

Shotgun metagenomic profiling of bacterial microbiomes, metagenome-assembled genomes and antimicrobial resistance in respiratory and blood samples from Gambian children with pneumonia

Dam Khan

dam.khan@lshtm.ac.uk

Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine

Josh L. Espinoza

J. Craig Venter Institute

Peggy-Estelle Tientcheu

Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine

Isaac Darko Otchere

Noguchi Memorial Institute for Medical Research

Nuredin Ibrahim Mohammed

Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine

Archibald Worwui

Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine

Mark P. Nicol

University of Western Australia

Brenda Kwambana-Adams

Liverpool School of Tropical Medicine

Martin Antonio

Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine

Chris L. Dupont

J. Craig Venter Institute

Article

Keywords: microbiome, metagenome-assembled genomes, antimicrobial resistance, *Streptococcus pneumoniae*, *Haemophilus* strains

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8724320/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Shotgun metagenomic profiling of bacterial microbiomes, metagenome-assembled genomes and antimicrobial resistance in respiratory and blood samples from Gambian children with pneumonia

Dam Khan^{1,2 *}, Josh L. Espinoza³, Peggy-Estelle Tientcheu¹, Isaac Darko Otchere^{1,4}, Nuredin Ibrahim Mohammed¹, Archibald Worwui¹, Mark P. Nicol⁵, Brenda Kwambana-Adams^{6,7,8}, Martin Antonio^{1,2,9}, and Chris L. Dupont³

Author Affiliations

¹Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine, Banjul, The Gambia

²Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom

³J. Craig Venter Institute, La Jolla, CA, United States of America

⁴Department of Bacteriology, Noguchi Memorial Institute for Medical Research, Legon, University of Ghana, Accra, Ghana

⁵Marshall Centre, Division of Infection and Immunity, School of Biomedical Sciences, University of Western Australia, Perth, Australia

⁶Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom

⁷Malawi Liverpool Wellcome Programme, Blantyre, Malawi

⁸NIHR Global Health Group on Vaccines to Control Respiratory Pathogens and AMR, Division of Infection and Immunity, University College London, London, United Kingdom

⁹Centre for Epidemic Preparedness and Response, London School of Hygiene & Tropical Medicine, London, United Kingdom

Corresponding author:

Dam Khan

dam.khan@lshtm.ac.uk

Abstract

Pneumonia is a leading cause of morbidity and mortality in children, with bacterial pathogens being important etiologic agents. Most microbiome studies in pneumonia use technologies with limited taxonomical resolution and few include lung aspirate or blood samples. In this study, we assessed the microbial communities of the nasopharynx, nasopharynx/oropharynx, induced sputum, lung aspirate and blood, and recovered metagenome-assembled genomes (MAGs) from the same sites using shotgun metagenomics sequencing of samples from children with severe and very severe pneumonia in The Gambia. Our data show that Proteobacteria and Firmicutes were the most common phyla across the body sites, and this was largely driven by *S. pneumoniae*, *H. influenzae/aegyptius* and *M. catarrhalis*. Blood and nasopharyngeal samples showed a mean Jaccard similarity of 58%, which was greater than expected under random pairing of specimens ($q = 0.046$). We recovered 60 medium-quality MAGs, of which 35 met the criteria for high-quality MAGs including 11 *S. pneumoniae*, 10 *H. influenzae* strains and a *Limosilactobacillus* with less than 95% Average Nucleotide Identity to the closest species in the Genome Taxonomy Database (GTDB). We also showed that antimicrobial resistance genes in our MAGs were highly species-specific; more than 70% of the unique AMR genes detected (11/15) were found only within MAGs of a single bacterial species rather than shared across multiple species. Together, these findings demonstrate the value of shotgun metagenomics in resolving species level detail across clinically relevant but understudied specimen types and provide a genomic resource to support future respiratory pathogen surveillance in low- and middle-income settings.

Keywords: microbiome, metagenome-assembled genomes, antimicrobial resistance, *Streptococcus pneumoniae*, *Haemophilus strains*.

Introduction

Lower respiratory tract infections (LRI) pose a major public health burden and remain one of the main infectious causes of death in low- and middle-income countries^{1,2}. LRI, defined by the Global Burden of Disease report as pneumonia or bronchiolitis, was responsible for more than 500,000 deaths in children under five years of age worldwide in 2021³. A wide range of organisms cause LRI including bacteria, viruses and fungi. Among the bacterial causes of LRI in 2021, *Streptococcus pneumoniae* was responsible for the highest number of episodes and deaths with an estimated 97.9 million episodes and 505 000 deaths globally across all ages³. The risk factors for pneumonia include poverty, malnutrition, lack of exclusive breastfeeding and inadequate vaccination^{4,5}.

Effective prevention and management strategies for pneumonia, including vaccination and antimicrobials, rely partly on the identification of the microbial etiologic agent of the pneumonia episode, which can be done using culture, nucleic acid amplification, or antigen testing. However, the use of these targeted detection methods limits our understanding of the complex interaction between microbes in the lungs whose interactions affect disease outcome⁶. Samples from the lower respiratory tract are difficult to obtain; therefore, the upper respiratory tract is often studied to gain insights into the microbiome and its role in disease⁷. Understanding the microbiome of the upper respiratory tract is important because it has been postulated that microaspiration is the primary route through which bacteria in the upper airways enter the lungs^{8,9}, although the temporal, spatial, and regulatory steps that lead to disease are unknown. Moreover, bacterial translocation across the lung epithelium into the blood has also been reported in pneumonia cases¹⁰, highlighting the significance of profiling the microbiome of both the respiratory tract and blood in disease.

The relationship between the upper respiratory airway microbiome and respiratory health is gaining attention^{11,12}. For instance, it has been shown that a high abundance of commensal bacterial species

of the *Corynebacterium* and *Dolosigranulum* genera prime the respiratory tract in children through stimulation of innate immunity and competition for resources to restrict colonisation by pathogenic bacteria¹³⁻¹⁵. Conversely, an increased nasopharyngeal density of *S. pneumoniae* may be associated with respiratory illnesses in children^{16,17}. This suggests colonisation resistance conferred by protective commensals may be overwhelmed once pathogen density exceeds a critical level tipping the balance from asymptomatic carriage toward symptomatic disease.

Our understanding of the microbiome has traditionally been based on findings from 16S rRNA sequencing technology¹⁸⁻²⁰. However, this method lacks species resolution. An untargeted microbial profiling approach, such as shotgun metagenomic sequencing, will provide deeper taxonomical insight into the microbial landscape of the respiratory tract and blood during disease. An added advantage of shotgun metagenomics over 16S rRNA sequencing is its ability to recover metagenome-assembled genomes (MAGs). MAGs are generated by clustering assembled contigs into bins that represent individual organisms, based on nucleotide composition (e.g., tetranucleotide frequency) and coverage patterns across samples²¹. This method has become increasingly important in microbiome research for reconstructing the draft genomes of community members, including uncultured microbes²²⁻²⁴. However, many MAGs reported in published studies are of low quality and completeness²⁴. Additionally, attempts to recover MAGs from respiratory specimens remain limited. Notably, *Li et al.* constructed a comprehensive respiratory genome catalogue (RGC) of 552 MAGs, although these were derived from only 20 studies²⁵.

Given the common use of antibiotic regimens in pneumonia patients, profiling antimicrobial resistance genes (AMR) of MAGs in respiratory and blood samples is essential for informing targeted antimicrobial therapy. Moreover, profiling AMR genes contributes to broader efforts in combating antimicrobial resistance which is a global health concern.

Despite the recognised importance of the lower respiratory tract and blood in understanding pneumonia pathogenesis, existing respiratory metagenomic genome catalogues remain built

predominantly from upper respiratory tract specimens across a limited number of studies and populations, with lung aspirate and blood rarely represented and populations from sub-Saharan Africa notably underrepresented. This study addresses these gaps by sampling across five specimen types, including lung aspirate and blood, from a Gambian paediatric pneumonia cohort.

We performed shotgun metagenomics and analysed nasopharyngeal swab (NPS), combined nasopharyngeal and oropharyngeal swabs (NP/OP), induced sputum (IS), lung aspirate (LA), and blood samples from children diagnosed with severe pneumonia to characterise the microbiomes across these sites. Moreover, we aimed to recover MAGs from sample metagenomes of the same specimen types and to characterise their antimicrobial resistance determinants.

Results

Study population demography, clinical features and specimen types

We analysed 71 samples from 37 children hospitalised with pneumonia and who were recruited as part of the PERCH cohort study²⁶. The demographic and clinical features of the children are presented in Table 1. The majority were male, 59% (22/37) and 41% (15/37) of the children were less than 6 months old. A total of 54% (20/37) and 51% (19/37) of the cases were fully vaccinated with the *Haemophilus influenzae* type b vaccine and the Pneumococcal Conjugate Vaccine (PCV13) respectively. Very severe pneumonia cases accounted for 57% (21/37) of the cases. NPS comprised the largest proportion of specimens accounting for 41% (29/71) (Supplementary Information, Table S1). Lung aspirates were the least represented sample type with seven specimens. Figure 1 shows specimens analysed from each of the 37 children. Overall, 62% (23/37) of the children provided more than one specimen type. The number of samples analysed in this study was based on availability.

Shotgun metagenomic sequencing depth

Shotgun metagenomic sequencing yielded a median of 19.4 million (Q1-Q3: 4,668,548-22,611,641) reads per sample. No detectable sequencing library was obtained from the blank swab processed during library preparation, and it was therefore not sequenced. The median proportion of reads classified as human-derived was 91.9% (Q1-Q3: 88.3-93.4%). After quality trimming and removal of human reads, the median read count per sample was 1.3 million (Q1-Q3: 802,821-1,693,183). Lung aspirate (median:1,659,428; Q1-Q3: 1,298,046–1,983,546) and blood (median:1,409,762; Q1-Q3: 1,200,388–2,148,311) samples had the highest read depth among the specimen types after quality filtering and removal of human-derived reads. Read depth by specimen type is summarised in Supplementary Information, Table S2.

Metagenomic Assembled Genomes

The MAGs metrics reported follow the Minimum Information about a Metagenome-Assembled Genome (MIMAG) guidelines²⁷. A total of 60 Medium-quality MAGs ($\geq 50\%$ completion and $< 10\%$ contamination) were recovered, of which 35 met the criteria for high-quality MAGs ($> 90\%$ completeness and $< 5\%$ contamination) based on MIMAG standards. Details of these 35 high-quality MAGs, including their contamination and completeness as calculated using *CheckM* are provided in Table 2.

Seven species were identified from the 35 high-quality MAGs recovered. These MAGs were obtained from 22 out of 71 samples (31%), and the highest number from a single sample was three. Most MAGs were recovered from blood specimens (11/35, 31%), whereas the fewest were recovered from IS (4/35, 11%). Notably, in 83% (29/35) of the high-quality MAGs, the recovered species corresponded to the species with the highest or second highest relative abundance in the sample from which the MAG was obtained. The number of MAGs recovered by specimen type is described in Supplementary Information, Table S2. Lung aspirate and blood samples showed the highest MAG recovery efficiency when normalised by sample number (1.0 and 1.14 high-quality MAGs per sample

respectively, versus 0.21–0.50 for the other three specimen types). The mean number of contigs across all 35 MAGs was 136.2 (Min:22, Max:339), and the mean contig length was 23,934 bp (Min: 1500, Max: 472,844). MAG19, which had the lowest number of contigs, had the highest genome completeness at 99.7%.

Species assignment of the genomes was based on GTDB-Tk, which integrates average nucleotide identity (ANI) against and phylogenetic placement to assign taxonomy. MAG25 (*Limosilactobacillus* sp.) was not assigned a species-level taxonomic classification. The closest match was *Limosilactobacillus gastricus* (GCF_001434365.1), with an ANI of 83.26%. The ANI values for the remaining 42 genetically related *Limosilactobacillus* genomes in the database ranged from 76.2% to 78.9% relative to MAG25. The most common species across all MAGs were *S. pneumoniae* (11/35, 31%) and *H. influenzae* (10/35, 29%). We obtained MAGs from three unique species of *Moraxella*, including five *M. catarrhalis* MAGs.

We performed in silico serotyping and MLST on *S. pneumoniae* and *H. influenzae* MAGs. There was 100% concordance between PneumoKITy and SeroBA *S. pneumoniae* serotype assignments. Serotype 19A was the most common *S. pneumoniae* serotype (45%, 5/11), as shown in supplementary Information, Table S3. Three of the five serotype 19A strains belonged to the same sequence type (ST847). Three *S. pneumoniae* strains could not be assigned a sequence type on the MLST database. In the only case where we recovered *S. pneumoniae* MAGs from two body sites in the same individual, the strain from blood sample was serotype 19F and ST925, while the strain from NPS was serotype 19A with an unassigned ST. All 10 of the *H. influenzae* strains were non-typeable. Five of them were assigned to distinct sequence types, while the other five could not be assigned a sequence type in the MLST database.

A total of 78 AMR genes, including 15 unique genes were detected across the 35 high-quality MAGs (Figure 2A). These genes conferred resistance to eight antibiotic classes. AMR genes conferring resistance to fluoroquinolones were the most prevalent with four unique genes detected, followed

by tetracycline resistance genes which included three unique genes across all MAGs. The average number of AMR genes per MAG was 2.2 (range: 0 – 6) (Figure 2B). Of the nine MAGs with no AMR genes detected, 78% (7/9) were *Haemophilus* species while the remaining two were MAG 25 (*Limosilactobacillus* sp.) from the NP/OP and MAG 35 (*M. lincolnii*) from the NPS.

The AMR genes were largely species specific. Up to 73% (11/15) of the unique AMR genes detected were found only within MAGs of a single bacterial species rather than shared across multiple species. The exceptions were AMR genes *tet(B)*, *tet(R)*, *BRO-1* and *ICR-Mc* each of which was found in MAGs from two different species. In terms of antibiotic class, AMR genes conferring resistance to tetracycline and peptide antibiotic class were present in MAGs of three different species.

Notably, all four *H. influenzae* MAGs from LA carried the *LpsA* gene while none from NPS, NP/OP and IS had the *LpsA* gene. Among the three *H. influenzae* MAGs from blood, two lacked the *LpsA* gene. All 11 *S. pneumoniae* MAGs carried the *patA* and the *patB* genes. The only non-fluoroquinolone or MLS conferring resistance genes in *S. pneumoniae* MAGs were *tet(M)* and *tet(B)* found in blood-derived MAGs. The *ICR-Mc* gene was found in all *M. catarrhalis* MAGs.

We compared the genomic diversity of our high-quality respiratory MAGs against the RGC of *Li et al.*²⁵ which contains 552 non-redundant medium- and high-quality MAGs using FastANI v1.32. The comparison was restricted to the 24 high-quality MAGs recovered from our respiratory specimens (NPS, NP/OP, IS, LA), as the RGC was derived exclusively from respiratory specimens. In total, 83% (20/24) of our respiratory MAGs shared $\geq 95\%$ ANI with an existing RGC genome and matched the same species classification (Supplementary Information, Table S4). The remaining four MAGs (17%) that did not meet the $\geq 95\%$ ANI threshold were MAG1 (*N. meningitidis*), MAG2 (*H. parahaemolyticus*), MAG8 (*H. parahaemolyticus*) and MAG25 (*Limosilactobacillus* Sp.). MAG1 (*N. meningitidis*) shared 85.6% ANI with a *Neisseria* RGC genome that was not resolved to species level. MAG2 and MAG8 (both *H. parahaemolyticus*) independently matched the same RGC genome

(annotated *Haemophilus parainfluenzae*) sharing 82.6% and 82.8% ANI with it respectively. MAG25 (*Limosilactobacillus* sp) had no detectable hit against any RGC genome.

Microbiome composition of children with pneumonia

All ecological analyses, including relative abundance bar plots, diversity measures, and differential abundance testing, were performed using the species-level cluster (SLC) relative abundance across samples. Each SLC was a cluster of MAGs with $\geq 95\%$ Average Nucleotide Identity (ANI). The *Haemophilus aegyptius* (a *H. influenzae* biotype) MAG obtained in our dataset clustered at the 95% ANI with *H. influenzae* MAGs. This SLC is designated as *Haemophilus influenzae/aegyptius*. In SLCs that lack species-level classification, the genus name was used along with the SLC identifier as a suffix. To minimize the effects of sequencing noise, taxa with a relative abundance of $< 0.1\%$ in any given sample were filtered out from those samples in the analysis.

The mean compositional relative abundance at the phylum and species levels across all samples is shown in Figure 3. The mean compositional relative abundance at the same taxonomic levels for individual samples is shown in Supplementary Information, Figure S1. Overall, the Proteobacteria phylum predominated in IS (66.0%), NPS (63.3%), LA (55.7%) and blood (66.6%) samples. In contrast, Firmicutes was the dominant coloniser in NP/OP samples (46.6%). Bacteroidota and Actinobacteriota contributed less than 5% of the bacterial DNA reads in all specimen types except NP/OP. Fusobacteriota was detected at low mean compositional relative abundance in NPS, NP/OP and IS (0.3% – 0.5%) samples and was absent in LA and blood samples. Patescibacteria (Candidate Phyla Radiation) was detected in one IS sample and one NP/OP sample, with compositional relative abundances of 0.6% and 0.1% respectively in those samples. *S. pneumoniae*, *H. influenzae/aegyptius* and *M. catarrhalis* consistently had the highest compositional mean relative abundances across all specimen types. *M. catarrhalis* predominated in IS samples with a compositional mean relative abundance of 25.0%; *S. pneumoniae* predominated in NP/OP (29.9%), NPS (26.4%) and LA (40%)

220 samples. In blood samples, *H. influenzae/aegyptius* predominated with 31.5%. Although *S.*
 221 *pneumoniae*, *H. influenzae/aegyptius* and *M. catarrhalis* were the most abundant based on the mean
 222 relative abundance across specimen types, they were not consistently the most dominant at the
 223 individual sample level, except in LA samples, where 57% (4/7) and 43% (3/7) of the samples had *H.*
 224 *influenzae/aegyptius* and *S. pneumoniae* as the dominant species (Supplementary Information,
 225 Figure S2). We defined a species as dominant in a sample if it had the highest relative abundance in
 226 that sample. In both IS and NP/OP specimens, eight different species dominated across the 12
 227 samples in each group.

228 We compared Shannon diversity across specimen types using a linear mixed-effects model (Figure 4).
 229 The NP/OP samples had the highest diversity (2.13), whereas the LA samples had the lowest diversity
 230 (1.19). Pairwise comparison showed significant differences in Shannon diversity between NP/OP and
 231 LA ($p = 0.001$), NP/OP and blood ($p = 0.008$), NPS and LA ($p = 0.01$), IS and Blood ($p = 0.038$), IS and LA
 232 ($p = 0.003$).

233 We assessed bacterial species overlap between blood and respiratory specimens from the same
 234 participants using the Jaccard similarity index (Figure 5). NPS and blood samples from the same
 235 participants had a mean Jaccard similarity of 58%, which was greater than the similarity observed
 236 under random pairing ($q = 0.046$). In contrast, participant-matched induced sputum and blood
 237 samples had mean Jaccard similarity 41%, which did not exceed that expected under random pairing
 238 ($q = 0.522$). Similarly, NP/OP and blood samples from the same participants had a mean Jaccard
 239 similarity of 34%, with no evidence that similarity exceeded that expected under random pairing ($q =$
 240 0.981).

241 We restricted our analyses to the nasopharynx and compared the microbial DNA profiles of children
 242 with severe and very severe pneumonia. The compositional mean relative abundance of *S.*
 243 *pneumoniae* in the nasopharynx was 33.1% in severe cases and 22.1% in very severe cases (Figure 6).
 244 *M. catarrhalis* was detected at a compositional mean relative abundance of 21.3% in severe cases

and 27.5% in very severe cases. The compositional mean relative abundance of *H. influenzae/aegyptius* was similar between the two groups (11.1% in severe cases vs. 11.6% in very severe cases). Notably, the relative abundance of *M. nonliquefaciens* in very severe cases (13.1%) was threefold higher than that in severe cases (3.9%). There was no significant difference in nasopharyngeal Shannon diversity between severe cases and very severe cases ($p=0.217$), (Supplementary Information, Figure S3). In addition, the multivariable model found no significant association in nasopharyngeal Shannon diversity between severe cases and very severe cases ($p=0.069$). Furthermore, PERMANOVA detected no significant difference in nasopharyngeal microbiome composition between the two groups ($P=0.634$), Supplementary Information, Figure S4. MaAsLin3 differential abundance testing of nasopharyngeal samples yielded no significant associations between bacterial species and pneumonia severity classification. We compared the relative abundance of *S. pneumoniae* between children with full ($n=14$) versus incomplete PCV vaccination status ($n=13$), as well as *H. influenzae/aegyptius* relative abundance between children with full ($n=15$) versus incomplete Hib vaccination status ($n=12$) in the nasopharynx of the children. *S. pneumoniae* relative abundance was similar between fully and incompletely PCV-vaccinated children (mean relative abundance: 27.5% vs. 27.8%; $q = 0.519$). *H. influenzae/aegyptius* relative abundance was lower in fully Hib-vaccinated children compared to those incompletely vaccinated (mean relative abundance: 7.1% vs. 13.6%), although this difference was not statistically significant after correction for multiple testing ($q = 0.106$).

Discussion

Few studies have characterised the respiratory or blood microbiome in children with pneumonia in such detail. Using shotgun metagenomics in this study, we found *S. pneumoniae*, *H. influenzae/aegyptius* and *M. catarrhalis* were the most common species from all sample types. Furthermore, we observed greater species overlap between blood and nasopharyngeal samples from the same participants than expected under random pairing. A major strength of our work lies in the

271 recovery of 35 high-quality MAGs in these niches, including what appears to be a novel
 272 *Limosilactobacillus* with less than 95% ANI to any species in GTDB.

273 Our analysis of the microbial profile of the respiratory tract and blood during infection achieved
 274 species-level resolution which is seldom achieved in respiratory microbiome studies. Proteobacteria
 275 and Firmicutes dominated in the respiratory samples, and this was largely driven by *S. pneumoniae*,
 276 *H. influenzae* and *M. catarrhalis*. These are clinically important pathogens, and vaccines targeting *S.*
 277 *pneumoniae* and *H. influenzae* type b are routinely administered in many countries including The
 278 Gambia²⁸⁻³⁰. Our findings are consistent with other studies in The Gambia where *S. pneumoniae* and
 279 *H. influenzae* were identified as the primary bacterial pathogens causing pneumonia in children,
 280 based on microbiology culture and PCR analyses of lung and pleural aspirates³¹. The high prevalence
 281 of *S. pneumoniae* detected in our analysis is also consistent with previous reports from The Gambia
 282 where carriage rates among children under one year of age were as high as 97%³². Although the
 283 introduction of Pneumococcal Conjugate Vaccines (PCVs) has led to the reduction of vaccine-type
 284 carriage, the overall carriage of *S. pneumoniae* remains unchanged. The reason for this persistence is
 285 likely due to the increase in the carriage of non-vaccine serotypes³³.

286 *S. pneumoniae* serotype 19A was the most common serotype in our dataset among the high-quality
 287 MAGs we recovered. This serotype increased in prevalence in many countries following the
 288 introduction of the seven valent pneumococcal vaccine (PCV7)³⁴. ST847 was the most frequently
 289 observed assigned sequence type in our dataset and has been associated with serotype 19A in both
 290 carriage³⁵ and disease³⁶. *H. influenzae* type b (Hib) disease and carriage have declined globally due to
 291 the introduction of the Hib vaccine^{37,38}. However, non-typeable *H. influenzae* which was the most
 292 common *H. influenzae* type recovered among our high-quality MAGs, has been reported elsewhere
 293 as responsible for the majority of otitis media, sinusitis, and pneumonia cases among children
 294 vaccinated with the Hib vaccine³⁹.

295 In our analysis, there were no significant differences in nasopharyngeal microbiome diversity
296 between severe and very severe cases. Differential abundance testing also revealed no significant
297 differences in any species by case type. This finding may reflect our small sample size limitation.
298 Previous studies have reported that the upper respiratory tract microbiome composition can serve as
299 a marker for severity of lower respiratory tract disease⁴⁰.

300 It is traditionally understood that the blood is sterile, largely from the fact that the presence of
301 bacteria as low as 1-10 bacterial cells per millilitre of whole blood is potentially life threatening⁴¹.
302 However, there is growing evidence that bacterial DNA of various taxa can be detected in blood
303 driven by 16S or shotgun sequencing ⁴²⁻⁴⁴. It remains unclear whether bacterial DNA detected in
304 blood represents viable organisms or remnants, and whether these arise from leakages from other
305 body sites. Proteobacteria, Firmicutes and Actinobacteria have been reported as predominant phyla
306 in blood across various conditions and age group⁴⁵⁻⁴⁷. Consistently, our analysis of children with
307 pneumonia revealed a similar dominance of the same phyla.

308 The fact that DNA from many different bacterial taxa can be detected in blood poses a problem in
309 using bacterial reads from metagenomic sequencing in determining etiology. *Barbeta, E. et al.* found
310 21 Operational Taxonomic Units in sixteen blood samples, even though only fifteen percent of those
311 samples had positive blood culture⁴⁸. A key finding from our study is that nasopharyngeal and blood
312 samples from the same individuals shared greater species overlap than expected under random
313 pairing. This participant-specific similarity was specific to the NPS–blood comparison, as the overlap
314 between induced sputum and blood and between NP/OP and blood did not exceed that observed
315 under random pairing. Whether the observed NPS–blood similarity reflects biological continuity
316 between the upper respiratory tract and bloodstream, or other factors influencing the detection of
317 bacterial DNA in blood, warrants further investigation

318 We conducted a comprehensive sample level recovery of metagenomes resulting in the identification
319 of 35 MAGs that met the MIMAG criteria for high-quality genomes. This was achieved using

320 consensus binning, a very stringent approach that leverages the strengths of the different binning
321 algorithms (MaxBin2, MetaBAT2 and CONCOCT) while minimising their individual limitations. This
322 strategy recovered high-quality bins that would have been missed by using a single algorithm alone.
323 However, the relatively limited microbial sequencing depth after host-read removal may have
324 constrained assembly and MAG recovery, particularly for low-abundance organisms. The 35 high-
325 quality MAGs recovered in this study are therefore not an exhaustive representation of the microbial
326 communities present in these specimens.

327 Recovery of MAGs from metagenomes enhances the identification of organisms that are not
328 routinely targeted for disease diagnostic purposes by other methods such as PCR. Furthermore, it is
329 a valuable for identifying species that are difficult to culture in the lab. Notably, we recovered a high-
330 quality genome belonging to the *limosilactobacillus* genus that did not meet the 95% ANI threshold
331 for species-level assignment in the Genome Taxonomy Database. An ANI range of 95–96% is
332 commonly used to delineate species boundary^{49,50}. *Limosilactobacillus* is a genus within the family
333 Lactobacillaceae. It was formally classified as lactobacillus until a reclassification in 2020 divided
334 lactobacillus into 25 novel genera including *Limosilactobacillus*⁵¹. To date, up to 23 species of
335 *Limosilactobacillus* have been described in both human and non-human hosts⁵². Our
336 *Limosilactobacillus* sp. MAG was recovered from the NP/OP, and it potentially represents a previously
337 undescribed or novel species within this genus.

338 To further contextualise the diversity of our recovered genomes, we compared our respiratory MAGs
339 against the Respiratory Genome Catalogue²⁵. While the majority (83%) of our respiratory MAGs
340 closely matched existing catalogued genomes, four of our MAGs were not represented in the
341 catalogue suggesting that, despite expected overlap with well-characterised respiratory taxa, our
342 dataset contributed additional genomic representation. This underscores the value of sampling
343 underrepresented populations.

While we used *checkM* metrics to ascertain the quality of our genomes, some studies have raised concern that the quality of MAGs may be overestimated by commonly used pipelines for assessing genome quality. *Meziti et al.*⁵³ explored this by comparing gene variability between pathogenic *Escherichia coli* isolates from diarrheal samples against their corresponding MAGs recovered from the same samples. Their analysis showed that completeness estimates near 95% as reported by the *MiGA* workflow captured only 77% of the population core genes and 50% of the variable genes. They were also able to find that about 5% of the genes from the MAGs were missing in the isolate and were of different taxonomic origin, even though pipeline-based contamination estimates were as low as 1.5%. However, these findings were based on a single bacterial species and may not apply to the species present in our dataset. *Meziti et al.* further outlined key criteria that contributes to MAGs reliability beyond standard pipeline-based completeness and contamination estimates. These include MAGs having fewer than 500 contigs, contig lengths greater than 1000 bp, N50 values above 20,000 bp and genome size deviations less than 0.5 Mb compared to reference strains. Our high-quality genomes meet these benchmarks in terms of contig number, contig length, genome size consistency. Additionally, more than 60% of our MAGs had N50 values exceeding 20,000 bp further supporting the overall high-quality of our MAGs. However, we still would recommend conducting gene content comparison and SNP analysis between paired *S. pneumoniae* isolate and MAG strains obtained from the same sample. This will be crucial for assessing the reliability of using sample driven *S. pneumoniae* MAGs for public health applications such as tracking vaccine impacts and in outbreak investigations. What we have not been able to achieve is the recovery of multiple high-quality MAGs of the same species from a single sample. In each case, only one strain of a single species is successfully binned. This limitation is particularly important for *S. pneumoniae* where multiple strain carriage in the upper respiratory tract is common^{54,55}. Co-carriage was observed in 15% of children in South Africa⁵⁶ using shotgun metagenomics sequencing and more than 48% of children in Southeast Asia by microarray⁵⁷. In our upper respiratory tract samples where *S. pneumoniae* was recovered, it is possible that more than one strain was present in some samples but only one dominant strain was

recovered. This suggests that our sequencing procedure or sample-specific binning procedures are limited in their ability to recover co-existing strains of the same species in a sample. The only pair of *S. pneumoniae* MAGs recovered from blood and NPS of the same individual belonged to different serotypes and sequence types. This highlights that using MAGs to study strain-level similarity of *S. pneumoniae* across body sites may be unreliable.

While the presence of an AMR genes does not equate to clinical resistance, monitoring their presence in the respiratory tract and blood could provide insights into which antibiotics the bacteria are resistant to. Additionally, given that some AMR genes can be transferred to other bacterial species in the same niche through horizontal gene transfer, it is important to identify which genes have the potential to spread to other bacteria. The *patA* and *patB* AMR genes were highly prevalent in our *S. pneumoniae* MAGs. These genes interact to form an ATP-binding cassette (ABC) antibiotic efflux pump that confers resistance to fluoroquinolones in *S. pneumoniae*^{58,59}. Resistance to the common fluoroquinolone, ciprofloxacin in *S. pneumoniae* mainly occurs through the acquisition of mutations in the quinolone resistance-determining region of the *ParC* and *gyrA* genes. However, *Lupien et al.* have shown that the *patA* and *patB* genes also contribute to low-level resistance to ciprofloxacin in a clinical isolate⁶⁰. The analysis we carried out on the CARD database couldn't infer penicillin resistance due to PBP mutations. Future analyses could complement CARD-based genotypic screening with CDC PBP-based pipeline to infer penicillin-binding protein-mediated β -lactam resistance phenotypes. *M. catarrhalis* is among the species that are intrinsically resistant to colistin. The ICR-Mc gene found in our *M. catarrhalis* MAGs encodes a chromosomally located colistin resistance phosphoethanolamine (PEtN) transferase⁶¹. The protein is the closest known ortholog to the well-known plasmid mediated colistin resistance genes *mcr-1* and *mcr-2*⁶¹. The *LpsA* gene was detected in *H. influenzae* MAGs in LA but absent in MAGs from NP/OP, NPS and IS. *LpsA* confers intrinsic resistance to peptide antibiotics⁶² but it is more widely known for its role in lipooligosaccharide biosynthesis⁶³ and it has been associated with adaptation to the lung environment and invasiveness⁶⁴.

A major limitation of our study was the sample size which limited statistical power and the generalisability of some of our findings. Secondly, we did not obtain controls at the sampling point to control for contamination during sampling. Controls at the DNA extraction stage were also not available for sequencing. Therefore, contamination introduced during sample collection or DNA extraction could not be directly assessed. However, we reduced the potential influence of taxa commonly associated with skin contamination by excluding *Cutibacterium acnes*, *Staphylococcus saprophyticus* and *Staphylococcus epidermidis* from all samples in this analysis. This decision was made because these bacteria primarily colonise the skin and not the specimen types we analysed and could have contaminated our samples during sample collection. They have also been implicated as contaminants in microbiome studies of low-biomass samples^{41,65}. A further limitation in the DNA extraction method was the confounding with specimen type. Because extraction kit and specimen type did not vary independently in this study, we could not statistically disentangle the effects of extraction method from those of specimen type. Consequently, differences in microbial composition may partly reflect differences in DNA extraction efficiency rather than biological differences between anatomical sites. The use of different extraction protocols was necessary to accommodate the distinct specimen matrices and their respective processing requirements, but this limits direct comparisons of microbial composition across specimen types. Future studies designed to compare microbial communities across specimen types should include appropriate cross-extraction controls to quantify methodological effects. Thirdly, blood, NP/OP, and induced sputum specimens were analysed exclusively from patients with very severe disease, resulting in strong confounding between specimen type and disease severity. Therefore, observed differences in microbial diversity between specimen types should be interpreted cautiously as they may reflect differences in disease severity rather than anatomical site alone. Fourth, a few children were administered antibiotics before sample collection, and this could have impacted their microbiome profiles. The fifth limitation was the substantial proportion of host-derived reads following microbial DNA enrichment. The relatively limited microbial sequencing depth may have constrained assembly and MAG recovery, particularly

for low-abundance organisms. Additionally, pre-enrichment sequencing was not performed. Therefore, we could not directly quantify the efficiency of the microbial DNA enrichment procedure or determine the reduction in host-derived DNA attributable to the enrichment step. Finally, our analysis could not establish whether the relative abundances of the different species in our samples reflected the existence of viable bacteria or bacteria-derived DNA. Therefore, the bacterial profiles of samples may not represent the viable microbiome at the time of sample collection, and as such our findings should be interpreted with caution.

Conclusion

Despite our limitations, we have been able to show that overall, the bacterial DNA from *S. pneumoniae*, *H. influenzae/aegyptius* and *M. catarrhalis* were the most common across the different body sites we studied in children with pneumonia in The Gambia. We also found greater bacterial species similarity between nasopharyngeal and blood samples from the same participants than expected under random pairing, suggesting a participant-specific signal in the bacterial DNA detected across these specimens. Thirdly, we have been able to recover high-quality MAGs from our samples further advancing the field of genomics of important pathogens like *H. influenzae* and *S. pneumoniae*. For future studies, it is essential to include sampling controls that undergo the same sequencing procedure as the samples. We also propose future studies in metatranscriptomics to study the activities of live bacteria present in blood and respiratory samples of pneumonia patients. Finally, we propose comparing MAGs and isolates of *S. pneumoniae* strains from the same respiratory or blood sample to assess how well the MAG genome reflects that of the isolate.

Methods

Study participants and sample collection

We used stored samples obtained between 2011 and 2014 from participants recruited in the Pneumonia Etiology Research for Child Health (PERCH) study²⁶. PERCH was a case-control study that sought to characterise the causes of severe childhood pneumonia in children living in high pneumonia burden and low-resource regions, including The Gambia. The samples analysed in this study were exclusively from children with pneumonia. In the PERCH study, children aged 1-59 months who presented and were admitted to the hospital in Basse, Upper River Region of The Gambia, with severe or very severe pneumonia were recruited. Severe pneumonia was defined as cough or difficulty breathing with lower chest wall indrawing, and very severe pneumonia was defined as cough or difficulty breathing and at least one of the following signs: vomiting everything, difficulty in drinking or breastfeeding, central cyanosis, lethargy, convulsions, unconsciousness, or head-nodding. Specimens, including NPS, NP/OP swabs, blood, IS and LA, were collected at enrolment. NPS were collected and stored as a distinct specimen. Separately, NP/OP specimen was collected by combining a second nasopharyngeal swab with an oropharyngeal swab into a single collection tube. Further details on participant recruitment, sample collection rationale, case definitions, inclusion criteria of the PERCH study are described elsewhere^{26,66}.

Ethics

Ethical approval for the Gambian samples used in the PERCH study was obtained from the Johns Hopkins Bloomberg School of Public Health Institutional Review Board (JHSPH IRB), and local ethical approval was obtained from the Joint MRC Unit The Gambia at The London School of Hygiene and Tropical Medicine and The Government of The Gambia Ethics Committee. Written informed consent was obtained from the patient's guardians for all procedures performed in the study. All procedures were performed in accordance with relevant guidelines and regulations.

Laboratory procedures

DNA was extracted from all samples using QIAGEN kits at the molecular diagnostics laboratory at the Medical Research Council unit The Gambia at the London School of Hygiene and Tropical Medicine,

based in The Gambia following the manufacturer's instructions. The laboratory is subscribed to the UK NEQAS External Quality Assessment (EQA) provider. DNA extraction methods differed according to specimen matrix. Total DNA was extracted from the stored NPS, NP/OP, and IS samples using the DNeasy PowerSoil Kit⁶⁷, whereas DNA from the LA and blood samples was extracted using the QIAamp MinElute Virus Spin kit⁶⁸ and the QiAamp DNA Blood Mini Kit⁶⁹, respectively. Library preparation and sequencing of the samples were carried out at the John Craig Venter Institute, San Diego, US. Firstly, extracted DNA samples were enriched for microbial DNA using the NEBNext Microbiome DNA kit (New England Biolabs) following the manufacturer's instructions before whole metagenomic sequencing using the Illumina NovaSeq 2X150bp. DNA extraction controls were not available for sequencing. Blank controls (sterile swabs) were included throughout the library preparation process to monitor for contamination in the laboratory reagents.

Sequencing reads processing

Preprocessing of reads, and the recovery, taxonomic classification and quality assessment of MAGs were performed using modules in the Viral (micro) Eukaryotic Bacterial Archaeal (VEBA) open-source software suite, v1.0.2a^{70,71}. Within VEBA's *preprocess.py* module, adapter removal and quality trimming were carried out using fastp v0.23.2⁷², and pairing of the trimmed reads was assessed and corrected using BBTools v38.92⁷³. Sequencing reads were mapped to the T2T-CHM13v2.0 human reference genome using Bowtie2 v2.3.5.1⁷⁴. All reads aligning to this reference were classified as human-derived and removed prior to downstream analysis. Cleaned reads were input into VEBA's *assembly.py* module, where they were assembled using metaSPAdes (SPAdes suite, v3.15.3)⁷⁵. Binning was performed using MetaBAT2 v2.15⁷⁶, CONCOCT v1.1.0⁷⁷ and MaxBin2 v2.2.7⁷⁸, followed by DAS Tool v1.1.2⁷⁹ for consensus binning. MAGs were quality assessed by CheckM v1.1.3⁸⁰. The genome statistics such as N50, number of contigs and genome size were calculated using *SeqKit* v2.0.0⁸¹. Taxonomic classification of the genomes was performed using the Genome Taxonomy

497 Database Toolkit *GTDB-TK*⁸² v2.1.1 with GTDB release R207v2, based on sequence similarity
498 (FASTANI) and phylogenetic placement.

499 MAGs with ≥95% Average Nucleotide Identity (ANI) were clustered into SLCs using VEBA's *cluster.py*
500 module. Sequencing reads were mapped to SLC representative genomes, and contig-level counts
501 were generated using *featureCounts* as part of the VEBA workflow. For samples with unbinned
502 contigs, pseudo- coassemblies were created to enable additional binning and taxonomic assignment.

503 ***S. pneumoniae* and *H. influenzae* serotyping and Multi-Locus Sequence Typing**

504 FASTA files from *S. pneumoniae* MAGs were analysed using PneumoKITy v1.0⁸³ to assign in silico *S.*
505 *pneumoniae* serotypes. SeroBA (v2.0.4)⁸⁴ was used to cross-validate the PneumoKITy serotype
506 assignments. In silico serotyping of *H. influenzae* MAGs was performed using hicap v1.0.3⁸⁵. Multi-
507 locus sequence types for both *S. pneumoniae* and *H. influenzae* were assigned using PubMLST⁸⁶. The
508 seven housekeeping genes analysed for *S. pneumoniae* were: *aroE*, *ddl*, *gdh*, *gki*, *recP*, *spi*, and *xpt*.
509 For *H. influenzae*, the seven genes were: *adk*, *atpG*, *frdB*, *fuck*, *mdh*, *pgi* and *recA*

510 **Antimicrobial Resistance determinants for MAGs**

511 The recovered MAGs were submitted to the Comprehensive Antimicrobial Resistance Database
512 (CARD) to identify and annotate AMR genes. Only AMR gene hits classified under “perfect” or “strict”
513 categories and a cutoff of >90% sequence identity of the matching region were considered.

514 **Statistical analysis**

515 Statistical analysis and visualisation were conducted using Rstudio v4.4.2. Shannon diversity indices
516 were compared across specimen types using a linear mixed-effects model with specimen type as a
517 fixed effect and participant as a random intercept to account for repeated sampling of the same
518 children across multiple specimen types. Pairwise comparisons between specimen types were
519 derived from the fitted model using estimated marginal means and Tukey's method was applied to
520 adjust for multiple comparisons.

521 We calculated the Jaccard similarity index to quantify bacterial species overlap between paired
522 blood and respiratory samples from the same participants. In addition, we performed a permutation
523 test to generate a null distribution. The null hypothesis was that participant-matched samples were
524 no more similar than randomly paired samples. For each permutation, we randomly shuffled the
525 specimens of one type across participants while keeping the specimens of the other type fixed,
526 thereby breaking the original participant-level pairing. We recalculated the mean Jaccard similarity
527 for each permutation and repeated this procedure 9,999 times to generate the null distribution. We
528 calculated one-sided P values as the proportion of permuted mean similarities that were greater
529 than or equal to the observed mean similarity. We performed separate tests for NPS versus blood,
530 induced sputum versus blood, and NP/OP versus blood. We used the Benjamini–Hochberg procedure
531 to control for false discovery rate and $q < 0.05$ was considered significant.

532 We evaluated differences in Shannon diversity between the nasopharynx of severe cases and very
533 severe cases using a two tailed Mann-Whitney U test. In addition, a multivariable linear model of
534 Shannon diversity with disease severity, age and PCV vaccination as covariates was fitted as a
535 sensitivity analysis. We computed the Aitchison distance (i.e., Euclidean distance on the centred log-
536 ratio transformed data) and used principal coordinates analysis (PCoA) to plot the beta diversity of
537 the nasopharynx in severe and very severe cases. Permutational multivariate analysis of variance
538 (PERMANOVA) was performed to determine whether beta diversity differed between the two groups
539 adjusting for age and vaccination status. Differences in *S. pneumoniae* relative abundance between
540 fully PCV vaccinated and incompletely vaccinated children (combining partial and no vaccination)
541 were restricted to nasopharyngeal swab (NPS) samples and evaluated using the two-tailed Mann-
542 Whitney U test. A similar analysis was carried out to test for differences in *H. influenzae* relative
543 abundance between fully Hib vaccinated and incompletely vaccinated children. P values were
544 adjusted using the Benjamini-Hochberg (q) false discovery rate correction. A significance threshold of
545 $q < 0.05$ was used.

We used Microbiome Multivariable Associations with Linear Models (MaAsLin3 v 0.99.2)⁸⁷ to identify differences in microbial species between the nasopharynx of severe cases and very severe cases. Abundance data were normalised using centred log-ratio (CLR) transformation implemented in MaAsLin3, with zero values handled internally by the package. Species were filtered to those detected in at least 15% of samples prior to model fitting to ensure adequate non-zero observations for stable effect estimation. Age was included as a covariate in the model, and all associations were considered significant at an FDR-corrected q-value threshold of 0.1, consistent with MaAsLin3 default.

All figures in the manuscript were generated using ggplot2 v 4.0.3 and microViz v 0.12.6⁸⁸.

References

- 1 Liu, L. *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* **388**, 3027-3035 (2016). [https://doi.org:10.1016/s0140-6736\(16\)31593-8](https://doi.org:10.1016/s0140-6736(16)31593-8)
- 2 Walker, C. L. F. *et al.* Global burden of childhood pneumonia and diarrhoea. *Lancet* **381**, 1405-1416 (2013). [https://doi.org:10.1016/s0140-6736\(13\)60222-6](https://doi.org:10.1016/s0140-6736(13)60222-6)
- 3 Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990-2021: a systematic analysis from the Global Burden of Disease Study 2021. *Lancet Infect Dis* **24**, 974-1002 (2024). [https://doi.org:10.1016/s1473-3099\(24\)00176-2](https://doi.org:10.1016/s1473-3099(24)00176-2)
- 4 Goyal, J. P. *et al.* Risk Factors for the Development of Pneumonia and Severe Pneumonia in Children. *Indian Pediatr* **58**, 1036-1039 (2021).
- 5 Nguyen, T. K. *et al.* Risk factors for child pneumonia - focus on the Western Pacific Region. *Paediatr Respir Rev* **21**, 95-101 (2017). <https://doi.org:10.1016/j.prrv.2016.07.002>
- 6 Pirrone, M., Pinciroli, R. & Berra, L. Microbiome, biofilms, and pneumonia in the ICU. *Curr Opin Infect Dis* **29**, 160-166 (2016). <https://doi.org:10.1097/qco.0000000000000255>
- 7 Kumpitsch, C., Koskinen, K., Schöpf, V. & Moissl-Eichinger, C. The microbiome of the upper respiratory tract in health and disease. *BMC Biol* **17**, 87 (2019). <https://doi.org:10.1186/s12915-019-0703-z>
- 8 Dickson, R. P. *et al.* Bacterial Topography of the Healthy Human Lower Respiratory Tract. *mBio* **8**, 10.1128/mbio.02287-02216 (2017). <https://doi.org:doi:10.1128/mbio.02287-16>
- 9 Mitsi, E. *et al.* Nasal Pneumococcal Density Is Associated with Microaspiration and Heightened Human Alveolar Macrophage Responsiveness to Bacterial Pathogens. *Am J Respir Crit Care Med* **201**, 335-347 (2020). <https://doi.org:10.1164/rccm.201903-0607OC>
- 10 Bhowmick, R. *et al.* Systemic Disease during Streptococcus pneumoniae Acute Lung Infection Requires 12-Lipoxygenase-Dependent Inflammation. *The Journal of Immunology* **191**, 5115-5123 (2013). <https://doi.org:10.4049/jimmunol.1300522>
- 11 de Steenhuijsen Piters, W. A., Sanders, E. A. & Bogaert, D. The role of the local microbial ecosystem in respiratory health and disease. *Philos Trans R Soc Lond B Biol Sci* **370** (2015). <https://doi.org:10.1098/rstb.2014.0294>
- 12 Camelo-Castillo, A. *et al.* Nasopharyngeal Microbiota in Children With Invasive Pneumococcal Disease: Identification of Bacteria With Potential Disease-Promoting and Protective Effects. *Front Microbiol* **10**, 11 (2019). <https://doi.org:10.3389/fmicb.2019.00011>
- 13 Biesbroek, G. *et al.* Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children. *Am J Respir Crit Care Med* **190**, 1283-1292 (2014). <https://doi.org:10.1164/rccm.201407-1240OC>
- 14 Stubbendieck, R. M., Hurst, J. H. & Kelly, M. S. Dolosigranulum pigrum: A promising nasal probiotic candidate. *PLoS Pathog* **20**, e1011955 (2024). <https://doi.org:10.1371/journal.ppat.1011955>
- 15 Raita, Y. *et al.* Maturation of nasal microbiota and antibiotic exposures during early childhood: a population-based cohort study. *Clin Microbiol Infect* **27**, 283.e281-283.e287 (2021). <https://doi.org:10.1016/j.cmi.2020.05.033>

621 16 Chochua, S. *et al.* Increased Nasopharyngeal Density and Concurrent Carriage of
622 *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis* Are
623 Associated with Pneumonia in Febrile Children. *PLoS One* **11**, e0167725 (2016).
624 <https://doi.org/10.1371/journal.pone.0167725>

625 17 Fan, R. R. *et al.* Nasopharyngeal Pneumococcal Density and Evolution of Acute Respiratory
626 Illnesses in Young Children, Peru, 2009-2011. *Emerg Infect Dis* **22**, 1996-1999 (2016).
627 <https://doi.org/10.3201/eid2211.160902>

628 18 Xu, L., Earl, J. & Pichichero, M. E. Nasopharyngeal microbiome composition associated with
629 *Streptococcus pneumoniae* colonization suggests a protective role of *Corynebacterium* in
630 young children. *PLoS One* **16**, e0257207 (2021).
631 <https://doi.org/10.1371/journal.pone.0257207>

632 19 de Steenhuijsen Piters, W. A. A. *et al.* Interaction between the nasal microbiota and *S.*
633 *pneumoniae* in the context of live-attenuated influenza vaccine. *Nat Commun* **10**, 2981
634 (2019). <https://doi.org/10.1038/s41467-019-10814-9>

635 20 Kelly, M. S. *et al.* Non-diphtheriae *Corynebacterium* species are associated with decreased
636 risk of pneumococcal colonization during infancy. *Isme j* **16**, 655-665 (2022).
637 <https://doi.org/10.1038/s41396-021-01108-4>

638 21 Han, H., Wang, Z. & Zhu, S. Benchmarking metagenomic binning tools on real datasets across
639 sequencing platforms and binning modes. *Nat Commun* **16**, 2865 (2025).
640 <https://doi.org/10.1038/s41467-025-57957-6>

641 22 León, M. J. *et al.* From metagenomics to pure culture: isolation and characterization of the
642 moderately halophilic bacterium *Spiribacter salinus* gen. nov., sp. nov. *Appl Environ Microbiol*
643 **80**, 3850-3857 (2014). <https://doi.org/10.1128/aem.00430-14>

644 23 Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for
645 rumen microbiome biology and enzyme discovery. *Nat Biotechnol* **37**, 953-961 (2019).
646 <https://doi.org/10.1038/s41587-019-0202-3>

647 24 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-
648 504 (2019). <https://doi.org/10.1038/s41586-019-0965-1>

649 25 Li, Y. *et al.* Comprehensive human respiratory genome catalogue underlies the high
650 resolution and precision of the respiratory microbiome. *Briefings in Bioinformatics* **26** (2024).
651 <https://doi.org/10.1093/bib/bbae620>

652 26 Causes of severe pneumonia requiring hospital admission in children without HIV infection
653 from Africa and Asia: the PERCH multi-country case-control study. *Lancet* **394**, 757-779
654 (2019). [https://doi.org/10.1016/s0140-6736\(19\)30721-4](https://doi.org/10.1016/s0140-6736(19)30721-4)

655 27 Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a
656 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725-
657 731 (2017). <https://doi.org/10.1038/nbt.3893>

658 28 Wahl, B. *et al.* Burden of *Streptococcus pneumoniae* and *Haemophilus influenzae* type b
659 disease in children in the era of conjugate vaccines: global, regional, and national estimates
660 for 2000-15. *Lancet Glob Health* **6**, e744-e757 (2018). [https://doi.org/10.1016/s2214-109x\(18\)30247-x](https://doi.org/10.1016/s2214-109x(18)30247-x)

661 29 Zaman, S. M. *et al.* Impact of routine vaccination against *Haemophilus influenzae* type b in
662 The Gambia: 20 years after its introduction. *J Glob Health* **10**, 010416 (2020).
663 <https://doi.org/10.7189/jogh.10.010416>

664 30 Mackenzie, G. A. *et al.* A cluster-randomised, non-inferiority trial of the impact of a two-dose
665 compared to three-dose schedule of pneumococcal conjugate vaccination in rural Gambia:
666 the PVS trial. *Trials* **23**, 71 (2022). <https://doi.org/10.1186/s13063-021-05964-5>

667 31 Howie, S. R. *et al.* Etiology of severe childhood pneumonia in the Gambia, West Africa,
668 determined by conventional and molecular microbiological analyses of lung and pleural
669 aspirate samples. *Clin Infect Dis* **59**, 682-685 (2014). <https://doi.org/10.1093/cid/ciu384>

670

671 32 Hill, P. C. *et al.* Nasopharyngeal carriage of *Streptococcus pneumoniae* in Gambian villagers.
672 *Clin Infect Dis* **43**, 673-679 (2006). <https://doi.org/10.1086/506941>

673 33 Kwambana-Adams, B. *et al.* Rapid replacement by non-vaccine pneumococcal serotypes may
674 mitigate the impact of the pneumococcal conjugate vaccine on nasopharyngeal bacterial
675 ecology. *Sci Rep* **7**, 8127 (2017). <https://doi.org/10.1038/s41598-017-08717-0>

676 34 Ruiz García, Y. *et al.* CIRCULATING CLONAL COMPLEXES AND SEQUENCE TYPES OF
677 *STREPTOCOCCUS PNEUMONIAE* SEROTYPE 19A WORLDWIDE: THE IMPORTANCE OF
678 MULTIDRUG RESISTANCE: A SYSTEMATIC LITERATURE REVIEW. *Expert Rev Vaccines* **20**, 45-57
679 (2021). <https://doi.org/10.1080/14760584.2021.1873136>

680 35 Roca, A. *et al.* Nasopharyngeal carriage of pneumococci four years after community-wide
681 vaccination with PCV-7 in The Gambia: long-term evaluation of a cluster randomized trial.
682 *PLoS One* **8**, e72198 (2013). <https://doi.org/10.1371/journal.pone.0072198>

683 36 Li, T. *et al.* Pan-Genome-Wide Association Study of Serotype 19A *Pneumococci* Identifies
684 Disease-Associated Genes. *Microbiol Spectr* **11**, e0407322 (2023).
685 <https://doi.org/10.1128/spectrum.04073-22>

686 37 Slack, M. *et al.* *Haemophilus influenzae* type b disease in the era of conjugate vaccines:
687 critical factors for successful eradication. *Expert Rev Vaccines* **19**, 903-917 (2020).
688 <https://doi.org/10.1080/14760584.2020.1825948>

689 38 Yang, Y. *et al.* *Haemophilus influenzae* type b carriage and burden of its related diseases in
690 Chinese children: Systematic review and meta-analysis. *Vaccine* **35**, 6275-6282 (2017).
691 <https://doi.org/10.1016/j.vaccine.2017.09.057>

692 39 Khattak, Z. E. & Anjum, F. in *StatPearls* (StatPearls Publishing

693 Copyright © 2025, StatPearls Publishing LLC., 2025).

694 40 Teo, S. M. *et al.* The infant nasopharyngeal microbiome impacts severity of lower respiratory
695 infection and risk of asthma development. *Cell Host Microbe* **17**, 704-715 (2015).
696 <https://doi.org/10.1016/j.chom.2015.03.008>

697 41 Whittle, E., Leonard, M. O., Harrison, R., Gant, T. W. & Tonge, D. P. Multi-Method
698 Characterization of the Human Circulating Microbiome. *Front Microbiol* **9**, 3266 (2018).
699 <https://doi.org/10.3389/fmicb.2018.03266>

700 42 Castillo, D. J., Rifkin, R. F., Cowan, D. A. & Potgieter, M. The Healthy Human Blood
701 Microbiome: Fact or Fiction? *Front Cell Infect Microbiol* **9**, 148 (2019).
702 <https://doi.org/10.3389/fcimb.2019.00148>

703 43 Schierwagen, R. *et al.* Circulating microbiome in blood of different circulatory compartments.
704 *Gut* **68**, 578-580 (2019). <https://doi.org/10.1136/gutjnl-2018-316227>

705 44 Potgieter, M., Bester, J., Kell, D. B. & Pretorius, E. The dormant blood microbiome in chronic,
706 inflammatory diseases. *FEMS Microbiol Rev* **39**, 567-591 (2015).
707 <https://doi.org/10.1093/femsre/fuv013>

708 45 Amar, J. *et al.* Blood microbiota dysbiosis is associated with the onset of cardiovascular
709 events in a large general population: the D.E.S.I.R. study. *PLoS One* **8**, e54461 (2013).
710 <https://doi.org/10.1371/journal.pone.0054461>

711 46 Liu, W. *et al.* Characterization of blood and urine microbiome temporal variability in patients
712 with acute myeloid leukemia. *Microb Pathog* **206**, 107734 (2025).
713 <https://doi.org/10.1016/j.micpath.2025.107734>

714 47 Khan, I. *et al.* Analysis of the blood bacterial composition of patients with acute coronary
715 syndrome and chronic coronary syndrome. *Front Cell Infect Microbiol* **12**, 943808 (2022).
716 <https://doi.org/10.3389/fcimb.2022.943808>

717 48 Barbeta, E. *et al.* Biological effects of pulmonary, blood and gut microbiome alterations in
718 patients with acute respiratory distress syndrome. *ERJ Open Res* **11** (2025).
719 <https://doi.org/10.1183/23120541.00667-2024>

- 49 Riesco, R. & Trujillo, M. E. Update on the proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* **74** (2024).
<https://doi.org/10.1099/ijsem.0.006300>
- 50 Ramasamy, D. *et al.* A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* **64**, 384-391 (2014).
<https://doi.org/10.1099/ijms.0.057091-0>
- 51 Zheng, J. *et al.* A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *Int J Syst Evol Microbiol* **70**, 2782-2858 (2020).
<https://doi.org/10.1099/ijsem.0.004107>
- 52 Ksiezarek, M., Grosso, F., Ribeiro, T. G. & Peixe, L. Genomic diversity of genus *Limosilactobacillus*. *Microb Genom* **8** (2022). <https://doi.org/10.1099/mgen.0.000847>
- 53 Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol* **87** (2021).
<https://doi.org/10.1128/aem.02593-20>
- 54 Valente, C. *et al.* Impact of the 13-valent pneumococcal conjugate vaccine on *Streptococcus pneumoniae* multiple serotype carriage. *Vaccine* **34**, 4072-4078 (2016).
<https://doi.org/10.1016/j.vaccine.2016.06.017>
- 55 Kamng'ona, A. W. *et al.* High multiple carriage and emergence of *Streptococcus pneumoniae* vaccine serotype variants in Malawian children. *BMC Infect Dis* **15**, 234 (2015).
<https://doi.org/10.1186/s12879-015-0980-2>
- 56 Manenzhe, R. I. *et al.* Characterization of Pneumococcal Colonization Dynamics and Antimicrobial Resistance Using Shotgun Metagenomic Sequencing in Intensively Sampled South African Infants. *Front Public Health* **8**, 543898 (2020).
<https://doi.org/10.3389/fpubh.2020.543898>
- 57 Turner, P. *et al.* Improved detection of nasopharyngeal cocolonization by multiple pneumococcal serotypes by use of latex agglutination or molecular serotyping by microarray. *J Clin Microbiol* **49**, 1784-1789 (2011). <https://doi.org/10.1128/jcm.00157-11>
- 58 Boncoeur, E. *et al.* PatA and PatB form a functional heterodimeric ABC multidrug efflux transporter responsible for the resistance of *Streptococcus pneumoniae* to fluoroquinolones. *Biochemistry* **51**, 7755-7765 (2012). <https://doi.org/10.1021/bi300762p>
- 59 Garvey, M. I., Baylay, A. J., Wong, R. L. & Piddock, L. J. Overexpression of patA and patB, which encode ABC transporters, is associated with fluoroquinolone resistance in clinical isolates of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **55**, 190-196 (2011).
<https://doi.org/10.1128/aac.00672-10>
- 60 Lupien, A. *et al.* Genomic characterization of ciprofloxacin resistance in a laboratory-derived mutant and a clinical isolate of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **57**, 4911-4919 (2013). <https://doi.org/10.1128/aac.00418-13>
- 61 Stogios, P. J. *et al.* Substrate Recognition by a Colistin Resistance Enzyme from *Moraxella catarrhalis*. *ACS Chem Biol* **13**, 1322-1332 (2018).
<https://doi.org/10.1021/acscchembio.8b00116>
- 62 Lin, J., Wang, Y., Lin, C., Li, R. & Wang, G. High Prevalence of Group III-Like Mutations Among BLPACR and First Report of *Haemophilus influenzae* ST95 Isolated from Blood in China. *Infect Drug Resist* **16**, 999-1008 (2023). <https://doi.org/10.2147/idr.S400207>
- 63 Hood, D. W. *et al.* Three genes, lgtF, lic2C and lpsA, have a primary role in determining the pattern of oligosaccharide extension from the inner core of *Haemophilus influenzae* LPS. *Microbiology (Reading)* **150**, 2089-2097 (2004). <https://doi.org/10.1099/mic.0.26912-0>
- 64 Cardoso, B. *et al.* Genomic insights of international clones of *Haemophilus influenzae* causing invasive infections in vaccinated and unvaccinated infants. *Microb Pathog* **150**, 104644 (2021). <https://doi.org/10.1016/j.micpath.2020.104644>

771 65 Cho, H. E. *et al.* Shotgun Metagenomic Sequencing Analysis as a Diagnostic Strategy for
772 Patients with Lower Respiratory Tract Infections. *Microorganisms* **13** (2025).
773 <https://doi.org:10.3390/microorganisms13061338>

774 66 Scott, J. A. *et al.* The definition of pneumonia, the assessment of severity, and clinical
775 standardization in the Pneumonia Etiology Research for Child Health study. *Clin Infect Dis* **54**
776 **Suppl 2**, S109-116 (2012). <https://doi.org:10.1093/cid/cir1065>

777 67 QIAGEN. *DNeasy® PowerSoil® Kit Handbook*,
778 <[https://www.qiagen.com/us/resources/resourcedetail?id=5a0517a7-711d-4085-8a28-](https://www.qiagen.com/us/resources/resourcedetail?id=5a0517a7-711d-4085-8a28-2bb25fab828a&lang=en)
779 [2bb25fab828a&lang=en](https://www.qiagen.com/us/resources/resourcedetail?id=5a0517a7-711d-4085-8a28-2bb25fab828a&lang=en)> (2017).

780 68 QIAGEN. *QIAamp® MinElute® Virus Spin Handbook*,
781 <[https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-](https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/multianalyte-and-virus/qiaamp-minelute-virus-kits)
782 [purification/multianalyte-and-virus/qiaamp-minelute-virus-kits](https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/multianalyte-and-virus/qiaamp-minelute-virus-kits)> (2020).

783 69 QIAGEN. *QIAamp DNA Mini and Blood Mini Handbook*,
784 <[https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-](https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/genomic-dna/qiaamp-dna-blood-kits)
785 [purification/dna-purification/genomic-dna/qiaamp-dna-blood-kits](https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/genomic-dna/qiaamp-dna-blood-kits)> (March 2024).

786 70 Espinoza, J. L. & Dupont, C. L. VEBA: a modular end-to-end suite for in silico recovery,
787 clustering, and analysis of prokaryotic, microeukaryotic, and viral genomes from
788 metagenomes. *BMC Bioinformatics* **23**, 419 (2022). [https://doi.org:10.1186/s12859-022-](https://doi.org:10.1186/s12859-022-04973-8)
789 [04973-8](https://doi.org:10.1186/s12859-022-04973-8)

790 71 Espinoza, J. L. *et al.* Unveiling the microbial realm with VEBA 2.0: a modular bioinformatics
791 suite for end-to-end genome-resolved prokaryotic, (micro)eukaryotic and viral multi-omics
792 from either short- or long-read sequencing. *Nucleic Acids Res* **52**, e63 (2024).
793 <https://doi.org:10.1093/nar/gkae528>

794 72 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
795 *Bioinformatics* **34**, i884-i890 (2018). <https://doi.org:10.1093/bioinformatics/bty560>

796 73 Bushnell, B. *BBMap: BBMap short read aligner, and other bioinformatic tools.*,
797 <<https://sourceforge.net/projects/bbmap/>> (2023).

798 74 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**,
799 357-359 (2012). <https://doi.org:10.1038/nmeth.1923>

800 75 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
801 metagenomic assembler. *Genome Res* **27**, 824-834 (2017).
802 <https://doi.org:10.1101/gr.213959.116>

803 76 Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
804 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
805 <https://doi.org:10.7717/peerj.7359>

806 77 Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods*
807 **11**, 1144-1146 (2014). <https://doi.org:10.1038/nmeth.3103>

808 78 Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to
809 recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605-607 (2016).
810 <https://doi.org:10.1093/bioinformatics/btv638>

811 79 Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication,
812 aggregation and scoring strategy. *Nat Microbiol* **3**, 836-843 (2018).
813 <https://doi.org:10.1038/s41564-018-0171-1>

814 80 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing
815 the quality of microbial genomes recovered from isolates, single cells, and metagenomes.
816 *Genome Res* **25**, 1043-1055 (2015). <https://doi.org:10.1101/gr.186072.114>

817 81 Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
818 Manipulation. *PLoS One* **11**, e0163962 (2016). <https://doi.org:10.1371/journal.pone.0163962>

819 82 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify
820 genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925-1927 (2019).
821 <https://doi.org:10.1093/bioinformatics/btz848>

- 83 Sheppard, C. L. *et al.* PneumoKITy: A fast, flexible, specific, and sensitive tool for
Streptococcus pneumoniae serotype screening and mixed serotype detection from genome
sequence data. *Microb Genom* **8** (2022). <https://doi.org:10.1099/mgen.0.000904>
- 84 Epping, L. *et al.* SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae
from whole genome sequence data. *Microb Genom* **4** (2018).
<https://doi.org:10.1099/mgen.0.000186>
- 85 Watts, S. C. & Holt, K. E. hicap: In Silico Serotyping of the Haemophilus influenzae Capsule
Locus. *J Clin Microbiol* **57** (2019). <https://doi.org:10.1128/jcm.00190-19>
- 86 Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb
software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124 (2018).
<https://doi.org:10.12688/wellcomeopenres.14826.1>
- 87 Nickols, W. A. *et al.* MaAsLin 3: Refining and extending generalized multivariable linear
models for meta-omic association discovery. *bioRxiv* (2024).
<https://doi.org:10.1101/2024.12.13.628459>
- 88 Rylance, J. *et al.* Household air pollution and the lung microbiome of healthy adults in
Malawi: a cross-sectional study. *BMC Microbiol* **16**, 182 (2016).
<https://doi.org:10.1186/s12866-016-0803-7>

Acknowledgements

We acknowledge all participants who contributed samples and their guardians. We also acknowledge the contributions of the laboratory staff at the West Africa Research Platform, Medical Research Council Unit The Gambia at LSHTM, as well as the J. Craig Venter Institute, La Jolla, California

Author contributions

The study was conceived and designed by MPN, BKA, MA and CLD. Data analysis was conducted by DK and JLE. Figures were prepared by DK. DK wrote the initial draft of the Manuscript. MPN, CLD, BKA, MA, IDO and JLE edited and revised the manuscript. All authors participated in the discussion and interpretation of the results.

Data Availability

Shotgun metagenomic sequencing data has been deposited in the NCBI Sequence Read Archive (SRA) associated with BioProject PRJNA727021. The accession numbers of the 71 samples used in the analysis in this paper are detailed in Supplementary Table S3. The 35 high-quality MAGs reported in this paper, together with the analysis scripts and supporting files are made available in Zenodo at <https://zenodo.org/records/22771565>.

855

856

857 **Funding**

858 This work is supported by the U.S. National Institutes of Health (NIH), Office of the Director (OD), the
859 National Institute of Environmental Health Sciences (NIEHS), and the National Human Genome
860 Research Institute (NHGRI) grant number U54HG009824.

861 **Additional information**

862 The authors declare no competing interests

863 **Figure legends**

864 Figure 1: List of specimen type analysed for each of the 37 participants. Rows represent individual
865 participants, ordered by shared sample availability profiles to visually group individuals with identical
866 patterns of missing or available sample.

867 Figure 2: (a) Prevalence of AMR genes in all 35 high-quality MAGs. Bars are coloured according to
868 antibiotic class of each corresponding gene. (b) AMR genes detected in the high-quality MAGs
869 grouped by specimen type. The y-axis of each plot lists all AMR genes detected within that specimen
870 type. The AMR gene names are coloured according to their antibiotic class. MLS: Macrolide-
871 Lincosamide-Streptogramins. Nm: *Neisseria meningitidis*, Hi: *Haemophilus influenzae*, Ls:
872 *Limosilactobacillus Sp.* Ml: *Moraxella lincolni*, Mc: *Moraxella catarrhalis*, Sp: *Streptococcus*
873 *pneumoniae*, Mn: *Moraxella nonliquefaciens*, Hp: *Haemophilus parahaemolyticus*

874 Figure 3: Barplots of the compositional mean relative abundances of bacteria DNA at the phylum (a)
875 and species (b) levels. All six phyla detected in our dataset are shown in the phylum-level plot. The
876 Top 14 bacterial species is shown in the species-level plot. All remaining species were grouped as
877 “Other”. The top 14 species were identified based on the sum of their compositional relative

878 abundances across all specimen type. Specimen type is shown on the y-axis, and the compositional
879 mean relative abundance of each taxonomic group is shown on the x-axis. The number of samples
880 (equal to the number of participating individuals contributing each specimen type) is indicated in
881 parantheses on the bar labels. DNA extraction methods for each specimen type: DNeasy PowerSoil
882 Kit (NPS, NP-OP, IS), QIAamp MinElute Virus Spin Kit (LA), and QIAamp DNA Blood Mini Kit (blood).
883 NPS: nasopharyngeal swab, NP-OP: nasopharyngeal/oropharyngeal swab, IS: induced sputum, LA:
884 lung aspirate

885 Figure 4: Shannon diversity of bacterial communities by specimen type. The horizontal line within
886 each box represents the median Shannon diversity, while the upper and lower bounds of each
887 boxplot represent the 75% and 25% quartiles. Pairwise comparisons were performed using a linear
888 mixed-effects model with participant as a random intercept, followed by Tukey's post-hoc adjustment
889 for multiple comparisons. Significant differences between specimen types are indicated by asterisks
890 (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). The number of samples (equal to the number of participating
891 individuals contributing each specimen type) is indicated in parentheses on the x-axis labels. DNA
892 extraction methods: DNeasy PowerSoil Kit (NPS, NP-OP, IS), QIAamp MinElute Virus Spin Kit (LA), and
893 QIAamp DNA Blood Mini Kit (blood). NPS: nasopharyngeal swab, NP-OP:
894 nasopharyngeal/oropharyngeal swab, IS: induced sputum, LA: lung aspirate.

895 Figure 5: Mean Jaccard similarity index (%) of bacterial species detected in paired blood and
896 respiratory specimens from the same participants. Error bars represent standard error of the mean.
897 Number of participant pairs is shown in parentheses on the x-axis. Asterisk indicates a permutation
898 test $p < 0.05$ comparing observed within-participant Jaccard similarity against a null distribution
899 generated from 9,999 random reassignments of paired samples. "ns" indicates $p > 0.05$.

900 Figure 6: Barplots of the compositional mean relative abundances of bacterial DNA at the species
901 level in the nasopharynx (NPS) of severe and very severe cases. Case type is shown on the Y-axis, and
902 the relative abundance of each species is shown on the X-axis. Taxa that are not in the top 14 most

903 abundant species are aggregated under "Other". The number of samples included in each category is
 904 added in the bar labels.

905

906 Table 1: Demographic and clinical characteristics of children. This table includes all children who
 907 contributed at least one specimen to any analysis.

	Classification	Frequency (%)
Sex	Male	22 (59)
	Female	15 (41)
Age	1-5 months	15 (41)
	6-11 months	4 (11)
	12-23 months	8 (22)
	24-59 months	10 (27)
Hib vaccination status	Not vaccinated	6 (16)
	Partially vaccinated	9 (24)
	Fully vaccinated	20 (54)
	No data	2 (5)
PCV vaccination status	Not vaccinated	8 (22)
	Partially vaccinated	8 (22)
	Fully vaccinated	19 (51)
	No data	2 (5)
Pneumonia severity	Severe	16 (43)
	Very severe	21 (57)
HIV status	Positive	2 (5)
	Negative	31 (84)

	Unconfirmed	4 (11)
--	-------------	--------

908

909

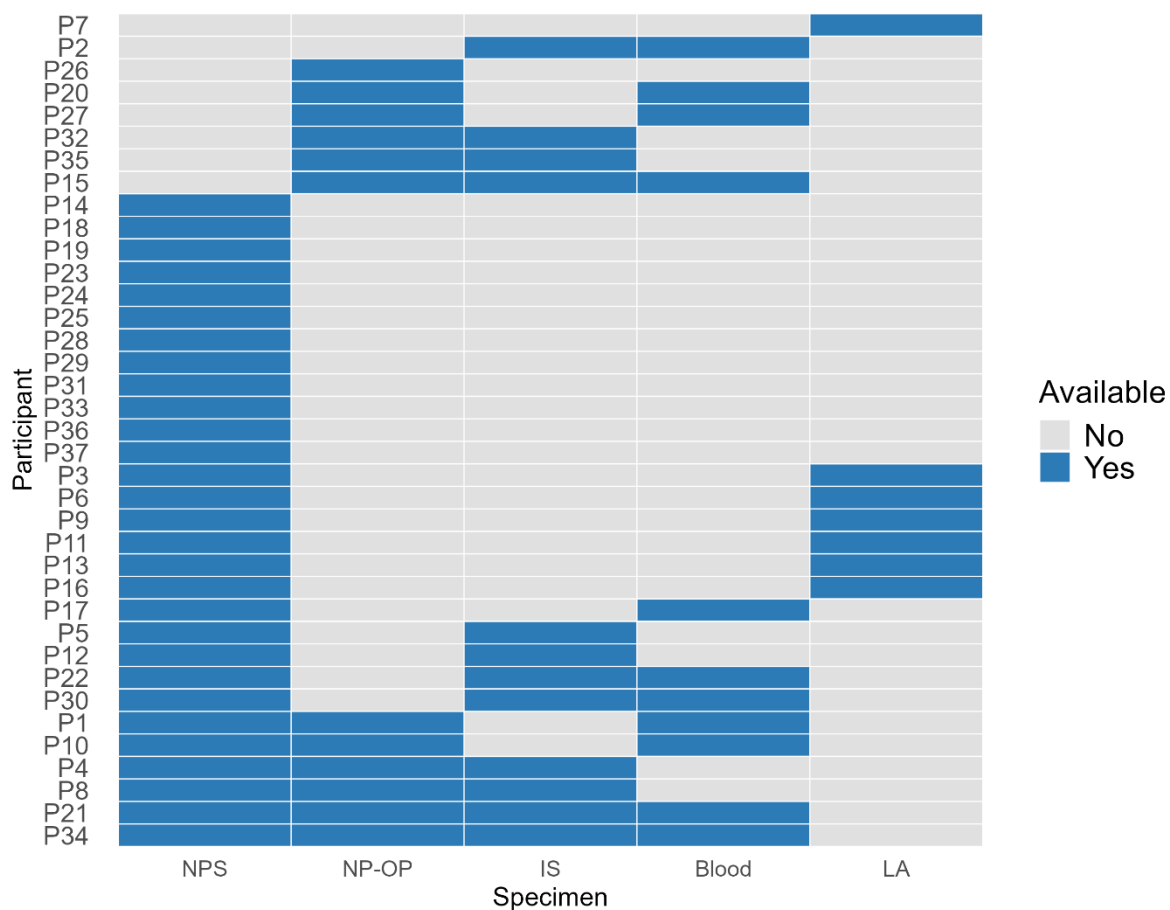
910 Table 2: Characterisation of high-quality metagenome-assembled genomes obtained from blood and

911 respiratory samples

	Species	ANI with type strain (%)	Number of contigs	N50	Genome Size (Mb)	Completeness (%)	Contamination (%)	Strain heterogeneity	Source specimen
MAG1	<i>Neisseria meningitidis</i>	97.58	227	11345	1.80	95.79	0.23	0	NP-OP
MAG2	<i>Haemophilus parahaemolyticus</i>	97.3	146	17420	1.78	90.94	1.24	66.67	IS
MAG3	<i>Moraxella catarrhalis</i>	99.12	273	8974	1.75	92.31	1.38	22.22	IS
MAG4	<i>Streptococcus pneumoniae</i>	98.53	114	26698	2.00	98.5	0	0	LA
MAG5	<i>Moraxella catarrhalis</i>	99.08	69	38190	1.77	95.76	0.42	33.33	NP-OP
MAG6	<i>Streptococcus pneumoniae</i>	98.62	69	44339	2.16	99.63	0.06	0	NP-OP
MAG7	<i>Haemophilus influenzae</i>	97.17	96	50042	1.92	99.37	0.25	50	IS
MAG8	<i>Haemophilus parahaemolyticus</i>	97.24	265	9977	1.83	90.75	1.05	28.57	IS
MAG9	<i>Streptococcus pneumoniae</i>	98.63	244	11618	1.80	93.6	0.41	20	NPS
MAG10	<i>Haemophilus influenzae</i>	97.14	56	93732	1.84	99.55	0.76	16.67	LA
MAG11	<i>Streptococcus pneumoniae</i>	99.52	76	67168	2.02	99.63	0	0	LA
MAG12	<i>Moraxella catarrhalis</i>	96.14	175	28060	2.00	97.56	3.54	10.53	LA
MAG13	<i>Haemophilus influenzae</i>	97.49	247	11400	1.83	92.84	3.45	56.52	LA
MAG14	<i>Streptococcus pneumoniae</i>	98.6	88	32376	2.05	99.63	0	0	Blood
MAG15	<i>Haemophilus influenzae</i>	97.39	167	20433	1.63	91.37	0.8	100	NP-OP
MAG16	<i>Streptococcus pneumoniae</i>	98.57	87	39492	2.07	99.44	0	0	LA
MAG17	<i>Haemophilus influenzae</i>	97.1	60	71164	1.80	98.98	0	0	LA
MAG18	<i>Haemophilus influenzae</i>	97.38	29	122554	1.79	99.55	0	0	LA
MAG19	<i>Haemophilus influenzae</i>	98.8	22	137938	1.85	99.66	0.23	0	Blood
MAG20	<i>Streptococcus pneumoniae</i>	98.45	197	16266	2.01	97.44	1.07	0	Blood
MAG21	<i>Moraxella catarrhalis</i>	99.23	139	28747	1.94	97.29	0.63	75	Blood
MAG22	<i>Moraxella catarrhalis</i>	99.12	26	120380	1.85	99.45	1.07	14.29	Blood
MAG23	<i>Streptococcus pneumoniae</i>	98.65	62	45294	2.08	99.63	0	0	Blood

MAG24	<i>Haemophilus influenzae</i>	97.11	26	186272	1.86	99.55	0	0	Blood
MAG25	<i>Limosilactobacillus Sp.</i>		228	8961	1.50	97.81	0.13	0	NP-OP
MAG26	<i>Neisseria meningitidis</i>	97.61	158	18553	1.97	99.18	0.38	0	Blood
MAG27	<i>Streptococcus pneumoniae</i>	98.56	64	48015	2.12	99.63	0	0	Blood
MAG28	<i>Moraxella nonliquefaciens</i>	98.47	56	343360	2.28	96.85	3.48	0	Blood
MAG29	<i>Haemophilus influenzae</i>	97.21	157	29106	1.93	95.55	1.05	28.57	Blood
MAG30	<i>Moraxella catarrhalis</i>	99.08	65	74719	1.98	98.88	0.44	25	NPS
MAG31	<i>Haemophilus influenzae</i>	97.28	156	19152	1.75	98.05	0.84	40	NPS
MAG32	<i>Streptococcus pneumoniae</i>	98.59	339	6926	1.83	91.90	1.37	83.33	NPS
MAG33	<i>Moraxella lincolni</i>	97.95	243	13659	2.07	92.60	0.5	66.67	NP-OP
MAG34	<i>Streptococcus pneumoniae</i>	98.86	238	10781	1.90	94.76	0.75	50	NPS
MAG35	<i>Moraxella nonliquefaciens</i>	98.22	103	57578	2.34	95.75	4.84	6.06	NPS

912



913

914 Figure 1

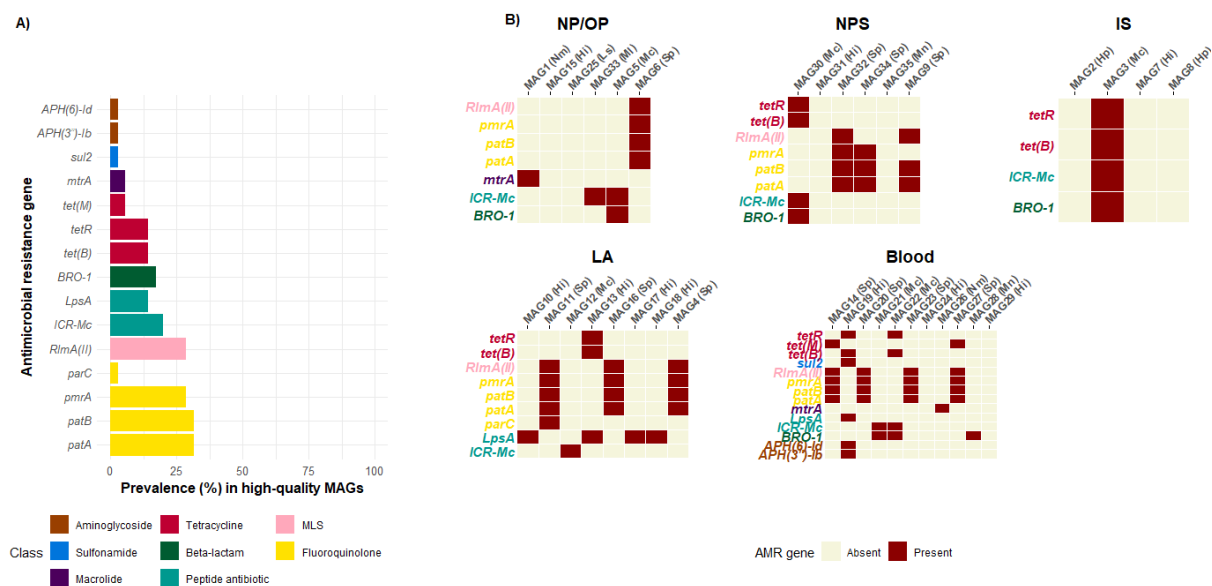


Figure 2

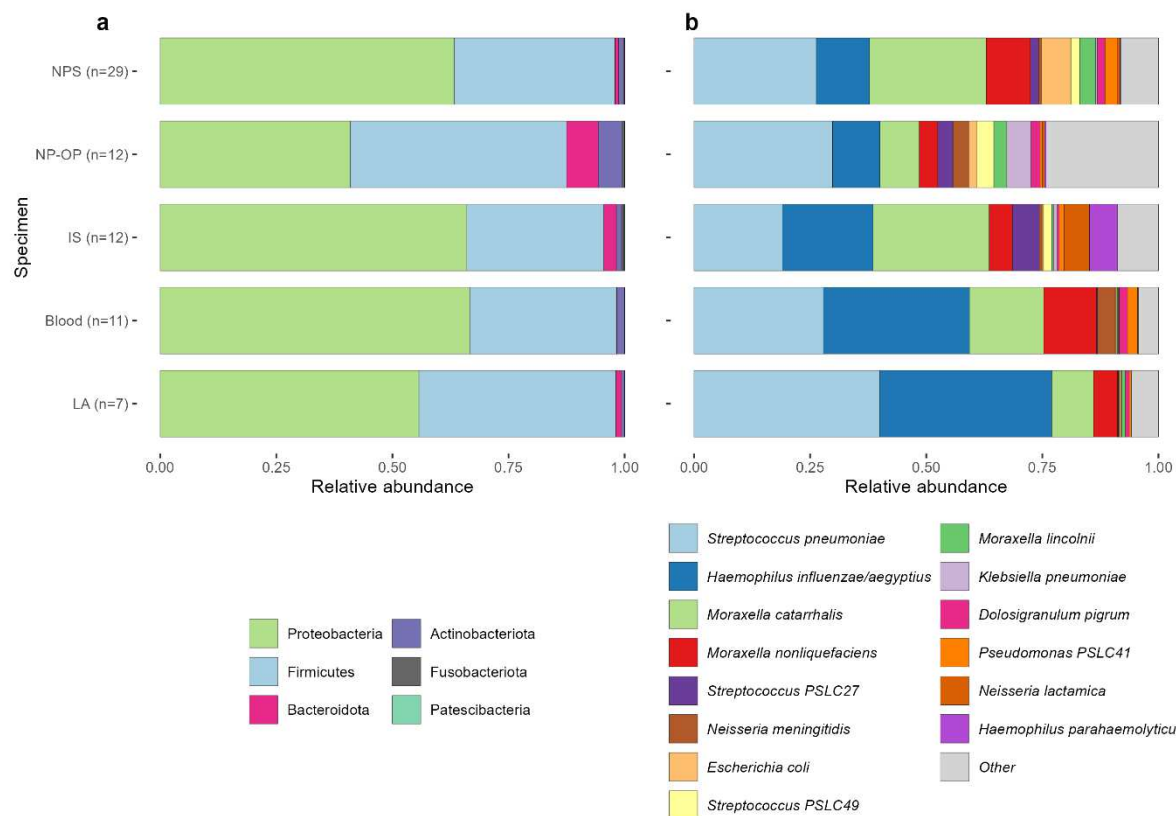
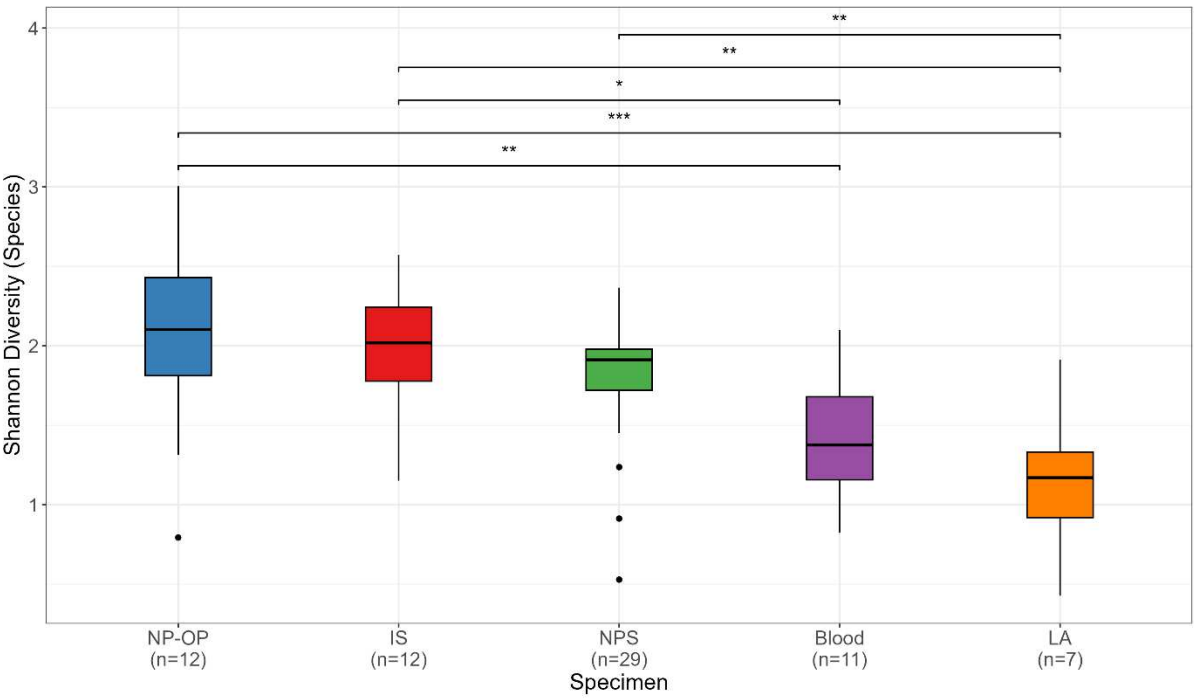


Figure 3

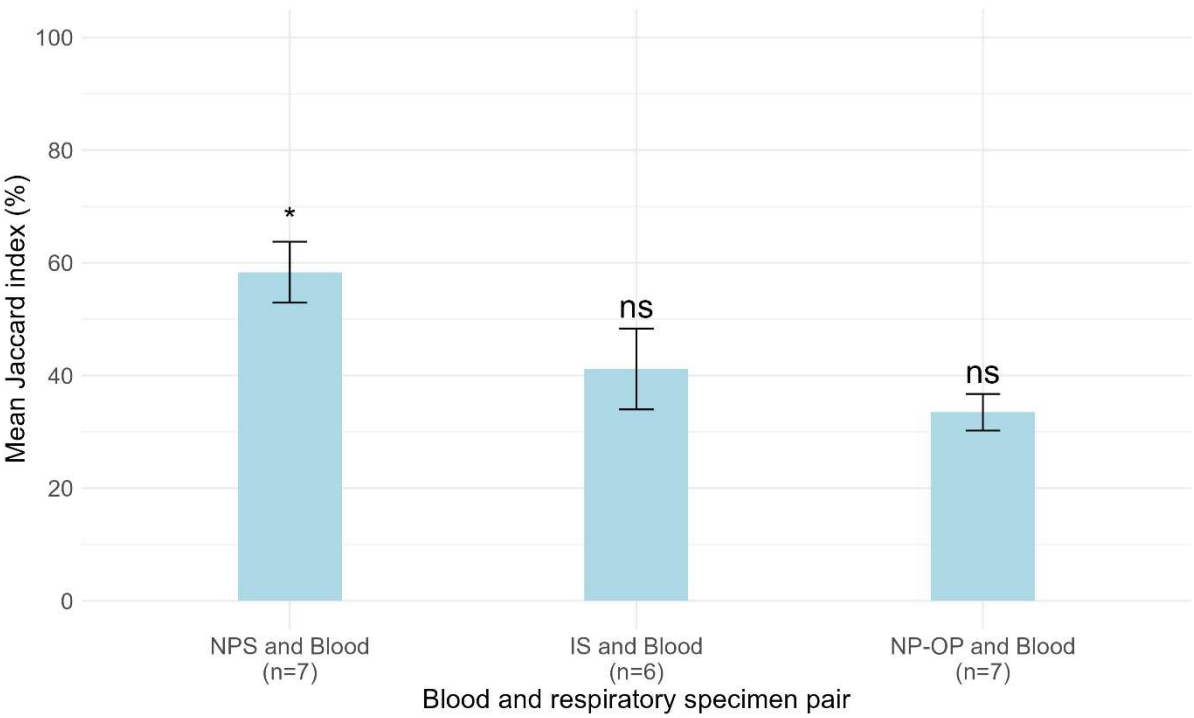
920



921

922 Figure 4

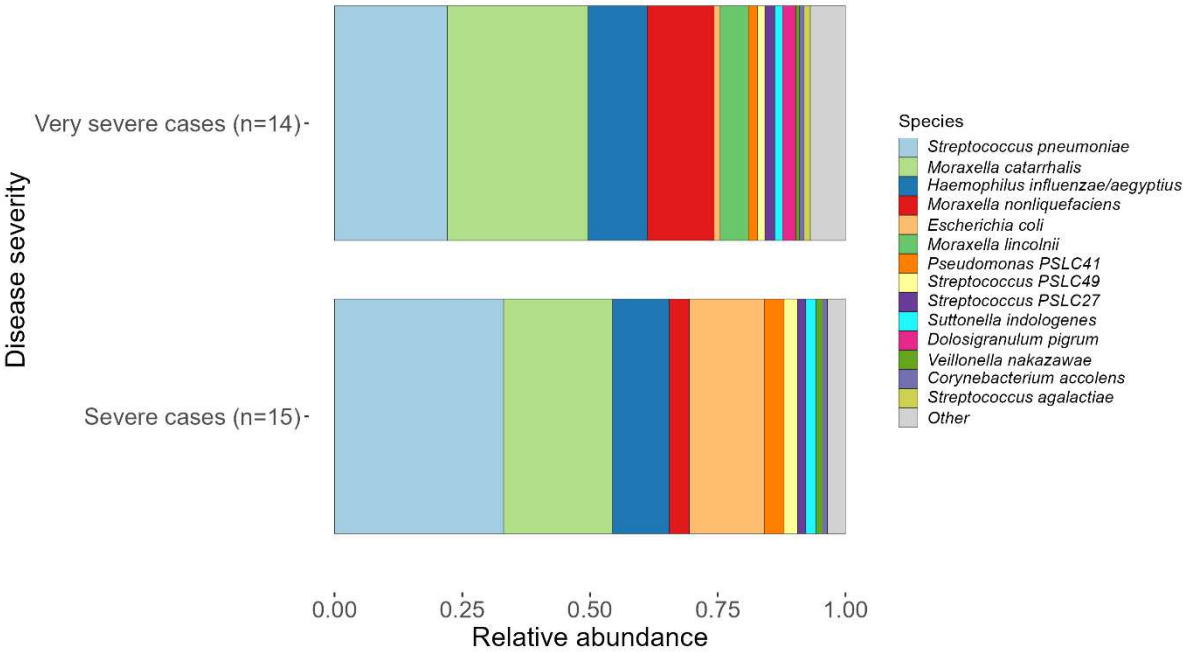
923



924

925 Figure 5

926



927

928 Figure 6

929

930

931

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplemetaryinformationclean.pdf](#)