Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The healthy pediatric oropharyngeal colonization landscape: Mycoplasma pneumoniae as a near-specific disease signal at shotgun depth and load-stratified DNA viruses in pneumonia

Chen Shen

shenchen1110@126.com

Beijing Children's Hospital

Research Article

Keywords: Mycoplasma pneumoniae, oropharyngeal colonization, healthy carriage landscape, shotgun metagenomics, P1 typing, adenovirus, pediatric pneumonia

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11702120/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

The healthy pediatric oropharyngeal colonization landscape: *Mycoplasma pneumoniae* as a near-specific disease signal at shotgun depth and load-stratified DNA viruses in pneumonia

Running title: Pediatric oropharyngeal colonization landscape

Authors: Chen Shen[a]*

[a] National Children's Medical Center, Beijing Children's Hospital, Capital Medical University, Beijing, China

† Deze Li, Hao Chen and Xiaotong Wang contributed equally to this work (co-first authors).

Corresponding author: Chen Shen, National Children's Medical Center, Beijing Children's Hospital, Capital Medical University, Nanlishi Road 56, Xicheng District, Beijing 100045, China. E-mail: shenchen1110@126.com

Abstract

Interpreting a pediatric pneumonia metagenome requires knowing which organisms colonize the healthy oropharynx. We re-screened 246 pediatric oropharyngeal shotgun metagenomes from a Shenzhen cohort (171 healthy, 75 with pneumonia) and 1,803 runs from twelve external pediatric cohorts using read-level, dual-verification mapping with an a priori ≥2-read positivity rule. The healthy landscape was streptococcus-dominated (*Streptococcus pneumoniae*–mitis complex 99.4%; *Haemophilus influenzae* 93.0%). Against this backdrop, *Mycoplasma pneumoniae* (MP) broke the colonization pattern as a near-specific disease signal: 29/51 (56.9%) oropharyngeal samples in MP pneumonia versus 1/170 (0.6%) in healthy children (Fisher $p=2.6\times10^{-21}$; odds ratio 214, exact 95% CI 32–9,147) and 4/24 (16.7%) in non-MP pneumonia (0/24 at ≥5 reads), and 0/802 quality-control-passed external carriage samples (95% CI 0–0.5%) — no detectable healthy carriage at shotgun depth. High-load samples were P1-1 and CARDS-toxin positive; the highest carried 23S A2063G in all 16 site-covering reads. Adenovirus was enriched in MP pneumonia (66.7% vs 1.2% healthy; most detection reflects oropharyngeal shedding), with no healthy detection above relative abundance 10^{-4}. At 56.9% sensitivity and 99.4% specificity, oropharyngeal detection is a strong rule-in signal (positive predictive value 97.6% at 30% prevalence) but cannot rule out disease. Read-level screening calibrated against healthy carriage distinguishes infection from background in public metagenomic archives.

Importance

Metagenomic sequencing is entering routine diagnosis of childhood pneumonia, but “detected” is too often misread as “causing disease”, because many microbes colonize the healthy airway. We mapped the healthy pediatric oropharyngeal colonization landscape and then asked which detections break it. Streptococci and *Haemophilus influenzae* were near-universal healthy colonizers, but *Mycoplasma pneumoniae* was not: across 802 community and carriage samples from healthy children on three continents, we found no detectable carriage at sequencing depth. At that depth, a metagenomic report of this organism in a sick child’s airway sample strongly supports infection over background carriage. The same colonization-calibrated logic separates clinically meaningful adenovirus detection from harmless trace shedding: only high-level detection signals meaningful virus. Because oropharyngeal sampling underestimates lower-airway infection, detection rules in but non-detection cannot rule out.

Keywords: *Mycoplasma pneumoniae*; oropharyngeal colonization; healthy carriage landscape; shotgun metagenomics; P1 typing; adenovirus; pediatric pneumonia

Introduction

The healthy pediatric oropharynx is a densely colonized habitat in which streptococci, *Haemophilus* and other commensals establish stable carriage early in life. As shotgun metagenomic sequencing enters routine diagnosis of childhood pneumonia, oropharyngeal swabs — the most accessible respiratory specimen — return a dense mixture of true pathogens, stable commensals and transient colonizers, and in this setting “detected” is too often misread as “causing disease”. The clinical meaning of any single detection depends on the colonization baseline against which it is read: which organisms healthy children carry, at what prevalence and at what load. Mapping this landscape — who colonizes, and how often — is therefore a prerequisite for interpreting any pneumonia metagenome.

That baseline is poorly mapped. No global benchmark describes the healthy pediatric oropharyngeal colonization landscape at read-level resolution, and the methodological landscape complicates every estimate. Standard taxonomic pipelines apply arbitrary relative-abundance thresholds under which low-load signals disappear: in the cohort analyzed here, the processed relative-abundance table reported species-level *Mycoplasma pneumoniae* above its 0.1% reporting threshold in only 6.1% of samples, and 19 of the 34 read-level positives fell below that threshold. A direct read-level mapping approach, with explicit controls for cross-mapping to related organisms and the rRNA operon, avoids both under-detection of low-load organisms and spurious calls; public shotgun archives are an underused resource for building such a colonization-calibrated landscape retrospectively.

Mycoplasma pneumoniae (MP; reclassified as *Mycoplasmoides pneumoniae* (1), a reclassification not universally adopted) is the sharpest test of any such landscape, because whether it colonizes the healthy upper respiratory tract at all remains debated. MP is a leading cause of community-acquired pneumonia in school-age children and young adults, responsible for an estimated 10–40% of cases in pediatric series, with epidemics recurring every 3–7 years (2–4), and a marked global resurgence of MP respiratory disease was documented from 2023 onward after the COVID-19 pandemic period of suppressed circulation (5–6). Yet the nature and frequency of asymptomatic upper-airway carriage remain unresolved. Culture- and PCR-based studies generally found low MP detection in healthy individuals, yet a widely cited Dutch study reported MP PCR detection in 21.2% of asymptomatic children during an epidemic, sparking debate on persistent carriage versus post-infectious shedding (7–8). Large population-based carriage studies using 16S rRNA profiling have detected the genus *Mycoplasma* at substantial prevalence in healthy oropharynges, but species-level attribution to MP is unreliable (9); MP is wall-less, and its single rRNA operon is poorly amplified by common "universal" primer sets, so 16S-based surveys systematically underestimate MP (10–11). MP pathogenesis and molecular epidemiology hinge on two well-characterized loci: the P1 adhesin, encoded by the *p1* gene (MPN141) and central to cytadherence, defines two major genotypes, P1-1 (exemplified by strain M129) and P1-2 (strain FH), whose prevalence shifts over epidemic cycles, P1-1 dominating recent Chinese epidemics (8, 12–13); and the CARDS (Community-Acquired Respiratory Distress Syndrome) toxin, an ADP-ribosylating and vacuolating toxin encoded by MPN372, is a key virulence determinant of airway injury (14–15). Shotgun re-screening of high-load samples therefore also permits in silico genotyping.

Here we re-analyzed a public shotgun metagenomic cohort of 246 oropharyngeal (OP) swabs from Shenzhen children — 171 healthy children and 75 children hospitalized with pneumonia with BALF-established lower-airway etiology (16) — using a validated, read-level dual-verification pipeline. We built the analysis colonization-first, beginning with the healthy oropharyngeal colonization landscape across 18 respiratory organisms and then asking how MP detection in pneumonia breaks that pattern — (i) how frequently is MP DNA detectable in the OP of healthy children versus children with MP pneumonia (MPP) or non-MPP pneumonia; (ii) how robust are these estimates to detection thresholds, subsampling and external replication; and (iii) what P1 type and CARDS status do high-load OP samples carry? Finally, we extended the same colonization-calibration logic to the DNA virome, where detection in pneumonia must likewise be read against the healthy-carriage baseline and stratified by load.

Materials and Methods

Cohort and group definitions

We used the public pediatric respiratory metagenomic cohort of Dai et al. (BioProject PRJNA413615) (16), consisting of oropharyngeal swabs from 171 healthy children and 75 children with pneumonia in Shenzhen, China, sequenced on Illumina HiSeq (paired-end 150 bp). Children with pneumonia were classified according to BALF pathogen testing reported in the source study: MPP (MP detected in BALF; n=51) and non-MPP (MP not detected in BALF; n=24). The non-MPP group comprised 12 children with a known alternative pathogen (*Streptococcus pneumoniae* 3, *Haemophilus influenzae* 2, respiratory syncytial virus 2, *H. parainfluenzae* 1, *Moraxella catarrhalis* 1, Epstein–Barr virus 1, CMV+RSV+*H. parainfluenzae* 1, RSV+*H. influenzae* 1) and 12 with no pathogen identified ("unknown"). Per-sample ages were recovered for 241/246 children through the source study's sample-identifier mapping: median age was 3.8 years in MPP, 1.67 years in non-MPP, and 4.3 years in healthy children. Group definitions rested on BALF pathogen testing in the source study — a different compartment and method from our OP mNGS re-screen — so this screen does not enter the case definition and no circularity arises. Pneumonia cases were sampled in July–August 2016, whereas healthy children were sampled predominantly in November 2016 (see Limitations). One healthy sample (SRR8369485) yielded zero screened reads and was excluded from the primary MP detection analysis (n=170); it remains in the denominators of the panel, AMR, and virome assays (n=171).

Data access and prefix subsampling

Complete downloads were bandwidth-limited; therefore mate-1 FASTQ prefixes (~3 MB, ~5×10^4 reads per sample) were analyzed for screening. FASTQ records are ordered by flowcell coordinates, so a file prefix is a spatially structured, mate-1-only subsample rather than a random read sample; we therefore describe it as prefix subsampling with empirically validated quantification, not an unbiased random subsample. Organism-level quantification from prefix subsamples was validated against the full-data ResMicroDb pipeline (Spearman r=0.78 across all 246 samples with non-detects as zeros, $p=5.1\times10^{-52}$; r=0.89 among the 72 samples with an abundance-layer MP assignment; Table S1). Per-sample depth was recorded; five samples yielded <10,000 reads, including one with zero reads (SRR8369485), flagged for detection-limit heterogeneity (see Limitations).

Read-level dual-verification mapping pipeline

Reads were mapped with minimap2 (via the mappy Python interface) (17) against the MP M129 reference genome (NC_000912.1). Alignments with MAPQ < 30 or alignment identity < 0.90 were

discarded, and reads mapping within the single rRNA operon (genome coordinates 118,000–123,100) were excluded. To control cross-mapping from related mollicutes, every candidate read was additionally mapped against the *M. genitalium* G37 genome (NC_000908.2) (18); a read was counted as MP-specific only if its alignment identity to M129 exceeded that to G37, which also excluded residual rRNA cross-mapping. Trace-level positives were manually verified: inspected reads showed 97–100% identity to M129, genome-wide scatter, and zero mapping to *M. genitalium*.

Positivity rule and sensitivity analysis

The a priori primary positivity rule was ≥2 MP-specific reads, chosen a priori because single-read calls are unreliable at low biomass and because read-level verification supported ≥2-read calls as genuinely MP-derived. Because conclusions for low-prevalence contrasts can depend on this threshold, we repeated all pairwise comparisons under positivity thresholds of ≥1, ≥2, ≥3, ≥5 and ≥10 reads as a sensitivity analysis.

Multi-pathogen panel

The dual-verification rule was generalized to a four-target panel using the first 25,000 mate-1 reads per sample: MP (control: *M. genitalium*), *S. pneumoniae* (control: *S. mitis*; read-level mapping cannot fully resolve the pneumoniae–mitis complex), *H. influenzae* (control: *H. parainfluenzae*), and human bocavirus (HBoV). The panel assay is shallower than the primary screen (25,000 reads vs the full prefix) and is used only for cross-pathogen comparison; the full-prefix screen is the primary assay for MP prevalence. Accordingly, MP prevalence in MPP is reported as 56.9% (primary screen) and 49.0% (25k panel), the difference reflecting panel depth; positivity thresholds were applied separately for each assay.

P1 typing and CARDS detection

For the highest-load MP-positive samples, reads were mapped across the P1 adhesin locus (MPN141; M129 coordinates 180,858–185,741) and scored by alignment identity against M129 (P1-1) and FH (P1-2) (19). CARDS toxin presence was assessed by reads mapping the MPN372 locus (M129 coordinates 444,341–446,116). SRR8369497 (MPP) was analyzed from the complete mate-1 file (1.97M reads); SRR8369395 and SRR8369523 from 60-MB prefixes.

External cohort screen and data-type verification

Candidate external cohorts were identified from public repositories. Because read-level screening presupposes untargeted shotgun data, the ENA library_strategy/library_selection of every run was verified: runs annotated AMPLICON (PRJDB39284, 100 runs; and 66 runs within PRJNA745160) were

excluded, yielding 1,803 WGS runs from twelve cohorts (Table 2; a thirteenth pediatric oropharyngeal cohort, SRP664982, was screened additionally and failed the biomass QC rule). Each run was screened with the same 3-MB prefix dual-verification rule (MP, S. pneumoniae, H. influenzae, HBoV); for cross-cohort carriage comparisons, S. pneumoniae–mitis and H. influenzae rates are reported at a ≥5-read threshold to suppress cross-mapping noise in the shallower external prefixes. MP-positive runs with sufficient load were genotyped from deeper (40-MB) prefixes: P1 as above; 23S rRNA macrolide-resistance sites A2063/A2064/A2067/C2617 (M129 coordinates 122,118/122,119/122,122/122,672) were called from site-covering reads with relaxed MAPQ, accepted only at read identity ≥0.90 with zero mismatches within ±8 bp flanking the site against M129 (confirming MP origin of each read); a site was called resistant when ≥3 site-covering reads were available and ≥80% carried the mutant base, and samples without site coverage are reported explicitly as "not covered" rather than silently dropped. Because the relaxed-MAPQ 23S assay lies outside the dual-verification layer, its specificity was verified post hoc: all reported calls were unchanged when the flanking rule was hardened from ≤1 to zero mismatches, site-covering reads showed full-length identity distributions of 97.6–100% to M129 with no 90–93% contaminant cluster, and no site read aligned better to the *M. genitalium* G37 23S homolog (Table S17). We additionally screened 32 severe pediatric pneumonia cases from an independent Shenzhen Children's Hospital cohort (PRJNA572371) (20) with the same panel.

Table 2. External validation cohorts (WGS runs only[a]; dual-verification screen, ≥2 MP-specific reads)

Cohort	Country	Design	n	MP+	Prevalence (95% CI)	Positive-control QC
PRJNA876769	China	Pediatric pneumonia BALF mNGS	103	7	6.8% (3.3–13.4%)	na (BALF)
PRJNA745160	USA	Pediatric CAP respiratory (WGS subset)	34	1	2.9% (0.5–14.9%)	pass
PRJNA727021	Gambia	Juvenile respiratory microbiome	100	0	0% (0–3.7%)	pass[c]
PRJEB25223	Thailand	Infant NP longitudinal carriage	128	0	0% (0–2.9%)	pass
PRJEB37312	South Africa	Healthy birth cohort NP	100	0	0% (0–3.7%)	pass
PRJEB71147	South Africa	Drakenstein birth cohort NP	150	0	0% (0–2.5%)	pass
PRJNA1078287	USA	Infant nasal/oral microbiome	250	0	0% (0–1.5%)	pass
PRJNA713420	USA	URT microbiome, influenza season (DNA-WGS)	174	0	0% (0–2.2%)	pass
PRJEB71847	Spain	Bronchiolitis + healthy controls	147	0	—	fail[b] (*S. pneumoniae*/*H. influenzae* ≈ 0–1%)
PRJNA1355199	Cameroon	Influenza-negative children NP	199	0	—	fail[b] (low biomass)
PRJNA1296491	USA	Clinical nasal swab surveillance	250	0	—	fail[b] (low biomass)
SRP664982	USA	PANDAS throat/oral microbiome	13	0	—	fail[b] (low biomass)
SRP503326	China	Pediatric ARI oropharyngeal mNGS (COVID-era, 2021–22)	168	0	0% (0–2.2%)	na (ARI disease cohort)
Pooled QC-passed carriage/community	—	—	**802**	**0**	**0% (0–0.5%)**	—

[a] Amplicon runs (PRJDB39284; 66 runs of PRJNA745160) were excluded before screening. [c] Gambia QC-passed but is a respiratory-disease cohort and is excluded from the pooled carriage/community denominator (802 = Thailand 128 + South Africa 100 + South Africa 150 + USA 250 + USA 174); the disease cohorts PRJNA876769, PRJNA745160 and SRP503326 are likewise not pooled.

[b] Cohorts failing the positive-control QC rule (see Methods) were judged to lack assayable bacterial biomass and were excluded from prevalence evidence.

Validation against published tNGS cohorts

Because this cohort is amplicon-based, it was used exclusively for locus-level genotyping (P1, 23S) at targeted positions, not for prevalence inference. All 132 runs of a published Australian targeted-NGS cohort from the 2023–2024 epidemic (21) and 38 UK 2023–2024 isolates (22) were re-analyzed with the same read-vote genotyping rules described above (P1 identity vote against M129/FH; 23S site calls at read identity ≥0.90 with zero flanking mismatches), and the resulting P1-1/P1-2 distribution and macrolide-resistance rates were compared with the published values. To test whether the 2016-era Shenzhen typing generalizes to the post-COVID Chinese resurgence, we additionally re-analyzed all 169 runs of a Beijing 2023 outbreak whole-genome cohort (PRJNA1149907) (23) with the identical read-vote rules applied to 3-MB mate-1 prefixes; runs without ≥80% single-lineage purity are reported separately as ambiguous rather than forced into either genotype.

Colonizer antimicrobial-resistance gene panel

Prefix subsamples were screened with minimap2 against a nine-reference AMR panel relevant to oropharyngeal colonizers — ermB, mefA, mefE, tetM, tetO, blaTEM-1, blaROB-1, bro-1 and bro-2 (reference accessions in Table S11), assigning each read to its best-identity reference without a MAPQ filter (multi-mapping among homologous resistance-gene families makes MAPQ uninformative here; a ≥1% runner-up identity margin sensitivity re-screen changed no group-level call, and a short-read-preset re-screen reproduced all group-level prevalences; Table S17). The initially used mefA accession proved to be a mef(E)-like sequence mislabeled in the source database; after replacement with canonical alleles (NG_047957.1 versus U83667, ~90% pairwise identity and resolvable by short reads), all mef detections assigned to mef(E) (Table 3 footnote). A gene was called detected at ≥3 reads with ≥90% identity over ≥80% of the read length. This gene-level assay does not assign resistance to a specific host species.

In silico validation of the positivity rule

Synthetic 150-bp reads drawn from the M129 genome were used as spike-ins in screening runs; the ≥2-read rule recovered 100% of spike-ins at ≥2 spiked reads and returned zero detections for 50-read spike-

ins of *M. genitalium*, *M. hominis*, *M. salivarium*, *Ureaplasma parvum* or *S. pneumoniae*, verifying the rule's behavior at its detection boundary and against related oropharyngeal mollicutes (design details and scope in Supplementary Methods). The theoretical 95% detection limit of the ≥2-read rule is a relative abundance of 4.74/N for N screened reads (Poisson λ_{95}=4.74; derivation in Supplementary Methods) — ≈9×10^{-5} at the median prefix depth (~5×10^{4} reads), roughly one MP genome per ~2,600 bacterial genomes.

DNA virome screening

DNA-virus detection used two complementary layers. Species-level detection prevalences were called from the processed ResMicroDb abundance tables of the same cohort, and read-level confirmation was performed against a custom reference panel covering human adenoviruses, WU and KI polyomaviruses, torque-teno viruses, and human herpesviruses (HHV-1–8). A sample was called positive for a given virus at ≥2 specific reads with ≥85% identity. Viral relative abundance was defined as viral reads per screened read, and enrichment analyses were repeated at fixed relative-abundance thresholds (>10^{-5} and >10^{-4}) to separate trace-level from higher-abundance detection. The Streptococcus relative-abundance covariate in the colonizer AMR regression was likewise taken from the ResMicroDb tables.

Adenovirus typing

Adenovirus-assigned reads were re-mapped against a 19-reference panel of human adenovirus complete genomes spanning species A–G (B3 MW748663, B7 MG696128, C5 KX868466, E4 KX868498, A12 KX868289, B14 JX892927, B16 JN860680, B21 KF528688, B55 MK123979, C1 MH183293, C2 MF044052, C6 OP871032, D8 AB746853, D19 JQ326209, D37 AB448777, F41 ON561778, G52 DQ923122, B11/34 KF906413, B35 OR735204; also listed in Table S13); reads were assigned to the highest-identity reference (≥85% identity, ≥1% margin required; hexon-based assignment parameters in Supplementary Methods). High-load samples were typed from 60-MB prefixes.

Community diversity and structure analysis

For community-level analyses, the processed species-level abundance tables of the 241 metadata-mapped samples were used. Species-level Shannon diversity was computed per sample and compared across groups by Kruskal–Wallis test, with Mann–Whitney U for the within-MPP contrast (read-level MP-positive versus MP-negative). Species were aggregated to genus level for Bray–Curtis dissimilarity, principal-coordinates analysis, and PERMANOVA (999 permutations). Community types were assigned per sample by the top-2 genus sequence (first- and second-ranked genera by relative abundance); the Prevotella-dominant state was compared across groups by two-sided Fisher exact test.

Statistics

Virome abundance thresholds (any detection, $>10^{-5}$, $>10^{-4}$) were fixed a priori to span trace through high-abundance detections and are reported in full. Group prevalences are reported with Wilson 95% confidence intervals. Pairwise comparisons used two-sided Fisher exact tests. Odds ratios are reported as conditional maximum-likelihood estimates with exact conditional 95% CIs (read-level denominators: MPP n=51, non-MPP n=24, healthy n=170); for contrasts containing cells with <5 events we additionally report Haldane–Anscombe-corrected ORs. Age stratification used within-stratum two-sided Fisher exact tests across age bands (<3, 3–6, 6–13 years) after per-sample age recovery (241/246 samples). The Benjamini–Hochberg procedure (24) controlled the false discovery rate within each test family (3 MP contrasts; 12 panel contrasts; 4 DNA-virus-family contrasts, MPP versus healthy), with Holm adjustment as a sensitivity check for the panel. The extended 26-target resistome panel (Table S16) was analyzed descriptively: per-gene group contrasts are reported with unadjusted two-sided Fisher exact p-values as descriptive statistics, and the panel was not included in the BH-controlled test families. For the colonizer AMR panel, logistic regression with Streptococcus relative abundance as covariate tested whether gene enrichment was independent of community composition. Quantitative agreement between the raw-read pipeline and ResMicroDb relative abundance used Spearman correlation. Analyses used Python (scipy, statsmodels).

Results

The healthy oropharyngeal colonization landscape

Interpreting any pathogen signal requires a healthy baseline, and that baseline is itself contingent, not universal. We profiled 171 healthy children — combining the processed species-level abundance tables with the fixed-depth 25,000-read read-level panel — against the two pneumonia groups (Figure 1a, b). Every detection rate below carries four dimensions: sampling era (2016), population (healthy Shenzhen children), specimen type (oropharyngeal swab) and age band. Here the oropharynx was streptococcus-dominated: the *S. pneumoniae*–mitis complex was near-saturated (99.4% in the fixed-depth panel; an upper bound, as read-level mapping cannot fully resolve the complex), and *H. influenzae* was carried by 93.0% in the panel and detected in every healthy sample in the abundance layer — a layer dependence we flag wherever the two disagree (per-organism assay layers in Table S10). Even within this cohort carriage varied with age (<3, 3–6, 6–13-year bands; Figure 1c): *S. pneumoniae*–mitis and *H. influenzae* were universal (100% per band), *M. catarrhalis* rose from 64% to 87% to 100%, and MP remained near-absent in every band (0/13, 1/139, 0/18 at read level). Cohorts differed comparably: the same screen in healthy children's nasopharyngeal swabs gave *S. pneumoniae*–mitis/*H. influenzae* rates of 85.2%/100% (Thai

infants), 96.0%/62.0% and 60.7%/62.0% (two South African birth cohorts), and 95.2%/82.4% (US infants), a ~1.5-fold spread among equally healthy children. These rates are contingent on era, cohort, specimen type, and age — they are not fixed biological constants.

Across 18 organisms screened in the three groups (171 healthy, 24 non-MPP, 51 MPP), four colonization regimes emerged as a descriptive framework — not fixed categories, given compositional prevalences and age/batch differences. Universal colonizers were near-saturated in all groups (*S. pneumoniae*–mitis complex 100%/95.8%/99.4% in MPP/non-MPP/healthy samples in the fixed-depth panel; 100% in all groups in the abundance layer, with oral streptococci). Health-associated organisms declined in disease: *H. influenzae* fell from 93.0% (healthy) through 79.2% (non-MPP) to 58.8% (MPP) in the fixed-depth panel, only the MPP-versus-healthy difference surviving multiple-testing control ($p=4.4\times10^{-8}$; OR 0.11) — although *H. influenzae* was detected in 100% of samples in every group in the abundance layer (Figure 1a), so this regime assignment is assay-layer-dependent; *M. catarrhalis* declined in both layers (86.5% to 12.5% to 3.9% in the abundance layer); *N. meningitidis* was likewise depleted (99.4% healthy versus 84.3% MPP, abundance layer; BH $q=8.2\times10^{-5}$). Disease-enriched opportunists — *K. pneumoniae*, *P. aeruginosa* and *A. baumannii* — concentrated in the pneumonia groups. MP alone stood outside the landscape: essentially absent from the healthy oropharynx (0.6%), a detection essentially confined to disease, with panel prevalence 49.0%/8.3%/0.6% (MPP/non-MPP/healthy), while HBoV was undetected everywhere (0%) — below the detection limit of the 25,000-read panel. Across the 18 organisms (assay layer per organism in Table S10), MP alone combined high disease prevalence with near-absence in health (among viruses, adenovirus showed the same architecture in its high-abundance subset); the *H. influenzae* gradient may partly reflect read competition and antibiotic exposure (see Discussion).

MP detection in pneumonia breaks the colonization pattern

We next quantified the landscape's outlier with the deeper, full-prefix primary screen. Under the primary rule (≥2 MP-specific reads; MAPQ ≥ 30; rRNA operon excluded; *M. genitalium* dual-mapping filter), MP was detected in 29/51 (56.9%; 95% CI 43.3–69.5%) MPP samples, 4/24 (16.7%; 95% CI 6.7–35.9%) non-MPP samples, and 1/170 (0.6%; 95% CI 0.1–3.3%) healthy samples (Table 1, Figure 2a). MPP differed strongly from healthy children (Fisher exact $p=2.6\times10^{-21}$; OR 214 by exact conditional estimation, 95% CI 32–9,147; Haldane–Anscombe-corrected OR 148, 95% CI 27–810) and from non-MPP pneumonia ($p=1.2\times10^{-3}$; OR 6.4, exact 95% CI 1.8–30); non-MPP exceeded healthy ($p=8.5\times10^{-4}$; OR 33, exact 95% CI 3.0–1,674). All three contrasts survived Benjamini–Hochberg adjustment, and after per-sample age recovery (241/246) the MPP-versus-healthy association remained significant within every age stratum (<3, 3–6 and 6–13 years; Figure 2b, d; Table S8). The detection gradient thus stands in direct opposition to the colonization regimes mapped above: MP appears essentially only with disease.

Table 1. Oropharyngeal *M. pneumoniae* detection by group, with detection-threshold sensitivity analysis

A. Primary analysis (≥2 MP-specific reads)			
Group	n	MP+	Prevalence (95% CI)[a]
Healthy children	170	1	0.6% (0.1–3.3%)
Non-MPP pneumonia	24	4	16.7% (6.7–35.9%)
MPP	51	29	56.9% (43.3–69.5%)
B. Pairwise contrasts (primary threshold)			
Contrast	OR (exact 95% CI)[b]	Haldane–Anscombe OR (95% CI)[b]	Fisher p
MPP vs healthy	214 (32–9,147)	148 (27–810)	2.6×10^{-21}
MPP vs non-MPP	6.4 (1.8–30)	6.0 (1.9–19.0)	1.2×10^{-3}
Non-MPP vs healthy	33 (3.0–1,674)	24.8 (3.7–166)	8.5×10^{-4}
C. Sensitivity to positivity threshold (Fisher exact p)			
Threshold	MPP vs healthy	MPP vs non-MPP	Non-MPP vs healthy
≥1 read	3.5×10^{-23}	0.142 (ns)	1.2×10^{-9}
≥2 reads (primary)	2.6×10^{-21}	1.2×10^{-3}	8.5×10^{-4}
≥3 reads	1.6×10^{-19}	2.6×10^{-4}	0.040
≥5 reads	8.8×10^{-18}	9.2×10^{-6}	1.0 (ns)
≥10 reads	1.1×10^{-13}	1.4×10^{-4}	1.0 (ns)

[a] Wilson 95% confidence interval. [b] Exact conditional confidence interval; Haldane–Anscombe-corrected odds ratios are reported for contrasts containing cells with <5 events (see Methods). ns, not significant.

The non-MPP excess was threshold-conditional. All four non-MPP positives were trace-level (2–4 reads). MPP-versus-healthy remained significant at every threshold from ≥1 to ≥10 reads ($p=3.5\times10^{-23}$ to 1.1×10^{-13}), and MPP-versus-non-MPP at every threshold from ≥2 reads upward ($p=1.2\times10^{-3}$ at ≥2 reads to 9.2×10^{-6} at ≥5 reads); at ≥1 read the two pneumonia groups did not yet differ (31/51 vs 10/24, p=0.142), delineating ≥2 reads as the boundary between trace-level noise and the disease signal, whereas the non-MPP-versus-healthy contrast fell from $p=1.2\times10^{-9}$ (≥1 read; 10/24 positive) through $p=8.5\times10^{-4}$ (≥2 reads, primary) to p=1.0 (≥5 reads; 0/24 positive) (Table 1). The non-MPP excess therefore reflects low-level detection near the assay limit. Three of the four trace-positive non-MPP samples were from the unknown-etiology subgroup (3/12 unknown vs 1/12 known-pathogen), and two carried CARDS-locus (MPN372) reads — consistent with, but not proving, sub-diagnostic MP involvement given the age confounding discussed below. The single healthy carrier (SRR8369588) had 10 reads scattered across 10 distinct genomic positions, consistent with genuine low-load carriage in a single child (0.6%) — bounding, not excluding, a healthy-carrier state.

Raw MP read fractions correlated strongly with ResMicroDb species-level MP abundance (Spearman r=0.78 across all 246 samples, $p=5.1\times10^{-52}$; r=0.89 among the 72 abundance-layer positives), and 19 of the 34 read-level positives fell below the pipeline's 0.1% relative-abundance threshold and were

recovered only by the read-level screen, empirically validating the prefix subsampling. A per-sample reconciliation of all 34 MP-positive samples across the two pipelines is provided in Table S9.

Two properties make the gradient clinically interpretable. First, read as a diagnostic against the healthy baseline, the OP screen achieved 56.9% sensitivity (95% CI 43.3–69.5%) and 99.4% specificity (95% CI 96.7–99.9%): across plausible MP-pneumonia prevalences the positive predictive value is 83.6% at 5%, 91.5% at 10%, 96.0% at 20%, 97.6% at 30% and 99.0% at 50%, whereas the negative predictive value reaches only 84.3% at 30% prevalence (Figure 2c, Table S17) — detection rules in; non-detection cannot rule out. Second, within the MPP group the OP MP load tracked illness severity: $\log_{10}$ MP read count correlated with fever duration (Spearman $\rho=0.42$, $p=2.9\times10^{-3}$) and C-reactive protein ($\rho=0.40$, $p=4.4\times10^{-3}$) (Figure 2f), and load-stratified fever duration rose across increasing load tiers (Kruskal–Wallis $p=0.037$), whereas length of stay and peak temperature showed no significant association; children with MP×adenovirus co-detection trended toward longer fever (median 10 vs 5 days; Mann–Whitney $p=0.21$), a hypothesis-generating signal limited by power (Table S17).

The external screen confirmed that the broken colonization pattern is not cohort-specific. After library-strategy verification excluded all amplicon runs and a cohort-level positive-control QC rule excluded three low-biomass cohorts (Table 2), MP was detected in 6.8% (7/103, 95% CI 3.3–13.4%) of a Chinese pediatric pneumonia BALF cohort (PRJNA876769) and 2.9% (1/34) of the WGS subset of a US pediatric CAP cohort (PRJNA745160), and 0% (0/100) in a Gambian pediatric respiratory cohort. The QC-passed community/carriage cohorts were uniformly MP-negative: 0/802 (Wilson 95% CI 0–0.5%) across cohorts from Thailand, South Africa and the USA — statistically indistinguishable from the Shenzhen healthy baseline (Fisher $p=0.18$). Across twelve cohorts, MP was thus confined to pneumonia respiratory samples, with no detectable healthy-carriage reservoir at shotgun depth. The finding extended across the sampled upper-airway compartments: read-level screening found zero MP in 378 QC-passed healthy nasopharyngeal samples (Wilson 95% CI upper bound 1.0%) and none in the healthy nasal/oral samples of the US infant cohort, so no sampled upper-airway compartment showed detectable carriage. Notably, a Chinese pediatric ARI oropharyngeal cohort sampled during COVID-era MP suppression (2021–22; SRP503326; median age 5.0 years, comparable to the Shenzhen healthy group) was uniformly MP-negative (0/168) at read level — consistent with a pathogen lacking a persistent carriage reservoir under interrupted transmission — while adenovirus remained detectable in 6.0% (10/168, predominantly species B3), the expected contrast between a reservoir-free bacterium and continuously circulating viruses.

P1 and 23S genotyping of high-load samples, with validation against published genotyping cohorts

Three high-load MPP samples were provisionally typed (Figure 2e, Table S2). SRR8369497 (full mate-1; 237,965 MP reads, 12.1% of the sample) gave a P1 read vote of 412:2 for P1-1 (99.90% vs 85.53% mean identity to M129 versus FH) and 492 CARDS-locus reads; SRR8369395 and SRR8369523 (8:0 and 9:0 votes from 15–19 P1-region reads each) were likewise P1-1 and CARDS-positive. At the 23S macrolide-resistance hotspot (call rules in Methods), SRR8369497 carried a homogeneous A2063G mutation (16/16 G at position 2063; reproduced 6/6 in a 60-MB prefix re-check) with A2064, A2067 and C2617 wild type. The ten other MP-positive Shenzhen samples had insufficient site depth (≤2 site-covering reads) and are reported explicitly as "not covered" rather than silently dropped (Table S3). The two deeply covered Chinese BALF positives from the external screen were likewise P1-1/CARDS-positive: SRR21425645 carried a predominant A2063G mutation (20 G : 1 C among site-covering reads), SRR21425688 had insufficient 23S depth and is reported as "not covered", and a third positive (SRR21425694) was macrolide-susceptible at the 23S locus but not P1-typable (only 3 discriminating P1 reads, 2:1).

Re-analyzing all 132 runs of a published Australian targeted-NGS cohort from the 2023–2024 epidemic (21), our read-vote calls reproduced the published P1-1/P1-2 distribution exactly (69%/31%) and the macrolide-resistance rate within sampling error (10.6% [12/113 callable] vs 13% published), with resistance observed mainly among P1-1 lineages (10/78 vs 2/35 P1-2; Fisher p=0.34, not statistically distinguishable). Applied to 38 UK 2023–2024 isolates (22), the same calls gave P1-1:P1-2 = 19:19 and 1/26 macrolide-resistant: both post-2023 epidemics were P1-2-enriched and predominantly macrolide-susceptible, whereas our 2016 Shenzhen pneumonia samples were exclusively P1-1 with A2063G predominating. In the Beijing 2023 outbreak cohort (PRJNA1149907) (23), the same pipeline called A2063G in 155 of 156 callable runs (99.4%; aggregate 1,114 of 1,121 site-covering reads carried G, zero susceptible calls) and resolved P1-1 in 126 runs versus P1-2 in 23 at ≥80% vote purity (149 cleanly typed; 20 further runs below the purity threshold counted as ambiguous) — confirming that the Chinese post-2023 resurgence, unlike the European and Australian epidemics, remained P1-1-dominant and near-universally macrolide-resistant (Figure 2e, Table S17).

In 32 severe pediatric pneumonia cases from an independent Shenzhen Children's Hospital cohort (PRJNA572371) (20), adenovirus was detected in 25/32 (78.1%, consistent with the original publication) and MP in 5/32, with 3/32 high-confidence MP×AdV co-detections. The two highest-load MP samples carried the same P1-1/CARDS-positive signature, and their 23S reads trended toward A2063G without reaching a definitive call (insufficient depth), consistent with P1-1/CARDS-positive MP co-circulating with adenovirus in severe pediatric pneumonia in Shenzhen.

Community-level structure: a structured healthy state that flips in disease

The landscape is not only a list of prevalences but a structured community state (Figure 3a). Healthy oropharyngeal communities were diversity-rich (species-level Shannon 3.59 ± 0.33) and predominantly *Prevotella*-led (top-2 genus sequence: *Prevotella*-dominant in 74.9%, 128/171). Both properties flipped in pneumonia: Shannon fell to 3.00 ± 0.68 in MPP and 2.60 ± 0.91 in non-MPP (Kruskal–Wallis $p=5.1\times10^{-13}$), genus-level community structure separated by group (Bray–Curtis PERMANOVA F=34.9, p=0.001; healthy versus MPP F=57.6, p=0.001), and *Prevotella*-dominant states collapsed to 4.1% (2/49) in MPP and 9.5% (2/21) in non-MPP (OR 70, Fisher $p=2.7\times10^{-20}$), replaced by *Streptococcus*-led (43–47%) and heterogeneous states. Critically, within the MPP group the 28 read-level MP-positive and 21 MP-negative samples showed no detectable Shannon-diversity difference (2.98 vs 3.03; Mann–Whitney p=0.55, powered only for large effects) — consistent with the community perturbation tracking pneumonia itself rather than MP detection (a null result, not proof that no MP-specific effect exists).

DNA viruses in pneumonia: adenovirus enrichment against a near-empty healthy baseline

The colonization-calibration logic extends to the DNA virome, where detection in pneumonia must likewise be read against the healthy baseline and stratified by load. Adenovirus DNA was detected in 66.7% of MPP versus 1.2% of healthy OP samples (34/51 vs 2/171; Fisher $p=5.5\times10^{-25}$; OR 169), with non-MPP intermediate (9/24, 37.5%). The enrichment persisted at relative abundance $>10^{-5}$ (58.8% vs 1.2%; OR 121) and $>10^{-4}$ (14/51, 27.5% vs 0/171; $p=2.4\times10^{-10}$) (Figure 3d) and survived Benjamini–Hochberg correction ($q=2.2\times10^{-24}$). WU-polyomavirus and torque-teno-virus enrichments were trace-driven (WU-polyomavirus 14/51 versus 1/171 — identical in the read-level screen and the abundance layer (Table S7)); herpesviruses showed no enrichment and served as a constant-carriage internal reference. Adenovirus read fractions were right-skewed (median 0.003%): most positives were trace-level, while a four-patient high-abundance cluster (1.8–8.6% of reads) indicated markedly higher fractional abundance, consistent with — but not proving — higher-level replication (Figure 3e). Only one OP-AdV-positive patient (P59, outside the high-abundance cluster) had adenovirus reported by clinical BALF testing, so most OP adenovirus detection reflects low-level oropharyngeal shedding rather than proven lower-airway co-infection. WU polyomavirus showed the same architecture at greater extremes: one unknown-etiology non-MPP sample (SRR8369652) reached 76.9% of screened reads — consistent with high-level respiratory replication or prolonged shedding — while the remaining positives were trace-level (Figure 3e). Detection structure across virus families was not independent (Table S14): WU-polyomavirus and torque-teno-virus detections co-occurred (OR 10.7, $p=9.1\times10^{-6}$), adenovirus co-occurred with TTV (OR 4.1, $p=2.9\times10^{-3}$) and trended toward mutual exclusion with WUPyV (OR 0.17, p=0.056), whereas herpesvirus detections were independent of every other family (all p>0.7) — again the

internal-control pattern. Adenovirus and TTV detection tracked read-level MP positivity (OR 21.8, $p=1.5\times10^{-13}$ and OR 6.4, $p=2.2\times10^{-4}$, respectively); herpesvirus detection did not (p=0.83). Hexon typing assigned 15 samples at ≥5 votes (B3 n=8, E4 n=5, C2 n=1, B7 n=1; Figure 3f); the two healthy trace-level positives could not be typed (1–2 votes each; below the ≥5-vote threshold), although both mapped closest to species-C references — the species group for which lymphocyte persistence has been documented (25–26) — whereas the typed disease-group detections were predominantly species B3/E4 (with one species-C2), more parsimoniously read as acute-infection shedding. A single high-load B7 (hexon identity 99.9%) predated the 2018–2019 southern-China B7-driven outbreak by two years — a single-sample, single-locus observation that requires prospective validation.

The colonizer resistome fingerprint

Finally, the colonizer reservoir mapped in the healthy landscape carries its own disease-group fingerprint. Against the 8-target colonizer AMR panel (Table 3), ermB was detected in 35/51 (68.6%) MPP versus 57/171 (33.3%) healthy samples (OR 4.38, Fisher $p=9.7\times10^{-6}$) and mef(E) in 35/51 (68.6%) versus 82/171 (48.0%) (OR 2.37, $p=1.1\times10^{-2}$); non-MPP was intermediate. Adjusting for Streptococcus relative abundance, the ermB association persisted (adjusted OR 2.49, p=0.012) whereas the mef(E) association did not (adjusted p=0.79), indicating that ermB enrichment was independent of streptococcal abundance, whereas mef(E) served as a composition-driven internal comparator. tetM was near-saturated (82.4% MPP, 77.8% healthy; p=0.56), while tetO, bro-1 and bro-2 were undetected (0/246 each) and blaTEM-1 and blaROB-1 were rare and group-independent (blaTEM-1 6.4% healthy vs 2.0% MPP, p=0.30). This colonizer reservoir links to the global AMR threat at a second level: the WHO GLASS high-burden, high-resistance pathogens are themselves oropharyngeal residents — *Klebsiella pneumoniae* (47% healthy, 90% MPP detection), *Staphylococcus aureus* (98–100%) and *Escherichia coli* (92–95% across groups in the abundance layer, Table S10) all colonize this niche, so the oropharynx could act as a standing reservoir from which resistance-prone nosocomial pathogens may descend to cause disease — a hypothesis-generating link, not a demonstrated pathway.

Table 3. Oropharyngeal colonizer AMR gene detection by disease group (≥3 gene reads)

Gene	Resistance class	MPP (n=51)	Non-MPP (n=24)	Healthy (n=171)	OR (MPP vs healthy)	Fisher p
ermB	Macrolide	35/51 (68.6%)	14/24 (58.3%)	57/171 (33.3%)	4.38	9.7×10^{-6}
mef(E)[a]	Macrolide	35/51 (68.6%)	17/24 (70.8%)	82/171 (48.0%)	2.37	1.1×10^{-2}
tetM	Tetracycline	42/51 (82.4%)	17/24 (70.8%)	133/171 (77.8%)	1.33	0.56
tetO	Tetracycline	0/51 (0%)	0/24 (0%)	0/171 (0%)	—	—
blaTEM-1	β-lactam	1/51 (2.0%)	4/24 (16.7%)	11/171 (6.4%)	0.29	0.30
blaROB-1	β-lactam	0/51 (0%)	0/24 (0%)	1/171 (0.6%)	—	—
bro-1 / bro-2	β-lactam	0/51 (0%)	0/24 (0%)	0/171 (0%)	—	—

[a] The mefA and mefE reference accessions initially used were nearly identical owing to a database annotation error (AY071836 is a mef(E)-like sequence mislabeled as mefA); re-analysis with canonical references (NG_047957.1 versus U83667, ~90% pairwise identity) assigns all detections to mef(E) and detects no mef(A).

Discussion

Principal findings. Starting from the healthy oropharyngeal colonization landscape, this read-level re-screening of 246 public pediatric OP metagenomes first established who colonizes the healthy child airway (near-universal streptococcal carriage, 93.0% *H. influenzae*) and then identified the detections that break it. MP broke it: detection rose steeply from healthy children (0.6%) through non-MPP pneumonia (16.7%) to MPP (56.9%); the primary contrasts were robust across all detection thresholds and, after age recovery, within every age stratum. Two caveats temper the single-cohort gradient: the OR of 214 rests on a single healthy event (exact 95% CI 32–9,147; Haldane-corrected 148), and cases and controls were seasonally mismatched (July–August versus predominantly November 2016) — a bias whose direction is conservative for the primary contrast, because November precedes the winter MP season in southern China and healthy sampling then would, if anything, overestimate background carriage relative to mid-summer. The external calibration pool removes the seasonal mismatch but carries its own constraints — predominantly infant nasopharyngeal cohorts sampled in unknown phases of the MP epidemic cycle — so we read internal and external evidence as complementary bounds rather than either one as definitive. With that framing, zero MP detection across 802 QC-passed community/carriage samples indicates no detectable healthy carriage at shotgun depth, so OP MP detection in a pediatric respiratory metagenome is a near-specific disease signal at that depth rather than detectable background colonization. Among DNA viruses, adenovirus enrichment in pneumonia was interpretable only once stratified by load and read against the near-empty healthy baseline.

A colonization-first framework with four regimes. Our healthy estimate (0.6%, Wilson CI 0.1–3.3%) agrees with PCR-based estimates of ~1–3% in asymptomatic children in non-epidemic settings (8); the processed abundance table of the same cohort, applying its 0.1% reporting threshold, reports species-level MP in 6.1% of samples (16). It is far below the 21.2% reported in asymptomatic Dutch children during an epidemic (7), consistent with that estimate reflecting epidemic-era exposure and post-infectious persistence rather than a stable carrier state; genus-level *Mycoplasma* signals in 16S-based healthy-cohort surveys cannot be resolved to MP at species level and provide no evidence for species-level carriage. The twelve-cohort external screen both replicates and bounds this result: MP was confined to pneumonia respiratory samples, while 802 QC-passed carriage samples across three continents were uniformly negative (95% CI 0–0.5%). Read colonization-first, the 18-organism landscape resolves into four

interpretive regimes — universal colonizers near-saturated in health and disease (*S. pneumoniae*–mitis complex, oral streptococci), health-associated colonizers whose detection declines in disease (*H. influenzae*, *M. catarrhalis*, *N. meningitidis*), disease-enriched opportunists (*K. pneumoniae*, *P. aeruginosa*, *A. baumannii*), and pathogen signals with no detectable healthy carriage at sequencing depth (MP). Which regime an organism occupies — calibrated against the healthy baseline — determines how a metagenomic detection should be read. The community-state layer sharpens this framework: the healthy baseline is itself a structured state — diversity-rich and *Prevotella*-led — that flips in both pneumonia groups (Figure 3a–c), echoing the anaerobe-depletion pattern reported in pediatric respiratory microbiome studies. A useful internal check comes from within the MPP group: read-level MP-positive and MP-negative samples showed no detectable community difference (Shannon p=0.55; powered only for large effects), so MP detection reads as a pathogen signal superimposed on an already-flipped community, not as a marker of community disturbance. All negative findings remain detection-limit-bounded: the ≥2-read rule detects MP only above a relative abundance of ≈4.74/N (≈9×10^{-5} at the median prefix depth; Poisson λ_{95}=4.74, spike-in-verified), so high-load, shotgun-detectable MP carriage in healthy children is globally rare, whereas PCR-detectable trace shedding below this limit is not excluded.

The unmapped compartment. One boundary of this calibration deserves emphasis as a finding in itself: healthy lower-airway metagenomes are essentially absent from public archives. Across all repositories screened, healthy BALF controls were found only in single digits and induced-sputum or healthy-smoker controls only in the low hundreds — all MP-negative — against thousands of disease samples. Because MP disease is a lower-airway disease, the healthy-carriage debate has been fought almost entirely on upper-airway specimens while the relevant compartment remains unmapped for ethical and logistical reasons: healthy pediatric BALF is virtually unobtainable. Our zero-carriage claim therefore extends only to the upper airway, and the few available controls are consistent with — but underpowered for — the lower airway. We flag this as the field's outstanding data need: ethically collected healthy lower-airway controls (e.g., bronchoscopy for non-respiratory indications, adult healthy volunteers) should be prioritized and shared.

Corollary 1: OP MP detection rules in, but absence does not rule out. Because healthy carriage is undetectable at shotgun depth, MP detection in a sick child's OP sample is a strong rule-in signal; the converse does not hold, because OP sampling, prefix depth and the ≥2-read limit mean that non-detection cannot exclude lower-airway infection. The non-MPP trace excess illustrates this boundary: three of four trace-positive non-MPP samples belonged to the unknown-etiology subgroup, and two carried CARDS-locus reads, compatible with sub-diagnostic MP involvement below the BALF testing limit; however, the group was small (n=24) and infant-skewed (median 1.67 vs 3.8 years in MPP), and the result is threshold-

conditional, so we present it as hypothesis-generating. The same calibration logic governs co-detected organisms: the graded decline of *H. influenzae* detection (93.0% to 79.2% to 58.8%), significant only for MPP versus healthy, is compatible with ecological displacement during MP infection, but displacement cannot be established with these data — panel prevalences are compositional, antibiotic exposure and age differed between groups, and absolute quantification is unavailable. Near-saturated *S. pneumoniae*–mitis detection in all groups shows the compositional artifact is not universal and partly mitigates, but does not eliminate, this concern.

The oropharynx does not represent the lower airway. Two compartments must not be conflated: we screened oropharyngeal swabs (upper airway, DNA mNGS), whereas the MPP/non-MPP case definitions and the clinical adenovirus confirmations rested on BALF testing (lower airway, targeted clinical assays) in the source study. OP detection systematically underestimates lower-airway infection: even in BALF-defined MPP, only 56.9% of OP samples were MP-positive at read-level depth, and among OP-adenovirus-positive children only one had adenovirus confirmed in the lower airway — different compartment and different method — while the remainder reflects oropharyngeal DNA detection without proven lower-airway involvement. Two consequences follow. Clinically, a negative OP metagenome cannot exclude lower-airway MP or viral infection, so OP screening is a rule-in tool, not a rule-out tool. Analytically, OP prevalence estimates of lower-airway pathogens are lower bounds, and cross-compartment comparisons are valid only when the compartment and method differences are stated.

Corollary 2: DNA-virus detection must be load-stratified against the healthy baseline. Three non-exclusive explanations for the adenovirus enrichment should be considered jointly: true co-infection in the high-load subset ($>10^{-4}$: 27.5% vs 0%), secondary oropharyngeal shedding generating trace detections, and the 2016 community-exposure background. Consistent with this reading, clinical BALF PCR confirmed adenovirus in only one OP-positive patient, and our species-B/E detections are interpreted as acute-infection shedding rather than the species-C lymphocyte persistence documented elsewhere (25–26). One exception tempers any species-label shortcut: an external COVID-era ARI sample (SRP503326) reached a 35.3% adenovirus read fraction and was provisionally assigned to HAdV-C6, showing that species-C viruses can occasionally reach high respiratory burdens. We therefore do not equate species C with persistence; persistence is inferred only for low-abundance species-C signals in the appropriate compartment. Whether pneumonia-associated immune perturbation reactivates persistent adenovirus is a hypothesis-generating question our cross-sectional DNA data cannot resolve; a persistence/reactivation mechanism has supporting evidence only for species C (25–26) — whereas disease-group detections were predominantly species B3/E4, more parsimoniously read as acute-infection shedding. The enrichment appears pediatric-specific — adenovirus detection was only 0–0.8% across 2,211 Chinese adult BALF

metagenomes (23 cohorts; Table S15) — and the WU-polyomavirus and torque-teno-virus enrichments were trace-driven, likely marking immune or compositional correlates of disease rather than etiology, while the flat herpesvirus detection across groups anchors the constant-carriage end of the viral landscape. High-load species-B adenovirus (B3, B7) in these 2016 samples raises a further hypothesis-generating question: might high-load OP adenovirus during pneumonia mark children at risk of post-infectious bronchiolitis obliterans? Testing this requires prospective designs linking acute-phase viral load to predefined small-airway outcomes; this retrospective re-analysis provides no risk estimate and serves only to motivate that test.

Corollary 3: a 2016 clone inventory, and implications for the post-2023 resurgence. All typeable high-load samples were P1-1 and CARDS-positive, consistent with P1-1 dominance in recent Chinese epidemics (8, 13); typing was restricted to three selected high-load samples and is provisional, not a cohort-level distribution. Our exclusively P1-1, A2063G-carrying 2016 Shenzhen samples match the P1-1/macrolide-resistant clones documented across East Asia through the pre-pandemic decades (Beijing 2008–2012 (13); Japan 1995–2005 (27); reviewed in (8)), consistent with P1-1/A2063G clones circulating in East Asia well before the post-2023 resurgence. Benchmarked against the validated post-2023 cohorts, however, the resurgence was dominated by different lineages in different regions: P1-1 dominated the Australian epidemic (69%) yet remained predominantly macrolide-susceptible (11–13% resistant), the UK resurgence was evenly split (P1-1:P1-2 = 19:19, 3.8% resistant), and the 2011 French upsurge was polyclonal, albeit mainly type 1 (28), whereas in the Chinese post-2023 resurgence our own re-analysis of the Beijing 2023 outbreak genomes showed P1-1 dominance (126 of 149 cleanly typed runs, 85%) with near-universal A2063G (155/156 callable, 99.4%) (23). These data indicate regional heterogeneity in the post-2023 resurgence rather than uniform amplification of a single seeded lineage. Sampling-era context likewise argues against a low-exposure artifact for the near-absent healthy carriage: 2016 was a documented national epidemic-peak year for MP (peak positivity 20.1% in the Anhui 2015–2023 surveillance series), and Chinese pediatric MP epidemics peak in September–November (29, 30) — the November-weighted healthy sampling therefore fell in a high-exposure period, so the 0.6% healthy estimate cannot be attributed to low-season sampling. Consistent with a carriage-reservoir-free pathogen, a Chinese pediatric ARI cohort sampled during COVID-era non-pharmaceutical interventions (SRP503326, 2021–22) was uniformly MP-negative at read level (0/168) — the absence of a detectable carrier reservoir is consistent with — though does not by itself establish — transmission having been interruptible to very low levels during that period.

Colonizer resistome. ermB was independently enriched in MPP (OR 4.38) beyond what the streptococcal expansion explains (adjusted OR 2.49, p=0.012), while the mef(E) enrichment was

composition-driven (adjusted p=0.79). These signals are compatible with macrolide selection pressure acting on the oropharyngeal streptococcal colonizer reservoir — although pre-sampling antibiotic exposure in hospitalized children cannot be separated from reservoir-level selection with these data — and are best read as a community-level resistome signal in colonizers — not as MP resistance, which we assessed separately at the MP 23S locus. Because the panel detects gene-level reads without host assignment, these are not species-specific resistance rates. Element-level attribution sharpened this signal (Table S16): after correcting a mislabeled reference, all mef detections resolved to mef(E) (MEGA-family elements, which are widespread in viridans-group streptococci), and ermB-positive samples co-occurred with Tn6002-family element flanks (OR 6.9; adjusted OR 6.0, $p=1.3\times10^{-8}$, after controlling for *Streptococcus* genus abundance) — consistent with, though not proving, an association of the enriched macrolide signal with transferable streptococcal elements (within-sample co-occurrence is not physical linkage). An extended 26-target panel mapped the wider reservoir: anaerobe-anchored genes (tetQ, cfxA2, ermF) tracked the *Prevotella* community flip (Figure 3c), with conditional carriage showing no residual group difference after host-abundance adjustment (a null result; power was limited) — while genes of last-line agents (blaKPC-2 and mcr-1) were undetected in all 246 children (blaNDM-1 was detected in a single MPP sample at 3 reads, unconfirmed). All extended-panel reference accessions were header-verified against the source database after the mef mislabel incident, and a ≥1% runner-up-margin re-screen left group-level calls unchanged (Table S17).

Methodological contributions. Three rules made low-load MP detection reliable: MAPQ ≥ 30 filtering, exclusion of the rRNA operon, and dual mapping against *M. genitalium*, with trace calls verified by inspection. The screen recovered 19 positives missed by a 0.1% relative-abundance pipeline while remaining quantitatively concordant with it (Spearman r=0.78 across all samples; r=0.89 among abundance-layer positives), and it generalized to a multi-pathogen panel with per-target controls. Re-mining public shotgun archives with target-specific, dual-verified read mapping can therefore recover clinically relevant low-load signals missed by general-purpose pipelines, particularly for wall-less organisms poorly served by 16S methods.

Limitations. First, group is confounded with age (the non-MPP group was infant-skewed, with no age-matched infant comparator) and season (cases sampled July–August 2016, controls predominantly November 2016), although the primary MPP-versus-healthy association held within every age stratum; the same seasonal mismatch applies to the DNA-virus enrichments, which are single-cohort observations without external replication and are therefore presented as hypothesis-generating. Second, all negative findings are upper-bound statements at a detection limit of ≈4.74/N relative abundance (Poisson $\lambda_{95}=4.74$; spike-in-verified): PCR-detectable trace shedding below this threshold is not excluded, five low-depth

samples introduce detection-limit heterogeneity, and uncovered resistance loci are reported as "not covered" rather than wild type. A biological asymmetry compounds this limit: MP is extracellular, attached to the mucosal surface via the P1 adhesin, so low-load colonization is sparsely distributed and easily diluted or missed at swab capture, whereas viruses replicate intracellularly and are captured together with the host cells a swab collects. Extracellular organisms therefore sit further below the detection limit than intracellular ones at equal true biomass, so our zero-carriage finding is an upper bound on load, not proof of absence. Finally, healthy lower-airway data are essentially absent from public repositories (single-digit healthy BALF and low-hundreds induced-sputum/healthy-smoker controls, all MP-negative), so our zero-carriage claim is established for the upper airway; healthy lower-airway carriage remains a data gap the field should fill. Third, prefix subsampling is spatially structured and mate-1-only — empirically validated against the full-data pipeline, but not a random subsample — and external prevalences are lower bounds at 3-MB prefix depth. Fourth, genotyping was provisional and restricted to a few high-load samples (n=3 Shenzhen + 3 external; three low-load external positives could not be genotyped), and the AMR panel is a gene-level assay without host assignment. Fifth, scope limits include a single hospital-based cohort from one city sampled in 2016; the external carriage cohorts were predominantly infant nasopharyngeal or nasal samples collected in unknown phases of the MP epidemic cycle, so the pooled 0/802 is an upper bound under those population constraints; OP swabs sample the upper airway and underestimate lower-airway infection, while group definitions rested on BALF testing in a different compartment with a different method; batch effects were not formally tested; and library-strategy verification plus rRNA-operon exclusion are required because read-level screening is invalid for amplicon data.

Conclusions

Mapped colonization-first against the healthy pediatric oropharyngeal landscape, stringent read-level re-screening of public metagenomes established streptococcus-dominated healthy carriage and revealed MP as the organism that breaks it: a steep, statistically robust detection gradient from healthy children (0.6%) to MP pneumonia (56.9%), with trace-level detection in some non-MPP cases concentrated among unknown etiologies. Healthy pediatric OP carriage of MP was near-absent and detection-limit-bounded (1/170 in-cohort; 0/802 across external QC-passed carriage cohorts), making OP MP detection a near-specific, rule-in disease signal at shotgun depth (56.9% sensitivity, 99.4% specificity; positive predictive value 97.6% at 30% prevalence); because the oropharynx underestimates the lower airway, detection rules in but non-detection cannot rule out. High-load MPP samples were provisionally P1-1 and CARDS positive, with A2063G macrolide resistance where genotyping was possible. Adenovirus enrichment in MPP was load-stratified against the healthy baseline — high-load signals distinct from trace-level OP

DNA detection — and a single-sample B7 detection predating the 2018–2019 southern-China B7 outbreak, a hypothesis-level observation resting on single-locus typing. The dual-verification mapping rules provide a transferable framework for detecting low-load, hard-to-amplify organisms in existing shotgun archives.

Data availability

All sequencing data are publicly available under NCBI BioProject PRJNA413615 (16); the independent severe-pneumonia cohort is available under PRJNA572371. Reference genomes: MP M129 (NC_000912.1), *M. genitalium* G37 (NC_000908.2). Supplementary tables include the per-sample primary-screen results (Table S1), the P1 typing of high-load samples (Table S2), the canonical per-sample 23S macrolide-resistance calls (Table S3), the per-run external screen results (Table S4), the colonizer AMR per-sample calls (Table S5), external MP genotyping (Table S6), per-sample DNA-virome results for the full cohort (Table S7; source data for Figure 3), the age-stratified analysis (Table S8), the bidirectional per-sample MP-positive reconciliation across pipelines (Table S9), the Figure 1a source data (Table S10), the colonizer AMR reference accessions (Table S11), the Figure 1c age-stratified source data (Table S12), the adenovirus typing reference panel (Table S13), the virus co-occurrence analysis (Table S14), the contextual datasets cited in the Discussion (Table S15), and the extended colonizer-resistome panel with host anchoring (Table S16), and the post-hoc assay-verification and clinical-association analyses (23S zero-mismatch re-calls, short-read-preset and ≥1%-margin AMR re-screens, extended mollicutes spike-in, diagnostic performance across prevalences, and MP load–severity associations) (Table S17). Analysis code, per-run screening results, panel matrices, and the analysis workbook will be deposited in a public repository with a permanent DOI upon submission of the revised version. All statistics were computed from the deposited per-run result files using scipy and statsmodels as described.

Funding

This work was supported by the Beijing Science and Technology Nova Program Interdisciplinary Project (20230484439) and the Beijing High-Level Public Health Technical Talents Project (2023-02-33). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Acknowledgements

We thank Dai et al. for generating and openly sharing the PRJNA413615 cohort, and the ResMicroDb project for standardized abundance profiles.

AI declaration. AI tools were used for language editing and for assistance with code generation and modification; the authors take full responsibility for the content.

Author contributions (CRediT)

Chen Shen: Conceptualization, Methodology, Formal analysis, Data curation, Writing, Supervision, Project administration, Funding acquisition, Validation; overall responsibility for the work.

Ethics statement

This study is a secondary analysis of de-identified, publicly available sequencing data (BioProject PRJNA413615 and the external cohorts listed in Table 2, including PRJNA572371 and PRJNA1149907). No new human specimens were collected and no identifiable participant information was accessed; institutional review board approval was therefore not required for this re-analysis.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Gupta RS, Sawnani S, Adeolu M, Alnajar S, Oren A. 2018. Phylogenetic framework for the phylum Tenericutes based on genome sequence data: proposal for the creation of a new order Mycoplasmoidales ord. nov., containing two new families Mycoplasmoidaceae fam. nov. and Metamycoplasmataceae fam. nov. Antonie Van Leeuwenhoek 111:1583–1630. https://doi.org/10.1007/s10482-018-1047-3
2. Waites KB, Talkington DF. 2004. *Mycoplasma pneumoniae* and its role as a human pathogen. Clin Microbiol Rev 17:697–728. https://doi.org/10.1128/cmr.17.4.697-728.2004
3. Waites KB, Xiao L, Liu Y, Balish MF, Atkinson TP. 2017. *Mycoplasma pneumoniae* from the respiratory tract and beyond. Clin Microbiol Rev 30:747–809. https://doi.org/10.1128/CMR.00114-16
4. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, Stockmann C, Anderson EJ, Grijalva CG, Self WH, Zhu Y, Patel A, Hymas W, Chappell JD, Kaufman RA, Kan JH, Dansie D, Lenny N, Hillyard DR, Haynes LM, Levine M, Lindstrom S, Winchell JM, Katz JM, Erdman D, Schneider E, Hicks LA, Wunderink RG, Edwards KM, Pavia AT, McCullers JA, Finelli L, CDC EPIC Study Team. 2015. Community-acquired pneumonia requiring hospitalization among U.S. children. N Engl J Med 372:835–845. https://doi.org/10.1056/NEJMoa1405870

5. Meyer Sauteur PM, Beeton ML, ESGMAC Study Group. 2024. *Mycoplasma pneumoniae*: delayed re-emergence after COVID-19 pandemic restrictions. Lancet Microbe 5:e100–e101. https://doi.org/10.1016/S2666-5247(23)00344-0
6. Meyer Sauteur PM, Chalker VJ, Berger C, Nir-Paz R, Beeton ML, ESGMAC Study Group. 2022. *Mycoplasma pneumoniae* beyond the COVID-19 pandemic: where is it? Lancet Microbe 3:e897. https://doi.org/10.1016/S2666-5247(22)00190-2
7. Spuesens EBM, Fraaij PLA, Visser EG, Hoogenboezem T, Hop WCJ, van Adrichem LNA, Weber F, Moll HA, Broekman B, Berger MY, van Rijsoort-Vos T, van Belkum A, Schutten M, Pas SD, Osterhaus ADME, Hartwig NG, Vink C, van Rossum AMC. 2013. Carriage of *Mycoplasma pneumoniae* in the upper respiratory tract of symptomatic and asymptomatic children: an observational study. PLoS Med 10:e1001444. https://doi.org/10.1371/journal.pmed.1001444
8. Meyer Sauteur PM, Unger WWJ, Nadal D, Berger C, Vink C, van Rossum AMC. 2016. Infection with and carriage of *Mycoplasma pneumoniae* in children. Front Microbiol 7:329. https://doi.org/10.3389/fmicb.2016.00329
9. Bogaert D, Keijser B, Huse S, Rossen J, Veenhoven R, van Gils E, Bruin J, Montijn R, Bonten M, Sanders E. 2011. Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. PLoS One 6:e17035. https://doi.org/10.1371/journal.pone.0017035
10. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 24:4420–4449. https://doi.org/10.1093/nar/24.22.4420
11. Hong S, Bunge J, Leslin C, Jeon S, Epstein SS. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J 3:1365–1373. https://doi.org/10.1038/ismej.2009.89
12. Dumke R, Lück PC, Noppen C, Schaefer C, von Baum H, Marre R, Jacobs E. 2006. Culture-independent molecular subtyping of *Mycoplasma pneumoniae* in clinical samples. J Clin Microbiol 44:2567–2570. https://doi.org/10.1128/JCM.00495-06
13. Zhao F, Liu G, Wu J, Cao B, Tao X, He L, Meng F, Zhu L, Lv M, Yin Y, Zhang J. 2013. Surveillance of macrolide-resistant *Mycoplasma pneumoniae* in Beijing, China, from 2008 to 2012. Antimicrob Agents Chemother 57:1521–1523. https://doi.org/10.1128/AAC.02060-12
14. Kannan TR, Baseman JB. 2006. ADP-ribosylating and vacuolating cytotoxin of *Mycoplasma pneumoniae* represents unique virulence determinant among bacterial pathogens. Proc Natl Acad Sci USA 103:6724–6729. https://doi.org/10.1073/pnas.0510644103

15. Becker A, Kannan TR, Taylor AB, Pakhomova ON, Zhang Y, Somarajan SR, Galaleldeen A, Holloway SP, Baseman JB, Hart PJ. 2015. Structure of CARDS toxin, a unique ADP-ribosylating and vacuolating cytotoxin from *Mycoplasma pneumoniae*. Proc Natl Acad Sci USA 112:5165–5170. https://doi.org/10.1073/pnas.1420308112
16. Dai W, Wang H, Zhou Q, Li D, Feng X, Yang Z, Wang W, Qiu C, Lu Z, Xu X, Lyu M, Xie G, Li Y, Bao Y, Liu Y, Shen K, Yao K, Feng X, Yang Y, Zhou K, Li S, Zheng Y. 2019. An integrated respiratory microbial gene catalogue to better understand the microbial aetiology of *Mycoplasma pneumoniae* pneumonia. GigaScience 8:giz093. https://doi.org/10.1093/gigascience/giz093
17. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191
18. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA, Venter JC. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403. https://doi.org/10.1126/science.270.5235.397
19. Xiao L, Ptacek T, Osborne JD, Crabb DM, Simmons WL, Lefkowitz EJ, Waites KB, Atkinson TP, Dybvig K. 2015. Comparative genome analysis of *Mycoplasma pneumoniae*. BMC Genomics 16:610. https://doi.org/10.1186/s12864-015-1801-0
20. Shenzhen Children's Hospital pediatric severe pneumonia metagenomic cohort. 2019. BioProject PRJNA572371. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA572371.
21. Tam KKG, Suster CJE, Fong W, Golubchik T, Sivalingam V, Jeoffreys N, Tay E, Ko D, Wehrhahn MC, Ginn AN, Robson J, Gardner I, Papanicolas LE, Kennedy K, Graham M, Tran T, Speers D, Cooley L, Baird RW, Meumann EM, Harbidge J, Campbell S, Basile K, Chen SCA, Sintchenko V, Kok J, Rockett RJ. 2026. Genomic surveillance reveals emergence and spread of macrolide-resistant *Mycoplasma pneumoniae* in Australia during the 2023–2024 epidemic. J Infect Dis 234:e121–e130. https://doi.org/10.1093/infdis/jiag163
22. Tewolde R, D'Aeth JC, Thombre R, Beeton ML, Rowlands RS, Sheppard SK, Hitchings MD, Pascoe B, Chalker VJ, Afshar B. 2025. Comparative genomic analysis of *Mycoplasma pneumoniae* isolated in the United Kingdom, between 2016 and 2024. BMC Genomics 26:893. https://doi.org/10.1186/s12864-025-12101-y
23. *Mycoplasma pneumoniae* genome sequencing, Beijing 2023 pediatric outbreak cohort. 2025. BioProject PRJNA1149907. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1149907.

24. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300.
25. Garnett CT, Erdman D, Xu W, Gooding LR. 2002. Prevalence and quantitation of species C adenovirus DNA in human mucosal lymphocytes. J Virol 76:10608–10616. https://doi.org/10.1128/jvi.76.21.10608-10616.2002
26. Lion T. 2014. Adenovirus infections in immunocompetent and immunocompromised patients. Clin Microbiol Rev 27:441–462. https://doi.org/10.1128/cmr.00116-13
27. Kenri T, Okazaki N, Yamazaki T, Narita M, Izumikawa K, Matsuoka M, Suzuki S, Horino A, Sasaki T. 2008. Genotyping analysis of *Mycoplasma pneumoniae* clinical strains in Japan between 1995 and 2005: type shift phenomenon of *M. pneumoniae* clinical strains. J Med Microbiol 57:469–475. https://doi.org/10.1099/jmm.0.47634-0
28. Pereyre S, Touati A, Petitjean-Lecherbonnier J, Charron A, Vabret A, Bébéar C. 2013. The increased incidence of *Mycoplasma pneumoniae* in France in 2011 was polyclonal, mainly involving *M. pneumoniae* type 1 strains. Clin Microbiol Infect 19:E212–E217. https://doi.org/10.1111/1469-0691.12107
29. Chen B, Gao LY, Chu QJ, Zhou TD, Tong Y, Han N, Wang AH, Zhou Q. 2024. The epidemic characteristics of *Mycoplasma pneumoniae* infection among children in Anhui, China, 2015–2023. Microbiol Spectr 12:e00651-24. https://doi.org/10.1128/spectrum.00651-24
30. Gao LW, Yin J, Hu YH, Liu XY, Feng XL, He JX, Liu J, Guo Y, Xu BP, Shen KL. 2019. The epidemiology of paediatric *Mycoplasma pneumoniae* pneumonia in North China: 2006 to 2016. Epidemiol Infect 147:e192. https://doi.org/10.1017/S0950268819000839

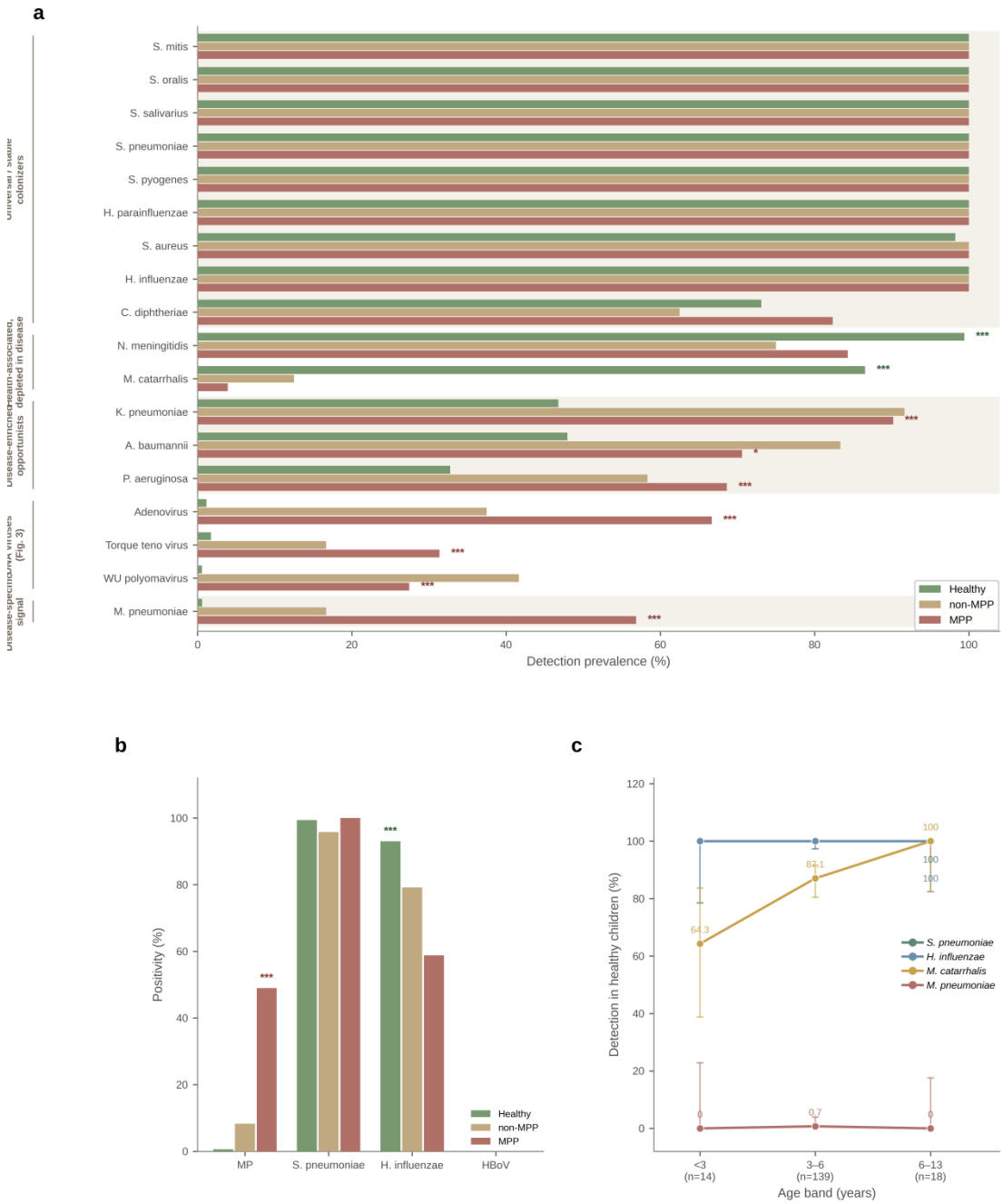


Figure 1. The healthy pediatric oropharyngeal colonization landscape and its disruption in disease.

(a) Detection prevalence of 18 respiratory organisms across healthy children (n=171), non-MPP pneumonia (n=24) and MPP (n=51), integrating the processed species-level abundance tables with read-level screening (per-organism assay layers in Table S10; M. pneumoniae is shown from the validated read-level screen, because the abundance layer assigns trace MP to nearly every sample). Organisms are grouped by interpretive regime (left brackets): universal or stable colonizers with no group difference; health-associated taxa depleted during disease; disease-enriched opportunists; DNA viruses (load-stratified in Figure 3); and M. pneumoniae as a near-specific disease signal. Stars mark MPP-versus-healthy contrasts significant after Benjamini–Hochberg correction across the 18 organisms (Fisher exact; \ q<0.05, \\\ q<0.001; placed on the healthy bar for depleted taxa and on the

MPP bar for enriched taxa). (b) Multi-pathogen panel in fixed-depth (25,000-read) subsamples. Detection prevalences for MP, the S. pneumoniae–mitis complex (upper-bound estimate), H. influenzae, and HBoV by group. MP: 49.0%/8.3%/0.6% (MPP/non-MPP/healthy); S. pneumoniae–mitis: 100%/95.8%/99.4%; H. influenzae: 58.8%/79.2%/93.0% (MPP vs healthy $p=4.4\times10^{-8}$, significant after multiple-testing correction); HBoV: 0% in all groups (below metagenomic detection limit). Panel prevalences are compositional. (c) Age-stratified detection among healthy children (bands <3 years, n=14; 3–6 years, n=139; 6–13 years, n=18) for four landscape-defining taxa, with Wilson 95% CIs (wide in the two small outer bands). Detection of the universal colonizers S. pneumoniae–mitis and H. influenzae was 100% in every band (abundance layer); M. catarrhalis was 64.3%/87.1%/100% across the bands; M. pneumoniae was near-absent in every band at read level (0/13, 1/139, 0/18; the <3-year read-level denominator is 13 because one zero-read sample in that band is excluded from the primary assay).

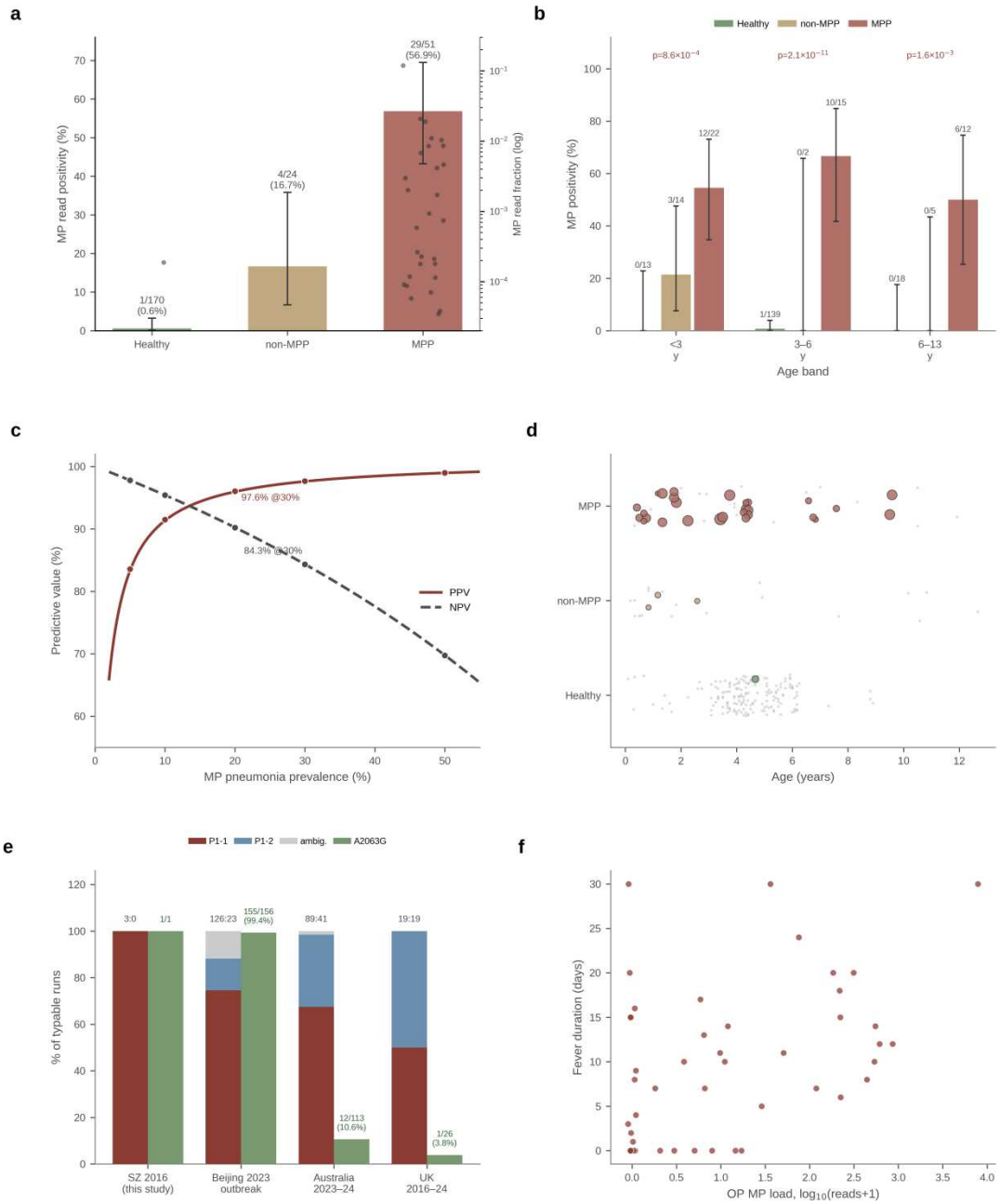


Figure 2. MP detection gradient, diagnostic performance, and strain background.

(a) Oropharyngeal M. pneumoniae detection across disease groups (primary rule: ≥2 MP-specific reads; MAPQ ≥ 30; rRNA operon excluded; M. genitalium dual-mapping filter): healthy 1/170 (0.6%), non-MPP 4/24 (16.7%), MPP 29/51 (56.9%); Wilson 95% CIs; per-sample MP read fractions overlaid on a log-scaled right axis. Pairwise Fisher p: MPP vs healthy 2.6×10^{-21}; MPP vs non-MPP 1.2×10^{-3}; non-MPP vs healthy 8.5×10^{-4} (threshold-conditional; Table 1C). (b) Age-stratified MP positivity with per-band MPP-versus-healthy Fisher p-values (Table S8); non-MPP cells in the 3–6 and 6–13 bands are small (n=2, n=5) and shown for completeness. (c) Diagnostic performance against the healthy baseline: positive (solid) and negative (dashed) predictive values versus MP-pneumonia prevalence (sens 56.9%, spec 99.4%); points mark the tabulated prevalences (Table S17). (d) Age-

continuous view of all metadata-mapped samples (n=241); colored markers denote read-level MP-positive samples, size scaled to $\log_{10}$ MP read fraction; no within-MPP age trend (logistic p=0.40). (e) Strain background across eras and regions at identical read-vote rules: P1 genotype composition (left bar of each pair; P1-1:P1-2 vote counts above) and A2063G macrolide resistance (right bar, green) for Shenzhen 2016 (this study, typable high-load samples), Beijing 2023 (PRJNA1149907), Australia 2023–24 (PRJNA1241475) and UK 2016–24 (PRJNA1231462). (f) Within-MPP oropharyngeal MP load versus fever duration (Spearman $\rho=0.42$, $p=2.9\times10^{-3}$) and C-reactive protein ($\rho=0.40$, $p=4.4\times10^{-3}$); n=49 BALF-confirmed MPP.

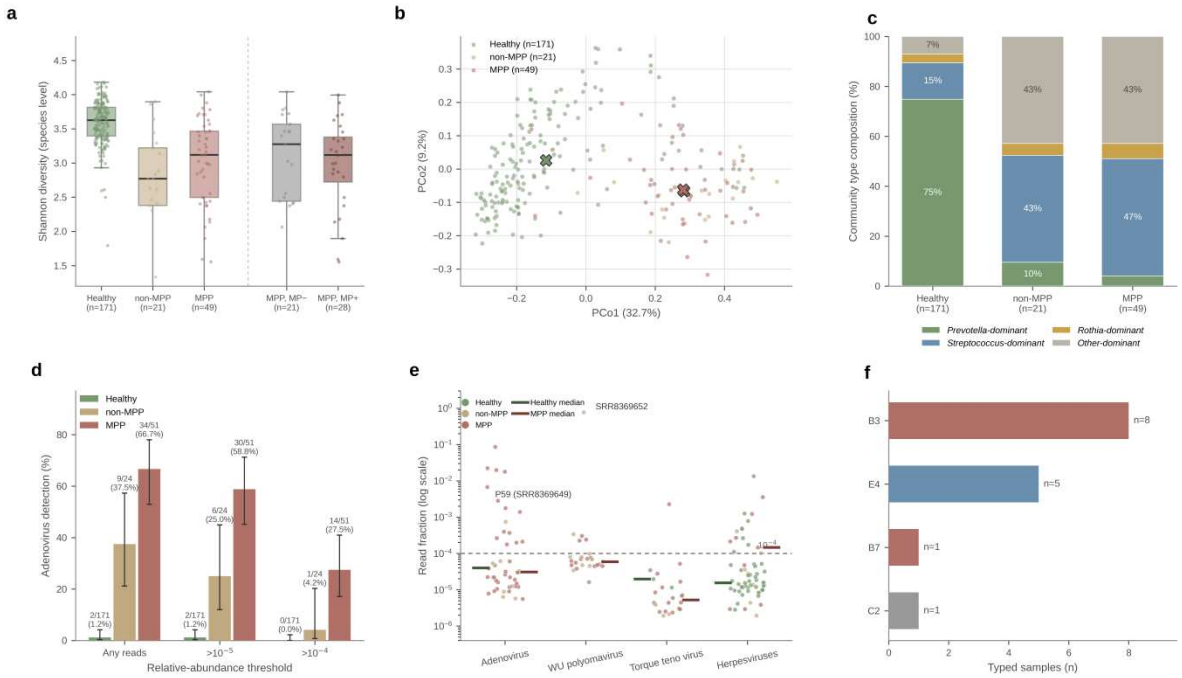


Figure 3. Load-stratified DNA viruses and community structure against the healthy-carriage baseline.

(a) Species-level Shannon diversity by group (healthy n=171, non-MPP n=21, MPP n=49; Kruskal–Wallis $p=5.1\times10^{-13}$) and within the MPP group by read-level MP status (MP-negative n=21 versus MP-positive n=28; Mann–Whitney p=0.55) — community perturbation tracks pneumonia itself, not MP detection. Boxes show median and IQR; points are individual samples. (b) Genus-level Bray–Curtis principal-coordinates analysis (PCo1 32.7%, PCo2 9.2% of variance), samples colored by group with group centroids (crosses); PERMANOVA (999 permutations): three groups F=34.9, p=0.001; healthy versus MPP F=57.6, p=0.001. (c) Community-type composition by group, assigned by the top-2 genus sequence. The Prevotella-dominant state comprised 74.9% of healthy communities versus 9.5% (non-MPP) and 4.1% (MPP) (OR 70, Fisher $p=2.7\times10^{-20}$); disease groups shifted to Streptococcus-led and heterogeneous states. (d) Oropharyngeal adenovirus detection by disease group at three relative-abundance thresholds (any reads; $>10^{-5}$; $>10^{-4}$), with Wilson 95% CIs; n/N and percentages above bars. (e) Per-sample read fractions (log scale) of the four screened DNA-virus groups, colored by disease group; only screen-positive samples are plotted. Short horizontal ticks mark group medians (healthy, left; MPP, right) where ≥2 positives exist. The dashed line marks relative abundance 10^{-4}; herpesvirus detections show no group gradient and serve as the constant-carriage internal reference. Annotated points: the WU-polyomavirus extreme sample SRR8369652; patient P59 (SRR8369649), the only OP-adenovirus-positive child with adenovirus also reported by clinical BALF testing; and the four-patient adenovirus high-abundance cluster (1.8–8.6% of reads). (f) Hexon-based type assignment of the 15 adenovirus-positive samples reaching ≥5 typing votes (B3 n=8, E4 n=5, B7 n=1, C2 n=1; bars colored by species: B, E and C). The single high-load B7 sample (hexon identity 99.9%) predated the 2018–2019 southern-China B7 outbreak by two years. The two healthy trace-level positives (1–2 votes each) were not typeable but mapped closest to species-C references. https://doi.org/10.1017/S0950268819000839it therefore did not require institutional review board approval. Consent to participate / for publication. Not applicable; no new human participants, specimens, or identifiable data were involved.

Supplementary Files

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

- MPcarriagesupplymentary2.zip