Supplementary Materials:

Materials and Methods:

Study design and participants

This study is an extension of the Childhood Asthma Study (CAS) – a prospective community-based birth cohort study of 234 infants at high risk of atopy (at least 1 parent with a doctor diagnosed history of hay fever, asthma or eczema) as previously described (*9, 22*). Parents of the subjects completed questionnaires prior to birth, and when each child reached 1, 2, 3, 4, 5 and 10 years of age. Parents were also asked to complete a daily diary for the first five years, recording information on breastfeeding, respiratory symptoms (fever, runny nose, cough, wheeze), and use of any medications by both mother and child. Healthy nasopharyngeal aspirates (NPAs) were collected from subjects by study clinicians during planned visits at approximately 2 months, 6 months and 12 months of age, after the child had been free from any symptoms of respiratory illness for a period of at least 4 weeks. Parents were also asked to report to the study clinicians whenever the child showed symptoms of an acute respiratory illness (ARI), at which point the family was visited within 48 hours by a study nurse who recorded clinical details of the infection and collected an NPA from the child. Clinical data recorded included the presence of fever, wheeze or rattly chest and any medications taken (including antibiotics). Each ARI was classified as either a lower respiratory illness (LRI; if wheeze or a rattly chest was present), or an upper respiratory illness (URI; otherwise). The material in each NPA was divided into four aliquots and stored at -80°C. Blood was collected from each child at age 6 months and 1, 2, 3, 4, 5 and 10 years.

Identification of viruses

One aliquot of each NPA (healthy and ARI) was screened for common causative agents via reverse transcriptase polymerase chain reactions (PCR) as previously reported (*23*). Target organisms were: human rhinoviruses (HRV); other picornaviruses (coxsackie, echo and enteroviruses); coronaviruses 229E and OC43; respiratory syncytial virus (RSV); influenza A and B; parainfluenzaviruses 1-3; adenoviruses and human metapneumovirus (HMPV). Viral profile data was successfully generated for 451 healthy, 327 LRI and 649 URI NPAs. Following the recent identification of HRV-C (*44*), and expansion in availability of HRV sequence information, a second aliquot of each LRI NPA was screened for picornaviruses using seminested PCR (*49*), which incorporated a wider diversity of HRV primers than the initial screen, and then partial sequencing was used to classify HRV into A, B and C species and type.

DNA extraction and bacterial 16S rRNA amplicon sequencing

One aliquot each of 1,021 NPAs was used for 16S rRNA microbiome profiling. These include 487/561 healthy NPAs collected from visits at 2 months, 6 months and 12 months of age; 380/381 LRI reported during the same period; and 154/782 URI (random selection of 0-2 URI per infant). Overall, 397 healthy NPAs, 326 LRIs and 101 URIs were profiled for both virus and bacteria.

Total DNA was extracted using a method combining homogenization and chemical lysis of cells. Extractions were performed in biosafety cabinets that were UV-sterilised, including all plastic-ware, for 30 min prior to the procedure. The NPA were thawed from -80°C storage, transferred into 1.5 mL sterile screw-capped tubes and

briefly micro-centrifuged. The saline storage buffer was removed and pellets were resuspended in 400 µL of lysis solution supplied with the Wizard SV Genomic DNA System (Promega, Victoria, Australia). Samples were mixed vigorously by pipetting and then transferred into a labelled Lysing Matrix B tube (MP Biomedicals, New South Wales, Australia). Suspensions were homogenized using a FastPrep-24 homogenizer for 40 s at 6.5 m/s. Following micro-centrifugation, homogenates were transferred into a 1.5 mL screw-capped tube. A further 200 µL of lysis solution was added into each lysing matrix tube and vortexed to wash off any residual homogenate, then transferred to the respective homogenate tube to retain the original lysis volume. Homogenates were then treated with nuclei lysis buffer/RNase A and DNA extraction was carried out using the Wizard SV Genomic DNA System as per manufacturer's instructions. Purified DNA was eluted in 100 µL of pre-warmed sterile low 1 X TE (Fisher Biotec, WA, Australia), aliquoted and stored at -80°C.

Amplicons were prepared for MiSeq sequencing using primers (prepared by Integrated DNA Technologies, Iowa, USA) spanning the V4 region of the 16S rRNA gene and containing barcoded reverse primers as published by Caporaso *et al.* (*50*). The forward universal primer included the 5′ Illumina adapter sequence, forward primer pad, linker and the 515F 16S rRNA sequence: 5′-AATGATACGGCGACCACCGAGATCTACACTATGGTAATTGTGTGCCAGC MGCCGCGGTAA-3′. The reverse primer included the 3′ Illumina adapter sequence, a 12-mer Golay barcode (denoted as N), reverse primer pad, linker and the 806R 16S rRNA sequence: 5′-CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNNNAGTCAGTCAGCCG GACTACHVGGGTWTCTAAT-3′. All laboratory equipment used was wiped with DNA Away (MBP, Mexico) before conducting each PCR procedure. Master mixes were prepared in a UV-treated PCR chamber before they were dispensed into 96-well plates using a Multiprobe II liquid handling system (Perkin Elmer, Victoria, Australia), followed by addition of samples and controls using the robot. Amplification of each sample was performed in quadruplicate to obtain enough amplicon for sequencing. To each plate, a positive control (gDNA from *S. enterica* strain LT2, a bacterium not normally associated with the respiratory system (ATCC#700720D-5, USA)) and water and TE negative controls (obtained from each extraction procedure) were included and assessed for amplification by agarose gel electrophoresis. Due to the high throughput nature of this study, we did not quantify and normalize sample DNA that was added into each PCR reaction, rather a fixed volume (4 µL) of DNA template was used per well. Amplification was conducted on a GeneAmp 9700 PCR System (Perkin Elmer) using the following conditions: an initial 94°C denaturation step for 2 min, followed by 30 cycles of 94°C denaturation for 30 s, 58°C annealing for 30 s and 72°C extension for 1 min.

Quadruplicate sample amplicons were combined into a single well on the PCR reaction plate, then transferred to a fresh round-bottom polystyrene plate where they were purified using Agencourt AMPure XP beads as directed by the manufacturer, with slight modifications (Beckman Coulter, USA). Purified amplicons were eluted in 25 µL sterile low 1 X TE buffer (Fisher Biotec). Quantitation of amplicon was performed using the Quant-iT PicoGreen dsDNA quantitation kit (Life Technologies, Victoria, Australia) and fluorescence was determined on a Wallac Victor^3 Multilabel counter (Perkin Elmer). PCR samples were equalized to 2 nM concentration (a neat

aliquot was used where a sample fell below this concentration) and pools of 48, 60 or 96 barcoded samples were generated and sent for sequencing.

Primer adaptors were removed from library pools using a 0.8x ratio of Agencourt AMPure XP beads (Beckman Coulter, USA). Library quantitation was determined by the high sensitivity Qubit kit (Life Technologies, USA) whilst library quality and average size distribution was assessed by the Bioanalyser (Agilent Technologies, USA) high sensitivity kit. Library pools were diluted to 2nM followed by NaOH denaturation as per manufacter's instructions (Illumina Inc., USA). Sequencing primers read 1: 5'- TATGGTAATTGTGTGCCAGCMGCCGCGGTAA -3', read 2: 5'-AGTCAGTCAGCCGGACTACHVGGGTWTCTAAT-3' and index: 5'-ATTAGAWACCCBDGTAGTCCGGCTGACTGACT-3' (Sigma, Australia) were spiked into the MiSeq Cartridge at a final concentration of 0.5μM. Denatured libraries were loaded at 6.5pM with a 5% PhiX spike for diversity and sequencing control, onto a v2 300 cycle cartridge for sequencing on the Illumina MiSeq. All sequencing runs yielded approximately 1100k clusters/mm^2 with an average of 80% clusters passing filter and an average 19M reads passing filter.

Sequence analysis, taxonomic assignment and phylogenetic analyses

Paired end reads were merged using Flash version 1.2.7 (*51*) with read length 151 base pairs (bp) and expected fragment length 253 bp. The merged sequences were quality filtered as follows: ≤3 low-quality bp (Phred quality score < 3) allowed before trimming, ≥189 consecutive high-quality bp with no uncalled bases (Ns) (*52*). A total of 33 million sequences were filtered out (13%), leaving 219 million for analysis. Quality filtered sequences were assigned to operational taxonomic units (OTUs) using the closed reference method in QIIME v1.8 (*53*) with the Greengenes 99% OTU reference set, version 13_5 (*54*). This reference set consists of >200,000 representative sequences obtained from clustering all sequences from the Greengenes reference database at 99% sequence similarity. Briefly, the closed reference OTU picking uses UCLUST (*55*) to search each sequence against the reference set, and assigns the sequence to an OTU based on the best hit at ≥99% sequence identity. Sequences that did not match the reference database (3%) were excluded from the analysis; these sequences could be chimeras, sequencing errors or novel sequences that are not well characterized in the database. This left an average >200,000 (interquartile range: 108,000 – 255,000; 8 samples with <1000 reads) taxonomy-assigned sequences for each PNA. Figures S1-6 show subtrees for each of the 6 common genera, extracted from the full Greengenes 99% OTU reference tree (v 13_05) using Dendroscope v2 (*56*).

Clustering into microbiome profile groups (MPGs)

For each NPA, the relative abundance of each OTU was calculated (i.e. reads matching the out, divided by total taxonomy-assigned reads for that sample). Most analyses were summarised at genus level, whereby all OTUs assigned to the same genus were collapsed into a single group for reporting. Samples were assigned to microbiome profile groups (MPGs) based on hierarchical clustering of the relative abundances of the six most common genera (distance metric: 1-Pearson's correlation; clustering method: Ward's minimum variance, implemented in *R hclust*). Previous studies of human microbiomes in various habitats have reported clinically or environmentally meaningful associations with community profiles or types (e.g. enterotypes) identified using similar clustering approaches (*57, 58*).

Statistical analyses

Classification of NPAs. Unless otherwise stated, the following definitions were used:

- Infection: Collected during a physician verified episode of symptoms of acute respiratory illness (ARI).
- Healthy: Collected in the absence of such symptoms for at least 4 weeks.
- LRI: Lower respiratory illness, defined as ARI with wheeze or rattly chest.
- URI: Upper respiratory illness, defined as ARI with no wheeze or rattly chest.
- 2-month samples: Age at collection 6-100 days (range 6-100 days for infections, 34-97 days for healthy).
- 6-month samples: Age at collection 150-210 days (interquartile range 165-196 days for infections, 179-190 days for healthy).
- 1-year samples: Age at collection 330-425 days (interquartile range 357-398 days for infections, 363-382 days for healthy).

Variable definitions. Unless otherwise stated, the following variable definitions were used in the statistical analyses:

- Gender (binary): Female (coded 0), Male (coded 1).
- Season (binary): Spring-summer (September to February, coded 1); autumn-winter (March to August, coded 0).
- Antibiotics intake (binary): Coded 1 if the child had taken antibiotics within the last 4 weeks before sample collection; 0 otherwise. Antibiotics data was curated from two sources of information; a daily diary kept by the mother, which reports any medication taken per day, and a questionnaire taken during each infection sampling. Positive antibiotics intake from any one source was sufficient. We considered only oral/intravenous antibiotics and excluded Flagyl (Metronidazole) and any topical antibiotics.
- Mother's antibiotics intake (binary): Coded 1 if the mother had taken antibiotics within the last 4 weeks before sample collection and she was breastfeeding on the day of antibiotics intake; 0 otherwise. Data was extracted from the daily diary, using the same group of antibiotics as for subjects.
- Breastfeeding (binary): Coded 1 if the child was breastfed on the day of sample collection; 0 otherwise.
- Siblings (binary): Coded 1 if any children less than 16 years of age were residing in the same house; 0 otherwise. Data was extracted from the 1-year questionnaire.
- Furry pets (binary): Coded 1 if household had any cat, dog, rabbit or guinea pig for the whole first year; 0 otherwise. Data was extracted from the 1-year questionnaire.
- Day-care (binary): Coded 1 if child has started attending day-care by sample collection; 0 otherwise.
- Delivery mode (binary): Coded 1 caesarean section delivery; 0 otherwise.
- Maternal/paternal history of atopic disease: Coded 1 if the mother/father had doctor diagnosed allergic hayfever, eczema or asthma.
- Atopy by 2 years (binary): Atopic status was assessed using serum IgE levels measured at 6 months, 1 year and 2 years. Positive atopy status was defined as any specific IgE to house dust mite, cat epithelium and dander, peanut, foodmix, couch grass, rye grass, mould mix, or infant phadiatop > 0.35 kU/L (*9*). Atopy by two years was defined as IgE above this cut-off at any of the three time points (6 months,

1 year, 2 years). Children atopic by 2 years are referred to as atopics; non-atopics otherwise.

- Chronic wheeze (binary): Presence of wheeze in the last 12 months, assessed during face-to-face interviews with the parents of subjects at 5 and 10 years of age.

Association of MPGs with healthy vs infection status. Odds ratios for each MPG with infection status (NPA taken during reported episode of respiratory illness, vs. NPA taken in the absence of such symptoms for at least 4 weeks prior) were estimated using generalized estimating equations (GEE) logistic regression with unstructured correlation and robust standard errors, to take into account samples from the same subject. The following variables were adjusted for by inclusion in the model: age at sample collection (days), gender, season, number of prior infections, antibiotics intake, mother's antibiotics intake, delivery mode and breastfeeding. ORs further adjusted for detection of common viruses (RSV, HRV) by PCR were also reported (**Fig. 1E**).

Association between microbes and classes of ARI. Amongst ARI samples, odds ratios for MPGs and markers of infection severity (LRI vs. URI, febrile vs. non-febrile LRI, and wheezy vs. non-wheezy LRI) were estimated using generalized estimating equations (GEE) logistic regression with unstructured correlation and robust standard errors. Models were adjusted for presence of the most common viruses (RSV and HRV), age, gender, and season. Samples with antibiotics intake within the last week were excluded from this analysis (**Table 1**). Febrile vs. non-febrile LRI comparison was also done separately for RSV-positive and RSV-negative subsets (**Fig. S13**).

Association of MPGs with otitis media symptoms. For each MPG, odds ratios for otitis media symptoms within (a) ARIs, and (b) LRIs were estimated using GEE logistic regression, adjusting for age at sample collection (**Table S2**).

Stability of MPGs within individuals. NP microbiome transitions were assessed by analysing the MPG assignments of consecutive samples within individual subjects (time point T_1 to next time point T_2). Transition into the same MPG (T_1 MPG = T_2 MPG) was considered a stable transition. Healthy and infection samples were analysed separately. The expected frequency of a stable transition for a given MPG, *i*, was calculated as the squared proportion of samples in that MPG at T_1, $(p_i(T_1))^2$. One thousand bootstrapped estimates of observed stability were generated by sampling with replacement from the observed transitions and calculating, for each bootstrapped sample, the proportion of stable transitions for each MPG. The estimated 95% confidence intervals for the observed frequency of respective stable transitions were taken to be the 2.5^{th} and 97.5^{th} quantiles of the 1,000 bootstrapped statistics. The MPG was considered significantly more stable than expected if the expected frequency of stable transitions $(p_i(T_1))^2$ fell below the lower limits of the 95% CI of observed transitions (and less stable if the expected value fell above the upper limits) (**Fig. S8**).

Association of infection and wheeze phenotypes with Alloiococcus/Moraxella stable colonization groups. Children were grouped by their colonization patterns: '*Alloiococcus* colonized', ≥1 healthy sample with *Alloiococcus* MPG and none with *Moraxella* MPG; '*Moraxella* colonized', ≥1 healthy sample with *Moraxella* MPG and

none with *Alloiococcus* MPG; 'Others', no *Moraxella* or *Alloiococcus* healthy sample or ≥1 of each. Association of the groups with number of infections was assessed using Wilcoxon rank sum test; other associations were assessed using Fisher's exact test (**Table S4**).

Association of genus abundance with environmental factors. We investigated the relationships between relative abundance of the six main genera and various environmental factors: attendance at day-care, living with siblings, antibiotics intake within the last 4 weeks, breastfeeding, number of prior infections (0, 1, ≥2; based on recorded ARI) and gender. As relative abundance values were highly skewed, they were log-transformed (base 10) prior to analysis. Odds ratios (ORs) for binary variables were estimated using logistic regression; odds ratios for association with prior infections were assessed using proportional odds ordinal logistic regression. Healthy and infection samples were analysed separately. Association with day-care was assessed for 12-month samples only, as few infants had started day-care by the time of their 6-month NPA; ORs for siblings, furry pets, antibiotics, season, and gender were adjusted for age at collection by inclusion in the model. Association with antibiotics was assessed for healthy samples only, as prescription due to infection severity could confound association with infection samples. Prior infections were assessed separately for each of the three age strata for healthy NPA sampling (2, 6 and 12 months) (**Fig. 2, Fig. S9**).

Association between early Streptococcus colonization and subsequent chronic wheeze. Odds ratios were calculated separately for atopics, non-atopics and all children, using logistic regression. Gender, age at sample collection, maternal history of atopic disease and paternal history of atopic disease were adjusted for by inclusion in the model. Samples with any prior infection or antibiotics intake within the last week were excluded from this analysis (**Table 2, Fig. 4A**).

Association between early colonization and time of first ARI. This analysis was restricted to children who contributed an asymptomatic NP sample between 5-9 weeks of age and prior to their first reported ARI (n=160). Children were grouped according to the MPG of the first pre-ARI healthy sample. Kaplan-Meier curves were plotted separately for each group, showing time of first infection (age in days) as recorded in the daily symptom diaries for (a) first ARI, (b) first URI and (c) first LRI. Cox proportional hazards models were fit to assess the significance of apparent differences in time of infection, adjusting for age, gender, season, virus status in the early healthy sample, and virus status in the subsequent infection (**Fig. S14**).

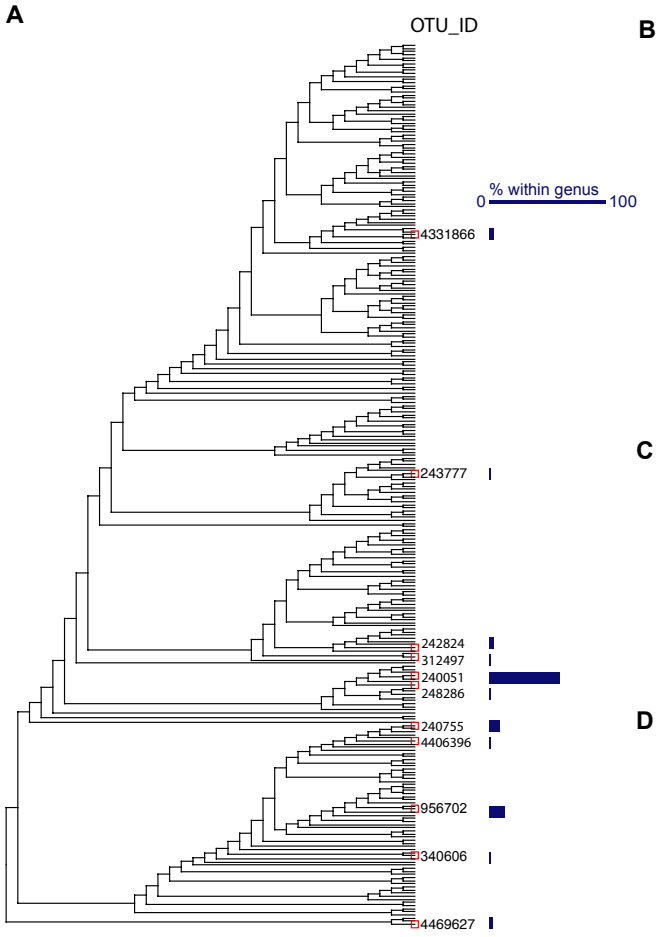
Association between LRI subtypes and subsequent chronic wheeze. Odds ratios were calculated separately for atopics, non-atopics and all children, using logistic regression. Gender, maternal history of atopic disease and paternal history of atopic disease were adjusted for by including them as covariates in the model (**Table 2, Table S3**).

Association between time of first LRI subtype and subsequent chronic wheeze. Children were grouped according to their chronic wheeze status at 5 years. Kaplan-Meier curves were plotted separately for each group, showing time of first infection (age in days) for (a) first fever LRI, (b) first wheezy LRI with detection of HRV-C. Cox proportional hazards models were fit to assess the significance of apparent

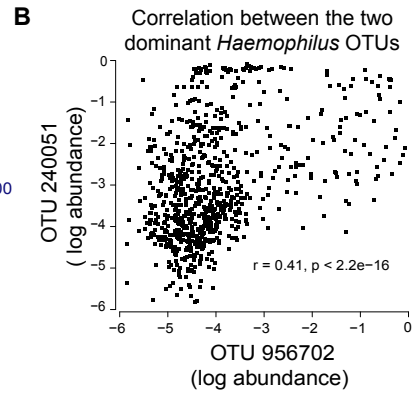
differences in time of infection, adjusting for gender, maternal history of atopic disease and paternal history of atopic disease (**Fig. 3A-B**).

Association between antibodies to species-specific surface proteins and number of NP samples containing specific OTUs. Logistic regression was used to estimate odds ratios for association of detectable IgG1/IgG4 antibodies (assessed from blood during collection at 12 months) to *H. influenza* P4/P6 surface proteins with (a) number of infection NP samples and (b) number of healthy NP samples with presence of the two most abundant *Haemophilus* OTUs (**Fig. S1**). The same was done for *S. pneumoniae* A1, A2 or C surface proteins and the two most abundant *Streptococcus* OTUs (**Fig. S3**).

Fig. S1. *Haemophilus* OTUs.
(**A**) Subtree of the Greengenes 99% sequence similarity OTU tree for genus *Haemophilus.1*. OTUs present at >0.1% frequency across our 1,021 NP samples are highlighted, and relative frequencies within these samples are shown with blue bars (right). (**B**) Correlation between the two most frequent *Haemophilus* OTUs. (**C-D**) Association between detectable IgG1/IgG4 antibodies to *H. influenzae* P4/P6 surface proteins (measured at 12 months of age) and number of (**C**) prior ARI NP samples or (**D**) prior asymptomatic NP samples containing the dominant *Haemophilus* OTUs.



*Greengenes tree contains two *Haemophilus* genera, *Haemophilus.1* and *Haemophilus.2*; none of our sequences match to *Haemophilus.2*.



C Association between detectable *H. influenzae* antibodies at 12 months and number of prior infections with *Haemophilus* OTUs.

Presence of OTU	OR (95% CI)
240051	2.0 (1.2 - 3.4)**
956702	3.3 (1.3 - 9.4)**

** p < 0.05

D Association between detectable *H. influenzae* antibodies at 12 months and number of prior asymptomatic NP samples with *Haemophilus* OTUs.

Presence of OTU	OR (95% CI)
240051	1.1 (0.92 - 1.2)
956702	1.2 (0.97 - 1.4)

Fig. S2. *Moraxella* OTUs.
Subtree of the Greengenes 99% sequence similarity OTU tree for genus *Moraxella,* indicating the position of the only *Moraxella* OTU identified across our 1,021 samples.

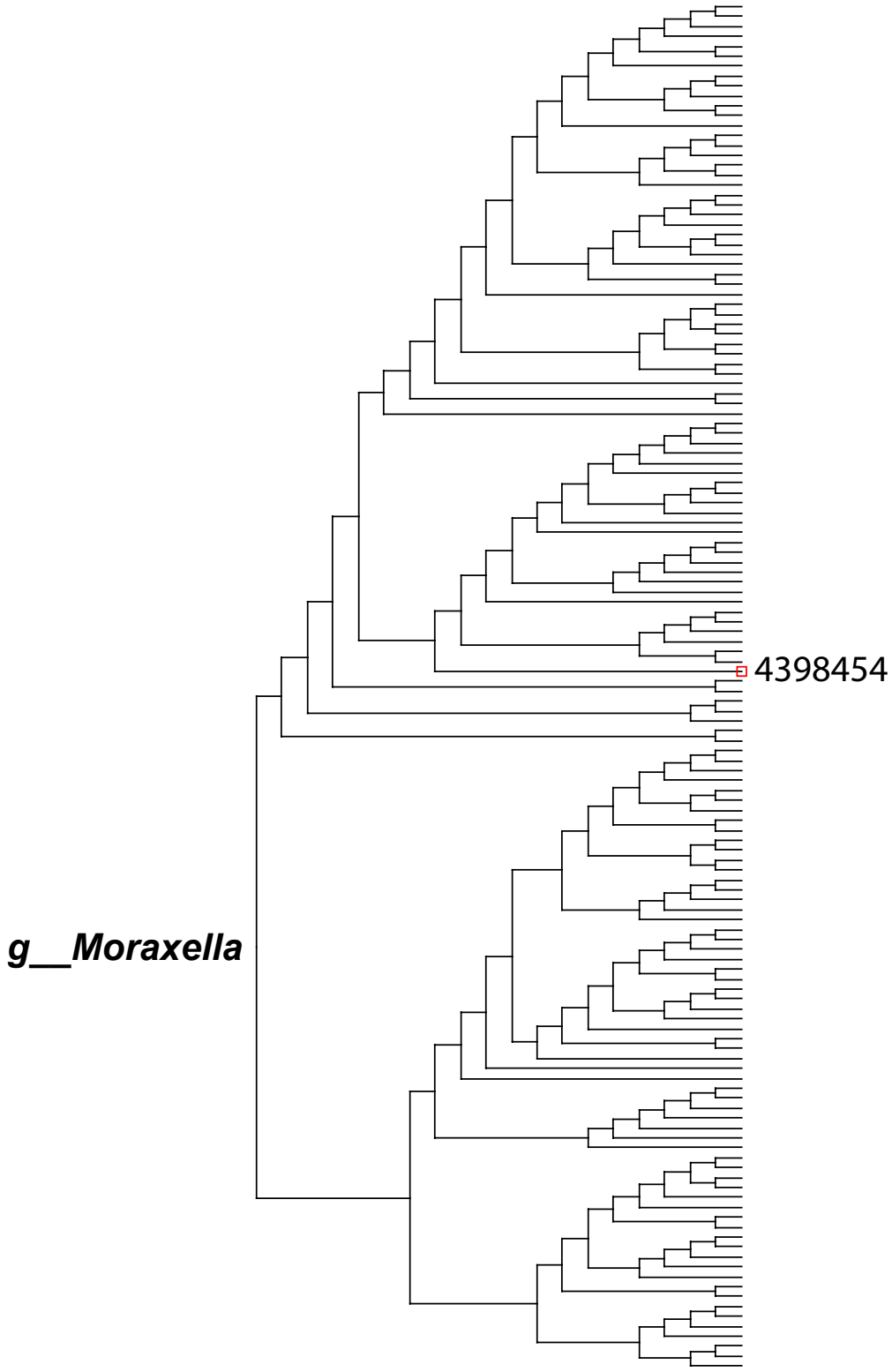
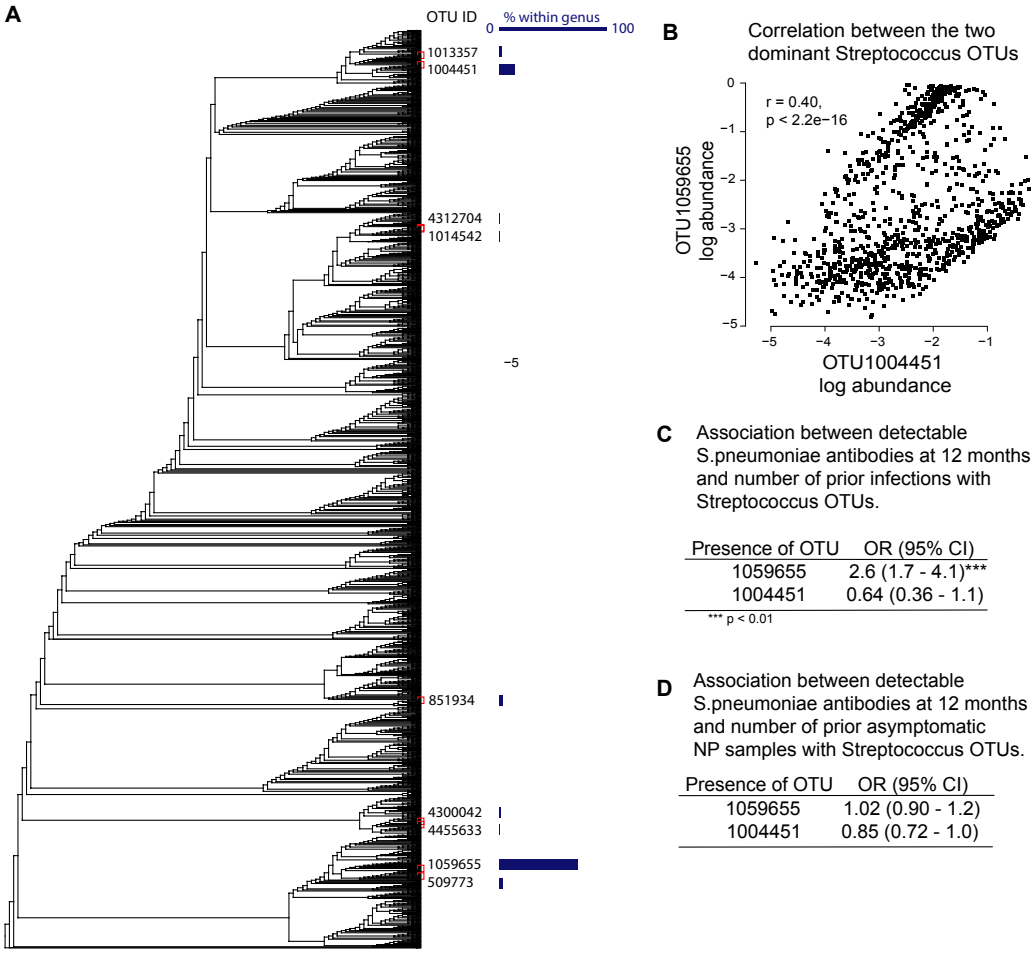


Fig. S3. *Streptococcus* OTUs.

(A) Subtree of the Greengenes 99% sequence similarity OTU tree for genus *Streptococcus*. OTUs present at >0.1% frequency across our 1,021 NP samples are highlighted, and relative frequencies within these samples are shown with blue bars (right). (**B**) Correlation between the two most frequent *Streptococcus* OTUs. (**B-C**) Association between detectable IgG1 antibodies to *S. pneumoniae* pneumococcal surface protein A1, A2, or C (measured at 12 months of age) and number of (C) prior ARI NP samples or (D) prior asymptomatic NP samples containing the dominant *Streptococcus* OTUs.



C Association between detectable S.pneumoniae antibodies at 12 months and number of prior infections with Streptococcus OTUs.

Presence of OTU	OR (95% CI)
1059655	2.6 (1.7 - 4.1)***
1004451	0.64 (0.36 - 1.1)

*** p < 0.01

D Association between detectable S.pneumoniae antibodies at 12 months and number of prior asymptomatic NP samples with Streptococcus OTUs.

Presence of OTU	OR (95% CI)
1059655	1.02 (0.90 - 1.2)
1004451	0.85 (0.72 - 1.0)

Fig. S4. *Alloiococcus* OTUs.
Subtree of the Greengenes 99% sequence similarity OTU tree for genus *Alloiococcus,* indicating the position of the only *Alloiococcus* OTU identified across our 1,021 samples.

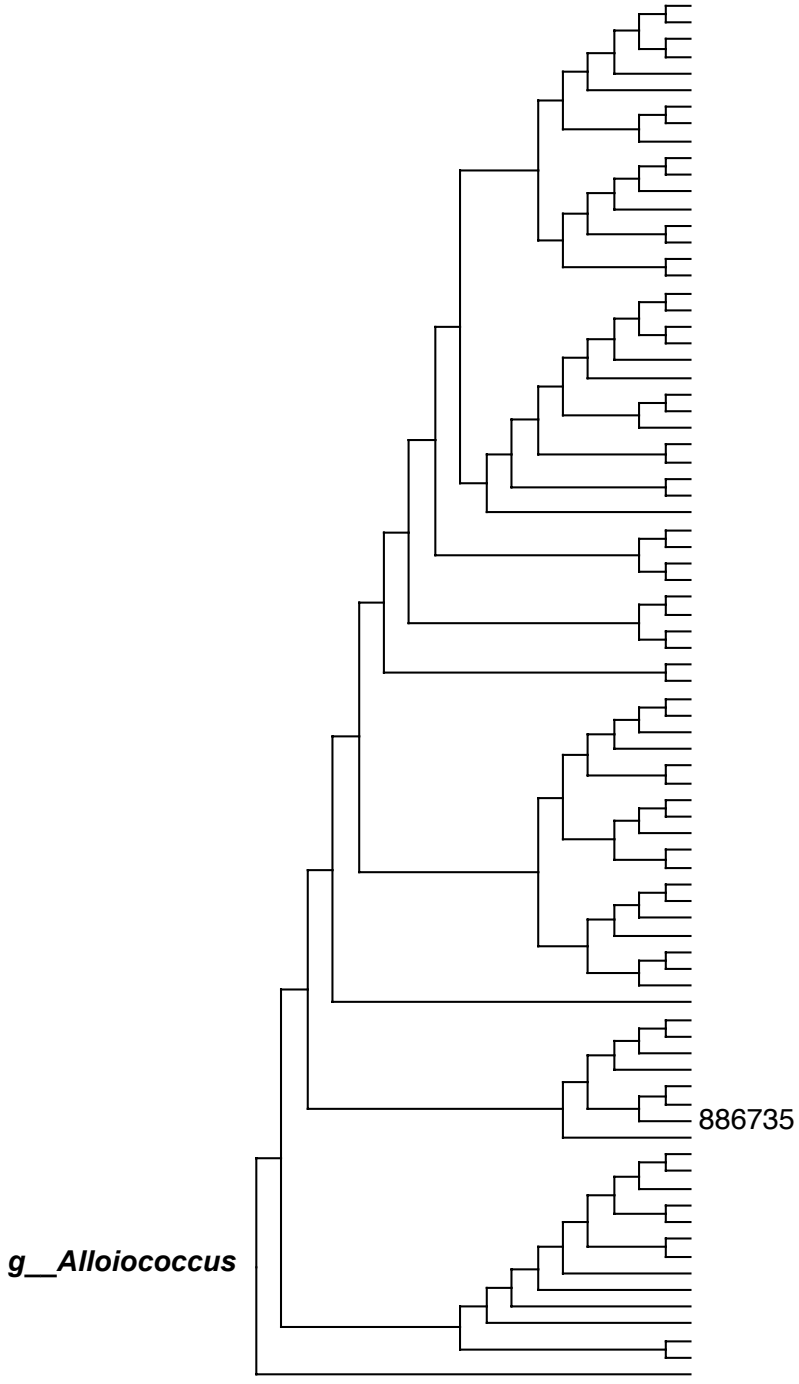


Fig. S5. *Corynebacterium* OTUs.
Subtree of the Greengenes 99% sequence similarity OTU tree for genus *Corynebacterium*. OTUs present at >0.1% frequency across our 1,021 samples are highlighted, and relative frequencies within these samples are shown with blue bars (right).

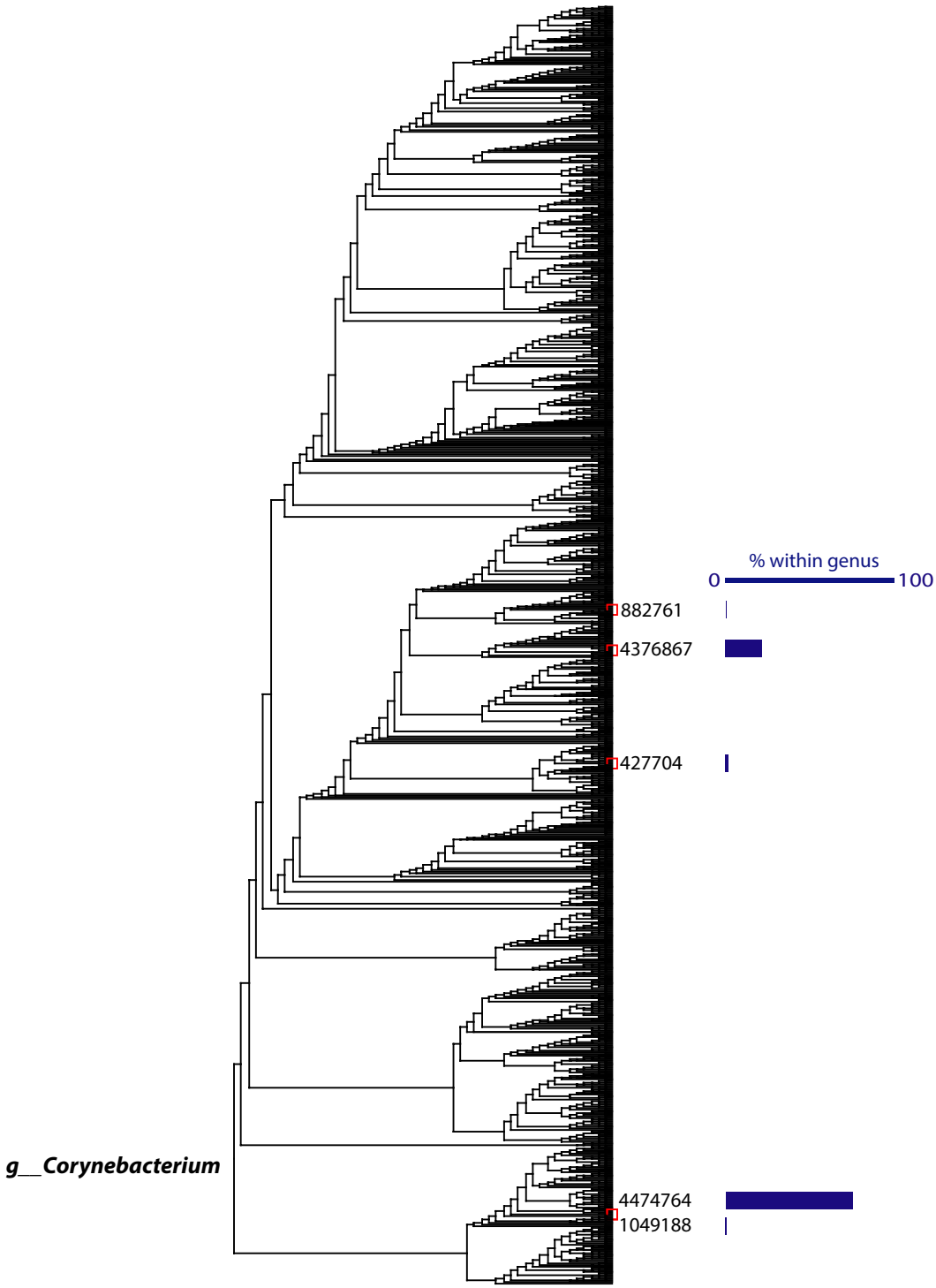


Fig. S6. *Staphylococcus* OTUs.
Subtree of the Greengenes 99% sequence similarity OTU tree for genus *Staphylococcus*. OTUs present at >0.1% frequency across our 1,021 samples are highlighted, and relative frequencies within these samples are shown with blue bars (right).

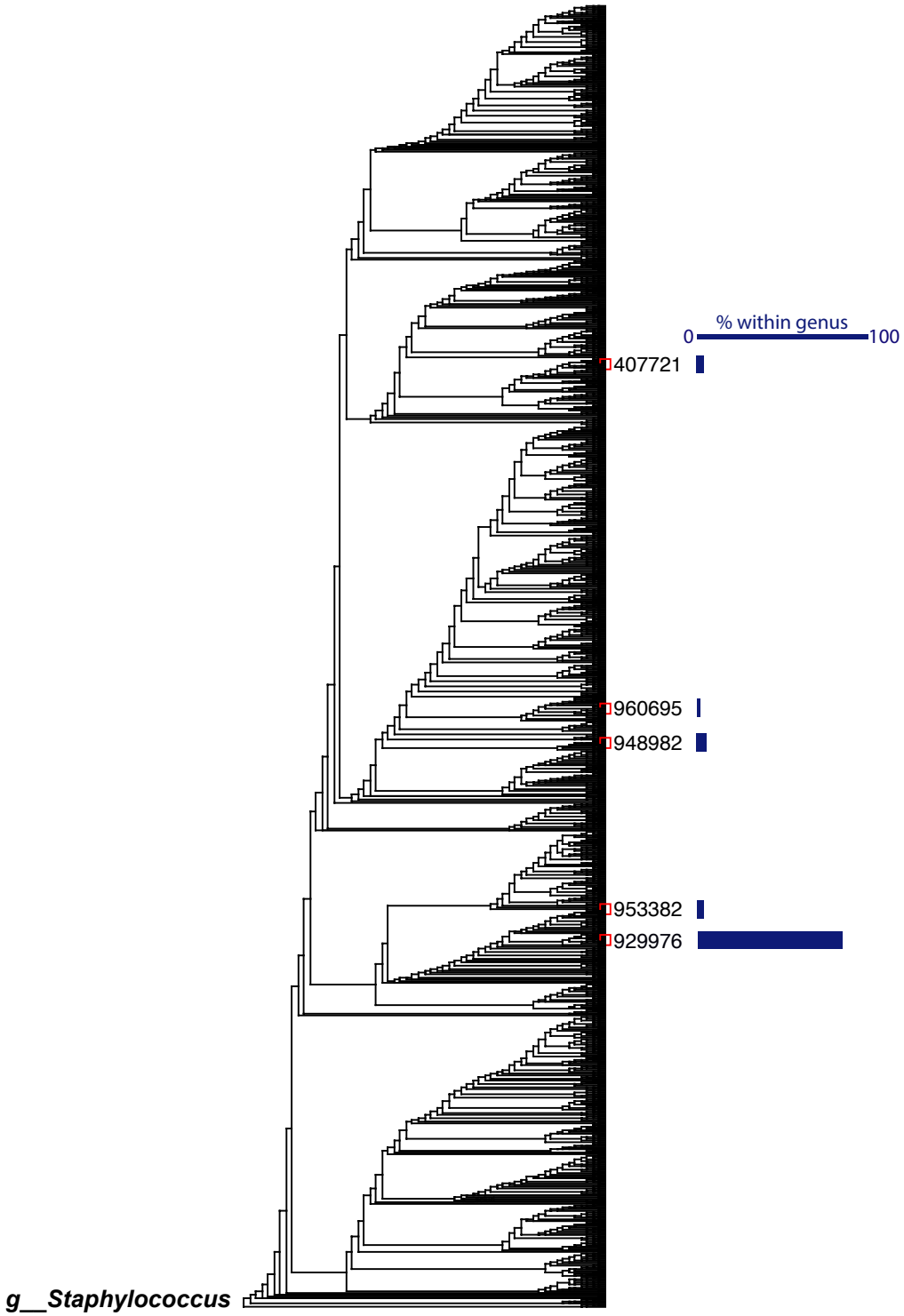


Fig. S7. Heatmap of OTUs with total abundance >0.1% across 1,021 NP samples. Samples are in columns, ordered by microbiome profile group (MPG) based on clustering of genus-level abundance for the six most common genera (**Fig. 2**).

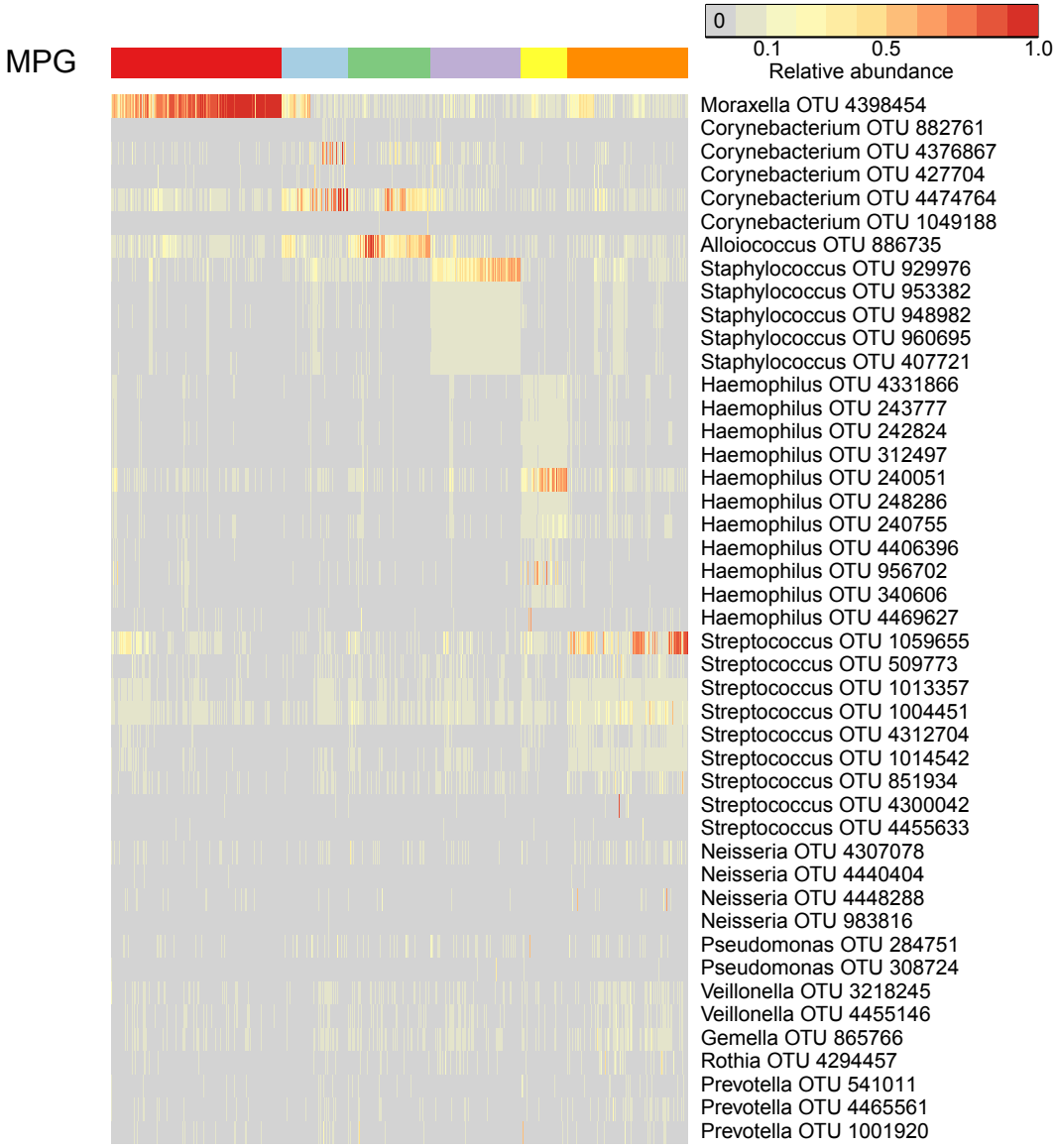
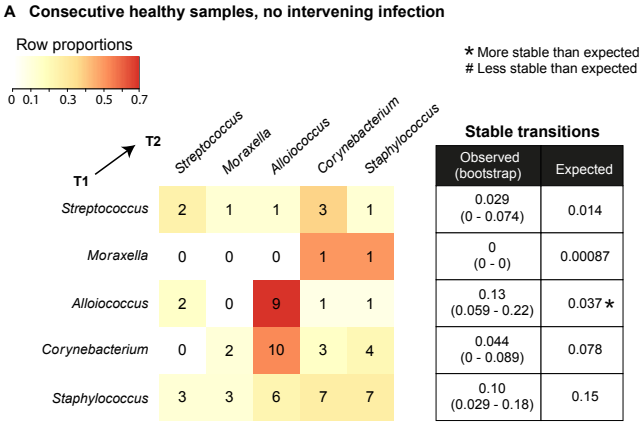
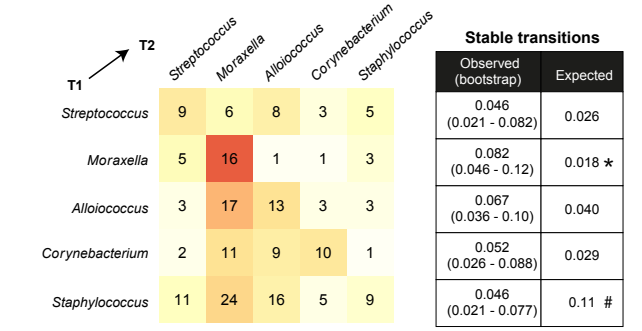


Fig. S8. Transitions between (A) consecutive healthy samples, with no intervening infections, (B) with intervening infections, and (C) between consecutive infection samples. T1, 1st timepoint; T2, next timepoint. Cell numbers indicate the number of cases in which the respective transition from T1 to T2 was observed; cells are coloured to indicate the row proportions as per legend. Stable transitions were defined as those with the same MPG at T1 and T2 (i.e. diagonals in the matrix). For each panel, the table at the right shows observed and expected frequencies of stable transitions for each genus MPG (note *Haemophilus* is not plotted in (A) and (B) because it was extremely rare in healthy samples). The expected frequency for each stable transition was taken to be the square of the proportion of samples in that MPG at T1 (*i.e.,* assuming constant frequencies and random transitions). Observed values are the proportion of stable transitions out of all observed transitions; 95% confidence intervals were calculated from 1,000 bootstraps of the real data, sampled with replacement.





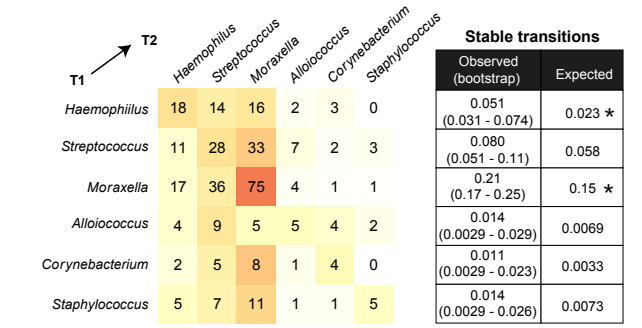


Fig. S9. Associations between log_{10} abundance of major genera and number of prior infections. Prior infections were categorized as 0, 1 or ≥2 ARI recorded prior to sample collection. Odds ratios, 95% confidence intervals and p-values (*, **, and *** indicate p-values ≤0.1, ≤0.05, ≤0.01) were estimated using proportional odds ordinal logistic regression. Healthy and infection samples in each age strata (2-month, 6-month and 1 year) were analysed separately.

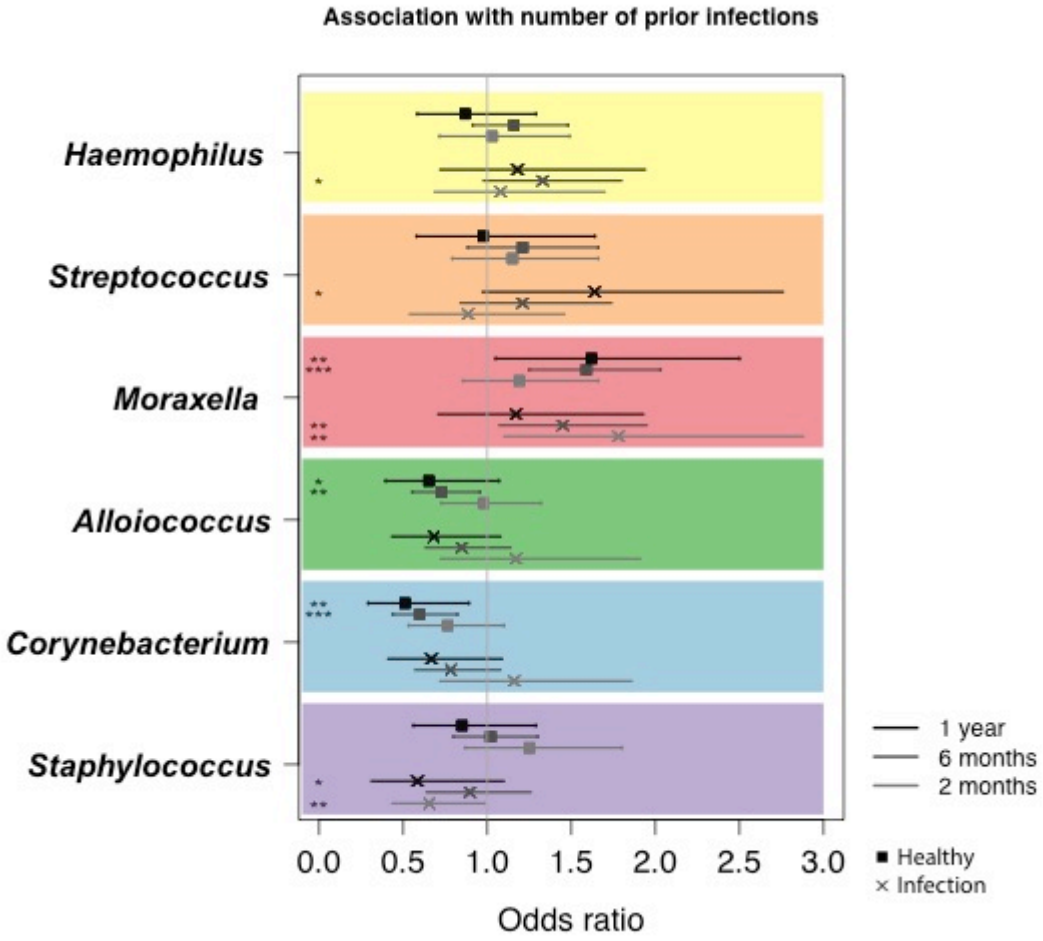


Fig. S10. Seasonal patterns. Top panel, mean maximum and minimum temperatures in the study location of Perth, Western Australia (data from Australian Government Bureau of Meteorology website, http://www.bom.gov.au/climate/averages/tables/cw_009225.shtml).Bottom panels, monthly proportions of samples in each microbiome profile group (MPG) for infection and healthy samples.

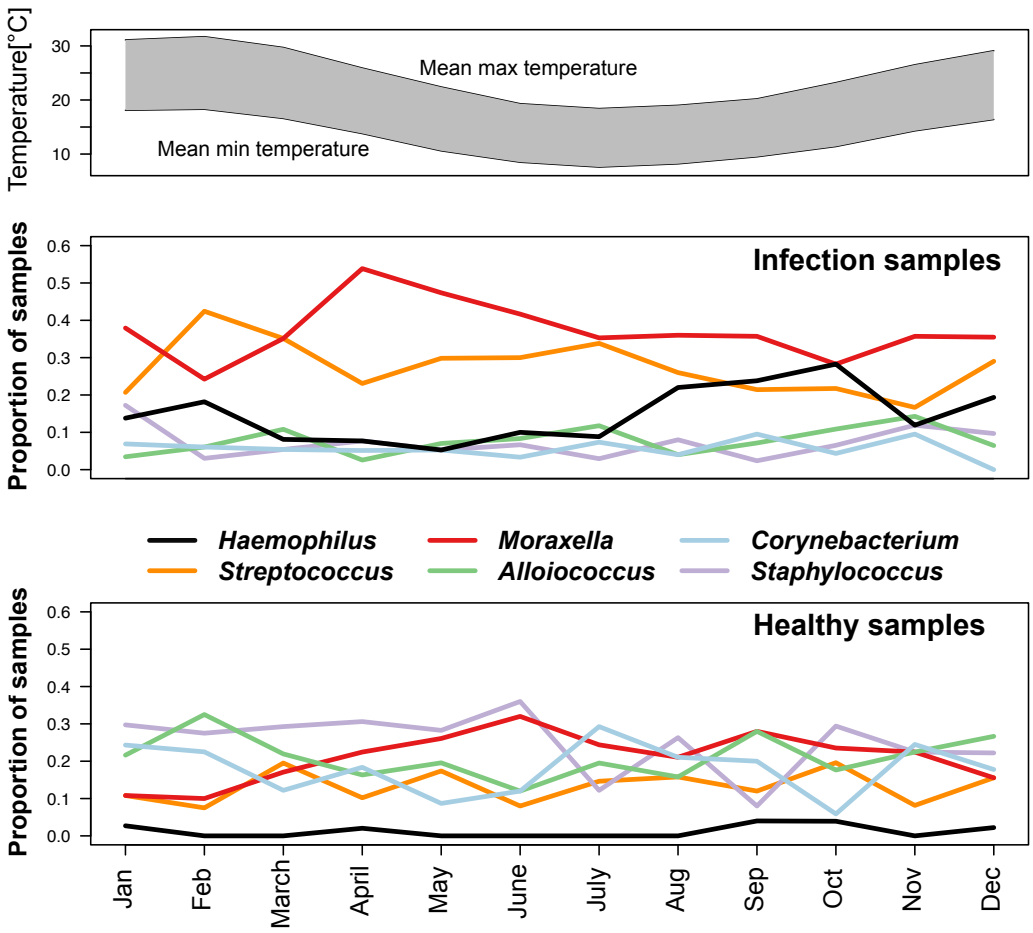


Fig. S11. Frequency of stable transitions between MPGs in consecutive ARIs, stratified by time between ARIs. Stable transitions were defined as pairs of consecutive ARIs from the same individual, that shared the same MPG.

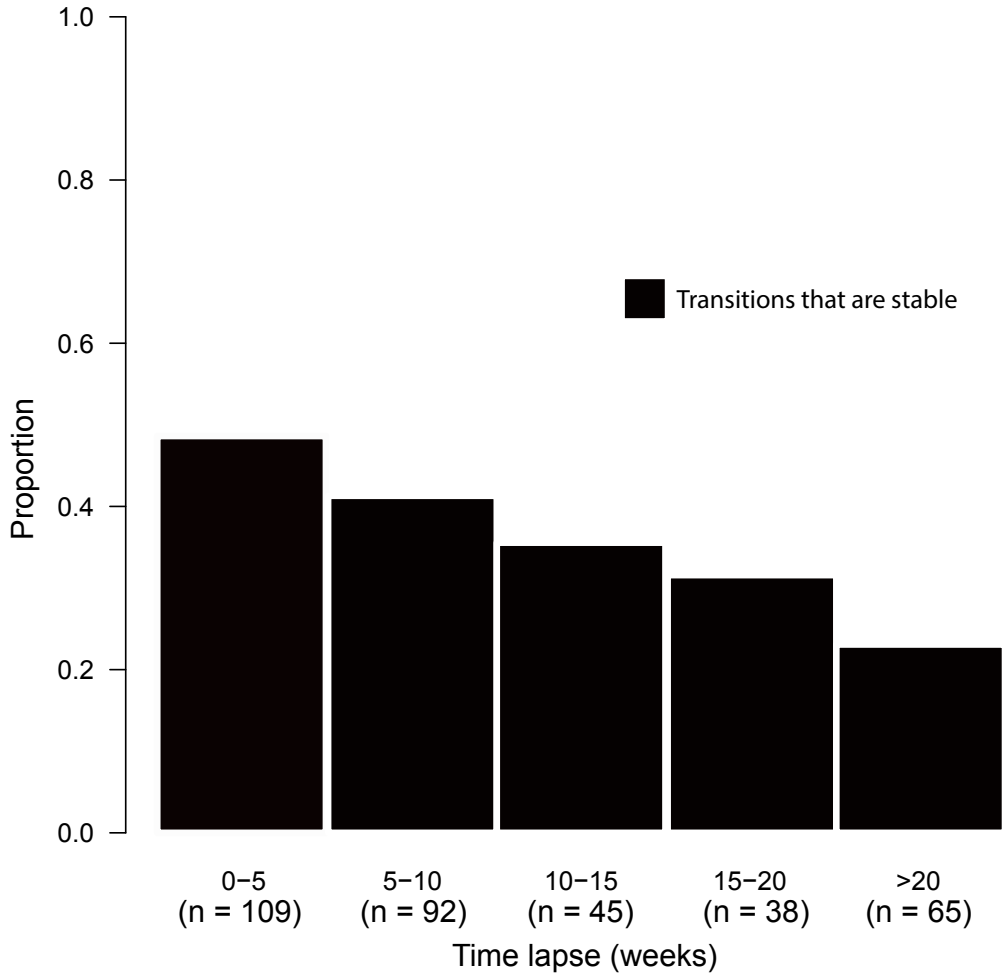


Fig. S12. Rates of bacterial microbiome profile groups (MPGs) and viruses for different types of NP samples. ARI, acute respiratory illness; URI, upper respiratory illness; LRI, lower respiratory illness; fLRI, febrile LRI; wLRI, wheezy LRI. Note HRV subtyping was performed for LRI samples only.

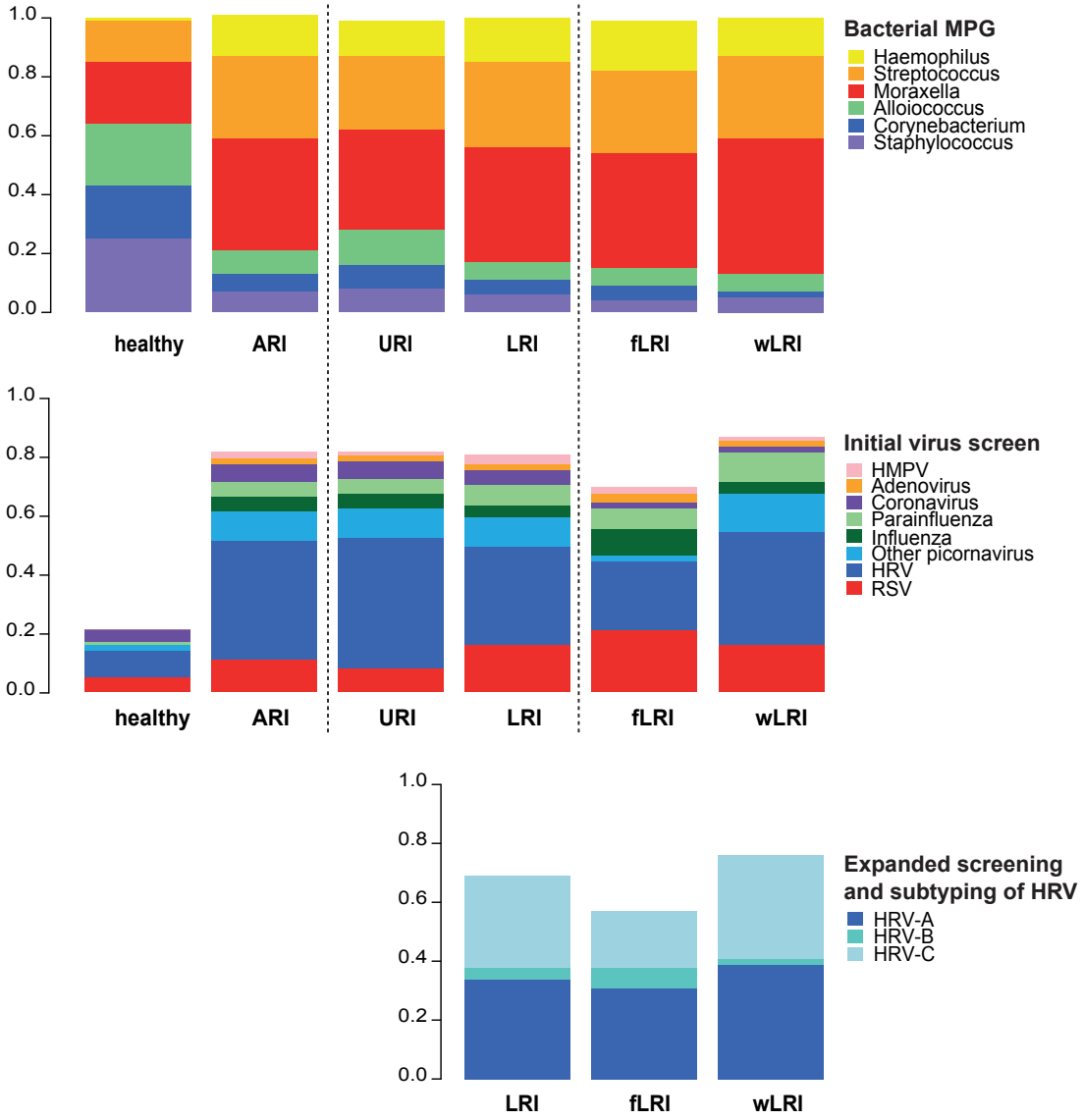


Fig. S13. Distribution of samples by MPG, amongst (A) RSV positive and (B) RSV negative samples. GEE logistic regression with unstructured correlation was used to assess, amongst (A) RSV-positive LRIs, and (B) RSV-negative LRIs, the association of fever status with *Moraxella*, *Streptococcus* and *Haemophilus* MPGs as independent covariates in the same model, adjusted for age, gender and season (p-values shown). Samples with antibiotics intake within the last week were excluded from this analysis.

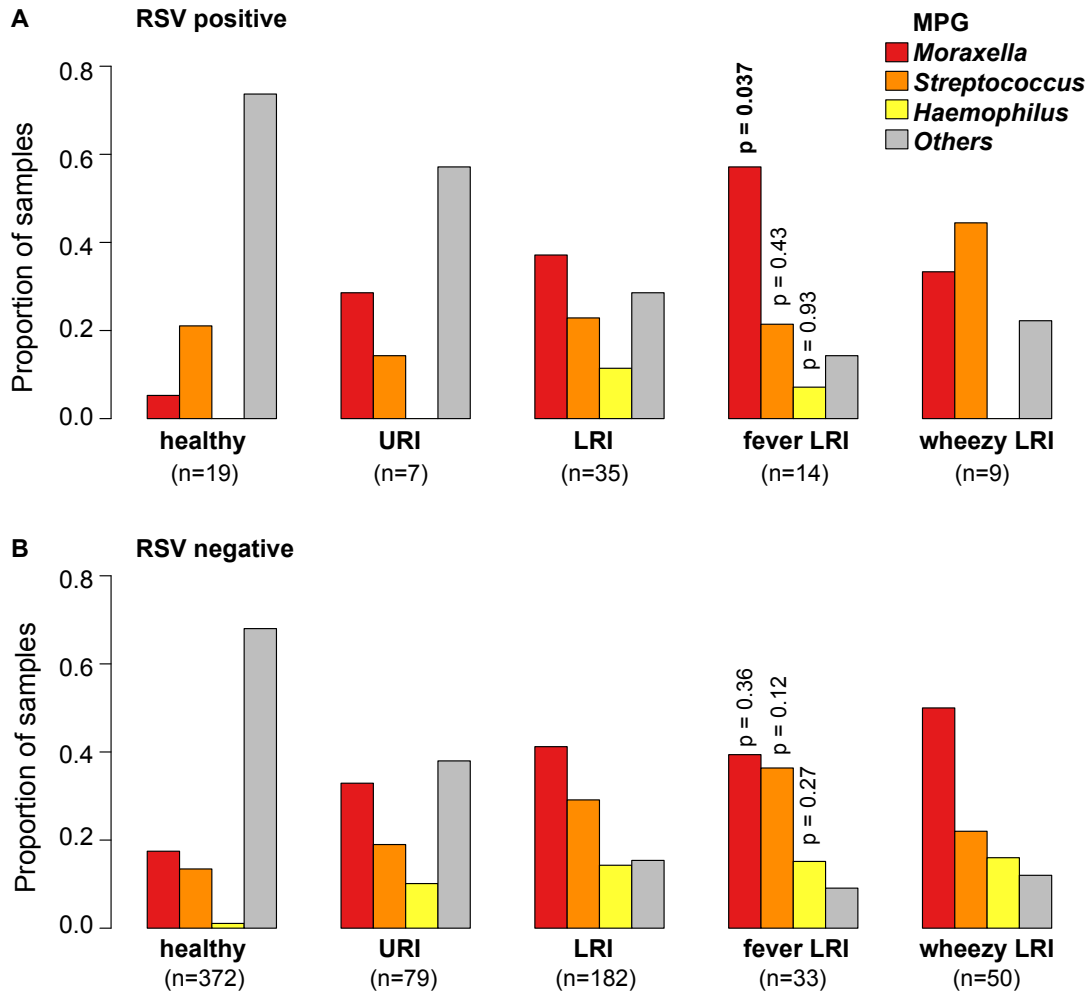


Fig. S14. Impact of early colonization on age at first respiratory infection. (A) MPG transitions between healthy samples (T1) and the next sequenced infection (T2). Cell numbers indicate the number of times the respective transition from T1 to T2 was observed in the data set; cells are coloured to indicate the row proportions as per legend. **(B-D)** Kaplan-Meier survival curves for age (days) of first **(B)** ARI, **(C)** URI and **(D)** LRI, stratified according to the microbiome profile groups (MPGs) of the first healthy sample, collected by 9 weeks of age and prior to any infection (n=160). Cox proportional hazards models were adjusted for age, gender, season, virus status in the early healthy sample, and virus status in the first ARI/URI/LRI.

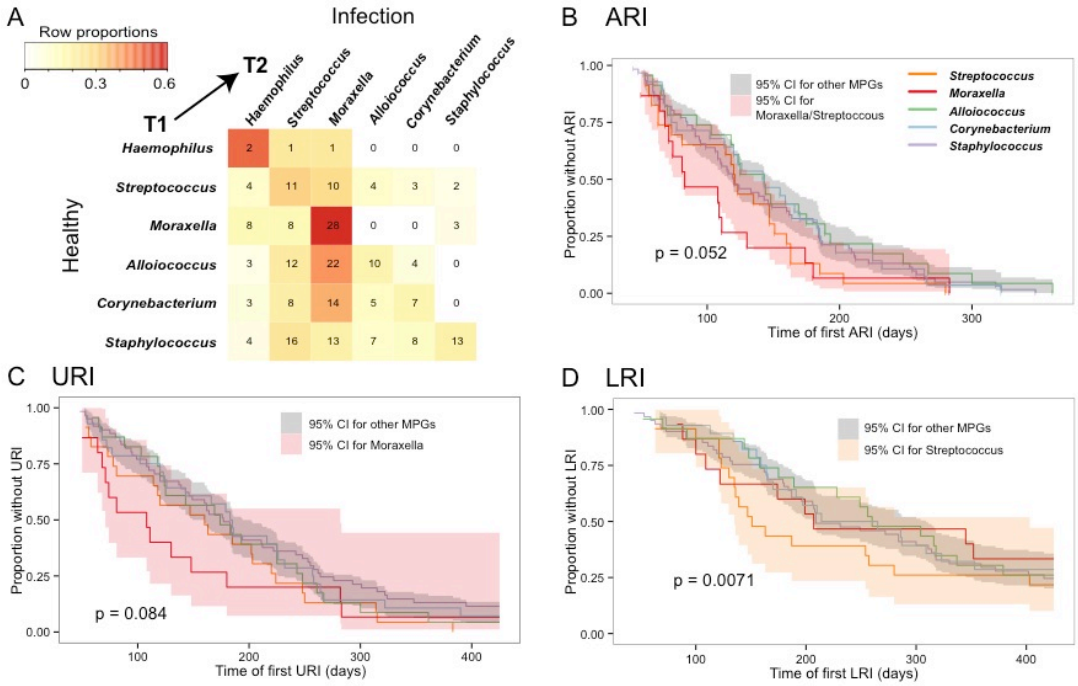
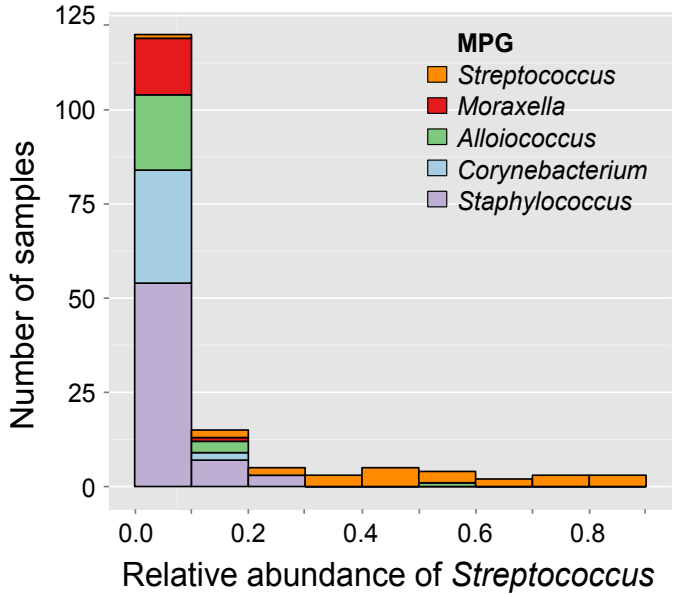


Fig. S15. Distribution of *Streptococcus* relative abundance amongst healthy samples collected by 9 weeks of age, broken down by microbiome profile group (MPG).



Characteristics of subjects	Number	Proportion
Total participants	234	100%
Male gender	132	56%
Paternal history of atopic disease	156	67%
Maternal history of atopic disease	196	84%
Vaginal delivery	167	72%
Breastfed at 3 months	196	84%
Smoking exposure at 1 year	44	19%
Day-care attendance at 1 year	67	29%
Current wheeze at 1 year	76	33%
Other children in house	108	47%
Furry pet in the first year	109	48%
≥1 LRI in first 425 days of life	167	75%
Atopic at 6 months of age	47	21%
Atopic at 2 years of age	88	41%
Sensitised to inhaled allergens[#] by 1 year of age	23	10%
Current wheeze at 5 years of age	56	28%

Table S1. Characteristics of infants in the study. [#] Inhaled allergens include house dust mite, cat epithelium and dander, couch grass, rye grass, or mould mix.

MPG	Number of OM ARI (LRI)	OM vs. non-OM ARI OR (95% CI)	OM vs. non-OM LRI OR (95% CI)
Staphylococcus	3 (3)	2.2 (0.52-8.9) p = 0.29	2.0 (0.51-8.0) p = 0.32
Corynebacterium	1 (0)	0.27 (0.036-2.1) p = 0.21	0.32 (0.041-2.4) p = 0.27
Alloiococcus	2 (2)	0.36 (0.089-1.5) p = 0.16	0.44 (0.1-1.9) p = 0.28
Moraxella	25 (16)	1.1 (0.61-2.0) p = 0.71	1.0 (0.58-1.8) p = 0.92
Streptococcus	21 (17)	1.2 (0.66-2.3) p = 0.50	1.2 (0.64-2.3) p = 0.55
Haemophilus	9 (5)	0.90 (0.33-2.5) p = 0.84	0.83 (0.31-2.2) p = 0.70

Table S2. Distribution of microbiome profile groups (MPGs) among ARI with otitis media (OM) diagnosis. Odds ratios (OR) for association of each MPG with OM, estimated using generalized estimating equations (GEE) logistic regression with unstructured correlation and robust standard errors, adjusted for age at infection. ARI, acute respiratory infection; LRI, lower respiratory illness.

	All children		Atopic by 2 years		Not atopic by 2 years	
	Wheeze at 5 years	**Wheeze at 10 years**	**Wheeze at 5 years**	**Wheeze at 10 years**	**Wheeze at 5 years**	**Wheeze at 10 years**
Any URI	0.81 (0.21-4) p = 0.78	0.71 (0.14-5.3) p = 0.69	0.42 (0.067-2.6) p = 0.33	0.48 (0.068-4.1) p = 0.46	NA	NA
Any LRI	0.95 (0.47-2) p = 0.89	1.9 (0.74-5.6) p = 0.2	1.4 (0.51-3.8) p = 0.55	1.9 (0.61-6.6) p = 0.29	0.57 (0.18-1.9) p = 0.35	2.6 (0.38-52) p = 0.41
Any fever LRI	**2.3 (1.2-4.5) p = 0.016**	2.2 (0.93-5.2) p = 0.071	**2.7 (1.1-7) p = 0.034**	2.7 (0.89-8.7) p = 0.083	1.5 (0.48-4.6) p = 0.46	1.9 (0.33-9.8) p = 0.45
Any wheezy LRI	1.6 (0.79-3) p = 0.2	1.4 (0.58-3.3) p = 0.45	1.3 (0.49-3.3) p = 0.61	1.6 (0.53-4.8) p = 0.4	1.9 (0.66-5.7) p = 0.23	1.2 (0.22-5.7) p = 0.83
Any mild LRI	0.48 (0.24-0.94) p = 0.035	0.92 (0.4-2.1) p = 0.84	0.51 (0.2-1.2) p = 0.14	0.76 (0.27-2.1) p = 0.61	0.44 (0.13-1.3) p = 0.16	1.6 (0.35-9.1) p = 0.54
Any HRV LRI	1.1 (0.59-2.2) p = 0.69	1.6 (0.69-3.8) p = 0.28	1.2 (0.49-2.9) p = 0.7	1.2 (0.43-3.3) p = 0.75	0.93 (0.31-2.7) p = 0.9	2.8 (0.59-16) p = 0.21
Any RSV LRI	1.1 (0.51-2.4) p = 0.76	1.4 (0.51-3.5) p = 0.5	2.2 (0.76-6.2) p = 0.14	2.7 (0.8-8.9) p = 0.11	0.56 (0.11-2.1) p = 0.42	0.33 (0.016-2.5) p = 0.35
Any risk bacteria LRI	0.89 (0.45-1.8) p = 0.73	2 (0.82-5.2) p = 0.14	0.96 (0.39-2.4) p = 0.93	1.7 (0.59-5.2) p = 0.34	0.73 (0.25-2.3) p = 0.57	4.9 (0.72-98) p = 0.16
Any HRV wheezy LRI	2 (0.93-4.2) p = 0.073	2.1 (0.81-5.4) p = 0.11	2.5 (0.86-7.2) p = 0.092	1.9 (0.55-6.4) p = 0.29	1.8 (0.49-6.3) p = 0.34	2.7 (0.46-14) p = 0.24
Any HRV-C wheezy LRI	2.4 (0.93-6.1) p = 0.064	**3.5 (1.1-11) p = 0.026**	**7.2 (1.7-35) p = 0.009**	**7.1 (1.6-40) p = 0.014**	1.1 (0.15-5) p = 0.95	1.2 (0.058-11) p = 0.86
Any HRV-A wheezy LRI	1.2 (0.42-3.1) p = 0.74	1.4 (0.37-4.7) p = 0.57	0.55 (0.11-2.2) p = 0.43	0.35 (0.018-2.2) p = 0.34	3.5 (0.74-16) p = 0.10	6.3 (0.94-44) p = 0.051
Any RSV wheezy LRI	2.6 (0.71-9.5) p = 0.15	1.7 (0.34-7.5) p = 0.47	7.2 (1-150) p = 0.084	2.5 (0.28-20) p = 0.37	0.95 (0.045-8) p = 0.97	1.2 (0.052-12) p = 0.89
***Streptococcus* colonization at 7 weeks**	**3.8 (1.3-12) p = 0.017**	2.7 (0.66-12) p = 0.18	4.0 (0.88-21) p = 0.077	3.9 (0.63-28) p = 0.15	3.4 (0.55-22) p = 0.17	NA
***Streptococcus* colonization at 8 weeks**	**3.1 (1.2-8.2) p = 0.023**	1.9 (0.46-7.1) p = 0.33	**4.4 (1.1-20) p = 0.040**	2.4 (0.46-12) p = 0.29	2.7 (0.47-14) p = 0.24	NA
***Streptococcus* colonization at 9 weeks**	**2.7 (1.0-7.1) p = 0.039**	1.9 (0.45-6.9) p = 0.36	3.3 (0.86-13) p = 0.080	2.4 (0.47-13) p = 0.28	2.6 (0.45-14) p = 0.26	NA

Table S3. Association between chronic wheeze at age 5 and 10 years, by LRI subtypes and early asymptomatic colonization with *Streptococcus*. Any mild LRI: Any LRI without wheeze or fever symptoms; Any risk bacteria LRI: any LRI in *Streptococcus*, *Moraxella* or *Haemophilus* MPG. *Streptococcus* colonization in the first healthy sample (collected by 7/8/9 weeks of age and prior to any respiratory infection) was assessed using >20% abundance cut-off. Odds ratios, 95% confidence intervals and p-values shown were estimated using logistic regression, adjusted for gender and maternal and paternal history of atopic disease, calculated separately for all children, those who were atopic by 2 years and those not atopic by 2 years.

Variable	*Alloiococcus* colonized (n = 63)	*Moraxella* colonized (n = 64)	Others (n = 107)	*Alloiococcus* vs. *Moraxella* (p-value)	*Alloiococcus* vs. all others (p-value)
Mean no. ARI	4.6	5.0	5.4	0.31	0.077
Mean no. LRI	1.3	1.7	1.9	0.32	0.10
Mean no. fLRI	0.44	0.49	0.51	0.53	0.47
Mean no. wLRI	0.32	0.62	0.53	0.14	0.11
Mean no. OM	0.44	0.57	0.47	0.29	0.54
Any RSV+ ARI	18	30	43	0.094	0.12
Any RSV+ LRI	5	17	27	0.017*	0.0050*
Any RSV+ fLRI	1	8	10	0.034*	0.050*
Wheeze at age 5 y **Atopic** **Non-atopic** **No wheeze**	8 4 44	11 6 38	15 10 56	0.45#	0.41#
Wheeze at age 10 y **Atopic** **Non-atopic** **No wheeze**	5 1 36	4 2 35	13 5 43	1#	0.46#
Atopic by age 2 y	33	32	51	1	0.88

Table S4. Infections and wheeze phenotypes for infants by colonization status. Groups are defined by microbiome profile group (MPG) clustering of healthy samples. '*Alloiococcus* colonized', ≥1 healthy sample with *Alloiococcus* MPG but none with *Moraxella* MPG; '*Moraxella* colonized', ≥1 healthy sample with *Moraxella* MPG but none with *Alloiococcus* MPG; 'Others', no *Moraxella* or *Alloiococcus* healthy sample or ≥1 of each. Infection types: ARI, acute respiratory infection; URI, upper respiratory illness; LRI, lower respiratory illness; fLRI, febrile LRI; wLRI, wheezy LRI; OM, otitis media; RSV+, PCR detection of respiratory syncytial virus; atopy status is defined by IgE > 0.35 kU/L for any antigen at 6, 12 or 24 months. Comparisons of rates of infection were assessed using Wilcoxon rank sum test; other comparisons are binary variables and were calculated using Fisher's exact test; # atopic wheeze vs no wheeze; *$p < 0.05$.