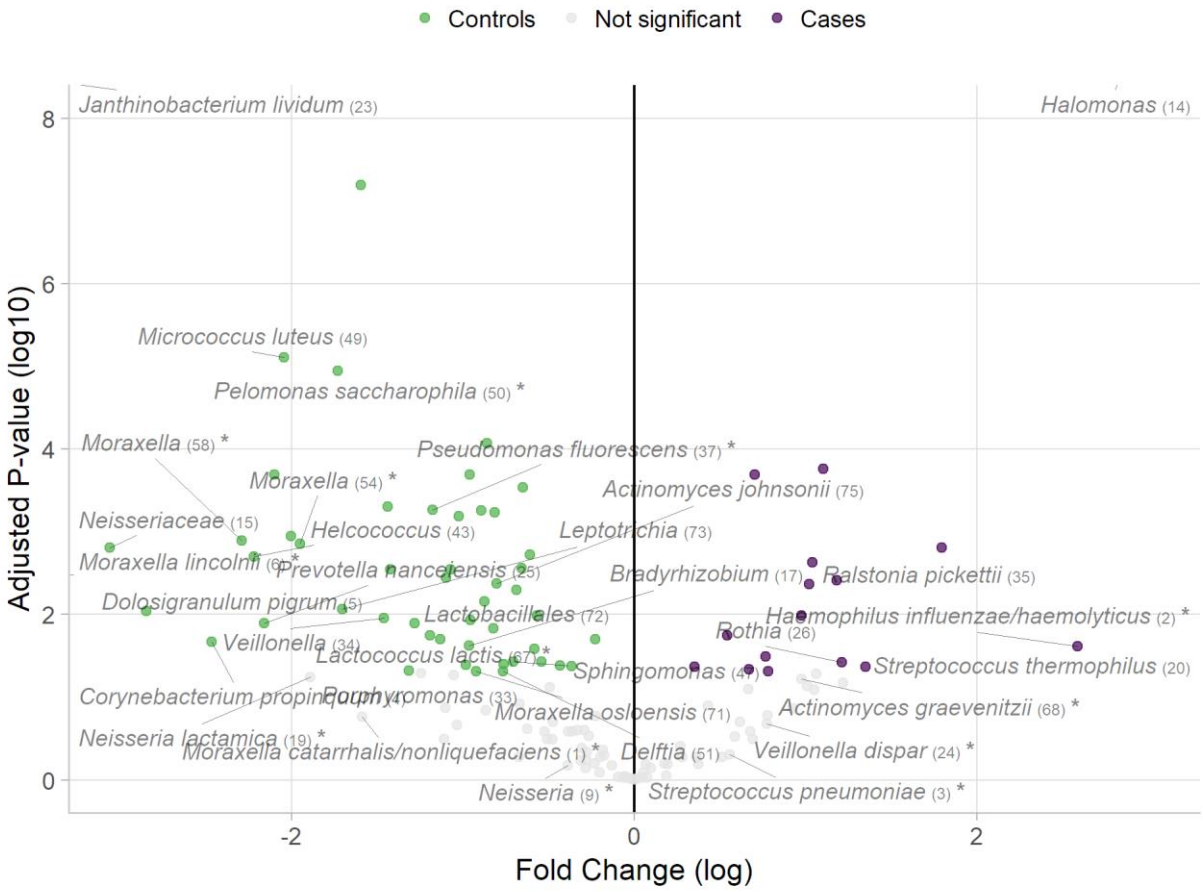

DNA (pg/ul)
1000
10
HAEMO
STREP
STAPH
MOR cat/nonl
COR/DOL
CORYNE
VEI/ACT
Cluster

Figure S5. Differential abundance of taxa that were significantly associated with health or acute disease. Taxa that were significantly increased in cases (right) or in matched controls (left) are depicted in a volcano plot. Log fold changes were obtained by *metagenomeSeq* analysis and corrected for multiple comparisons (Benjamini-Hochberg). The results are depicted from the pairwise comparison of cases and controls of the entire case-control cohort (n=457, **A**), the PICU cohort (n=288, **B**), the subcohorts stratified by phenotype (pneumonia, orange, n=108, **C**; bronchiolitis, purple, n=171, **D**; wheezing illness, green, n=100, **E**; mixed-phenotype, pink, n=78, **F**) and the subcohort stratified for need for antibiotic treatment (to-be-treated, red, n=126 **G**; not-to-be-treated, grey, n=331, **H**).

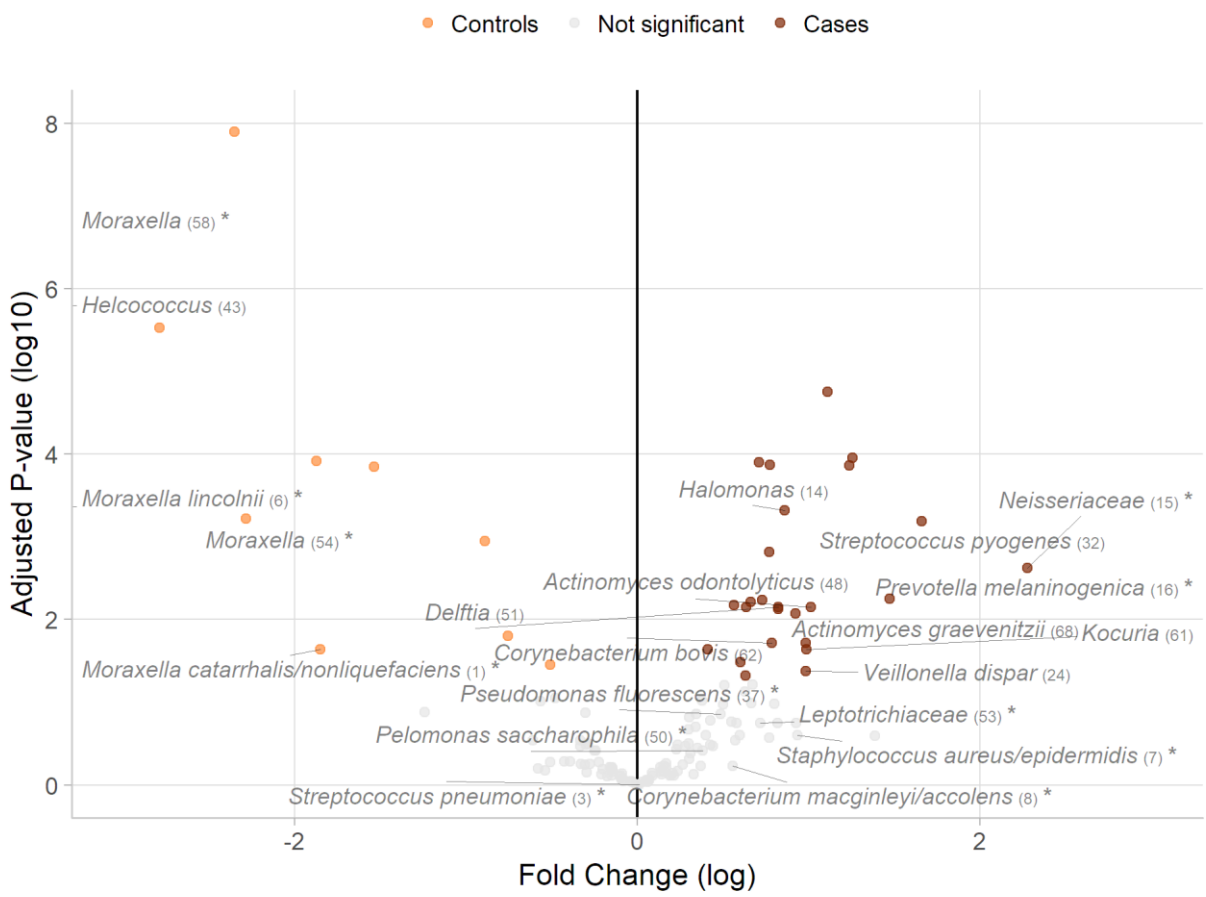
A



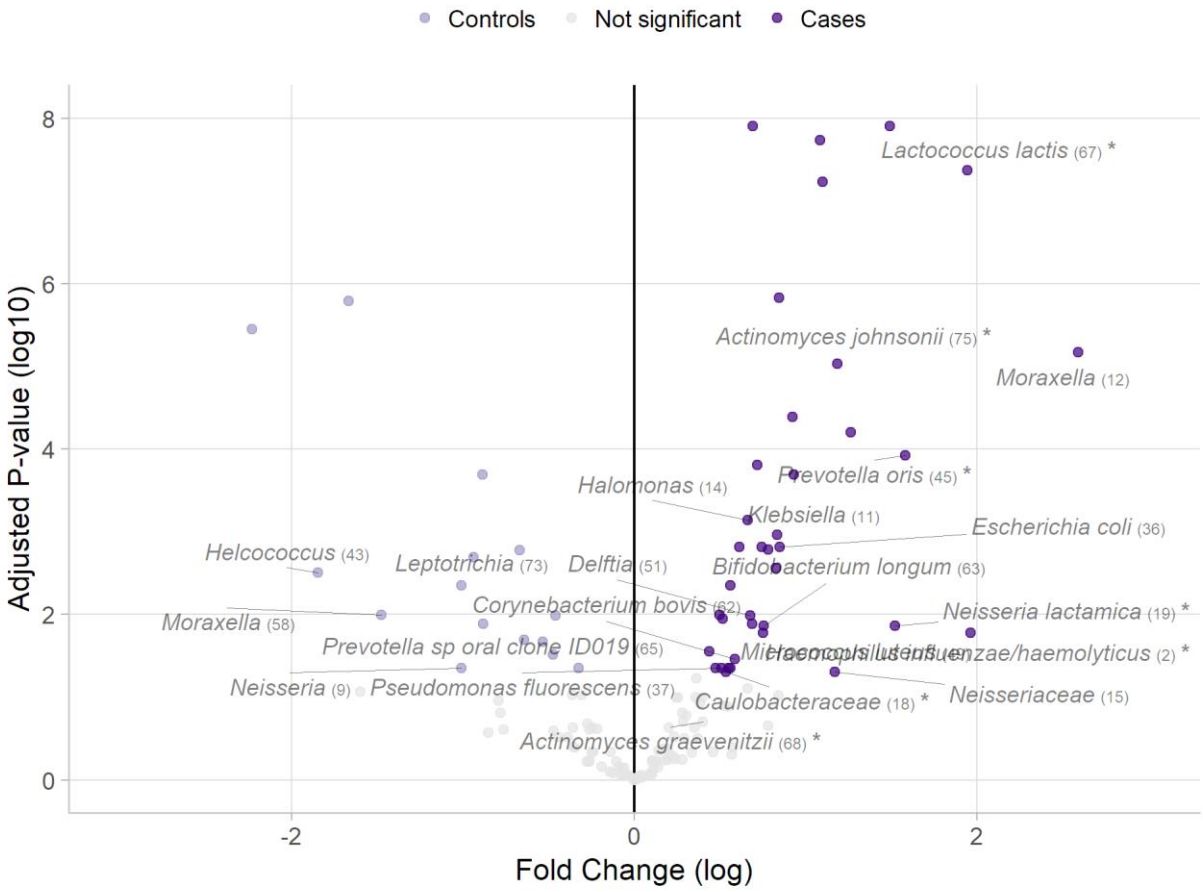
B


Controls
Not significant
Cases
Janthinobacterium lividum (23)
Halomonas (14)
Micrococcus luteus (49)
Pelomonas saccharophila (50) *
Moraxella (58) *
Pseudomonas fluorescens (37) *
Moraxella (54) *
Actinomyces johnsonii (75)
Neisseriaceae (15)
Helcococcus (43)
Leptotrichia (73)
Moraxella lincolnii (6) *
Prevotella nanceiensis (25)
Bradyrhizobium (17)
Ralstonia pickettii (35)
Dolosigranulum pigrum (5)
Lactobacillales (72)
Haemophilus influenzae/haemolyticus (2) *
Rothia (26)
Veillonella (34)
Lactococcus lactis (67) *
Sphingomonas (41)
Streptococcus thermophilus (20)
Corynebacterium propinquum (4)
Porphyromonas (33)
Moraxella osloensis (71)
Actinomyces graevenitzii (68) *
Neisseria lactamica (19) *
Moraxella catarrhalis/nonliquefaciens (1) *
Delftia (51)
Veillonella dispar (24) *
Neisseria (9) *
Streptococcus pneumoniae (3) *
Adjusted P-value (log10)
Fold Change (log)
8
6
4
2
0
-2
0
2

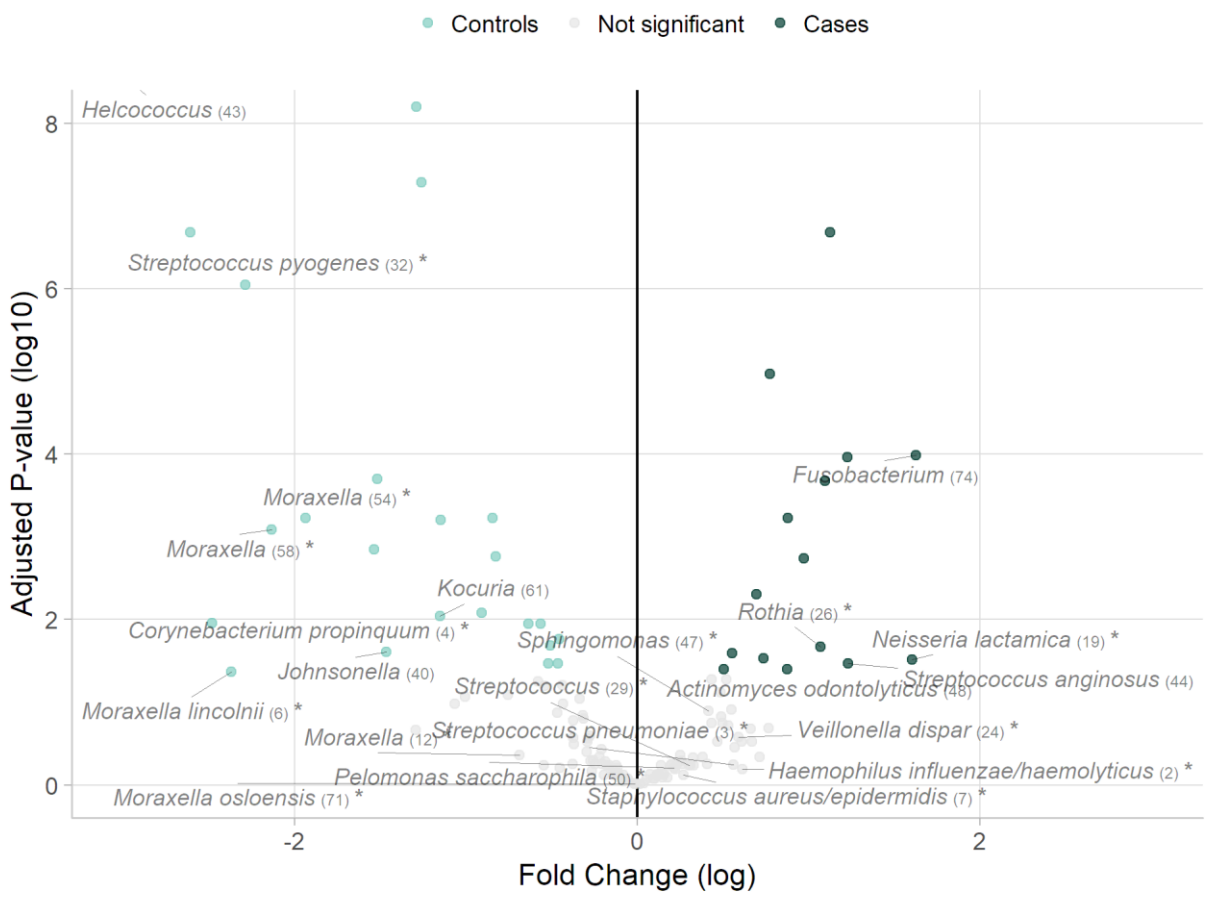
C


Controls
Not significant
Cases
Adjusted P-value (log10)
Fold Change (log)
8
6
4
2
0
-2
0
2
Moraxella (58) *
Helcococcus (43)
Moraxella lincolnii (6) *
Moraxella (54) *
Moraxella catarrhalis/nonliquefaciens (1) *
Halomonas (14)
Neisseriaceae (15) *
Streptococcus pyogenes (32)
Actinomyces odontolyticus (48)
Prevotella melaninogenica (16) *
Delftia (51)
Actinomyces graevenitzii (68)
Kocuria (61)
Corynebacterium bovis (62)
Veillonella dispar (24)
Pseudomonas fluorescens (37) *
Leptotrichiaceae (53) *
Pelomonas saccharophila (50) *
Staphylococcus aureus/epidermidis (7) *
Streptococcus pneumoniae (3) *
Corynebacterium macginleyi/accolens (8) *

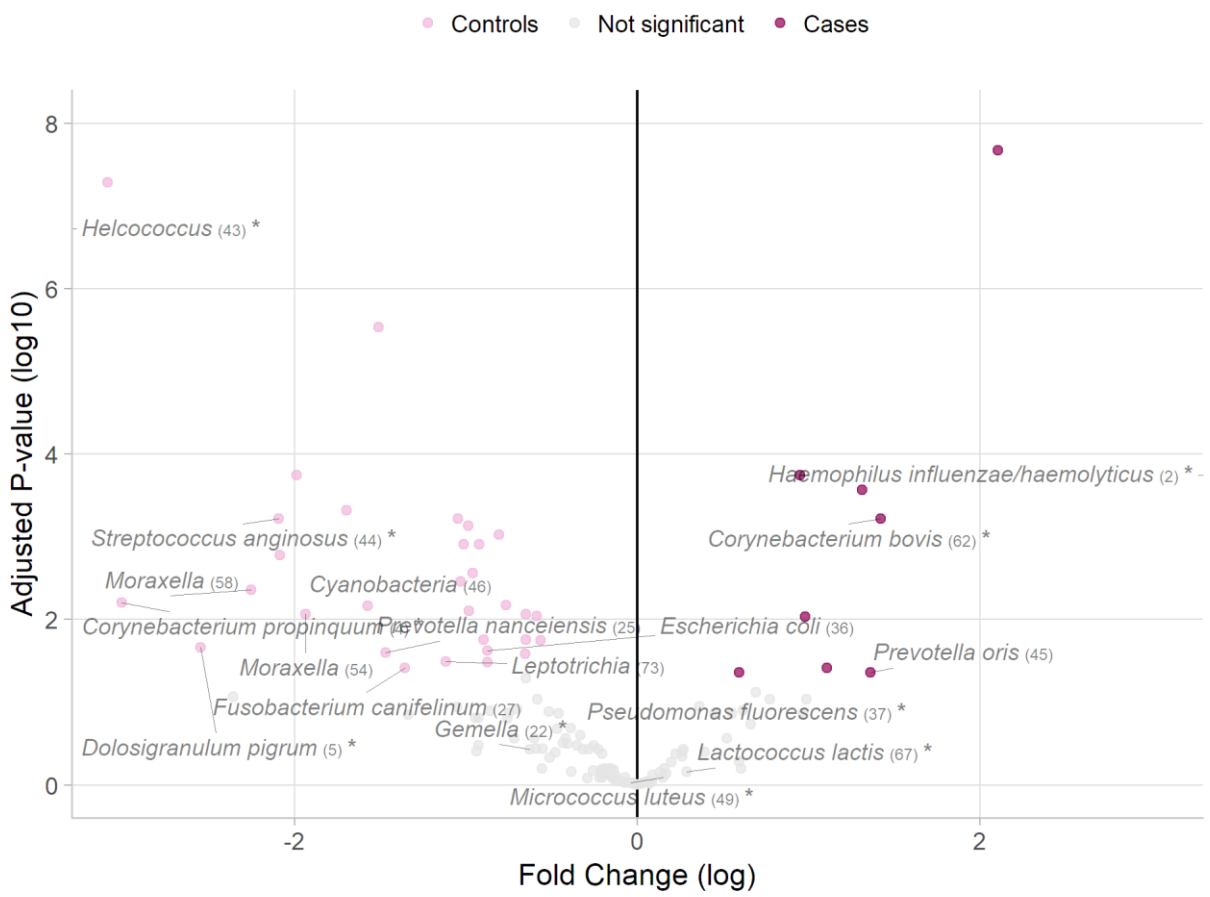
D


Controls
Not significant
Cases
Adjusted P-value (log10)
Fold Change (log)
8
6
4
2
0
-2
0
2
Lactococcus lactis (67) *
Actinomyces johnsonii (75) *
Moraxella (12)
Prevotella oris (45) *
Halomonas (14)
Klebsiella (11)
Escherichia coli (36)
Helcococcus (43)
Leptotrichia (73)
Delftia (51)
Bifidobacterium longum (63)
Corynebacterium bovis (62)
Moraxella (58)
Neisseria lactamica (19) *
Prevotella sp oral clone ID019 (65)
Haemophilus influenzae/haemolyticus (2) *
Neisseria (9)
Pseudomonas fluorescens (37)
Caulobacteraceae (18) *
Neisseriaceae (15)
Actinomyces graevenitzii (68) *

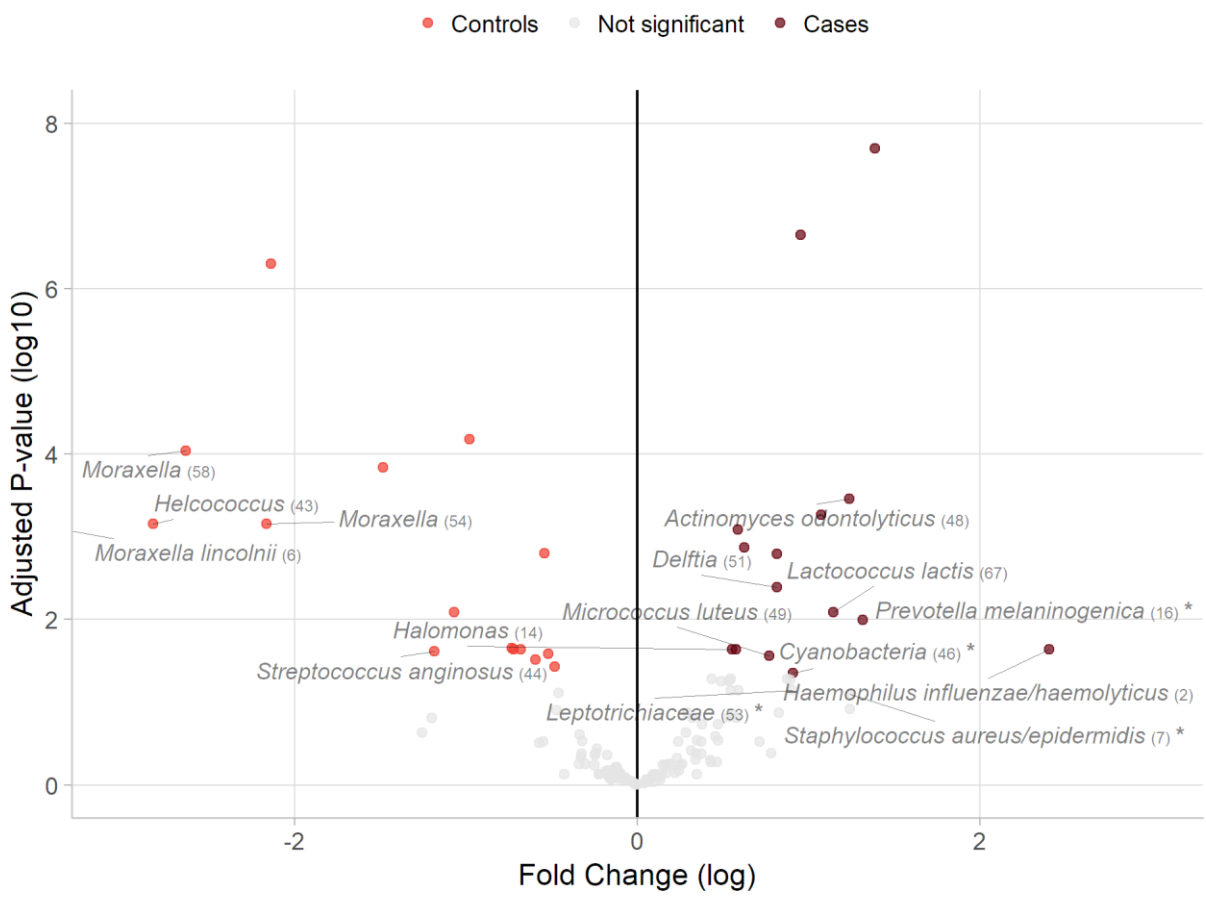
E


Controls
Not significant
Cases
Helcococcus (43)
Streptococcus pyogenes (32) *
Moraxella (54) *
Moraxella (58) *
Kocuria (61)
Corynebacterium propinquum (4) *
Johnsonella (40)
Moraxella lincolnii (6) *
Moraxella (12)
Pelomonas saccharophila (50) *
Moraxella osloensis (71) *
Fusobacterium (74)
Rothia (26) *
Sphingomonas (47) *
Neisseria lactamica (19) *
Streptococcus (29) *
Actinomyces odontolyticus (48)
Streptococcus anginosus (44)
Streptococcus pneumoniae (3) *
Veillonella dispar (24) *
Haemophilus influenzae/haemolyticus (2) *
Staphylococcus aureus/epidermidis (7) *
Adjusted P-value (log10)
Fold Change (log)
8
6
4
2
0
-2
0
2

F


Controls
Not significant
Cases
Adjusted P-value (log10)
Fold Change (log)
8
6
4
2
0
-2
0
2
Helcococcus (43) *
Streptococcus anginosus (44) *
Moraxella (58)
Cyanobacteria (46)
Corynebacterium propinquum (6) *
Prevotella nanceiensis (25)
Escherichia coli (36)
Moraxella (54)
Leptotrichia (73)
Fusobacterium canifelinum (27)
Gemella (22) *
Dolosigranulum pigrum (5) *
Haemophilus influenzae/haemolyticus (2) *
Corynebacterium bovis (62) *
Prevotella oris (45)
Pseudomonas fluorescens (37) *
Lactococcus lactis (67) *
Micrococcus luteus (49) *

G


Controls
Not significant
Cases
Adjusted P-value (log10)
Fold Change (log)
8
6
4
2
0
-2
0
2
Moraxella (58)
Helcococcus (43)
Moraxella (54)
Moraxella lincolnii (6)
Actinomyces odontolyticus (48)
Delftia (51)
Lactococcus lactis (67)
Micrococcus luteus (49)
Prevotella melaninogenica (16) *
Halomonas (14)
Cyanobacteria (46) *
Streptococcus anginosus (44)
Haemophilus influenzae/haemolyticus (2)
Leptotrichiaceae (53) *
Staphylococcus aureus/epidermidis (7) *

H

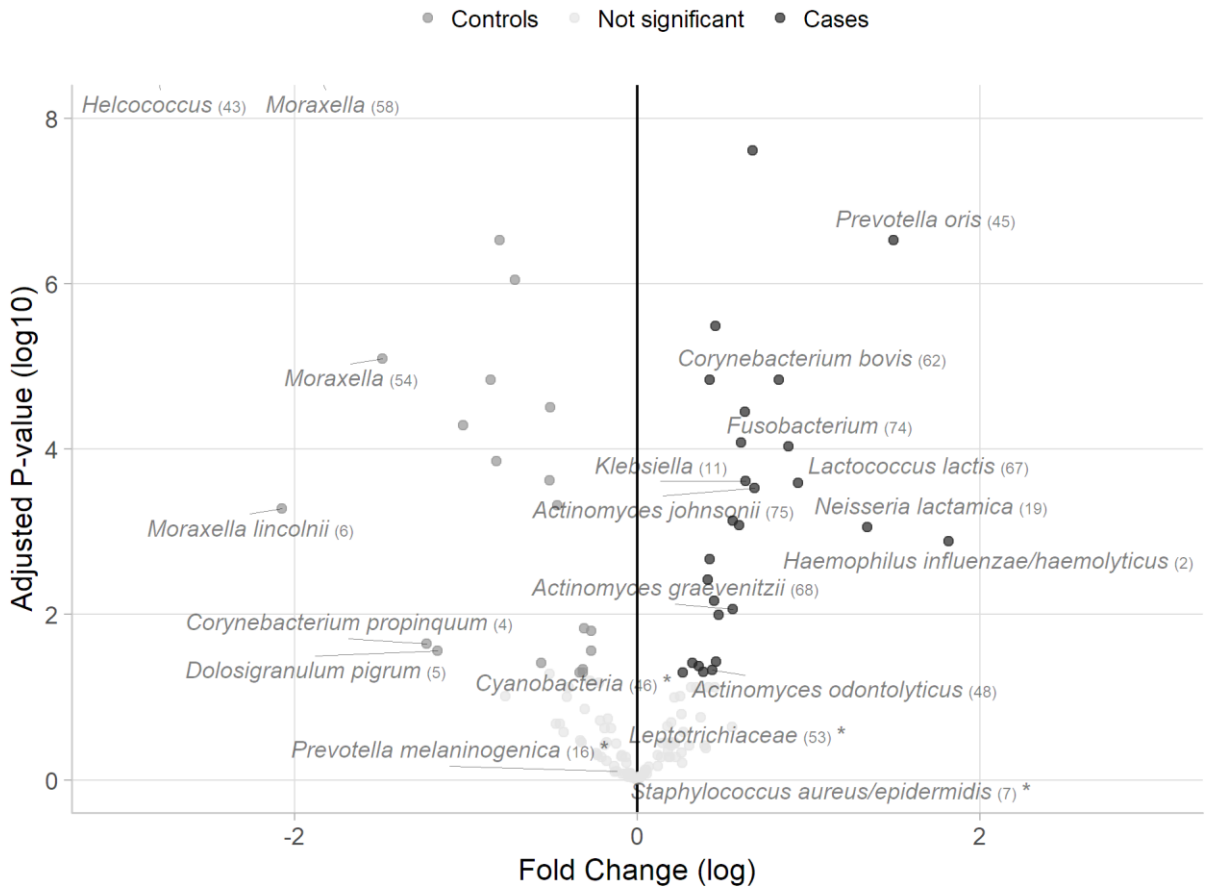

Controls
Not significant
Cases
Helcococcus (43)
Moraxella (58)
Prevotella oris (45)
Corynebacterium bovis (62)
Moraxella (54)
Fusobacterium (74)
Klebsiella (11)
Lactococcus lactis (67)
Actinomyces johnsonii (75)
Neisseria lactamica (19)
Moraxella lincolnii (6)
Haemophilus influenzae/haemolyticus (2)
Actinomyces graevenitzii (68)
Corynebacterium propinquum (4)
Dolosigranulum pigrum (5)
Cyanobacteria (46) *
Actinomyces odontolyticus (48)
Leptotrichiaceae (53) *
Prevotella melaninogenica (16) *
Staphylococcus aureus/epidermidis (7) *
Adjusted P-value (log10)
Fold Change (log)
8
6
4
2
0
-2
0
2

Figure S6. Results of viral qPCR in cases and controls.

(A) Two expert pediatricians independently classified all cases of the case-control cohort in three major disease phenotypes, i.e. pneumonia, bronchiolitis, and wheezing illness. Cases with a mixed or unclear phenotype were deemed mixed-phenotype. The figure visualized the proportions of samples positive for any virus, and the specific viruses, stratified for the pneumonia cohort (upper left), bronchiolitis cohort (upper right), wheezing illness cohort (lower left) and mixed-phenotype cohort (lower right). **(B)** The proportions for the to-be-antibiotic treated cohort (left) and not-to-be treated cohort (right). P values were calculated with conditional logistic regression accounting for the matched nature of the data.

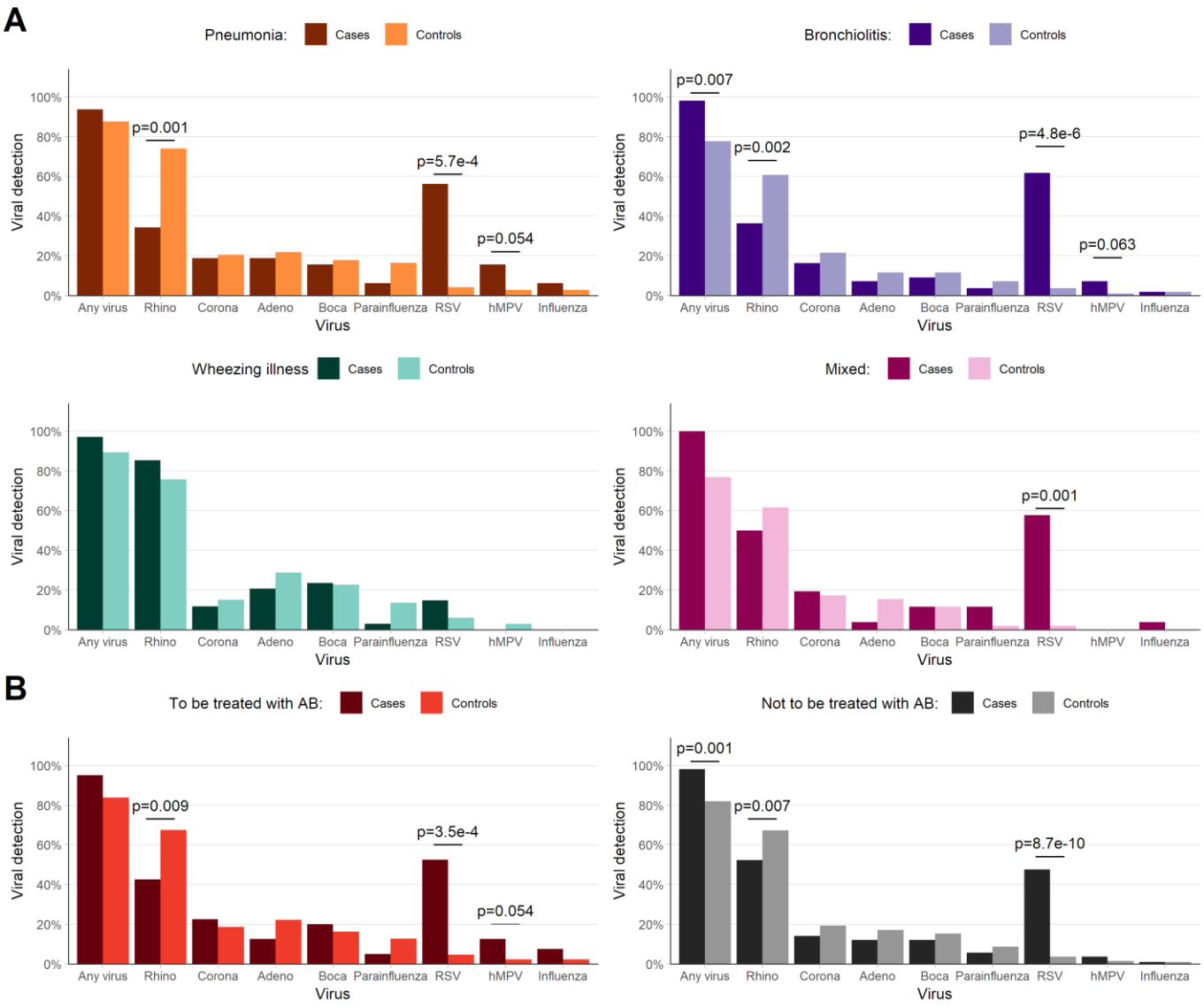


Figure S7. NMDS plots showing cases and controls stratified for the different phenotypes.
NMDS plots stratified by disease phenotypes visualizing the differences in composition of the LRTI-cases at admission compared to their matched controls (pneumonia [upper left, n=108], bronchiolitis [upper right, n=171], wheezing illness [lower left, n=100], mixed-phenotype [lower right, n=78]). Ellipses represent the standard deviation of all points within a cohort. Stress: 0·269.

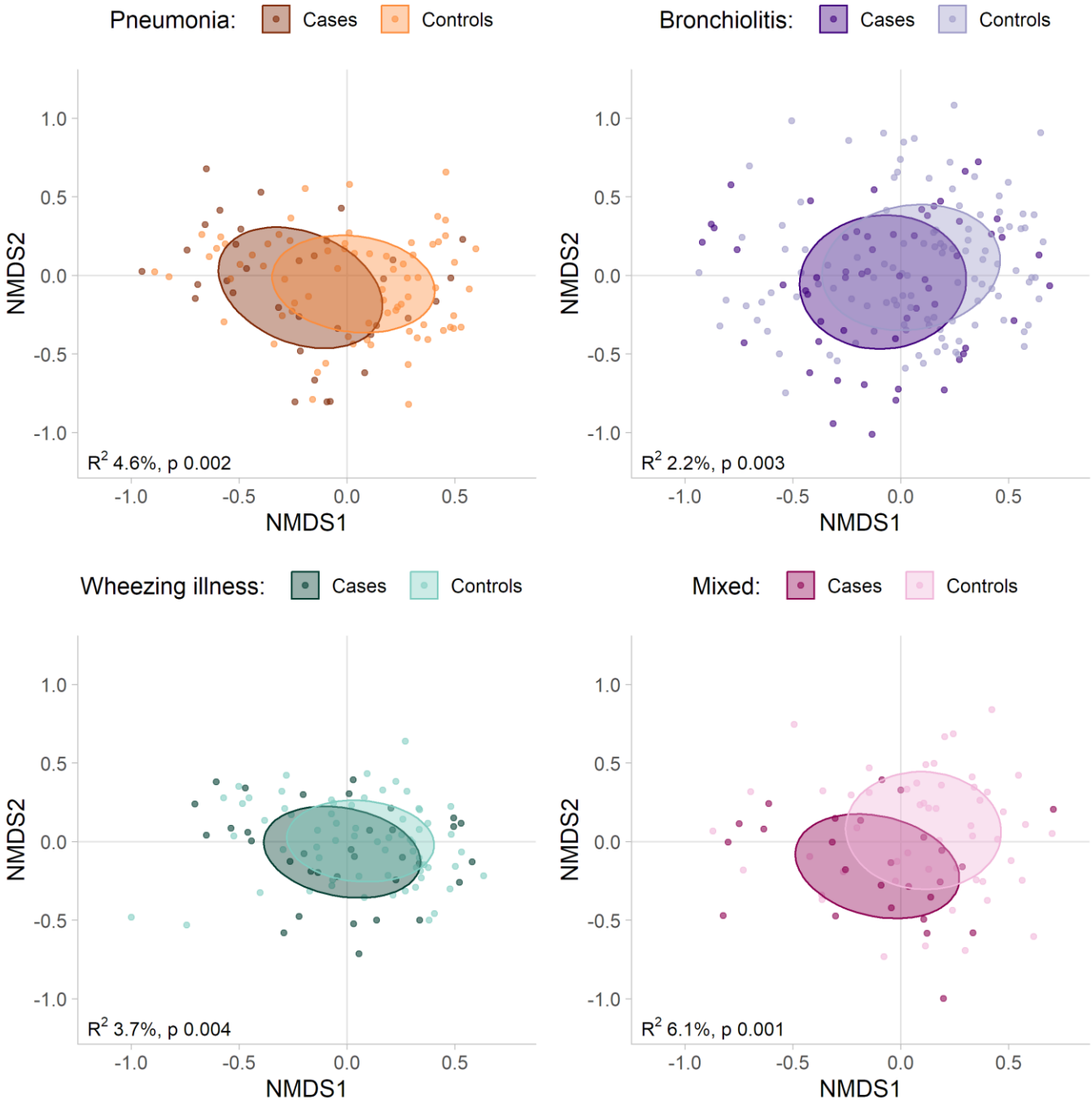


Figure S8. Random forest models classifying disease and health based on (clustered) genus level bacterial microbiota data, viral presence and patient characteristics combined.

Nineteen variables in total were discriminating cases from controls in the unstratified cohort (n=457; **B**) leading to a sparse classification model with an AUC of 0·92 (**A**). Variables are ranked in descending order based on their importance to the accuracy of the model. Variable importance was estimated by calculating the mean decrease in Gini after randomly permuting the values of each given variable (mean ± standard deviation, 100 replicates). The direction of the associations was estimated *post-hoc* using point biserial correlations (green = associated with health; red = associated with disease). The disease-discriminatory variables for the pneumonia cases (brown, n=108; **C**), bronchiolitis cases (purple, n=171; **D**), wheezing illness cases (dark green, n=100; **E**), and mixed-phenotype cases (pink, n=78; **F**) versus their matched controls are depicted in figures **C**-**F** (light colored bars are positively associated with health). The ROC curves for distinguishing disease from health of these stratified sparse random forest classifying models are depicted in **A**.

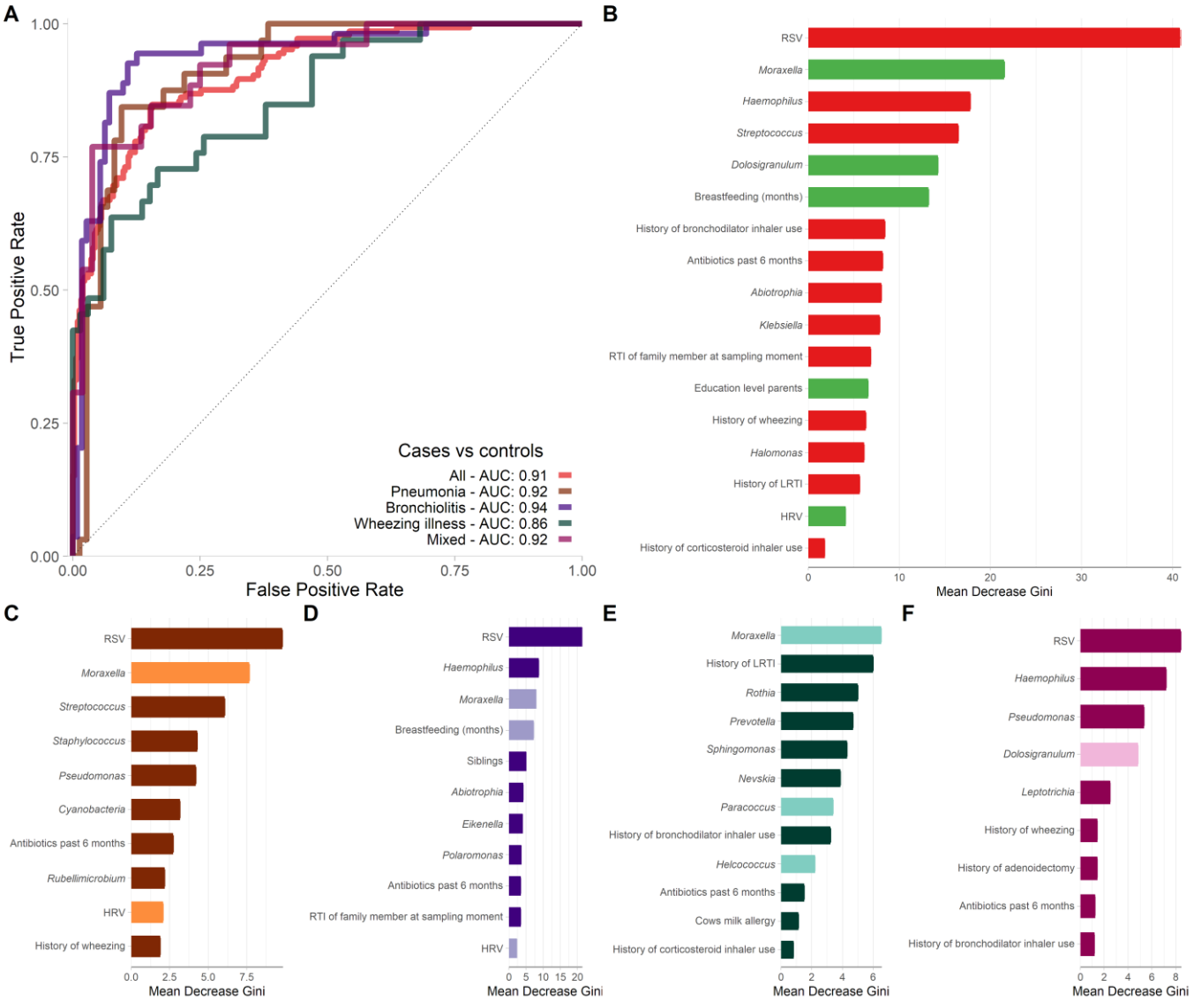
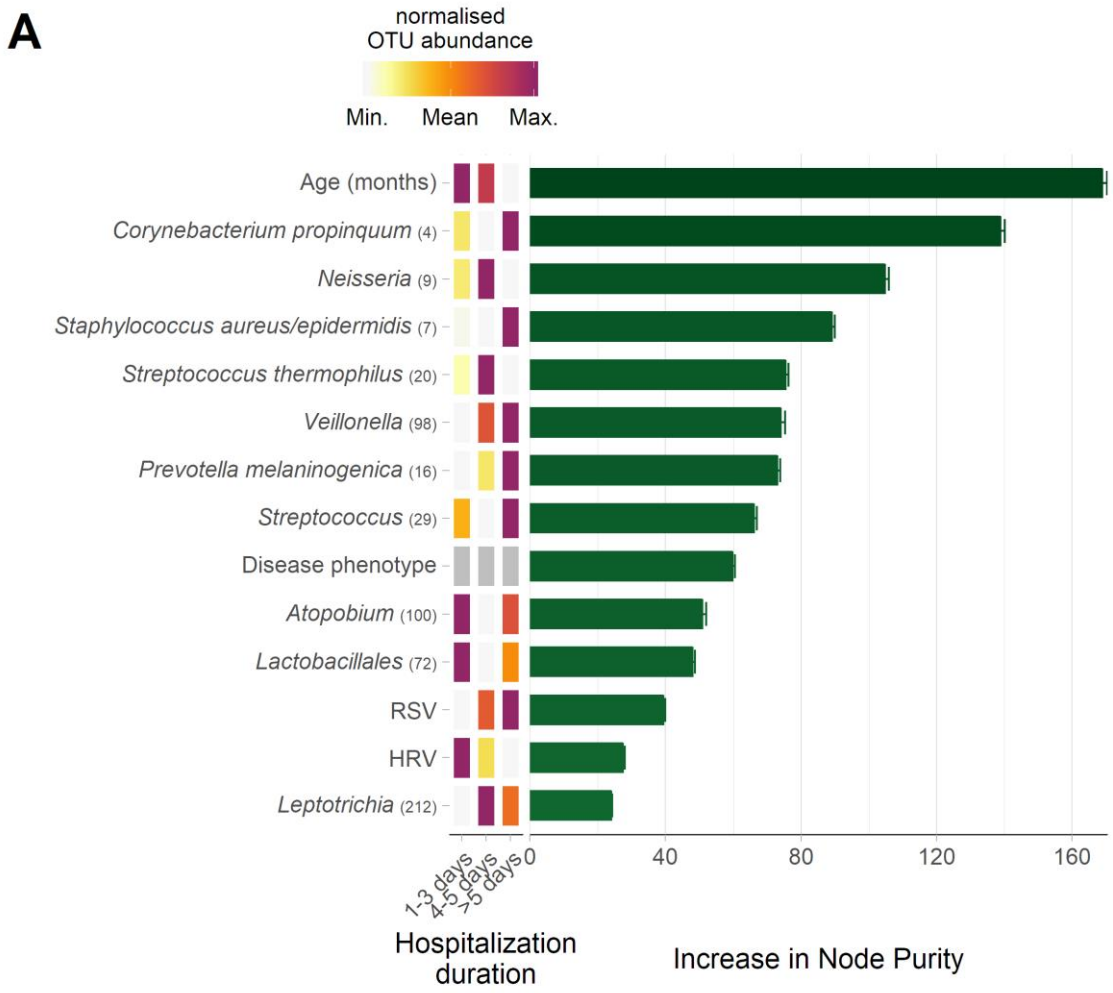
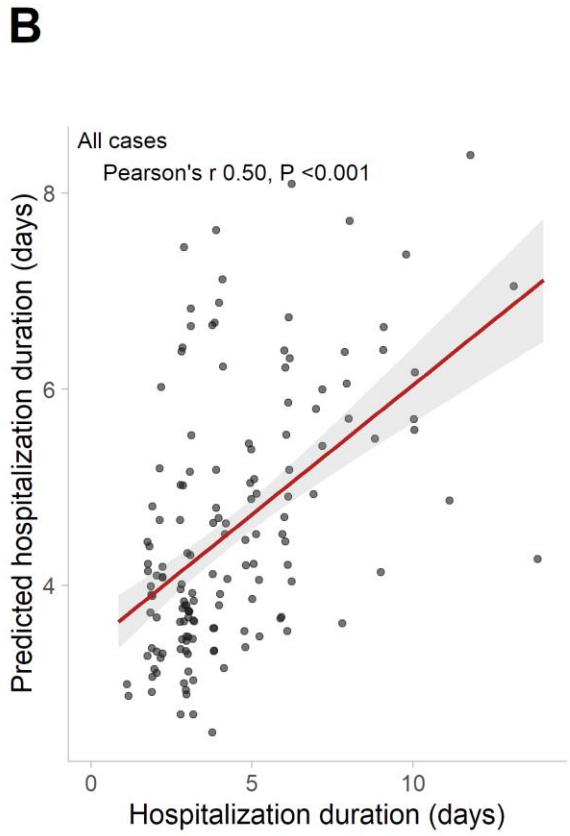


Figure S9. NMDS plots showing cases and controls stratified by future antibiotic treatment.
NMDS plots of total microbiota composition in samples stratified by LRTI-cases who were to-be-treated with antibiotics during admission (left, n=126) and those who were not-to-be-treated with antibiotics (right, n=331) compared to their respective matched controls (lighter tint). Ellipses represent the standard deviation of all points within a subcohort. Stress: 0·269.

Figure S10. Sparse random forest prediction model of the hospitalization duration.
(A) Thirteen duration-predictive variables were selected by a cross-validated stepwise ascending variable introduction strategy (VSURF) on the random forest regression model of the full dataset combining viral presence and bacterial abundance data as well as patient characteristics. The individual association of these duration-predictive variables with hospitalization duration is depicted as a heatmap of their mean values against short (1-3 days), medium (4-5 days) and long (>5 days) hospitalization, except for disease phenotype as this is a categorical variable. Colours correspond with row wise normalized values (i.e. white indicates the overall minimum value of that variable, purple indicates the overall maximum value). Next to the heatmap, the importance of the duration-predictive variables to the accuracy of the sparse model are depicted as bars (increase in node purity, mean ± standard deviation of 100 random replicates). **(B)** The sparse model comprising these 14 variables predicted duration of hospitalization with a Pearson's r of 0·50. **(C)** Remarkably, the model was only accurate for patients not-to-be-treated with antibiotics (green, n=111) but not for the to-be-treated cohort (red, n=40, p=0·033 comparing the coefficients of the linear models) underlining that antibiotic treatment is interfering with the natural process of recovery.


A
normalised
OTU abundance
Min.
Mean
Max.
Age (months)
Corynebacterium propinquum (4)
Neisseria (9)
Staphylococcus aureus/epidermidis (7)
Streptococcus thermophilus (20)
Veillonella (98)
Prevotella melaninogenica (16)
Streptococcus (29)
Disease phenotype
Atopobium (100)
Lactobacillales (72)
RSV
HRV
Leptotrichia (212)
1-3 days
4-5 days
>5 days
Hospitalization duration
0
40
80
120
160
Increase in Node Purity


B
All cases
Pearson's r 0.50, P <0.001
Predicted hospitalization duration (days)
Hospitalization duration (days)

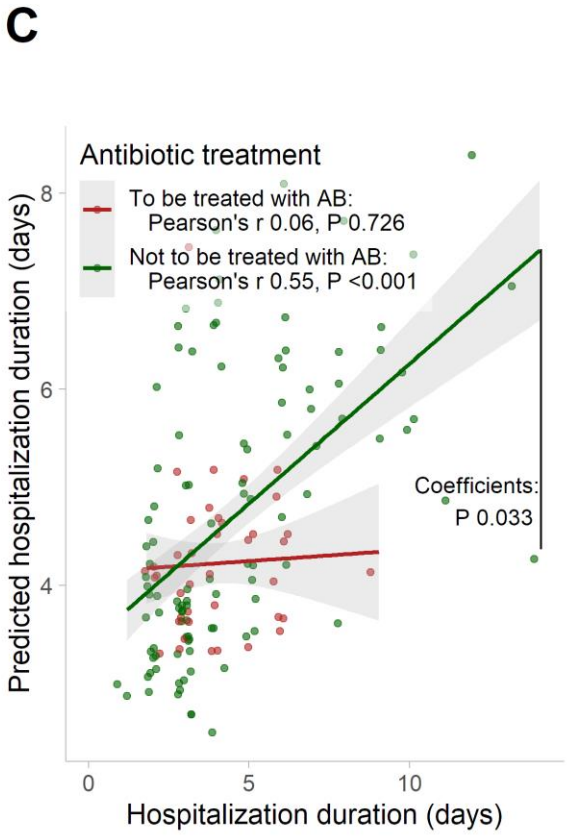
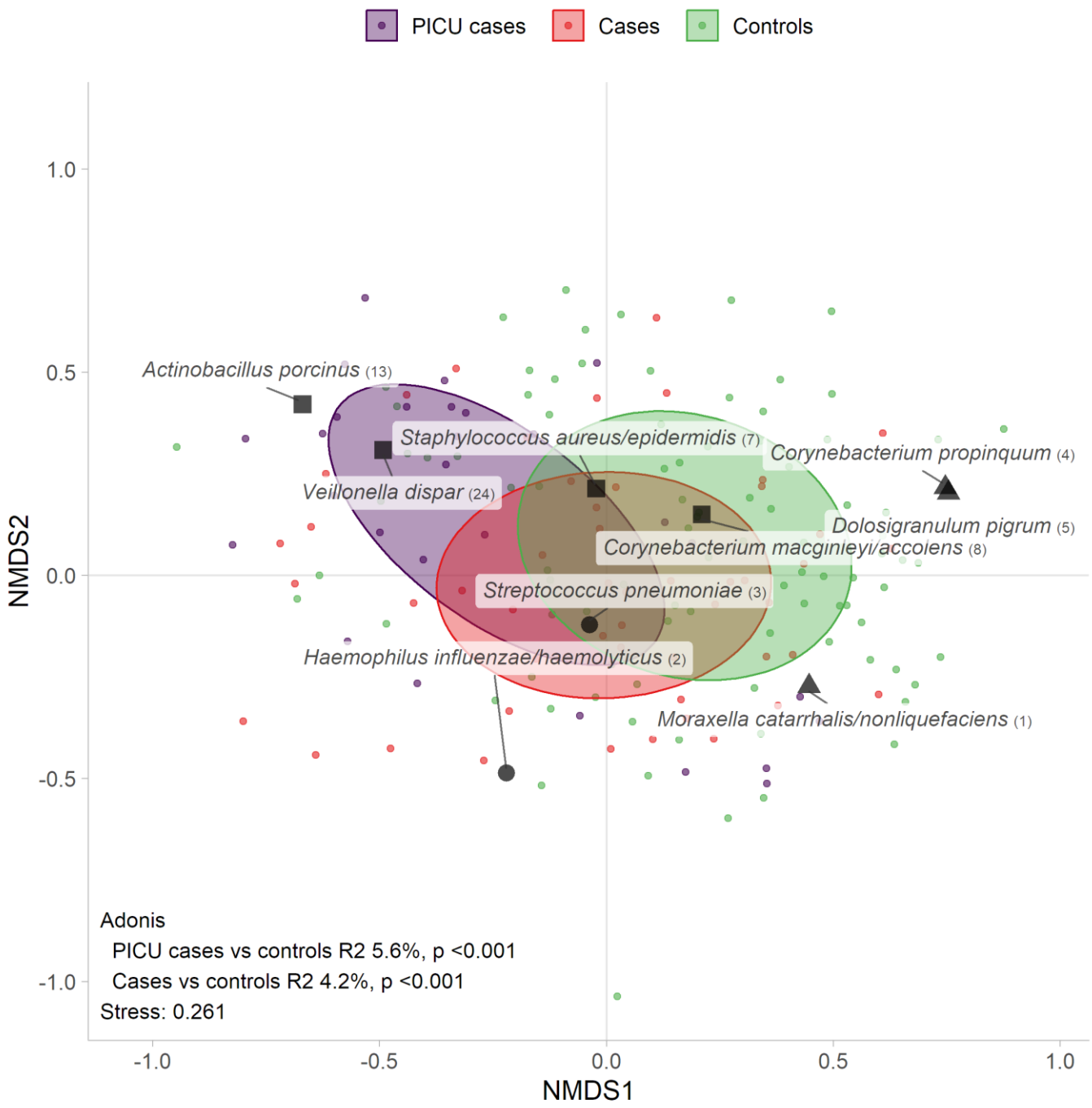

C
Antibiotic treatment
To be treated with AB:
Pearson's r 0.06, P 0.726
Not to be treated with AB:
Pearson's r 0.55, P <0.001
Coefficients:
P 0.033
Predicted hospitalization duration (days)
Hospitalization duration (days)

Figure S11. Associations between microbiota composition and disease severity and recovery using the PICU cohort.

(A) NMDS biplot depicting the individual microbiota composition colored by subcohort: PICU-cohort cases at admission (purple), case-control cohort cases at admission (red) and controls (green). Ellipses represent the standard deviation of all points within a subcohort. Total microbiota composition of PICU-cases is more deviated from that of healthy controls than cases admitted in a general pediatric ward and **(B)** shifts towards a composition with a higher abundance of H. influenzae/haemolyticus and S. pneumoniae, and a lower abundance of *Moraxella*, *Dolosigranulum*, and *Corynebacterium* spp. Both NMDS biplots also depict the 9 biomarkers determined in **Supplementary Figure 4**.

A



B

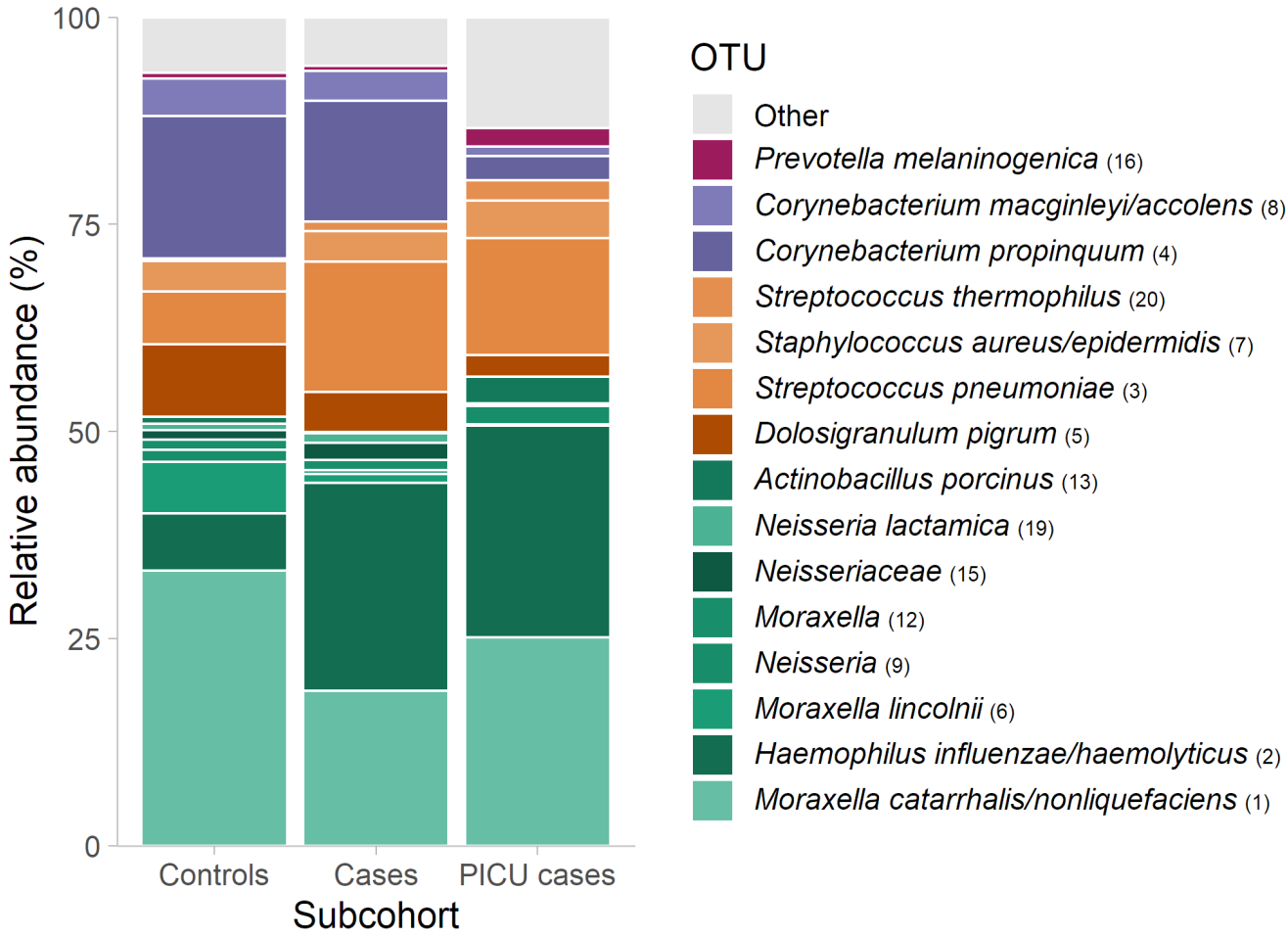

100
75
50
25
0
Relative abundance (%)
Controls
Cases
PICU cases
Subcohort
OTU
Other
Prevotella melaninogenica (16)
Corynebacterium macginleyi/accolens (8)
Corynebacterium propinquum (4)
Streptococcus thermophilus (20)
Staphylococcus aureus/epidermidis (7)
Streptococcus pneumoniae (3)
Dolosigranulum pigrum (5)
Actinobacillus porcinus (13)
Neisseria lactamica (19)
Neisseriaceae (15)
Moraxella (12)
Neisseria (9)
Moraxella lincolnii (6)
Haemophilus influenzae/haemolyticus (2)
Moraxella catarrhalis/nonliquefaciens (1)

Figure S12. Rarefaction curves for all study samples on raw count data approached plateau.
Results of rarefaction analyses are depicted for the PICU cohort (top) and case-control cohort samples (bottom), separated for sampling cohort.

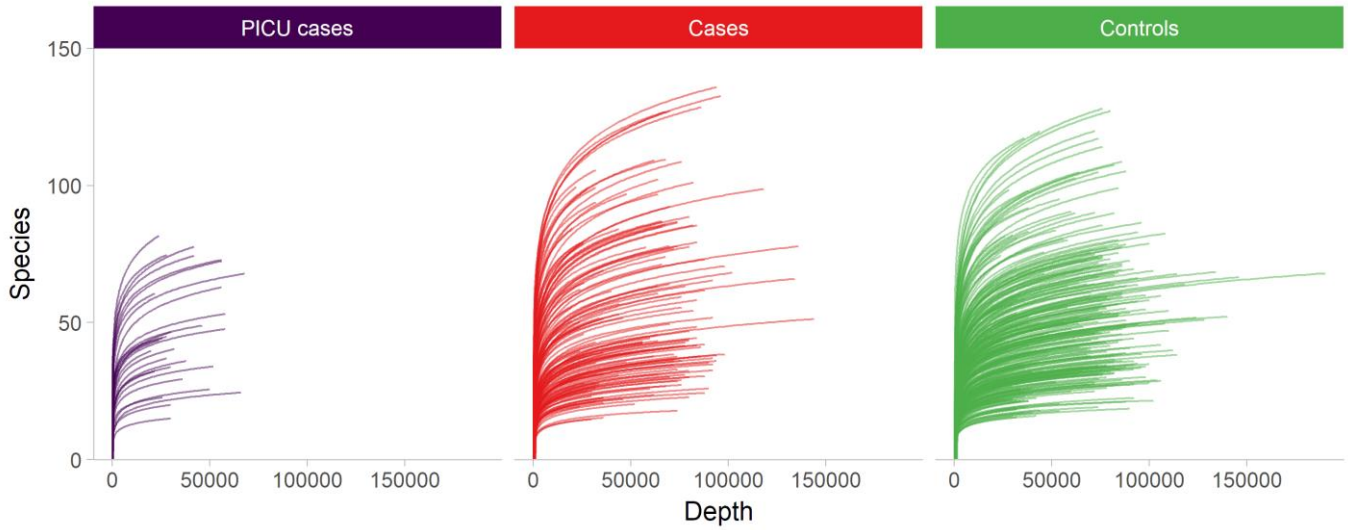
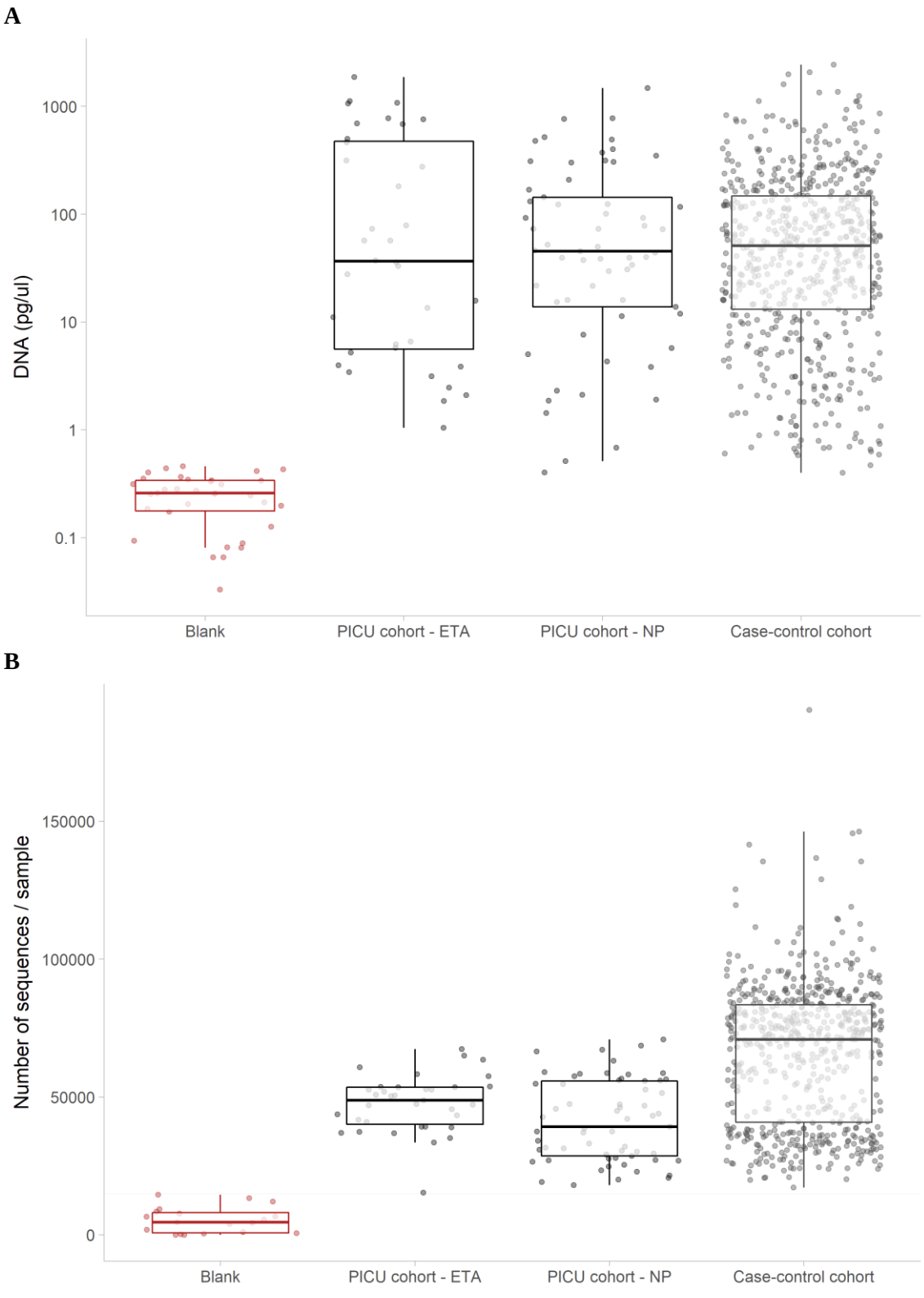
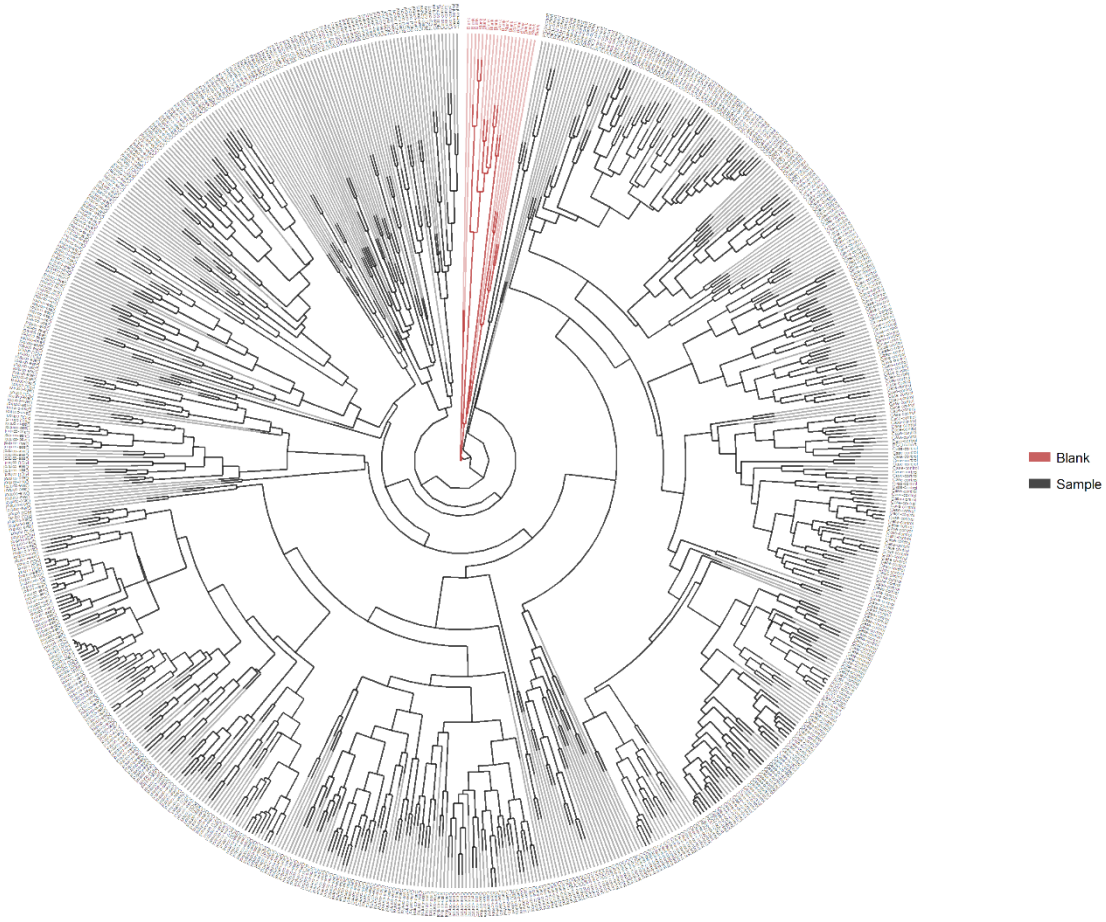


Figure S13. Samples are distinct from blanks.
Both the DNA load and the number of reads in the DNA isolation blanks and PCR blanks (red) were at least an order of magnitude lower compared to the samples (black; **A** & **B**). Boxplots of samples are divided by cohort and sample type. (**C**) visualizes the hierarchical clustering dendrogram, which clearly separates the blanks (red) from the samples (grey). ETA = endotracheal aspirate, NP = nasopharynx.



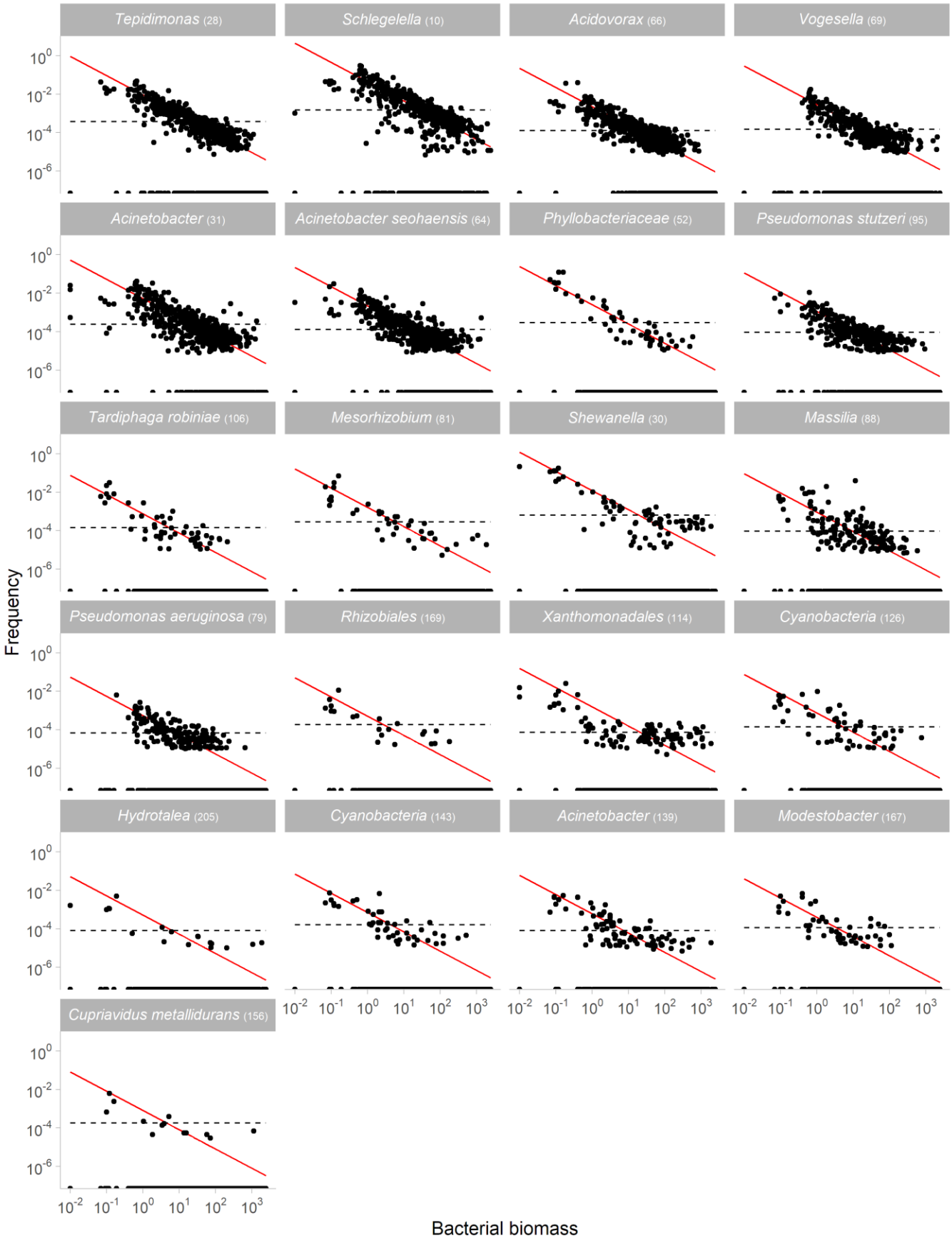
C



358

359

Figure S14. Twenty-one OTUs were identified as contaminants.
The frequency of each OTU is depicted as a function of the bacterial biomass. The dashed black line shows the model of a noncontaminant sequence feature for which frequency is expected to be independent of the input DNA concentration. The red line shows the model of a contaminant sequence feature, for which frequency is expected to be inversely proportional to input DNA concentration, as contaminating DNA will make up a larger fraction of the total DNA in samples with very little total DNA.


Tepidimonas (28)
Schlegelella (10)
Acidovorax (66)
Vogesella (69)
Acinetobacter (31)
Acinetobacter seohaensis (64)
Phyllobacteriaceae (52)
Pseudomonas stutzeri (95)
Tardiphaga robiniae (106)
Mesorhizobium (81)
Shewanella (30)
Massilia (88)
Pseudomonas aeruginosa (79)
Rhizobiales (169)
Xanthomonadales (114)
Cyanobacteria (126)
Hydrotalea (205)
Cyanobacteria (143)
Acinetobacter (139)
Modestobacter (167)
Cupriavidus metallidurans (156)
Frequency
Bacterial biomass

366

References

1 Liu L, Oza S, Hogan D, *et al.* Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *Lancet* 2016; **388**: 3027–35.

2 Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015; **386**: 743–800.

3 Rudan I. Epidemiology and etiology of childhood pneumonia. *Bull World Health Organ* 2008; **86**: 408–16.

4 Jackson S, Mathews KH, Pulanic D, *et al.* Risk factors for severe acute lower respiratory infections in children: a systematic review and meta-analysis. *Croat Med J* 2013; **54**: 110–21.

5 Tsolia MN, Psarras S, Bossios A, *et al.* Etiology of Community-Acquired Pneumonia in Hospitalized School-Age Children: Evidence for High Prevalence of Viral Infections. *Clin Infect Dis* 2004; **39**: 681–6.

6 Michelow IC, Olsen K, Lozano J, *et al.* Epidemiology and clinical characteristics of community-acquired pneumonia in hospitalized children. *Pediatrics* 2004; **113**: 701–7.

7 Man WH, de Steenhuijsen Piters WAA, Bogaert D. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat Rev Microbiol* 2017; **15**: 259–70.

8 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**: 704–15.

9 Biesbroek G, Tsivtsivadze E, Sanders EAM, *et al.* Early Respiratory Microbiota Composition Determines Bacterial Succession Patterns and Respiratory Health in Children. *Am J Respir Crit Care Med* 2014; **190**: 1283–92.

10 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**: 1582–90.

11 Laufer AS, Metlay JP, Gent JF, Fennie KP, Kong Y, Pettigrew MM. Microbial communities of the upper respiratory tract and otitis media in children. *MBio* 2011; **2**: e00245-10.

12 Pettigrew MM, Laufer AS, Gent JF, Kong Y, Fennie KP, Metlay JP. Upper respiratory tract microbial communities, acute otitis media pathogens, and antibiotic use in healthy and sick children. *Appl Environ Microbiol* 2012; **78**: 6262–70.

13 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**: 1104–15.

14 WHO. IMCI chart booklet. World Health Organization, 2014.

15 Scott JAG, Wonodi C, Moïsi JC, *et al.* The definition of pneumonia, the assessment of severity, and clinical standardization in the Pneumonia Etiology Research for Child Health study. *Clin Infect Dis* 2012; **54 Suppl 2**: S109-16.

16 Marsh RL, Kaestli M, Chang AB, *et al.* The microbiota in bronchoalveolar lavage from young children with chronic lung disease includes taxa present in both the oropharynx and nasopharynx. *Microbiome* 2016; **4**: 37.

17 van de Pol AC, Wolfs TFW, van Loon AM, *et al.* Molecular quantification of respiratory syncytial virus in respiratory samples: reliable detection during the initial phase of infection. *J Clin Microbiol* 2010; **48**: 3569–74.

18 Perkins SM, Webb DL, Torrance SA, *et al.* Comparison of a Real-Time Reverse Transcriptase PCR Assay and a Culture Technique for Quantitative Assessment of Viral Load in Children Naturally Infected with Respiratory Syncytial Virus. *J Clin Microbiol* 2005; **43**: 2356–62.

19 Bassis CM, Erb-Downward JR, Dickson RP, *et al.* Analysis of the Upper Respiratory Tract Microbiotas as the Source of the Lung and Gastric Microbiotas in Healthy Individuals. *MBio* 2015; **6**: e00037-15.

20 Dickson RP, Erb-Downward JR, Freeman CM, *et al.* Bacterial Topography of the Healthy Human Lower Respiratory Tract. *MBio* 2017; **8**: e02287-16.

21 Rhedin S, Lindstrand A, Hjelmgren A, *et al.* Respiratory viruses associated with community-acquired pneumonia in children: matched case–control study. *Thorax* 2015; **70**: 847–53.

22 Jain S, Williams DJ, Arnold SR, *et al.* Community-Acquired Pneumonia Requiring Hospitalization among U.S. Children. *N Engl J Med* 2015; **372**: 835–45.

23 Dickson RP, Erb-Downward JR, Huffnagle GB. Towards an ecology of the lung: New conceptual models of pulmonary microbiology and pneumonia pathogenesis. *Lancet Respir Med* 2014; **2**: 238–46.

24 Said HS, Suda W, Nakagome S, *et al.* Dysbiosis of Salivary Microbiota in Inflammatory Bowel Disease and

Its Association With Oral Immunological Biomarkers. *DNA Res* 2014; **21**: 15–25.
25 Larsen JM. The immune response to Prevotella bacteria in chronic inflammatory disease. *Immunology* 2017; **151**: 363–74.
26 Marks LR, Davidson BA, Knight PR, Hakansson AP. Interkingdom signaling induces Streptococcus pneumoniae biofilm dispersion and transition from asymptomatic colonization to disease. *MBio* 2013; **4**: e00438-13.
27 O'Donnell PM, Aviles H, Lyte M, Sonnenfeld G. Enhancement of in vitro growth of pathogenic bacteria by norepinephrine: importance of inoculum density and role of transferrin. *Appl Environ Microbiol* 2006; **72**: 5097–9.
28 Clarke MB, Hughes DT, Zhu C, Boedeker EC, Sperandio V. The QseC sensor kinase: A bacterial adrenergic receptor. *Proc Natl Acad Sci* 2006; **103**: 10420–5.
29 Sakwinska O, Schmid VB, Berger B, *et al.* Nasopharyngeal microbiota in healthy children and pneumonia patients. *J Clin Microbiol* 2014; **52**: 1590–4.
30 Pettigrew MM, Gent JF, Kong Y, *et al.* Association of sputum microbiota profiles with severity of community-acquired pneumonia in children. *BMC Infect Dis* 2016; **16**: 317.
31 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**: 140612135546007.
32 Brook I. Prevotella and Porphyromonas infections in children. *J Med Microbiol* 1995; **42**: 340–7.
33 Prevaes SMPJ, de Winter-de Groot KM, Janssens HM, *et al.* Development of the Nasopharyngeal Microbiota in Infants with Cystic Fibrosis. *Am J Respir Crit Care Med* 2016; **193**: 504–15.
34 Ramsey MM, Freire MO, Gabrilska RA, Rumbaugh KP, Lemon KP. Staphylococcus aureus Shifts toward Commensalism in Response to Corynebacterium Species. *Front Microbiol* 2016; **7**: 1230.
35 Bomar L, Brugger SD, Yost BH, Davies SS, Lemon KP. Corynebacterium accolens Releases Antipneumococcal Free Fatty Acids from Human Nostril and Skin Surface Triacylglycerols. *MBio* 2016; **7**: e01725-15.
36 Pendleton KM, Erb-Downward JR, Bao Y, *et al.* Rapid Pathogen Identification in Bacterial Pneumonia Using Real-Time Metagenomics. *Am J Respir Crit Care Med* 2017; **196**: 1610–2.
37 Bosch AATM, Levin E, van Houten MA, *et al.* Development of Upper Respiratory Tract Microbiota in Infancy is Affected by Mode of Delivery. *EBioMedicine* 2016; **9**: 336–45.
38 Luna PN, Hasegawa K, Ajami NJ, *et al.* The association between anterior nares and nasopharyngeal microbiota in infants hospitalized for bronchiolitis. *Microbiome* 2018; **6**: 2.
39 Salter SJ, Turner C, Watthanaworawit W, *et al.* A longitudinal study of the infant nasopharyngeal microbiota: The effects of age, illness and antibiotic use in a cohort of South East Asian children. *PLoS Negl Trop Dis* 2017; **11**: e0005975.
40 Kanmani P, Clua P, Vizoso-Pinto MG, *et al.* Respiratory Commensal Bacteria Corynebacterium pseudodiphtheriticum Improves Resistance of Infant Mice to Respiratory Syncytial Virus and Streptococcus pneumoniae Superinfection. *Front Microbiol* 2017; **8**: 1613.
41 Eggleston PA, Ward BH, Pierson WE, Bierman CW. Radiographic abnormalities in acute asthma in children. *Pediatrics* 1974; **54**: 442–9.
42 Domínguez J, Blanco S, Rodrigo C, *et al.* Usefulness of urinary antigen detection by an immunochromatographic test for diagnosis of pneumococcal pneumonia in children. *J Clin Microbiol* 2003; **41**: 2161–3.
43 Bosch AATM, Biesbroek G, Trzcinski K, Sanders EAM, Bogaert D. Viral and bacterial interactions in the upper respiratory tract. *PLoS Pathog* 2013; **9**: e1003057.
44 Hasegawa K, Mansbach JM, Ajami NJ, *et al.* Association of nasopharyngeal microbiota profiles with bronchiolitis severity in infants hospitalised for bronchiolitis. *Eur Respir J* 2016; **48**: 1329–39.
45 Beigelman A, Bacharier LB. Early-life respiratory infections and asthma development. *Curr Opin Allergy Clin Immunol* 2016; **16**: 172–8.
46 Bogaert D, Keijser B, Huse S, *et al.* Variability and Diversity of Nasopharyngeal Microbiota in Children: A Metagenomic Analysis. *PLoS One* 2011; **6**: e17035.
47 Yatsunenko T, Rey FE, Manary MJ, *et al.* Human gut microbiome viewed across age and geography. *Nature* 2012; **486**: 222–7.
48 Markle JGM, Frank DN, Mortin-Toth S, *et al.* Sex Differences in the Gut Microbiome Drive Hormone-Dependent Regulation of Autoimmunity. *Science* 2013; **339**: 1084–8.
49 Mueller S, Saunier K, Hanisch C, *et al.* Differences in Fecal Microbiota in Different European Study Populations in Relation to Age, Gender, and Country: a Cross-Sectional Study. *Appl Environ Microbiol* 2006; **72**: 1027–33.

50 Pillet S, Lardeux M, Dina J, *et al.* Comparative Evaluation of Six Commercialized Multiplex PCR Kits for the Diagnosis of Respiratory Infections. *PLoS One* 2013; **8**: e72174.

51 Carvalho M da GS, Tondella ML, McCaustland K, *et al.* Evaluation and improvement of real-time PCR assays targeting lytA, ply, and psaA genes for detection of pneumococcal DNA. *J Clin Microbiol* 2007; **45**: 2460–6.

52 Wyllie AL, Chu MLJN, Schellens MHB, *et al.* Streptococcus pneumoniae in saliva of Dutch primary school children. *PLoS One* 2014; **9**: e102045.

53 Joshi N, Fass J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. 2011. https://github.com/najoshi/sickle.

54 Nikolenko SI, Korobeynikov AI, Alekseyev MA. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* 2013; **14**: S7.

55 Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 2012; **13**: 31.

56 Caporaso JG, Kuczynski J, Stombaugh J, *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; **7**: 335–6.

57 Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011; **27**: 2194–200.

58 Rognes T, Mahé F, Flouri T, Quince C, Nichols B. VSEARCH. 2014. https://github.com/torognes/vsearch.

59 Westcott SL, Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 2015; **3**: e1487.

60 Quast C, Pruesse E, Yilmaz P, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2012; **41**: D590–6.

61 Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 2010; **26**: 266–7.

62 Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010; **5**: e9490.

63 Good IJ. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* 1953; **40**: 237–64.

64 Subramanian S, Huq S, Yatsunenko T, *et al.* Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* 2014; **510**: 417.

65 Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 2013; **10**: 1200–2.

66 Anderson MJ, Crist TO, Chase JM, *et al.* Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol Lett* 2011; **14**: 19–28.

67 Team RC. R: A Language and Environment for Statistical Computing. 2015.

68 Wickham HEG for DA. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag New York, 2009 DOI:10.1007/978-0-387-98141-3.

69 Legendre P, Anderson MJ, Egendre PIL. Distance-Based Redundancy Analysis: Testing Multispecies Responses in Multifactorial Ecological Experiments. Ecol. Monogr. 1999; **69**: 1–24.

70 Blanchet FG, Legendre P, Borcard D. Forward selection of explanatory variables. *Ecology* 2008; **89**: 2623–32.

71 Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognit Lett* 2010; **31**: 2225–36.

72 Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002; **2**: 18–22.

73 Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw* 2008; **28**: 1–26.

74 Robin X, Turck N, Hainard A, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; **12**: 77.

75 Vatanen T, Kostic AD, D'Hennezel E, *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* 2016; **165**: 842–53.